Imputation of assay bioactivity data using deep learning

Intellegens

optibrium

Tom Whitehead
Ben Irwin, Peter Hunt, Matt Segall, Gareth Conduit

Unique deep learning algorithm

Utilise chemical descriptors, assay bioactivities, and simulations **in combination**

Understand and exploit **uncertainties** and noise to improve confidence in predictions

Broadly applicable algorithm with **proven** applications in drug design, materials discovery, patient analytics, …

Deep learning

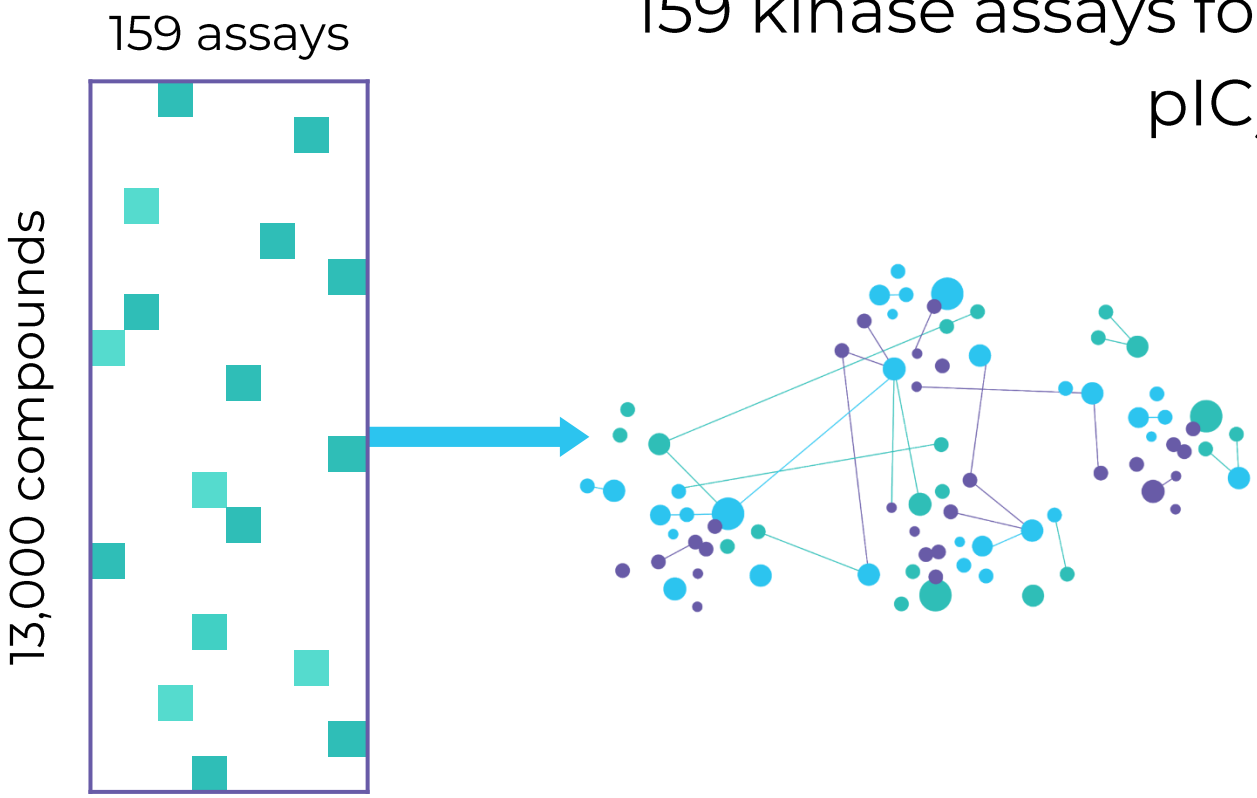
Alchemite™ deep learning

Whitehead *et al.*, J. Chem. Inf. Model. 59, 1197 (2019)

