Predicting off-target binding profiles with confidence using Conformal Prediction

Samuel Lampa [1,*] Jonathan Alvarsson [1] Staffan Arvidsson Mc Shane [1]
Arvid Berg [1] Ernst Ahlberg [2] Ola Spjuth [1]

[1]Pharmaceutical Bioinformatics group, Department of Pharmaceutical Biosciences, Uppsala University, Uppsala, Sweden
[2]Predictive Compound ADME & Safety, Drug Safety & Metabolism, AstraZeneca IMED Biotech Unit, Mölndal, Sweden

Abstract

Ligand-based models can be used in drug discovery to obtain an early indication of potential off-target interactions that could be linked to adverse effects. Another application is to combine such models into a panel, allowing to compare and search for compounds with similar profiles. Most contemporary methods and implementations however lack valid measures of confidence in their predictions, and only providing point predictions. We here describe the use of conformal prediction for predicting off-target interactions with models trained on data from 31 targets in the ExCAPE dataset, selected for their utility in broad early hazard assessment. Chemicals were represented by the signature molecular descriptor and support vector machines were used as the underlying machine learning method. By using conformal prediction, the results from predictions come in the form of confidence p-values for each class. The full pre-processing and model training process is openly available as scientific workflows on GitHub, rendering it fully reproducible. We illustrate the usefulness of the methodology on a set of compounds extracted from DrugBank. The resulting models are published online and are available via a graphical web interface and an OpenAPI interface for programmatic access.

Introduction

Drug-target interactions are central to the drug discovery process [1]. It is common that drugs interact with multiple targets [2], and off-target pharmacology and polypharmacology has important implications on drug efficacy and safety [3, 4]. Organizations involved in drug discovery, such as pharmaceutical companies and academic institutions, use many types of experimental techniques and assays to determine target interactions, including in vitro pharmacological profiling [5]. However an attractive complementary method is to use computational (in silico) profiling of binding profiles for ligands [6], which also opens the possibility to predict hypothetical compounds. A common approach to the target prediction problem is to use a panel of structure-activity relationship (QSAR) models [7], where chemicals in a knowledge base with known interaction values (numerical or categorical) are described numerically by descriptors and a statistical learning model is trained to predict numerical values (regression) or categorical values (classification). The recent increase in the number of available data points in interaction databases such as ChEMBL [8] and PubChem [9] makes it possible to using ligand-based models to predict not only targets but also panels of targets.

Several methods and tools are available for target prediction and for constructing and using target profiles. Bender *et al.* use a Bayesian approach to train models for 70 selected targets and use these for target profiling to classify adverse drug reactions [10]. Yao *et al.* describes TargetNet [11], a web service for multi-target QSAR models; an online service that uses Naïve Bayes. Yu *et al.* use Random Forest (RF) and Support Vector Machines (SVM) to predict drug-target interactions from heterogeneous biological data [12]. TargetHunter [13] is another online tool that uses chemical similarity to predict targets for ligands, and show how training models on ChEMBL data can enable useful predictions on examples taken from PubChem bioassays. The polypharmacology browser [14] is a web-based target prediction tool that queries ChEMBL bioactivity data using multiple fingerprints.

We observe three important shortcomings among previous works. Primarily, available methods for ligand-based target profiling often do not offer valid measures of confidence in predictions, leaving the user uncertain

about the usefulness of predictions. Secondly, the majority of the web tools lack an open and standardized API, meaning that it is not straightforward (and in most cases not possible at all) to consume the services programmatically, such as from a script, or a scientific workflow management system used in drug discovery like KNIME [15]. Thirdly, previous works do not publish the preprocessing and modeling workflows in reproducible formats, rendering it hard to update the models as data changes, and limits the portability of methods. In fact, most implementations are only accessible from a website without the underlying implementations being openly available for inspection.

We here present an approach for ligand-based target profiling using a confidence framework, delivering target profiles with confidence scores for the predictions of whether a query compound interacts with each target. The confidence scores were calculated using the Conformal Prediction methodology [16], which has been successfully demonstrated in several recent studies [17, 18, 19, 20]. The goal of this study was to create an automated and reproducible approach for generating a predicted target profile based on QSAR binding models, with the models making up the profile published online as microservices and the profile accessible from a web page. Although the models give a confidence measure we also set out to evaluate them on a test set to see how well they performed on representative data. We exemplified the process by creating a profile for the targets for broad early hazard assessment as suggested by Bowes *et al.* [5].

Methods

Training data

A scientific workflow was constructed to automate the entire data pre-processing pipeline. The first step comprises extracting data on binding association between ligands and targets from the ExcapeDB dataset [21], more specifically the columns Gene symbol, Original entry ID (PubChem CID or CHEMBL ID), SMILES and Activity flag. This was performed early in the workflow to make subsequent data transformation steps less time-consuming, given the relatively large size of the uncompressed ExcapeDB data file (18 GB). From the extracted dataset, all rows for which there existed rows with a conflicting activity value for the same target (gene symbol) and SMILES string, were completely removed.

A subset of the panel of 44 binding targets as suggested in [5] was selected for inclusion in the study. The selection was based on the criteria that targets should have at least 100 active and at least 100 non-active compounds. In addition some targets were excluded for which data was not found in ExcapeDB. This is described in detail below. Some of the gene symbols used in [5] were not found in their exact form in the ExcapeDB dataset. To resolve this, PubMed was consulted to find synonymous gene symbols with the following replacements being done: *KCNE1* was replaced with *MINK1* which is present in ExcapeDB. *CHRNA1* (coding for the α1 sub-unit of the Acetylcholine receptor) was excluded, as it is not present in the dataset (*CHRNA4*, coding for the α4 sub-unit of the Acetylcholine receptor, is present in the dataset). We note though, that both *MINK1* and *CHRNA4* were removed in the filtering step mentioned above, since the dataset did not contain more than 100 active and 100 non-active compounds for *MINK1* nor *CHRNA*. However, since one aim of the study is to present and publish an automated and reproducible data processing workflow, these targets could potentially be included in subsequent runs on later versions of the database with additional data available.

