A Structural and Functional View of Polypharmacology

Aurelio A. Moya-Garcia[1], Natalie L. Dawson[1], Felix A. Kruger[2#], John P. Overington[2‡], Christine Orengo[1] and Juan A.G. Ranea[3*]

[1] Institute of Structural and Molecular Biology, University College of London, Gower Street, London WC1E 6BT UK

[2] European Molecular Biology Laboratory - European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, CB10 1SD UK

[3*] Department of Molecular Biology and Biochemistry – CIBER de Enfermedades Raras, University of Malaga, 29071, Malaga, Spain

[#] Current Address: UCL Farr Institute, Euston Road 222, London NW1 2DA UK

[‡] Current Address: Stratified Medical. 40 Churchway - London, NW1 1LW

*Corresponding author. E-mail: ranea@uma.es

keywords: drug polypharmacology, protein domains, drug targets

Abstract

The similarity property principle states that similar compounds have similar properties. In this study, we demonstrate that validity this principle holds well when the drug targets are protein domains. We leverage the similarity property principle to explore the druggability of CATH-FunFams, a type of protein domain and we use the associations between drugs and CATH-FunFams to explore drug polypharmacology by considering how the drugs' pharmacological effects arise from the molecular targets they interact with. Our results demonstrate that drug protein interactions are mediated by drug-domain interactions and that CATH-FunFams provide a reasonable annotation level for drug-target interactions, opening a new research direction in target identification.

Author Summary

Similar drugs tend to target the same proteins and therefore tend to have the same biological action. In this work we assess this general trend, the Similarity Property Principle, and use it to investigate the potential of CATH Functional Families, a structurally and functionally coherent protein domain definition, as drug targets. We show that the interactions between drugs and their targets are mediated by these Functional Families, and that they provide a useful mean to identify new drug targets.

Introduction

We are facing the so called "data deluge" in almost every aspect of scientific research. At the same time, we witness the increasing gap between the amount of raw biological data available and processed biological information, as well as the gap between this biological information and the actual new knowledge [1]. The concept of "dark matter" illustrates our current situation: only a small part of the available data has been integrated into new knowledge [2]. Systems approaches aim to cover this gap, leveraging data from disparate "omics" analyses to build network (and other) models for analysing and predicting drug action [3,4], and aim to develop tools to uncover new knowledge from this growing "dark matter" [5]. Such approaches have highlighted the potential limitations of viewing drug action from the perspective of a single target and have provided some support for the need for multi-target approaches in drug discovery [4,6,7]. The field provides a growing body of evidence against two principles that guide drug discovery: (i) compounds specifically target one particular and critical biological agent (often a protein); and (ii) this molecular target is involved in one function, namely a critical point or step in a disease process. Therefore, drugs act as "magic bullets" acting on one molecular target, affecting one biological process and thus effecting a cure with few other consequences. However, many drugs bind to multiple targets and molecular targets are involved in multiple processes and perform multiple biological functions [8,9]. Hence the ability of drugs to bind multiple molecular targets and to affect multiple biological processes (a characteristic referred to as polypharmacology) should not be considered as an exception [6,10,11].

Polypharmacological behaviour has a mechanistic aspect, in which a drug binds several molecular targets, and a functional aspect in which the drug perturbs several biological processes. It is often recognised as an unintended phenomenon [12] and the rational

design of polypharmacological ligands is less frequently undertaken and remains a challenging task [6,13]. A major difficulty is the analysis of the relationship between the structure and the biological activity of molecules interacting with different biological targets (structure-activity relationship; SAR) along with pharmacokinetic/pharmacodynamic modelling, in which drug concentrations (and the corresponding modulated target spectrum) vary over time.

The similarity property principle (which states that similar compounds should have similar biological activity) considers molecular similarity as a guide to the biological action of a small molecule [14,15] and is at the core of any quantitative structure-activity relationship (QSAR) and chemogenomic approaches [16]. The general validity of this principle has however been questioned on the basis of a systematic exploration of the relationships amongst drug structures and their targets [17,18].

A druggable target is generally a protein with activity that can be modulated by a drug [19], i.e. a target is the mediator of a drug's activity. The activity of a compound can be considered at different levels in the biological hierarchy, ranging from macromolecules, to organelles, cellular or tissue types, organs, even at the species level; nevertheless, the biological activity of a small molecule is the result of its interaction with one or several biological targets at the molecular level. Target identification is a crucial task when considering application of polypharmacological compounds and it is important to identify synergistic combinations of targets, rather than single targets [4]. This analysis is often complicated by the fact that many binding events will be silent with respect to phenotypic modulation and emergent drug efficacy.

Most human targets are proteins that are composed of more than one domain [20,21], but we lack a unified definition of protein domain. Under the accepted and general definition

that a protein domain is a functional and structural module within a protein, there are several ways to identify and classify protein domains [22]: classification based on structure, SCOP [23] and CATH [24]; classification based on sequence, Pfam [25]; and function oriented domain classifications such as the functional families classified in CATH, CATH-FunFams [24,26]. In general terms, domains are compact and functional structural units that can be considered the evolutionary and structural building blocks of proteins. Since domains are units of structure [27] and there is a limited repertoire of domain types [28], they are combined to form different proteins with different overall functions [29]. Furthermore, protein-protein interactions are dominated by discrete domain-domain interactions [30]. Recent research suggests that protein domains mediate the interactions between a drug and its targets [16,31,32]: protein domains are a major factor in the polypharmacology of approved and experimental drugs [33]; binding sites tend to lie completely in a domain or at the interface of multiple domains [32]; there are privileged druggable protein domains [34]. These results support the idea that a particular structural domain can be the druggable entity in a protein target. Since proteins have a modular structure and domains are repeatedly found in different proteins, the reason why a compound binds different protein targets can be that they share a domain that is the actual target for the compound. Furthermore, since protein domains determine protein function, the association of drugs with domains will inform on the biological processes they perturb, offering a rich perspective of drug polypharmacology.

In this study, we assess the pharmacology of these structural and functional building blocks of protein targets. We leverage the similarity property principle to show that protein domains are the druggable entities within targets and direct the biochemical interactions and biological functions of drugs. Furthermore, we show that these units of protein structure and protein function explain the polypharmacology of approved drugs, enabling a

joint exploration of drugs pharmacological effects, and their mechanisms of action –i.e. the domains they bind.

