Practical Applications of Deep Learning to Imputation of Drug Discovery Data

Benedict Irwin* ben@optibrium.com

Julian Levell [†], Thomas Whitehead [‡], Matthew Segall*, Gareth Conduit [‡]

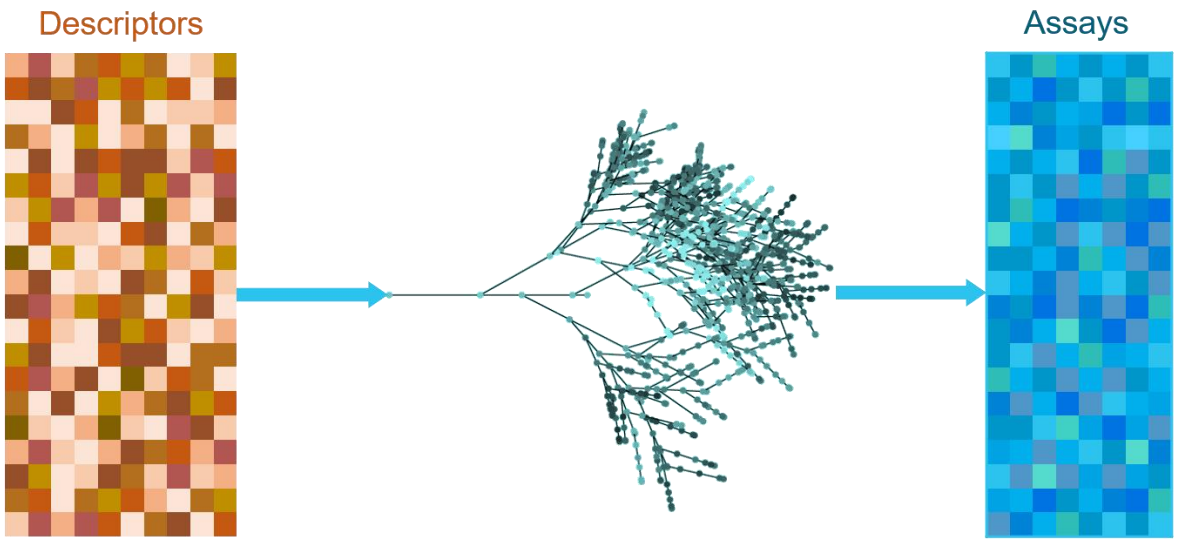
*Optibrium Limited, Cambridge UK. [†]Constellation Pharmaceuticals, Cambridge MA. [‡]Intellegens Limited, Cambridge UK.

Overview

- **Problems** with pharma data:
  - Define solutions to these problems
- **Alchemite**: A novel deep learning algorithm for *imputation*
  - *Imputation = Filling in the blanks*
- **Walkthrough** deep learning imputation on a **real project**:
  - Early screen data
  - Validation
  - Late stage models
  - Comparison with standard QSAR methods
- Larger applications and **future prospects**

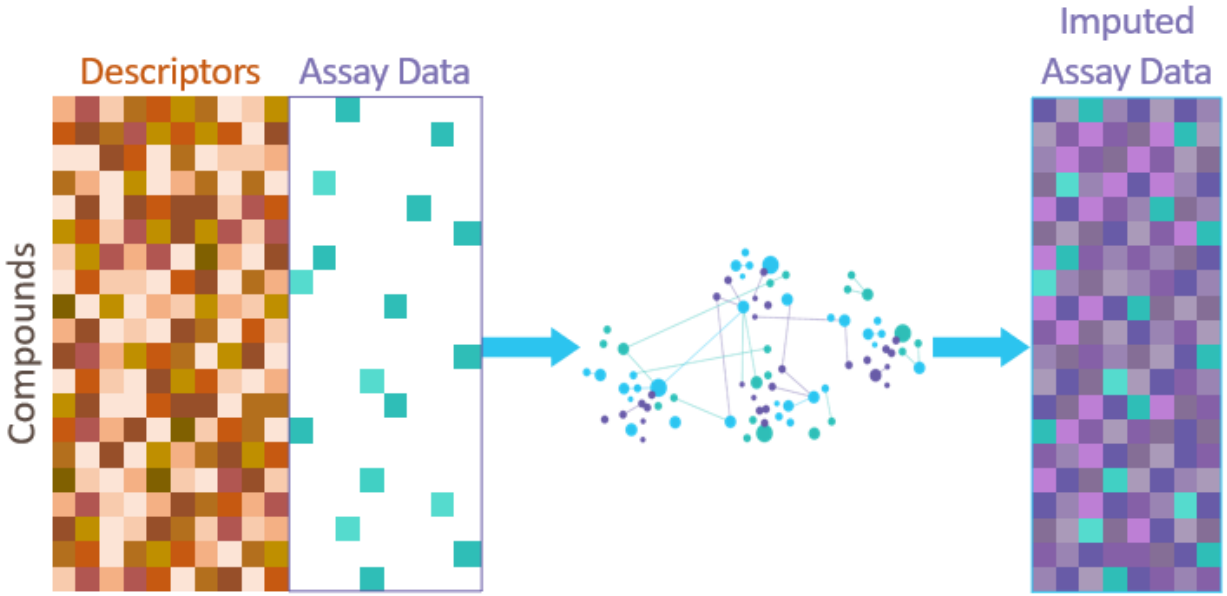
Imputation goes beyond QSAR!

QSAR



e.g.
Random forest

Imputation



Alchemite

Problems with Pharma Data

optibrium

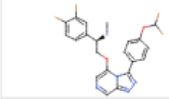
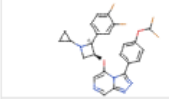
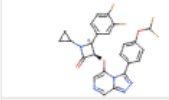
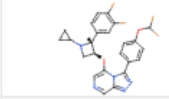
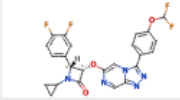
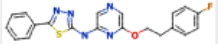
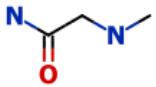
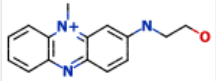
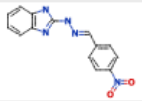
Problems with Pharma Data

For a machine learning method to be **practically** useful in QSAR it should handle:

Missing Values	Noisy Data
Multiple Endpoints	Data Changing with Time

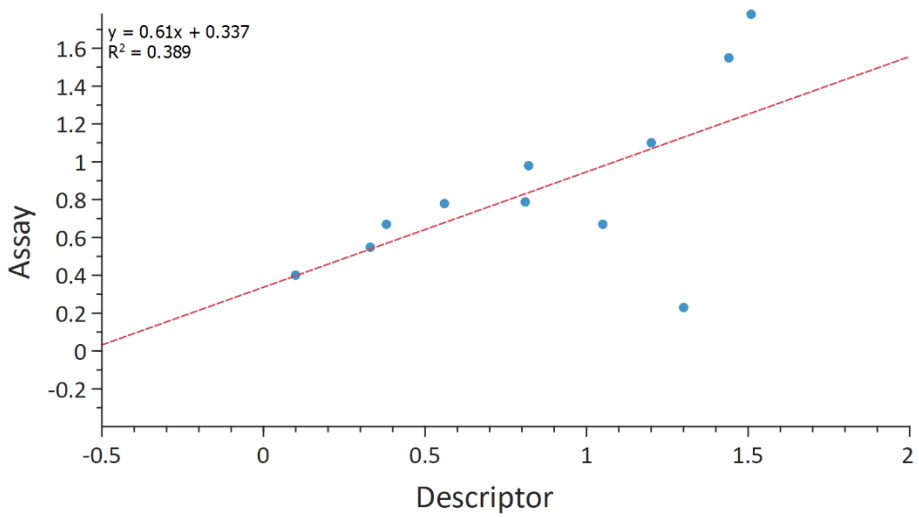
Missing Values

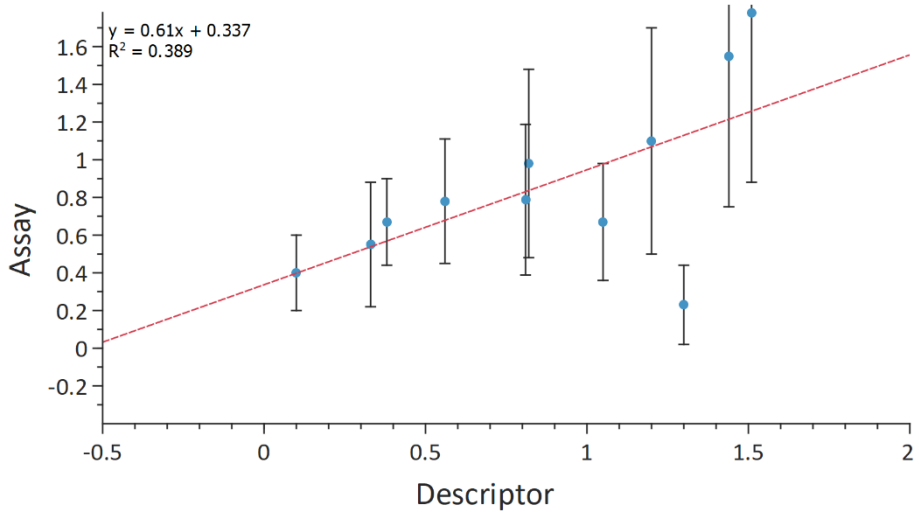
- Problem:
  - Most algorithms cannot handle missing inputs
  - $y = f(x_1, ?, x_3, x_4, ?)$
  - Simple methods to impute give poor quality results e.g. imputation via mean
  - $y \neq f(x_1, \overline{x_2}, x_3, x_4, \overline{x_5})$
- Solution:
  - Algorithm should make the most of data present
  - “Fill in” the missing values with sensible predictions