Novartis dataset to benchmark machine learning

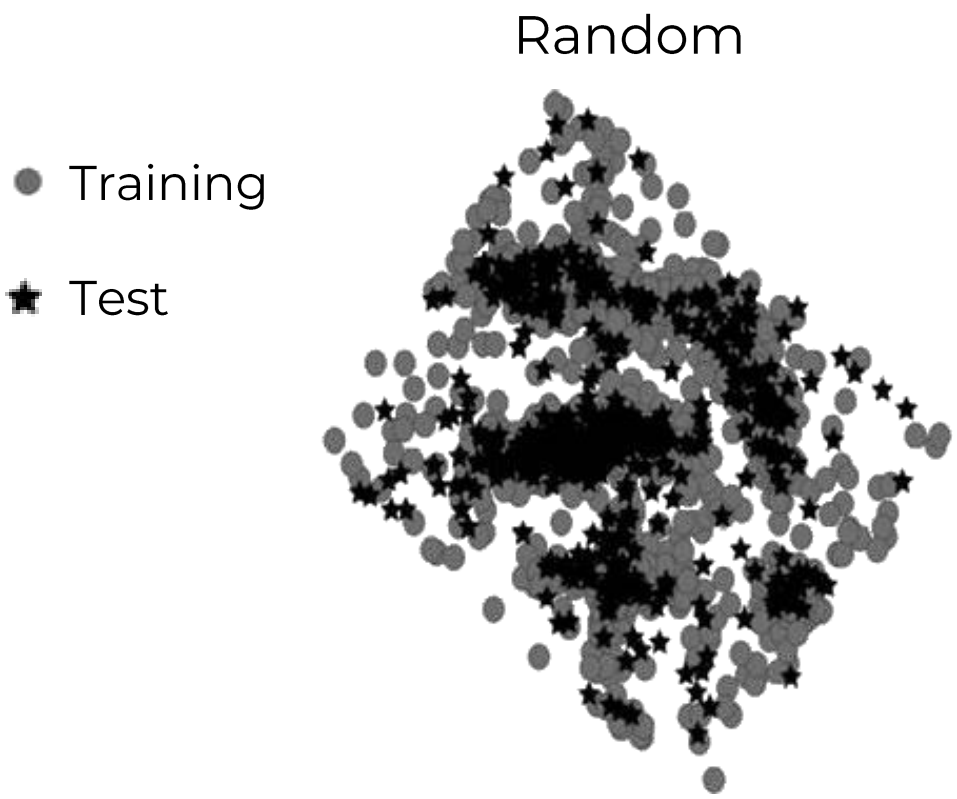
159 kinase assays for 13,000 compounds

pIC_{50} data **5%** complete



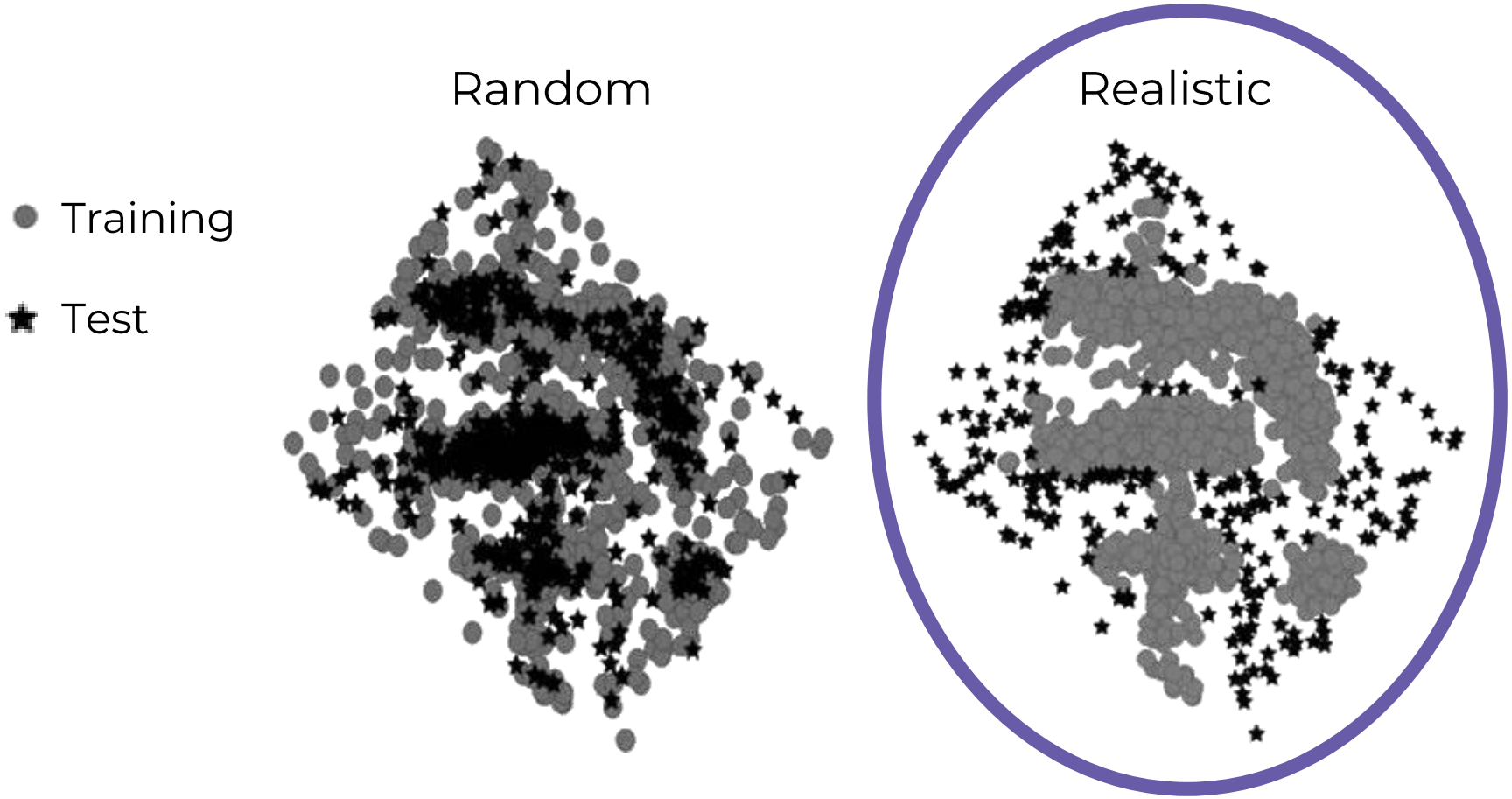
Data from ChEMBL
Martin, Polyakov, Tian, and Perez, J. Chem. Inf. Model. 57, 2077 (2017)

Novartis dataset distribution



Data from ChEMBL
Martin, Polyakov, Tian, and Perez, J. Chem. Inf. Model. 57, 2077 (2017)

Novartis dataset is realistically distributed



Data from ChEMBL
Martin, Polyakov, Tian, and Perez, J. Chem. Inf. Model. 57, 2077 (2017)