Results

Similarity property principle

The similarity property principle (SPP) establishes that drugs with similar molecular structure are likely to have the same properties. For clarity, these properties are the modes of action and the mechanisms of action. The mode of action of a drug describes the functional changes produced by the drug on a living system [35], that is the drug's pharmacological effect, while the mechanism of action usually refers to the targets through which a drug produces its pharmacologic effect [36].

Similar drugs have similar mechanisms of action

We considered two different types of protein domain definitions: Pfam-A and CATH-FunFams to investigate the relationship between drug similarity and similarity in the mechanism of action.

Fig 1 shows the similarities of the interaction profiles of drugs as a function of their molecular similarity, for the three types of molecular targets analysed (proteins, Pfam-A domains and CATH-FunFams). High values of the Jaccard association index indicate that a pair of drugs have similar interaction profiles, thus where the association index = 1, the two drugs have the same targets. For proteins and CATH-FunFams similar drugs (i.e. Tc ≥ 0.65, see Supporting Information) tend to have similar interaction profiles, that is they tend to bind the same targets, while different drugs show different interaction profiles. For Pfam-A domains the interactions profile similarities are relatively flat, with no marked

difference based on drug similarity showing low variation between different and similar drugs.

This is expected, as related domains in many different proteins would collapse into a single Pfam-A family and result in higher Jaccard values. To a lesser degree, this applies to CATH-FunFams, which on average show more intermediate Jaccard values. This is because CATH-FunFams tend to separate domains according to their functional similarity and multi-domain context.

Similar drugs are involved in the same biological processes

Gene Ontology (GO) is a comprehensive vocabulary, representing molecular functions, biological processes and cellular locations that has become the standard to describe and annotate the cellular functions of genes and proteins [37,38]. We used the available GO annotations of proteins and CATH-FunFams as a proxy of the drugs modes of action. Although GO Biological Process (GOBP) annotations do not represent proper modes of action, they are a useful approximation of the biological processes that are perturbed by the action of drugs. To our knowledge there is no GO annotation of Pfam-A families, thus we performed the analysis of the biological processes affected by drugs with Proteins and CATH-FunFams. We produced datasets of drug-GOBP associations for CATH-FunFams and proteins to evaluate the similarity property principle in terms of the drugs mode of action. Fig 2 shows that GOBP terms are correlated with drug molecular similarity, both in the protein (left panels) and the CATH-FunFam levels (right panels). We observe the same behaviour in the drug-GOBP association as observed above in the analysis of drug-targets associations, although the similarity property principle is more evident with GOBP inherited annotations. Drugs with similar molecular structure perturb the same biological processes, while different drugs tend to act on different biological processes.

Our data complies with the similarity property principle, regardless of how we measure the molecular similarity between drugs and how we calculate the similarity of their target profiles (see S4 Fig and S5 Fig). This illustrates that the tendency of similar drugs to have the same targets and functions is an inherent property of drugs and targets. Our drug-domain associations are almost as good as the known drug-protein associations; therefore we consider it as a validation of our drug-domain associations, and further analysed them to study drug polypharmacology.

Drug polypharmacology through protein domains

Led by the idea that the modes of action of a drug (i.e. its pharmacological effects) can be understood through its mechanisms of action (i.e. its targets), we analysed the modes of action that stem from the drug's associations with proteins and with CATH-FunFams. We defined the polypharmacology potential of each drug as the number of different GOBP terms the drug is associated with, as an approximation of the drugs capacity to perturb biological processes and thus get an insight into the functional aspect of drug polypharmacology. Fig 3 suggests that the association of drugs with CATH-FunFams unveils a polypharmacology potential that is not evident from the annotation of drugs with protein targets. It is therefore possible that polypharmacology at the domain level can be used as an indicator to flag drugs affecting multiple biological pathways.

We further looked into the relationship between the mechanistic and functional aspects of drug polypharmacology by analysing the correlation between the number of targets (proteins or CATH-FunFams) and the number of GOBP terms a drug is associated with. There are two possible scenarios: drugs can affect multiple biological processes because they target multiple proteins (or protein domains) which are associated with specific biological processes, or they can alter several biological processes because they have targets

that are involved in multiple functions. Fig 4 shows that the latter is the case, drugs target a few proteins and CATH-FunFams, which are involved in multiple biological processes, to affect several functions. This also implies that even drugs specific to a single target can have a considerable functional polypharmacology potential.

This effect is more marked for CATH-FunFams than for proteins. This means that drugs tend to have less domain targets than protein targets, which is a consequence of the multi domain architecture of proteins (a drug that targets several proteins is associated with the common domain in these proteins). Furthermore, CATH-FunFams are annotated with more functions than proteins, which is a consequence of the limited repertoire of domains that combine to form proteins. The annotation for a single protein might be informative for one process, but the domain within that protein that binds the drug occurs in other paralogous protein contexts that are used in several other biological processes, and therefore has a richer functional annotation.

Mapped CATH-FunFams contain drug binding sites

As outlined in Materials and Methods, we mapped small molecule binding to CATH-FunFams, aiming to identify the CATH-FunFam domain that mediates small molecule binding. To evaluate our mapping, we examined the resulting set of CATH-FunFams for potential binding sites. That is, if these CATH-FunFams mediate drug action at the level of protein domains, they should contain drug-binding sites.

Out of the 70 CATH-FunFams identified as drug targets by our drug-domain association scheme (see S1 Table), only 29 have a crystal structure in PDB. For comparison, we also assembled a set of CATH-FunFams that have a crystal structure but were not among the 70 CATH-FunFams resulting from our mapping. We found that 90% of the 29 CATH-FunFams

our mapping identified as targets have cavities where binding of prodrugs or drug-like molecules is possible. Out of a set of 5054 CATH-FunFams not identified as drug targets and with defined structure, only 51.4% have cavities capable of binding drug-like molecules Thus, when comparing the set of CATH-FunFams that resulted from our mapping with all other CATH-FunFams, we found that the former has a greater proportion of CATH-FunFams with druggable cavities (p-val $= 6.2 \cdot 10^{-4}$, Fisher exact test). This suggests that the set of CATH-FunFams we identified using our mapping is enriched for potential drug targets.