	SMILES	Potency vs Parasite (uMc	Ion Regulation Activity	SSI%	EC50Chembl(uM)	ertl-39	aminoethanol1
1		10	?	?	?	0	1
2		0.6095	?	?	?	0	0
3		1.121	?	?	?	0	0
4		0.7308	?	?	?	0	0
5		10	?	?	?	0	0
6		?	?	?	?	0	0
7		?	?	?	?	0	1
8		0.296	0	?	?	0	1
9		0.142	0	?	0.4809	0	0

Noisy Data and Confidence in Predictions

- Problem:
  - **Pharma data is inherently noisy**
  - Input data may not be “true”
  - Model outputs a number with no context
- Solution:
  - Input noise accounted for
  - Predictions should come with confidence values!
  - Highly confident predictions are more valuable than weak ones
  - Provide a big error bar if model doesn’t know the answer





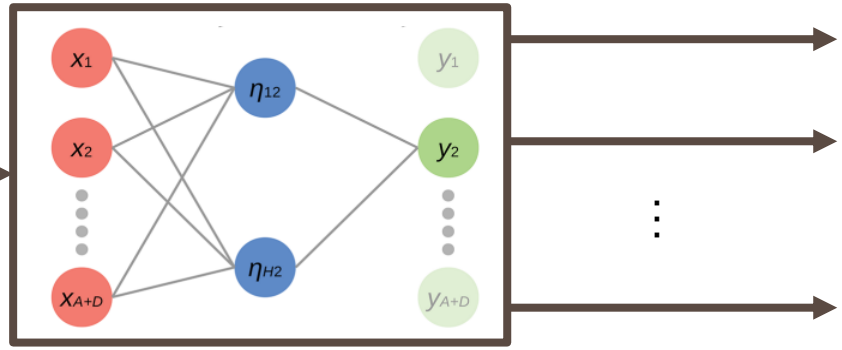
Multiple Endpoints – One Model

- Problem:
  - **Many columns in project data:** can't train a model for each one…
  - Activity IC50, EC50: protein, supersome, cell
  - Multiple targets: related, unrelated
  - (ADME) Absorption, distribution, metabolism, and excretion
  - Plasma protein binding, intrinsic clearance, CYP inhibition, permeability, solubility

- Solution:
  - One model to handle everything

(Noisy Fragmented Data)

X_{ij}



Protein Assay $\pm\ \sigma_{i1}$

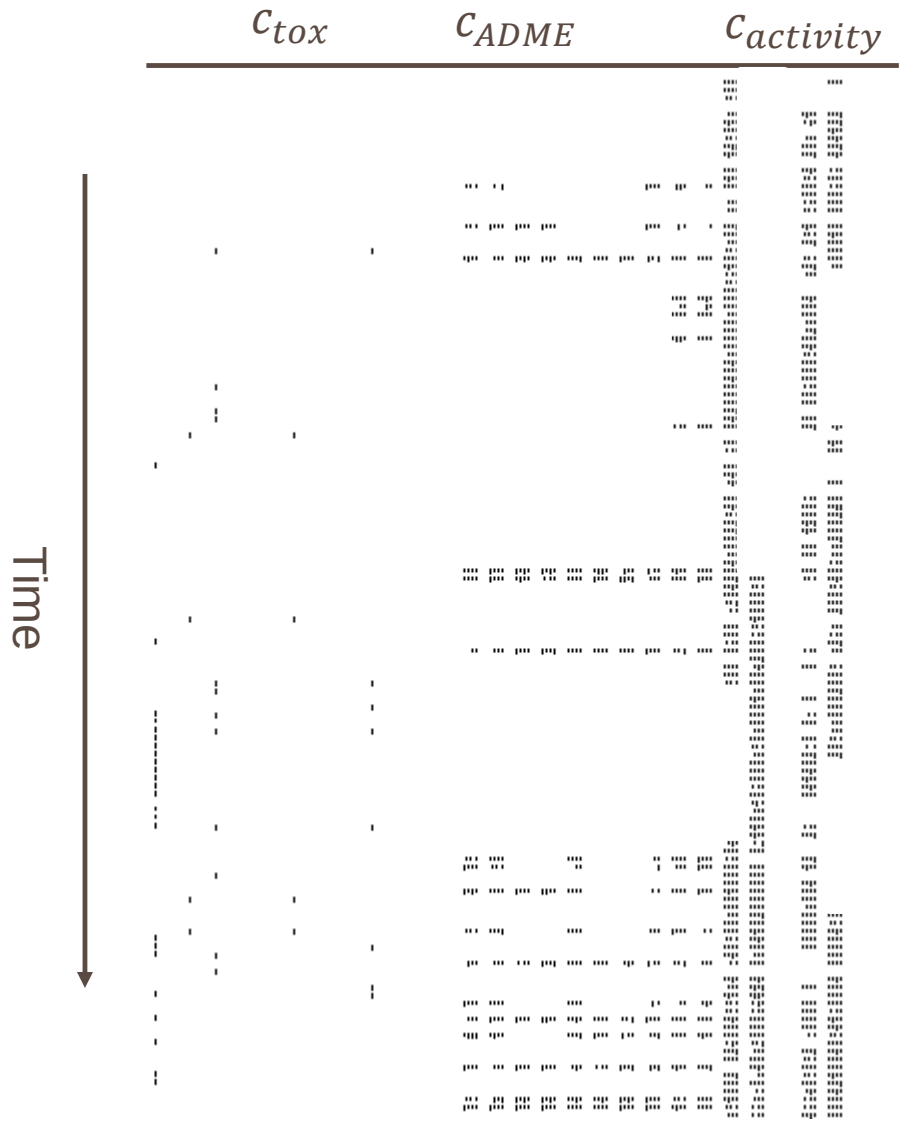
Cell Assay $\pm\ \sigma_{i2}$

⋮

ADME $\pm\ \sigma_{iN}$

Changing with Time

- Problem:
  - Data are evolving as project continues
  - Chemical space changes
  - Activity changes i.e. increasingly active
  - Data sparsity changes
    (more ADME, less HTS)
  - Uncertainty changes
    (new assay concentration, finer resolution)
- Solution:
  - Model which extrapolates well
  - Retraining the model as appropriate



Alchemite – A Method for Deep Multiple Imputation

optibrium

Optibrium Collaboration with Intellegens

optibrium Intellegens

Optibrium and Intellegens Collaborate to Apply Novel Deep Learning Methods to Drug Discovery

Partnership combines Intellegens' proprietary AI technology with Optibrium's expertise in predictive modelling and compound design

optibrium Intellegens CATAPULT Medicines Discovery

Novel deep learning drug discovery platform gets £1 million innovation boost

Optibrium™, Intellegens and Medicines Discovery Catapult awarded funding to apply machine learning in drug discovery

JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

Cite This: J. Chem. Inf. Model. 2019, 59, 1197−1204

pubs.acs.org/jcim

Imputation of Assay Bioactivity Data Using Deep Learning

T. M. Whitehead,*,† B. W. J. Irwin,‡ P. Hunt,‡ M. D. Segall,‡ and G. J. Conduit†,¶

†Intellegens, Eagle Labs, Chesterton Road, Cambridge CB4 3AZ, United Kingdom

‡Optibrium, F5-6 Blenheim House, Cambridge Innovation Park, Denny End Road, Cambridge CB25 9PB, United Kingdom

¶Cavendish Laboratory, University of Cambridge, J.J. Thomson Avenue, Cambridge CB3 0HE, United Kingdom

Supporting Information

ABSTRACT: We describe a novel deep learning neural network method and its application to impute assay pIC_{50} values. Unlike conventional machine learning approaches, this method is trained on sparse bioactivity data as input, typical of that found in public and commercial databases, enabling it to learn directly from correlations between activities measured in different assays. In two case studies on public domain data sets we show that the neural network method outperforms traditional quantitative structure−activity relationship (QSAR) models and other leading approaches. Furthermore, by focusing on only the most confident predictions the accuracy is increased to $R^2 > 0.9$ using our method, as compared to $R^2 = 0.44$ when reporting all predictions.