Accuracy metrics

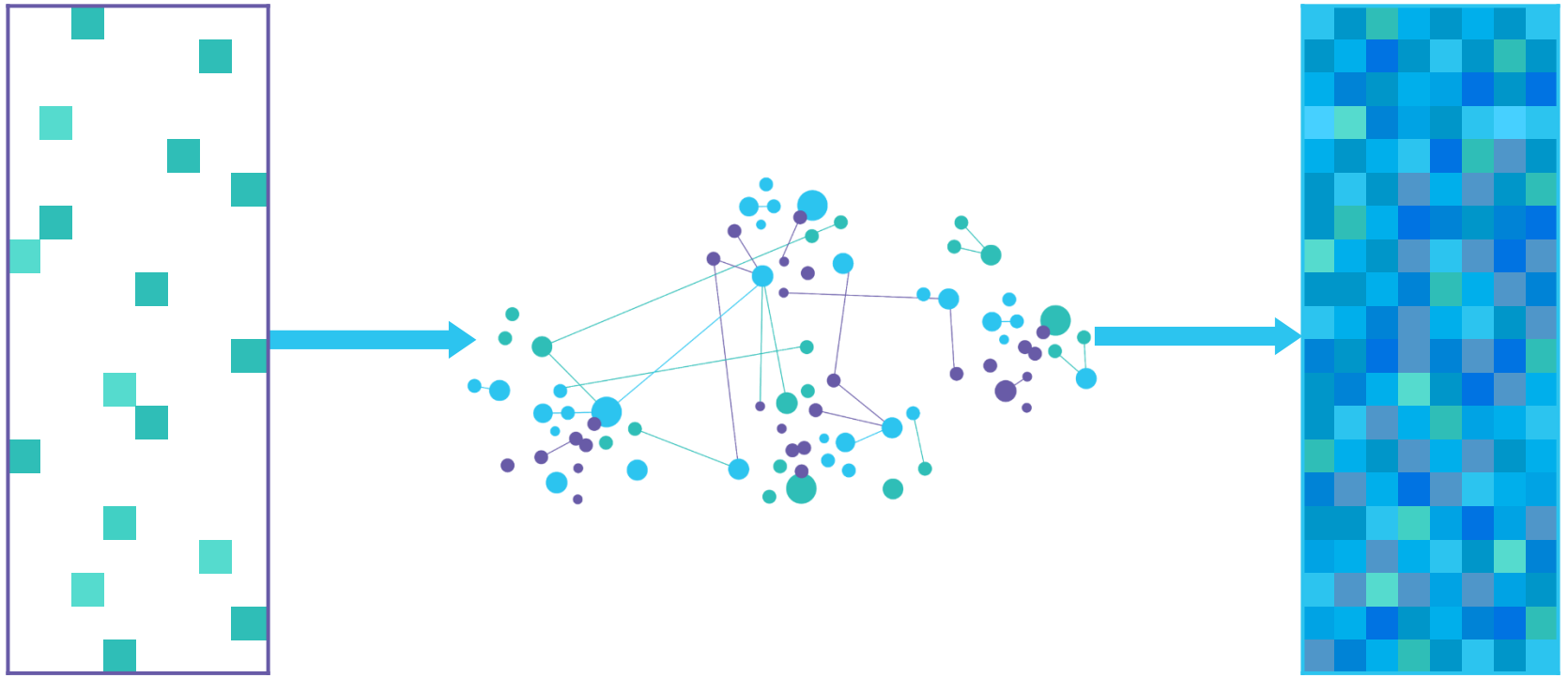
Coefficient of Determination, R^2

Root Mean Square Error, RMSE

Measure R^2 and RMSE per assay against realistic test set, then report mean across assays

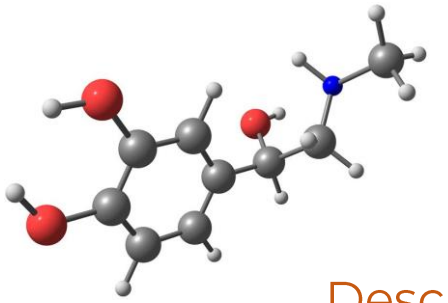
Aim: impute missing assay values

Validate against
realistically-split holdout set

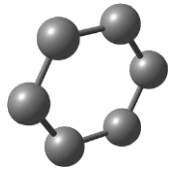


Data from ChEMBL
Martin, Polyakov, Tian, and Perez, J. Chem. Inf. Model. 57, 2077 (2017)

