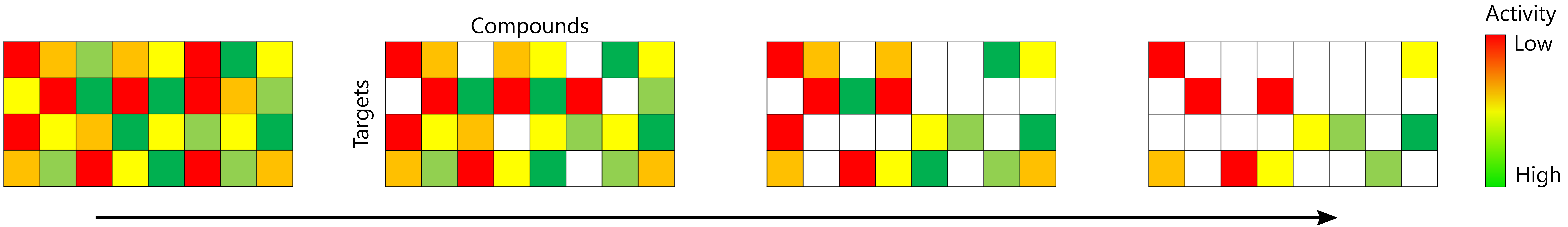
Deep learning application to aid phenotypic assay campaigns with public chemical data

Antonio de la Vega de León & Valerie J. Gillet
Information School, University of Sheffield, Regent Court, 211 Portobello, S1 4DP Sheffield, United Kingdom

Introduction

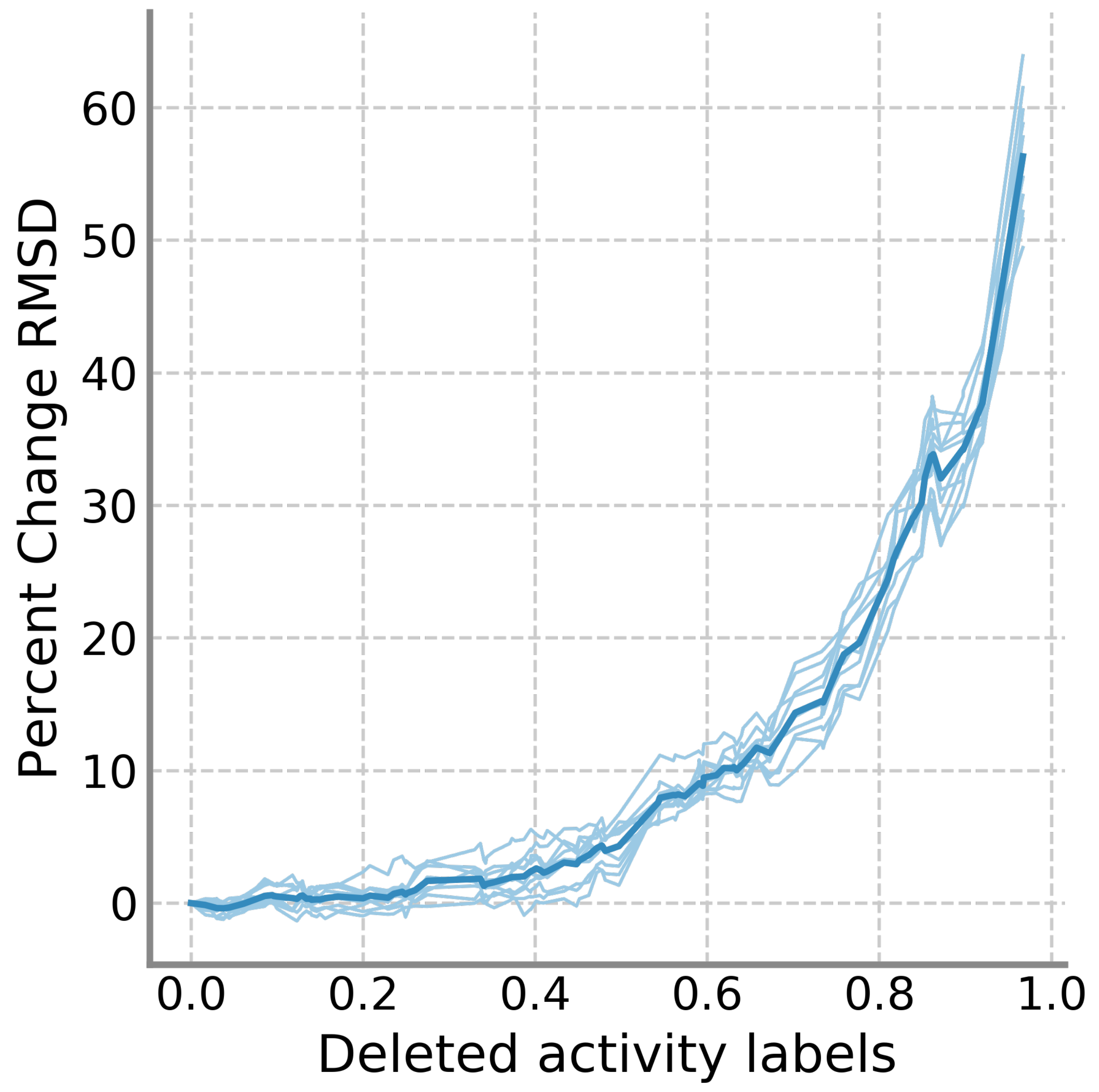
Current modelling of biological pathways using public data sources yields data matrices that are incomplete; that is, not all compounds have been tested against all targets. Predictive models built from these data sets are expected to have lower accuracy compared to a complete data set, but little attention has focused on by how much. Here we investigate the loss of performance through progressive deletion of data using deep neural networks (DNNs). DNNs have previously shown good performance in multitask prediction on chemical data.[1,2] DNNs were implemented in Tensorflow[3]. The PKIS[4] data set was used as a large complete data matrix, which contains 367 compounds with percent inhibition values for 454 kinase assays.



