Molecular topology, a novel descriptor for compound quality assessment

Rurika Oka

Department of Biology, Lund University, SE-22100 Lund, Sweden

Abstract

The pharmaceutical industry is currently facing a high clinical attrition rate. In order to prevent the late-stage clinical failure, many investigations on compound quality and drug-likeness of compounds have been carried out. It has been widely accepted that molecular size and lipophilicity plays an important role in compound quality. Many attempts have been done to find out other factors which can influence compound quality beyond size and lipophilicity. Recently, a molecular topology concept has been put forward and its influence on compound quality has been investigated. It has been shown that drugs have higher fraction of compounds with only one ring system compared to clinical candidate and bioactive compounds. As an extension to the previous studies, the aim of this project is to further investigate how the molecular topology influences some of the most important physicochemical properties of molecules as well as the compound potency efficiency indices in general. Our results show that among reported molecules in the literature, compounds with only one ring system are smaller in size, less lipophilic and therefore has a higher probability to be less toxic. Interestingly compounds which have a simple topology also show advantage in terms of potency efficiency such as ligand efficiency (LE), ligand lipophilic efficiency (LLE) and ligand-efficiency-dependent lipophilicity index (LELP) compared with compounds which have a more complex topology. Thus a novel hypothesis why compounds with only one ring system are abundant among drugs has been proposed. On average molecules with only one ring system seems to bind more strongly to its protein target; this might reduce the

necessary size of the molecule to reach a certain potency level. The reduction in size and lipophilicity reduces the risk of failure in clinical trials.

Introduction

Drug discovery and development is a time consuming process which typically takes 15 to 20 years from the target identification until a drug makes it to the market. During this lengthy process, numerous compounds are tested, synthesized and validated in order to achieve the optimal efficacy and safety profile. Firstly, a potential target causing a disease is identified; the target may be over-expressed, under-expressed or mutated in a patient. Then, many compounds are screened for an interaction to the target to discover "hit" compounds, which will be starting molecules for the whole drug development process. The identified "hits" may have undesirable side-effects or may not have good bioavailability; the potency and bioavailability of compounds are improved and promiscuity is minimized. The compounds are now called "lead" compounds and their efficacy of lead compounds are examined in animal models before they are further optimized. Once the compounds' safeties are assured, they can be tested in clinical trials: Phase I assess the safety in a few healthy individuals, Phase II assesses the efficacy in a small group of patients, and Phase III extends the assessment to a larger group of patients. Only compounds passed phase III will be registered as a drug and will be on market.

Historically drug discovery was an iterative process of compound synthesis and *in vivo* screening. This paradigm has changed by the advancement of *in vitro* high-throughput screening technology and *in silico* techniques. Now computational methods are commonly used from target identification to lead optimization: gene sequence analysis and gene expression analysis find out potential targets, virtual screening discovers hit compounds and structural analysis suggests possible ways of compound optimization, for example. Another

use of a computational method is to predict compound quality; by estimating physicochemical properties and bioavailability in an early stage and avoiding the advancement of compounds with poor basic properties, it is possible to reduce the number of late stage failure.

The paradigm shift has largely improved hit identification efficiency; however the pharmaceutical industry still faces a high attrition rate. It was estimated by U.S. Food and Drug Administration (FDA) that eventually only 8 % of the compounds entering to Phase I can reach the market.[1] In order to decrease the late-stage clinical attrition, it is critical to identify compounds which are unlikely to succeed and to terminate the development of these compounds as early as possible. The most important properties for a drug molecule are so called ADMET (absorption, distribution, metabolism, excretion and toxicity) properties, and they are investigated at an early stage in the drug development process. Only compounds with a good ADMET property profile should be advanced to the clinical trial.

The well-known Lipinski's Rule of Five was proposed to assess the oral bioavailability of a compound.[2] It defines cutoffs for molecular mass, lipophilicity, number of hydrogen bond donors and acceptors which captures 90% of oral drugs and clinical candidates.[2-4] Despite the fact that Rule of Five is often used as a filter, a number of drug compounds violates one or more criteria,[2] therefore this should be kept as a guideline for bioavailability estimation. As mentioned in the Lipinski's Rule of Five, the importance of lipophilicity and molecular size is widely accepted. Leeson and Springthorpe observed that high promiscuity, as well as other ADMET liabilities including low solubility, high plasma protein binding (PPB), and high hERG (human ether-ago-go-related potassium channel protein) inhibition,[5] can be caused by high lipophilicity. Low solubility and high PPB leads to low free plasma concentration of the compounds, which is critical to get desired pharmacology effect. hERG is a voltage gated potassium ion channel, and high inhibition of this protein can cause QT prolongation which may result in fatal cardiac arrhythmia.[6] Comparison of several simple physicochemical

properties analysis done by Wenlock *et al.* showed that drugs have significantly lower lipophilicity and molecular size than general bioactive compounds, compounds which have effects on living systems, and clinical candidates.[7]

Another important factor considered during the drug development process is potency. Potency has an influence on the amount of dose in most cases; low potency requires high concentration of the drug compound which increases the risk of toxic response. It has been observed that there is a strong correlation between the molecular size and the potency,[8-10] and Ligand Efficiency (LE) has been introduced to correct the potency for the molecular size by dividing by the heavy atom numbers. The use of this index enables to prioritize small molecules with lower potency rather than large molecules with higher potency.[11,12] Even though LE still has some size bias,[8,9,13] this index is widely accepted to assess the ligand affinity and practically used. Another useful factor related to the potency is Ligand Lipophilicity Efficiency (LLE) investigated by Leeson and Springthorp.[5] There is a general trend of increase in potency during the drug development process which also increases the lipophilicity.[10,14] As discussed earlier, lipophilicity should be kept low to avoid the ADMET liabilities and potential promiscuity; LLE is a useful index to improve potency while controlling the lipophilicity at the same time. While LE and LLE are two indices that complement each other, Ligand-Efficiency-dependent LipoPhilicity (LELP) is another potency efficiency index which correlates both lipophilicity and molecular size. It has been shown a single score combined molecular mass and lipophilicity provides better prioritization of compounds than individual parameter cut-off rules.[10] Drugs display significantly lower LELP value than general hits, leads and phase II clinical candidates, while it is not always the case for LLE.[14]

Several investigations have been carried out to find out other compound quality criteria beyond molecular size and liphophilicity. Ritchie *et al*. observed the negative impact of a

large number of aromatic rings in a molecule on the compound quality,[15] and Lovering *et al.* reported higher fraction of sp3 hybridized carbon atoms and the presence of chiral centers increases the clinical success rate.[16] These two observations complement each other.

Recently molecular topology class was proposed[17] as an extension of molecular framework proposed by Bemis and Murcko.[18] The advantage of this molecular topology class is that this classification divides compounds into only a few topological classes compared to the earlier molecular framework, and therefore it is possible to apply rigorous statistical tests on the results.[17] This classification defines the terminal rings (TR), molecular bridge (B) and side chains, and the number of terminal rings and molecular bridge determines the molecular topology class of the molecule. It has been reported that the fraction of compounds with only one ring system is higher in drugs compared to clinical candidates and general bioactive compounds.[17] This trend is still valid after correcting for lipophilicity, molecular size, and potential target bias in the data set.[19] f_{MF} is another descriptor which related to the concept of molecular topology; it is the fraction of molecular framework, which consists of terminal rings and molecular bridge. It has been shown that f_{MF} correlates to the ADMET properties, such as solubility, permeability and Cytochrome P450 isoform 3A4 inhibition, as well.[20]

The previous study results have demonstrated that the molecular topology has some influence on the compound quality and drug-likeness of the compounds; however, the influence of molecular topology on ADMET properties and potency is still unknown. This project examined the correlation between molecular topology and ADMET experimental data as well as the three potency efficiency indices to better understand the earlier observation of 1TR compounds' enrichment in drugs.

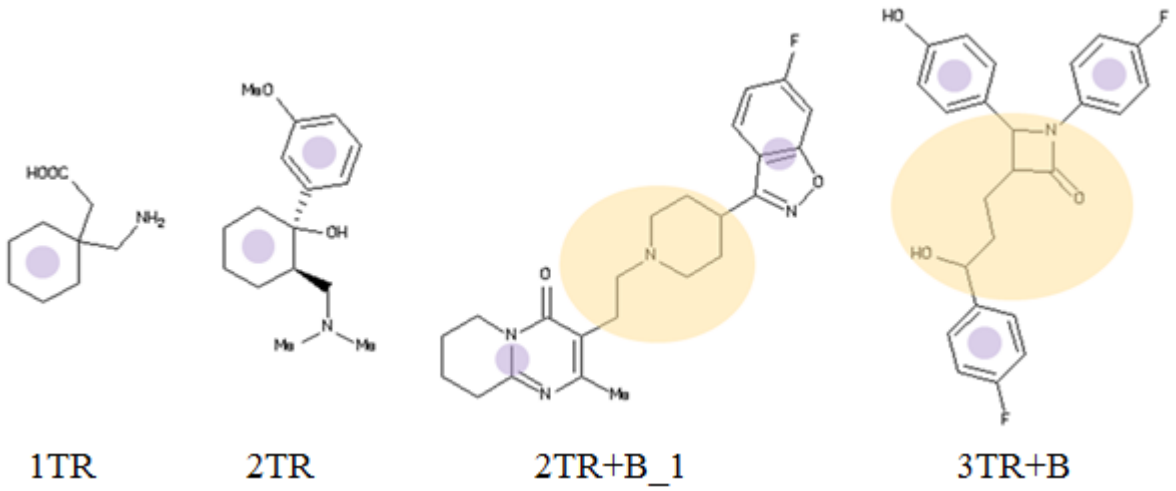
Methods

Data Sets:

For the Cytochrome P450 isoform 3A4 (CYP3A4) inhibition data analysis, a dataset of 15,887 compounds was used from the AstraZeneca in-house collection.

For the compound quality indices analysis, the data set was retrieved from GVKBIO, a commercial database where activity data has been extracted from medicinal chemistry journals and patents.[21] Only compounds associated with human targets were selected. Altogether, the data set includes in total 1,022,057 data points which correspond to bioactivity data of 695,322 unique structures for 972 human targets.

Definition of molecular topology classes:



1TR 2TR 2TR+B_1 3TR+B

Figure 1. Examples of different topological classes. (TR = Terminal Rings, B = Bridge). Purple is showing the terminal rings and orange is showing the molecular bridges in each compound.

The concept of molecular topology class[17] is based on an earlier molecular framework analysis.[18] A molecule can be divided into three subunits: terminal ring system (TR), molecular bridge (B), and side chain. A terminal ring system is defined as a cyclic structure which has only one connection to other ring structures in the molecule. The molecular bridge

is a linker that connects the terminal ring systems and it may include ring systems. Side chains refer to the parts of the molecule which do not belong to neither of the two other subunits. The topology class for any given compound is therefore depending on the number of terminal ring systems and the presence or absence of a molecular bridge. Some examples of different topological classes are shown in Figure 1. The 2TR+B class was subdivided according to the number of ring systems in molecular bridge to further enhance the analysis. In this study, the classes of 1TR, 2TR, 2TR+B and 3TR+B were considered. More complicated classes such as 4TR+B, 5TR+B were excluded due to their low occurrence in the dataset.

Property Calculations:

For all compounds, the ionization state was determined by substructure matching of a set of predefined acidic, basic and cationic functional groups with an in-house program. ClogP was calculated using the commercial available Biobyte ClogP program.[22] Molecular topology classes and the number of heavy atoms (*i.e.*, non-hydrogen atoms) were calculated with an in-house C++ program based on the OpenEye toolkit.[23] The ring system complexity (RSC) is calculated by subtracting number of ring systems from the number of smallest set of smallest rings (SSSR).[17,24] The descriptor f_{MF} is defined as:

$$f_{MF} = HEV_{MF} / HEV_{total} \qquad \text{(Equation 1)}$$

Where HEV_{MF} is the number of heavy atoms in the molecular framework and HEV_{total} is the number of heavy atoms in the whole molecule.

The calculation steps are also shown in Figure 2 for f_{MF}.

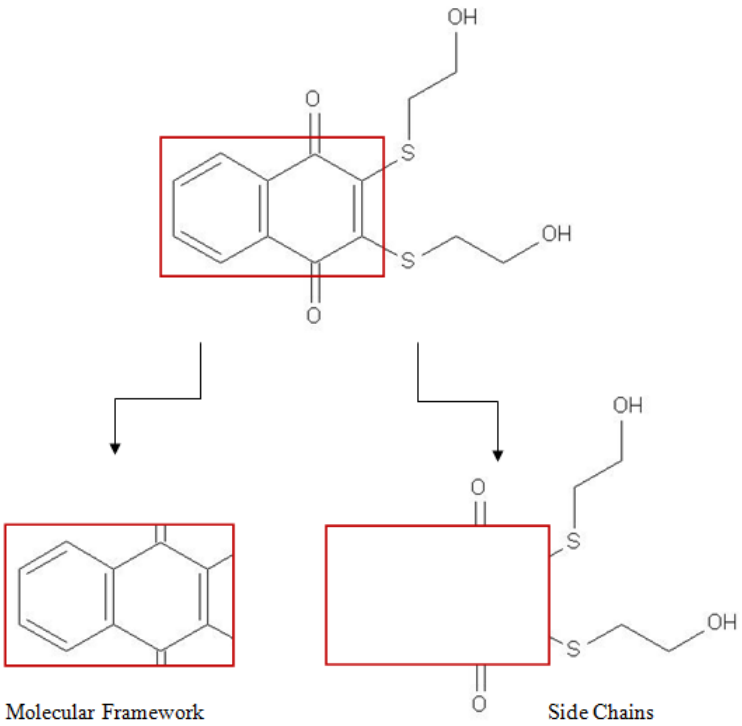


Figure 2. f_{MF} can be calculated with the following steps: identify the molecular frame work from the original molecule and count the number of heavy atom(HEV), and divide the number by the total heavy atom number. With this example in the figure, 10 HEV for molecular framework, and 20 for the whole molecule; thus f_{MF} = 10 / 20 = 0.5.

Ligand Efficiency (LE), Ligand Lipophilic Efficiency (LLE) and Ligand-Efficiency-dependent LipoPhilicity (LELP) are defined and calculated using Equation 2, 3 and 4 accordingly.

$$LE = \Delta G / HEV \qquad \text{(Equation 2)}$$

Where ΔG = -RT ln IC50, and HEV is the number of heavy atoms.

$$LLE = pIC50 - ClogP \qquad \text{(Equation 3)}$$

$$LELP = ClogP / LE \qquad \text{(Equation 4)}$$

All statistical analysis of Wicoxon rank-sum test[25] was performed with the open source statistic package, Python-statlib.[26]

Results and Discussion

CYP3A4 inhibition analysis:

Cytochrome P450 is a family of enzymes which is involved in drug metabolism and other reactions in cholesterols, steroids, and lipids synthesis. CYP3A4 is one of the most abundant isoforms[27] and it metabolizes approximately half of the drugs used today.[28] Multidrug administration can cause competing drug-metabolism (drug-drug interaction) which can lead to a high drug concentration with a higher probability of toxic response.[29] It is therefore important to monitor and screen for CYP3A4 inhibition. Previously, it has been shown high lipophilicity, large molecular size and large f_{MF} correlates with CYP3A4 inhibition.[20]

This study investigated how the CYP3A4 inhibition correlates with the molecular topology classes. 1TR compounds showed lower CYP3A4 inhibition compared to more complicated compounds, such as 2TR+B and 3TR+B. Further analysis showed the correlation of molecular size and f_{MF} with the molecular topology class, and the cross-correlation of the descriptors. From these observations, it was concluded that 1TR compounds' lower CYP3A4 inhibition is a combinatorial effect of small molecular size and low f_{MF}.

For further details of results and analyses, see the manuscript "An investigation of the relationship between molecular topology and CYP3A4 inhibition for drug-like compounds" which is appended. The manuscript is submitted for publication.

Compound quality indices analysis:

Even though CYP3A4 inhibition is an important property in the drug development process, CYP3A4 inhibition plays probably only a minor role in explaining why 1TR compounds are abundant among drugs compared to other compound collections. For a better understanding

of the higher fraction of 1TR compounds among drugs, the potency distribution over molecular topology classes and the correlation between three compound quality indices, LE, LLE and LELP, and molecular topology classes was investigated.

In this investigation, it was observed that the reported potency data in the literature is independent from both lipophilicity and molecular size. It is shown that 1TR compounds have better values of LE, LLE and LELP compared to the other topological classes; this conclusion is still valid after correcting for any potential bias due to that the compounds from different topological classes might be reported to be active on different targets. Two different factors contribute to this novel result. 1TR compounds are much more frequent among the reported compounds with small size and low lipophilicity. Thus when a potent compound is identified with low lipophilicity and small size in a drug discovery project, it is often compounds from the 1TR topology class. The second factor is that at constant size 1TR compounds have higher potency than other compounds. Thus it is easier to reach a desired potency with a compound belonging to the 1TR topology class that has small size and lipohilicity. The reduced size and lipophilicity will then contribute to a lower clinical attrition rate and accordingly to an enrichment of compounds with a 1TR topology among drugs.

For further details of results and analyses, see the manuscript "The influence of molecular topology on compound quality indices like ligand efficiency and ligand lipophilic efficiency" which is appended. The manuscript will be submitted for publication soon.

Acknowledgement

I would like to thank Dr. Ola Engkvist and Dr. Hongming Chen for their great help in this project.

Reference:

1. Innovation or Stagnation: Challenge and Opportunity on the Critical Path to New Medical Products. *U.S. Food and Drug Administration* [online]. [Published March 16, 2004; Updated July 20, 2010; Cited March 8, 2012] Available from: http://www.fda.gov/ScienceResearch/SpecialTopics/CriticalPathInitiative/CriticalPathOpportunitiesReports/ucm077262.htm
2. Lipinski, C. et al. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, **2001**, 46, 3-26.
3. Hann, M. Molecular obesity, potency and other addictions in drug discovery. *Med. Chem. Commun.*, **2011**, 2, 349-355.
4. Walter, P. Going further than Lipinski's rule in drug design. *Expert Opin. Drug Discov*, **2012**, 7, 99-107.
5. Leeson, P. and Springthorpe B. The Influence of drug-like concepts on decision-making in medicinal chemistry. *Nat. Rev. Drug. Discov.*, **2007**, 6, 881-890.
6. Gleeson, MP. Generation of a set of simple, interpretable ADMET rules of thumb. *J. Med. Chem.*, **2008**, 51:4, 817-834.
7. 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.
8. Raynolds, C., Tounge, B., and Bembenek, S. Ligand Binding Efficiency: Trends, Physical Basis, and Implications. *J. Med. Chem.*, **2008**, 51, 2432-2438.
9. Bembenek, S., Tounge, B., and Reynolds, C. Ligand efficiency and fragment-based drug discovery. *Drug Discov. Today*, **2009**, 14:5/6, 278-283.
10. Gleeson, P.: Hersey, A.; Montanari, D.; Overington, J. Probing the links between in vitro potency, ADMET and physicochemical parameters. *Nat. Rev. Drug. Discov.* **2011**, 10, 197-208.
11. Carr, R., Congreve, M., Murray C., and Rees D. Fragment-based lead discovery: leads by design. *Drug Discov. Today*, **2005**, 10:14, 987-992.
12. Mortenson, P., and Murray C., Assessing the lipophilicity of fragments and early hits. *J Comput. Aided Mol. Des.*, **2011**, 25, 663-667.
13. Nissink, J. Simple Size-Independent Measure of Ligand Efficiency. *J. Chem. Inf. Model*., **2009**, 49, 1617-1622.

