A Global Deep Learning Model for Global Health Drug Discovery

27th April 2021

Samar Mahmoud – Optibrium Limited

Overview

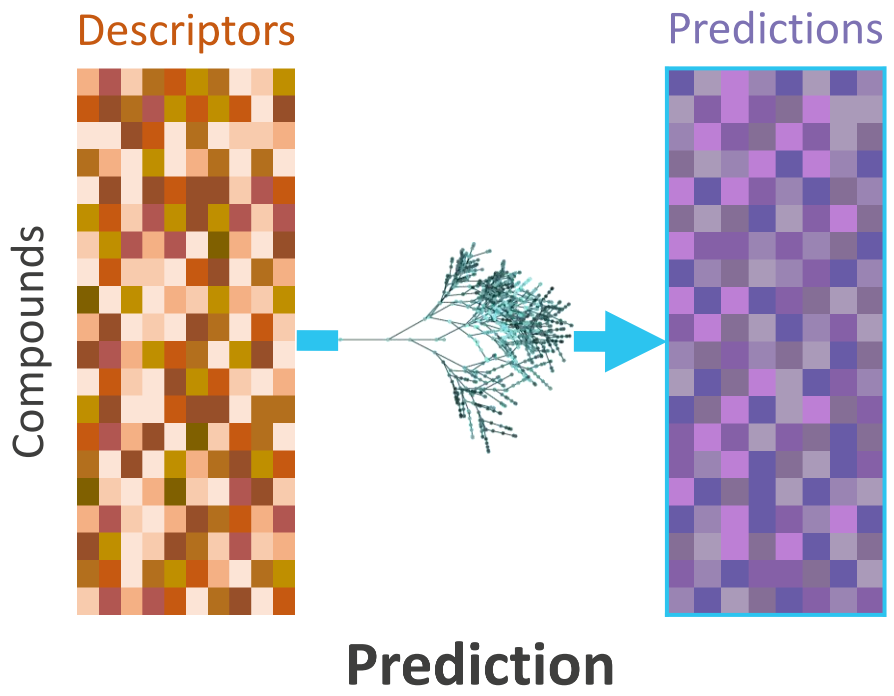
- Introduction to deep learning imputation using Alchemite™
- Data set and objectives
- Model validation
  - Comparing global and project-specific models
  - Assessing model confidence estimates
- Application of a global deep learning model to project optimisation
  - Multi-parameter optimisation for an anti-TB therapeutic objective
- Conclusions

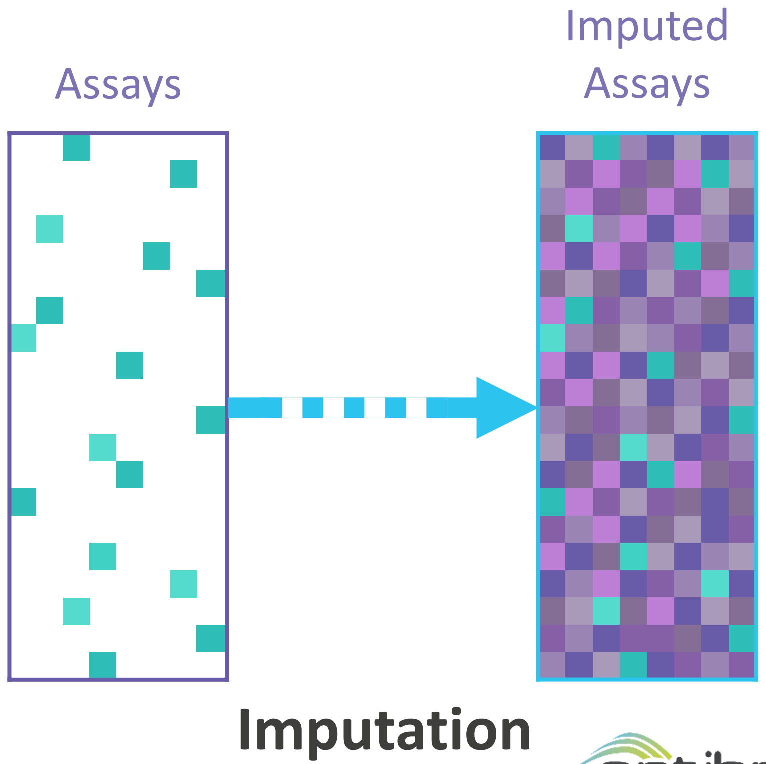
Introduction to Deep Learning Imputation using Alchemite™

Prediction vs. Imputation

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

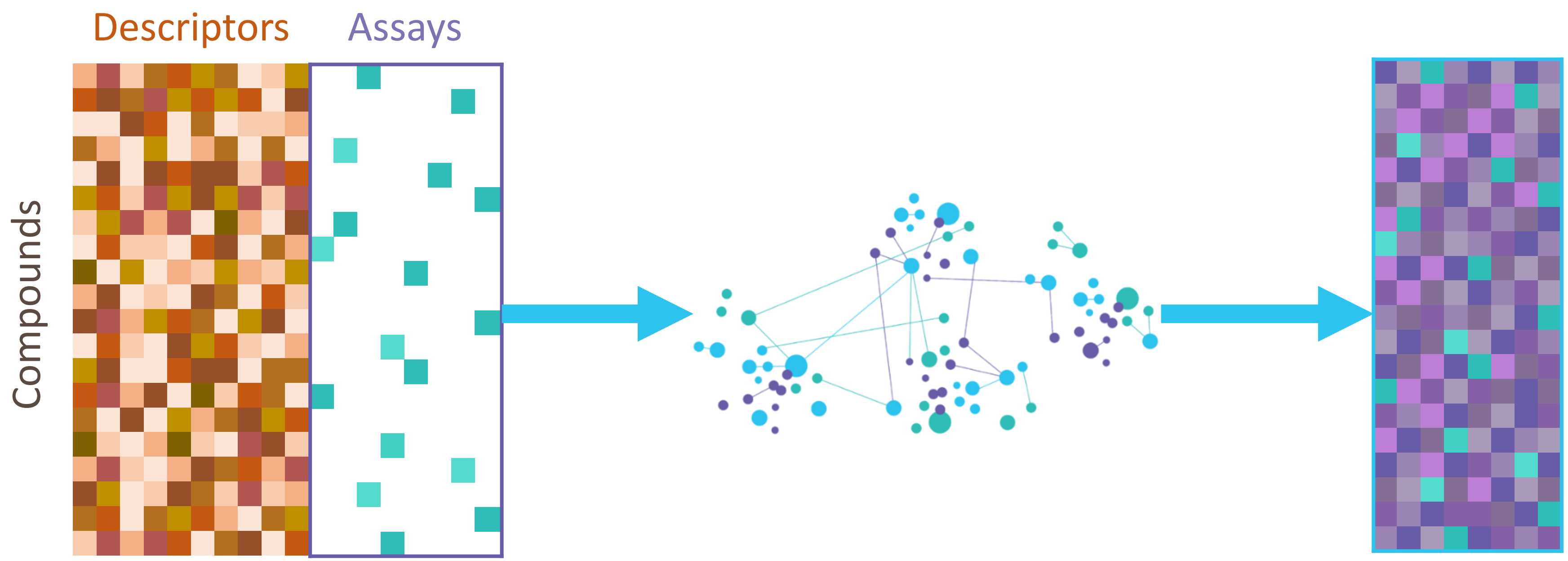
- Imputation is the process of filling in the gaps in sparse experimental data using the limited results that are already available





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

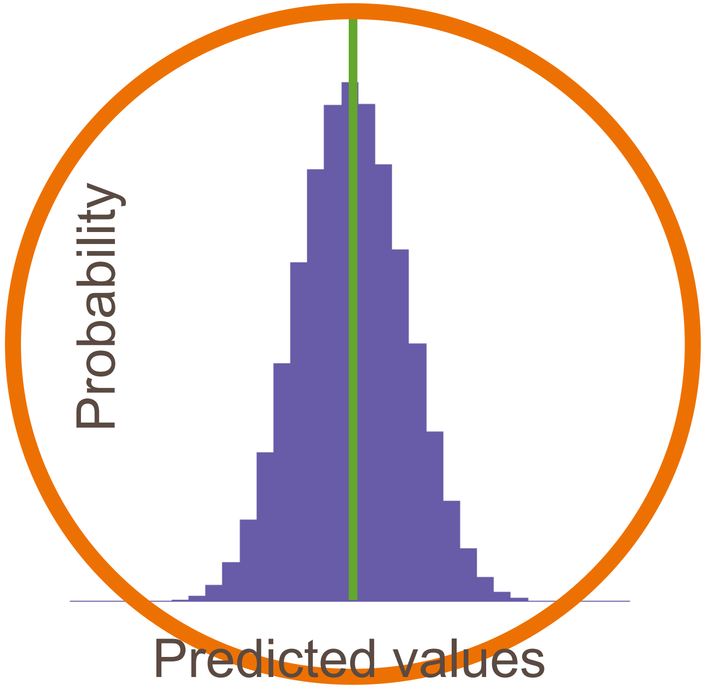
- Learns directly from relationships between experimental endpoints as well as SAR
  - Makes better use of sparse and noisy experimental data than conventional QSAR models
- 'Fills in' the gaps in your data and makes predictions for 'virtual' compounds
  - Generates more accurate predictions to target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

