Systematic Identification of Scaffolds Representing Different Types of Structure-Activity Relationships

Kumulative Dissertation

zur Erlangung des Doktorgrades (Dr. rer. nat.)

der Mathematisch-Naturwissenschaftlichen Fakultät

der Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von
Ye Hu
aus Jiangsu, China

Bonn
March, 2011

Angefertigt mit Genehmigung
der Mathematisch-Naturwissenschaftliche Fakultät
der Rheinischen Friedrich-Wilhelms-Universität Bonn

1. Referent: Univ.-Prof. Dr. rer. nat. Jürgen Bajorath
2. Referent: Univ.-Prof. Dr. rer. nat. Michael Gütschow

Abstract

In medicinal chemistry, it is of central importance to understand structure-activity relationships (SARs) of small bioactive compounds. Typically, SARs are analyzed on a case-by-case basis for sets of compounds active against a given target. However, the increasing amount of compound activity data that is becoming available allows SARs to be explored on a large-scale. Moreover, molecular scaffolds derived from bioactive compounds are also of high interest for SAR analysis. In general, scaffolds are obtained by removing all substituents from rings and from linkers between rings.

This thesis aims at systematically mining compounds for which activity annotations are available and investigating relationships between chemical structure and biological activities at the level of active compounds, in particular, molecular scaffolds. Therefore, data mining approaches are designed to identify scaffolds with different structural and/or activity characteristics. Initially, scaffold distributions in compounds at different stages of pharmaceutical development are analyzed. Sets of scaffolds that overlap between different stages or preferentially occur at certain stages are identified. Furthermore, a systematic selectivity profile analysis of public domain active compounds is carried out. Scaffolds that yield compounds selective for communities of closely related targets and represent compounds selective only for one particular target over others are identified. In addition, the degree of promiscuity of scaffolds is thoroughly examined. Eighty-three scaffolds covering 33 chemotypes correspond to compounds active against at least three different target families and thus are considered to be promiscuous. Moreover, by integrating pairwise scaffold similarity and compound potency differences, the propensity of scaffolds to form multi-target activity or selectivity cliffs and, in addition, the global scaffold potential of individual targets are quantitatively assessed, respectively. Finally, structural relationships between scaffolds are systematically explored. Most scaffolds extracted from active compounds are found to be involved in substructure relationships and/or share topological features with others. These substructure relationships are also compared to, and combined with, hierarchical substructure relationships to facilitate activity prediction.

Acknowledgments

I would like to first thank my supervisor Prof. Dr. Jürgen Bajorath for his inspirational guidance, great patience, continuous support and encouragement during my PhD study. I would also like to thank Prof. Dr. Michael Gütschow for being co-referent and taking time to review my thesis.

I would like to express my gratitude to all my colleagues of the LSI research group for providing the helpful, friendly, interactive and collaborative working atmosphere. Especially, many thanks are given to José Batista, Eugen Lounkine, Lisa Peltason, Anne Mai Wassermann and Preeti Iyer for pleasant collaborations and fruitful discussions.

Finally, I would like to thank all my friends, especially Qiong Lin, for any kind of help and encouragement they have ever done for me and special thanks to my family for their support during the past few years since I have been in Germany.

Contents

Introduction **1**

1 Scaffold Distributions in Bioactive Molecules, Clinical Trials Compounds and Drugs **9**

Introduction 9
Publication 11
Summary 15

2 Systematic Analysis of Public Domain Compound Potency Data Identifies Selective Molecular Scaffolds Across Druggable Target Families **17**

Introduction 17
Publication 19
Summary 27

3 Exploring Target-Selectivity Patterns of Molecular Scaffolds **29**

Introduction 29
Publication 31
Summary 37

4 Polypharmacology Directed Compound Data Mining: Identification of Promiscuous Chemotypes **39**

Introduction 39
Publication 41
Summary 49

5 Molecular Scaffolds with High Propensity to Form Multi-Target Activity Cliffs **51**

Introduction 51
Publication 53
Summary 65

6 Global Assessment of Scaffold Hopping Potential for Current Pharmaceutical Targets 67
Introduction 67
Publication 69
Summary 75

7 Structural and Potency Relationships Between Scaffolds of Compounds Active Against Human Targets 77
Introduction 77
Publication 79
Summary 85

8 Combining Horizontal and Vertical Substructure Relationships in Scaffold Hierarchies for Activity Prediction 87
Introduction 87
Publication 89
Summary 99

Conlcusion 101

Introduction

After the completion of human genome project, it has been predicted that products of ∼3,000 genes might represent druggable targets and ∼600 to ∼1,500 of these targets might be directly linked to diseases [1, 2]. On the other hand, chemical space consisting of all possible small molecules is estimated to contain more than 10^{60} molecules with at most 30 heavy atoms [3]. However, "biologically relevant chemical space" that consists of chemical compounds acting on biological system represents only a small fraction of theoretically possible chemical space [4]. In chemical biology and medicinal chemistry research, it is of high importance, and also challenging, to understand structure-activity relationships (SARs) of such bioactive compounds that bind to one or more individual targets and trigger biological responses and therapeutic effects.

Structure-Activity Relationships

Traditionally, SARs have been explored on a case-by-case basis, i.e. for individual compound series active against a given target. For this purpose, a number of computational approaches are available such as classical quantitative SAR models [5], pharmacophore [6] or machine learning techniques [7]. Recently, several new methodologies have also been developed. For example, different numerical functions have been designed to quantitatively characterize SAR features contained in a data set [8, 9]. In addition, computational activity landscapes have also been utilized to graphically represent both structure and potency relationships between compounds having similar biological activity [10]. Activity landscapes of different design and complexity have been introduced such as two-dimensional Network-like Similarity Graphs [11] or three-dimensional landscape views [12], where regions displaying different global and local SAR characteristics can be identified. Furthermore, SAR contributions of substitution sites and their combinations have also been quantitatively analyzed for series of analogous compounds [13]. However, such SAR determinants can also be explored at the level of molecular scaffolds.

Molecular Scaffolds

Molecular scaffolds or frameworks extracted from active compounds have been, and continue to be, of high interest in medicinal chemistry. Different definitions of scaffolds are available. For example, scaffolds might be generated by breaking predefined bonds in compounds on the basis of retrosynthetic rules [14]. Alternatively, scaffolds might also be obtained by removing all substituents from rings and from linkers between rings, forming molecular frameworks, also called Bemis-Murcko scaffolds [15].

A number of scaffold analyses have been carried out from rather different points of view. For example, possible scaffold topologies have been exhaustively enumerated for up to eight rings and the structural complexity of chemical databases was analyzed on the basis of these scaffold topologies [16, 17]. Furthermore, the relationship between aromatic ring count and compound developability was analyzed on the basis of different physicochemical properties and ring types [18, 19]. Moreover, scaffold distributions and diversity have been examined for different data sources such as screening libraries [20], large databases of organic compounds [21], natural products [22, 23], and drugs or drug candidates [24].

In addition, a classification scheme has been introduced that organizes scaffolds and their derivatives in hierarchies and facilitates the identification of new ligand types [25, 26]. Moreover, the notion of *"privileged substructures"*, originally introduced by Evans *et al* [27], is highly attractive for drug discovery. Privileged substructures are scaffolds thought to preferentially, or exclusively, bind to a specific target family [28–30].

Availability of Public Compound Data

With the advant of high-throughput screening techniques, compounds can be effectively assayed against an array of targets [31]. Therefore, increasing numbers of active compounds become available, which enable SARs to be analyzed on a large scale, instead of by conventional case-by-case investigations. Efforts have been dedicated to build publicly accessible databases that are composed of compounds annotated with targets and measured binding affinities. For example, *PubChem* [32] is a pioneering initiative that has organized millions of compound structures and substance information. Furthermore, it also contains more than 490,000 bioassays including high-throughput screening data. In addition, *BindingDB* currently stores 284,206 small molecules with 648,915 binding records for 5,662 different protein targets [33], and *ChEMBLdb* deposits 658,075 compounds with more than 3,000,000 activity measurements against 8,091 targets [34].

These databases grow steadily. Therefore, the design of effective computational methods for mining large databases and extracting available SAR information becomes critically important for the identification of potential hits and predictive model of biological activities.

Thesis Outline

The major goal of my doctoral studies has been the systematic analysis of SARs in publicly available compound data at the level of molecular scaffolds. A series of studies have been designed to identify sets of scaffolds with different SAR characteristics and information content. This dissertation consists of eight individual chapters:

(a) Analysis of scaffold distributions in compounds at different stages of pharmaceutical development and exploration of scaffolds that might preferentially occur at early- and/or late-stage (*Chapter 1*).

(b) Identification of scaffolds selective for communities of closely related targets (*Chapter 2*).

(c) Identification of scaffolds yielding compounds that are always selective for a particular target over one or more others (*Chapter 3*).

(d) Search for promiscuous scaffolds and chemotypes representing compounds active against multiple target families (*Chapter 4*).

(e) Exploration of scaffolds with high propensity to yield compounds forming activity or selectivity cliffs against different targets (*Chapter 5*).

(f) Assessment of scaffold hopping potential of pharmaceutical targets at a global level (*Chapter 6*).

(g) Investigatation of structural diversity of scaffolds representing currently available active compounds (*Chapter 7*).

(h) Examination of a hierarchical scaffold classification scheme and search for additional structural information between scaffolds (*Chapter 8*).

Bibliography

[1] Hopkins A. L., Groom C. R. The druggable genome. *Nat. Rev. Drug Discov.* **2002**, 1, 727-730.

[2] Russ A. P., Lampel S. The druggable genome: an update. *Drug Discov. Today* **2005**, 10, 1607-1610.

[3] Bohacek R. S., McMartin C., Guida W.C. The art and practice of structure-based drug design: a molecular modeling perspective. *Med. Res. Rev.* **1996**, 16, 3-50.

[4] Dobson C. M. Chemical space and biology. *Nature* **2004**, 432, 824-828.

[5] Esposito E. X., Hopfinger A. J., Madura J. D. Methods for applying the quantitative structure-activity relationship paradigm. *Methods Mol. Biol.* **2004**, 275, 131-214.

[6] Leach A. R., Gillet V. J., Lewis R. A., Taylor R. Three-dimensional pharmacophore methods in drug discovery. *J. Med. Chem.* **2010**, 53, 539-558.

[7] 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.

[8] Peltason L., Bajorath J. SAR index: quantifying the nature of structure-activity relationships. *J. Med. Chem.* **2007**, 50, 5571-5578.

[9] Guha R., Van Drie J. H. Structure–activity landscape index: identifying and quantifying activity cliffs. *J. Chem. Inf. Model.* **2008**, 48, 646-658.

[10] Wassermann A. M., Wawer M., Bajorath J. Activity landscape representations for structure-activity relationship analysis. *J. Med. Chem.* **2010**, 53, 8209-8223.

[11] Wawer M., Peltason L., Weskamp N., Teckentrup A., Bajorath J. Structure-activity relationship anatomy by network-like similarity graphs and local structure-activity relationship indices. *J. Med. Chem.* **2008**, 51, 6075-6084.

[12] Peltason L., Iyer P., Bajorath J. Rationalizing three-dimensional activity landscapes and the influence of molecular representations on landscape topology and the formation of activity cliffs. *J. Chem. Inf. Model.* **2010**, 50, 1021-1033.

[13] Peltason L., Weskamp N., Teckentrup A., Bajorath J. Exploration of structure-activity relationship determinants in analogue series. *J. Med. Chem.* **2009**, 52, 3212-3224.

[14] 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.

[15] Bemis G. W., Murcko M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, 39, 2887-2893.

[16] 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.

[17] 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, 1311-1324.

[18] Ritchie T. J., Macdonald S. J. The impact of aromatic ring count on compound developability–are too many aromatic rings a liability in drug design? *Drug Discov. Today* **2009**, 14, 1011-1020.

[19] Ritchie T. J., Macdonald S. J., Young R. J., Pickett S. D. The impact of aromatic ring count on compound developability: further insights by examining carbo- and hetero-aromatic and -aliphatic ring types. *Drug Discov. Today* **2011**, 16, 164-171.

[20] Krier M., Bret G., Rognan D. Assessing the scaffold diversity of screening libraries. *J. Chem. Inf. Model.* **2006**, 46, 512-24.

[21] Lipkus A. H., Yuan Q., Lucas K. A., Funk S. A., Bartelt W. F. III, Schenck R. J., Trippe A. J. Structural diversity of organic chemistry. A scaffold analysis of the CAS Registry. *J. Org. Chem.*, **2008**, 73, 4443-4451.

[22] Grabowski K., Schneider G. Properties and architecture of drugs and natural products revisited. *Curr. Chem. Biol.* **2007**, 1, 115-127.

[23] Grabowski K., Baringhaus K. H., Schneider G. Scaffold diversity of natural products: inspiration for combinatorial library design. *Nat. Prod. Rep.* **2008**, 25, 892-904.

[24] Wang J., Hou T. Drug and drug candidate building block analysis. *J. Chem. Inf. Model.* **2010**, 50, 55-67.

[25] 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.

[26] 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.

[27] 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.

[28] Horton D. A., Bourne G. T., Smythe M. L. The combinatorial synthesis of bicyclic privileged structures or privileged substructures. *Chem. Rev.* **2003**, 103, 893-930.

[29] Constantino L., Barlocco D. Privileged substructures as leads in medicinal chemistry. *Curr. Med. Chem.* **2006**, 12, 65-85.

[30] Schnur D. M., Hermsmeier M. A., Tebben A. J. Are target-family-privileged substructures truly privileged? *J. Med. Chem.* **2006**, 49, 2000-2009.

[31] Bleicher K. H., Böhm H. J., Müller K., Alanine A. I. A guide to drug discovery: hit and lead generation: beyond high-throughput screening. *Nat. Rev. Drug Discov.* **2003**, 2, 369-378.

[32] *PubChem*; National Center for Biotechnology Information: Bethesda, 2010; http://pubchem.ncbi.nlm.nih.gov/

[33] 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.

[34] *ChEMBL*; European Bioinformatics Institute (EBI): Cambridge, 2010; http://www.ebi.ac.uk/chembl/.

Chapter 1

Scaffold Distributions in Bioactive Molecules, Clinical Trials Compounds and Drugs

Introduction

The frequency of occurrence of molecular frameworks (or scaffolds) has been explored in many studies in order to associate scaffolds with different biological activities and investigate lead- or drug-like properties of active compounds. Here, we have analyzed scaffold distributions in compounds at different stages of pharmaceutical development, i.e. biologically active molecules, compounds in clinical trials, and registered or approved drugs. Subsets of scaffolds that overlapped across different stages were extracted and their inter- and intra-subset structural diversity was examined. In addition, scaffolds that preferentially occurred during certain development stages were identified.

DOI: 10.1002/cmdc.200900419

Scaffold Distributions in Bioactive Molecules, Clinical Trials Compounds, and Drugs

Ye Hu and Jürgen Bajorath*[a]

Molecular building blocks including scaffolds (core structures)[1–4] and fragments of varying origin and size[5–8] have been intensely investigated in the search for target-class-directed structural motifs[9–11] and in fragment-based drug discovery.[12–15] In the context of these studies, it has often been possible to associate scaffolds, fragments, or combinations of fragments with specific or multiple biological activities.[4–11] The majority of these studies have surveyed distributions of fragments in biologically relevant compounds on the basis of frequency analysis.[5–10] Furthermore, scaffolds present in known drugs[1] or compounds directed against different target classes[10,11] have been analyzed in order to evaluate structural features that distinguish drugs from non-drugs or that are characteristic of certain drug classes. These structure-oriented investigations are conceptually related to other studies of lead-like or drug-like compound character that have predominantly focused on analyzing molecular property distributions with the aid of various molecular descriptors.[16–19] Taken together, these and other studies have substantially aided in elucidating structural signatures of various biological activities and in identifying molecular property distributions consistent with drug- or lead-likeness.

We have been interested in analyzing scaffold distributions from a different perspective, that is, in order to better understand how structural features in compounds at different stages of pharmaceutical development might compare. Therefore, we carried out a comparative molecular scaffold analysis of three sets of compounds representing different stages in drug discovery: biologically active molecules (hits or leads), compounds in clinical trials, and registered/approved drugs. With this analysis, we attempted to explore several questions. For example, would there be notable differences in the composition of scaffold populations at different development stages? Might some scaffolds preferentially occur in early- but not late-stage compounds or drugs? Or would certain scaffolds be consistently found in these types of compounds? Clearly, such questions are, to some extent, inspired by the high clinical attrition rates of drug candidates.[20,21]

Initially, we assembled suitable compound data sets. As a pool of hits and leads, we retrieved all active molecules directed against human targets from BindingDB,[22] which contains active compounds taken from original literature sources and their target information. A total of 17 837 BindingDB molecules were collected, which can be regarded as a representative sample of biologically relevant chemical space. Because no publicly accessible repository exists for compounds that are or have been in clinical trials, we extracted clinical trials compounds from the MDL Drug Data Report (MDDR),[23] obtaining a total of 1586 molecules. The situation is different for approved drugs that are available in DrugBank.[24] Hence, 1493 approved drugs were taken from DrugBank and combined with 1491 registered or launched drugs extracted from the MDDR, giving a set comprising 2980 unique drugs; compounds producing identical SMILES strings were considered duplicates. The limited overlap between drugs currently available in DrugBank and the MDDR has been a rather surprising finding. Small numbers of compounds that occurred in more than one of the accessed databases were omitted such that there was no compound overlap between our sets of bioactive compounds, clinical trials compounds, and drugs. Table 1 summarizes the composition of these compound data sets.

Table 1. Data sets.[a]

Data Set	# Molecules	# Scaffolds	# Carbon Skeletons
BindingDB	17 837	6451	2910
Clinical Trials	1586	1270	842
Drugs	2980	1233	603

[a] The number of molecules, unique hierarchical scaffolds, and unique carbon skeletons is reported for each compound set.

A few aspects of the design of the compound sets are worth considering. For the purpose of our scaffold analysis, the bioactive compound set was intended to be larger than the clinical trials and drug sets simply because biologically relevant chemical space is larger than drug candidate/drug space. Also, the clinical trials compounds we could access are transient; that is, they either reach drug status at some point or fail (however, in light of the high clinical attrition rates, the majority of these compounds are expected to fail). Hence, these compounds are a snapshot of current trials and represent a smaller sample than known drugs that have accumulated over time. Furthermore, it should also be considered that compound selectivity/specificity criteria differ between these sets. Whereas BindingDB contains many molecules with reported activity against multiple targets, for clinical trials compounds and drugs, a specific mode of action must usually be demonstrated as part of the drug approval process. Hence, specificity requirements become increasingly stringent over different de-

---

[a] *Y. Hu, Prof. Dr. J. Bajorath*
Department of Life Science Informatics, B-IT, LIMES
Program Unit Chemical Biology and Medicinal Chemistry
Rheinische Friedrich-Wilhelms-Universität Bonn
Dahlmannstr. 2, 53113 Bonn (Germany)
Fax: (+49) 228-2699-341
E-mail: bajorath@bit.uni-bonn.de

Supporting information for this article is available on the WWW under http://dx.doi.org/10.1002/cmdc.200900419.

© 2010 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

velopment stages, and it is currently unclear how generally increasing compound specificity might be reflected at the structural level.

To systematically extract molecular scaffolds from our compound sets, we applied the scaffold generation scheme of Bemis and Murcko[1] that was first used to study building blocks of drugs. These scaffolds are derived from compounds by removing all substituents from ring systems but retaining non-substituted aliphatic linkers between rings.[1] As a further level of generalization, we also generated carbon skeletons from scaffolds by setting all atom types to carbon and all bond orders to single bonds. Thus, several unique scaffolds can correspond to the same carbon skeleton, and unique skeletons represent different molecular topologies. Figure 1 illustrates the relationship between compounds, scaffolds, and carbon skeletons.

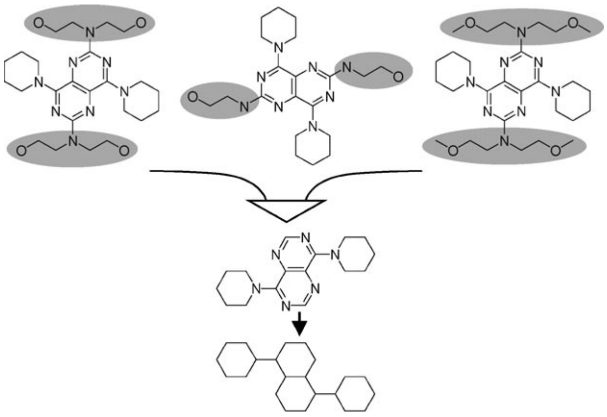


Figure 1. Scaffolds and carbon skeletons: Three molecules, the scaffold they share, and the corresponding carbon skeleton are shown. Parts of the molecules removed to generate the scaffold are displayed on a gray background. All calculations required for our analysis were carried out using in-house-generated Perl scripts and the PipelinePilot environment (version 6.1.5).[25]

Table 1 lists the number of scaffolds and skeletons extracted from each compound data set. In each case, a relatively large number of scaffolds was obtained, with a ratio of ~2.7 compounds per scaffold for bioactive molecules, ~2.4 for drugs, and ~1.3 for clinical trials compounds. Furthermore, these scaffolds displayed a surprising degree of diversity, as indicated by the large number of carbon skeletons generated from them, ranging from ~2.2 scaffolds per skeleton for bioactive molecules and ~2 for drugs to ~1.5 for clinical trials compounds.

We next determined the overlap between these three scaffold sets. The results are shown in Figure 2. There was comparably small overlap between these scaffold ensembles. A total of 65 scaffolds were found in all three sets (subset BCD); 79 scaffolds were found in bioactive molecules and clinical trials compounds, but not drugs (BC); 85 scaffolds in bioactive molecules and drugs, but no clinical trials compounds (BD); and 90 scaffolds in clinical trials compound and drugs, but not in bioactive molecules (CD). The intra- and inter-subset diversity of these scaffolds was comparable, as revealed by a similarity-based scaffold network calculated for these four subsets, shown in Figure 3 (generated with Cytoscape[26]). Scaffolds from different overlap subsets are represented as gray-scaled nodes. Edges are drawn between nodes if their pairwise Tanimoto coefficient[27] calculated using MACCS structural keys[28] is greater than 0.8. Thus, structurally similar scaffolds are connected. In this network representation, no large central network component is observed as well as no preferential clustering of intra-subset scaffolds. Only a limited number of similarity-based scaffold clusters are formed. These clusters are mostly composed of scaffolds from different subsets. These observations reflect significant structural diversity among overlap scaffolds. Similar observations were made when the scaffold subsets were analyzed on the basis of a similarity-based scaffold

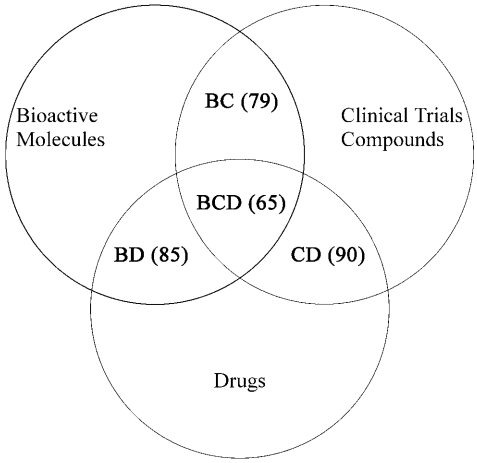


Figure 2. Scaffold Venn diagram: The comparison of scaffold ensembles extracted from the three compound data sets is shown. Four subsets of scaffolds are identified that represent different overlaps. The number of scaffolds in each subset is given in parentheses.

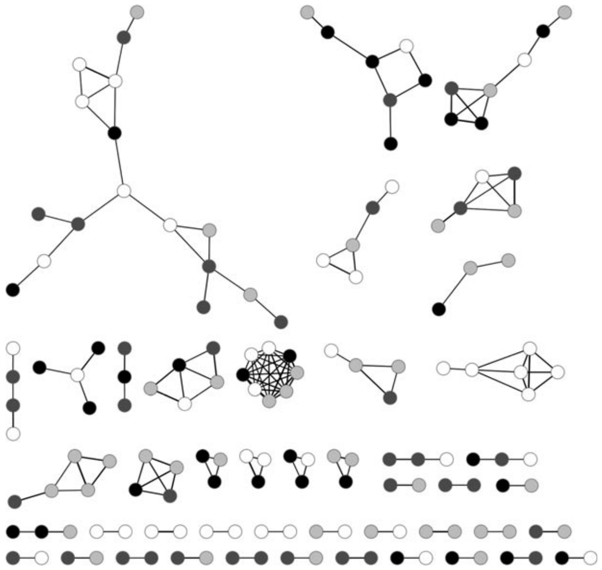


Figure 3. Similarity-based scaffold network: The four scaffold overlap subsets shown in Figure 2 are organized in a network representation (generated with Cytoscape[26]). Scaffolds are represented as nodes: BCD (black), BC (dark gray), BD (light gray), and CD (white). Edges are drawn between nodes representing structurally similar scaffolds.

© 2010 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

network of the entire set of 6451 BindingDB scaffolds (Supporting Information figure S1). This network was calculated as the one shown in Figure 3 and subset scaffolds were then mapped. Furthermore, we also mapped the scaffold subsets on a target-based network of BindingDB scaffolds (Supporting Information figure S2). Different from the similarity-based scaffold networks, in this case, nodes represent BindingDB scaffolds that are connected if compounds containing these scaffolds are shared by at least two target proteins. This network displayed target-directed clustering of BindingDB scaffolds, but only very little clustering of subset scaffolds was observed when they were mapped onto the network. Thus, there were no systematic structural relationships detectable within or between these scaffold overlap subsets and no target cluster preferences. Supporting Information figure S3 reports all scaffolds comprising the four overlap subsets.

We also isolated scaffolds from drugs withdrawn from the market. These represent an interesting subset of approved drugs, owing to severe side effects associated with them that were identified in the course of regular patient treatment. However, only a small number of withdrawn drugs were found in DrugBank (62) and the MDDR (33), yielding a set of 95 unique compounds from which a total of 43 scaffolds were isolated. Of these scaffolds, 11, 7, and 24 were also found in BindingDB compounds, clinical trials compounds, and drugs, respectively. Five of these scaffolds consistently occur in all compound sets. The four scaffold overlap sets containing scaffolds from withdrawn drugs are shown in Supporting Information figure S4. Given the small number of scaffolds isolated from withdrawn drugs, it is difficult to draw conclusions from their distributions in other compound sets. However, approximately half of these scaffolds also appear in non-withdrawn drugs, and thus cannot be directly responsible for severe side effects.

Table 2 lists the number of compounds in the different data sets that contain scaffolds from overlap subsets BCD, BC, BD, and CD. The BCD scaffolds represented 1503 bioactive molecules (8.4 %), 213 clinical trials compounds (13.4 %), and 552 drugs (18.5 %). Many but not all of these scaffolds are small aromatic and heteroaromatic rings and are thus rather generic in nature (Supporting Information figure S3 a). Accordingly, the BCD subset produced fewer carbon skeletons than the other scaffold subsets (Table 2). However, the BCD subset also contained a number of large and complex scaffolds that were recurrent in bioactive molecules, clinical trials compounds, and drugs (Supporting Information figure S3 a). Thus, compounds having such scaffolds are likely to pass through different stages of pharmaceutical development. Furthermore, the presence of 90 CD scaffolds indicated that bioactive molecules currently available in the public domain are an incomplete sample of structural classes present in clinical trials compounds and drugs. The BD subset consists of 85 drug scaffolds that are also available in bioactive molecules, but not current clinical trials compounds. Compounds containing these scaffolds might often not be subjected to clinical evaluation because they already exist in established drugs and thus lack novelty. The BC subset of 79 scaffolds is also of interest because scaffolds present in bioactive molecules and clinical trials compounds, but not drugs, might contain chemotypes that preferentially undergo attrition during clinical evaluation. The availability of the BC subset makes it possible to further analyze whether individual scaffolds contained in compounds of interest have previously failed during clinical evaluation, which requires follow-up studies of patent literature and clinical trials reports.

Table 2. Comparison of four distinct sets of scaffolds.[a]

Set Identifier	# Scaffolds	# CSK	# Molecules		
			BindingDB	Clinical Trials	Drugs
BCD	65	28	1503	213	552
BC	79	68	520	86	NA
BD	85	50	357	NA	192
CD	90	78	NA	146	250

[a] Systematic comparison of the scaffold ensembles extracted from the three compound data sets according to Table 1 yields four distinct scaffold subsets that are either shared by all three (BCD) or two of three (BC, BD and CD) ensembles. The number of scaffolds in each subset and the number of corresponding carbon skeletons (# CSK) is reported. The number of molecules in each data set containing these scaffolds is also provided.

We also searched our data sets for scaffolds with characteristic frequencies of occurrence. Representative examples are shown in Figure 4. For example, scaffolds that occur with high frequency in all compound data sets are generally small and generic (including the benzene ring), as one would anticipate. Furthermore, only few scaffolds were found that had a steadily decreasing frequency from bioactive molecules to clinical compounds and drugs, which would be consistent with attrition

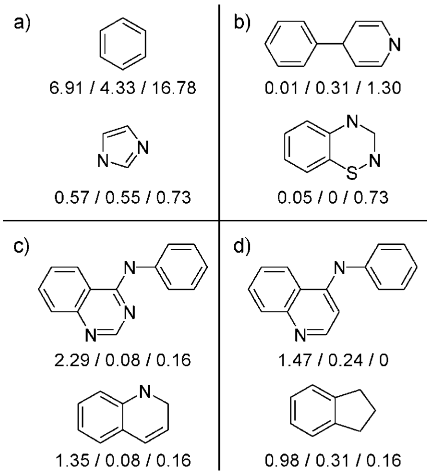


Figure 4. Scaffolds with various frequencies of occurrence: Shown are representative examples of scaffolds with distinct distributions in bioactive molecules, clinical trials compounds, and drugs. The frequencies of compounds (in %) containing each scaffold in the three compound data sets are reported as bioactive/clinical trials/drugs. a) scaffolds with overall moderate to high frequency of occurrence; b) scaffolds enriched in drugs; c) scaffolds frequently occurring in bioactive molecules, but rarely in late-stage compounds or drugs; and d) scaffolds with decreasing frequency of occurrence over different development stages.

along the pathway. In contrast, 90 scaffolds were found that were detectably enriched in drugs over both bioactive molecules and clinical compounds. Moreover, 114 scaffolds were identified that occurred with high frequency in bioactive molecules but rarely in both clinical trials compounds and drugs.

In summary, we have systematically analyzed and compared scaffolds contained in bioactive molecules, compounds in clinical trials, and known drugs. The analysis provides insight into differences in scaffold distributions, the degree of scaffold diversity, and the occurrence of overlap between scaffolds contained in compounds at different pharmaceutical development stages. The scaffolds comprising our four overlap subsets have been made available and can be readily used as markers to analyze newly discovered active compounds and to determine whether the scaffolds they contain are known to preferentially occur in early-stage molecules, compounds in clinical trials, and/or drugs. Upon publication of our analysis, the scaffold information can also be freely obtained via the following URL: http://www.lifescienceinformatics.uni-bonn.de (Downloads section).

Keywords: bioactive molecules · clinical trials compounds · drugs · molecular scaffolds · scaffold distributions · statistical analysis

[1] G. W. Bemis, M. A. Murcko, *J. Med. Chem.* **1996**, *39*, 2887 – 2893.
[2] A. Schuffenhauer, P. Ertl, S. Roggo, S. Wetzel, M. A. Koch, H. Waldmann, *J. Chem. Inf. Model.* **2007**, *47*, 47 – 58.
[3] A. M. Clark, P. Labute, *J. Med. Chem.* **2009**, *52*, 469 – 483.
[4] S. Renner, W. A. L. Van Otterlo, M. D. Seoane, S. Möcklinghoff, B. Hoffmann, S. Wetzel, A. Schuffenhauer, P. Ertl, T. I. Oprea, D. Steinhilber, L. Brunsveld, D. Rauh, H. Waldmann, *Nat. Chem. Biol.* **2009**, *5*, 585 – 592.
[5] R. P. Sheridan, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1037 – 1050.
[6] J. Batista, J. Bajorath, *J. Chem. Inf. Model.* **2007**, *47*, 1405 – 1413.
[7] J. J. Sutherland, R. E. Higgs, I. Watson, M. Vieth, *J. Med. Chem.* **2008**, *51*, 2689 – 2700.
[8] E. Lounkine, J. Auer, J. Bajorath, *J. Med. Chem.* **2008**, *51*, 5342 – 5348.
[9] G. Müller, *Drug Discovery Today* **2003**, *8*, 681 – 691.
[10] D. M. Schnur, M. A. Hermsmeier, A. J. Tebben, *J. Med. Chem.* **2006**, *49*, 2000 – 2009.
[11] A. M. Aronov, B. McClain, C. S. Moody, M. A. Murcko, *J. Med. Chem.* **2008**, *51*, 1214 – 1222.
[12] D. A. Erlanson, R. S. McDowell, T. O'Brien, *J. Med. Chem.* **2004**, *47*, 3463 – 3482.
[13] P. J. Hajduk, J. Greer, *Nat. Rev. Drug Discovery* **2007**, *6*, 211 – 219.
[14] M. G. Siegel, M. Vieth, *Drug Discovery Today* **2007**, *12*, 71 – 79.
[15] M. Fischer, R. E. Hubbard, *Mol. Interventions* **2009**, *9*, 22 – 30.
[16] T. I. Oprea, *J. Comput. Aided Mol. Des.* **2000**, *14*, 251 – 264.
[17] I. Muegge, *Med. Res. Rev.* **2003**, *23*, 302 – 321.
[18] C. A. Lipinski, *Drug Discovery Today Technol.* **2004**, *1*, 337 – 341.
[19] T. I. Oprea, T. K. Allu, D. C. Fara, R. F. Rad, L. Ostopovici, C. G. Bologa, *J. Comput. Aided Mol. Des.* **2007**, *21*, 113 – 119.
[20] I. Kola, J. Landis, *Nat. Rev. Drug Discovery* **2004**, *3*, 711 – 716.
[21] P. D. Leeson, B. Springthorpe, *Nat. Rev. Drug Discovery* **2007**, *6*, 881 – 890.
[22] T. Liu, Y. Lin, X. Wen, R. N. Jorissen, M. K. Gilson, *Nucleic Acids Res.* **2007**, *35*, D198 – D201.
[23] MDL Drug Data Report (MDDR), Symyx Software: San Ramon, CA (USA) **2007**, http://www.symyx.com (accessed December 3, 2009).
[24] D. S. Wishart, C. Knox, A. C. Guo, D. Chen, S. Shrivastava, D. Tzur, B. Gautam, M. Hassanali, *Nucleic Acids Res.* **2008**, *36*, D901 – D906.
[25] Scitegic Pipeline Pilot, Accelrys, Inc.: San Diego, CA (USA) **2009**, http://accelrys.com/products/scitegic (accessed December 3, 2009).
[26] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, *Genome Res.* **2003**, *13*, 2498 – 2504.
[27] P. Willett, J. M. Barnard, G. M. Downs, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983 – 996.
[28] MACCS structural keys, Symyx Software: San Ramon, CA (USA) **2008**, http://www.symyx.com (accessed December 3, 2009).

Received: October 7, 2009
Revised: December 2, 2009
Published online on December 11, 2009

Summary

Structural features of compounds at different stages of pharmaceutical development were analyzed and compared on the basis of molecular scaffolds. The overlap between these scaffold sets was rather limited. Four subsets of overlapping scaffolds were assembled, which revealed to what extent compounds were likely to pass through different development stages. These subsets of scaffolds displayed significant inter- and intra- structural diversity. Furthermore, scaffolds with different frequencies at development stages were also identified, i.e. scaffolds that preferentially occurred in early- and/or late-stage compounds. These ensembles of scaffolds having different characteristics can be utilized as structural markers to analyze other active compounds.

Having analyzed scaffold distributions in different sets of active compounds, the next step has been to study the relationship between molecular selectivity and target families or individual targets at the level of molecular scaffolds. Specificaly, we aimed at examining whether chemical frameworks exist with inherent selectivity against certain target families.

Chapter 2

Systematic Analysis of Public Domain Compound Potency Data Identifies Selective Molecular Scaffolds Across Druggable Target Families

Introduction

The concept of *"privileged substructures"* has been a focal point in searching for fragments that are recurrent in and unique to ligands of a given target family for decades. Such target-class privileged chemotypes were usually identified on the basis of frequency of occurrence analysis of pre-selected substructures. Different from traditional case-by-case studies, we carried out a systematic selectivity profile analysis of public domain compounds and explored molecular scaffolds that were selective for given target families. More than 200 scaffolds were found in publicly available compounds that were active against only one community of closely related targets. The majority of these scaffolds displayed significant target-selective tendencies within a community. These scaffolds were found to be underrepresented in approved drugs.

752 J. Med. Chem. 2010, 53, 752–758
DOI: 10.1021/jm9014229

Journal of
Medicinal
Chemistry
Article

Systematic Analysis of Public Domain Compound Potency Data Identifies Selective Molecular Scaffolds across Druggable Target Families

Ye Hu, Anne Mai Wassermann, Eugen Lounkine, 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*

Received September 24, 2009

Molecular scaffolds that yield target family-selective compounds are of high interest in pharmaceutical research. There continues to be considerable debate in the field as to whether chemotypes with a priori selectivity for given target families and/or targets exist and how they might be identified. What do currently available data tell us? We present a systematic and comprehensive selectivity-centric analysis of public domain target−ligand interactions. More than 200 molecular scaffolds are identified in currently available active compounds that are selective for established target families. A subset of these scaffolds is found to produce compounds with high selectivity for individual targets among closely related ones. These scaffolds are currently underrepresented in approved drugs.

Introduction

Twenty years ago Evans et al.[1] first put forward the idea that chemotypes might exist that preferentially bind to a given target class, and the characterization of molecular scaffolds active against individual target classes has ever since been a topic of intense research in pharmaceutical settings.[2] The notion of "privileged substructures"[1] is highly attractive for drug discovery and chemical biology because they might ultimately be evolved into chemical entities that are selective for individual targets. However, it has been shown that substructures thought to be target class-characteristic typically also appeared in compounds active against other target families[3] and exclusive binding of known chemotypes to given target classes has not been confirmed to this date.

The concept of privileged substructures touches upon a much more general question in molecular probe and drug discovery, namely, how to generate small molecules that are selective for a target of interest within a target family.[4] Currently, only little is known about the relationship between molecular selectivity at the level of target families and individual targets[5] and it is not understood what the likelihood might be to discover selective compounds for different target classes.

Target selectivity (TS[a]) is typically explored on a case-by-case or family basis, and systematic analyses of compound selectivity data across different families are currently not available. With the growing availability of small molecule structure−activity data in the public domain, we are now in a position to explore molecular selectivity in a way that fundamentally differs from traditional case-by-case studies. This is accomplished by focusing, in an unbiased manner, on what data currently available for different target families might tell us about the selectivity of known molecular scaffolds and compounds. Such an analysis also provides a basis for the identification of new selective compounds.

