Research Article

How Promiscuous Are Pharmaceutically Relevant Compounds? A Data-Driven Assessment

Ye Hu¹ and Jürgen Bajorath^{1,2}

Received 20 August 2012; accepted 5 October 2012; published online 23 October 2012

Abstract. Given the increasing notion of target promiscuity of bioactive compounds and polypharmacological drug behavior, a detailed analysis of publicly available compound activity data from medicinal chemistry sources was carried out to determine and quantify the degree of promiscuity of active compounds across all known human target families. The results are surprising. Approximately 62% of currently available compounds with high-confidence activity data are only annotated with a single biological target, whereas 36% are known to act against multiple targets within the same family (*i.e.*, closely related targets). However, only ~2% of bioactive compounds are promiscuous across different target families. Thus, despite general data sparseness, these findings indicate that highly promiscuous bioactive compounds only rarely occur. Because pharmaceutically relevant active compounds represent the pool from which drug candidates emerge, one might extrapolate from these results and conclude that there is a low statistical probability to obtain drugs that act against multiple targets belonging to distinct families.

KEY WORDS: activity data; compound promiscuity; database mining; pharmaceutically relevant compounds; polypharmacology; target families; target proteins.

INTRODUCTION

Compound promiscuity and polypharmacology are emerging themes in pharmaceutical research (1–4), with potentially far-reaching consequences. Increasing evidence is accumulating that bioactive compounds, drug candidates, and current drugs often specifically interact with multiple targets (1–4). Such insights depart from the single-target-specificity paradigm that has long governed drug discovery strategies. Multitarget drug activities give rise to polypharmacological drug behavior, which is thought to be of critical relevance for therapeutic effects of many (but not all) drugs (5,6). Current estimates are that a drug might on average interact with about six different targets (7). It is therefore not surprising that the notion of polypharmacology is beginning to influence drug discovery and design strategies (8,9). It also

provides the conceptual basis of drug repurposing (10,11), *i.e.*, attempting to find novel targets and therapeutic applications for existing drugs; another much-discussed topic in pharmaceutical research and development. Taking the issue of data sparseness into account, *i.e.*, not all compounds have been tested against all targets (7), one might anticipate even further increasing relevance of compound promiscuity and polypharmacology for drug discovery, although target-specific drugs will continue to be of prime interest for many therapeutic applications (including, for example, the treatment of chronic and infectious diseases). Target promiscuity of pharmaceutically relevant compounds provides the basis for polypharmacology. Compound promiscuity has been assessed at a structural level, for example, by characterizing molecular scaffolds that occur in compounds active against multiple targets (12). In addition, a number of computational approaches have been introduced to predict off-target effects and secondary targets of known drugs (9,13,14). However, although off-target effects and multitarget activities have been explored for a number of drugs and compound classes, it has thus far not been determined how frequently target promiscuity is generally observed among currently available bioactive compounds; a question that should be of considerable interest to better understand the basis of polypharmacological drug behavior. We have analyzed this question in detail and report in part surprising results herein.

Electronic supplementary material The online version of this article (doi:10.1208/s12248-012-9421-y) contains supplementary material, which is available to authorized users.

¹Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstr. 2, 53113 Bonn, Germany.

²To whom correspondence should be addressed. (e-mail: bajorath@bit.uni-bonn.de)

MATERIALS AND METHODS

Data Collection

The ChEMBL database has its origins in pharmaceutical research and is currently the major public domain source of compound structure and activity data originating from medicinal chemistry projects (15). Compound activity data are systematically extracted from original literature and patent resources and curated. Hence, the database has a strong focus on pharmaceutically relevant compounds. From the current release of ChEMBL (release 13), bioactive compounds that were reported to have direct interactions (*i.e.*, target relationship type “D”) with human targets at the highest confidence level (*i.e.*, target confidence score of 9) were assembled. Only equilibrium constants (K_i values) were considered as activity data. Approximate K_i measurements such as “>,” “<,” or “~” were discarded. For compounds with multiple K_i values for the same target, the geometric mean of all values was calculated to yield the final potency annotation. Furthermore, all targets were organized into target families following the protein classification hierarchy of ChEMBL and UniProt family annotations (16).

