Deep Learning Based on Sparse and Noisy Data to Improve Predictions of Compound Activities

Thomas Whitehead*, Benedict Irwin†, Peter Hunt†, Matthew Segall†, Gareth Conduit‡

*Intellegens Limited, Cambridge, UK †Optibrium Limited, Cambridge, UK ‡Cavendish Laboratory, University of Cambridge

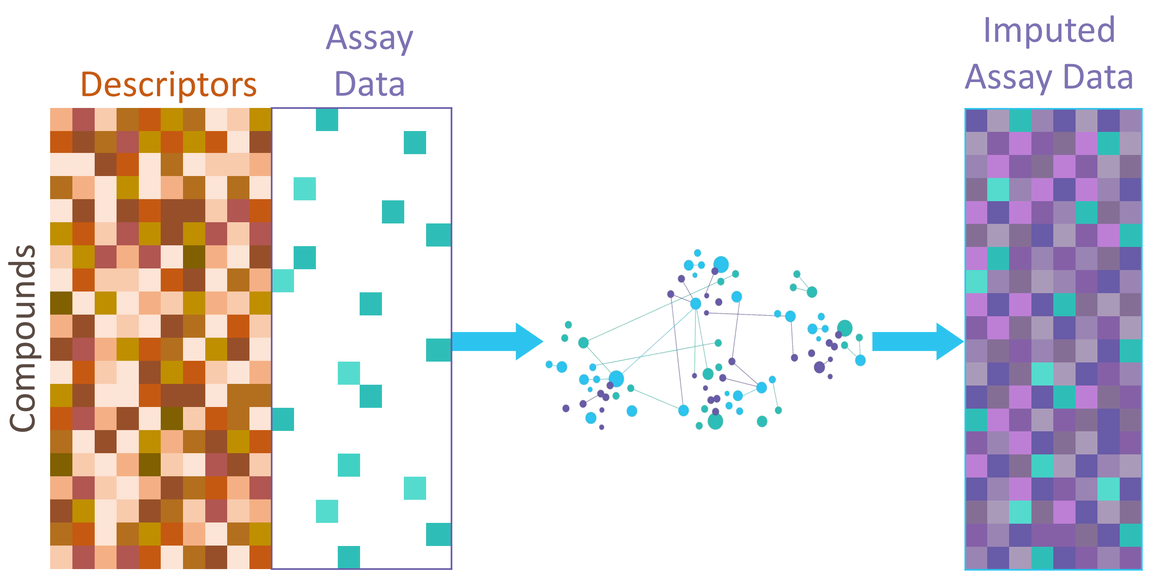
Introduction

The data available in drug discovery are sparse; a large pharma company's collection may contain millions of compounds and thousands of experimental endpoints, however only a small fraction of the possible compound-assay combinations will have been measured.

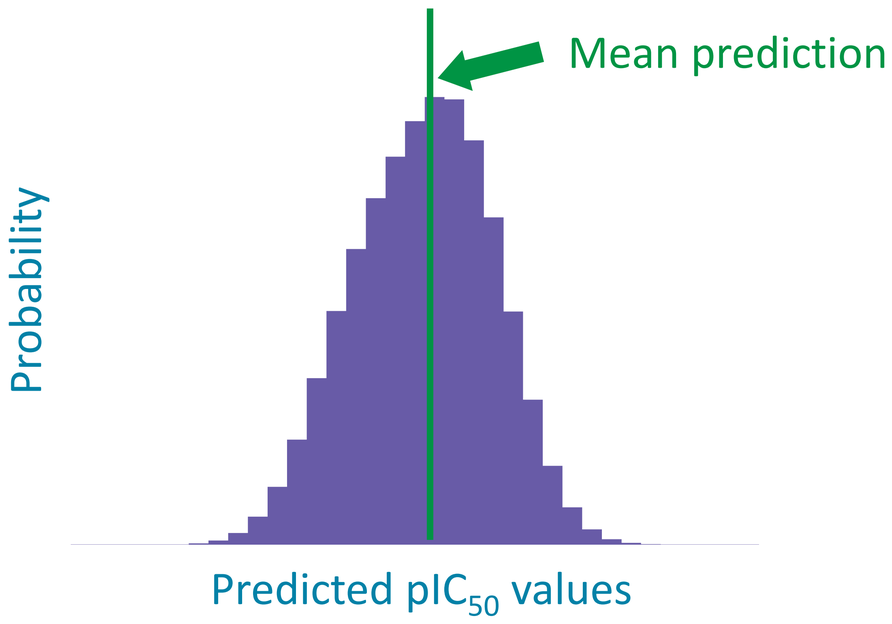
We introduce Alchemite™, a novel deep learning neural network method that, unlike conventional machine learning approaches, can train directly on sparse, noisy bioactivity data [1]. In combination with molecular descriptors, this enables it to learn immediately from correlations between activities measured in different assays as well as structure-activity relationships. Furthermore, the model provides a robust estimate of the confidence in each prediction, enabling attention to be focused on only the most accurate results.

Method: Alchemite

A novel deep neural network is trained on molecular descriptors and sparse experimental bioactivity data as inputs with which to impute the missing bioactivity values.



An ensemble of networks generates a probability distribution for each individual prediction, accounting for uncertainties in both the experimental data and due to extrapolation of the training data. From this, a confidence in each prediction can be assessed.