To these ends, we have designed and carried out a systematic computational selectivity profile analysis of the BindingDB database,[6] a major public domain repository of activity information of small molecules, which we have found to represent by far the currently most comprehensive source of activity annotations that can be transformed into compound selectivity data. BindingDB contains ~31000 compound entries with ~57000 activity measurements taken from the scientific literature. Because of the ensuing high level of accuracy of the activity annotations, BindingDB is particularly suitable for a large-scale exploration of molecular selectivity. It represents an up-to-date view of the current scientific literature and knowledge in the field. The results of our analysis are reported herein and offer some surprising insights into the availability of target class-selective molecular scaffolds that might be evolved into target-selective compounds.

Results

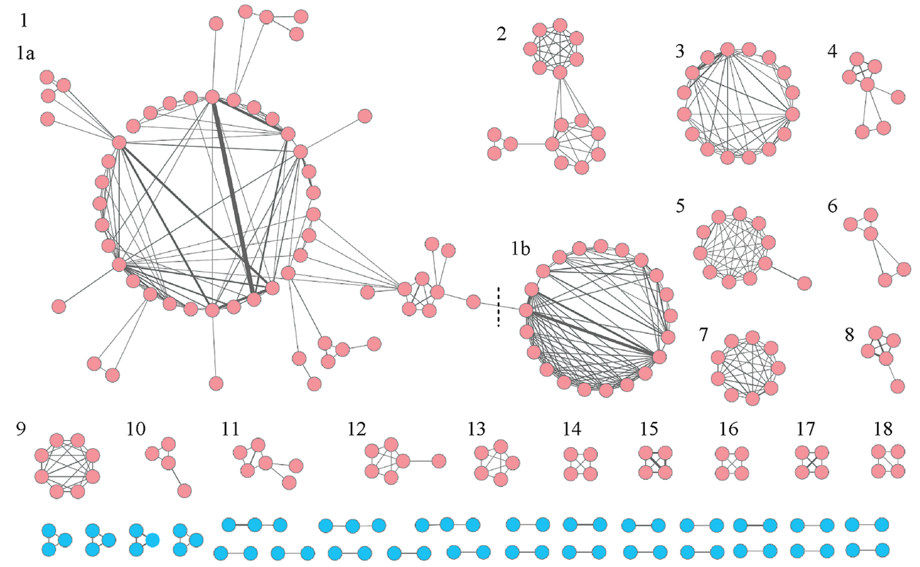
Compounds, Targets, and Selectivity Sets. A total of 6343 compounds active against 259 human targets (Supporting Information Table S1) were extracted from BindingDB. Many of these compounds were active against multiple targets, yielding a total of 17 929 compound−target combinations, and we identified 520 target pairs that shared at least five active compounds (with an average of 34 molecules per pair). For each molecule active against a target pair, its target selectivity was calculated as TS = $pK_i^A - pK_i^B$ (where pK_i^A and pK_i^B refer to the logarithmic potency value of the compound against targets A and B, respectively). Absolute TS values of selected compounds ranged from 0 to 6.86, i.e., from equal potency (and thus no selectivity) to potency differences of nearly 7 orders of magnitude (i.e., highest selectivity for one of two targets). Each pair of targets and the compounds they shared represented 1 of 520 selectivity sets for further analysis.

---

*To whom correspondence should be addressed. Phone: +49-228-2699-306. Fax: +49-228-2699-341. E-mail: bajorath@bit.uni-bonn.de.

[a] Abbreviations: TS, target selectivity.

Published on Web 12/09/2009 © 2009 American Chemical Society



1a. Tyrosine kinases; CytochromeP450 enzymes
1b. Carbonic anhydrases; Matrixmetalloproteinases
2. Protein kinases
3. Serine proteinases
4. GPCRs
5. Phosphodiesterases
6. Inhibitor of apoptosis proteins
7. Caspases
8. Dipeptidyl peptidases
9. Histone deacetylases
10. Protein-Tyrosine phosphatases
11. Purinergic receptors
12. Phosphoinositide 3-kinases
13. GABA$_A$ receptors
14. Sphingosine 1-Phosphate receptors
15. Cathepsins
16. Steroid receptors
17. Esterases
18. Polo-like kinases

Figure 1. Target pair network. Nodes represent targets, and edges are drawn between nodes if they share at least five compounds. The network representation reveals a total of 18 communities containing at least four targets. Community 1 is subdivided (dashed vertical line) on the basis of target family membership. Nodes in communities are colored light-red and others light-blue.

Target Pair Network and Target Communities. The 259 human targets participated in multiple target pairs, and a network representation was generated to analyze target relationships (Figure 1). In the network, nodes represent targets and edges are drawn between nodes if they share at least five molecules. This number of molecules was chosen to control network noise and ensure the reliability of selectivity profiling. The width of edges is scaled according to the number of active compounds shared by a target pair. The network reveals the presence of 18 separate and in part densely connected communities containing at least four targets (smaller communities were not considered). These communities are found to represent different target families (Figure 1). Thus, known biological activities of small molecules organize targets into functional families, as has been observed in drug–target networks based on chemical drug similarity.[7,8] For the purpose of our selectivity studies, network analysis was only required to organize and preselect target communities. The largest community identified in our network (community 1) contains 82 targets that mainly belong to three target families, i.e., tyrosine kinases, carbonic anhydrases (CAs), and matrix metalloproteinases (MMPs). Tyrosine kinases form a large subset (1a) on the left in Figure 1, while CAs and MMPs form a densely connected subset (1b) on the right (i.e., they share many active compounds). These two subsets are linked by cytochrome P450 enzymes and steroid sulfatase. By removal of the edge connecting steroid sulfatase and CA2, community 1 was divided into subsets 1a and 1b, hence producing a total of 19 communities for further analysis. These communities consisted of 4–59 targets and 8–2252 active compounds. Details for each community are provided in Supporting Information Figure S1 and Supporting Information Table S2.

Scaffolds and Selectivity Profiles. From the initial pool of 6343 active compounds, hierarchical molecular scaffolds[9] were isolated that represented at least five active compounds, yielding a total of 210 distinct scaffolds, listed in Supporting Information Table S3. For each target within a community with at least five ligands having the same scaffold, the active compounds were collected. The TS values for target pairs containing this target and the active compounds were calculated. The median of these TS values is an indicator of scaffold selectivity for the particular target. A high median TS value means that a scaffold shows high selectivity toward the target over other targets within the community. A negative median TS value indicates that the scaffold produces compounds that are selective for other members of the community. On the basis of median TS values, a scaffold–target heat map was generated to represent the *target selectivity profile* of each scaffold within a community. Furthermore, for each scaffold found in a community, all relevant compounds used in the generation of the target–scaffold heat map were pooled, and the median of their absolute TS values was calculated. In this case, high median values indicate that a scaffold produces many compounds with different potency against individual targets and hence a differentiated selectivity profile within a community. A scaffold–community heat map was also generated to represent the *community selectivity profile* of each scaffold. Supporting Information Figure S1 reports the number of scaffolds in each community. For two communities (6 and 13), no relevant scaffolds were found. For the other communities, the number of scaffolds ranged from 1 to 102. For individual targets, between 1 and 32 scaffolds were found.

Target and Community Selectivity of Scaffolds. The scaffold–target heat map for community 3 representing serine

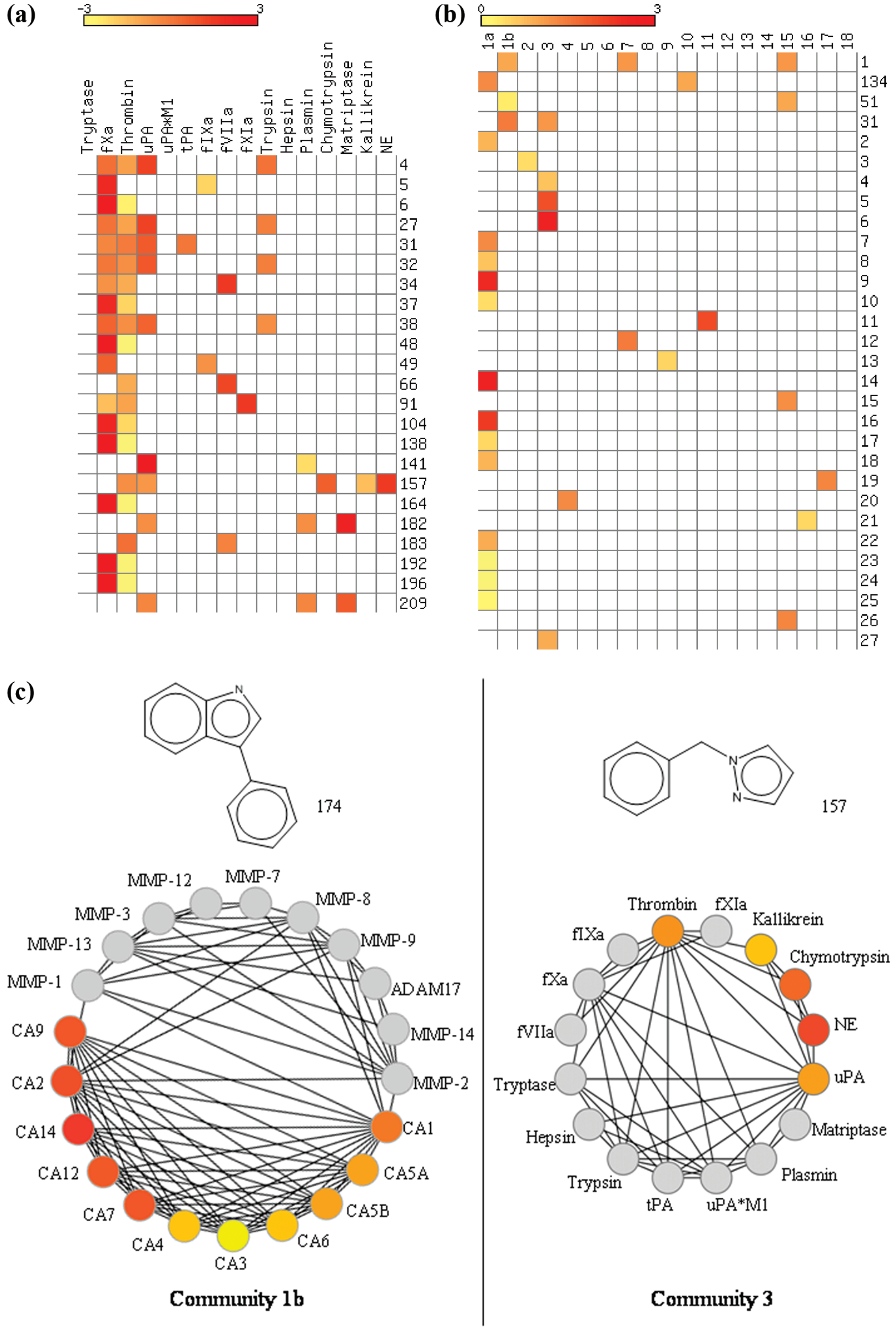


Figure 2. Target and community selectivity profiles. (a) The heat map representing the target selectivity profile of community 3 is shown. Targets form columns and scaffolds rows. A cell corresponding to a scaffold−target combination is filled if the scaffold is present in at least five compounds active against the target and color-coded according to median TS values. (b) A section of the community selectivity profiles is shown. Here, columns represent communities and rows scaffolds. Cells are color-coded according to absolute median TS values. (c) Shown are community-centric target selectivity profiles for two representative scaffolds (174 and 157) that are selective for communities 1b and 3, respectively. Nodes are color-coded by median TS values of active compounds according to part a. Thus, for targets with red nodes, the scaffold has highest potential to produce selective compounds. Targets for which fewer than five active compounds containing the scaffold exist are depicted as gray nodes. Edges between nodes are drawn according to Figure 1.

proteases is shown in Figure 2a as an example (Supporting Information Figure S2 shows the corresponding heat maps for all communities). Median TS values are represented via a continuous color spectrum ranging from −3 (yellow) to 3 (red). A key observation in Figure 2a is that individual scaffolds mostly display different selectivity against related targets, and this trend is observed for all communities (Supporting Information Figure S2). For example, scaffold 6 represents compounds that are active against factor Xa and thrombin but these inhibitors are much more potent against factor Xa and thus highly selective for this target. Similar observations are made for scaffolds 48, 104, 138, 164, 192, and 196, all of which differentiate between these two proteases. Other scaffolds represent compounds that inhibit proteases more broadly. For example, scaffold 157 represents inhibitors of five proteases. However, the compounds are more potent against neutrophil elastase than against the other targets. Supporting Information Figure S2 shows that selectivity-conferring scaffolds were found for many targets across all communities, and Supporting Information Table S4 lists the scaffolds that are most selective for individual targets. The number of scaffolds per target varies in part significantly, but for many targets only a single scaffold is found that yields selective compounds relative to the other targets of the communities.

Figure 2b shows a heat map representing the community selectivity profile of a subset of scaffolds (and Supporting Information Figure S3 shows the corresponding profiles for all 210 scaffolds). Here, median of absolute TS values are represented via a continuous color spectrum ranging from 0 (yellow) to 3 (red). A value of 0 means that the scaffold does not generate selective compounds across the community, and a value of 3 means that compounds containing the scaffold display at least a 1000-fold difference in potency against targets within the community. Figure 2c shows two representative examples of scaffolds that act on multiple targets within a community yielding substantial differences in compound selectivity. A key observation in Figure 2b is that only four scaffolds (1, 31, 51, and 134) are active against multiple communities. These scaffolds mainly correspond to compounds that are nonselective. By contrast, all other scaffolds are found to specifically act on only one community. However, these community-selective scaffolds display a distinctly different potential to yield target-selective compounds. Supporting Information Table S5 reports the potential of community-selective scaffolds to produce target-selective compounds. A total of 111 scaffolds display a target-selective tendency (median $|TS| \geq 1$), and 37 of these scaffolds represent compounds with at least 100-fold potency differences against other community targets.

Taken together, the results of the target and community selectivity profile analysis reveal that community-selective scaffolds are consistently found and that a subset of these scaffolds has in part significant potential to yield target-selective compounds within their communities. Figure 3 shows examples of scaffolds having high potential to produce target-selective compounds for major drug targets including, among others, receptor tyrosine kinases, G-protein-coupled receptors, or caspases.

Community-selective scaffolds can also be utilized to identify new target-selective compounds, as illustrated in Figure 4. For example, the community and target selectivity profiles suggest that compounds containing scaffold 37 should have high potential to produce inhibitors that are selective for factor Xa over thrombin. When a nonpublic domain database was searched,[10] two compounds containing this scaffold were identified that are currently not available in BindingDB and both of these compounds are indeed

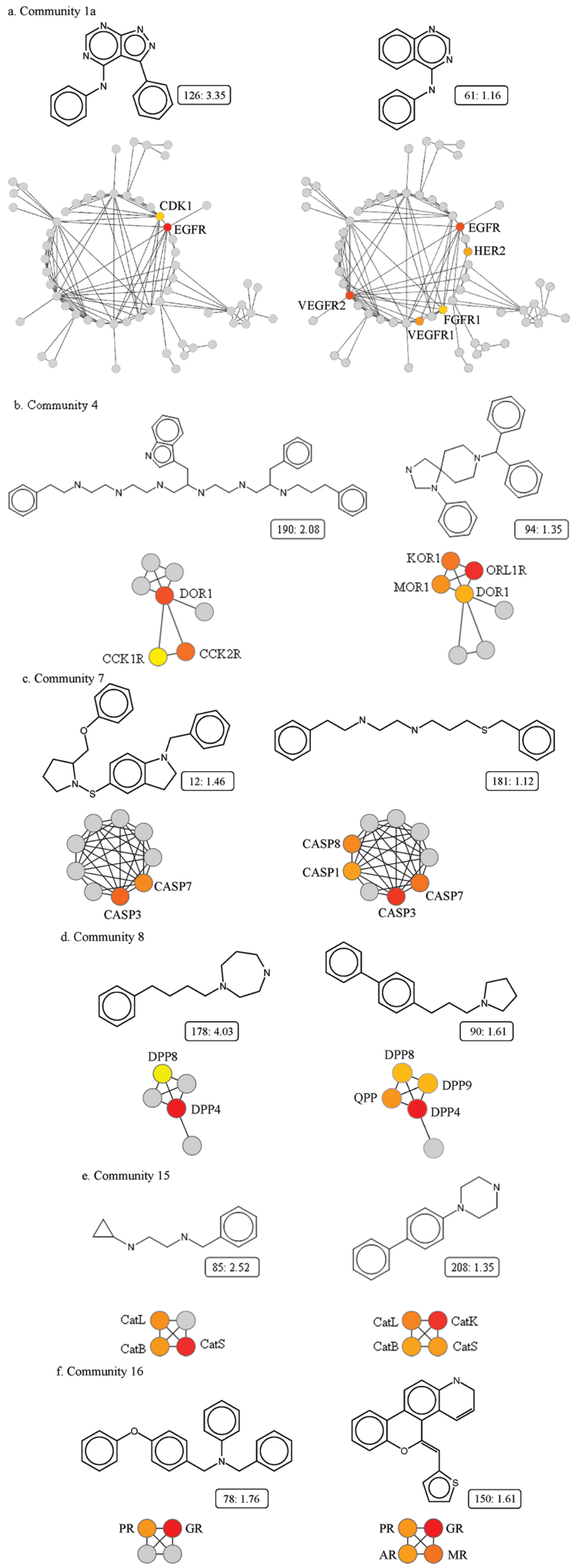


Figure 3. Community-selective scaffolds. For different target communities, selective scaffolds are shown that have high potential to yield target-selective compounds. Scaffolds have "scaffold number: median TS value" annotations. On the left of each figure, the scaffold with the highest median TS value in the community is shown. On the right, another scaffold with a broader selectivity profile is shown.

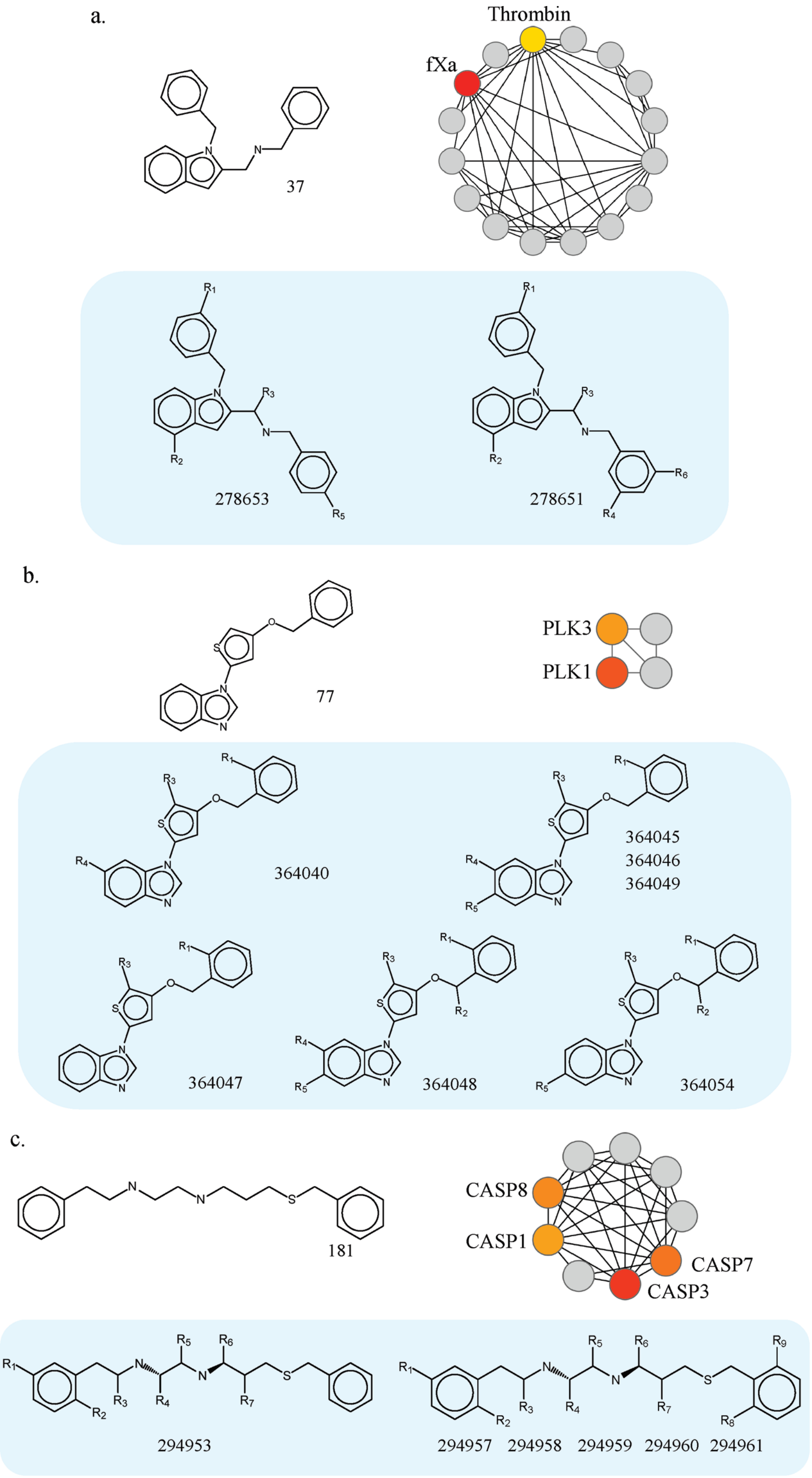


Figure 4. Searching for selective compounds. Examples of scaffolds (and their community selectivity profiles) are shown that were utilized to search the MDDR database. Compounds found to have the predicted selectivity are shown on a blue background. MDDR compounds are license-protected and therefore represented as Markush structures. Each Markush structure is annotated with MDDR identifiers of the compounds it represents.

reported to be highly selective for factor Xa (Figure 4a). Similarly, compounds were found containing scaffold 77 (Figure 4b) and 181 (Figure 4c) that were inhibitors of polo-like kinase 1 and caspase 3, respectively, with no reported activity against other community targets. The target selectivity profile for the caspase community also suggested that compounds containing scaffold 12 should have comparable potency for caspase 3 and 7 but not for other members of the caspase family. This prediction is confirmed by a recent study aiming at the development of isatin sulfonamides as caspase inhibitors.[11] Four compounds containing scaffold 12 were reported that inhibited both caspase 3 and 7 with nanomolar potency and were ~200-fold selective over caspases 1, 6, and 8.

Distribution of Community-Selective Scaffolds in Drugs. We have also determined the distribution of community-selective scaffolds in known drugs. Therefore, 1247 approved drugs were retrieved from DrugBank[12] and a total of 726 unique scaffolds were isolated from them. Only 11 of these drug scaffolds were also found within the set of 206 target community-selective scaffolds, illustrating that these scaffolds are currently underrepresented in approved drugs. Because a subset of community-selective scaffolds is target-selective, as discussed above, chemical exploration of these scaffolds might be expected to provide further opportunities for drug discovery.

Discussion

The focal point of our study has been the exploration of small molecule selectivity on a large scale. Ligand preferences of target families have thus far been explored by calculating the frequency of occurrence of selected substructures in compounds active against individual target families. Such statistical approaches are based on a binary formulation of biological activity (i.e., active vs inactive) and do not take selectivity into account. The approach reported herein is specifically focused on exploring the selectivity of active compounds at the level of molecular scaffolds. It is data-driven and does not employ any preconceived notions of structure-selectivity relationships or target family assignments. Rather, the target pair network provides a data structure to organize known targets into communities based on shared ligands. Moreover, community and target selectivity profiles make it possible to assign molecular scaffolds to communities and explore their potential to produce target-selective compounds. Key findings of our analysis include the following: more than 200 scaffolds exist in currently available public domain compounds that are selective for communities of closely related targets, and a majority of these scaffolds yield compounds that are either selective for individual targets or display a target-selective tendency. These scaffolds can also be utilized to search for other active compounds having a desired selectivity profile. Hence, currently available data suggest that a substantial molecular knowledge base exists to generate target class- or target-selective small molecular probes or leads. Because we focus on currently available activity data of small molecules, the scaffold and selectivity information we report should provide many alternative starting points for a further experimental evaluation of scaffold selectivity profiles and the chemical exploration of molecular selectivity.

Methods

In order to comprehensively cover public domain compound data that could be utilized to extract target selectivity information relevant for our analysis, we also analyzed bioassays available in Pubchem[13] as a potential source. Compound screens were analyzed for appropriate selectivity information. However, given the target pair criteria applied in our study, only three relevant target pairs could be identified in Pubchem. The results of our analysis are reported in Supporting Information Figure S4. From BindingDB, compounds with reported activity (IC_{50} and/or K_i values) against human targets were extracted. If multiple potency measurements were reported in a BindingDB entry, their geometric mean was calculated to yield a single potency value. For each molecule active against a target pair, its target selectivity (TS) was calculated as $TS = pK_i^A - pK_i^B$. TS and median TS values are simple, intuitive, and continuous measures of target selectivity that do not require the definition of selectivity thresholds. Conventional hierarchical scaffolds were derived according to Bemis and Murcko.[9] These scaffolds represent ring systems connected by linkers after removal of substituents. Compounds and scaffolds were recorded and processed in SMILES format.[14] The target pair network was generated using Cytoscape.[15] Target communities connected only by intra- but no intercommunity edges and comprising a minimum of four targets were isolated. Community- and target-selective scaffolds were searched in the MDDR database[10] and compared to scaffolds extracted from DrugBank.[12] Nonselective scaffolds were not further analyzed. The community- and target-based selectivity profile analysis was carried out with in-house generated Pipeline Pilot[16] and Perl programs.

Supporting Information Available: Tables S1−S5 listing all investigated targets, target communities, scaffolds, target-selective scaffolds, and community-selective scaffolds, respectively; Table S6 listing the results of scaffold overlap analysis between selectivity-conferring and current drug scaffolds; Figures S1−S3 showing target and scaffold distributions, target selectivity profiles, and community selectivity profiles, respectively; Figure S4 showing the results of target pair analysis in PubChem. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) 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.

(2) Horton, D. A.; Bourne, G. T.; Smythe, M. L. The Combinatorial Synthesis of Bicyclic Privileged Structure or Privileged Substructures. *Chem. Rev.* **2003**, *103*, 893–930.

(3) Schnur, D. M.; Hermsmeier, M. A.; Tebben, A. J. Are Target-Family-Privileged Substructures Truly Privileged? *J. Med. Chem.* **2006**, *39*, 2000–2009.

(4) Bajorath, J. Computational Analysis of Ligand Relationships within Target Families. *Curr. Opin. Chem. Biol.* **2008**, *12*, 352–358.

(5) Karaman, M. W.; Herrgard, S.; Treiber, D. K.; Gallant, P.; Atteridge, C. E.; Campbell, B. T.; Chan, K. W.; Ciceri, P.; Davis, M. I.; Edeen, P. T.; Faraoni, R.; Floyd, M.; Hunt, J. P.; Lockhart, D. J.; Milanov, Z. V.; Morrison, M. J.; Pallares, G.; Paterl, H. K.; Pritchard, S.; Wodicka, L. M.; Zarrinkar, P. P. A Quantitative Analysis of Kinase Inhibitor Selectivity. *Nat. Biotechnol.* **2008**, *26*, 127–132.

(6) 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.

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

(8) 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.

(9) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.

(10) *MDL Drug Data Report* (*MDDR*), version 2005.2; Symyx Software: San Ramon, CA, 2005.

(11) Smith, G.; Glaser, M.; Perumal, M.; Nguyen, Q. D.; Shan, B.; Arstad, E.; Aboagye, E. O. Design, Synthesis, and Biological Characterization of a Caspase 3/7 Selective Isatin Labeled with 2-[^{18}F]fluoroethylazide. *J. Med. Chem.* **2008**, *51*, 8057–8067.

(12) Wishart, D. S.; Knox, C.; Guo, A. C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M. DrugBank: A

Knowledgebase for Drugs, Drug Actions and Drug Targets. *Nucleic Acids Res.* **2008**, *36*, D901–D906.

(13) PubChem. http://pubchem.ncbi.nlm.nih.gov/ (accessed September 1, **2009**).

(14) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.

(15) 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.

(16) *Scitegic Pipeline Pilot, Student Edition*, version 6.1; Accelrys, Inc.: San Diego, CA, 2007.

Summary

We have presented a systematic and rather comprehensive selectivity profile analysis of publicly available active compounds at the level of molecular scaffolds. In this data-driven analysis, a compound-based network representation was utilized to organize targets into communities based on currently available target-ligand interactions. A total of 206 scaffolds were found to be specifically active against only one target community. Community and target selectivity profiles of each scaffold were analyzed in order to assess the target selective tendency. The majority of these community-selective scaffolds displayed a notable potential to produce compounds selective for certain target(s). Community-selective scaffolds could also be used to search for new selective compounds and provide further opportunities for chemical exploration.

In light of these findings, a logical follow-up question has been whether truly target-selective scaffolds exist that yielded compounds solely selective for one particular target against others. This question was investigated in the next study.

Chapter 3

Exploring Target-Selectivity Patterns of Molecular Scaffolds

Introduction

In the previous study, we have thoroughly analyzed selectivity profiles of public domain compounds and identified target class-selective molecular scaffolds. Here, we further refined the approach and explored the presence of target-selective scaffolds that would exclusively produce compounds selective for a particular target over others. Although currently available selectivity data is sparsely distributed, small sets of target-selective scaffolds were identified at different selectivity threshold levels. Furthermore, we also explored selectivity patterns formed by these target-selective scaffolds. These patterns and corresponding scaffolds can aid in the design of new compounds with desired target selectivity.

ACS Medicinal Chemistry Letters
pubs.acs.org/acsmedchemlett

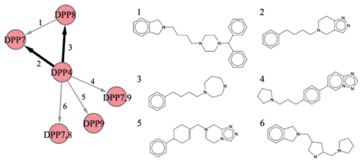
Exploring Target-Selectivity Patterns of Molecular Scaffolds

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 We investigate the question of whether target-selective molecular scaffolds can be identified on the basis of currently available compound activity data. Starting from a pool of 17745 public domain compounds with activity annotations for 433 human targets, we ultimately identify, through a selectivity classification and database-mining approach, 42 molecular scaffolds represented by multiple compounds that are highly selective for a particular target over one or more others. In many other cases, individual compounds representing unique scaffolds are target-selective. Hence, currently available public domain compound selectivity data are sparse. However, we also identify selectivity patterns that evolve around specific targets and are formed by multiple target-selective scaffolds. These scaffolds should provide interesting starting points for further chemical exploration.

KEYWORDS Molecular selectivity, privileged substructures, target family selective molecular scaffolds, target-selective scaffolds, selectivity patterns, compound database mining



In medicinal chemistry, "privileged substructures",[1] that is, chemotypes that bind with high preference to a family of targets, have been—and continue to be—intensely studied. In many instances, substructures considered to be target class-selective on the basis of frequency of occurrence analysis have also been detected in compounds active against other target families;[2] hence, the existence of truly privileged structural motifs has been controversial.[2]

Recently, we have carried out a large-scale analysis of public domain compound data to investigate whether target class-selective molecular scaffolds exist.[3] To avoid potential caveats of occurrence frequency-based analysis, we searched for compounds with multiple activity annotations and formed pairs of biological targets that were "connected" by at least five active compounds. This target pair information was then organized in a compound-based target network that enabled the identification of different target communities. From these compounds, conventional hierarchical scaffolds[4] were isolated, and scaffolds were determined that exclusively occurred in one of the target communities formed by the network. The approach is summarized on the left side in Figure 1.

For this target pair-based analysis, BindingDB[5] was found to be a comprehensive public domain source of bioactivity data. For example, by systematically analyzing currently available PubChem[6] confirmatory bioassays, only three target pairs were identified that met the selection criterion. Of 17745 compounds available in BindingDB with activity annotations against a total of 433 human targets, 6343 compounds active against 259 targets met our target pair selection criterion (i.e., five or more shared ligands), yielding a total 520 target pairs organized into 18 target communities. From these 6343 compounds, a total of 206 target community-selective scaffolds were identified, that is, scaffolds that only occurred in one of 18 communities (Figure 1). We also calculated a pairwise potency-based selectivity ratio for compounds representing these scaffolds, which indicated that a subset of these scaffolds had the potential to yield selective compounds, at least at the level of target pairs.[3]

In light of these findings, a logical follow-up question thus became if there might also be truly target-selective scaffolds present among community-selective ones. Target-selective scaffolds, that is, scaffolds that exclusively yield target-selective compounds, would be of high interest for medicinal chemistry research. Hence, we have investigated this question and report the results herein.

To make the analysis of target-selective scaffolds as comprehensive as possible, we decided to revise the previous target pair and scaffold selection approach, as illustrated on the right side in Figure 1. Therefore, we applied a more stringent target pair selection criterion by requiring not only at least five shared ligands but also at least five scaffolds representing shared ligands. A total of 220 human targets yielding 428 target pairs met these requirements and are reported in Table S1 of the Supporting Information with their BindingDB target IDs. This target pair information was

Received Date: December 21, 2009
Accepted Date: January 26, 2010
Published on Web Date: February 01, 2010

© 2010 American Chemical Society

DOI: 10.1021/ml900024v | *ACS Med. Chem. Lett.* **2010,** *1,* 54–58

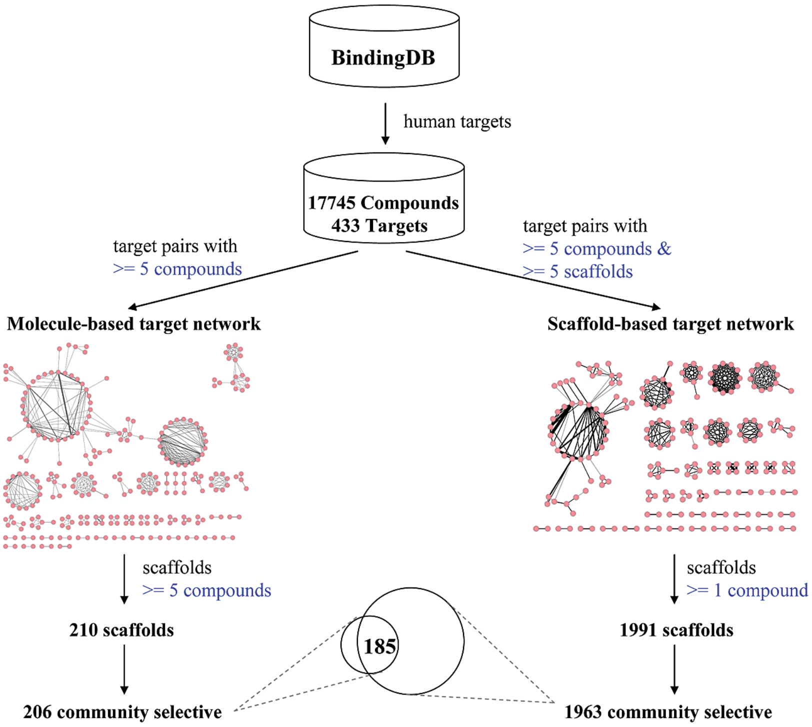


Figure 1. Target communities and community-selective scaffolds. Shown is an overview of alternative approaches to establish target communities on the basis of compound activity data and isolate community-selective scaffolds, which provide a basis for the identification of target-selective scaffolds.

then organized in a scaffold-based target network, as illustrated on the right in Figure 1. In this network, targets are connected if compounds active against them represent at least five scaffolds. We found that this scaffold-based target network further refined the formation of target communities as compared to the previous compound-based target network. In the scaffold-based network, 21 well-defined communities containing at least three targets were found (rather than 18). The target, compound, and scaffold composition of these 21 communities are reported in Table 1, and the scaffold-based network with target community annotations is shown in Figure S1 of the Supporting Information. After a more stringent target pair criterion was applied, we then relaxed the scaffold selection criterion by accepting any scaffold (and not only scaffolds represented by at least five compounds), which yielded a total of 1991 scaffolds, 1963 of which occurred in only one of 21 communities. These community-selective scaffolds were active against 174 targets in 405 target pairs and also included 185 of the 206 community-selective scaffolds previously identified from the compound-based target network (Figure 1, left side). The remaining 21 scaffolds occurred in more than one of the 21 communities in the scaffold-based network.

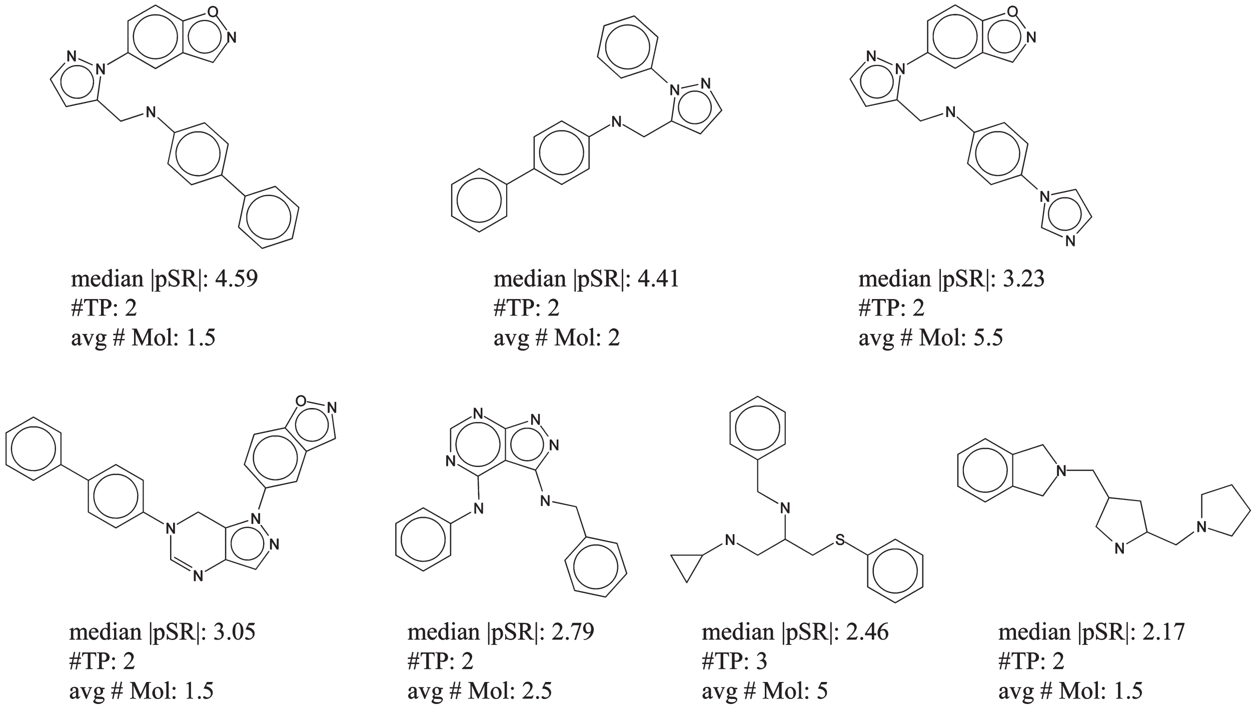
The 1963 community-selective scaffolds were then ranked on the basis of the median absolute selectivity ratio (|pSR|) of compounds that they represent for established target pairs. The absolute selectivity ratio of a compound for a target pair is simply given by the positive difference of its logarithmic potency values against the two targets. Accordingly, median |pSR| values ≥1 and ≥2 indicate that at least half of the compounds represented by a scaffold have at least a 10- and 100-fold potency difference for one target over another, respectively. Figure S2 of the Supporting Information shows the distribution of scaffolds over median |pSR| values, the number of compounds that they represent, and the target pairs that these compounds are active against. Of the 1963 community-selective scaffolds, 1026 scaffolds had a median |pSR| ≥ 1, and 329 scaffolds had a median |pSR| ≥ 2. Thus, a significant number of scaffolds corresponded to highly selective compounds. However, 1350 scaffolds were found to represent a single compound, 1049 scaffolds were found to be active against a single target pair, and 785 scaffolds were found to correspond to both a single molecule and a target pair. Thus, this distribution reflects a notable degree of data incompleteness, which generally affects the systematic analysis of target−ligand interactions.[7] Hence, when more compounds representing individual scaffolds and more measurements become available, the number of selective scaffolds is expected to decrease. However, among the 329 highly selective scaffolds with median |pSR| ≥ 2, there were also 50 scaffolds that represented multiple compounds active against multiple target pairs (Figure S2 of the Supporting Information), which represented particularly interesting scaffolds for further analysis.

