Imputation of assay activity data using deep learning

Intellegens

Tom Whitehead

Unique deep learning algorithm

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

Impute assay bioactivity levels from sparse data

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

Broadly applicable algorithm with **proven** applications in drug design and materials discovery

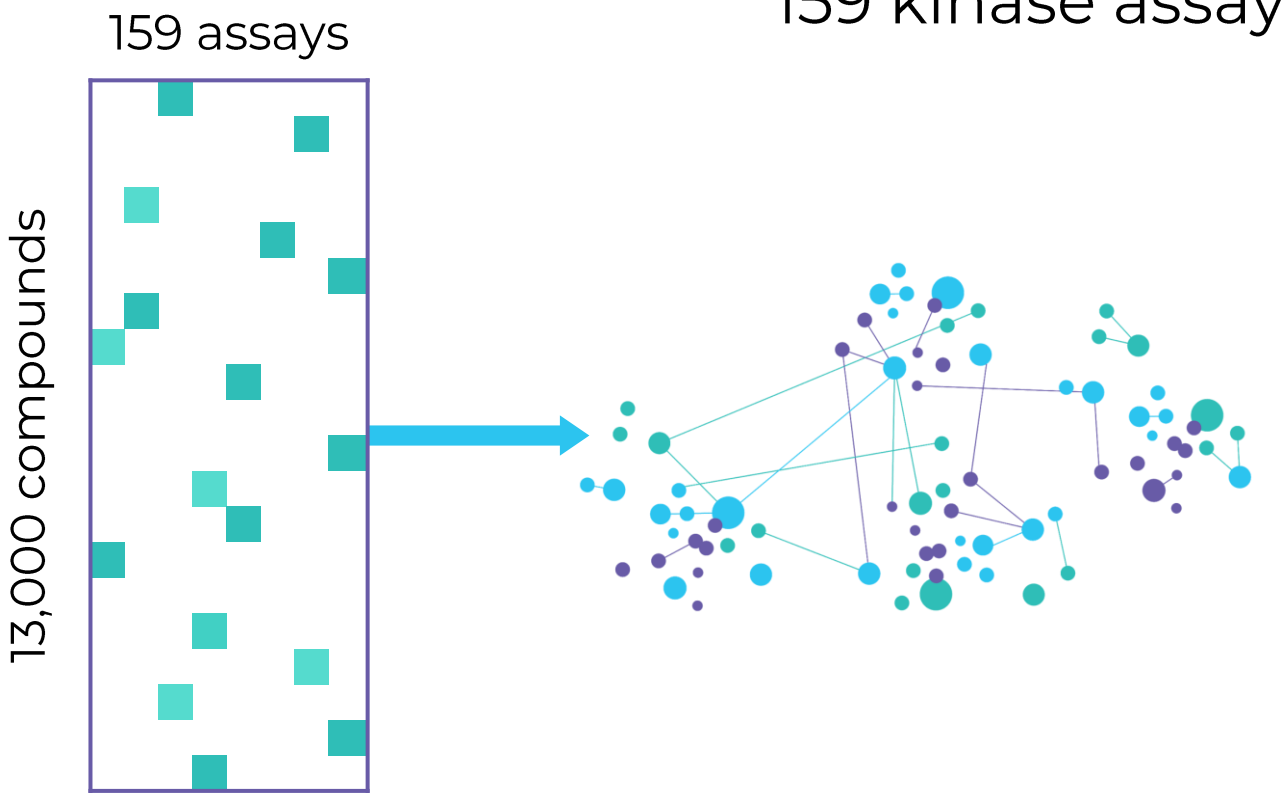
Deep learning

Alchemite™ deep learning

Novartis dataset to benchmark machine learning

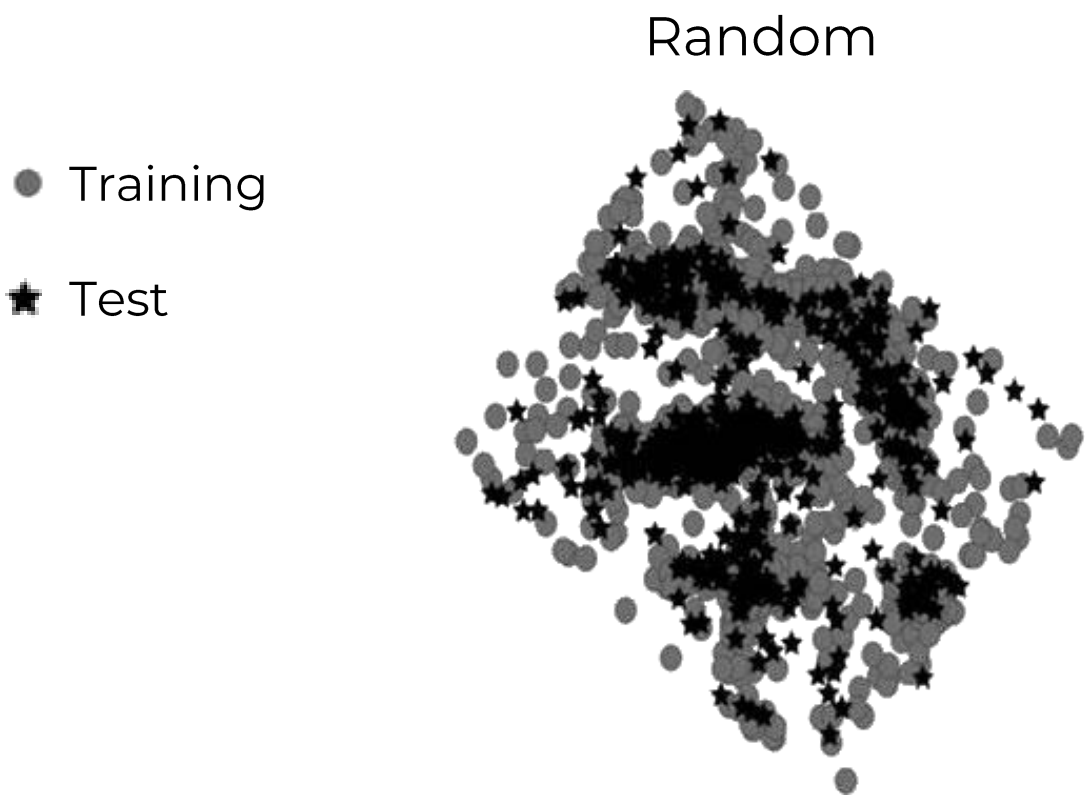
159 kinase assays for 13,000 compounds

