Molecular Topology Analysis of the Differences between Drugs, Clinical Candidate Compounds, and Bioactive Molecules

Hongming Chen,* Yidong Yang, and Ola Engkvist

DECS Global Compound Sciences, Computational Chemistry, AstraZeneca R&D Mölndal, SE-43183 Mölndal, Sweden

Received July 5, 2010

A new method to decompose molecules is proposed and used to analyze drugs, clinical candidate compounds and bioactive molecules. The method classifies a set of molecules into a few well-defined classes based on their molecular framework. It is then possible to use these classes to investigate differences between drugs, clinical candidates and bioactive molecules. The analysis shows that in comparison with clinical candidates and bioactive compounds, drugs have a higher fraction of compounds with only one ring system. This conclusion is still valid after correcting for lipophilicity (ClogP) and molecular size, as well as any potential protein target bias in the data sets. Furthermore the molecular bridge part of compounds in the drug set has on average fewer ring systems than molecules from the other sets. The ring system complexity (RSC) was also investigated and for most topological classes drugs have a lower RSC than the clinical candidates and bioactive molecules. Hence, this study highlights differences in topology between drugs, clinical candidate compounds and bioactive molecules.

INTRODUCTION

Since Lipinski et al. proposed the "rule-of-five" criteria for drug permeability,[1] similar efforts[2−6] have been done by many other research groups to examine the relationship between the molecular properties of pharmaceutical relevant compounds and their potential to become drugs. The purpose has been to discover trends in physicochemical properties for compounds in a particular development stage or in a certain disease area and to identify key factors for compound related attritions. For example, Oprea et al.[2] has investigated the property differences between lead compounds and drugs; Wenlock et al.[3] compared the physicochemical property profiles for compounds in different clinical development phases and for marketed drugs to identify property trends that favor a drug passing through clinical development to reach the market; Vieth[4] et al. compared the molecular weight (MW) and ClogP for different target families of marketed drugs. Recently,[5] a comparison of several simple physicochemical properties for marketed drugs, clinical candidates and bioactive compounds at the individual protein target level was done. The results showed that the ClogP and MW for drugs are significantly lower than for clinical candidates and bioactive compounds. These results were consistent with earlier findings.[3] Lovering et al.[7] reported that the fraction of sp3 carbon atoms and the presence of chiral centers correlate with clinical success. Also, Ritchie et al.[8] showed that larger number of aromatic rings can negatively affect several drug-like properties. The efforts published in the literature have so far been focused on fairly simple molecular physicochemical properties such as MW, ClogP, polar surface area (PSA), etc.

However, it would be interesting to identify new descriptors which differentiate drugs from bioactive compounds and are at the same time different from simple descriptors such as lipophilicity and molecular size. These factors could provide new ways to improve the compound quality leading to higher success rates in the drug development process. Here we report a novel method for comparing drugs, clinical candidates, and bioactive compounds. This method is based on analyzing the molecular topology. The relationship between the identified trends and known differentiators between drugs and bioactive molecules like ClogP and molecular size is also investigated. Any potential target bias in the different databases is also taken into account in the analysis.

Molecular framework analysis goes back to the 1990s, when Bemis and Murcko[9] reported their framework study based on 5120 drugs from the Comprehensive Medicinal Chemistry (CMC) database. They showed that these drugs can be represented by 1179 frameworks and that the top 42 frameworks account for 24% of the drugs. Inspired by their pioneering work, other research groups [10−13] used similar methodologies to analyze different sets of pharmaceutical relevant compounds to extract representative molecular scaffolds and ring systems. Xu and Johnson developed molecular equivalence indices (MEQI) to be able to cluster compounds based on their molecular graphs.[14,15] In this paper, a new framework analysis method is adopted and applied to identify differences in the molecular frameworks between marketed drugs, clinical candidate compounds and bioactive molecules. To be able to compare the different classes of molecules, we applied a framework analysis method that only divides up the framework in a few distinct topology classes. The rationale is that with this division it is possible to compare molecular framework distributions between the three data sets. This can also be done when

* To whom correspondence should be addressed, Tel: +46-31-7065285, Fax: +46-31-7763792. Email: hongming.chen@astrazeneca.com.

10.1021/ci1002558
Published on Web 11/15/2010

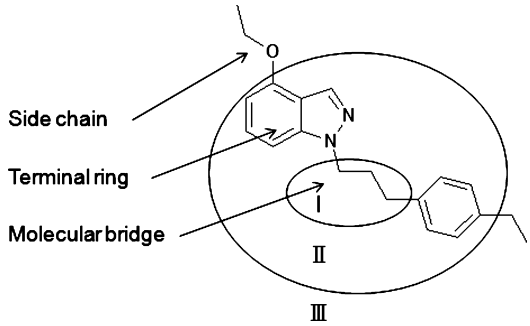

Figure 1. Definition of the different topological layers.

taking into account potential protein target bias as well as differences in the molecular framework distributions related to ClogP and molecular size. Murcko[9] and Xu's[14] molecular framework classification schemes create many more molecular framework classes. The type of analysis done here would be much more difficult with their schemes since the number of molecules in each topological class would be much smaller and it would therefore be more difficult to obtain reliable statistics in the analysis.

METHODS

Molecular Framework Analysis. Murcko et al.[9] fragmented a molecular structure into four different subunits: ring systems, side chains, molecular framework, and linkers. Ring systems are defined as cycles or cycles sharing an edge within the graph representation of a molecule. Linkers are atoms connecting two ring systems. Ring systems and relevant linkers in a compound constitute the molecular framework. All atoms not belonging to the molecular framework are part of the side chains. We propose here a modification to Murcko's molecular framework (MF) concept. A molecule is dissected into three parts: side chains, terminal rings (TR), and a molecular bridge (as shown in Figure 1). Our side chain definition is the same as used by Murcko et al. Terminal ring systems refer to ring systems which have only one connection to other ring systems. The molecular bridge connects all of the terminal rings. The difference between the definition of a molecular bridge and Murcko's definition of a linker is that a molecular bridge can also include ring systems, while a linker does not include any ring system at all. Any ring system that is directly connected through linkers to more than one other ring system is regarded as part of the molecular bridge. Thus, the molecular framework is the combination of the terminal rings and the molecular bridge.

The procedure to decompose the molecules is as follows: First, all the side chains in a compound are identified and removed analogous to earlier studies.[9] Then the ring systems, which are only connected to one other ring system, are identified and labeled as terminal rings. The rest is labeled as the molecular bridge. Any compound can therefore be classified according to the number of terminal ring systems it contains. Some examples for the different topological classes are shown in Figure 2. For example, compound **1** belongs to the class, which has only one ring system and no molecular bridge (1TR). Compound **2** belongs to the class which has two ring systems directly connected without a molecular bridge (2TR). Compounds **3** and **4** have two ring systems and a molecular bridge (2TR+B). Compound **5** has three terminal ring systems and a molecular bridge (3TR+B). More complex molecules can be assigned to classes 4TR+B, 5TR+B, etc. The 2TR+B class can be further divided into subclasses depending on the number of ring systems included in the molecular bridge. For example, compound **4** belongs to the subclass which does not have ring in the molecular bridge (2TR+B_0), while compound **3** is assigned to the subclass that has one ring system in the bridge (2TR+B_1). In the current study, the three major subclasses (2TR+B_0, 2TR+B_1, 2TR+B_2) of the 2TR+B class are considered. Compounds having more than 2 ring systems in the molecular bridge are not included because of their low occurrence in the drug set. Any subdivision of the 3TR+B class is not considered here because of the limited number of compounds in this class for the drug set.