Data Set

The methods described above were applied to a challenging data set, published by Martin et al. [2], in which the training set compounds are not representative of the test set. This contains ~13,000 compounds and pIC_{50} values from 159 kinase assays, but only 6.3% of the possible compound-assay pairs have measured data.

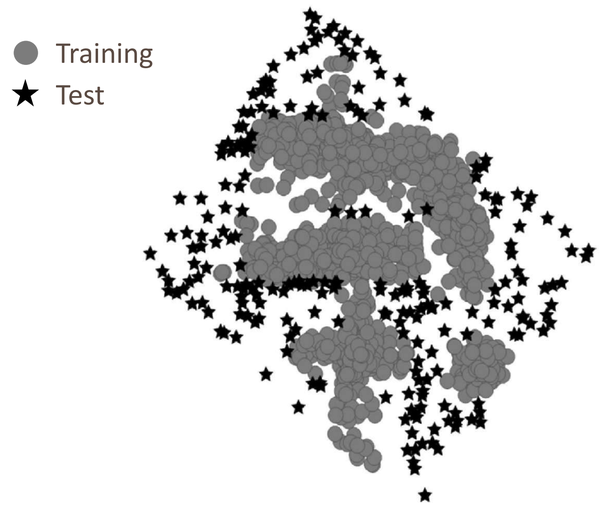
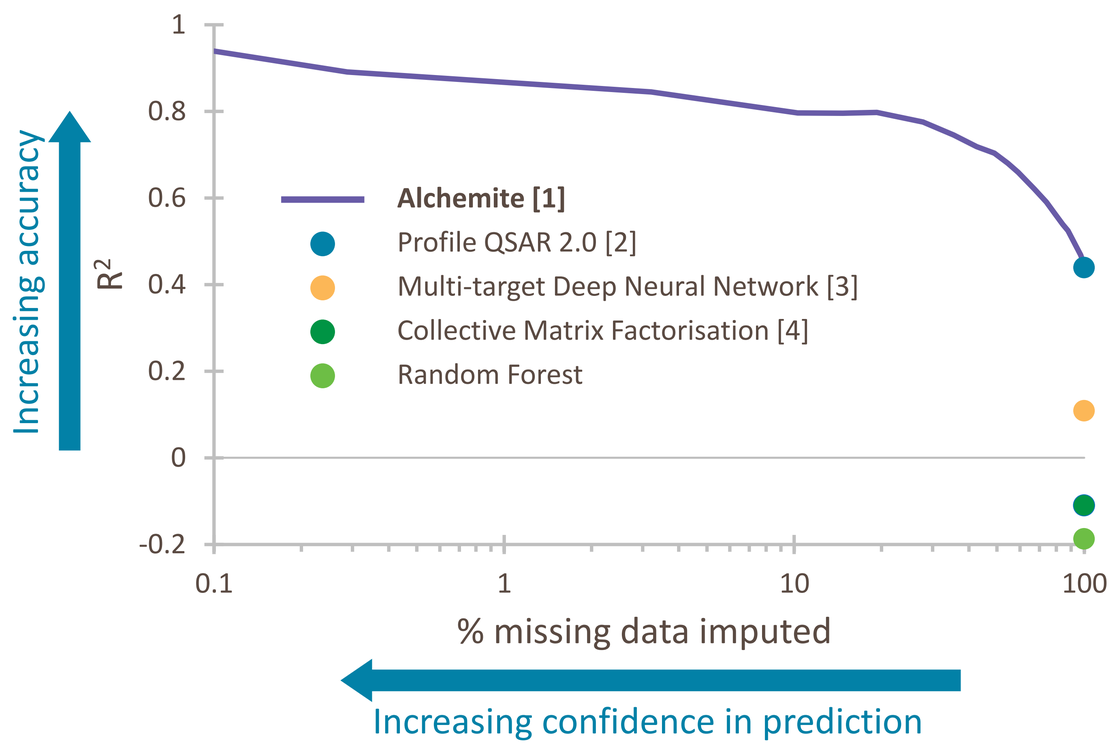


Figure reproduced from Martin *et al.* [2]

Results

The accuracy of five machine learning methods when applied to the realistically split, independent test set is shown below.



Alchemite and pQSAR 2.0 significantly outperform the other methods.

Furthermore, Alchemite can discard the predictions with the highest uncertainty, resulting in an increase in the accuracy of the remaining predictions. For example, the most confident 50% of the neural network's predictions have $R^2 > 0.7$, a typical threshold for accurate, reliable predictions.

Conclusion

We have presented a new neural network imputation technique, Alchemite [1], which can learn simultaneously from incomplete bioactivity data and molecular descriptors, resulting in a significant improvement in accuracy over conventional QSAR models. It can also accurately estimate the confidence in each individual prediction, identifying the most accurate results. In the example presented, this delivered a nine-fold increase in the number of accurate predictions, relative to the original sparse experimental measurements.

References

[1] Whitehead et al. (2019) J. Chem. Inf. Model. DOI: 10.1021/acs.jcim.8b00768
[2] Martin *et al.* (2017) J. Chem. Inf. Model. **57**(8), pp. 2077-2088
[3] Abadi *et al.* (2015) https://www.tensorflow.org/
[4] Cortez (2018) CoRR arXiv:1809.00366