Papyrus - A large scale curated dataset aimed at bioactivity predictions

O. J. M. Béquignon[1†], B. J. Bongers[1†], W. Jespers[1], A. P. IJzerman[1], B. van der Water[1], G. J. P. van Westen[1*]

1 Division of Drug Discovery and Safety, Leiden Academic Centre for Drug Research, Leiden University, Leiden, The Netherlands

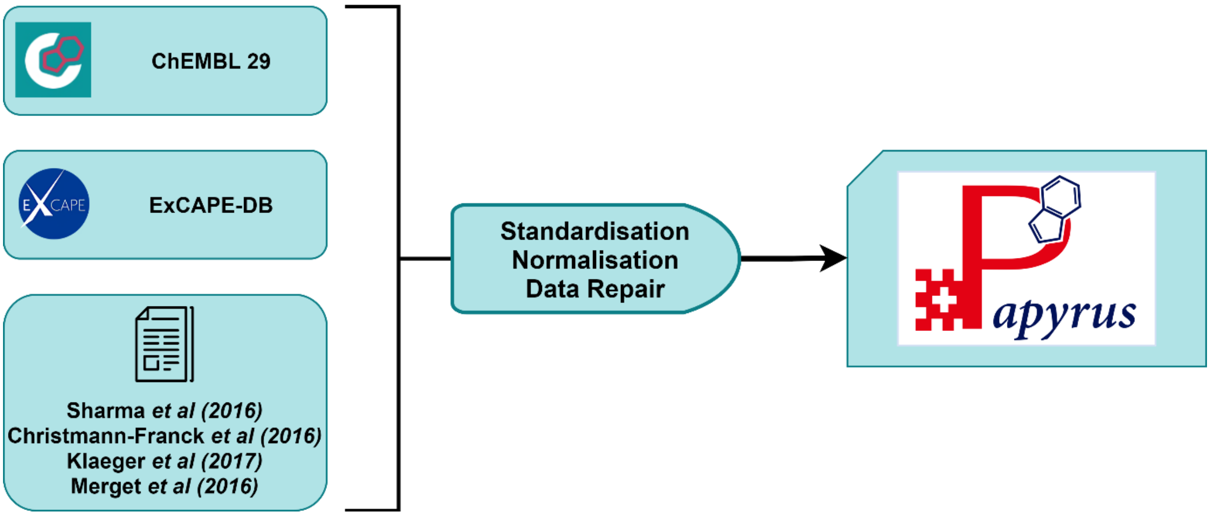
* Corresponding author

Email: gerard@lacdr.leidenuniv.nl (GJPW)

† These authors contributed equally to this work

Abstract

With the recent rapid growth of publicly available ligand-protein bioactivity data, there is a trove of viable data that can be used to train machine learning algorithms. However, not all data is equal in terms of size and quality, and a significant portion of researcher’s time is needed to adapt the data to their needs. On top of that, finding the right data for a research question can often be a challenge on its own. As an answer to that, we have constructed the Papyrus dataset (DOI: [10.4121/16896406](https://doi.org/10.4121/16896406)), comprised of around 60 million datapoints. This dataset contains multiple large publicly available datasets such as ChEMBL and ExCAPE-DB combined with several smaller datasets containing high quality data. The aggregated data has been standardised and normalised in a manner that is suitable for machine learning. We show how data can be filtered in a variety of ways, and also perform some baseline quantitative structure-activity relationship analyses and proteochemometrics modeling. Our ambition is this pruned data collection constitutes a benchmark set that can be used for constructing predictive models, while also providing a solid baseline for related research.



Keywords: machine learning, cheminformatics, bioactivity, curated dataset, Papyrus, standardisation, normalisation.

Introduction

Academic computational drug discovery has gained a massive boost with the growth of publicly available data[1,2]. One of the areas where this has led to improvement is the prediction of bioactivity, specifically ligand-protein affinity. Databases such as ChEMBL and BindingDB provide a wealth of information and relationships between ligands, proteins, and their interaction[3,4]. However, public data has a diverse quality range and is subject to experimental error[5,6]. In contrast to large datasets like ChEMBL, there are also smaller more focused datasets available. These typically focus on a single protein family and usually from a single set of literature such as the Klaeger clinical kinase drugs dataset[7]. Such collections contain a trove of high quality data, but are limited in their scope and are usually not viable as sole data sources in a more general study.

In previous work we compared the performance of established bioactivity prediction methods versus deep neural networks[8]. In order to publish the results from this benchmark the creation of a public dataset (to accompany the publication) was required. ChEMBL (version 20) was used and a high quality subset was extracted and made available[9]. There were several problems we ran into. Firstly there was the amount of work needed to prepare the ChEMBL dataset, leading to an inability to include the separate smaller scale datasets we were planning to add.. In addition, there was a large reduction in size due to the selection of high quality data with the final dataset being 2.5% of the total ChEMBL data.

The current research aims to address these issues and produce standardized diverse dataset. This dataset, named Papyrus[10] (in reference to Leiden Papyrus X), is created with ease-on-use and filtering in mind. We want to remove some of the limitations mentioned above and provide a dataset that does not need further curating. Aside from ChEMBL, we implemented data from the ExCAPE-DB[11] database, and added Sharma *et al*'s[12], Christmann-Franck *et al*'s[13], Klaeger *et al*'s[7] and Merget *et al*'s[14] data.

This current work contributes by providing a standardised and normalised dataset that can be used ‘out-of-the-box’. On top of that, we provide multiple sets of integrated descriptors that are widely used in literature. We also show how to manipulate and model the data using proteochemometrics and provide the Python scripts that we used[15]. Lastly, as the focus in Papyrus is on filtering as well, we provide several Python scripts for ease of querying the dataset. This includes filters for organism, activity type, and accession numbers, and a link to the scripts can be found at the end of this document (under ‘Data Availability’).

Material and Methods

Construction of Papyrus

ChEMBL

Three levels of quality were defined in the data: high, medium, and low. Data from the difference sources were all classified in one of these three classifications. ChEMBL version 29 (ChEMBL29) data[16] were first split between high and low quality data In total 18,635,916 activity data points measured on 2,105,464 compounds and 14,554 targets were extracted. The following data were deemed as low quality: Data flagged as potential duplicates, of questioned validity (Supplementary Table 1) - unless errors were confirmed by authors, in which case they were entirely disregarded -, censored, not associated with any pChEMBL value, or of questioned activity (Supplementary Table 2). Remaining activity data were temporarily regarded as high quality. In the ExCAPE-DB dataset[11], if the data source was PubChem[17] (source identifier 7), then the data was flagged as high quality. Protein targets were retrieved along with their classifications if they had Uniprot[18] accessions defined. Accession Q8MMZ4, corresponding to the secondary accession for the Plasmodium falciparum (isolate NF54) cGMP-dependent protein kinase was associated with a secondary UniProt accession and was manually replaced by its primary accession Q8MMZ4. Using accessions, protein sequences were then obtained from UniProt. Only molecules identified as small molecules and associated with a molecular registry number were kept, then parsed from connection tables and standardised (see Molecular structure standardisation). Activity data of high quality were reclassified as low quality if the target type was other than ‘single protein’ or the assay confidence score was 0, 1, 2, 3, 4 or 6 (Supplementary Table 3). Activity data with assay confidence scores of 5 and 8 were reclassified of medium quality.