Community-selective scaffolds were further classified according to different selectivity threshold levels of the compounds that they represent, that is, at least 10-, 50-, or 100-fold selectivity. The classification scheme is illustrated in

Table 1. Composition of Target Communities[a]

		no. of			
community	target family	targets	target pairs	compounds	scaffolds
1	tyrosine kinases and cytochrome P450 enzymes	50	100	2128	782
2	serine proteinases	12	34	545	229
3	protein kinase C	8	22	72	34
4	carbonic anhydrases	11	55	327	87
5	phosphodiesterases	11	39	117	47
6	matrix metalloproteinases	10	24	187	56
7	protein kinase B and serine protein kinases	6	11	109	78
8	caspases	9	31	114	49
9	histone deacetylases	8	22	121	68
10	purinergic receptors	6	7	107	54
11	phosphoinositide 3-kinases (PI3Ks)	6	10	46	26
12	GABAA receptors	5	9	8	7
13	opioid receptors	4	6	84	27
14	cathepsins	4	6	307	152
15	dipeptidyl peptidases	4	6	287	105
16	esterases	4	6	238	110
17	polo-like kinases	4	5	35	21
18	sphingosine 1-phosphate (S1P) receptors	3	3	20	9
19	peroxisome proliferator-activated receptors	3	3	61	16
20	steroid receptors	3	3	35	9
21	β-secretases and cathepsin D	3	3	127	66

[a] Target communities extracted from the scaffold-based target network are characterized by the number and nature of the targets and, in addition, by the number of compounds active against pairs of targets and the corresponding scaffolds.



median |pSR|: 4.59
#TP: 2
avg # Mol: 1.5

median |pSR|: 4.41
#TP: 2
avg # Mol: 2

median |pSR|: 3.23
#TP: 2
avg # Mol: 5.5

median |pSR|: 3.05
#TP: 2
avg # Mol: 1.5

median |pSR|: 2.79
#TP: 2
avg # Mol: 2.5

median |pSR|: 2.46
#TP: 3
avg # Mol: 5

median |pSR|: 2.17
#TP: 2
avg # Mol: 1.5

Figure 2. Scaffolds contained in highly selective compounds. Seven scaffolds are shown for which corresponding compounds had median |pSR| ≥ 2 and for which each compound was 50-fold selective for a target over another. For each scaffold, the median median |pSR| value is reported as well as the number of target pairs in which it occurs and the average number of molecules per pair.

Table 2. Target-Selective Scaffolds[a]

scaffolds		no. of scaffolds	no. of targets	no. of target pairs
community-selective		1963	174	405
target-selective	10-fold	**472**	83	66
	50-fold	**250**	65	43
	100-fold	**191**	58	38

[a] Reported is the number of target-selective scaffolds (bold) represented by one or more compounds at different selectivity levels. In addition, the corresponding numbers of targets and target pairs for all community- and target-selective scaffolds are also reported.

Figure S3 of the Supporting Information, and further details are provided in the Methods of the Supporting Information. If a scaffold was always selective for a target over one or more others (in different pairs), it was termed "purely" selective (i.e., a scaffold can be purely selective for more than one target). For the 10-, 50-, and 100-fold selectivity levels, a total of 499, 252, and 191 purely selective scaffolds were identified, respectively. These scaffold sets were compared to the 50 scaffolds with median $|pSR| \geq 2$ that represent multiple compounds active against multiple target pairs (Figure S4 of the Supporting Information), revealing an overlap of 11 (10-fold), 7 (50-fold), and 3 (100-fold) scaffolds, respectively. Figure 2 shows the seven purely selective scaffolds for the 50-fold selectivity level. These scaffolds and the corresponding compounds are provided in Table S2 of the Supporting Information.

Having found that community-selective scaffolds had rather different distributions and selectivity profiles, we searched for target-selective scaffolds among the purely selective ones. We considered a scaffold target-selective if it was selective for an individual target over one or more others. Because complex pairwise selectivity relationships can exist for scaffolds in multiple target pairs, the identification of target-selective scaffolds can be complicated. Hence, it was facilitated through a directed graph type method illustrated in Figure S5 of the Supporting Information. Details are provided in Methods of the Supporting Information. In Table 2, the number of target-selective scaffolds is reported. For the 10-, 50-, and 100-fold selectivity levels, 472, 250, and 191 target-selective scaffolds were identified. Hence, most purely selective scaffolds were also target-selective scaffolds. For the 100-fold selectivity level, 149 of 191 target-selective scaffolds only corresponded to a single compound selective for one target over one or two others. The remaining 42 scaffolds were represented by 2−21 compounds and were selective for an individual target over one or two others. These scaffolds are displayed in Figure S6 of the Supporting Information and their target annotations are provided.

Going beyond target selectivity of individual scaffolds, we also asked the question of which target relationships, or selectivity patterns, might be formed by target-selective scaffolds. Therefore, we analyzed the three sets of selective scaffolds reported in Table 2. For the 10-, 50-, and 100-fold selectivity levels, 28 (50), 18 (31), and 19 (23) well-defined target relationships were formed by single (multiple) scaffolds, respectively. As shown for the 50-fold selectivity level in Figure 3a, these relationships can be viewed in a directed target network where nodes (targets) are connected by directed edges if they share one or more target-selective scaffolds. In this case, all scaffolds correspond to selective compounds, and the directionality of the edges indicates target (A over B) selectivity. In addition, the width of the edges is scaled according to the number of target-selective scaffolds. In Figure 3a, different selectivity patterns are observed. Figure S7 of the Supporting Information shows the corresponding networks for the 10- and 100-fold selectivity levels where similar observations can be made. As shown in Figure 3a, in addition to binary selectivity relationships, there are inverse relationships (where some scaffolds are selective for target A over B and others for B over A) and

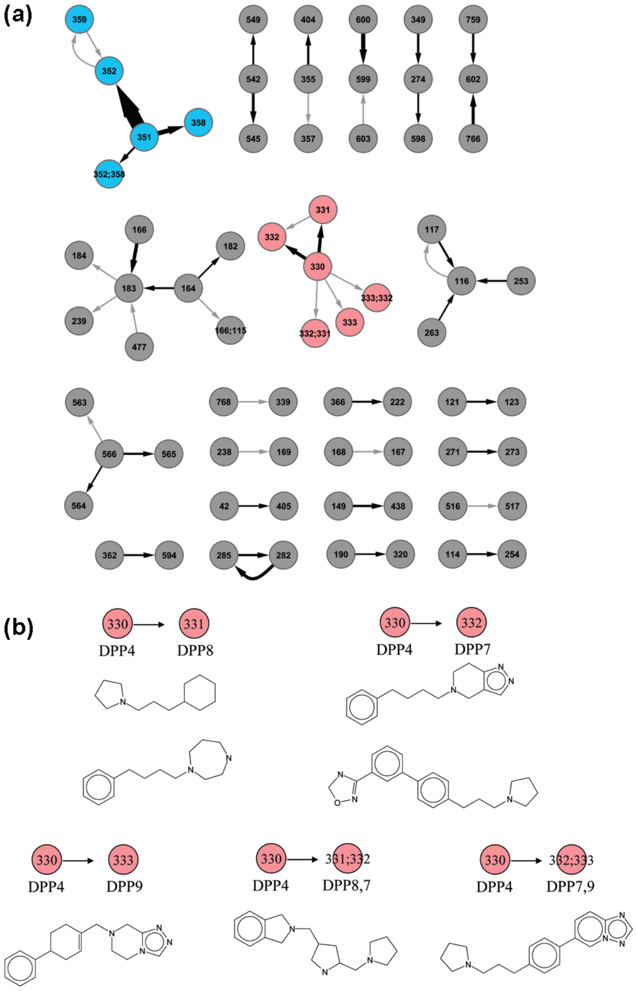


Figure 3. Selectivity patterns. (a) The directed target network for the 50-fold selectivity level is shown displaying different scaffold-based target relationships. The width of directed edges is scaled according to scaffold numbers. When a relationship is formed by a single scaffold, the edge is shown in gray. (b) Scaffolds that yield compounds selective for DDP4 over other DDPs, corresponding to the target cluster with pink nodes in panel a. The two relationships at the top are formed by 10 and 11 scaffolds, respectively, and two representative scaffolds are shown in each case. The three selectivity relationships at the bottom each involve a single scaffold.

also complex selectivity patterns. In addition, "selectivity hubs" become apparent, that is, individual targets with scaffold selectivity over several others. For example, the cluster formed by blue nodes at the upper left in Figure 3a represents selectivity relationships among the closely related serine proteases factor Xa (target ID 351), thrombin (352), and factor IXa (358) where multiple scaffolds generate compounds that are at least 50-fold selective for factor Xa over the other two proteases. Moreover, the cluster of pink nodes in the center corresponds to closely related dipeptidyl peptidases (DPPs) where single or multiple scaffolds are selective for DPP4 over related DPPs or pairs of DPPs. Figure 3b shows seven representative scaffolds that produce compounds selective for DDP4 over other DDPs and the selectivity relationships that they form. These scaffolds and the corresponding compounds are provided in Table S3 of the Supporting Information. Such scaffolds can be collected as starting points for generating compounds that are highly selective for a particular target over other closely related ones.

In summary, systematic mining of a publicly available compound data has revealed that small sets of target-selective scaffolds represented by multiple compounds exist, although selectivity data are sparsely distributed. These target-selective scaffolds are represented by up to 21 compounds that are highly selective for an individual target over one or two others. However, the majority of currently available target-selective scaffolds (at different selectivity levels) are only represented by individual compounds. Thus, many scaffolds are available for further experimental evaluation that might yield target-selective compounds. Importantly, however, selectivity patterns can be observed around specific targets that are formed by multiple target-selective scaffolds and establish different target relationships, which can also be exploited in the design of target-selective compounds.

EXPERIMENTAL PROCEDURES

From BindingDB,[5] compounds with reported activity against human targets were extracted. If multiple potency measurements were reported in a BindingDB entry, their geometric mean was calculated as the final single potency value. Hierarchical scaffolds[4] were extracted from active compounds that represent ring systems and rings connected by linkers after removal of substituents. Compounds and scaffolds were represented in SMILES format[8] for processing. Network representations were generated with Cytoscape.[9] The method to determine target-selective scaffolds and the selectivity level assignments of scaffolds are detailed in the Methods of the Supporting Information. Scaffold and target selectivity analysis were carried out using in-house Pipeline Pilot[10] and Perl programs. These programs are described in the Methods of the Supporting Information and are available via the following URL: http://www.lifescienceinformatics.uni-bonn.de ("Downloads").

SUPPORTING INFORMATION AVAILABLE

Details of scaffold selectivity analysis (Methods); tables reporting all targets investigated in this study (Table S1) and the scaffolds and corresponding compounds that are discussed (Tables S2 and S3); and figures presenting the annotated scaffold-based target network (Figure S1), the distribution of community-selective scaffolds (Figure S2), the selectivity-based scaffold classification scheme (Figure S3), the distribution of purely selective scaffolds (Figure S4), the methodology applied to identify scaffolds with exclusive target selectivity (Figure S5), the structures of target-selective scaffolds (Figure S6), and the network representations of selectivity patterns at different selectivity levels (Figure S7). This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author: *To whom correspondence should be addressed. Tel: +49-228-2699-306. Fax: +49-228-2699-341. E-mail: bajorath@bit.uni-bonn.de.

REFERENCES

(1) 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.

(2) Schnur, D. M.; Hermsmeier, M. A.; Tebben, A. J. Are Target-Family-Privileged Substructures Truly Privileged?. *J. Med. Chem.* **2006,** *39*, 2000–2009.

(3) 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.

(4) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996,** *39*, 2887–2893.

(5) 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.

(6) PubChem. http://pubchem.ncbi.nlm.nih.gov/ (accessed September 1, 2009).

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

(8) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988,** *28*, 31–36.

(9) 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.

(10) *Scitegic Pipeline Pilot*, Student Edition, Version 6.1; Accelrys, Inc.: San Diego, CA, 2007.

© 2010 American Chemical Society

Summary

From an initial pool of 17,745 compounds with activity annotations reported for 433 human targets, target-selective scaffolds that exclusively yielded compounds selective for one target over one or two others at different selectivity levels were identified. However, most scaffolds were only represented by a single compound, which reflected a high degree of data sparseness. Hence, only small sets of these target-selective scaffolds were represented by multiple compounds. Moreover, selectivity patterns formed by these scaffolds were derived, which support the design of compounds selective for a given target over other closely related ones.

In the previous two associated studies (*Chapter 2 and 3*), community- or target-selective scaffolds were identified representing by compounds active against at least two targets. In the next study, we asked the question whether scaffolds also exist that are promiscuous in nature, rather than target-selective. Such polypharmacological behavior at the level of molecular scaffolds had so far not been systematically explored. Therefore, we designed an analysis to identify promiscuous chemotypes across different targets or target families.

Chapter 4

Polypharmacology Directed Compound Data Mining: Identification of Promiscuous Chemotypes

Introduction

Although the conventional paradigm of target specificity has dominated in drug discovery over the past decades, it has also been shown that many drugs act against multiple targets, rather than a single target. Therefore, systematical exploration of such polypharmacology is of particularly high interest. Here we designed a large-scale data mining approach to analyze activity annotations of public domain compounds at three hierarchical levels of abstraction, i.e. active compounds, hierarchical scaffolds, and carbon skeletons. A target family classification scheme was applied to identify promiscuous scaffolds that represented compounds active against targets in multiple families. For these scaffolds, target family relationships were analyzed to evaluate their degree of promiscuity. In addition, activity profiles were also compared and scaffolds were mapped to approved drugs.

2112 *J. Chem. Inf. Model.* **2010,** *50,* 2112–2118

Polypharmacology Directed Compound Data Mining: Identification of Promiscuous Chemotypes with Different Activity Profiles and Comparison to Approved Drugs

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

Received September 13, 2010

Increasing evidence that many pharmaceutically relevant compounds elicit their effects through binding to multiple targets, so-called polypharmacology, is beginning to change conventional drug discovery and design strategies. In light of this paradigm shift, we have mined publicly available compound and bioactivity data for promiscuous chemotypes. For this purpose, a hierarchy of active compounds, atomic property based scaffolds, and unique molecular topologies were generated, and activity annotations were analyzed using this framework. Starting from ~35 000 compounds active against human targets with at least 1 μM potency, 33 chemotypes with distinct topology were identified that represented molecules active against at least 3 different target families. Network representations were utilized to study scaffold−target family relationships and activity profiles of scaffolds corresponding to promiscuous chemotypes. A subset of promiscuous chemotypes displayed a significant enrichment in drugs over bioactive compounds. A total of 190 drugs were identified that had on average only 2 known target annotations but belonged to the 7 most promiscuous chemotypes that were active against 8−15 target families. These drugs should be attractive candidates for polypharmacological profiling.

INTRODUCTION

A single-target focus has traditionally dominated compound optimization efforts in medicinal chemistry, and a high degree of target specificity is usually considered a hallmark of drug candidates. However, in recent years, there has been increasing evidence of polypharmacological drug behavior.[1−4] It has been shown that many known drugs elicit their therapeutic effects by acting on multiple targets,[1−4] with protein kinase inhibitors being an extreme and well-studied example.[5,6] Hence, polypharmacology is beginning to be regarded as a general principle in drug discovery that influences compound design, optimization, and evaluation.[6−9]

Given this increasing focus on polypharmacology, we have been interested in exploring promiscuous chemotypes in compounds active against currently available targets. In previous compound data mining exercises, we have extended the concept of ‘privileged substructures’[10] by identifying target community selective molecular scaffolds,[11] studied activity cliff formation by selected structural classes,[12] and established topological and potency relationships between bioactive scaffolds.[13]

In addition to target (class) selectivity of active compounds, leading to the identification of community selective molecular scaffolds,[11] the potential promiscuity of chemotypes is another important topic for compound design. In a survey of the relevant literature, only one conceptually related study was found.[14] In this investigation, Cases and Mestres collected 214 drug targets implicated in cardiovascular diseases and studied polypharmacological relationships of ligands active against these targets. In the context of this analysis, the authors also extracted scaffolds from ligands of cardiovascular targets and determined the five most promiscuous scaffolds. In addition, scaffolds isolated from polypharmacological compounds were utilized to establish a relationship between molecular weight and cardiovascular promiscuity.[14]

Herein, we report the results of a large-scale data mining effort designed to identify highly promiscuous chemotypes on the basis of 19 target families. In addition, we have analyzed the activity profiles associated with these structures and studied their distribution in approved drugs.

MATERIALS AND METHODS

Compounds active against human targets with at least 1 μM potency were extracted from two major public domain repositories, ChEMBLdb (CDB)[15] and BindingDB (BDB).[16] These compounds were pooled and organized into target sets. Only target sets containing at least 10 active compounds were further considered. Targets were organized into 19 target families following the CDB classification scheme,[15] which contained between 3 and 130 individual targets, as summarized in Table 1. From all target set compounds, atomic property based Bemis and Murcko (B-M) scaffolds[17] were isolated. These scaffolds were obtained by removing all substituents from compounds and retaining only ring systems and linkers between them. B-M scaffolds were then transformed into carbon skeletons (CSKs) by converting all bond orders to one and all atom types to carbon. CSKs correspond to graph-based B-M scaffolds.[17] B-M scaffolds and CSKs are illustrated in Figure 1. We have selected B-M scaffolds and CSK representations for our analysis because they represent a straightforward structural hierarchy.

---

* Address correspondence to E-mail: bajorath@bit.uni-bonn.de. Telephone: +49-228-2699-306.

10.1021/ci1003637 © 2010 American Chemical Society
Published on Web 11/12/2010

Table 1. Target Families[a]

family ID	family	targets
1	Tyr protein kinase	31
2	Ser_Thr protein kinase	47
3	Ser_Thr_Tyr protein kinase	12
4	phosphadiesterase	9
5	protein phosphatase	3
6	aspartic protease	6
7	cysteine protease	10
8	metallo protease	20
9	serine protease	26
10	carbonic anhydrase	12
11	histone deacetylases	8
12	cytochromeP450 enzyme	13
13	transferase	5
14	ion channel	17
15	GPCR	130
16	cytosolic other	9
17	electrochemical transporter	14
18	nuclear receptor	17
19	other	69

[a] Nineteen target families were assembled following the CHEMBLdb classification scheme. For each family, the number of targets is reported.

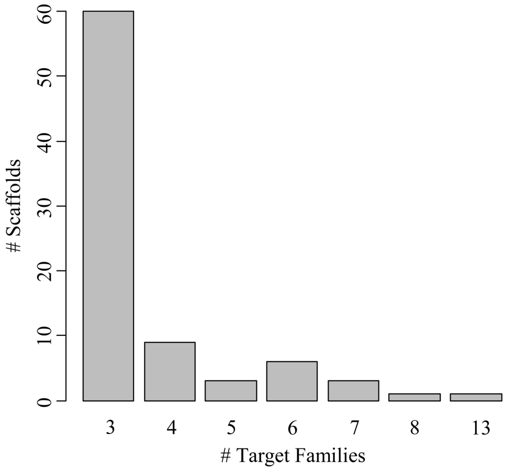


Figure 2. Scaffolds with multiple activities. Shown is the distribution of 83 B-M scaffolds that are active against 3 or more target families.

Target family nonspecific (promiscuous) scaffolds were determined, and their target-based activity profiles were analyzed. Promiscuous scaffolds were mapped to approved drugs available in DrugBank.[18] Our data mining effort did not involve predictive model building and was hence not amenable to external validation. Furthermore, it should be noted that the scaffold promiscuity mining we present does not involve a conventional analysis of structure−activity relationships. All calculations required for our analysis were carried out with in-house generated Scientific Vector Language (SVL)[19] or Perl scripts and Pipeline Pilot[20] programs. CDB and BDB compounds, B-M scaffolds, and CSKs were stored as SMILES strings.[21] Scaffold−target family networks were drawn with Cytoscape.[22]

RESULTS AND DISCUSSION

Molecular Promiscuity Analysis. In order to study bioactivity promiscuity we analyzed activity annotations of public domain compounds at three levels of abstraction, including active compounds with target annotations, atomic property based B-M scaffolds, and corresponding carbon skeletons. These levels represent a hierarchy: Different active compounds yield the same B-M scaffold and different scaffolds the same CSK. Importantly, unique CSKs represent different molecular topologies, and our ultimate goal has been to identify topologically distinct CSKs that represent promiscuous scaffolds and compounds. Here topologically distinct CSKs are considered general 'chemotypes'. Promis-

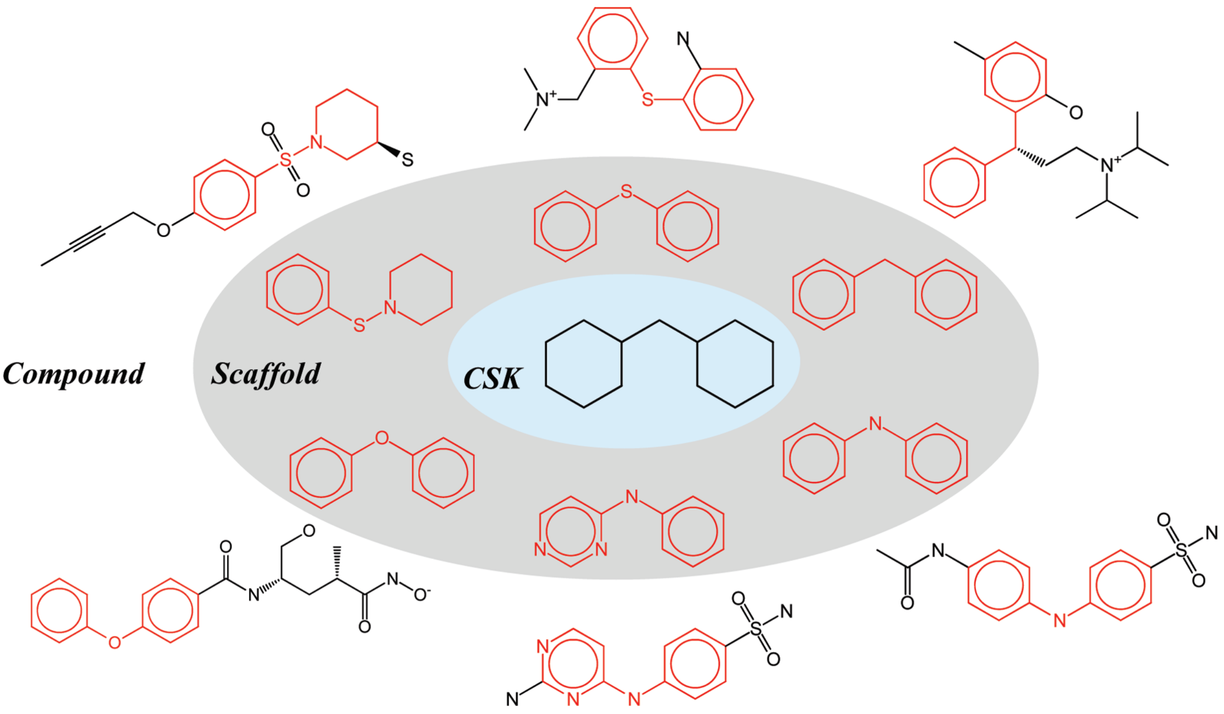


Figure 1. Generation of scaffolds and carbon skeletons. Shown are three layers of structural representations, i.e., compound (outer), scaffold (middle), and carbon skeleton (CSK; inner). B-M scaffolds in compounds are generated by only retaining ring systems and linkers between them (highlighted in red). All six different B-M scaffolds correspond to the same CSK.

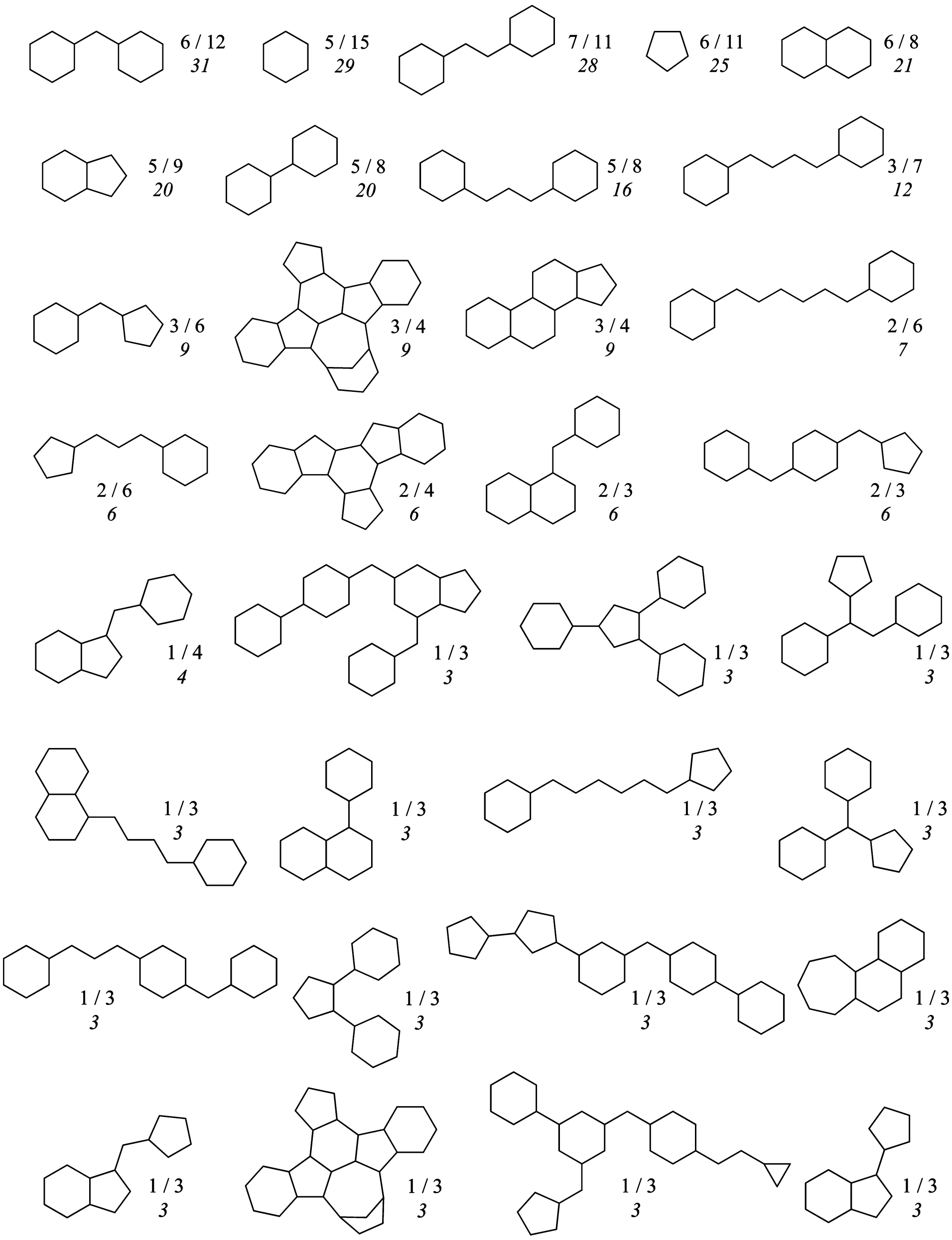


Figure 3. Promiscuous chemotypes. The set of 33 topologically distinct CSKs covering 83 promiscuous scaffolds is shown. For each CSK, the number of its B-M scaffolds and the number of target families these scaffolds are active against are reported. For example, "6/12" means that the CSK covers 6 promiscuous scaffolds that are active against 12 target families. Below this annotation, the total number of scaffold−target family relationships is reported for each CSK.

cuous chemotypes have a high potential to display polypharmacological behavior. The knowledge of their structures is useful for drug design. For example, promiscuous chemotypes can be selected for polypharmacological applications. However, promiscuous structural classes might also be avoided if target specificity is desirable.

For our analysis, we collected compounds active against human targets with at least 1 μM potency and organized them into target sets that had to contain a minimum of 10 compounds for further consideration. A compound active against multiple targets was a member of multiple target sets. B-M scaffolds and CSKs were derived from these compounds. To each scaffold, the activity annotations of the compounds it represented were assigned, and to each CSK, the activity annotations of the scaffolds it covered hence providing a hierarchical analysis frame.

On the basis of our selection criteria, a total of 34 906 CDB and BDB compounds active against 458 human targets were obtained that yielded 13 462 unique B-M scaffolds. These 458 targets were divided into 19 families according to Table 1.

Promiscuous Scaffolds. Initially, we searched for scaffolds that represented compounds active against targets in multiple families. A total of 435 B-M scaffolds were found with activity against targets in at least 2 different families. Of these 435 scaffolds, 83 were active against 3 or more target families, ranging from 3 to 13, as shown in Figure 2. Thus, there was a significant decline in the number of active scaffolds proceeding from 2 to 3 or more target families. Therefore, we considered scaffolds promiscuous that were active against at least three target families. These 83 promiscuous scaffolds corresponded to 33 topologically distinct CSKs, shown in Figure 3, each of which covered between 1 and 7 unique B-M scaffolds. Figure 3 reveals that these CSKs were of very different chemical complexity and represented rather different topologies. The CSKs ranged

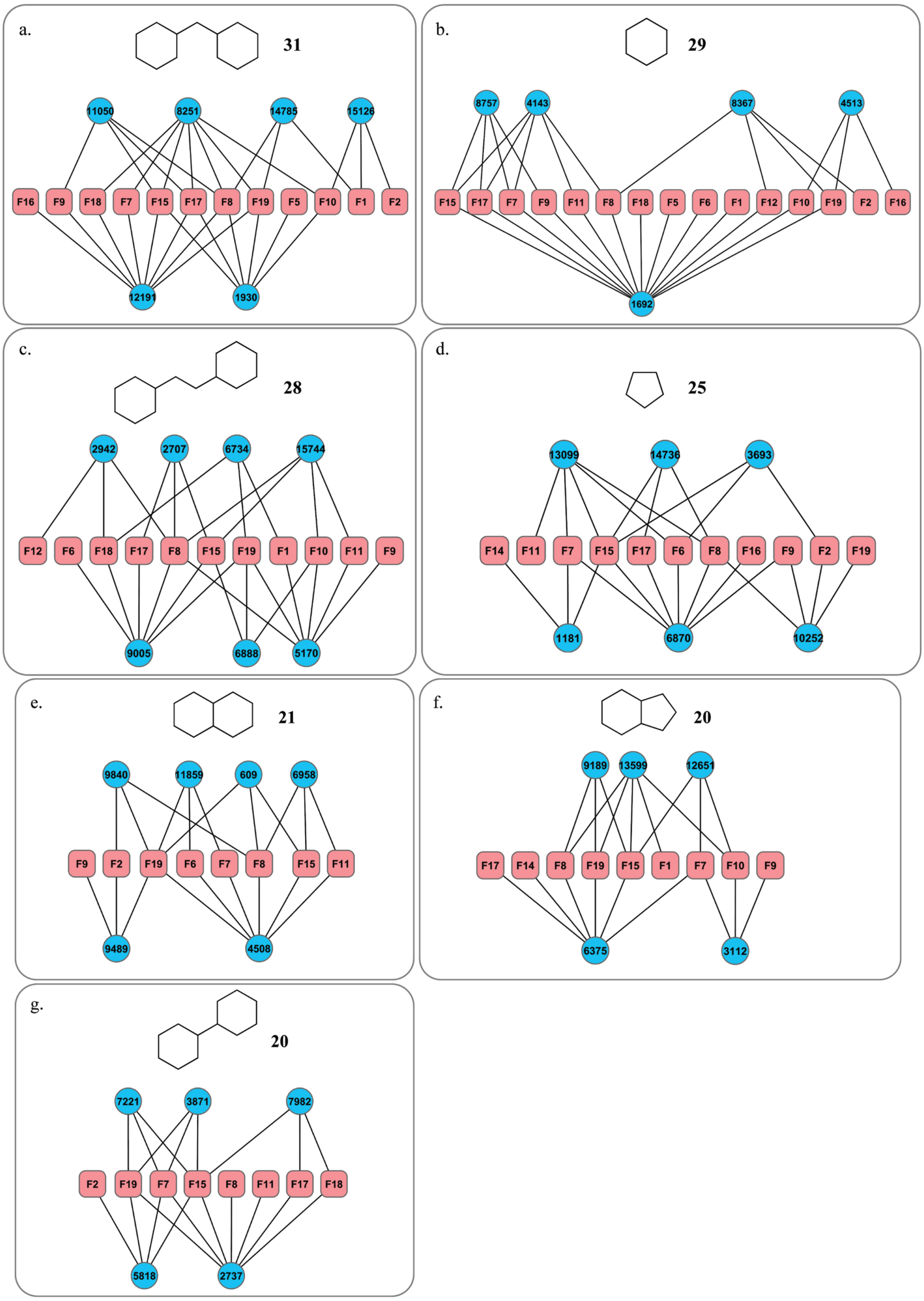


Figure 4. Scaffold-target family networks. In a−g, scaffold−target family relationships are displayed in a network representation for B-M scaffolds of each of the 7 most promiscuous chemotypes. At the top of each figure, the CSK structure is shown, and the total number of scaffold-family relationships is reported (bold). Circular nodes represent B-M scaffolds (labeled with scaffold IDs), and rectangular nodes represent target families (labeled with family IDs). An edge connects a scaffold and a family if compounds containing the scaffold are active against target(s) of this family.

from simple five- and six-membered rings to highly condensated ring systems and flexible structures containing multiple rings in diverse topological arrangements. Thus, promiscuity was clearly not limited to chemotypes of low complexity.

Promiscuous Chemotypes. For each of these 33 CSKs, the number of activity relationships formed between its B-M scaffolds and members of the 19 target families was determined (e.g., an individual B-M scaffold with compounds active against three target families accounted for three relationships). For each CSK, the number of its scaffold−target family relationships is also reported in Figure 3. 7 CSKs displayed at least 20 scaffold−family relationships and were the most promiscuous chemotypes we identified.

For each of these 7 CSKs, their scaffold−family relationships were analyzed in detail in a network representation,

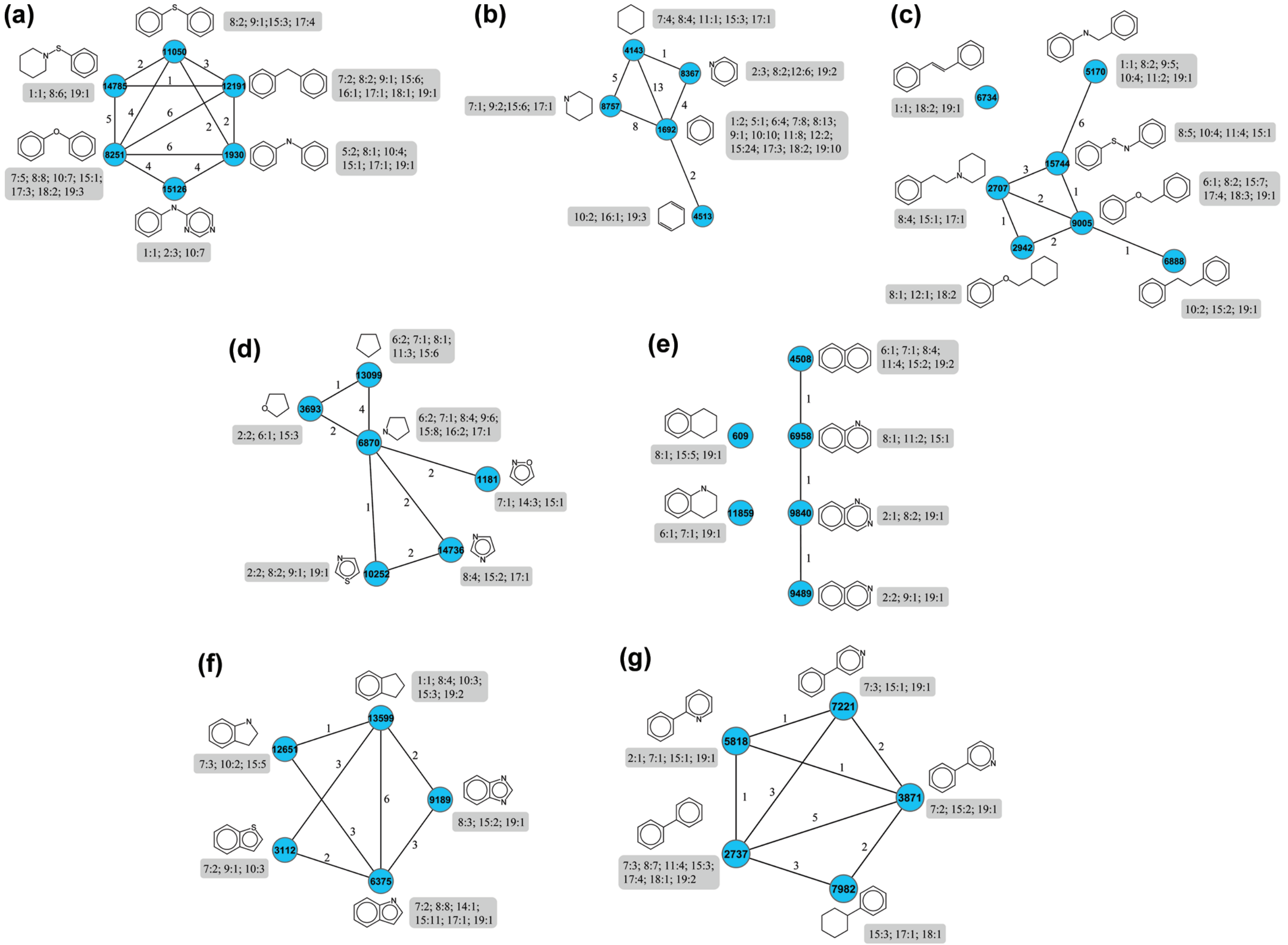


Figure 5. Activity profiles. In a−g, activity profiles for B-M scaffolds covered by each of the 7 most promiscuous chemotypes are displayed in a network representation. Nodes represent scaffolds, and an edge connects two scaffolds if they are active against the same target(s). Edge labels report the number of shared targets. The structure of each B-M scaffold is shown and annotated with its activity profile consisting of "target family:number of relevant targets" expressions. For example, the scaffold at the bottom in figure a has the activity profile "1:1; 2:3; 10:7". This means that this scaffold is present in compounds that are active against 1 target in target family 1, 3 targets in family 2, and 7 targets in family 10. In a−g, the sequence of CSKs corresponds to Figure 4.