The resulting set (named Dataset1) consisted of 31 targets (marked as included in Table 1). For 21 of these targets, the dataset contained less than 10 000 non-active compounds, leading to imbalanced datasets (marked with a ✓ in the 'Assumed non-actives added' column of Table 1). These 21 targets are referred to as Dataset2, and their respective target datasets was expanded with randomly selected examples from the ExcapeDB dataset which were not reported to be active for the target, thus being 'assumed non-active'. The number of new examples was chosen such that the total number of non-actives and assumed non-actives added up to twice the number of actives, for each target respectively. The compounds for the remaining 10 targets, which were not extended with assumed non-actives, were named Dataset3. All the targets, with details about their respective number of active and non-active compounds, and whether they are included or not, are summarized in table 1.

Table 1: The panel of targets used in this study, identified by gene symbol. Actives and non-actives refer to the number of ligand interactions marked as active and non-active in ExcapeDB. Included indicates if the target was included in the study or excluded, based on whether it did pass the filtering criteria of at least 100 actives and 100 non-actives.

	Gene symbol	Actives	Non-actives (before adding assumed non-actives & deduplication)	Non-actives (after adding assumed non-actives & deduplication)	Assumed non-actives added	Remarks
Included	ACHE	3 160	1 152	5 824	✓	
	ADORA2A	5 275	593	10 092	✓	
	ADRB1	1 306	149	2 544	✓	
	ADRB2	1 955	342 282	341 925		
	AR	2 593	4 725	4 866	✓	
	AVPR1A	1 055	321 406	321 098		
	CCKAR	1 249	132	2 458	✓	
	CHRM1	2 776	417 549	358 330		
	CHRM2	1 817	152	3 440	✓	
	CHRM3	1 676	144	3 234	✓	
	CNR1	5 336	400	10 220	✓	
	CNR2	4 583	402	8 676	✓	
	DRD1	1 732	356 201	355 909		
	DRD2	8 323	343 206	342 958		
	EDNRA	2 129	124	4 050	✓	
	HTR1A	6 555	64 578	64 468		
	HTR2A	4 160	359 962	359 663		
	KCNH2	5 330	350 773	350 452		
	LCK	2 662	283	5 246	✓	
	MAOA	1 260	1 083	2 452	✓	
	NR3C1	2 525	4 382	4 804	✓	
	OPRD1	5 350	826	9 580	✓	
	OPRK1	3 672	303 335	303 111		
	OPRM1	5 837	2 872	11 252	✓	
	PDE3A	197	110	392	✓	
	PTGS1	849	729	1 634	✓	
	PTGS2	2 862	827	5 162	✓	
	SCN5A	316	119	624	✓	
	SLC6A2	3 879	218	7 498	✓	
	SLC6A3	5 017	106 819	106 594		
	SLC6A4	7 228	382	13 660	✓	
Not Included	ADRA1A	1 782	24			
	ADRA2A	839	39			
	CACNA1C	166	20			
	CHRNA1	-	-			Not in ExcapeDB
	CHRNA4	256	17			
	GABRA1	112	5			
	GRIN1	555	92			
	HRH1	1 218	65			
	HRH2	394	56			
	HTR1B	1 262	86			
	HTR2B	1 159	66			
	HTR3A	584	65			
	KCNQ1	37	303 466			
	MINK1	929	8			Synonym to KCNE1
	PDE4D	484	98			

Table 2: Summary of datasets discussed.

Name	Description
Dataset1	All 31 targets selected from the set of 44 targets in [5].
Dataset2	Targets from Dataset1 with at least 10 000 non-actives.
Dataset3	The remaining 10 "large" target datasets, from Dataset1 which were not included in Dataset2
Dataset4	The external test set created by extracting rows from ExcapeDB for a selected set of 1 000 compounds in DrugBank (All withdrawn, and randomly sampled approved, drugs, until reaching 1000 drugs).

Conformal prediction

Conformal Prediction (CP) [16] provides a layer on top of existing machine learning methods and produces valid prediction regions for test objects. This contrasts to standard machine learning that delivers point estimates. In conformal prediction a prediction region contains the true value with probability equal to a selected significance level . Such a prediction region can be obtained under the assumption that the observed data is exchangeable. An important consequence is that the size of this region directly relates to the 'strangeness' of the test example, and is an alternative to the concept of a model's 'applicability domain' [22]. For the classification case, a prediction is represented as a set of p-values (one for each class), and in the Mondrian setting the classes are handled independently, which has attractive properties when dealing with imbalanced datasets [23, 24].

In this study we used the Mondrian conformal prediction implementation in the software CPSign [25], leveraging the liblinear SVM implementation [26] together with the signatures molecular descriptor [27]. This descriptor is based on the neighboring of atoms in a molecule and has been shown to work well for QSAR studies [28, 29] and for ligand-based target prediction [30]. Signatures were generated with height 3, which means that molecular subgraphs including all atoms of distance 3 from initial atoms, are generated. Support vector machines is a machine learning algorithm which is commonly used in QSAR studies [31, 32] together with molecular signatures and similar molecular descriptors, *e.g.*, the extended connectivity fingerprints [33]. In order to not use the assumed non-active compounds in Dataset2 in the calibration step in the Conformal Prediction method, these additional compounds were treated separately, by providing them to the CPSign software with the `--proper-train` parameter, see the CPSign documentation [25].