optibrium Constellation PHARMACEUTICALS Intellegens

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

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

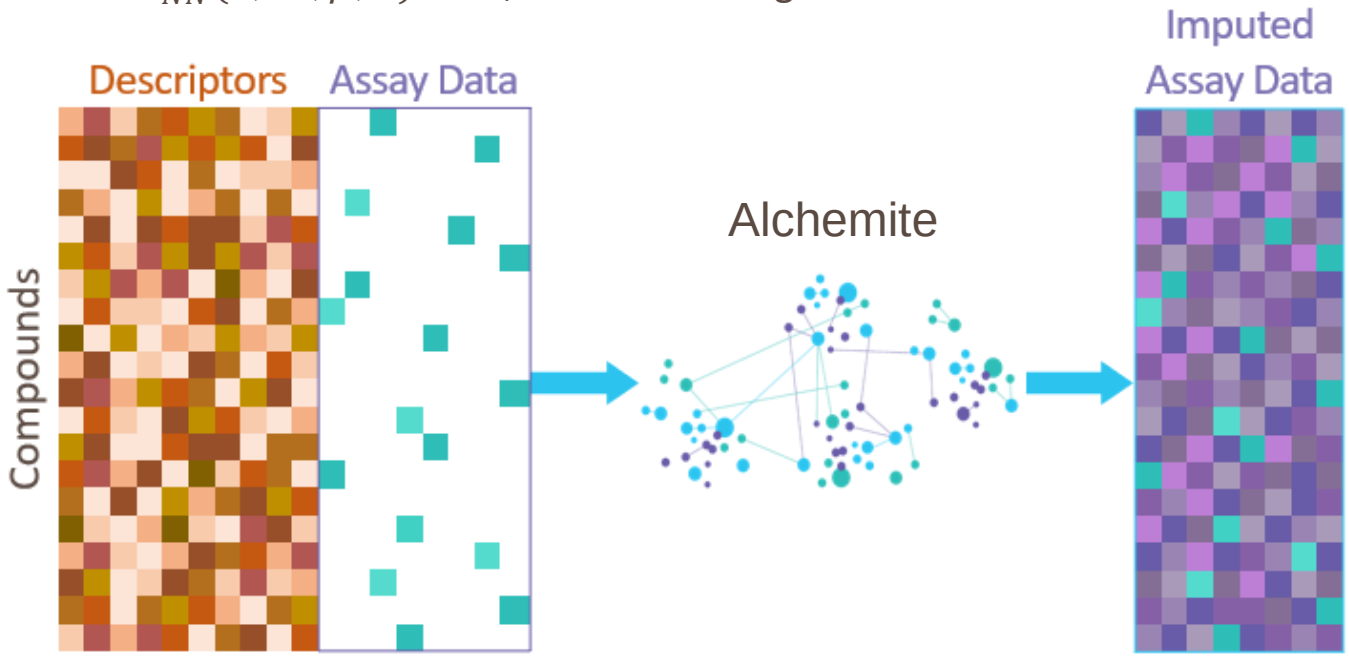
*Optibrium Limited, Cambridge UK. †Constellation Pharmaceuticals, Cambridge MA. ‡Intellegens Limited, Cambridge, UK.

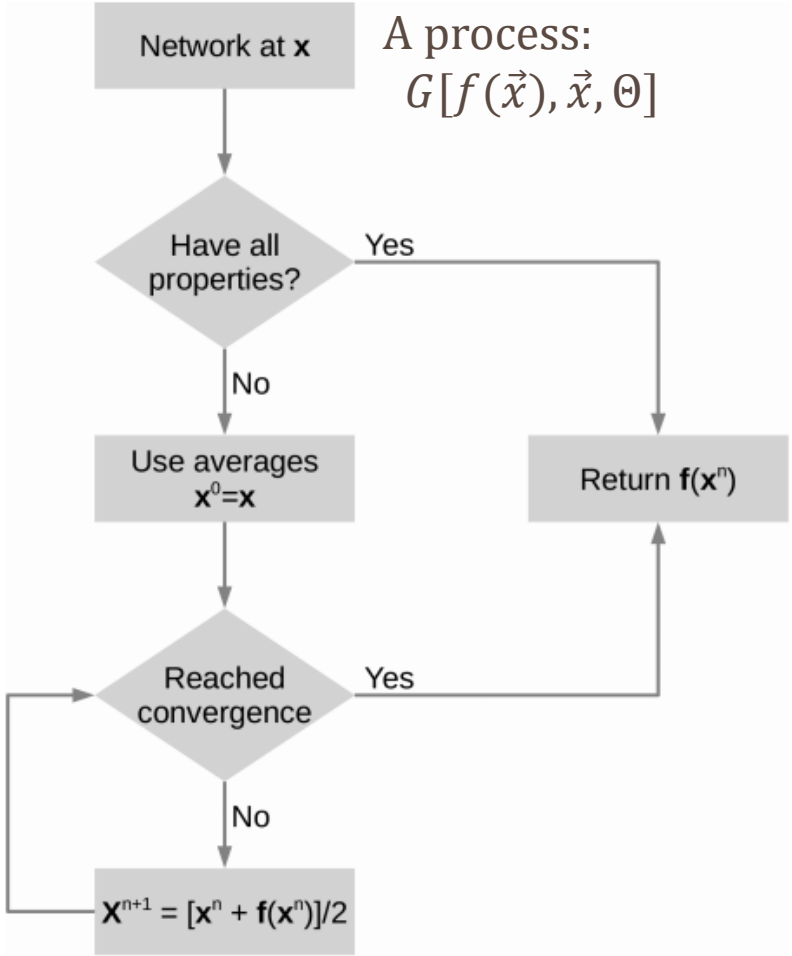
Whitehead et al.
J. Chem. Inf. Model. 2019, 59, 1197-1204

Alchemite – A Method for Deep Multiple Imputation

- Originally used to design new materials at the University of Cambridge, UK
  - Design alloys, identify errors in databases
  - Optimising algorithm and applying to drug discovery data

- Take solution of deep neural network $D_{NN}(\vec{x})$ under fixed point iteration
  - $D_{NN}(\vec{x}; W, \beta, \theta) = \vec{x}$, for $\vec{x}$ in training set.



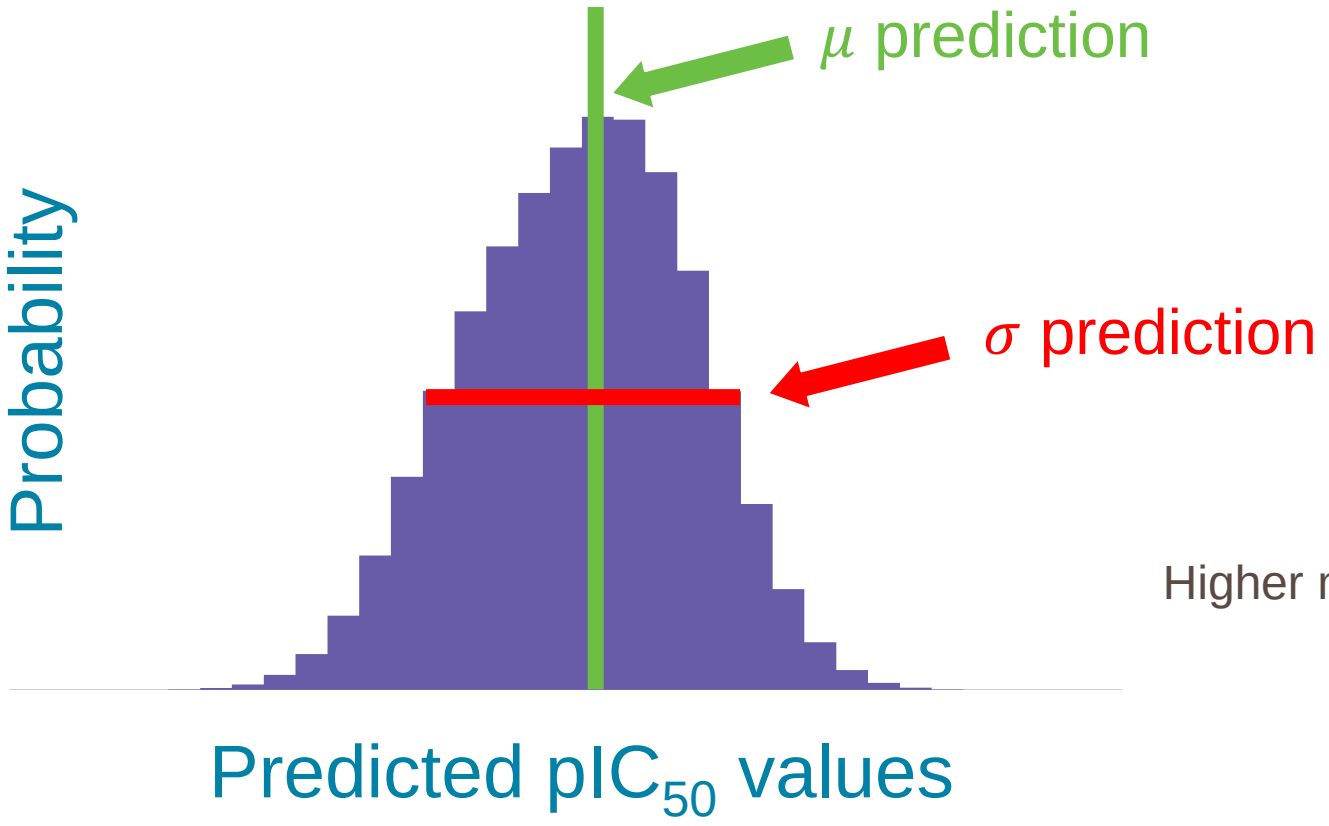


A process:
$G[f(\vec{x}), \vec{x}, \Theta]$

Imputation of Assay Bioactivity Data Using Deep Learning, T. M. Whitehead*, B. W. J. Irwin, P. Hunt, M. D. Segall, G. J. Conduit, JCIM, 2019