Data **5%** complete



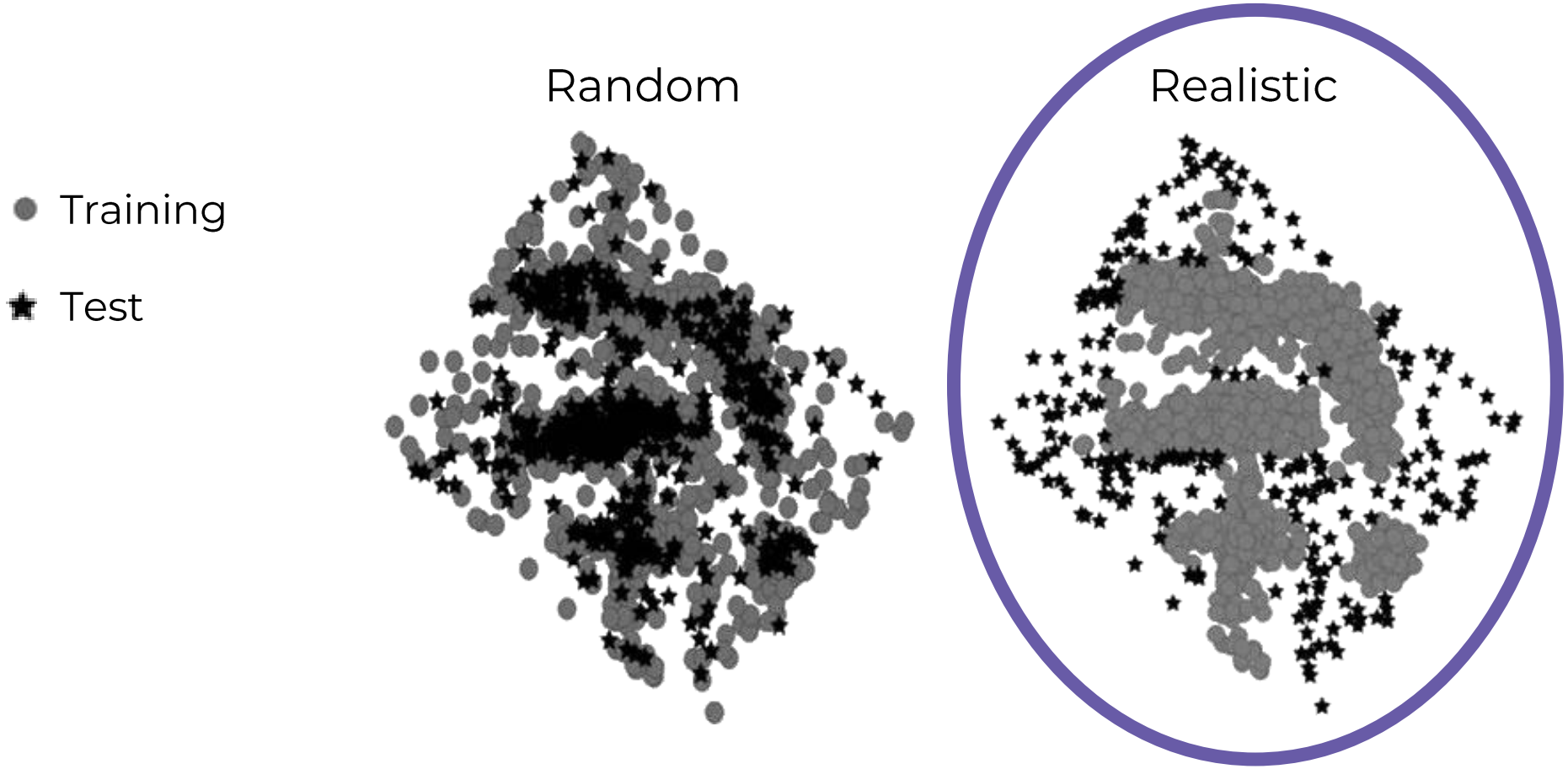
Data from ChEMBL
Martin, Polyakov, Tian, and Perez, J. Chem. Inf. Model. 57, 2077 (2017)

Novartis dataset distribution



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

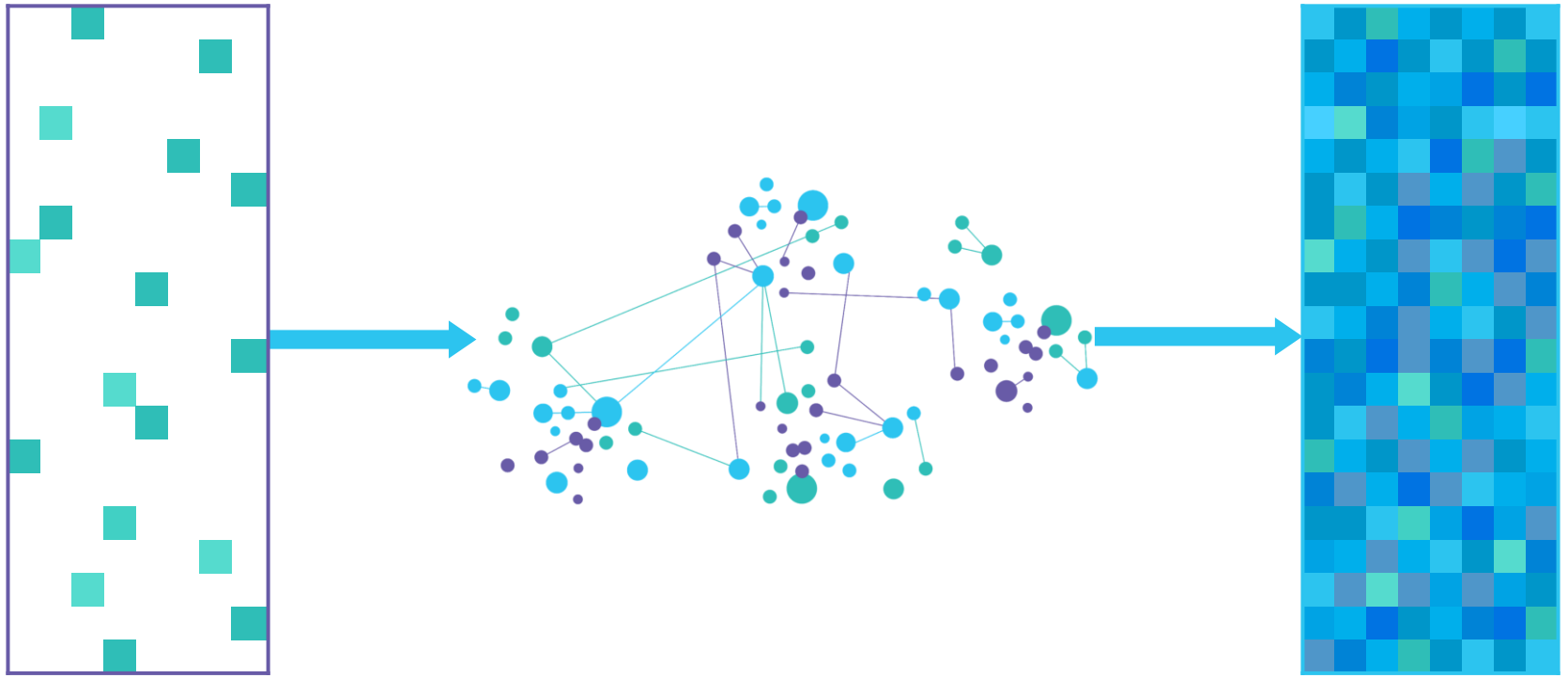
Novartis dataset is realistically distributed



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

Aim: impute missing assay values

Validate against realistically-split holdout set



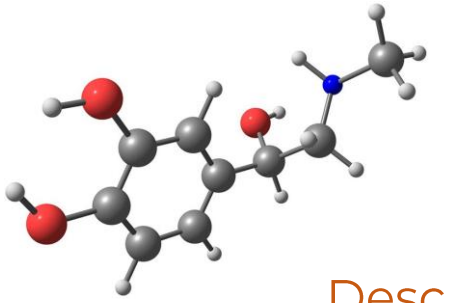
Data from ChEMBL
Martin, Polyakov, Tian, and Perez, J. Chem. Inf. Model. 57, 2077 (2017)