If low quality data were censored, inequality signs of the standard relation were reversed (Supplementary Table 4) unless expressing an approximation with a tilde, in which case the data were dropped. Standard values of low quality data with unassigned pChEMBL values were only considered if they had case insensitive standard type of either GI50, Ki, Kd, IC50, EC50 or XC50 and if standard units denoted molar or mass concentrations. Scaling factors were appropriately applied to standard activity values (Supplementary Table 5), mass concentrations were transformed to molar concentrations and log-scale transformation applied to all concentrations. Only exceptions to data with unassigned pChEMBL values were records with derivatives of the following standard types: pKi, pKd, pIC50, pEC50 or pXC50 (Supplementary Table 6). For those records, no transformations were applied. Finally, data were flagged on whether activities were derived from IC50, EC50, Ki, Kd or any other data. The preprocessed ChEMBL29 high quality data consisted of 1,097,673 activity values, 549,140 compounds and 4,644 targets, the medium quality data of 489,315 activity values, 263,824 compounds and 2,886 targets, and low quality data of 1,510,494 activity values, 514,302 compounds and 4,711 targets.

ExCAPE-DB

The ExCAPE-DB dataset, consisting of 70,850,163 activity data points of 998,131 compounds measured on 1,667 targets, was first discarded of records originating from ChEMBL version 20 or whose assay identifiers were present in the PubChem flagged data of the preprocessed ChEMBL29. Gene Entrez identifiers were mapped, using the identifier mapping tool of UniProt, to protein unique Swiss-Prot sequences. Only four genes were manually mapped, three genes as they resolved to multiple reviewed entries and one gene as it resolved to multiple unreviewed entries(Supplementary

Table 7). ChEMBL29 protein classifications were then assigned to the previously mapped sequences. Deposition dates of the assays were retrieved from PubChem. Data with numeric activity values were considered of high quality and binary data of low quality. Molecular structures failing standardisation (see Molecular structure standardisation) were downloaded from PubChem and standardised afterwards. Finally low quality activity data with compound-target pairs present in the high quality subset were disregarded. The preprocessed ExCAPE-DB high quality data consisted of 278,226 activity values, 201,644 compounds and 1,535 targets, and low quality data of 58,445,354 activity values, 650,217 compounds and 646 targets.

Sharma et al

Sharma *et al*'s dataset[12], consisting of 258,060 activity data points of 76,017 compounds measured on 8 targets was considered of high quality. Gene names were mapped to unique Swiss-Prot sequences using the identifier mapping tool of UniProt and protein classifications retrieved from ChEMBL29. A set of 14 custom reactions (Supplementary Table 8) were applied to molecular structures failing standardisation (see Molecular structure standardisation), mostly fixing aromaticity-related issues. Years of filing of patents were collected using Google Cloud BigQuery API patents public data and manual mapping (Supplementary Table 9 & 10) after having fixed erroneous patent numbers (Supplementary Table 11). Digital object identifiers or PubMed identifiers of source articles were added when missing (Supplementary Table 12). If activity values were associated with multiple sources, only the first published article or filed patent was recorded. Censored activity values or values not associated with case insensitive standard types GI50, Ki, Kd, IC50 or EC50 and their logarithmically-derived counterparts were disregarded. Mass concentrations were transformed to molar concentrations and log-scale transformation applied to all concentrations but already log-transformed. Finally infinite or activity values lower than 3 and higher than 14 log units were discarded. The preprocessed Sharma data consisted of 77,562 activity values, 40,738 compounds and 8 targets.

Christmann-Franck et al

Christmann-Franck *et al*'s dataset[13], consisting of 344,788 activity data points of 2,065 compounds measured on 448 targets was considered of high quality. The wrongly assigned Cryptococcus neoformans mitogen-activated protein kinase (CPK1) with accession code P0CP66 was corrected to the Plasmodium falciparum calcium-dependent protein kinase 1 (CDPK1) with accession code P62344. Swiss-Prot sequences were retrieved using accessions and protein classifications retrieved from ChEMBL29. Sequence mutations of the hepatocyte growth factor receptor (MET) and the serine/threonine-protein kinase (B-raf) were corrected to M1250T and V600E respectively and that of the Fibroblast growth factor receptor 1 (FGFR1) was reverted to wildtype. Activity data expressed as proportion of reference activities were discarded. Finally molecular structures were standardised (see Molecular structure standardisation). The preprocessed Christmann-Franck data consisted of 135,948 activity values, 1,669 compounds and 485 targets.

Klaeger et al

Klaeger *et al*'s dataset[7], consisting of 5,916 activity data points of 229 compounds measured on 520 targets was considered of high quality. Swiss-Prot sequences were retrieved using HUGO Gene Nomenclature Committee (HGNC) identifiers. If multiple identifiers were assigned the measurement was discarded. Protein classifications were retrieved from ChEMBL29. Apparent Kd values were log-transformed and infinite results disregarded. Finally molecular structures were standardised (see Molecular structure standardisation), with only RDEA-436 failing for its structure was not disclosed. The preprocessed Klaeger data consisted of 5,721 activity values, 228 compounds and 500 targets.

Merget et al

Merget *et al*'s dataset[14], consisting of 260,757 activity data points of 47,774 compounds measured on 241 targets was considered of high quality, except for activity values originating from ChEMBL22, which were disregarded. Data originating from the

Published Kinase Inhibitor Set (PKIS) of GlaxoSmithKline (doi:10.1038/nbt.3374) with activity values of 5 log units were considered as censored and as such reclassified as low quality data. Swiss-Prot sequences were retrieved using HUGO Gene Nomenclature Committee (HGNC) identifiers[19], a few of which were manually fixed (Supplementary Table 13). Protein classifications were retrieved from ChEMBL29. Finally molecular structures were standardised (see Molecular structure standardisation). The Merget preprocessed high quality data consisted of 127,441 activity values, 1,666 compounds and 239 targets, and low quality data of 62,642 activity values, 360 compounds and 195 targets.