Output Predictions and Uncertainty

- Outputs a probability distribution by multiple imputation (1000's of samples).
  - Network is very quick to train/evaluate: train thousands of networks



Higher moments $\gamma_1, \gamma_2, \ldots$

Practical Application of Deep Learning to Project Data

optibrium

Initial Project Data

- Two Projects
  - A: Completed project
  - B: Ongoing project that had recently commenced

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

- Small number of additional data points for Project B compounds were measured for imputed data points after completion of the models

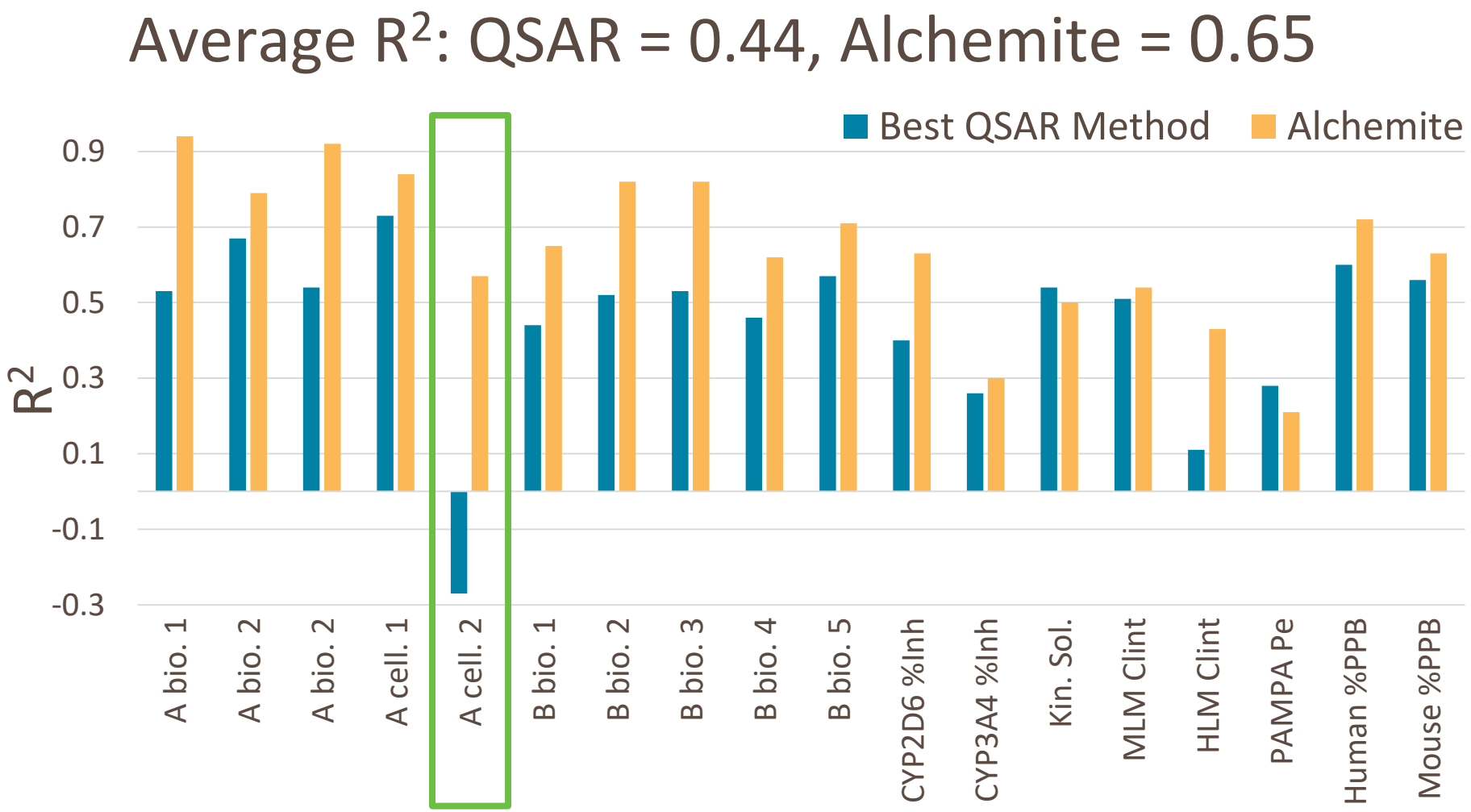
* After removal of qualifiers

Overview

- Objectives
  - Compare accuracy of Alchemite model to conventional QSAR models
  - Compare models built on all data simultaneously with those build on individual projects and subsets of data
  - Evaluate Alchemite's ability to estimate confidence in individual predictions and target the most accurate results
- Three sets of models generated:
  - Two Alchemite models of the individual project data sets
  - A single Alchemite model covering the combined activity and ADME data from both projects
  - Conventional QSAR models of the individual endpoints
    - Random forest, Gaussian processes, radial basis functions and partial least squares