Data Sets. Three data sets, marketed drugs, clinical candidate compounds and bioactive compounds, are used in this study. All three sets are from GVKBio.[16] Only compounds with a number of heavy atoms between 5 and 60 are used in the analysis. Compounds associated to human targets and a reported potency lower than or equal to 10 μM for that target were selected for the bioactive set. Altogether 741 464 active compounds associated with 1486 human gene targets were used as the bioactive compound set to represent the biologically relevant chemical space. For the drug and clinical candidate sets, compounds labeled as antiseptics, antivirals, disinfectants, and vitamins were removed resulting in a set of 957 drugs and a set of 4354 clinical candidate compounds. The number of compounds for each topological class and data set is given in Table 1.

Property Calculations. ClogP were calculated with the Biobyte ClogP program.[17] Molecular frameworks were generated by a Pipeline Pilot[18] component. An in-house C++ program based on the Openeye toolkit[19] was used to identify the terminal ring systems and the molecular bridge for the molecules. Another in-house program was used to calculate the number of heavy atoms, rotatable bonds, fused rings and ring systems. Here the rotatable bonds refers to the nonterminal acyclic bonds, which exclude the double bonds, triple bonds, and amide bonds and the number of fused rings is calculated by subtracting the number of ring systems from the number of smallest set of smallest rings (SSSR).[20] All statistical analysis was done with JMP.[21]

RESULTS AND DISCUSSION

Previous studies[5,11] have showed that marketed drugs are generally smaller and less lipophilic than clinical candidates and bioactive compounds. But how drugs differ from other compounds in terms of topology has not, to the best of our knowledge, been studied systematically before. Molecular framework analysis is carried out on the three data sets (drugs, clinical candidates and bioactive compounds) to identify the molecular framework, terminal ring systems, and molecular bridge in each compound. The percentage of compounds with a molecular framework is above 96% (Figure 3) for all the three data sets. Thus most of the drugs, clinical candidates and bioactive compounds have at least one ring system. A molecular framework can be further divided into terminal rings and a molecular bridge and classified according to the scheme described in the methods section. As seen in Figure 3, there is a clear difference between drugs, clinical candidates and bioactive compounds

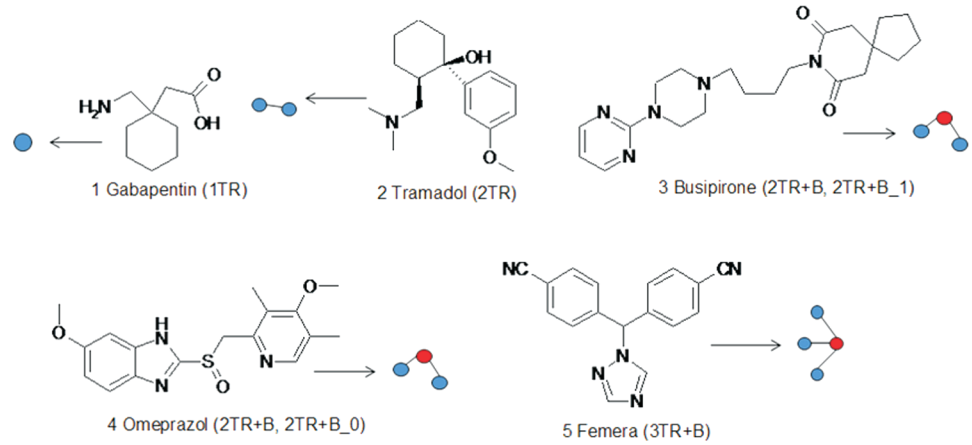

Figure 2. Examples of the different topology classes, 1TR, 2TR, 2TR+B, and 3TR+B. Also included are the subdivision of the 2TR_B class into the 2TR+B_0 and 2TR+B_1 classes.

Table 1. Comparison of Number of Compounds (NoC) and the Median of Number of Heavy Atoms (HEV) and ClogP for Drugs, Clinical Candidates, and Bioactive Compounds for the Different Topology Classes

	drugs			clinical candidates			bioactive cmpds		
	NoC	HEV	ClogP	NoC	HEV	ClogP	NoC	HEV	ClogP
1TR	340	20	2.2	729	22	2.4	32378	22	2.7
2TR	105	21	2.9	463	23	3.1	40064	25	3.6
2TR+B	423	26	3.2	2593	29	3.6	537988	31	4.1
2TR+B_0	289	24	2.9	1051	26	3.1	104731	27	3.6
2TR+B_1	113	28	3.6	1145	30	3.8	265634	30	4.0
2TR+B_2	17	34	5.7	343	35	4.4	141198	35	4.4
3TR+B	56	27	4.1	453	37	4.6	118755	38	5.1

with respect to the presence of a molecular bridge. Only 50.2% of the drugs have a molecular bridge in comparison to 71.3% of the clinical candidates and 86.7% of the bioactive compounds.

On the basis of the number of terminal rings and the presence/absence of a molecular bridge in the framework, compounds in the three data sets were further subdivided into different topological classes. As can be seen in Figure 4, 36.8% of the drugs and only 4.4% of the bioactive compounds belong to the 1TR (one terminal ring and without a molecular bridge) class. However, 45.7% of drugs and as many as 72.8% of bioactive compounds belong to the 2TR+B class. Also, 6.2% of drugs and as many as 16.1% of the bioactive compounds belong to the 3TR+B class. For the topological classes that have more than three terminal rings, the number of compounds is very low for all the three data sets, and they are therefore excluded from this study.

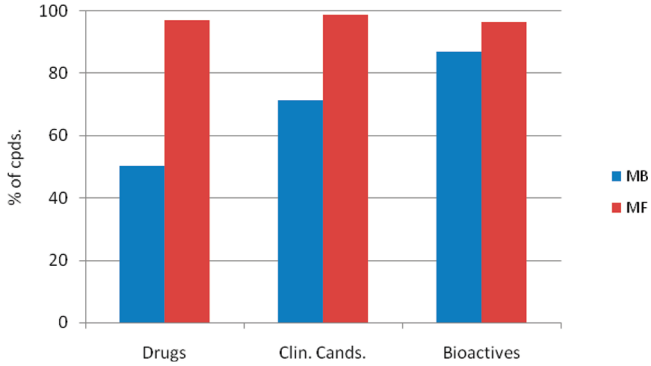

Figure 3. Percentage of compounds in the drugs, clinical candidates, and bioactive compound sets, which have a molecular framework (MF, red) and which have a molecular bridge (MB, blue).