Molecular Structure Standardisation

During the preprocessing of each original dataset parent molecular structures were gathered after a first standardisation using the ChEMBL structure pipeline[20]. Then canonical tautomers were determined using the Pipeline Pilot tautomer enumerator[21] with tautomerization of amides enabled. The canonical tautomers were then standardised once again with the ChEMBL structure pipeline after which the parent structures were obtained. Any molecule not parsable from simplified molecular input line entry specification (SMILES) by the RDKit[22] at any step of the previous workflow was considered failing the standardisation process.

After the individual datasets were processed and aggregated into the Papyrus dataset, molecular structures were standardised. This last standardisation ensured normalisation across different sources that each applied different prior standardisation. For instance, tautomerization tools can alter a compound’s stereochemistry by removing or introducing a chiral centre. To limit the effects of having bioactivity values relating to the same molecular compound having different stereochemistry across sources, a set with removed stereochemistry was created and deemed of higher quality than the set with conserved stereochemistry. Molecular structures in the latter, after having removed stereochemistry, were first neutralized with the RDKit by adding or removing hydrogen atoms. Subsequently they were standardised with the ChEMBL structure pipeline after which parent structures were obtained. OpenBabel[23,24] was then used to recreate dative bonds and to neutralize molecules that were not during the previous step. Tetravalent negatively charged boron atoms were overlooked in the latter stage, making them erroneously pentavalent. This was corrected by detecting these pentavalent boron atoms, removing the newly introduced implicit hydrogen atom and reassigning them a negative formal charge. SMILES of molecules with the same connectivity differing by overall charge were converted to InChi[25] with OpenBabel, a step that consists in incorporating the normalisations after the InChI canonicalization process. For these molecules the canonicalization process removed the dative bonds, these were then recreated using OpenBabel. Then Dimorphite-DL[26] was used to deprotonate molecules by setting minimum pH to 14.0. This ensured that after the last standardisation step, equivalent to that applied to individual datasets, in which molecules are neutralised, only one charge state of the same molecular species was present in the set.

Papyrus data aggregation

The processed ChEMBL29 high, medium and low quality, ExCAPE-DB high and low quality, Sharma, Christmann-Franck, Klaeger and Merget datasets were aggregated together. The first step consisted in ensuring that the activity of any compound-target pair was contained within 3 to 14 log units. Then compound-target pairs were uniquely identified by a concatenation of the compound's connectivity and of the target accession along with its mutations if any. All activities relating to the same compound-target pair were then filtered depending on the highest data quality available for that pair. For instance, if high quality activities were identified, any data point deemed of medium to low quality was filtered out. Considering censored activity values, the data was filtered out if contradictory relations were identified, if not the highest recorded activity was retained for lower bounds, and lowest for higher bounds. During this filtering step, all patents and journal articles associated with the activity of a compound-target pair were gathered whatever the quality and only the first published or filed was retained. Finally, activity values were aggregated and mean averages, medians, standard errors of the mean, standard deviations and mean average distances were calculated for each unique compound-target pair.

Use of Papyrus

Data extraction

The first subset that was extracted from Papyrus consists in adenosine receptors. Using the Papyrus Python scripts, data of high quality with protein classification level 5 being “Adenosine receptor” was extracted. This subset, consisting of 15,941 activity points, 24 protein targets and 7,967 compound structures.

Data visualization

Unique molecules of Papyrus were collected based on the uniqueness of their connectivity. Each molecule was encoded using MinHash fingerprint (MHFP6)[27] and then visualized using TMAP[28]. Molecules were labelled using their fraction of carbon atom. Unique proteins of Papyrus were collected based on their unique target identifier. Each sequence was encoded using UniRep[29] 64, 256 and 1,900 average hidden states, final cell states and final hidden states. The 6660 dimensions were then MinHashed and visualized with TMAP. Proteins were labelled using organisms they originate from.

Bioactivity modelling: Quantitative Structure Activity-Relationships

Each protein target in the subset was modelled independently using the Papyrus Python scripts. Targets for which less than 30 activity values or associated with activity values spanning less than 2 log units were disregarded for modelling. Then for each target, a temporal split between training and test sets was performed: datapoints associated with year 2013 and above constituted the test set. If no activity data was available after year 2013, then the target was disregarded. The 777 Mold2 molecular descriptors[30] were calculated for each molecule and were centered and scaled to unit variance. Extreme Gradient Boosting (XGBoost version 1.4.2) regressors and classifiers were trained on the training set using random seed 1234 and default parameters. Regressors were trained to predict mean pChEMBL values using 5-fold cross-validation, while classifiers were trained to predict a binary label of activity class with threshold set at 6.5 log units using 5-fold stratified cross-validation.

Bioactivity modelling: Proteochemometrics

No subsequent filtering of the subsets was carried out since proteochemometrics (PCM) handles multiple targets all at once. A temporal split on year 2013 was employed to split the training and test set. The 777 Mold2 molecular descriptors were calculated for compounds, UniRep 64, 256 and 1,900 average hidden states, final cell states and final hidden states were used as 6,660 protein descriptors and were calculated for each protein. An XGBoost classifier and an XGBoost regressor were trained using the same protocol as for QSAR models.

Results and Discussion

A new dataset, called Papyrus of bioactivities, resulting from the aggregation and extensive standardisation of data from six sources, was created. Unless mentioned otherwise, only the extensively standardised Papyrus set without stereochemistry is considered in this section.

Papyrus dataset statistics

Papyrus consists of 59,763,781 compound-protein pairs, each associated with at least either one activity value or activity class. Additionally, this represents the data of 1,268,606 unique compounds and 6,996 proteins across 496 different organisms. In terms of data quality, 1,236,296 datapoints are of high quality, i.e., representing exact bioactivity values measured and associated with a single protein or complex subunit. 335,854 datapoints are of medium quality, i.e., exact bioactivity values associated with either potentially multiple proteins or a homologous single protein. 58,191,631 datapoints are of low quality, i.e., exact bioactivity values associated with either multiple homologous proteins or homologous complex subunits, censored bioactivity values and binary activity classes. When considering datapoints across all quality types, 2,585,248 are associated with exact bioactivity values, 354,981 with censored data and 56,823,552 with binary activity classes. The repartition of data quality across the ten organisms with most data (Table 1**Error! Reference source not found.**) indicates a clear bias towards human, with more than 93% of the data related to it, but also emphasizes the interest towards rodent targets with more than 4% of the data associated with mouse and 2% with rats.