Hyper-parameter tuning

For each of the 31 targets in Dataset 1, a parameter sweep was run to find the optimal value of the cost parameter of liblinear, optimizing modeling efficiency using 10-fold cross validation. The training approach used an Aggregated Conformal Predictor (ACP) with 10 aggregated models. The parameter sweep evaluated three values for the cost parameter for each target; 1, 10 and 100. The efficiency measure used for the evaluation was the observed fuzziness (OF) score described in [34] as:

$$OF = \frac{1}{m} \sum_{i=1}^{m} \sum_{y_i \neq y} p_i^y, \tag{1}$$

where p_i^y is the p-value of the i^{th} test case for class y, and m is the number of test examples, or in our case with only two classes:

$$OF = \frac{\sum_{i,\ y_i=A} p_i^N \ + \sum_{i,\ y_i=N} p_i^A}{m_A + m_N} \tag{2}$$

where p_i^N is the i^{th} p-value for class N, p_i^A is the i^{th} p-value for class A and m_A and m_N is the number of test examples in class A and N respectively. *OF* is basically an average of the p-values for the wrong class, *i.e.*, lower fuzziness means better prediction.

To study the effect of imbalanced datasets on efficiency, we also implemented a modified version of *OF*, due to the fact that *OF* is influenced more by values in the larger class in case of imbalanced datasets, referred to as "class-averaged observed fuzziness" (*CAOF*) as:

$$CAOF = \frac{\sum_{i,\ y_i=A} p_i^N}{m_A} + \frac{\sum_{i,\ y_i=N} p_i^A}{m_N} \tag{3}$$

with the same variable conventions as above. Where *OF* is only an average for the p-values in the test set, *CAOF* averages the contribution from each class separately, meaning that for very unbalanced cases *OF* is mostly affected by the larger class, while for *CAOF*, both classes contribute equally much, regardless of their respective number of p-values. *CAOF* was not used for cost selection, but is provided for information in the results from the workflow.

A commonly used efficiency measure in CP is the size of the prediction region or set given by the predictor. In the classification setting, this is expressed as the fraction of *multi-label* predictions. This measure is denoted as the *M criterion* (MC) and described in [34]:

$$M\ criterion = \frac{1}{m}\sum_{i=1}^{m} 1_{\{|\Gamma_i|>1\}} \tag{4}$$

where 1_E denotes the indicator function of event E, returning the value 1 if E occurs and 0 otherwise, and Γ_i denotes the prediction set for test example i. A smaller value is preferable.

Modeling workflow

Before the training, the CPSign `precompute` command was run, in order to generate a sparse representation of each target's dataset. ACPs consisting of 10 models were then trained for each target using the CPSign `train` command. The cost value used was the one obtained from the hyper-parameter tuning. The observations added as 'assumed non-actives' were not included in the calibration set to avoid biasing the evaluation. The computational workflows for orchestrating the extraction of data, model building, and the collection of results for summarizing and plotting were implemented in the Go programming language using the SciPipe workflow library that is available as open source software at scipipe.org. The cost values for each target is stored in the workflow code, available on GitHub [35]. A graphical overview of the modeling workflow is shown in figure 1. More detailed workflow graphs are available in the supplementary material, figures S4 and S5).

Results

Published models

Models for all targets in Dataset1 were produced in the form of portable Java Archive (JAR) files, which were also built into similarly portable Docker containers, for easy publication as micro services. The model JAR files, together with audit log files produced by SciPipe, containing trace of the workflow (all the shell commands and parameters) used to produce them, are available for download at [36]

Validity of models

To check that the conformal prediction models are valid (i.e. that they predict with an error rate in accordance to the selected significance level), calibration plots were generated in the cross validation step of the workflow.

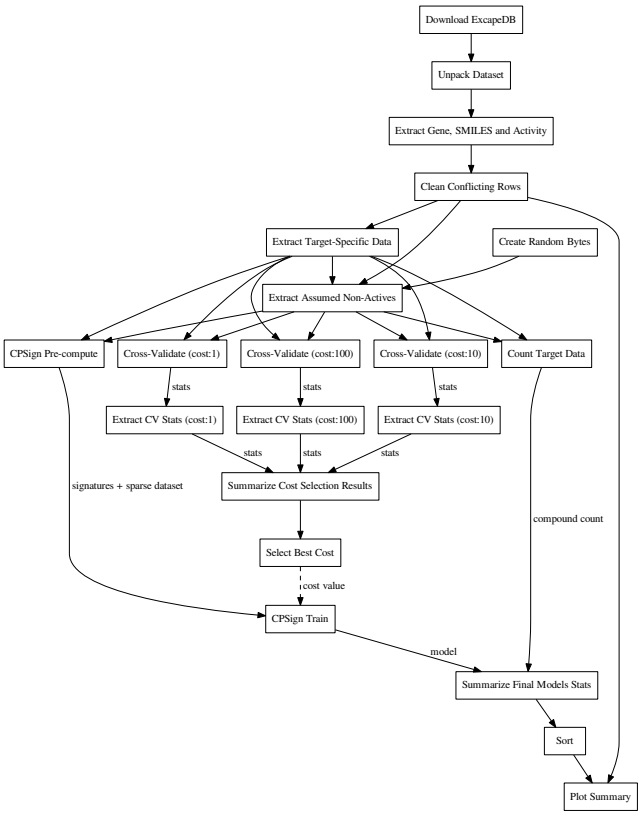
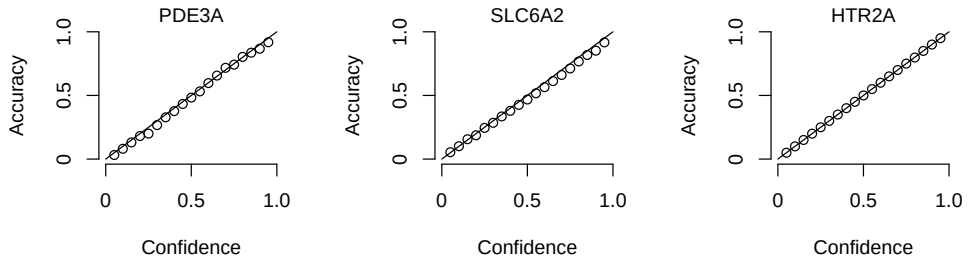


Figure 1: Schematic directed graph of processes and their data dependencies in the the modeling workflow used in the experiments in this study. Boxes represent processes, while edges represent data dependencies between processes. The direction of the edges show in which direction data is being passed between processes. The order of execution is here from top to bottom, of the graph. Each experiment contains additions and modifications to the workflow, but the workflow shown here, exemplifies the basic structure, common among most of the workflows. For more detailed workflow plots, see the supplementary material, figures S4 and S5.