The high structural and functional coherence of the CATH-FunFams enables us to explore the idea that members of a CATH-FunFam protein family are potential targets of the drug that is associated with that CATH-FunFam. We analysed four cases of drugs that: (i) have been associated with CATH-FunFams, (ii) the CATH-FunFams associated to the drug include human proteins among their members, and (iii) are present in the PDB as drug-target complexes. We selected four examples of complexes between drugs and CATH-FunFams shown in Fig 5. All 5, 8, 62 and 494 structural domains within the CATH-FunFams associated with the drugs: exemestane, epinephrine, vorinostat and acetazolamide, respectively, were pairwise aligned with SSAP and superposed. The drug-binding residues inferred onto members in each of the FunFams are highly conserved in their amino acid residue type and structural location. The mean RMSD for the aligned domain pairs across all four families is 0.64Å ± 0.62, illustrating a high structural coherence.

Despite the limited structural data, our analysis shows a high structural conservation in the binding sites of all the proteins comprising CATH-FunFam, as illustrated by the examples

shown in Fig 5. Therefore, our analysis suggests that the protein members of a CATH-FunFam are potential targets of the drug that has been associated with that CATH-FunFam.

Discussion

Drug-domain associations from drug-target data

The design of new drugs is often based on the development of molecules that are similar to previously known drugs. They are screened against a limited number of proteins selected on the basis of safety concerns and phylogenetic relationships with the primary targets relevant for a particular drug discovery project. This often results in biased drug target datasets that could affect our investigation into the application of drug polypharmacology through drug-domain associations. We show that the drug-target data compiled from ChEMBL, one of the most relevant resources of bioactivity data for drug-like molecules, is not substantially biased towards phylogenetically close proteins (see Supporting Information). Therefore, the association of drugs with protein domains is a consequence of the modular structure of proteins rather than the result of a bias of drugs to target proteins with a particular domain composition.

Mechanisms of action and modes of action

In this study, we have demonstrated that similar drugs tend to affect the same biological processes (same mode of action) and tend to target the same proteins (same mechanism of action), complying with the similarity property principle. Furthermore, when we consider CATH-FunFams as drug targets this observation is still apparent. We propose that this is due to the natural grouping of domain targets into families of evolutionary relatives sharing similar structural and functional properties. Our results suggest that the similarity property principle applies well to our drug-domain association scheme, using CATH-FunFams, and

the idea that protein domains are the drug targets within proteins. We showed that Pfam-A domains are less suitable for applying this principle. This is largely because these are much broader families which sometimes group together relatives having rather different structures and functions.

We would like to emphasize that whilst we have demonstrated that the similarity property principle applies to our set of drugs as a whole, there are many examples of structurally similar drugs that do not share similar target profiles and, vice-versa, structurally dissimilar drugs that have the same targets. Others have reported this already and, for this reason, were led to challenge the similarity property principle for approved drugs. It is difficult to compare these earlier studies and the current study directly as the former used approved and experimental drugs and target annotations from DrugBank [17], while in this study, we examine approved drugs and target annotations derived from experimental data in ChEMBL. Moreover, while we use the full structure of drugs for our pairwise similarity analysis and a statistically significant Tc similarity threshold, previous studies transformed drug structures to scaffolds and evaluated molecular similarity in terms of matching substructures and topological equivalence. Finally, we evaluate the similarity of target profiles using association indices while other studies use another statistic, AOF, to measure the similarity of target profiles. All these factors may contribute to the differences in our observations and conclusions. Nevertheless, with regard to generality, we provide in this study a strong case for the similarity property principle, which we evaluate on the level of individual proteins, and evolutionary meaningful groupings based on CATH-FunFams and Pfam-A domains.

Drug promiscuity and polypharmacology

Our analysis shows that drug promiscuity is greatest on the level of individual proteins, and much reduced on the level of protein domains. This is not unexpected, because the three-dimensional structure of protein domains is largely conserved, providing a shared recognition element for small molecule binding. We have previously observed this using a probabilistic drug-domain mapping [33], and using a heuristic small molecule-domain mapping [32]. In this study, we further show that FunFams provide a useful abstraction to rationalise drug promiscuity. This is because, as mentioned already, CATH-FunFams cluster together proteins sharing similar sequence patterns reflecting similar structures and functions. Structural and functional similarity of the domains within a CATH-FunFam tends to be preserved even when the domain occurs in different multi-domain contexts (i.e. different proteins). Hence, CATH-FunFams capture the multiple biological functions affected by promiscuous drugs.

Nevertheless, we find that 41% of drugs keep their multi-target behaviour even when the target considered is a protein domain, as in the case of the CATH-FunFams (see S1 Table). Roughly half of them (47% of drugs with more than one CATH-FunFam target) are associated with two or three CATH-FunFams that belong to the same CATH superfamily. This may be due to the conservative nature of clustering domain sequences into CATH-FunFams. The functional classification protocol used to cluster sequences into CATH-FunFams applies rather cautious and generic thresholds regardless of the superfamily and this can sometimes result in relatives with very similar functions being assigned to separate CATH-FunFams in some superfamilies. In such cases where the functions of a CATH-FunFam can be defined in more specific terms, it seems clear that the over-splitting of a functional family might overlook its drug binding function.

CATH-FunFams as drug targets

We have provided fundamental support to the idea suggested by previous research that protein domains provide a useful level of abstraction for a systematic understanding of small molecule bioactivity and drug action [16,31-33,39,40]. We examined further this idea to test whether protein domains can be druggable and we show in this work that CATH-FunFams have the potential to be the druggable entities within drug targets. Moreover, because of their high structural and functional coherence, our assessment of the druggability of CATH-FunFams suggest that the relatives within a CATH-FunFam are potential targets of the drugs that are associated to that CATH-FunFam.

In summary, our work supports the idea that drug protein interactions are mediated by drug-domain interactions. The identification of CATH-FunFams as a reasonable annotation level for drug-target interactions opens a new research direction in target identification with potential application in drug repurposing.

Materials and Methods

Gathering drug-target datasets

We compiled an initial drug-protein target dataset with 531 drugs and 557 human targets by querying ChEMBL release 20. ChEMBL allows us to define the drug-target relationship based on the concentration at which the compound affects the target, providing us a way to restrict our dataset to biologically meaningful drug-protein associations. We considered a drug as a small molecule with therapeutic application (THERAPEUTIC_FLAG =1), not currently known to be a pro-drug, reporting a direct binding interaction with single protein (ASSAY_TYPE = 'B'; RELATIONSHIP_TYPE = 'D'; TARGET_TYPE = 'SINGLE PROTEIN'), with a maximum phase of development reached for the compound of 4 (meaning an approved drug). In order to exclude non-specific interactions between small molecules and

biological targets we impose the filter that the activity against a human protein target should be stronger than 1 µM, where activity includes IC50, EC50, XC50, AC50, Ki, Kd, pchembl_value ≥ 6 [41].

