A Single Deep Learning Model for Confident Imputation of Heterogeneous Drug Discovery Endpoints

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

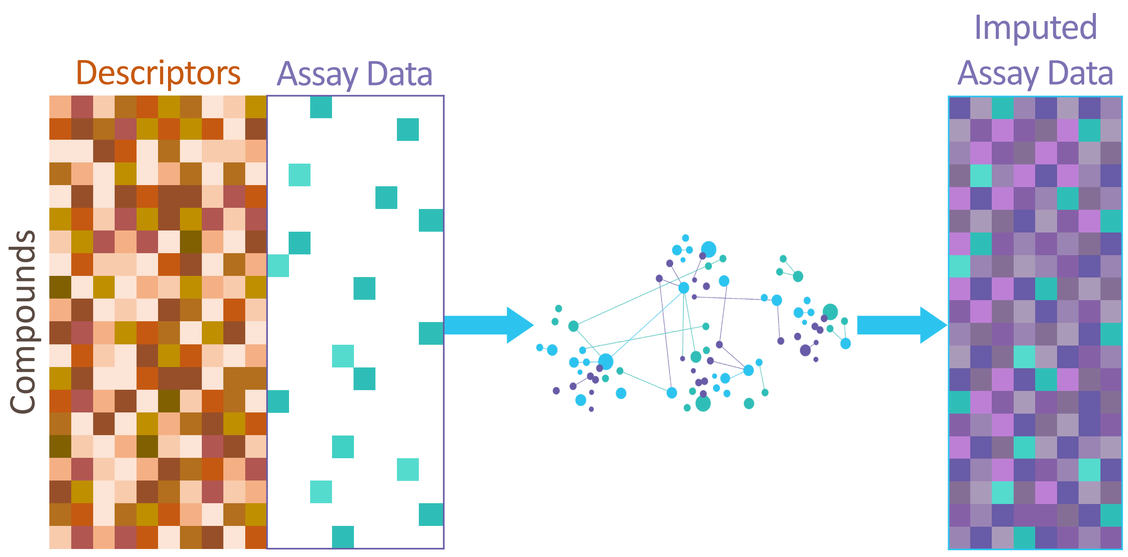
*Optibrium Limited, Cambridge UK. †Constellation Pharmaceuticals, Cambridge MA. ‡Intellegens Limited, Cambridge UK.

Introduction

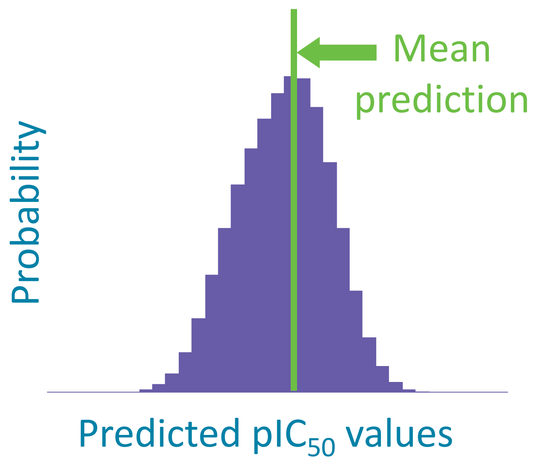
We have previously described a novel deep learning method for data imputation, Alchemite™ (Whitehead *et al.* J. Chem. Inf. Model. (2019) **59** pp. 1197-1204). This accepts both molecular descriptors and sparse experimental data as inputs, to exploit the correlations between experimentally measured endpoints, as well as structure-activity relationships (SAR). It has been demonstrated to outperform quantitative SAR (QSAR) models, including multi-target deep learning methods, on a challenging benchmark data set of compound bioactivities. Here we will describe the application and validation of this method on drug discovery data covering two projects and **diverse endpoints**, including activities in both biochemical and cellular assays and absorption, distribution, metabolism and elimination (ADME) endpoints.

Methods

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



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.



Objectives

- Compare Alchemite to conventional QSAR models on practical, project data sets
- Evaluate the ability of Alchemite to identify the most accurate predictions
- Investigate the potential to apply Alchemite to heterogenous data across multiple projects

Data Sets

Data from two projects (A and B) were used to build and validate models. Project A was a completed project while Project B had recently commenced. The data for each project are summarised below.

