JOURNAL OF
CHEMICAL INFORMATION
AND MODELING

PERSPECTIVE

pubs.acs.org/jcim

Lessons Learned from Molecular Scaffold Analysis

Ye Hu, Dagmar Stumpfe, and Jürgen Bajorath*

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

1. INTRODUCTION

The scaffold concept is one of the most frequently applied concepts in medicinal chemistry and virtual screening.[1] The term scaffold is used to describe molecular core structures that are utilized in drug design or detected in virtual screening and, in addition, building blocks for synthetic efforts. For a series of analogs, a scaffold might be derived by determining their maximum common substructure, but there are many other ways to define scaffolds (vide infra). Unfortunately, in chemoinformatics, the scaffold concept is often applied in a rather subjective manner, without adhering to clear, formal, and consistent definitions.[1] For scaffold hopping,[2] i.e., the identification of different scaffolds with similar activity that represents the "holy grail" of virtual screening, the frequent lack of formal consistency presents a substantial problem[1] and makes it often impossible to compare different studies and methods.[3] In fact, the absence of generally accepted evaluation standards for benchmarking and the inconsistency in assessing scaffold hopping analyses currently are major roadblocks for the further development of the virtual screening field.[3]

To further complicate matters, the terms scaffolds, substructures, and fragments are often used to refer to similar or the same structures. Substructures and fragments are rather general designations and are applied to describe small or large structural moieties, scaffolds, parts of scaffolds, or R-groups. Moreover, many different substructures are utilized in drug design applications[4] and different molecular fragmentation schemes have been introduced.[5−8] These fragmentation methods include knowledge-based approaches such as the generation of fragment dictionaries to flag reactive and toxic compounds[5] or predict ADME properties[6] as well as systematic fragmentation schemes that are based on synthetic or retrosynthetic criteria.[7,8] Such fragmentation and fragment organization approaches have also provided a basis for the design of fragment libraries in the context of fragment-based drug discovery.[9−11] Also, even random fragmentation approaches have been introduced to generate structural signatures of compounds with certain biological activities.[12,13]

In addition to knowledge-based and synthetically oriented fragmentation methods, fragments can also be systematically derived on the basis of a defined molecular hierarchy and such approaches have become particularly relevant for scaffold generation and analysis. Regardless of how scaffolds are ultimately rationalized, general aims of scaffold analysis include the assessment of structural diversity of small molecules, the generation of structural classes and structural organization schemes, and the evaluation of biological activities or other molecular properties that are associated with different structural motifs.

In this Perspective, we largely, but not exclusively, focus on studies that have analyzed scaffold distributions in selected compound data sets (such as drugs or screening libraries) or in currently available bioactive compounds. A number of these investigations have explored different types of relationships between scaffolds and the biological activities of compounds they represent.

2. SCAFFOLDS

2.1. Definitions. In 1996, a seminal article by Bemis and Murcko[14] introduced a hierarchical molecular organization scheme by dividing small molecules into R-groups, linkers, and frameworks (Figure 1a). In this context, the terms framework and scaffold can be interchangeably used. Originally, Bemis and Murcko utilized so-defined frameworks as a proxy for molecular shape in their analysis of drug structures. Frameworks containing heteroatom and bond order information in the literature often are referred to as "Bemis and Murcko scaffolds" (BM scaffolds). From these scaffolds, one can further abstract to "cyclic skeletons"[15] (CSKs) and "reduced cyclic skeletons"[15] (RCSs) (Figure 1a). CSKs are derived from BM scaffolds by changing each heteroatom to carbon and all double and triple bonds to single bonds. Thus, CSKs generalize BM scaffolds (CSKs are referred to as "graph frameworks" by Bemis and Murcko[14]) and organize them according to different topologies (i.e., different CSKs represent sets of topologically distinct BM scaffolds). Topological considerations often become an important issue in scaffold analysis and classification. RCSs are obtained from CSKs by ignoring differences in ring size and linker length such that all rings are of unit size and all linkers are of unit length. However, these representations are not often used in scaffold analysis. Synthetically oriented fragmentation schemes are distinct from hierarchical scaffold generation and, for comparison, the RECAP[7] approach to decompose compounds into synthetically accessible moieties is outlined in panel (b) of Figure 1.

The hierarchical framework/scaffold definition following Bemis and Murcko provides a consistent and generally applicable reference frame for scaffold generation. However, it is not without problems. For example, distinct BM scaffolds might actually describe very similar structures that are only distinguished, for example, by a single heteroatom position or bond order. Hence, it is usually important to take CSKs in scaffold analysis into account (vide supra). Furthermore, the addition of any ring to an existing BM scaffold creates per definitionem a new

Received: April 19, 2011
Published: July 14, 2011

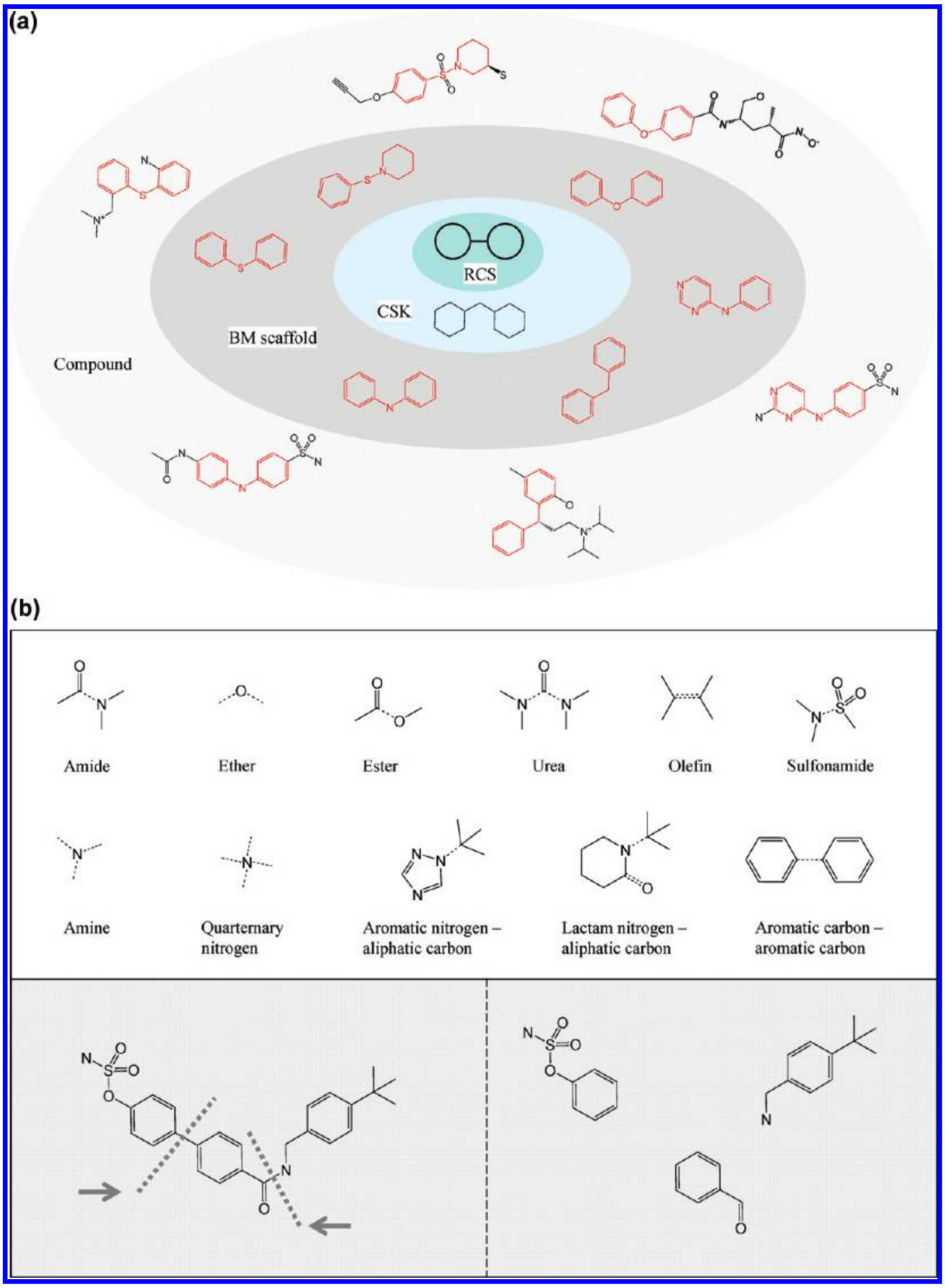
(a)
RCS
CSK
BM scaffold
Compound
(b)
Amide
Ether
Ester
Urea
Olefin
Sulfonamide
Amine
Quarternary nitrogen
Aromatic nitrogen – aliphatic carbon
Lactam nitrogen – aliphatic carbon
Aromatic carbon – aromatic carbon