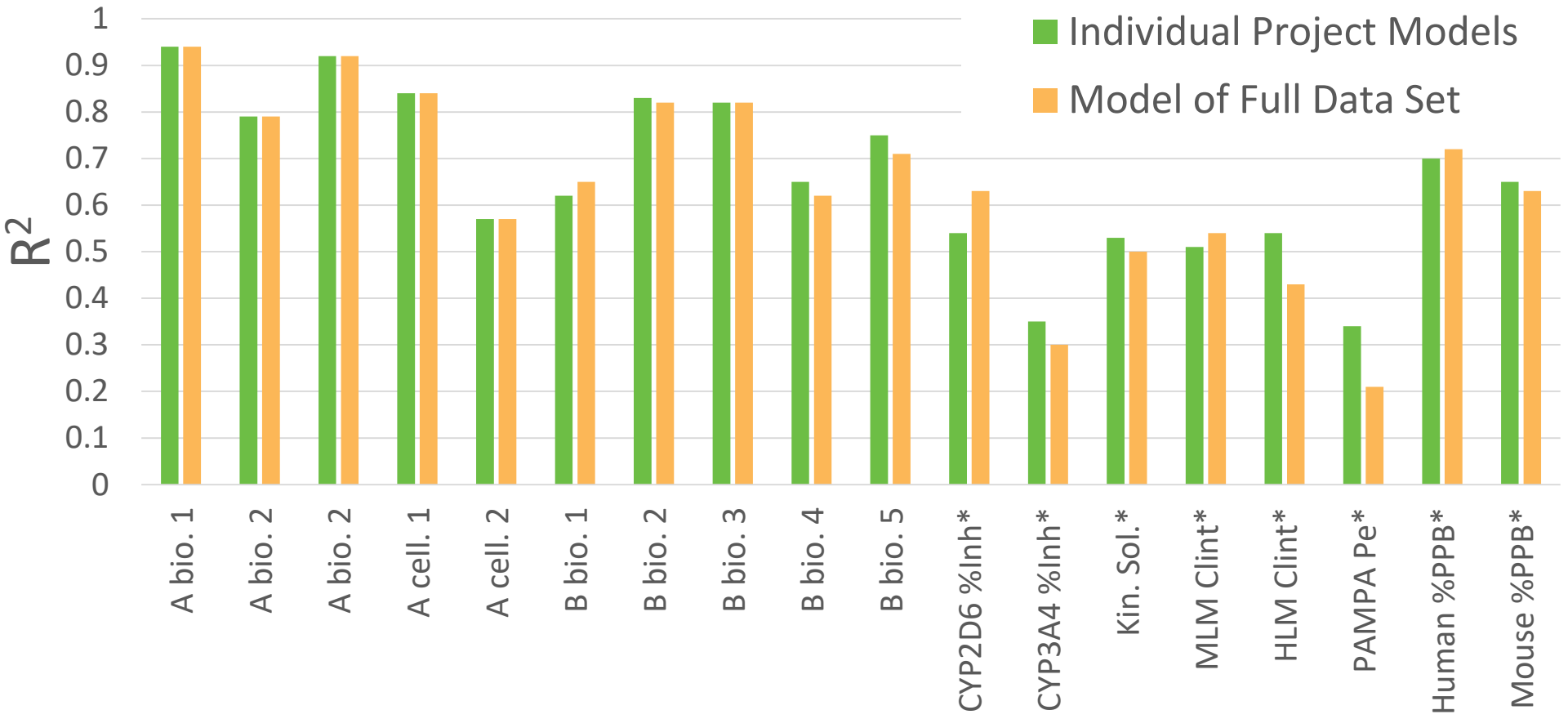
Comparison of Alchemite and QSAR

Single Alchemite model of combined data set



Single Model vs Individual Project Models

Single model performs equivalently to individual project models

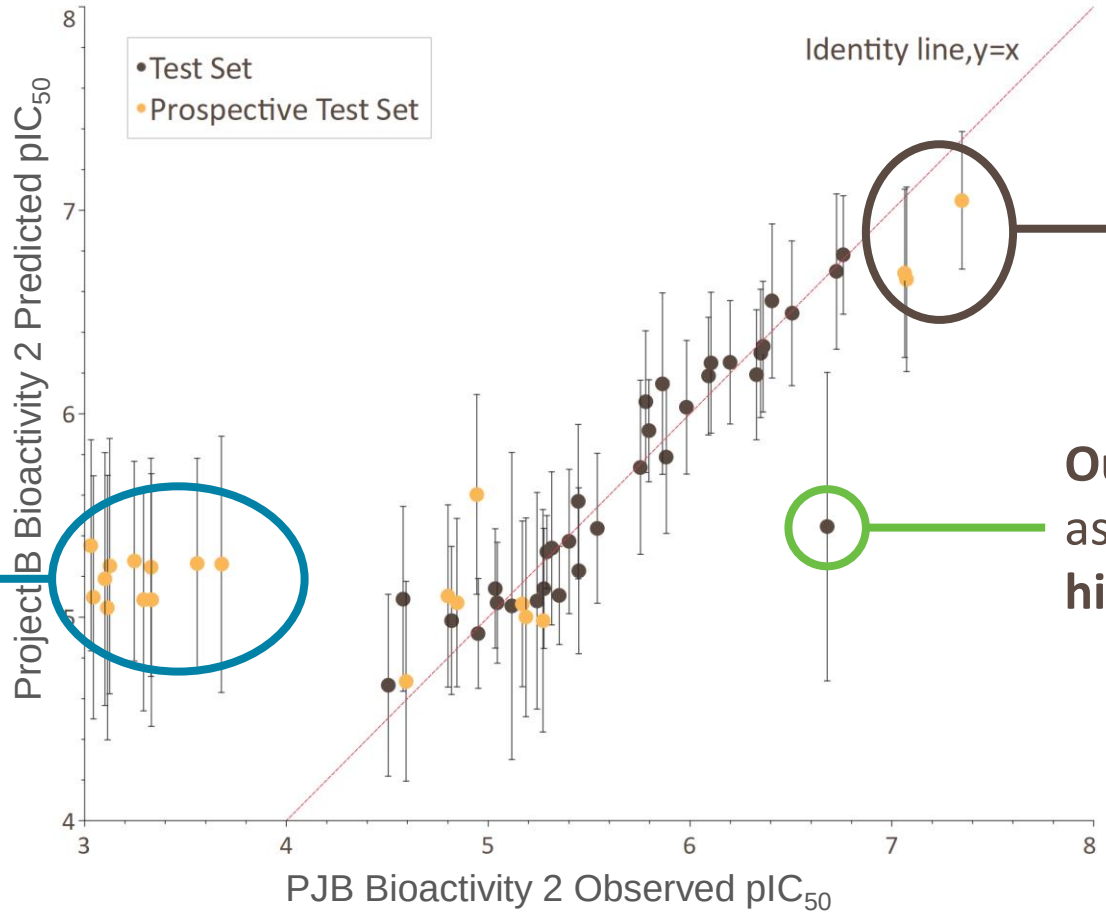


* Individual project model for ADME properties built and tested on Project A only. Full data set model tested against both projects.

