How reliable are ligand-centric methods for Target Fishing?

Antonio Peón, Cuong C. Dang and Pedro J. Ballester*

Cancer Research Center of Marseille, INSERM U1068, F-13009 Marseille, France; Institut Paoli-Calmettes, F-13009 Marseille, France; Aix-Marseille Université, F-13284 Marseille, France; and CNRS UMR7258, F-13009 Marseille, France

Correspondence to: pedro.ballester@inserm.fr

Computational methods for Target Fishing permit the discovery of new targets of a drug, which may result in its reposition in a new indication or improving our current understanding of its efficacy and side effects. Being a relatively recent class of methods, there is still a need to improve their validation, which is technically difficult, often limited to a small part of the targets and not easily interpretable by the user. Here we propose a new validation approach and use it to assess the reliability of ligand-centric techniques, which by construction provide the widest coverage of the proteome. On average over approved drugs, we find that only five predicted targets will have to be tested in order to find at least two true targets with submicromolar potency, although a strong variability in performance is observed. Also, we identify an average of eight known targets in approved drugs, which suggests that polypharmacology is a common and strong event. In addition, we observe that many known targets of approved drugs are currently missed by these methods. Lastly, by using a control group of randomly-selected molecules, we discuss how the data generation process confounds this analysis and its implications to method validation.

1. Introduction

Target Fishing, also known as Target Prediction, consists in predicting the macromolecular targets of a query molecule. This problem is the reverse of Virtual Screening[1], where the goal is to predict the ligands of a query target. Computational methods for Target Fishing are of great interest, as identifying previously unknown targets of a molecule is the basis of a number of important drug design and chemical biology applications. Indeed, discovering a new target in a drug could lead to its reposition in a new indication as well as an enhanced understanding of its efficacy and side-effects. Furthermore, these tools can be used for target deconvolution of phenotypic screening hits[2], which is a prerequisite to gain mechanistic understanding of phenotypic activity and helpful for drug development. This two-stage process, phenotypic screening followed by target deconvolution, constitutes an attractive alternative strategy for the discovery of molecularly targeted therapies.

The fast growth of freely-available bioactivity resources (e.g. PubChem[3] or ChEMBL[4]) has sparked a new generation of powerful data-driven methods for Target Fishing. This growth is exemplified by the ChEMBL database[5], which in a few years has assembled and fully curated chemical structures and bioactivities from more than 50,000 scientific publications. Moreover, this database is periodically updated and will eventually incorporate a flood of new data that is being extracted from the patent literature (https://www.surechembl.org/search/).

There are two major classes of methods for Target Fishing: target-centric and ligand-centric. Target-centric methods build a predictive model for each target, which is used to determine whether the query molecule has activity against the target. Thereafter, the query molecule is evaluated by this panel of models to provide its set of predicted targets. Each method adopt a

particular model type, e.g. Quantitative Structure–Activity Relationship (QSAR)[6], chemical similarity to an ensemble of cognate ligands[7,8] or structure-based[9] models. On the other hand, ligand-centric methods are based on calculating the similarity of a very large number of target-annotated molecules to the query molecule. These molecular similarities are also of different types, e.g. Two-Dimensional (2D)[10] based on Morgan fingerprints[11] or Three-Dimensional (3D)[9,12] based on Ultrafast Shape Recognition (USR)[13] variants.

Both classes of methods are complementary. In cases where the targets of interest are known to have many ligands, more accurate target-centric models will be possible and thus these tools are likely to be more suitable. By contrast, in cases where evaluating as many targets as possible is preferable, ligand-centric tools will be more suitable, as these provide the widest coverage of the proteome. Indeed, whereas ligand-centric methods can interrogate any target that has at least one known ligand, target-centric models can only evaluate the much smaller set of targets for which a model can be built. For instance, building a structure-based model usually requires the availability of a suitable X-ray crystal structure for the target and methods based on QSAR/similarity-ensemble can only build a model in those targets with a sufficiently high number of known ligands.

