Effect of missing data on multitask prediction methods

Antonio de la Vega de León

RSC-NSFC Symposium 2019

Sheffield Chemoinformatics Research Group

The University Of Sheffield.

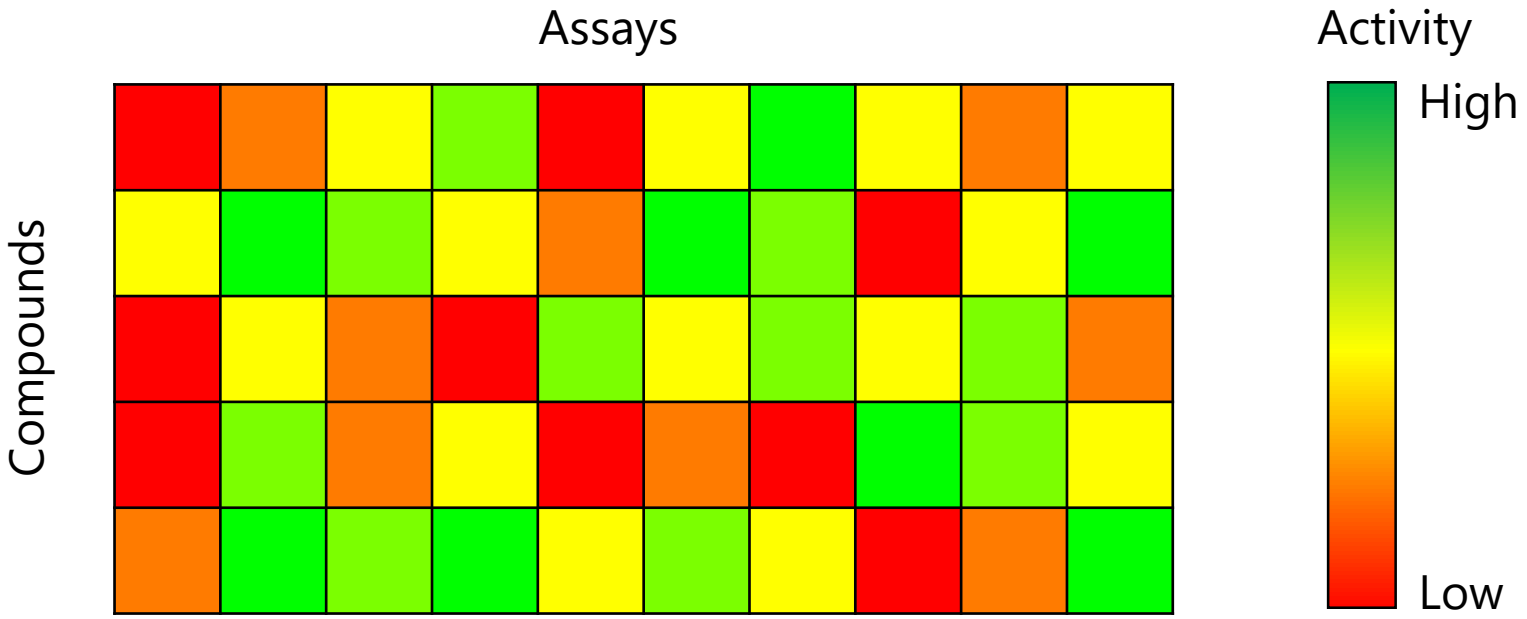
Introduction

Modelling of complex biological systems requires bioactivity data from several targets

When using public data sources, the resulting data sets are sparse, not all compounds are tested on all targets