14. Tarcsay, Á., Nyíri, K., and Keserü G. Impact of Lipophilic Efficiency on Compound Quality. *J. Med. Chem.*, **2012**, 55:3, 1252-1260.
15. Ritchie, T. and Macdonald, S. The impact of aromatic ring count on compound developability – are too many aromatic rings a liability in drug design? *Drug Discov. Today*, **2009**, 14:21/22, 1011-1020.
16. Lovering, F., Bikker, J. and Humblet C. Escape from Flatland: Increasing Saturation as an Approach to Improving Clinical Success. *J. Med. Chem.*, **2009**, 52, 6752-6756.
17. Chen, H., Yang, Y., and Engkvist, O. Molecular Topology Analysis of the Differences between Drugs, Clinical Candidate Compounds, and Bioactive Molecules. *J. Chem. Inf. Model.*, **2010**, 50, 2141–2150.
18. Bemis, G., and Murcko M. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.*, **1996**, 39, 2887-2893.
19. Chen, H., Engkvist, O., Blomberg, N., and Li, J. A comparative analysis of the molecular topologies for drugs, clinical candidates, natural products, human metabolites and general bioactive compounds. *Med. Chem. Comm.*, **2012**, 3, 312-321.
20. Yang, Y., Engkvist, O., Llinàs A., and Chen H. Beyond Size, Ionixation State and Lipophilicity: Influence of Molecular Topology on ADMET Properties for drug-like compounds. *J. Med. Chem*. in press, DOI: 10.1021/jm201548z
21. *GVKBio databases*; GVK Biosciences Private Ltd.: Hyderabad, India.
22. *BioByte ClogP* version 4.3, http://www.biobyte.com (accessed Sep. 6, 2011)
23. *OEChem toolkit* 1.7.2; OpenEye Scientific Software, Inc., Santa Fe, NM.
24. Sorkau, E. Ringerkennung in chemischen Strukturen mit dem Computer. *Wiss. 2. Tech. Hochsch. Leuna-Meuseburg*. **1985**, 27, 765-770.
25. F. Wilcoxon, *Biometrics Bulletin,* **1945,** 1, 80-83.
26. *Python-statlib* version 1.1, http://python-statlib.googlecode.com, **2007**.
27. Gleeson, MP. Generation of a set of simple, interpretable ADMET rules of thumb. *J. Med. Chem.*, **2008**, 51:4, 817-834.
28. Guengerich, F. P. CYTOCHROME P-450 3A4: Regulation and Role in Drug Metabolism. *Annu. Rev. Pharmacol. Toxicol.* **1999,** 39, 1-17
29. Ogu, C., and Maxa, J. Drug interactions due to cytochrome P450. *Proc (Bayl Univ Med Cent)*, **2000**, 13, 421-423.

An investigation of the relationship between molecular topology and CYP3A4 inhibition for drug-like compounds

Rurika Oka[*€], Ola Engkvist[€] ,Hongming Chen[€Δ]

* Department of Biology, Lund University, SE-22100 Lund, Sweden

€ Computational Sciences, Discovery Sciences, AstraZeneca R&D Mölndal, SE-43183 Mölndal, Sweden

Δ Corresponding author

Abstract

Earlier studies have shown that CYP3A4 inhibition is influenced by lipophilicity, size and f_{MF}, where f_{MF} is the size of the molecular framework divided by the total size of the molecule. Other previous study has shown that there is an enrichment of compounds with only one ring system among drugs in comparison to clinical candidates and general bioactive compounds as well as enrichment among natural products and human metabolites compared to general bioactive compounds. Thus a logical extension of the earlier investigations would be to study how ring systems influence CYP3A4 inhibition. It is shown here that compounds with only one ring system have lower CYP3A4 inhibition compared to compounds with several ring systems. This is due to that they are both smaller and have smaller f_{MF} compared to compounds with several ring systems. The investigation provides additional insights how a molecule should be constituted to have low CYP3A4 inhibition and may influence library design and compound acquisitions.

Keywords: CYP3A4 inhibition; Molecular topology class; ADMET; f_{MF}; Fsp3; Drug-likeness

The pharmaceutical industry faces a high attrition rate for compounds entering clinical development. Only a small fraction of them passes all the development hurdles to become approved drugs.[1] In order to reduce the number of costly late-stage clinical failures, it is critical to filter out compounds which have ADMET (absorption, distribution, metabolism, excretion, and toxicity) liabilities as early as possible. For instance, Lipinski proposed the well known "rule of five" as a useful filter to assess the likelihood of oral bioavailability.[2,3] The lipophilicity and the molecular size are widely regarded as important factors influencing the drug-likeness[4-7]. High lipophilicity increases the promiscuity in addition to other liabilities such as low solubility, high plasma protein binding and high hERG inhibition.[4] Recently a comparison of several simple physicochemical properties for marketed drugs, clinical candidates and bioactive compounds was performed, where any potential bias due to target class differences was removed. The study confirmed that the lipophilicity is lower and the molecular size is smaller for drugs than for clinical candidates and general bioactive compounds.[5]

Further investigations have been carried out to identify additional factors affecting the compound quality. For example, the fraction of sp3 carbon atoms (Fsp3) and the presence of chiral centers are correlated with clinical success.[8] Also a large number of aromatic rings can negatively impact several drug-like properties.[9] It has also been shown that the molecular topology influences the promiscuity.[10] In another recent study it was investigated how ADMET is influenced by molecular size, lipophilicity, ionization state, Fsp3 and f_{MF}.[11] f_{MF} is defined as the size of the molecular framework divided the total size of the molecule.[10] It was shown that high lipophilicity, large size and large f_{MF} correlate with Cytochrome P450 isoform 3A4 (CYP3A4) inhibition. However, Fsp3 did not show any significant correlation with CYP3A4 inhibition. While it is important to correlate simple descriptors to CYP3A4

inhibition, the descriptors give only a rudimentary understanding of how a molecule should be constituted to have a low CYP3A4 inhibition. The descriptors like size, lipophilicity and f_{MF} can be used in lead optimization to modify a molecule which leads to reduced CYP3A4 inhibition. However, these descriptors are rather abstract and it would be useful to more thoroughly understand what types of molecular topologies that have lower CYP3A4 inhibition. A more thorough understanding of the molecular topology can guide library designs and compound acquisitions to lower potential CYP3A4 liabilities in a compound collection. It is not always possible to do major changes to the molecular topology during lead generation and optimization so it is important that the starting molecule has a suitable molecular topology. It is therefore of interest to correlate CYP3A4 inhibition to different molecular topologies in order to enhance the understanding of the relationship between molecular structure and CYP3A4 inhibition.

An extension of the molecular framework analysis has recently been proposed. In the new method, a molecule is partitioned into only a few different classes based on the molecular topology.[12-13] A topology class analysis comparing drugs, clinical candidates and general bioactive compounds has shown that in comparison with clinical candidates and bioactive compounds, drugs have a higher fraction of compounds with only one ring system. [12] This conclusion is still valid after correcting for any bias originating from lipophilicity (ClogP), molecular size and target class.[12] The main advantage of the topology class analysis is because it classifies a molecule into only a few different classes, while in comparison earlier molecular framework analysis created many more different topologies.[14] Thus it is possible to apply rigorous statistical methods to the obtained results, while still providing important information of the relationship between the molecular topology and CYP3A4 inhibition.

According to the definition of topology class, a molecule is partitioned into three different subunits: terminal ring systems (TR), a molecular bridge (B), and side chains. A terminal ring system is defined as a cyclic structure which has only one connection to other ring structures. The molecular bridge is the linker which connects the terminal ring systems. The molecular bridge might also include additional ring systems. Side chains refer to the parts that do not belong to the other two parts. The molecular framework consists of the terminal ring systems and the molecular bridge. The topology class for a given compound depends on the number of terminal ring systems and the presence (or absence) of a molecular bridge. Examples of different topology classes are shown in Figure 1. The 2TR+B class can be further subdivided according to the number of ring systems in the molecular bridge. For example, the 2TR+B_1 class compound consists of two terminal rings and a molecular bridge and there is additionally one ring system in the molecular bridge. 1TR, 2TR, 2TR+B and 3TR+B classes were used in the analysis. More complicated classes such as 4TR+B and 5TR+B were excluded due to their low occurrence in the used dataset.

Earlier investigations has shown that 1TR compounds are more common among drugs in comparison to clinical candidates and general bioactive compounds, this result was still valid after correcting for size, lipophilicity and any bias due to differences in activities on different targets.[12] Additionally it has been shown that 1TR compounds is also common among natural products and human metabolites.[13] Thus it seems that 1TR is a preferred molecular topology in nature. It is of interest to broaden the investigation of molecular topology into ADME properties to further enhance the knowledge between molecular topology and drug-likeness. Further studies can give additional understanding if compounds with 1TR are suitable starting points for drug discovery projects, which is suggested by the earlier results.

CYP3A4 is a gene encoding a member of the Cytochrome P450 superfamily of enzymes. The CYP3A4 enzyme catalyzes many reactions involved in drug metabolism and synthesis of cholesterol, steroids and other lipids and is the most abundant Cytochrome P450 isoform[14] and is involved in the metabolism of approximately half of the drugs in use today.[15] Multidrug administration can cause competing metabolism of drugs (so called drug-drug interaction), which might lead to an increased blood concentration of the drugs with an accompanying higher probability of toxic side effects.[16] It is therefore important to investigate CYP3A4 inhibition early in a drug discovery project. A dataset of 15,887 in-house compounds was used in the current study. The relationships between size, lipophilicity and f_{MF} and CYP3A4 inhibition for this dataset have already been reported.[11] However, since the format of the reported figures is different than the format used in this study, the results are shown in the format used here in Figure S1.

For all compounds in the data set, the ionization state was determined by substructure matching of a set of predefined acidic, basic and cationic functional groups. ClogP was calculated with a commercial program.[17] Molecular topology class and the number of heavy atoms were determined with an in-house C++ program based on the OpenEye toolkit.[18] All statistical analysis was done with the open source statistics package Python-statlib.[19]

In the current study, the relationship between the topology class and CYP3A4 inhibition has been investigated. The results show that the molecular topology class has a distinct influence on CYP3A4 inhibition. The CYP3A4 inhibition for each topology class is shown in Figure 2. 1TR compounds have on average lower inhibition than compounds from the other topological classes. The Wilcoxon rank-sum test[20] shows that the differences in CYP3A4 inhibition between the compounds in the 1TR class and the compounds from the other topological classes are statistically significant (Table S1). Molecules with two ring systems (2TR and 2TR+B_0) have also on average lower inhibition than molecules with three ring

systems. To exclude any potential bias in the results originating from differences in ionization state, the whole dataset was partitioned into four groups according to the ionization state (acidic, basic, neutral and zwitterionic). The most common ionic species are basic and neutral compounds (Figure 3). For both basic and neutral compounds, 1TR compounds have lower CYP3A4 inhibition than compounds from the other topology classes. Thus any potential bias due to the ionization state for the obtained results in this study is excluded. Again, the Wilcoxon rank-sum test shows that the differences in CYP3A4 inhibition between compounds in the 1TR class and the compounds from the other topological classes are statistically significant (Table S2).

As shown in an earlier study, [11] the CYP3A4 inhibition is influenced by ClogP, size and f_{MF}. CYP3A4 inhibition increases with increasing ClogP, size and f_{MF}. The relationship between the three descriptors and the topology classes were therefore thoroughly investigated. As is seen in Figure 4, there are only small differences in lipophilicity between the different topological classes. However, there is a significant difference in size; 1TR compounds are on average significantly smaller than compounds from the other topological classes. Compounds with two ring systems are also on average smaller than compounds with three ring systems. Thus there is a correlation between the number of ring systems and the molecular size. Compounds in the 1TR class have on average smaller f_{MF} than compounds from the other topological classes, *i.e*. compounds in the 1TR class have larger and/or more side chains than compounds from the other topological classes. Again compounds with two ring systems are in between compounds with one and three ring systems. As it has been earlier observed that an increase in molecular size and f_{MF} also increases the CYP3A4 inhibition,[11] the small size and small f_{MF} for 1TR compounds might explain their low CYP3A4 inhibition. The statistical significance tests are shown in Table S3.

It is of interest to deepen the understanding of how the three descriptors analyzed influence CYP3A4 inhibition and to investigate if there are some cross correlations between them that give rise to the low CYP3A4 inhibition for compounds in the 1TR class. First it was investigated if the observed differences in CYP3A4 inhibition between the topological classes are independent of the three descriptors individually. If the differences are independent, we need to investigate if there is a cross correlation between the descriptors that explains the observed relationships. As is shown in Figures S2-S4, the differences in CYP3A4 inhibition between the topology classes are mainly independent of the ClogP, size and f_{MF}, respectively. However, there is an increased level of noise in the data for bins populated with only a small number of compounds. There is a strong correlation between size and f_{MF} for a topological class (Figures S5-S6). For a fixed molecular size, f_{MF} is much smaller for compounds with fewer ring systems. Since compounds in the 1TR topological class have only one ring system, they have on average also the smallest f_{MF}. Thus the most likely reason that compounds in the 1TR topological class have low CYP3A4 inhibition is because they are both smaller and have a lower f_{MF} than compounds from the other topological classes. In conclusion it has been shown that the size is correlated with the number of ring systems and accordingly 1TR compounds are smaller than 2TR+B_0 compounds, which are smaller than 2TR + B_1 and 3TR+B compounds. Additionally, for a constant molecular size, 1TR compounds have a lower f_{MF} i.e. they are likely to have larger part of the molecule that are not part of the molecular framework in comparison with compounds from the other topology classes,

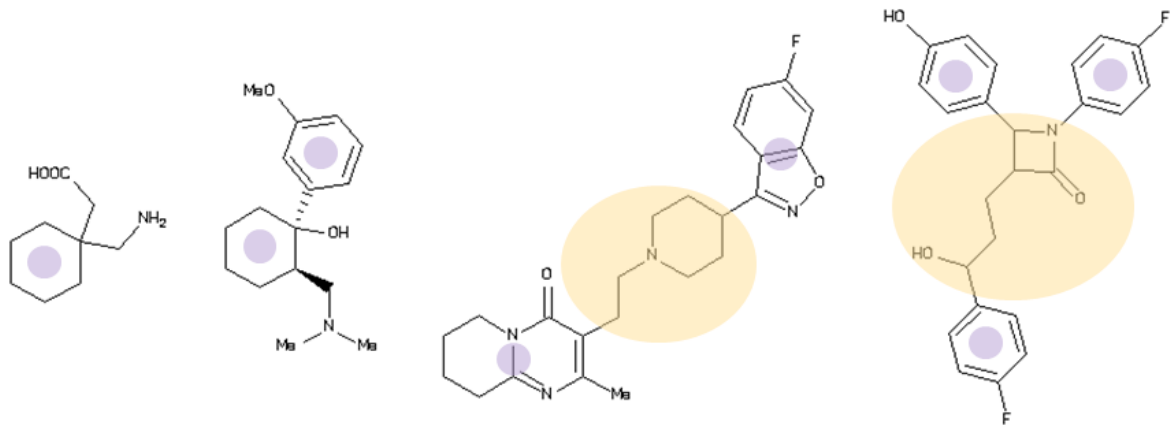
It is important for future drug discovery projects to further enhance the understanding of how a molecular should be constituted to have better ADME properties. In this study we have therefore investigated how the molecular topology affects CYP3A4 inhibition. It is essential in this type of analysis to balance the need to better understand the relationship between CYP3A4 inhibition and molecular topology with the need to be able to perform rigorous

statistical analysis. We have therefore decided to divide our dataset into just a few topological classes which allows a statistically valid comparison of CYP3A4 inhibition between different topology classes. In conclusion it was found that compounds in the 1TR class have lower CYP3A4 inhibition than compounds from the other topological classes. This is due to a combination of smaller size and lower f_{MF} for compounds in the 1TR topology class. The same explanation can be used to rationalize why compounds with two ring systems have lower CYP3A4 inhibition than compounds with three ring systems. Further investigations will be needed to understand the relationship between the molecular topology class and other properties related to drug-likeness. In conclusion, descriptors found earlier to influence CYP3A4 inhibition has here been interpreted from a topology perspective, it was found that molecules with fewer ring systems, preferably only one, is most common among compounds with low CYP3A4 inhibition. As has been reported earlier 1TR compounds are over-represented among drugs and human metabolites, the found results here gives additional support in prioritizing 1TR type compounds in library design and compound acquisitions to enrich a compound collection.

Supporting Information available: Figures S1-S6 and Tables S1-S3 are appended as Supporting Material.