These results show that the drug set is enriched with compounds containing only one ring system or two ring systems directly connected to each other (1TR and 2TR classes). In contrast, the bioactive compound set comprise of more compounds with a molecular bridge (2TR+B and 3TR+B). Furthermore looking into the details for 2TR+B class, it is noted that the drug set has a higher percentage of compounds belonging to the 2TR+B_0 subclass than the two other sets and accordingly the drug set has a lower percentage of compounds in the other two 2TR+B (2TR+B_1, 2TR+B_2) subclasses than the clinical candidate and bioactive sets. The differences in the topological class distributions show that drugs have generally fewer terminal ring systems and for the 2TR+B class fewer ring systems in the

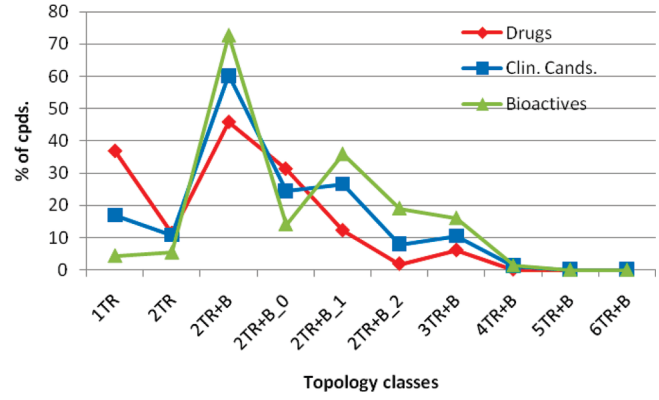

Figure 4. Topology class distributions for drugs, clinical candidates and bioactive compounds are depicted. The distributions for the three subclasses (2TR+B_0, 2TR+B_1, 2TR+B_2) of the 2TR+B class are also included.

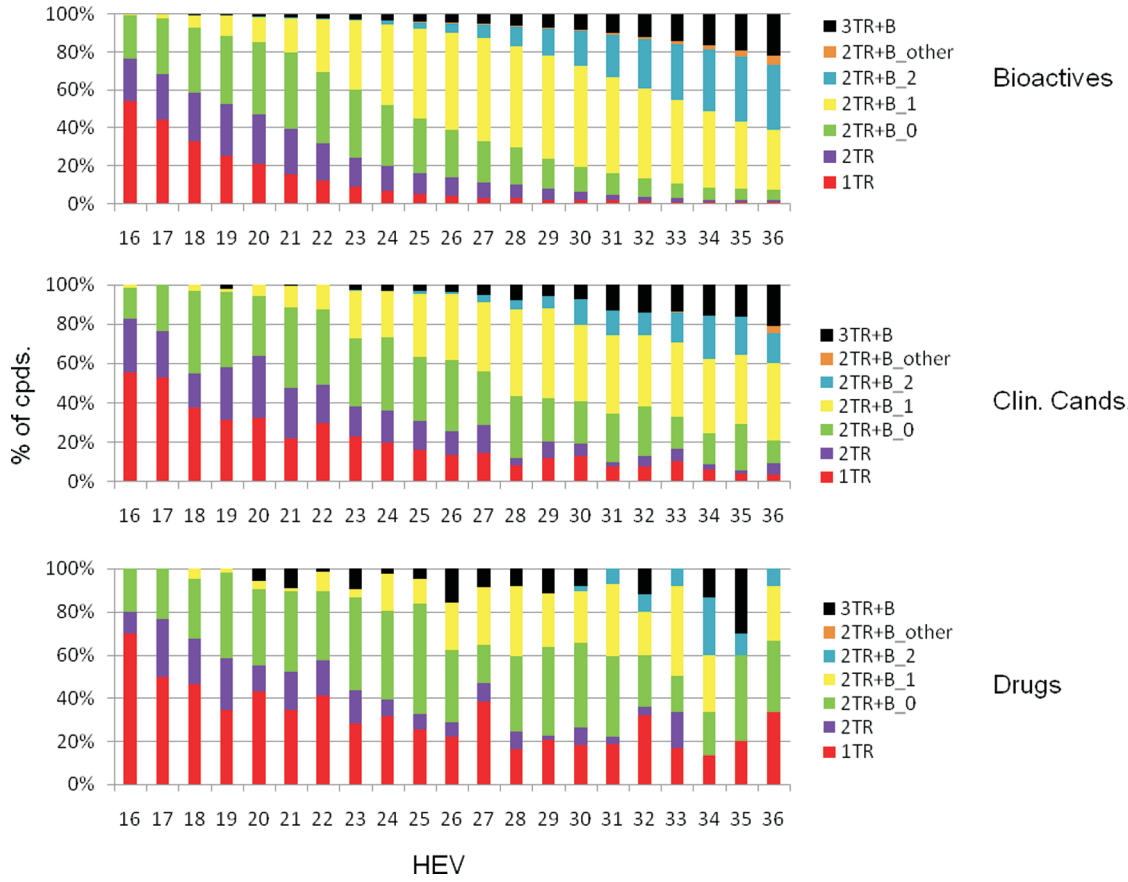

Figure 5. Topology class distributions for drugs, clinical candidates, and bioactive compounds depicted at different heavy atom counts (HEV).

molecular bridge than both the clinical candidate and bioactive molecule sets.

Earlier studies have shown that drugs, clinical candidates, and bioactive compounds have significantly different average molecular size and lipophilicity. Hence, a relevant question is therefore if the observed trends described above are due to the effect of size alternatively lipophilicity or if the trends are still present after correcting the results for these properties? The topology distribution for drugs, clinical candidates, and bioactive compounds were therefore compared for compounds with the same number of heavy atoms. Compounds in all the three sets were sorted based on the number of heavy atoms and the distributions for the four main topology classes were calculated. Data from 16 to 36 heavy atoms are displayed in Figure 5. This interval covers most of the compounds in the three data sets. The average percentage of compounds for each topological class (averaged over all the different number of heavy atoms) is shown in Figure 6a. Hence, excluding the molecular size effect, the drug set still have a significantly higher fraction of compounds in the 1TR class compared to the clinical candidate and bioactive compound sets. Although the bioactive set has the highest percentage of compounds in the 2TR+B class, the drug set has a higher percentage of compounds belonging to the 2TR+B_0 subclass compared to the clinical candidate and bioactive sets. All three data sets have similar percentage of compounds in the 2TR and 3TR+B classes. Thus drugs tend to be overrepresented among the molecules that have only one ring system compared to the other data sets and also enriched with 2TR+B compounds having simple bridges even when the molecular size effect is taken into account.