It has been argued that the difficult validation of Target Fishing methods hinders comparisons of the performance of each method.[14] Furthermore, these comparisons are often restricted to a few tens of information-rich targets, using benchmarks borrowed from Virtual Screening, and thus tell us very little about how well the methods perform on the remaining thousands of targets. This common preference is probably due to the technical difficulties associated to mining large-scale chemogenomics data for benchmark design. More importantly, Target Fishing is often posed as a

single classification problem (whether the combination of any target and any ligand is active or not), which is not well suited to analyse how performance changes depending on the query molecule. Furthermore, commonly applied performance measures, such as the ROC (Receiver Operating Characteristic) AUC (Area Under Curve), do not precisely answer the question of how well the method will work prospectively. For example, how do you estimate how many true targets of a query molecule you are likely to find if the validation shows that the average ROC AUC over 40 targets is 0.7? As a result, these measures do not offer guidance on pragmatic questions such as how many predicted targets have to be tested on average to find a true target or how many known targets of the query molecule are typically missed.

In this study, we propose a new approach to validating methods for Target Fishing, which naturally lends itself to answer such questions. This approach is based on formulating an identical classification problem for each query molecule. From this new perspective, we provide a lower-bound for the current performance of ligand-centric methods with which to analyse the minimum that can be expected nowadays from these methodologies. As a byproduct, our analysis gives an update for the degree of polypharmacology observed in approved drugs. The rest of the paper is organised as follows. Section 2 describes the experimental setup, including data selection, data partitions, target fishing method and performance metrics. Section 3 discusses the results. Section 4 presents the conclusions.

2. Experimental setup

This section describes the setup of all the numerical experiments carried out in this study. This setup is composed of the following elements: data selection, data partitions, target-fishing method

and measures of predictive performance. All molecular data processing is done with Structured Query Language (SQL) queries from Python 2.7.9 on a local copy of the ChEMBL database running PostgreSQL 9.4.3, with molecular similarity searches using in addition the RDKit PostgreSQL cartridge (2015.03.1 release).

2.1. *Data selection*

The first step is constructing datasets from the ChEMBL database[4]. We started by downloading release 20 as a PostgreSQL dump[15], which contains data for 10,774 targets, 1,456,020 molecules with disclosed chemical structure and 13,520,737 bioactivities.

Single-protein was the most common target type (6,018 of the 10,774 targets). In order to provide the most specific target prediction, we restricted to single-protein targets, which incidentally constitutes the largest molecular target type in the database (the 'protein complex', 'protein family' and 'nucleic-acid' types only have 261, 217 and 29 targets respectively). The remaining general constraints were confidence_score = 9 (i.e. direct single-protein target assigned by the data curator), activities.published_relation = '=', assay_type = 'B' and standard_units = 'nM'. As a result of this process, 888,354 molecules were found to be associated to the 6,018 single-protein targets through 4,871,527 bioactivities.

Further requirements are commonly imposed for the measured bioactivity of a ligand against a target to be counted as a known target for that ligand. First, the bioactivity measurement must be of relatively high quality, activities.standard_type IN ('EC50','Ki','Kd','IC50'), which discards percentages of inhibition among other lower-quality measurements. Second, only target-ligand complexes with a sufficiently potent bioactivity are retained (common activity thresholds are 1μM

and 10μM meaning that a ligand hitting any target with an activity higher than 10μM will not be considered to be a target in neither of these two scenarios). Third, only targets with at least n qualifying ligands are considered. For target-centric methods, a sufficiently high number of ligands are requested to build a model for the target, e.g. those methods based on similarity-ensemble approaches (n=5)[7] or classical QSAR (n=40)[6]. In this study, we analyse ligand-centric methods, which can evaluate any target with at least a known ligand (i.e. n=1), with two common activity thresholds (the more potent is this threshold, the fewer targets will be included in the search). Thus, this study exclusively analyses the first two rows of Table 1.

Table 1. Dataset size depending on modelling constraints.

Dataset	Only targets with at least	#Targets
Thres1μM	1 ligand	2,592
Thres10μM	1 ligand	3,046
Thres1μM	5 ligands	1,788
Thres10μM	5 ligands	2,158
Thres1μM	40 ligands	725
Thres10μM	40 ligands	917

This table also shows the number of targets that would be searched by methods requiring n=5 and n=40. For example, a method that builds models requiring at least 40 ligands per target and defining a target with an activity threshold of 10μM would be predicting whether the query molecule has activity against any of the 917 available single-protein targets. In contrast, a ligand-centric method with the same activity threshold will be able to evaluate 2,109 additional targets, for which the first method is unable to provide any prediction by construction. Of course, the advantage of target-centric over ligand-centric methods is that the former will tend to perform

better on those targets with a higher number of ligands, which highlights the complementarity of both approaches.