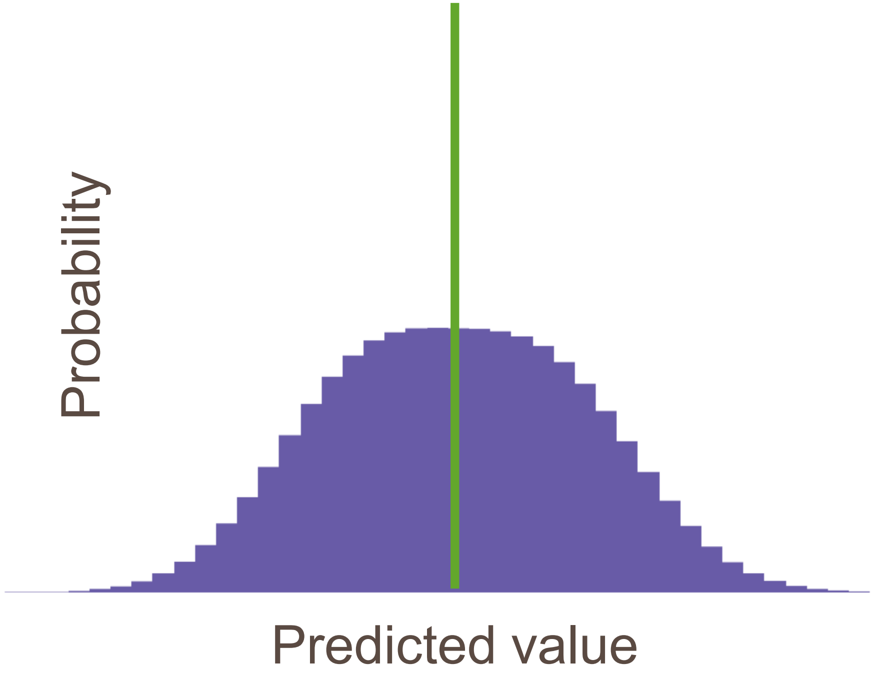


Whitehead *et al.* J. Chem Inf. Model. (2019) **59**(3) pp. 1197-1204, B. Irwin *et al.* J. Chem. Inf Model. (2020) **60**(6), pp. 2848–2857

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

- Estimates uncertainty in each individual prediction
  - 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, B. Irwin *et al.* J. Chem. Inf Model. (2020) **60**(6), pp. 2848–2857

Objectives and Data Set

Overview

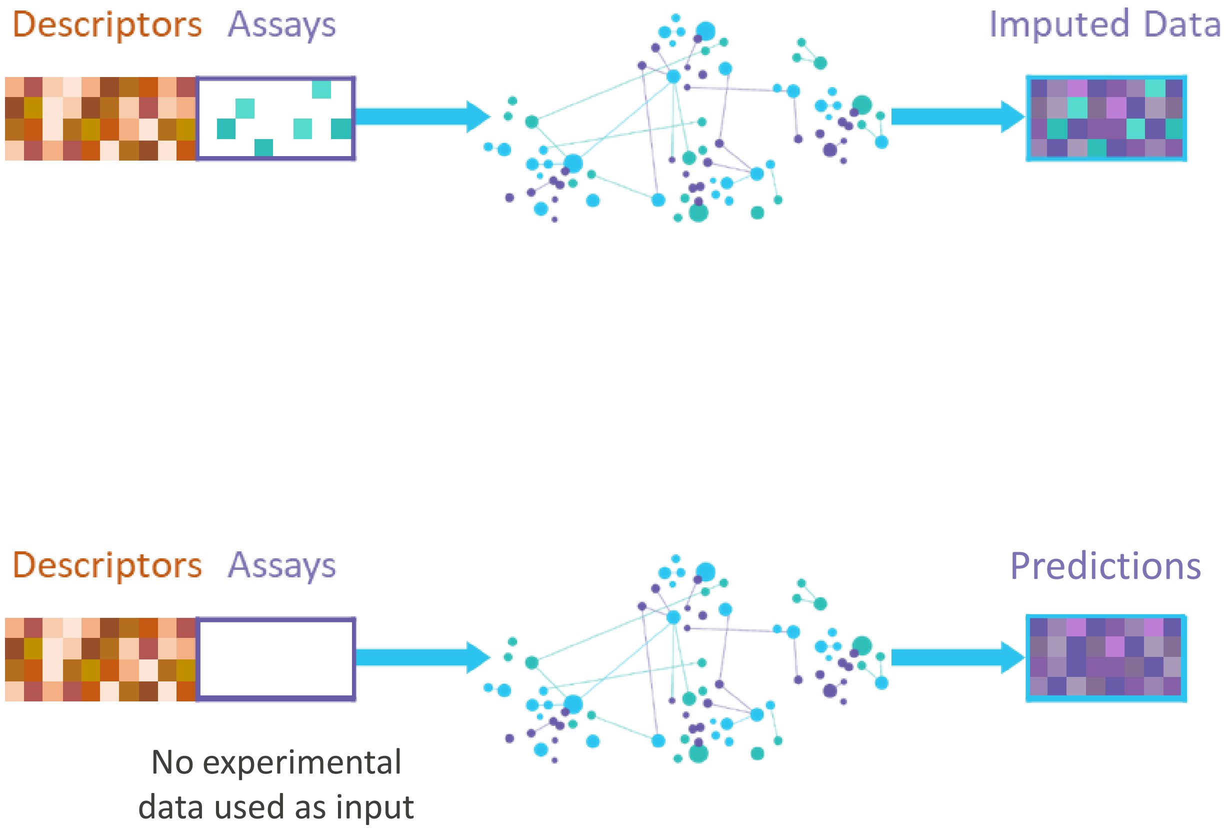
- Goal: More accurately predict TB activities and ADME properties to guide optimisation of compounds in a project context
  - Compare project-specific versus ‘global’ models
  - Compare imputation and virtual models

- Summary of Data
  - Global data set
    - 300,000 compounds x 468 experimental endpoints across several developing-world/neglected diseases
    - 3.1% complete
  - Project data set – a subset of global data set corresponding to a single TB project
    - 495 compounds x 34 experimental endpoints
    - 40.6% complete