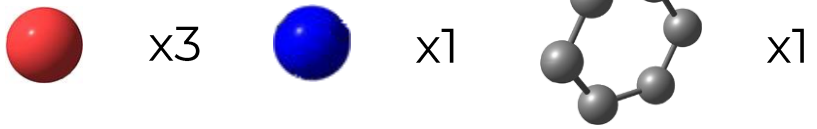
Accuracy metric

Coefficient of determination, R^2

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

Random forest regression

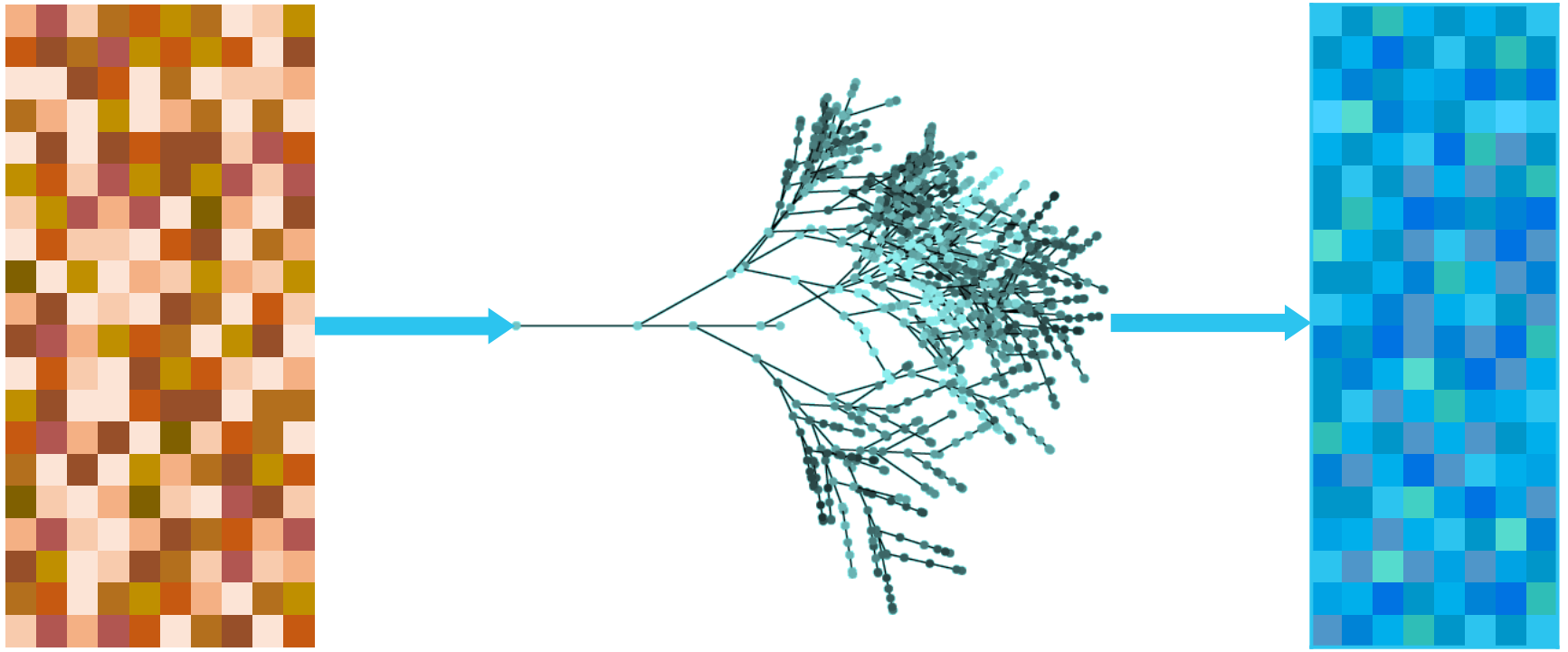




Molecular weight = 183 Da

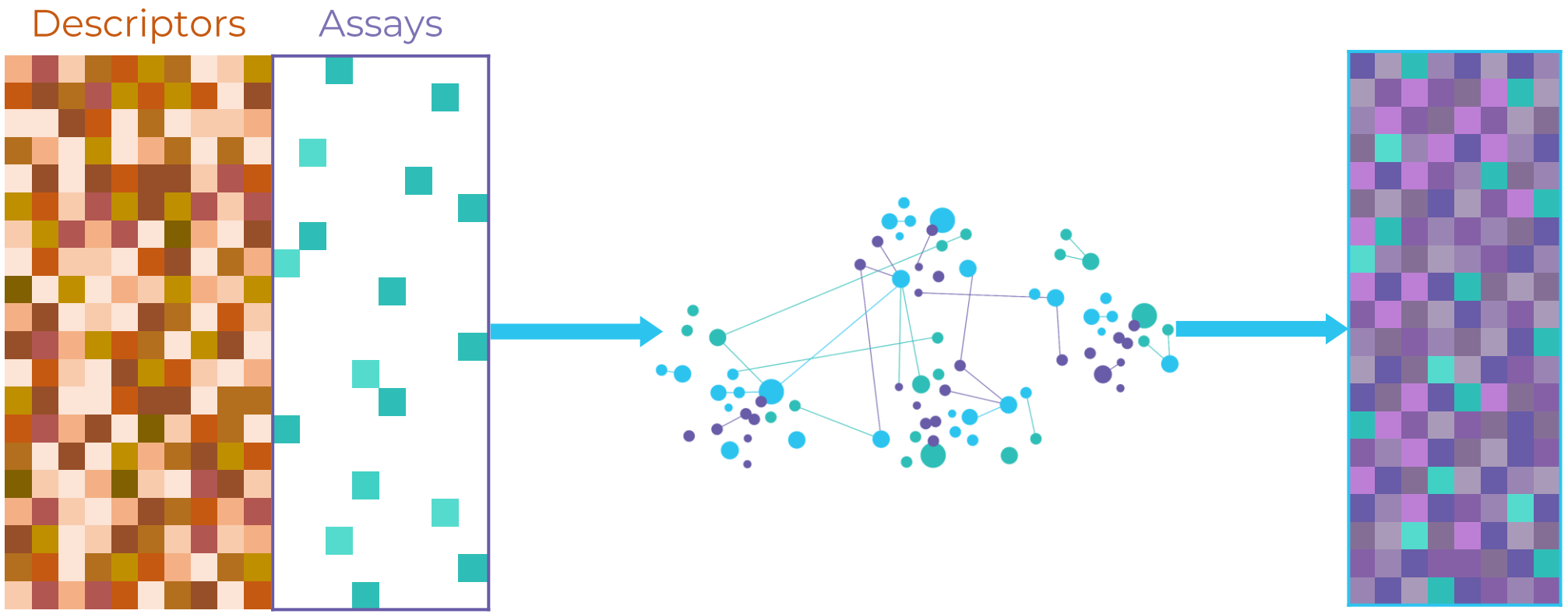
$R^2 = -0.19$

Descriptors



intellegens.ai

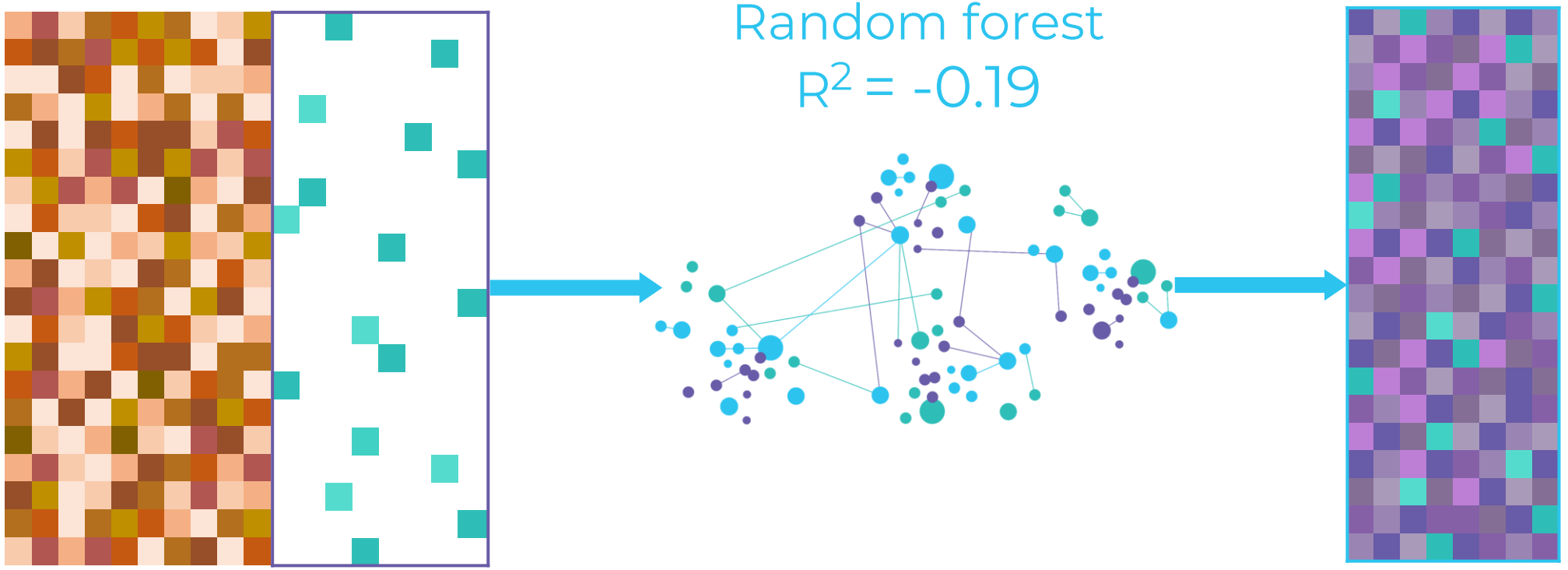
Descriptors and bioactivity values



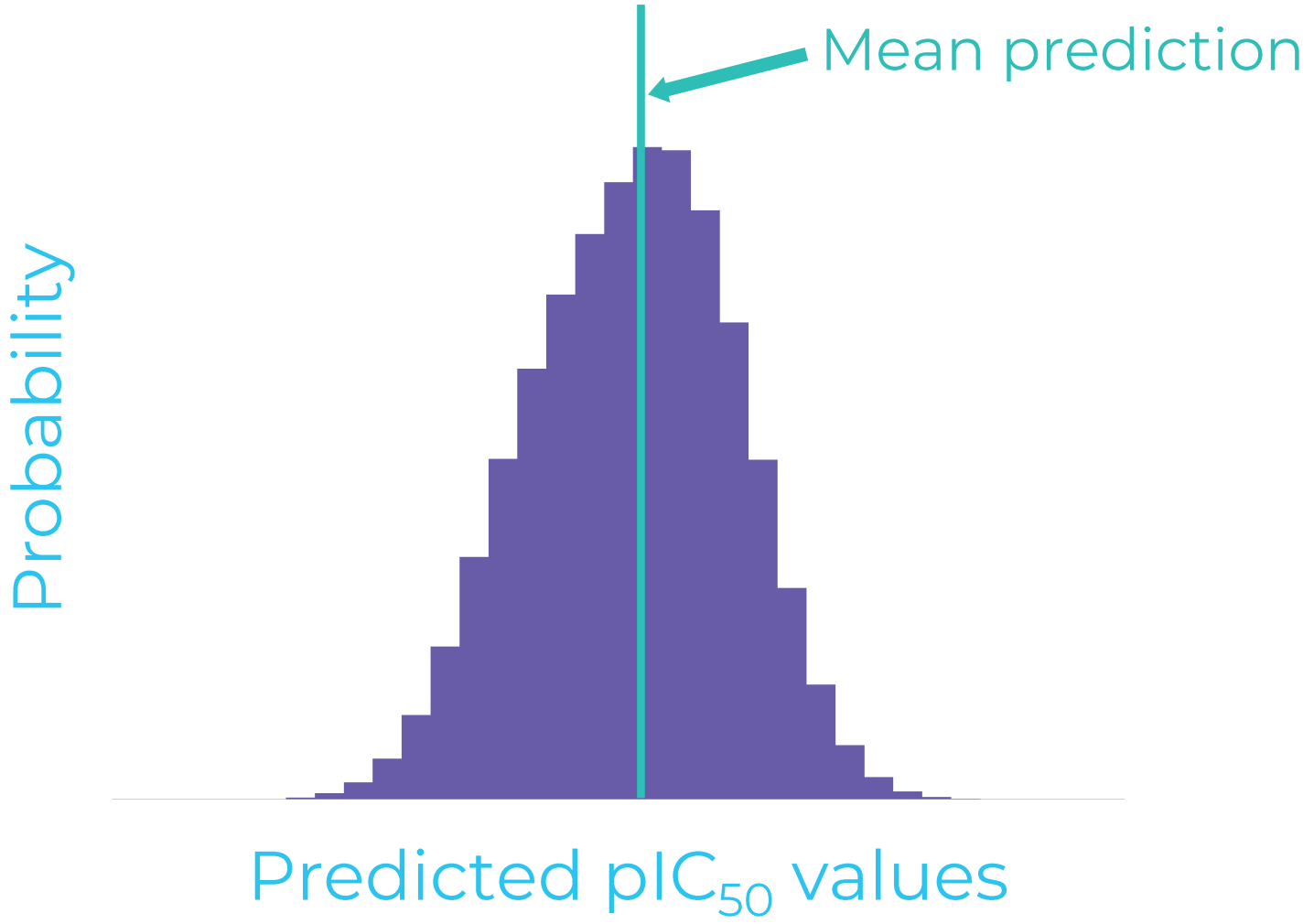
intellegens.ai

Deep learning predictions

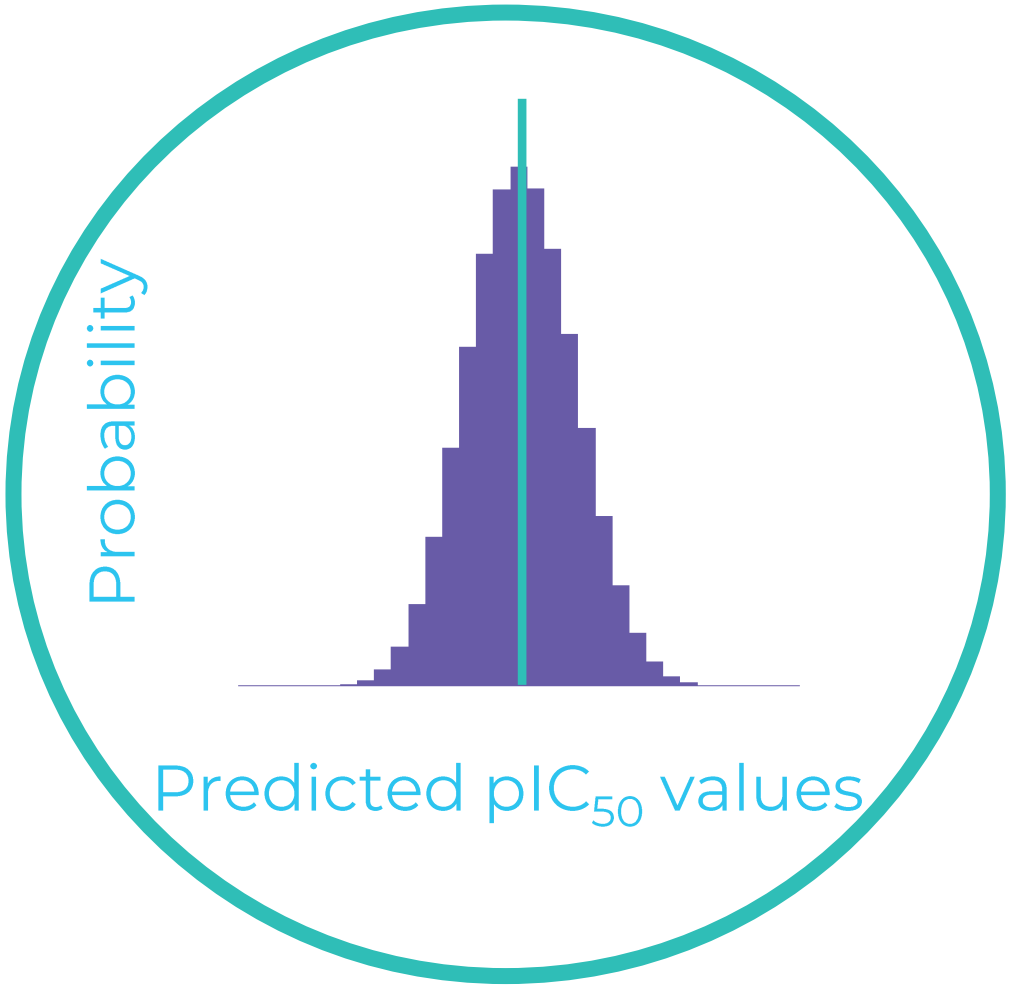