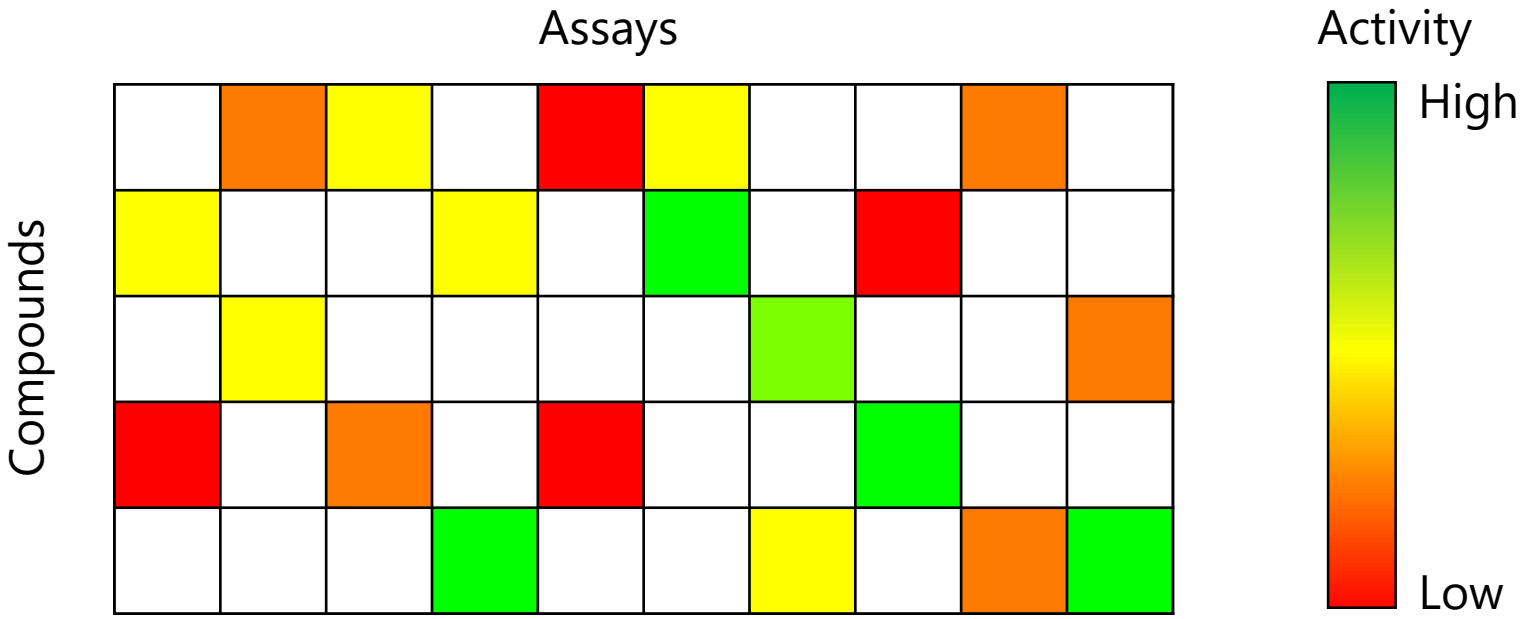
Data matrices

Data set is represented as a matrix



Data sparseness

How does increasingly sparser data matrices affect the performance of machine learning methods?



Aim of the study

Our aim was:

Measure how much performance is lost in multitask prediction when data is sparse

Estimate at which point there is enough data that measuring more is no longer efficient use of resources

Complete data sets

Complete data sets are used as a baseline

PKIS (Published kinase inhibitor set)

- Regression task, predict percent inhibition
- 367 compounds, 454 assays

HTSFP

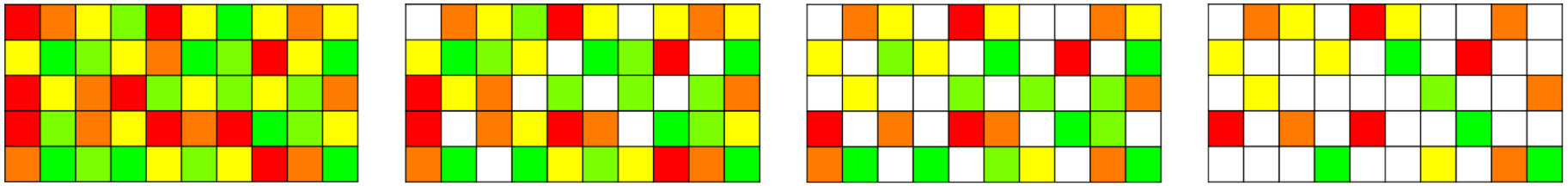
- Classification task, predict active/inactive in PubChem assays
- 49,715 compounds, 5 assays