2.2. *Data partitions*

Next, we partition each of the two n=1 datasets as follows. First, we identify the subset of approved drugs. Second, we search for all those approved drugs in the ChEMBL database meeting the criteria introduced in the previous section and hitting any of the targets in Table 1 (these are 620 drugs in the Thres1μM set and 750 drugs in the Thres10μM set). These are the two approved-drugs sets of query molecules. Third, we pick at random two further sets of molecules of the same size, which we called random-molecules sets. This will serve as a control group to understand how target predictions for marketed drugs differ from those made for other types of molecules. The rest of ligands form the set of database molecules, which is the same for both sets of query molecules but different between thresholds (this is shown in Table 2).

Table 2. Dataset size depending on modelling constraints.

ID	Dataset	#query-molecules	#database-molecules
A	approved-drugs_Thres10μM	750	184,163
B	random-molecules_Thres10μM	750	184,163
C	approved-drugs_Thres1μM	620	147,751
D	random-molecules_Thres1μM	620	147,751

2.3. *A simple method for Target Fishing*

For our analysis, we selected a standard 2D chemical similarity search[14] in order to obtain a lower-bound for the performance of ligand-centric target-fishing methods. This is a two-step procedure starting with the generation of MACCS fingerprints[17] for all query and database molecules in Table 2. For each molecule, MACCS (for Molecular ACCess System) codifies 166 predetermined

chemical groups as a binary string of the same size. These fingerprints were generated using the RDKit PostgreSQL cartridge[18] and inserted in the database for its use.

As usual, fingerprint generation could not be generated for a few unusual molecules and consequently queries could not be performed for these (e.g. 5 out of 750 in approved-drugs_Thres10μM). This is the case of Gramidicin (CHEMBL1201469), which is actually not a molecule but a mixture of three antibiotic compounds. Other examples are some organometallic compounds such as the anti-rheumatic agent Auranofin (CHEMBL1366).

Using their MACCS fingerprints, the Dice score was used to measure the similarity between a query molecule and all the database molecules. The Dice score is defined as:

$$Dice = 2c/(a + b) \quad (1)$$

where a is the number of on bits in molecule A, b is number of on bits in molecule B, while c is the number of bits that are on at the same positions in both molecules. For each query, the top k hits can be identified from the corresponding ranking of database molecules (these are the k database molecules with the most similar chemical structure to that of the query molecule). We consider here k = 1, 5, 10 and 15 to investigate the dependence of the method with its only control parameter k.

Finally, the known targets for the k hits are retrieved from the ChEMBL database and returned as predicted targets for the considered query molecule. Thus, a set of predicted targets is obtained for each combination of query molecule and k value. Note that the known targets are not any target annotated in the ChEMBL database, but those that comply with the requirements set in section 2.1. for each of the four cases in Table 2.

2.4. *Measuring predictive performance*

Each performed query can be posed as a separate classification problem. For validation purposes, the known targets of the query molecule are taken as a ground truth. Thus, we assume that the known targets are all the qualifying targets of the molecule, whereas the rest of considered targets are non-targets for that molecule. However, as the query molecule has only been tested against less than 0.1% of the ChEMBL targets on average, it is expected that many unconfirmed targets, especially those coming from molecules similar to the query molecule, would be actually targets if only these could be thoroughly tested. As a result, any positive prediction on a non-target will have to be rejected as false, despite an unknown part of these being true targets of the molecule. Therefore, we must keep in mind that we will be looking at a lower-bound for performance also in this sense.

Table 3 shows the confusion matrix arising from assessing target predictions against experimental evidence for each query molecule. After the assessment, each target prediction can be classed in one of four categories: TP for True Positive (the predicted target is a known target); TN for True Negative (the target was not predicted but anyway is not known to be a target); FP for False Positive (the predicted target is not known to be a target, i.e. a false discovery or Type I error); and FN for False Negative (the target was not predicted and it is actually a target, i.e. missed discovery or Type II error).