$R^2 = 0.44$

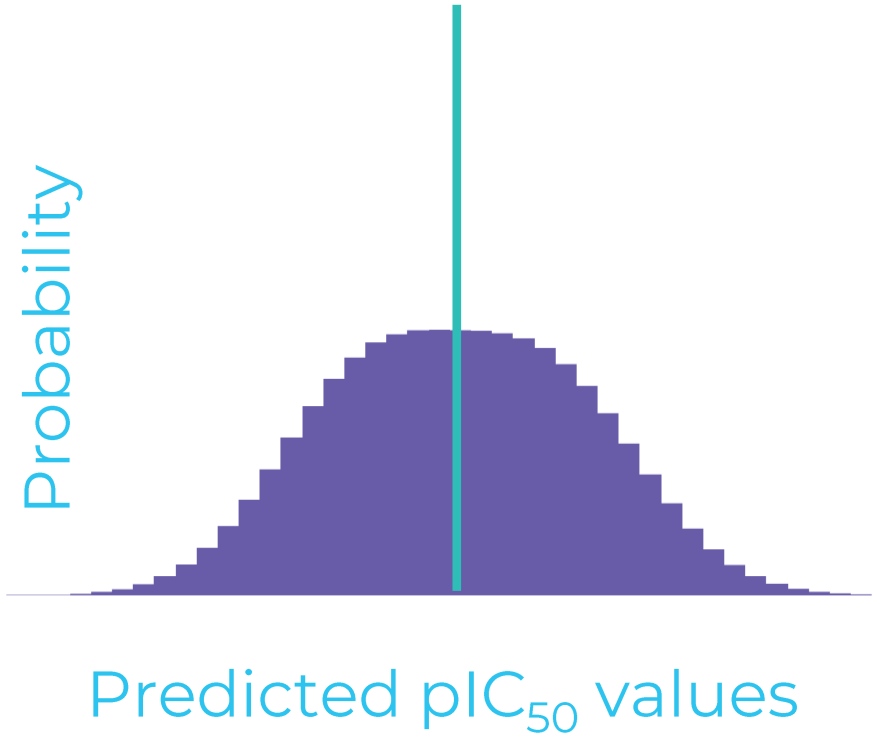


Calculate probability distribution



Focus on most confident predictions





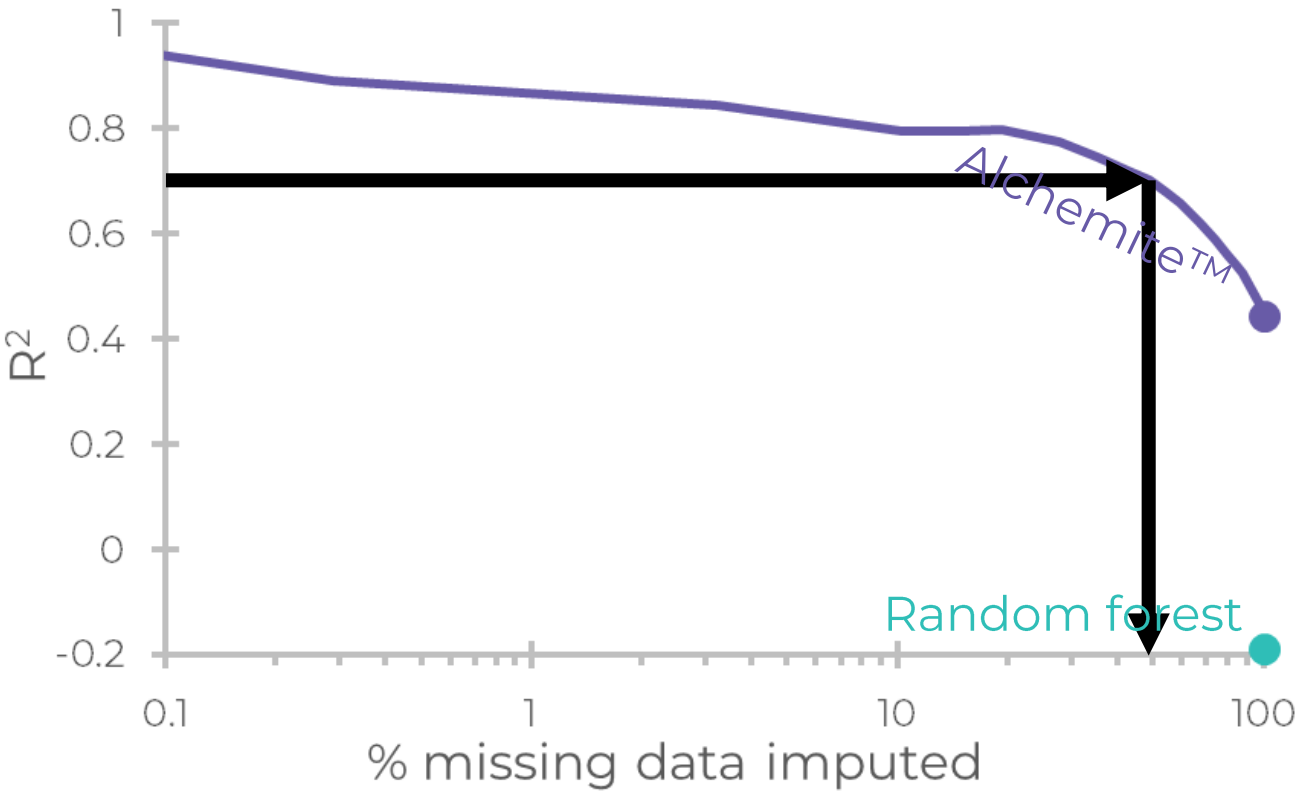
Reporting only most confident predictions

Reporting only most confident predictions

Increasing confidence

intellegens.ai

Reporting only most confident predictions



Increasing confidence

intellegens.ai

Summary

Train across all endpoints simultaneously to capture **activity-activity** correlations