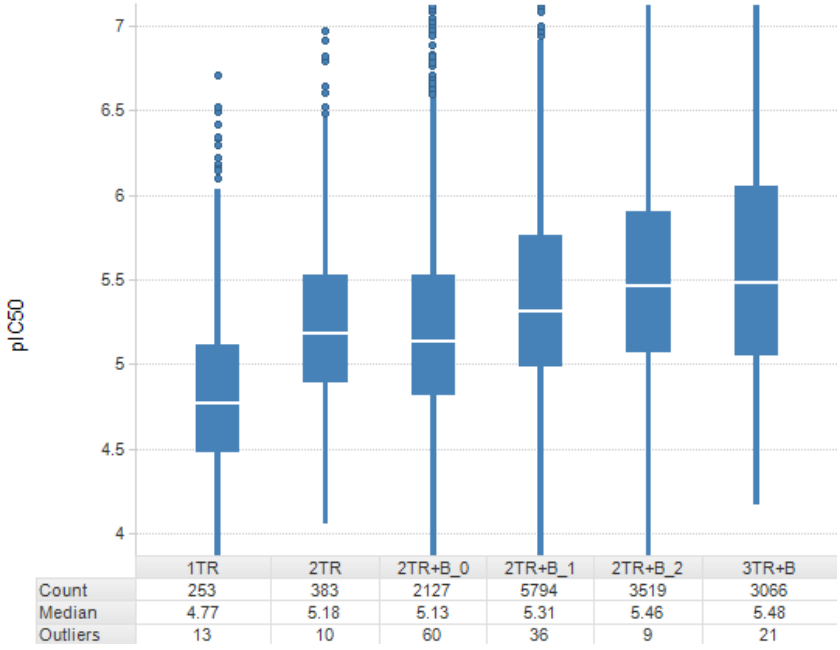
References

1. I. Kola, J. Landis, *Nat. Rev. Drug. Discov.* **2004**, 3, 711-715.
2. M. Hann, *MedChemCommun.* **2011**, 2, 349-355.
3. C. A. Lipinski, F. Lombardo, B. Dominy, P. J. Feeney, *Adv. Drug Delivery Rev.* **2001**, 46, 3-26.
4. P. D. Leeson, B. Springthorpe, *Nat. Rev. Drug. Discov.* **2007**, 6, 881-890.
5. C. Tyrchan, N. Blomberg, O. Engkvist, T. Kogej, S. Muresan, *Bioorg. Med. Chem. Lett*. **2009**, 19, 6943-6947.
6. M. C. Wenlock, R. P. Austin, P. Barton, A. M. Davis, P. D. Leeson, *J. Med. Chem.* **2003**, **46**, 1250-1256.
7. M. Vieth, J. J. Sutherland, *J. Med. Chem.* **2006**, **49**, 3451-3453.
8. F. Lovering, J. Bikker, C. Humblet, *J. Med. Chem.* **2009**, 52, 6752-6756.
9. T. Ritchie, S. J. Macdonald, *Drug Discov. Today* **2009**, 14, 1011-1020.
10. Y. Yang, H. Chen, I. Nilsson, S. Muresan, O. Engkvist, *J. Med. Chem*. **2010**, 53, 7709-7714.
11. Y. Yang, O. Engkvist, A. Llinàs, H. Chen, *J. Med. Chem*. **2012**, 55, 3667-3677.
12. H. Chen, Y. Yang, O. Engkvist, *J. Chem. Inf. Model.* **2010**, 50, 2141-2150.
13. H. Chen, O. Engkvist, N. Blomberg, J. Li, *MedChemCommun*. **2012**, 3, 312-321.
14. G. Bemis, M. Murcko, *J. Med. Chem.* **1996**, 39, 2887-2893.
15. M. P. Gleeson, *J. Med. Chem.* **2008**, 51, 817-834.
16. F. P. Guengerich, *Annu. Rev. Pharmacol. Toxicol.* **1999,** 39, 1-17.
17. *BioByte ClogP* version 4.3, http://www.biobyte.com, (accessed March 23, 2012)
18. *OEChem toolkit* 1.7.2, OpenEye Scientific Software, Inc., Santa Fe, NM.
19. *Python-statlib* version 1.1, http://code.google.com/p/python-statlib/, (accessed March 23, 2012)
20. F. Wilcoxon, *Biometrics Bulletin* **1945,** 1, 80-83.



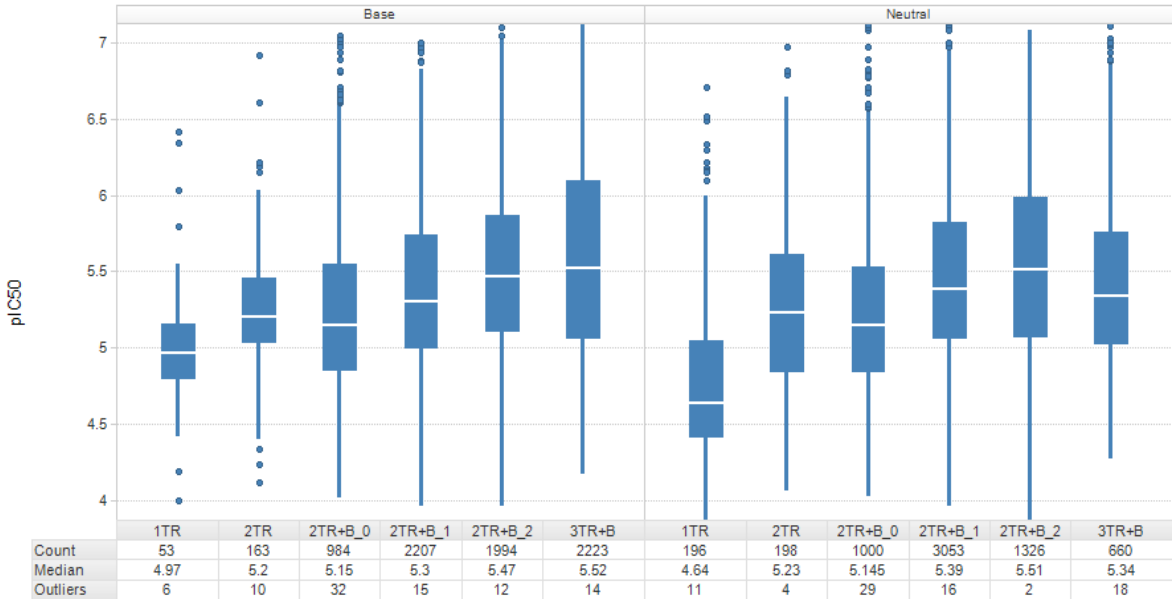
1TR 2TR 2TR+B_1 3TR

Figure 1. Examples of different topological classes. (TR = Terminal Rings, B = Bridge). The purple color is emphasizing the terminal rings and the orange color is emphasizing the molecular bridge for each compound.

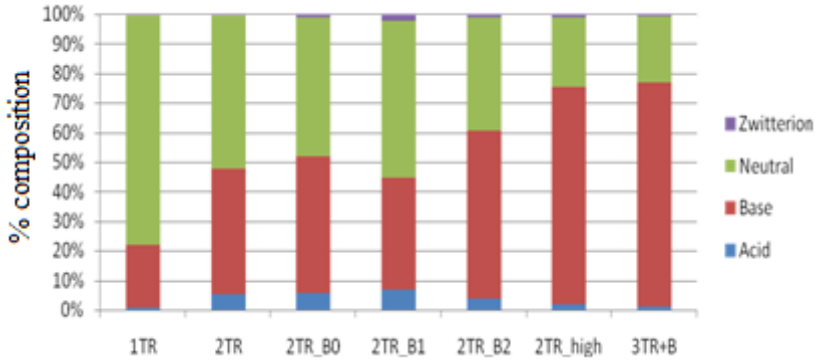


	1TR	2TR	2TR+B_0	2TR+B_1	2TR+B_2	3TR+B
Count	253	383	2127	5794	3519	3066
Median	4.77	5.18	5.13	5.31	5.46	5.48
Outliers	13	10	60	36	9	21

Figure 2. The relationship between CYP3A4 inhibition and the different topological classes.

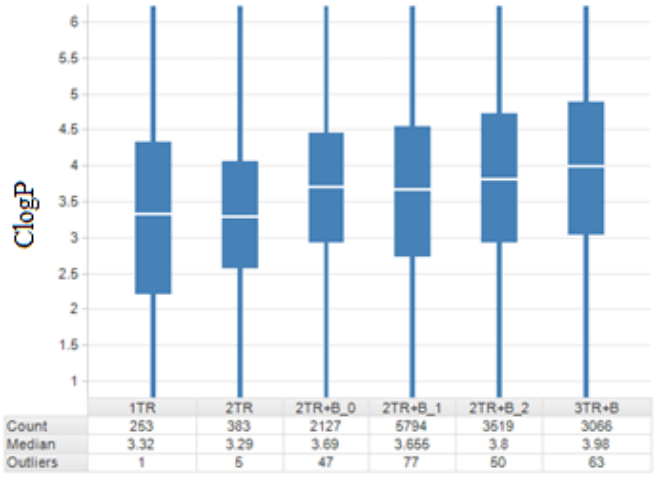


(a)

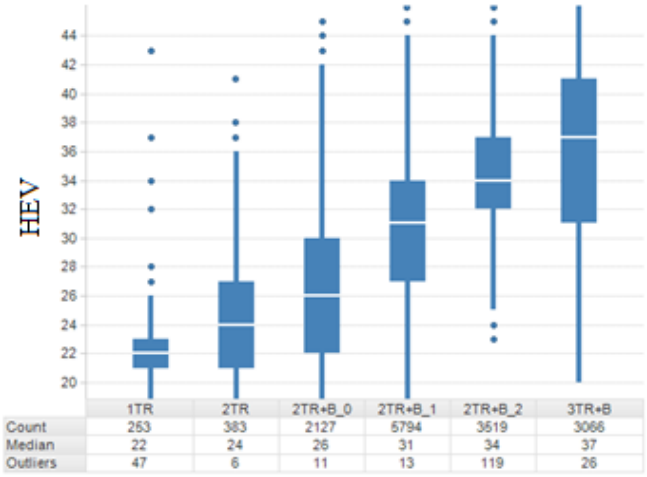


(b)

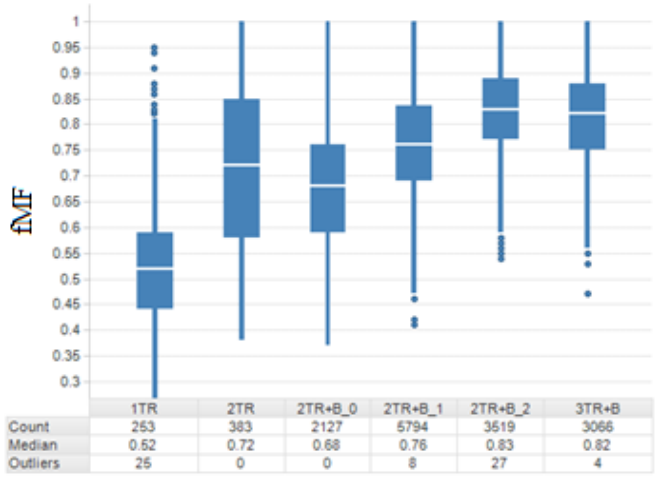
Figure 3. (a) Relationship between the topological classes and CYP3A4 inhibition for basic and neutral compounds, and (b) the distribution of different ionization states for each topological class.



(a) ClogP

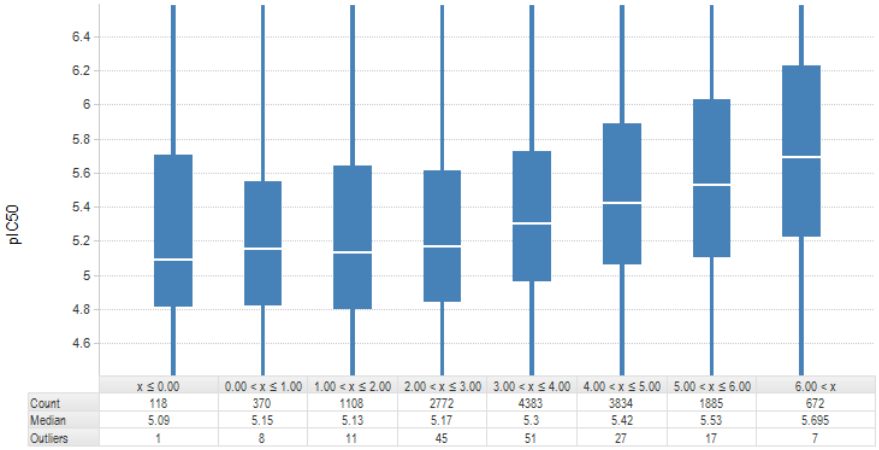


(b) HEV



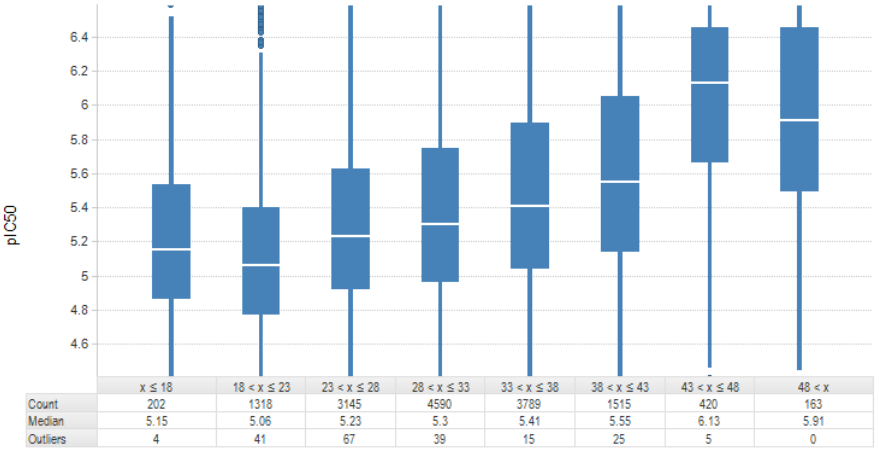
(c) f_{MF}

Figure 4. The relationship between the molecular topology classes and (a) ClogP, (b) Heavy atom count (HEV) and (c) f_{MF}.



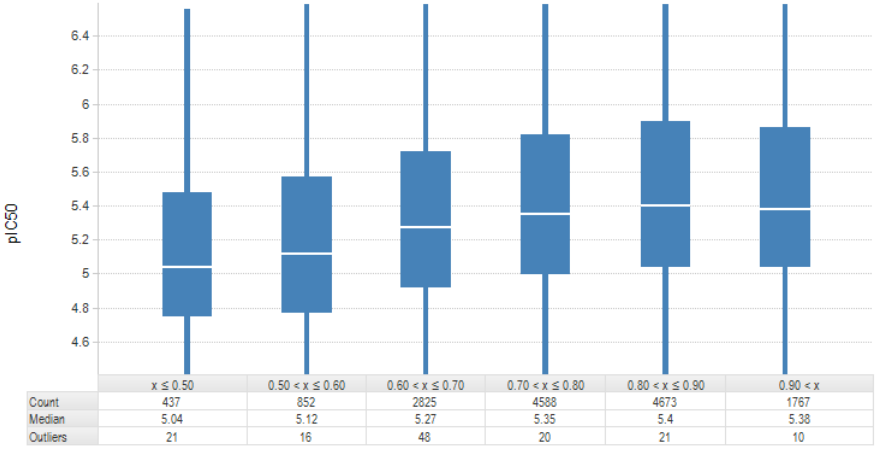
	x ≤ 0.00	0.00 < x ≤ 1.00	1.00 < x ≤ 2.00	2.00 < x ≤ 3.00	3.00 < x ≤ 4.00	4.00 < x ≤ 5.00	5.00 < x ≤ 6.00	6.00 < x
Count	118	370	1108	2772	4383	3834	1885	672
Median	5.09	5.15	5.13	5.17	5.3	5.42	5.53	5.695
Outliers	1	8	11	45	51	27	17	7

(a) ClogP



	x ≤ 18	18 < x ≤ 23	23 < x ≤ 28	28 < x ≤ 33	33 < x ≤ 38	38 < x ≤ 43	43 < x ≤ 48	48 < x
Count	202	1318	3145	4590	3789	1515	420	163
Median	5.15	5.06	5.23	5.3	5.41	5.55	6.13	5.91
Outliers	4	41	67	39	15	25	5	0

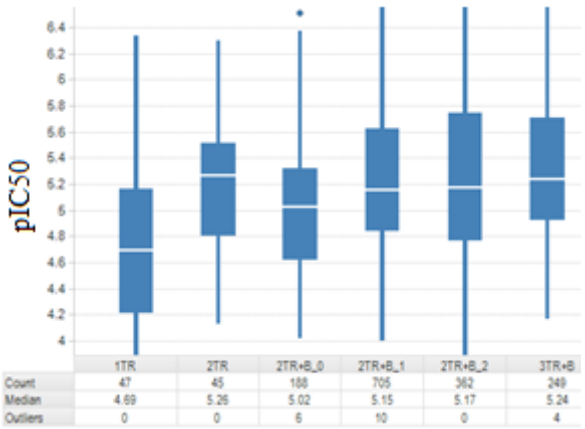
(b) HEV



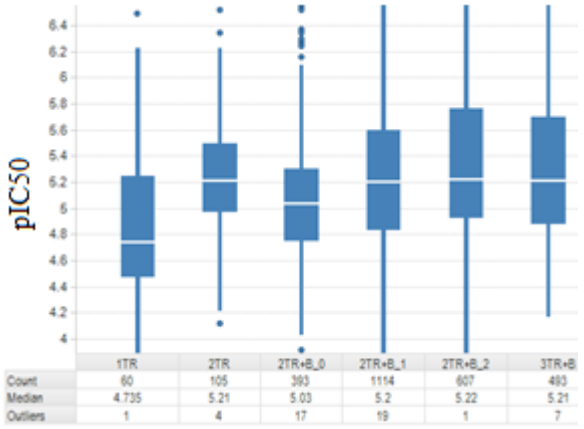
	x ≤ 0.50	0.50 < x ≤ 0.60	0.60 < x ≤ 0.70	0.70 < x ≤ 0.80	0.80 < x ≤ 0.90	0.90 < x
Count	437	852	2825	4588	4673	1767
Median	5.04	5.12	5.27	5.35	5.4	5.38
Outliers	21	16	48	20	21	10

(c) f_{MF}

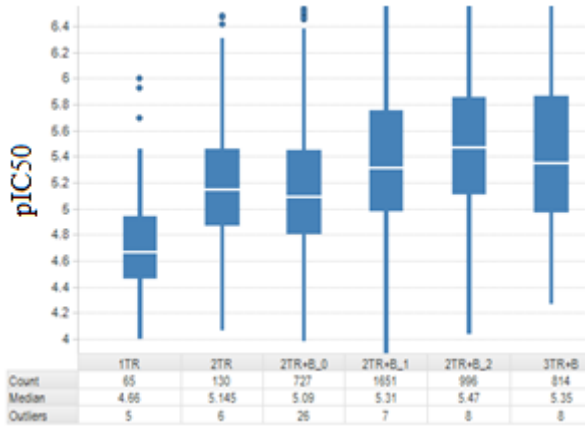
Figure S1. The relationship between CYP3A4 inhibition (pIC50) and (a) ClogP, (b) Heavy atom count (HEV) and (c) f_{MF}.