Table 3. Confusion matrix arising from assessing target predictions against experimental evidence for each query molecule**.**

Target		Predicted	Non-predicted
Yes	(experimentally tested)	TP	FN
No	(not tested/tested)	FP	TN

From these quantities, we will calculate four performance measures per query molecule. The Accuracy (ACC) is the proportion of correct target predictions:

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{2}$$

The Positive Predictive Value (PPV) which is essentially the proportion of new targets that would be obtained after experimentally validating the predictions of the method:

$$PPV = \frac{Number\ of\ known\ targets\ correctly\ predicted}{Number\ of\ predicted\ targets} = \frac{TP}{TP + FP} \tag{3}$$

The False Negative Rate (FNR) accounts for the proportion of true targets that the method has missed:

$$FNR = \frac{Number\ of\ known\ targets\ incorrectly\ predicted}{Number\ of\ known\ targets} = \frac{FN}{FN + TP} \tag{4}$$

The Matthews Correlation Coefficient (MCC) captures both types of error in a single metric, with higher values being better up to +1 (perfect classification):

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{5}$$

Lastly, the Number of Predicted Targets (NPT) will be also reported to investigate how this varies with the method's control parameter k. The entire workflow is sketched in Fig. 1.

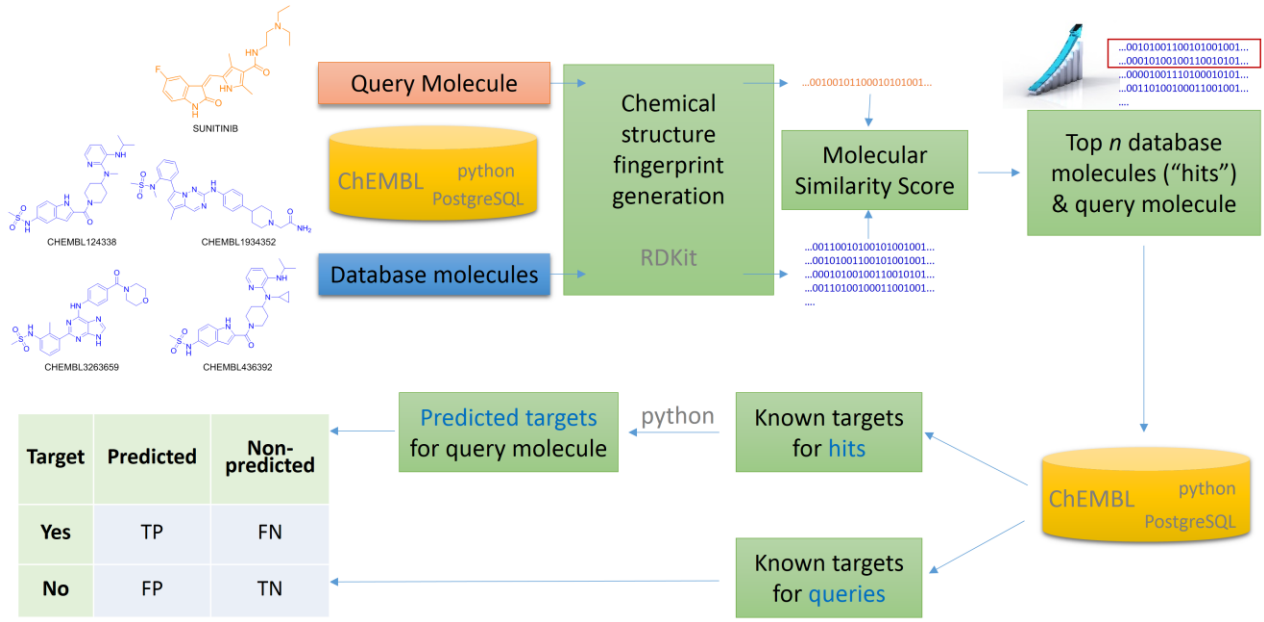


Figure 1. Generic workflow to apply and validate a ligand-centric method for target fishing.

