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
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.

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]

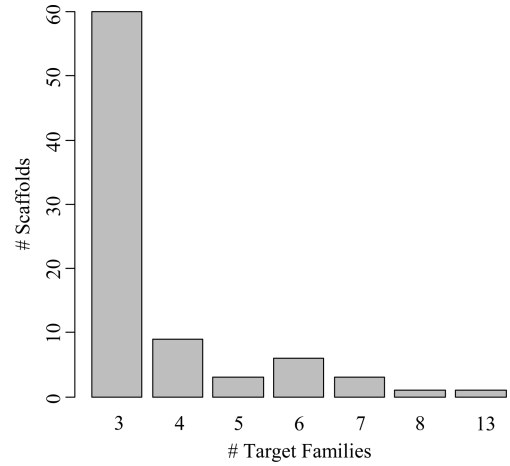


Figure 2. Scaffolds with multiple activities. Shown is the distribution of 83 B-M scaffolds that are active against 3 or more target families.

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-

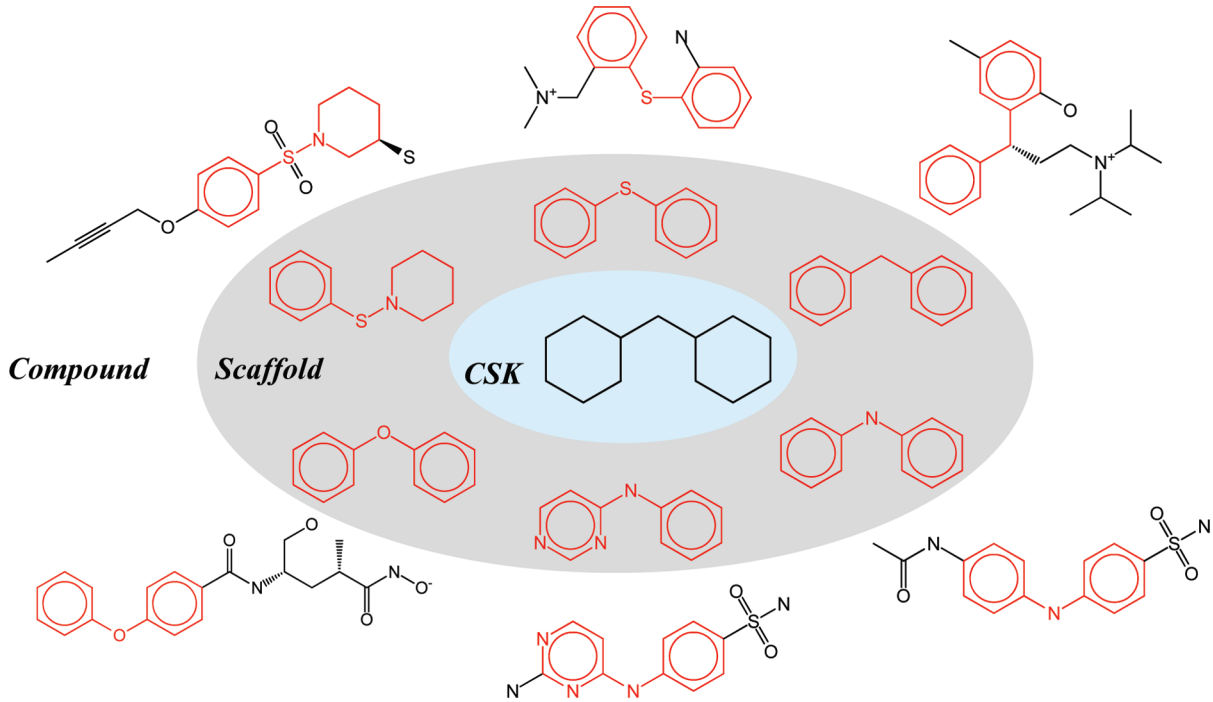


Figure 1. Generation of scaffolds and carbon skeletons. Shown are three layers of structural representations, i.e., compound (outer), scaffold (middle), and carbon skeleton (CSK; inner). B-M scaffolds in compounds are generated by only retaining ring systems and linkers between them (highlighted in red).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