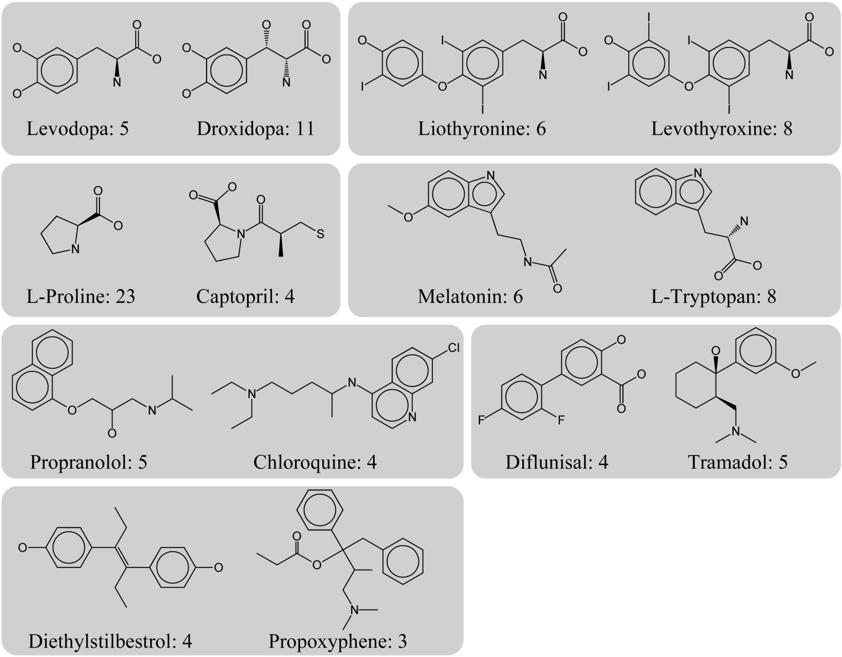
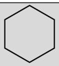
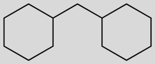
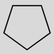
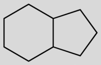
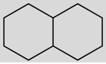
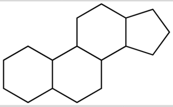
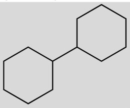
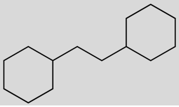
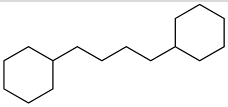
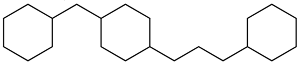
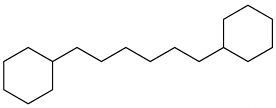
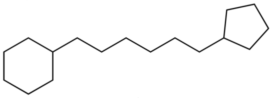
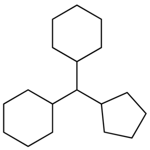
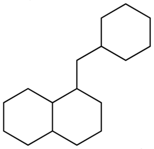
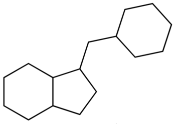
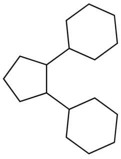
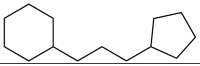


Figure 6. Drugs represented by promiscuous CSKs. For each of the 7 most promiscuous CSKs, two representative drugs are shown. Drug names and the number of annotated targets are reported.

Table 2. Promiscuous Chemotypes in Drugs[a]

CSK	Scaffolds	Drugs	Targets
	4	119	161
	4	25	48
	4	16	33
	5	8	31
	4	12	18
	3	11	17
	2	4	13
	3	6	10
	1	2	7
	1	1	5
	1	2	5
	1	2	3
	1	1	3
	2	2	3
	1	1	2
	1	1	2
	1	2	2

[a] Seventeen promiscuous CSKs that occur in both bioactive compounds and drugs are ranked according to the number of drug target annotations. For each CSK, the number of corresponding drugs, drug scaffolds, and target annotations are reported. There is no overlap between drugs assigned to different CSKs, i.e., the assignments are unique. The 7 most promiscuous scaffolds from bioactive compounds are shown on a grey background.

as shown in Figure 4. In these graphs, circular nodes represent unique B-M scaffolds corresponding to a given CSK and rectangular nodes represent target families. An edge connects a scaffold and a family if compounds represented by the scaffold were active against target(s) of this family. The total number of relationships per CSK ranged from 20 to 31. The comparison of CSK networks revealed that scaffolds corresponding to each of the most promiscuous CSKs generally formed rather different relationships. However, only for the CSK shown in Figure 4g, two scaffolds (3871 and 7221) displayed the same scaffold−family relationships. For all other CSKs, scaffolds were involved in only partly overlapping or distinct relationships. In addition, the degree of promiscuity of scaffolds corresponding to the same CSK also varied, from 3 to 13 target families per scaffold. Consequently, there was substantial target family coverage by scaffolds of promiscuous chemotypes, ranging from 8 to 15 target families per CSK.

Activity Profiles. We next analyzed the target activity profiles of B-M scaffolds representing each of the 7 most promiscuous CSKs. These profiles were generated by collecting the target annotations of active compounds containing each scaffold and assigning them to the scaffold. The results are shown in Figure 5. Most B-M scaffolds of a promiscuous chemotype were chemically very similar, often only distinguished by a single heteroatom substitution. Scaffold pairs covered by promiscuous CSKs shared varying numbers of targets. Moreover, Figure 5 reveals the presence of in part strikingly different activity profiles for closely related scaffolds. For example, this can be observed for biphenyl thioether and related scaffolds (Figure 5a), cyclohexane, pyrimidine, piperidine, cyclohexane, and cyclohexadiene (Figure 5b) or naphthalene, quinoline, and related scaffolds (Figure 4e). Regardless of their chemical complexity, scaffolds representing each of the promiscuous CSKs were found to display different target activity profiles, even if differences between these scaffolds were only subtle. Compounds containing these scaffolds had different or only in part overlapping bioactivities and displayed different degrees of promiscuity, which is also evident in Figure 5. Hence, the activity profiles were highly differentiated, and these findings further corroborate significant degree of target family coverage by promiscuous chemotypes.

Promiscuous Chemotypes in Drugs. Based on the results of our systematic analysis of bioactive compounds, we then searched for promiscuous scaffolds and chemotypes in current drugs. For this purpose, we utilized the set of 83 unique B-M scaffolds that were active against 3 or more target families and mapped these scaffolds to 1247 approved drugs taken from DrugBank. We found that a subset of 39 of these scaffolds was present in a total of 215 drugs. Thus, promiscuous scaffolds from bioactive compounds were present in ~17% of approved drugs. By contrast, these 83 scaffolds were only present in ~6% of the bioactive compounds we analyzed. Therefore, the proportion of promiscuous chemical entities was much higher in drugs than in bioactive compounds, although only a subset of these scaffolds occurred in drugs. These 39 scaffolds corresponded to 17 distinct CSKs. In Table 2, these 17 CSKs are ranked according to the number of reported drug targets (for the drugs they represented). The 7 most promiscuous CSKs we identified in bioactive compounds also occurred most frequently in drugs. Only the steroid skeleton, ranked sixth in Table 2, was not part of this set but had more assigned drug targets than two of the 7 most promiscuous CSKs. For each of the 7 most promiscuous CSKs, two representative drugs are shown in Figure 6. Each of the remaining nine CSKs had fewer target annotations and was only found in one or two drugs. For drugs containing the 7 most promiscuous CSKs, the average drug target-to-drug ratio was ∼2.2. However, as reported above, the most promiscuous chemotypes covered CDB and BDB compounds that were active against targets from 8 to 15 different families. Thus, these findings suggest that drugs corresponding to these 7 CSKs might be more polypharmacological in nature than it appears on the basis of their current drug target annotations. A total of 190 drugs were covered by the 7 most promiscuous CSKs (Table S1, Supporting Information). These drugs are thought to be good candidates for experimental polypharmacological profiling. The target family relationships for promiscuous CSKs and the corresponding scaffolds reported in Figure 4 can be used as guidelines to prioritize target families for a further analysis of the polypharmacological behavior of these drugs.

CONCLUSIONS

In this study, we have systematically searched currently available bioactive compounds for promiscuous structural classes. A total of 458 targets belonging to 19 target families provided the basis for our analysis. Promiscuity was explored at the level of active compounds, atomic property based scaffolds, and carbon skeletons (topologically distinct chemotypes). A total of 83 scaffolds and 33 chemotypes were found to be active against 3 or more target families. Similar scaffolds typically displayed very different target family relationships and activity profiles, resulting in broad target family coverage among promiscuous chemotypes. Subtle chemical differences among scaffolds of promiscuous chemotypes were often accompanied by significant changes in activity profiles. The 7 most promiscuous chemotypes were found to be active against 8−15 different target families. Seventeen promiscuous chemotypes covering 39 unique scaffolds were also found in 17% of approved drugs, whereas all 33 promiscuous chemotypes covering 83 scaffolds only occurred in 6% of the bioactive compounds, hence revealing a clear enrichment of a subset of promiscuous chemotypes in drugs. Moreover, 190 drugs with on average only 2 known target annotations were found to belong to the 7 most promiscuous bioactive chemotypes, suggesting that these drugs might display a higher degree of polypharmacology than is currently known.

Supporting Information Available: Table S1 reports 190 approved drugs that are likely to exhibit a high degree of polypharmacology. This information is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) 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.

(2) 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.

(3) Hopkins, A. L. Network Pharmacology: the Next Paradigm in Drug Discovery. *Nat. Chem. Biol.* **2008**, *4*, 682–690.

(4) Keiser, M. J.; Setola, V.; Laggner, C.; Abbas, A. I.; Hufeisen, S. J.; Jensen, N. H.; Kuijer, M. B.; Matos, R. C.; Tran, T. B.; Whaley, R.; Glennon, R. A.; Hert, J.; Thomas, K. L.; Edwards, D. D.; Shoichet, B. K.; Roth, B. L. Predicting New Molecular Targets for Known Drugs. *Nature* **2009**, *462*, 175–181.

(5) Karaman, M. W.; Herrgard, S.; Treiber, D. K.; Gallant, P.; Atteridge, C. E.; Campbell, B. T.; Chan, K. W.; Ciceri, P.; Davis, M. I.; Edeen, P. T.; Faraoni, R.; Floyd, M.; Hunt, J. P.; Lockhart, D. J.; Milanov, Z. V.; Morrison, M. J.; Pallares, G.; Paterl, H. K.; Pritchard, S.; Wodicka, L. M.; Zarrinkar, P. P. A Quantitative Analysis of Kinase Inhibitor Selectivity. *Nat. Biotechnol.* **2008**, *26*, 127–132.

(6) Morphy, R. Selectively Nonselective Kinase Inhibition: Striking the Right Balance. *J. Med. Chem.* **2010**, *53*, 1413–1437.

(7) Bajorath, J. Computational Analysis of Ligand Relationships within Target Families. *Curr. Opin. Chem. Biol.* **2008**, *12*, 352–358.

(8) Mestres, J.; Gregori-Puigjané, E. Conciliating Binding Efficiency and Polypharmacology. *Trends Pharmacol. Sci.* **2009**, *30*, 470–474.

(9) 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.

(10) 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.

(11) Hu, Y.; Wassermann, A. M.; Lounkine, E.; Bajorath, J. Systematic Analysis of Public Domain Compound Potency Data Identifies Selective Molecular Scaffolds across Druggable Target Families. *J. Med. Chem.* **2010**, *53*, 752–758.

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

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

(14) Cases, M.; Mestres, J. A Chemogenomic Approach to Drug Discovery: Focus on Cardiovascular Diseases. *Drug Discovery Today* **2009**, *14*, 479–485.

(15) *ChEMBLdb*; European Bioinformatics Institute (EBI): Cambridge, U.K., 2010; http://www.ebi.ac.uk/chembl/. Accessed May 11, 2010.

(16) 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.

(17) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.

(18) Wishart, D. S.; Knox, C.; Guo, A. C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M. DrugBank: A Knowledgebase for Drugs, Drug Actions and Drug Targets. *Nucleic Acids Res.* **2008**, *36*, D901–D906.

(19) *MOE (Molecular Operating Environment)*; Chemical Computing Group Inc.: Montreal Quebec, Canada, 2007.

(20) *Pipeline Pilot*, student ed., version 6.1; Accelrys, Inc.: San Diego, CA, 2007.

(21) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.

(22) 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.

CI1003637

Summary

A total of 34,906 compounds active against human targets with at least 1 μM potency were organized into 458 targets from 19 families and yielded 13,462 scaffolds. Of these scaffolds, 83 were found to be active against targets from three to thirteen families. These promiscuous scaffolds corresponded to 33 topologically distinct carbon skeletons (CSKs) of different structural complexity. Seven of 33 CSKs were involved in at least 20 scaffold-target family relationships and thus identified as the most promiscuous chemotypes. The scaffold-target family network we designed indicated that scaffolds covering the same promiscuous chemotype mostly displayed rather different activity profiles, although structural differences between them were often quite subtle. Thirty-nine promiscuous scaffolds were also found to be present in ~17% of approved drugs. These findings were consistent with the observation that many drugs elicit their therapeutic effects by binding to multiple targets.

In addition to target selectivity and promiscuity analysis, the potency distribution of compounds representing the same scaffold active against multiple targets has thus far not been investigated on a large scale. For example, it would be of considerable interest to explore whether scaffolds can be found that display a general tendency to produce compounds forming activity cliffs, i.e. structurally similar compounds having significant difference in potency. Thus far, activity cliffs have only been studied for sets of compounds active against a given target. Therefore, in the next study, we searched for molecular scaffolds yielding compounds that formed activity cliffs across different targets.

Chapter 5

Molecular Scaffolds with High Propensity to Form Multi-Target Activity Cliffs

Introduction

Activity cliffs are formed by structurally similar compounds with large potency differences and are a focal point of SAR analysis. We further extended the concept to compounds representing a given scaffold and analyzed whether scaffolds exist that might have a high propensity to form cliffs against multiple targets. We systematically analyzed compound activity data in two major public repositories and identified a number of scaffolds that were represented by multiple compounds forming activity or selectivity cliffs against multiple targets. These findings provide useful information for chemical optimization efforts.

Molecular Scaffolds with High Propensity to Form Multi-Target Activity Cliffs

Ye Hu and Jürgen Bajorath*

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

Received February 8, 2010

In target-dependent activity landscapes of compound series, cliffs are formed by pairs of molecules that are structurally analogous but display significant differences in potency. The detection and analysis of such activity cliffs is a major task in structure−activity relationship analysis and compound optimization. In analogy to activity cliffs, selectivity cliffs can be defined that are formed by structural analogs having significantly different potencies against two targets. The formation of activity cliffs by analogs is generally a consequence of different R-group patterns; e.g., a specific substitution of a given scaffold might increase and another substitution decrease potency. Therefore, activity (or selectivity) cliffs are typically analyzed for a given scaffold representing an analog series, and it has thus far not been explored whether certain scaffolds might display a general tendency to yield compounds forming activity cliffs against different targets. We have exhaustively analyzed scaffolds and associated compound activity data in the ChemblDB and BindingDB databases in order to compare the availability of target-selective scaffolds in these databases and determine whether multi-target activity and multi-target selectivity cliff scaffolds exist. Perhaps unexpectedly, we have identified 143 scaffolds that are represented by multiple compounds and form activity or selectivity cliffs against different targets. These scaffolds have varying chemical complexities and are in part promiscuous binders (i.e., compounds containing these scaffolds bind to distantly related or unrelated targets). However, analogs derived from these scaffolds form steep activity cliffs against different targets. A catalog of scaffolds with high propensity to form activity or selectivity cliffs against multiple targets is provided to help identify potentially promiscuous candidate scaffolds during compound optimization efforts.

INTRODUCTION

Molecular scaffolds (core structures) are of high interest in pharmaceutical research as building blocks or markers of drug-like compounds.[1−5] Scaffolds are often defined in different ways, which makes it difficult to assess and compare studies that explore scaffold distributions or scaffold hopping.[3] For example, scaffolds might be systematically derived by breaking predefined bonds in compounds following a hierarchy or on the basis of retro-synthetic criteria,[6] i.e., by separating groups in molecules according to chemical reactions carried out to synthesize them. Organic ring systems have thus far been a major focal point of scaffold analysis and design.[7,8] However, scaffolds have been analyzed from rather different points of view. For example, scaffold distributions have been determined for screening libraries,[9] large databases of synthetic molecules,[10] or compounds at different pharmaceutical development stages.[11] Frequency analysis has typically been applied to identify molecular scaffolds that are recurrent in synthetic molecules[12] or in compounds active against different targets.[13] In addition, attempts have also been made to systematically organize scaffold populations derived from active compounds on the basis of structural and activity criteria and thereby establish scaffold systems and hierarchies.[14,15] Scaffold analysis has also received much attention in the context of fragment-based drug discovery[16−18] where small weakly active compounds are combined in order to generate potent leads.

In addition to the scaffold analysis schemes described above, the high interest in “privileged substructures”[19] thought to preferentially bind to a given target class has triggered intense scaffold analysis efforts.[20−22] For the evaluation of privileged substructures, frequency analysis has also been carried out to compare the occurrence of proposed privileged substructures in different compound activity classes.[22] Although evidence for a preferential enrichment of certain scaffolds in specific activity classes has been accumulating,[21] the existence of truly privileged substructures has remained controversial.[22]

In order to thoroughly explore the presence of target class-selective molecular scaffolds beyond frequency calculations, we have previously carried out a large-scale analysis of public domain compounds with multiple activity annotations.[23] In this study, target communities were defined via a compound-based target network where individual targets were connected if they shared at least five active compounds. For different target communities, active compounds were collected, and the community selectivity of scaffolds derived from them was explored. This analysis has led to the identification of a total of 206 scaffolds that were selective for one of 18 target communities.[23] On the basis of these findings, we subsequently also searched for target-selective, rather than target class-selective, scaffolds. For this purpose, we modified the network analysis approach and introduced a scaffold-based target network where targets were connected if they shared at least five “active” scaffolds (rather than compounds).[24] Ultimately, we identified 42 scaffolds, each

---

* To whom correspondence should be addressed. Tel.: +49-228-2699-306. Fax: +49-228-2699-341. E-mail: bajorath@bit.uni-bonn.de.

10.1021/ci100059q © 2010 American Chemical Society
Published on Web 04/02/2010

of which was represented by multiple compounds that were highly selective (at least 100-fold) for a given target over one or more others.[24] However, we also found that currently available selectivity data were in general only sparse. For example, many scaffolds that formally met the criteria for target selectivity were only represented by individual compounds. Our scaffold analyses were based on public domain compound data available in BindingDB (BDB),[25] a major source of activity information of small molecules, in addition to PubChem.[26] From BDB data, a total of 520 pairs of human targets were identified that shared at least five ligands.[23] By contrast, in PubChem confirmatory bioassays, only three target pairs could be identified that met our selection criterion.

Herein, we address as of yet unexplored questions in scaffold analysis, namely, whether scaffolds exist that have a high propensity to introduce “activity cliffs”[27] in structure−activity landscapes. Activity cliffs are formed by structurally similar compounds having large differences in potency and are the major source of SAR discontinuity,[28] which provides opportunities for compound optimization but often hinders QSAR modeling and activity predictions.[27,28]

It is indeed difficult to speculate about the question of whether scaffolds might have significantly different abilities to form activity cliffs. This is the case because strong activity cliffs are formed by close structural analogs with large potency differences, i.e., compounds that usually contain the same scaffold. Therefore, the formation of target-specific activity cliffs is primarily attributed to different substitution patterns of the same or very similar scaffolds, and it has thus far not been investigated whether scaffolds might exist that have an intrinsic ability to form activity cliffs across different targets. This analysis can also be extended to study the relationship between scaffolds and “selectivity cliffs” that are formed by structurally similar compounds having different selectivities against two targets.[29]

To investigate these questions, we have systematically explored relationships between active compounds, corresponding scaffolds, activity cliffs, and selectivity cliffs. Recently, the ChemblDB (CDB) database[30] has become available as another major public domain source of compound activity data, in addition to PubChem and BDB. In contrast to PubChem bioassays, we found sufficient CDB compound data for meaningful scaffold-based target network analysis and exploration of target-selective scaffolds. Therefore, we have carried out this analysis also for CDB to compare the results with BDB. The scaffold information extracted from these databases has then been utilized to search for scaffolds forming activity (and selectivity) cliffs for multiple targets.

METHODS

Scaffold Definition and Extraction. Scaffolds were derived from active compounds according to Bemis and Murcko.[1] Following this approach, scaffolds are isolated from synthetic compounds by removing R groups from ring systems but retaining linkers between rings.[1] Thus, these hierarchical scaffolds comprise individual, condensated, or linked ring systems. The Bemis and Murcko approach has been a major origin of systematic analyses of scaffold distributions in drugs.[1] It should also be noted that the information required for ligand-centric scaffold distribution analysis is exclusively obtained from two-dimensional molecular graphs. Hence, aspects of three-dimensional structure or protein−ligand interactions do not play a role in the context of this analysis.

BDB compounds with reported activity against human targets were collected. For compounds with multiple potency measurements against the same target, the geometric mean was calculated as the final potency value. From CDB, compounds active against human targets were selected that had the highest target confidence level (CDB target confidence score 9) for direct interactions (target relationship type “D”). From all selected BDB and CDB compounds, scaffolds were isolated and represented as SMILES strings[31] for further analysis.

Scaffold-Based Target Network. For BDB and CDB scaffolds, separate scaffold-based target networks were generated for comparison. Target pairs were formed if two targets shared multiple active compounds representing at least five unique scaffolds. In such networks, targets are represented as nodes and connected by an edge if the target pair criterion is met. The width of edges is scaled according to the number of shared scaffolds. Network representations were drawn with Cytoscape.[32]

Community- and Target-Selective Scaffolds. Scaffolds that exclusively occurred in compounds active against only one of *n* target communities in the network were determined and termed “community-selective” scaffolds. For each compound active against a target pair, its selectivity ratio (SR) was calculated as follows:

$$\mathrm{SR} = \mathrm{pot}_{\mathrm{A}}(i) - \mathrm{pot}_{\mathrm{B}}(i)$$

Here, $\mathrm{pot}_A(i)$ and $\mathrm{pot}_B(i)$ represent the negative logarithm of potency values of compound i for targets A and B, respectively. Compounds and corresponding scaffolds were classified according to different selectivity levels, i.e., corresponding to at least a 10-fold, 50-fold, or 100-fold potency difference. If all compounds representing a unique scaffold were found to be selective for one particular target over one or more others (e.g., selective for A over B, A over C, etc.), the scaffold was classified as “target-selective”.

Classification of Scaffolds. Scaffolds were further classified according to three different criteria reflecting the formation of activity or selectivity cliffs by compounds derived from these scaffolds.

Compound Potency Value Ranges. For each scaffold, the potency range of its compounds against each target was determined, and the maximum potency was recorded. If a scaffold was represented by a single compound, the potency interval was set to 0. If multiple compounds existed that displayed the same potency, the scaffold was assigned to interval [0, 1], i.e., 0 to 1 order of magnitude difference in potency. Scaffolds were assigned to a total of six potency intervals, i.e., 0, [0, 1], [1, 2], [2, 3], [3, 4], and [4, Max]. Max designates the highest potency value range. For example, a scaffold was assigned to interval [2, 3] if the potency range of its compounds spanned 2 to 3 orders of magnitude. For each potency interval, the total number of scaffolds, average number of unique compounds per scaffold, average number of targets, and average maximum potency were calculated.

Compound Selectivity Ratio Ranges. In analogy to compound potency-based scaffold classification, selectivity ratio ranges for target pairs were determined for compounds representing each scaffold and corresponding selectivity

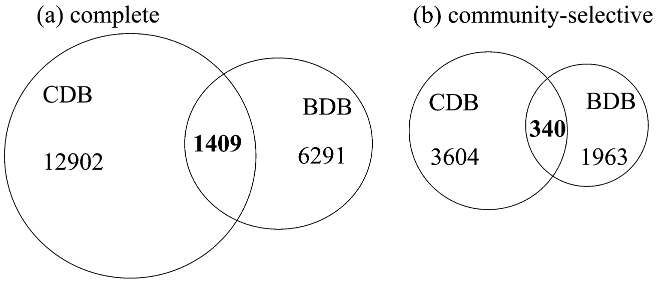


Figure 1. Scaffold overlap between CDB and BDB. Scaffold sets extracted from CDB and BDB are compared in Venn diagrams: (a) all scaffolds, (b) community-selective scaffolds. The number of shared scaffolds is shown in bold.

interval assignments were made, i.e. 0, [0, 1], [1, 2], [2, 3], [3, 4], and [4, Max]. Here, Max designates the highest selectivity value range. For example, a scaffold was assigned to selectivity interval [3, 4] if the potency ratios of its compounds against their target pairs spanned 3 to 4 orders of magnitude. For each selectivity interval, the number of scaffolds, average number of unique compounds per scaffold, average number of target pairs, and average maximum potency ratio were calculated.

Scaffold Discontinuity Scores. A local SAR discontinuity score was originally developed to quantify the SAR contributions of individual compounds in data sets and identify compounds forming activity cliffs.[33] We have adapted this scoring scheme to assess the propensity of scaffolds to yield compounds forming activity or selectivity cliffs. For each scaffold, all compounds were collected for all targets (activity cliff assessment) or target pairs (selectivity cliff assessment),

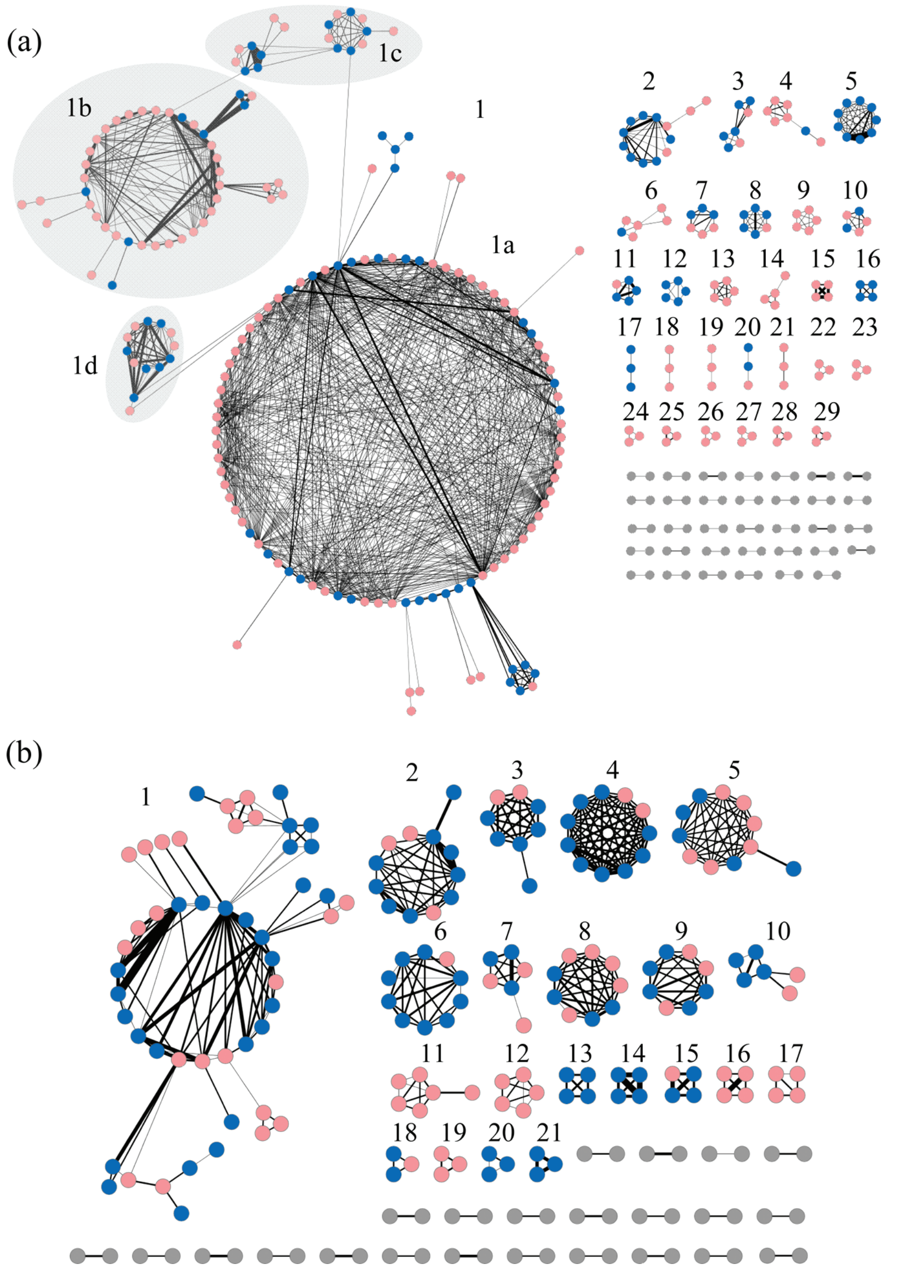


Figure 2. Scaffold-based target networks. Nodes represent targets that are connected by an edge if they share at least five scaffolds. Edge width is scaled according to the number of shared scaffolds. Target communities that contain at least three targets are considered in our analysis and consecutively numbered. Target classes are described in Table 1. Nodes representing targets that do not belong to these communities are colored gray. Network representations are shown for (a) CDB and (b) BDB. Targets common to CDB and BDB are colored blue.

Table 1. Composition of Target Communities[a]

		number of			
community	target family	targets	target pairs	compounds	scaffolds
	(a) CDB				
1a	tyrosine kinases, serine/threonine protein kinases	99	696	1283	605
1b	GPCRs	43	151	2685	1090
1c	GPCRs, cytochrome P450 enzymes	18	45	779	348
1d	matrix metalloproteinases	13	47	577	256
2	serine proteases	13	27	345	202
3	phosphodiesterases	7	12	145	53
4	prostanoid receptors	7	11	120	51
5	carbonic anhydrases	9	36	462	173
6	phosphatases	6	8	35	20
7	dipeptidyl peptidases	6	10	118	67
8	steroid receptors	6	15	399	89
9	GABAA receptors	5	10	48	9
10	sphingosine 1-phosphate (S1P) receptors	5	10	152	39
11	cathepsins	5	9	270	154
12	histone deacetylases	5	7	21	11
13	somatostatin receptors	5	10	75	39
14	cytochrome P450 enzymes	4	4	95	36
15	melanocortin receptors	4	6	317	179
16	caspases	4	6	126	61
17	β-secretases and cathepsin D	3	1	12	6
18	matrix metalloproteinases	3	2	10	10
19	fatty acid binding proteins	3	2	10	6
20	dehydragenases	3	2	8	7
21	vasopressin/oxytocin receptors	3	2	91	32
22	excitatory amino acid transporters	3	3	32	10
23	retinoic acid receptors	3	3	8	6
24	steroid reductases/isomerases	3	3	37	12
25	peroxisome proliferator-activated receptors	3	3	154	65
26	guanine nucleotide-binding protein G	3	3	28	9
27	neuropeptide Y receptors	3	3	10	10
28	adrenergic receptors	3	3	101	46
29	nitric-oxide synthase	3	3	88	42
	(b) BDB				
1	tyrosine kinases and cytochrome P450 enzymes	50	100	2128	782
2	serine proteinases	12	34	545	229
3	protein kinase C	8	22	72	34
4	carbonic anhydrases	11	55	327	87
5	phosphodiesterases	11	39	117	47
6	matrix metalloproteinases	10	24	187	56
7	protein kinase B and serine protein kinases	6	11	109	78
8	caspases	9	31	114	49
9	histone deacetylases	8	22	121	68
10	purinergic receptors	6	7	107	54
11	phosphoinositide 3-kinases (PI3Ks)	6	10	46	26
12	GABAA receptors	5	9	8	7
13	opioid receptors	4	6	84	27
14	cathepsins	4	6	307	152
15	dipeptidyl peptidases	4	6	287	105
16	esterases	4	6	238	110
17	polo-like kinases	4	5	35	21
18	sphingosine 1-phosphate (S1P) receptors	3	3	20	9
19	peroxisome proliferator-activated receptors	3	3	61	16
20	steroid receptors	3	3	35	9
21	β-secretases and cathepsin D	3	3	127	66

[a] Target communities extracted from scaffold-based target networks are reported for (a) CDB and (b) BDB. For each community, the target family annotation, the number of targets and target pairs, and the number of active compounds and corresponding scaffolds are reported.

and the potency-based scaffold discontinuity score (PScS) or selectivity-based score (SScS) was calculated as follows:

$$\mathrm{PScS}(s) = \frac{\sum (|p_i - p_j| \times \mathrm{sim}(i,j))}{|ij|}$$

$$\mathrm{SScS}(s) = \frac{\sum (|\mathrm{SR}_i - \mathrm{SR}_j| \times \mathrm{sim}(i,j))}{|ij|}$$

Here, $|p_i - p_j|$ and $|\mathrm{SR}_i - \mathrm{SR}_j|$ indicate the absolute potency value and selectivity ratio difference of compounds i and j represented by scaffold s, respectively, $\mathrm{sim}(i,j)$ is the structural similarity of compounds i and j, assessed by MACCS[34] Tanimoto similarity,[35] and $|ij|$ is the number of all compound pairs. Scores were normalized with respect to scaffold scores. Raw scores were first transformed into conventional z scores and then mapped to a cumulative probability function assuming a normal value distribution in order to obtain final scores between 0 and 1.[33] Scaffolds were initially ranked on the basis of these scores that reflect their general propensity to form activity and/or selectivity cliffs.

For each scaffold representing at least three compounds active against more than one target, the score calculations over all targets described above were then repeated for each individual target using the compounds active against the target. These calculations identify activity/selectivity cliffs on a per target basis.

Scaffold analysis and classification was carried out with in-house generated Perl and Pipeline Pilot[36] programs.

RESULTS AND DISCUSSION

Comparison of BDB and CDB Scaffolds. Given our selection criteria for compounds active against human targets, 17 745 compounds with activity annotations against 433 human targets were taken from BDB. These compounds produced 6291 unique scaffolds. From CDB, 32 848 compounds active against 671 human targets were selected yielding 12 902 unique scaffolds. There was limited compound and scaffold overlap between BDB and CDB; only 3589 compounds and 1409 scaffolds were shared by both databases (Figure 1). Hence, the scaffold information in both databases was complementary and a total of 47 004 unique compounds and 17 784 unique scaffolds were available for further analysis.

Scaffold-Based Target Network. The CDB and BDB scaffold sets were used to build scaffold-based target networks in order to establish target communities (classes) for the analysis of community- and target-selective scaffolds. In these network representations, targets (nodes) are connected if they share active compounds yielding at least five unique scaffolds. The scaffold-based target networks are shown in Figure 2, and the resulting communities are designated in Table 1. The CDB network in Figure 2a displays a total of 29 separate communities each consisting of at least three targets. The major network component 1 can be further subdivided into four distinct target communities (1a−1d), hence yielding a total of 32 target communities. However, the network is clearly dominated by community 1a, representing tyrosine kinases, and, to a lesser extent, by community 1b, representing G protein coupled receptors (GPCRs). Thus, kinase inhibitors and GPCR antagonists account for much of the information contained in CDB. Target communities in the corresponding BDB network in Figure 2b are more evenly distributed. A total of 21 communities with at least three targets are formed. Here, the largest community 1 is formed by kinases and cytochrome P450 isoforms (that share many active compounds in BDB). However, this community is much smaller than community 1a in the CDB network that contains much more kinase (but no cytochrome P450) information. There is significant overlap between a number of multi-target communities in CDB and BDB (as indicated by blue nodes in Figure 2), but both networks also contain several distinct small communities. A particularly noteworthy case of complementarity between these databases is provided by the GPCR community 1b in the CDB network, its second largest community. GPCR information is clearly under-represented in BDB (see communities 10 and 13) where GPCR ligands correspond to a total of fewer than 100 unique scaffolds, whereas 1090 GPCR ligand scaffolds are found in CDB (Table 1). Moreover, there are also relative differences between target and scaffold information in CDB and BDB, the most striking case again being the kinase communities. In CDB, community 1a represents 696 target pairs that are connected by 605 scaffolds. In BDB, the combined kinase/cytochrome P450 community 1 only represents 100 target pairs that are, however, connected by 782 scaffolds. Hence, for kinases, CDB contains more target and BDB more scaffold/chemical information. Similar observations are made for other corresponding target communities.

Table 2. Comparison of Target and Scaffold Numbers[a]

		CDB	BDB
network	targets	371	220
	target pairs	1188	428
	scaffolds	4167	2467
communities	number	32	21
	targets	303	174
	target pairs	1154	405
	community-selective scaffolds	3604	1963
target-selective scaffolds	10-fold	695 (112)	472 (100)
	50-fold	343 (43)	250 (55)
	100-fold	244 (24)	191 (42)
selectivity patterns	10-fold	184 (104)	78 (50)
	50-fold	114 (64)	49 (31)
	100-fold	88 (51)	42 (23)

[a] The numbers of targets, community- and target-selective scaffolds, and selectivity patterns are reported for CDB and BDB. Target-selective scaffolds and selectivity patterns are provided at different selectivity levels. The numbers of target-selective scaffolds and selectivity patterns that are represented by multiple compounds are given in parentheses. Selectivity patterns are target relationships evolving around specific targets that are formed by multiple target-selective scaffolds.

Community- and Target-Selective Scaffolds. We determined the number of community-selective scaffolds for each database and the number of target-selective scaffolds at different selectivity levels. The results are reported in Table 2. Of 4167 CDB and 2467 BDB scaffolds that were extracted from the scaffold-based target networks, 3658 and 1991, respectively, were found to be community-selective (i.e., the compounds represented by each scaffold were only active against targets in a single community). However, only 340 community-selective scaffolds were common to CDB and BDB, and hence the overlap was limited (Figure 1). We then compared target-selective scaffolds in CDB and BDB. Previously, we reported that a total of 191 target-selective scaffolds were available in BDB at a 100-fold selectivity level but that only 42 of those were represented by multiple compounds.[24] Similar results were also obtained for CDB. The number of target-selective scaffolds declined from 695 at the 10-fold selectivity level to 244 at the 100-fold level. However, at the 50- and 100-fold selectivity level, only 43 and 24 of these CDB scaffolds, respectively, were represented by multiple compounds, i.e., fewer than for BDB (Table 2). Processing the entire CDB only added 23 unique multiple-compound scaffolds at the 100-fold selectivity level to the 42 previously identified BDB scaffolds, hence supporting the conclusion that public domain selectivity data is currently sparse.

