# Practical Applications of Deep Learning to Imputation of Drug Discovery Data

Webinar: 28th April 2020

**Presenters: Ben Irwin – Optibrium and Julian Levell – Constellation Pharmaceuticals**

**Host: Matt Segall – Optibrium**

# Today's Webinar Presenters and Host

**Ben Irwin**
Senior Scientist
Optibrium

**Julian Levell**
Vice President of Drug Discovery
Constellation Pharmaceuticals

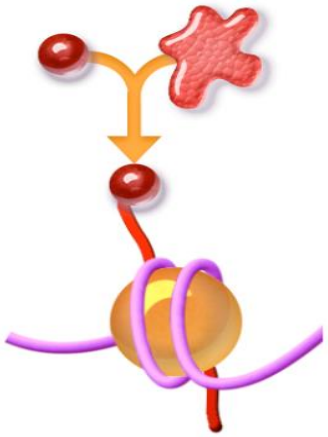
**Matt Segall**
CEO
Optibrium

# Constellation Pharmaceuticals

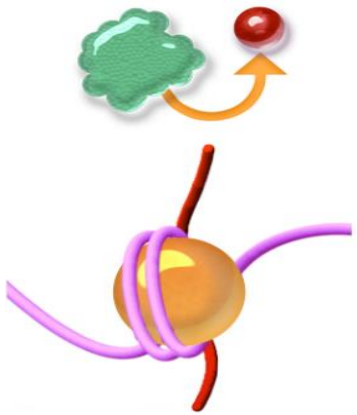
Cancer therapeutics via manipulation of transcriptional programs in tumor cells and immune cells

## Focused on Three Distinct Protein Target Classes That Operate on Chromatin

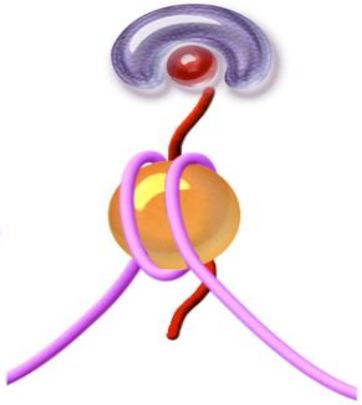
Writers



Erasers



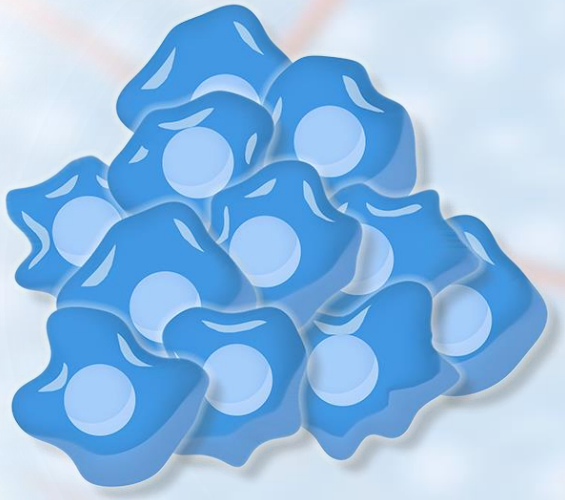
Readers



Trancriptional Control to Turn Genes On or Off

## Oncology Applications

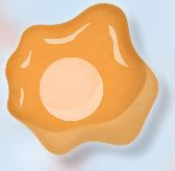
Tumor Cells



Target Transcriptional Networks That Result in Cell Death

T Cell

NK Cell



MDSC

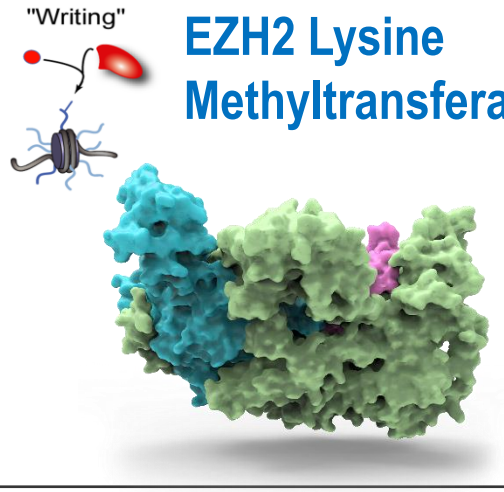
Re-program Immune Cells to Overcome Resistance to Cancer Immunotherapies

***Stellar Science, Breakthrough Medicine***

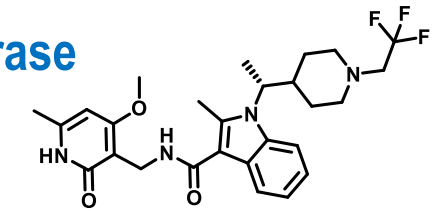
Constellation PHARMACEUTICALS

# Constellation Pipeline

Clinical programs and preclinical development candidates




## EZH2 Lysine Methyltransferase

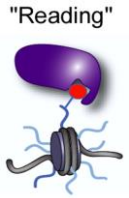


CPI-1205

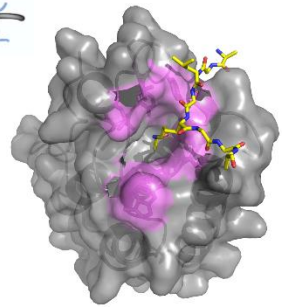
**Phase II**
mCRPC

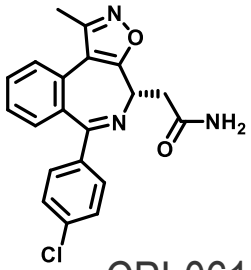
CPI-0209
Undisclosed

**FIH**



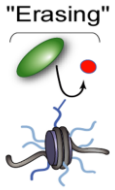

## BET Proteins



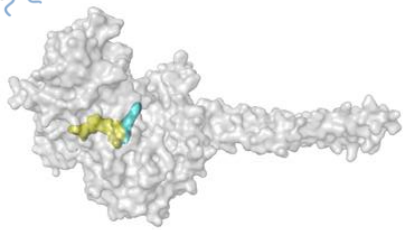


CPI-0610

**Phase II**
MF

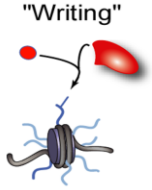



## LSD1 Lysine Demethylase

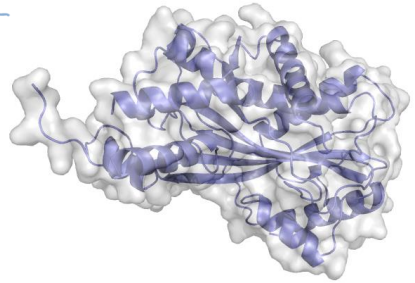


CPI-482
Undisclosed

**Pre-IND**




## CBP and EP300 Lysine Acetyltransferases



CPI-2429
Undisclosed

**Pre-IND**

***Stellar Science, Breakthrough Medicine***

Constellation PHARMACEUTICALS

# Alchemite™ Applied to Constellation Programs

## Scope of deep learning & data sharing collaboration

### Inhibitors of CBP and EP300 Lysine Acetyltransferase

**Completed program**
**Mostly closed, complete dataset**

### Ongoing Undisclosed Discovery Program

**Ongoing hit-to-lead program**
**Modest initial dataset, plus batchwise new datasets**

***No structures disclosed : shared StarDrop molecular descriptors plus all primary biochemical, cellular and ADME data***

***Recent Publications covering aspects of the CBP/EP300-HAT program:***

- Make the right measurement: discovery of an allosteric inhibition site for p300-HAT
  Gardberg *et al*, *Struct. Dyn.* **2019**, *6*, 054702
  [https://doi.org/10.1063/1.5119336]

- Early Drug-Discovery Efforts towards the Identification of EP300/CBP Histone Acetyltransferase (HAT) Inhibitors
  Huhn *et al*, *ChemMedChem* **2020** (in press)
  [https://doi.org/10.1002/cmdc.202000007]

- Discovery of CPI-1612: A Potent, Selective, and Orally Bioavailable EP300/CBP Histone Acetyltransferase (HAT) Inhibitor
  Wilson *et al*, *ACS Med. Chem. Lett.* **2020** (in press)
  [https://doi.org/10.1021/acsmedchemlett.0c00155]