Compound Decomposition

Following the Bemis and Murcko definition (17), each compound was decomposed into two structural components, *i.e.*, a molecular scaffold and R-groups. The union of all R-groups extracted from a compound comprised its *R-group set*.

Definition of Promiscuity and Compound Categories

For each qualifying compound, its *activity profile* was generated by assembling all available target activity annotations. On the basis of the target family classification scheme, the corresponding *target family profile* was also determined. On this basis, three categories of compounds were introduced to represent different levels of *target promiscuity*:

1. Compounds active against a single target, indicating apparent target specificity, were classified as single-target compounds (*ST compounds*). These compounds were determined but not further considered in our promiscuity analysis.
2. Compounds active against multiple targets from the same family, representing *intra-family promiscuity*, were considered as single-family compounds (*SF compounds*).
3. Compounds active against multiple targets that belonged to different target families, reflecting *inter-family promiscuity*, the highest level of promiscuity, were designated multifamily compounds (*MF compounds*).

Clearly, MF compounds have a higher degree of promiscuity than SF compounds, and one might argue that MF compounds should be a primary focal point of promiscuity analysis. However, in our view, it is also

Table I. Compound and Activity Data

ChEMBL release 13	Number
Compounds	33,778
Targets	529
Target families	130
Activity profiles	1,144
Scaffolds	12,044
R-groups	2,825
R-group sets	13,334

The number of active compounds meeting the selection criteria defined in the text is reported together with the number of their targets, target families, and unique activity profiles. In addition, the number of molecular scaffolds, individual R-groups, and unique R-group sets is given

relevant to consider different levels of intra-family compound promiscuity, in particular, since significant differences are detectable. Hence, we have aimed to provide a comprehensive picture (as much as possible on the basis of currently available data) of compound promiscuity within and across different target families.

Molecular Properties

For compounds of different promiscuity categories, their molecular weight and octanol/water (o/w) partition coefficient (logP) values were calculated using the Molecular Operating Environment (18). In addition, for each category, the numbers of represented scaffolds and R-group sets were determined.

Fig. 1. Compound classification. A total of 33,778 compounds were classified according to their target profiles. For compounds active against multiple targets, the number of target families was determined. Three sets of compounds were obtained that were active against a single target (ST compounds), multiple targets from the same family (SF compounds), or multiple targets belonging to different families (MF compounds). These compound categories represented increasing degrees of promiscuity

Fig. 2. Promiscuous compounds. Four representative MF compounds that are active against 15 to 25 targets belonging to three or five families are shown. For each compound, the ChEMBL compound ID and the numbers of targets and families are provided. Detailed activity profiles of these compounds are reported in Table S1 in the ESM

Compound Data Mining

All data mining calculations were carried out with in-house generated programs.

RESULTS AND DISCUSSION

Compound Data

High-Confidence Activity Data

The use of high-confidence activity data is of critical importance for a meaningful assessment of compound promiscuity. For example, many drugs are annotated with much larger numbers of targets in ChEMBL than in DrugBank (19). We have previously shown that the number of drug targets is generally reduced when approximate and/or low-confidence activity measurements are discarded and only precise measurements are considered (20), leading to similar numbers of targets for drugs in ChEMBL and DrugBank. For instance, the approved drug buspirone was annotated with 72 potential protein targets in ChEMBL. However, when applying high-

confidence data selection criteria, as specified above, the number of targets was reduced to two, consistent with the number of target annotations reported in DrugBank. Furthermore, other noteworthy studies addressing compound promiscuity have consistently used high potency and/or high-confidence data (1,21,22). Given high-confidence data requirements, screening data have not been used to assess promiscuity. One cannot exclude that approximate activity measurements such as IC_{50} values are indicative of true targets of bioactive compounds and drugs. However, their use also increases the tendency of false-positive or nonrelevant assignments, especially for weakly active compounds. Therefore, we have applied the most rigorous data integrity criteria, as described above, and exclusively utilized equilibrium constants as activity measurements. The exclusive use of explicit equilibrium constants also alleviates the likely requirement for introducing potency threshold values for very weakly active compounds (that might be false-positives). This is the case because explicit equilibrium constants are rarely reported for borderline active compounds.