Random forest regression



x3



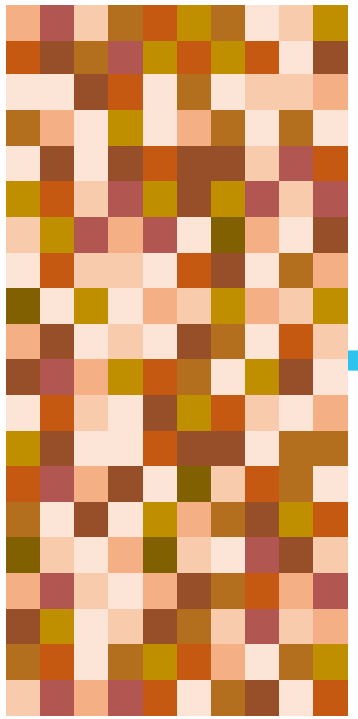
x1

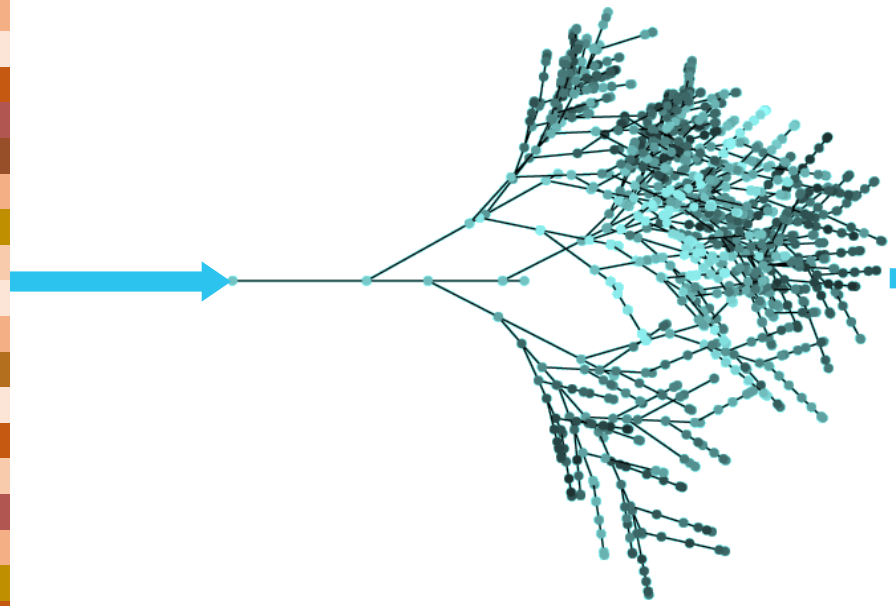
Molecular weight = 183 Da

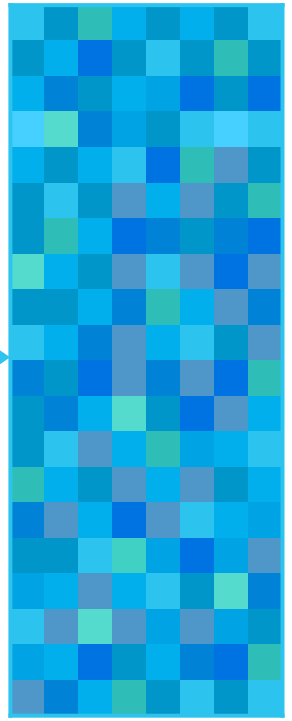
R^2 = -0.19

RMSE = 0.89

