Application of the Alchemite deep-learning methodology to categorical modelling of PK endpoints

26th March 2023

Charlotte Wharrick

Overview

- Introduction
- Alchemite™ - the unique deep learning method
- Alchemite™ Proven Success
    - Regression model applications and case studies
- Categorically modelling using Alchemite™
- Conclusions

Challenges of Using Data in Drug Discovery

- It is impossible to measure all of the compounds in all assays - how to make the most of the data available?

- The sparse and noisy nature of the data causes common methods for predictions to struggle

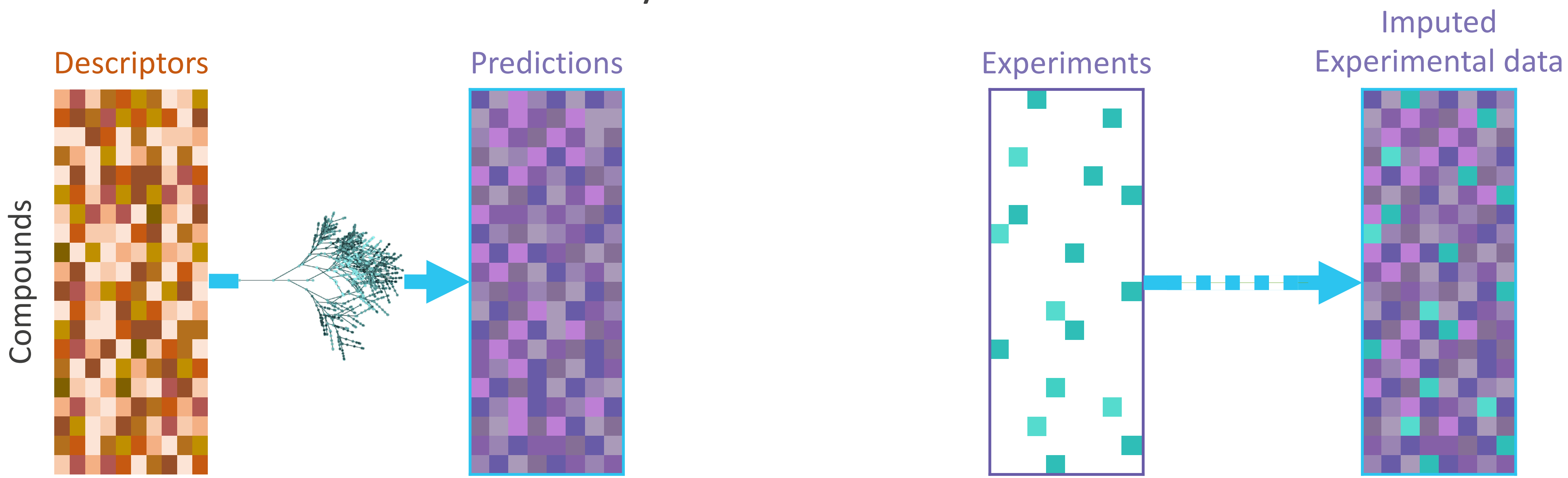
- How can the limited data be used to make better predictions for new compound designs?

Augmented Chemistry

Prediction vs. Imputation

- Prediction uses input ‘features’ to predict one or more property values for a compound, e.g. QSAR models

- Imputation is the process of filling in the gaps in sparse experimental data using the limited results that are already available