Three example plots, for three representative targets (the smallest, the median-sized and the largest, in terms of compounds in ExcapeDB) can be seen in figure 2, while calibration plots for all targets can be found in the supplementary material (figure S1). From these calibration plots we conclude that all models produce valid results over all significance levels.

Efficiency of models

The efficiency metrics OF, CAOF and MC for Dataset2 (without adding assumed non-actives) are shown in figure 3a. In figure 3b, the same metrics are shown for when all target datasets in Dataset2 have been extended with assumed non-actives, to compensate these datasets' imbalanced nature. We observe that by adding assumed non-actives for small, imbalanced datasets we improve the efficiency of models trained on these datasets. Thus, this strategy of extending the "small" target datasets in Dataset2 was chosen for the subsequent analysis workflows.



(a) (b) (c)

Figure 2: Three representative calibration plots, for models PDE3A (2a), SLC6A2 (2b), and HTR2A (2c), based on the smallest, the median, and the largest target data sets in terms of total number of compounds. The plots show accuracy versus confidence, for the confidence values between 0.05 and 0.95 with a step size of 0.05.

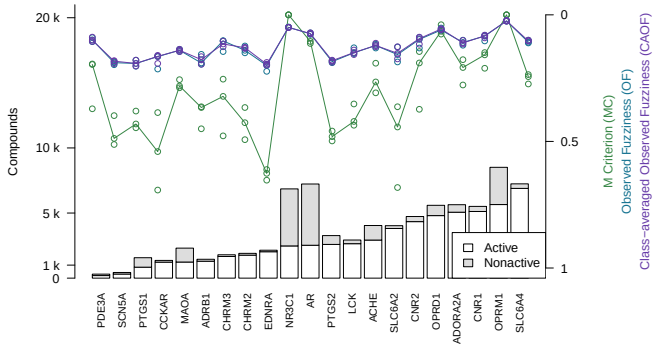
External validation

In order to validate the predictive ability of the trained models, a new dataset was created (Dataset4) by withholding 1 000 compounds from the ExcapeDB dataset, to form an external validation dataset. The compounds chosen to be withheld were the following: i) all small molecules in DrugBank (version 5.0.11) with status "withdrawn", for which we could find either a PubChem ID or a CHEMBL ID, ii) a randomly selected subset of the remaining compounds in DrugBank 5.0.11, with status "approved", for which we could also find PubChem or CHEMBL IDs, until a total number of 1 000 compounds was reached. No regard was paid to other drug statuses in DrugBank such as "investigational".

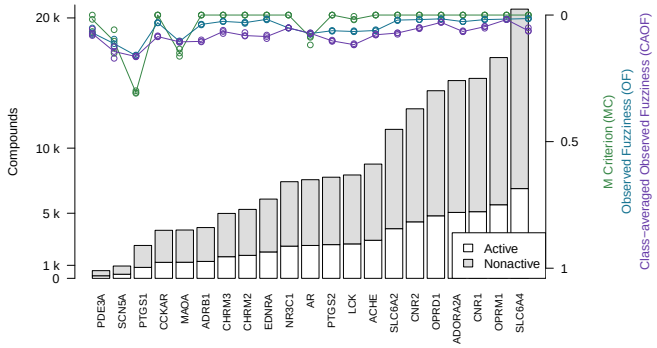
The models built were validated by predicting the binding activity against each of the 31 targets for all compounds for which there existed known binding data for a particular target in ExcapeDB. The validation was done with CPSign's `validate` command, predicting values at confidence levels 0.8 and 0.9. In figure 4 predicted versus observed labels for Dataset4 is shown, for confidence levels 0.8 and 0.9 respectively. It can be seen how the number of prediction of "Both" labels increase when the confidence level increases from 0.8 to 0.9. This is as expected, as this means that fewer compounds could be predicted to only one label, with the higher confidence level. The number of "Null" predictions decreases at the higher confidence, which is also as expected. The reason is that with a higher confidence, the predictor must consider less probable (in the Conformal Prediction ranking sense) predictions to be part of the prediction region. This behavior might seem backwards, but at a higher confidence the predictor has to include less likely predictions in order to reach the specified confidence level, which leads to larger prediction sets. For predicted versus observed labels for each target individually, see the supplementary material, figure S2 and S3.

Target Profile-as-a-Service

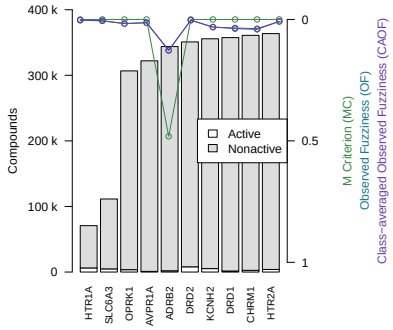
All models based on Dataset2 were published as micro-services with REST APIs publicly made available using the OpenAPI specification [37] on an OpenShift [38] cluster. A web page aggregating all the models was also created. The OpenAPI specification is a standardization for how REST APIs are described, meaning that there is a common way for looking up how to use the REST API of a web service and that greatly simplifies the process of tying multiple different web services together. It simplifies calling the services from scripts as well as from other web pages, such as the web page (figure 5) that generates a profile image out of the multiple QSAR models. At the top of the web page (see figure 5) is an instance of the JSME editor [39] in which the user can draw a



(a) Dataset2 without extending with assumed non-actives. Circles show individual results from the three replicate runs that were run, while the lines show the median value from the individual replicate results. Targets are here sorted by number of active compounds.