There is a possibility that the drugs act on a different set of protein targets compared to the bioactive molecules and that it is this target bias that give rise to the differences observed in Figures 5 and 6a. To investigate this possibility the target Entrez Gene id annotations for the compounds in all the three data sets were used to identify compounds interacting with the same target. Altogether 214 targets have compounds annotated across the drugs, clinical candidates, and bioactive compound sets. The target distributions (for these 214 common targets) for the drugs, clinical candidates and bioactive compounds are compared for each heavy atom count (Figure 7). For all the three compound sets the variation of target coverage across all heavy atom count is small. However, the results could still be influenced by a skewed compound distribution among the different targets. If most of the drugs bind to a small group of targets and most of the bioactive compounds bind to another set of targets, there could still be a target bias. Therefore, a target corrected topology class distribution was calculated by the following formula:

$$P_{ij} = \sum_{k=1}^{N_i} \frac{F_{ij}^k}{N_i} \quad (1)$$

where P_{ij} refers to the average fraction of compounds in topology class j for heavy atom interval i; F_{ij}^k corresponds to the fraction of compounds (belonging to topology class j and heavy atom interval i) associated with target k (among the 214 common targets that are present in all three databases for heavy atom interval i); N_i is the number of common targets for the heavy atom interval i. In this way, the

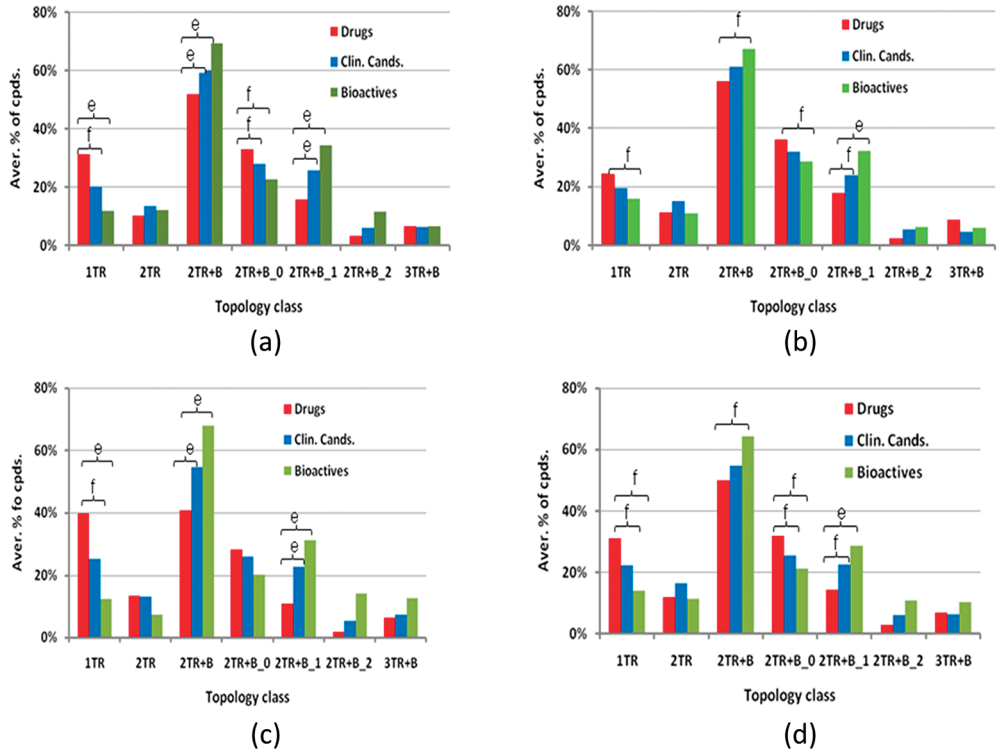

Figure 6. Averaged topology class distribution for drugs, clinical candidates, and bioactive compounds: (a) percentage value for each topological class averaged over all the different number of heavy atoms, (b) percentage value for each topological class averaged over all targets present in all three categories and all number of heavy atoms, (c) percentage value for each topological class averaged over all the different ClogP intervals, and (d) percentage value for each topological class averaged over all targets present in all three categories and all ClogP intervals. The differences are statistical significant at the (e) $p < 0.001$ and (f) $p < 0.05$ levels.

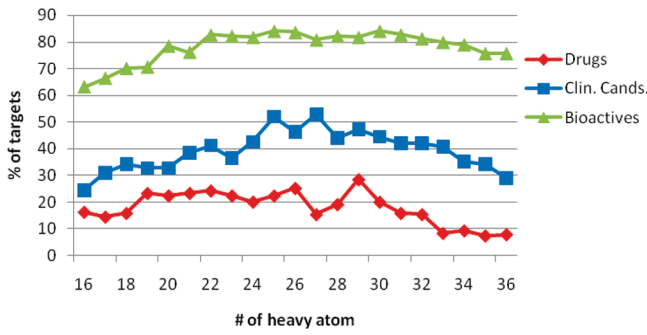

Figure 7. Percentage of the 214 common targets present in the different data sets as a function of the number of heavy atoms.

calculated fraction for topology class j and heavy atom interval i is averaged over the common targets. This correction for any potential target bias in the data sets will give a more fair comparison between the drugs, clinical candidates and bioactive compounds. The average value for the target corrected topology class distribution summed over all heavy atom intervals can be calculated with formula 2

$$\bar{P}_j = \frac{\sum_{i=1}^{M} P_{ij}}{M} \qquad (2)$$

Here, $\bar{P}_j$ is the averaged fraction of compound for topology class j summed over all heavy atom intervals i; P_{ij} is defined as in (**1**); M is the total number of intervals. The comparison for the three data sets is shown in Figure 6b and the detailed information for the target corrected topology class distributions for different heavy atom count is plotted in the Supporting Material (Figure S1). Although the difference between drugs and bioactive molecules for the 1TR and 2TR+B classes has decreased after adjusting for the target bias, drugs still have the highest percentage of 1TR and 2TR+B_0 compounds and lowest percentage of total 2TR+B compounds of the three data sets. Thus even after correcting for potential target bias in the different databases, drugs are still overrepresented among compounds with only one ring system.

In a similar way the potential influence of ClogP on the observed differences between the three data sets was investigated. All compounds with a ClogP in between −3.0 and 6.0 were selected and the topology class distributions were calculated for the nine different ClogP intervals in two different scenarios, that is, with or without the correction for potential target bias. The topology class distribution (no target correction) for the different ClogP intervals are displayed in Figure 8 and the topology class distribution summed over all ClogP intervals are shown in Figure 6c (no target correction) and d (with target correction). The target correction is done in the same way as for the target correction for the heavy atom count. In both scenarios, the drug set has the largest fraction of the 1TR class and the smallest fraction of the 2TR+B class, while the bioactive set has the reverse distribution and the clinical candidate set is in the middle. Within the 2TR+B class, the drug set has the highest percentage of compounds in the 2TR+B_0 subclass. All three sets have similar fractions for the 2TR and 3TR+B classes. The results show that for compounds in the same ClogP interval, drugs are overrepresented among the compounds with only one ring system and among 2TR+B compounds that have a simple bridge. The detailed