	Quality			
Species	High	Medium	Low	Total
Homo sapiens (Human)	985,579	246,723	54,363,214	55,595,516
Mus musculus (Mouse)	41,986	6,682	2,465,153	2,513,821
Rattus norvegicus (Rat)	60,374	32,075	1,151,936	1,244,385
Escherichia coli (strain K12)	539	11,283	54,800	66,622
Equus caballus (Horse)	18,326	32	27,987	46,345
Influenza A virus (A/WSN/1933(H1N1))	23,813	-	9,143	32,956
Trypanosoma cruzi	5,886	30	23,927	29,843
Schistosoma mansoni (Blood fluke)	13,916	-	14,473	28,389
Bacillus subtilis	12,106	-	11,693	23,799
Bos taurus (Bovine)	5,923	5,107	8,913	19,943

Table 1: Activity data of organisms in Papyrus with the most datapoints.

When it comes to the activity types Papyrus is derived from (Table 2), most of the data is either associated with untraceable data types, such as for binary data, or with types derived from others, for instance the KIBA scores were derived from IC_{50}, K_i and K_D data[31] present in the Merget source dataset.

Activity type	Original datapoints
K_i	507,821
K_D	118,773
IC_{50}	1,070,430
EC_{50}	141,672
Other	58,315,137

Table 2: Number of original datapoints in Papyrus for each activity type.

Papyrus protein space (Figure 1A) is largely dominated by human proteins, reflecting the abundance of activity values measured for these. Nevertheless, clusters of homologous proteins can be observed, mostly aggregating human rat and mouse protein. As a comparison, the compound space was also visualized (Figure 1B) with carbon fraction evenly spread across clusters.

Concerning protein classification, enzymes represent nearly half of the classified and annotated proteins, with more than 25 million data points, and membrane receptors 21% with more than 11 million (Figure 2A). Family A G protein-coupled receptors represent 37% of proteins annotated with a second level class (Figure 2B), consisting in more than 9 million datapoints, proteases 23%, more than 5 million, and kinases, long thought undruggable targets, represent 18% of the data with more than 4.5 million datapoints.

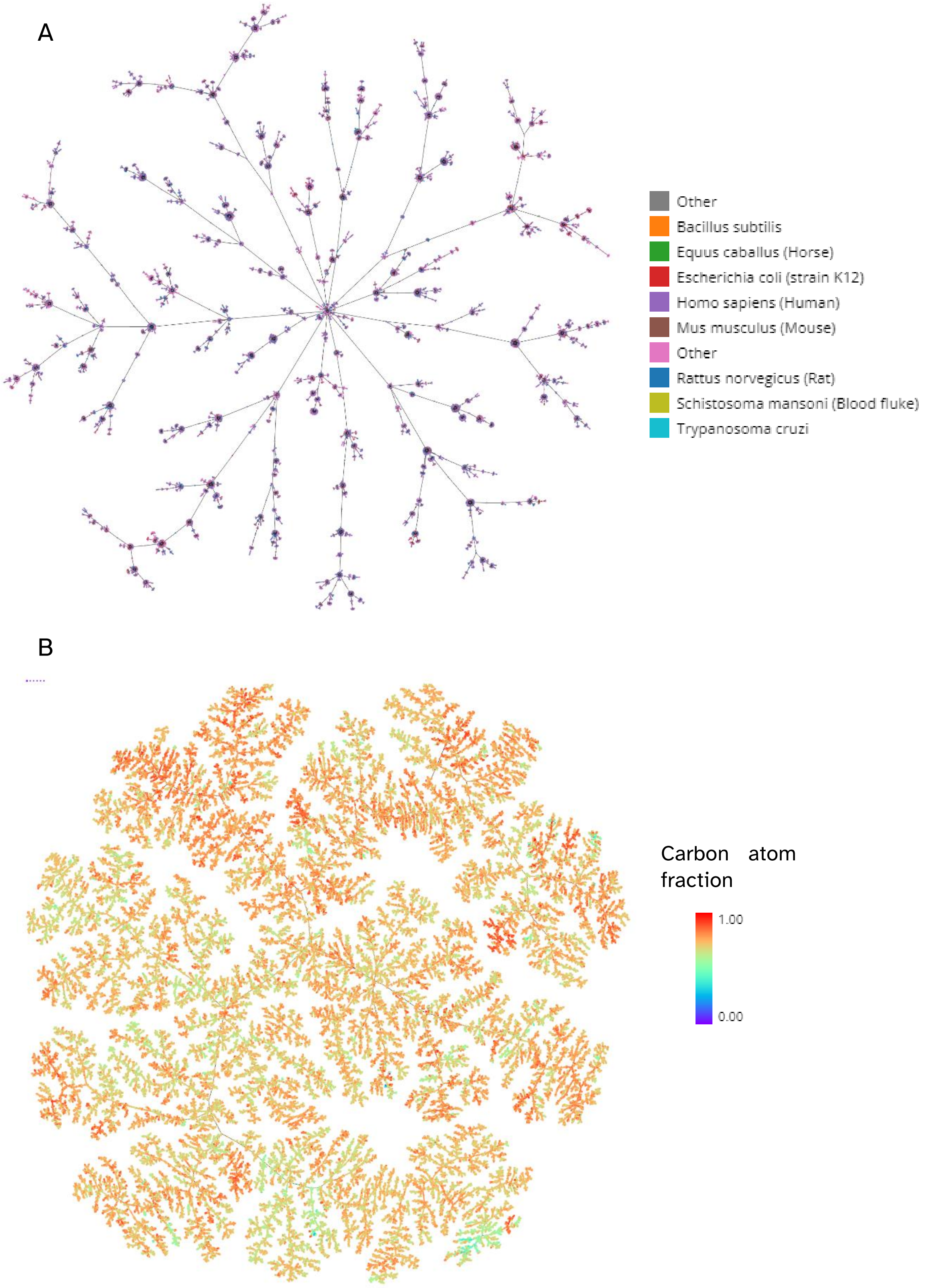


Figure 1: Papyrus protein (A) and chemical (B) spaces.