Performance of predictive models?

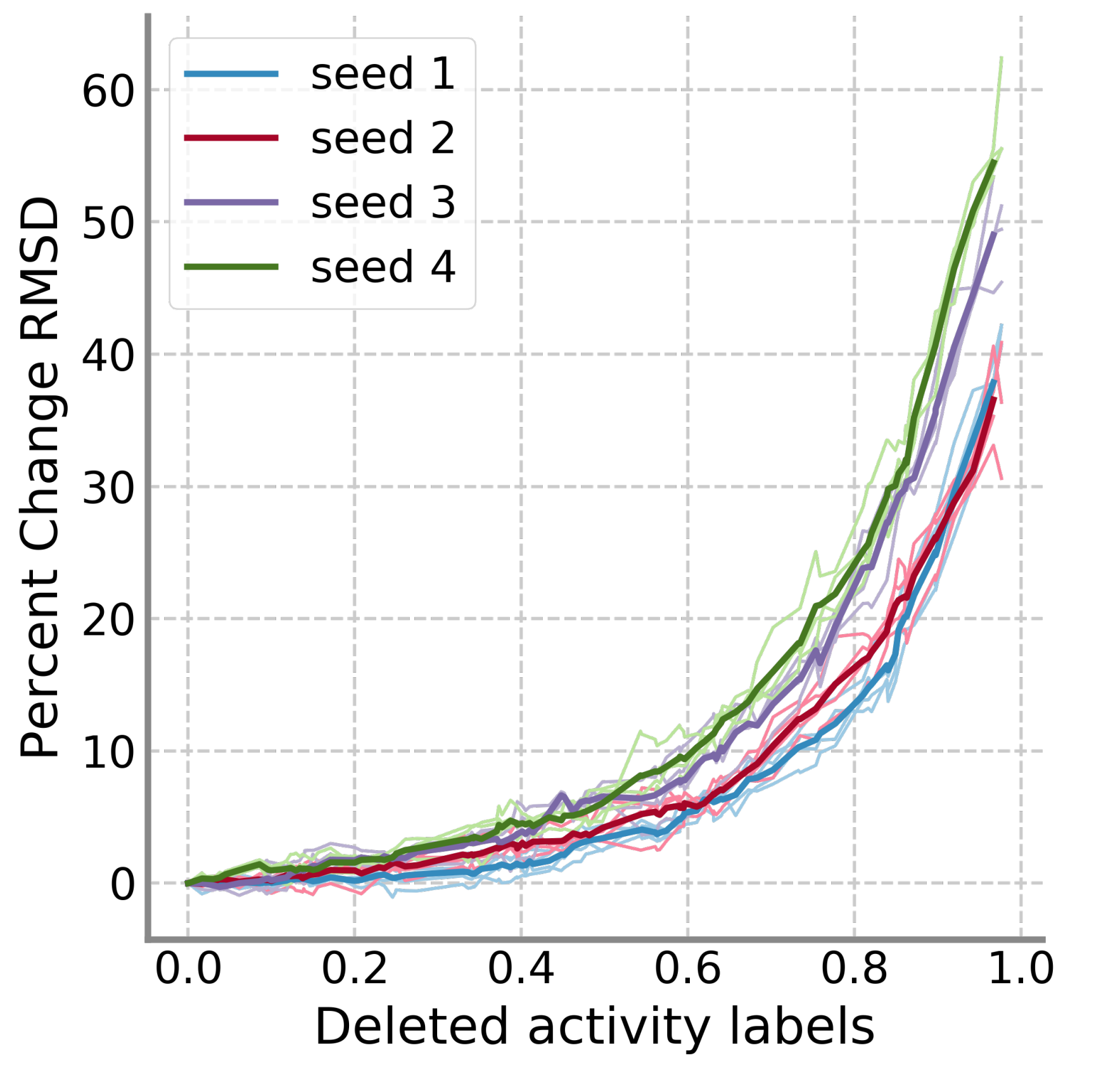
Simulating sparse data sets

Increasingly large amounts of activity labels in the training set were removed. The performance decrease is at first very slow. The median RMSD increases only by 10% when 60% of training activity labels are deleted. The same trend in performance loss can be seen in models trained on ten different sets of hyper-parameters.