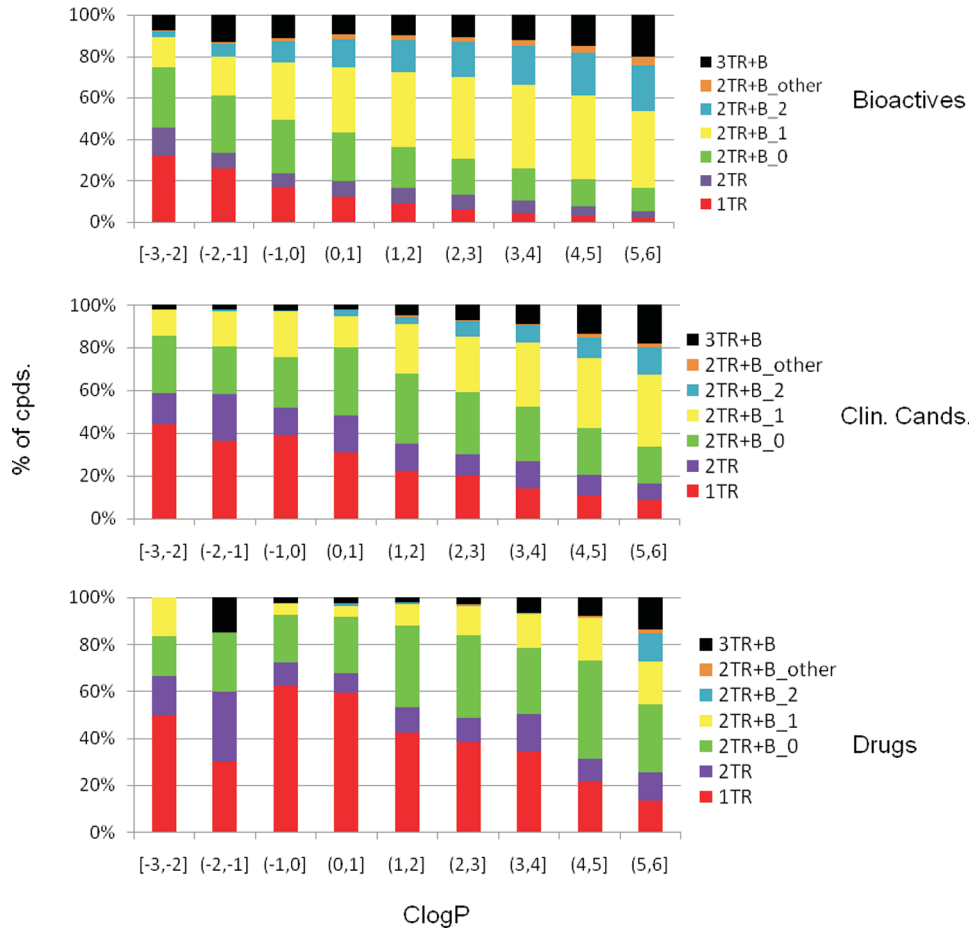

Figure 8. Topology class distributions for drugs, clinical candidates and bioactive compounds depicted for different ClogP intervals.

information for the target corrected topology class distributions for different ClogP interval is plotted in the Supporting Material (Figure S2). In conclusion, after correcting for differences in molecular size and lipophilicity, there are still differences in molecular topology between drugs, clinical candidates and bioactive molecules.

The serotonin transporter (SERT) and the beta-1 adrenergic receptor are chosen as examples to illustrate the differences observed between drugs, clinical candidates, and bioactive compounds. There are 18 drugs, 37 clinical candidates and 9175 bioactive compounds annotated with an activity of less than or equal to 10 μm on the SERT. There are 62 drugs, 52 clinical candidates, and 1650 bioactive compounds with an activity on the beta-1 adrenergic receptor below the 10 μm cutoff. The topology class distributions are displayed in Figure 9a and b. The drug set has the highest fraction of 1TR compounds and the lowest fraction of 2TR+B and 3TR+B compounds for both targets. Representative 1TR, 2TR, and 2TR+B SERT inhibitors and beta-1 receptor antagonists are displayed in Figure 9c and d. Clomipramine and Fluvoxamine in Figure 9c are SERT inhibitors and therefore used as antidepressants. Metoprolol and Propranolol in Figure 9d are beta-1 antagonists and therefore used to lower the blood pressure.

The heavy atom count and ClogP distributions for compounds in the 1TR, 2TR, 2TR+B, and 3TR+B classes are shown in Figures 10 and 11, respectively. Drugs are left shifted for all the four topological classes (i.e., drugs are generally smaller than clinical candidates and bioactive compounds) with respect to heavy atom count. This is consistent with earlier reports.[5] The median heavy atom count for the four classes is listed in Table 1. The difference in number of heavy atoms between drugs and the two other data sets is statistically significant (shown in the Supporting Materials, Table S5) and gradually increase with the number of terminal ring systems. Drugs have on average approximately two heavy atoms less than bioactive compounds for the 1TR class. However, for the 3TR+B class, the difference is approximately 11 heavy atoms. Hence, when the molecular framework becomes larger with a molecular bridge and more terminal ring systems, the size difference between drugs and bioactive compounds becomes larger. We see the same trend for the ClogP difference between drugs and bioactive compounds. Drugs have statistically lower ClogP than bioactive compounds across all four topological classes. The ClogP difference between drugs and clinical candidates is not statistically significant for the 1TR and 2TR classes, but are significant for the 2TR+B and 3TR+B classes.

For compounds with a molecular bridge (2TR+B etc.), the heavy atom count and number of rotatable bonds in the molecular bridge are compared for the three data sets. As shown in Figure 12a, small molecular bridges (less than three heavy atoms) are most frequent in the drug set (40%), followed by the clinical candidate set (19.7%) and bioactive compound set (9.1%). The distribution of number of rotatable bonds in the molecular bridge is shown in Figure 12b. It is observed that compounds in the drug set has significantly less number of rotatable bonds than compounds in the clinical candidate and bioactive compound sets. These results indicate