Regardless of compound numbers, target-selective scaffolds are in general interesting from another perspective because they often form “selectivity patterns” around individual targets, i.e., inter-target relationships constituted by

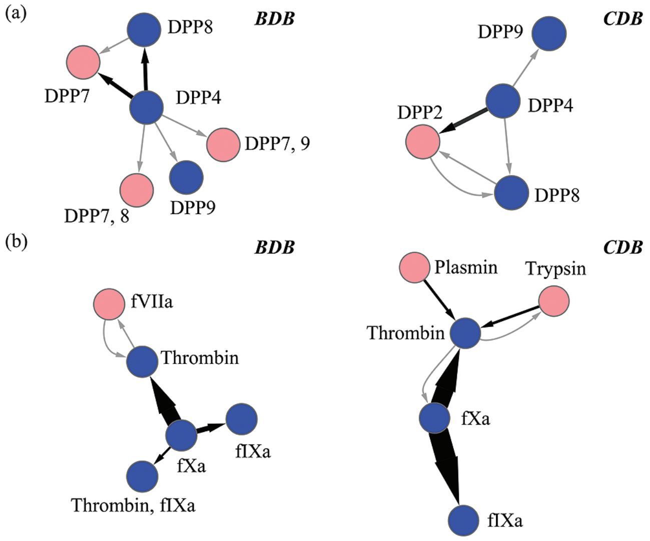


Figure 3. Comparison of selectivity patterns. Shown are four representative selectivity patterns at the 50-fold selectivity level for two target families, (a) dipeptidyl peptidases and (b) serine proteases. For each family, the pattern derived from BDB is shown on the left and the corresponding CDB pattern on the right. Selectivity patterns are displayed in a directed network representation where nodes represent targets that are connected by an arrow if they share at least one target-selective scaffold. The arrow points from the selective target to the non-selective target, thus representing a "selective over" relationship. The width of the arrows is scaled according to the number of shared target-selective scaffolds. Arrows representing selectivity relationships formed by a single scaffold are colored gray. Nodes are annotated with target names. "DPP" stands for dipeptidyl peptidase and "f" for factor. Nodes of targets shared between BDB and CDB are colored blue.

multiple target-selective scaffolds. As examples, Figure 3 shows corresponding BDB and CDB selectivity patterns that evolve around dipeptidyl peptidase 4 and factor Xa, respectively. In such selectivity patterns, target-selective scaffolds establish different selectivity relationships between related targets. Table 2 shows that many selectivity patterns can be extracted from CDB and BDB. As can be seen in Figure 3, selectivity patterns for given targets often differ in CDB and BDB and can be combined to further increase their information content, which again points at a notable degree of complementarity between these databases. The analysis of scaffold-based selectivity patterns is of practical relevance because these patterns that can be exploited, for example, in the design of compounds that are target-selective, have inverse selectivity, or display a desired selectivity profile for a group of related targets.

Potency and Selectivity Ranges. We next classified the complete CDB and BDB scaffold sets according to the potency and selectivity ranges of the compounds they represent. The potency- and selectivity-based classifications are reported in Table 3a and b, respectively. Increasing potency ranges of compounds representing a given scaffold provide an indication of activity cliff potential. Table 3a reveals some clear trends for both CDB and BDB scaffolds. With increasing potency ranges, the number of scaffolds decreases, as one would expect, but the number of compounds per scaffold increases and also the number of targets the compounds are active against. For the largest potency ranges of more than 4 orders of magnitude, 278 CDB scaffolds are found that are, on average, represented by ~17 compounds active against ~7 targets and 165 BDB scaffolds each represented by ~18 compounds active against ~5 targets. Equivalent trends are observed for the selectivity-based classification in Table 3b. Here, the number of scaffolds also decreases for increasing selectivity ranges, but the number of compounds they represent and the number of target pairs these compounds are active against increase. For the largest target selectivity ranges, 52 CDB and 42 BDB scaffolds exist that are, on average, represented by ~18 and ~9 compounds active against ~209 and ~31 target pairs, respectively. There are generally fewer scaffolds for selectivity- than potency-based classification because selectivity results from activity against a minimum of two targets, which only applies to a subset of compounds and scaffolds. Many of the prioritized scaffolds have already been explored rather extensively, as suggested by, in part, large numbers of compounds corresponding to individual scaffolds and multiple targets they have been tested against. This suggests that more extensive chemical exploration of other scaffolds might further increase the number of scaffolds yielding compound potency differences of more than 4 orders of magnitude. Taken together, these findings strongly indicated that several hundred scaffolds already exist in public domain compound data that generate compounds with differential activity

Table 3. Scaffold Classification Based on Potency and Selectivity Ranges[a]

(a) potency-based

potency range	# scaffolds	# cpds	# targets	max pot
CDB				
0	6431	1.0	1.0	6.82
[0, 1]	2433	2.1	1.7	7.07
[1, 2]	1881	3.2	2.2	7.62
[2, 3]	1212	4.5	2.8	8.17
[3, 4]	667	7.2	3.6	8.68
[4, 11.72]	278	16.9	6.9	9.30
BDB				
0	2823	1.0	1.0	6.59
[0, 1]	1232	1.9	1.7	6.96
[1, 2]	989	3.1	2.2	7.47
[2, 3]	658	4.8	2.7	7.90
[3, 4]	424	8.0	3.2	8.48
[4, 8.29]	165	17.7	5.4	9.18

(b) selectivity-based

selectivity range	# scaffolds	# cpds	# tps	max |sr|
CDB				
0	1642	1.0	1.0	1.11
[0, 1]	948	2.2	2.8	0.93
[1, 2]	861	2.6	4.5	1.85
[2, 3]	560	4.0	11.8	2.70
[3, 4]	253	4.9	26.0	3.61
[4, 7.13]	52	15.6	208.6	4.63
BDB				
0	1033	1.0	1.0	1.17
[0, 1]	491	2.6	2.4	1.14
[1, 2]	459	3.6	5.0	1.95
[2, 3]	295	4.7	6.9	2.70
[3, 4]	106	6.1	15.8	3.51
[4, 6.83]	42	9.2	31.3	4.74

[a] Scaffolds extracted from CDB and BDB are classified into six sets based on (a) potency and (b) selectivity value ranges. Reported are the number (#) of scaffolds, average number of compounds (cpds) and targets or target pairs (tps) per scaffold, and the average maximal logarithmic potency (max pot) or selectivity ratio (max |sr|).

Table 4. Scaffolds from CDB with High Potency-Based Discontinuity Scores[a]

rank	scaffold ID	PScS	#cpds	#targets	#targets ⟨MultiCpds⟩	#TargetCliffs
1	6445	1	30	24	**10**	**4**
2	9793	1	5	28	**1**	**1**
3	2269	1	4	3	**3**	**1**
4	1901	1	4	6	**1**	**1**
5	3866	1	3	9	**3**	**2**
16	2623	0.98	24	9	**2**	**1**
17	5821	0.98	9	3	**3**	**3**
27	5749	0.97	15	2	**1**	**1**
33	3659	0.96	27	3	**3**	**2**
34	11047	0.96	8	10	**4**	**2**
39	10707	0.95	46	22	**11**	**7**
40	8996	0.95	13	2	**2**	**1**
41	1927	0.95	10	8	**7**	**3**
48	5794	0.94	52	6	**3**	**3**
59	347	0.92	28	3	**3**	**2**
60	10775	0.92	14	2	**2**	**1**
61	8115	0.92	9	3	**3**	**3**
68	4196	0.90	16	15	**6**	**1**
73	2712	0.89	27	4	**4**	**2**
84	11159	0.87	21	2	**2**	**2**
85	9992	0.87	19	4	**4**	**2**
90	10539	0.86	24	10	**7**	**1**
91	3062	0.86	11	9	**1**	**1**
92	1105	0.86	9	3	**3**	**2**
94	2306	0.85	18	7	**1**	**1**
95	10235	0.85	12	2	**1**	**1**
96	12449	0.85	12	5	**2**	**3**
97	10483	0.85	12	3	**3**	**2**
98	2831	0.85	9	4	**2**	**1**
106	5285	0.84	10	3	**3**	**1**
111	3355	0.83	27	2	**2**	**2**
112	9749	0.83	9	5	**4**	**3**
117	6783	0.82	14	5	**4**	**1**
118	9066	0.82	14	3	**3**	**2**
126	5314	0.81	20	3	**3**	**2**
127	2849	0.81	19	3	**2**	**1**

[a] A total of 36 CDB scaffolds with PScS greater than 0.8 that represent more than two compounds that are active against more than one target are listed. For each scaffold, the score-based rank position (rank), discontinuity score (PScS), the number of unique compounds it represents (#cpds), the total number of targets (#targets), the number of targets with multiple active compounds (#targets <MultiCpds>), and the number of targets for which it forms activity cliffs (#TargetCliffs) are reported.

Table 5. Scaffolds from BDB with High Potency-Based Discontinuity Scores[a]

rank	scaffold ID	PScS	#cpds	#targets	#targets ⟨MultiCpds⟩	#TargetCliffs
1	1990	1	6	2	**1**	**1**
2	413	1	4	2	**2**	**2**
3	455	1	4	2	**2**	**2**
4	1161	1	3	2	**2**	**2**
5	851	0.99	32	6	**2**	**2**
7	363	0.98	17	4	**3**	**2**
11	312	0.97	9	3	**3**	**3**
14	1348	0.96	8	8	**7**	**3**
15	274	0.95	13	5	**5**	**5**
16	1304	0.95	13	4	**1**	**1**
20	1144	0.94	9	3	**3**	**1**
27	1506	0.92	13	4	**2**	**2**
28	1922	0.92	8	4	**1**	**1**
31	1120	0.91	13	2	**2**	**1**
35	204	0.90	31	2	**1**	**1**
36	606	0.90	24	5	**5**	**1**
41	819	0.89	9	3	**1**	**1**
42	2389	0.88	60	22	**7**	**4**
46	857	0.87	8	5	**3**	**3**
48	1106	0.86	8	5	**4**	**3**
51	1457	0.85	14	2	**2**	**1**
52	39	0.85	12	4	**2**	**1**
54	1169	0.84	10	6	**6**	**4**
56	1257	0.83	148	7	**6**	**2**
57	1109	0.83	40	9	**4**	**3**
58	193	0.83	35	2	**1**	**1**
59	972	0.83	27	2	**2**	**1**
60	833	0.83	14	2	**2**	**2**
61	1385	0.83	14	3	**2**	**1**
62	2363	0.83	11	2	**2**	**2**
63	1152	0.82	63	12	**7**	**4**
64	810	0.82	13	2	**2**	**1**
65	1425	0.82	9	3	**2**	**1**
68	1851	0.81	29	2	**1**	**1**
69	720	0.81	20	2	**2**	**1**
70	787	0.81	15	2	**2**	**1**

[a] A total of 36 BDB scaffolds with PScS greater than 0.8 that represent more than two compounds that are active against more than one target are listed. For each scaffold, the score-based rank position (rank), discontinuity score (PScS), the number of unique compounds it represents (#cpds), the total number of targets (#targets), the number of targets with multiple active compounds (#targets < MultiCpds>), and the number of targets for which it forms activity cliffs (#TargetCliffs) are reported.

against several targets and the tendency to produce activity (or selectivity) cliffs.

Cliff-Forming Scaffolds. The search for activity cliff-forming scaffolds was further refined by calculating a discontinuity score for each scaffold. This calculation involves systematic pairwise similarity and potency comparison of compounds containing the scaffold. High discontinuity scores approaching 1 indicate the presence of significant activity cliffs within the compound set. This scoring formalism is also applicable to selectivity-based score calculation because selectivity is expressed as a pairwise potency ratio.

Scaffold discontinuity scores were first calculated over all targets against which compounds with a particular scaffold were active (i.e., global scores), which provides a general measure for the propensity of a scaffold to form cliffs. Then, the scores were recalculated on a per-target basis, thus identifying target-dependent activity cliffs, if present, or target pair-dependent selectivity cliffs. For both global and target-based calculations, discontinuity scores of greater than 0.8 were considered. In our experience, this score level reliably indicates the presence of activity cliffs.

In CDB and BDB, 137 and 75 scaffolds were found, respectively, that achieved global potency-based discontinuity scores of greater than 0.8 and were represented by more than two compounds active against more than one target. The complete sets of these CDB and BDB scaffolds with potency values and intervals are provided in Tables S1 and S2 (Supporting Information), respectively. In Tables 4 and 5, 36 of these CDB and BDB scaffolds are reported, respectively. Furthermore, in Tables 6 and 7, all 34 CDB and all 23 BDB scaffolds are listed that produced selectivity-based discontinuity scores of greater than 0.8 and were represented by more than two compounds active against more than one target pair. Tables S3 and S4 (Supporting Information) list these scaffolds with associated selectivity ratios and selectivity intervals. Only two scaffolds formed both strong activity

Table 6. Scaffolds from CDB with High Selectivity-Based Discontinuity Scores[a]

rank	scaffold ID	#cpds	SScS	#TPs	#TPs ⟨MultiCpds⟩	#TPCliffs
1	9230	5	1	3	**1**	**1**
2	10683	4	1	34	**6**	**5**
3	4991	3	1	3	**3**	**3**
4	10707	16	0.99	27	**12**	**9**
5	2840	7	0.99	3	**3**	**2**
6	6582	3	0.99	10	**10**	**7**
7	12673	3	0.98	6	**3**	**3**
8	3211	3	0.97	3	**3**	**2**
9	3754	7	0.96	3	**2**	**1**
10	5304	5	0.96	6	**3**	**2**
11	572	5	0.95	3	**3**	**2**
12	7198	4	0.95	3	**3**	**2**
13	4266	3	0.95	3	**1**	**1**
14	8848	3	0.95	6	**3**	**2**
15	143	7	0.94	3	**3**	**2**
16	5834	6	0.93	6	**3**	**3**
17	7991	21	0.92	10	**10**	**3**
18	3153	3	0.92	3	**1**	**1**
19	10439	5	0.91	6	**5**	**3**
20	973	3	0.91	21	**1**	**1**
21	12298	4	0.90	3	**3**	**2**
22	1779	3	0.90	6	**6**	**3**
23	12627	3	0.90	3	**3**	**2**
24	2634	6	0.89	6	**6**	**3**
25	2712	8	0.88	6	**5**	**2**
26	1087	3	0.86	6	**3**	**1**
27	6611	13	0.84	3	**3**	**1**
28	9649	4	0.84	3	**2**	**2**
29	6332	17	0.83	35	**21**	**3**
30	12747	4	0.83	3	**3**	**1**
31	2008	3	0.82	3	**3**	**2**
32	6576	3	0.82	6	**6**	**2**
33	9011	66	0.81	3	**1**	**1**
34	347	23	0.81	3	**3**	**2**

[a] All 34 CDB scaffolds with SScS greater than 0.8 that represent more than two compounds that are active against more than one target are listed. For each scaffold, the score-based rank position (rank), discontinuity score (SScS), the number of unique compounds it represents (#cpds), the total number of target pairs (#TPs), the number of target pairs with multiple active compounds (#TPs ⟨MultiCpds⟩), and the number of selectivity cliffs it forms (#TPCliffs) are reported.

Table 7. Scaffolds from BDB with High Selectivity-Based Discontinuity Scores[a]

rank	scaffold ID	SScS	#cpds	#TPs	#TPs ⟨MultiCpds⟩	#TPCliffs
1	2115	1	3	3	**3**	**3**
2	312	0.99	9	3	**3**	**3**
3	381	0.99	6	3	**3**	**3**
4	424	0.99	5	3	**3**	**3**
5	196	0.97	4	2	**1**	**1**
6	857	0.97	4	7	**1**	**1**
7	1425	0.96	6	3	**1**	**1**
8	733	0.96	3	3	**1**	**1**
9	1615	0.94	6	4	**1**	**1**
10	1008	0.94	3	4	**3**	**2**
11	1169	0.93	8	13	**8**	**4**
12	1001	0.92	4	4	**2**	**2**
13	1100	0.90	6	3	**3**	**2**
14	2005	0.88	8	3	**3**	**2**
15	511	0.87	10	3	**1**	**1**
16	466	0.87	9	3	**3**	**3**
17	1238	0.87	5	3	**2**	**1**
18	1152	0.86	21	9	**4**	**2**
19	362	0.85	6	3	**3**	**1**
20	1319	0.85	3	3	**3**	**3**
21	1106	0.84	5	10	**6**	**4**
22	2208	0.81	45	6	**6**	**2**
23	446	0.81	19	26	**9**	**4**

[a] All 23 BDB scaffolds with SScS greater than 0.8 that represent more than two compounds that are active against more than one target are listed. For each scaffold, the score-based rank position (rank), discontinuity score (SScS), the number of unique compounds it represents (#cpds), the total number of target pairs (#TPs), the number of target pairs with multiple active compounds (#TPs ⟨MultiCpds⟩), and the number of selectivity cliffs it forms (#TPCliffs) are reported.

and selectivity cliffs, scaffold 10 707 (Table 4 and 6) and scaffold 1152 (Table 5 and 7).

Many of the scaffolds in Tables 4 and 5 form activity cliffs against multiple targets. For example, the top-scoring scaffold in Table 4 (rank 1) corresponds to 30 compounds that are active against 24 targets and form significant activity cliffs against four of these targets. The scaffold at score rank 17 corresponds to nine compounds that are active against three targets and form activity cliffs in each case. Furthermore, the scaffold at score rank 39 is represented by 46 compounds active against 22 targets, forming activity cliffs for seven of these targets. In Table 5, the scaffold at rank 5 corresponds to 32 compounds active against six targets and forming activity cliffs against two of them. Moreover, the scaffold at rank 15 is represented by 13 compounds that form activity cliffs for all five targets they are active against. Similar observations were made for selectivity cliffs. For example, in Table 6, the scaffold at rank 4 is represented by 16 compounds forming nine selectivity cliffs, and the scaffold at rank 2 in Table 7 corresponds to nine compounds forming three selectivity cliffs. A total of 25 CDB (Table 6) and 15 BDB (Table 7) scaffolds were found to yield compounds forming multiple selectivity cliffs. In many instances, fewer than 10 compounds representing a particular scaffold produced activity or selectivity cliffs for multiple targets. Thus, thorough chemical exploration of these scaffolds was not required for cliff formation.

Multi-target Activity Cliffs. Which types of scaffolds form multi-target activity cliffs? Figure 4a and b show representative CDB and BDB scaffolds that are represented by more than two compounds and form at least three activity cliffs for distinct targets. The target annotations of the scaffolds are shown in Figures S1 and S2 (Supporting Information). These scaffolds are of different sizes and chemical natures ranging from small generic structures, e.g., simple aliphatic rings such as cyclohexane, tetrahydrofuran, or pyrrolidine, to complex multiring scaffolds. Hence, multi-target activity cliff scaffolds were diverse, and there were no apparent preferences for specific chemotypes. Importantly, many of these scaffolds formed activity cliffs for targets belonging to different communities. Similar observations were made for multi-target selectivity cliff scaffolds. These BDB and CDB scaffolds and their selectivity annotations are shown in Figures S3 and S4 (Supporting Information), respectively. Thus, the formation of multiple-target activity or selectivity cliffs was not limited to closely related targets but also involved different classes of targets. Figure 5 shows representative multi-target activity cliffs. Different pairs of compounds representing a scaffold introduce activity cliffs of varying magnitudes against different targets. Such

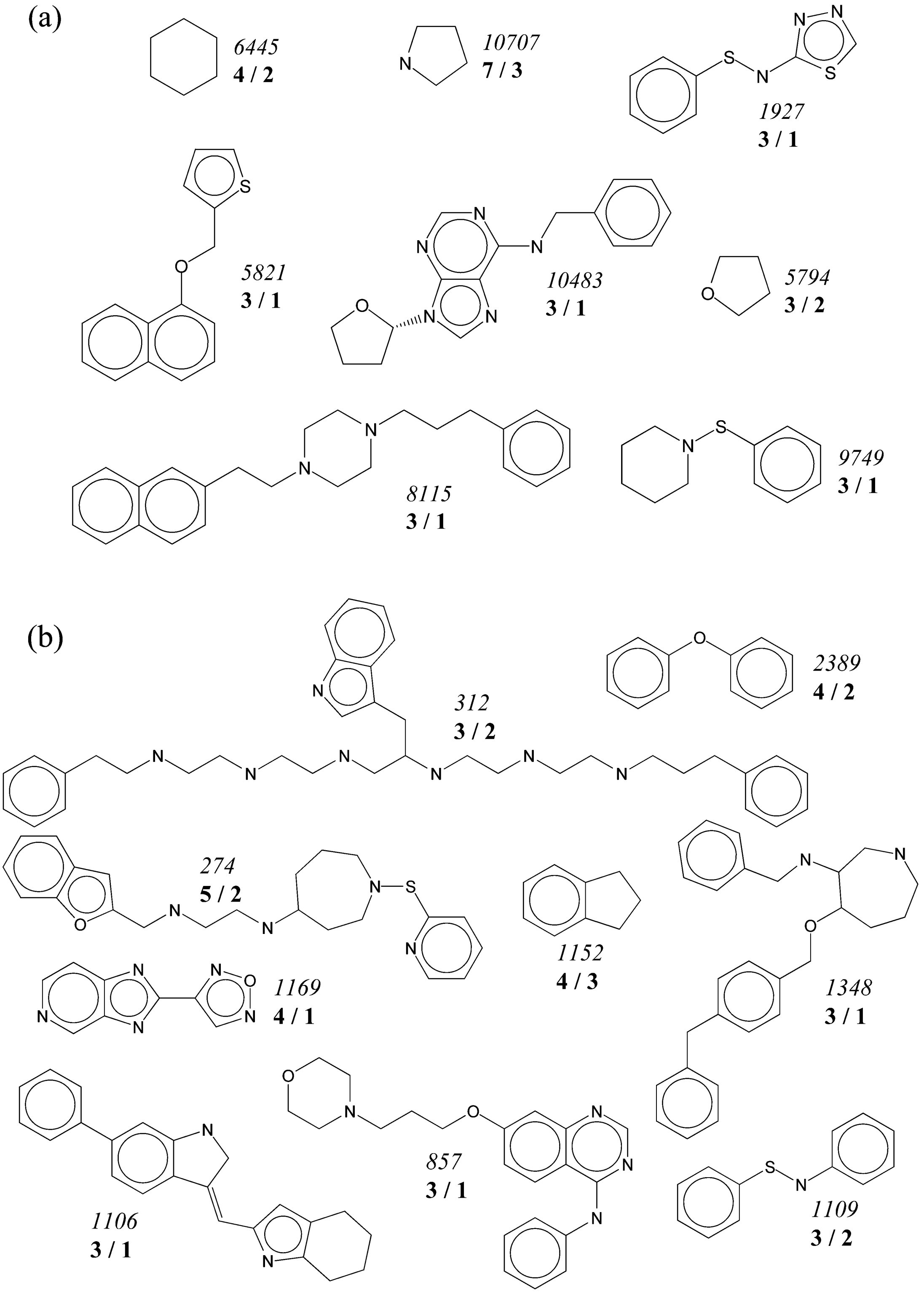


Figure 4. Scaffolds with high propensity to form activity cliffs. Scaffolds are shown that produce activity cliffs for at least three targets (a, CDB; b, BDB). Scaffolds IDs (in italics) and target/community numbers (bold) are provided. For example, "3/2" means that a scaffold forms activity cliffs for three targets in two communities.

patterns are representative for many multi-target activity cliff scaffolds. In Table S5 (Supporting Information), we provide SMILES representations of the complete set of multi-target activity cliff scaffolds represented by multiple compounds. Table S6 (Supporting Information) provides a corresponding list of multi-target selectivity cliff scaffolds.

CONCLUSIONS

In this study, we have primarily explored the question of whether molecular scaffolds exist that display a general tendency to form activity or selectivity cliffs against different targets. From the ChemblDB and BindingDB databases, scaffolds and associated compound activity data were systematically extracted, target communities were estab-

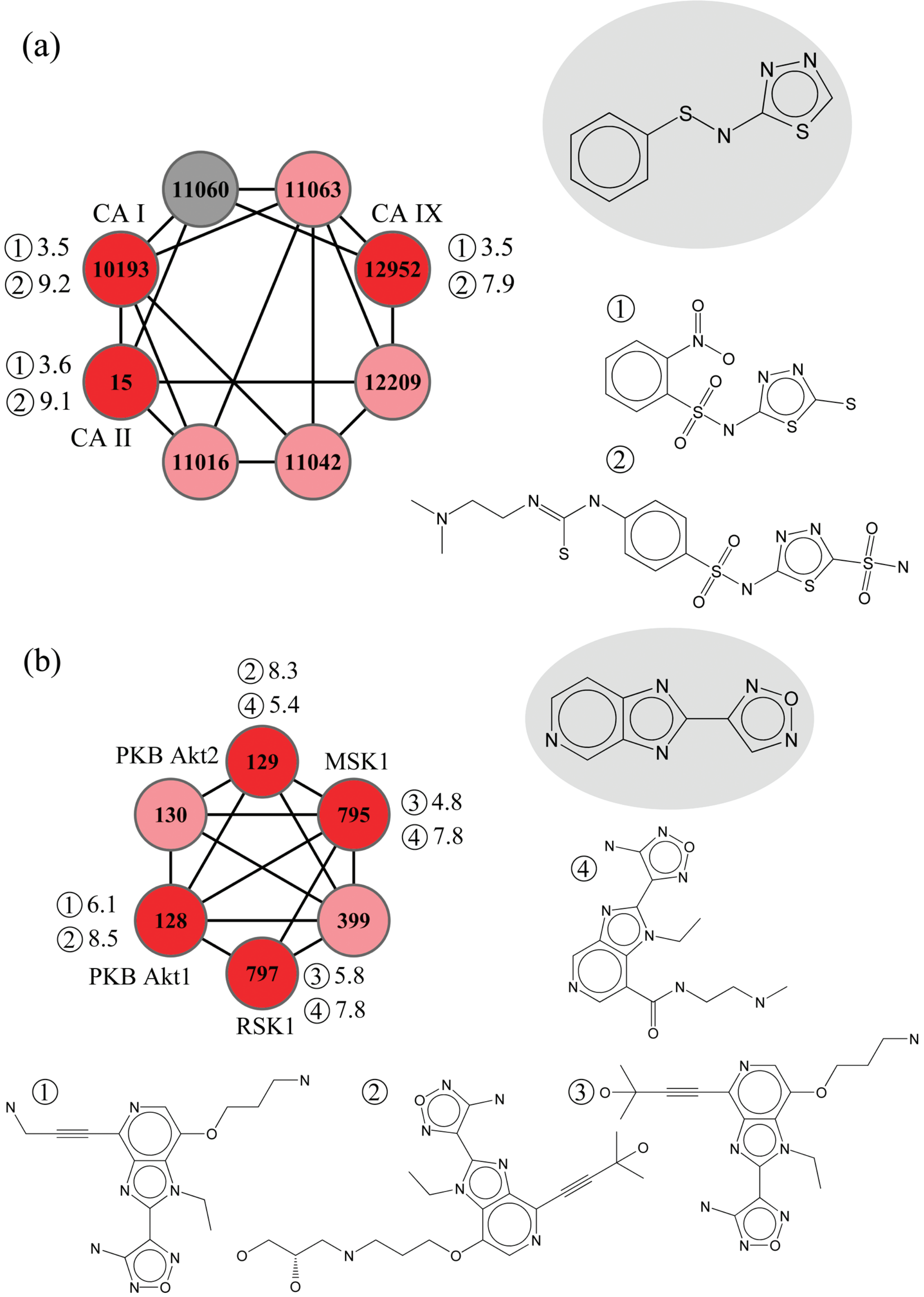


Figure 5. Representative multi-target activity cliffs. Two representative scaffolds are shown that form multi-target activity cliffs (a, CDB; b, BDB). The scaffolds are shown on a gray background. Nodes represent targets. Two nodes are connected if they share compounds containing the scaffold. A target is colored gray if only a single compound is reported to be active against it (hence, for such targets, activity cliffs cannot be detected). A node is colored red if compounds active against the target yield a PScS value greater than 0.8 (indicating strong discontinuity). Representative compounds containing the scaffold are shown and labeled. For these compounds, negative logarithmic potency values for individual targets are reported. The differences between these values indicate the magnitude of the target-dependent activity cliffs the compounds form. Target abbreviations: CA, carbonic anhydrase; MSK, mitogen- and stress-activated protein kinase; PKB, protein kinase B; RSK, ribosomal S6 kinase.

lished, and target-selective and cliff-forming scaffolds were identified. Consistent with our early findings, compound selectivity data is only sparse in public domain compound databases, which currently limits a reliable assignment of target-selective scaffolds. By contrast, we have identified a significant number of scaffolds that are represented by compounds forming activity or selectivity cliffs against multiple targets. These targets are often unrelated and occur in different target communities. Many multi-target cliff scaffolds have not yet been extensively explored; i.e., they are currently represented by only fewer than 10 compounds, yet they already display strong tendencies of cliff formation. Multi-target activity cliff scaffolds are in part promiscuous in nature and able to bind to different types of targets. However, these scaffolds yield compounds that form substantial activity cliffs for different targets and are thus phenotypically distinct from “non-specific” molecules. Thus, multi-target activity cliff scaffolds might be interesting candidates for compound optimization when considered on a per-target basis. Yet it should be taken into account that compounds derived from such scaffolds might often be highly potent against multiple targets. However, depending on the therapeutic application, the use of multi-target activity cliff scaffolds might also be desirable, for example, when optimizing compounds for series of closely related targets having similar or overlapping functions. The collection of multi-target cliff scaffolds we have identified and provide as part of this study should be helpful in order to evaluate and prioritize scaffolds for compound optimization efforts.

Supporting Information Available: Tables S1−S4 list scaffolds with high potency- or selectivity-based discontinuity scores, and Tables S5 and S6 provide SMILES representations for multi-target activity and selectivity cliff scaffolds that are represented by at least three compounds. Figures S1−S4 show target and target pair annotations of multi-target activity and selectivity cliff scaffolds. This information is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.

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

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

(4) Zhao, H. Scaffold Selection and Scaffold Hopping in Lead Generation: A Medicinal Chemistry Perspective. *Drug Discovery Today* **2007**, *12*, 149–155.

(5) Wang, J.; Hou, T. Drug and Drug Candidate Building Block Analysis. *J. Chem. Inf. Model.* **2010**, *50*, 55–67.

(6) 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.

(7) 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.

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

(9) Krier, M.; Bret, G.; Rognan, D. Assessing the Scaffold Diversity of Screening Libraries. *J. Chem. Inf. Model.* **2006**, *46*, 512–524.

(10) Lipkus, A. H.; Yuan, Q.; Lucas, K. A.; Funk, S. A.; Bartelt, W. F., III; Schenck, R. J.; Trippe, A. J. Structural Diversity of Organic Chemistry. A Scaffold Analysis of the CAS Registry. *J. Org. Chem.* **2008**, *73*, 4443–4451.

(11) Hu, Y.; Bajorath, J. Scaffold Distributions in Bioactive Molecules, Clinical Trials Compounds, and Drugs. *ChemMedChem* **2010**, *5*, 187–190.

(12) Lameijer, E.; Kok, J. N.; Bäck, T.; Ijzerman, A. P. Mining a Chemical Database for Fragment Co-Occurrence: Discovery of “Chemical Clichés”. *J. Chem. Inf. Model.* **2007**, *46*, 553–562.

(13) 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.

(14) 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.

(15) 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.

(16) Siegel, M. G.; Vieth, M. Drugs in Other Drugs: A New Look at Drugs as Fragments. *Drug Discovery Today* **2007**, *12*, 71–79.

(17) Villar, H. O.; Hansen, M. R. Computational Techniques in Fragment-Based Drug Discovery. *Curr. Top. Med. Chem.* **2007**, *7*, 1509–1513.

(18) Congreve, M.; Chessari, G.; Tisi, D.; Woodhead, A. J. Recent Developments in Fragment-based Drug Discovery. *J. Med. Chem.* **2008**, *51*, 3661–3680.

(19) 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.

(20) Horton, D. A.; Bourne, G. T.; Smythe, M. L. The Combinatorial Synthesis of Bicyclic Privileged Structures or Privileged Substructures. *Chem. Rev.* **2003**, *103*, 893–930.

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

(22) Schnur, D. M.; Hermsmeier, M. A.; Tebben, A. J. Are Target-Family-Privileged Substructures Truly Privileged. *J. Med. Chem.* **2006**, *49*, 2000–2009.

(23) 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.

(24) Hu, Y.; Bajorath, J. Exploring Target-Selectivity Patterns of Molecular Scaffolds. *ACS Med. Chem. Lett.* [Online] DOI: 10.1021/ml900024v.

(25) 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.

(26) PubChem. http://pubchem.ncbi.nlm.nih.gov/ (accessed October 1, 2009).

(27) Maggiora, G. M. On Outliers and Activity Cliffs - Why QSAR Often Disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535.

(28) Bajorath, J.; Peltason, L.; Wawer, M.; Guha, R.; Lajiness, M. S.; Van Drie, J. H. Navigating Structure-Activity Landscapes. *Drug Discovery Today* **2009**, *14*, 698–705.

(29) Peltason, L.; Hu, Y.; Bajorath, J. From Structure-Activity to Structure-Selectivity Relationships: Quantitative Assessment, Selectivity Cliffs, and Key Compounds. *ChemMedChem* **2009**, *4*, 1864–1873.

(30) ChemblDB. http://www.ebi.ac.uk/chembl/ (accessed January 2, 2010).

(31) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.

(32) 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.

(33) Wawer, M.; Peltason, L.; Weskamp, N.; Teckentrup, A.; Bajorath, J. Structure-Activity Relationship Anatomy by Network-Like Similarity Graphs and Local Structure-Activity Relationship Indices. *J. Med. Chem.* **2008**, *51*, 6075–6084.

(34) *MACCS Structural Keys*; Symyx Software: San Ramon, CA, 2005.

(35) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.

(36) *Scitegic Pipeline Pilot*, Student ed., version 6.1; Accelrys, Inc.: San Diego, CA, 2007.

CI100059Q

Summary

We have systematically searched for scaffolds forming multi-target activity or selectivity cliffs. The analysis was facilitated by designing a scoring scheme that integrated pairwise similarity and potency difference of compounds to quantitatively assess the tendency of scaffolds to introduce activity or selectivity cliffs. In case of activity cliffs, 103 scaffolds were found that were represented by more than two compounds and had a high propensity to form cliffs against multiple targets. Moreover, 46 scaffolds were identified to form selectivity cliffs against multiple target pairs. These cliff-forming scaffolds were often represented by fewer than 10 compounds. Therefore, such scaffolds might be further explored in the design of compounds with desired activity or selectivity against closely related targets.

Going beyond activity cliff formation, scaffold hopping refers to the identification of new compounds having distinct scaffolds but comparable activity. Scaffold hopping potential has been intensely studied in both computational and medicinal chemistry. However, a systematic and general evaluation of scaffold hopping potential across different targets has not yet been explored. Thus, an analysis of scaffold hops has been carried out on a large scale for individual compound activity classes.

Chapter 6

Global Assessment of Scaffold Hopping Potential for Current Pharmaceutical Targets

Introduction

In chemoinformatics and drug discovery, it is often difficult to identify different structure classes having the same activity, which is commonly referred to scaffold hopping. We present a systematic survey of global scaffold hopping potential across different pharmaceutical targets. Therefore, we analyzed topologically distinct scaffolds in active compounds and designed a scoring scheme that incorporated structural similarity of scaffolds and potency information to quantitatively assess scaffold hopping potential for individual target sets and scaffold pairs. Target sets were ranked according to the frequency of distinct scaffolds sharing the same activity and scaffold hopping score. Furthermore, scaffold pairs with low structure similarity yielding comparably potent compounds were prioritized.

View Online

CONCISE ARTICLE www.rsc.org/medchemcomm | MedChemComm

Global assessment of scaffold hopping potential for current pharmaceutical targets

Ye Hu and Jürgen Bajorath*

Received 9th September 2010, Accepted 22nd September 2010
DOI: 10.1039/c0md00156b

Downloaded on 08 February 2011
Published on 13 October 2010 on http://pubs.rsc.org | doi:10.1039/C0MD00156B

Scaffold hopping is an intensely investigated topic, both in the context of computational method evaluation and practical compound screening applications. Scaffold hopping refers to the identification of different compound classes having similar biological activity and is typically explored on a case-by-case basis. However, how frequently scaffold hops occur across different targets is presently not well understood. We have investigated global scaffold hopping potential by systematically analyzing topologically distinct scaffolds in currently available bioactive compounds with defined target and activity annotations. The analysis reveals that for the majority of target proteins, active compounds representing between five and 49 topologically distinct scaffolds are available. Moreover, for 70 targets, between 50 and more than 300 distinct scaffolds are found. Thus, scaffold hops occur with rather high frequency among active compounds.

In medicinal chemistry, the search for different structural classes (chemotypes) having similar activity is generally of high interest,[1,2] for example, to support chemical optimization efforts or secure intellectual property positions. Moreover, the demonstration of scaffold hopping potential has become the "holy grail" of computational screening methods.[3–9] The "value" of any virtual screening approach is essentially judged upon its ability to identify different chemotypes having similar activity, mostly in benchmark calculations. Beyond the often rather artificial scenario provided by typical benchmark studies, in prospective applications, a virtual screen is generally claimed to be a "success" if at least one or a few novel compounds with different core structures (scaffolds) and desired biological activity have been identified. Unfortunately, the assessment of scaffold hopping potential often suffers from the lack of clear scaffold definitions and inconsistent analysis of scaffold hops.[9] Moreover, it is currently unclear how "difficult" scaffold hopping really might be. No studies are available at present that provide a general assessment of scaffold hopping potential across different targets, although such insights would be of general interest, both for the evaluation of computational screening methods and practical medicinal chemistry applications.

General scaffold hopping potential might be estimated by systematically analyzing, on a per-target basis, how many well-defined scaffold hops are "encoded" by currently available bioactive compounds. Accordingly, we have carried out a large-scale analysis of scaffold hops among publicly available active compounds. All calculations reported herein were carried out with in-house Perl and Scientific Vector Language (SVL)[10] programs and Pipeline Pilot[11] tools.

From two major public repositories of bioactive compounds, CHEMBLdb (CDB)[12] and BindingDB (BDB),[13] 31,158 and 17,745 molecules with activity annotations (K_i or IC_{50} values) against human targets were selected, respectively. These compounds were organized in 586 and 433 individual target sets and 12,047 and 6,291 atomic property-based Bemis & Murcko scaffolds[14] were extracted from them, respectively. CDB and BDB currently show limited compound overlap[15] and we therefore merged the CDB and BDB compound and scaffold sets, yielding a total of 795 individual target sets containing 45,263 compounds and 16,873 unique scaffolds.