Constellation PHARMACEUTICALS

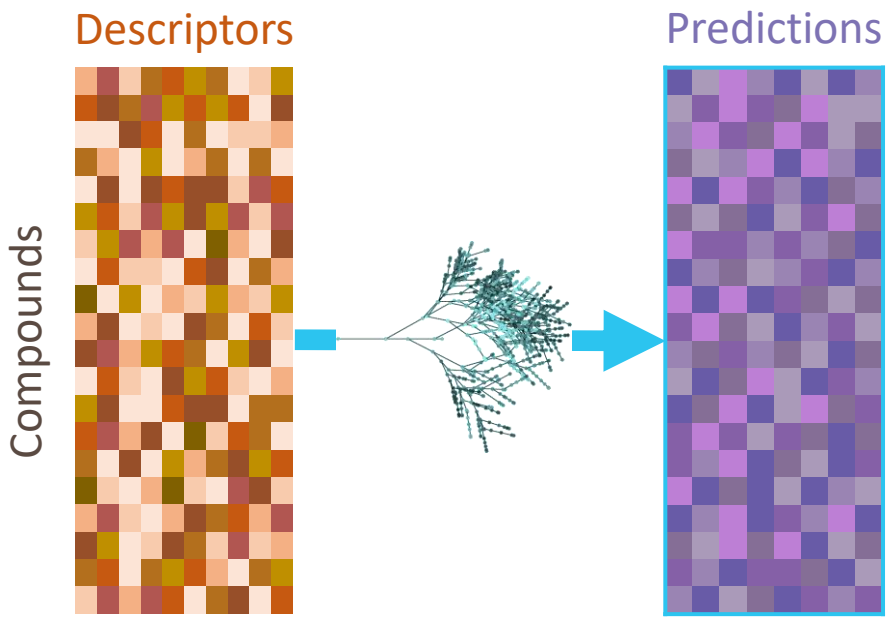
# Overview

- **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**

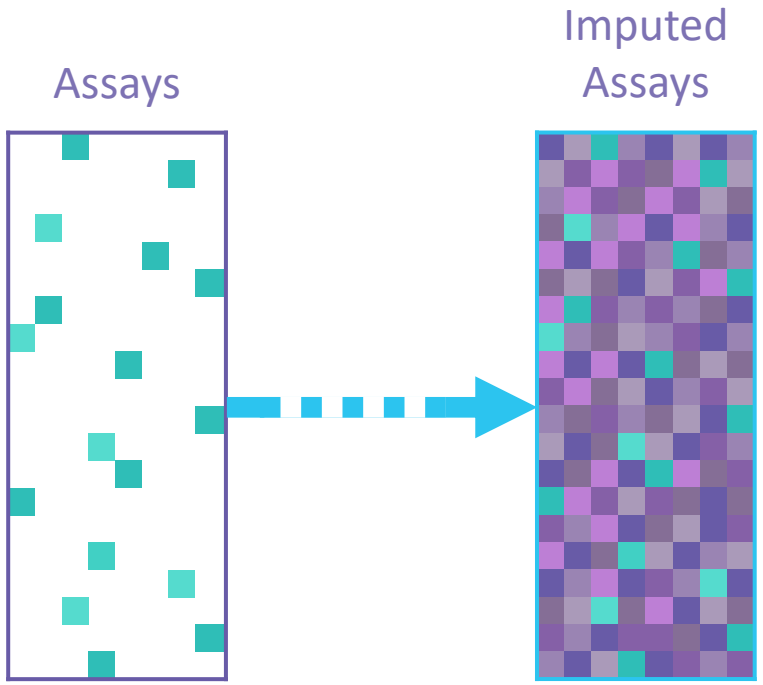
# Prediction vs. Imputation

- 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 missing data in sparse data using the limited data that are already available




**Prediction**




**Imputation**

# Problems with Pharma Data

optibrium

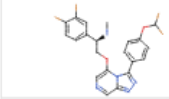
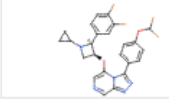
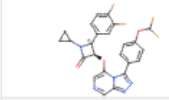
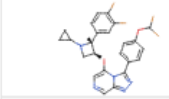
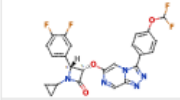
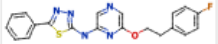
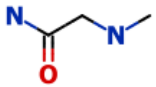
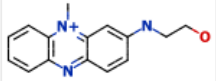
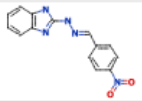
# Problems with Pharma Data

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

| | |
|---|---|
| Missing Values | Noisy Data |
| Multiple Endpoints | Data Changing with Time |

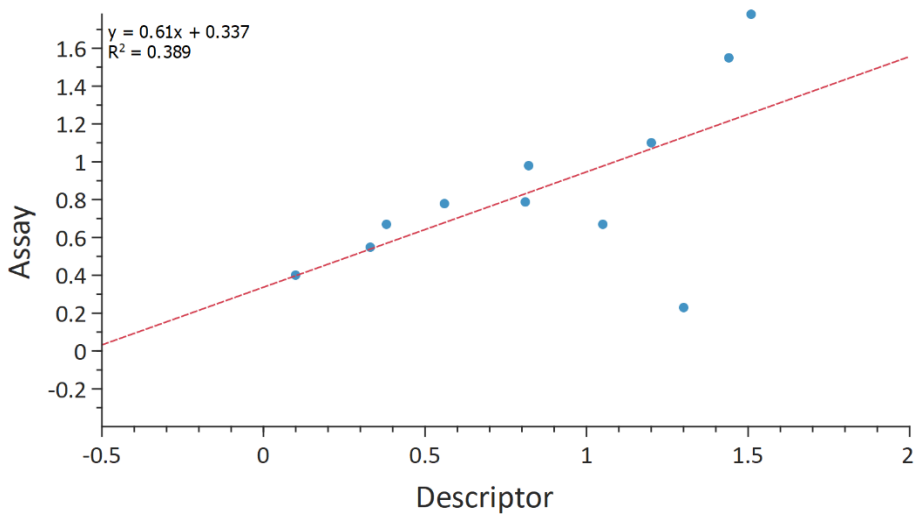
# Missing Values

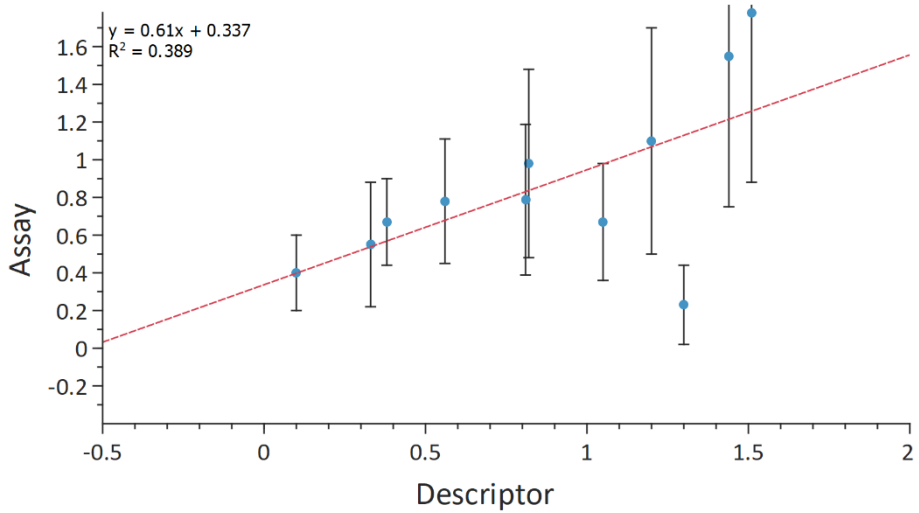
- Problem:
  - Most algorithms cannot handle missing inputs
  - $y = f(x_1, ?, x_3, x_4, ?)$
  - Simple methods to impute give poor quality results e.g. imputation via mean
  - $y \neq f(x_1, \overline{x_2}, x_3, x_4, \overline{x_5})$
- Solution:
  - Algorithm should make the most of data present
  - “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”
  - Models output numbers with no context
- Solution:
  - Account for input noise
  - Predictions should come with confidence values!
  - Highly confident predictions are more valuable than weak ones
  - Provide a big error bar if the model doesn’t know the answer



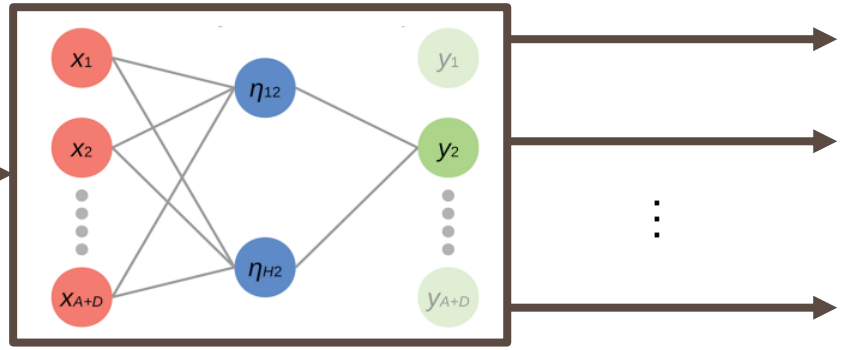
# Multiple Endpoints – One Model

- Problem:
  - **Many columns in project data:** can't train a model for each one…
  - Activity $IC_{50}$, $EC_{50}$: protein, supersome, cell
  - Multiple targets: related and unrelated
  - Absorption, distribution, metabolism, and excretion (ADME)
    - Plasma protein binding, intrinsic clearance, CYP inhibition, permeability, solubility

- Solution:
  - One model to handle everything

(Noisy Fragmented Data)