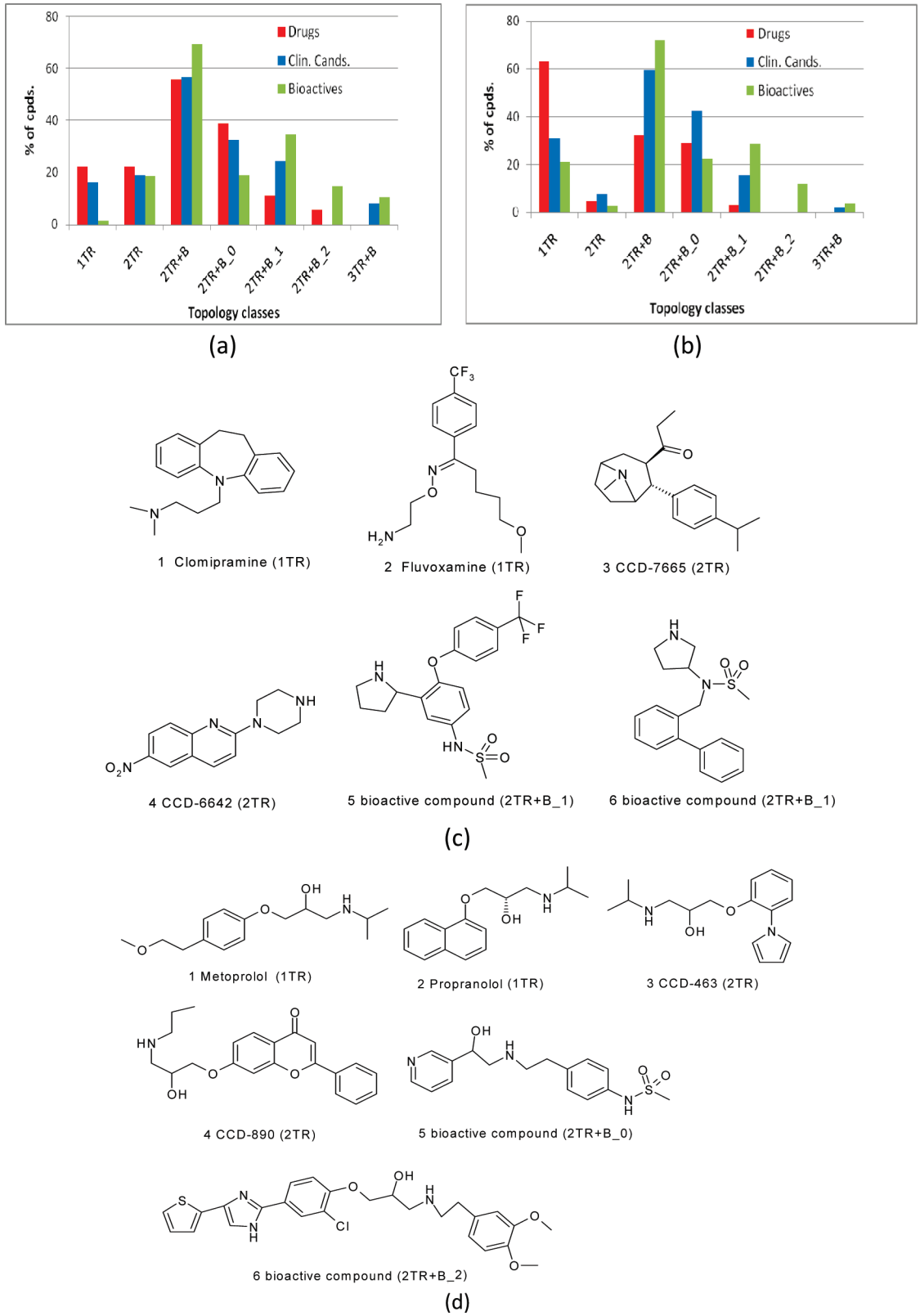

Figure 9. Topology class distribution for drugs, clinical candidates, and bioactive compounds annotated as (a) serotonin transporter inhibitors and (b) beta-1 adrenergic receptor antagonists. Examples of topological classes for drugs, clinical candidates (CCD) and bioactive compounds annotated as (c) serotonin transporter inhibitors and (d) beta-1 adrenergic receptor antagonists.

that drugs tend to have smaller and more rigid molecular bridges than other types of compounds. The complexity of the ring systems for the three data sets is compared in Figure 13. Here the concept of Ring System Complexity (RSC) is introduced and defined as the number of smallest set of smallest rings (SSSR)[20] minus the number of ring systems in a molecule. Generally, when the number of ring system increases, the RSC decreases (Figure 13 and Table S4 in Supporting Material). An exception is for drugs that belongs the 2TR+B_2 subclass. However, there are only 17 compounds in this subclass, which makes any interpretation difficult. The trend is reasonable. It is expected that molecules

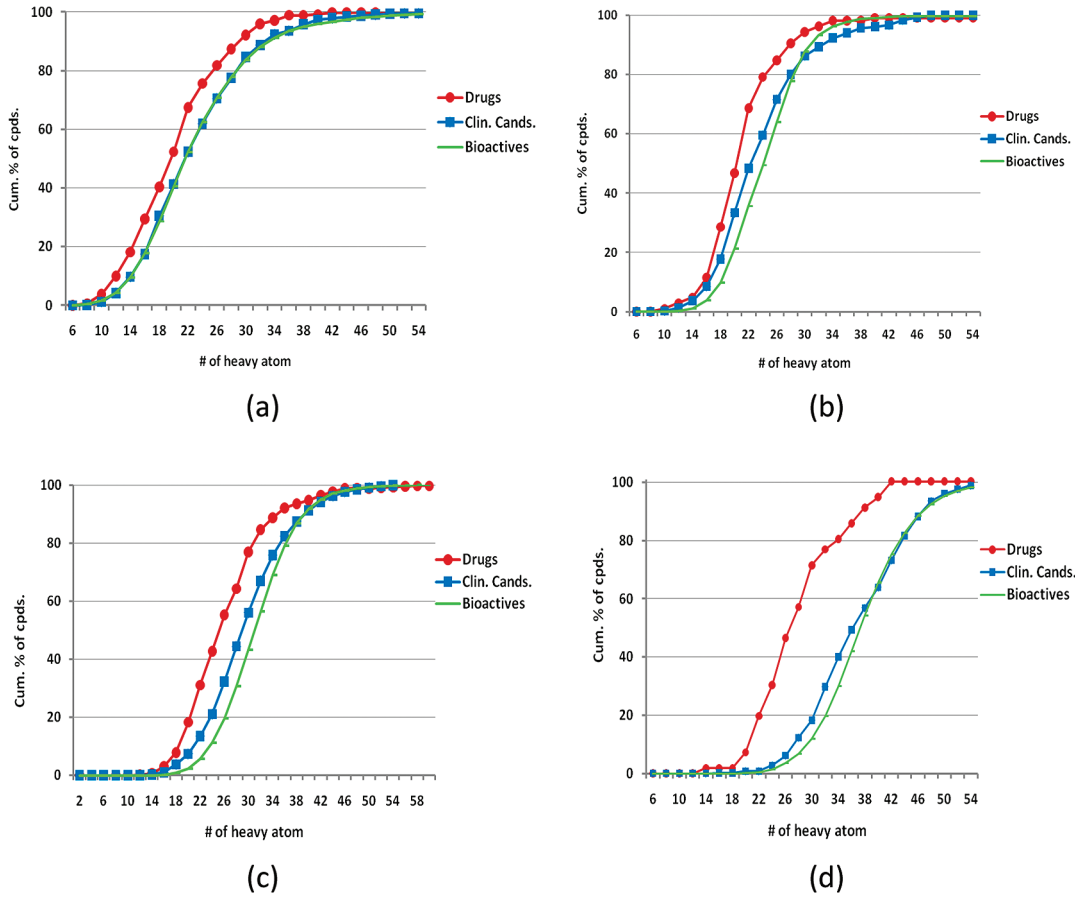

Figure 10. Cumulative percentage of compounds versus number of heavy atoms for drugs, clinical candidates and bioactive compounds for the (a) 1TR class, (b) 2TR class, (c) 2TR+B class, and (d) 3TR+B class.