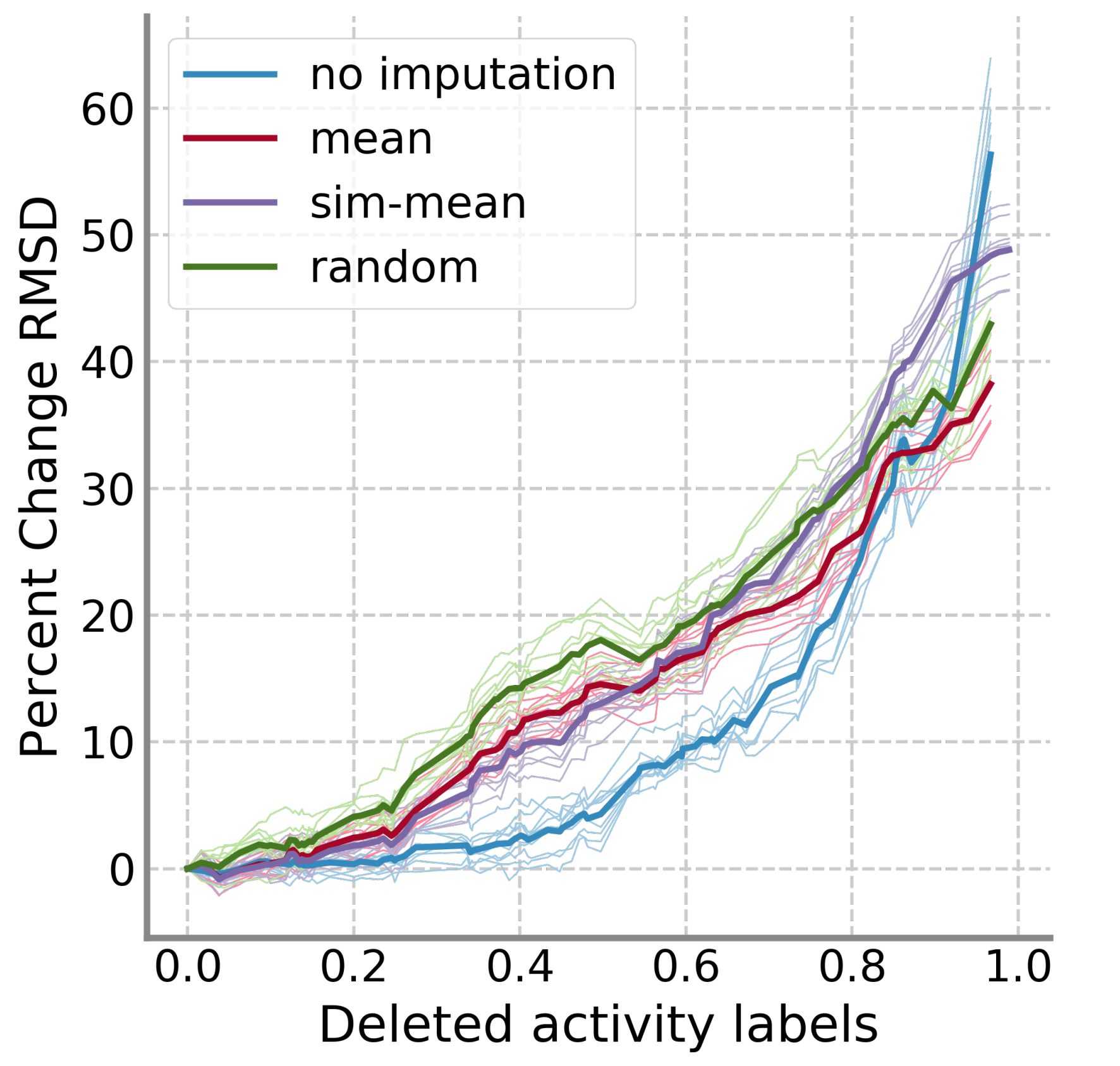
Effect of data split

The effect is not dependent on random compound selection. The performance change trend is very similar even if we vary the train/test split and which labels are chosen to be deleted.