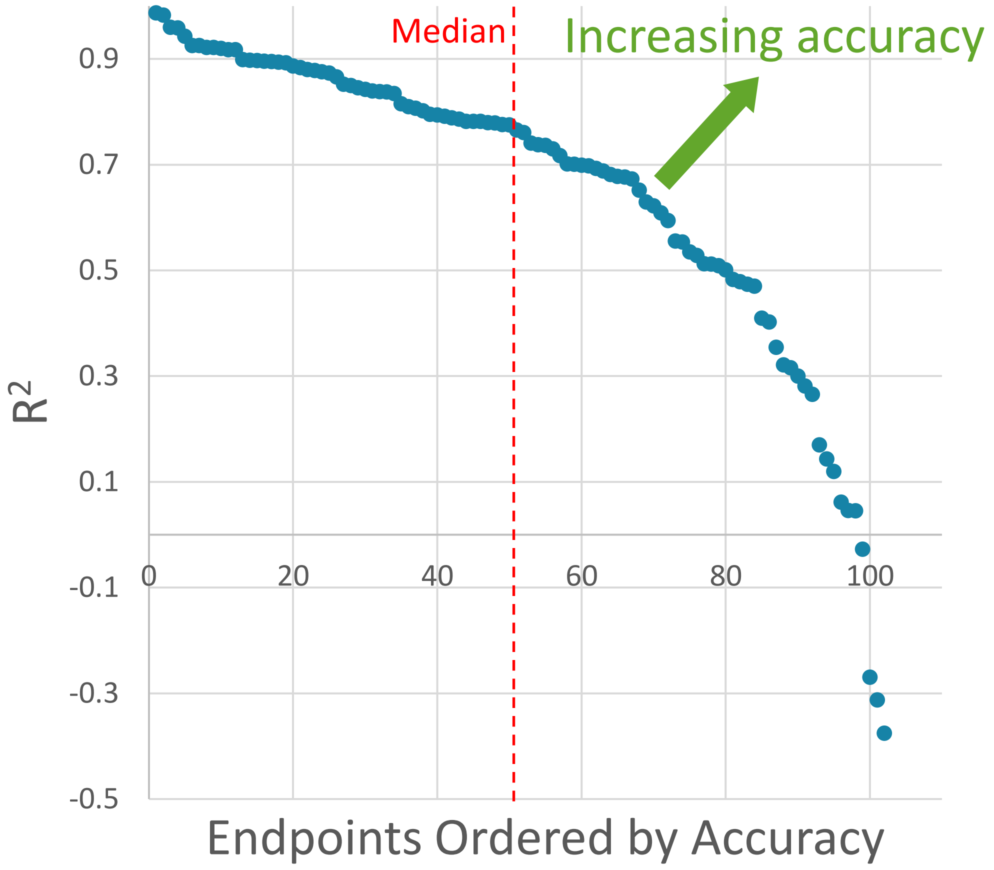
Imputation vs Virtual Models

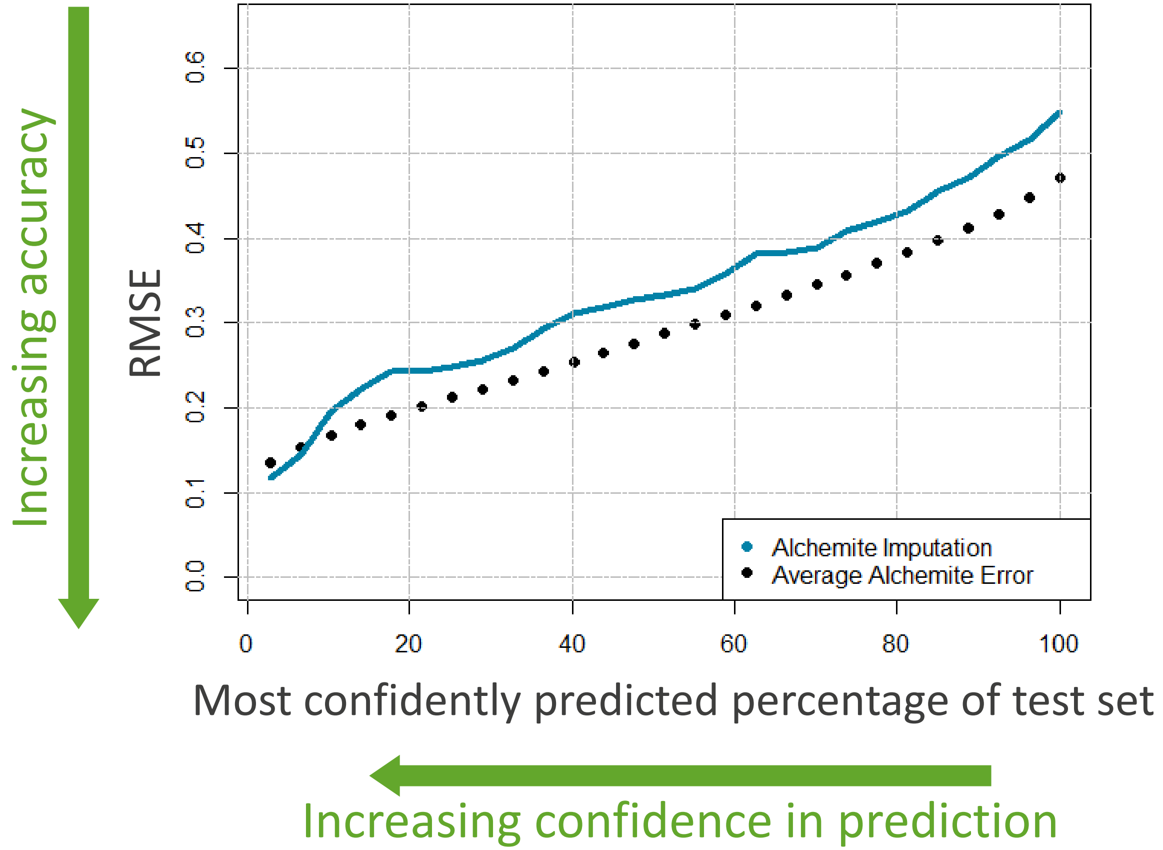
- Imputation: These models generate predictions for the test data points using sparse assay data as input, in addition to molecular descriptors
  - These models test an Alchemite model's ability to 'fill in the gaps' in the experimental data for compounds that have been synthesised and tested in some assays
- Virtual: These models are built to expect only molecular descriptors as input.
  - These test an Alchemite model's ability to make predictions based only on compound structure, i.e., for a compound that has not yet been synthesised or tested

Application to Test Set



Assessment of Results

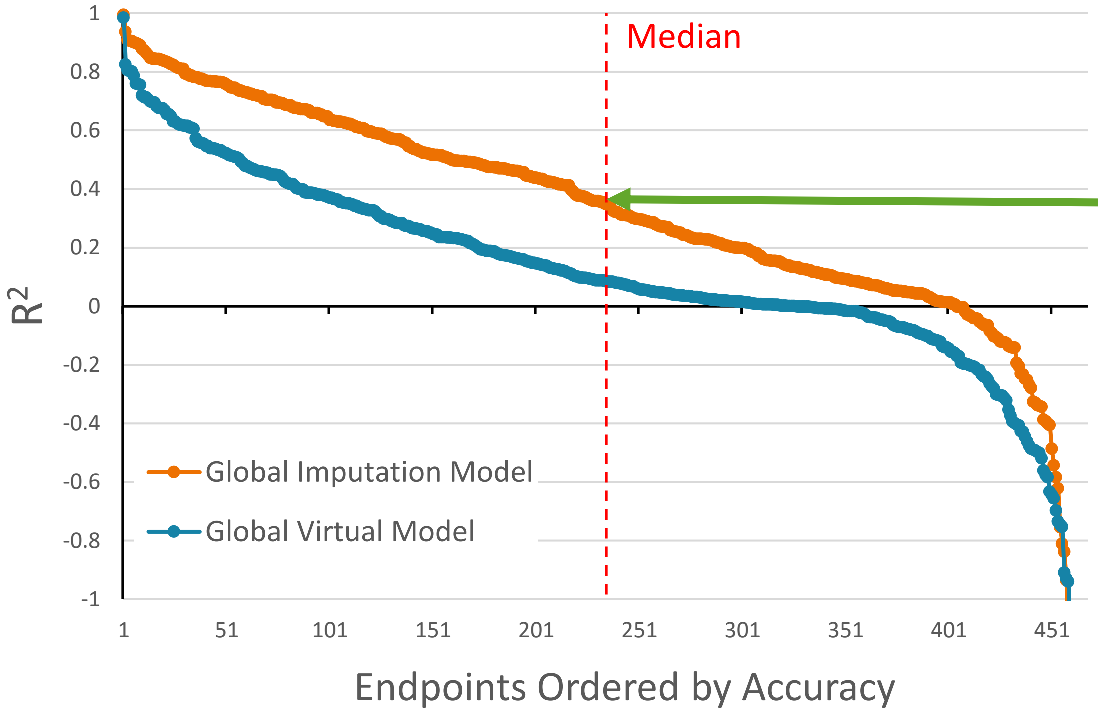




R^2 – Coefficient of Determination. RMSE – Root-Mean-Square Error

Model Validation

Global Models Test Set Results

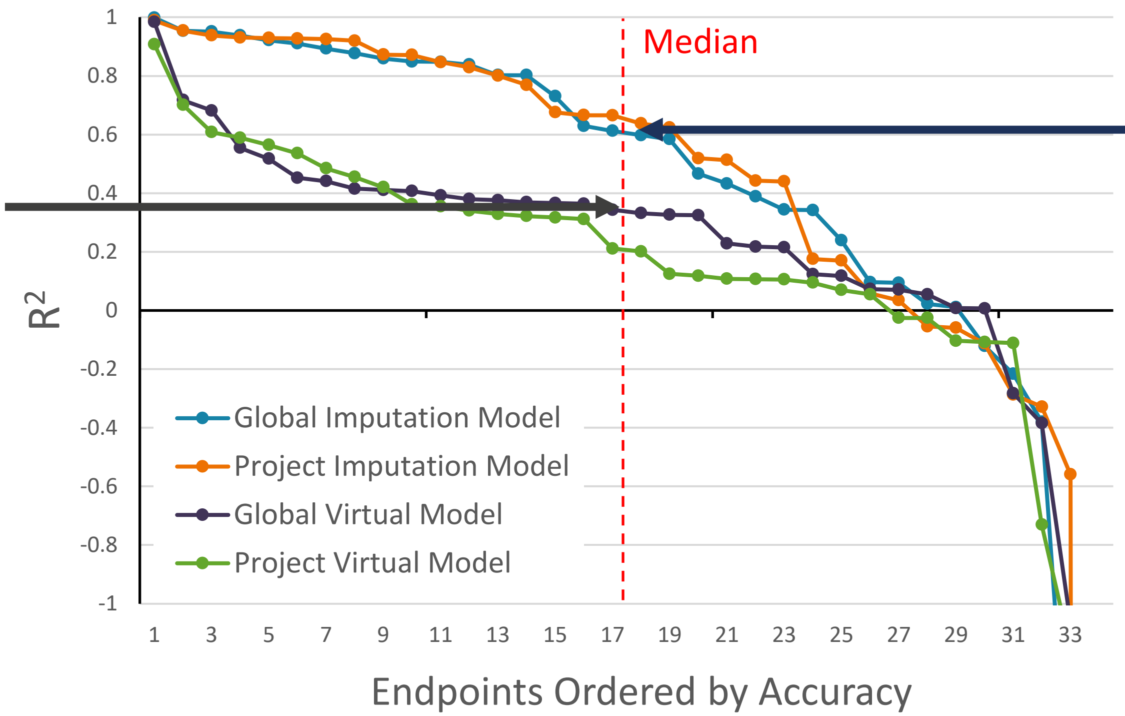


The Imputation model clearly outperforms the Virtual model

	Median R^2	Number with $R^2 > 0.5$	Number with $R^2 > 0.3$
Alchemite Imputation	0.35	159	248
Alchemite Virtual	0.10	44	137