$X_{ij}$




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

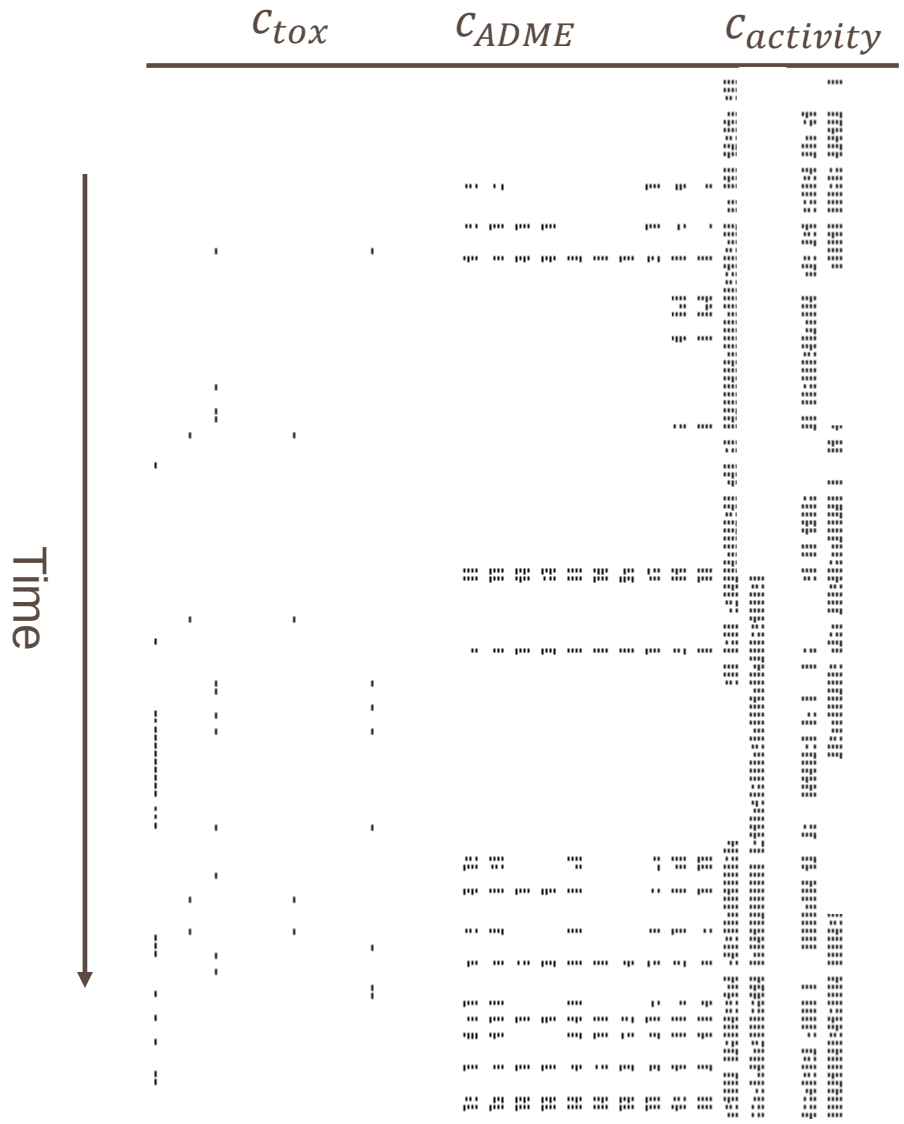
Cell Assay $\pm\ \sigma_{i2}$

⋮

ADME $\pm\ \sigma_{iN}$

# Changing with Time

- Problem:
  - Data are evolving as project continues
  - Chemical space changes
  - Activity changes i.e. increasingly active compounds are discovered
  - Data sparsity changes (more ADME, less HTS)
  - Uncertainties change (multiple replicates, finer resolution)
- Solution:
  - Models which extrapolate well
  - Retraining the models as appropriate
  - Temporal validation



# 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.


## Imputation versus prediction: applications in machine learning for drug discovery

Benedict W J Irwin*[,1,2], Samar Mahmoud[1], Thomas M Whitehead[3], Gareth J Conduit[2,3] & Matthew D Segall[1]

[1]Optibrium Limited, Cambridge, CB25 9PB, UK

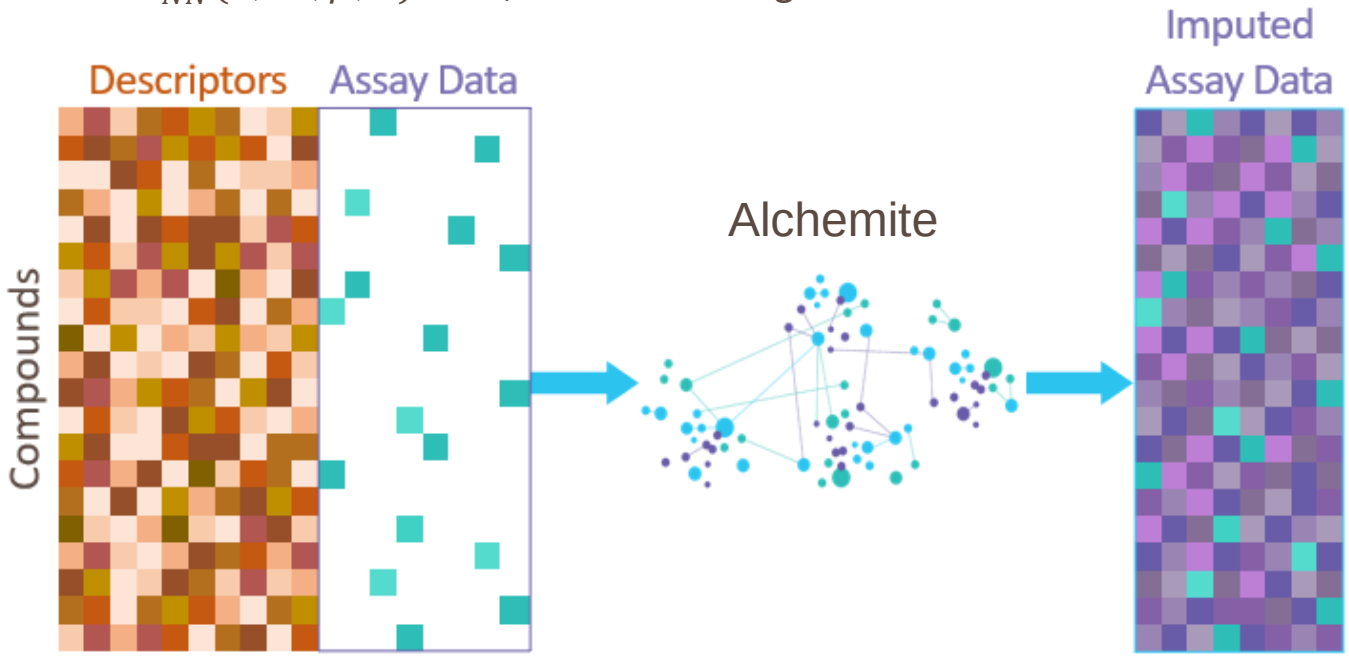
Future Drug Discovery

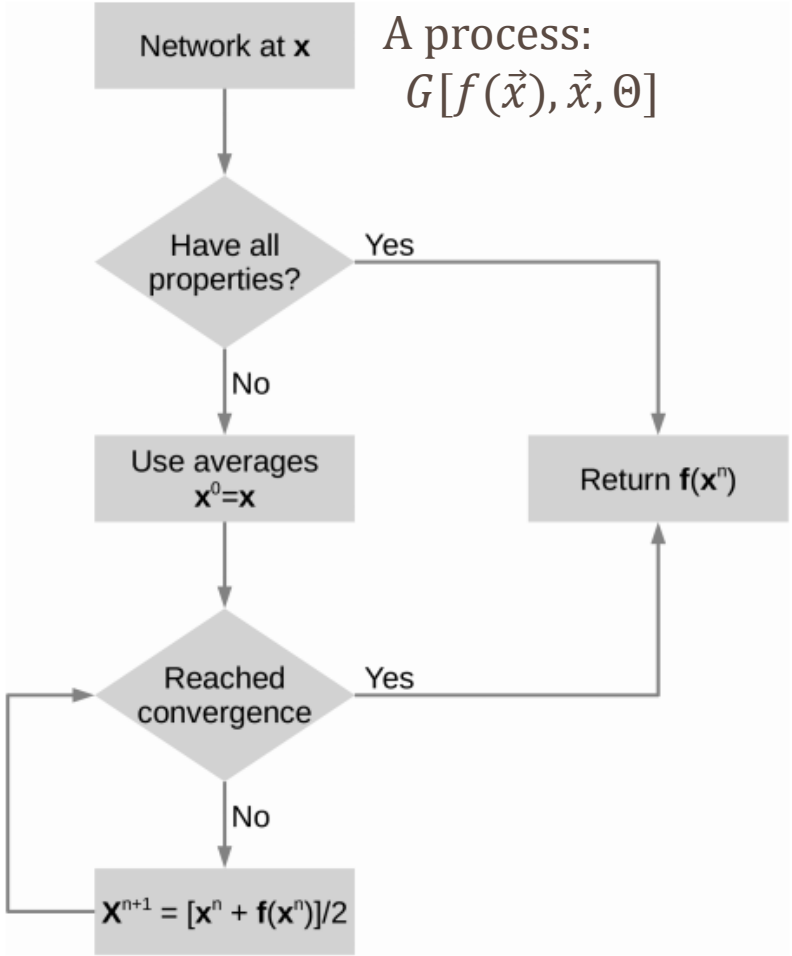
Whitehead et al.
J. Chem. Inf. Model. 2019, **59**, pp. 1197-1204

# Alchemite – A Method for Deep Multiple Imputation

intellegens

- 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.