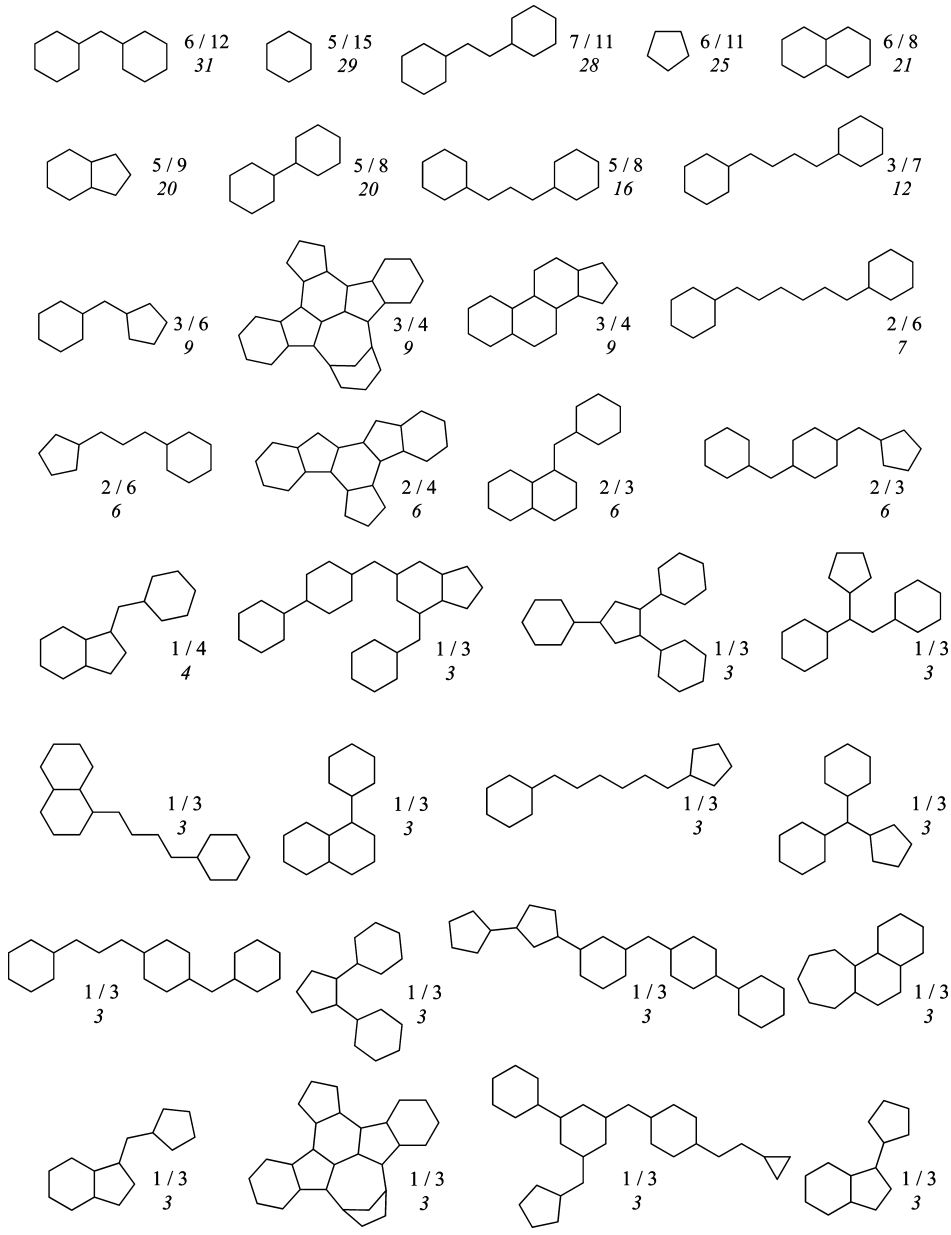


Figure 3. Promiscuous chemotypes. The set of 33 topologically distinct CSKs covering 83 promiscuous scaffolds is shown. For each CSK, the number of its B-M scaffolds and the number of target families these scaffolds are active against are reported. For example, "6/12" means that the CSK covers 6 promiscuous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