Global and Project-specific Model Performance on Project Test Set

Global Virtual model outperforms project-specific Virtual model

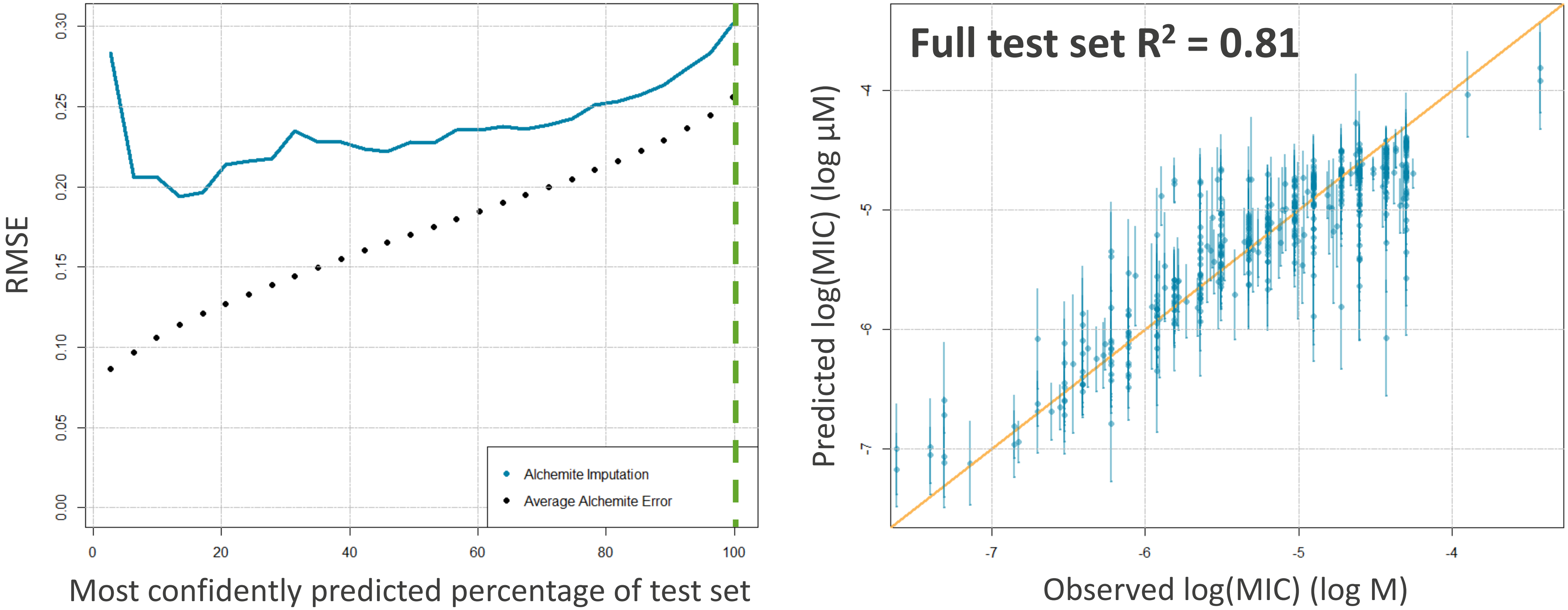


Global and project-specific Imputation models achieve almost identical performance

	Median R^2	Number with $R^2 > 0.5$	Number with $R^2 > 0.3$
Project Imputation	0.65	21	23
Project Virtual	0.21	6	16
Global Imputation	0.61	19	24
Global Virtual	0.33	5	20