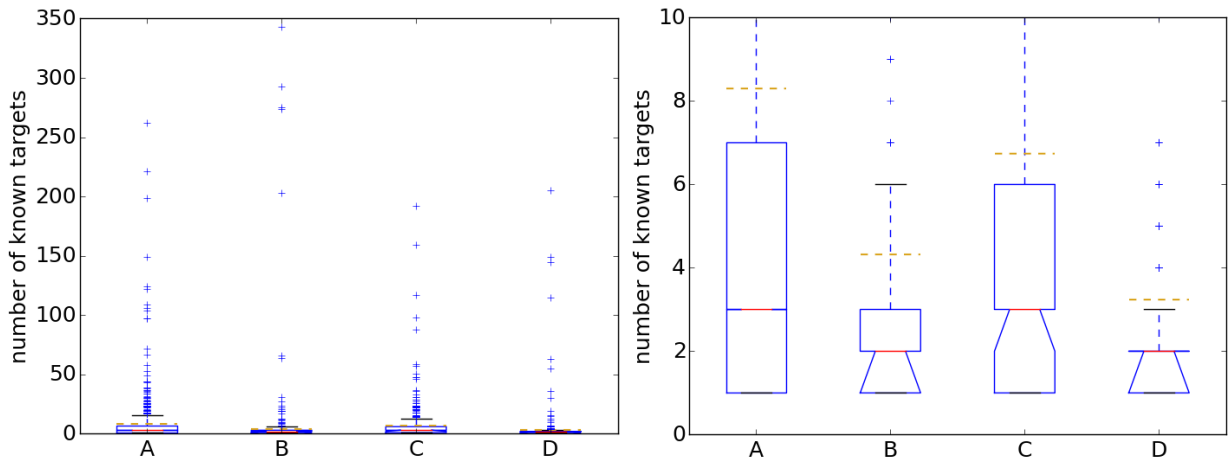
3. Results and Discussion

Three questions are addressed in this section, each in a different subsection. This analysis is based on the performance obtained by the query molecules in the four datasets in Table 2, which will be summarised with boxplots of PPV, FNR, MCC and NPT.

3.1. *How many targets are typically hit by a molecule?*

For each of the four cases in Table 2, Fig. 2 shows Tukey boxplots summarising the distribution of the number of known single-protein targets across query molecules. On the left, a substantial number of strong outliers are appreciated. These correspond to promiscuous query molecules such as sunitinib, which has 192 submicromolar targets. In contrast, there are also selective drugs like

the antiretroviral agent Nelfinavir with only one known target below 1µM (HIV-1 protease; although there are also many non-molecular targets annotated in ChEMBL for this drug). On the right, we can appreciate that approved drugs have an average of eight known targets with potency better than 1µM (dashed yellow line in set A), although the median number is three targets (continuous red line in the same set). However, the boxplot's lower quartile value indicates that at least 25% of these drugs have just one known target and thus seem very selective. It is also noteworthy that the number of annotated targets for the set of random molecules is smaller than that for approved drugs, with four targets on average (set B) instead of eight. This substantial difference is likely to be due to a much higher number of targets being tested during the process of developing a drug.



(a) (b)

Figure 2. (**a**) Tukey boxplots with the number of known targets across query molecules (**b**) A zoom of the same boxplots on the left (the mean is the yellow line and the median is the red line). A = (Approved, 10µM), B = (Random, 10µM), C = (Approved, 1µM), D = (Random, 1µM).