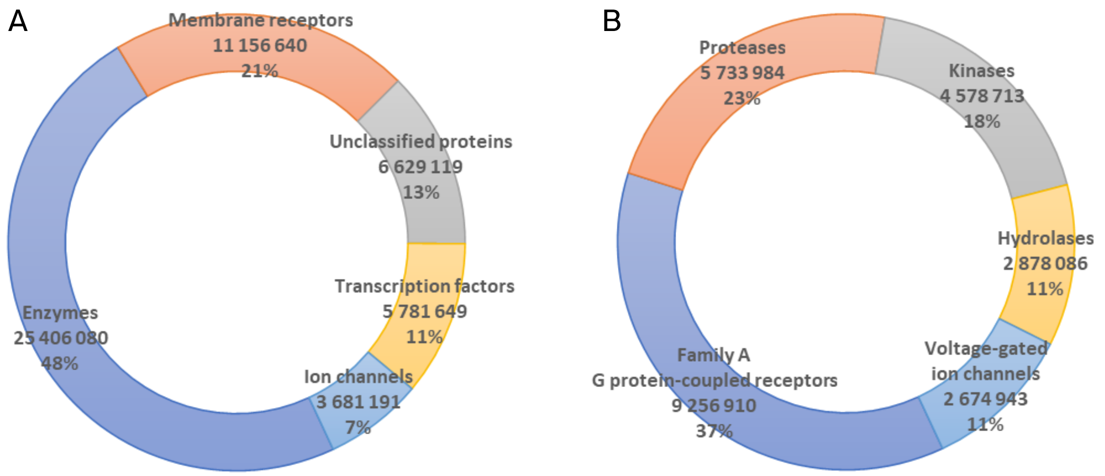


Figure 2: Protein classification levels 1 (A) and 2 (B) of protein targets in Papyrus.

Bioactivity modeling: Results

To exemplify the potential of Papyrus, an Adenosine receptor (AR) subset was extracted, considering only the high-quality data. Quantitative structure-activity relationships (QSAR) and proteochemometrics (PCM) regression and classification models were trained. A temporal split scheme was chosen, to better asses the prediction performance of the models[32] and minimize congeneric series being split between training and test sets.

QSAR models were trained on protein targets with sufficient data. This resulted in only 11 of 24 the adenosine receptors in the subset to be suitable for QSAR modelling. PCM models allowing the use of all related targets, all 24 adenosine receptors could be modelled.

QSAR regression models for human ADORA2b, rat ADORA1, ADORA2a and ADORA3 and mouse ADORA3 performed well with coefficient of determination R^2 between 0.6 and 0.7 during cross-validation (Figure 3A). Surprisingly human ADORA1, ADORA2a, mouse ADORA3 and bovine ADORA1 performed quite bad with median R^2 lower than 0.5 with fold performance reaching -2.3 to -2.1 for the first three. Nonetheless, the maximum error associated with these first three was significantly lower than that of other QSAR models although the root-mean-square error (RMSE) and mean absolute error (MAE) of all models were not significantly different. With regards to external validation (Figure 4B), the performance of the predictions on the temporally split test set performed as expected with RMSE around 1 log unit for most models, only human ADORA1 and mouse ADORA3 performing noticeably badly. The PCM regression model performed quite well at cross-validation in terms of R^2 and RMSE with values of 0.59 and 0.72 but had higher median maximum error than all QSAR models. These results were reflected in the temporal validation.

QSAR classification models of human ADORA2a performed very well with Matthews correlation coefficient (MCC) ranging between 0.62 and 0.70 for cross-validation and 0.48 on the temporal test set (Figure 4). Except for the human ADORA3 and rat ADORA3 that performed bad both during cross-validation and testing due to the imbalance of the datasets (ratios of 1:7 to 2:6 of actives to inactives for human ADORA3 and 4:1 to 5:1 for rat ADORA3) and showing very variable sensitivity and specificity, most models performed equally well at cross-validation and on the test set. The human ADORA1 and ADORA2a, rat ADORA1, ADORA2a and ADORA2b, mouse ADORA1 and bovine ADORA1 had balanced accuracy (BAcc) over 0.70 and area under the receiver operator characteristic curve (AUC) over 0.65, which showed very good predictive performance in a prospective setting. It is worth noting that the bovine ADORA1 QSAR regression model R^2 was one of the lowest (-0.27). The PCM classification model showed performance on par with well performing QSAR models during cross-validation but showed lower performance on test set with MCC of 0.25 and BAcc of 0.62.

Overall models on the AR subset showed similar performance between regression and classification. It is no surprise that the receptors that performed best, i.e. most of rat and human receptors, were those with the most datapoints.