Figure 1. Scaffolds. In panel (a), the generation of hierarchical scaffolds at different levels of abstraction is illustrated. From the periphery to the center, the degree of structural abstraction increases from compounds to BM scaffolds, CSKs, and reduced cyclic skeletons (RCS). Six exemplary compounds are shown that yield six different BM scaffolds (red) and only a single CSK (and RCS). The figure is adapted from ref 53. In panel (b), the retrosynthetic RECAP fragmentation scheme is outlined for comparison. In the upper panel, dashed lines indicate the 11 cleavable bonds upon which the RECAP scheme is based. The lower panel illustrates a RECAP fragmentation.

scaffold, which often complicates the assessment of alternative scaffolds for combinatorial chemistry applications[16] and renders the boundaries between scaffolds and substituents rather fluid. Despite these limitations, the Bemis and Murcko approach currently is probably most widely applied to characterize molecular scaffolds. It has provided the basis for many of the scaffold mining studies discussed herein and, in addition, for general data structures to organize scaffolds such as the scaffold tree (ST).[17] Here BM-like scaffolds serve as a starting point for further rule-based decomposition of scaffolds along tree branches until only individual rings remain, hence representing a hierarchical organization scheme for scaffold populations (vide infra).

2.2. Sources. In hit-to-lead projects, scaffolds are often compared for only a few different series of compounds, and alternatives are considered on a case-by-case basis. However, scaffold analysis is usually more interesting when applied on a large scale. Pharmaceutical companies have their in-house screening libraries and compound decks, and one can assume that many scaffold studies carried out in these environments are not reported. Of course, there are also other commercially available collections of bioactive compounds (that are not considered here). Importantly, however, public domain repositories of compounds and activity data including BindingDB,[18] the ChEMBL database,[19] and the confirmatory bioassay collection of PubChem[20] have become indispensable resources for both academic and pharmaceutical research, including large-scale compound data and scaffold mining, and the unrestricted communication of scientific results generated by these studies. On the basis of our most recent calculations, more than 76,000 unique BM scaffolds (corresponding to ~29,000 unique CSKs) can be obtained from compounds active against human targets (with defined activity measurements) that are currently available in these three databases. These structures provide a fully accessible pool of scaffold information for method evaluation, data mining, and knowledge extraction.

3. SCAFFOLD DIVERSITY

In their original scaffold study,[14] Bemis and Murcko analyzed a set of ~5000 drugs and found that ~25% of these drugs were represented by the 42 most frequently occurring BM scaffolds and ~50% by the 32 most frequent CSKs, which indicated that the diversity of drug "shapes" was rather limited. Similar analyses focusing on the systematic extraction of scaffolds from drugs or drug-like compounds and frequency of occurrence analysis have remained popular to this date[21−23] and have also been expanded to monitor scaffold distributions in compounds at different stages of pharmaceutical development.[24] For scaffold generation, algorithms have been introduced to isolate heterocyclic sytems[21] or all possible scaffolds[22] from compound data sets or systematically fragment compounds and classify the resulting fragment populations.[13,23] Scaffold diversity has also been assessed for organic compounds[25] and screening libraries.[26,27] In addition, scaffold distributions of screening libraries have been compared to those of drugs and drug-like compounds in order to estimate the relevance of screening collections for pharmaceutical development.[27] Furthermore, scaffolds have also been extracted from natural products and classified.[28] Such natural product-derived scaffolds are of particular interest for diversity-oriented synthesis.[29,30] The generation of screening libraries incorporating compounds derived from naturally occurring scaffolds is thought to increase the likelihood of identifying attractive leads[29,30] because many of these natural scaffolds have evolved to adopt various binding functions.

4. IN SILICO SCAFFOLD ENUMERATION

An extension of scaffold diversity analysis of available bioactive compounds or drugs is the in silico enumeration of possible ring structures and their subsequent mapping to known compounds. For example, Ertl and colleagues[31] created a large virtual library containing ~580,000 heteroaromatic scaffolds consisting of one to three fused rings. A variety of molecular properties were calculated for these enumerated scaffolds. In addition, these properties were calculated for a set of 780 aromatic scaffolds isolated from bioactive compounds in order to distinguish between enumerated and bioactive scaffolds. Scaffolds were then clustered in the resulting property reference space through self-organizing neural networks. It was found that bioactive scaffolds were confined to very small regions in this scaffold property space. Furthermore, Oprea and colleagues developed a graph−theoretic approach to enumerate topologically distinct scaffolds containing up to eight rings and classified different topologies.[32] Subsequently, scaffold topology distributions were analyzed in various compound databases. Theoretically derived topologies of scaffolds containing six or more rings were found to be only rarely represented in known compounds.[33] In a recent study, Pitt et al. generated a set of ~25,000 small heteroaromatic ring systems and determined that less than 2000 of these rings had already been synthesized.[34] It was estimated that ~3000 of the remaining heteroaromatic systems would be synthetically accessible,[34] hence providing opportunities for the evaluation of previously unconsidered heterocycles. Despite differences in scope and methodology, a common theme of these studies has been that bioactive and drug-like compounds only represent a small fraction of principally available "scaffold space", in line with the earlier findings of Bemis and Murcko in their analysis of frequent drug scaffolds.[14] In addition, a very large enumeration effort has been reported by Blum and Reymond[35] who generated more than 970 million small organic molecules in the course of their GDB-13 database project. Structures of organic compounds of up to 13 carbon and heteroatoms were enumerated taking synthetic criteria into account. Thus, the GDB-13 database represents a large virtual source of fragment information.

5. PRIVILEGED SUBSTRUCTURES

The concept of privileged substructures (PSS) was originally introduced in 1988 by Evans et al.,[36] who recognized that the benzodiazepine framework was contained in many ligands of different G-protein coupled receptors (GPCRs) and ion channels. PSS were originally defined as molecular frameworks representing compounds having a high propensity to bind to different proteins.[36] The PSS concept has ever since been of high interest in medicinal chemistry.[37,38] Over the years, PSS have been increasingly understood as scaffolds or parts of scaffolds that would preferentially or exclusively bind to a given target family (i.e., closely related proteins).[37,38] A number of PSS have been proposed, especially for high-profile target families such as GPCRs[39] or protein kinases.[40] Figure 2 shows examples. However, the existence of "truly" privileged substructures has remained controversial because PSS proposed for different target families have also been found to frequently occur in compounds active against other targets.[41] This is understandable if one considers that putative PSS have generally been put forward on the basis of medicinal chemistry knowledge and the analysis of series of active compounds. Frequency of occurrence or other statistical analyses[40,41] were subsequently carried out to investigate the presence of PSS in ligands of other target families. In this context, it is reasonable to consider the relative enrichment of putative PSS in different target families[40] instead of their exclusive occurrence.