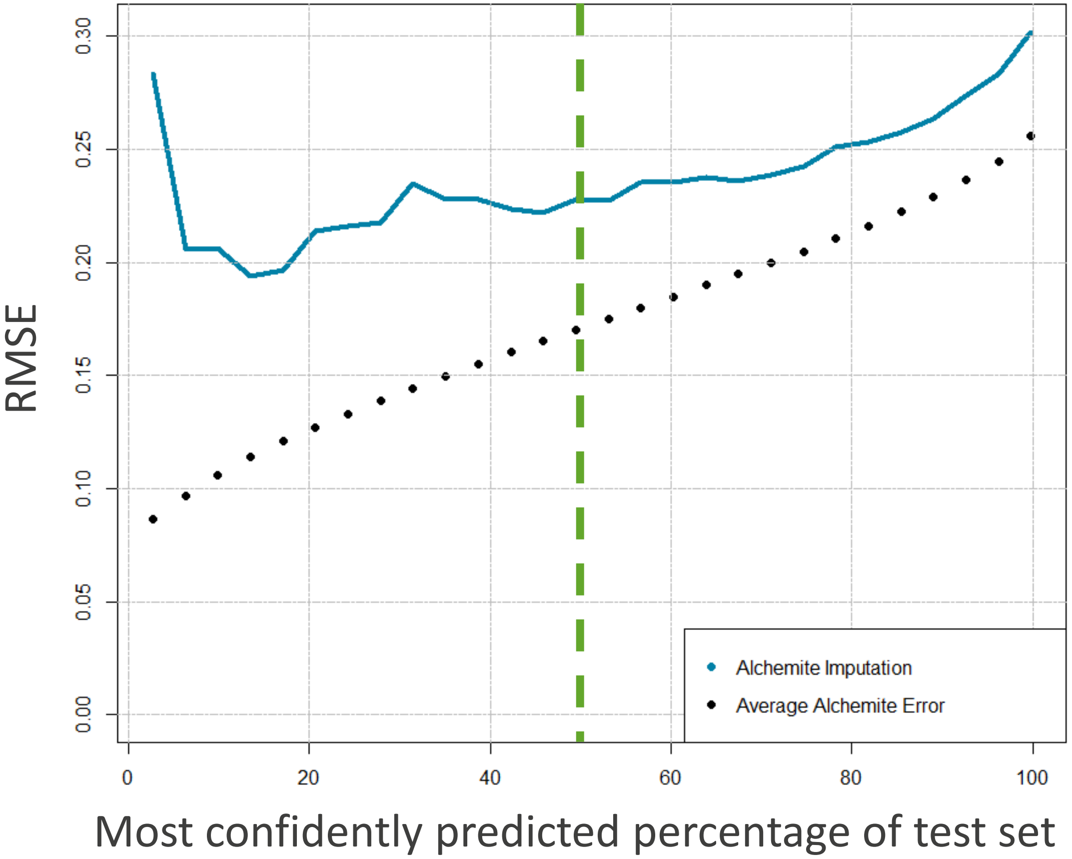
Focusing on the Most Confident Results

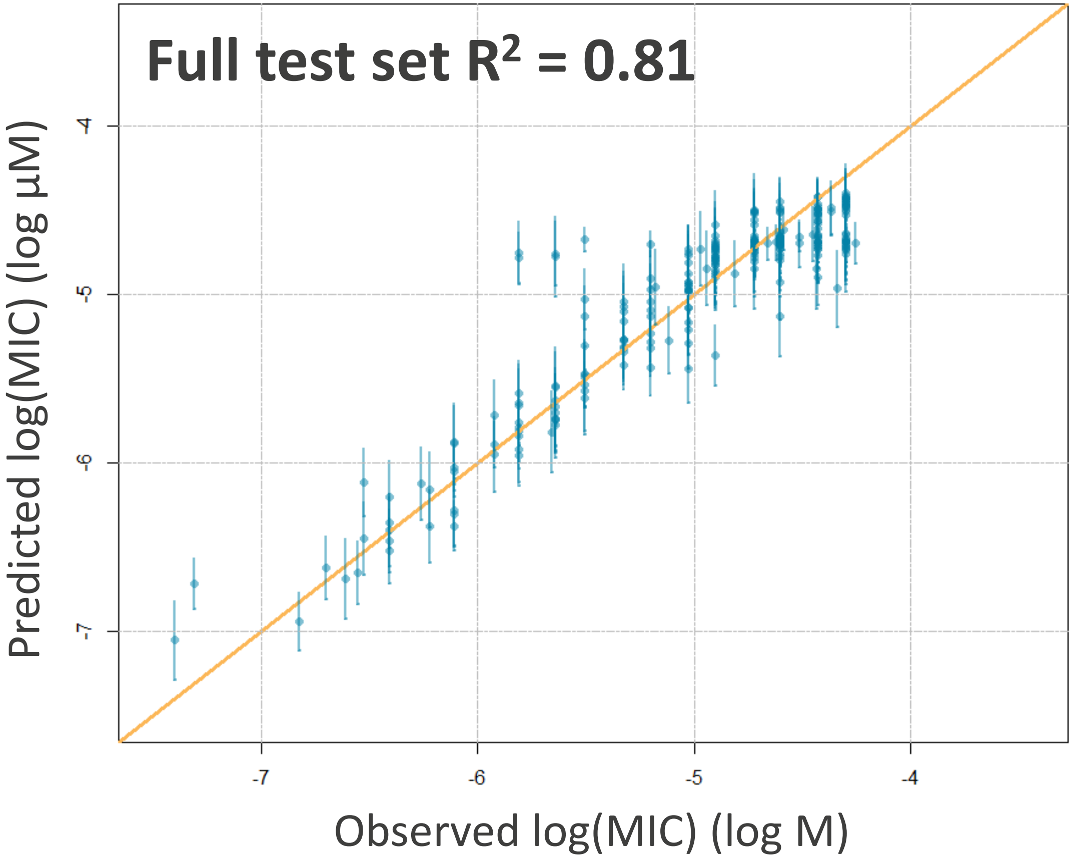
TB Activity Endpoint



Focusing on the Most Confident Results

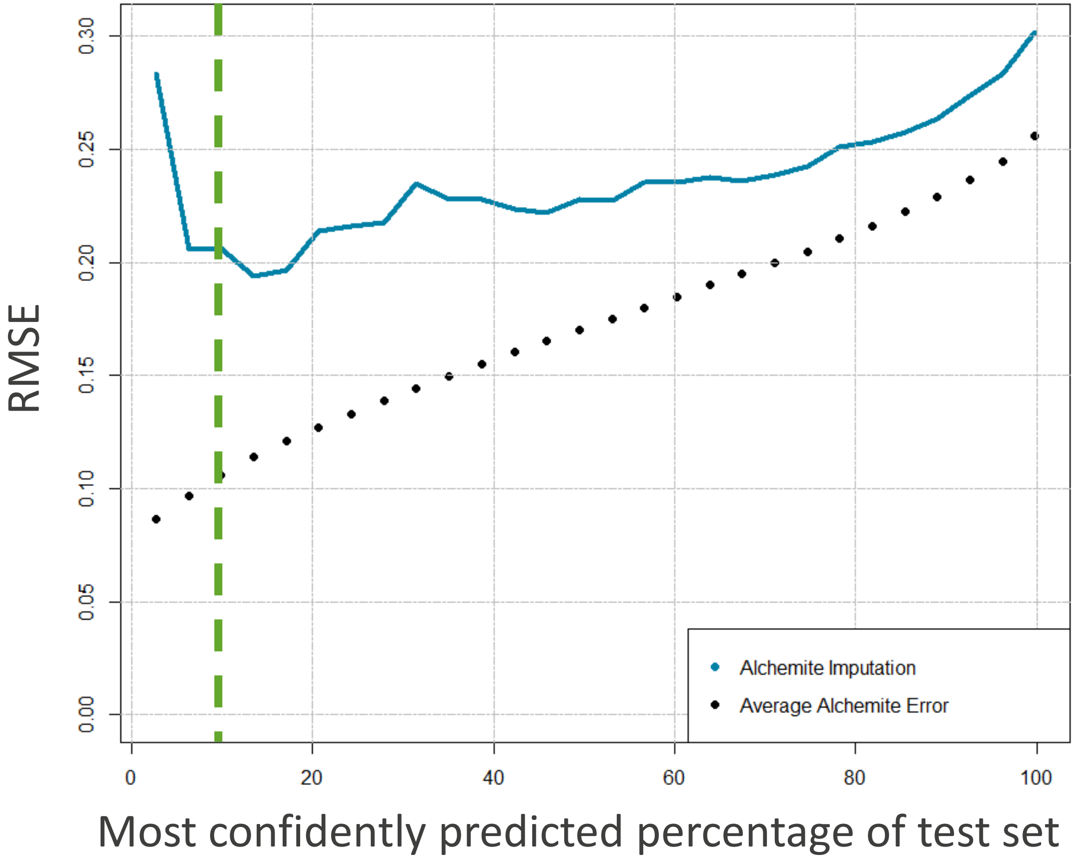
TB Activity Endpoint

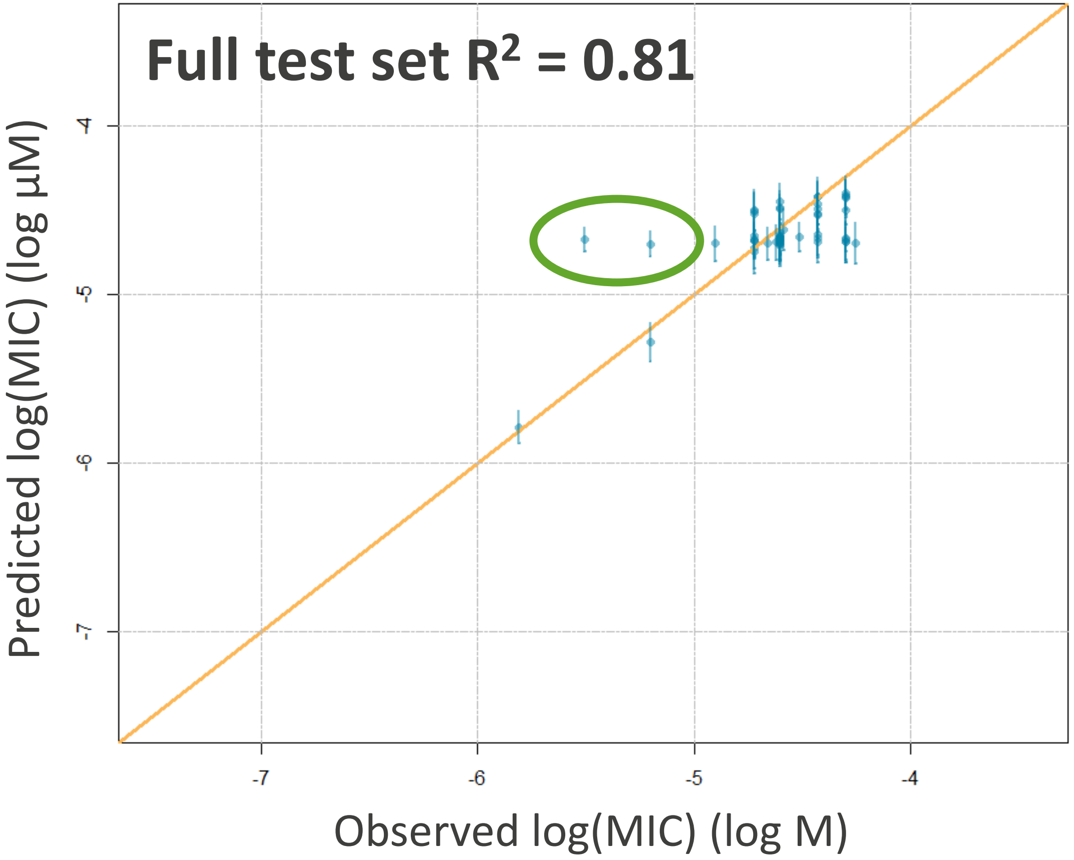