Descriptors

Compounds

Predictions

Experiments

Imputed Experimental data

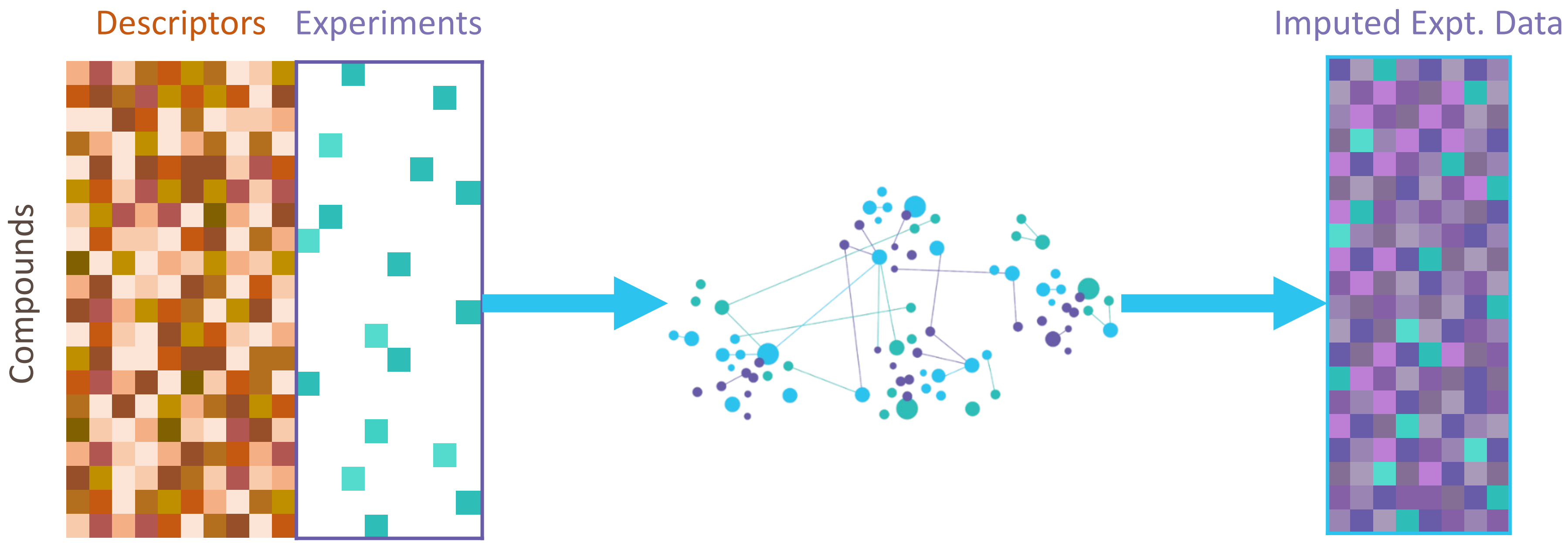
Prediction

Imputation

Alchemite™ Deep Learning Imputation

Optibrium's exclusive partnership with Intellegens

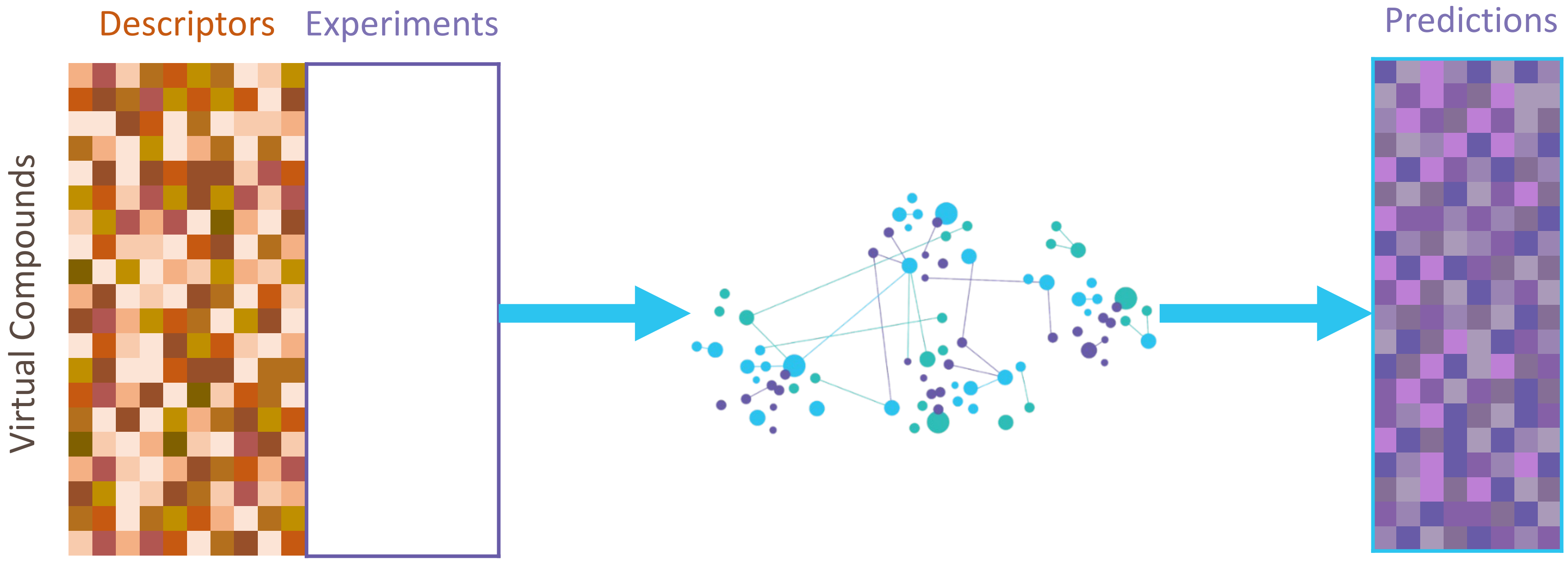
- Learns directly from relationships between experimental endpoints as well as SAR
  - Makes better use of sparse and noisy experimental data than conventional QSAR models
- 'Fills in' the gaps in your data and makes predictions for 'virtual' compounds
  - Generates more accurate predictions to target high-quality compounds



Whitehead *et al.* J. Chem Inf. Model. (2019) **59**(3) pp. 1197-1204, Irwin *et al.* J. Chem. Inf Model. (2020) **60**(6), pp. 2848–2857

Alchemite™ Deep Learning Imputation
Optibrium's exclusive partnership with Intellegens

- Learns directly from relationships between experimental endpoints as well as SAR
  - Makes better use of sparse and noisy experimental data than conventional QSAR models
- 'Fills in' the gaps in your data and makes predictions for 'virtual' compounds
  - Generates more accurate predictions to target high-quality compounds