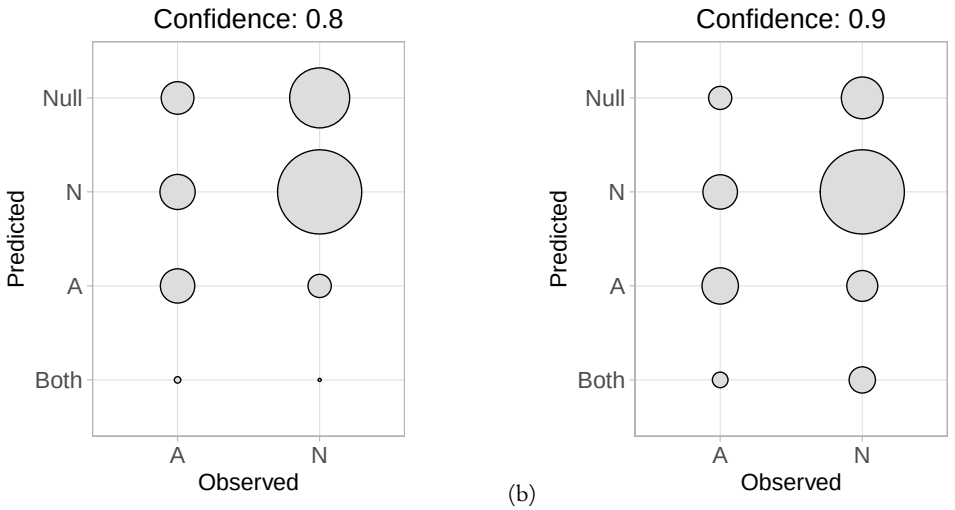
(b) Dataset2 after extending with assumed non-actives. Circles show individual results from the three replicate runs that were run, while the lines show the median value from the individual replicate results. Targets are here sorted by number of active compounds.



(c) Dataset3, the 10 largest target datasets, which were not extended with assumed non-actives. Targets are here sorted by total number of compounds.

Figure 3: Efficiency metrics (M Criterion, Observed Fuzziness and Class-Averaged Observed Fuzziness) for Dataset1, Dataset2 and Dataset3.

molecule. As the user draws the molecule, the web page extracts the SMILES from the editor and sends it to the individual model services to get predictions based on all available models. The user can set a threshold for the confidence and get visual feedback on whether the models predict the drawn molecule as active or non-active for each of the targets, at the chosen confidence. In figure 5 on the right side is a graphical profile in the form of a bar plot where confidence of the active label is drawn in the upward direction and the confidence for non-active is drawn in the downward direction. Hovering a bar in the plot will give information about which model a the bar corresponds to. The web page is accessible at http://ptp.service.pharmb.io/



(a) (b)

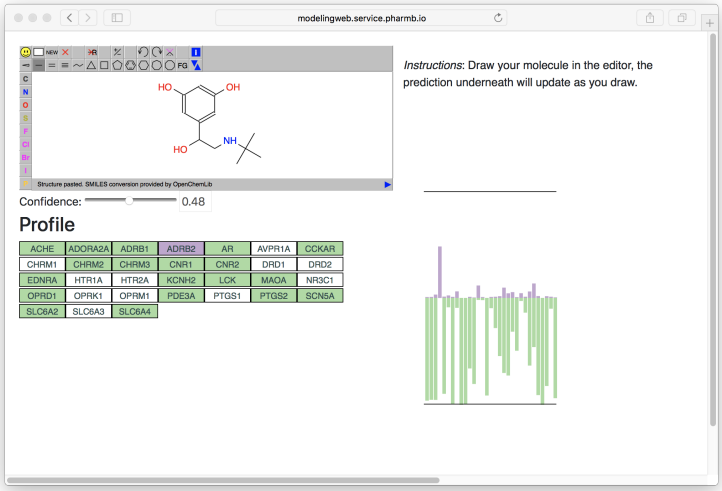
Figure 4: Predicted versus observed labels, for all targets, for the prediction data, at confidence level 0.8 (4a) and 0.9 (4b). The x-axis show observed labels (as found in ExcapeDB), while the y-axis show the set of predicted labels. The areas of the circles are proportional to the number of compound-target binding data points, for each observed label/predicted label combination. For predicted versus observed labels for each target individually, see the supplementary material, figure S2 and S3.

Example predictions

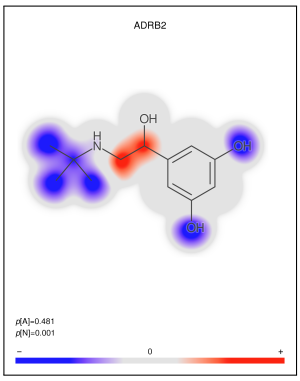
Using the models built without the external validation dataset (Dataset4), target profiles were predicted for three molecules from the test set (figure 6), *i.e.*, the profiles were made for drugs that the models have not seen before. Figure 6a shows the target profile for Tacrine, a centrally acting anticholinesterase, with a distinct peak for the ACHE gene, es expected. Further, we note that most other targets are predicted as non-active with high p-values (green color) or predicted as active with relatively low p-values (purple color). Figure 6b shows the target profile for Pilocarpine, a muscarinic acetylcholine receptor M_1 agonist, with a target profile consisting of mostly non-active predictions, and only mildly two active targets (CHRM1 and LCK). We note that LCK has a similar p-value for active and non-active. For a conformal prediction in binary classification setting, the *confidence* of a prediction is defined as $1 - p_2$ where p_2 is the lower p-value of the two, or in other words, the p-value for the class that was not predicted [40]. This means that even if a prediction has one high p-value, its confidence and hence usefulness in a decision setting might be low. Figure 6c shows the target profile for Pergolide, an agonist for DRD1, DRD2, HTR1A, and HTR2A which shows up as the four highest positive predictions in the profile.

