Evaluation of the Papyrus dataset for kinase activity classification

Rachael Skyner & Ben Tehan

OMass Therapeutics, Building 4000, Chancellor Court, John Smith Drive, ARC Oxford, OX4 2GX

Introduction

- One of the most challenging parts of building models in AI/ML is the selection or construction of a suitable dataset.
- Publicly available datasets contain a trove of information on proteins, ligands, and their interactions, but the quality of data varies in quality and is subject to experimental error
- More focused datasets tend to be limited to a much smaller selection of information that may not be suitable for models requiring lots of data, such as neural networks.
- Béquignon et al. [1] have recently released the Papyrus dataset that aims to alleviate problems in data quality and range applied to bioactivity predictions. It consists of around 60 million data points, which have been standardized and normalized for application to machine learning, combining multiple large and small publicly available datasets.
- In this work, we investigate the utility of this new dataset to generating models for kinase binding classification.
- Kinases play a central role in virtually all signal transduction networks, and so are common targets in drug discovery. Molecules designed as kinase inhibitors often exhibit off-target binding for seemingly unrelated kinases, meaning an understanding of selectivity across the kinome for any potential selective inhibitor is crucial.
- In addition to evaluation of the dataset for predicting kinase inhibition, we also present an open-source adaptation of the dataset as a postgres database for integration with other frameworks, such as web-based tools via. integration with the RDKit cartridge [2] and razi [3].

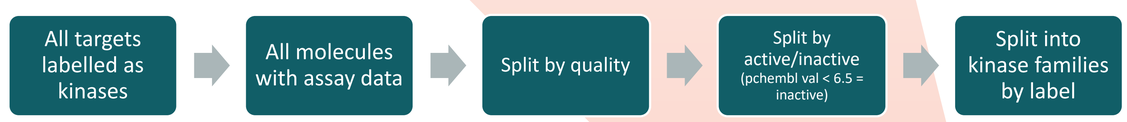
Methods

Database Preparation

- Papyrus data (version 05.6) was downloaded from zenodo [4]
- A data schema for a postgreql database was developed to remove repetition in the data and make it accessible with the rdkit cartridge. Models were implemented in sqlalchemy to allow programmatic access to the data (**github.com/reskyner/Papyrus_scripts** – src/papyrus_scripts/postgres)
- Scripts to reproduce the datasets and work presented here are available on GitHub, and data dump linked to on Github to Zenodo

Kinase Classification Dataset

- Summarised in the flowchart below (all data obtained from Papyrus postgres)



Comparative kinase classification dataset

- Difficult to find datasets with inactive/decoy molecules that won't all be accounted for in Papyrus
- Used KLIFs [5] database for examples of kinase binders in the pdb, and DUD-E [6] kinase decoys as approximations of inactive molecules

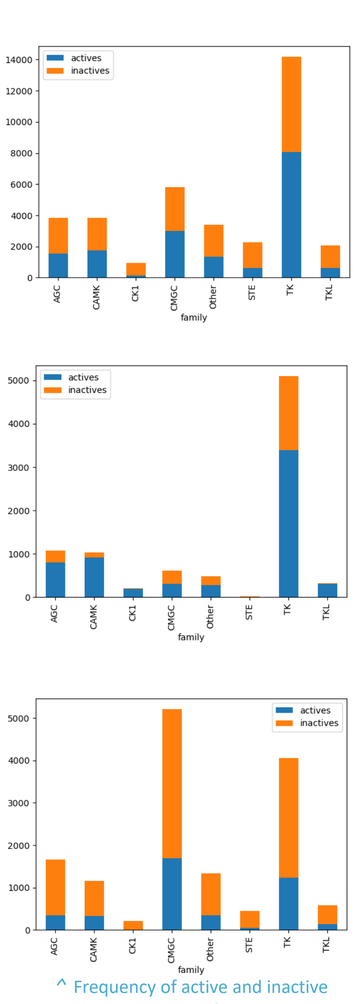
Classification Models

- All datasets were run through 9 classification estimators with hyperparameter selection done through a halving Cross-Validated grid search. (70:30 train/test splits), implemented in Sklearn
- Balancing of data by random selection of samples from major category to match no of datapoints in minor category)
- The 'best' models, selected by f1 score, were taken forward for analysis