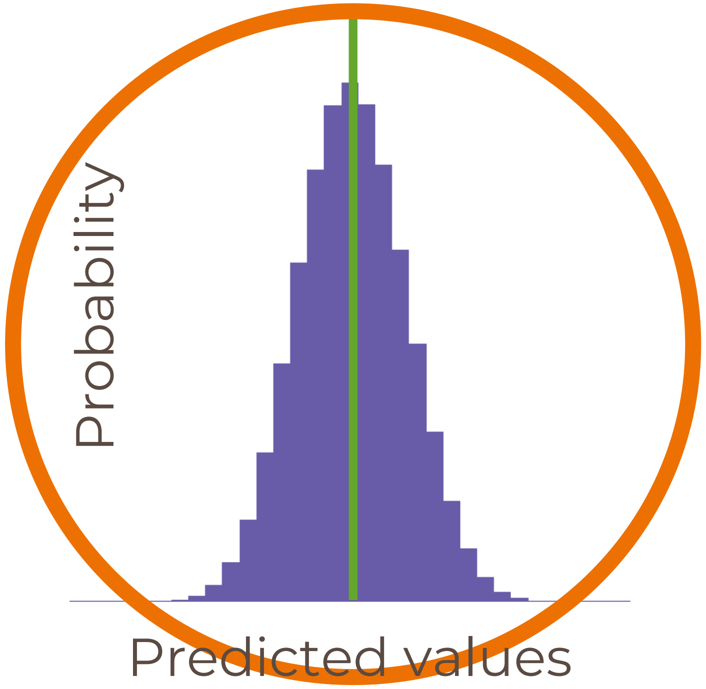


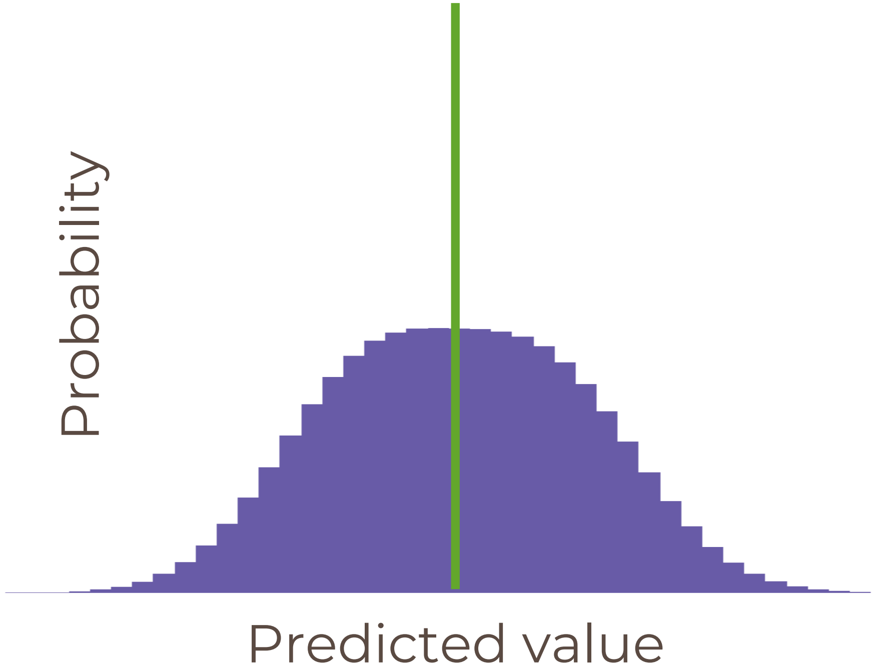
Whitehead *et al.* J. Chem Inf. Model. (2019) **59**(3) pp. 1197-1204, Irwin *et al.* J. Chem. Inf Model. (2020) **60**(6), pp. 2848–2857

Alchemite™ Deep Learning Imputation

Optibrium's exclusive partnership with Intellegens

- Estimates uncertainty in each individual prediction
  - Strong correlation between uncertainty estimates and observed accuracy on independent test sets
  - Highlights the most accurate predictions on which to base decisions
- Confidently targets high-quality compounds and prioritise experimental resources