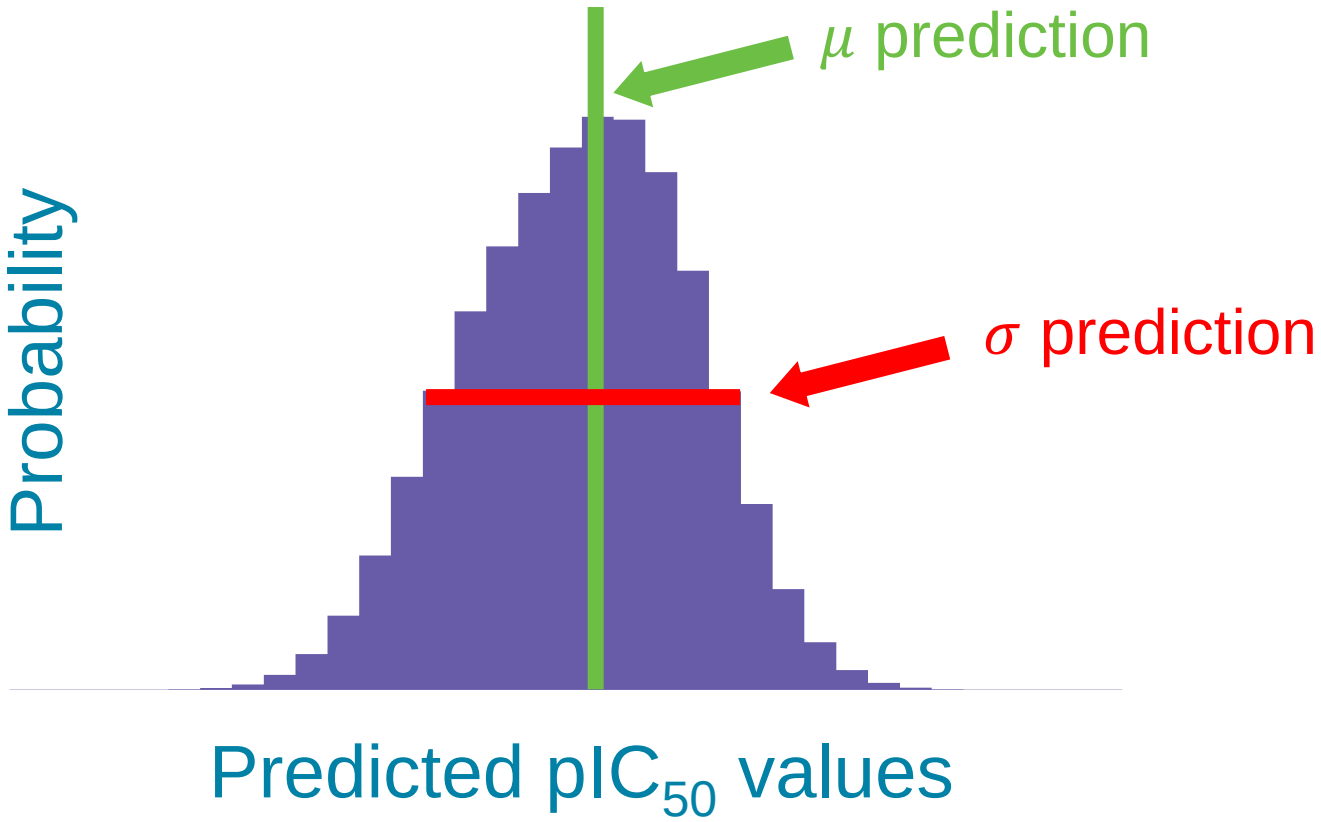




A process:
$G[f(\vec{x}), \vec{x}, \Theta]$

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



# Practical Application of Deep Learning to Project Data

optibrium

# Initial Project Data

- Two Projects
  - A: Completed project (CBP/EP300-HAT)
  - B: Ongoing project that had recently commenced

Constellation PHARMACEUTICALS

| 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 |

- Additional data points for Project B compounds were measured for imputed data points after completion of the models

# Initial Models – Objectives

- Compare accuracy of Alchemite model to conventional QSAR models
  - Does Alchemite add value in the limit of small data sets?
- Compare models built on all data simultaneously with those built on individual projects and subsets of data
  - Can deep learning handle the complexity of different chemical spaces and endpoints in a single model?
- Evaluate Alchemite's ability to estimate confidence in individual predictions and target the most accurate results