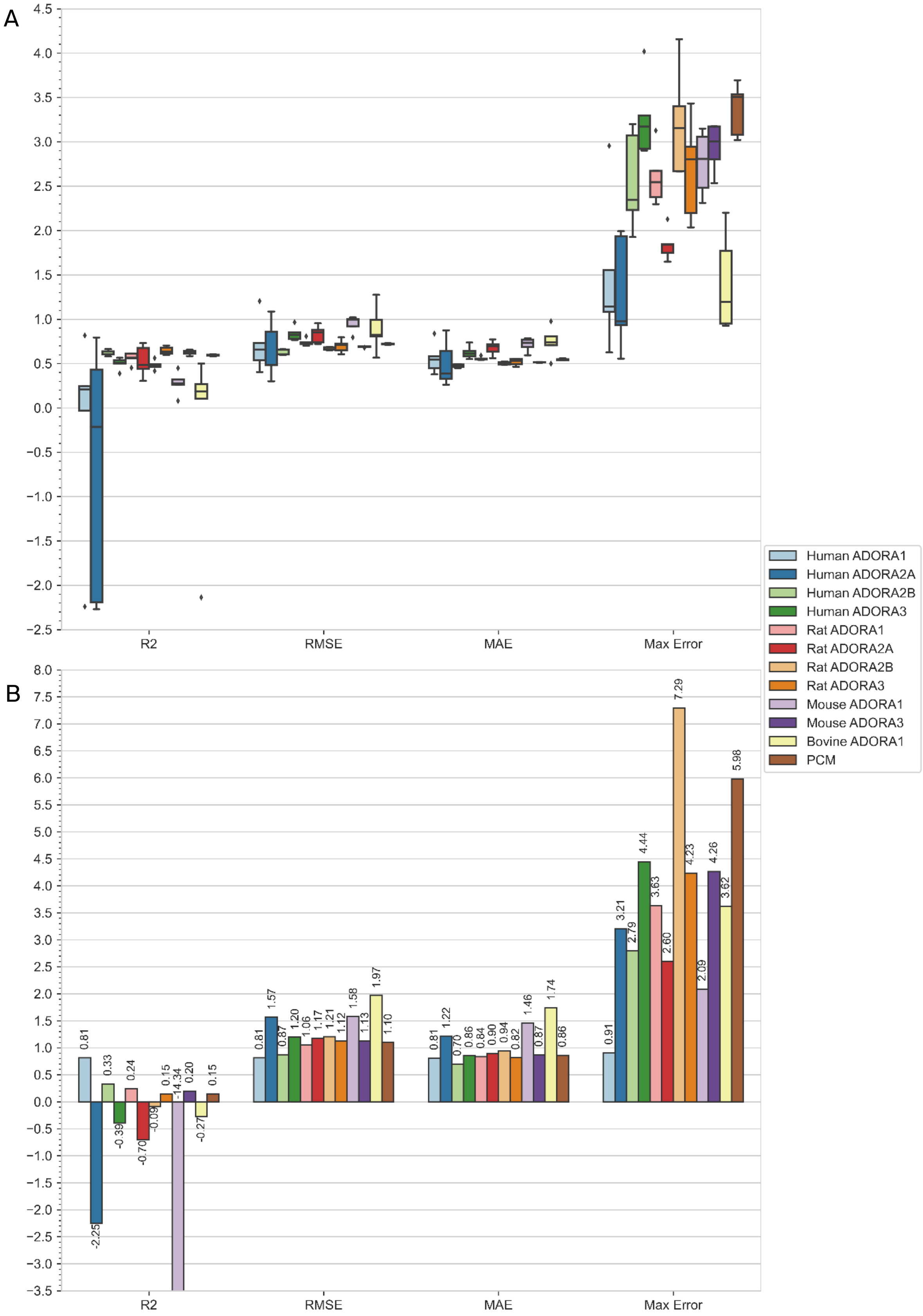


Figure 3: Cross-validation performance (A) and temporally split test set performance (B) of regression QSAR and PCM models. R^2: coefficient of determination, RMSE: root-mean-square error, MAE: mean absolute error, Max Error: Maximal error.

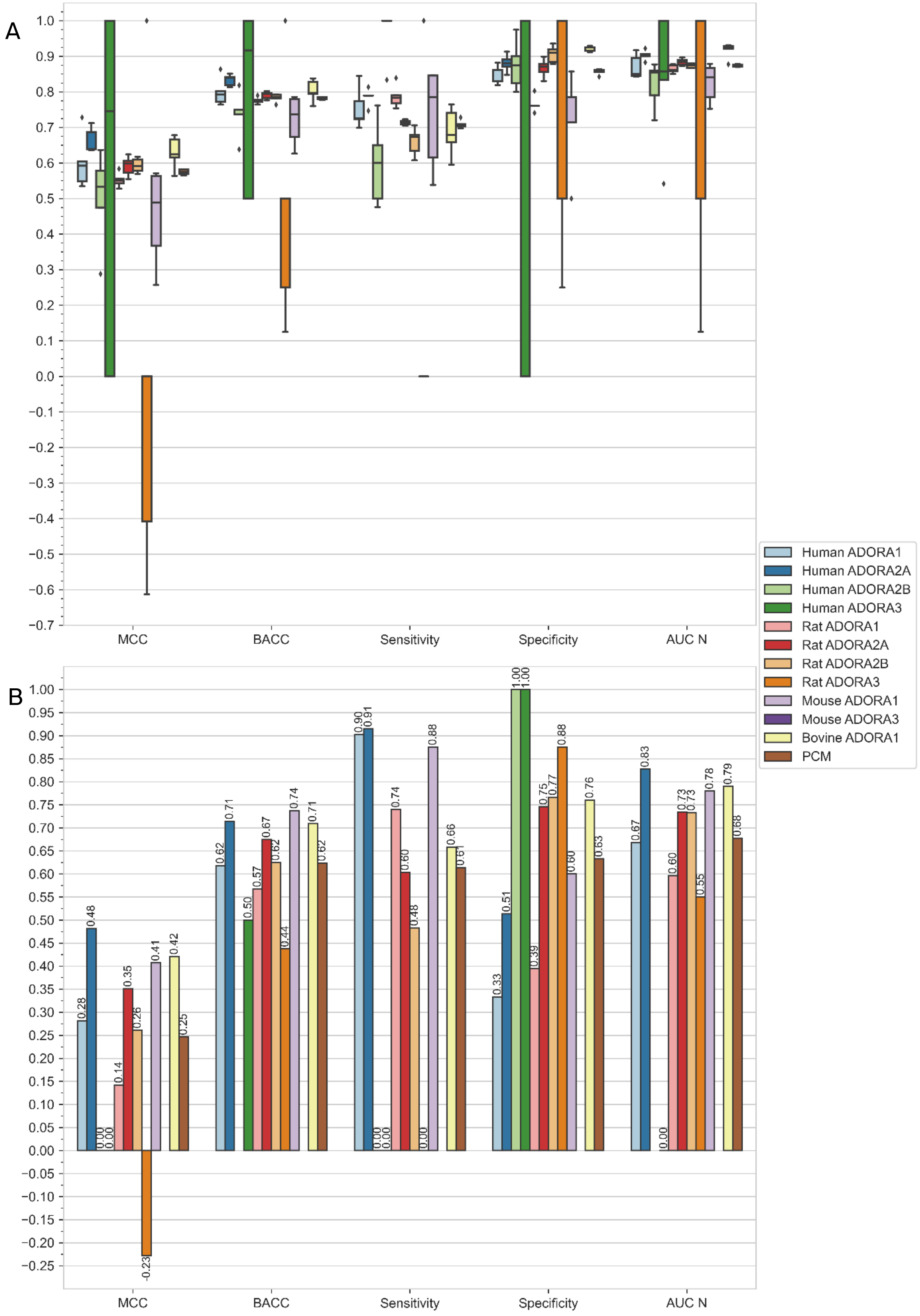


Figure 4: Cross-validation performance (A) and temporally split test set performance (B) of regression QSAR and PCM models. R^2: coefficient of determination, RMSE: root-mean-square error, MAE: mean absolute error, Max Error: Maximal error.

Discussion

We have shown the format of the Papyrus bioactivity dataset, as well as a few examples of baseline models that could be created from this data. While we are confident that is a reliable publicly available dataset, there are still some limitations present.

First, the most important limitation in our eyes is that although Papyrus consists of nearly 60 million activity data points, the data it contains is extremely sparse as only 0.67% of the activity data matrix is represented by the 1.25 million compounds and almost 7,000 proteins. This is unfortunately hard to avoid as many of the compounds have simply not been tested on all proteins. Only relatively popular proteins will appear in the data that is aggregated here. This makes it hard to model proteins that are understudied, however if data are available they can be added to the Papyrus dataset to create a more comprehensive set.

As Papyrus is a static dataset, updates (or corrections) are possible but are reliant on the aggregated datasets. While this is always a restriction on static datasets, there is a second degree of reliance here as all the data needs to be updated in their respective datasets. While we do not think this will pose an issue for modeling, as data freeze often occurs when a dataset is used in research, updates to Papyrus will have more time between them than the respective aggregated datasets.