Molecules represented with hashed Morgan radius 2 (ECFP4)

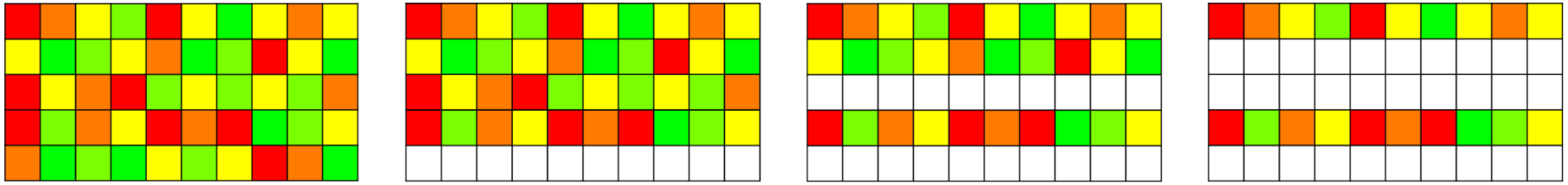
Data removal models

Data is removed to generate sparse data sets

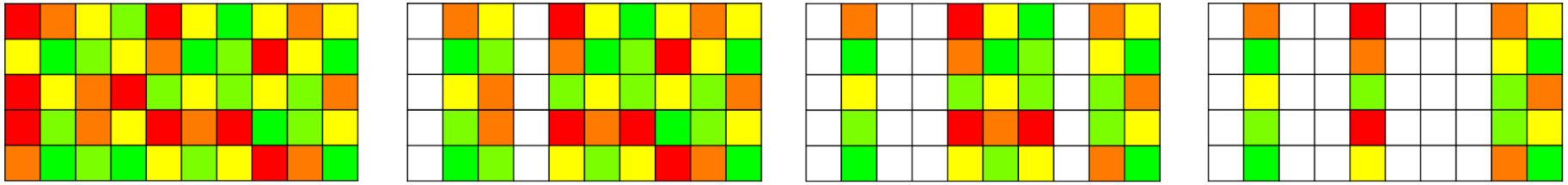
- Label removal



- Compound removal



- Assay removal (only on PKIS)



Multitask machine learning methods

Deep neural networks

- Fully connected, feedforward architecture
- Cost function is insensitive to missing data

Macau

- Based on Bayesian probabilistic matrix factorization
- Fingerprints added as side information to compounds

Random Forest

- Only partial comparison, as it can't train on sparse data

Experimental setup

For each method, a first exploration of hyperparameter effect on performance is used to select good value ranges

Based on these, 10 sets of hyperparameters are randomly chosen

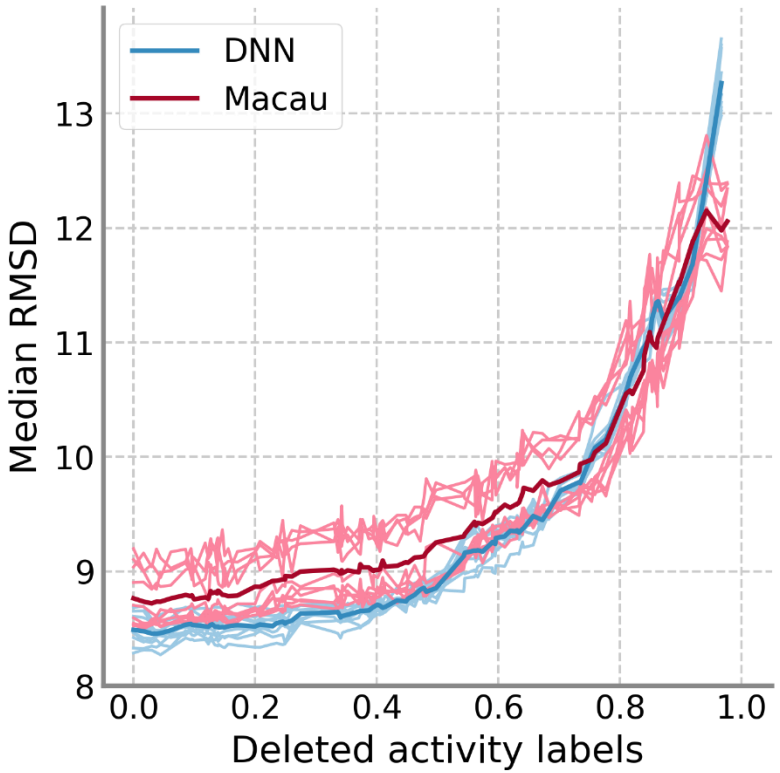
¼ of data taken as test data and rest as training data