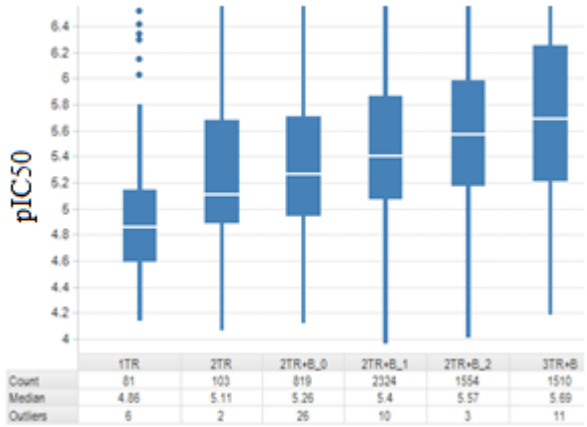
(a) ClogP ≤ 2.00



(b) 2.00 < ClogP ≤ 3.00

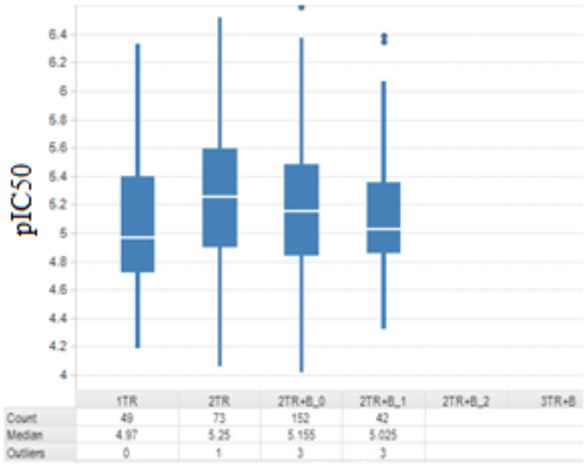


(c) 3.00 < ClogP ≤ 4.00

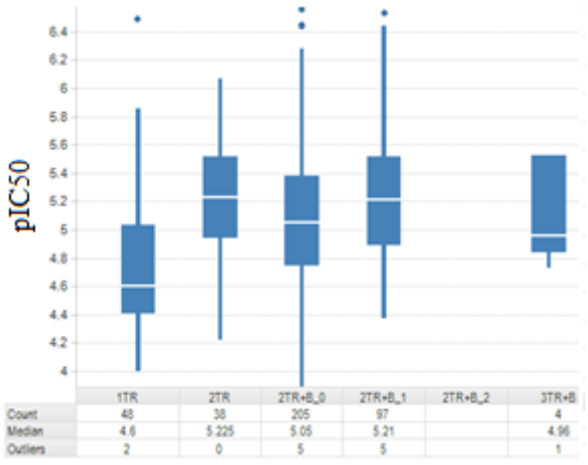


(d) ClogP > 4.00

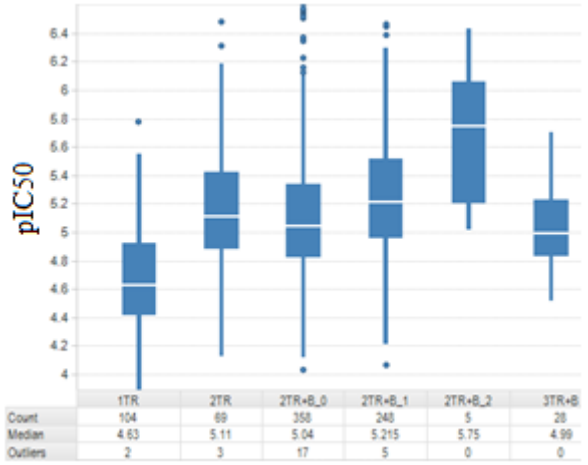
Figure S2. The relationship between the molecular topology classes and CYP3A4 inhibition (pIC50) for different ClogP intervals.



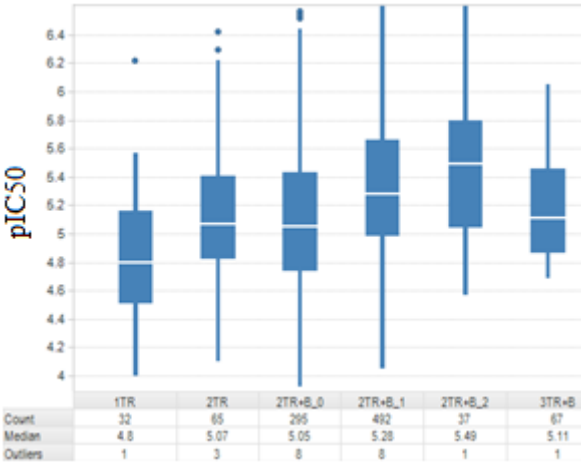
(a) HEV ≤ 19



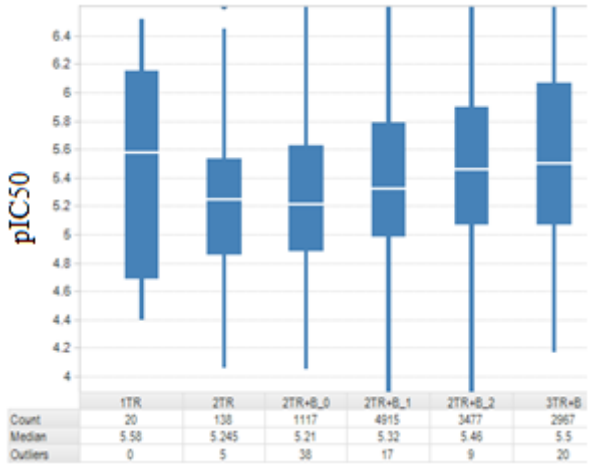
(b) 19 < HEV ≤ 21



(c) 21 < HEV ≤ 23

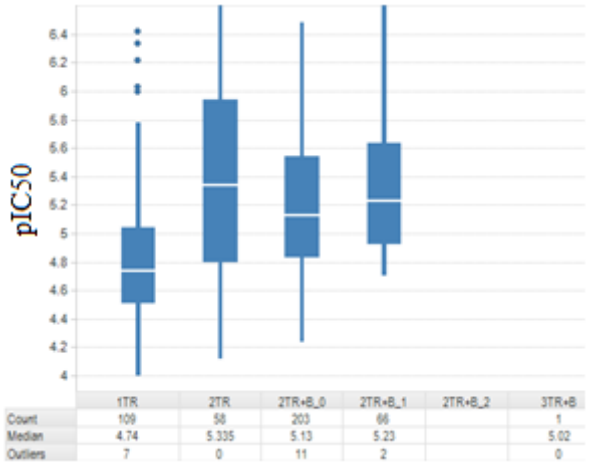


(d) 23 < HEV ≤ 25

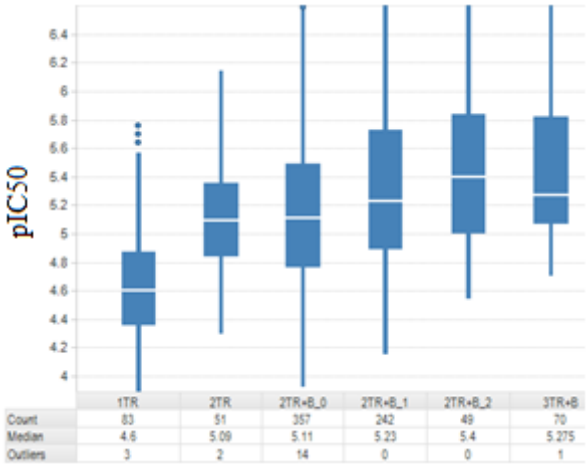


(e) HEV > 25

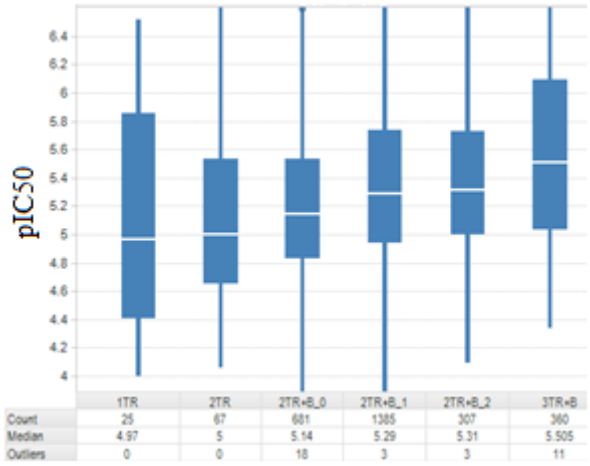
Figure S3. The relationship between the molecular topology classes and CYP3A4 inhibition (pIC50) for different Heavy atom count (HEV) intervals.



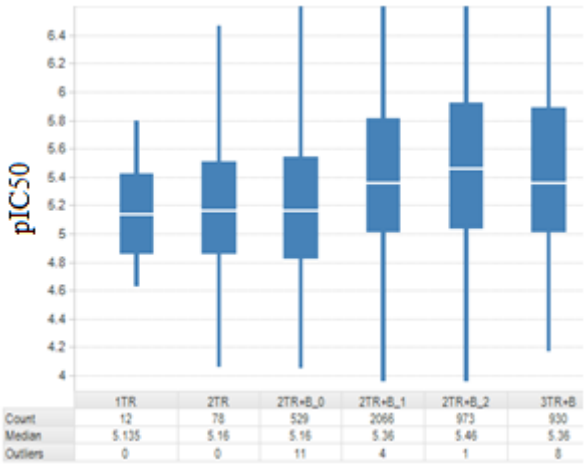
(a) $f_{MF} \leq 0.50$



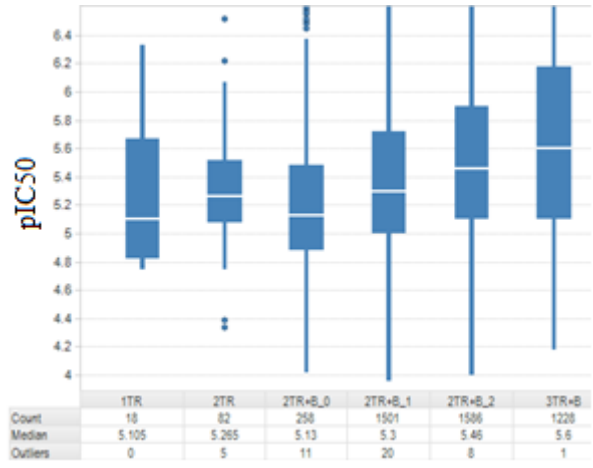
(b) $0.50 < f_{MF} \leq 0.60$



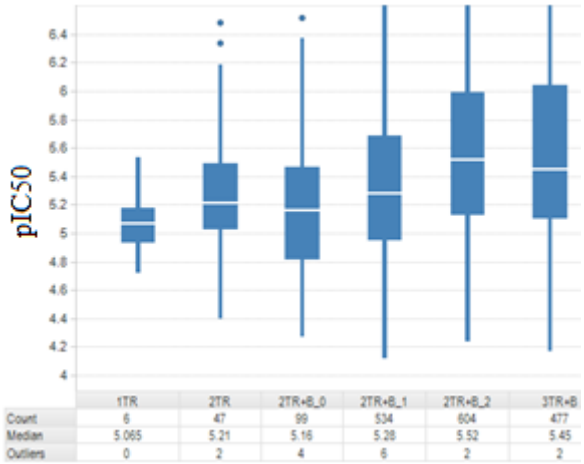
(c) $0.60 < f_{MF} \leq 0.70$



(d) $0.70 < f_{MF} \leq 0.80$

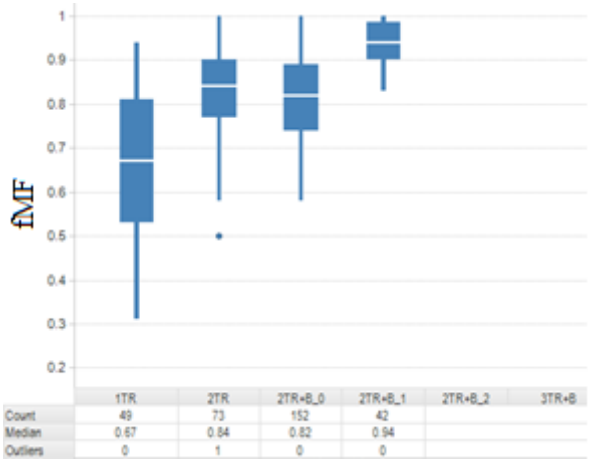


(e) $0.80 < f_{MF} \leq 0.90$

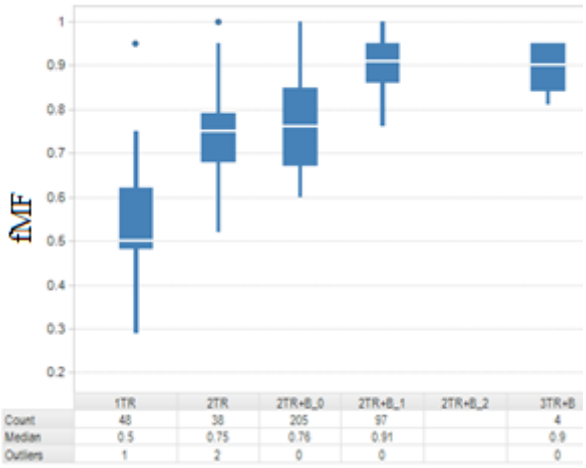


(f) $f_{MF} > 0.90$

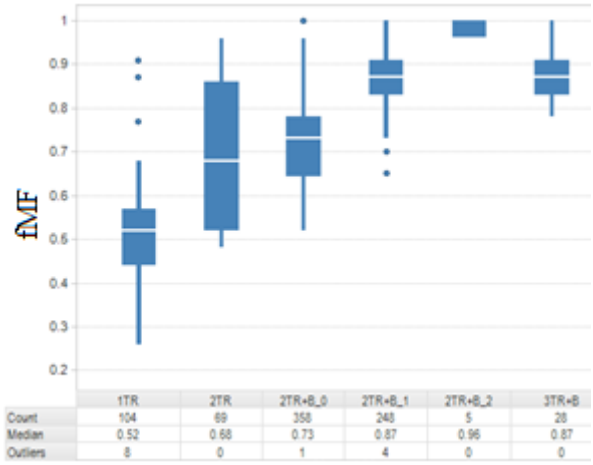
Figure S4. The relationship between the molecular topology classes and CYP3A4 inhibition (pIC50) for different f_{MF} intervals.



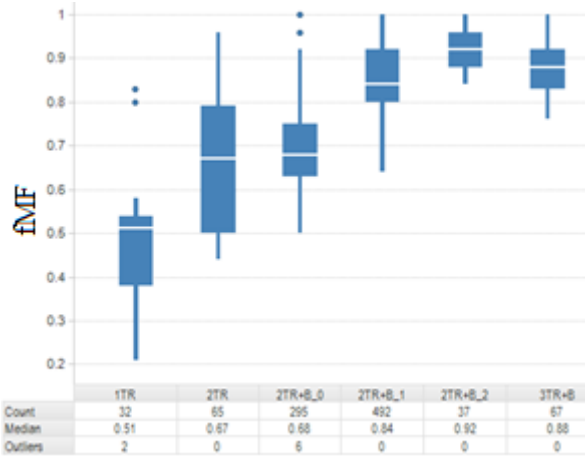
(a) HEV ≤ 19



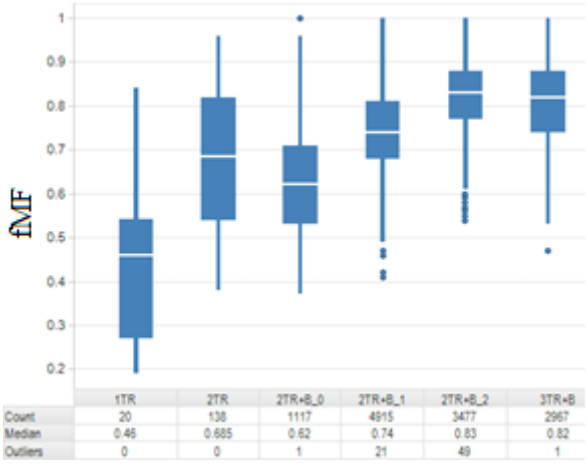
(b) 19 < HEV ≤ 21



(c) 21 < HEV ≤ 23

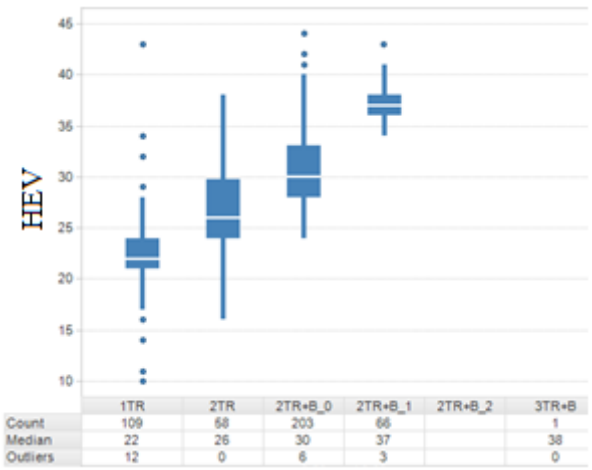


(d) 23 < HEV ≤ 25

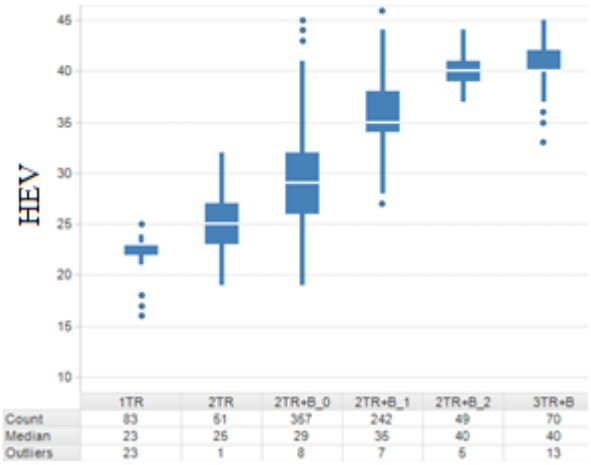


(e) HEV > 25

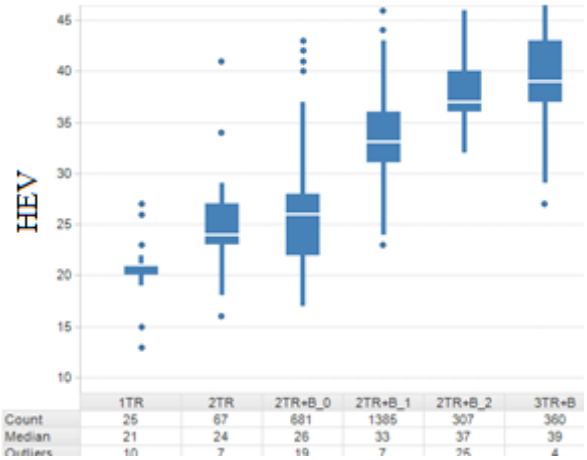
Figure S5. The relationship between the molecular topology classes and f_{MF} for different Heavy atom count (HEV) intervals.