Another limitation is the choice for the specific datasets that were aggregated into Papyrus in addition to ChEMBL. Firstly we have implemented ExCAPE-DB, like we did in previous work. The single article datasets are a reflection of some of the interests of our group, and we value the high quality data present. There are enough arguments to include a certain dataset that we did not mention here, which would improve the quality of the dataset even more. We do provide the full dataset and all the descriptors that are used. So if anyone wants to add a certain dataset the tools are available to do so. Our goal was to create a benchmark set to set a reliable standard to perform bioactivity modeling on, and we think that Papyrus meets that goal.

Additionally repetition of data in the source datasets was scrutinized and where possible only the most recent bioactivity data was kept. This is for instance the case of the KIBA scores of Tang *et al*[6] that used a combination of activity types of ChEMBL to increase the quality of single measurements, or of ExCAPE-DB or the ChEMBL subset in Merget *et al*[14] aggregating data from ChEMBL versions 20 and 21 respectively. While Tang's data was kept as is and subsets in ExCAPE-DB and Merget *et al* were not, this phenomenon is not isolated and several activity values in Papyrus might have originated from the same source. This over representation of the same values would, in turn, bias the aggregated mean and standard deviation for specific compound-target pairs.

Stereochemical aspects were discarded in Papyrus to ensure that differing molecular standardisation processes of sources would not have an impact on the aggregation of activity values. However, the procedure of removing stereochemistry completely overlooked the potential of chiral molecules having opposing therapeutic or toxic effects and does not allow for the modelling of activity cliffs.

Another related shortcoming of Papyrus is its disregard for peptides and nucleic acids, even more so since one of the most abundant protein classes of level 3 are family A GPCR peptide receptors. This means that, though many drugs and compounds have been designed for these receptors, they do not have a single related data point in the dataset. In turn, related peptide derived models will only show limited performance. In a future version we would like to explore the possibility of increasing peptide representation in the Papyrus dataset, but for now this is what we settled on.

In a similar vein as the datasets, the descriptors that were added are a selection of descriptors that we frequently use. We believe that the provided set will be sufficient for anyone investigating a specific protein (family), and that high quality results can be obtained. However, we understand the need to tinker with all options of the process, and we separated the descriptors from the main dataset instead of adding them together. This gives the option for researchers to implement their own descriptors if so desired, while keeping the format of the original Papyrus dataset.

We have provided several implementations of filters, to narrow down the data for use in modeling (or perhaps other purposes). Using the entirety of Papyrus is not feasible without adequate computational resources, and we recommend users to reduce the data using the provided filters or in their own manner. It should be noted that the quality annotation filter does not imply that only high-quality data should be used, especially since classification models can leverage both the censored and binary data, the latter constituting more than 95% of the dataset.

Conclusion

We created an extensive benchmark set named Papyrus, that contains high quality data aggregated from multiple data sources. This standardised set is primarily used as a reliable data source for modeling ligand-protein interactions. We have shown the statistics of the Papyrus dataset and several classification and regression models using QSAR and PCM, with performance on par with prior results. We anticipate that the Papyrus dataset can be exploited in a myriad of ways and filtered or altered for specific research questions. We believe the strength of the dataset lies in its standardisation, normalisation and quality, while providing the necessary tools for further manipulation to specific needs.

Author’s Contributions

OJMB, BJB and GJPvW conceived the study. OJMB, BJB and WJ performed the experimental work and analysis. APIJ, BvdW and GJPvW provided feedback and critical input. All authors read, commented on and approved the final manuscript.

Data Availability

The Papyrus dataset can be found at https://doi.org/10.4121/16896406.v1. Python scripts can be found at https://github.com/OlivierBeq/Papyrus-scripts.

Acknowledgements

We would like to thank Marc Boef and Remco van den Broek for their input during creation. We would also like to thank Marina Gorostiola González for her input during the discussion about the work.

Funding

This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 777365 (“eTRANSAFE”). This Joint Undertaking receives support from the European Union’s Horizon 2020 research and innovation programme and EFPIA.

Competing Interests

The authors declare that they have no competing interests.

References

1. Hu, Y. & Bajorath, J. Growth of Ligand–Target Interaction Data in ChEMBL Is Associated with Increasing and Activity Measurement-Dependent Compound Promiscuity. *Journal of Chemical Information and Modeling* **52**, 2550–2558 (2012).

2. Cook, C. E. *et al.* The European Bioinformatics Institute in 2016: Data growth and integration. *Nucleic Acids Research* **44**, D20–D26 (2016).

3. Bento, A. P. *et al.* The ChEMBL bioactivity database: an update. *Nucleic Acids Research* **42**, D1083–D1090 (2014).

4. Gilson, M. K. *et al.* BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Research* **44**, D1045–D1053 (2016).

5. Papadatos, G., Gaulton, A., Hersey, A. & Overington, J. P. Activity, assay and target data curation and quality in the ChEMBL database. *Journal of Computer-Aided Molecular Design 2015 29:9* **29**, 885–896 (2015).

6. Tang, J. *et al.* Making Sense of Large-Scale Kinase Inhibitor Bioactivity Data Sets: A Comparative and Integrative Analysis. *Journal of Chemical Information and Modeling* **54**, 735–743 (2014).

7. Klaeger, S. *et al.* The target landscape of clinical kinase drugs. *Science* **358**, (2017).

8. Lenselink, E. B. *et al.* Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set. *Journal of Cheminformatics* **9**, 45 (2017).

9. Lenselink, E. B. *et al.* Beyond the Hype: Deep Neural Networks Outperform Established Methods Using A ChEMBL Bioactivity Benchmark Set. (2019) doi:10.4121/uuid:b64986dd-3203-445e-9b93-13a5ac7ef999.

10. Béquignon, O. *et al.* Papyrus - A large scale curated dataset aimed at bioactivity predictions. (2021) doi:10.4121/16896406.v1.

11. Sun, J. *et al.* ExCAPE-DB: an integrated large scale dataset facilitating Big Data analysis in chemogenomics. *Journal of Cheminformatics 2017 9:1* **9**, 1–9 (2017).

12. Sharma, R., Schürer, S. C. & Muskal, S. M. High quality, small molecule-activity datasets for kinase research. *F1000Research* **5**, (2016).

13. Christmann-Franck, S. *et al.* Unprecedently Large-Scale Kinase Inhibitor Set Enabling the Accurate Prediction of Compound–Kinase Activities: A Way toward Selective Promiscuity by Design? *Journal of Chemical Information and Modeling* **56**, 1654–1675 (2016).

14. Merget, B., Turk, S., Eid, S., Rippmann, F. & Fulle, S. Profiling Prediction of Kinase Inhibitors: Toward the Virtual Assay. *Journal of Medicinal Chemistry* **60**, 474–485 (2017).

15. van Westen, G. J. P., Wegner, J. K., IJzerman, A. P., van Vlijmen, H. W. T. & Bender, A. Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets. *MedChemComm* **2**, 16 (2011).