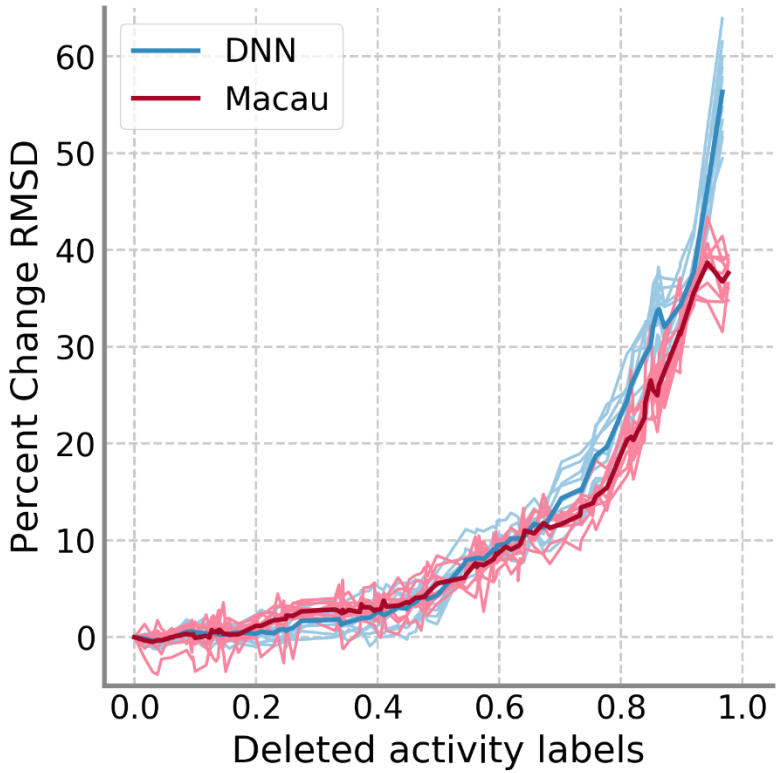
Several models (40 for HTSFP and 101 for PKIS) trained with increasing amounts of training data removed

Performance measured as RMSD (PKIS) and MCC (HTSFP)

PKIS with label removal

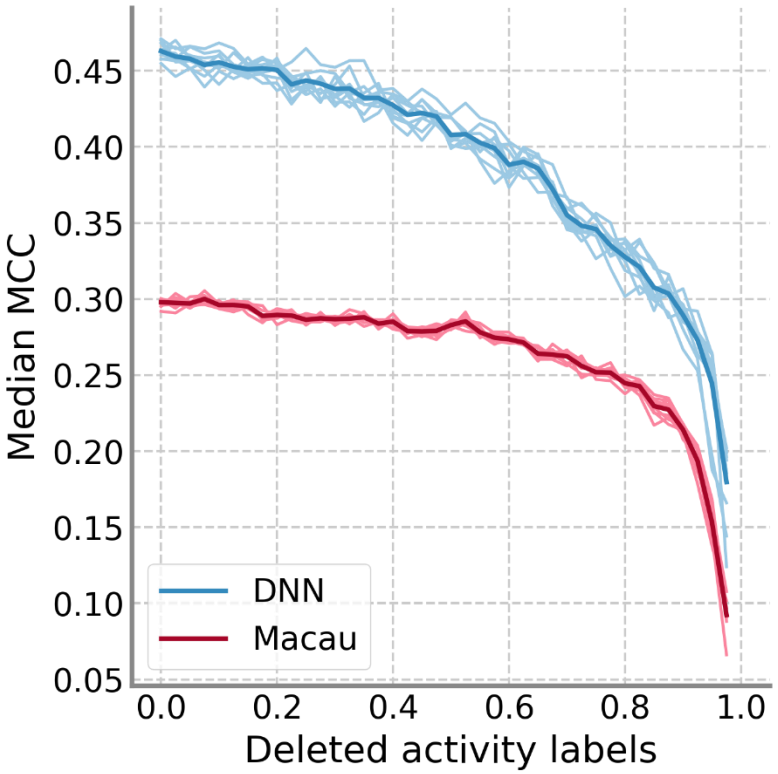
Both multitask methods see large drops in performance only after large amounts of training data were removed