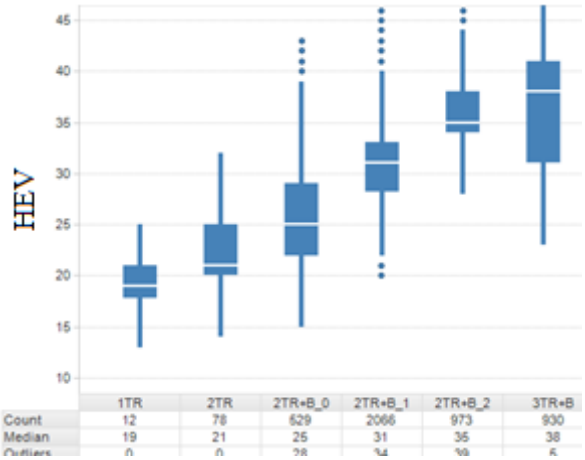
(a) $f_{MF} \leq 0.50$



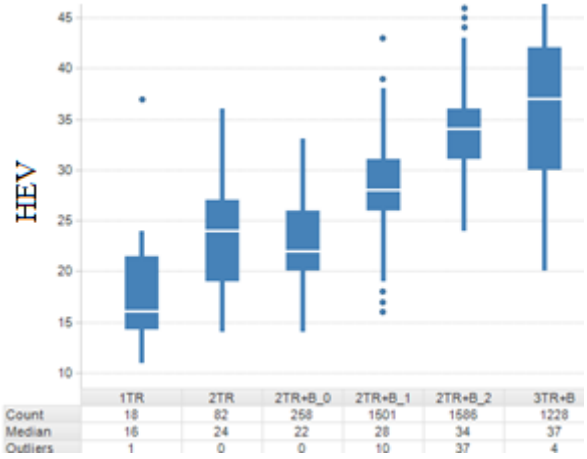
(b) $0.50 < f_{MF} \leq 0.60$



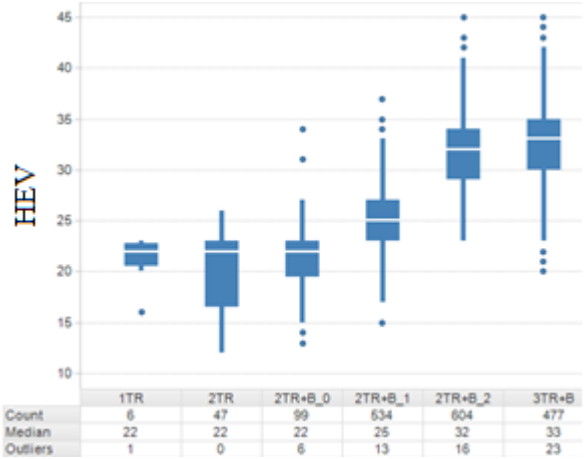
(c) $0.60 < f_{MF} \leq 0.70$



(d) $0.70 < f_{MF} \leq 0.80$



(e) $0.80 < f_{MF} \leq 0.90$



(f) $f_{MF} > 0.90$

Figure S6. The relationship between the molecular topology classes and Heavy atom count (HEV) for different f_{MF} intervals.

Table S1. p-values calculated with the Wilcoxon rank-sum test displaying the statistical significance of the differences in CYP3A4 inhibition between the different molecular topological classes.

	2TR	**2TR+B_0**	**2TR+B_1**	**2TR+B_2**	**3TR**
1TR	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
2TR	-	0.1863	<0.0001	<0.0001	<0.0001
2TR+B_0	-	-	<0.0001	<0.0001	<0.0001
2TR+B_1	-	-	-	<0.0001	<0.0001
2TR+B_2	-	-	-	-	0.0013

Table S2. p-values calculated with the Wilcoxon rank-sum test displaying the statistical significance of the differences in CYP3A4 inhibition between the different molecular topological classes for basic and neutral compounds.

Basic	**2TR**	**2TR+B_0**	**2TR+B_1**	**2TR+B_2**	**3TR**
1TR	<0.0001	0.0049	<0.0001	<0.0001	<0.0001
2TR	-	0.1127	0.0307	<0.0001	<0.0001
2TR+B_0	-	-	<0.0001	<0.0001	<0.0001
2TR+B_1	-	-	-	<0.0001	<0.0001
2TR+B_2	-	-	-	-	0.0010

Neutral	**2TR**	**2TR+B_0**	**2TR+B_1**	**2TR+B_2**	**3TR**
1TR	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
2TR	-	0.201	<0.0001	<0.0001	0.0016
2TR+B_0	-	-	<0.0001	<0.0001	<0.0001
2TR+B_1	-	-	-	<0.0001	0.1472
2TR+B_2	-	-	-	-	<0.0001

Table S3. p-values calculated with the Wilcoxon rank-sum test displaying the statistical significance of differences in ClogP between the different molecular topological classes. All differences between the topological classes for Heavy atom count and f_{MF} were statistically significant ($p<0.0001$).

ClogP	**2TR**	**2TR+B_0**	**2TR+B_1**	**2TR+B_2**	**3TR**
1TR	0.6726	<0.0001	0.0001	<0.0001	<0.0001
2TR	-	<0.0001	<0.0001	<0.0001	<0.0001
2TR+B_0	-	-	0.4048	0.0001	<0.0001
2TR+B_1	-	-	-	<0.0001	<0.0001
2TR+B_2	-	-	-	-	<0.0001

Investigation of the influence of molecular topology on ligand binding

Rurika Oka[*€], , Ola Engkvist[€], Hongming Chen[€Δ]

* Department of Biology, Lund University, SE-22100 Lund, Sweden

€ Discovery Sciences, Computational Sciences, AstraZeneca R&D Mölndal, SE-43183 Mölndal, Sweden

Δ Corresponding author

Abstract

The influence of molecular topology on physico-chemical properties and on the probability of success in clinical trials has been proposed in earlier studies. Compounds with only one ring system were more frequently observed among drugs in comparison to compounds with reported biological activity in the literature and to clinical candidate compounds. Another study has identified compounds with only one ring system are much more frequent among human metabolites compared to other sources of molecules. To better understand how the molecular topology is influencing the drug discovery process, we have investigated the frequency of different molecular topologies in published patents. Our study shows that patent compounds with low lipophilicity and small size have on average fewer ring systems. The influence of the molecular topology on ligand efficiency (LE), ligand lipophilic efficiency (LLE) and Ligand-efficiency-dependent lipophilicity (LELP) was also analyzed and it was found that on average compounds with fewer ring system and in particular compounds with only one ring system show consistently better LE, LLE and LELP. The results suggest that compounds with fewer ring system have better properties that might contribute to a higher

clinical success rate even after taken into account differences in target distribution between the different topological classes.

Introduction

Since the beginning of the genomics era in the 1990s, the drug discovery strategy has been largely transformed from low throughput *in vivo* based phenotypic screening cascade to a target centric high-throughput *in vitro* screening cascade. This paradigm shift has so far not been able to increase the productivity of pharmaceutical industry as a whole and recent data indicate that, only 10% of the compounds that are selected as clinical candidates will survive to the market.[1,2] The complexity of the drug discovery process is well recognized. A successful drug discovery project involves optimization of multiple parameters, such as potency on the primary target, selectivity and absorption, distribution, metabolism, excretion and toxicity (ADMET) properties. It is a very difficult path beginning with hit identification, lead generation and followed by lead optimization to achieve the optimal molecular properties needed by a clinical candidate. For a molecule to reach clinical candidate selection, it should have a high potency on the primary target, high selectivity, a good ADMET property profile as well as a good PKPD (pharmaco-kinetic and pharmaco-dynamic) profile. High potency on the primary target is an important factor to optimize during the drug development process to reach an optimal PKPD profile. It contributes, in most cases, to a low required dose for reaching the desired pharmacological effect. Thus the risk of having compound related safety liabilities decreases. However, other factors like ADMET properties, target engagement and residence time are very important as well. In many cases high potency and good ADMET properties are very challenging to combine in one molecule. For example, the dependence of potency on molecular size and lipophilicity has been well described in the literature.[3-6]

Finding the right balance between them[3,4] is still a very challenging task for medicinal chemists.

As a new descriptor for compound quality assessment, molecular topology class has been proposed. Molecular topology class is an extension of Murcko's molecular framework concept.[7,8] Comparing to Murcko molecular framework, it consists of only a few classes, which makes it possible to apply rigorous statistical methods to analyze the results in an informative way. In comparison a large dataset of molecules may consist of thousands of different molecular frameworks. Previously it has been observed that there is an enrichment of compounds with only one ring system (1TR) among drugs compared to clinical candidates and general bioactive compounds. This conclusion is still valid after correcting for any potential bias originating from lipophilicity (ClogP), molecular size and differences in target distribution among the different molecular topology classes.[7] The study was further extended with the inclusion of natural product and human metabolites and it was found that they also were enriched with compounds having only one ring system. Further analysis showed that published compounds are getting less similar to human metabolites over time. These results might be due to the change in medicinal chemistry practice during the last decades. Analyzing the shift in how often different topological classes occur in medicinal chemistry publications indicated that this is the case. The influence of the molecular topology on the selectivity of drugs has also been investigated.[9] Compounds with only one ring system were found to be more selective than compounds with two or more ring systems. To further understand the differences between molecular topology classes, it was decided to investigate a large set of compounds extracted from a commercial patent database. Pharmaceutical patents were chosen as the data source, since these compounds should have been reasonably well characterized by the organization filing the patent. It is also the major source for analyzing compounds synthesized and utilized in the pharmaceutical industry. It is acknowledged that an analysis of

data from a patent database mixes different type of targets and assays, so the data needs to be analyzed carefully to remove this type of bias. The patent data was used to investigate the frequency of different topological classes as a function of size and lipophilicity and to investigate the influence of topology on efficiency indices for ligand-target binding which are commonly applied in the drug discovery process, such as ligand efficiency (LE), ligand lipophilic efficiency (LLE) and Ligand-efficiency-dependent lipophilicity (LELP).

As it has been discussed earlier, screening of compounds purely on the potency could mean selecting compounds with poor ADMET properties.[3,4] Ligand efficiency (LE)[10] has been proposed as a measure for selecting lead compounds. Per definition LE is a method to normalize the potency with respect to the molecular size. It is therefore possible to compare potencies of compounds in a wide range of molecular size. LE is defined as:

$$LE = \Delta G / HEV \qquad \text{(Equation 1)}$$

Where $\Delta G = -RT \ln Ki$, $Ki = XC50 / 2$ and HEV is the number of heavy atoms.

There are criticisms that LE does not completely eliminate the size bias.[5,6,11] Despite of this, LE is still commonly used to assess the affinity of compounds for a target. Especially in fragment based drug design, LE is widely accepted and practically used, since this index gives priority to small molecules with relatively lower potency rather than large molecules with higher potency.[12,13]

Another useful ligand efficiency index is Ligand Lipophilicity Efficiency (LLE) [14], which is defined as:

$$LLE = pXC50 - ClogP \qquad \text{(Equation 2)}$$

LLE can be used to improve the potency while keeping the lipophilicity low.[14] It was shown by Tarcsay *et al.*[15] that increasing potency during hit identification and lead optimization also

increases lipophilicity in general. LLE is a useful index to control the lipophilicity while improving the potency, since high lipophilicity tends to cause adverse ADMET properties and high promiscuity.[14] It has been shown that LLE differentiates between hits, leads, and drugs.[15]

LE and LLE highlights low potency compounds with small size and low lipophilicity, which might otherwise be overlooked when analyzing hits from a high-throughput screening. The use of a small and less lipophilic molecule as a starting point is beneficial as the size and lipophilicity generally will be increased during the lead optimization process.[4,16]

As ligand efficiency does not consider the lipophilicity and ligand lipophilicity efficiency does not consider the molecular size, Ligand-Efficiency-dependent LipoPhilicity index (LELP) has recently been proposed to address this issue.[16] LELP is a descriptor which combines the three important factors, lipophilicity, molecular size, and potency into one descriptor. LELP is calculated using the following formula:

$$LELP = ClogP / LE \qquad \text{(Equation 3)}$$

Where LE is the ligand efficiency

It has been observed that LELP can differentiate between drugs from hits, leads and even from compounds that had entered phase II clinical trials, which was not possible to do using LLE.[15] It has also been suggested that the combination of molecular mass and lipophilicity provides a better prioritization of compounds during the screening process.[3] While LELP has not been used as widely in publications as LE and LLE, we still felt that it would be useful to include an efficiency indices that combines both size and lipophilicity in one measure in the analysis. Also other compound efficiency metrics were investigated like size-independent ligand efficiency (SILE).[6] The results were found to be consistent with that of LE, LLE and LELP and therefore they will not further be discussed here.

Thus LE, LLE and LELP have been shown to be useful indices to judge compound quality. In the current study, the influence of the molecular topology on LE, LLE and LELP for a set of compounds from the patent literature is investigated. The aim is to better understand why certain molecular topologies are more frequent among drugs than among general bioactive compounds.

Material and Methods

Data Sets

All the data used in current study was retrieved from a commercial source.[17] Compounds with a reported potency for at least one human target were extracted. In total 1,022,057 data points which correspond to bioactivity data (pXC50) for 695,322 unique structures and 972 human targets. Target classes represented in the dataset were GPCRs, Kinases, Ion Channels, Transporters, Proteases, Nuclear Hormone Receptors, Phosphotases, Oxidoreductases and Hydrolases. All compounds with a substructure composed of more than three amino acids were removed to assure there was no influence of patents covering peptides.

Property Calculations

ClogP was calculated using a commercially available program.[18] Molecular topology classes and the number of heavy atoms (*i.e.*, non-hydrogen atoms) were calculated with an in-house C++ program.[19] Wilcoxon rank-sum test[21] was performed with the open source statistic package, Python-statlib.[22]

Definition of molecular topology classes

The molecular topology class definition[7] is an extension of the molecular framework (MF) concept.[8] A molecule is divided into three subunits: terminal ring systems (TR), a molecular bridge (B), and side chains. A terminal ring system refers to a ring system which has only one connection to other ring systems in the molecule. The molecular bridge connects all of the terminal rings. The difference between the definition of a molecular bridge and the definition of a linker is that a molecular bridge might include additional ring systems, while a linker does not include any ring system at all. Any ring system which is directly connected through linkers to more than one other ring system is regarded as part of the molecular bridge. Thus, the MF is the combination of the terminal rings and the molecular bridge. Side chains refer to atoms that do not belong to the MF. Some examples of different topological classes are shown in Figure 1. The 2TR+B class is subdivided according to the number of ring systems in the molecular bridge to further enhance the analysis. In this study, 1TR, 2TR, 2TR+B and 3TR+B topology classes were considered. More complicated classes such as 4TR+B, 5TR+B were excluded due to their low occurrence in the dataset.

Result and Discussion

It is important to first investigate if there are differences in the activity data reported for the different topological classes. The distributions of the reported potencies (pXC50) for the six molecular topology classes are shown in Figure 2. The activity distribution is fairly similar for all classes, approximately 30% of the compounds have high potency ($pXC50 \leq 7.0$), approximately 28 % have an pXC50 between 7.0 and 6.0 and the rest has mainly potencies in the range between 6.0 and 5.0. Compounds having a 1TR topology has slightly higher reported averaged potency than compounds from the other topology classes. There are also significant differences in the number of reported molecules. The class 2TR+B_1 has the most

number of reported molecules, roughly 274,000 molecules, while 1TR has only 24,000 reported molecules. It reinforces the observations from an earlier study, which indicated a lot of 2TR+B and 3TR+B compounds are synthesized in modern medicinal chemistry programs. These compounds are then identified in screening campaigns and thereafter followed-up in drug discovery projects and accordingly patented. In order to investigate the influence of the lipophilicity for the dataset, the same set was binned according to ClogP for each topological class. As is seen in Figure 3a, distribution of the number of compounds differs a lot between the different topological classes. The 1TR class has a much higher proportion of molecules with a reported low ClogP compared to the other classes. Compounds from patents will reflect what type of compounds that have been found as hits in screening campaigns and followed-up with medicinal chemistry resources. The data shows that compounds identified from the 1TR class are on average less lipophilic. Similarly it was investigated how the distribution of molecules for the different topological classes varies as a function of their size. The trend is clear; compounds with fewer ring systems are generally smaller. Compounds with one ring system (1TR) are the smallest, while compounds with two ring system are in the middle (2TR and 2TR+B_0). This result is not surprised since in general the molecular size increases with increasing number of ring systems. Thus reported 1TR compounds are both less lipophilic and smaller than compounds from the other topological classes. However, in the Figures 3, the potency is not taken into account. Thus the identified high frequency of compounds belonging to the 1TR topological class might consist of weakly active compounds, even though 1TR compounds are on average slightly more active than compounds from other topological classes as shown in Figure 2. The average potency as a function of lipophilicity and size is plotted in Figure 4a and 4b. For ClogP values above 2.0, compounds belonging to the 1TR class are the most potent and accordingly have the best LLE. For ClogP below 2.0, 3TR+B has the highest potency; visual inspection of the compounds revealed that for the 3TR+B,

there exist a set of highly potent peptidomimetic compounds with low ClogP published. However, as already shown in Figure 3a the fraction of compounds for the 3TR+B class with a low ClogP is very low. To better understand the interrelationship between activity, size and lipophilicity, the LE was plotted against ClogP. It is seen that, for all ClogP values, the compounds from the 1TR topological class has the highest LE values followed by compounds from the 2TR topological class. Compounds from the 3TR topological class have the lowest LE. Thus the (peptidomimetic) compounds identified form the 3TR+B class with high potency and low ClogP has a low LE, due to their large size. In conclusion, compounds from the 1TR class have overall the best properties taking size and lipophilicity into account. More elaborated graphs describing the relationship between potency, topological class, size and lipophilicity are given in Figures S1-S4.

In this current study, it is of our interest to further explore what is the impact of molecular topology class on binding efficiency indices such as LE, LLE and LELP (Equations 1-3). LE, LLE and LELP values for different molecular topology class were compared (Figure 5). As already shown in Figure 2 compounds with a 1TR topology has slightly higher reported potencies, however, the differences in LE, LLE and LELP are significantly larger. Besides being slightly more potent, 1TR compounds are on average significantly smaller and less lipophilic which gives rise to the differences in LE, LLE and LELP.