Project	No. of Cmpds.	Biochemical Activity Endpoints		Cell-based Activity Endpoints		ADME Endpoints	
		Number	Sparsity (% Filled)	Number	Sparsity (% Filled)	Number	Sparsity (% Filled)
A	1241	3	45	2	15	8	16
B	338	5	55	0	N/A	8	3

The data sets were split into independent training and test sets (80:20) using a stratified selection method that ensures the average sparsity is the same in the training and test sets.

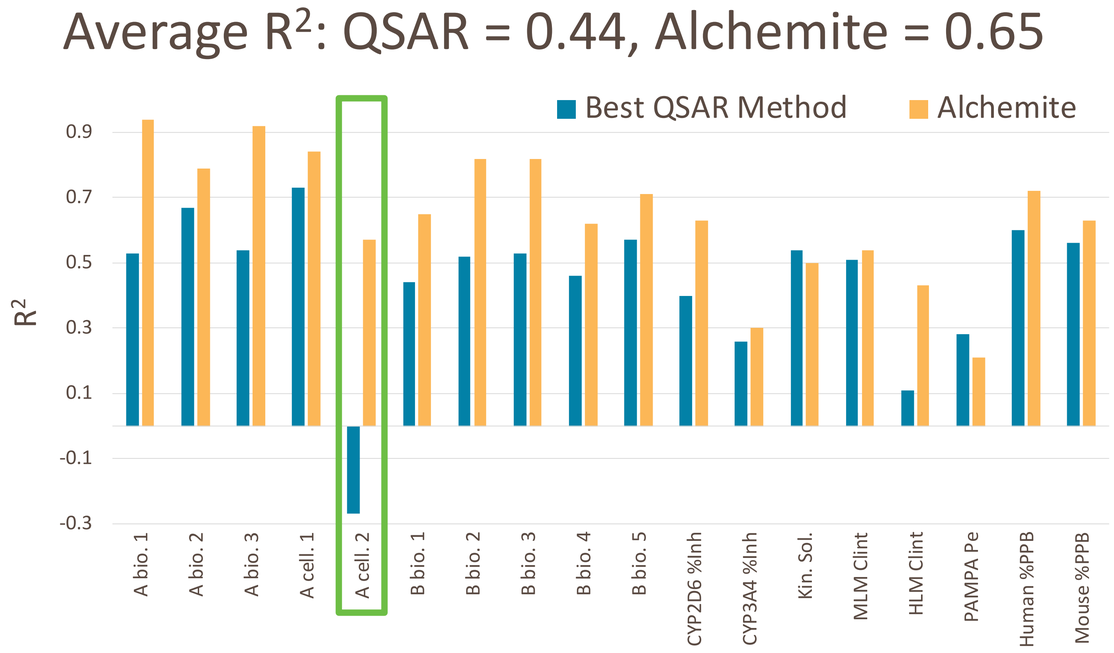
These data were used to build and test the following models:

- Two Alchemite models of the individual project data sets
- A single Alchemite model covering the combined activity and ADME data from both projects
- QSAR models of the individual endpoints.

After completion of the modelling, a small number of new data points were obtained for the Project B compounds included in the model and used as a prospective test of the imputed values.

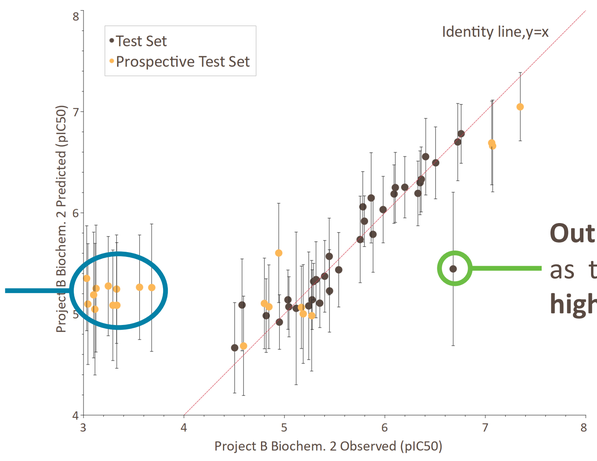
Results

An Alchemite model of the full data set, combining **compound activities** and **ADME properties** in a **single model**, was compared with four QSAR modelling methods: partial least squares, random forests, Gaussian processes and radial basis functions. The improvement in prediction of cellular activity (green box), illustrates the impact of learning directly from correlations between experimental endpoints, even based on sparse data.