All data processing, statistics analysis and results plots were produced using the R computing environment [42] and the R library ggplot2 [43].

Domain family resources

We used two different definitions of protein domains: Pfam-A domains from Pfam release 27.0 [41], and CATH-FunFams from CATH-Gene3D v12.0 [26,42]. Pfam-A entries in the Pfam database are based on manually curated sequence alignments and can be used to recognise family members even for remote phylogenetic relationships. CATH-Gene3D is a large collection of CATH [24] domain predictions for genome sequences ~20 million [43]. CATH is a protein domain classification system that makes use of a combination of manual and automated structure- and sequence-based procedures to decompose proteins into their constituent domains and then classify these domains into homologous superfamilies (groups of domains that are related by evolution); domain regions in CATH are more clearly defined than in other domain resources by the use of structural data which is more highly conserved than the sequence. CATH superfamilies map to at least 60% of predicted domain sequences in completed genomes using in-house HMM protocols –and as high as 70-80% if more sophisticated threading-based protocols are used [44]. CATH-Gene3D is the starting point to derive functionally coherent families by clustering domain sequences within a CATH superfamily using an in-house protocol [26]. The most recent version of this method (FunFHMMer) distinguishes functional families on the basis of differences in specificity determining residues [45]. CATH-FunFams group together relatives likely to

have highly similar structures [46]and functions and have been highly ranked in the CAFA international Critical Assessment of Functional Annotation [47].

We used the heuristic developed by Kruger et al. [32] to map drugs in ChEMBL to protein domains, to obtain our drug-domain datasets from the drug-protein data.

Molecular similarity calculation

We retrieved the chemical table representing the chemical structure record of 2015 approved drugs (regardless of their targets) from ChEMBL release 20 and we obtained their MACCS molecular fingerprints. We computed each pairwise Tanimoto similarity coefficients (Tc) using the RDKit package RDKit: Cheminformatics and Machine Learning Software; the Tc similarity quantifies the fraction of features common to the molecular fingerprints of the pair of drugs to the total number of features of the molecular fingerprints of each drug in the pair [48].

We performed a significance analysis of the molecular similarity for our set of multi-target drugs, in order to choose a threshold Tc which will define a statistically significant level of similarity between any pair of drugs in our dataset. Based on the shape of the Tc curves (see S1 Fig), we assumed that the data fit a normal distribution. For all the drugs we could gather from ChEMBL (2015 drugs), we computed the Tc similarity between each drug and the remaining 2014 drugs. From these distributions of Tc values, we extracted the cumulative distribution function $F(t)$ that gives the probability of having a similarity less or equal than a given Tc value. A significance level (p-value) defined as $p = 1 - F(t)$ was assigned to every drug for each Tc value, according to Maggiora et al. [14].

Measuring pairwise associations of drug interaction profiles

We used the Python module Networkx to transform each dataset into a bipartite graph that connects drugs with targets (protein or domains) in order to compute the similarity of the interaction profiles of any pair of drugs. For each drug in the bipartite graph, its interaction profile is the set of targets (proteins or domains) the drug is linked to. We analysed the interaction profile similarity between two drugs by means of the Jaccard (J_{ab}) association indices, which measure the similarity between the interaction profiles of each pair of drugs [49], defined as:

$$J_{ab} = \frac{n_a \cap n_b}{n_a \cup n_b}$$

where n_a is the set of targets of drug a and n_b is the set of targets of drug b.

GO Functional annotations

We produced two datasets of drug-GOBP associations corresponding to each type of target with available GO annotation, by inheriting the GOBP annotations of proteins and CATH-FunFams to the drugs associated with them. We extracted GOBP terms for protein targets from the Gene Ontology Annotation (UniProt-GOA) Database [50]. To ensure we were using high quality annotations we restricted to manually curated annotations derived from experimental evidence in published scientific literature. Protein relatives within a functional family (CATH-FunFam) are likely to share highly similar functions. Therefore, CATH-FunFams are annotated with GO terms probabilistically in order to ensure their functional coherence. We obtained the most significant GOBP terms annotated to CATH-FunFams (Benjamini-Hochberg FDR corrected p-val 0.05).

The semantic similarity among GOBP terms was evaluated with the graph-based algorithm described in [51] and implemented in the R package GOSemSim [52]. Two GOBP terms were defined to be different if their semantic similarity was below 0.4.

Druggability screening

We used the Fpocket platform [53] to detect cavities in the structure of selected domains that can bind drug-like molecules. Fpocket is a fast protein pocket prediction algorithm that identifies cavities on the surface of proteins and ranks them according to their ability to bind drug-like small molecules. Thus, Fpocket assesses the ability of a given binding site to host drug-like organic molecules in terms of a druggability scoring function described in [54]. Fpocket is released under an open source license at fpocket.sourceforge.net.

Structural alignment

To explore whether CATH-FunFams associated with drug binding consist of members with a similar binding pocket and similar amino acid residues, we looked in detail at four CATH-FunFams associated with binding the compounds acetazolamide, epinephrine, exemestane, and vorinostat. Structural domains from these four different CATH-FunFams were pairwise structurally aligned using SSAP [55]. SSAP scores were used to construct a distance matrix and maximum spanning tree which was then used to derive a multiple superposition of the structural relatives. Data on residues involved in binding each of the drugs of interest were extracted from the NCBI IBIS resource [56] using the following PDB IDs as queries: 3ML5 for acetazolamide; 4LDOA for epinephrine; 3S7S for exemestane; 4LXZ for vorinostat. These four PDB IDs were chosen as they were the only PDBs in each CATH-FunFam with drug binding information. These drug-binding residue positions were mapped onto the other structural domains using the SSAP alignment data. When producing the figures in PyMOL (www.pymol.org), the number of redundant structural domains in the acetazolamide and vorinostat alignments was reduced to improve clarity.

Acknowledgments

Aurelio A. Moya-Garcia has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme (FP7/2007-2013) under REA Grant Agreement No 623543.