On the basis of our compound selection criteria, a total of 33,778 compounds were obtained from the current release

Table II. Molecular Properties of Compounds

Statistics	Molecular weight			logP		
	ST	SF	MF	ST	SF	MF
1st quartile	360.5	334.5	352.7	2.4	2.02	3.5
Median	432.6	408.6	409.9	3.7	3.6	4.2
Mean	459.0	431.1	397.9	3.6	3.5	4.1
3rd quartile	515.9	490.4	448.6	4.9	5.0	5.0

For each compound category (ST, SF, and MF compounds), the molecular weight and o/w partition coefficient (logP) value distributions are reported

of ChEMBL that were active against a total of 529 human targets, as reported in Table I.

Compound Promiscuity

Categorization of Compounds

Activity Profiles and Compound Structures

The targets of these compounds belonged to 130 human target families and formed 1,144 unique activity profiles. In addition, the compounds represented 12,044 unique molecular scaffolds and 13,334 R-group sets involving 2,825 individual R-groups.

For all qualifying compounds, the promiscuity was evaluated on the basis of their target and family profiles, as shown in Fig. 1. For each compound, the number of targets against which it was reported to be active was determined. Approximately 61.6% of all compounds were found to be active against a single target (ST compounds)

Fig. 3. Top-ranked scaffolds. For each compound category, the top six scaffolds representing the largest number of compounds are shown. For each scaffold, the number of corresponding compounds (*above the line*) and the number of targets against which these compounds were active (*below the line*) are given. **a** ST, **b** SF, and **c** MF compounds

group sets are shown in Figs. 3 and 4, respectively. The majority of these scaffolds contained three to four rings. The benzene ring, the most generic scaffold, was found to represent the largest number of ST and SF compounds. Furthermore, generic substituents including the methyl and carbonyl groups as well as chlorine and fluorine were frequently observed in highly ranked R-group sets. Different combinations of these R-groups were found in all three compound categories.

Overlap Between Target Families, Scaffolds, and R-group Sets

There were different degrees of overlap between target families, scaffolds, and R-group sets belonging to the ST, SF, and MF compound categories, as reported in Fig. 5.

Target Families

Figure 5a shows that nearly half of all target families (45%) exclusively hosted ST compounds. The list of the 59 target families is provided in Table S6 in the ESM. However, these target families usually consisted of only one or two targets with fewer than 20 active compounds. In addition, for a total of four and eight families, only SF and MF compounds were found, respectively. Again, these target families only consisted of one to three targets, as reported in Tables S7 and S8 in the ESM. Hence, exclusive binding of ST, SF, or MF compounds to target families essentially corresponded to binding to individual targets. Furthermore, 35 target families hosted compounds belonging to two promiscuity categories, as reported in Tables S9 and S10 in the ESM. These target families often contained multiple targets for which larger numbers of active compounds were available. Moreover, for 24 target families, ST, SF, and MF compounds were found (Table S11 in the ESM). These families included prominent therapeutic target families with many targets and large numbers of active compounds. Thus, for well-explored target families, many compounds with different levels of intra-family promiscuity were found, but compounds with promiscuity across different families were rare.

Scaffolds and R-group Sets

In Fig. 5b, the overlap between scaffolds representing ST, SF, and MF compounds is reported. Surprisingly, more than 90% of all scaffolds were found to exclusively represent compounds belonging to an individual category. Sixty percent of the scaffolds exclusively represented ST compounds and 31% SF compounds. By contrast, only 2% of the scaffolds exclusively occurred in MF compounds and an additional 0.9% in combinations of MF and SF or ST compounds. Equivalent observations were made for R-group sets in Fig. 5c. Thus, scaffolds and R-group sets representing MF compounds were very rare. In Fig. 6, the most promiscuous scaffolds identified in our analysis are shown.