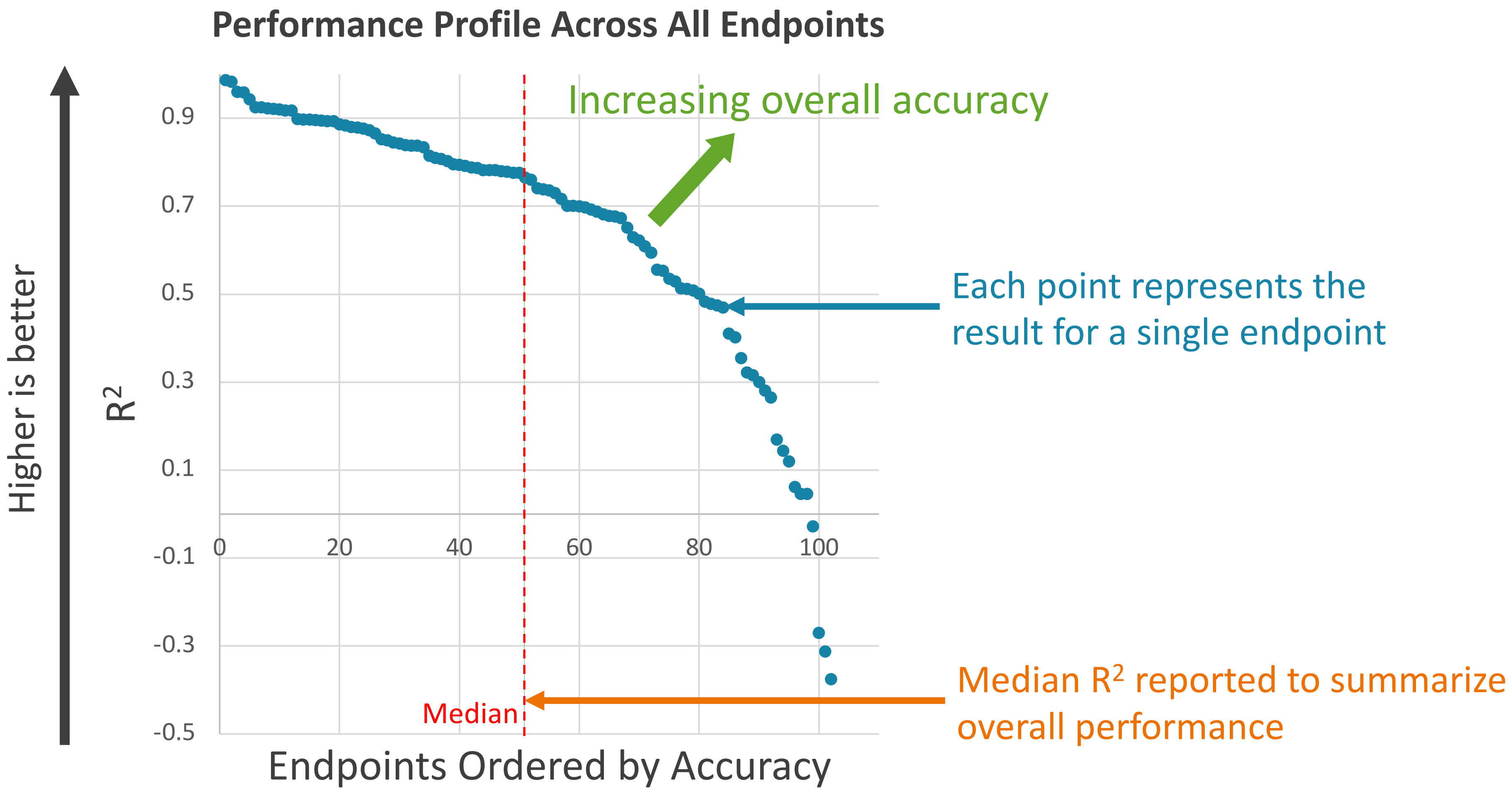


Whitehead *et al.* J. Chem Inf. Model. (2019) **59**(3) pp. 1197-1204, Irwin *et al.* J. Chem. Inf Model. (2020) **60**(6), pp. 2848–2857

Definitions

- Endpoint: An experimental measurement that may be made on a compound
  - E.g. IC_{50} against a target, solubility, Cl_{int} in human liver microsomes, C_{max} in rat PK
- Imputation Model: These models generate predictions for compounds using sparse assay data as input, in addition to molecular descriptors
  - These models 'fill in the gaps' in the experimental data for compounds that have been synthesised and tested in some assays
- Virtual Model: These models generate predictions for compounds using only molecular descriptors as input
  - These models make predictions based only on compound structure, i.e., for a compound that has not yet been synthesised or tested

Assessment of Results
Performance Profile



Increasing overall accuracy

Each point represents the result for a single endpoint

Median R^2 reported to summarize overall performance

R^2 – Coefficient of Determination (1 = perfect prediction, 0 = random, <0 = worse than random)