The PSS concept has recently been revisited from a different perspective by focusing on systematic data mining[42,43] rather than the assessment of knowledge-based proposals. For all BindingDB and PubChem confirmatory bioassay compounds active against human targets, activity annotations were collected,

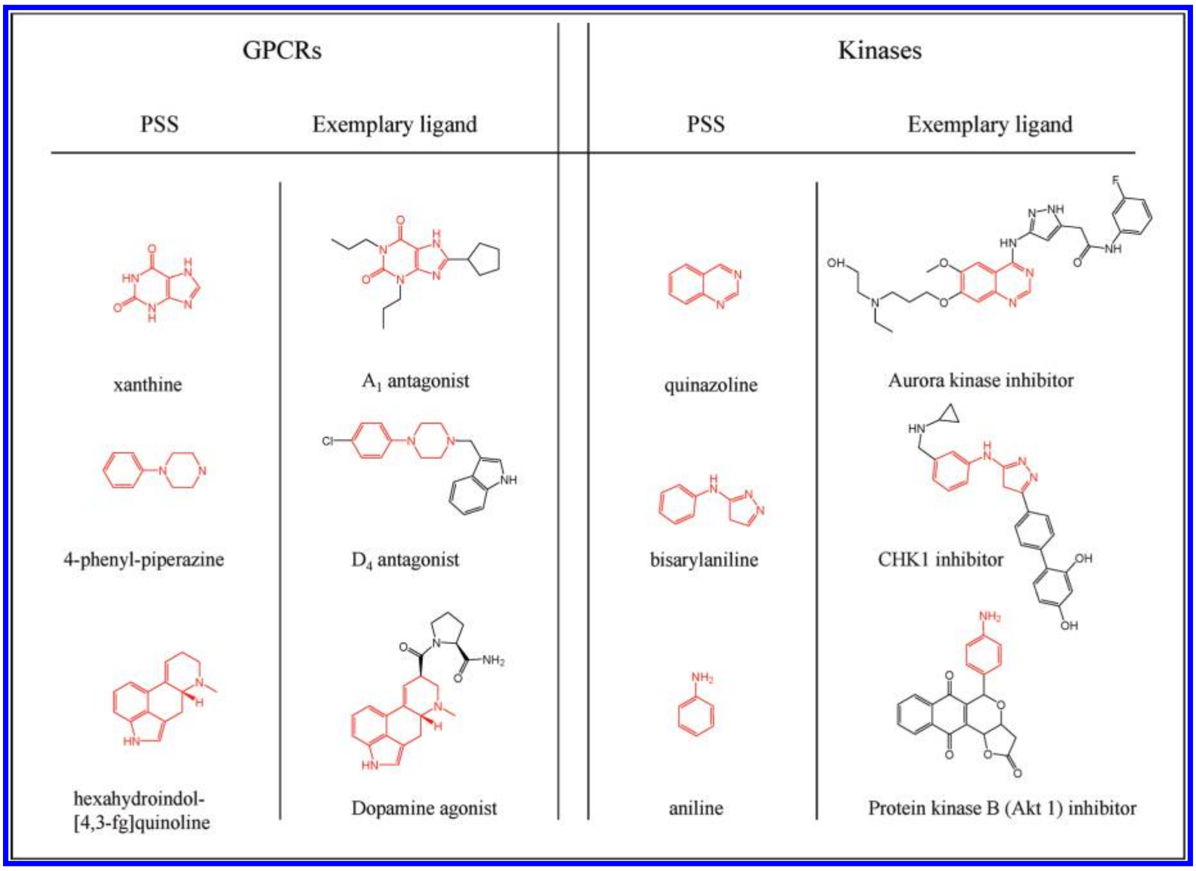

Figure 2. Privileged substructures. Exemplary ligands of GPCRs and protein kinases are shown that contain privileged substructures (PSS, red). Privileged substructures for GPCRs and kinases were taken from ref 39 and refs 68−70, respectively.

and all pairs of targets that shared at least five known active compounds were formed. A compound-based target pair network representation, shown in panel (a) of Figure 3, was generated from BindingDB data that organized 259 human targets forming 520 target pair sets into 18 target communities, which mostly represented known protein families. From PubChem bioassays, only three additional target pair sets were obtained. In this analysis, more than 200 BM scaffolds were identified that represented at least five compounds exclusively active against an individual target community.[42] These community-selective scaffolds yielded ∼150 topologically distinct CSKs. Panel (b) of Figure 3 shows community selectivity profiles of exemplary scaffolds. Only 11 of these community-selective scaffolds were found in approved drugs, suggesting that there should be significant opportunities for further pharmaceutical exploration of community-selective scaffolds. Panel (c) of Figure 3 shows a scaffold-based target network that compares the BindingDB and ChEMBL databases and nicely illustrates the complementarity of their compound information for major target families such as GPCRs and kinases. From ChEMBL, more than 300 BM scaffolds selective for one of 21 target communities were identified, and these community-selective scaffolds corresponded to more than 200 distinct CSKs, only ∼80 of which were also found in BindingDB. Hence, on the basis of systematic compound data mining, original proposals of privileged substructures have been well complemented. Furthermore, target selectivity information was added to the analysis by calculating target pair potency ratios for compounds represented by community-selective scaffolds.[42,43] This enabled the generation of target selectivity profiles and the identification of scaffolds representing compounds selective for one particular target over one or more others (at different potency-derived selectivity levels). However, less than 20% of target-selective scaffolds that were identified when no compound-per-scaffold constraints were applied (different from the analysis of community-selective scaffolds) represented multiple compounds. For example, at the level of 100-fold target selectivity, only ∼60 scaffolds were found to represent more than one compound. Thus, due to the sparseness of selectivity data for bioactive compounds, there currently is no sound basis for the proposal of truly target-selective scaffolds. Similarly, data incompleteness also affects the information content and predictive ability of drug-target networks.[44] In this context, data incompleteness essentially means that drugs or active compounds have not been tested on all potential targets, which likely causes false-negative target annotations. Nevertheless, when organizing scaffolds according to target selectivity, regardless of the number of compounds they represent, interesting target selectivity patterns were observed,[43] as illustrated in panel (d) of Figure 3. Despite data incompleteness, such selectivity patterns can guide attempts to design selective compounds, for example, by generating analogs of scaffolds that already display desired selectivity relationships with two or more targets.

6. POLYPHARMACOLOGY AT THE LEVEL OF SCAFFOLDS

Systematic analyses of ligand-target interactions have provided substantial evidence that many bioactive compounds and drugs elicit their effects by acting on multiple targets.[45−47]