3.2. *How many predicted targets have to be tested to find a true target?*

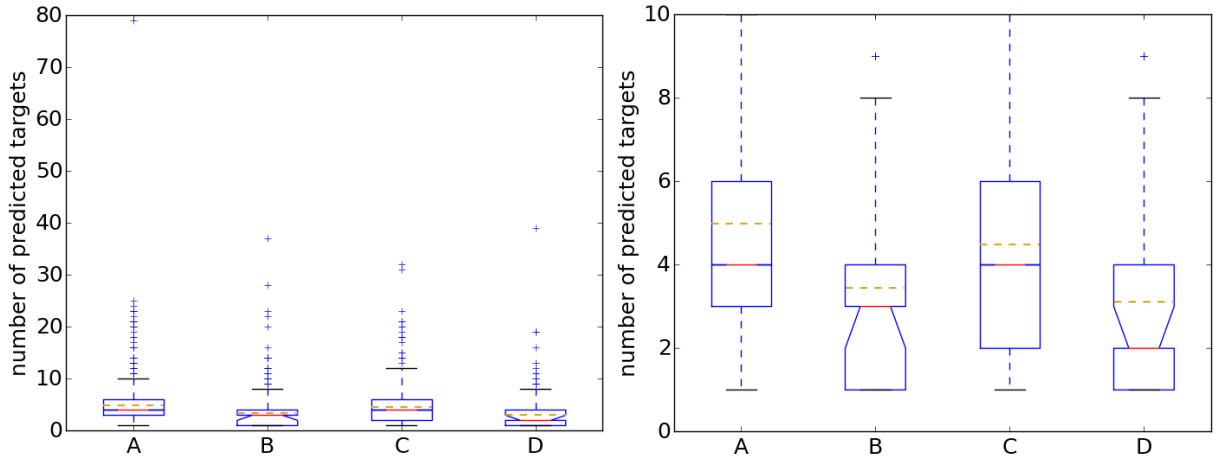
Table 4 presents average performance results for approved drugs, with the method using four different k values. As k increases, Type I errors increase (i.e. lower PPVs) and Type II errors decrease (i.e. lower FNRs). In other words, as more top hits are used to provide predicted targets, fewer known targets are missed. However, this comes at the cost of having more false positives, as target inferences are made using increasingly less similar database molecules. Using the top 5 hits to predict targets (k=5) provides the best compromise between these two conflicting objectives (i.e. the highest average MCC). This setting leads to 4.57 predicted targets on average over the query molecules (note that each top hit may have more than one known target, but collectively provide fewer targets because some of these are repeated). Lastly, the very high average ACC values are due to each classification problem being highly unbalanced and the method correctly discarding the vast majority of non-targets. However, unlike PPV and FNR, ACC does not precisely measure Type I and II errors and hence is not helpful to address the investigated questions.

Table 4. Average (av) performance for query molecules from set C= (Approved, 1μM).

k	avNTP	avACC	avPPV	avFNR	avMCC
1	1.83	0.997	0.49	0.76	0.32
5	4.57	0.997	0.41	0.62	0.36
10	7.08	0.996	0.32	0.56	0.33
15	9.18	0.995	0.30	0.53	0.34

Fig. 3 shows the distribution of the number of predicted targets across query molecules using k=5. By comparing it with Fig. 2, it is observed that there are substantially more known targets than predicted targets for approved drugs using the top 5 hits for predictions (this is not the case

for the sets of random molecules, where most molecules have a higher number of predicted targets than of known targets).



(a) (b)

Figure 3. (**a**) Tukey boxplots with the number of predicted targets across query molecules using k=5; (**b**) A zoom of the same boxplots on the left (the mean is the yellow line and the median is the red line). A = (Approved, 10µM), B = (Random, 10µM), C = (Approved, 1µM), D = (Random, 1µM).

Fig. 4 (a) summarises the distribution of PPV results across the query molecules. For approved drugs, the mean PPV is between 0.35 and 0.41 depending on the adopted threshold. That is, despite the simplicity of the method and thanks to the wealth of data on which relies, on average only five predicted targets will have to be tested in order to find at least two true targets with potency better than 1µM. In all cases, there is strong performance variability across the query molecules, as it can be appreciated by the large interquartile range (IQR) of each boxplot. For instance, in set C, the predicted targets of 150 approved drugs are 100% correct (PPV=1), whereas those for other 170 approved drugs are 100% wrong (PPV=0). Also, the cases with a tighter

activity threshold of 1µM are on average better predicted than their counterparts using 10µM. On the other hand, the sets with random molecules obtained much better results than those with approved drugs. Thus, if we order the four cases by average PPV (yellow line in Fig. 4), this gives the following performance hierarchy D>B>C>A (i.e. D obtains higher average PPV than B, B better than C and C better than A). Interestingly, this is the opposite ranking for both the number of known targets (A>C>B>D). In other words, those sets with a higher number of known targets tend to be harder to predict.