Natalie L. Dawson acknowledges funding from the Wellcome Trust (Award number: 104960/Z/14/Z).

Juan A.G. Ranea was funded by EU-FP7-Systems Microscopy NoE (Grant Agreement 258068), SAF2012-33110 and CTS-486 (Spanish Ministry of Economy and Competitiveness, Andalusian Government and Fondos Europeos de Desarrollo Regional). The CIBERER is an initiative of the Carlos III Health Institute.

The authors thank Dr. Ian Sillitoe and Dr. Tony E. Lewis from the Orengo Group at UCL for their valuable help in obtaining and analysing the structural data; and Dr. Ian Morilla from the Laboratoire Analyse Géométrie et Applications (LAGA), Université Paris 13 - Sorbonne Paris Cité for his help with the statistics analyses.

References

1. Medina MA. Systems biology for molecular life sciences and its impact in biomedicine. Cell Mol Life Sci. 2012;70: 1035–1053. doi:10.1007/s00018-012-1109-z

2. Ranea JAG, Morilla I, Lees JG, Reid AJ, Yeats C, Clegg AB, et al. Finding the “dark matter” in human and yeast protein network prediction and modelling. Rzhetsky A, editor. PLoS Comput Biol. 2010;6: e1000945. doi:10.1371/journal.pcbi.1000945

3. Zhao S, Iyengar R. Systems Pharmacology: Network Analysis to Identify Multiscale Mechanisms of Drug Action. Annu Rev Pharmacol Toxicol. 2012;52: 505–521. doi:10.1146/annurev-pharmtox-010611-134520

4. Berg EL. Systems biology in drug discovery and development. Drug Discov Today. 2014;19: 113–125. doi:10.1016/j.drudis.2013.10.003

5. Ma'ayan A, Rouillard AD, Clark NR, Wang Z, Duan Q, Kou Y. Lean Big Data integration in systemsbiology and systems pharmacology. Trends in Pharmacological Sciences. Elsevier Ltd; 2014;35: 450–460. doi:10.1016/j.tips.2014.07.001

6. Anighoro A, Bajorath J, Rastelli G. Polypharmacology: Challenges and Opportunities in Drug Discovery. J Med Chem. 2014. doi:10.1021/jm5006463

7. Moya-García AA, Morilla I, Ranea JAG. Oncogenic Signalling Networks and Polypharmacology as Paradigms to Cope with Cancer Heterogeneity. Current Proteomics. 2014;11: 1–8.

8. Yildirim MA, Goh K-I, Cusick ME, Barabási A-L, Vidal M. Drug-target network. Nat Biotechnol. 2007;25: 1119–1126. doi:10.1038/nbt1338

9. Mestres J, Gregori-Puigjané E, Valverde S, Solé RV. Data completeness--the Achilles heel of drug-target networks. Nat Biotechnol. 2008;26: 983–984. doi:10.1038/nbt0908-983

10. Hu Y, Bajorath J. Polypharmacology directed compound data mining: identification of promiscuous chemotypes with different activity profiles and comparison to approved drugs. J Chem Inf Model. 2010;50: 2112–2118. doi:10.1021/ci1003637

11. Hopkins AL. Network pharmacology: the next paradigm in drug discovery. Nat Chem Biol. 2008;4: 682–690. doi:10.1038/nchembio.118

12. Geppert T, Koeppen H. Biological networks and drug discovery-where do we stand? Drug Dev Res. 2014;75: 271–282. doi:10.1002/ddr.21207

13. Hopkins AL, Mason JS, Overington JP. Can we rationally design promiscuous drugs? Curr Opin Struct Biol. 2006;16: 127–136. doi:10.1016/j.sbi.2006.01.013

14. Maggiora G, Vogt M, Stumpfe D, Bajorath J. Molecular similarity in medicinal chemistry. J Med Chem. 2014;57: 3186–3204. doi:10.1021/jm401411z

15. Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, Shoichet BK. Relating protein pharmacology by ligand chemistry. Nat Biotechnol. Nature Publishing Group; 2007;25: 197–206. doi:10.1038/nbt1284

16. Yamanishi Y, Pauwels E, Saigo H, Stoven V. Extracting Sets of Chemical Substructures and Protein Domains Governing Drug-Target Interactions. J Chem Inf Model. 2011. doi:10.1021/ci100476q

17. Hu Y, Bajorath J. Many structurally related drugs bind different targets whereas distinct drugs display significant target overlap. RSC Advances. Royal Society of Chemistry; 2012;2: 3481–3489.

18. Hu Y, Bajorath J. Rationalizing structure and target relationships between current drugs. AAPS J. 2012;14: 764–771. doi:10.1208/s12248-012-9392-z

19. Rask-Andersen M, Almén MS, Schiöth HB. Trends in the exploitation of novel drug targets. Nat Rev Drug Discov. 2011;10: 579–590. doi:10.1038/nrd3478

20. Chothia C, Gough J, Vogel C, Teichmann SA. Evolution of the protein repertoire. Science. American Association for the Advancement of Science; 2003;300: 1701–1703. doi:10.1126/science.1085371

21. Apic G, Gough J, Teichmann SA. An insight into domain combinations. Bioinformatics. 2001;17 Suppl 1: S83–9.

22. Koonin EV, Wolf YI, Karev GP. The structure of the protein universe and genome evolution. Nature. 2002;420: 218–223. doi:10.1038/nature01256

23. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol. 1995;247: 536–540. doi:10.1006/jmbi.1995.0159

24. Sillitoe I, Lewis TE, Cuff A, Das S, Ashford P, Dawson NL, et al. CATH: comprehensive structural and functional annotations for genome sequences. Nucleic Acids Res. Oxford University Press; 2015;43: D376–D381. doi:10.1093/nar/gku947

25. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic Acids Res. 2014;42: D222–30. doi:10.1093/nar/gkt1223

26. Rentzsch R, Orengo CA. Protein function prediction using domain families. BMC Bioinformatics. BioMed Central Ltd; 2013;14: S5. doi:10.1186/1471-2105-14-S3-S5

27. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH – a hierarchic classification of protein domain structures. 1997;5: 1093–1108.