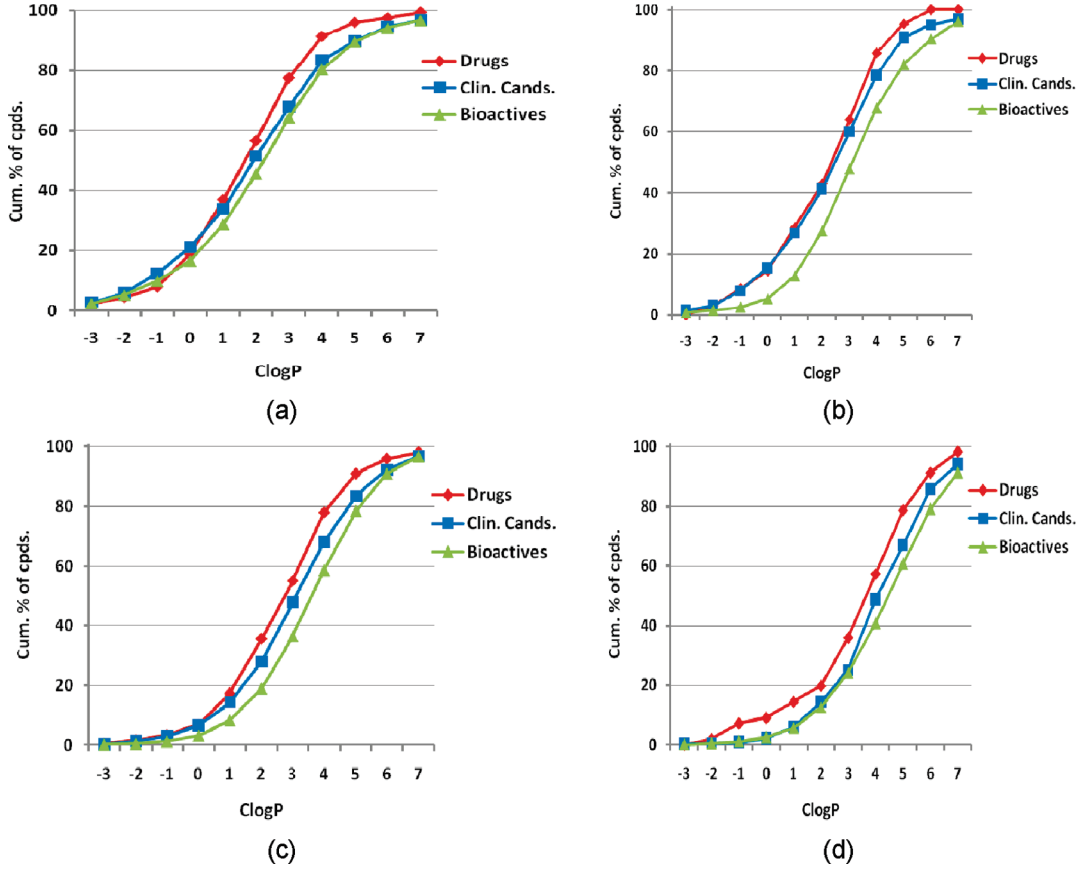

Figure 11. Cumulative percentage of compounds versus ClogP for drugs, clinical candidates, and bioactive compounds for the (a) 1TR class, (b) 2TR class, (c) 2TR+B class, and (d) 3TR+B class.

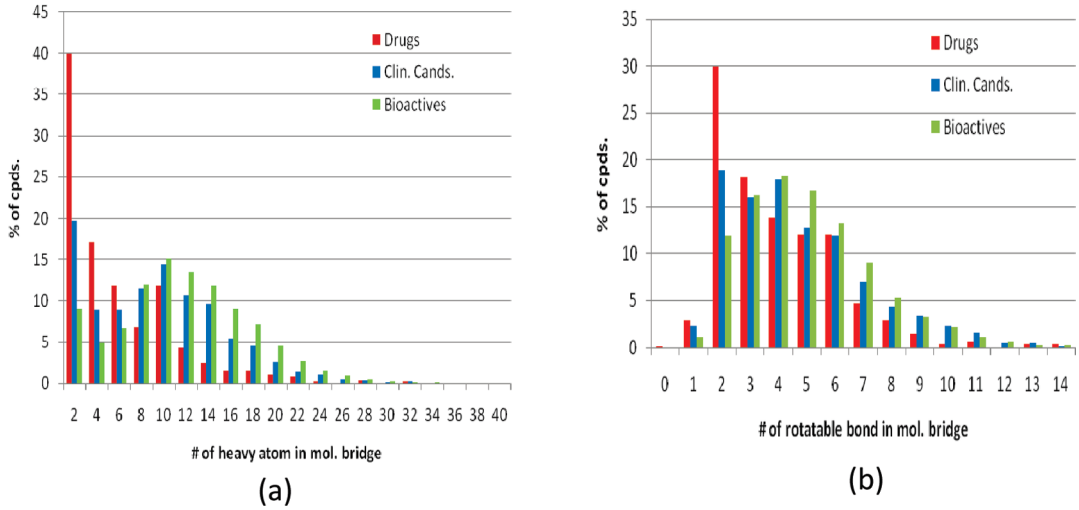

Figure 12. Number of heavy atoms (a) and number of rotatable bonds (b) in the molecular bridge for drugs, clinical candidates, and bioactive compounds.

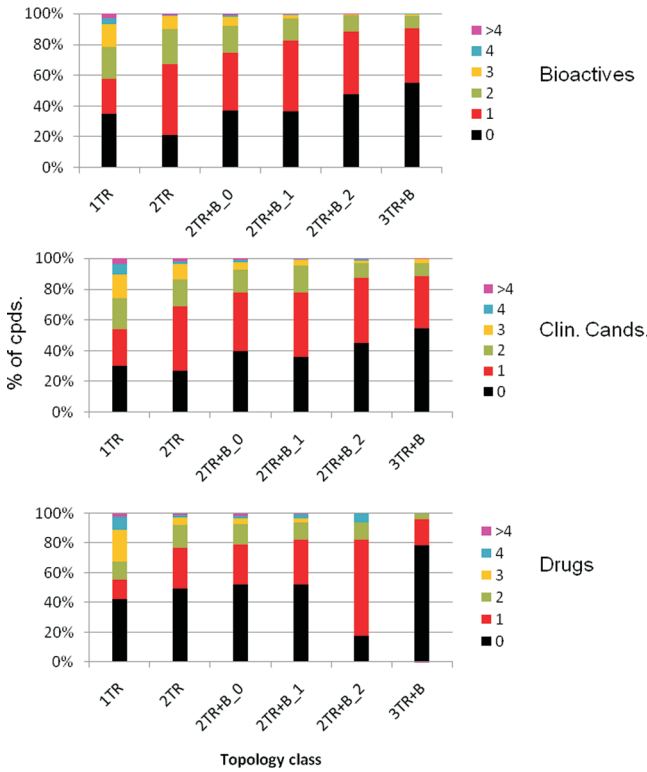

Figure 13. Ring system complexity (RSC) for the different topology classes depicted for the drugs, clinical candidate and bioactive compound sets.