Descriptors

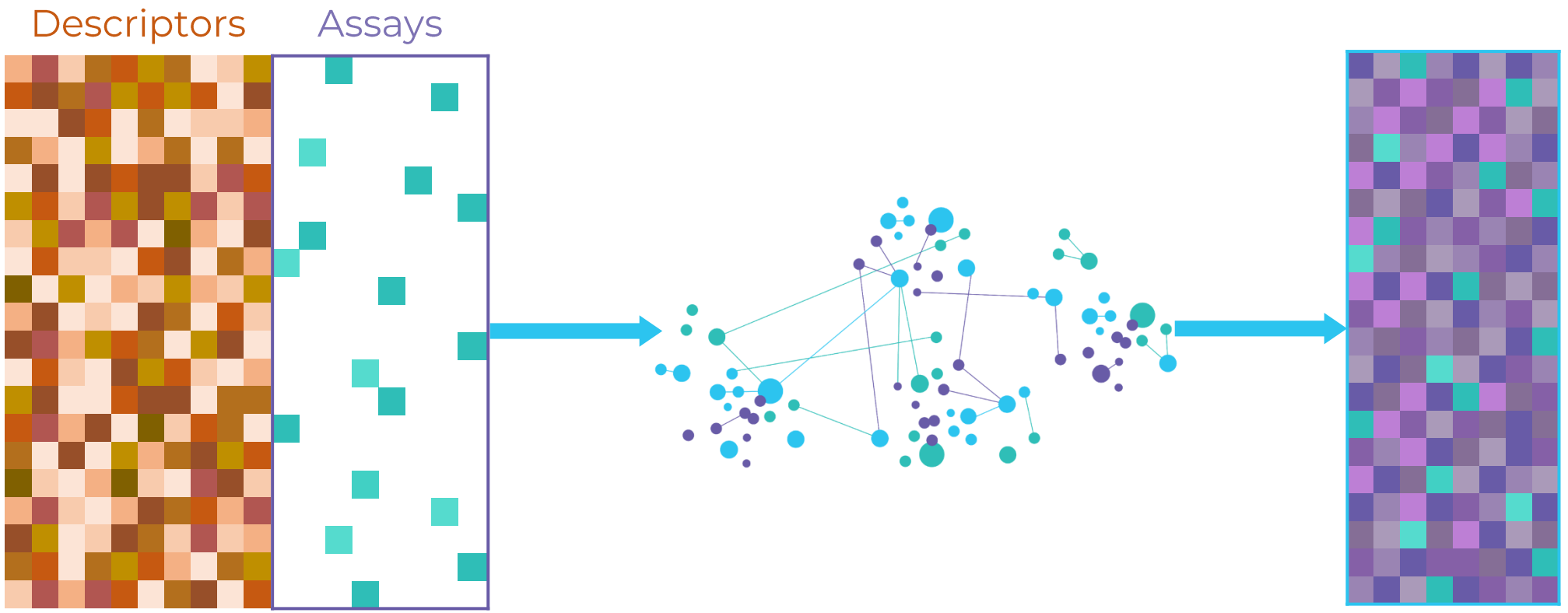






intellegens.ai

Descriptors and bioactivity values



Deep learning predictions

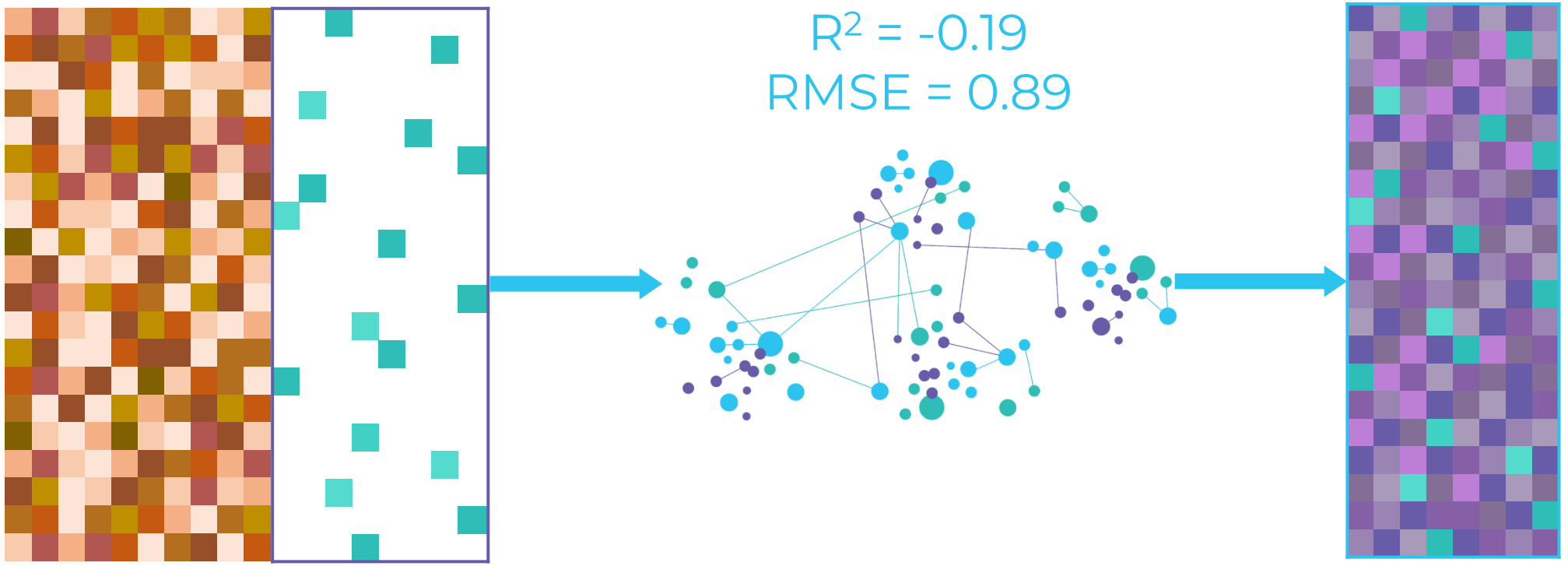
$R^2 = 0.46$

RMSE = 0.59

Random forest