CONCLUSIONS

Herein, we have systematically analyzed the target promiscuity of pharmaceutically relevant compounds on the

Fig. 5. Overlap between target families, scaffolds, and R-group sets. In (a–c), target families, scaffolds, and R-group sets representing different compound categories are compared in Venn diagram representations, respectively. Table inserts report the numbers (percentages) of target families, scaffolds, and R-group sets for different subsets consisting of compounds belonging to one, two, or all three promiscuity categories

basis of currently available high-confidence activity data and molecular graphs. The ligand-centric approach is not the only way to analyze compound promiscuity, which

Fig. 6. Scaffolds exclusively occurring in MF compounds. Six scaffolds are shown that represent the largest numbers of MF compounds. For each scaffold, the number of corresponding compounds (*above the line*) and the number of targets against which these compounds were active (*below the line*) are given

might also be attempted, for example, by taking 3D target structure and ligand information into account (23). In our analysis, different compound categories were introduced that represented different levels of promiscuity, from single-target compounds (that are, by definition, non-promiscuous) over single-family compounds to multifamily compounds that are promiscuous across different target families (representing the highest level of promiscuity in our analysis). In this context, it should also be noted that SF compounds might present a source of privileged structure information for individual target families (24). Compounds with activity annotations against single targets were most frequently observed. However, for popular target families, large numbers of compounds with different levels of promiscuity were found. We have determined that approximately a third of current bioactive compounds acted against different targets belonging to the same family, whereas only ~2% of all compounds displayed promiscuity across different target families. Thus, compound promiscuity was mostly confined to targets belonging to individual families (*i.e.*, closely related targets). By contrast, compounds acting against targets from different families were very rare. In addition, we also found that more than 90% of currently available molecular scaffolds exclusively represented compounds belonging to an individual promiscuity category. Thus, there are many opportunities to further explore scaffolds for the design of compounds with different levels of target selectivity. When currently available active compounds will be tested against increasing numbers of targets, the ratio of promiscuous to apparently target-specific compounds is expected to further increase. However, if one extrapolates from our findings that compounds with high-level promiscuity are rare, there is only a low statistical probability to select polypharmacological drug candidates that act against different types of targets. Thus, such compounds might have been

preferentially selected because of favorable therapeutic effects and high efficacy, which would rationalize the apparent enrichment of polypharmacological compounds among drugs.

REFERENCES

- Paolini GV, Shapland RHB, van Hoorn WP, Mason JS, Hopkins AL. Global mapping of pharmacological space. *Nat Biotechnol.* 2006;24(7):805–15. doi:10.1038/nbt1228.
- Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, Shoichet BK. Relating protein pharmacology by ligand chemistry. *Nat Biotechnol.* 2007;25(2):197–206. doi:10.1038/nbt1284.
- Hopkins AL. Network pharmacology: the next paradigm in drug discovery. *Nat Chem Biol.* 2008;4(11):682–90. doi:10.1038/nchembio.118.
- Keiser MJ, Setola V, Laggner C, Abbas AI, Hufeisen SJ, Jensen NH, Kuijter MB, Matos RC, Tran TB, Whaley R, Glennon RA, Hert J, Thomas KL, Edwards DD, Shoichet BK, Roth BL. Predicting new molecular targets for known drugs. *Nature.* 2009;462(7270):175–81. doi:10.1038/nature08506.
- Mestres J, Gregori-Puigjané E. Conciliating binding efficiency and polypharmacology. *Trends Pharmacol Sci.* 2009;30(9):470–4. doi:10.1016/j.tips.2009.07.004.
- Merino A, Bronowska AK, Jackson DB, Cahill DJ. Drug profiling: knowing where it hits. *Drug Discov Today.* 2010;15(17–18):749–56. doi:10.1016/j.drudis.2010.06.006.
- Mestres J, Gregori-Puigjané E, Valverde S, Sole RV. Data completeness—the Achilles heel of drug-target networks. *Nat Biotechnol.* 2008;26(9):983–4. doi:10.1038/nbt0908-983.
- Koutsoukas A, Simms B, Kirchmair J, Bond PJ, Whitmore AV, Zimmer S, Young MP, Jenkins JL, Glick M, Glen RC, Bender A. From *in silico* target prediction to multi-target drug design: current databases, methods and applications. *J Proteomics.* 2011;74(12):2554–74. doi:10.1016/j.jprot.2011.05.011.
- Xie L, Xie L, Kinnings SL, Bourne PE. Novel computational approaches to polypharmacology as a means to define responses to individual drugs. *Annu Rev Pharmacol Toxicol.* 2012;52:361–79. doi:10.1146/annurev-pharmtox-010611-134630.