However, it should be noted that so far the analysis does not take into account that compounds from different topological classes might be reported from different targets. In order to exclude that the identified results are affected by different topological classes modulates different targets, a separate analysis needs to be done where the median LE, LLE and LELP are calculated for each target and topological class. We selected only targets where there are at least 10 compounds known for each of the six topological classes. In total, 190 targets fulfilled this requirement. For each of these selected targets, the median values of LE, LLE

and LELP of every molecular topology class were calculated (Figure 6). Figure 6 shows that compounds belonging to the 1TR class has the highest median values for both LE and LE and the lowest median value for LELP. The topological classes with several ring systems have the lowest LE and LLE and highest LELP. Thus the results here mirror the results in Figure 5, when comparing the median LE, LLE and LELP values for the 190 targets with at least 10 compounds from each target class. Each pair of topological classes was also compared for all targets where at least one compound from each of the two topological classes is reported. The results are given in Table S1 and are consistent with the results of Figure 6. The most common target classes (Enzymes, GPCRs, Ion Channels and NHRs) were also investigated individually for each pair of topological classes and the results are shown in Table S2-S5 confirming the earlier results. Compounds belonging to the 1TR class have on average better LE, LLE and LELP for each target class. Thus the observed trends of LE, LLE and LELP with molecular topology in Figure 5 are still valid after correcting for any potential target bias. Some structure examples are shown in Figure 7. Here we have chosen a GPCR, Histamine receptor type 3 (H3), to illustrate how representative molecules look like. It should be noted that H3 has histamine as an endogenous agonist, which belongs to the 1TR molecular topology class; so it is not surprising that also synthetic 1TR compounds show high pXC50 towards this receptor.

Conclusions

It has been shown in earlier studies that the molecular topology influences clinical success and selectivity and the distribution among the topological classes are very different for different data sources. To better understand how clinical success is influenced by topological class, an analysis has been performed on bioactivity data reported in the patent literature. It

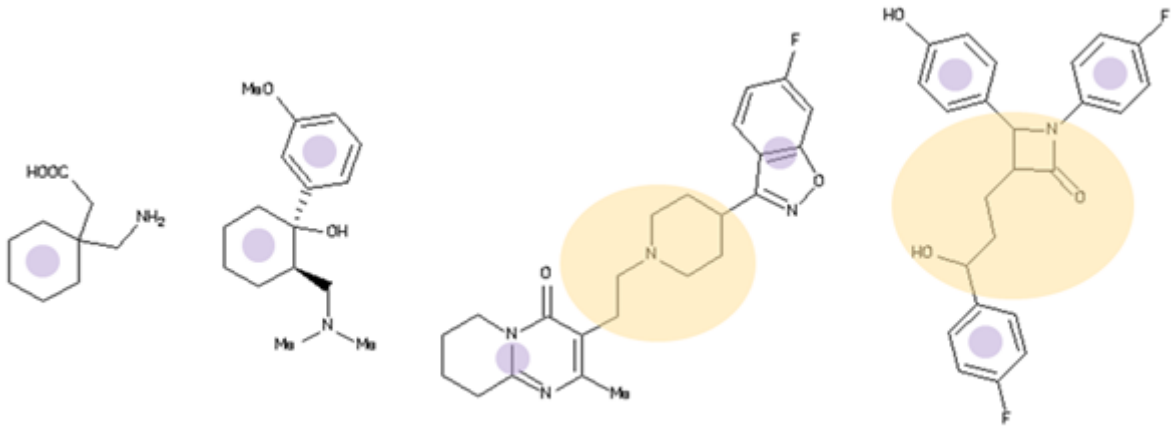
was first observed that compounds from the 1TR topological class are much less reported in patents than several other topological classes. Thereafter how the distribution of the different topological classes changes with ClogP and HEV was further analyzed and it was found that compounds belonging to the 1TR class are more likely to be smaller and less lipophilic than compounds with two ring systems (2TR and 2TR+B_0) and even more so compared to compounds with three or more ring systems (2TR+B_1, 2TR+B_2 and 3TR). It was also shown that at a constant size or lipophilicity, compounds with a 1TR ring system were generally slightly more active. An exception is compounds belonging to the 3TR+B class with low lipophilicity. They are mainly peptidomimetic compounds, and are the most potent compared to compounds belonging to other topology classes. However, only a small fraction of the 3TR+B class compounds has low lipophilicity. Most of the compounds have a high lipophilicity. Our studies also show that compounds with fewer ring systems have better LE, LLE and LELP. These trends are not due to differences in target distribution for compounds belonging to different topological classes. Thus the conclusions still hold even after taking into account that the distribution of active compounds for different molecular topology classes might be different for different targets. The results are also valid for all the major target classes. The reason that compounds from the 1TR topological class have better LE, LLE and LELP is two-fold, first, they have on average a slightly higher potency than compounds from the other topological classes even after taking size and lipophilicity into account. Second, compounds belonging to the 1TR topological class have generally lower lipophilicity and smaller size comparing with other topology classes among patented compounds. The stronger affinity for compounds with 1TR topology might be related to the prominence of compounds with 1TR topology among natural products and human metabolites. Small molecules present in nature might have evolved to interact strongly with their target and to accomplish that a 1TR topology might have been preferred by the evolution. As a consequence, synthetic

molecules with a 1TR topology might also interact more strongly with their protein target. However, the observation that compounds with a 1TR topology has better LE and LLE is empirical. Modeling of ligand-protein interactions is still very difficult as evidenced by the difficulties to accurately calculate the free energy for the ligand-protein complex. It is also an interesting observation that the fraction of compounds reported with a 1TR topology is much higher among small and hydrophilic compounds. This might be related to the modern medicinal chemistry practice that mainly produces 2TR+B and 3TR+B compounds. These compounds are then identified as hits in high-throughput screening and used in drug discovery project and thereafter reported in the patents. However, the results in this article indicate that screening more human metabolites and other natural products might be an approach to improve the success in drug discovery projects. These sets are enriched with compounds belonging to the 1TR topological class and might therefore produce screening hits that can be optimized to compounds with good LE and LLE. It should be noted that all retrospective analysis of patented compounds will per definition reflect the historical target space that has been exploited in drug discovery projects, thus there is no guarantee that the identified conclusions will hold for future drug discovery projects that might pursue different types of targets.

Reference:

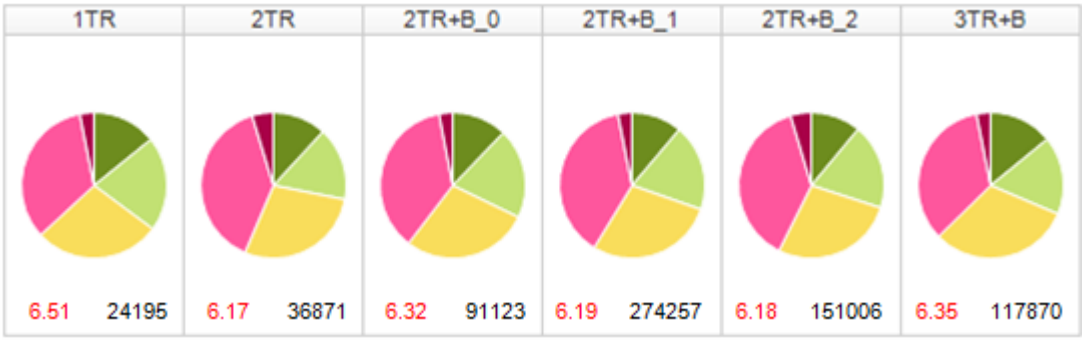
1. Peck, R. Driving earlier clinical attrition: if you want to find the needle, burn down the haystack. Considerations for biomarker development. *Drug Discov. Today.* **2007**, 12, 289-294.
2. Paul, S.; Mytelka D.; Dunwiddie C.; Persinger, C.; Munos, B.; Lindborg, S.; Schacht, A. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat. Rev. Drug. Discov.* **2010**, 9, 203-214.
3. Gleeson, P.; Hersey, A.; Montanari, D.; Overington, J. Probing the links between in vitro potency, ADMET and physicochemical parameters. *Nat. Rev. Drug. Discov.* **2011**, 10, 197-208.
4. Hann, M.; Keserü, G. Finding the sweet spot: the role of nature and nurture in medicinal chemistry. *Nat. Rev. Drug. Discov.* **2012**, 11, 355-365.
5. Raynolds, C.; Tounge, B.; Bembenek, S. Ligand Binding Efficiency: Trends, Physical Basis, and Implications. *J. Med. Chem.* **2008**, 51, 2432-2438.
6. Bembenek, S.; Tounge, B.; Reynolds, C. Ligand efficiency and fragment-based drug discovery. *Drug Discov. Today* **2009**, 14, 278-283.
7. Chen, H.; Yang, Y.; Engkvist, O. Molecular Topology Analysis of the Differences between Drugs, Clinical Candidate Compounds, and Bioactive Molecules. *J. Chem. Inf. Model.* **2010**, 50, 2141–2150.
8. Bemis, G.; Murcko M. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, 39, 2887-2893.
9. Yang, Y.; Engkvist, O.; Llinàs, A.; Chen, H. Beyond Size, Ionization State and Lipophilicity: Influence of Molecular Topology on ADMET Properties for drug-like compounds. *J. Med. Chem.* **2012**, 55, 3667-3677.
10. Hopkins, A.; Groom, C.; Alex A. Ligand efficiency: a useful metric for lead selection. *Drug Discov. Today* **2004**, 9, 430-431.
11. Nissink, J. Simple Size-Independent Measure of Ligand Efficiency. *J. Chem. Inf. Model*. **2009**, 49, 1617-1622.
12. Carr, R.; Congreve, M.; Murray C.; Rees D. Fragment-based lead discovery: leads by design. *Drug Discov. Today* **2005**, 10, 987-992.
13. Mortenson, P.; Murray C.; Assessing the lipophilicity of fragments and early hits. *J Comput. Aided Mol. Des.* **2011**, 25, 663-667.

14. Leeson, P.; Springthorpe B. The Influence of drug-like concepts on decision-making in medicinal chemistry. *Nat. Rev. Drug. Discov.* **2007**, 6, 881-890.
15. Tarcsay, Á.; Nyíri, K.; Keserü G. Impact of Lipophilic Efficiency on Compound Quality. *J. Med. Chem.* **2012**, 55, 1252-1260.
16. Keserü, G.; Makara, G. The influence of lead discovery strategies on the properties of drug candidates. *Nat. Rev. Drug. Discov.* **2009**, 8, 203-212.
17. *GVKBio databases*; GVK Biosciences Private Ltd.: Hyderabad, India.
18. *BioByte ClogP* version 4.3, http://www.biobyte.com (accessed Sep. 6, 2011).
19. *OEChem toolkit* 1.7.2; OpenEye Scientific Software, Inc., Santa Fe, NM.
20. Sorkau, E. Ringerkennung in chemischen Strukturen mit dem Computer. *Wiss. 2. Tech. Hochsch. Leuna-Meuseburg*. **1985**, 27, 765-770.
21. Wilcoxon, F. Individual Comparisons by Ranking Methods. *Biom. Bull.,* **1945,** 1, 80-83.
22. *Python-statlib* version 1.1, http://python-statlib.googlecode.com, **2007**.
23. Chen, H.; Engkvist, O.; Blomberg, N.; Li, J. A comparative analysis of the molecular topologies for drugs, clinical candidates, natural products, human metabolites and general bioactive compounds. *MedChemComm.* **2012**, 3, 312-321.



1TR 2TR 2TR+B_1 3TR+B

Figure 1. Examples of compounds from different topological classes. (TR = Terminal Ring(s), B = Bridge). Purple is showing the terminal rings and orange is showing the molecular bridge in each compound.



Colored by potency (pXC50)

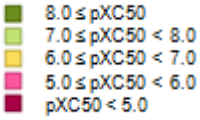
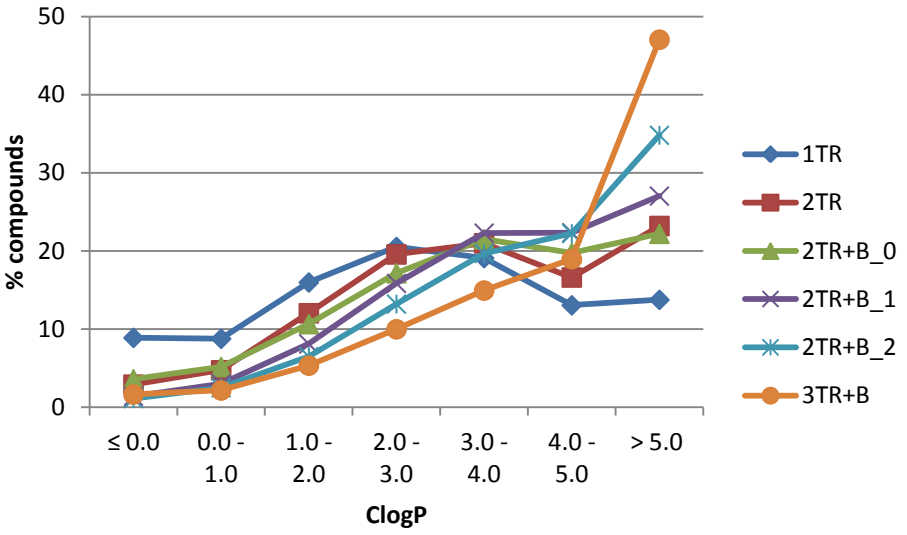
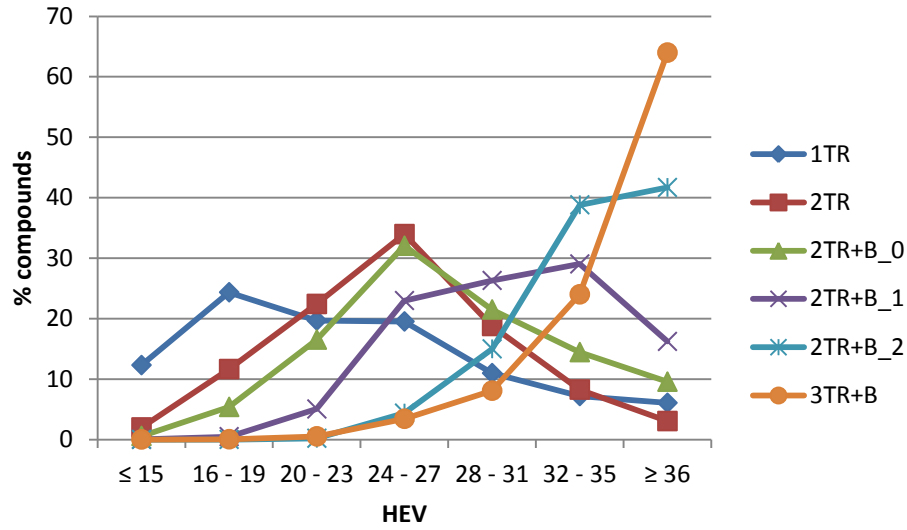


Figure 2. The potency distribution for the six topological classes. The number of molecules represented by each pie-chart is shown in black and the average pXC50 values are shown in red.

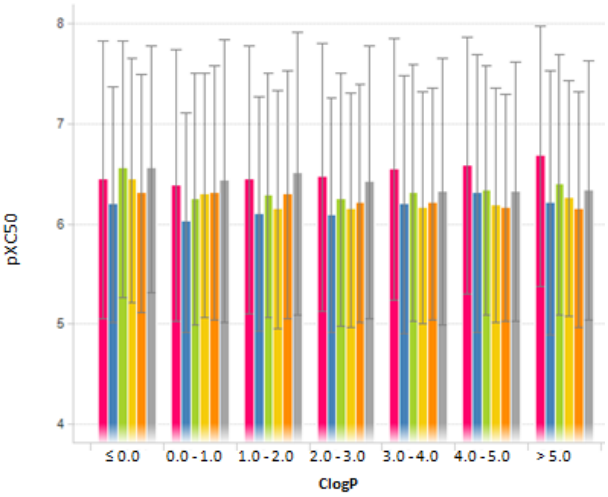


(a)

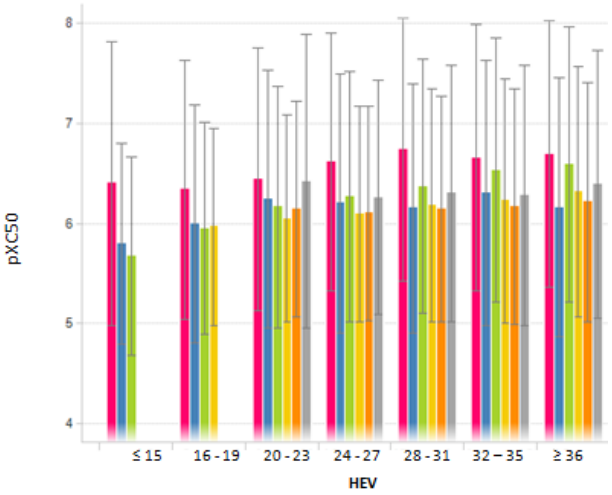


(b)

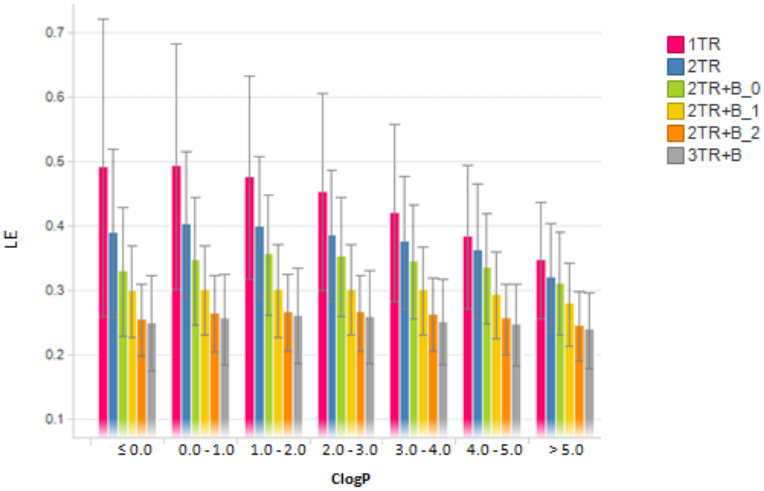
Figure 3. The percentage of compounds in the six topological classes versus (a) ClogP and (b) heavy atom count (HEV).



(a)

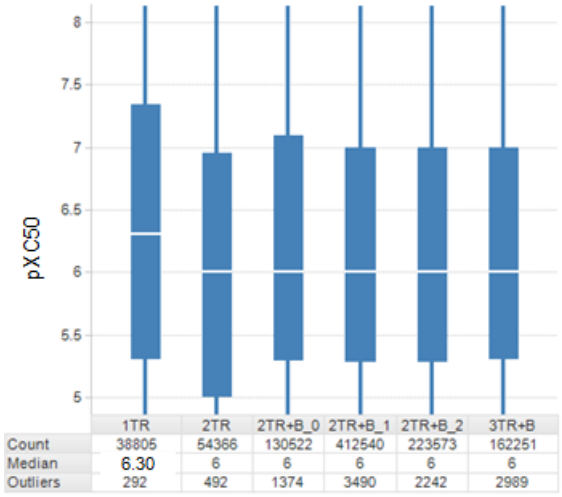


(b)



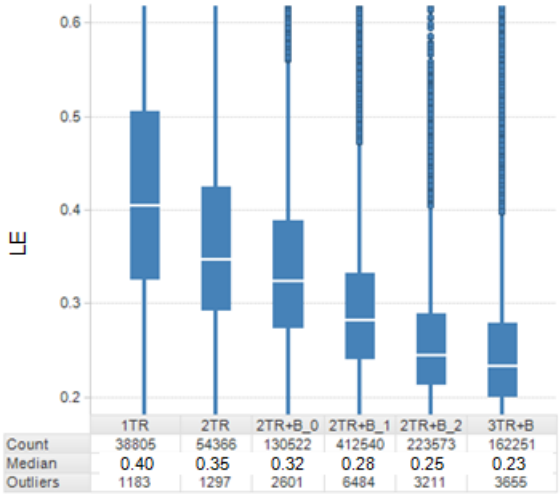
(c)

Figure 4. The relationship between (a) ClogP and mean pXC50, (b) Number of heavy atoms (HEV) and mean pXC50, and (c) ClogP and mean LE for the different topological classes.