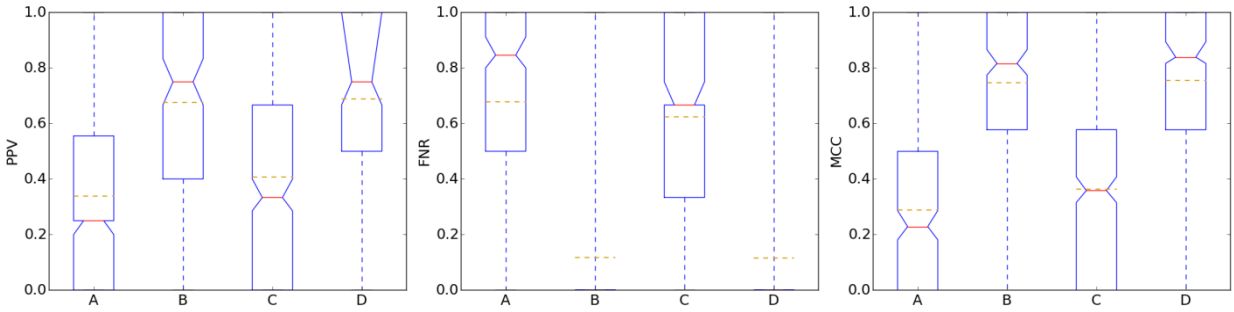


Figure 4. Tukey boxplots for PPV, FNR and MCC for the four datasets using k=5. A = (Approved, 10µM), B = (Random, 10µM), C = (Approved, 1µM), D = (Random, 1µM).

However, the cause of obtaining lower predictive accuracy with approved drugs is not their higher number of known targets *per se*, but an underlying factor correlated with it: the query drug and its top hits, which should include some of the chemical derivatives that eventually led to this drug, often have a lower overlap in terms of known targets. A possible explanation for a low overlap is that some of the top hits could have been tested against a range of targets in other studies, which might not have included the drug and thus this was not tested against the targets (a lower PPV for this query molecule would be consequently obtained, as such targets would be perceived as false positives). Importantly, while these are not known targets of the drug, some are expected to

become a known target once tested. In contrast, a molecule from the randomly-chosen set often has a larger overlap with its top hits (e.g. in set D, the predicted targets of 294 randomly-chosen molecules are 100% correct, whereas those for just 40 randomly-chosen molecules are 100% wrong). Many of these cases are likely to arise from a situation where a chemical series is investigated against a set of related targets to be later abandoned (e.g. because of these molecules having insufficient whole-cell activity). This would explain the lower number of known targets and the smaller predictive errors for these sets.

3.3. *How many known targets of the query molecule are typically missed?*

This question is rarely analysed in the context of Target Fishing, but it is necessary to estimate how many discoveries are being missed by the method. Fig. 4 presents the results in terms of FNR and MCC. Looking at the FNR boxplots, only about 10% of the targets are on average missed in the sets of random molecules (i.e. FNR~0.1), whereas the mean of missed targets for approved drugs is about 65%. A large part of these missed targets might be due to more intense research on the drug after approval than on its chemical derivatives, leading to many targets being tested in the former but not the latter. On the other hand, the MCC boxplots show the distribution of the total error across query molecules, i.e. a high MCC necessarily means that query molecule obtains low levels of both Type I and II errors. The latter occurs to most random molecules regardless of the activity threshold (almost 75% of these query molecules have MCCs higher than 0.6). In contrast, only a small proportion of approved drugs are in this category. Again, the performance hierarchy is D>B>C>A for both FNR and MCC. Here, the higher number of known targets in the query

molecules is also correlated with the difficulty of predicting their targets, but this is also explained by the different ways in which the query molecules and their hits were tested against targets.

4. Conclusions

We have seen that ligand-centric techniques for Target Fishing are capable of considering up to thousands of targets more than target-centric techniques. However, it was not clear how accurate the prediction by ligand-centric techniques is in practice. Using the proposed benchmark, we have been able to analyse the performance of a representative of these techniques using approved drugs as the query molecules. Despite the simplicity of the adopted method and owing to the wealth of data on which relies, on average only five predicted targets will have to be tested in order to find at least two true targets with potency better than 1μM. This level of performance is already useful for prospective applications and it is encouraging that there is plenty of scope for methodological improvement. The latter will be particularly needed to reduce the high number of false negatives, or known targets that are currently missed by ligand-centric techniques. Note that, while this issue has not been investigated yet for target-centric techniques, the many targets not considered by these are by construction false negatives of any molecule that hits them.