As illustrated in Fig. 1, property-based Bemis & Murcko scaffolds consist of core ring structures and linkers between them.[14] Scaffolds only distinguished by heteroatom substitutions and bond orders display the same topology, as reflected by carbon skeletons (CSKs; *i.e.* scaffolds with all atom types set to carbon and all bond orders to one), as also illustrated in Fig. 1. We deliberately focused our analysis on topologically distinct scaffolds that are more relevant for scaffold hopping than scaffolds that are only distinguished by minor heteroatom substitutions or bond order alterations. Therefore, for each target set, we determined all Bemis & Murcko scaffolds yielding the same CSKs. In each of these cases, we only retained the scaffold that was represented by the largest number of compounds or, if several scaffolds had the same number of compounds, the scaffold represented by the largest number of compounds with highest median potency. An individual scaffold was retained instead of the CSK because compounds representing the scaffold were required for score calculations, as described below. Importantly, by retaining one Bemis & Murcko scaffold per CSK, all scaffolds selected for a target set at this stage were topologically distinct. This selection scheme yielded 10,989 topologically distinct scaffolds corresponding to 35,004 compounds. In order to further streamline the collection of target sets for meaningful scaffold hopping analysis, we only retained target sets containing at least five compounds with at least 1 μM potency (*i.e.*, pKi or $pIC_{50} >= 6$) and at least two

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

View Online

Published on 13 October 2010 on http://pubs.rsc.org | doi:10.1039/C0MD00156B
Downloaded on 08 February 2011

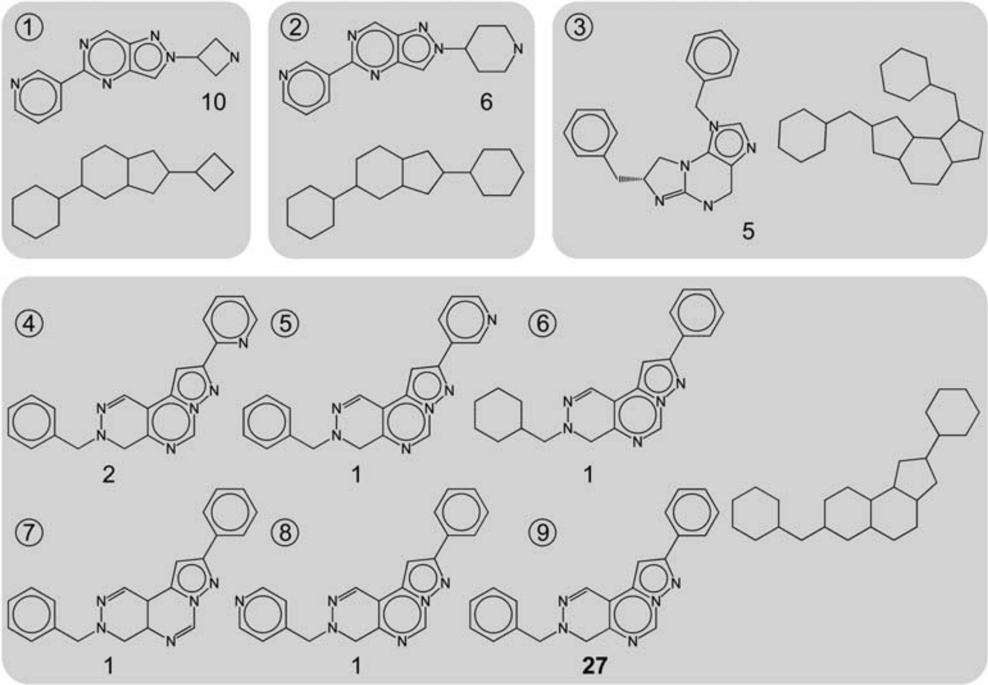


Fig. 1 Topologically distinct scaffolds. Nine representative scaffolds extracted from phosphodiesterase 5A inhibitors are shown. For each scaffold, the corresponding carbon skeleton (CSK) is shown and the number of compounds each scaffold represents is reported. Scaffolds 1 to 3 yield distinct CSKs, whereas scaffolds 4 to 9 share the same CSK. Scaffold 9 is selected for further analysis because it represents the largest number of compounds (*i.e.*, 27), and the other five scaffolds are not further considered. This selection scheme ensures that only topologically distinct scaffolds are analyzed.

Table 1 Target families and scaffold distribution.[a]

		# Targets						
		Source			# Scaffolds			
FamilyID	Target Family	BDB	CDB	Total	< 5	[5, 50)	[50, 100)	> = 100
1	Tyr protein kinases	30	32	38	4	28	2	4
2	Ser_Thr protein kinases	37	38	49	6	37	4	2
3	Ser_Thr_Tyr kinases	9	7	13	4	9	0	0
4	Phosphadiesterases	8	7	9	0	9	0	0
5	Protein phosphatases	1	3	3	0	3	0	0
6	Aspartic proteases	4	7	7	2	3	2	0
7	Cysteine proteases	11	12	14	1	9	2	2
8	Matrix metalloproteases	14	17	19	2	11	5	1
9	Serine proteases	18	21	25	2	20	0	3
10	Carbonic anhydrases	12	10	12	0	9	2	1
11	Histone deacetylases	8	5	8	0	7	1	0
12	CytochromeP450 enzymes	8	9	13	1	12	0	0
13	Transferases	4	4	8	1	7	0	0
14	Ion channels	4	20	22	8	14	0	0
15	GPCRs	45	129	137	13	92	19	13
16	Cytosolic-others	9	7	14	8	6	0	0
17	Electrochemical transporters	6	15	15	5	8	2	0
18	Nuclear receptors	15	16	20	5	13	2	0
19	Others	44	57	76	16	57	1	2

[a] Nineteen target families are listed following the CHEMBLdb classification scheme. For each family, the numbers of targets taken from CHEMBLdb, BindingDB, and the total number of targets are reported (taking target overlap between these databases into account). In addition, for each family, the number of targets is reported whose compound sets contain different numbers of scaffolds. Target family abbreviations: GPCR, G-Protein Coupled Receptor; Others, all none classified targets.

scaffolds. Accordingly, our analysis was ultimately based on 8,693 topologically distinct scaffolds represented by 26,664 compounds organized into 502 different target sets. For the assignment of targets to families, we followed the CDB classification scheme and combined targets available in CDB and BDB. Table 1 reports the 19 target families considered in our

This journal is © The Royal Society of Chemistry 2010

View Online

Published on 13 October 2010 on http://pubs.rsc.org | doi:10.1039/C0MD00156B
Downloaded on 08 February 2011

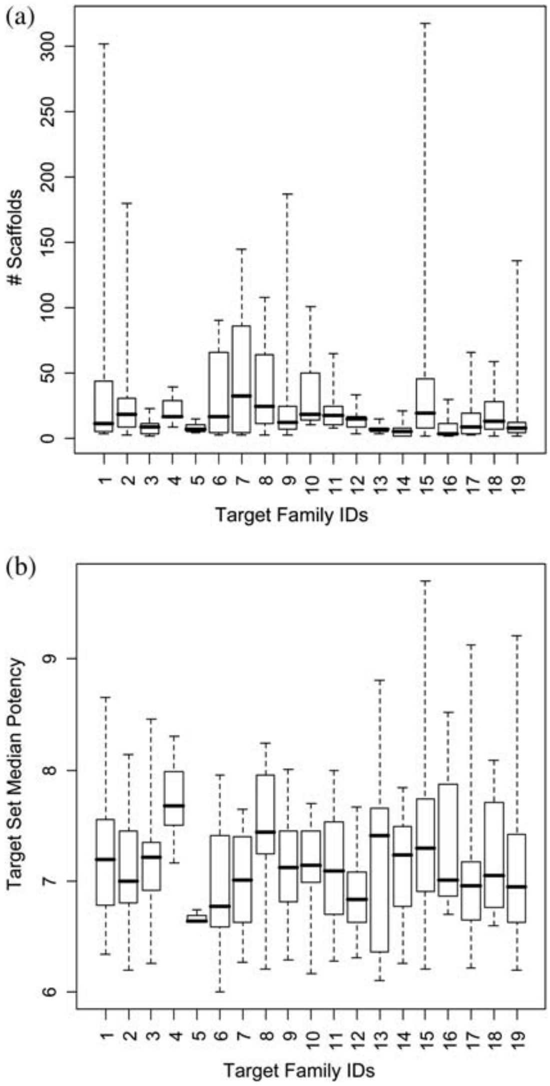


Fig. 2 Target family statistics. (**a**) Scaffold distribution and (**b**) target set median potency; presented as box plots. Target family IDs are according to Table 1. The box plots report the smallest value (bottom line), lower quartile (lower boundary of the box), median (thick horizontal line), upper quartile (upper boundary of the box), and the largest value (top line).

analysis that contained between three and 137 individual targets.

We first determined the number of distinct scaffolds present in each target set. The results are reported in Table 1 on a target family basis. Surprisingly, the majority of target sets were found to contain significant numbers of distinct scaffolds. A total of 354 target sets contained between five and 49 scaffolds, 42 target sets between 50 and 99, and 28 sets at least 100 scaffolds. Thus, the range of five to 49 scaffolds represents "average" scaffold diversity across current targets corresponding to average scaffold hopping potential. This is further illustrated by monitoring the scaffold distributions within target families (Fig. 2a). Many of these scaffolds were represented by compounds with in part very large potency differences (Fig. 2b).

A total of 70 target sets from twelve different target families (covering ~14% of the current target spectrum) were characterized by what we considered high to very high scaffold diversity, each containing between 50 and more than 300 topologically distinct scaffolds. We next analyzed these sets in more detail. Table 2 shows the top 30 targets ranked by scaffold numbers. Well-known pharmaceutical targets appear high on the ranking. These targets, which are also popular for virtual compound screening studies, include, for example, different adenosine and dopamine receptor subtypes and other GPCRs, protein kinases, and various proteases. These targets are chemically well explored. We have recently shown that more than 80% of scaffolds from currently available bioactive compounds are topologically equivalent and/or display substructure relationships.[16] Here we have exclusively focused on topologically distinct scaffolds, but we also determined substructure relationships between them, as reported in Table 2. For the target sets containing most scaffolds, at least approx. half of these scaffolds, but often more than 70% or 80% were found to be involved in substructure relationships (*i.e.* one scaffold is a substructure of another in the same set). From this point of view, it might not be very surprising that these targets have high scaffold hopping probability, also in

Table 2 Target sets ranked by scaffold number[a]

Target Name	#Sc	FamilyID	%Sc-in-Sub
Melanin-concentrating hormone receptor 1	318	15	57.2
Vascular endothelial growth factor receptor 1	302	1	66.2
Melanocortin receptor 4	207	15	63.3
Factor Xa	187	9	59.4
Cyclin-dependent kinase 2	180	2	70.6
Src tyrosine kinase	174	1	46.0
Thrombin	162	9	41.4
Adenosine receptor A3	160	15	73.1
Mu opioid receptor	157	15	86.6
Kappa opioid receptor	155	15	80.6
Delta opioid receptor	154	15	89.0
Cathepsin K	145	7	53.1
Serotonin receptor 5HT 1a	136	15	81.6
Acetylcholinesterase	136	19	53.7
Endothelial growth factor receptor	134	1	76.1
Dopamine receptor D2	134	15	50.7
Adenosine receptor A1	129	15	74.4
Mitogen-activated protein p38 alpha	129	2	64.3
Cathepsin S	129	7	55.8
Dipeptidyl peptidase 4	119	9	81.5
Adenosine receptor A2A	115	15	80.0
Serotonin transporter	110	15	43.6
Matrix metalloproteinase 3	108	8	61.1
Leukocyto-specific tyrosine kinase	106	1	69.8
Butyrylcholinesterase	104	19	51.9
Carbonic anhydrase II	101	10	55.4
Nociceptin receptor 1	100	15	72.0
Histamine H3 receptor	100	15	75.0
Protein kinase B Akt1	95	2	75.8
Matrix metalloproteinase 2	94	8	75.5

[a] The top 30 target sets ranked according to scaffold numbers are reported. For each set, the number of scaffolds (#Sc) and the percentage of these scaffolds (%Sc-in-Sub) that are involved in substructure relationships are reported.

View Online

Downloaded on 08 February 2011
Published on 13 October 2010 on http://pubs.rsc.org | doi:10.1039/C0MD00156B

benchmark calculations, and we would hence consider them “easy” virtual screening targets.

In order to assess scaffold hopping potential in quantitative terms, beyond scaffold numbers, we have also designed a function yielding a “hopping score” that incorporates compound potency information and is calculated over individual scaffold pairs in target sets. For a scaffold pair *ij* in target set *T*, all possible compound pairs C_{ij} are enumerated (*i.e.*, compounds in a pair contain scaffold *i* and *j*, respectively). For each scaffold pair *ij*, a “raw” score is calculated as:

$$score_{raw}(i,j) = (1 - sim(i,j)) * \frac{\sum \frac{P_{Ci} * P_{Cj}}{1 + |P_{Ci} - P_{Cj}|}}{|C_{ij}|}$$

Here *sim*(*i*,*j*) reports the Tanimoto similarity[17] of the two scaffolds in a pair calculated using MACCS structural keys[18] and $(1 - sim(i,j))$ is a measure of their dissimilarity. Because similarity calculations are only carried out for topologically distinct scaffolds, a topologically insensitive molecular representation such as MACCS keys can be used here. P_{Ci} and P_{Cj} are the potency values of compound C_i and C_j and $|C_{ij}|$ is the total number of all compound pairs representing the scaffold pair. Raw scores are transformed into conventional *Z*-scores by subtracting the sample mean and dividing standard deviation of the sample of all original raw scores. The *Z*-scores are then normalized with respect to a cumulative probability function in order to obtain final scores between 0 and 1.

It should be noted that the large-scale analysis of compound data inevitably involves at this stage the risk of comparing IC_{50} and Ki values, which represents a potential error source. However, for compounds from a series representing an individual scaffold, as used for our raw score calculations, consistent potency measurements are usually reported. In addition, it should also be noted that IC_{50} values are generally assay-dependent and hence often less reliable than Ki measurement. However, the potency weighting factor emphasizes large potency differences and the score is balanced by multiple pairwise contributions. Furthermore, the raw scores are converted into *Z*-scores. Taken together, these procedures make the scoring scheme fairly insensitive to limited fluctuations or inaccuracies of potency values.

On the basis of this scoring scheme, scaffold pairs will be prioritized (and obtain scores close to 1) that consist of scaffolds with low similarity yielding comparably potent compounds; identifying such scaffolds is a primary goal of scaffold hopping analysis.[9] By contrast, it is *a priori* not desired to facilitate scaffold transitions from highly potent to only weakly potent molecules. Therefore, not only target annotations, but also compound potency should be taken into consideration when assessing scaffold hopping potential on a large scale. For a target set *T*, the hopping score is then calculated as the median of all normalized scaffold pair scores:

$$score(T) = median\{score_{norm}(i,j) | i,j \in T; i < j\}$$

This score was calculated for the 70 target sets that were then ranked on the basis of decreasing scores, as reported in Table 3.

Table 3 Target sets ranked by scaffold score.[a]

Target Name	#Sc	FamilyID	MedianPot	PotRange	Score
Carbonic anhydrase II	101	10	7.7	3.7	0.849
Carbonic anhydrase IX	84	10	7.4	3.8	0.839
Carbonic anhydrase I	67	10	7.2	3.2	0.744
Matrix metalloproteinase 8	53	8	8.0	4.0	0.741
Cannabinoid receptor 1	84	15	7.6	3.8	0.719
Matrix metalloproteinase 13	76	8	7.9	4.8	0.705
Neurokinin receptor 1	70	15	8.9	4.7	0.698
Estrogen receptor alpha	59	18	7.4	3.6	0.693
Histone deacetylase 1	65	11	7.2	3.0	0.689
Matrix metalloproteinase 2	94	8	7.9	4.0	0.665
Matrix metalloproteinase 9	79	8	7.7	3.6	0.663
Cannabinoid receptor 2	74	15	7.4	3.9	0.660
Estrogen receptor beta	57	18	7.7	3.8	0.659
Matrix metalloproteinase 3	108	8	7.3	3.6	0.628
Norepinephrine transporter	51	17	7.1	3.2	0.584
Matrix metalloproteinase 6	68	15	7.8	3.4	0.568
Acetylcholinesterase	136	19	7.3	5.1	0.567
Dopamine transporter	66	17	7.1	3.4	0.550
Cyclin-dependent kinase 1	80	2	6.9	4.0	0.546
Vascular endothelial growth factor receptor 2	302	1	7.3	3.3	0.545
Histamine H3 receptor	100	15	7.9	4.1	0.538
Beta-secretase 1	89	6	7.4	3.5	0.519
Protein kinase B Akt1	95	2	7.5	3.8	0.517
Alpha-1a adrenergic receptor	73	15	8.3	4.3	0.516
Poly (ADP-ribose) polymerase-1	75	19	7.5	3.0	0.513
Adenosine receptor A3	160	15	7.6	3.9	0.501
Matrix metalloproteinase 1	90	8	7.3	4.0	0.490
Checkpoint kinase	62	2	7.7	3.9	0.486
Cyclin-dependent kinase 2	180	2	7.2	3.5	0.483
Serotonin transporter	110	15	7.9	4.4	0.477

[a] The top 30 target sets ranked according to scaffold hopping scores are reported. For each set, the number of scaffolds (**#**Sc), median compound potency (MedianPot), potency range (PotRange), and hopping score are reported.

This ranking differed from the one in Table 2 and highest scores were in this case obtained for carbonic anhydrases. Most of the target sets with significant scaffold hopping potential reported in Table 3 contained fewer than 100 scaffolds. Matrix metalloproteases and various GPCRs were also highly ranked. The rankings in Tables 2 and 3 were also combined on the basis of rank fusion. Table 4 shows the top 30 targets organized by increasing sum of ranks. These targets include many popular GPCRs, kinases, and proteases. Hence, on the basis of currently available compound data, these targets have highest scaffold hopping potential.

Vascular endothelial growth factor receptor-2 is the top-ranked target in Table 4 followed by carbonic anhydrase II.

This journal is © The Royal Society of Chemistry 2010

Published on 13 October 2010 on http://pubs.rsc.org | doi:10.1039/C0MD00156B
Downloaded on 08 February 2011

Table 4 Combined target set ranking[a]

						Rank		
Target Name	#Sc	FamilyID	MedianPot	PotRange	Score	Scaffold	Score	Sum
Vascular endothelial growth factor receptor 2	302	1	7.3	3.3	0.545	2	20	22
Carbonic anhydrase II	101	10	7.7	3.7	0.849	26	1	27
Acetylcholinesterase	136	19	7.3	5.1	0.567	13	17	30
Adenosine receptor A3	160	15	7.6	3.9	0.501	8	26	34
Cyclin-dependent kinase 2	180	2	7.2	3.5	0.483	5	29	34
Matrix metalloproteinase 3	108	8	7.3	3.6	0.628	23	14	37
Carbonic anhydrase IX	84	10	7.4	3.8	0.839	37	2	39
Matrix metalloproteinase 2	94	8	7.9	4.0	0.665	30	10	40
Cannabinoid receptor 1	84	15	7.6	3.8	0.719	37	5	42
Cathepsin K	145	7	7.6	5.5	0.464	12	32	44
Histamine H3 receptor	100	15	7.9	4.1	0.538	27	21	48
Cathepsin S	129	7	7.4	3.9	0.467	17	31	48
Src tyrosine kinase	174	1	7.3	3.8	0.406	6	42	48
Melanin-concentrating hormone receptor 1	318	15	7.6	4.0	0.397	1	48	49
Matrix metalloproteinase 13	76	8	7.9	4.8	0.705	44	6	50
Thrombin	162	9	7.1	6.0	0.404	7	44	51
Protein kinase B Akt1	95	2	7.5	3.8	0.517	29	23	52
Serotonin transporter	110	15	7.9	4.4	0.477	22	30	52
Mitogen-activated protein p38 alpha	129	2	7.6	4.3	0.436	17	36	53
Factor Xa	187	9	7.9	5.3	0.395	4	49	53
Matrix metalloproteinase 9	79	8	7.7	3.6	0.663	43	11	54
Endothelial growth factor receptor	134	1	7.3	5.5	0.435	15	39	54
Neurokinin receptor 1	70	15	8.9	4.7	0.698	49	7	56
Carbonic anhydrase I	67	10	7.2	3.2	0.744	54	3	57
Beta-Secretase 1	89	6	7.4	3.5	0.519	35	22	57
Cannabinoid receptor 2	74	15	7.4	3.9	0.660	46	12	58
Leukocyto-specific tyrosine kinase	106	1	7.9	5.0	0.436	24	36	60
Cyclin-dependent kinase 1	80	2	6.9	4.0	0.546	42	19	61
Matrix metalloproteinase 1	90	8	7.3	4.0	0.490	34	27	61
Dipeptidyl peptidase 4	119	9	7.5	4.0	0.408	20	41	61

[a] Target sets are ranked according to the sum of the scaffold number- and scaffold score-based rankings. The top 30 targets are listed. For each set, the number of scaffolds (#Sc), median compound potency (MedianPot), potency range (PotRange), scaffold hopping score, and individual ranks (Scaffold and Score) and sum (SUM) are given.

Fig. 3 shows scaffold pairs for these targets that yield high or low hopping scores. The top-scoring scaffold pairs display an astonishing degree of structural diversity, whereas low-scoring pairs are involved in close structural relationships. These observations are representative for many target sets that were found to contain a spectrum of topologically distinct scaffolds, ranging from closely related to virtually unrelated structures.

Finally, we also determined scaffold overlap between different target sets. The results are reported in Fig. 4 as a scaffold-based target network (drawn with Cytoscape[19]). Sixty of the 70 target sets shared one or more scaffolds with others. A total of 142 pair-wise target set relationships were detected among the 70 target sets; 106 of these relationships were formed exclusively within target families and 36 across different families. Substantial scaffold overlap between target sets was observed within the GPCR, kinase, and matrix metalloprotease target families. By contrast, inter-target family scaffold overlap was rather limited. These 142 relationships involved 1,298 scaffolds of a total of 5,232 scaffolds contained in the 70 target sets, *i.e.* ~25%. Hence, scaffold overlap was generally limited and the majority of scaffolds belonged to individual target sets.

In summary, in order to better understand how frequently scaffold hops occur in compounds active against different targets, we have systematically derived topologically distinct scaffolds for sets of compounds representing 502 targets belonging to 19 target families. The occurrence of different scaffolds in target sets provides an estimate for the likelihood that scaffold hops can be identified for given targets. In 354 of our target sets, between five and 49 distinct scaffolds were detected, providing a range for average scaffold hopping frequency. In 70 target sets, between 50 and 318 different scaffolds were found. A subset of these scaffolds was structurally highly diverse but yielded similarly potent compounds, thus meeting "ideal" scaffold hopping criteria. However, many other scaffolds (on average ~60% of all scaffolds in a target set) displayed well-defined substructure relationships. Thus, in these cases, it is not surprising that similarity-based virtual screening methods often display scaffold hopping potential (although scaffold hopping ability is usually considered the ultimate "proof" that a computational screening method is useful). By contrast, identifying scaffolds that are truly distinct is much more difficult, given the observed distributions of structurally related and unrelated scaffolds. However, on the

Downloaded on 08 February 2011
Published on 13 October 2010 on http://pubs.rsc.org | doi:10.1039/C0MD00156B

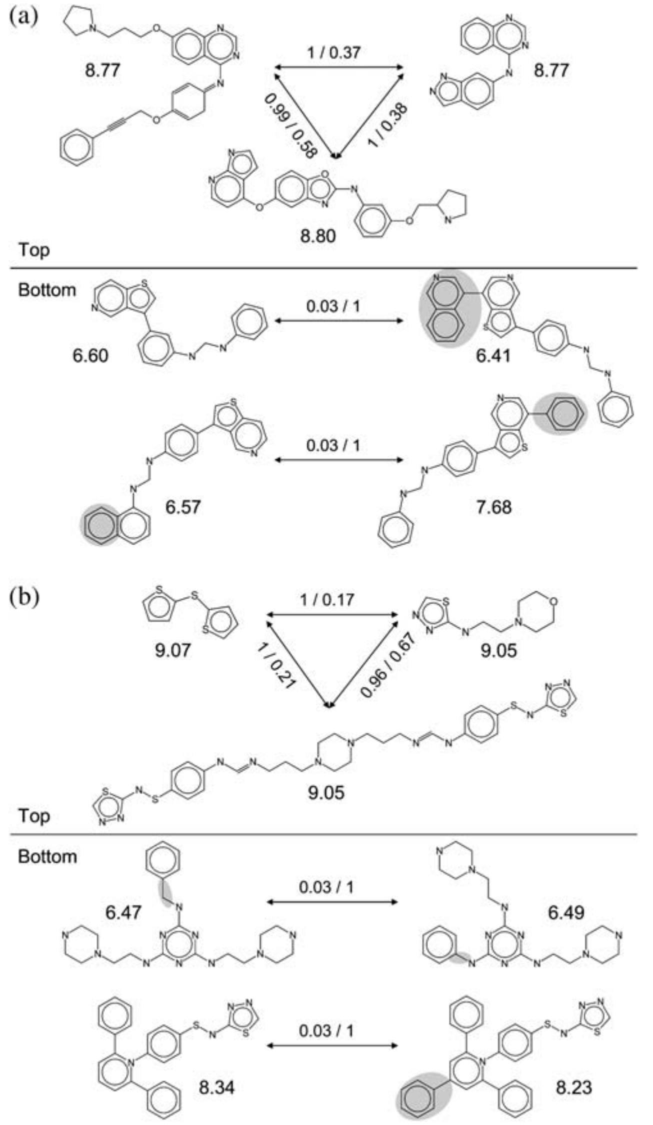


Fig. 3 Highly ranked target sets. Scaffold pairs with high (Top) and low (Bottom) hopping scores are shown for two top-ranked target sets; (**a**) vascular endothelial growth factor receptor 2 ligands and (**b**) carbonic anhydrase II inhibitors. For each set, three high scoring and two low scoring scaffold pairs are shown. For each scaffold, the median potency of the compounds it represents is reported. For each scaffold pair, the hopping score and MACCS Tanimoto similarity are reported. For example, 1/0.17 means that the scaffold pair has score of 1 and their Tanimoto similarity is 0.17. For low-scoring scaffold pairs, structural differences are highlighted.

basis of our analysis, we conclude that there is considerable scaffold hopping potential across the spectrum of currently available targets. Thus, searching for structurally diverse active compounds should be promising in many cases.

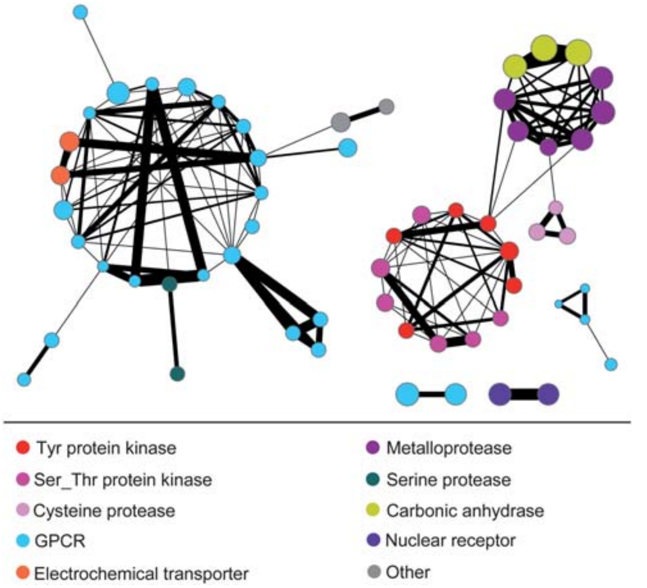


Fig. 4 Scaffold-based target network. Scaffold overlaps between target sets are viewed in a network representation. Nodes represent target sets that are connected by an edge if they share one or more scaffolds. The width of edges is scaled by scaffold numbers. Nodes are colored to reflect target family membership and their size is scaled by median scaffold hopping scores.

References

1 J. Brown and E. Jacoby, *Mini-Rev. Med. Chem.*, 2006, **6**, 1217–1229.
2 H. Zhao, *Drug Discovery Today*, 2007, **12**, 149–155.
3 S. Renner and G. Schneider, *ChemMedChem*, 2006, **1**, 181–185.
4 E. J. Barker, D. Buttar, D. A. Cosgrove, E. J. Gardiner, V. J. Gillet, P. Kitts and P. Willett, *J. Chem. Inf. Model.*, 2006, **46**, 503–511.
5 K. Tsunoyama, A. Amini, M. J. E. Sternberg and S. H. Muggleton, *J. Chem. Inf. Model.*, 2008, **48**, 949–957.
6 N. Wale, I. A. Watson and G. Karypis, *J. Chem. Inf. Model.*, 2008, **48**, 730–741.
7 S. Senger, *J. Chem. Inf. Model.*, 2009, **49**, 1514–1524.
8 M. Vogt, D. Stumpfe, H. Geppert and J. Bajorath, *J. Med. Chem.*, 2010, **53**, 5707–5715.
9 H. Geppert, M. Vogt and J. Bajorath, *J. Chem. Inf. Model.*, 2010, **50**, 205–216.
10 *MOE (Molecular Operating Environment)*; Chemical Computing Group Inc.: Montreal, Quebec, Canada, 2007.
11 *Scitegic Pipeline Pilot*, Student Edition, Version 6.1; Accelrys, Inc.: San Diego, CA, 2007.
12 CHEMBLdb. http://www.ebi.ac.uk/chembl/(accessed May 11, 2010).
13 T. Liu, Y. Lin, X. Wen, R. N. Jorissen and M. K. Gilson, *Nucleic Acids Res.*, 2007, **35**, D198–D201.
14 G. W. Bemis and M. A. Murcko, *J. Med. Chem.*, 1996, **39**, 2887–2893.
15 Y. Hu, A. M. Wassermann, E. Lounkine and J. Bajorath, *J. Med. Chem.*, 2010, **53**, 752–758.
16 Y. Hu and J. Bajorath, *ChemMedChem*, 2010, **5**, 1681–1685.
17 P. Willett, J. M. Barnard and G. M. Downs, *J. Chem. Inf. Comput. Sci.*, 1998, **38**, 983–996.
18 *MACCS Structural Keys*; Symyx Software: San Ramon, CA, 2005.
19 P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski and T. Ideker, *Genome Res.*, 2003, **13**, 2498–2504.

Summary

In this study, we have investigated how frequently scaffold hops occur across different targets on the basis of currently available compound data. The majority of compounds that were active against different targets contained between five to 49 topologically distinct scaffolds, representing average scaffold hopping frequency. A total of 70 targets were found to contain between 50 to 318 scaffolds and were further ranked according to the scaffold hopping scores. The top-scoring scaffold pairs displayed a significant degree of structural diversity, but comparable high potency. By contrast, low-scoring scaffold pairs were often involved in substructure relationships, indicating the limited structural variations.

Based on the observation that many scaffolds were structurally related, we next also investigated the scaffold diversity of currently available active compounds. The following chapters explored structural relationships among scaffolds in different ways.

Chapter 7

Structural and Potency Relationships Between Scaffolds of Compounds Active Against Human Targets

Introduction

In order to systematically assess the degree of structural diversity among scaffolds of active compounds, we performed a large-scale analysis to detect two types of structural relationships, i.e. a scaffold is a substructure of another scaffold and a scaffold shares the same topology with another one. Scaffolds involved in substructure relationships were further organized into sequential scaffold paths, whereas scaffolds yielding the same carbon skeletons were analyzed to identify compound activity cliff pairs. In addition, the potency direction among compound cliff pairs was also analyzed.

DOI: 10.1002/cmdc.201000272

Structural and Potency Relationships between Scaffolds of Compounds Active against Human Targets

Ye Hu and Jürgen Bajorath*[a]

The analysis of molecular scaffolds as core structures for drug-like compounds has traditionally played an important role in medicinal chemistry.[1,2] Scaffolds have been defined, for example, on the basis of synthetic and retrosynthetic criteria,[3] by focusing on ring systems,[4] or by applying a molecular hierarchy of R-groups, ring structures, and linkers between rings,[5] which represents the currently most widely applied definition. A variety of statistical analyses have been carried out to characterize scaffold distributions in drugs and pharmaceutically relevant active compounds[5–10] or screening libraries.[11] For example, we have previously analyzed the distribution of scaffolds in compounds at different pharmaceutical development stages.[10] In this study, we identified sets of scaffolds that preferentially occur in bioactive molecules, clinical trials compounds, or drugs. However, structural relationships between scaffolds found in these compounds were not explored.

Of particular interest in scaffold analysis is the ability to identify different molecular scaffolds that have the same specific activity, a task often referred to as scaffold hopping.[12–15] The exploration of scaffold hopping ability is a major focal point in both medicinal chemistry[12,13] and computational design.[14,15] For computational compound screening methods, the assessment of scaffold hopping potential has become one of the most important criteria.[15] The search for different scaffolds sharing the same activity, through chemical and/or computational means, is based on assumed scaffold diversity among specifically active compounds. However, the degree of scaffold diversity among currently available active compounds has not yet been investigated in a systematic manner. Therefore, we asked the question as to what currently available compound data might tell us about scaffold diversity. To these ends, a large-scale analysis of scaffolds in compounds that are active against currently available human drug targets was carried out. Structural relationships between scaffolds were systematically explored at different levels and related to compound potency distributions.

For our analysis, ChEMBL db (CDB)[16] represented the most relevant public domain compound source. CDB is a well-curated database that contains >500000 compounds with more than two million activity annotations and broad target coverage. The majority of CDB entries represent compound optimization data, that is, high-confidence activity annotations,[16] which we considered an important criterion for a global target-based assessment of scaffold diversity.

From CDB, 31158 compounds active against human targets were selected that had the highest target confidence level (i.e., CDB target confidence score 9) for direct interactions (target relationship type "D"). These compounds represented 577 different target sets. From these sets, a total of 12047 scaffolds were extracted according to Bemis and Murcko.[5] Hence, all R-groups were removed from ring systems, but linkers between rings were retained. Figure 1a shows the distribution of Bemis and Murcko scaffolds over all target sets. These sets contained from one to 615 scaffolds, but the majority of sets contained fewer than 50 scaffolds. Only 44 target sets consisted of compounds representing only a single scaffold. Figure 1b shows the distribution of compound potency ranges over target sets. Approximately 75% of the target sets contained active compounds with a potency spread of more than two orders of magnitude (the extreme case being renin inhibitors, with potency up to the attomolar (10^{-18} M) level). We compared the potency of compounds representing each scaffold. The median potency of all compounds representing a unique scaffold was calculated as the so-called "scaffold potency".

Scaffolds were also transformed into carbon skeletons (CSKs) by converting all heteroatoms to carbon atoms and all non-single bonds to single bonds. Thus, scaffolds producing the same CSK are topologically equivalent. On the basis of scaffolds and CSKs, two types of structural relationships were explored for each target set: 1) a scaffold is a substructure of another scaffold; 2) different scaffolds yield the same CSK. Accordingly, scaffold diversity was assessed at two levels. We considered scaffolds to be structurally diverse if they were 1) not involved in substructure relationships with others and 2) yielded unique CSKs (i.e., were topologically distinct). Figure 1c reports the ratio of scaffolds involved in two types of structural relationships over the total number of scaffolds for all target sets. For the analysis of substructural and CSK relationships between scaffolds, the benzene ring, the most generic scaffold, was not considered because benzene was a substructure of the majority of CDB scaffolds. For 465 of 533 target sets containing multiple scaffolds (~87%), structural relationships were detected, which was a rather unexpected finding. Moreover, for 107 target sets, all scaffolds were found to be involved in substructural and/or CSK relationships.

For those target sets containing at least two scaffolds and one structural relationship, we first determined the number of scaffold pairs that were involved in substructure relationships. Figure 2a shows that one or more substructural scaffold relationships were observed in ~85% of the target sets. A total of 261 sets contained more than five substructural relationships. Among these, vascular endothelial growth factor receptor (VEGFR) 2 antagonists contained 475 substructure pairs, the overall largest number.

---

[a] Y. Hu, Prof. Dr. J. Bajorath
Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität Bonn, Dahlmannstr. 2, 53113 Bonn (Germany)
Fax: (+49) 228-2699-341
E-mail: bajorath@bit.uni-bonn.de

© 2010 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
View this journal online at wileyonlinelibrary.com

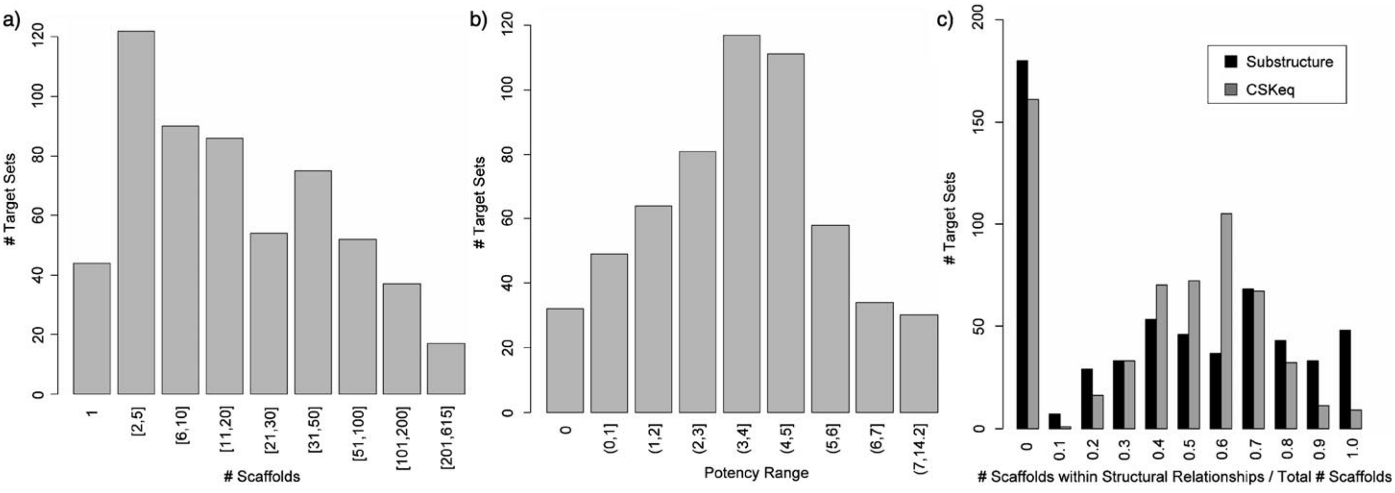


Figure 1. Target set statistics. For all 577 target sets, distributions are reported for a) the number (#) of scaffolds, b) compound potency range, and c) percentage of scaffolds that are either involved in substructural relationships (black) or that have topologically equivalent scaffolds (light grey); that is, [# Scaffolds with Structural Relationships]/[Total # Scaffolds].

In 397 target sets with one or more substructure relationships, a total of 9020 unique substructure relationships were detected. The corresponding scaffold pairs were analyzed for sequential substructure relationships on a per-target basis (i.e., A is a substructure of B, B of C, and C of D, etc.). It was found that ~80% of the substructure relationships had a path length of 1 (i.e., A is a substructure of B, but B is not a substructure of another scaffold), ~18% had a path length of 2 (i.e., A is a substructure of B and B of C), and 1.5% had of length of 3. Only five substructure paths of length 4 were identified. Three of these paths were overlapping and are shown in Figure 2b. Hence, the majority of scaffolds were found to be involved in substructure relationships, but sequential relationships involving more than three scaffolds were rare.

