Imputation Versus Prediction and Applications in Drug Discovery

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

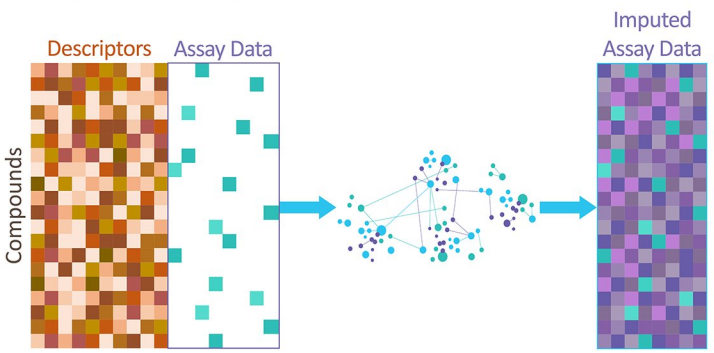
* Optibrium Limited †Intellegens Limited ‡Cavendish Laboratory, University of Cambridge

Introduction

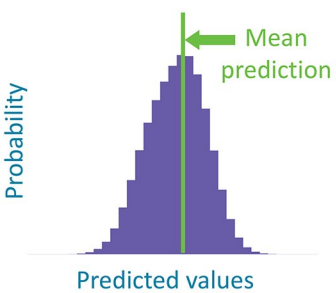
'Imputation' describes the process of filling in the gaps in a data set comprising multiple experimental endpoints, where values have not yet been measured, using the limited data that are already present. By taking these sparse data as inputs, imputation models can directly 'learn' from correlations between experimental endpoints. This **approach gains more value from the sparse and noisy data available in drug discovery** than conventional quantitative structure – activity relationship (QSAR) methods that use only descriptor – endpoint correlations. We have previously demonstrated a method for data imputation using deep learning and compared it with other methods for both imputation and prediction using public domain data sets [1]. However, benchmarking data sets are not representative of the real data available in drug discovery organisations and ongoing projects. In this poster, we will describe some **practical applications** of the Alchemite™ method for deep learning imputation.

Methods

A novel deep neural network is trained using **molecular descriptors and sparse experimental data as inputs** with which to impute the missing values [1].



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



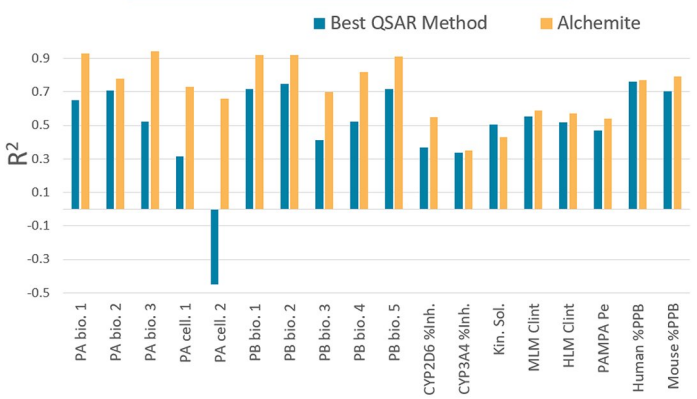
Example 1: Discovery Project Data

Alchemite™ was applied to **heterogeneous data** for 2,453 compounds, including target and phenotypic activities and ADME endpoints from two projects.

The performance of Alchemite™ was compared with four QSAR methods and achieved an **average coefficient of determination (R^2)** of **0.72** compared with 0.50 for the best QSAR method for each endpoint.

Full details of this study, including analysis of the chemical space and uncertainty estimates can be found in reference [2], or watch a webinar describing this project in detail at

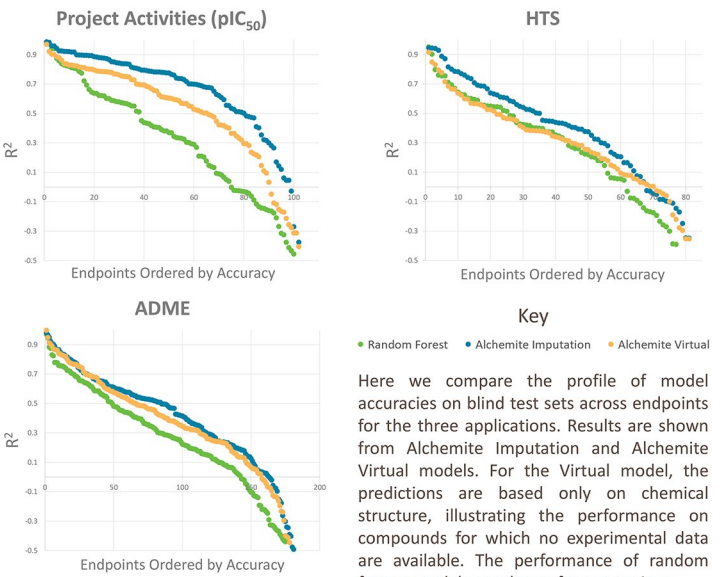
bit.ly/practical_deep_learning



Example 2: Global Pharma Data

Alchemite™ can also be applied to much larger data sets. In this example, a total of **678,994 compounds and 1,166 heterogenous endpoints** were modelled, covering applications to prediction of **project target activities, high-throughput screening (HTS) data and ADME assays**. For full details of this project, you can view a webinar at

bit.ly/large_scale_imputation



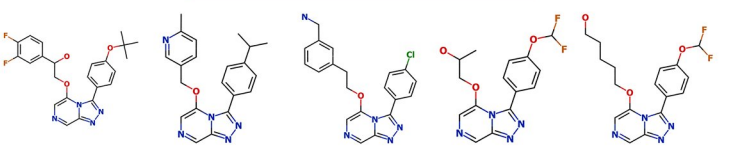
Here we compare the profile of model accuracies on blind test sets across endpoints for the three applications. Results are shown from Alchemite Imputation and Alchemite Virtual models. For the Virtual model, the predictions are based only on chemical structure, illustrating the performance on compounds for which no experimental data are available. The performance of random forests models are shown for comparison.

Example 3: Combined with Generative Methods

In this example, an Alchemite™ model was generated based on **sparse anti-malarial data**, provided by the Open Source Malaria (OSM) project. The resulting model was used to prioritise novel compound ideas generated with the Nova™ module in StarDrop™ [3]. A compound **confidently predicted** to have a good activity profile and physicochemical properties was synthesised and tested by OSM and **experimentally confirmed to be active**.

More details of this project are provided in reference [4], or you can watch a webinar at

bit.ly/AI_guided_design



Optibrium/ Intellegens	Davy Guan	Exscientia	Molomics	Molomics
0.65 µM	>25 µM	11 µM	>25 µM	14 µM

Conclusion

We have outlined three practical examples of the application deep learning imputation to drug discovery and demonstrated a substantial improvement over conventional predictive modelling in a wide range of scenarios, including project optimisation based on small data sets, modelling large pharma-scale data and guiding the design of new compounds based on sparse, noisy data.

The ability to extract more information from the sparse, noisy data that are typical in drug discovery, and robust estimates of the confidence in each prediction, enable better decision-making to quickly target high-quality compounds and prioritise experimental effort.

References

[1] Whitehead *et. al.* J. Chem. Inf. Model (2019) **59**(3) pp. 1197–1204

[2] Irwin *et al.* J. Chem. Inf Model. (2020) **60**(6), pp. 2848–2857

[3] https://www.optibrium.com/stardrop/stardrop-nova.php

[4] Irwin *et al.* Future Drug Discovery (2020) **2**(2) DOI: 10.4155/fdd-2020-0008