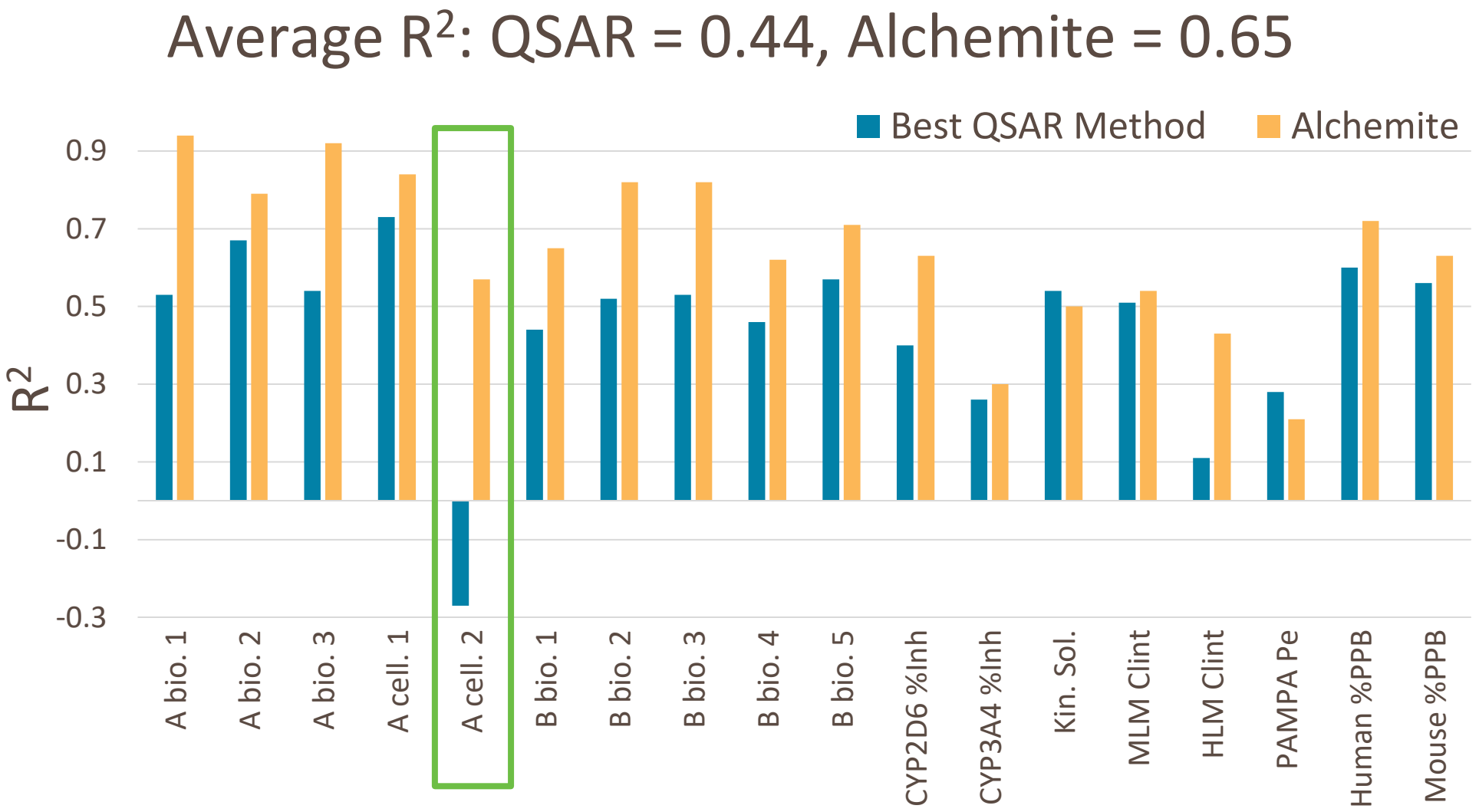
# Initial Models – Approach

## Compare three types of models

- 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

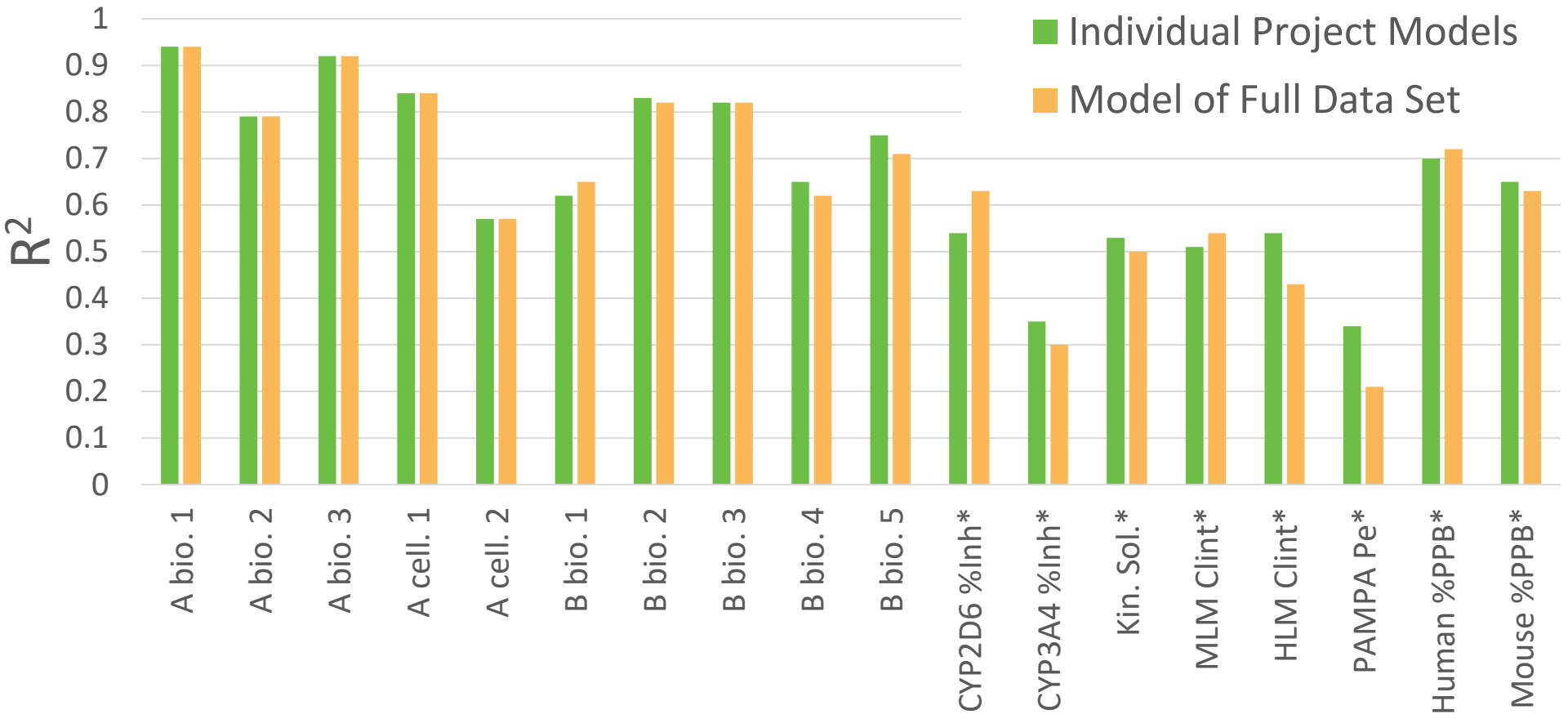
# Comparison of Alchemite and QSAR

## Single Alchemite model of combined data set



# Single Model vs Individual Project Models

## Single model performs equivalently to individual project models

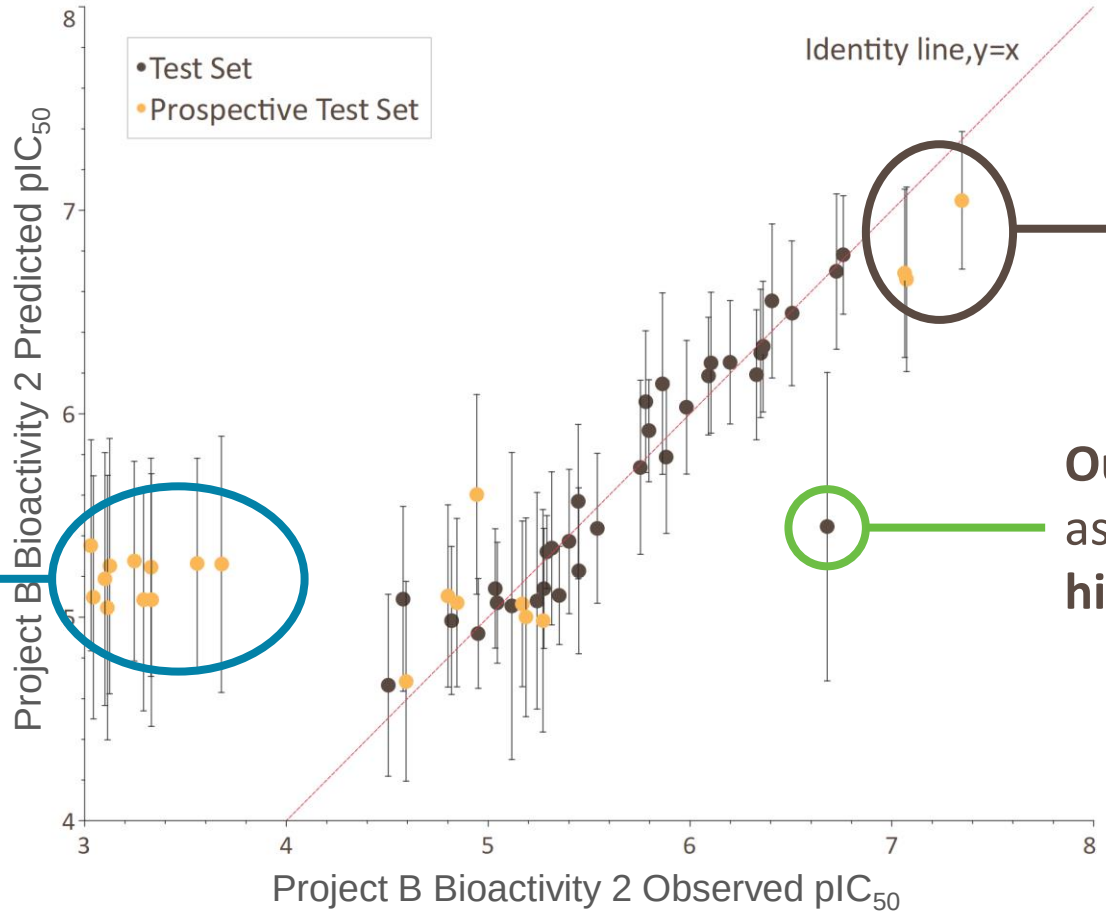



* Individual project model for ADME properties built and tested on Project A only. Full data set model tested against both projects.