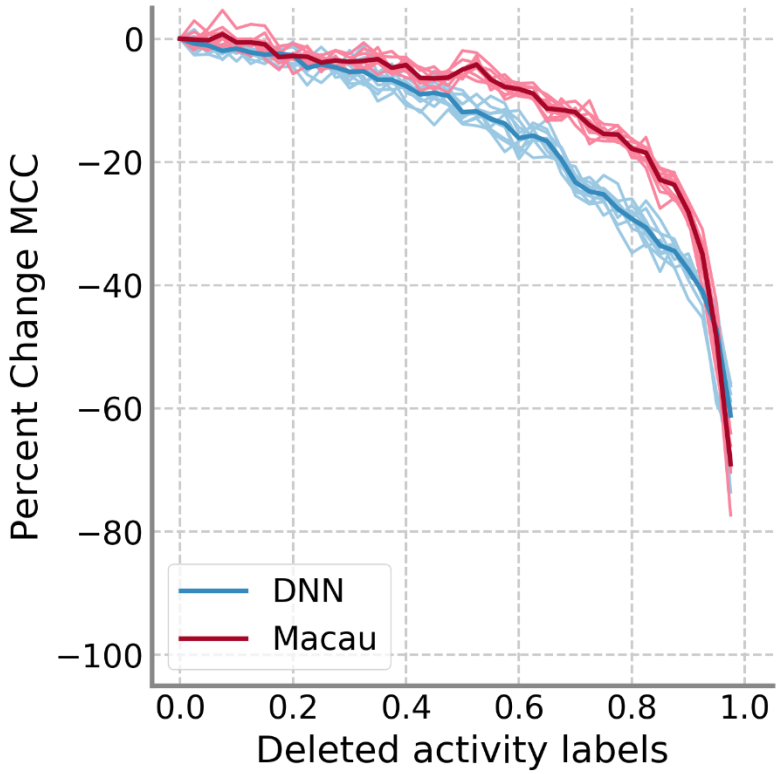




HTSFP with label removal

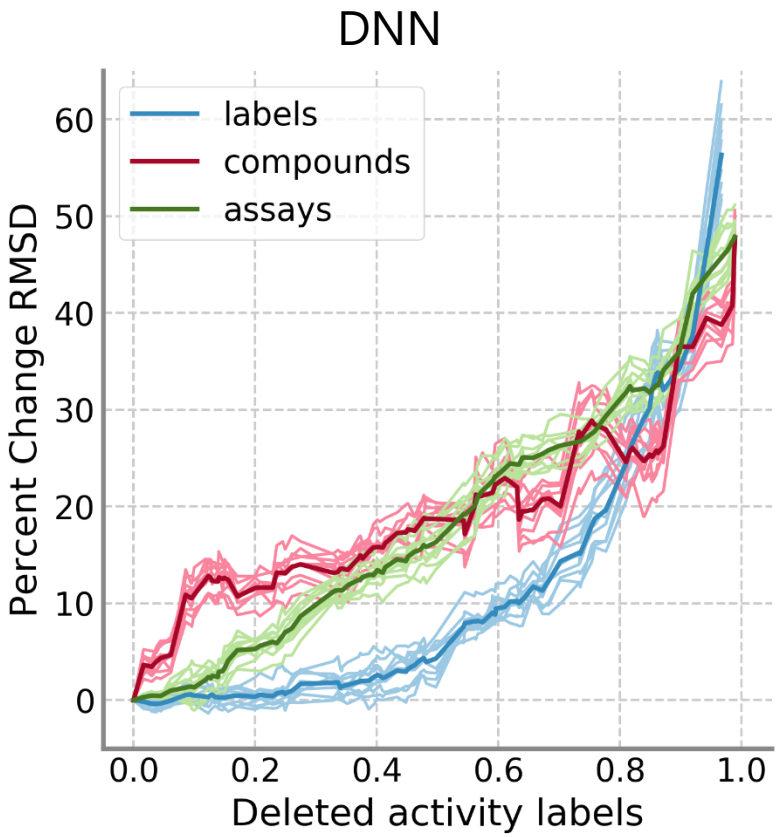
Although absolute values are quite different, performance progression still shows similar behaviour