Regression Models

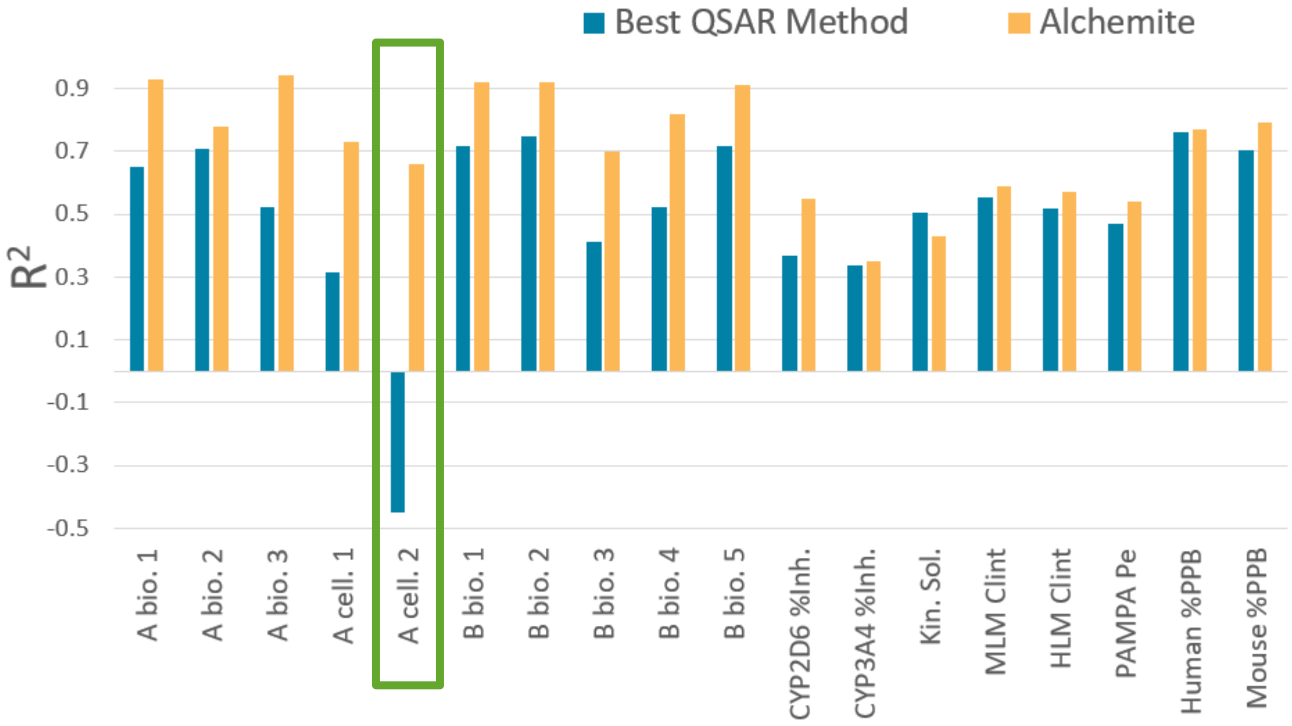
Alchemite Application to Project Data

- Application to **heterogeneous** data across two projects
  - Target and phenotypic activities and ADME endpoints
  - 2453 compounds across 18 endpoints
- Significant improvement in accuracy

	Average R^2
Best QSAR	0.50
Alchemite™	**0.72**

- Example of value delivered:
  - Few false negatives among confidently-predicted inactives – could have saved >$600,000 in unnecessary synthesis

Constellation PHARMACEUTICALS



Irwin *et al.* J. Chem. Inf Model. (2020) **60**(6), pp. 2848–2857
Watch our webinar: http://bit.ly/practical_deeplearning

Alchemite Application to Global Pharma Data

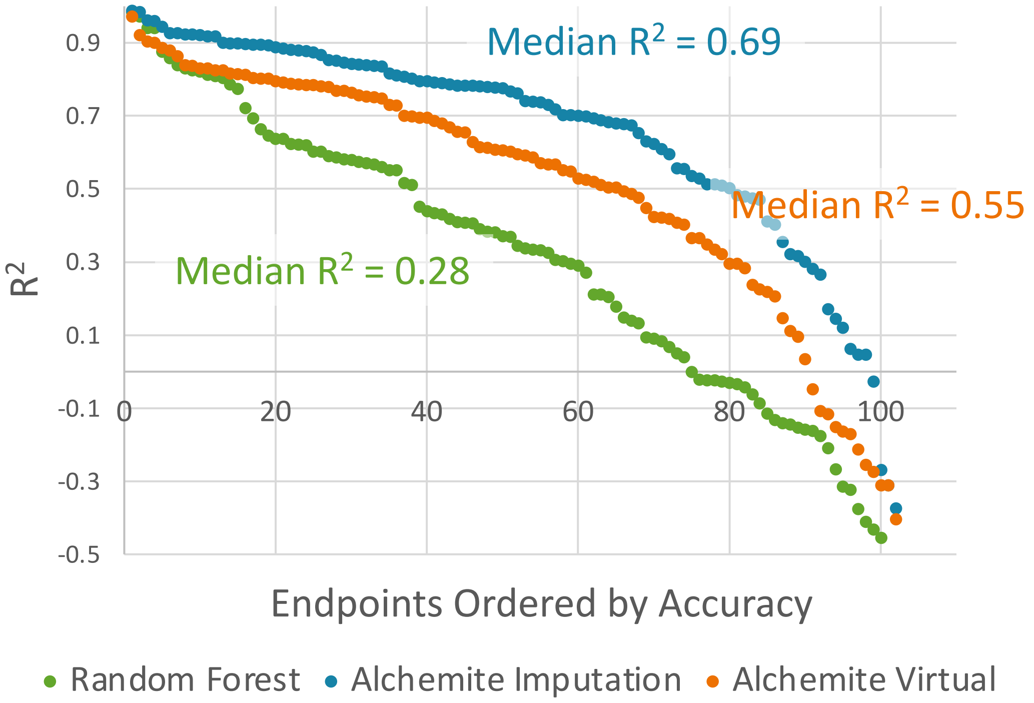
- Application to large data set
  - **678,994** compounds
  - 1,116 experimental endpoints
  - 2% complete

- Covering a **full range** of drug discovery assays, including compound activities and ADME properties

- Example of value delivered:
  - "…an extension of what medicinal chemists… do in a discovery project, but at much larger scale than would be possible for a person."