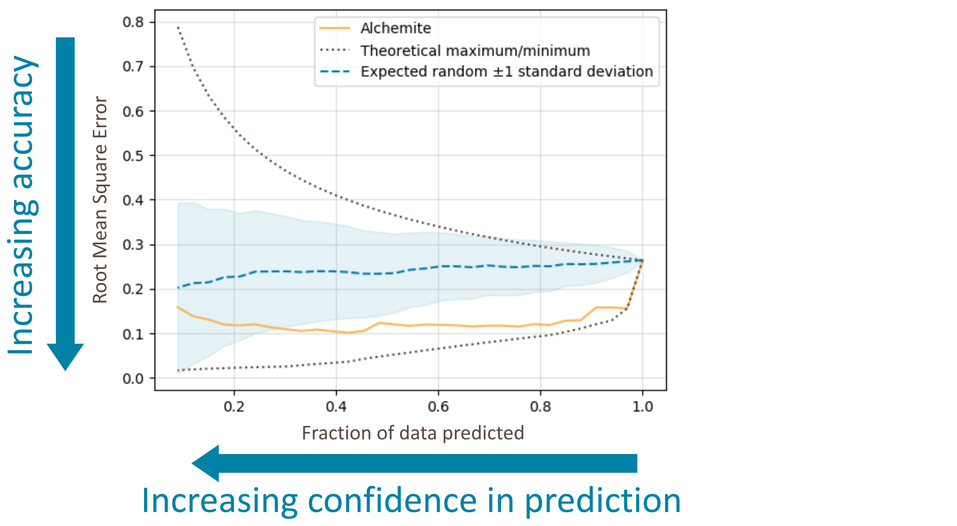
Example Correlation for Project B Bioactivity 2

Observed values are outside the range of the initial training set; yet, they are correctly predicted to be inactive.

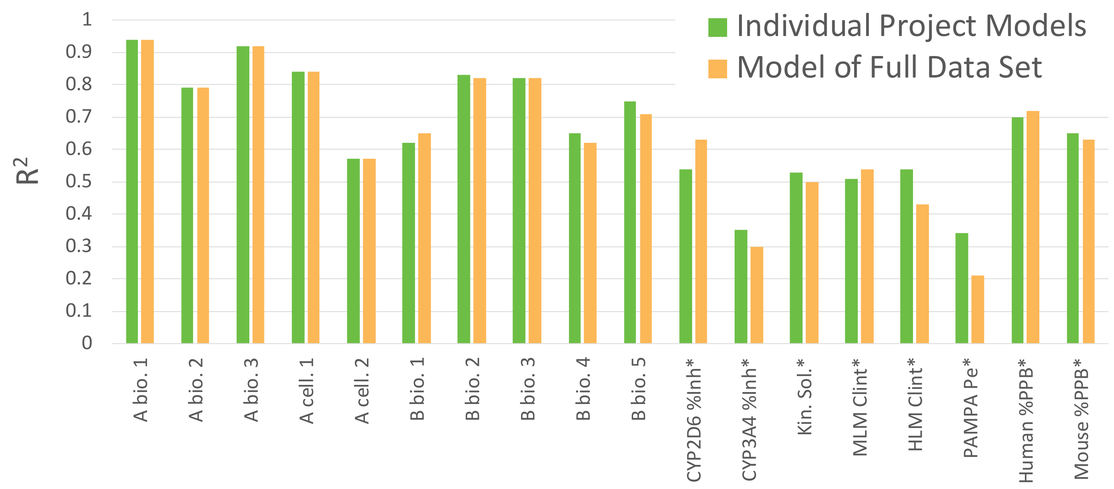


Outlier correctly identified as the prediction with the **highest uncertainty**.

Alchemite can **identify and discard the least-confident predictions**, resulting in an increased accuracy of the remaining predictions, as shown below for biochemical activity 2 for Project B.



The **Alchemite model of the combined data sets performs equivalently** to those built on individual project data sets.



* Individual project model for ADME properties built and tested on Project A only. Full data set model tested against both projects.