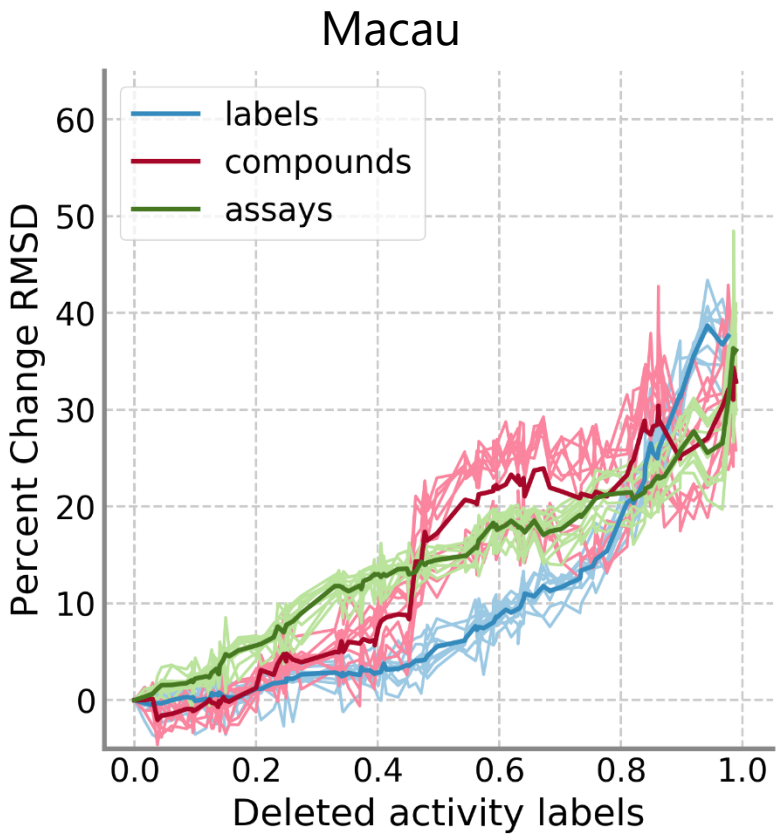




Comparison of data removal models

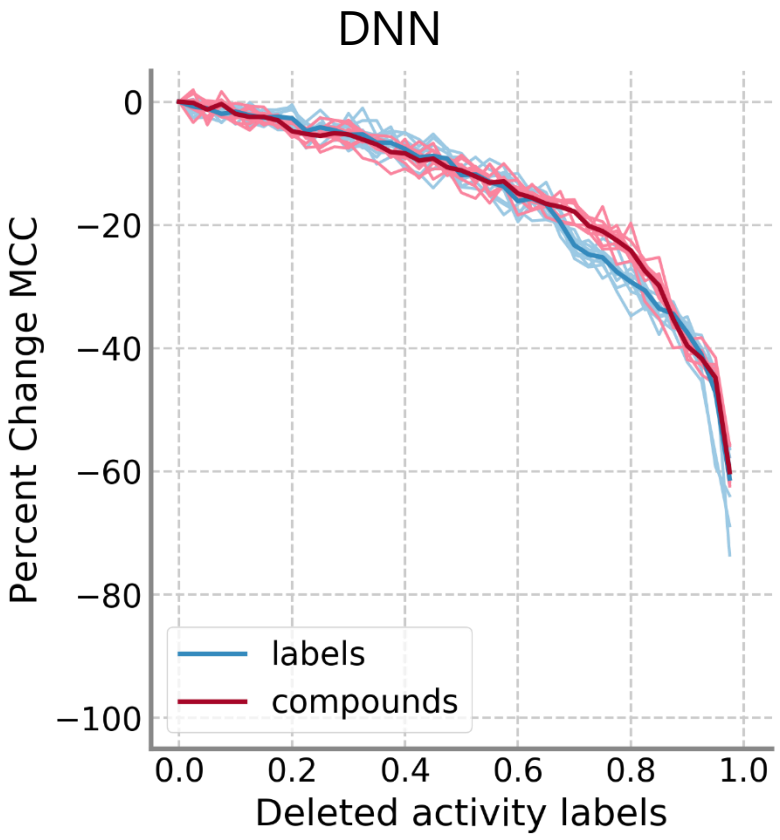
On PKIS, compound and assay removal models lead to worse performance progression