Focusing on the Most Confident Results

TB Activity Endpoint

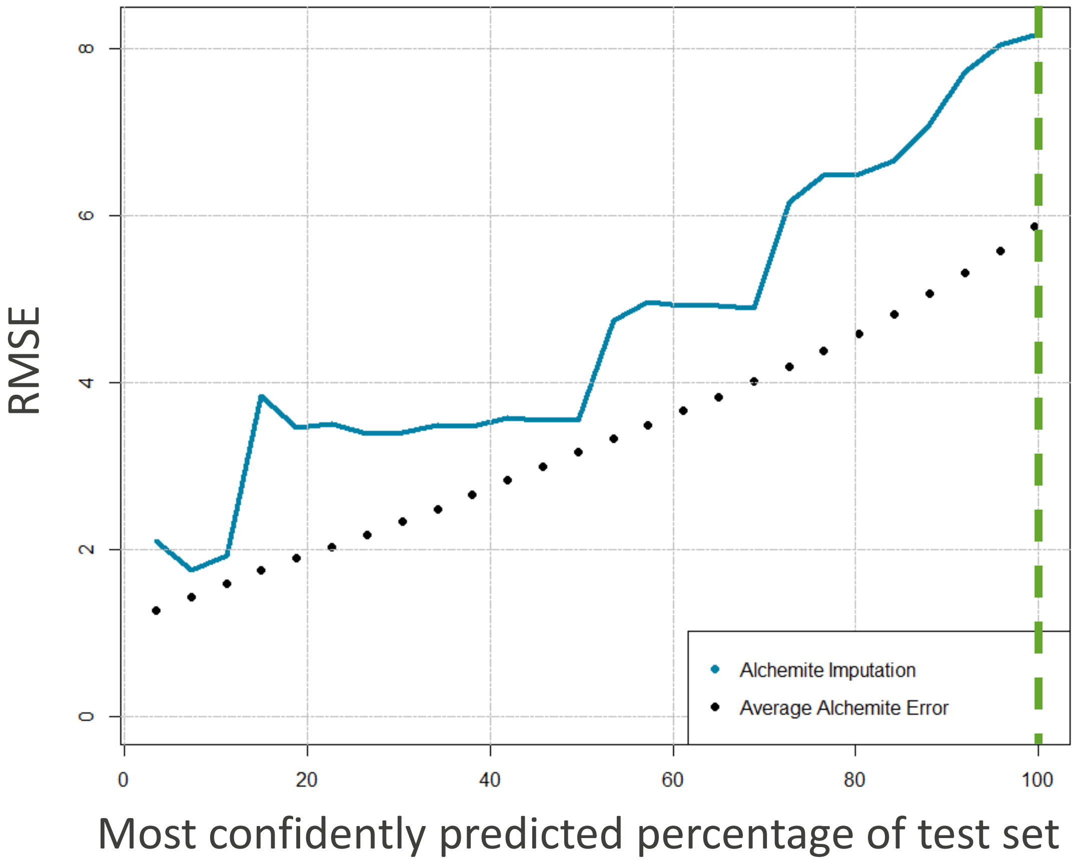


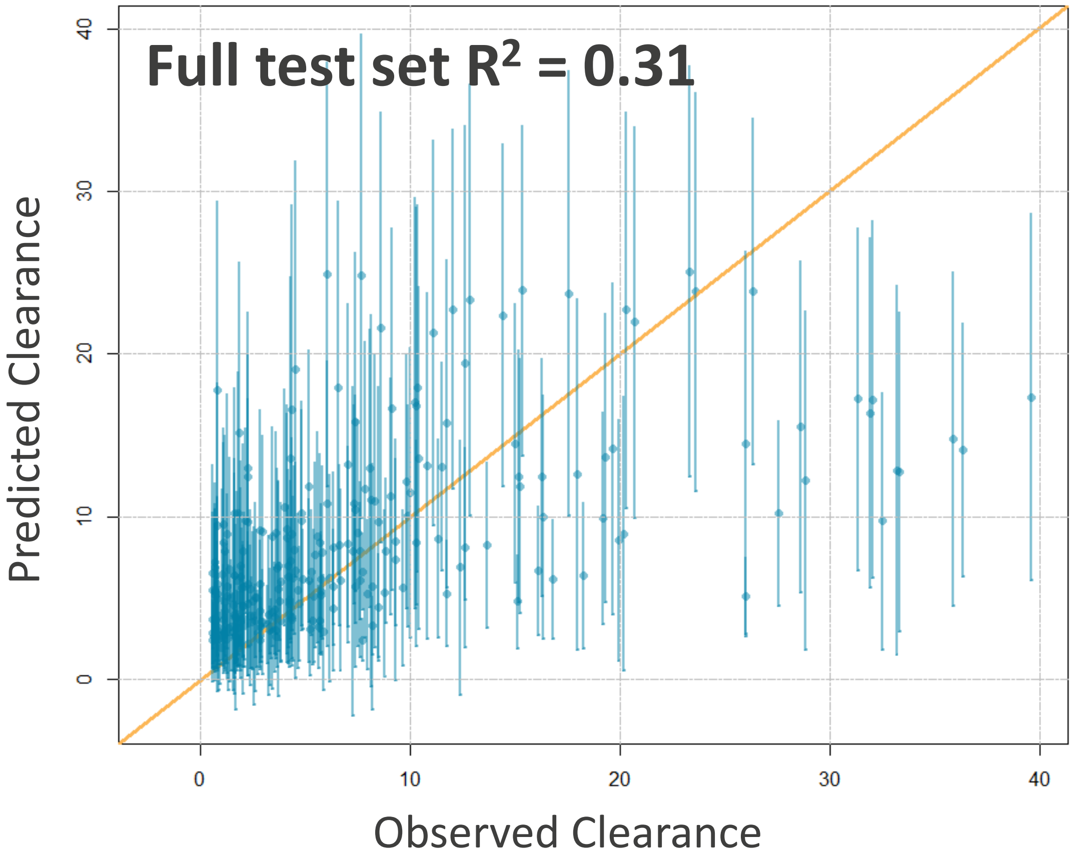


- Excellent correlation between model confidence (error bars) and observed accuracy
- Outliers clearly identified for further investigation