Discussion

The use of workflows to automate pre-processing and model training and make it completely reproducible has several implications. Primarily, the entire process can be repeated as data change, e.g. new data is made available or data is curated. In our case, the pre-processing can be re-run when a new version of ExCAPE-DB is released, and new models trained on up-to-date data can be deployed and published without delay. The components of the pre-processing workflow are however general, and can be re-used in other settings as well. Further, a user can select the specific targets that will be pre-processed, and focus the analysis on smaller subsets without having to pre-process and train models on all targets, which could be resource-demanding. With a modular workflow it is also easy to replace specific components, such as evaluating different strategies and modeling methods.



(a) The profile as seen on the web page (on the right hand in the figure). To show the profile, the user draws a molecule and selects a confidence level, whereafter the profile will update underneath. The profile is shown as a bar plot with two bars for each target: A purple bar, pointing in the upward direction, indicating the size of the p-value of the "Active" label, and a green bar, pointing downwards, indicating the size of the p-value for the "Non-active" label.



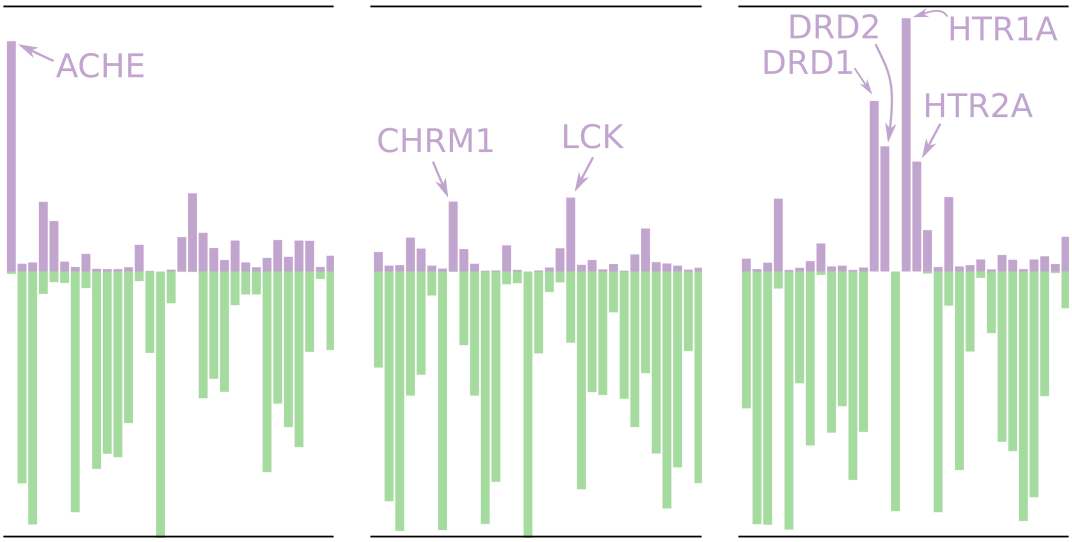
(b) Coloring of which parts of the molecule contributed the most to the prediction for ADBR2. Red color indicates the centers of molecular fragments (of height 3) that contributed most to the larger class, while blue color indicates center of fragments contributing most to the smaller class. In this case the larger class is "Active", which can be seen in the size of the p-values in the bottom left of the figure (p[A]=0.481 >p[N]=0.001).

Figure 5: The prediction profile for Terbutaline, a known selective beta-2 adrenergic agonist used as a bronchodilator and tocolytic.

The packaging of models as JAR-files and Docker containers makes them portable, to be transferred and deployed on different systems, including servers or laptops on public and private networks without cumbersome dependency management. We chose to deploy our services inside RedHat OpenShift container orchestration system, which has the benefit of providing a resilient and scalable service, but any readily available infrastructure provider is sufficient. The use of OpenAPI for deploying an interoperable service API means that the service is simple to integrate and consume in many different ways, including being called from a web page, (such as our reference page on https://ptp.service.pharmb.io/) but also into third party applications and workflow systems. With the flexibility to consume models on individual level comes the power to put together custom profiles (panels) of targets. In this work we have selected targets based on usefulness in a drug safety setting, but it is easy to envision other types of panels for other purposes. While there has been some previous research on the use of predicted target profiles [14, 11], further research is needed to maximize their usefulness and to integrate with other types of in vitro and in silico measures. Our methodology and implementation facilitates such large-scale and integrative studies, and paves the way for target predictions that can be integrated in different stages of the drug discovery process.

Conclusion

We developed a methodology and implementation of target prediction profiles, with fully automated and reproducible data pre-processing and model training workflows to build them. Models are packaged as portable



(a) The profile for Tacrine, a centrally acting anticholinesterase, with a distinct peak for the ACHE gene.

(b) The profile for Pilocarpine, a muscarinic acetylcholine receptor M_1 agonist, with only two moderately higher peaks for positive prediction, CHRM1 and LCK.

(c) The profile for Pergolide, a DRD1, DRD2, HTR1A, and HTR2A agonist, which is reflected by the four highest p-values for a positive prediction.

Figure 6: Profiles for a few of the removed drugs using the validation models, *i.e.*, these molecules are not in the training sets for the models. The profiles are shown as bar plots with two bars for each target: A purple bar pointing in the upward direction, indicating the size of the p-value of the "Active" label, and a green bar pointing downwards, indicating the size of the p-value for the "Non-active" label.

Java Archive (JAR) files, and as Docker containers that can be deployed on any system. We trained data on 31 targets related to drug safety, from the ExCAPE dataset and published these as a predictive profile, using conformal prediction to deliver prediction intervals for each target. The example profile is deployed as an online service with an interoperable API.