Results

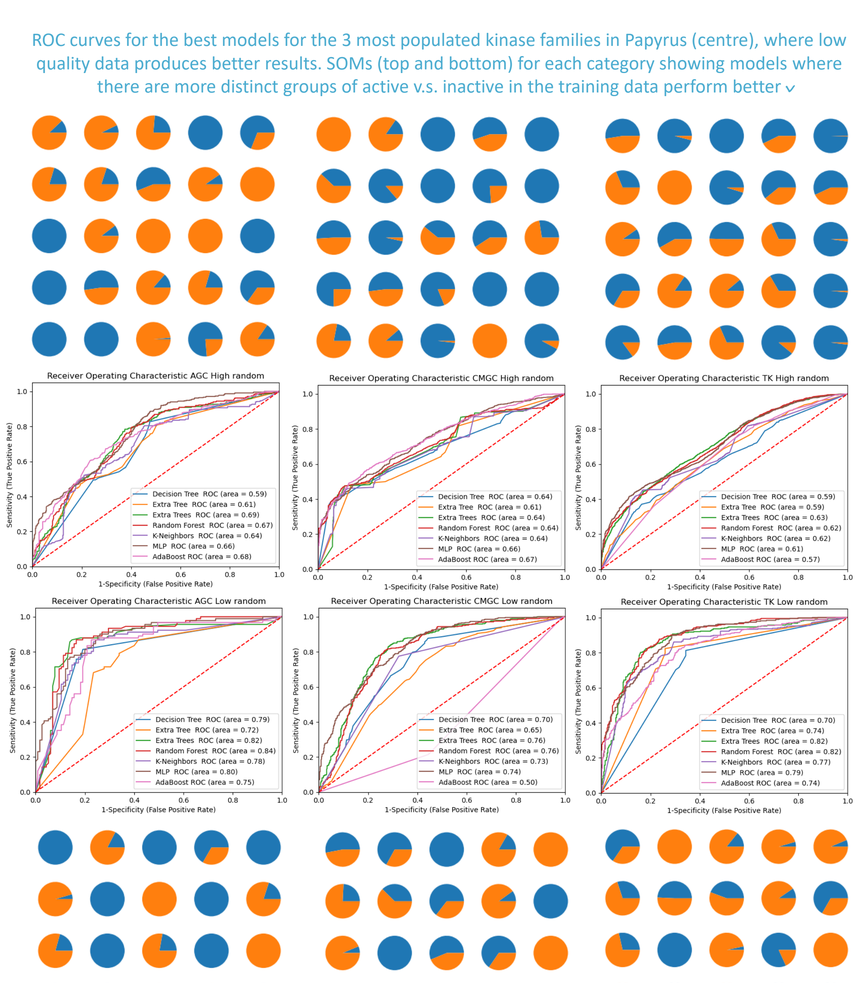
Improved data quality as labelled in Papyrus does not correspond to improved classification

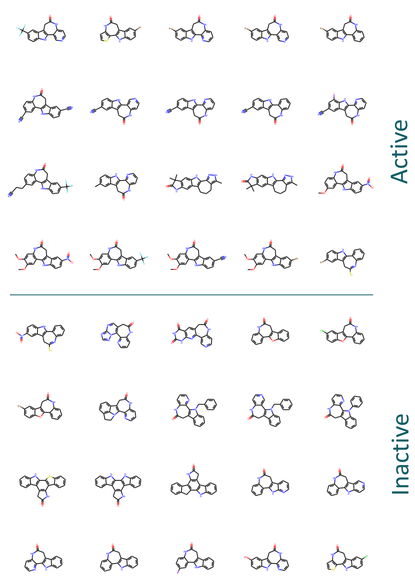
Data labelled as high quality does not produce better classification models than data labelled as low quality. **Better models are produced when there is more diversity between actives and inactives**. The SOMs for active(blue) and inactive (yellow) group molecules together into similar chemical space. Models where there are more distinct groups of active v.s. inactive in the training data perform better



^ Frequency of active and inactive datapoints by kinase family, split into High (top), Medium (middle) and Low (bottom) quality

ROC curves for the best models for the 3 most populated kinase families in Papyrus (centre), where low quality data produces better results. SOMs (top and bottom) for each category showing models where there are more distinct groups of active v.s. inactive in the training data perform better v





< An example of a group of molecules that are very similar to each other and contain both actives and inactives(by pchembl value threshold 6.5). This demonstrates where there are large areas of assay data for typical SAR-like explorations.