Takeda | ONCOLOGY

Prospective Prediction of Project Target Activities



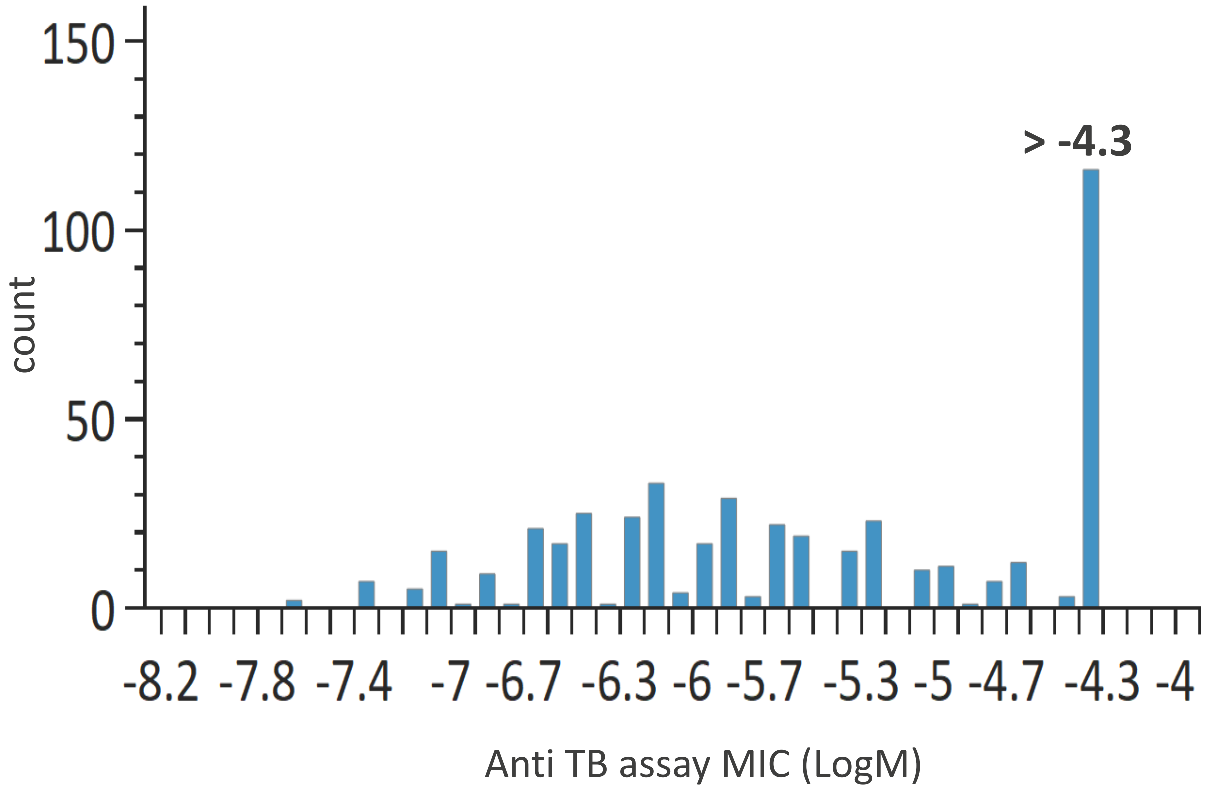
Irwin *et al.* App. AI Lett. (2021) DOI: 10.1002/ail2.31
Watch our webinar: http://bit.ly/largescale_imputation

Limitations of Regression Models

- Qualified values (continuous values with <, > signs) are removed prior to building the model to prevent a skewed distribution

- Noisy data as input can lead to low-quality predictions

- Labelled data or inherently categorical endpoints cannot be modelled



Application of Categorical Modelling

Categorical Modelling Methods

- Handling qualified data
  - Continuous data may contain qualified data, e.g. <, >
  - Define cut-offs to “bin” the data into classes and include these values in the model