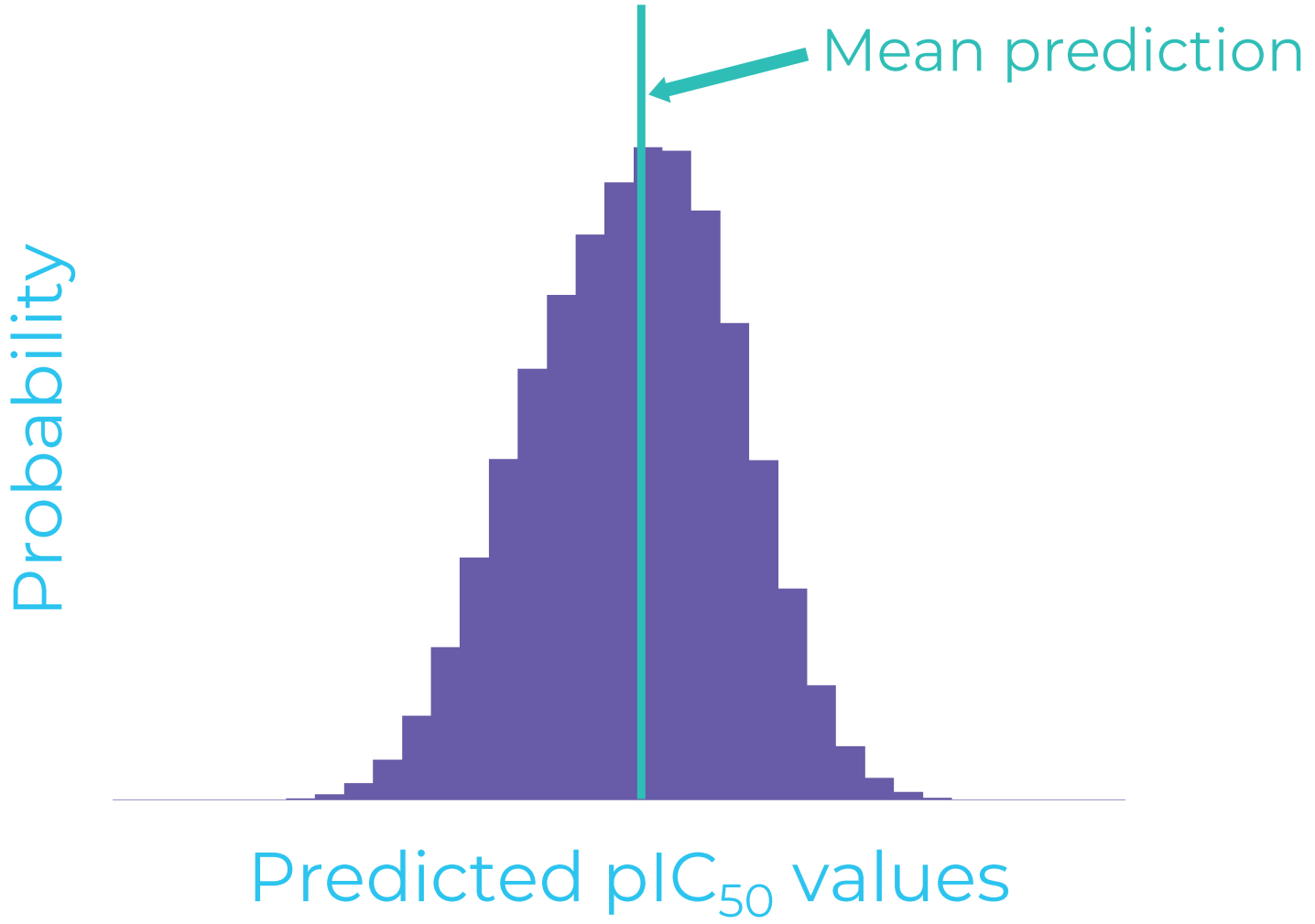
$R^2 = -0.19$

RMSE = 0.89

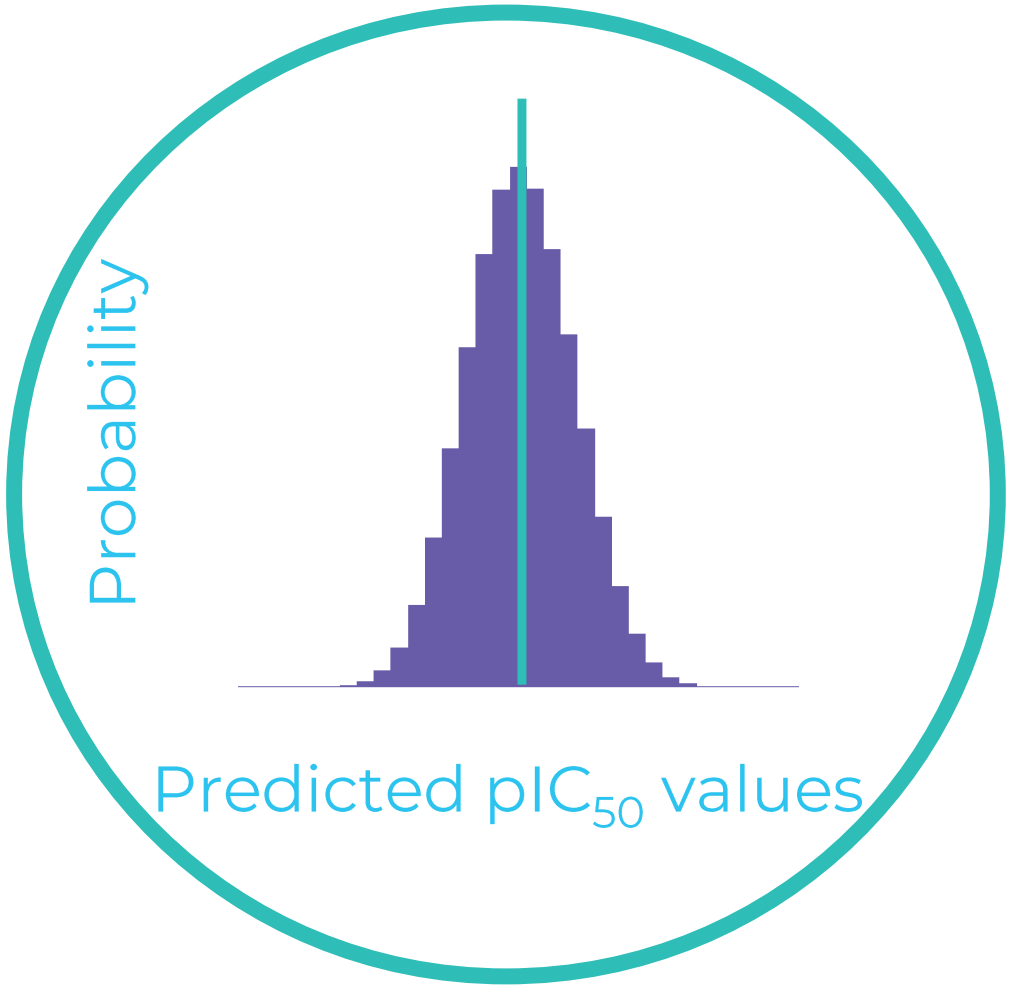


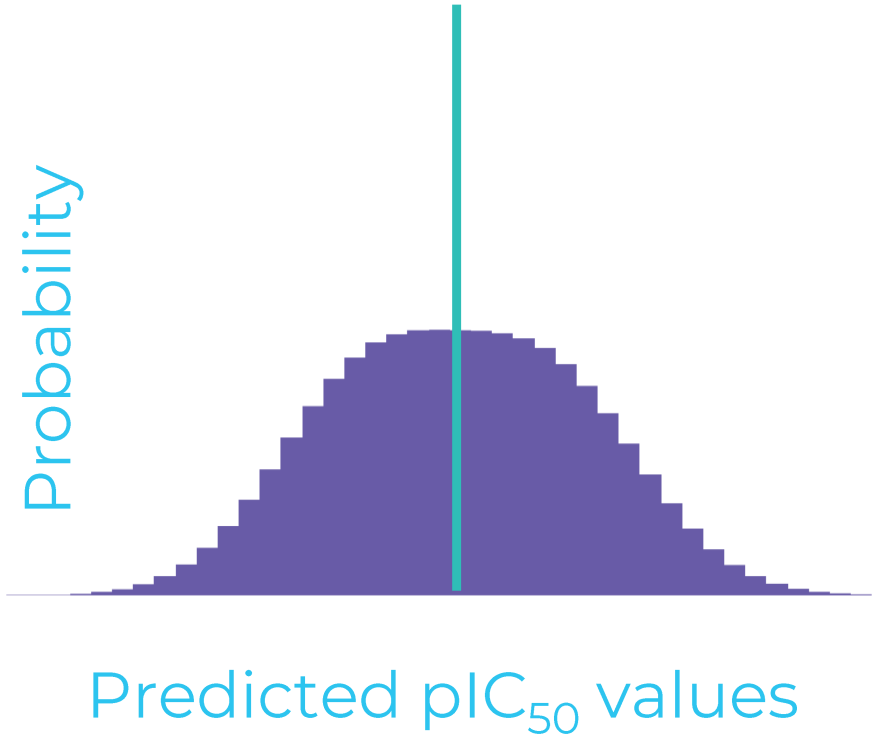
Whitehead *et al.*, J. Chem. Inf. Model. 59, 1197 (2019)