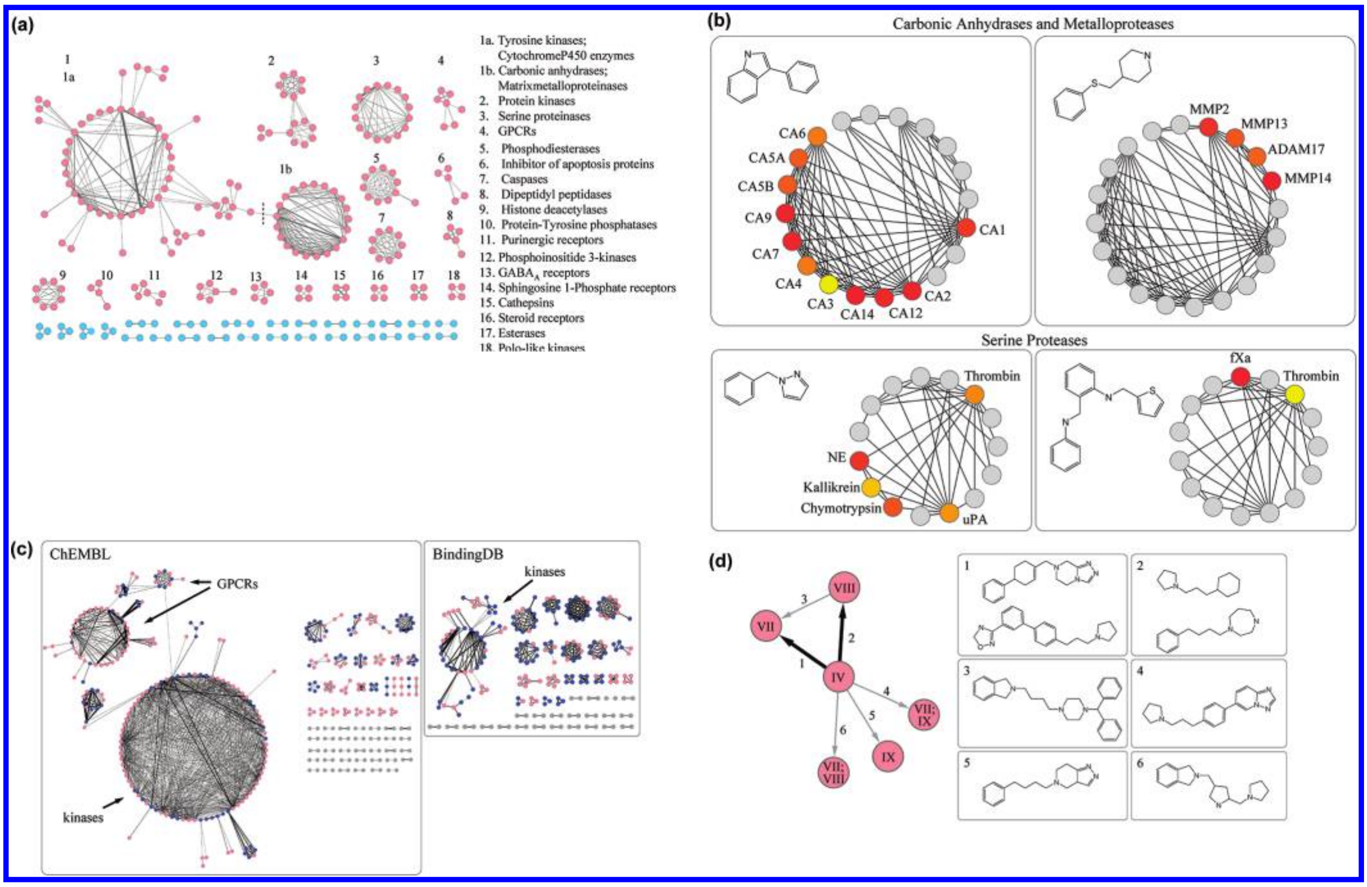

Figure 3. Target networks and community-selective scaffolds. In panel (a), a compound-based target network is shown (for compounds taken from BindingDB). Pink nodes represent targets that are organized in communities (consisting of at least four targets) on the basis of shared ligands, and blue nodes represent the remaining targets. Nodes are connected by edges if both targets share at least five active compounds. Network representations were drawn with Cytoscape.[71] In panel (b), exemplary BM scaffolds and their community-centric target selectivity profiles are shown. Node colors reflect the median target selectivity values of active compounds represented by the given scaffold, ranging from red (highest potential to produce target-selective compounds) to yellow (lowest potential). Gray nodes indicate targets for which less than five active compounds containing the scaffold are available. Panel (c) shows the comparison of scaffold-based target networks for BindingDB and ChEMBL. Different from compound-based target networks, nodes are connected here if the targets share active compounds yielding at least five distinct BM scaffolds. Pink nodes indicate targets in communities that are unique to BindingDB or ChEMBL, blue nodes shared community targets, and gray nodes target pairs or singletons. In this case, a community is defined to consist of at least three targets. Panel (d) shows target selectivity patterns of scaffolds in a directed target network and the corresponding scaffolds. The nodes represent different dipeptidyl peptidases (labeled with Roman numbers). The direction of the arrow indicates a "selective over" relationship for the scaffolds. Black directed edges represent two or more scaffolds and gray edges one (edges and scaffolds are correspondingly numbered). Panels (a) and (b) are adapted from ref 42, panel (c) is adapted from ref 67, and panel (d) from ref 43.

It is therefore not surprising that "polypharmacological" drug behavior has become an emerging and widely investigated paradigm in drug discovery.[48−50] For the study of polypharmacology, ligand-target networks and other types of network representations have been extensively utilized.[48−51] In these investigations, drugs and their target annotations have been systematically related to each other utilizing network representations. On the basis of near neighbor analysis, new targets and therapeutic applications for existing drugs have been proposed, which explains the high level interest in such studies. In addition to drug collections, PubChem bioassay data have also been mined for compounds displaying polypharmacological behavior and for target-selective molecules.[51] Polypharmacology has also been analyzed at the level of scaffolds. In this case, the analysis scheme is essentially orthogonal to data mining for community-selective scaffolds because the potential promiscuity of compounds that a scaffold represents must be evaluated. In a study focused on polypharmacology in a specific therapeutic area, more than 200 drug targets implicated in cardiovascular diseases and their known ligands were assembled and polypharmacological relationships were analyzed.[52] In the course of this investigation, scaffolds were extracted from ligands of cardiovascular targets, and the five most promiscuous scaffolds were determined.[52] Moreover, in a general data mining effort, all BindingDB and ChEMBL compounds active against human targets with at least 1 μM potency were selected, and target sets were assembled consisting of at least 10 compounds. From these target sets, scaffolds were systematically isolated, yielding more than 13,000 BM scaffolds active against ~450 targets belonging to 19 families.[53] From this pool, more than 400 scaffolds were identified that were active against at least two target families, and 83 of these scaffolds were active against targets in three to 13 families. These 83 scaffolds yielded 33 topologically distinct CSKs representing promiscuous chemotypes. Panel (a) of Figure 4 shows examples of promiscuous CSKs, and panel (b) of Figure 4 show the scaffold-target family relationships