Example Validation
Project B - Bioactivity 2

- We then received more data on the Project B compounds

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

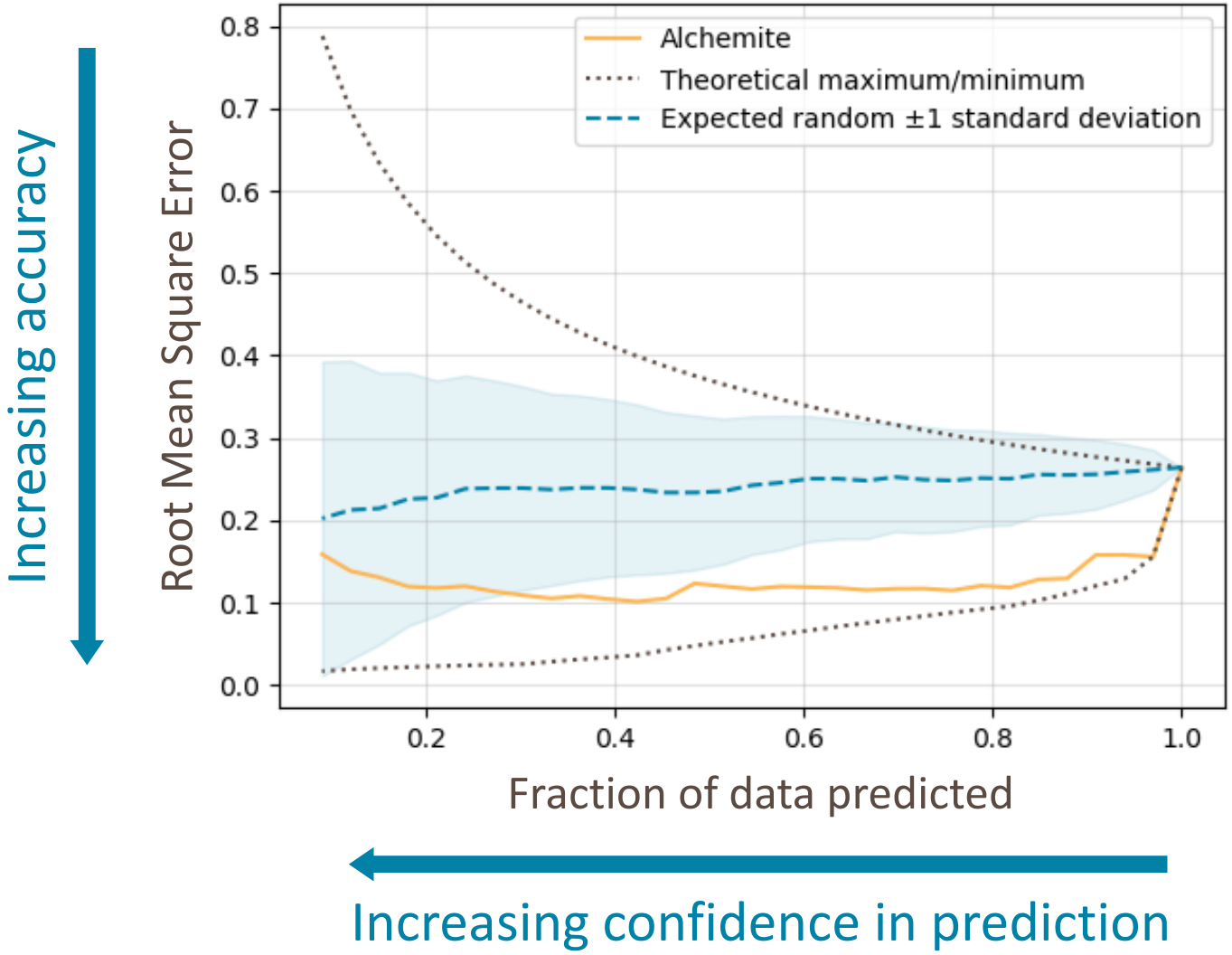


New active compounds correctly identified as active

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

Identify and Discard the Least-Confident Predictions

Project B – Bioactivity 2

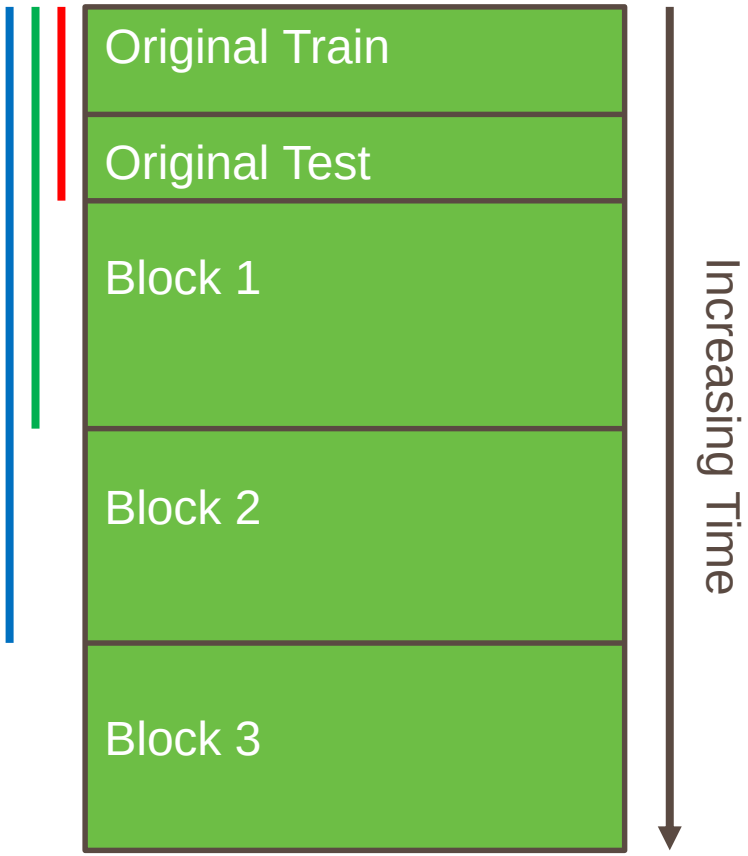


Part 1 - Conclusions

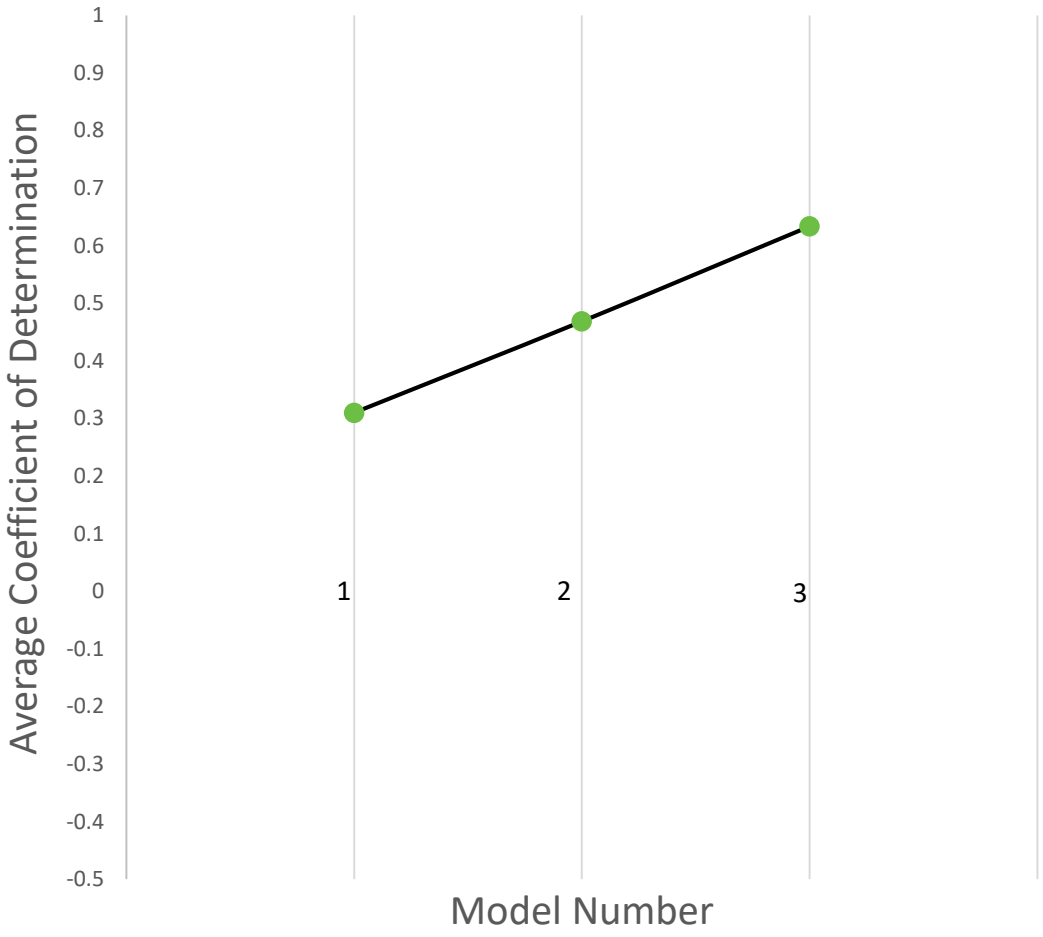
- The single Alchemite model of data for both projects, including biochemical and cell-based activities, and ADME properties significantly outperforms QSAR models
- The performance on independent and prospective test sets is very good and consistent.
- The single Alchemite model performs equivalently to models of individual projects and subsets of the data
  - Can combine data from multiple chemistries and types of endpoints in a single model
- Alchemite can target focus on the most confident and accurate results to use as the basis for decisions
- Next steps… Application to new compounds and data as project progresses

Part 2 - Temporal Prospective Validation

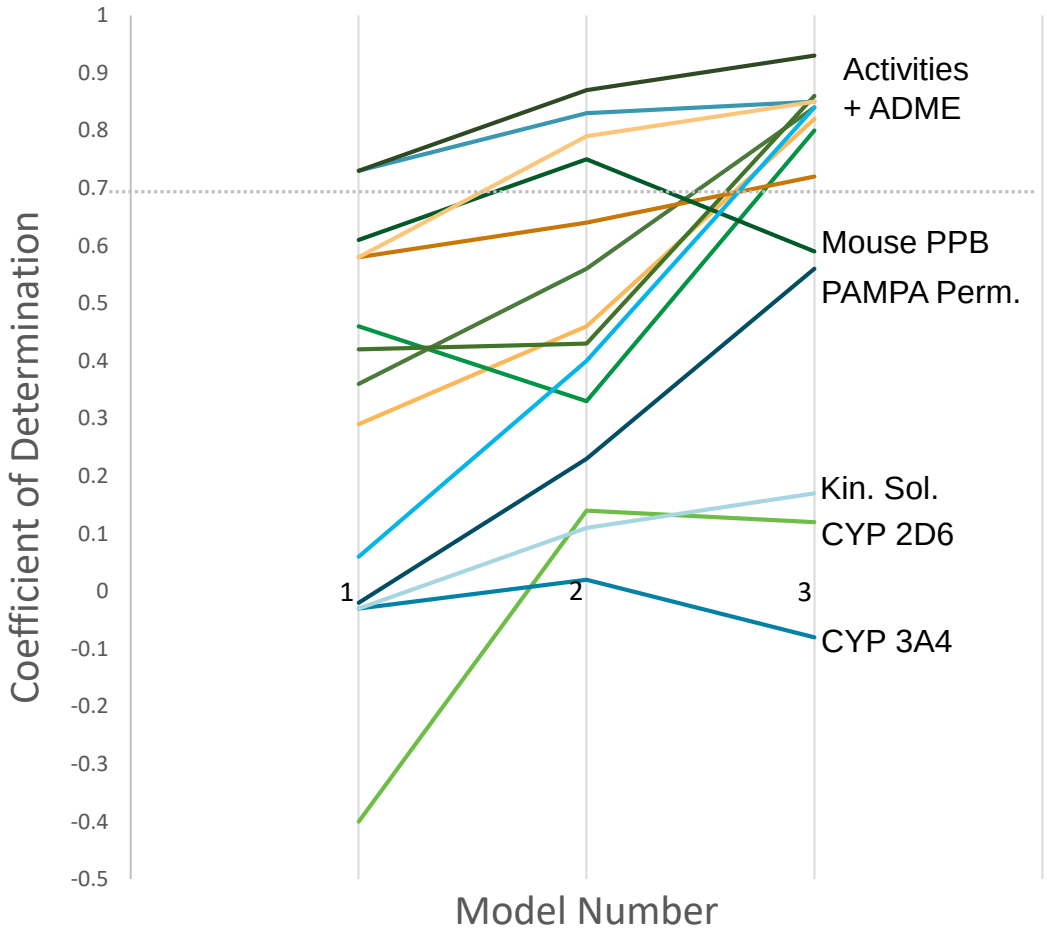
- Received an **additional 874 compounds** for project B
  - Sparse results from real experiments
  - Many additional ADMET datapoints
- Three blocks of temporally coordinated data, B1,2,3:
  - Model 1 : Trained on all of the original data
  - Model 2 : Original + B1
  - Model 3 : Original + B1 + B2
  - Test each model on B3