Data Availability

- The datasets analyzed for this study can be found on Zenodo (for ExcapeDB) [41], and on the DrugBank website (for DrugBank datasets) [42].
- The data (i.e. the predictive models) generated in this study are available on Zenodo at [36].
- Source code used in this study, is available on GitHub at [35].

Abbreviations

- A: Active
- ACP: Aggregated Conformal Predictor
- CAOF: Class-Averaged Observed Fuzziness
- CP: Conformal Prediction
- JAR: Java Archive (A file format)
- MC: M Criterion (Fraction of multi-label predictions)
- N: Non-active
- OF: Observed Fuzziness
- QSAR: Quantitative Structure-Activity Relationship
- RF: Random Forest
- SMILES: Simplified molecular-input line-entry system (A text-based representation of chemical structures)
- SVM: Support Vector Machines

Conflict of Interest Statement

OS, JA, AB, and SA are involved in Genetta Soft AB, a Swedish based company developing the CPSign software.

Author Contributions

OS conceived the study. OS, JA, SA and SL designed the study, interpreted results, and wrote the manuscript. SL implemented the workflow and carried out the analysis. SA extended CPSign with new features. JA, SA and AB contributed with model deployment and APIs. EA contributed with expertise in target profiles and modeling. All authors read and approved the manuscript.

Funding

This study was supported by OpenRiskNet (Grant Agreement 731075), a project funded by the European Commission under the Horizon 2020 Programme.

Acknowledgments

The computations were performed on resources provided by SNIC through Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) under Project SNIC 2017/7-89.

Bibliography

[1] Muhammed A Yildirim, Kwang-Il Goh, Michael E Cusick, Albert-László Barabási, and Marc Vidal. Drug-target network. *Nat Biotechnol*, 25(10):1119–26, Oct 2007.

[2] Andrew L Hopkins. Network pharmacology: the next paradigm in drug discovery. *Nature chemical biology*, 4(11):682, 2008.

[3] Jens-Uwe Peters. Polypharmacology - foe or friend? *J Med Chem*, 56(22):8955–71, Nov 2013.

[4] Balaguru Ravikumar and Tero Aittokallio. Improving the efficacy-safety balance of polypharmacology in multi-target drug discovery. *Expert Opin Drug Discov*, 13(2):179–192, Feb 2018.

[5] Joanne Bowes, Andrew J Brown, Jacques Hamon, Wolfgang Jarolimek, Arun Sridhar, Gareth Waldron, and Steven Whitebread. Reducing safety-related drug attrition: the use of in vitro pharmacological profiling. *Nature Reviews Drug Discovery*, 11(12):909–922, 2012.

[6] Adrià Cereto-Massagué, María José Ojeda, Cristina Valls, Miquel Mulero, Gerard Pujadas, and Santiago Garcia-Vallve. Tools for in silico target fishing. *Methods*, 71:98–103, Jan 2015.

[7] Hansch C. A Quantitative Approach to Biochemical Structure-Activity Relationships. *Acc. Chem. Res.*, 2:232–239, 1969.

[8] Anna Gaulton, Anne Hersey, Micha l Nowotka, A Patrícia Bento, Jon Chambers, David Mendez, Prudence Mutowo, Francis Atkinson, Louisa J Bellis, Elena Cibrián-Uhalte, Mark Davies, Nathan Dedman, Anneli Karlsson, María Paula Magariños, John P Overington, George Papadatos, Ines Smit, and Andrew R Leach. The ChEMBL database in 2017. *Nucleic Acids Res*, 45(D1):D945–D954, Jan 2017.

[9] Yanli Wang, Stephen H Bryant, Tiejun Cheng, Jiyao Wang, Asta Gindulyte, Benjamin A Shoemaker, Paul A Thiessen, Siqian He, and Jian Zhang. PubChem BioAssay: 2017 update. *Nucleic Acids Res*, 45(D1):D955–D963, Jan 2017.

[10] Andreas Bender, Josef Scheiber, Meir Glick, John W Davies, Kamal Azzaoui, Jacques Hamon, Laszlo Urban, Steven Whitebread, and Jeremy L Jenkins. Analysis of pharmacology data and the prediction of adverse drug reactions and off-target effects from chemical structure. *ChemMedChem*, 2(6):861–73, Jun 2007.

[11] Zhi-Jiang Yao, Jie Dong, Yu-Jing Che, Min-Feng Zhu, Ming Wen, Ning-Ning Wang, Shan Wang, Ai-Ping Lu, and Dong-Sheng Cao. TargetNet: a web service for predicting potential drug-target interaction profiling via multi-target SAR models. *J Comput Aided Mol Des*, 30(5):413–24, 05 2016.

[12] Hua Yu, Jianxin Chen, Xue Xu, Yan Li, Huihui Zhao, Yupeng Fang, Xiuxiu Li, Wei Zhou, Wei Wang, and Yonghua Wang. A systematic prediction of multiple drug-target interactions from chemical, genomic, and pharmacological data. *PLoS One*, 7(5):e37608, 2012.

[13] Lirong Wang, Chao Ma, Peter Wipf, Haibin Liu, Weiwei Su, and Xiang-Qun Xie. TargetHunter: an in silico target identification tool for predicting therapeutic potential of small organic molecules based on chemogenomic database. *AAPS J*, 15(2):395–406, Apr 2013.

[14] Mahendra Awale and Jean-Louis Reymond. The polypharmacology browser: a web-based multi-fingerprint target prediction tool using ChEMBL bioactivity data. *J Cheminform*, 9:11, 2017.

[15] Michael P Mazanetz, Robert J Marmon, Catherine B T Reisser, and Inaki Morao. Drug discovery applications for KNIME: an open source data mining platform. *Curr Top Med Chem*, 12(18):1965–79, 2012.

[16] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.

[17] Isidro Cortés-Ciriano, Andreas Bender, and Thérèse Malliavin. Prediction of PARP Inhibition with Proteochemometric Modelling and Conformal Prediction. *Mol Inform*, 34(6-7):357–66, 06 2015.