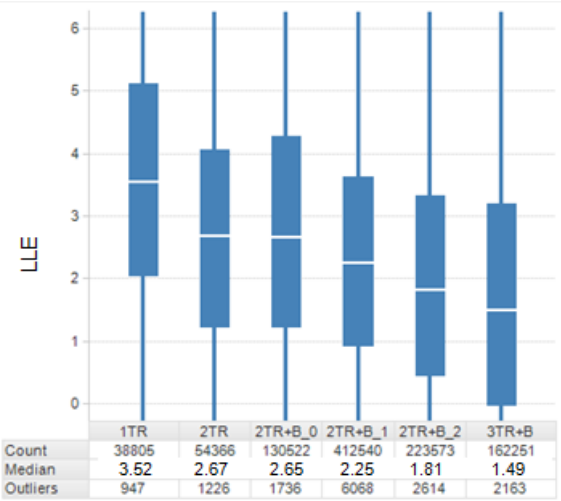
	1TR	2TR	2TR+B_0	2TR+B_1	2TR+B_2	3TR+B
Count	38805	54366	130522	412540	223573	162251
Median	6.30	6	6	6	6	6
Outliers	292	492	1374	3490	2242	2989

(a) Potency (pXC50)



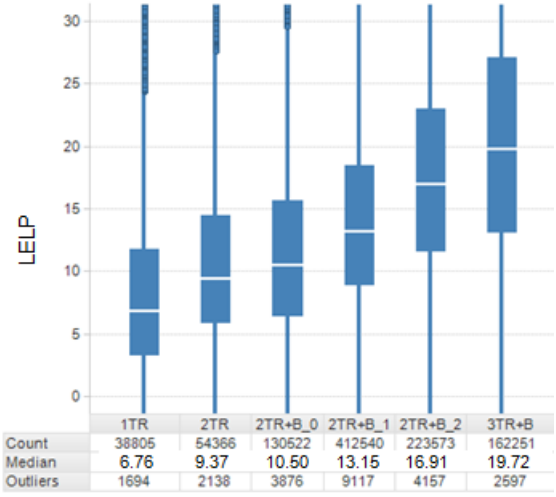
	1TR	2TR	2TR+B_0	2TR+B_1	2TR+B_2	3TR+B
Count	38805	54366	130522	412540	223573	162251
Median	0.40	0.35	0.32	0.28	0.25	0.23
Outliers	1183	1297	2601	6484	3211	3655

(b) Ligand Efficiency (LE)



	1TR	2TR	2TR+B_0	2TR+B_1	2TR+B_2	3TR+B
Count	38805	54366	130522	412540	223573	162251
Median	3.52	2.67	2.65	2.25	1.81	1.49
Outliers	947	1226	1736	6068	2614	2163

(c) Ligand Lipophilic Efficiency (LLE)



	1TR	2TR	2TR+B_0	2TR+B_1	2TR+B_2	3TR+B
Count	38805	54366	130522	412540	223573	162251
Median	6.76	9.37	10.50	13.15	16.91	19.72
Outliers	1694	2138	3876	9117	4157	2597

(d) Ligand-efficiency-dependent lipophilicity (LELP)

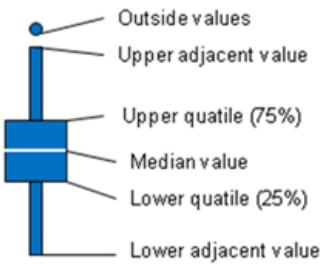
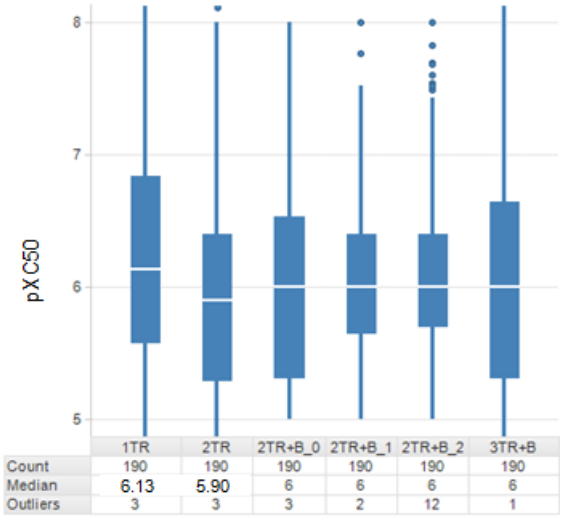
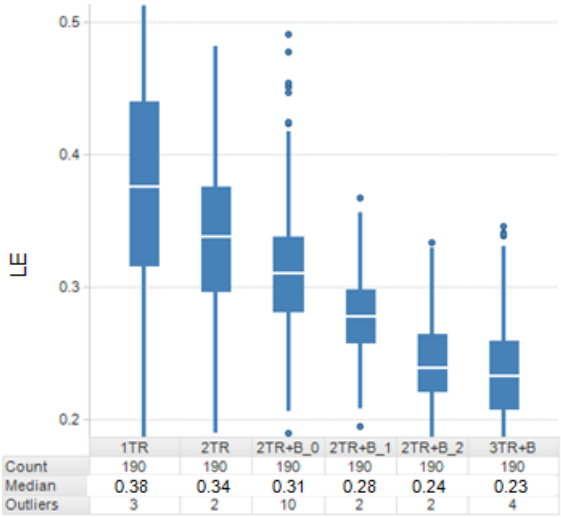


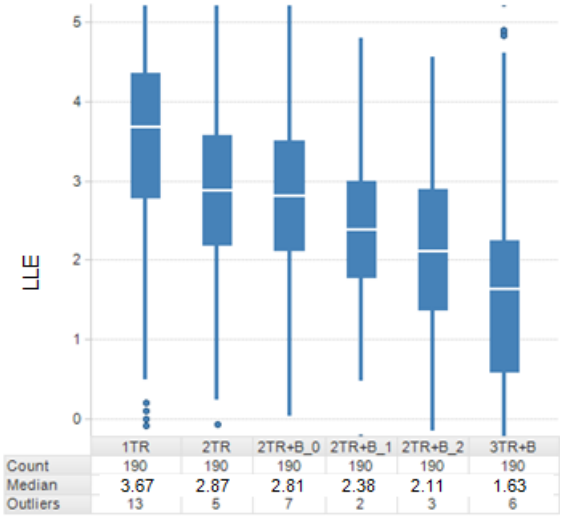
Figure 5. The (a) pXC50, (b) LE, (c) LLE and (d) LELP distributions for each topological class. Count is the number of molecules for each topological class.



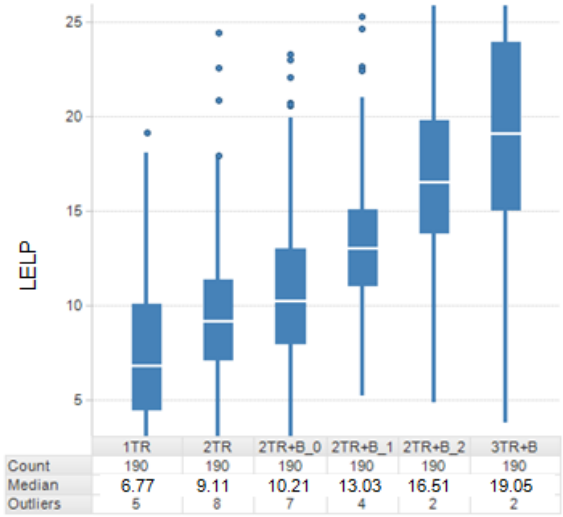
(a) Potency (pXC50)



(b) Ligand Efficiency (LE)

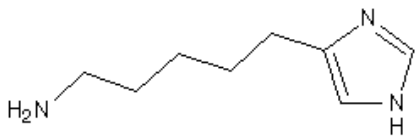


(c) Ligand Lipophilic Efficiency (LLE)

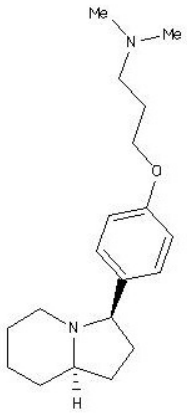


(d) Ligand-efficiency-dependent lipophilicity (LELP)

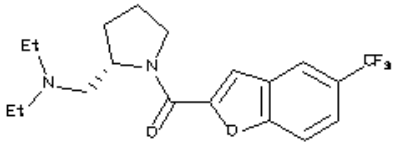
Figure 6. The distribution of the median (a) pXC50, (b) LE, (c) LLE and (d) LELP for each target and topological class. Count is the number of targets which have more than 10 compounds for each topological class.



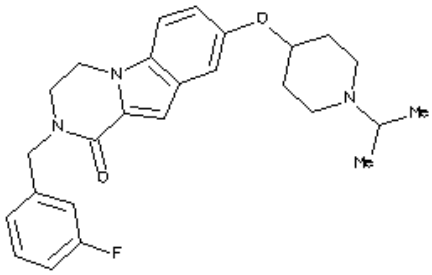
(a) 1TR
pXC50: 7.70, LE: 0.95,
LLE: 7.22, LELP: 0.49



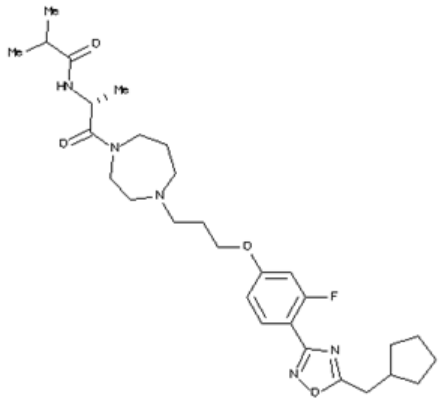
(b) 2TR
pXC50: 7.72, LE: 0.48,
LLE: 3.84, LELP: 8.11



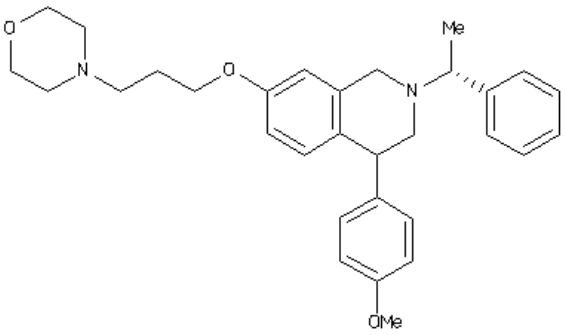
(c) 2TR+B_0
pXC50: 7.55, LE: 0.40,
LLE: 3.20, LELP: 11.0



(d) 2TR+B_1
pXC50: 7.64, LE: 0.33,
LLE: 3.03, LELP: 14.2



(e) 2TR+B_2
pXC50: 7.55, LE: 0.26,
LLE: 2.92, LELP: 17.6



(f) 3TR+B
pXC50: 7.85, LE: 0.30,
LLE: 1.70, LELP: 20.7

Figure 7. Representative structures for Histamine receptor H3. The median values for each topology are: 1TR: pXC50 5.85, LE 0.52, LLE 3.67, LELP 4.35, 2TR: pXC50 5.70, LE 0.37, LLE 2.94, LELP 8.88, 2TR+B_0: pXC50 5.70, LE 0.33, LLE 2.76, LELP 8.48, 2TR+B_1: pXC50 5.70, LE 0.31, LLE 2.86, LELP 10.7, 2TR+B_2: pXC50 5.70, LE 0.25, LLE 2.65, LELP 12.3, 3TR+B: pXC50 5.70, LE 0.24, LLE 2.77, LELP 14.2.

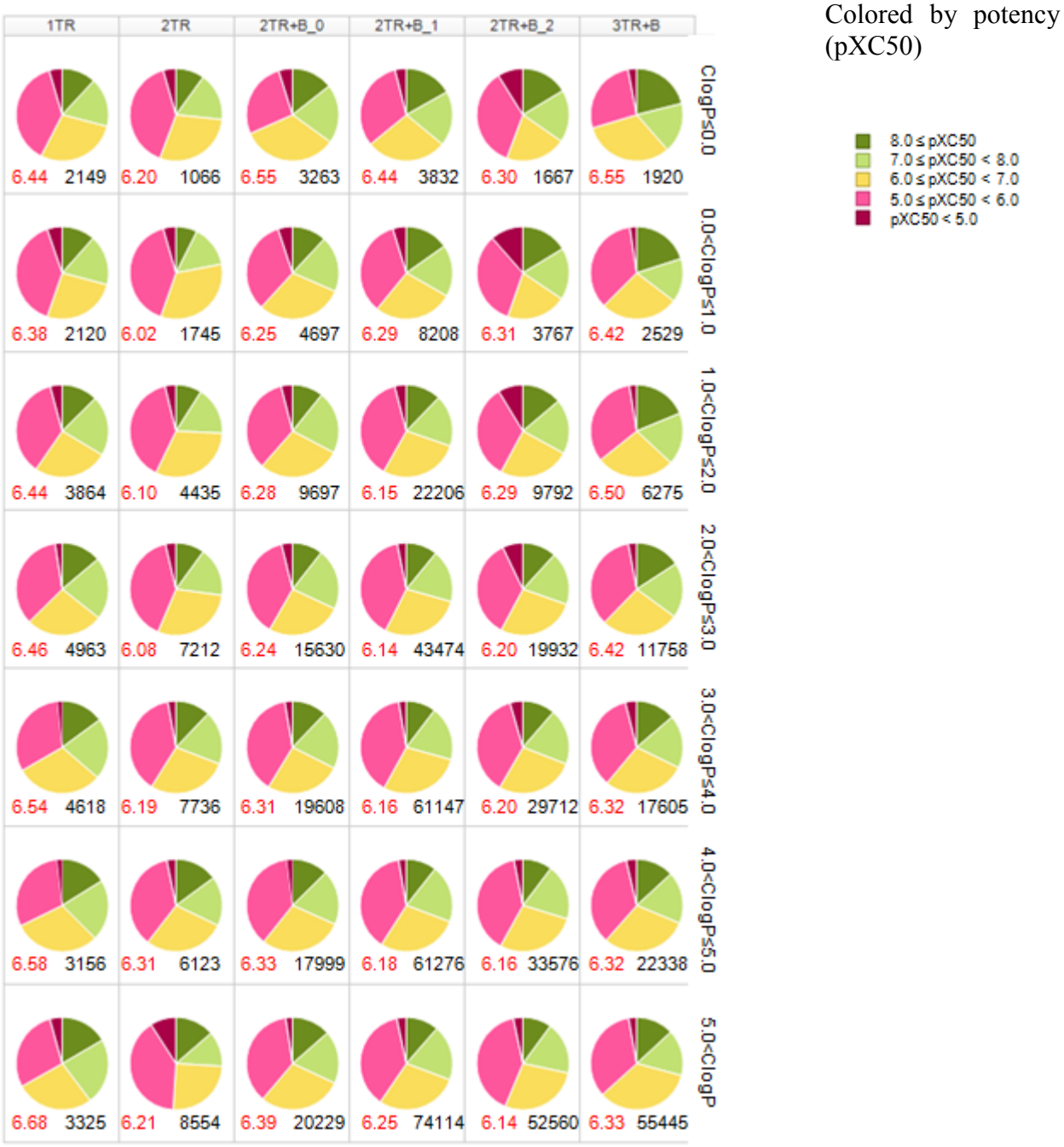


Figure S1. The potency distribution of the six topological classes for different ClogP intervals. The number of molecules represented by each pie-chart is shown in black and the average pXC50 values are shown in red.

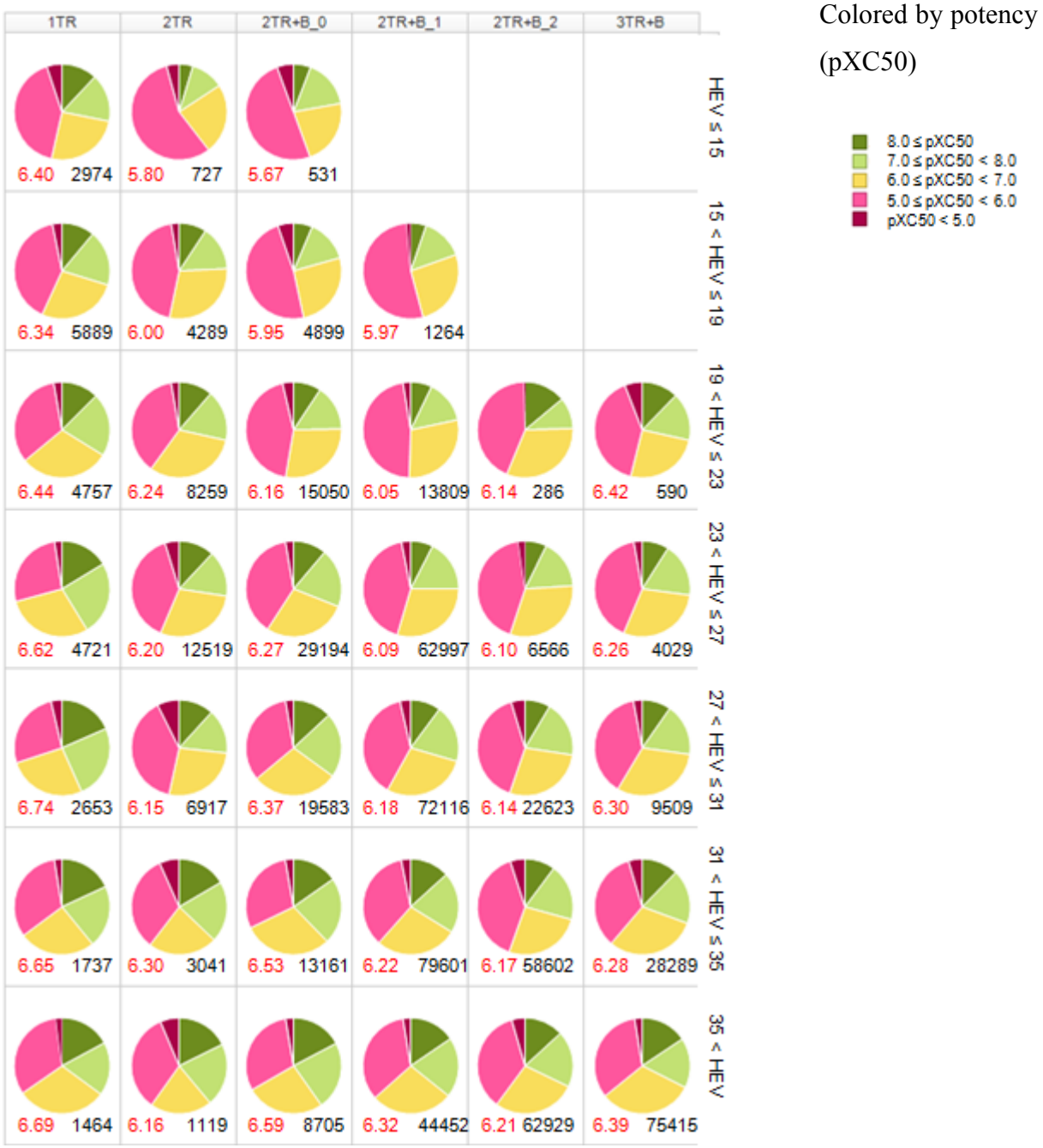


Figure S2. The potency distribution for the six topological classes for different heavy atom count (HEV). The number of molecules represented by each pie-chart is shown in black and the average pXC50 values are shown in red.