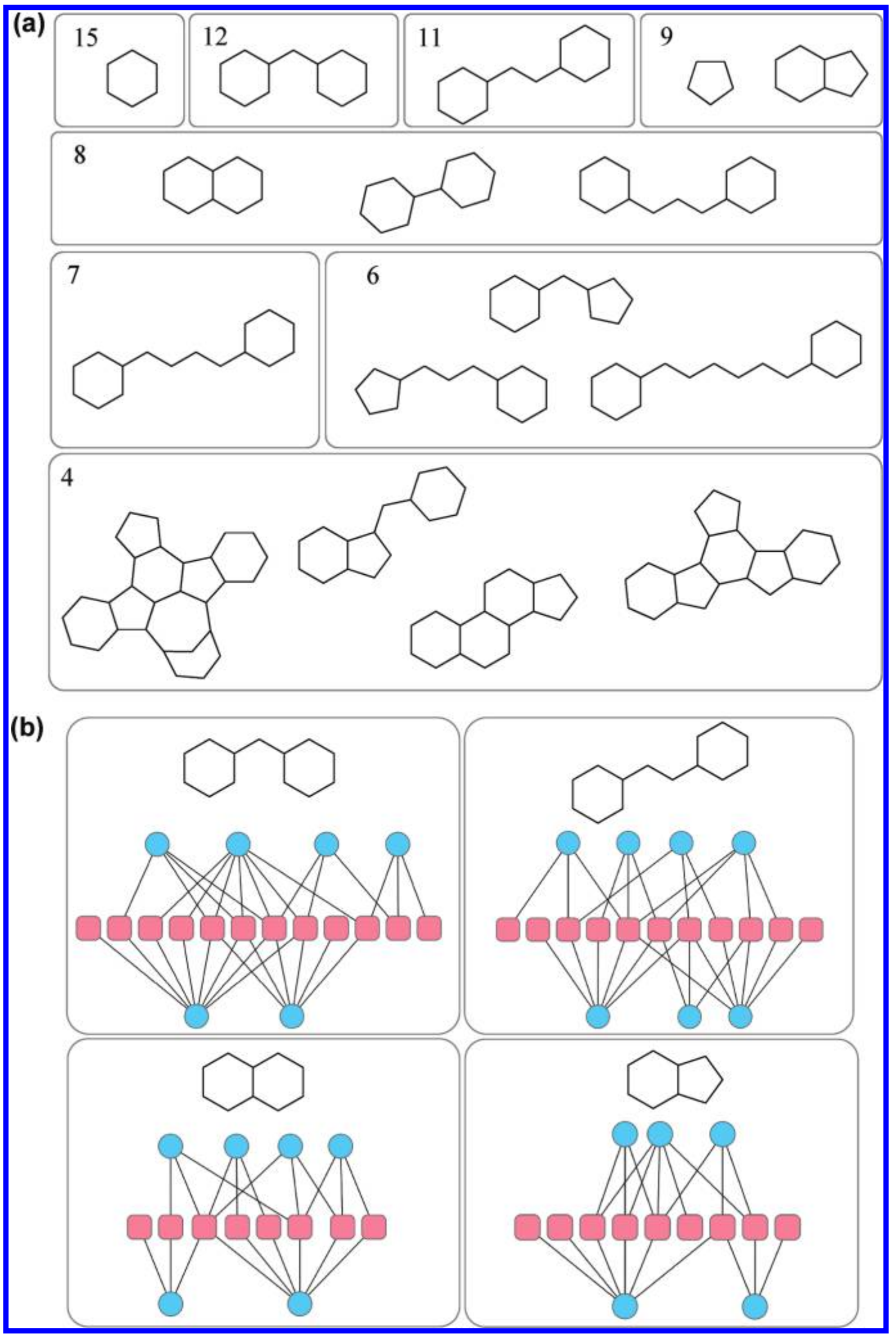

Figure 4. Promiscuous chemotypes. In panel (a), representative promiscuous CSKs are depicted, and the number of target families they are active against is reported (ranging from 4 to 15). In panel (b), scaffold-target family relationships are shown in a bipartite network representation for four CSKs. Pink nodes represent different target families, and blue nodes represent different BM scaffolds yielding the same CSK. This figure is adapted from ref 53.

they represent. Seventeen of these CSKs were found in more than 200 approved drugs, seven of which occurred in at least 10 drugs with at least 17 target annotations.[53] Thus, promiscuous chemotypes isolated from bioactive compounds were well represented in current drugs.

7. PREDICTION OF ACTIVE SCAFFOLDS

One of the most attractive applications of compound data mining is activity prediction, especially the identification of new active scaffolds. Regardless of methodological details, this ultimately involves an extrapolation from known active scaffolds in order to prioritize other candidates predicted to share the same activity. In principle, this task can either be accomplished by applying similarity-based computational methods for scaffold hopping (vide supra) or by utilizing a canonical organization of scaffolds with known and unknown activity from which the activity of new scaffolds might be predicted, for example, on the basis of nearest neighbor relationships.

7.1. Scaffold Hopping. Although scaffold hopping is the primary focal point of retrospective and prospective virtual screening applications, the diversity of known active scaffolds at the level of individual targets has only recently been analyzed. Assessing scaffold diversity on a per-target basis was thought

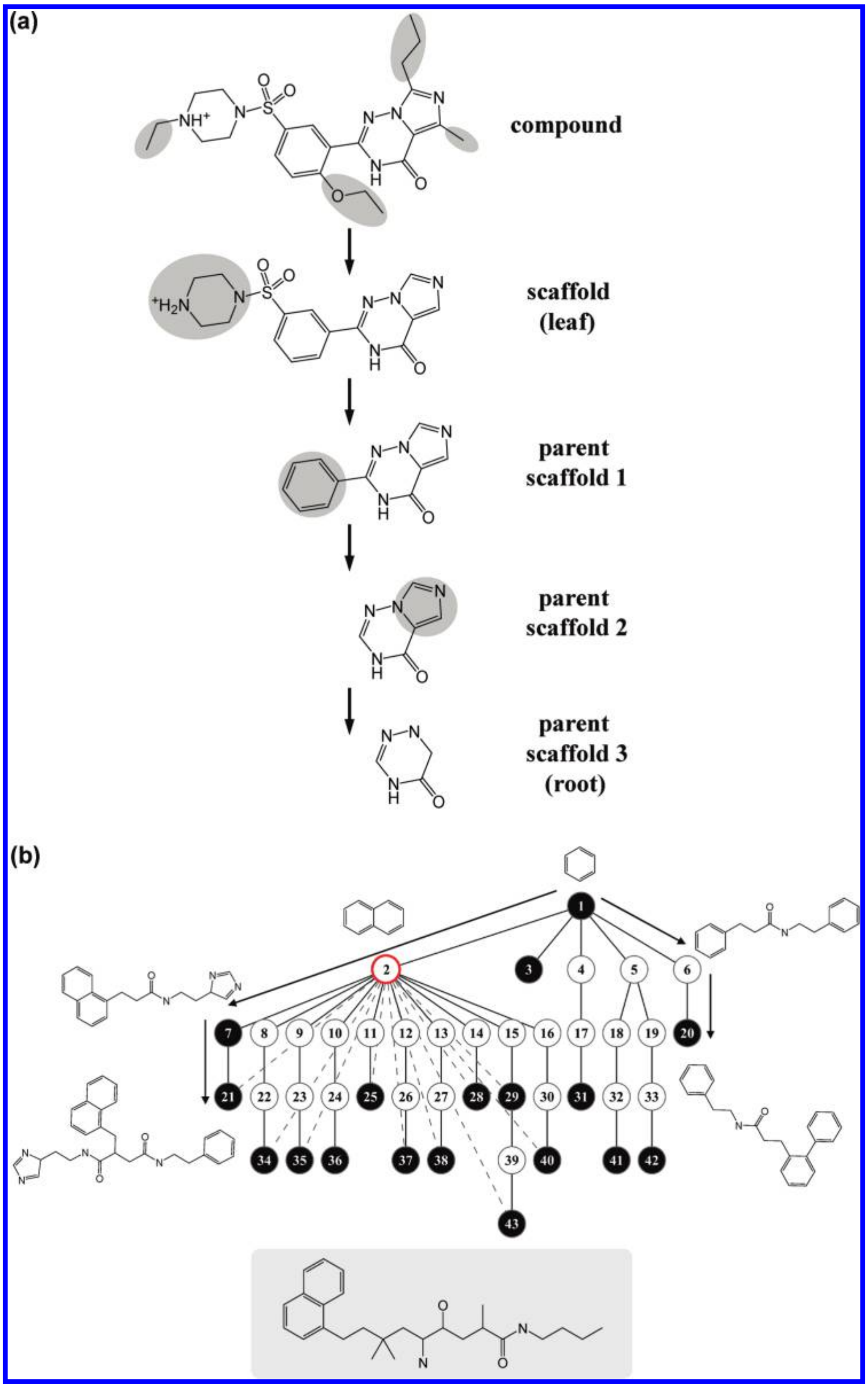

Figure 5. Scaffold tree. (a) An exemplary ST branch generated from an active compound is shown. The deletion of all single-bond substituents from rings or linkers between rings produces the leaf scaffold. Each of the following decomposition steps removes a ring (according to 13 chemical priority rules) and leads to a smaller (parent) scaffold until only a single ring remains (root scaffold). (b) Representative ST branches are shown for renin inhibitors. Nodes represent scaffolds found in active compounds (black) or virtual scaffolds (white). Two virtual scaffolds (2 and 6) are involved in "real-virtual-real" (R-V-R) patterns that are indicated by directed edges. Different from scaffold 6, scaffold 2 (with red border) is also involved in substructure relationships with 11 known active scaffolds (gray dashed lines). Therefore, scaffold 2 is assigned a higher priority for activity prediction. A known renin inhibitor containing scaffold 2 is shown at the bottom. The figure is adapted from ref 63.

to help evaluate how likely it might be to identify new active scaffolds for different targets. For example, if many different scaffolds are already present in compounds having a target-specific activity, it might be quite likely to find additional ones not only because of the available knowledge base for extrapolation but also because the target is likely to be a rather permissive small molecule target. By contrast, if known scaffold diversity is limited, the identification of new active scaffolds might be more difficult. In a study designed to explore target-based scaffold diversity,[54] ~500 target sets, each containing at least five compounds with at least 1 μM potency, were assembled from BindingDB and ChEMBL. Nearly 400 of these target sets were found to contain between five and 99 different active CSKs. In addition, for 28 targets, 100 or more distinct CSKs were already available (each of which representing one or more BM scaffold).[54] In addition to scaffold diversity, structural relationships between scaffolds have also been analyzed. Surprisingly, more than 80% of all bioactive scaffolds were found to be involved in well-defined structural relationships, i.e., either BM scaffolds were topologically equivalent, yielding the same CSK, or one scaffold was a substructure of another.[55] Importantly, ~70% of all scaffolds displayed substructure relationships, and 20% of these substructure relationships involved series of three of more scaffolds. This unexpectedly high degree of structural relatedness was also detected for scaffolds extracted from ZINC[56] compounds.[55] Taken together, the high occurrence of structural relationships between known active and other scaffolds would suggest that prior knowledge governs most compound design and synthesis efforts in medicinal chemistry. This means that most scaffolds are designed on the basis of others, rather than de novo. For scaffold hopping, these findings also have important implications. On the one hand, a large number of active scaffolds is already known for many targets, but on the other hand, many of these scaffolds are structurally related. Thus, from a statistical and structural point of view, scaffold hopping might be less challenging than often assumed. However, structurally unrelated scaffolds sharing the same activity might indeed be difficult to identify.