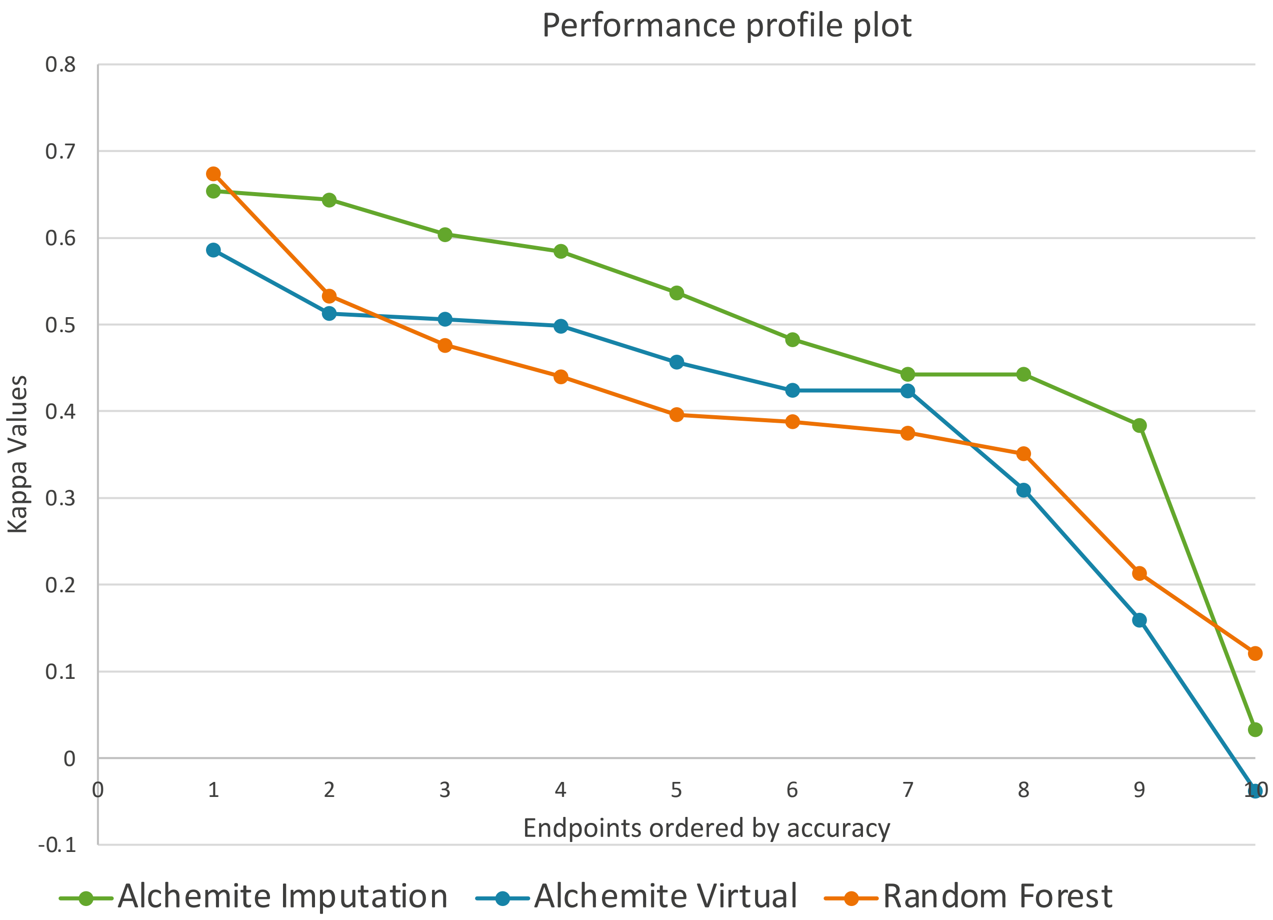
- Model building
  - The library of descriptors were provided by StarDrop and consists of 10 whole molecule Descriptors and 320 Auto-Modeller descriptors based on 2D SMARTS, logP, TPSA, MW, charge etc
  - Training and test sets consist of discrete values (0s and 1s) for binary categorical models
  - The predictions are discrete values
  - Cohen’s Kappa values are used to indicate performance

- Alchemite (imputation and virtual) categorical models were built and compared with the categorical QSAR model
  - Consistent training and test sets
  - Consistent cut-offs for the same assay in the different model

Deep learning methods *Vs* QSAR

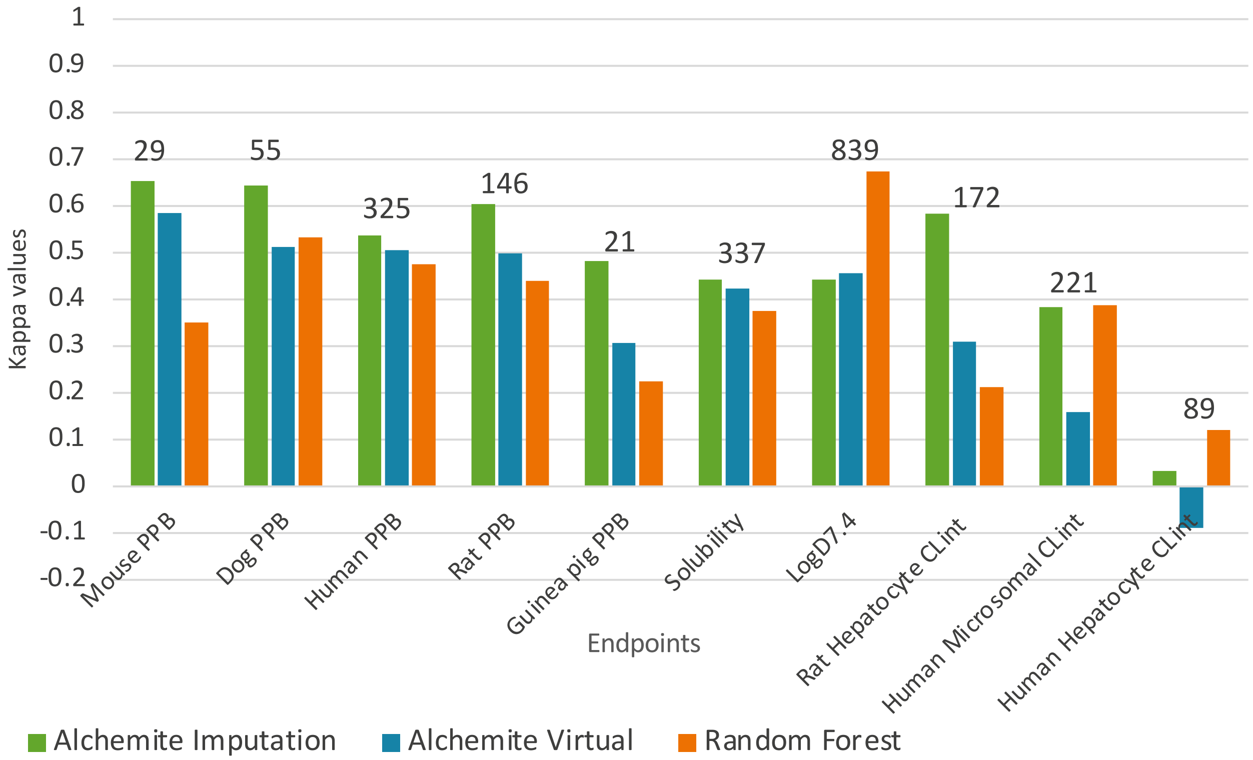
- Application to the publicly available **AZ data** set: *document_chembl_id:CHEMBL3301361*
  - 5788 compounds
  - 10 PK assays from different species
  - 13% complete
- Model building
  - The continuous data were "binned"
  - Alchemite imputation and virtual categorical models *Vs* Random Forest categorical model
- Improvements in accuracy

