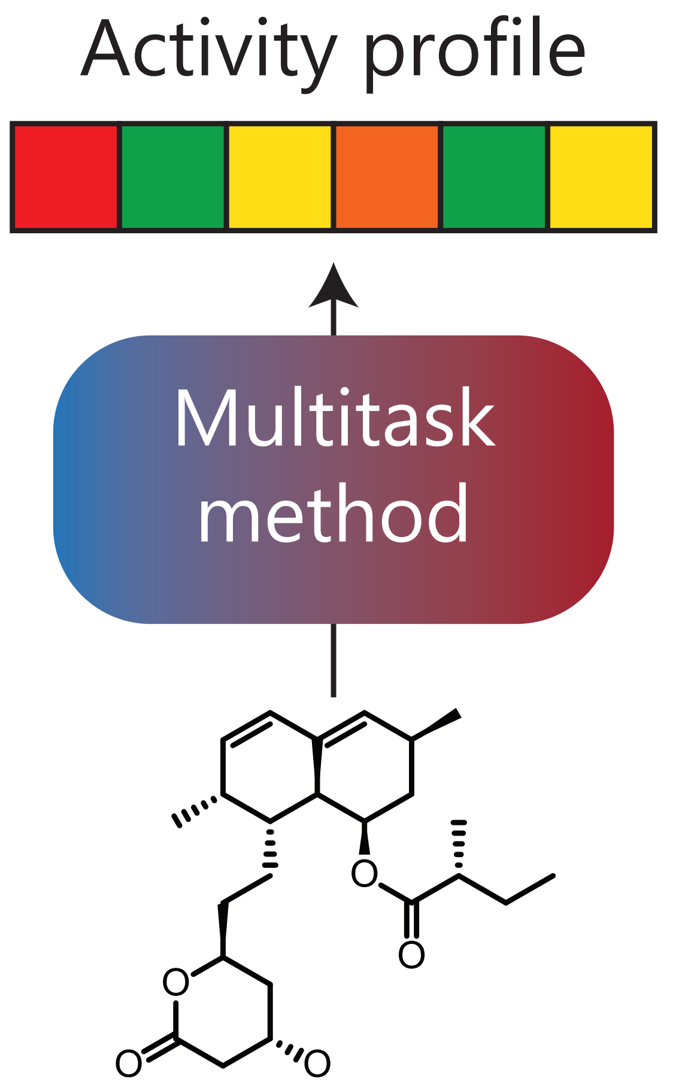
# Multitask machine learning for sparse chemical data sets How much data is enough?

Antonio de la Vega de León, Valerie J. Gillet   Information School, University of Sheffield

## Introduction

Multitask machine learning techniques are able to predict with a single model an activity profile, e.g.: a set of activity values against different kinases. They have become popular in chemoinformatics because of the rising interest in deep neural networks.[1] These techniques are frequently trained on sparse data sets, where not all compounds have been tested across all assays. The effect of the sparsity of the data on the performance of multitask prediction performance has seen little research.