7.2. Structural Organization. Several structural organization schemes have been introduced that generate molecular hierarchies and capture scaffolds in different ways, including the molecular equivalence number (Meqnum) structural classification system,[15,57] HierS,[22] and ST.[17] Following the Meqnum approach, different structural feature classes are derived in a hierarchical manner, including scaffolds and R-groups, and are utilized to organize compound data sets. The HierS algorithm begins with BM scaffolds and generates all possible derivative scaffolds by systematic removal of fused rings. The ST methodology, which evolved from a tree structure specifically designed to organize natural products,[58] has been extensively applied for compound library design and activity predictions. ST is based on a structural "leaf-to-root" decomposition scheme that involves the iterative removal of rings from BM-like scaffolds (the difference from the original definition being that double-bonded substituents at rings and linkers are retained in this case) until only a single ring remains, as illustrated in panel (a) of Figure 5. Because scaffold decomposition follows a set of predefined chemical rules, the ST hierarchy often contains "virtual scaffolds" that do not occur in the original active compounds. These scaffolds represent prime candidates for activity prediction,[59,60] i.e., virtual scaffolds are predicted to have the same activity as their nearest neighbors in the hierarchy. Furthermore, STs generated for different compound sets might also be merged by identifying scaffolds that are shared between them and by combining the leaf-to-root pathways these scaffolds form.[61] On the basis of ST analysis, several successful activity predictions have been reported.[60,61] In addition, utilizing the ST structure, a statistical methodology termed compound set enrichment (CSE) has been introduced to identify series of analogs in screening data whose activity distribution differs from the one of the entire data set.[62] The link between statistical analysis and the ST concept is intuitive: individual scaffolds represent chemical series and their activities can be mapped on the ST structure. This makes it possible to monitor how activities are distributed across the tree and prioritize other scaffolds for further chemical exploration.

The observation that ~70% of all currently known active scaffolds are involved in substructure relationships (vide supra) has provided a basis to further extend the ST data structure by incorporating additional substructure relationships,[63] as illustrated in panel (b) of Figure 5. Formally, these additional relationships are "leaf-to-leaf" (nonhierarchical) substructure relationships identified by systematic scaffold comparison. Some of these relationships might already be part of the ST structure, if the corresponding scaffolds also occur in leaf-to-root (hierarchical) decomposition pathways. The addition of independently derived nonhierarchical substructure relationships further increases the information content of the ST structure and helps to select preferred virtual scaffolds for activity prediction,[61] as also shown in panel (b) of Figure 5. Virtual scaffolds have high priority if their nearest neighbors are active scaffolds and, in addition, if they are involved in many substructure relationships with other active scaffolds. In addition, the ST and HierS organization schemes have recently been combined to further increase the ST information content and generate "scaffold networks" for CSE applications.[64]

8. ACTIVITY CLIFFS

In general, activity cliffs are formed by structurally similar compounds (often analogs) with large differences in potency.[65] Such activity cliffs are focal points of SAR analysis because they represent the most prominent and often most informative features of target-specific activity landscapes.[66] An unusual question in scaffold analysis has been whether there might be general scaffold preferences for activity cliffs because the formation of cliffs is usually attributed to different substitution patterns of the same or closely related scaffolds. Accordingly, scaffolds have thus far not been primarily considered in the study of activity cliffs. However, when analyzing compound activity data associated with currently available scaffolds, more than 100 BM scaffolds were found that represented compounds forming large-magnitude activity cliffs against different targets often belonging to different families.[67] Examples of multi-target activity cliff scaffolds are shown in panel (a) of Figure 6. A characteristic feature of these cliff-forming scaffolds is that they are not always small and generic, as one might expect, but of rather different chemical complexity. Scaffolds with high propensity to form multi-target cliffs include both community-selective and promiscuous scaffolds. Details of exemplary multi-target activity cliffs are shown in panel (b) of Figure 6. In addition, 40 other scaffolds were identified that formed multi-target "selectivity cliffs".[67] These scaffolds represented structurally similar compounds having different selectivity against a pair of targets.

9. KEY MESSAGES

Scaffold analysis has been carried out from different points of view. For example, knowledge-based approaches have long dominated scaffold selection and compound design efforts.