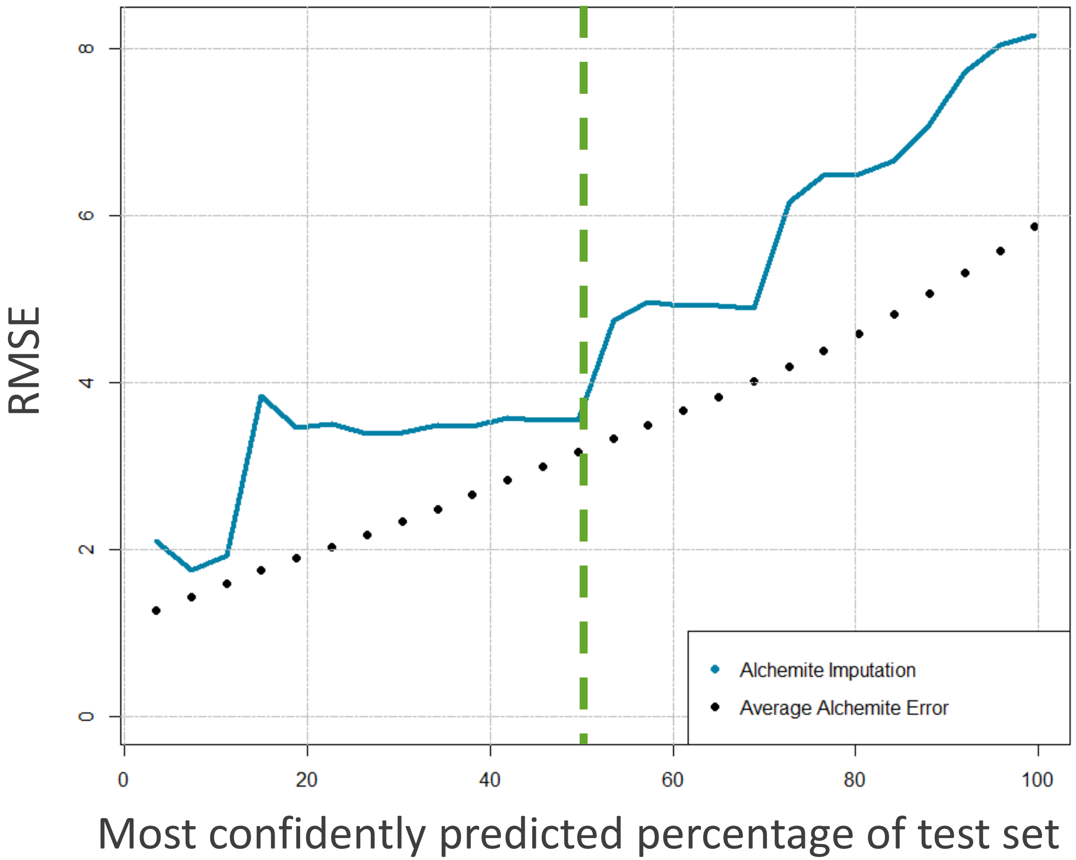
Focusing on the Most Confident Results

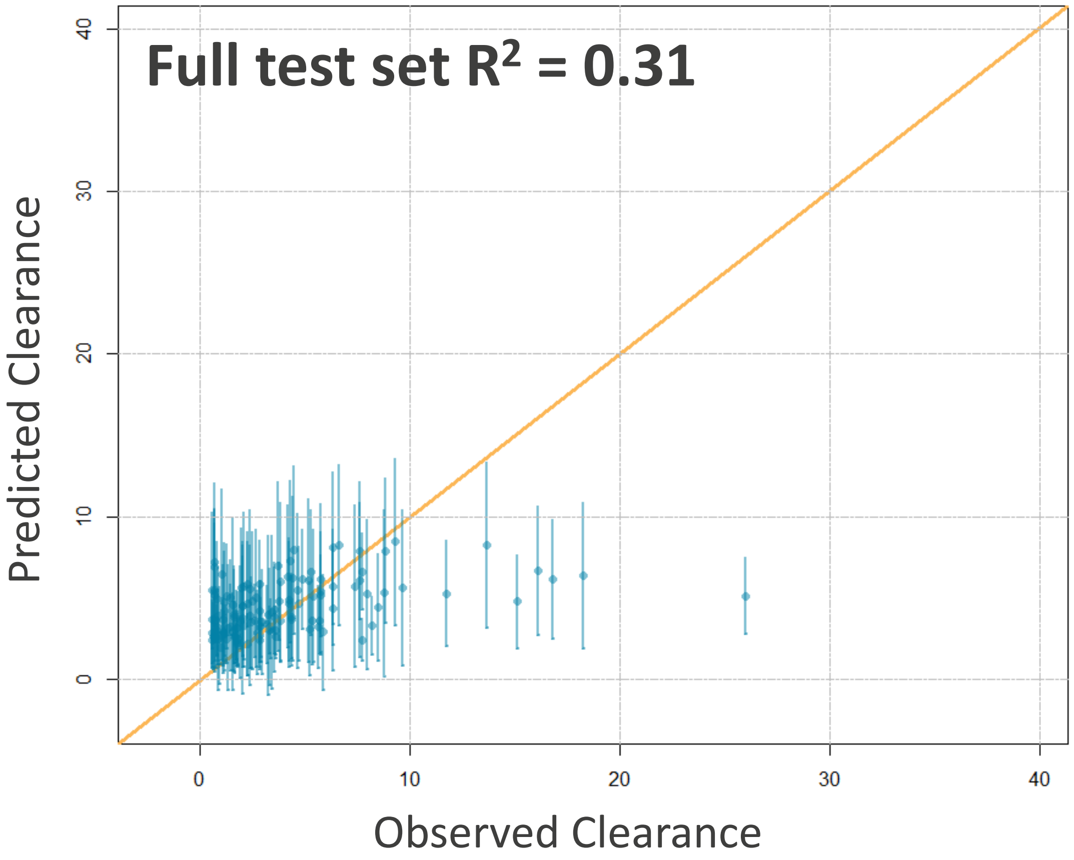
Hepatocyte Clearance



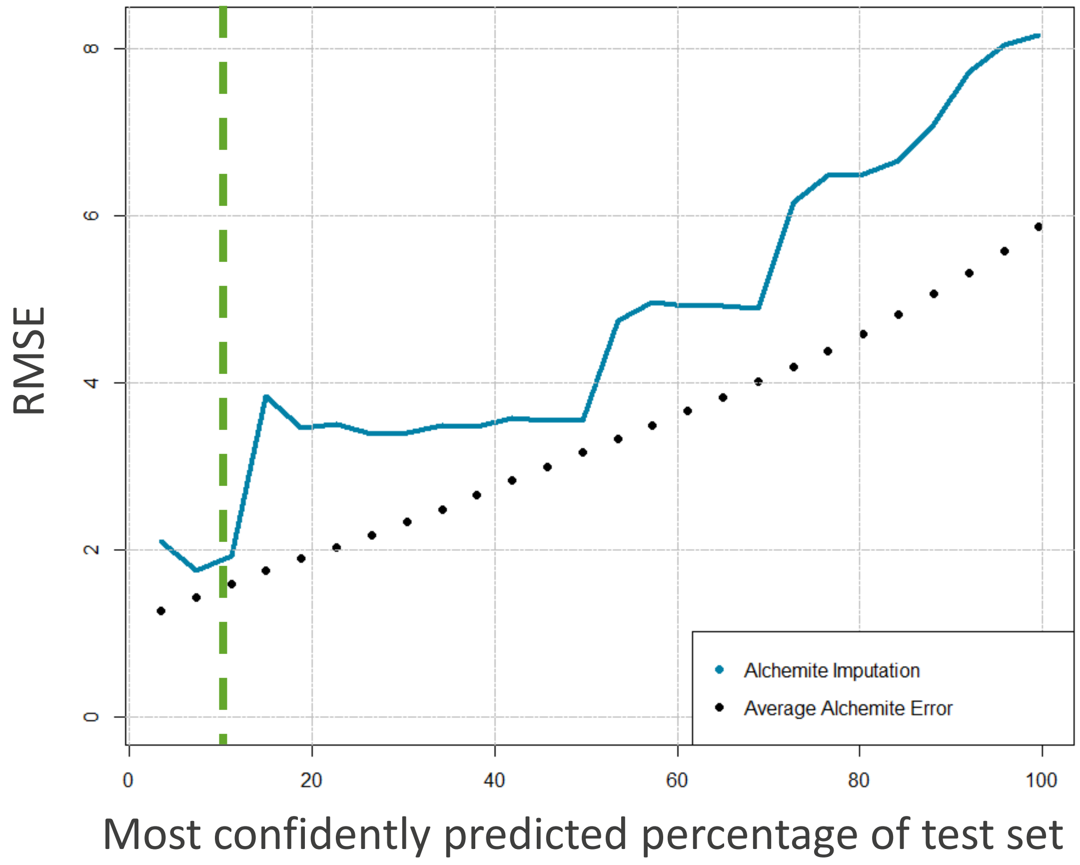


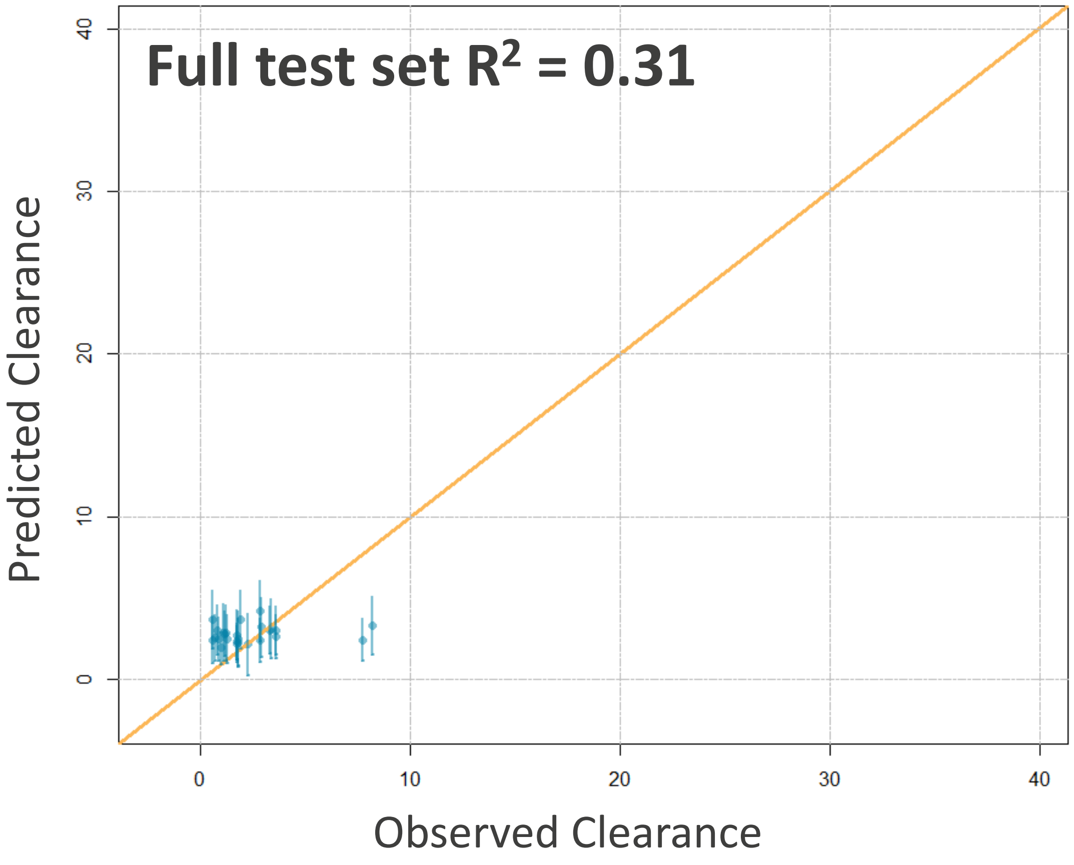
Focusing on the Most Confident Results
Hepatocyte Clearance





Focusing on the Most Confident Results
Hepatocyte Clearance





- Even for model with poor overall performance, we can identify accurate predictions that can be used with confidence

Application of the Global Deep Learning Model to TB Project Optimisation

TB Project Objectives

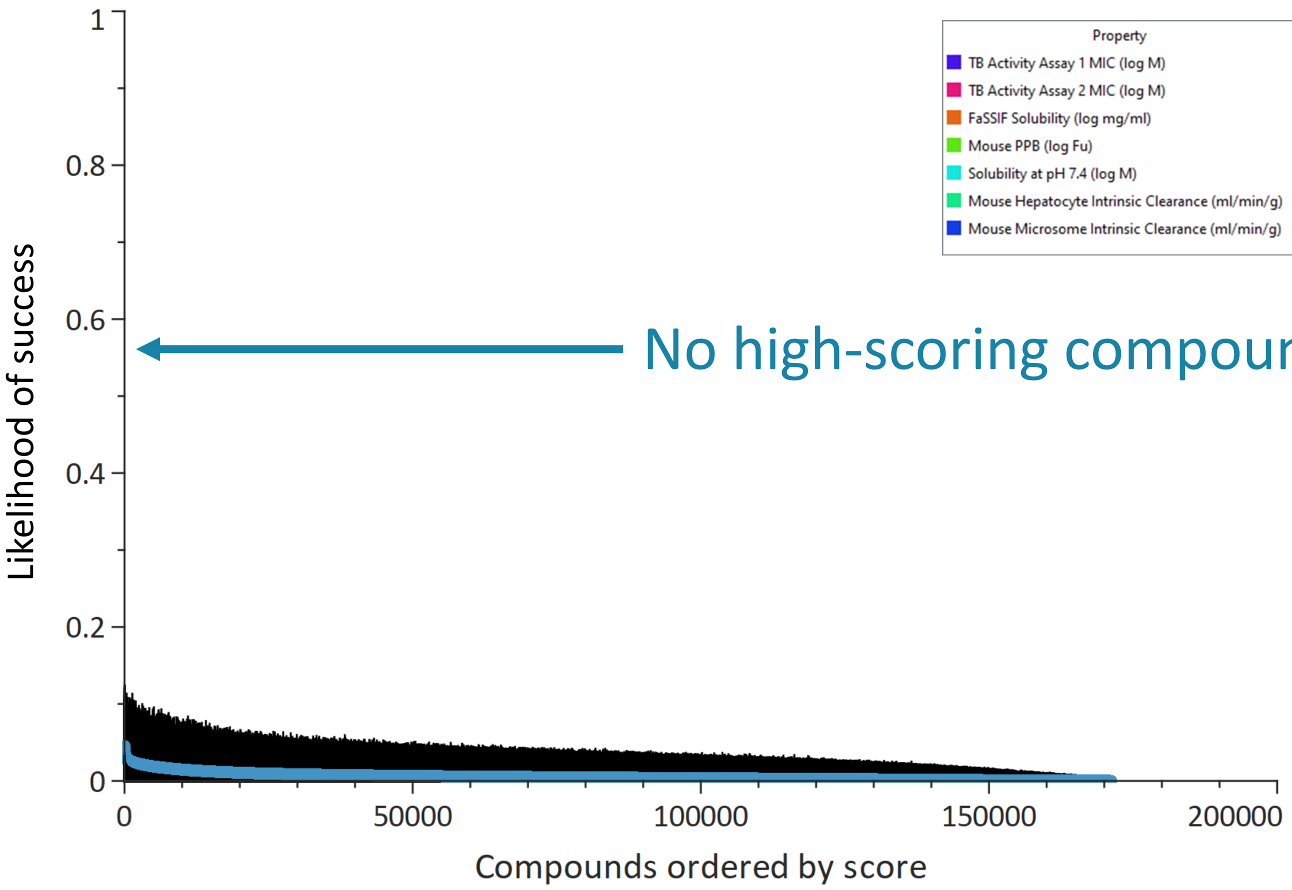
- Desired compound property criteria:

Property	Desired Value	Importance
TB Activity Assay 1 MIC (log M)	-inf -> -6.4	
TB Activity Assay 2 MIC (log M)	-inf -> -6.4	
FaSSIF Solubility (log mg/ml)	-1 -> inf	
Mouse PPB (log Fu)	-1.3 -> -0.3	
Solubility at pH 7.4 (log M)	-4 -> inf	
Mouse Hepatocyte Intrinsic Clearance (ml/min/g)	-inf -> 1.5	
Mouse Microsome Intrinsic Clearance (ml/min/g)	-inf -> 1.5	

- Challenges achieving a balance of activity with hepatocyte stability and solubility
- Strategy: Explore a large virtual library enumerated around the series core
- Apply the global Alchemite Virtual model to all compounds to determine if the desired *balance* of properties is likely to be accessible in this series

Multi-Parameter Scores for TB Project

TB Project Profile Score Distribution



Property	Desired Value	Importance
TB Activity Assay 1 MIC (log M)	-inf -> -6.4	
TB Activity Assay 2 MIC (log M)	-inf -> -6.4	
FaSSIF Solubility (log mg/ml)	-1 -> inf	
Mouse PPB (log Fu)	-1.3 -> -0.3	
Solubility at pH 7.4 (log M)	-4 -> inf	
Mouse Hepatocyte Intrinsic Clearance (ml/min/g)	-inf -> 1.5	
Mouse Microsome Intrinsic Clearance (ml/min/g)	-inf -> 1.5	