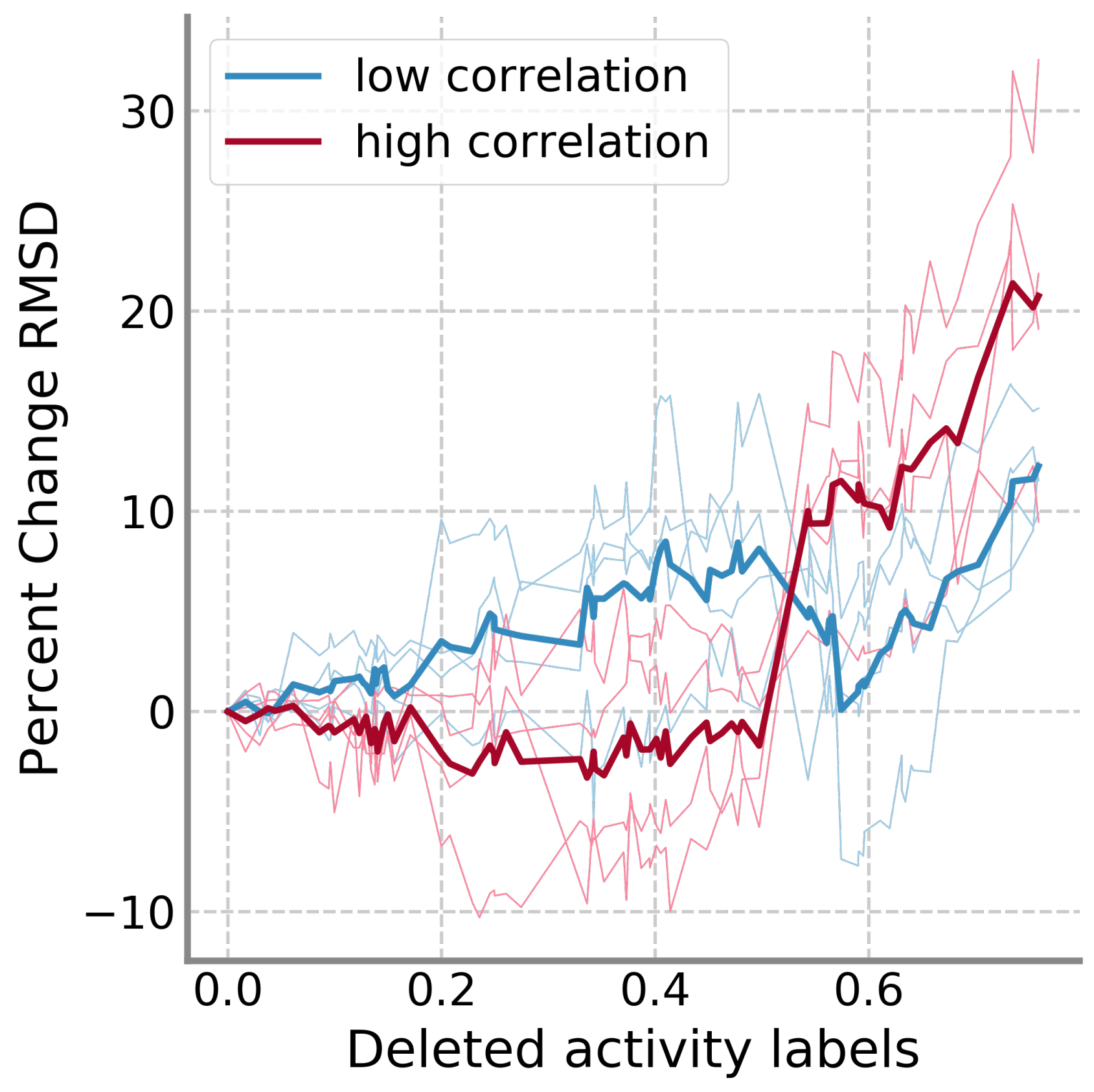
Imputing missing data

Different imputation methods (the mean, a similarity-weighted mean, or random values from the assay) used to fill deleted activity labels did not improve the predictive performance.



Subset analysis

Recent analysis has shown correlation between assays to be important to multitask DNN performance.[5] Subsets of assays based on high or low average correlation show very different trends in the slope of the performance change.



Conclusions

Deep neural networks are resistant to loss of performance due to missing data, irrespective of network parameters or data partitioning. This effect could be related to correlation between assays, following results published recently.[5] In our testing, imputation of missing data was counterproductive. Overall, deep neural networks constitute a promising technique in modelling biological pathways using chemical data.

References

(1) Ramsundar *et al.* J. Chem. Inf. Model., 2017, 57 (8), 2068–2076
(2) Lenselink *et al.* J. Cheminf. 2017, 9:45
(3) www.tensorflow.org (accessed 16/10/17)
(4) www.ebi.ac.uk/chembldb/extra/PKIS (accessed 16/10/17)
(5) Xu *et al.* J. Chem. Inf. Model., 2017, 57 (10), 2490–2504

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.