Project B - Temporal Prospective Validation

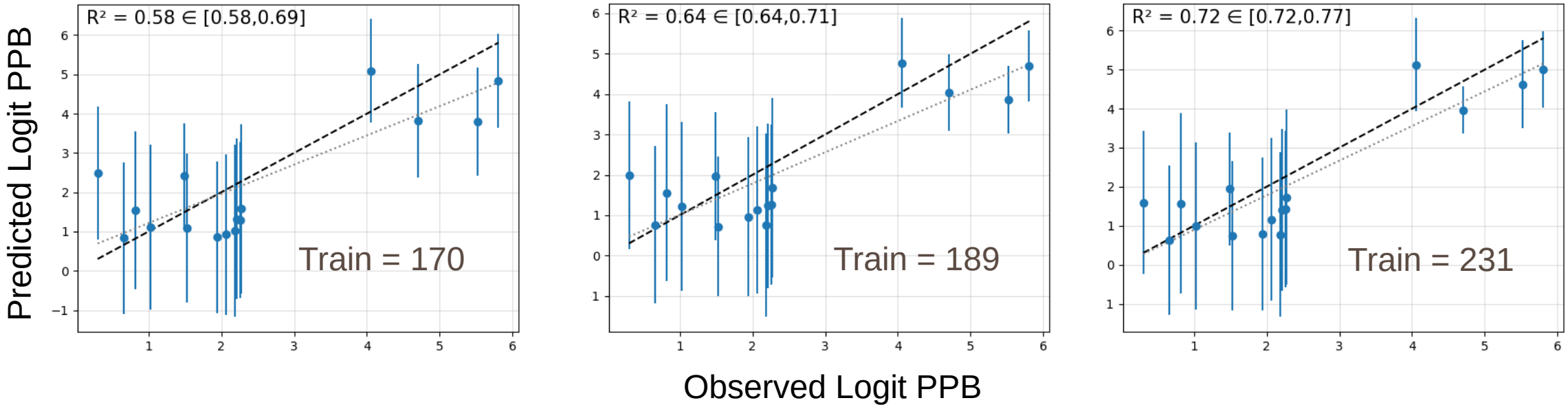


Increasing Data



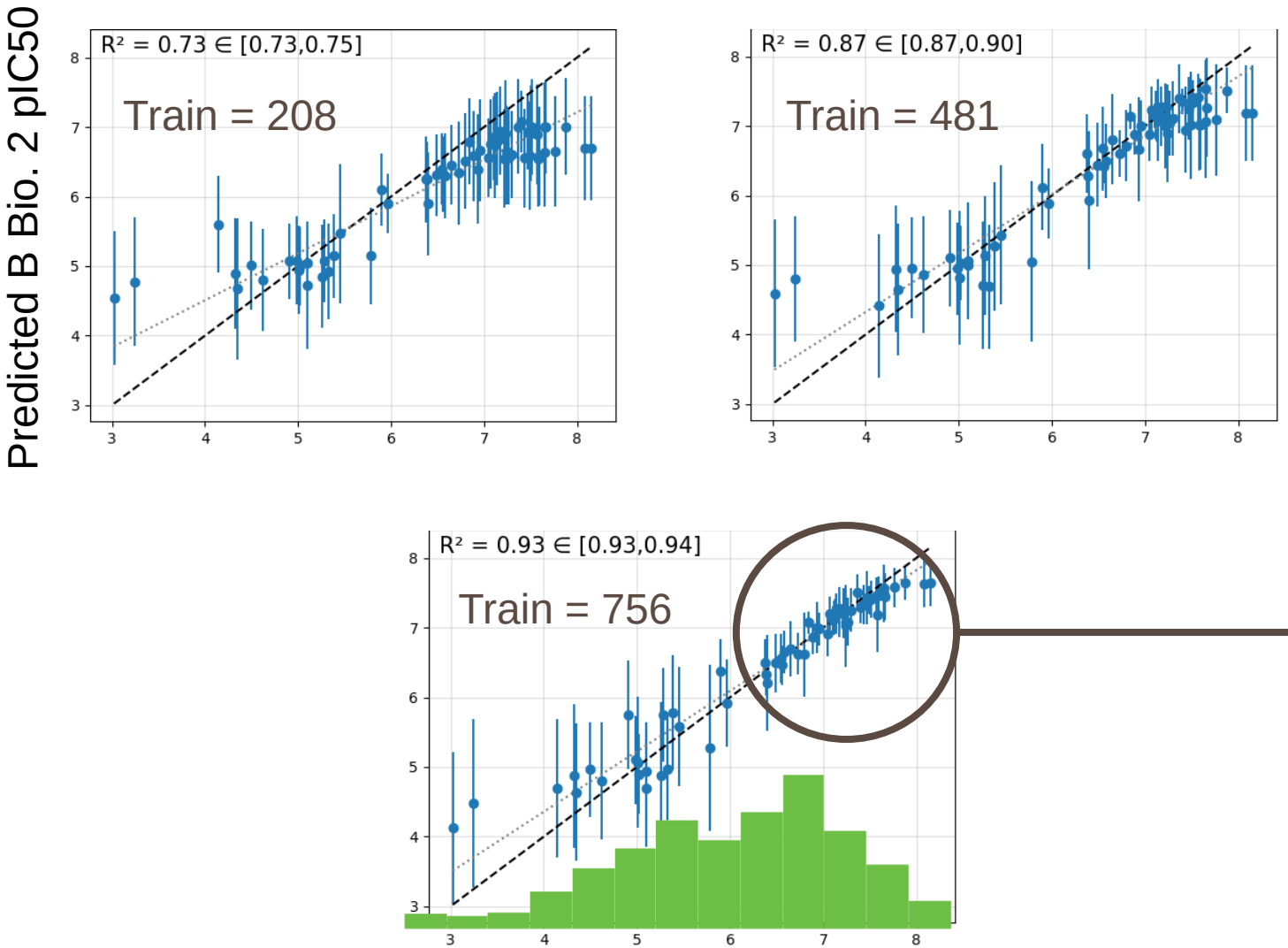
Increasing Data

ADME Human Plasma Protein Binding: Predicting Block 3



- Initial models can tell high from low
- Quality of predictions and error models improves with more data

Example of Activity Improving



- Activity
- Good model gets better
- Last model confident identifying **active compounds** better than µM