# 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

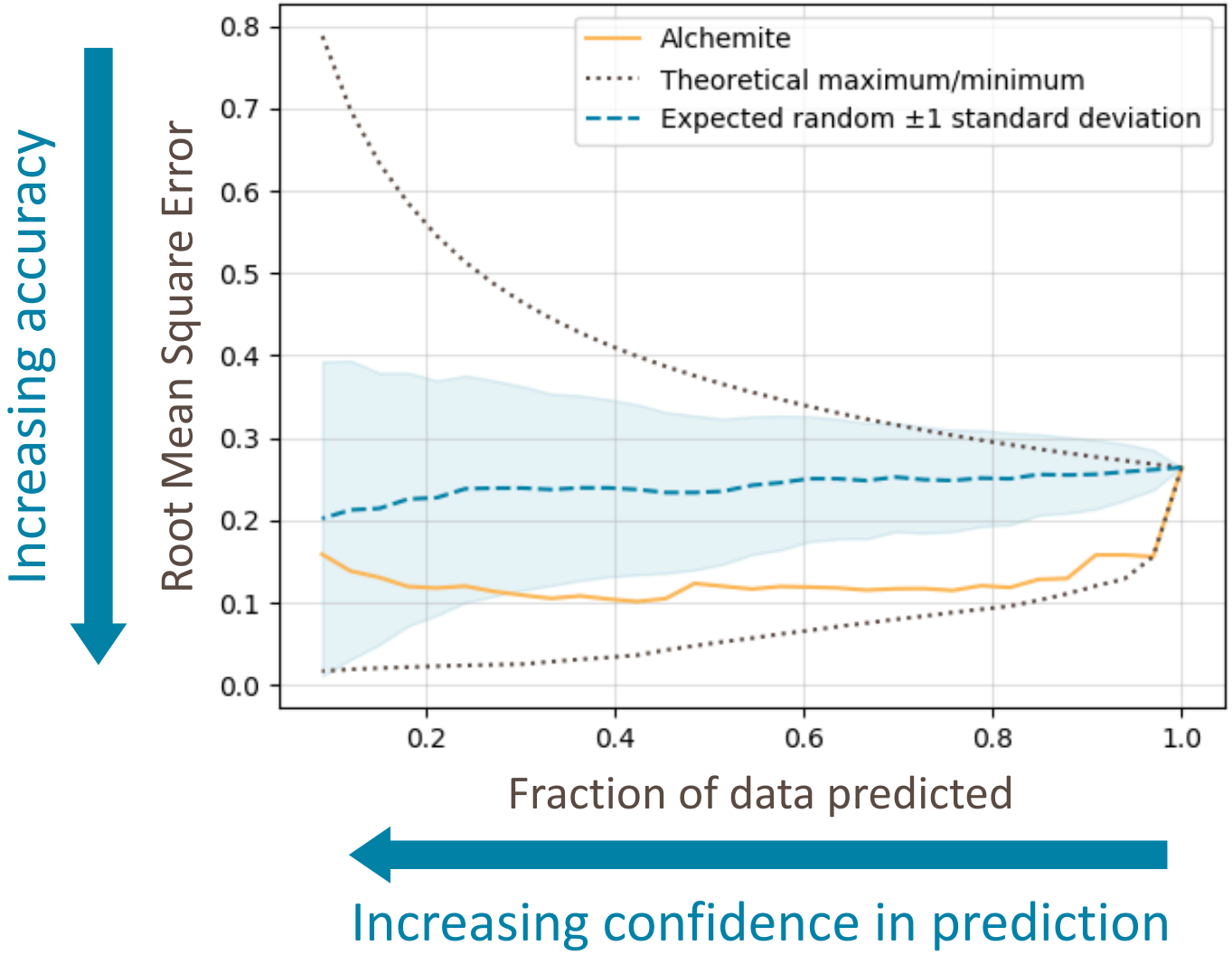



**New active compounds** correctly identified as active

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

# Identify and Discard the Least-Confident Predictions

## Project B Bioactivity 2

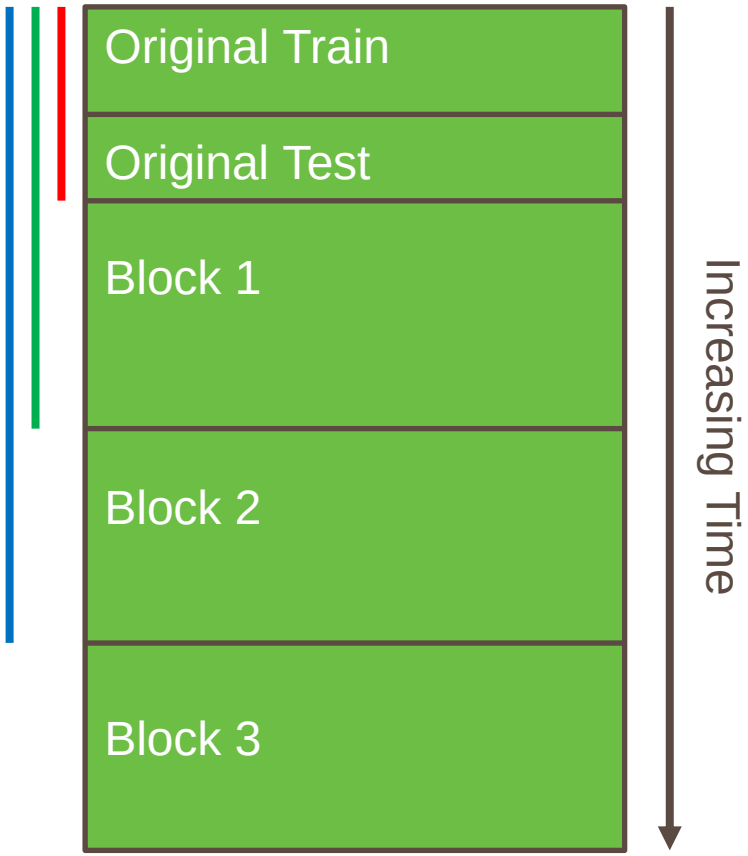


# Conclusions from Initial Models

- Alchemite significantly outperforms QSAR models
- Independent and prospective test set performance 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
- **Next steps… Application to new compounds and data as project progresses**