16. Gaulton, A. *et al.* ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Research* **40**, 1100–1107 (2012).

17. Kim, S. *et al.* PubChem in 2021: New data content and improved web interfaces. *Nucleic Acids Research* **49**, D1388–D1395 (2021).

18. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Research* **45**, D158–D169 (2017).

19. Tweedie, S. *et al.* Genenames.org: The HGNC and VGNC resources in 2021. *Nucleic Acids Research* **49**, D939–D946 (2021).

20. Bento, A. P. *et al.* An open source chemical structure curation pipeline using RDKit. *Journal of Cheminformatics* **12**, 51 (2020).

21. Sayle, R. A. & Jack Delany. Canonicalization and Enumeration of Tautomers. in *EuroMUG99* 28-29 October 1999 (1999).

22. RDKit: Open-source cheminformatics (version 2021.03.5). doi:10.5281/zenodo.5242603.

23. O'Boyle, N. M. *et al.* Open Babel: An Open chemical toolbox. *Journal of Cheminformatics* (2011).

24. The Open Babel Package, version 3.0.1.

25. O'Boyle, N. M. Towards a Universal SMILES representation - A standard method to generate canonical SMILES based on the InChI. *Journal of Cheminformatics* **4**, 1–14 (2012).

26. Ropp, P. J., Kaminsky, J. C., Yablonski, S. & Durrant, J. D. Dimorphite-DL: An open-source program for enumerating the ionization states of drug-like small molecules. *Journal of Cheminformatics* **11**, 1–8 (2019).

27. Probst, D. & Reymond, J.-L. A probabilistic molecular fingerprint for big data settings. *Journal of Cheminformatics 2018 10:1* **10**, 1–12 (2018).

28. Probst, D. & Reymond, J.-L. Visualization of very large high-dimensional data sets as minimum spanning trees. *Journal of Cheminformatics* **12**, (2020).

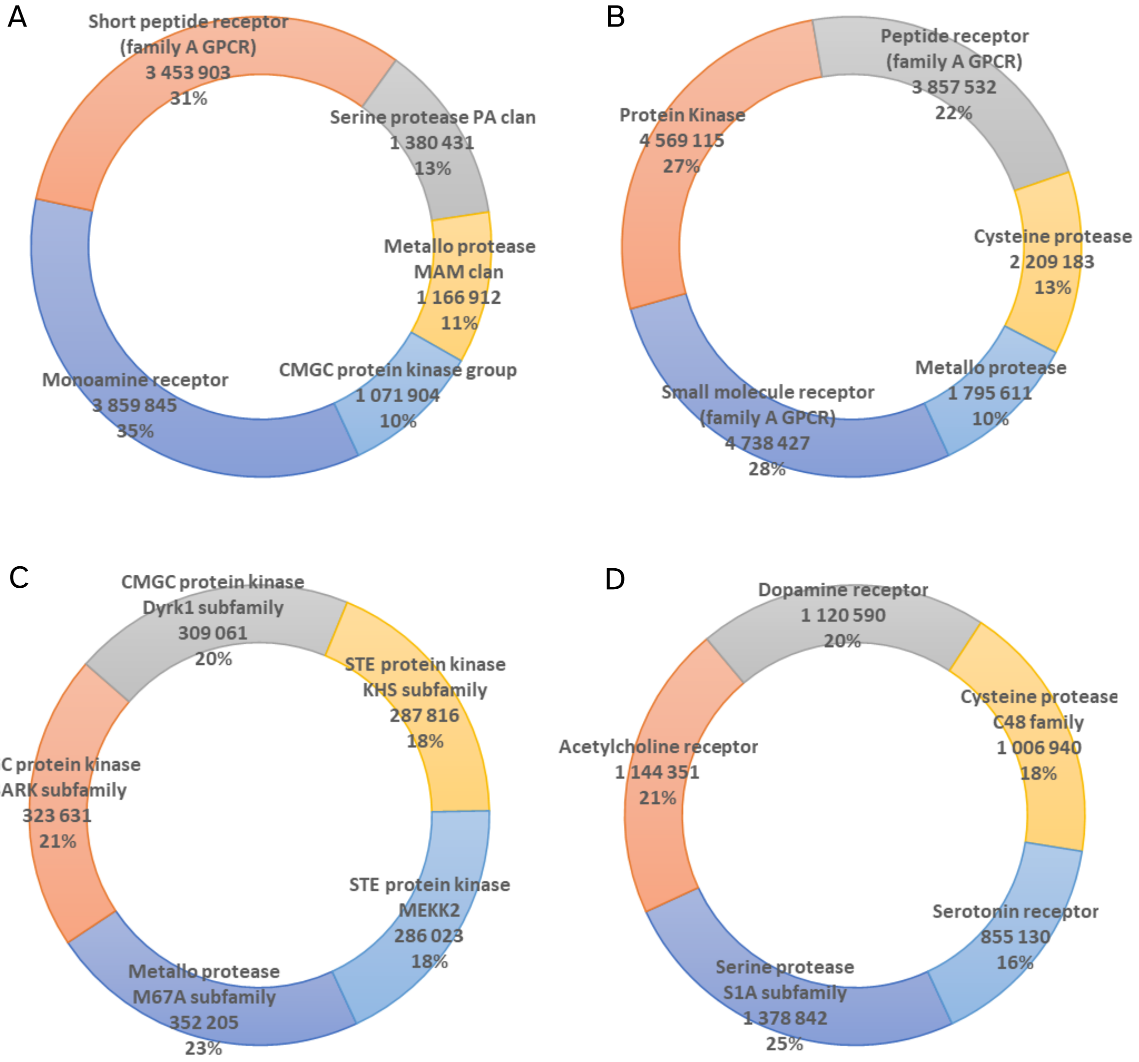
29. Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M. & Church, G. M. Unified rational protein engineering with sequence-based deep representation learning. *Nature Methods* **16**, 1315–1322 (2019).

30. Hong, H. *et al.* Mold2 , molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics. *Journal of Chemical Information and Modeling* **48**, 1337–1344 (2008).

31. Tang, J. *et al.* Making sense of large-scale kinase inhibitor bioactivity data sets: A comparative and integrative analysis. *Journal of Chemical Information and Modeling* **54**, 735–743 (2014).

32. Sheridan, R. P. Time-Split Cross-Validation as a Method for Estimating the Goodness of Prospective Prediction. *Journal of Chemical Information and Modeling* **53**, 783–790 (2013).

Additional information



Additional figure 1: Protein classification levels 3 (A), 4(B), 5 (C) and 6 (D) of targets in Papyrus.