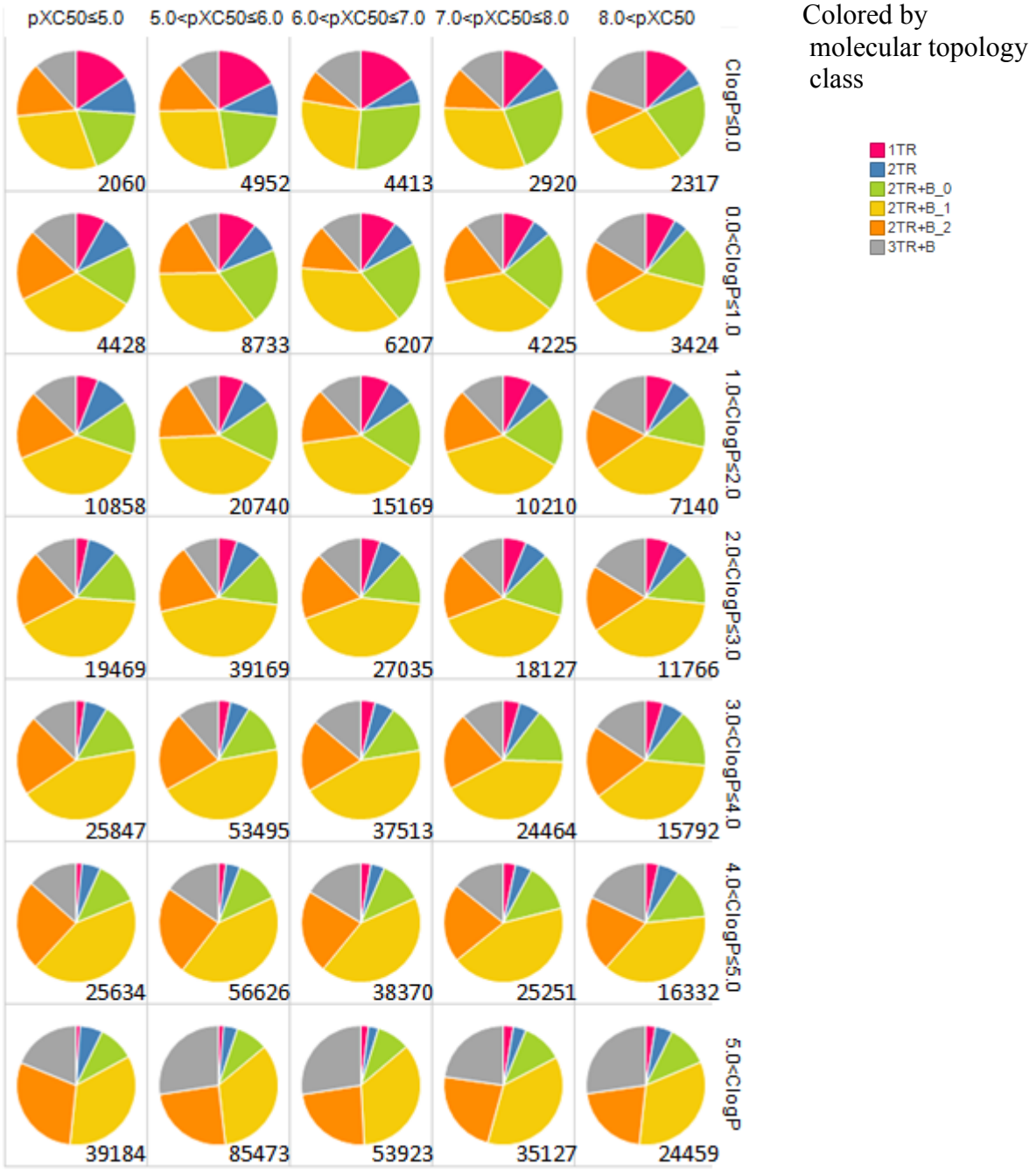


Figure S3. The topological class distribution for the five potency ranges and different ClogP intervals. The number of molecules for each pie-chart is indicated in the figure.

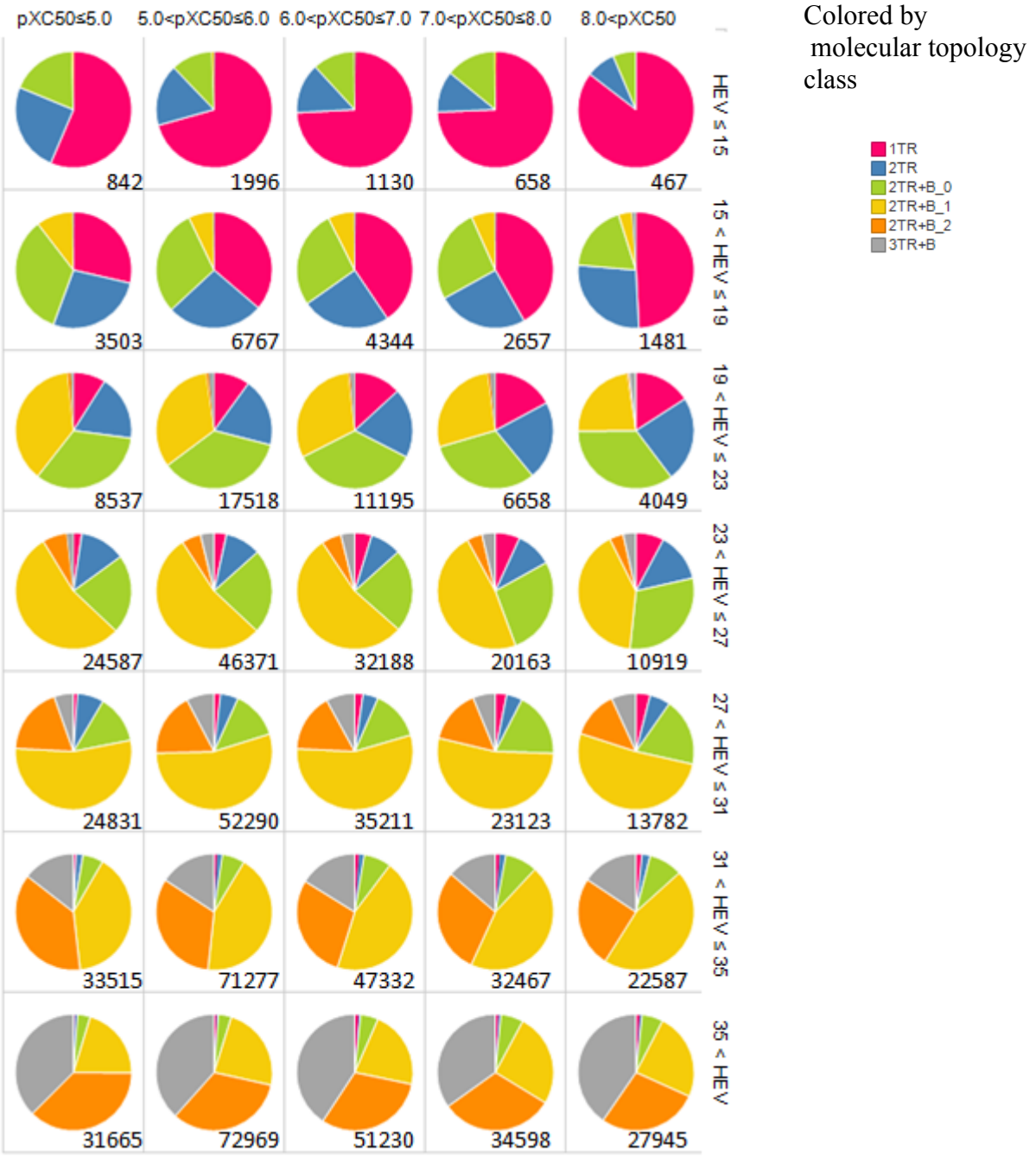


Figure S4. The topological class distribution for five different potency ranges and different number of heavy atoms (HEV). The number of molecules represented by each pie-chart is also shown.

Table S1. The differences between the median of LE, LLE and LELP between two molecular topology classes for all targets that have compounds from both topological classes. The numbers in the parentheses are the number of targets, which the two topological classes have in common.

ΔLE	**2TR**	**2TR+B_0**	**2TR+B_1**	**2TR+B_2**	**3TR+B**
1TR	-0.03 (466)	-0.06 (553)	-0.09 (548)	-0.11 (465)	-0.12 (442)
2TR	-	-0.03 (505)	-0.06 (521)	-0.10 (472)	-0.10 (434)
2TR+B_0	-	-	-0.03 (641)	-0.06 (534)	-0.07 (510)
2TR+B_1	-	-	-	-0.03 (584)	-0.04 (543)
2TR+B_2	-	-	-	-	-0.01 (490)

ΔLLE	**2TR**	**2TR+B_0**	**2TR+B_1**	**2TR+B_2**	**3TR+B**
1TR	-0.73 (466)	-0.77 (553)	-1.27 (548)	-1.62 (465)	-2.04 (442)
2TR	-	0.01 (505)	-0.56 (521)	-0.89 (472)	-1.13 (434)
2TR+B_0	-	-	-0.52 (641)	-0.80 (534)	-1.12 (510)
2TR+B_1	-	-	-	-0.35 (584)	-0.73 (543)
2TR+B_2	-	-	-	-	-0.37 (490)

ΔLELP	**2TR**	**2TR+B_0**	**2TR+B_1**	**2TR+B_2**	**3TR+B**
1TR	2.05 (466)	3.08 (553)	6.02 (548)	9.78 (465)	11.52 (442)
2TR	-	0.99 (505)	4.12 (521)	7.90 (472)	9.80 (434)
2TR+B_0	-	-	3.15 (641)	6.69 (534)	8.36 (510)
2TR+B_1	-	-	-	3.46 (584)	5.33 (543)
2TR+B_2	-	-	-	-	2.34 (490)

Table S2. The differences between the median of LE, LLE and LELP between two molecular topology classes for enzymes that have compounds from both topological classes. The numbers in the parentheses are the number of targets.

ΔLE	**2TR**	**2TR+B_0**	**2TR+B_1**	**2TR+B_2**	**3TR+B**
1TR	-0.02 (294)	-0.05 (362)	-0.08 (361)	-0.09 (298)	-0.10 (275)
2TR	-	-0.02 (310)	-0.06 (323)	-0.10 (286)	-0.10 (255)
2TR+B_0	-	-	-0.04 (406)	-0.06 (323)	-0.07 (300)
2TR+B_1	-	-	-	-0.03 (360)	-0.04 (319)
2TR+B_2	-	-	-	-	-0.01 (285)

ΔLLE	**2TR**	**2TR+B_0**	**2TR+B_1**	**2TR+B_2**	**3TR+B**
1TR	-0.81 (294)	-0.61 (362)	-1.17 (361)	-1.45 (298)	-1.85 (275)
2TR	-	0.29 (310)	-0.52 (323)	-0.80 (286)	-1.07 (255)
2TR+B_0	-	-	-0.55 (406)	-0.89 (323)	-1.19 (300)
2TR+B_1	-	-	-	-0.35 (360)	-0.73 (319)
2TR+B_2	-	-	-	-	-0.35 (285)

ΔLELP	**2TR**	**2TR+B_0**	**2TR+B_1**	**2TR+B_2**	**3TR+B**
1TR	1.83 (294)	2.44 (362)	5.13 (361)	8.26 (298)	10.39 (275)
2TR	-	0.28 (310)	3.77 (323)	7.52 (286)	9.59 (255)
2TR+B_0	-	-	3.38 (406)	6.71 (323)	8.32 (300)
2TR+B_1	-	-	-	3.44 (360)	5.40 (319)
2TR+B_2	-	-	-	-	2.14 (285)

Table S3. The differences between the median of LE, LLE and LELP between two molecular topology classes for GPCRs that have compounds from both topological classes. The numbers in the parentheses are the number of targets.

ΔLE	**2TR**	**2TR+B_0**	**2TR+B_1**	**2TR+B_2**	**3TR+B**
1TR	-0.06 (116)	-0.10 (126)	-0.15 (123)	-0.18 (119)	-0.20 (114)
2TR	-	-0.04 (130)	-0.06 (136)	-0.10 (134)	-0.10 (125)
2TR+B_0	-	-	-0.03 (158)	-0.06 (153)	-0.07 (146)
2TR+B_1	-	-	-	-0.03 (166)	-0.04 (155)
2TR+B_2	-	-	-	-	-0.01 (151)

ΔLLE	**2TR**	**2TR+B_0**	**2TR+B_1**	**2TR+B_2**	**3TR+B**
1TR	-0.73 (116)	-1.22 (126)	-1.83 (123)	-1.87 (119)	-2.74 (114)
2TR	-	-0.55 (130)	-0.59 (136)	-0.99 (134)	-1.26 (125)
2TR+B_0	-	-	-0.28 (158)	-0.70 (153)	-1.07 (146)
2TR+B_1	-	-	-	-0.34 (166)	-0.63 (155)
2TR+B_2	-	-	-	-	-0.42 (151)

ΔLELP	**2TR**	**2TR+B_0**	**2TR+B_1**	**2TR+B_2**	**3TR+B**
1TR	2.78 (116)	5.16 (126)	8.33 (123)	11.70 (119)	15.10 (114)
2TR	-	2.43 (130)	4.41 (136)	8.28 (134)	10.09 (125)
2TR+B_0	-	-	2.79 (158)	6.61 (153)	8.85 (146)
2TR+B_1	-	-	-	3.05 (166)	4.91 (155)
2TR+B_2	-	-	-	-	2.68 (151)

Table S4. The differences between the median of LE, LLE and LELP between two molecular topology classes for ion channels that have compounds from both topological classes. The numbers in the parentheses are the number of targets.

ΔLE	**2TR**	**2TR+B_0**	**2TR+B_1**	**2TR+B_2**	**3TR+B**
1TR	-0.06 (35)	-0.11 (43)	-0.13 (41)	-0.19 (29)	-0.17 (30)
2TR	-	-0.03 (44)	-0.06 (40)	-0.09 (33)	-0.09 (32)
2TR+B_0	-	-	-0.03 (49)	-0.07 (37)	-0.06 (38)
2TR+B_1	-	-	-	-0.04 (36)	-0.03 (41)
2TR+B_2	-	-	-	-	0.00 (33)

ΔLLE	**2TR**	**2TR+B_0**	**2TR+B_1**	**2TR+B_2**	**3TR+B**
1TR	-0.71 (35)	-0.91 (43)	-1.46 (41)	-2.70 (29)	-1.84 (30)
2TR	-	-0.09 (44)	-0.61 (40)	-1.03 (33)	-1.19 (32)
2TR+B_0	-	-	-0.77 (49)	-1.21 (37)	-1.42 (38)
2TR+B_1	-	-	-	-0.36 (36)	-0.83 (41)
2TR+B_2	-	-	-	-	-0.41 (33)

ΔLELP	**2TR**	**2TR+B_0**	**2TR+B_1**	**2TR+B_2**	**3TR+B**
1TR	2.94 (35)	4.80 (43)	6.76 (41)	10.39 (29)	12.66 (30)
2TR	-	0.40 (44)	4.25 (40)	8.72 (33)	10.34 (32)
2TR+B_0	-	-	3.76 (49)	7.99 (37)	8.64 (38)
2TR+B_1	-	-	-	4.26 (36)	5.79 (41)
2TR+B_2	-	-	-	-	1.44 (33)

Table S5. The differences between the median of LE, LLE and LELP between two molecular topology classes for NHRs that have compounds from both topological classes. The numbers in the parentheses are the number of targets.

ΔLE	**2TR**	**2TR+B_0**	**2TR+B_1**	**2TR+B_2**	**3TR+B**
1TR	0.01 (21)	-0.05 (22)	-0.08 (23)	-0.09 (19)	-0.10 (23)
2TR	-	-0.05 (21)	-0.07 (22)	-0.09 (19)	-0.08 (22)
2TR+B_0	-	-	-0.04 (28)	-0.05 (21)	-0.04 (26)
2TR+B_1	-	-	-	-0.02 (22)	-0.02 (28)
2TR+B_2	-	-	-	-	0.00 (21)

ΔLLE	**2TR**	**2TR+B_0**	**2TR+B_1**	**2TR+B_2**	**3TR+B**
1TR	0.32 (21)	-0.22 (22)	-0.55 (23)	-1.10 (19)	-1.42 (23)
2TR	-	-0.64 (21)	-1.09 (22)	-0.63 (19)	-1.32 (22)
2TR+B_0	-	-	-0.51 (28)	-0.38 (21)	-0.86 (26)
2TR+B_1	-	-	-	-0.11 (22)	-0.71 (28)
2TR+B_2	-	-	-	-	-0.38 (21)

ΔLELP	**2TR**	**2TR+B_0**	**2TR+B_1**	**2TR+B_2**	**3TR+B**
1TR	-0.03 (21)	3.93 (22)	5.65 (23)	11.90 (19)	11.88 (23)
2TR	-	3.13 (21)	5.79 (22)	11.36 (19)	10.02 (22)
2TR+B_0	-	-	4.17 (28)	5.95 (21)	6.54 (26)
2TR+B_1	-	-	-	3.70 (22)	3.34 (28)
2TR+B_2	-	-	-	-	1.63 (21)