Calculate probability distribution

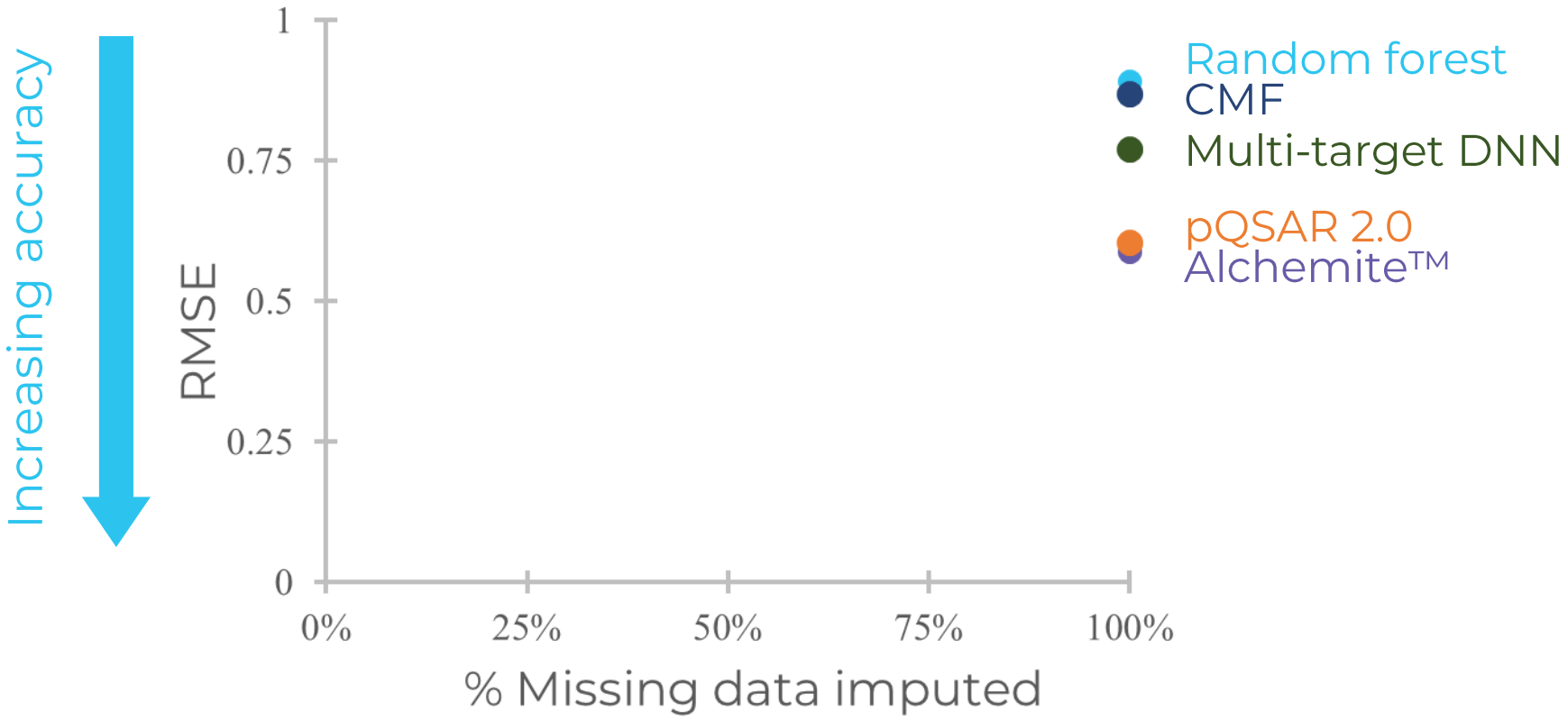


Focus on most confident predictions

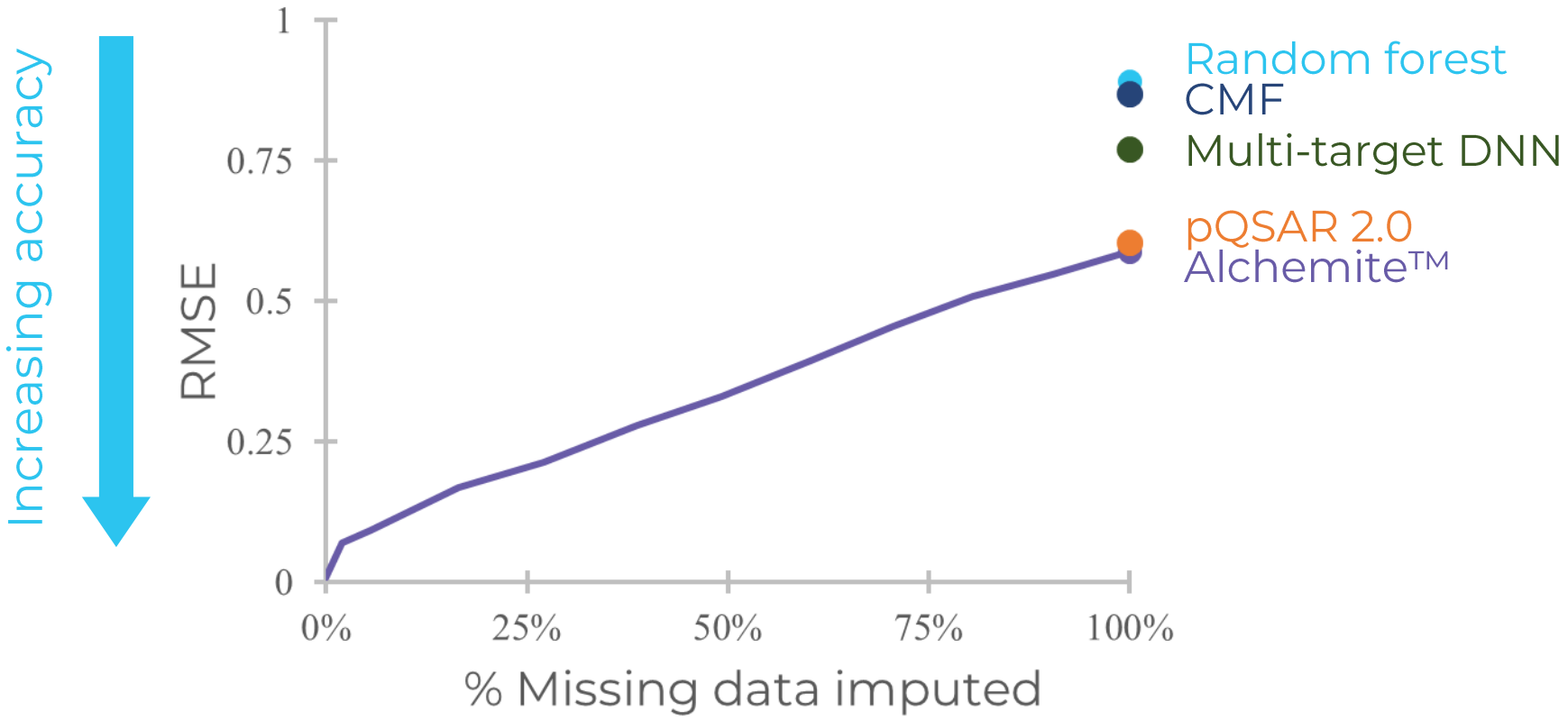




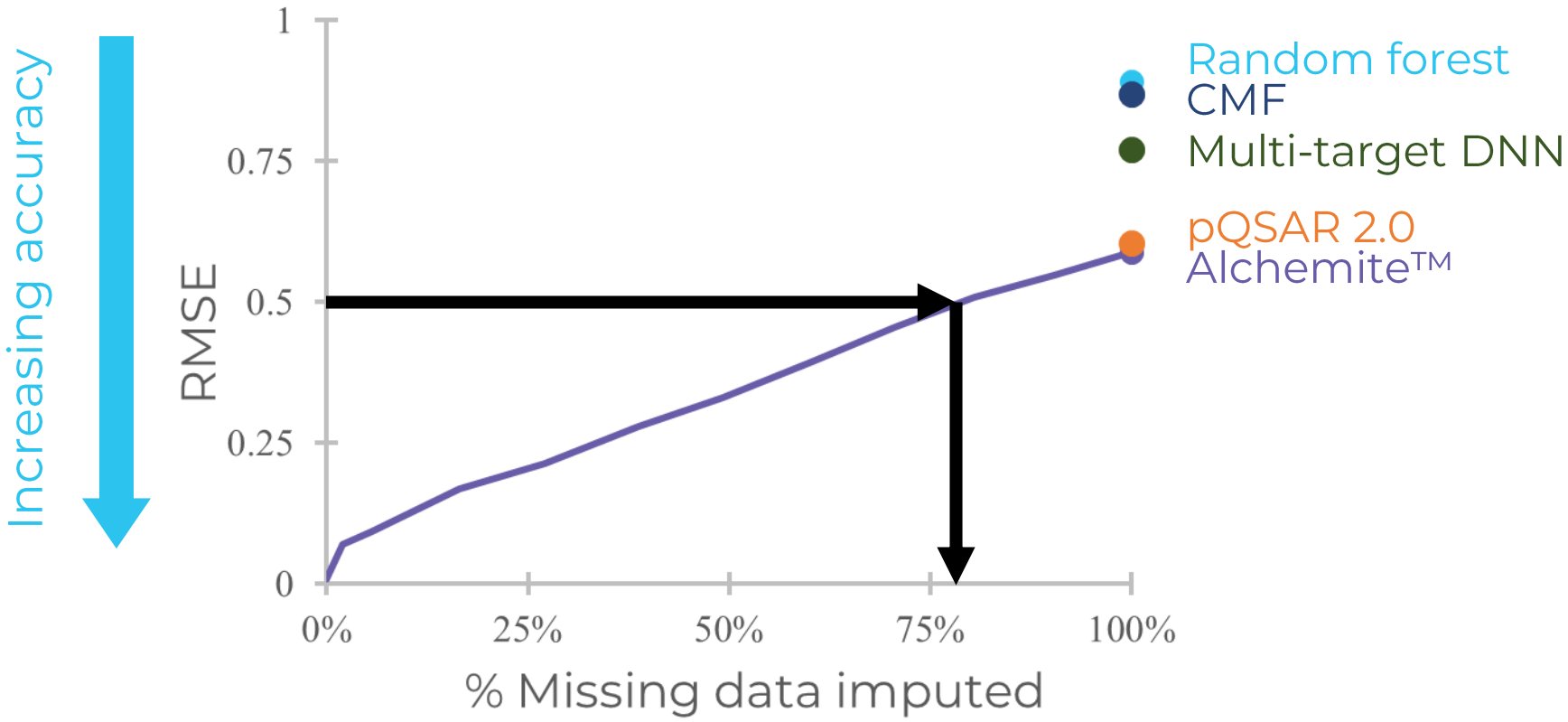
Reporting only most confident predictions