10. Ashburn TT, Thor KB. Drug repositioning: identifying and developing new uses for existing drugs. *Nat Rev Drug Discov*. 2004;3(8):673–83. doi:10.1038/nrd1468.
11. Chong CR, Sullivan DJ. New uses for old drugs. *Nature*. 2007;448(7154):645–6. doi:10.1038/448645a.
12. 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(12):2112–8. doi:10.1021/ci1003637.
13. Bender A, Scheiber J, Glick M, Davies JW, Azzaoui K, Hamon J, Urban L, Whitebread S, Jenkins JL. Analysis of pharmacology data and the prediction of adverse drug reactions and off-target effects from chemical structure. *ChemMedChem*. 2007;2(6):861–73. doi:10.1002/cmdc.200700026.
14. Brown JB, Okuno Y. Systems biology and systems chemistry: new directions for drug discovery. *Chem Biol*. 2012;19(1):23–8. doi:10.1016/j.chembiol.2011.12.012.
15. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington JP. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res*. 2012;40(D1):D1100–7. doi:10.1093/nar/gkr777.
16. UniProtConsortium. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res*. 2012;40(D1):D142–8. doi:10.1093/nar/gkr981.
17. Bemis GW, Murcko MA. The properties of known drugs. 1. Molecular frameworks. *J Med Chem*. 1996;39(15):2887–93. doi:10.1021/jm9602928.
18. *Molecular Operating Environment (MOE)*, 2011.10; Chemical Computing Group Inc., 1010 Sherbrooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2011.
19. Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, Pon A, Banco K, Mak C, Neveu V, Djoumbou Y, Eisner R, Guo AC, Wishart DS. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res*. 2011;39(Database issue):D1035–41. doi:10.1093/nar/gkq1126.
20. Hu Y, Bajorath J. Many structurally related drugs bind different targets whereas distinct drugs display significant target overlap. *RSC Adv*. 2012;2(8):3481–9. doi:10.1039/C2RA01345B.
21. Hopkins AL, Mason JS, Overington JP. Can we rationally design promiscuous drugs? *Curr Opin Struct Biol*. 2006;16(1):127–36. doi:10.1016/j.sbi.2006.01.013.
22. Fabian MA, Biggs III WH, Treiber DK, Atteridge CE, Azimioara MD, Benedetti MG, Carter TA, Ciceri P, Edeen PT, Floyd M, Ford JM, Galvin M, Gerlach JL, Grotzfeld RM, Herrgard S, Insko DE, Insko MA, Lai AG, Lélias JM, Mehta SA, Milanov ZV, Velasco AM, Wodicka LM, Patel HK, Zarrinkar PP, Lockhart DJ. A small molecule-kinase interaction map for clinical kinase inhibitors. *Nat Biotechnol*. 2005;23(3):329–36. doi:10.1038/nbt1068.
23. Sturm N, Desaphy J, Quinn RJ, Rognan D, Kellenberger E. Structural insights into the molecular basis of the ligand promiscuity. *J Chem Inf Model*. 2012;52(9):2410–21. doi:10.1021/ci300196g.
24. Evans BE, Rittle KE, Bock MG, Dipardo RM, Freidinger RM, Whitter WL, Lundell GF, Veber DF, Anderson PS. Methods for drug discovery: development of potent, selective, orally effective cholecystokinin antagonists. *J Med Chem*. 1988;31(12):2235–46. doi:10.1021/jm00120a002.