28. Wolf YI, Grishin NV, Koonin EV. Estimating the number of protein folds and families from complete genome data. J Mol Biol. 2000;299: 897–905. doi:10.1006/jmbi.2000.3786

29. Kummerfeld SK, Teichmann SA. Protein domain organisation: adding order. BMC Bioinformatics. 2009;10: 39. doi:10.1186/1471-2105-10-39

30. Pang E, Tan T, Lin K. Promiscuous domains: facilitating stability of the yeast protein-protein interaction network. Mol Biosyst. 2012;8: 766–771. doi:10.1039/c1mb05364g

31. Wang X, Wei X, Thijssen B, Das J, Lipkin SM, Yu H. Three-dimensional reconstruction of protein networks provides insight into human genetic disease. Nat Biotechnol. Nature Publishing Group; 2012;30: 159–164. doi:10.1038/nbt.2106

32. Kruger FA, Rostom R, Overington JP. Mapping small molecule binding data to structural domains. BMC Bioinformatics. 2012;13 Suppl 17: S11. doi:10.1186/1471-2105-13-S17-S11

33. Moya-García AA, Ranea JAG. Insights into polypharmacology from drug-domain associations. Bioinformatics. 2013;29: 1934–1937. doi:10.1093/bioinformatics/btt321

34. Hopkins AL, Groom CR. The druggable genome. Nat Rev Drug Discov. 2002;1: 727–730. doi:10.1038/nrd892

35. Croset S, Overington JP, Rebholz-Schuhmann D. The functional therapeutic chemical classification system. Bioinformatics. 2014;30: 876–883.

doi:10.1093/bioinformatics/btt628

36. Petrone PM, Simms B, Nigsch F, Lounkine E, Kutchukian P, Cornett A, et al. Rethinking molecular similarity: comparing compounds on the basis of biological activity. ACS Chem Biol. 2012;7: 1399–1409. doi:10.1021/cb3001028

37. Gene Ontology Consortium. Gene Ontology Consortium: going forward. Nucleic Acids Res. Oxford University Press; 2015;43: D1049–56. doi:10.1093/nar/gku1179

38. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25: 25–29. doi:10.1038/75556

39. Kruger FA, Gaulton A, Nowotka M, Overington JP. PPDMs-a resource for mapping small molecule bioactivities from ChEMBL to Pfam-A protein domains. Bioinformatics. 2014. doi:10.1093/bioinformatics/btu711

40. Pardo EP, Godzik A. Analysis of individual protein regions provides novel insights on cancer pharmacogenomics. PLoS Comput Biol. 2015;11: e1004024. doi:10.1371/journal.pcbi.1004024

41. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic Acids Research. 2014;42: D222–D230.

42. Lee DA, Rentzsch R, Orengo C. GeMMA: functional subfamily classification within superfamilies of predicted protein structural domains. Nucleic Acids Res. 2010;38: 720–737. doi:10.1093/nar/gkp1049

43. Lees J, Yeats C, Redfern O, Clegg A, Orengo C. Gene3D: merging structure and function for a Thousand genomes. Nucleic Acids Research. Oxford University Press; 2010;38: D296–D300. doi:10.1093/nar/gkp987

44. Lees J, Yeats C, Perkins J, Sillitoe I, Rentzsch R, Dessailly BH, et al. Gene3D: a domain-based resource for comparative genomics, functional annotation and protein network analysis. Nucleic Acids Res. Oxford University Press; 2012;40: D465–71. doi:10.1093/nar/gkr1181

45. Das S, Lee D, Sillitoe I, Dawson NL, Lees JG, Orengo CA. Functional classification of CATH superfamilies: a domain-based approach for protein function annotation. Bioinformatics. 2015;31: 3460–3467. doi:10.1093/bioinformatics/btv398

46. Dessailly BH, Dawson NL, Mizuguchi K, Orengo CA. Functional site plasticity in domain superfamilies. Biochim Biophys Acta. 2013;1834: 874–889. doi:10.1016/j.bbapap.2013.02.042

47. Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, Sokolov A, et al. A large-scale evaluation of computational protein function prediction. Nat Methods. 2013;10: 221–227. doi:10.1038/nmeth.2340

48. Willett P, Barnard JM, Downs GM. Chemical Similarity Searching. J Chem Inf Model. American Chemical Society; 1998;38: 983–996. doi:10.1021/ci9800211

49. Fuxman Bass JI, Diallo A, Nelson J, Soto JM, Myers CL, Walhout AJM. Using networks to measure similarity between genes: association index selection. Nat Methods. 2013;10: 1169–1176. doi:10.1038/nmeth.2728

50. Huntley RP, Sawford T, Mutowo-Meullenet P, Shypitsyna A, Bonilla C, Martin MJ, et al. The GOA database: gene Ontology annotation updates for 2015. Nucleic Acids Res. Oxford University Press; 2015;43: D1057–63. doi:10.1093/nar/gku1113

51. Wang JZ, Du Z, Payattakool R, Yu PS, Chen C-F. A new method to measure the semantic similarity of GO terms. Bioinformatics. 2007;23: 1274–1281. doi:10.1093/bioinformatics/btm087

52. Yu G, Li F, Qin Y, Bo X, Wu Y, Wang S. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. Bioinformatics. Oxford University Press; 2010;26: 976–978. doi:10.1093/bioinformatics/btq064

53. Le Guilloux V, Schmidtke P, Tuffery P. Fpocket: an open source platform for ligand pocket detection. BMC Bioinformatics. BioMed Central Ltd; 2009;10: 168. doi:10.1186/1471-2105-10-168

54. Schmidtke P, Barril X. Understanding and predicting druggability. A high-throughput method for detection of drug binding sites. J Med Chem. 2010;53: 5858–5867. doi:10.1021/jm100574m

55. Taylor WR, Orengo CA. Protein structure alignment. Journal of molecular biology. 1989;208: 1–22. doi:10.1016/0022-2836(89)90084-3

56. Shoemaker BA, Zhang D, Tyagi M, Thangudu RR, Fong JH, Marchler-Bauer A, et al. IBIS (Inferred Biomolecular Interaction Server) reports, predicts and integrates multiple types of conserved interactions for proteins. Nucleic Acids Res. 2012;40: D834–40. doi:10.1093/nar/gkr997