[18] Ulf Norinder, Lars Carlsson, Scott Boyer, and Martin Eklund. Introducing conformal prediction in predictive modeling. a transparent and flexible alternative to applicability domain determination. *J Chem Inf Model*, 54(6):1596–603, Jun 2014.

[19] Andy Forreryd, Ulf Norinder, Tim Lindberg, and Malin Lindstedt. Predicting skin sensitizers with confidence - Using conformal prediction to determine applicability domain of GARD. *Toxicol In Vitro*, 48:179–187, Apr 2018.

[20] U Norinder, A Rybacka, and P L Andersson. Conformal prediction to define applicability domain - A case study on predicting ER and AR binding. *SAR QSAR Environ Res*, 27(4):303–16, Apr 2016.

[21] Jiangming Sun, Nina Jeliazkova, Vladimir Chupakhin, Jose-Felipe Golib-Dzib, Ola Engkvist, Lars Carlsson, Jörg Wegner, Hugo Ceulemans, Ivan Georgiev, Vedrin Jeliazkov, Nikolay Kochev, Thomas J. Ashby, and Hongming Chen. ExCAPE-DB: an integrated large scale dataset facilitating big data analysis in chemogenomics. *Journal of Cheminformatics*, 9(1):17, Mar 2017.

[22] Ulf Norinder, Lars Carlsson, Scott Boyer, and Martin Eklund. Introducing conformal prediction in predictive modeling. a transparent and flexible alternative to applicability domain determination. *Journal of chemical information and modeling*, 54(6):1596–1603, 2014.

[23] Ulf Norinder and Scott Boyer. Binary classification of imbalanced datasets using conformal prediction. *J Mol Graph Model*, 72:256–265, Mar 2017.

[24] Jiangming Sun, Lars Carlsson, Ernst Ahlberg, Ulf Norinder, Ola Engkvist, and Hongming Chen. Applying mondrian cross-conformal prediction to estimate prediction confidence on large imbalanced bioactivity data sets. *J Chem Inf Model*, 57(7):1591–1598, 07 2017.

[25] Staffan Arvidsson. CPSign Documentation. http://cpsign-docs.genettasoft.com, 2016. [Online; Accessed 28-February-2018].

[26] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A library for large linear classification. *Journal of machine learning research*, 9(Aug):1871–1874, 2008.

[27] Jean-Loup Faulon, Donald P Visco, and Ramdas S Pophale. The signature molecular descriptor. 1. Using extended valence sequences in QSAR and QSPR studies. *Journal of chemical information and computer sciences*, 43(3):707–720, 2003.

[28] Maris Lapins, Staffan Arvidsson, Samuel Lampa, Arvid Berg, Wesley Schaal, Jonathan Alvarsson, and Ola Spjuth. A confidence predictor for logD using conformal regression and a support-vector machine. *Journal of cheminformatics*, 10(1):17, 2018.

[29] Jonathan Alvarsson, Samuel Lampa, Wesley Schaal, Claes Andersson, Jarl E S Wikberg, and Ola Spjuth. Large-scale ligand-based predictive modelling using support vector machines. *J Cheminform*, 8:39, 2016.

[30] Jonathan Alvarsson, Martin Eklund, Ola Engkvist, Ola Spjuth, Lars Carlsson, Jarl ES Wikberg, and Tobias Noeske. Ligand-based target prediction with signature fingerprints. *Journal of chemical information and modeling*, 54(10):2647–2653, 2014.

[31] Ulf Norinder. Support vector machine models in drug design: applications to drug transport processes and qsar using simplex optimisations and variable selection. *Neurocomputing*, 55(1-2):337–346, 2003.

[32] Xi Bin Zhou, Wen Jing Han, Jing Chen, and Xiao Quan Lu. QSAR study on the interactions between antibiotic compounds and DNA by a hybrid genetic-based support vector machine. *Monatshefte für Chemie-Chemical Monthly*, 142(9):949–959, 2011.

[33] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.

[34] Vladimir Vovk, Valentina Fedorova, Ilia Nouretdinov, and Alexander Gammerman. Criteria of efficiency for conformal prediction. In Alexander Gammerman, Zhiyuan Luo, Jesús Vega, and Vladimir Vovk, editors, *Conformal and Probabilistic Prediction with Applications*, pages 23–39, Cham, 2016. Springer International Publishing.

[35] PTP Project Source Code Repository. https://github.com/pharmbio/ptp-project, 2018. [Online; Accessed 20-June-2018].

[36] Samuel Lampa, Jonathan Alvarsson, Staffan Arvidsson Mc Shane, Arvid Berg, Ernst Ahlberg, and Ola Spjuth. Predictive models for off-target binding profiles generation, June 2018. [Dataset].

[37] OpenAPI Origin - Open Source Container Application Platform. https://www.openapi.org/, 2018. [Online; Accessed 11-June-2018].

[38] OpenShift Origin - Open Source Container Application Platform. https://www.openshift.org/, 2018. [Online; Accessed 11-June-2018].

[39] Bruno Bienfait and Peter Ertl. JSME: a free molecule editor in JavaScript. *Journal of Cheminformatics*, 5(1):24, May 2013.

[40] C. Saunders, A. Gammerman, and V. Vovk. Transduction with confidence and credibility. 1999.

[41] Jiangming Sun, Nina Jeliazkova, Vladimir Chupakin, Jose-Felipe Golib-Dzib, Ola Engkvist, Lars Carlsson, Jörg Wegner, Hugo Ceulemans, Ivan Georgiev, Vedrin Jeliazkov, Nikolay Kochev, and Hongming Chen. ExcapeDB: An integrated large scale dataset facilitating Big Data analysis in chemogenomics, November 2016. [Dataset].

[42] DrugBank Release Version 5.1.0. https://www.drugbank.ca/releases/latest, 2018. [Online; Accessed 20-June-2018].