Reporting only most confident predictions



Reporting only most confident predictions



Real project application

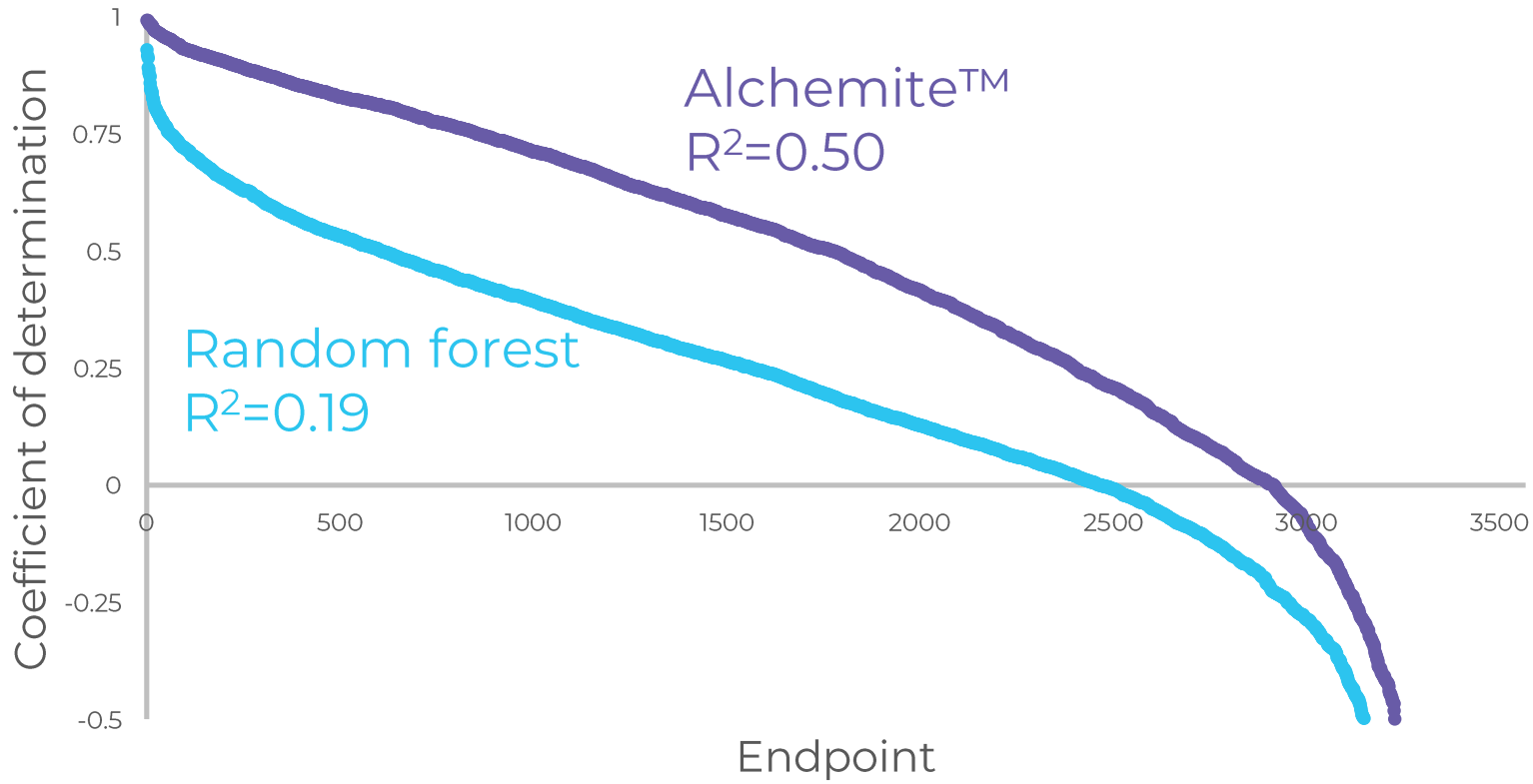
Big Pharma company

710,305 compounds

2,171 assays totalling **3,568** endpoints

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

Results


Alchemite™
R^2=0.50
Random forest
R^2=0.19
Coefficient of determination
1
0.75
0.5
0.25
0
-0.25
-0.5
0
500
1000
1500
2000
2500
3000
3500
Endpoint

Summary

Train across all endpoints simultaneously to capture **activity-activity** correlations using sparse data as **input**

Understand and exploit **probability distribution** to focus on most confident results

Applicable to pharma-scale data sets **today** through Alchemite platform

For more details: Whitehead *et al.*, J. Chem. Inf. Model. 59, 1197 (2019)

tom@intellegens.ai

Collaborate using Alchemite

Optibrium and **Intellegens** are working together on applying Alchemite to drug discovery and compound optimisation

If you'd like to discuss how to use Alchemite on your project or data, please contact

info@optibrium.com

intellegens.ai