Figures

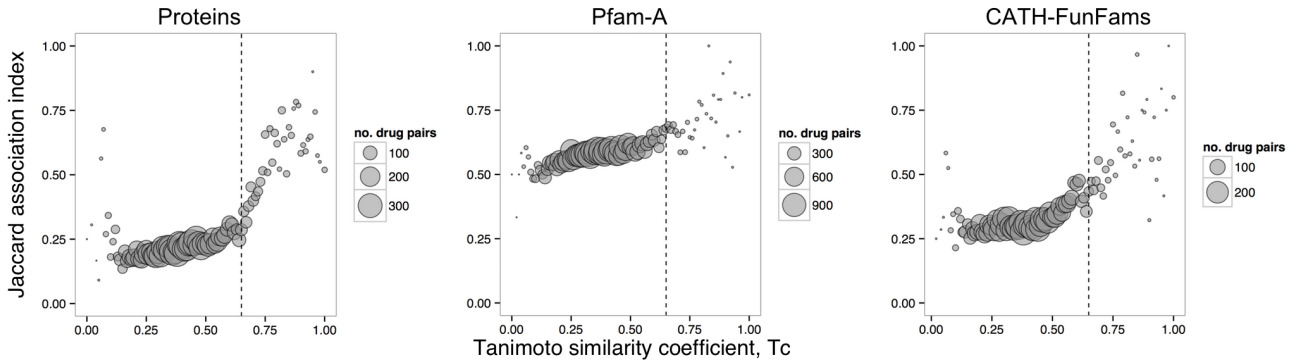

Fig 1. Correlation of drugs interactions profiles with drugs molecular similarity. Each circle is the average Jaccard index for the three drug-target datasets at a given bin of Tc similarity (bin size 0.01). The size of the circles is proportional to the number of drug pairs in the corresponding Tc bin. The vertical dashed line indicates the drug similarity threshold, Tc = 0.65 (see Supporting Information).

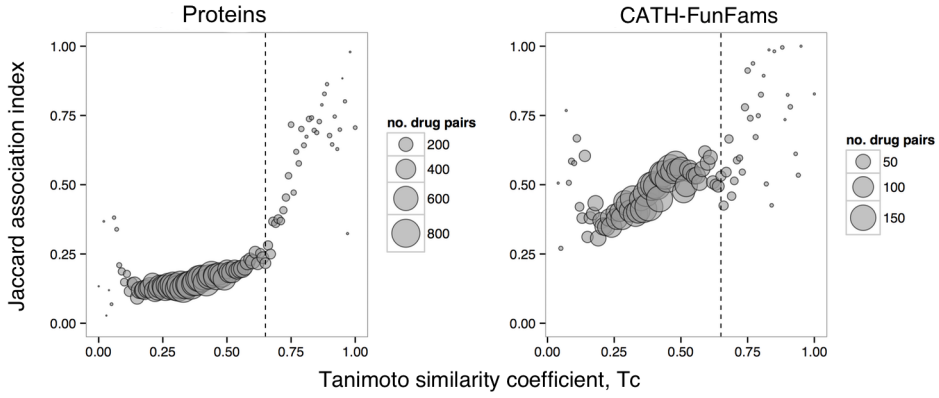

Fig 2. Correlation of drugs inherited GO annotations with drugs molecular similarity. Each circle is the average Jaccard index for the inherited GOBP annotations derived from proteins (left panel) and CATH-FunFams (right panel) at a given bin of Tc similarity (bin size 0.01). The size of the circles is proportional to the number of drug pairs in the corresponding Tc bin.

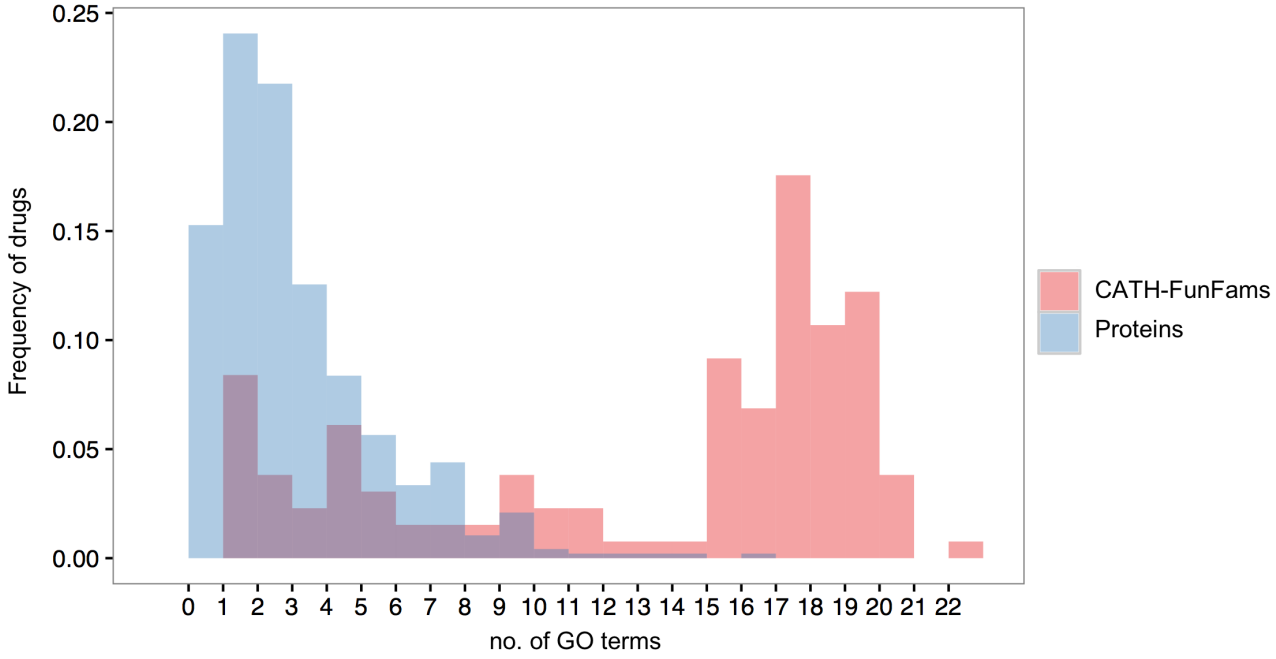

Fig 3. Drug polypharmacology potential of approved drugs. Distribution of drug polypharmacology potential for the two types of targets with GO annotations measured as the frequency of the number of GOBP terms associated to each drug through drug-protein mapping and drug-CATH-FunFam mapping. The vertical dashed line indicates the drug similarity threshold, Tc = 0.65 (see Supporting information).