that have many ring systems have lower RSC, due to restrictions on the molecular size for compounds aimed to be drugs. An interesting observation is that drugs tend to have significantly lower RSC than clinical candidate and bioactive sets for most of the topology classes. An exception is the 1TR class. For the 1TR class all three sets have a similar RSC (Figure S3 in Supporting Material). This might at a first glance to be counterintuitive, since a lot of steroid drugs belong to the 1TR class. Steroids have a complex ring structure and therefore a large RSC value. One could therefore expect that the RSC for drugs belonging to the 1TR class would be higher than for the 1TR class of the other two sets, however, as is shown in Figures 13 and S3, this is not the case. The RSC is zero for more than 40% of the drugs in the 1TR class. Thus for these compounds the ring system consists of only one ring.

CONCLUSIONS

All our results show that drugs are overrepresented among molecules with only one ring system compared to the clinical candidate and bioactive compound sets. This trend is to a large extent independent of size and lipophilicity and remains when taking into account that the three data sets have different target bias. The drug set tends to have fewer heavy atoms and rotatable bonds in the molecular bridge than the two other sets for compounds with a molecular bridge. Most compounds that belong both to the drug set and the 2TR+B class have no ring system in the molecular bridge. However, compounds in the clinical candidate and bioactive molecule sets that belong to the 2TR+B class have larger molecular bridges that contain one or more ring systems. The heavy atom count and ClogP distributions for drugs, clinical candidates, and bioactive compounds are compared for the most common topological classes and the results show that drugs are generally smaller and less lipophilic than clinical candidates and bioactive compounds for all the topological classes. The difference between drugs and bioactive compounds regarding the topology might indicate that a compound with only one ring system is on average more drug-like than molecules with a molecular bridge. The ring system complexity (RSC) tends also to be lower for drugs. Preliminary results show that the molecular topology is related to promiscuity.[22] However, further work need to be done to investigate the relationship between the molecular topology and important physicochemical, ADME and pharmacokinetic properties. Also, the relationship between the topology class distribution of drugs and the drug approval year will be investigated in the future. Another reason for the observed differences might be due to the chemical reactions in current use by medicinal chemists. Many of these reactions like amide coupling might be predominantly used for producing compounds of the 2TR+B topological class.

ACKNOWLEDGMENT

The authors would like to thank Dr. Sorel Muresan and Dr. Ingemar Nilsson for valuable discussions, Dr. Péter

Várkonyi for providing the GVKBio databases, Dr. Jens Sadowski for providing the source code of SSSR perception algorithm, and Dr. Niklas Blomberg and Dr. Paul Leeson for critical revisions of the manuscript.

Supporting Information Available: Figures S1 and S2 showing the target bias corrected topology class distributions for different heavy atom count and different ClogP intervals, respectively, Figure S3 showing the comparison and statistical confidences of the RSC for drugs, clinical candidates and bioactive compounds, Table S4 listing the Wilcoxon statistical test of the differences in the average RSC for topological classes, and Table S5 listing the Student *t* test results for the differences in the average values for ClogP and heavy atom count for the topological classes listed. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3–25.
(2) Oprea, T. I.; Davis, A. M.; Teague, S. J.; Leeson, P. D. Is there a difference between leads and drugs? A historical perspective. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1308–35.
(3) Wenlock, M. C.; Austin, R. P.; Barton, P.; Davis, A. M.; Leeson, P. D. A comparison of physiochemical property profiles of development and marketed oral drugs. *J. Med. Chem.* **2003**, *46*, 1250–1256.
(4) Vieth, M.; Sutherland, J. J. Dependence of molecular properties on proteomic family for marketed oral drugs. *J. Med. Chem.* **2006**, *49*, 3451–3453.
(5) Tyrchan, C.; Blomberg, N.; Engkvist, O.; Kogej, T.; Muresan, S. Physicochemical property profiles of marketed drugs, clinical candidates and bioactive compounds. *Bioorg. Med. Chem. Lett.* **2009**, *19*, 6943–6947.
(6) Leeson, P. D.; Springthorpe, B. The influence of drug-like concepts on decision-making in medicinal chemistry. *Nat. Rev. Drug Discovery* **2007**, *6*, 881–890.
(7) Lovering, F.; Bikker, J.; Humblet, C. Escape from Flatland: Increasing saturation as an approach to improving clinical success. *J. Med. Chem.* **2009**, *52*, 6752–6756.
(8) 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 Discovery Today* **2009**, *14*, 1011–1020.
(9) Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.
(10) Grabowski, K.; Schneider, G. Properties and architecture of drugs and natural products revisited. *Curr. Chem. Biol.* **2007**, *1*, 115–127.
(11) Lee, M. L.; Schneider, G. Scaffold architecture and pharmacophoric properties of natural products and trade drugs: Application in the design of natural product-based combinatorial libraries. *J. Comb. Chem.* **2001**, *3*, 284–289.
(12) Hu, Y.; Bajorath, J. Scaffold distributions in bioactive molecules, clinical trials compounds, and drugs. *ChemMedChem* **2009**, *5*, 187–190.
(13) Wang, J.; Hou, T. Drug and drug candidate building block analysis. *J. Chem. Inf. Model.* **2010**, *50*, 55–67.
(14) Xu, Y. J.; Johnson, M. Algorithm for naming molecular equivalence classes represented by labeled pseudographs. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 181–185.
(15) Xu, Y. J.; Johnson, M. Using molecular equivalence numbers to visually explore structural features that distinguish chemical libraries. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 912–926.
(16) *GVKBio databases 2009*; GVK Biosciences Private Ltd.: Hyderabad, India, 2009.
(17) BioByte ClogP, version 4.3, http://www.biobyte.com (accessed Feb. 20, 2010).
(18) *Pipeline Pilot*, version 7.5; Accelrys: San Diego, CA.
(19) *OEChem toolkit 1.7.2*; OpenEye Scientific Software, Inc.: Santa Fe, NM.
(20) Sorkau, E. Ringerkennung in chemischen Strukturen mit dem Computer. *Wiss. Z. Tech. Hochsch. Leuna-Meuseburg.* **1985**, *27*, 765–770.
(21) *JMP*, version 7.0; SAS Institute Inc.: Cary, NC.
(22) Yang, Y.; Chen, H.; Nilsson, I.; Sorel, M.; Engkvist, O. Investigation of the relationship between the topology and selectivity for druglike molecules. *J. Med. Chem.* **2010**, *53* (21), 7709–7714.

CI1002558