Dataset size does not correspond to improved classification

Larger datasets do not necessarily correspond to improved classification. For example in the AGC, CMCG and TK family datasets, there are more datapoints in the lower quality dataset. This means that neither data quality or size can explain model performance for the classification models in this work.

This is possibly because **large areas of assay data relating to kinase activity are data dumps from large companies**. This makes it likely that there are large areas of relevant chemical space that are not accounted for, but also **many molecules that are very similar to each other from typical exploration of SAR**. An example of a large number of similar molecules (grouped by morgan fingerprints by a SOM) is shown on the left, with both active an inactive examples. **If the majority of the datasets in this work consist of similar isolated areas of chemical space, they will not be generally applicable.**

Dataset	No. of molecules	No. of unique targets	Mols/targets	Best model	F1 score	ROC AUC
All Papyrus kinases HML	26,581	358	74.24	ExtraTrees	0.83	0.82
All Papyrus kinases HM	19,351	341	56.75	ExtraTrees	0.71	0.72
AGC Papyrus kinases H	2786	85	32.78	ExtraTrees	0.68	0.69
CMGC Papyrus kinases H	4092	82	49.90	AdaBoost	0.66	0.67
TK Papyrus kinases H	9977	171	58.35	ExtraTrees	0.64	0.63
AGC Papyrus kinases L	1526	73	20.90	Random Forest	0.82	0.84
CMGC Papyrus kinases L	4572	81	56.44	Extra Trees	0.77	0.76
TK Papyrus kinases L	3632	107	33.94	ExtraTrees	0.81	0.82
KLIFS/DUDE decoys	7812	319	24.44	RandomForest	0.93	0.93

^ Table showing all datasets produced in this work. For Papyrus datasets, quality labels are indicated by H (high) M (medium) and L (low). An estimation of molecules per target is given simply by division of molecules in the dataset over targets in the dataset. In all but one case, tree-based models outperform other types of classification estimator. Inclusion of low-quality data often improves model quality.

Structural evidence of binding is a better metric for binary classification than pchembl value

For comparison, a custom dataset consisting of the KLIFs dataset as actives, and the DUDE kinase as decoys was constructed, and models generated in the same way as for the papyrus data. This model was by far the best model. **However, the DUDE decoys are generated with respect to specific targets, and there are far fewer targets in DUDE than any of the Papyrus datasets, so the models may have learned the differences between the two original datasets it was constructed from**. Selection of a suitable threshold for pchembl value (active/inactive) needs further investigation.

Conclusions

- Improved data quality as labelled in Papyrus does not correspond to improved classification
- Better models are produced when there is more diversity between actives and inactives
- Dataset size does not correspond to improved classification
- A lot of the data in the original dataset are large areas of assay data relating to kinase activity from data dumps from large companies
- This sort of data means that there are many molecules in each dataset that are very similar to each other from typical exploration of SAR
- If most of the datasets in this work consist of similar isolated areas of chemical space, they will not be generally applicable.
- Structural evidence of binding is a better metric for binary classification than pchembl value
- However, the DUDE decoys are generated with respect to specific targets, and there are far fewer targets in DUDE than any of the Papyrus datasets, so the models may have learned the differences between the two original datasets it was constructed from
- Selection of a suitable threshold for pchembl value (active/inactive) needs further investigation.

References

1. Béquignon, O.J.M., Bongers, B.J., Jespers, W. et al. Papyrus: a large-scale curated dataset aimed at bioactivity predictions. J Cheminform 15, 3 (2023). https://doi.org/10.1186/s13321-022-00672-x
2. https://www.rdkit.org/docs/Cartridge.html
3. https://github.com/rvianello/razi/tree/master
4. Dataset - Papyrus - A large scale curated dataset aimed at bioactivity predictions 10.5281/zenodo.7373213
5. Kanev, G. K., de Graff, C., Westerman, B.A., de Esch, I.J.P and Kooistra, A.J. KLIFS: an overhaul after the first 5 years of supporting kinase research, Nucleic Acids Research, Volume 49, Issue D1, 8 January 2021, Pages D562–D569, https://doi.org/10.1093/nar/gkaa895
6. Mysinger MM, Carchia M, Irwin JJ, Shoichet BK J. Med. Chem., 2012, Jul 5. doi 10.1021/jm300687e