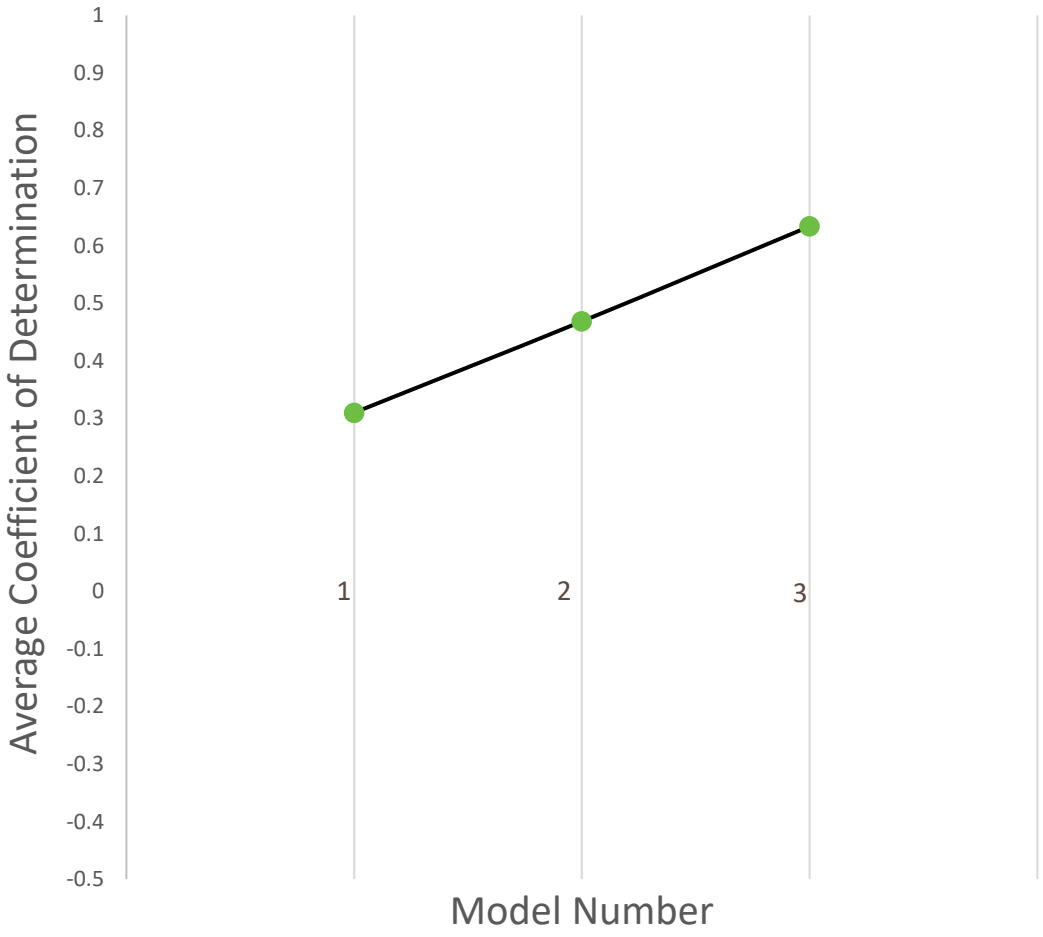
# Temporal Prospective Validation

- 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

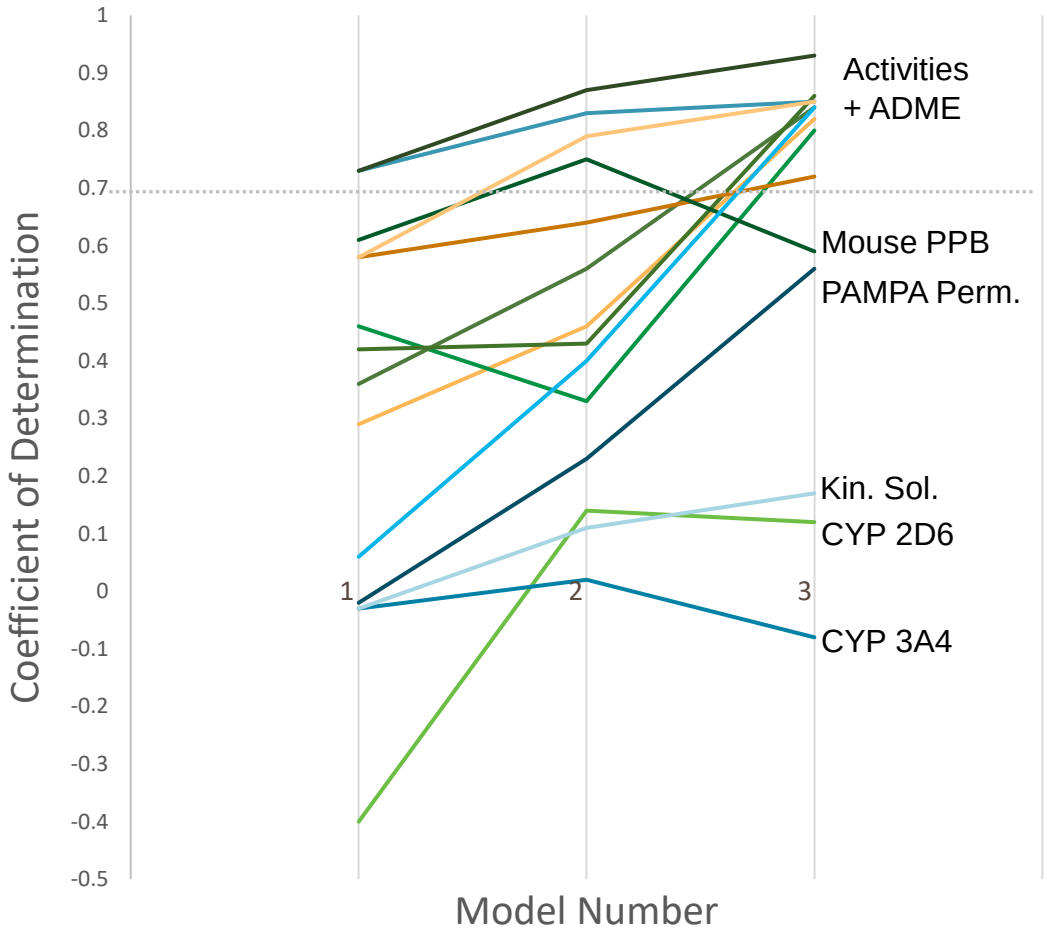


# Project B - Temporal Prospective Validation

## Performance on Block 3 (most recent) data

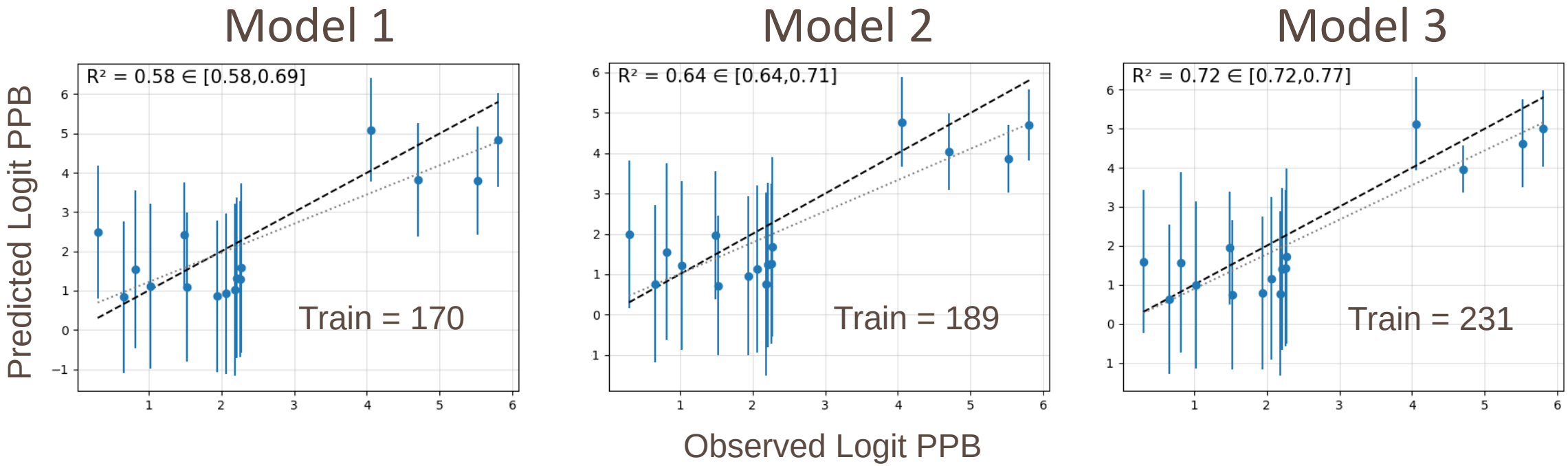



Increasing Data



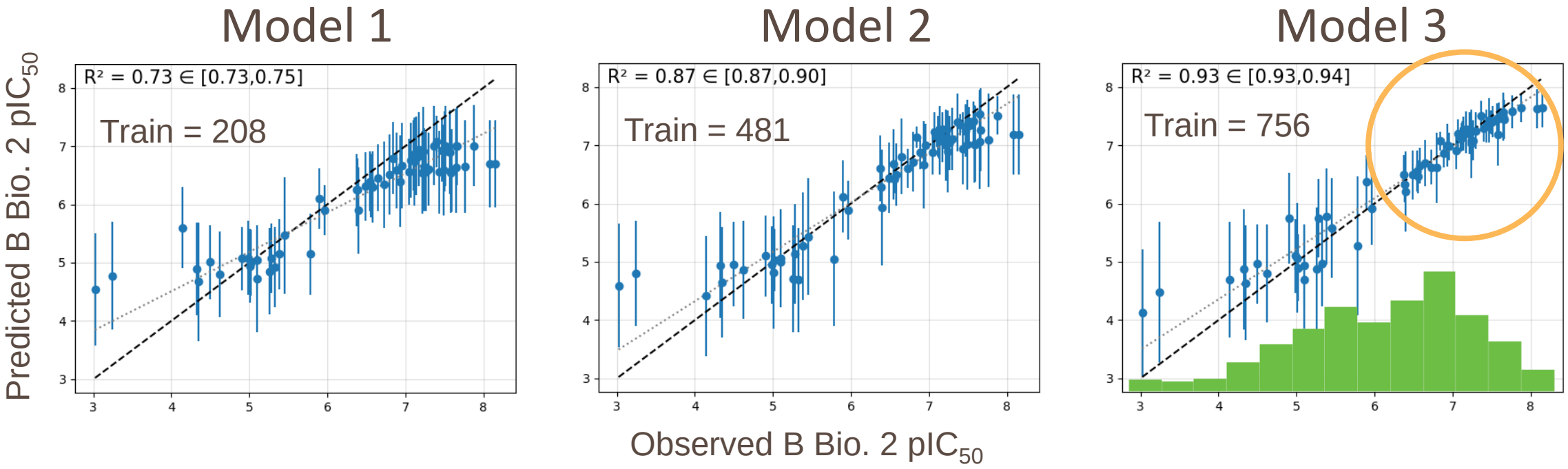

Increasing Data

# ADME Human Plasma Protein Binding: Predicting Block 3




- Initial models can't tell high from low
- Quality of predictions and error models improves with more data