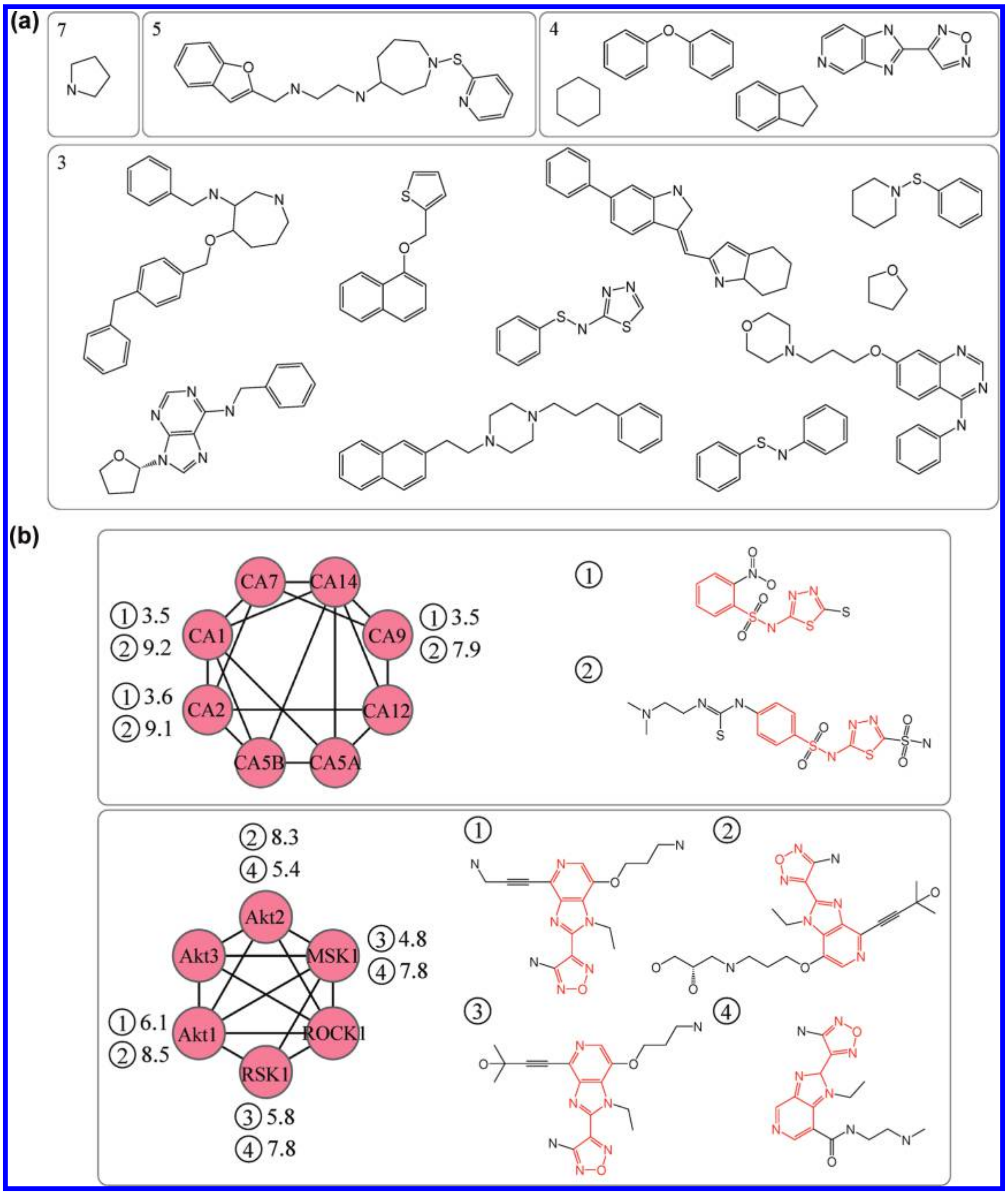

Figure 6. Activity cliff-forming scaffolds. (a) Exemplary BM scaffolds forming multi-target activity cliffs are shown. For each scaffold set, the number of targets is reported. (b) Representative multi-target activity cliffs are shown. Compounds represented by two multi-target activity cliff-forming BM scaffolds (red) are depicted. The compounds are numbered, and their negative logarithmic potency values for individual targets are reported. Nodes represent different targets and are connected by an edge if the targets share compounds containing the scaffold. Target abbreviations: CA, carbonic anhydrase; Akt1, RAC-alpha serine/threonine-protein kinase; Akt2, RAC-beta serine/threonine-protein kinase; Akt3, RAC-gamma serine/threonine-protein kinase; MSK, mitogen- and stress-activated protein kinase; ROCK1, Rho-associated protein kinase 1; and RSK, ribosomal S6 kinase. The figure is adapted from ref 67.

Regardless of the applied methods, a major motivation for scaffold analysis has been (and continues to be) the search for scaffolds that are associated with specific biological activities. However, although scaffolds can be clearly defined, the term scaffold is often loosely and subjectively used in the literature, in particular in scaffold hopping analysis, which renders many studies incomparable. For the chemoinformatics field, this situation presents one of the major problems going forward. Currently, most widely utilized in the literature is the hierarchical scaffold definition following Bemis and Murcko. However, it is generally important to relate BM scaffolds to CSKs in order to emphasize topological differences between scaffolds.

Since the beginnings of systematic scaffold analysis, much emphasis has been put on analyzing scaffold distributions in drugs and comparing them to other compound collections. Here, a key finding has been that scaffold diversity in drugs is generally limited. It has also been observed that bioactive scaffolds are only sparsely distributed in theoretically available scaffold space and that many synthetically available scaffolds have not yet been explored in pharmaceutical research.

Furthermore, PSS have been focal points of knowledge-based scaffold analysis. Although PSS are usually not truly target family specific, a notable enrichment of PSS in ligands of some target families has been observed, in particular for GPCRs and protein kinases. The PSS concept has recently been revisited from a general data mining perspective, leading to the identification of an unexpectedly large number of target community-selective scaffolds, only few of which are present in current drugs.

While PSS and community-selective scaffolds form one end of the current scaffold selectivity spectrum, polypharmacological scaffolds form the other. Systematic data mining has also identified a number of scaffolds that represent compounds with strong polypharmacological behavior. In contrast to community-selective scaffolds, polypharmacological scaffolds were found to be well represented in drugs.

A number of scaffolds of varying chemical complexity have also been detected to form multi-target activity cliffs across different families. Although many of these scaffolds are currently only represented by a few active compounds, these compounds already form large-magnitude activity cliffs against different targets. These findings suggest that many opportunities exist for further chemical exploration of selected scaffolds that might yield compounds with varying potency or selectivity across different targets.

Similar to the large-scale analysis of drug-target interactions, systematic scaffold explorations must also be viewed taking the issue of data incompleteness into account. There are different potential consequences to consider and not all of them are a priori negative. For example, as more assay data and activity annotations become available for known active compounds, the number of community-selective scaffolds is likely to decrease, whereas the numbers of polypharmacological and activity cliff forming scaffolds are likely to further increase. Thus, some of the trends revealed by scaffold analysis based on currently available data are expected to become even stronger in the future.

Another key observation has been that the majority of scaffolds, known to be active or not, is involved in topological and/or substructure relationships. This observation and the finding that for many targets, scaffold hops are already abundant in currently available bioactive compounds have significant implications for scaffold hopping analysis. Furthermore, the presence of structural relationships involving the majority of known active scaffolds suggests that activity prediction utilizing data structures such as ST is a promising approach, consistent with experimental findings obtained thus far.

10. CONCLUSIONS AND FUTURE PERSPECTIVE

Herein, we have discussed different approaches to the analysis of molecular scaffold distributions and the study of scaffold-activity relationships. Scaffold distributions are assessed by extracting scaffolds from known compounds in a consistent manner. The analysis of scaffold—activity relationships additionally requires the comparison of biological activity data of compounds represented by distinct scaffolds. For activity cliff analysis at the level of scaffolds, potency ratios and structural comparisons of compounds containing a given scaffold must also be taken into account. Scaffold distributions in drugs, drug-like compounds, and other molecules have been extensively studied. In addition, privileged substructures have been evaluated in a number of investigations including frequency of occurrence analysis for different target families. Increasing emphasis has been put on the systematic structural organization of scaffold populations and on compound data mining to evaluate different types of scaffold—activity relationships. As the amount of publicly available compound activity data further grows, a variety of opportunities also exist for future research, for example, to study polypharmacological behavior of scaffolds in greater detail or their tendency to yield target- or target family selective compounds. In addition, it should be interesting to further explore synthetically accessible scaffolds that have not yet been evaluated and their structural relationships to known bioactive scaffolds. Another attractive area for future research might be the design of other scaffold hierarchies that focus, for example, on specific therapeutic areas and/or utilize different structural criteria for the organization of scaffold populations.

AUTHOR INFORMATION

Corresponding Author

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

REFERENCES

(1) Brown, N.; Jacoby, E. On scaffolds and hopping in medicinal chemistry. *Mini-Rev. Med. Chem.* **2006**, *6*, 1217–1229.

(2) 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.

(3) Geppert, H.; Vogt, M.; Bajorath, J. Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation. *J. Chem. Inf. Model.* **2010**, *50*, 205–216.

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

(5) Oprea, T.; Davis, A.; Teague, S.; Leeson, P. Is there a difference between leads and drugs? A historical perspective. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1308.

(6) Clark, M. Generalized fragment-substructure based property prediction method. *J. Chem. Inf. Model* **2005**, *45*, 30–38.

(7) 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.

(8) Sutherland, J. J.; Higgs, R. E.; Watson, I.; Vieth, M. Chemical fragments as foundations for understanding target space and activity prediction. *J. Med. Chem.* **2008**, *51*, 2689–2700.

(9) Siegel, M. G.; Vieth, M. Drugs in other drugs: A new look at drugs as fragments. *Drug Discovery Today* **2007**, *12*, 71–79.

(10) Hajduk, P. J.; Greer, J. A decade of structure-based drug design: Strategic advances and lessons learned. *Nat. Rev. Drug Discovery* **2007**, *6*, 211–219.

(11) Congreve, M.; Chessari, G.; Tisi, D.; Woodhead, A. J. Recent developments in fragment-based drug discovery. *J. Med. Chem.* **2008**, *51*, 3661–3680.

(12) Graham, D. J.; Malarkey, C.; Schulmerich, M. V. Information content in organic molecules: Quantification and statistical structure via brownian processing. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1601–1611.

(13) Batista, J.; Bajorath, J. Mining of randomly generated molecular fragment populations uncovers activity-specific fragment hierarchies. *J. Chem. Inf. Model.* **2007**, *47*, 1405–1413.

(14) Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.

(15) Xu, Y.-J.; Johnson, M. Algorithm for naming molecular equivalence classes represented by labeled pseudographs. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 181–185.

(16) Katritzky, A. R.; Kiely, J. S.; Hebert, N.; Chassaing, C. Definition of templates within combinatorial libraries. *J. Comb. Chem.* **2000**, *2*, 2–5.

(17) 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.

(18) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: A web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.* **2007**, *35*, D198–D201.