Impute results of missing assays to high accuracy, enabling identification of **new hits** and computational screening of compounds

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

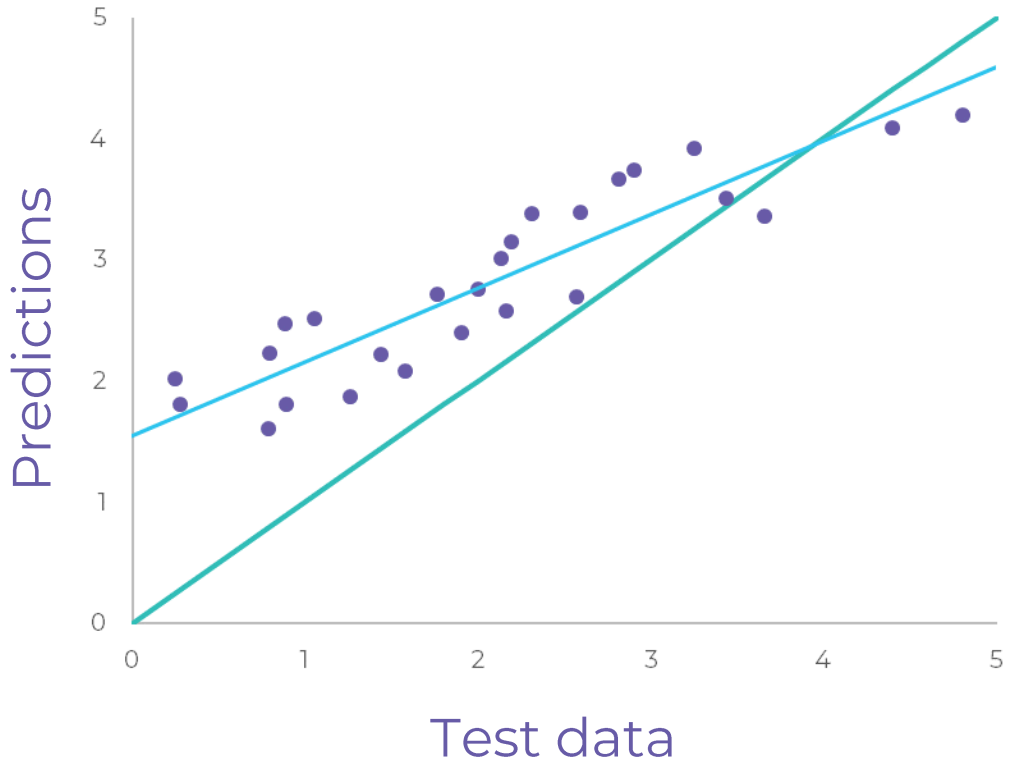
Any questions?

Intellegens

Dr Tom Whitehead

tom@intellegens.ai

Statistical interlude: Coefficient of determination, R^2



Squared correlation coefficient, r^2, compares to best fit line

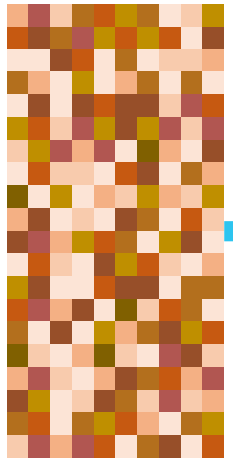
$r^2 = 0.94$

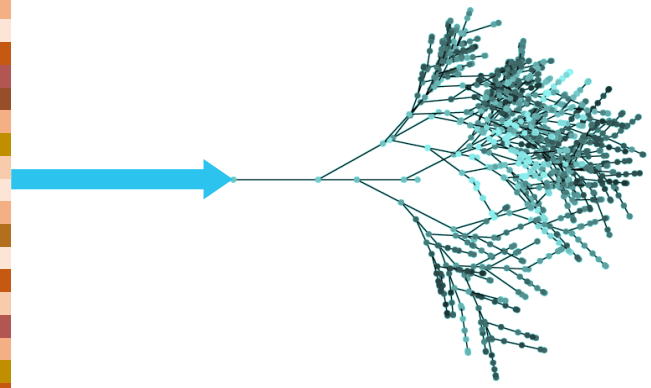
Coefficient of determination, R^2, compares to identity line