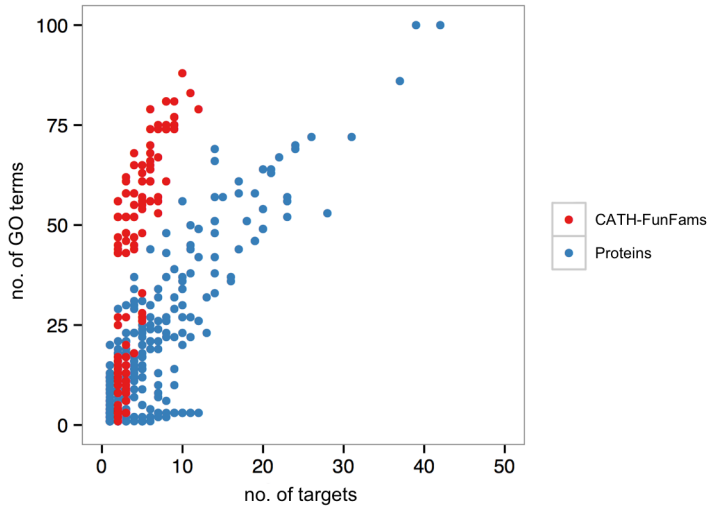

Fig 4. Correlation between mechanistic and functional polypharmacology. The number of GOBP terms of each drug is plotted as a function of its number of targets, for CATH-FunFam and protein targets.

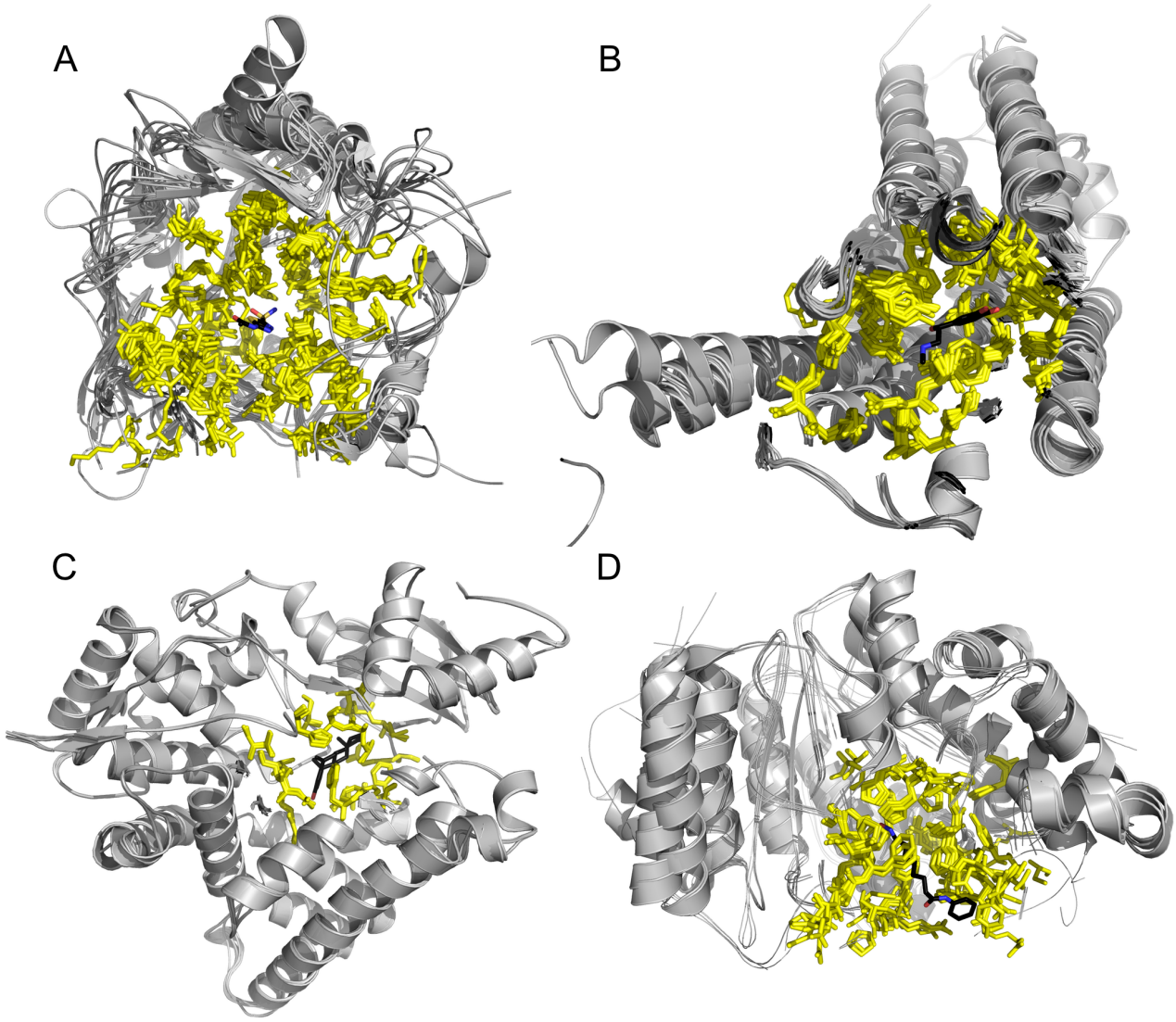

Fig 5. Conservation of the binding sites within CATH-FunFams. Structural alignment of the CATH-FunFams associated with: A) acetazolamide (CATH ID: 3.10.200.10-FF1430), B) epinephrine (CATH ID: 1.20.1070.10-FF44570), C) exemestane (CATH ID: 1.10.630.30-

FF29451) and D) vorinostat (CATH ID: 3.40.800.20-FF2860), and the drug-target complexes of these four drugs. The protein domains are all in grey except for the ligand binding residues, which have been mapped across the domains, coloured yellow. The drug molecules are in black.

Supporting Information

S1 Appendix. Additional analysis performed on the drug-protein and drug-domain datasets.

S1 Table. Topology parameters of the bipartite drug-target graphs.

S1 Fig. Threshold Tanimoto similarity of approved drugs.

S2 Fig. Structural variability of drugs and bioactive compounds.

S3 Fig. Fraction of drug targets in each k-core of the domain co-occurrence networks.

S4 Fig. Association indices with MACCS fingerprints.

S5 Fig. Association indices with ECFP4 fingerprints.