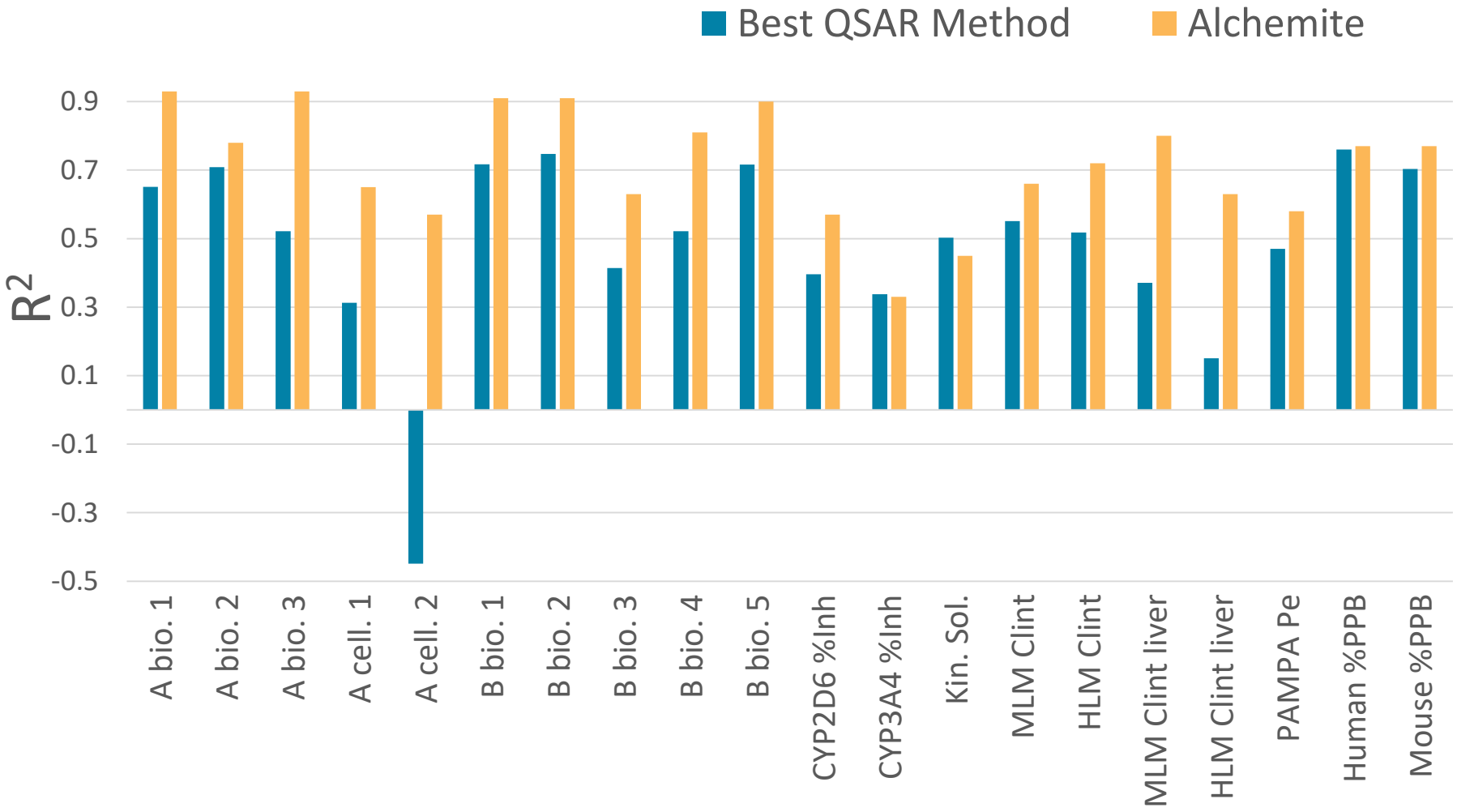
Comparison of Alchemite and QSAR

Single Alchemite model of Model 3 data set

Average R^2

<u>QSAR</u>
was 0.44
now 0.48

<u>Alchemite</u>
was 0.65
now 0.72

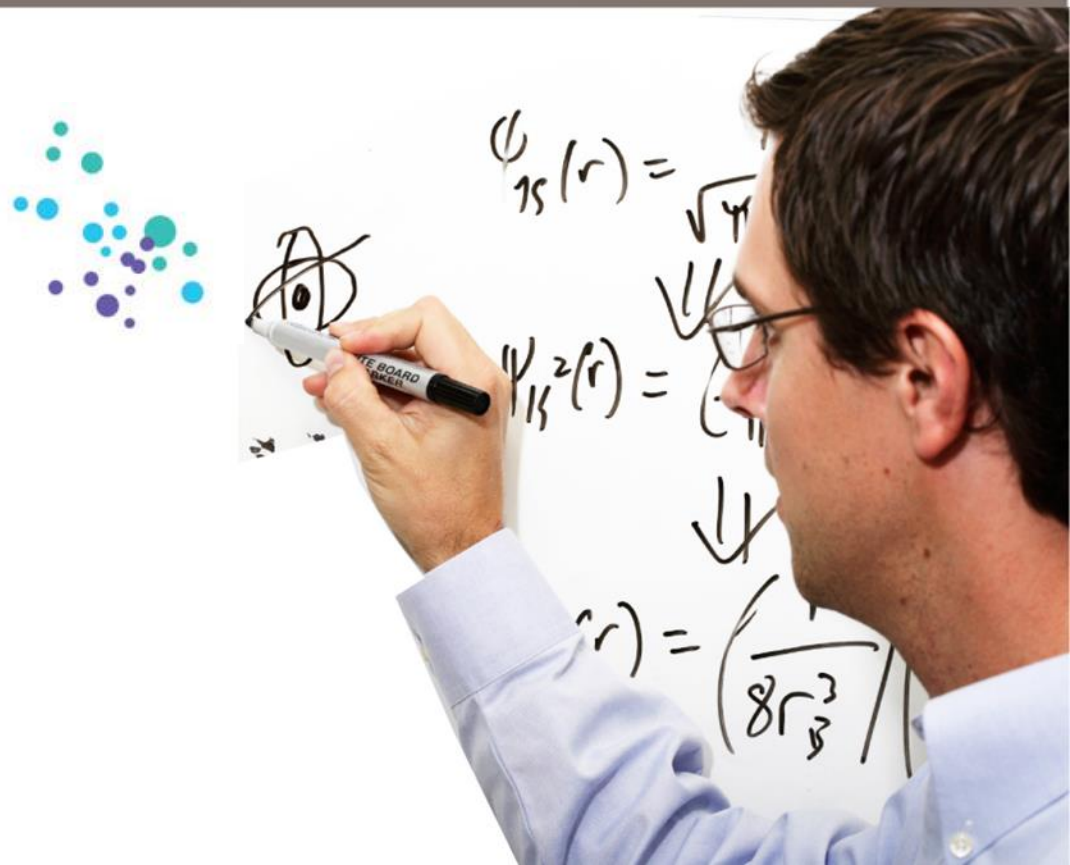


Part 2 - Conclusions

- Alchemite: Practical application of deep learning
  - Handles missing data and makes the most of extreme levels of sparsity
  - Provides robust uncertainty estimates on predictions
  - One model trained for all project data simultaneously, exploits assay-assay correlations
  - Retrainable to handle all stages of project which changes in time
- Alchemite can focus on the most confident and accurate results
- Alchemite models improve as data is added in a realistic chronological project series

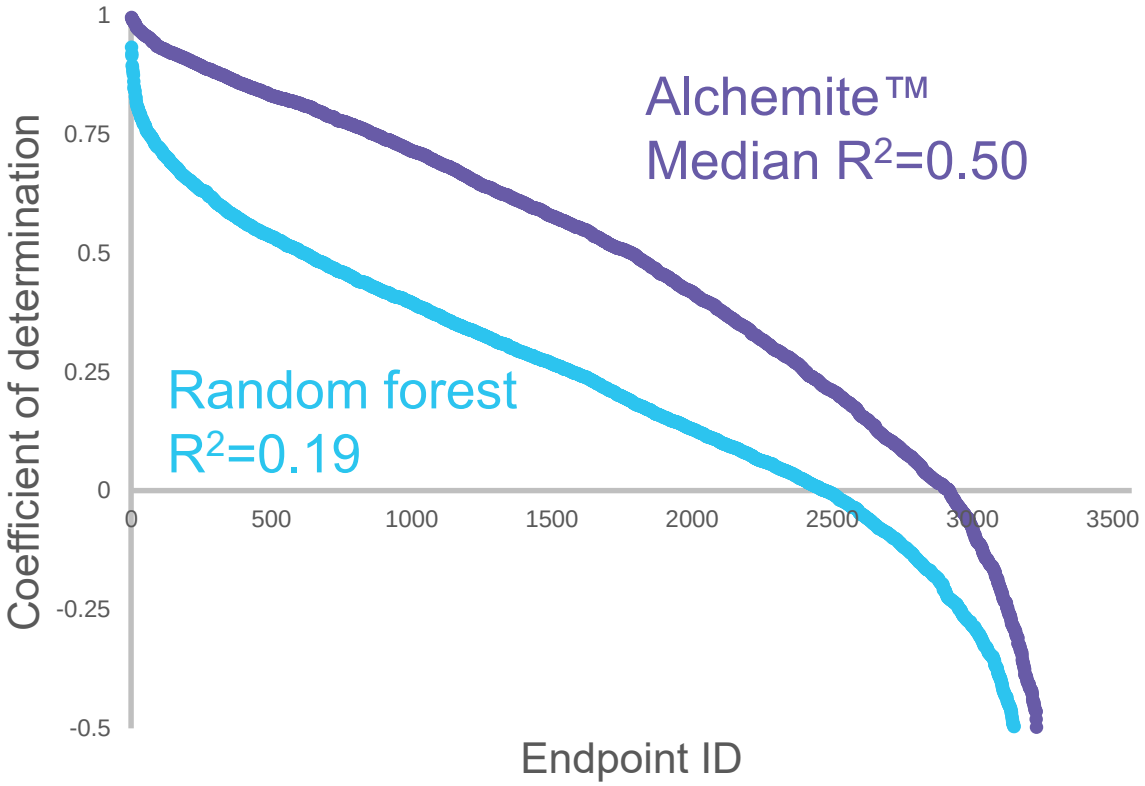
Application to Larger Datasets

optibrium



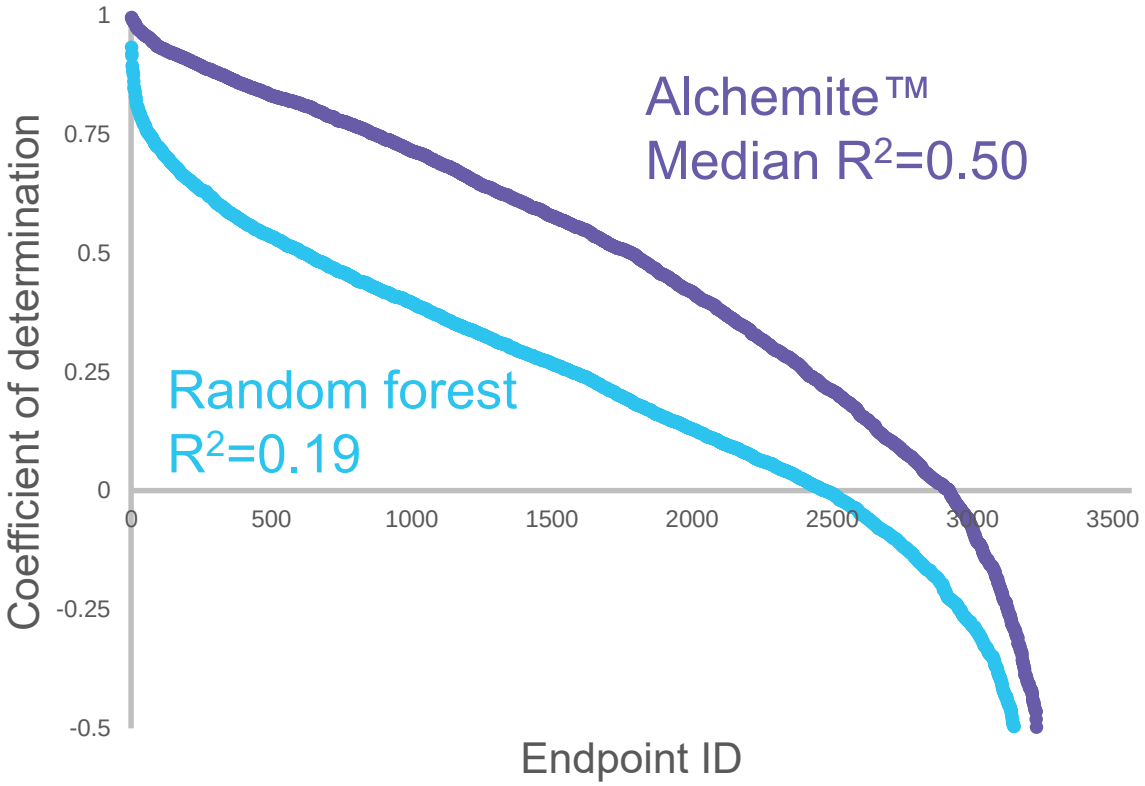
Global Pharma Collaboration

- **710,305** compounds
- 2,171 assays totaling **3,568** endpoints
- Covering a **full range** of drug discovery assays, including compound activities and ADME properties



Thank you for Listening!

- **Thanks to:**
  - Tom Whitehead, Gareth Conduit
  - Julian Levell
  - Matthew Segall, Peter Hunt
- **If you want to find out more:**
  - ben@optibrium.com
  - info@optibrium.com



optibrium

Constellation PHARMACEUTICALS

Intellegens