(19) ChEMBL. European Bioinformatics Institute (EBI): Cambridge, 2011. http://www.ebi.ac.uk/chembl/ (accessed May 2, 2011).

(20) PubChem. National Center for Biotechnology Information: Bethesda, 2011. http://pubchem.ncbi.nlm.nih.gov/ (accessed May 2, 2011).

(21) Broughton, H. B.; Watson, I. A. Selection of heterocycles for drug design. *J. Mol. Graph. Modell.* **2004**, *23*, 51–58.

(22) Wilkens, S. J.; Janes, J.; Su, A. I. HierS: Hierarchical scaffold clustering using topological chemical graphs. *J. Med. Chem.* **2005**, *48*, 182–193.

(23) Wang, J.; Hou, T. Drug and drug candidate building block analysis. *J. Chem. Inf. Model* **2010**, *50*, 55–67.

(24) Hu, Y.; Bajorath, J. Scaffold distributions in bioactive molecules, clinical trials compounds, and drugs. *ChemMedChem* **2010**, *5*, 187–190.

(25) 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*, 4443–4451Lipkushttp://pubs.acs.org/doi/abs/10.1021/jo8001276.

(26) Krier, M.; Bret, G.; Rognan, D. Assessing the scaffold diversity of screening libraries. *J. Chem. Inf. Model.* **2006**, *46*, 512–524.

(27) Shelat, A. A.; Guy, R. K. Scaffold composition and biological relevance of screening libraries. *Nat. Chem. Biol.* **2007**, *3*, 442–446.

(28) Koch, M. A.; Schuffenhauer, A.; Scheck, M.; Wetzel, S.; Casaulta, M.; Odermatt, A.; Ertl, P.; Waldmann, H. Charting biologically relevant chemical space: A structural classification of natural products (SCONP). *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 17272–17277.

(29) Schreiber, S. L. Target-oriented and diversity-oriented organic synthesis in drug discovery. *Science* **2000**, *287*, 1964–1969.

(30) Tan, D. S. Diversity-oriented synthesis: Exploring the intersections between chemistry and biology. *Nat. Chem. Biol.* **1**, *2005*, 74–84.

(31) 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.

(32) 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.

(33) Wester, M. J.; Pollock, S. N.; Coutsias, E. A.; Allu, T. K.; Muresan, S.; Oprea, T. I. Scaffold topologies 2. Analysis of chemical databases. *J. Chem. Inf. Model.* **2008**, *48*, 1304–1310.

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

(35) Blum, L. C.; Reymond, J.-L. 970 Million druglike small molecules for virtual screening in the chemical universe database GDB-13. *J. Am. Chem. Soc.* **2009**, *131*, 8732–8733.

(36) 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.

(37) Müller, G. Medicinal chemistry of target family-directed masterkeys. *Drug Discovery Today* **2003**, *8*, 681–691.

(38) Constantino, L.; Barlocco, D. Privileged substructures as leads in medicinal chemistry. *Curr. Med. Chem.* **2006**, *13*, 65–85.

(39) Klabunde, T.; Hessler, G. Drug design strategies for targeting G-protein-coupled receptors. *ChemBioChem* **2002**, *3*, 928–944.

(40) 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.

(41) Schnur, D. M.; Hermsmeier, M. A.; Tebben, A. J. Are target-family-privileged substructures truly privileged? *J. Med. Chem.* **2006**, *49*, 2000–2009.

(42) 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.

(43) Hu, Y.; Bajorath, J. Exploring target-selectivity patterns of molecular Scaffolds. *ACS Med. Chem. Lett.* **2010**, *1*, 54–58.

(44) Mestres, J.; Gregori-Puigjané, E.; Valverde, S.; Solé, R. V. Data completeness: The Achilles heel of drug-target networks. *Nat. Biotechnol.* **2008**, *26*, 983–984.

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

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

(47) Morphy, R. Selectively nonselective kinase inhibition: Striking the right balance. *J. Med. Chem.* **2010**, *53*, 1413–1437.

(48) Hopkins, A. L. Network pharmacology: The next paradigm in drug Discovery. *Nat. Chem. Biol.* **2008**, *4*, 682–690.

(49) Metz, J. A.; Hajduk, P. J. Rational approaches to targeted polypharmacology: Creating and navigating protein-ligand interaction networks. *Curr. Opin. Chem. Biol.* **2010**, *14*, 498–504.

(50) Keiser, M. J.; Irwin, J. J.; Shoichet, B. K. The chemical basis of pharmacology. *Biochemistry* **2010**, *49*, 10267–10276.

(51) Chen, B.; Wild, D.; Guha, R. PubChem as a source of polypharmacology. *J. Chem. Inf. Model* **2009**, *49*, 2044–2055.

(52) Cases, M.; Mestres, J. A. Chemogenomic approach to drug discovery: Focus on cardiovascular diseases. *Drug Discovery Today* **2009**, *14*, 479–485.

(53) Hu, Y.; Bajorath, J. Polipharmacology directed data mining: Identification of Promiscuous chemotypes with different activity profiles and comparison to approved drugs. *J. Chem. Inf. Model.* **2010**, *50*, 2112–2118.

(54) Hu, Y.; Bajorath, J. Global assessment of scaffold hopping potential for current pharmaceutical target. *Med. Chem. Commun.* **2010**, *1*, 339–344.

(55) Hu, Y.; Bajorath, J. Structural and potency relationships between scaffolds of compounds active against human targets. *ChemMedChem* **2010**, *5*, 1681–1685.

(56) Irwin, J. J.; Shoichet, B. K. ZINC—A free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 177–182.

(57) Xu, Y.-J.; Johnson, M. Using molecular equivalence numbers to visually explore structural features that distinguish chemical libraries. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 912–926.

(58) Koch, M. A.; Schuffenhauer, A.; Scheck, M.; Wetzel, S.; Casaulta, M.; Odermatt, A.; Ertl, P.; Waldmann, H. Charting biologically relevant chemical space: A structural classification of natural products (SCONP). *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 17272–17277.

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

(60) 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.

(61) 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.

(62) Varin, T.; Gubler, H.; Parker, C. N.; Zhang, J.-H.; Raman, P.; Ertl, P.; Schuffenhauer, A. Compound set enrichment: A novel approach to analysis of primary HTS data. *J. Chem. Inf. Model.* **2010**, *50*, 2067–2078.

(63) Hu, Y.; Bajorath, J. Combining horizontal and vertical substructure relationships in scaffold hierarchies for activity prediction. *J. Chem. Inf. Model.* **2011**, *51*, 248–257.

(64) Varin, T.; Schuffenhauer, A.; Ertl, P.; Renner, S. Mining for bioactive scaffolds with scaffold networks: Improved compound set enrichment from primary screening data. *J. Chem. Inf. Model.* **2011**, *51*, 1528–1538.

(65) Maggiora, G. M. On outliers and activity cliffs—Why QSAR often disappoints. *J. Chem. Inf. Model,* **2006**, *46*, 1535.

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

(67) Hu, Y.; Bajorath, J. Molecular scaffolds with high propensity to form multi-target activity cliffs. *J. Chem. Inf. Model* **2010**, *50*, 500–510.

(68) Salaski, E. J.; Krishnamurthy, G.; Ding, W.-D.; Yu, K.; Insaf, S. S.; Eid, C.; Shim, J.; Levin, J. L.; Tabei, K.; Toral-Barza, L.; Zhang, W.-G.; McDonald, L. A.; Honores, E.; Hanna, C.; Yamashita, A.; Johnson, B.; Li, Z.; Laakso, L.; Powell, D.; Mansour, T. S. Pyranonaphthoquinone lactones: A new class of AKT selective kinase inhibitors alkylate a regulatory loop cysteine. *J. Med. Chem.* **2009**, *52*, 2181–2184.

(69) Yan, A.; Wang, L.; Xu, S.; Xu, J. Aurora-A kinase inhibitor scaffolds and binding modes. *Drug Discovery Today* **2011**, *16*, 5–6.

(70) *Kinase Inhibitor Database* http://www.gvkbio.com/database_pdf/Kinase_Inhibitor.pdf (assessed April 4, 2011).

(71) 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.