$R^2 = 0.77$

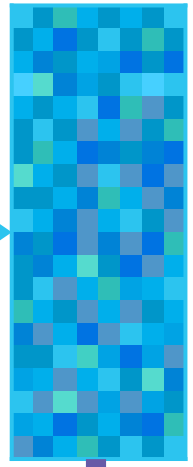
pQSAR 2.0 method

Descriptors

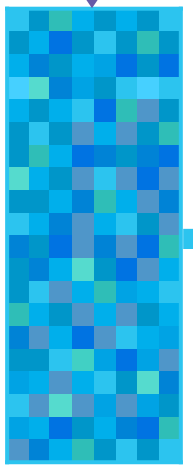


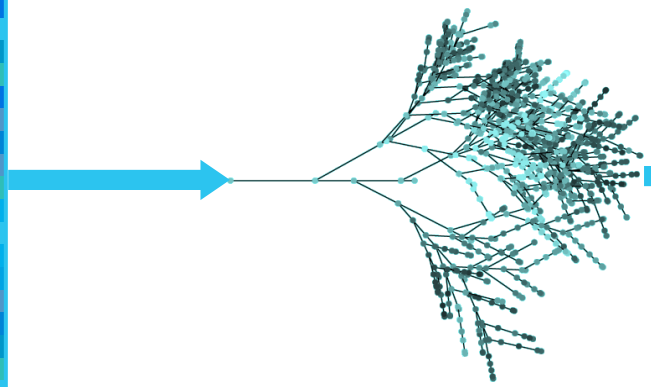


Assay predictions

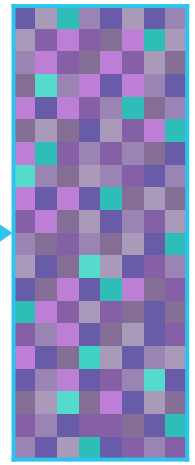


$R^2 = 0.43$





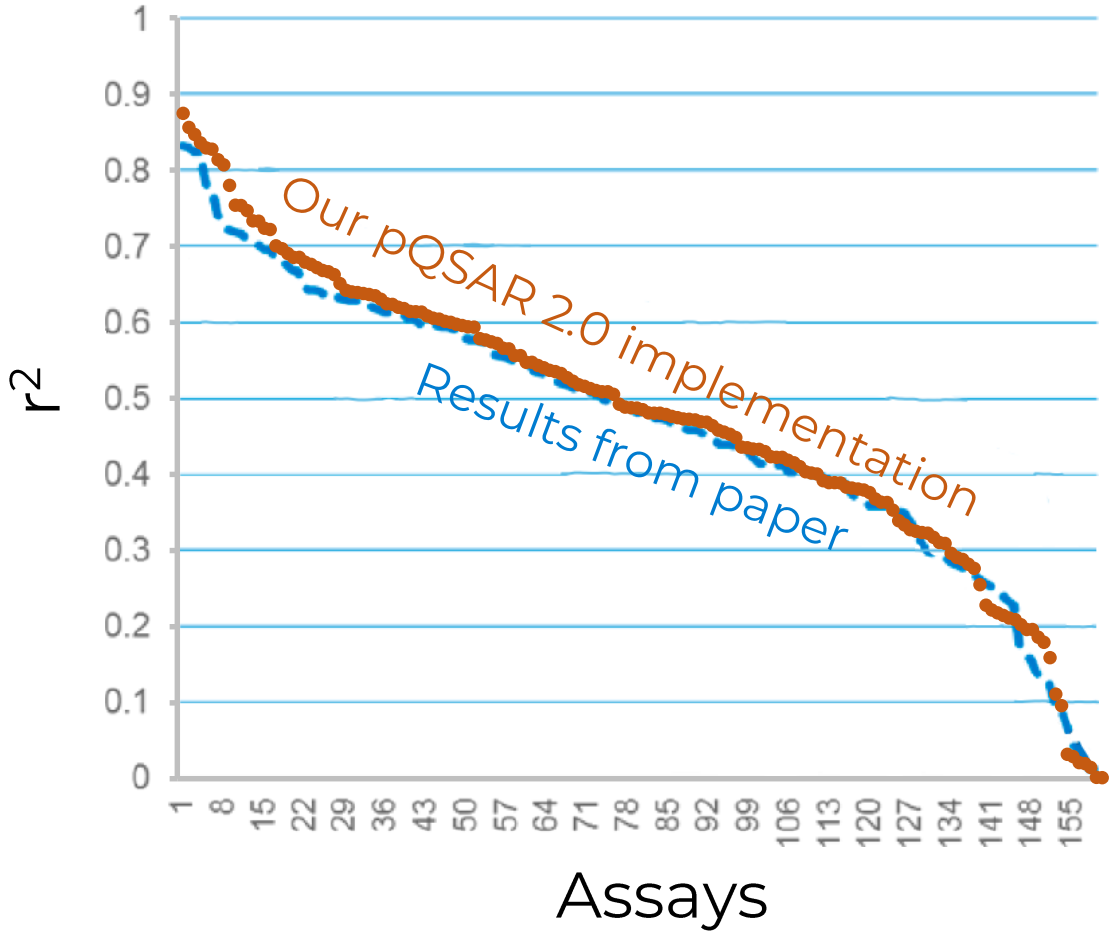
Final assay predictions



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

intellegens.ai

pQSAR 2.0 results



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

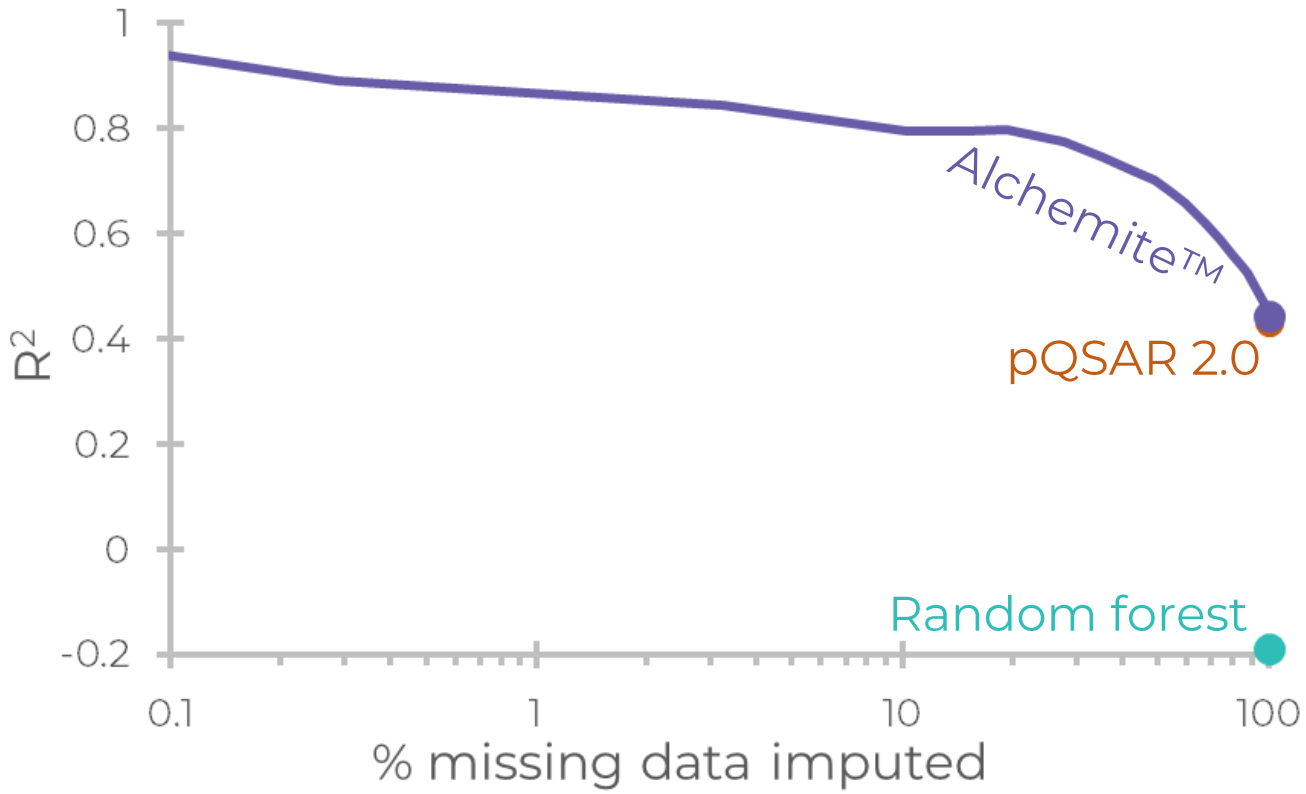
Reporting only most confident predictions