# Example of Activity Improving: Predicting Block 3




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

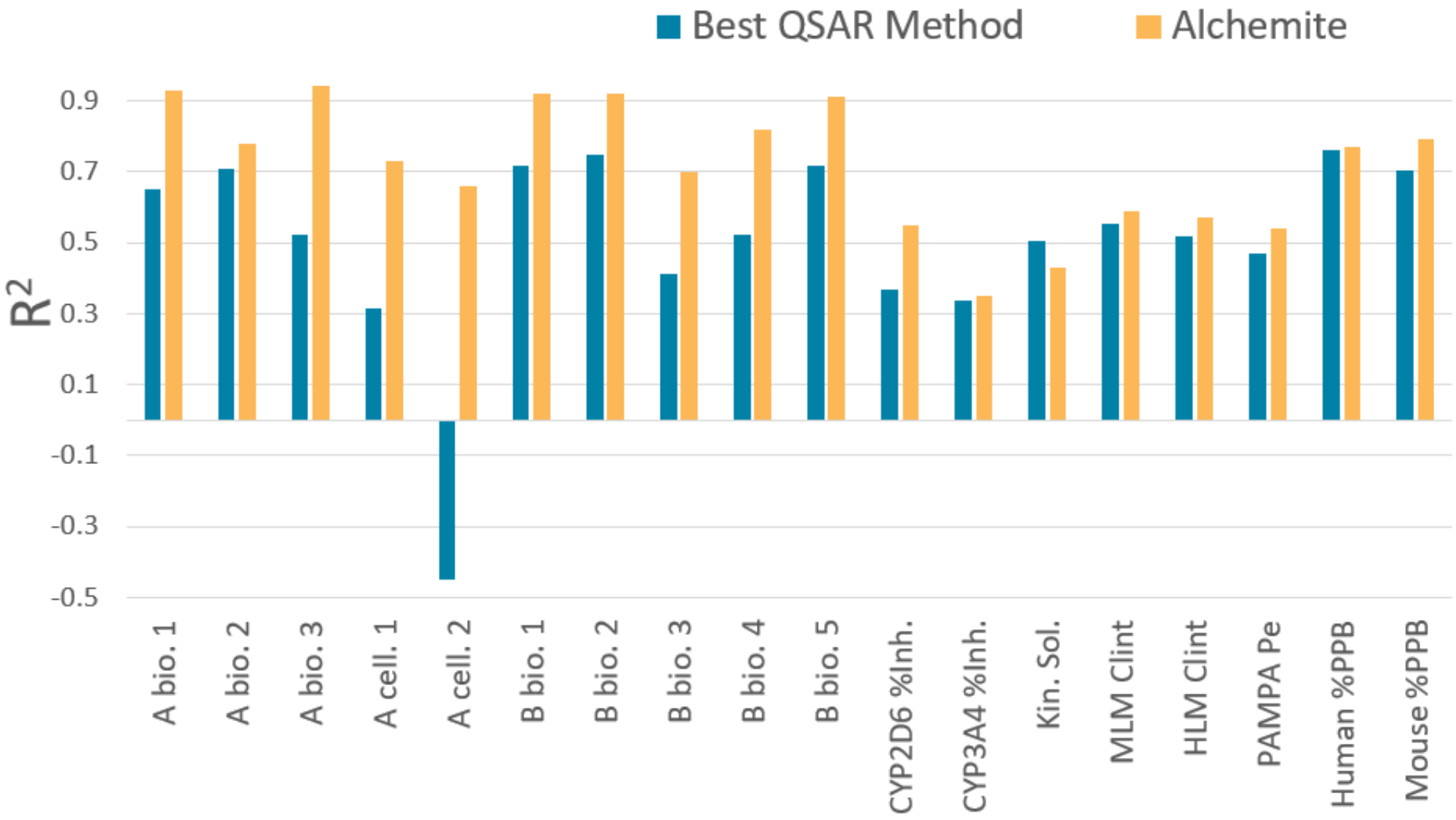
# Comparison of Alchemite and QSAR

## Single Alchemite Model – 20% independent test set

Average $R^2$

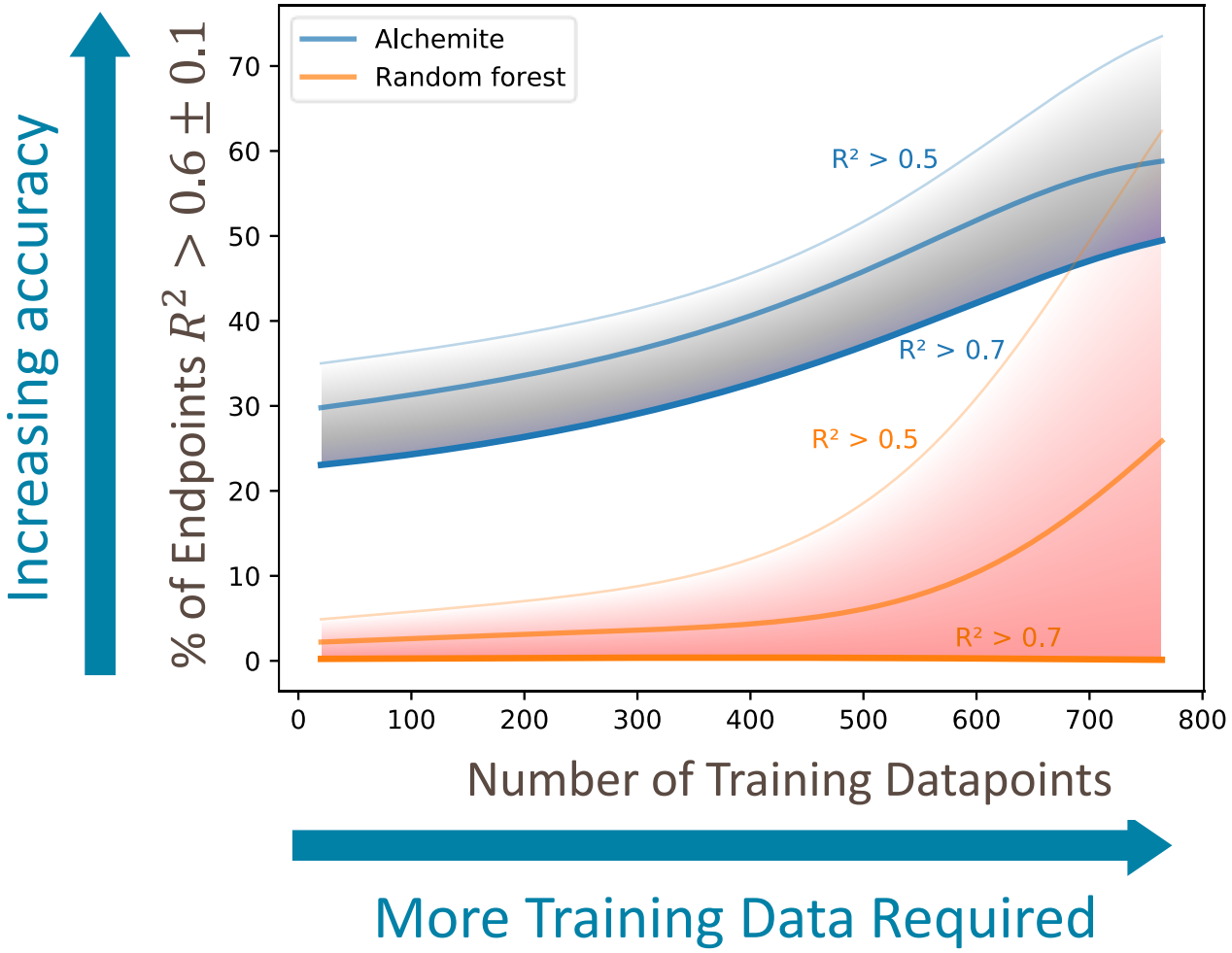
<u>QSAR</u>
was 0.44
now 0.50

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



# Make Better Use of Data

## Averaged over all Endpoints




More Training Data Required

# Part 2 - Conclusions

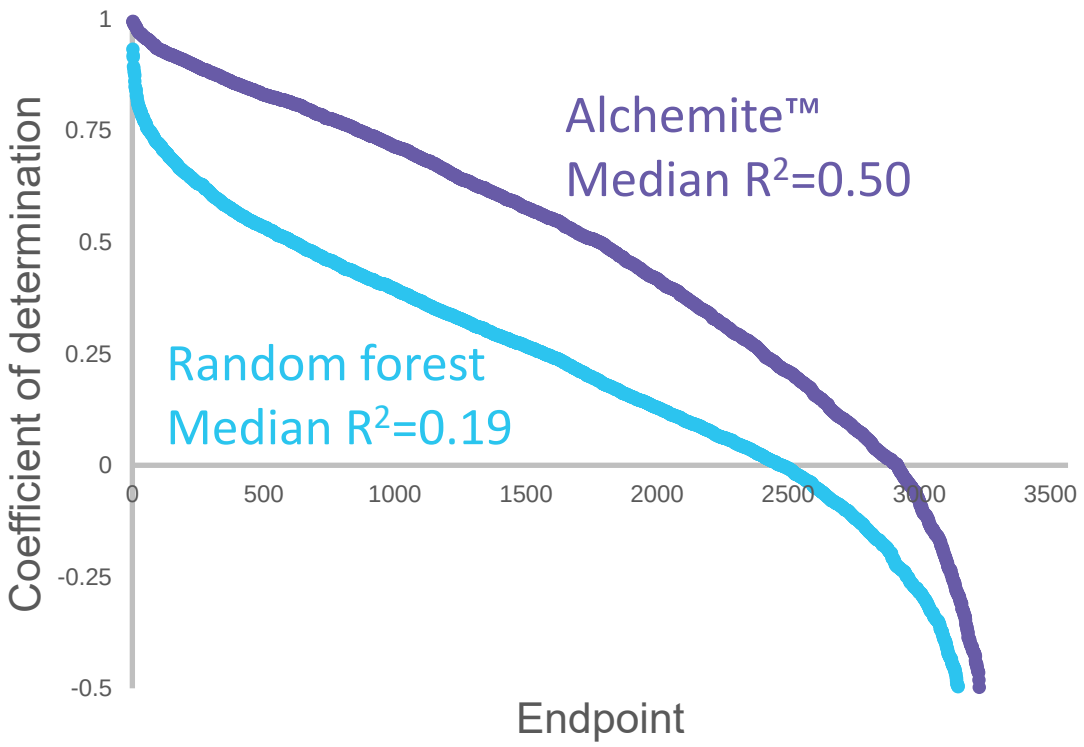
- Alchemite: Practical application of deep learning
  - Handles missing data and makes the most of extreme levels of sparsity
  - Provides robust uncertainty estimates on predictions
  - One model trained for all project data simultaneously, exploits assay-assay correlations
  - Retrainable to handle 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

# Alchemite™ Application to Global Pharma Data

- Application to large data set
  - **710,305** compounds
  - 2,171 assays totaling **3,568** endpoints
  - Less than 1% complete
- Covering a **full range** of drug discovery assays, including compound activities and ADME properties
- Join our webinar on Tuesday 26th May to learn more:
  - "Large scale imputation of drug discovery data using deep learning"

Takeda | ONCOLOGY



# Non-Proprietary Value Aspects of Alchemite™

## Some overarching learnings and caveats

**Confidently deprioritizing the synthesis of new target molecules**

- Confidently predicted inactives: few false negatives
  - Can save substantial resources or repurpose to higher value targets by limiting the number of predicted inactive compounds made
  - Still need to make the compounds with structurally distinct changes, but overall could avoid ~10 to 20% of irrelevant target molecules.
- Activity prediction improved with potency but false negatives observed, mostly in predictions with low confidence
  - All false negatives were structurally outside of the SAR for the training set
  - Not comfortable to only make predicted active compounds, so also explored compounds predicted to be inactive with low confidence

**Identifying outliers in measured datasets**

- Empty well data, and (for example) solubility driven artifacts in permeability & off-target datasets can be identified
  - Important to pay attention to the confidence in the predicted data (eg. color plots by error and only pay attention to outliers with high confidence)
  - Testing or data for close structural analogs, and / or retesting confirmed the issues is several cases
- Avoid discarding good molecules for further profiling, or discarding subseries for further exploration due to incorrect measured data

**Caveats**

- Need at least some base datasets to build the initial model – could need over a hundred molecules to reach a good level of confidence
- Chirality : descriptors used did not include a chirality factor & many compounds were not assigned absolute stereochemistry due to achiral synthesis (and separation to test multiple isomers)
  - Obviously, stereochemistry can have a profound effect upon on- and off-target activity as well as ADME profiles.
  - Could add stereochemistry descriptors to explore if this solves for the problem. However, this will not solve for data which is based on unknown stereochemistry (eg. R and S enantiomers across the series are separated by a variety of different columns / methods but absolute stereochem is either not known or not unambiguously assigned in the database)

Constellation PHARMACEUTICALS

# Deep Learning Collaborations

Augmented Chemistry™

## www.augmentedchemistry.ai

- Application of Alchemite offered on a collaborative basis
- Example applications include:
  - 'Fill in the gaps' in your database with confident results to target high-quality compounds
  - Identify your most valuable compounds and the most important experiments to perform
  - Run virtual screens to find new starting points for your projects
- Based on a discussion of your data and objectives, we can provide a tailored project proposal