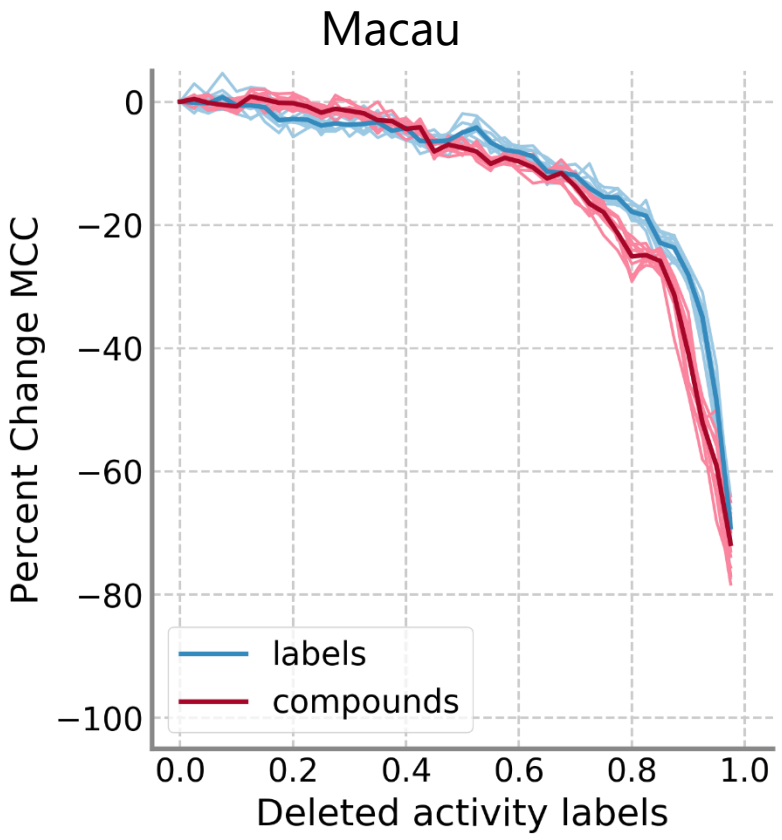




Comparison of data removal models

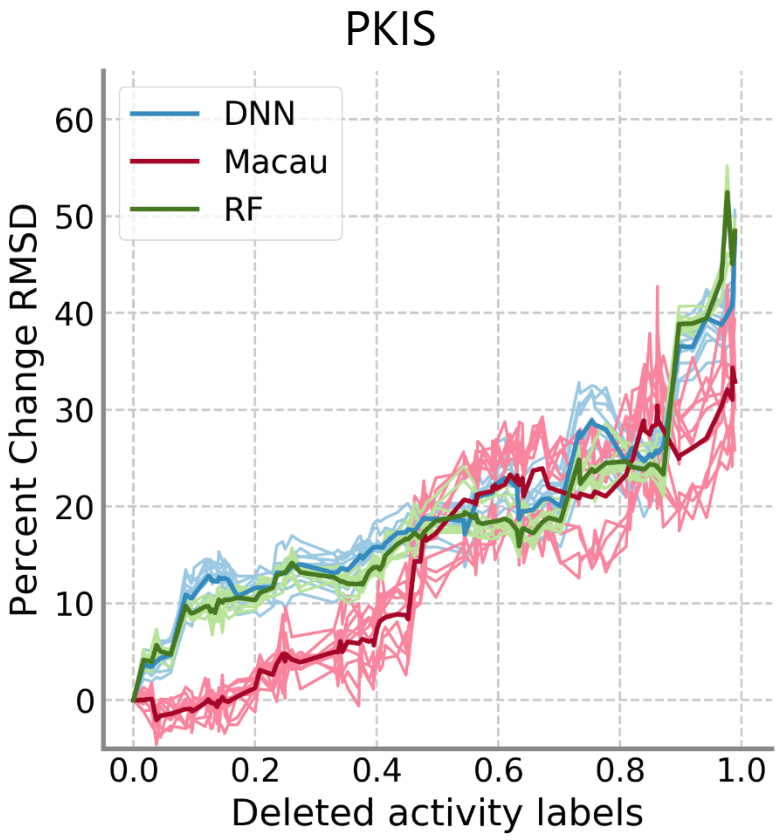
On HTSFP, compound removal models follows a very similar performance progression

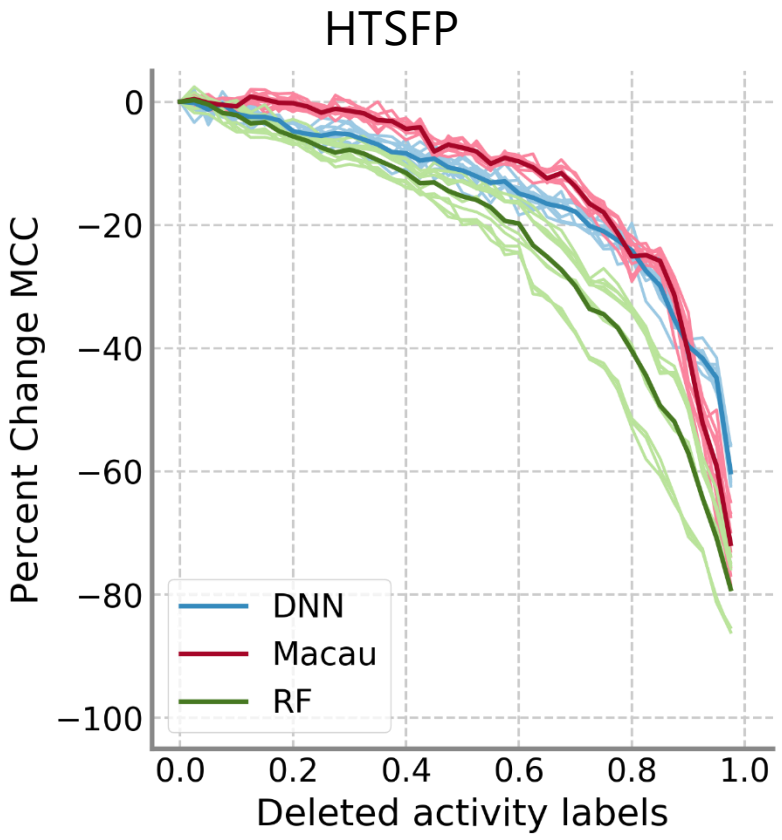




Random Forest with compound removal

The performance progression of Random Forest is quite similar to that of DNN





Conclusions

Multitask prediction methods are resistant to loss of performance because of missing data

This is a first estimate of how much performance could be gained by additional data

On regression, DNN and Macau performed very similar on absolute terms

Research was published in J Cheminf (DOI 10.1186/s13321-018-0281-z); the data, code, and results are publicly available at our Github, and archived in Zenodo (DOI 10.5281/zenodo.1230488)

Acknowledgements

University of Sheffield:

- Prof. Dr. Val Gillet
- Prof. Dr. Beining Chen

Eli Lilly:

- Dr. David Evans
- Dr. Matthew Baumgartner

Colleagues from the Information School and the Chemistry Department at Sheffield

Funding for this research has been provided by the European commission under IAPP Grant No. 612347