Validating against realistically-split holdout set

Increasing confidence

intellegens.ai

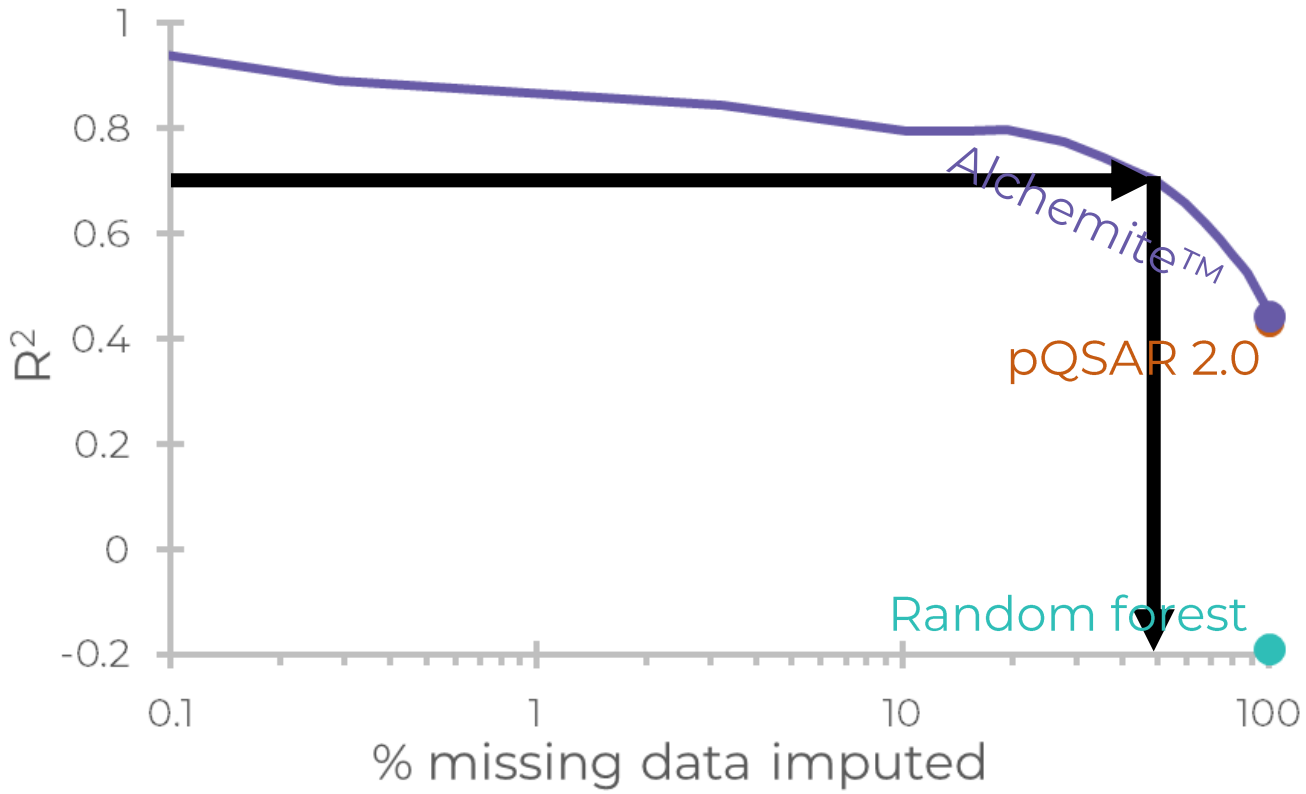
Reporting only most confident predictions



Increasing confidence

intellegens.ai

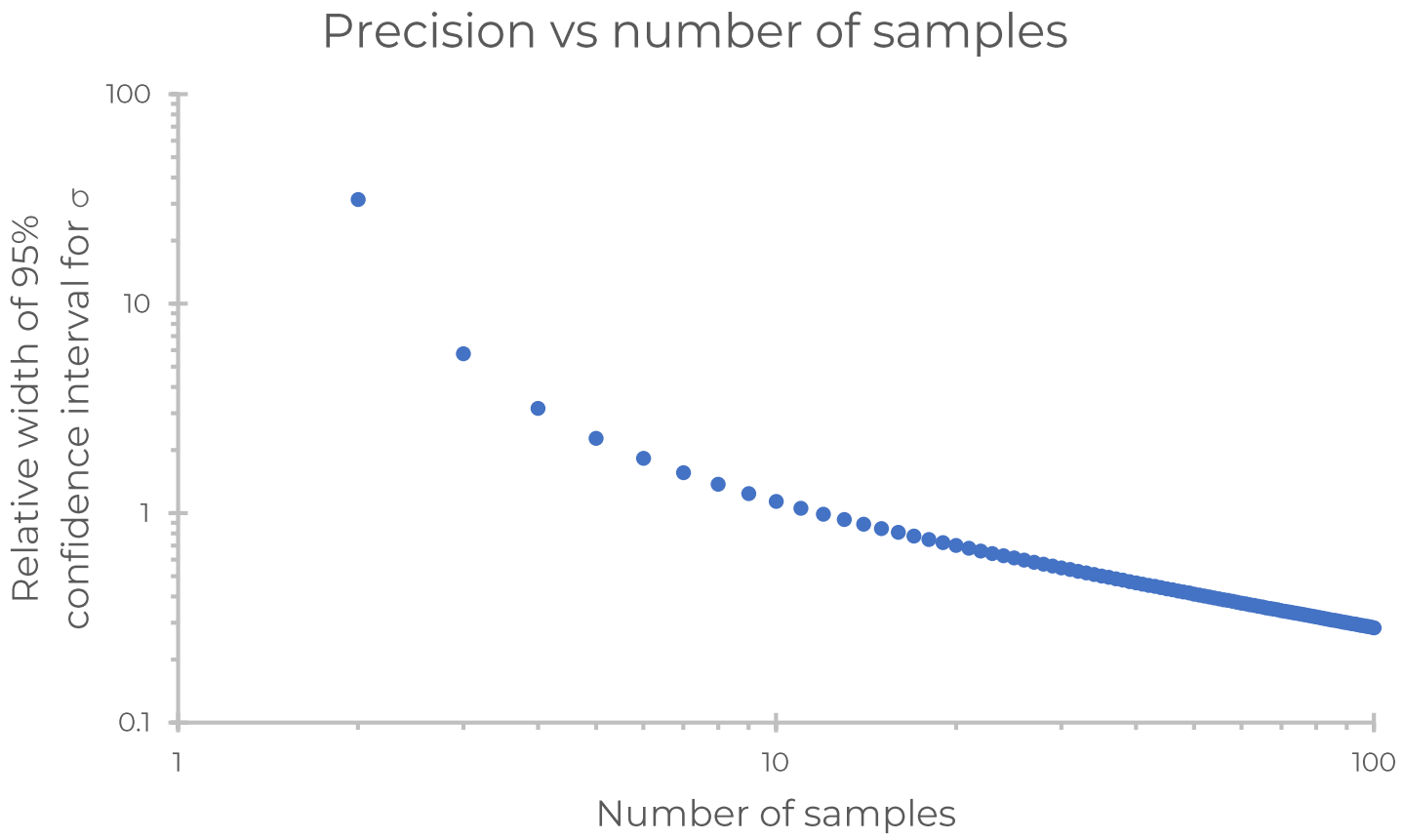
Reporting only most confident predictions