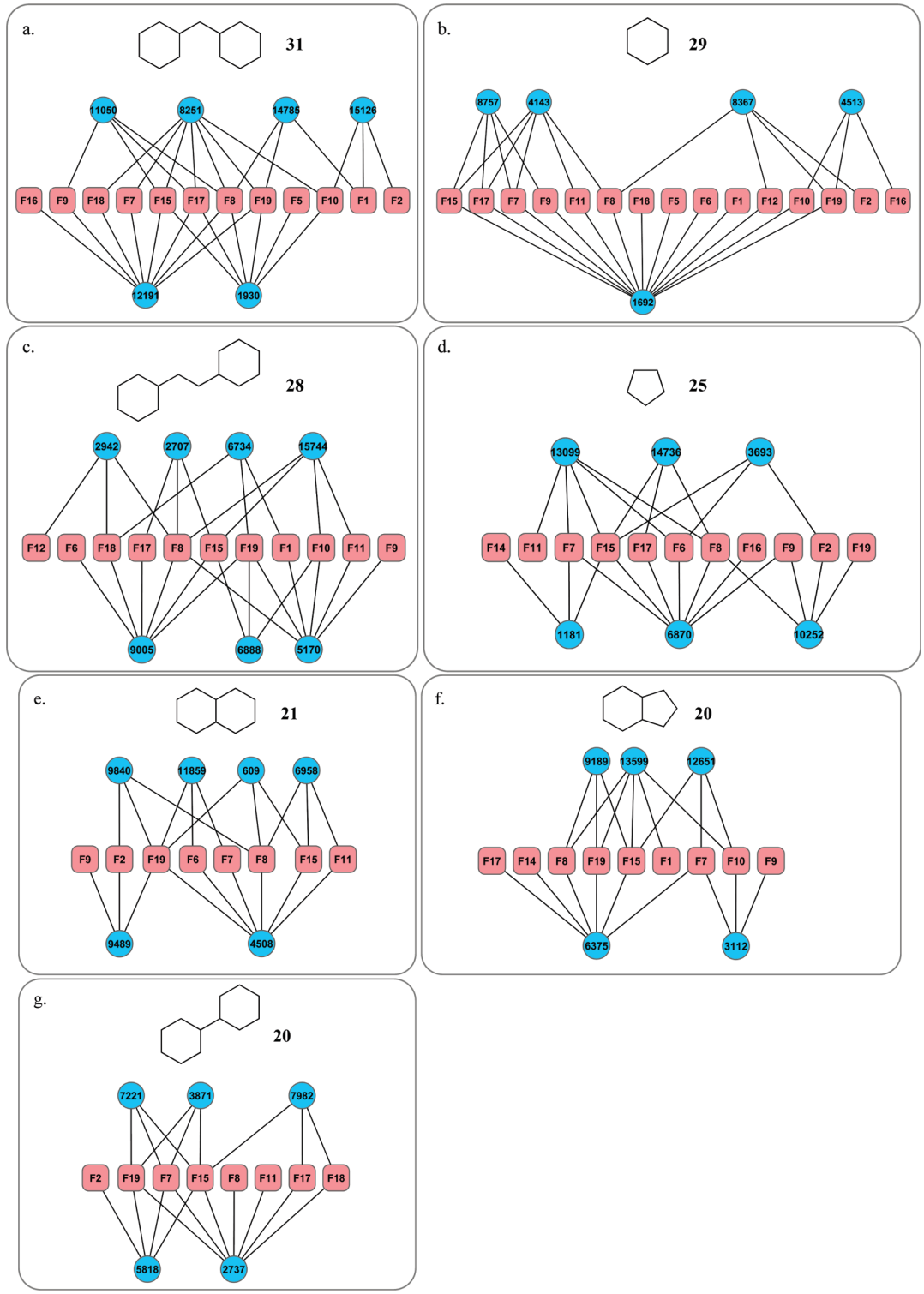


Figure 4. Scaffold-target family networks. In a−g, scaffold−target family relationships are displayed in a network representation for B-M scaffolds of each of the 7 most promiscuous chemotypes. At the top of each figure, the CSK structure is shown, and the total number of scaffold-family relationships 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,