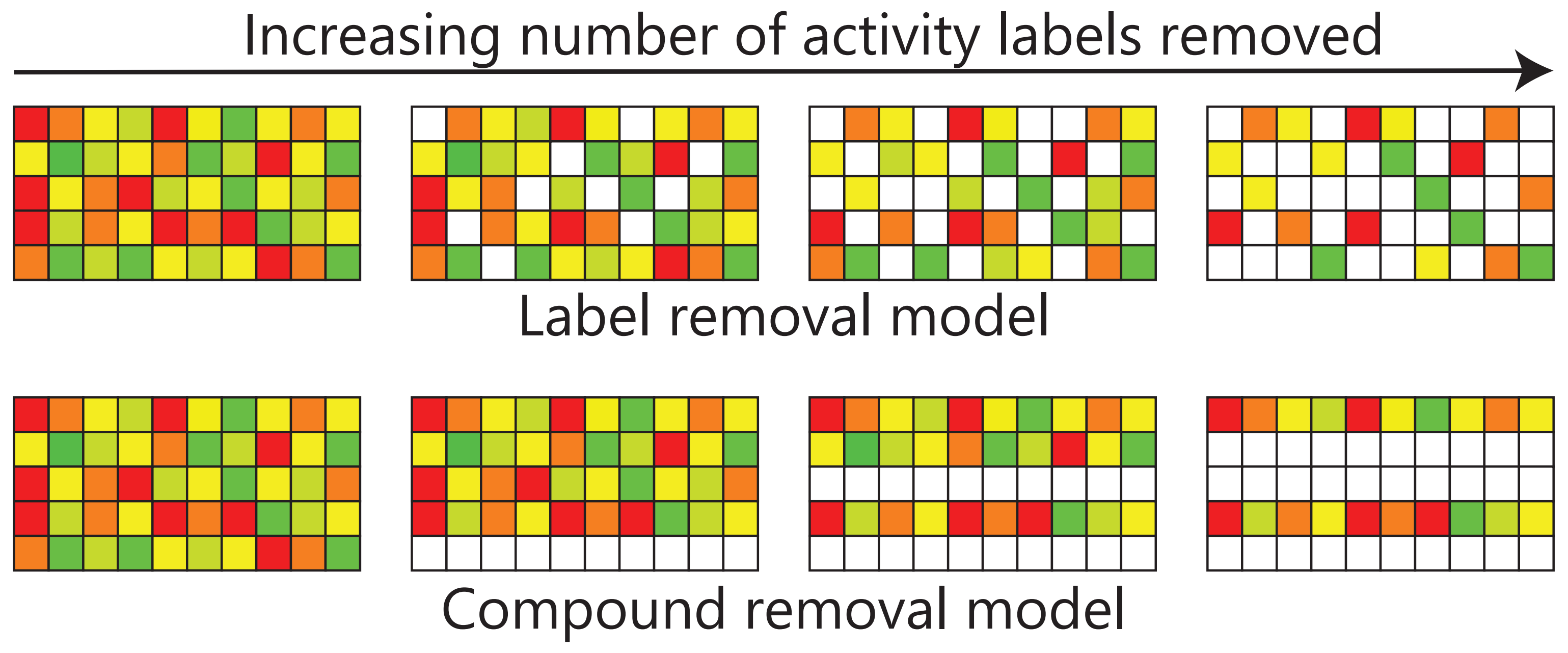
Activity profile

Multitask method

### Our main goal was: How performance changes as increasingly large amounts of training data are removed

## Datasets

Two complete datasets were assembled: PKIS[2], a kinase profiling data set (regression task, 367 cpds, 454 assays); and HTSFP, a set of PubChem assays based on a previous publication[3] (classification task, 49 713 cpds, 5 assays). The data was randomly split into training and test data with a 3:1 ratio. Sparse training sets were generated by removing increasing numbers of activity labels from the complete training data. Two removal models were compared: removing individual labels (label removal model) or whole compounds (compound removal model).