Next we identified scaffold pairs yielding the same CSKs. Figure 2c shows their distribution. In nearly 90% of all target sets, scaffold pairs yielding the same CSKs were detected. For 270 of these sets, more than five scaffold pairs were topologically equivalent (the maximum being 662 pairs for ligands of the melanin-concentrating hormone receptor 1). Thus, most of the

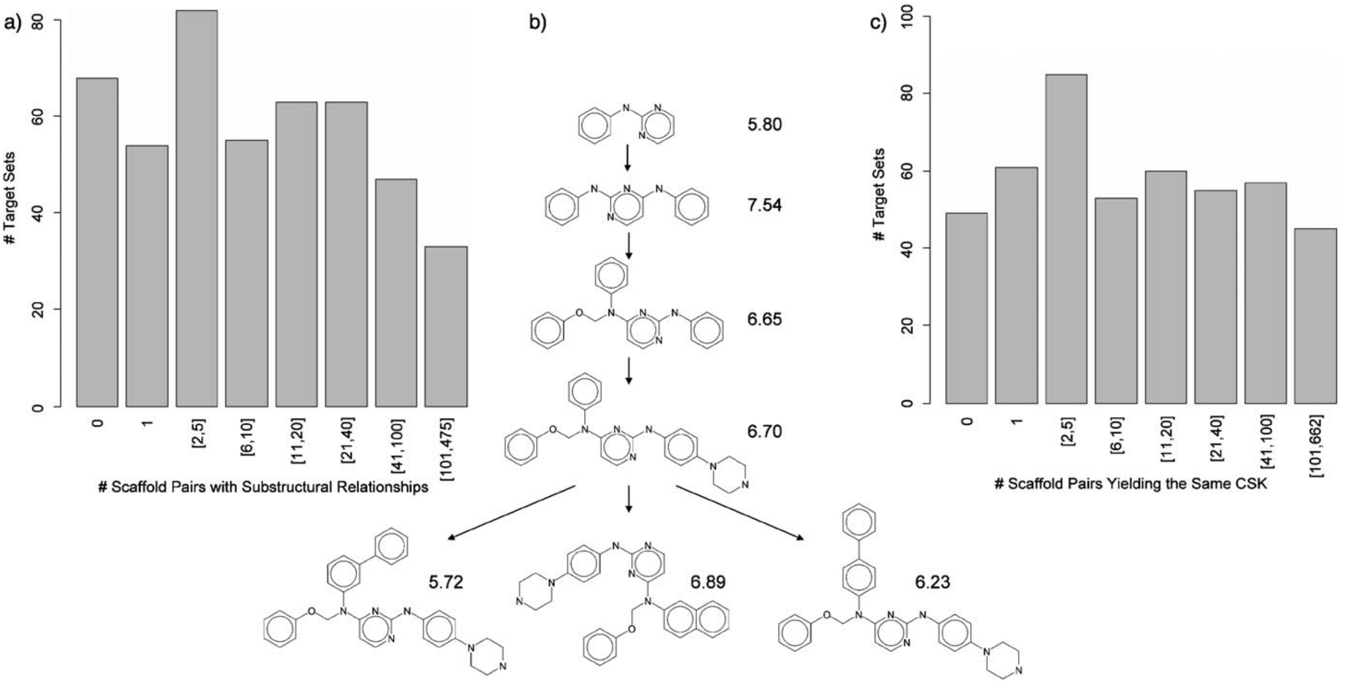


Figure 2. Distribution of scaffold relationships. a) For 465 target sets containing at least two scaffolds and one structural relationship, the number (#) of scaffold pairs with substructural relationships is reported. b) Three overlapping sequential substructural scaffold paths of length 4 are shown for the VEGFR2 antagonist set; the median compound potency value is reported for each scaffold. c) Distribution of the number of scaffold pairs yielding the same CSKs.

target sets contained a substantial number of topologically equivalent scaffolds.

For 348 of 533 target sets consisting of multiple scaffolds, both substructural relationships and CSK equivalences were observed. In 49 target sets, only substructural relationships were found, and in 68 sets, only CSK equivalences. In only 68 other target sets, no structural relationships were detected. Figure 3 a reports the distribution of scaffolds involved in different types of relationships over all target sets containing multiple scaffolds. As can be seen, sets with both substructure and CSK relationships typically consisted of many scaffolds. In contrast, targets with no scaffold relationships mostly contained only very few scaffolds. Figure 3 b shows an example of a target set with both substructural scaffold relationships and CSK equivalences. Figure 3 c and 3 d show representative examples of the only 14 target sets with more than five scaffolds but no structural relationships. In these cases, scaffolds were either completely unrelated (Figure 3 c) or related by symmetry and/or polymer character (Figure 3 d). Hence, the limited number of target sets with no substructural or CSK relationships included rather unusual active compounds. However, the majority of target sets were characterized by well-defined structural scaffold relationships involving most, if not all scaffolds.

We also carried out an analysis of structural relationships between scaffolds isolated from a set of 1586 clinical trials compounds extracted from the MDL Drug Data Report (MDDR),[17] a set of 2980 registered or launched drugs extracted from DrugBank[18] and the MDDR, as described previously,[10] and a set of 50 000 synthetic compounds randomly collected from ZINC.[19] These calculations were carried out without target set constraints because these compounds could not be systematically organized into defined target sets based on activity annotations. Thus, as a control, we also calculated scaffold relationships for the entire collection of bioactive CDB compounds without applying the target set organization. For all of these sets of medicinal chemistry relevant compounds, $>90\%$ of the scaffolds were found to be involved in structural relationships. Hence, scaffold diversity, as defined herein, among these compounds was generally much lower than we anticipated, and bioactive compounds mirrored this low level of structural diversity. When target set constraints were applied, structural relationships among scaffolds were decreased; importantly, however, still $>80\%$ of scaffolds extracted from specifically active CDB compounds displayed defined structural relationships.

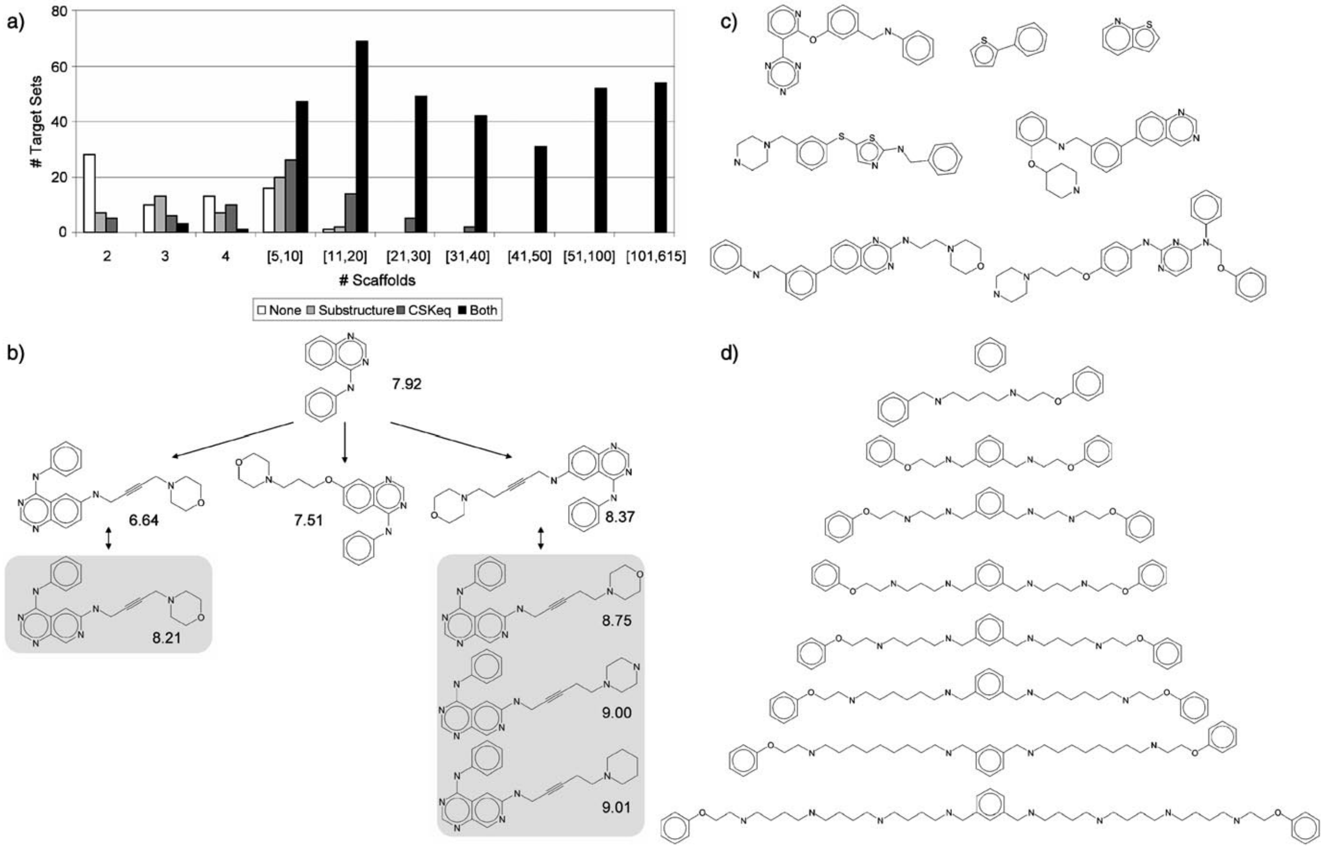


Figure 3. Target sets with scaffolds having defined structural relationships. a) The number of scaffolds involved in different types of relationships over all target sets containing multiple scaffolds; no structural relationship ("None"; white), only substructural relationships ("Substructure"; light grey), only CSK equivalence ("CSKeq"; dark grey), or substructure and CSK relationships ("Both"; black). b) Scaffolds from an exemplary target set (protein tyrosine kinase erbB-4 inhibitors) are shown that contain both types of structural relationships. Arrows denote substructure relationships, and grey shading indicates CSK equivalence. Median compound potency values are reported for all scaffolds. c), d) Scaffolds are shown for two representative target sets without structural relationships including inhibitors of c) tyrosine protein kinase BTK and d) glutathione S-transferase A1.

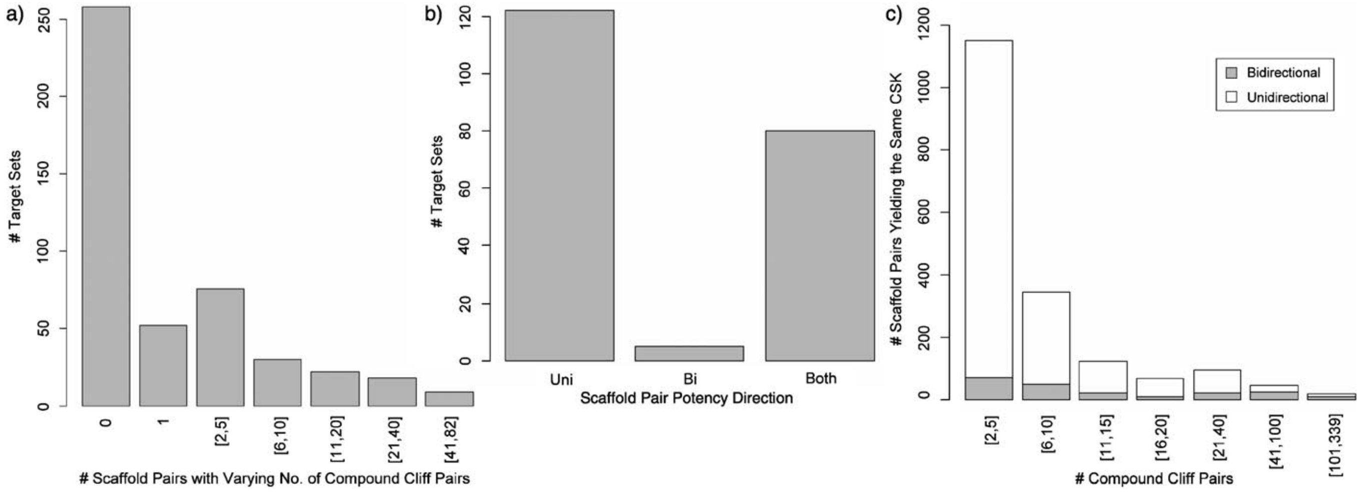


Figure 4. Distribution of compound cliff pairs. a) The target set distribution of scaffold pairs with varying numbers of compound cliff pairs is shown. b) The number of target sets that contain only scaffold pairs with unidirectional potency, bidirectional potency, or both. c) The distribution of uni- and bidirectional scaffold pairs represented by multiple compound cliff pairs.

Therefore, these structurally related scaffolds corresponded to compounds with well-defined activity against currently available targets.

We previously showed that different scaffolds have different propensities to form activity cliffs[20] (an activity cliff is formed by two or more structurally similar compounds with marked difference(s) in potency[21]). Therefore, we also investigated activity cliff formation among topologically equivalent scaffolds. For each scaffold pair yielding the same CSK, all possible compound pairs were selected that represented one or the other scaffold. If the potency of compounds in a pair differed by at least two orders of magnitude, it was considered to form an activity cliff (and termed “compound cliff pair”). Figure 4a reports the number of topologically equivalent scaffold pairs that were represented by multiple compound cliff pairs. In 258 target sets, no such scaffold pairs were found. However, 79 target sets contained more than five scaffold pairs represented by multiple compound cliff pairs.

We then analyzed the potency “direction” among compound cliff pairs; that is, we examined whether compounds representing one of two topologically equivalent scaffolds were always more potent than compounds representing the other (“unidirectional”) or not (“bidirectional”). Figure 4b reports the potency direction among these pairs on a target set basis. A total of 122 target sets were exclusively unidirectional in compound cliff pair potency distribution, and only five sets were exclusively bidirectional; the remaining 80 targets contained both uni- and bidirectional scaffold pairs. Thus, a notable tendency for unidirectional potency among topologically equivalent scaffold pairs with activity cliff potential was observed. Figure 4c shows the distribution of compound cliff pairs over topologically equivalent scaffold pairs, which further corroborates this trend. A total of 1564 unique scaffold pairs with at least two compound cliff pairs were found in 202 target sets, and 1398 of these scaffold pairs displayed unidirectional potency; 1218 unidirectional pairs occurred in single target sets, and the remaining 180 pairs in multiple sets. In contrast, only 166 bidirectional scaffold pairs were found in 85 target sets, 145 of which only occurred in single sets. The uni- and bidirectional scaffold pair sets shared 46 pairs. Thus, cliff-forming compounds representing these 46 pairs showed target-specific differences in activity. The prevalence of topologically equivalent unidirectional scaffolds indicates that the choice of these scaffolds plays an essential role for achieving high compound potency.

In Figure 5, exemplary scaffolds with defined structural relationships are shown that correspond to compounds forming significant activity cliffs. These scaffolds display substructural relationships (Figure 5a) or are topologically equivalent (Figure 5b).

In summary, we have shown that the majority of currently available molecular scaffolds representing compounds active against human targets display substructure relationships and/or are topologically equivalent. Of the 12047 scaffolds analyzed herein, a total of 9993 scaffolds were involved in these well-defined structural relationships. Thus, true scaffold diversity among active compounds was limited. However, we have also shown that this applies, in the absence of target set constraints, to randomly selected medicinal chemistry relevant and druglike compounds or drugs. Hence, bioactive compounds essentially mirror general scaffold relationships in currently available small molecules. These findings might suggest that biologically relevant chemical space is small, perhaps even smaller than previously thought, and that the identification of distinctly different compound structures sharing a specific target activity might often be difficult. However, the high frequency of substructural and topological relationships between currently available medicinal chemistry relevant scaffolds might also suggest that many medicinal chemistry efforts build upon prior knowledge and modify known structural motifs. Regardless, among topologically equivalent scaffolds with the potential to form activity cliffs, many scaffolds display clear preferences

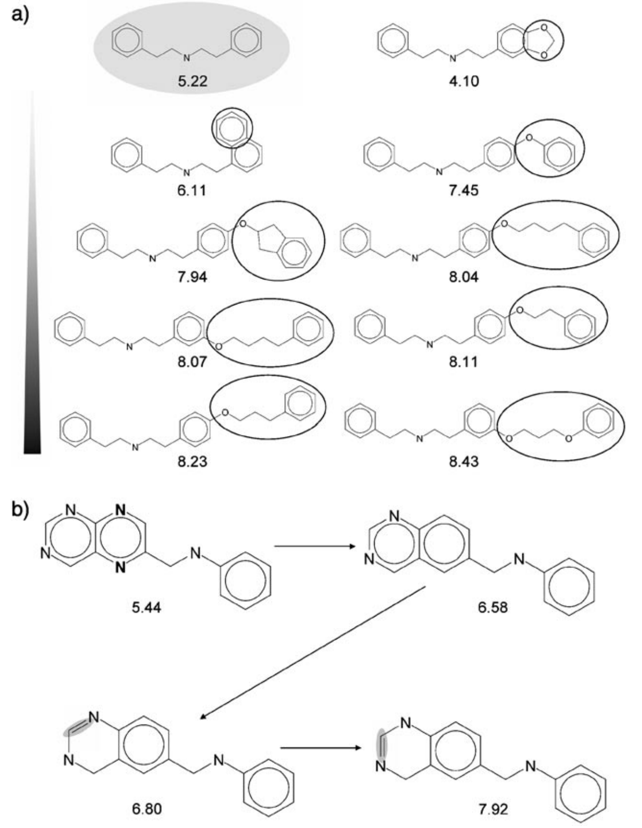


Figure 5. Activity-cliff-forming scaffolds with defined structural relationships. a) Ten scaffolds from inhibitors of the β-2 adrenergic receptor are shown that are involved in substructural relationships. The root scaffold is shown on a grey background, and nine scaffolds containing the root as a substructure are arranged according to increasing scaffold potency values (i.e., the median potency of compounds corresponding to each scaffold). Structural regions outside the root substructure are circled. b) Four scaffolds extracted from thymidylate synthase inhibitors that yield the same carbon skeleton are shown. Differences in heteroatom positions and bond orders are highlighted, and scaffold potency values are reported.

over others to yield highly potent compounds. These observations indicate that structurally related scaffolds might yield rather different target-dependent compound activity. Thus, minor differences in scaffold structures might substantially affect compound optimization efforts.

All calculations were carried out with in-house written Perl or Scientific Vector Language (SVL)[22] scripts and Pipeline Pilot[23] programs. For CDB compounds with multiple potency measurements reported for the same target (either K_i or IC_{50} values), the geometric mean was calculated to yield the final potency value. Compounds, scaffolds, and CSKs were represented in SMILES[24] format for processing.

Keywords: activity cliffs · compound potency · database mining · molecular scaffolds · substructure relationships · topological equivalence

[1] C. Merlot, D. Domine, C. Cleva, D. J. Church, *Drug Discovery Today* **2003**, *8*, 594 – 602.

[2] L. Costantino, D. Barlocco, *Curr. Med. Chem.* **2006**, *13*, 65 – 85.

[3] X. Q. Lewell, D. B. Judd, S. P. Watson, M. M. Hann, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511 – 522.

[4] P. Ertl, S. Jelfs, J. Mühlbacher, A. Schuffenhauer, P. Selzer, *J. Med. Chem.* **2006**, *49*, 4568 – 4573.

[5] G. W. Bemis, M. A. Murcko, *J. Med. Chem.* **1996**, *39*, 2887 – 2893.

[6] D. M. Schnur, M. A. Hermsmeier, A. J. Tebben, *J. Med. Chem.* **2006**, *49*, 2000 – 2009.

[7] J. J. Sutherland, R. E. Higgs, I. Watson, M. Vieth, *J. Med. Chem.* **2008**, *51*, 2689 – 2700.

[8] Y. Hu, A. M. Wassermann, E. Lounkine, J. Bajorath, *J. Med. Chem.* **2010**, *53*, 752 – 758.

[9] J. Wang, T. Hou, *J. Chem. Inf. Model.* **2010**, *50*, 55 – 67.

[10] Y. Hu, J. Bajorath, *ChemMedChem* **2010**, *5*, 187 – 190.

[11] M. Krier, G. Bret, D. Rognan, *J. Chem. Inf. Model.* **2006**, *46*, 512 – 524.

[12] N. Brown, E. Jacoby, *Mini-Rev. Med. Chem.* **2006**, *6*, 1217 – 1229.

[13] H. Zhao, *Drug Discovery Today* **2007**, *12*, 149 – 155.

[14] S. Renner, G. Schneider, *ChemMedChem* **2006**, *1*, 181 – 185.

[15] H. Geppert, M. Vogt, J. Bajorath, *J. Chem. Inf. Model.* **2010**, *50*, 205 – 216.

[16] ChEMBL db, http://www.ebi.ac.uk/chembl/ (accessed May 11, 2010).

[17] MDL Drug Data Report (MDDR), Symyx Software, San Ramon, CA (USA), **2007**.

[18] D. S. Wishart, C. Knox, A. C. Guo, D. Chen, S. Shrivastava, D. Tzur, B. Gautam, M. Hassanali, *Nucleic Acids Res.* **2008**, *36*, D901 – D906.

[19] J. J. Irwin, B. K. Shoichet, *J. Chem. Inf. Model.* **2005**, *45*, 177 – 182.

[20] Y. Hu, J. Bajorath, *J. Chem. Inf. Model.* **2010**, *50*, 500 – 510.

[21] G. M. Maggiora, *J. Chem. Inf. Model.* **2006**, *46*, 1535.

[22] MOE (Molecular Operating Environment), Chemical Computing Group Inc., Montreal, QC (Canada), **2007**.

[23] Scitegic Pipeline Pilot, Student Edition, v. 6.1, Accelrys Inc., San Diego, CA (USA), **2007**.

[24] D. Weininger, *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31 – 36.

Received: June 30, 2010
Revised: July 27, 2010
Published online on August 18, 2010

© 2010 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

Summary

We have carried out a systematic analysis of structural relationships between scaffolds of compounds active against human targets. The majority of target sets (∼87%) contain scaffolds involved in substructure relationships and/or having a topology equivalent to other scaffolds in the same set, suggesting limited structural diversity. This also indicated that currently utilized chemical space might be smaller than previously anticipated and that chemical modifications were usually made on the basis of known chemical classes. Furthermore, we have detected compound cliff pairs formed by topologically equivalent scaffolds in ∼50% of target sets. Many of these scaffolds displayed a tendency to produce highly potent compounds.

Further extending the scaffold classification scheme introduced here, we also adopted a well-known hierarchical scaffold classification scheme, i.e. the *Scaffold Tree*, which also represents substructure relationships along the tree branches. By comparing substructure relationships identified in our analysis with those implemented in scaffold trees, we increased information content of this data structure.

Chapter 8

Combining Horizontal and Vertical Substructure Relationships in Scaffold Hierarchies for Activity Prediction

Introduction

A rule-based *Scaffold Tree* classification scheme organizes original scaffolds of active compounds and their derivative scaffolds (virtual scaffolds) that are not contained in original compounds in hierarchies. Leaf-to-root substructure relationships originating from the *Scaffold Tree* structure were compared with leaf-to-leaf substructure relationships that were often not described by tree hierarchies. These two substructure relationships were found to be complementary in nature and therefore combined to prioritize virtual scaffolds for further activity prediction. Scaffolds having high-priority on the basis of these complementary relationships were mapped to external sets of active compounds and a number of correct activity predictions were obtaind.

JOURNAL OF
CHEMICAL INFORMATION
AND MODELING

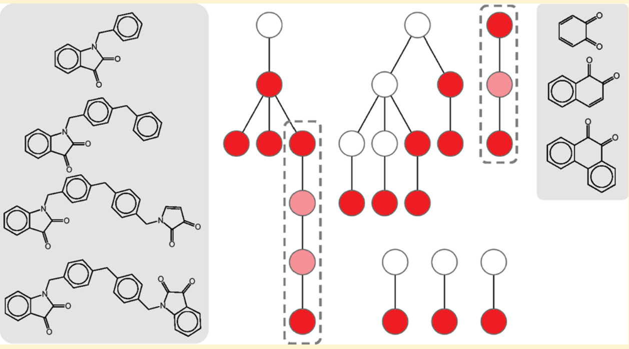
ARTICLE

pubs.acs.org/jcim

Combining Horizontal and Vertical Substructure Relationships in Scaffold Hierarchies for Activity Prediction

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: For a systematic exploration of structural relationships between molecular scaffolds, ~24,000 unique scaffolds were extracted from 458 different target sets. Substructure relationships between these scaffolds were systematically determined. The scaffold tree data structure was utilized to study structural relationships between original scaffolds and derivative scaffolds obtained by rule-based decomposition. Leaf-to-root substructure relationships that resulted from rule-based decomposition were compared to leaf-to-leaf relationships between original scaffolds most of which were not part of the scaffold tree hierarchy. Decomposed scaffolds not contained in active target set compounds were prioritized on the basis of hierarchical scaffold patterns and additional substructure relationships. For high-priority virtual scaffolds, activity predictions were carried out, and these scaffolds were often found in external test compounds having the predicted activity. Taken together, our results suggest that leaf-to-root substructure relationships in scaffold trees should best be complemented with additional substructure relationships to determine high-priority virtual scaffolds for activity prediction.

■ INTRODUCTION

The analysis of molecular scaffolds or frameworks of bioactive compounds is highly relevant for a number of tasks in medicinal chemistry including, for example, the identification of potentially privileged substructures[1,2] or of target class-selective chemotypes.[3] Furthermore, the search for different scaffolds yielding compounds with similar activity, often referred to as scaffold hopping, is another important goal.[4,5]

A scaffold is often obtained from an active compound by removing all substituents from ring systems and from linker segments between rings, following the Bemis and Murcko definition.[6] Alternatively, scaffolds might also be defined on the basis of retrosynthetic criteria[7] or other chemical rules. Large-scale scaffold analyses have been carried out, for example, to assess the structural diversity of synthetic compounds[8] and screening libraries,[9] survey heteroaromatic scaffolds in bioactive compounds,[10,11] or study the distribution of scaffolds in compounds at different pharmaceutical development stages.[12]

Since conventional scaffolds[6] contain all rings and linkers between rings, the addition of any ring to a compound (e.g., a phenyl substituent) always constitutes a new scaffold, given the underlying hierarchical scaffold definition (although the compound might in such cases better be considered an analog). This is often considered a potential caveat in scaffold analysis.[13] Accordingly, scaffold classification schemes have been introduced that do not predominantly focus on core structures, but chemical transformations,[13] similar to the matched molecular pair concept,[14] or that organize ring systems after removal of linkers.[15]

Another scaffold organization scheme was introduced that iteratively removes rings from initially derived Bemis and Murcko-like scaffolds, starting at peripheral and moving to more central positions until only a single ring remains.[16] Here rings are not only removed that are connected by linkers but also from condensed ring systems by dividing them into individual (parental) rings. A set of generally applicable chemical rules is applied to prioritize rings for iterative removal. For scaffolds from any source, these procedures generate a hierarchy where initially derived scaffolds ("leaves") are systematically reduced until an individual "root" ring remains. For sets of active compounds, the resulting pathways of this "leaf-to-root" hierarchy are displayed as so-called Scaffold Trees[16] (STs) that currently probably represent the most general data structure to hierarchically organize scaffold populations. ST "leaf" scaffolds differ from Bemis and Murcko scaffolds only in that double bonded atoms (e.g., carbonyl oxygens) attached to rings or linkers are retained as part of the scaffold. Given the rule-based decomposition of ring systems, the ST hierarchy typically contains scaffolds that are not contained in the original set of active compounds, so-called virtual scaffolds.[16] Thus, STs can be utilized to predict biological activities of such scaffolds.[16,17] For activity prediction, STs of different compound sets can also be merged by mapping shared scaffolds and combining the pathways they are involved in.[18]

Received: November 15, 2010

ACS Publications © XXXX American Chemical Society

dx.doi.org/10.1021/ci100448a | J. Chem. Inf. Model. XXXX, XXX, 000–000

We have been interested in scaffold hierarchies to systematically analyze substructure relationships between scaffolds and their relevance for biological activity. This analysis was inspired by a previous finding that 71% of bioactive scaffolds were involved in defined substructure relationships (i.e., A is a substructure of B).[19] We reasoned that it might be possible to reconcile and further explore these structural relationships on the basis of scaffold hierarchies. In this context, STs have been of particular interest because they capture substructure relationships along decomposition pathways (i.e., from leafs to roots), which we, for the purpose of our analysis, term "vertical" relationships. However, the substructure relationships we identified previously are, in the context of the ST hierarchy, leaf-to-leaf relationships, which we term "horizontal". Such relationships have thus far not been explicitly considered in ST analysis. Therefore, we have systematically analyzed to what extent vertical and horizontal substructure relationships between scaffolds complement each other. For this purpose, a large-scale analysis of target set-dependent scaffold hierarchies has been carried out. Prioritized candidate scaffolds that were not contained in target set compounds have been mapped to external compound sources and their biological activity has been predicted.

■ MATERIALS AND METHODS

For scaffold generation, bioactive compounds were extracted from the ChEMBL[20] database (CDB) and BindingDB[21] (BDB). These databases are two major publicly available repositories of active compounds from medicinal chemistry sources with defined target and activtiy annotations. ST scaffolds were generated using the Scaffold Tree Generator program.[16] Figure 1 illustrates the rule-based decomposition of ST scaffolds and the formation of a tree branch. Resulting STs were drawn with Cytoscape.[22] Hierarchically organized scaffolds were prioritized on the basis of defined scaffold patterns and substructure relationships and mapped to scaffolds of active compounds from the Molecular Drug Data Report[23] (MDDR) and approved drugs from DrugBank.[24] The scaffold analysis reported herein was carried out with in-house generated Molecular Operating Environment[25] (MOE) Scientific Vector Language (SVL), Perl, and Pipeline Pilot[26] scripts.

Upon publication the scaffold hierarchies generated for our analysis can be freely obtained via the following URL: http://www.lifescienceinformatics.uni-bonn.de (please, see the "Downloads" section).

■ RESULTS AND DISCUSSION

Compound and Scaffold Statistics. From the pool of CDB and BDB compounds, we extracted compound activity classes (with specific target annotations) under the condition that each activity class (target set) had to contain at least 10 compounds with at least 1 μM potency. On the basis of these criteria, we obtained 458 target sets containing a total of 34,916 active compounds that yielded 23,879 unique ST scaffolds.

Scaffold Hierarchies. For each of our 458 target sets, an ST was generated. Figure 2 shows a representative example and illustrates how different leaf scaffolds form (or do not form) converging scaffold pathways toward a root scaffold, i.e. an individual ring. A consequence of this hierarchical scaffold decomposition scheme is that not all ST scaffolds are represented by active compounds from which leaf scaffolds are derived. This leads to the distinction of "real" ST scaffolds (R) that are contained in active compounds and "virtual" scaffolds (V) that do not occur in source compounds, as illustrated in Figure 2. Following this classification scheme, the 23,879 unique scaffolds comprising 458 target set STs yielded 13,377 real scaffolds and 10,502 virtual scaffolds.

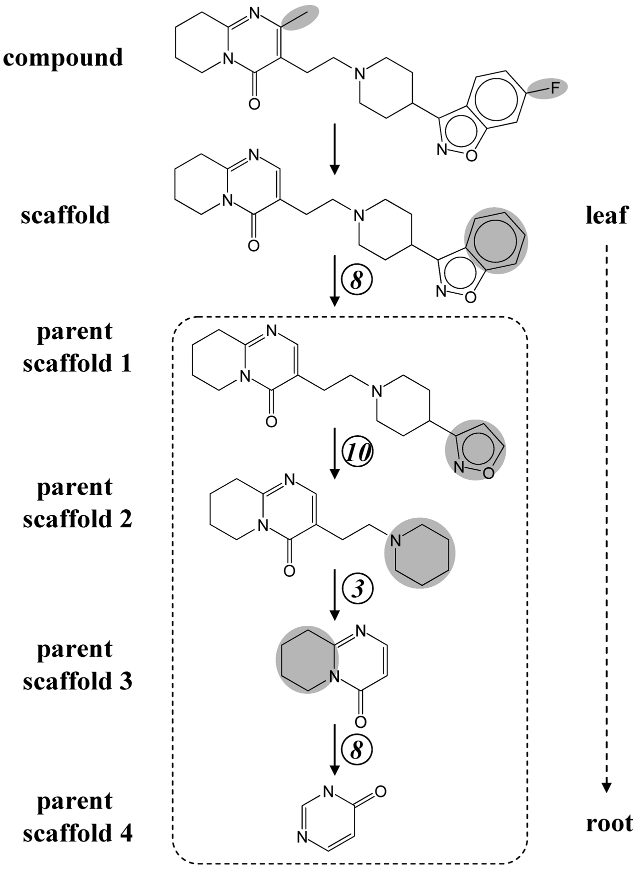


Figure 1. Scaffold hierarchy. Shown is an exemplary scaffold branch[16] generated from an active compound. The leaf scaffold is extracted from the compound by removing all single-bond substituents from rings or linkers between rings. During each decomposition step, a smaller (parent) scaffold is generated. One ring is iteratively removed per step from the current scaffold according to 13 predefined chemical rules[16] until only a single ring remains as the root scaffold. In this example, parent scaffold 1 was generated by removing a benzene ring on the basis of rule 8 (i.e., "remove rings with the least number of heteroatom first"). In the following steps, rule 3 ("choose a parent scaffold having the smallest number of acyclic linker bonds") and rule 10 ("smaller rings are removed first") were applied and, finally, rule 8 again to yield the root scaffold.

Substructure Relationships. Scaffold trees capture hierarchical leaf-to-root substructure relationships between scaffolds along decomposition pathways but do not explicitly account for horizontal leaf-to-leaf substructure relationships. A central point of our study has been to determine to what extent such horizontal substructure relationships are implicitly captured by the ST data structure. Therefore, we first identified all pairs of leaf scaffolds that represented a defined substructure relationship. For this analysis, the most generic scaffold, the benzene ring, was not considered. As reported in Table 1, we detected 13,181 pairs that involved a total of 9712 leaf scaffolds, i.e. 73% of all original

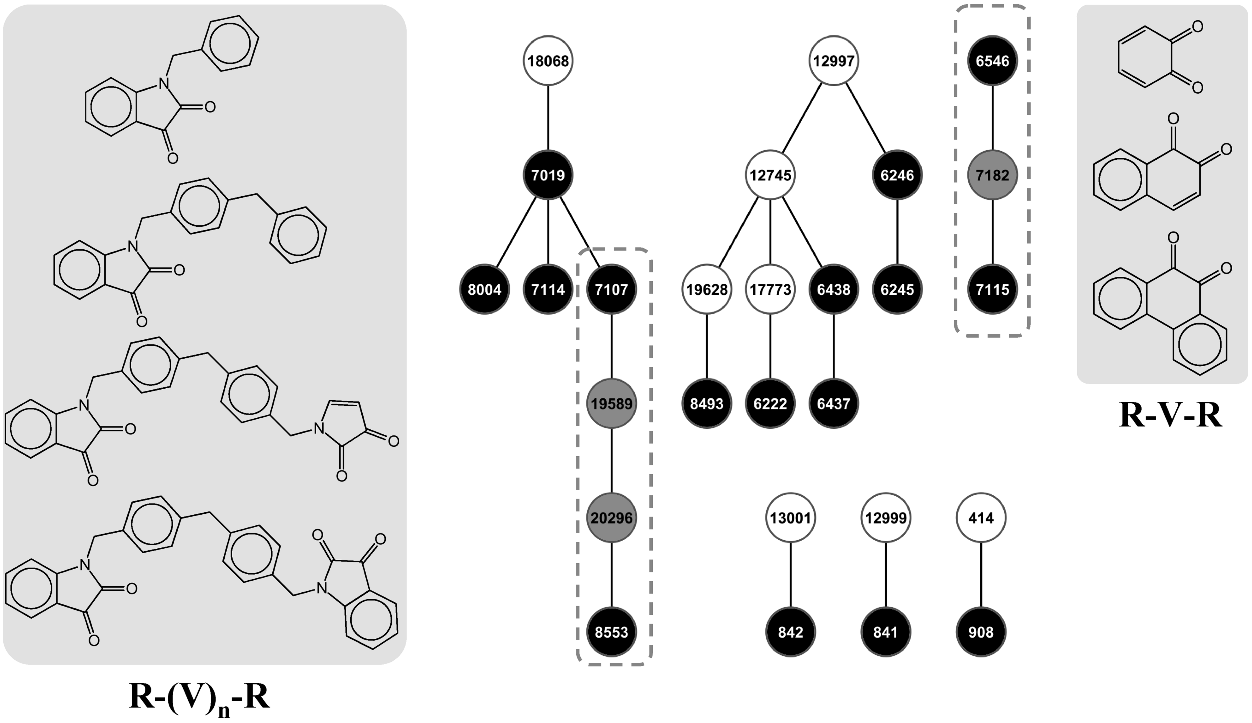


Figure 2. Real and virtual scaffolds. An exemplary scaffold tree is shown for the carboxylesterase-2 inhibitor set. Nodes represent scaffolds that are labeled with IDs and gray-scaled according to different scaffold types including "real" scaffolds (black), "virtual" scaffolds (white), and "prioritized virtual" scaffolds (gray). Edges connect scaffolds in a tree branch (decomposition pathway). In this orientation, leaf scaffolds are at the bottom and root scaffolds at the top. Two scaffold branches are highlighted using dashed rectangles. In the left branch, virtual scaffolds 19589 and 20296 are located between two "real" scaffolds 7107 and 8553, forming an R-(V)$_2$-R pattern. In the right branch, virtual scaffold 7182 has two neighboring real scaffolds (6546 and 7115), forming an R-V-R pattern. Scaffold sequences comprising these patterns are displayed on a light gray background. Virtual scaffolds from such patterns are considered prioritized virtual scaffolds for activity prediction.

Table 1. Substructure Relationships[a]

substructure relationship	scaffold pairs	scaffolds
horizontal	13,181	9712 (73%)
vertical	4217	5205 (39%)

[a] Each scaffold pair represents a substructure relationship. Horizontal relationships represent leaf-to-leaf and vertical leaf-to-root substructure relationships between scaffolds. Vertical relationships are determined by the ST hierarchy.

scaffolds. Thus, the majority of all leaf ST scaffolds were involved in pairwise substructure relationships (similar to 71% of all Bemis and Murcko scaffolds extracted from CDB compounds[19]). We then determined how many of these scaffold pairs also represented vertical ST substructure relationships. Only 4217 of all 13,181 pairs (32%) were detected in ST pathways. These pairs involved a total of 5205 scaffolds, i.e. 39% of all leaf scaffolds (Table 1). Thus, for all 458 target sets, the STs only contained about one-third of the substructure relationships that were present between the original scaffolds. On the basis of these findings, we then asked the question how the additional substructure relationship information might be utilized for tree analysis.

Substructure Information Content. Substructure relationships can be added to the ST structure by annotating trees with nonpathway substructure pair information, as illustrated in Figure 3. The hierarchy of the exemplary ST on the left in Figure 3 contains three pairwise substructure relationships, and scaffolds 3 and 4 are each involved in two pairs. Within a branch, a scaffold can be a part of at most two pairs and hence these scaffolds cannot be further distinguished by pair numbers. However, on the right in Figure 3, the tree is annotated with all additional substructure relationships involving leaf scaffolds (two in this case). Now scaffold 2 is also involved in two pairs, and leaf scaffold 1 and root scaffold 4 are each involved in three pairs. Thus, taking this additional information into account, scaffolds can be further differentiated by the number of substructure pairs they participate in and scaffolds involved in most pairs can be prioritized on the basis of substructure information content. We next evaluated how added substructure information might affect activity predictions. For this purpose, all possible pairs between virtual and real scaffolds were systematically analyzed.