	Median Kappa Value
Random Forest	0.39
Alchemite Virtual	0.44
Alchemite Imputation	**0.51**



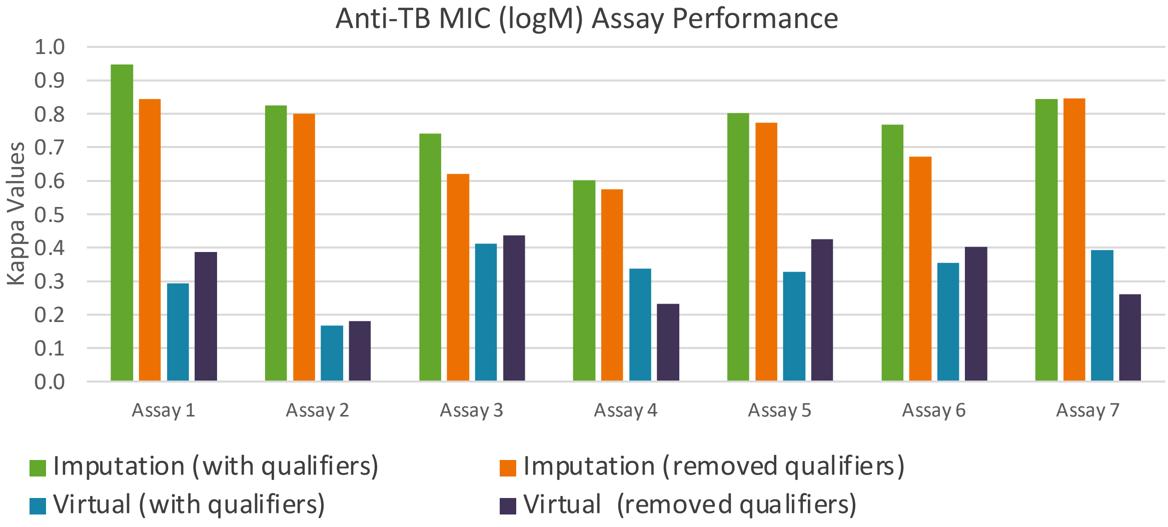
PK Endpoint Performance - Deep learning methods *vs* QSAR

- Analysing the performance for each ADME/PK endpoint
- PPB: Plasma protein binding
  - 5 species
  - Alchemite Imputation model is consistently outperforming the virtual and RF models
- CL int: Intrinsic clearance
  - 2 different species
  - Hepatocyte and Microsomal
- The number of data points in the test set are included



Application of Categorical Modelling to Qualified Data

- Categorical modelling on **a global health data set** with qualified data
- The data set
  - **495** compounds
  - **34** endpoints (in-vitro and in-vivo activity, PK and ADME data)
- Including qualified data changes the sparsity of the overall data set from **20%** to **30%** data points present
- More data leads to a wider chemical spaces and a more accurate model
  - Alchemite imputation and virtual categorical models were built on the datasets with and without the qualified data included
  - Anti-TB MIC (LogM) assays showed the greatest improvements for the imputation methods with the additional qualified data included



Conclusions

- Advantages of Alchemite deep learning imputation
  - Gains more value than prediction from experimental data
  - Outperforms traditional QSAR methods

- We have demonstrated the successful application of Alchemite in a range of categorical modelling scenarios
  - Heterogenous data across multiple drug discovery endpoints
  - Sparse data sets
  - Large data sets with qualified data

- The categorical feature of Alchemite has shown success where regression models struggle
  - Qualified data
  - Labelled or classified data

Acknowledgements

optibrium

Matt Segall

Samar Mahmoud

Bailey Montefiore

Peter Hunt

Ben Irwin

intellegens

Gareth Conduit

Tom Whitehead

Thomas Bridge

For more information visit www.optibrium.com