Increasing number of activity labels removed

Label removal model

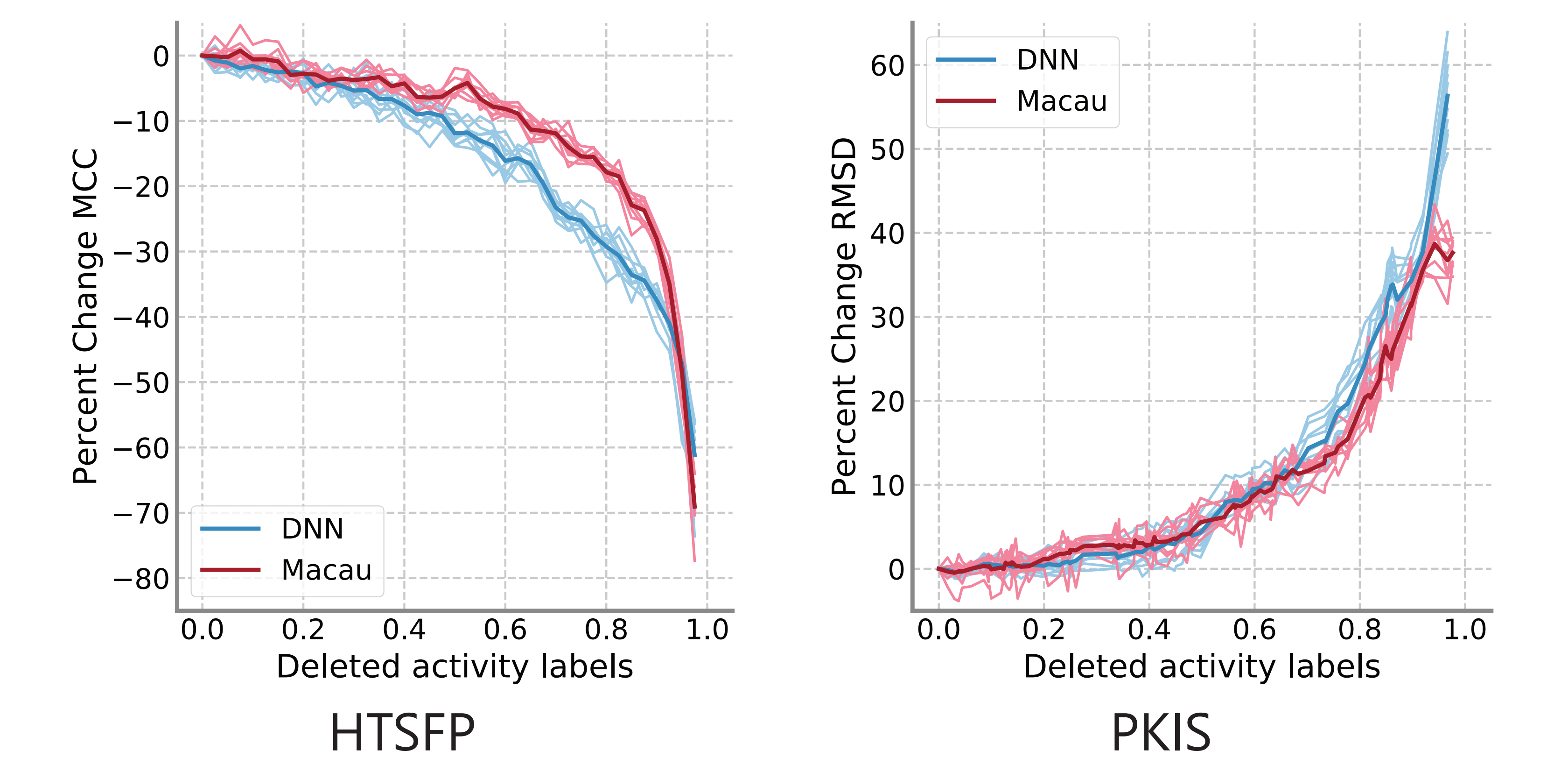
Compound removal model

## Methods

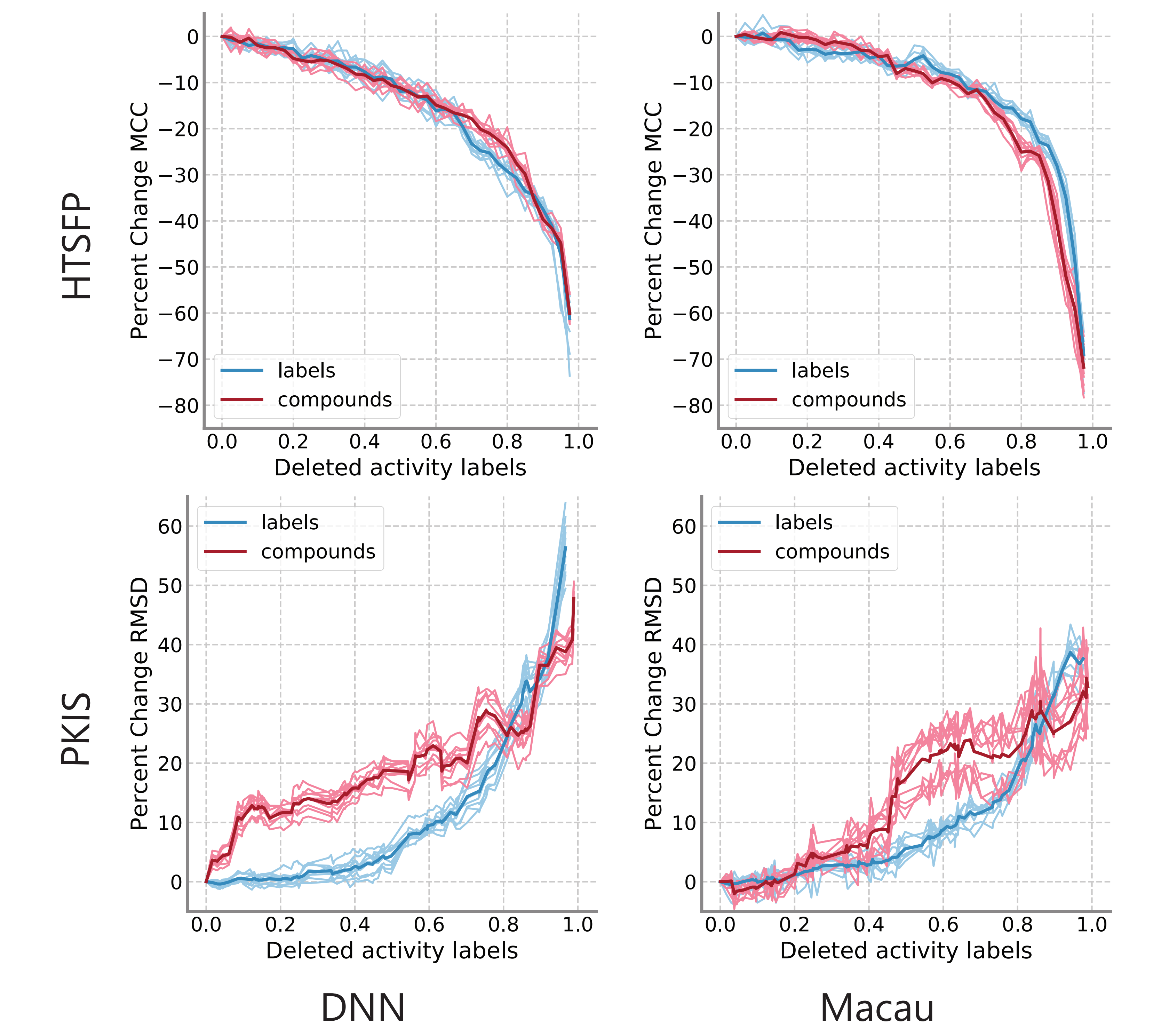
Two different multitask methods were tested: deep neural networks (DNNs) and Macau[4]. Ten sets of hyperparameters were chosen for each method based on a large parameter space random search. Performance was measured as median RMSD and median MCC across all assays for regression and classification data, respectively. Performance values were scaled relative to the performance on the complete data set.

## Results

### The performance progression of both techniques is very similar. The decrease in performance is at first slow but accelerates after 80% of the training data is removed.



### The effect of the data removal model is different in each dataset. On PKIS there is a large effect while on HTSFP there is little difference.



## Conclusion

Multitask prediction methods are very promising in chemoinformatics. The removal of up to 60% of the training data decreased performance only by 10%. Macau and DNN had very similar performance progression as training data was removed.

Data sets, methods and results are available at https://github.com/SheffieldChemoinformatics/missing-data-multitask-methods

### References

1) LeCun et al. Nature 521, 436-444, 2015
2) https://www.ebi.ac.uk/chembldb/extra/PKIS
3) Helal et al. JCIM 56, 390-398, 2016
4) Simm et al. arXiv 1509.04610, 2015

### Funding

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n°612347.