Pairs of Virtual and Real Scaffolds. One of the most interesting aspects of the ST data structure is the opportunity to predict the activity of virtual scaffolds.[16,17] Prime candidates for activity prediction are virtual scaffolds that are proximal to real scaffolds in the tree because of their structural relatedness,[17] which represents a rather intuitive approach, leading to a number of successful predictions.[17,18]

In order to systematically explore relevant scaffold pairings, we isolated all V-R scaffold pairs (i.e., pairs formed by a virtual and a real scaffold) from all target set STs. As reported in Table 2, 53,220 V-R pairs were found in 442 target sets. When we limited the magnitude of structural differences within a pair to a maximum of two rings, the number of V-R pairs was reduced to

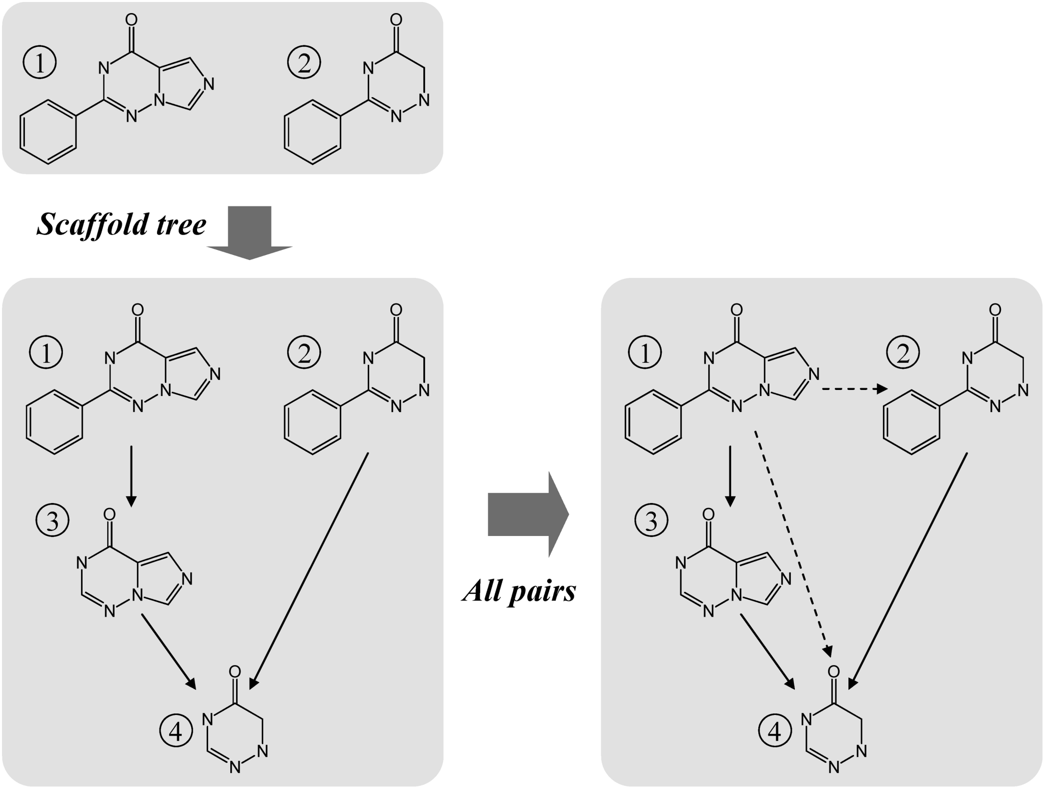


Figure 3. Scaffold pairs representing substructure relationships. For two scaffolds 1 and 2, the parent scaffolds 3 and 4 are generated by iteratively removing one ring from each child. The resulting scaffold tree contains three scaffold pairs that form substructure relationships, i.e. 1−3, 3−4, and 2−4 (solid arrows). However, by examining additional substructure relationships for leaf scaffolds, two additional pairs, i.e. 1−2 and 1−4 (dashed arrows), are identified. Thus, scaffolds 1 and 4 are now each involved in three substructure relationships, and this additional information can be used to further prioritize scaffolds.

Table 2. V-R Scaffold Pairs[a]

V-R pair category	V-R pairs	virtual scaffolds	real scaffolds	target sets
$\|V - R\| = n$ rings	53,220	9668	12,541	442
$\|V - R\| \leq 2$ rings	25,664	8750	11,799	442
$\|V - R\| = 1$ ring	10,886	6584	8933	440

[a] Reported are the total number of V-R scaffold pairs within ST hierarchies ($|V - R| = n$ rings, i.e. no structural constraint), the number of V-R pairs where the virtual and real scaffold differed by at most two ring systems, and the number of pairs where V and R differed by only one ring. In addition, for each V-R pair category, the number of virtual and real scaffolds and the number of target sets from which the pairs originated are provided.

approximately half. Moreover, when structural deviations in pairs were limited to one ring, 10,886 V-R pairs were obtained from 440 target sets that involved 6584 virtual and 8933 real scaffolds (Table 2). These V-R pairs provided a pool of scaffold pairs for pattern definition and activity prediction, as discussed in the following.

Prioritized Scaffold Patterns. We next defined scaffold patterns formed by V-R pairs that were of increasing attractiveness for activity prediction. Virtual scaffolds were considered attractive candidates if they were "framed" by real scaffolds. In Figure 2, such R-$(V)_n$-R patterns are highlighted. For example, two virtual scaffolds appear in a pathway between two real scaffolds (i.e., $n = 2$). However, the most attractive pattern is formed when a virtual scaffold has two real scaffolds as neighbors (i.e., $n = 1$), yielding an R-V-R pattern. In this case, a virtual scaffold is involved in substructure relationships to a known active child and parent scaffold (and each of these real scaffolds differs in one ring from the virtual scaffold). For the purpose of our analysis, we regarded virtual scaffolds involved in R-$(V)_n$-R and R-V-R patterns as "prioritized virtual" scaffolds.

As illustrated in Figure 3, nonterminal scaffolds in tree branches are always involved in at least two substructure relationships in the ST hierarchy, and these structural relationships correspond to two scaffold pairs (for example, pairs 1−3 and 3−4 in Figure 3). However, scaffolds might also participate in additional substructure relationships/pairs with leaf (or other) scaffolds that are not a part of the hierarchy. These pairwise relationships can be systematically detected and added to the tree structure, as also illustrated in Figure 3. Hence, by further extending the ST hierarchy scaffolds can also be evaluated by taking additional substructure information into account. For prioritized virtual scaffolds in R-V-R patterns, nonhierarchy substructure pairings (i.e., relationships to leaf scaffolds and other real scaffolds) add further activity-relevant structural information. Moreover, for many virtual scaffolds that are not part of R-V-R patterns, other V-R pairs might also be found that provide additional substructure information. Thus, the likelihood of activity would be expected to further increase for virtual scaffolds involved in additional substructure relationships to known active scaffolds.

Scaffold Mapping. Different categories of virtual scaffolds were mapped to scaffolds extracted from MDDR compounds and approved drugs. We reasoned that prioritized virtual scaffolds

Table 3. Scaffold Mapping[a]

scaffold type		no. of scaffolds		
		total	MDDR	drugs
real		13,377	2658 (20%)	226 (1.7%)
virtual		10,502	1005 (10%)	59 (0.6%)
prioritized virtual	R - (V)$_n$ - R	997	174 (17%)	30 (3%)
	R - V - R	544	120 (22%)	23 (4.2%)
	R - V - R or ≥2 V-R pairs	1678	449 (27%)	75 (4.5%)

[a] Five sets of CDB/BDB scaffolds extracted from ST hierarchies including real and virtual scaffolds and prioritized virtual scaffolds in different patterns were mapped to scaffolds of MDDR compounds and of approved drugs from DrugBank. The pattern designated "R-V-R or ≥2 V-R pairs" combines all prioritized virtual scaffolds of R-V-R patterns with other virtual scaffolds that were involved in at least two additional substructure relationships (i.e. nonhierarchy pairs). The total number of scaffolds comprising each set is given, and the number of these scaffolds that matched MDDR or drug scaffolds is reported.

should display an increasing tendency to match scaffolds from bioactive compounds. From 157,522 MDDR compounds and 1247 drugs, 71,649 and 722 scaffolds were obtained, respectively. Mapping of CDB/BDB scaffolds from ST hierarchies to the MDDR was considered a meaningful exercise because, as reported in Table 3, only 20% of all real and 10% of all virtual scaffolds were found to match MDDR scaffolds. For approved drugs, the numbers of matching scaffolds were much smaller.

We first considered virtual scaffolds from R-(V)$_n$-R patterns, which reduced the number of candidate scaffolds from 10,502 to 997 (∼9%). However, the match rate of these scaffolds increased to 17% in the MDDR and 3% in drugs (Table 3). We then focused on virtual scaffolds from R-V-R patterns, which further reduced the number of candidate scaffolds to 544. In this case, 22% of these scaffolds matched MDDR scaffolds. Thus, compared to nonprioritized virtual scaffolds, the use of R-V-R virtual scaffolds essentially doubled the match rate. Furthermore, we also found that nearly all virtual scaffolds from R-V-R patterns were also involved in additional substructure relationships. Therefore, we complemented the set of R-V-R scaffolds with other virtual scaffolds that were involved in at least two V-R substructure pairs outside the ST hierarchy, which resulted in a total of 1,678 prioritized virtual scaffolds. This extended scaffold set produced a match rate of 27% in the MDDR (and 4.5% in approved drugs). Thus, scaffold mapping revealed that prioritized virtual scaffolds were generally more likely to match bioactive scaffolds than nonprioritized virtual scaffolds.

Activity Prediction. On the basis of these findings, we went a step further and predicted the activity of high-priority virtual scaffolds. For this purpose, we ranked the two sets of matching virtual scaffolds from R-V-R patterns on the basis of additional substructure information content, i.e. additional V-R pairs these scaffolds were involved in. Tables 4 and 5 report the rankings for prioritized virtual scaffolds matching MDDR compounds (120 scaffolds) and approved drugs (23 scaffolds), respectively. For these scaffolds, up to 58 additional substructure relationships (V-R pairs) were detected. Each prioritized V-scaffold was then predicted to have the same activity as the neighboring R-scaffolds, and this prediction was compared to the target annotations of matching MDDR compounds or drugs. Correct predictions were found at a high rate for 26 of 120 virtual scaffolds in the MDDR and for six of 23 virtual scaffolds in DrugBank, although these compound sources have different target distributions.

Table 4. Activity Prediction for Prioritized Virtual Scaffolds in the MDDR[a]

ScafID	#additional V-R pairs	correct prediction	SMILES
11893	*58*	*cathepsin B*	*c1ccc(cc1)c2ccccc2*
12745	*53*	*serotonin receptor 2c, renin*	*c1ccc2ccccc2(c1)*
10815	38		c1ccc(cc1)Cc2ccccc2
12771	*22*	*matrix metalloproteinases 1, 3*	*c1ccc2ncccc2(c1)*
10620	*17*	*matrix metalloproteinases 2, 3, 9, 13*	*c1ccc(cc1)COc2ccccc2*
12537	17		c1ccc2CCCc2(c1)
12634	17		c1ccc2[nH]cnc2(c1)
12772	16		c1ccc2ncncc2(c1)
22794	16		c1ccc2occc2(c1)
11896	13		c1ccc(cc1)c2ccccn2
21195	13		c1ccc(cc1)C2CCNCC2
15362	12		O=C(CC(c1ccccc1)-c2ccccc2)N3CCCC3
21850	*12*	*matrix metalloproteinases 3, 8*	*c1ccc(cc1)OCc2ccnc3ccccc23*
11700	11		c1ccc(cc1)Oc2ccccc2
9220	9		O=S(=O)(Nc1ccccc1)-c2ccccc2
9823	9		c1ccc(cc1)C2CCCCC2
12828	*9*	*serotonin receptor 1d, 1b*	*c1ccc3c(c1)[nH]cc3-(C2CC[N+]CC2)*
13145	*9*	*adenosine receptor A2A*	*c1ncc2nc[nH]c2(n1)*
22561	9		c1ccc2[n+]cccc2(c1)
4470	8		O=C(NCCc1ccccc1)c2ccccc2
6312	8		O=C(c1ccccc1)c2ccccc2
9366	8		O=S(=O)(c1ccccc1)N2CCCC2
9812	8		c1ccc(cc1)C2CC2
22733	8		c1ccc2nc(ccc2(c1))N3CC-[N+]CC3
600	7		C1COC(C1)n3cnc2cncnc23
2054	7		O=C(CCc1ccccc1)-NCCc2ccccc2
10321	7		c1ccc(cc1)CCc2ccccc2
9597	6		c1[nH]cc(n1)C2CC2
10813	*6*	*aldose reductase*	*c1ccc(cc1)Cc2ccc3ccccc3(c2)*
11486	*6*	*adenosine receptor A1*	*c1ccc(cc1)Nc4ncnc3c4-(ncn3(C2CCCO2))*
11769	*6*	*serotonin receptor 2a*	*c1ccc(cc1)c2cc3ccccc3-([nH]2)*
11909	6		c1ccc(cc1)c2ccncn2

[a] The 120 virtual scaffolds from R-V-R patterns that matched MDDR scaffolds were ranked according to the number of additional nonhierarchy V-R substructure pairs they were involved in and their activity was predicted. Thirty-two scaffolds were involved in more than five additional pairs and are listed. Prioritized virtual scaffolds with correct activity prediction are shown in italics and their activities are reported. In addition, SMILES[27] representations of ranked scaffolds are provided.

As shown in Tables 4 and 5, correct predictions were preferentially observed for highly ranked virtual scaffolds having high substructure information content.

In Table 4, the top-ranked prioritized virtual scaffold is the biphenyl scaffold, which occurred in the cathepsin B scaffold tree and was hence predicted to be present in compounds active against cathepsin B. Figure 4 shows the scaffold tree environment of the biphenyl scaffold and its two immediate real scaffold neighbors, representing an R-V-R pattern. Also shown is a cathepsin B inhibitor that was found to contain this prioritized scaffold. The second-ranked scaffold in Table 4 is naphthalene, which was

Table 5. Activity Prediction for Prioritized Virtual Scaffolds in DrugBank[a]

ScafID	#additional V-R pairs	correct prediction	SMILES
11893	58		c1ccc(cc1)c2ccccc2
12745	*53*	*beta-2 adrenergic receptor, cyclooxygenase 2*	*c1ccc2ccccc2(c1)*
10815	*38*	*dopamine transporter*	*c1ccc(cc1)Cc2ccccc2*
12771	22		c1ccc2ncccc2(c1)
10620	17		c1ccc(cc1)COc2ccccc2
12537	17		c1ccc2CCCc2(c1)
11700	11		c1ccc(cc1)Oc2ccccc2
12828	*9*	*serotonin receptor 1d, 1b, 2a*	*c1ccc3c(c1)[nH]cc3(C2CC[N+]CC2)*
13145	*9*	*adenosine receptor A3*	*c1ncc2nc[nH]c2(n1)*
22561	9		c1ccc2[n+]cccc2(c1)
4470	8		O=C(NCCc1ccccc1)c2ccccc2
6312	8		O=C(c1ccccc1)c2ccccc2
9812	8		c1ccc(cc1)C2CC2
600	*7*	*purine nucleoside phosphorylase (PNP)*	*C1COC(C1)n3cnc2cncnc23*
10321	7		c1ccc(cc1)CCc2ccccc2
11909	6		c1ccc(cc1)c2ccncn2
154	5		C1CC2CCC(C1)[N+]2
11248	*3*	*alpha-2a adrenergic receptor*	*c1ccc(cc1)Nc2ccccc2*
17561	3		O=C(Nc1nccs1)c2ccccc2
10075	2		c1ccc(cc1)CC2CCCC2
21088	2		c1ccc(cc1)C(CCC[N+]2CCCCC2)c3ccccc3
21181	2		c1ccc(cc1)C2CCCC2
9415	0		O=S(=O)(c1ccccc1)c2ccccc2

[a] The 23 virtual scaffolds from R-V-R patterns that matched approved drugs were ranked according to the number of additional nonhierarchy V-R substructure pairs they were involved in and their activity was predicted. Prioritized virtual scaffolds with correct activity prediction are shown in italics and their activities are reported. In addition, SMILES[27] representations of ranked scaffolds are provided.

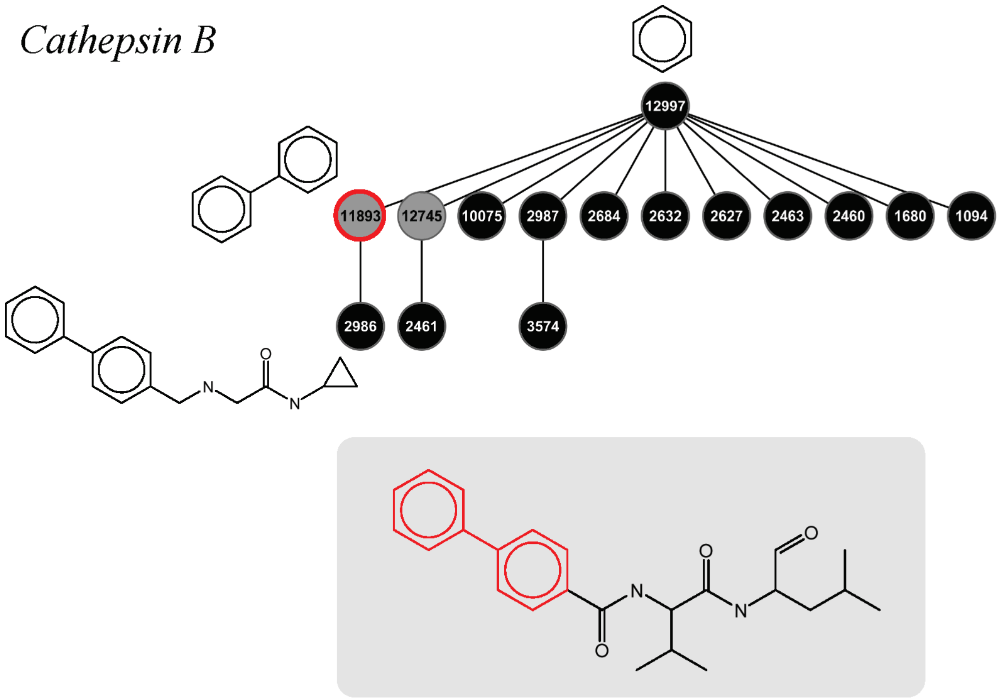


Figure 4. Activity prediction for the biphenyl scaffold (bioactive compounds). Scaffold tree branches for the cathepsin B inhibitor set contained the prioritized "virtual" biphenyl scaffold (labeled with ID 11893 and identified by a red circle). Neighboring "real" scaffolds of the same branch are shown. A representative cathepsin B inhibitor containing the biphenyl scaffold (red) is shown on a light gray background.

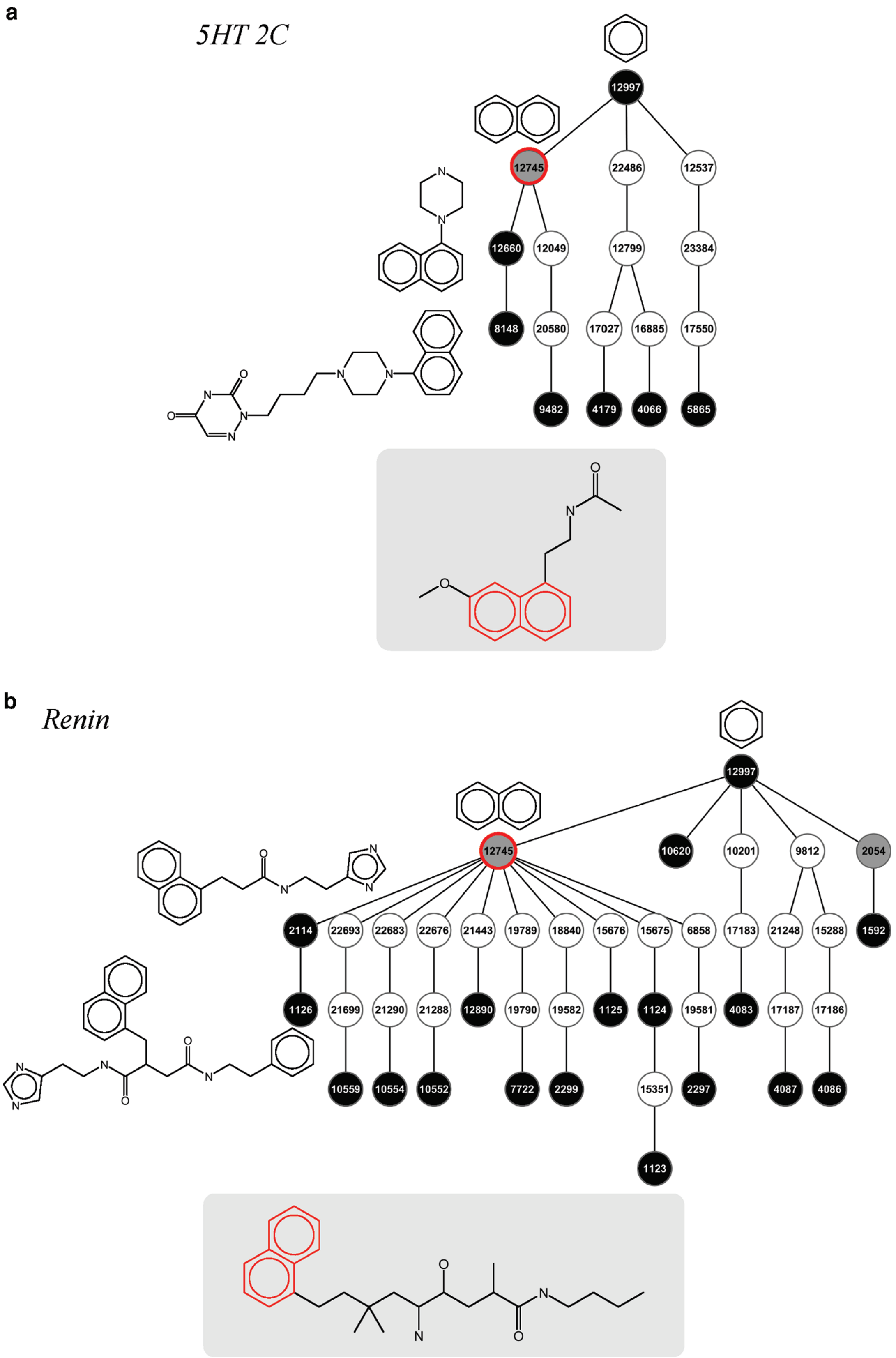


Figure 5. Activity prediction for the naphthalene scaffold (bioactive compounds). Scaffold tree branches for (**a**) 5HT 2C antagonists and (**b**) renin inhibitors contained naphthalene as a prioritized virtual scaffold. The presentation is according to Figure 4. Matching bioactive compounds containing this scaffold (red) are shown.

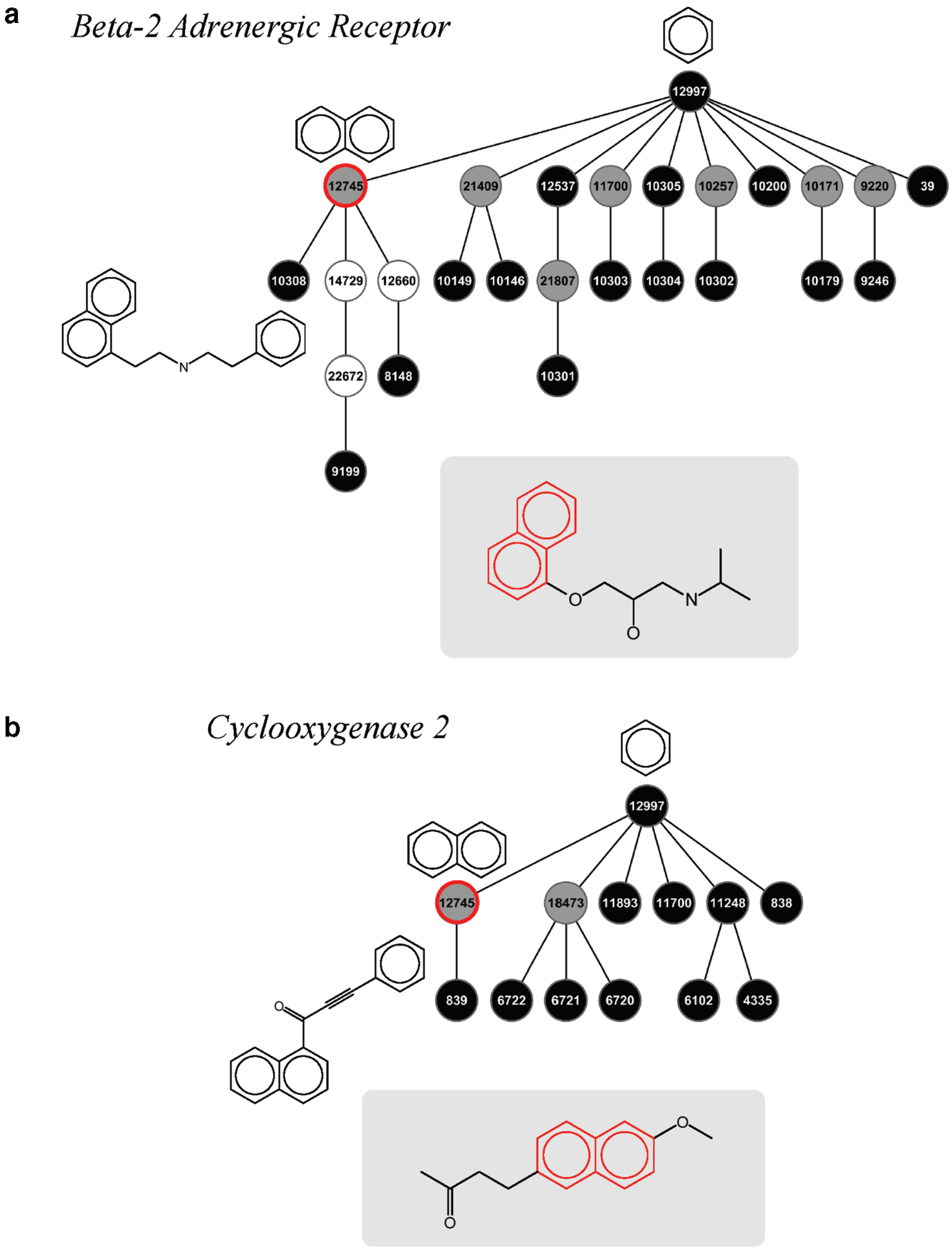
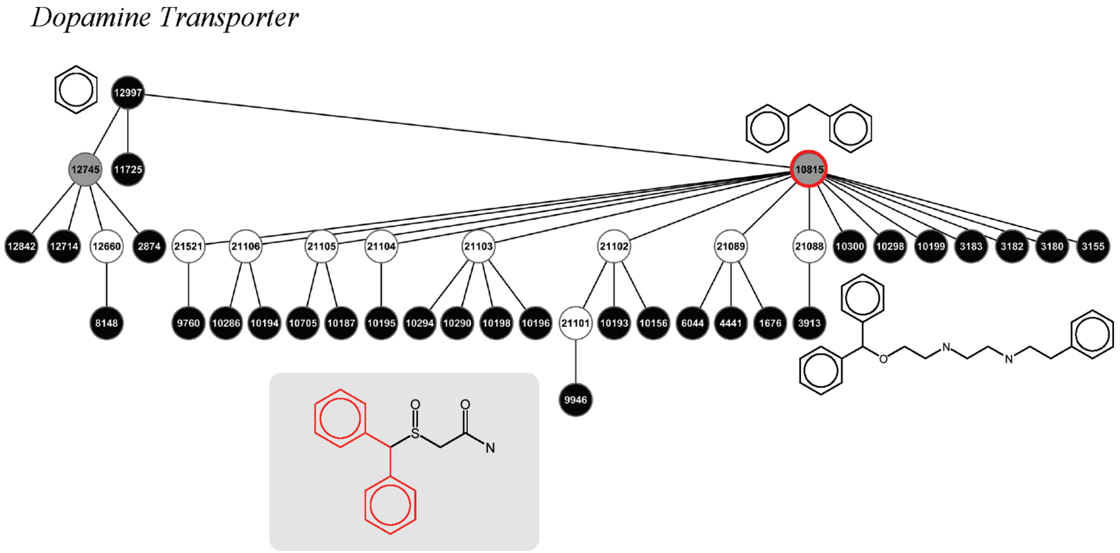


Figure 6. Activity prediction for the naphthalene scaffold (drugs). Scaffold tree branches for (**a**) beta-2 adrenergic receptor antagonists and (**b**) cyclooxygenase 2 inhibitors contained naphthalene as a prioritized virtual scaffold. The presentation is according to Figure 4. Matching drugs are shown.

prioritized in two different scaffold trees originating from 5HT 2C serotonin receptor antagonists and renin inhibitors, respectively, both of which correctly matched MDDR compounds. In Figure 5, the different scaffold tree environments of the naphthalene scaffold are shown. In the 5HT 2C scaffold tree in Figure 5a, this high-priority scaffold occurs in a peripheral branch, whereas in the renin inhibitor-derived tree in Figure 5b it is the most central scaffold from which 10 sub-branches originate. Hence, these target-set derived scaffold tree environments of naphthalene differed substantially. However, for both targets, active compounds containing the naphthalene moiety were identified. As reported in Table 5, the biphenyl scaffold did not yield a correct match, i.e. no drug was found directed against a target representing one of the trees where the biphenyl scaffold was prioritized. However, for the naphthalene scaffold, drugs acting against two of its targets were identified, the beta-2 adrenergic receptor and cyclooxygenase 2. The corresponding scaffold tree environments and matching drugs are shown in Figure 6a and Figure 6b, respectively. Thus, these targets differed from those for which matching MDDR compounds were identified. In Table 5, the third-ranked scaffold is diphenylmethane, which is closely related to the biphenyl scaffold. In this case, no matching MDDR compound was identified. However, the diphenylmethane scaffold was found to correctly match a drug active against the dopamine transporter. In Figure 7, the corresponding scaffold tree environment of diphenylmethane is shown. Here, this prioritized virtual scaffold is also the most central scaffold and involved in seven partly



Dopamine Transporter

Figure 7. Activity prediction for the diphenylmethane scaffold (drugs). Scaffold tree branches for dopamine transporter inhibitors contained diphenylmethane as a prioritized virtual scaffold. The presentation is according to Figure 4. A matching drug is shown.

overlapping R-V-R patterns, making it a prime candidate for activity prediction.

CONCLUSIONS

In this study, we have systematically explored the overlap between horizontal and vertical substructure relationships in scaffold hierarchies of many different target sets. Only about a third of all leaf-to-leaf substructure relationships detected in our large-scale analysis were found to be implicitly covered by scaffold hierarchies. Hierarchical and nonhierarchical substructure relationships are complementary in nature. Thus, the additional substructure information was included in the analysis to further differentiate between scaffolds. On the basis of our findings, virtual scaffolds were successfully prioritized for scaffold mapping and activity prediction by combining scaffold pattern and substructure pair information. Given the wealth of available scaffold substructure relationships, scaffold prioritization scheme introduced herein should also be useful for practical applications of scaffold hierarchies to predict novel active compounds.

AUTHOR INFORMATION

Corresponding Author

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

REFERENCES

(1) 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.

(2) Schnur, D. M.; Hermsmeier, M. A.; Tebben, A. J. Are Target-Family-Privileged Substructures Truly Privileged? *J. Med. Chem.* **2006**, *39*, 2000–2009.

(3) 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.

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

(5) Zhao, H. Scaffold Selection and Scaffold Hopping in Lead Generation: A Medicinal Chemistry Perspective. *Drug Discovery Today* **2007**, *12*, 149–155.

(6) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.

(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) Lipkus, A. H.; Yuan, Q.; Lucas, K. A.; Funk, S. A.; Bartelt, W. F., III; Schenck, R. J.; Trippe, A. J. Structural Diversity of Organic Chemistry. A Scaffold Analysis of the CAS Registry. *J. Org. Chem.* **2008**, *73*, 4443–4451.

(9) Krier, M.; Bret, G.; Rognan, D. Assessing the Scaffold Diversity of Screening Libraries. *J. Chem. Inf. Model.* **2006**, *46*, 512–524.

(10) 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.

(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) Hu, Y.; Bajorath, J. Scaffold Distributions in Bioactive Molecules, Clinical Trials Compounds, and Drugs. *ChemMedChem* **2010**, *5*, 187–190.

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

(14) Wilkens, S. J.; Janes, J.; Su, A. I. HierS: Hierarchical Scaffold Clustering Using Topological Chemical Graphs. *J. Med. Chem.* **2005**, *48*, 3182–3193.

(15) Kenny, P. W.; Sadowski, J. Structure Modification in Chemical Databases. In *Chemoinformatics in Drug Discovery*; Oprea, T. I., Ed.; Wiley-VCH: Weinheim, Germany, 2004; pp 271−285.

(16) 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.

(17) 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.

(18) 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.

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

(20) *ChEMBL*; European Bioinformatics Institute (EBI): Cambridge, 2010. http://www.ebi.ac.uk/chembl/ (accessed May 11, 2010).

(21) 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.

(22) 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.

(23) *Molecular Drug Data Report (MDDR)*; Symyx Software: San Ramon, CA, 2008.

(24) Wishart, D. S.; Knox, C.; Guo, A. C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M. DrugBank: A Knowledgebase for Drugs, Drug Actions and Drug Targets. *Nucleic Acids Res.* **2008**, *36*, D901–D906.

(25) *Molecular Operating Environment (MOE)*; Chemical Computing Group Inc.: Montreal, Quebec, Canada, 2009.

(26) *Pipeline Pilot*, Student ed., version 6.1; Accelrys, Inc.: San Diego, CA, 2007.

(27) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.

Summary

Herein, we have systematically compared horizontal (nonhierarchical) and vertical (hierarchical) substructure relationships among 13,377 scaffolds isolated from 34,916 compounds active against 458 different targets. Only ∼32% of scaffold pairs having horizontal substructure relationships were detected in scaffold hierarchies of the *Scaffold Tree* structure. Therefore, additional nonhierarchical substructure relationships were combined with *Scaffold Tree* data to further prioritize virtual scaffolds that were not contained in original compounds for activity prediction. By mapping virtual scaffolds to external sets of bioactive compounds and approved drugs, prioritized scaffolds were more likely to match known active compounds than non-prioritized ones. Moreover, the activity of high-priority virtual scaffolds was predicted and more than 20% of these scaffolds were found to match the predicted activity in external compounds. Therefore, horizontal and vertical substructure relationships were complementary for activity prediction, which further increased information content of the Scaffold Tree data structure.

Conclusion

In medicinal chemistry, it is often difficult to identify structure-activity relationship determinants of bioactive compounds. Therefore, the aim of this thesis has been to explore different types of relationships between chemical structure and biological activities at the level of molecular scaffolds through systematically mining of currently available compound data. A series of large-scale analyses has been presented.

First, scaffold distributions at different stages of pharmaceutical development were analyzed and compared. Scaffolds with characteristic frequencies of occurrence at different stages have been identified, indicating a likelihood of compounds passing through different development stages (*Chapter 1*).

Next, by exploring relationships between compound selectivity and targets or target families, a set of community-selective scaffolds was identified that represented compounds selective for sets of closely related targets. The identification of community-selective molecular scaffolds has revised the conventional view of privileged substructures (*Chapter 2*). Moreover, a subset of community-selective scaffolds was found to be target-selective, i.e. yielding compounds consistently selective for one particular target over others (*Chapter 3*).

On the other hand, 83 scaffolds corresponding to 33 chemotypes were found to be promiscuous, i.e. representing compounds active against at least three different target families. Despite subtle structural differences, promiscuous scaffolds sharing the same chemotype displayed rather different activity profiles. Also, these scaffolds were found to be enriched in approved drugs (*Chapter 4*).

In addition to analyzing scaffold frequencies, selectivity, and promiscuity, potency distributions of compounds representing the same scaffold were systematically investigated. Scaffolds having a high propensity to form multi-target activity or selectivity cliffs were identified (*Chapter 5*). Moreover, scaffold hopping potential was systematically assessed for current pharmaceutical targets (*Chapter 6*).

Finally, structural relationships were exhaustively explored between scaffolds of compounds active against human targets. Approximately 87% of scaffolds displayed substructure relationships and/or shared the same topology with others. Activity cliffs could be also identified among topologically equivalent scaffolds (*Chapter 7*). These substructure relationships were further compared to, and combined with, a hierarchical scaffold classification scheme called the *Scaffold Tree* to facilitate activity prediction of virtual scaffolds that were not yet found in active compounds contained in the tree structure (*Chapter 8*).

In summary, departing from conventional case-by-case SAR analysis, data mining approaches presented in this thesis were designed to better understand relationships between molecular scaffolds and biological activities on a large scale. On the basis of publicly available compounds, scaffolds were isolated. Sets of scaffolds displaying different characteristics were identified by taking frequencies of occurrence, target selectivity, promiscuity, potency distribution, and structural relationships into consideration, which have provided useful information for lead optimization and compound design.

Additional Publications

Lounkine E., Hu Y., Batista J., Bajorath J. Relevance of feature combinations for similarity searching using general or activity class-directed molecular fingerprints. *J. Chem. Inf. Model.* **2009**, 49, 561-570.

Hu Y., Lounkine E., Bajorath J. Improving the performance of extended connectivity fingerprints through activity-oriented feature filtering and application of a bit density-dependent similarity function. *ChemMedChem* **2009**, 4, 540-548.

Hu Y., Lounkine E., Bajorath J. Filtering and counting of extended connectivity fingerprint features maximizes compound recall and the structural diversity of hits. *Chem. Biol. Drug Des.* **2009**, 74, 92-98.

Peltason L., Hu Y., Bajorath J. From structure-activity to structure-selectivity relationships: quantitative assessment, selectivity cliffs, and key compounds. *ChemMedChem* **2009**, 4, 1864-1873.

Iyer P., Hu Y., Bajorath J. SAR monitoring of evolving compound data sets using activity landscapes. *J. Chem. Inf. Model.* **2011**, 51, in press; available online, DOI 10.1021/ci100505m.

Lebenslauf

Ye Hu
Von-Weichs-Str.20
53121 Bonn

Born on 05. August 1981 in Jiangsu, China

1999-2004: Bachelor of Clinical Medicine, Southeast University, Nanjing

2004-2005: Medical Intern, Nanjing Railway Hospital

2006-2008: Master of Life Science Informatics, Bonn-Aachen International Center for Information Technology (B-IT), Universität Bonn

2008-2011: Doctoral studies in Computational Life Sciences, Bonn-Aachen International Center for Information Technology (B-IT), Universität Bonn

Eidesstattliche Erklärung

An Eides statt versichere ich hiermit, dass ich die Dissertation "Systematic Identification of Scaffolds Representing Different Types of Structure-Activity Relationships" selbst und ohne jede unerlaubte Hilfe angefertigt habe, dass diese oder eine ähnliche Arbeit noch an keiner anderen Stelle als Dissertation eingereicht worden ist und dass sie an den nachstehend aufgeführten Stellen auszugsweise veröffentlicht worden ist:

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.

Hu Y., Bajorath J. Scaffold distributions in bioactive molecules, clinical trials compounds, and drugs. *ChemMedChem* **2010**, 5, 187-190.

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

Hu Y., Bajorath J. Exploring target-selectivity patterns of molecular scaffolds. *ACS Med. Chem. Lett.* **2010**, 1, 54-58.

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

Hu Y., Bajorath J. Global assessment of scaffold hopping potential for current pharmaceutical targets. *Med. Chem. Commun.* **2010**, 1, 339-344.

Hu Y., Bajorath J. Polypharmacology-directed compound data mining: identification of promiscuous chemotypes with different activity profiles and comparison to approved drugs. *J. Chem. Inf. Model.* **2010**, 50, 2112-2118.

Hu Y., Bajorath J. Combining horizontal and vertical substructure relationships in scaffold hierarchies for activity prediction. *J. Chem. Inf. Model.* **2011**, 51, in press; available online, DOI 10.1021/ci100448a.

Ye Hu
March 2011
Bonn