No high-scoring compounds

D. Segall, *Curr. Pharm. Des.* **2012**, *18* (9), 1292–1310

Multi-Parameter Profiles

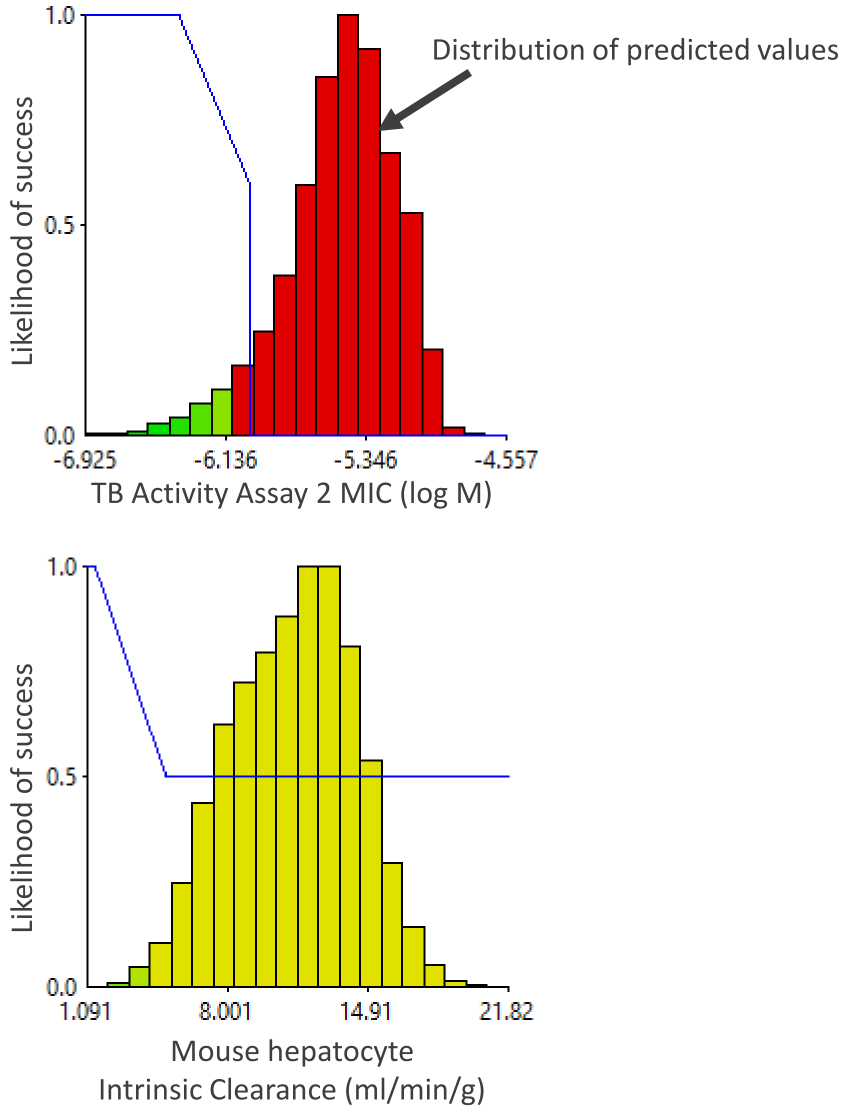
Balancing activity and hepatocyte stability

	TB Project Profile	ID	TB Activity Assay 1 M	TB Activity Assay 2 MI	FaSSIF Solubility (log r	Mouse PPB (log Fu)	Solubility at pH 7.4 (l	Mouse Hepatocyte Int	Mouse Microsome Int
1	0.045	300K-ARRAY-CMPD-289935	-6.307	-6.241	-1.358	-0.514	-5.052	9.635	0.5078
2	0.04161	300K-ARRAY-CMPD-245199	-6.437	-6.42	-1.519	-0.449	-5.216	7.956	0.5727
3	0.04158	300K-ARRAY-CMPD-285557	-6.623	-6.488	-1.706	-0.4166	-5.547	5.029	0.3847
4	0.04147	300K-ARRAY-CMPD-244311	-6.802	-6.682	-1.685	-0.3381	-5.732	7.308	0.4943
5	0.04085	300K-ARRAY-CMPD-144354	-6.408	-6.051	-1.212	-0.646	-5.17	7.259	0.4914
6	0.04085	300K-ARRAY-CMPD-299356	-6.2	-6.135	-1.275	-0.4994	-4.939	10.74	
7				-6.811	-1.632	-0.4453	-6.087	8.681	
8	0.04054	300K-ARRAY-CMPD-264575	-6.313	-6.19	-1.458	-0.3545	-4.82	7.095	0.4775
9	0.04031	300K-ARRAY-CMPD-247585	-6.544	-6.41	-1.65	-0.5082	-5.552	5.911	0.3992
10	0.0402	300K-ARRAY-CMPD-299704	-6.592	-6.672	-1.683	-0.5709	-5.834	5.94	0.5107
11	0.04009	300K-ARRAY-CMPD-244865	-6.849	-6.925	-1.587	-0.6351	-6.103	10.4	0.6447
12	0.03995	300K-ARRAY-CMPD-299690	-6.399	-6.546	-1.574	-0.6154	-5.635	7.061	0.441
13	0.03955	300K-ARRAY-CMPD-246489	-6.361	-6.181	-1.442	-0.7116	-4.959	7.044	0.5143

TB Activity Assay 2

Mouse Hepatocyte Intrinsic Clearance

Mouse Hepatocyte Stability vs TB Activity Assay 2



Multi-Parameter Profiles

Balancing activity and solubility

UNIVERSITY OF DUNDEE

	TB Project Profile	ID	TB Activity Assay 1 M	TB Activity Assay 2 MI	FaSSIF Solubility (log r	Mouse PPB (log Fu)	Solubility at pH 7.4 (lo	Mouse Hepatocyte Int	Mouse Microsome Int
1	0.045	300K-ARRAY-CMPD-289935	-6.307	-6.241	-1.358	-0.514	-5.052	9.635	0.5078
2	0.04161	300K-ARRAY-CMPD-245199	-6.437	-6.42	-1.519	-0.449	-5.216	7.956	0.5727
3	0.04158	300K-ARRAY-CMPD-285557	-6.623	-6.488	-1.706	-0.4166	-5.547	5.029	0.3847
4	0.04147	300K-ARRAY-CMPD-244311	-6.802	-6.682	-1.685	-0.3381	-5.732	7.308	0.4943
5	0.04085	300K-ARRAY-CMPD-144354	-6.408	-6.051	-1.212	-0.646	-5.17	7.259	0.4914
6		300K-ARRAY-		-6.135	-1.275			.74	0.4748
7	0.04056	300K-ARRAY-CMPD-244258	-6.845	-6.811	-1.632	-0.4453	-6.087	8.681	0.581
8	0.04054	300K-ARRAY-CMPD-264575	-6.313	-6.19	-1.458	-0.3545	-4.82	7.095	0.4775
9	0.04031	300K-ARRAY-CMPD-247585	-6.544	-6.41	-1.65	-0.5082	-5.552	5.911	0.3992
10	0.0402	300K-ARRAY-CMPD-299704	-6.592	-6.672	-1.683	-0.5709	-5.834	5.94	0.5107
11	0.04009	300K-ARRAY-CMPD-244865	-6.849	-6.925	-1.587	-0.6351	-6.103	10.4	0.6447
12	0.03995	300K-ARRAY-CMPD-299690	-6.399	-6.546	-1.574	-0.6154	-5.635	7.061	0.441
13	0.03955	300K-ARRAY-CMPD-246489	-6.361	-6.181	-1.442	-0.7116	-4.959	7.044	0.5143

TB Activity Assay 2

FASSIF Solubility

optibrium™

Project Application Conclusions

- Compounds are predicted to achieve good activity **or** hepatocyte stability **or** good solubility
- However, it is **unlikely that compounds in this series will be able to achieve all three criteria simultaneously**
- The application of a high-quality multi-parameter model enables a very rigorous exploration of chemical space around the series of interest
- Synthesis of a small number of selected compounds will enable the validation of this predicted hypothesis – **saving time and resources**

Summary

- Alchemite was used to build Imputation and Virtual models using a sparse data of 300,000 compounds across approximately 500 experimental endpoints
  - No loss of accuracy over project-specific models, even for unrelated endpoints and project chemistries
  - Consistent with findings in collaboration with Constellation Pharmaceuticals on a smaller-scale data set (J. Chem. Inf Model. (2020) 60(6), pp. 2848–2857)
  - The global Virtual model was more accurate due to additional chemical diversity in training set
  - **Build once, run everywhere…**
    - Save time – No need to build multiple, individual project models
    - Maximise information – Learn across multiple projects, chemistries and therapeutic areas simultaneously
- Strong agreement confirmed between model confidence and observed accuracy
  - **Focus on the most valuable results for decision-making**, even for models with poor headline accuracy
- Example application to a TB project
  - Combined with multi-parameter optimisation
  - Unwelcome result for the project, but saves expending time and effort with a low probability of success

Acknowledgements

- UK-QSAR and Cheminformatics Group

- University of Dundee
  - Paul Wyatt
  - Fabio Zuccotto
  - Laura Cleghorn
  - Simon Green
  - James Burkinshaw
  - And colleagues…

- Intellegens
  - Gareth Conduit
  - Tom Whitehead

- Optibrium
  - Matt Segall
  - Ben Irwin