The results for the set of randomly-selected molecules used as a control group are far better than those for approved drugs. We have argued that the different way in which targets are tested against the query molecules and their top hits is the reason for this marked difference. Since approved drugs have been presumably tested against many more targets, we consider that their performance level is more realistic than that of the control set. Interestingly, we have identified an average of eight known targets under 10μM in approved drugs, which suggest that

polypharmacology is a common and strong event. Lastly, high performance variability across query molecules has been observed in all cases. Thus, a promising avenue for future research consists in investigating which features make the molecule more difficult to predict in order to assign a confidence score to each prediction.

Acknowledgments

This work has been carried out thanks to the support of the A*MIDEX grant (n° ANR-11-IDEX-0001-02) funded by the French Government «Investissements d'Avenir» program awarded to P.J.B.

Additional information

The authors declare no competing financial interests.

References

1. Schneider, G. Virtual screening: an endless staircase? *Nat Rev Drug Discov* **9,** 273–276 (2010).

2. Lee, J. & Bogyo, M. Target deconvolution techniques in modern phenotypic profiling. *Curr. Opin. Chem. Biol.* **17,** 118–126 (2013).

3. Cheng, T., Pan, Y., Hao, M., Wang, Y. & Bryant, S. H. PubChem applications in drug discovery: a bibliometric analysis. *Drug Discov. Today* **19,** 1751–1756 (2014).

4. Gaulton, A. *et al.* ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **40,** D1100–1107 (2012).

5. Bento, A. P. *et al.* The ChEMBL bioactivity database: an update. *Nucleic Acids Res.* **42,** D1083–1090 (2014).

6. Martínez-Jiménez, F. *et al.* Target prediction for an open access set of compounds active against Mycobacterium tuberculosis. *PLoS Comput. Biol.* **9,** e1003253–1003257 (2013).

7. Keiser, M. J. *et al.* Predicting new molecular targets for known drugs. *Nature* **462,** 175–181 (2009).

8. Liu, X. *et al.* In Silico target fishing: addressing a 'Big Data' problem by ligand-based similarity rankings with data fusion. *J. Cheminform.* **6,** 33–47 (2014).

9. Gao, Z. *et al.* PDTD: a web-accessible protein database for drug target identification. *BMC Bioinformatics* **9,** 104–111 (2008).

10. Cortés-Cabrera, A., Morris, G. M., Finn, P. W., Morreale, A. & Gago, F. Comparison of ultra-fast 2D and 3D ligand and target descriptors for side effect prediction and network analysis in polypharmacology. *Br. J. Pharmacol.* **170,** 557–567 (2013).

11. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J.Che.Inform.Model.* **50,** 742–754 (2010).

12. Gfeller, D. *et al.* SwissTargetPrediction: a web server for target prediction of bioactive small molecules. *Nucleic Acids Res.* **42,** W32–38 (2014).

13. Ballester, P. J. Ultrafast shape recognition: method and applications. *Future Med. Chem.* **3,** 65–78 (2010).

14. Cereto-Massagué, A. *et al.* Tools for in silico target fishing. *Methods* **71,** 98–103 (2015).

15. *ChEMBL 20 release*. (2015) Available at: ftp://ftp.ebi.ac.uk/pub/databases/chembl/ChEMBLdb/releases/chembl_20/. (Accessed: 4th March 2015).

16. Willett, P. The Calculation of Molecular Structural Similarity: Principles and Practice. *Mol. Inform.* **33,** 403–413 (2014).

17. Durant, J. L., Leland, B. A., Henry, D. R. & Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **42,** 1273–1280 (2002).

18. Lamdrum, G., *RDKit: Open-source cheminformatics*. (2015) Available at: http://www.rdkit.org. (Accessed: 3rd April 2015).

Author contributions

P.J.B. designed the study and wrote the manuscript. A.P. implemented the software and carried out the numerical experiments with the assistance of C.C.D. All authors reviewed the manuscript.