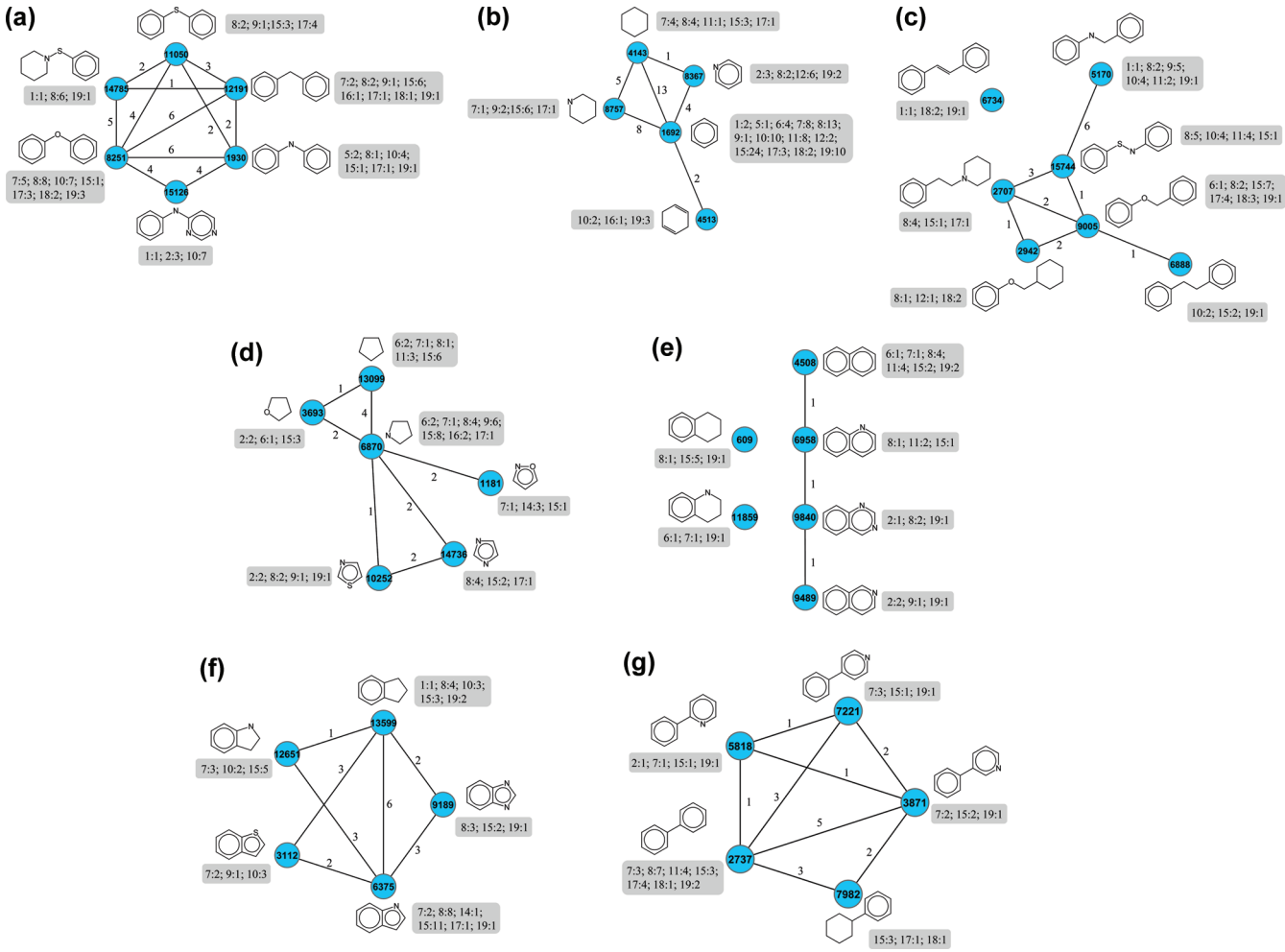


Figure 5. Activity profiles. In a−g, activity profiles for B-M scaffolds covered by each of the 7 most promiscuous chemotypes are displayed in a network representation. Nodes represent scaffolds, and an edge connects two scaffolds if they are active against the same target(s). Edge labels report the 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.

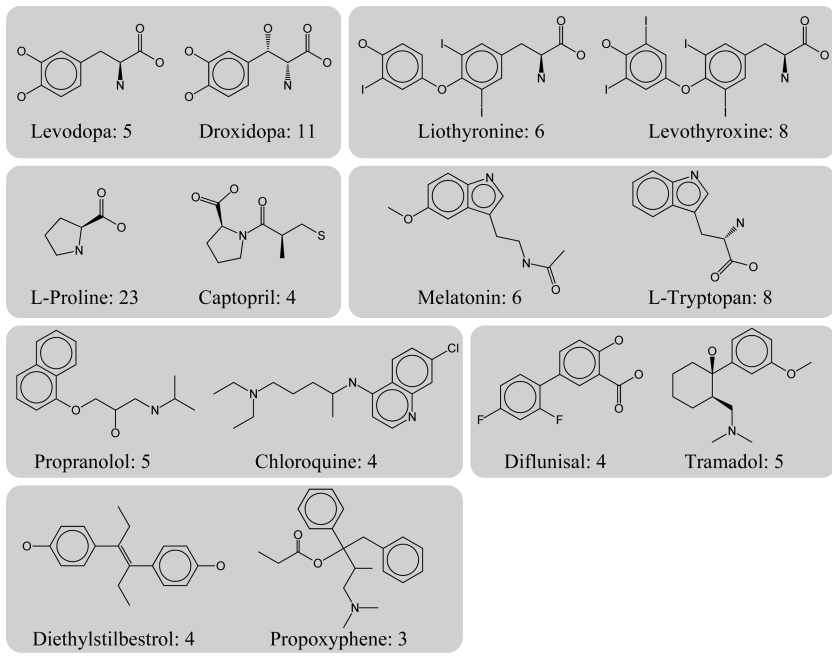
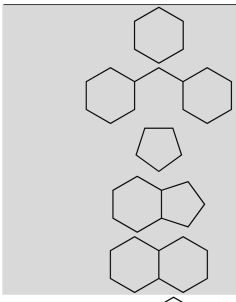
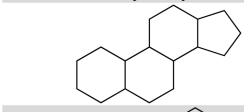
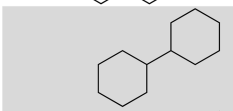
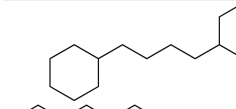
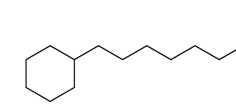
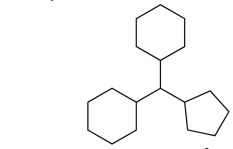
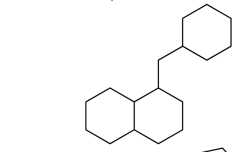


Figure 6. Drugs represented by promiscuous CSKs. For each of the 7 most promiscuous CSKs, two representative drugs are shown. Drug names and the number of annotated targets are reported.

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