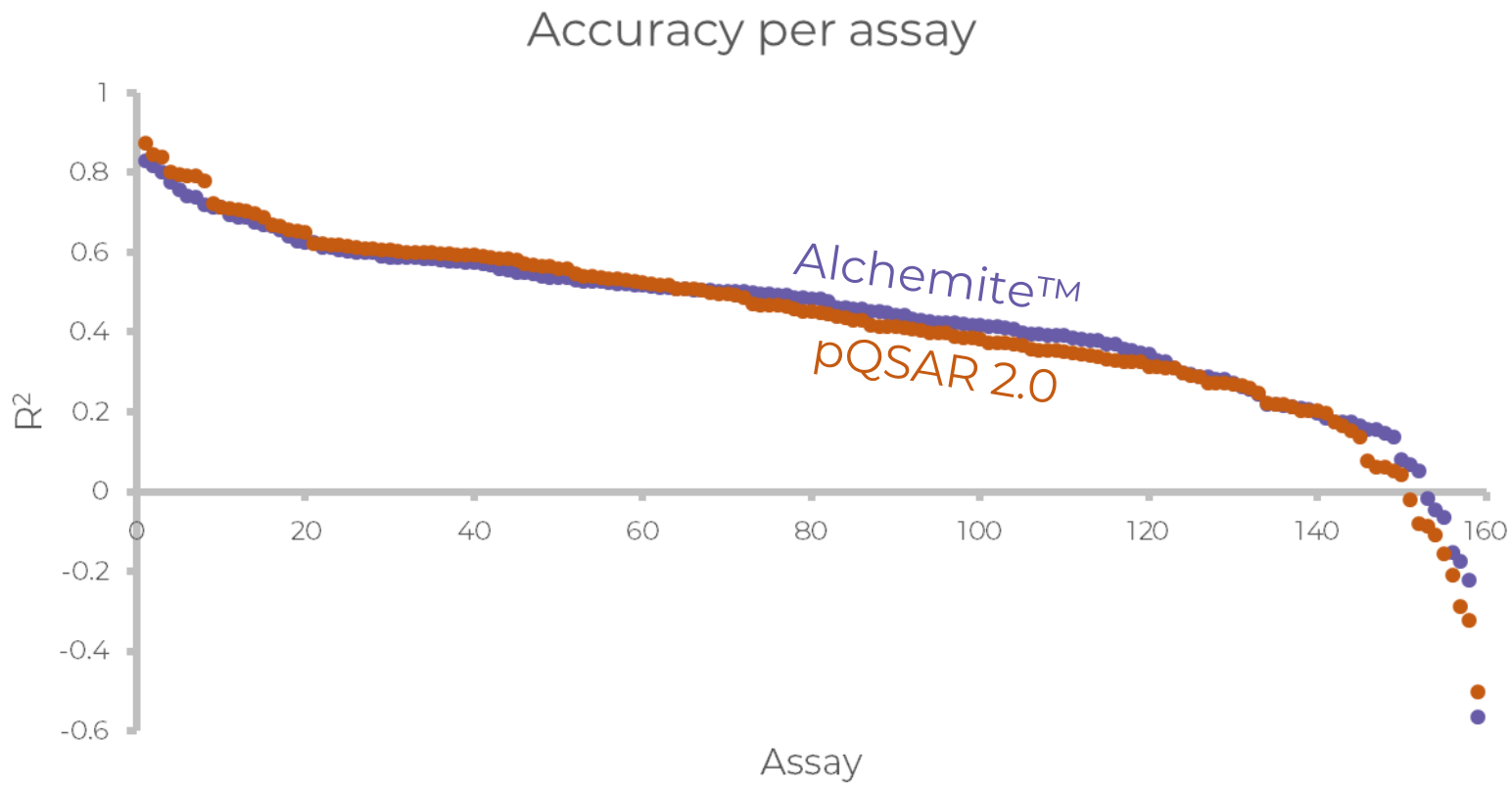
Increasing confidence

intellegens.ai

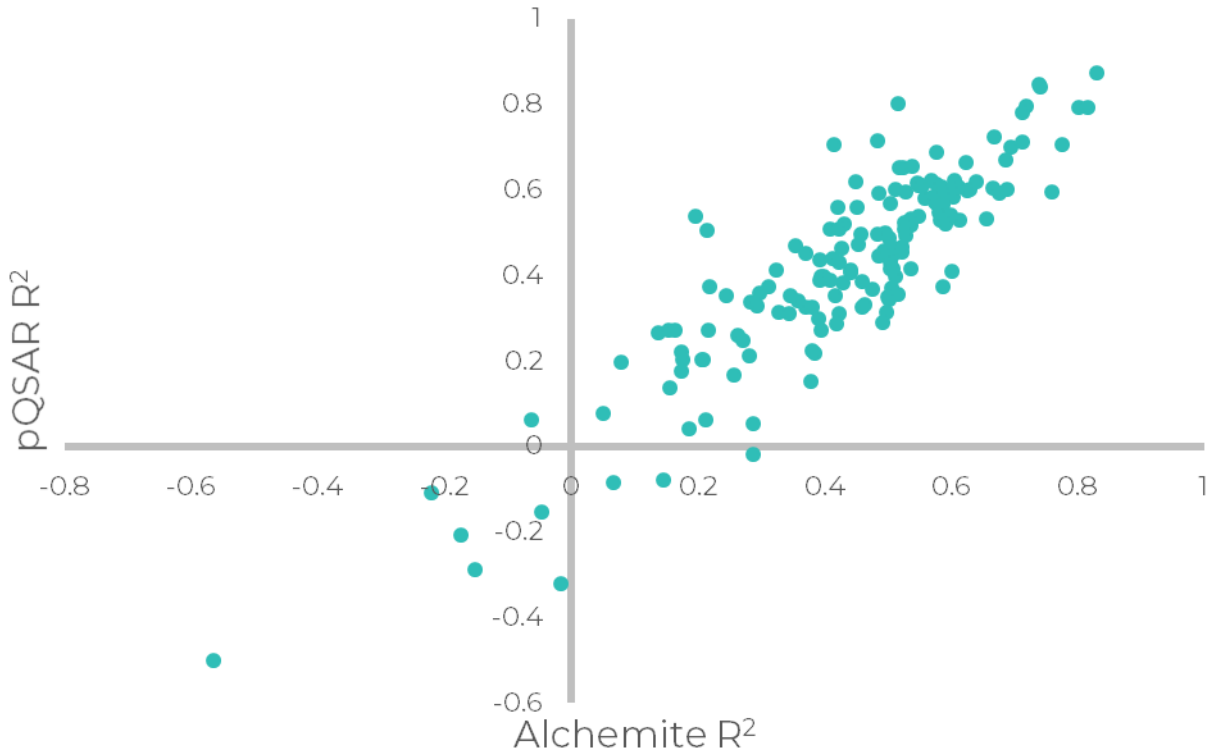
Samples from probability distribution



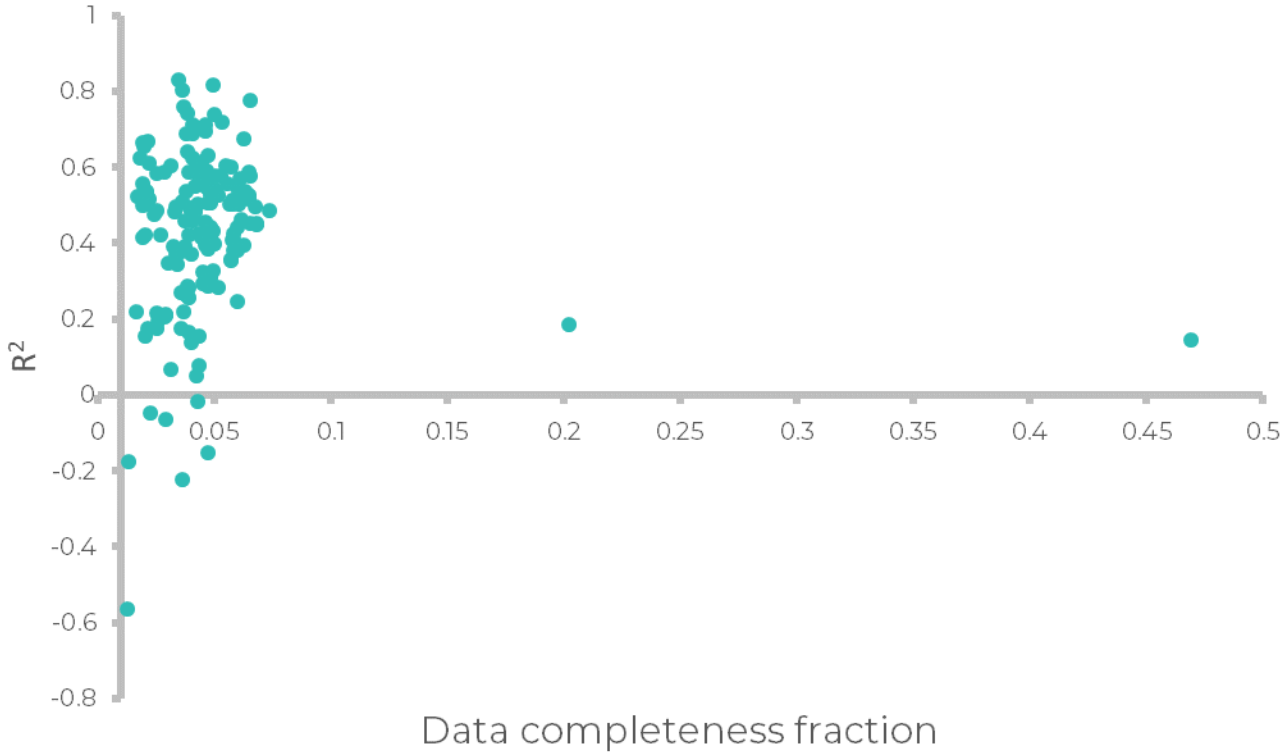
Accuracy per assay



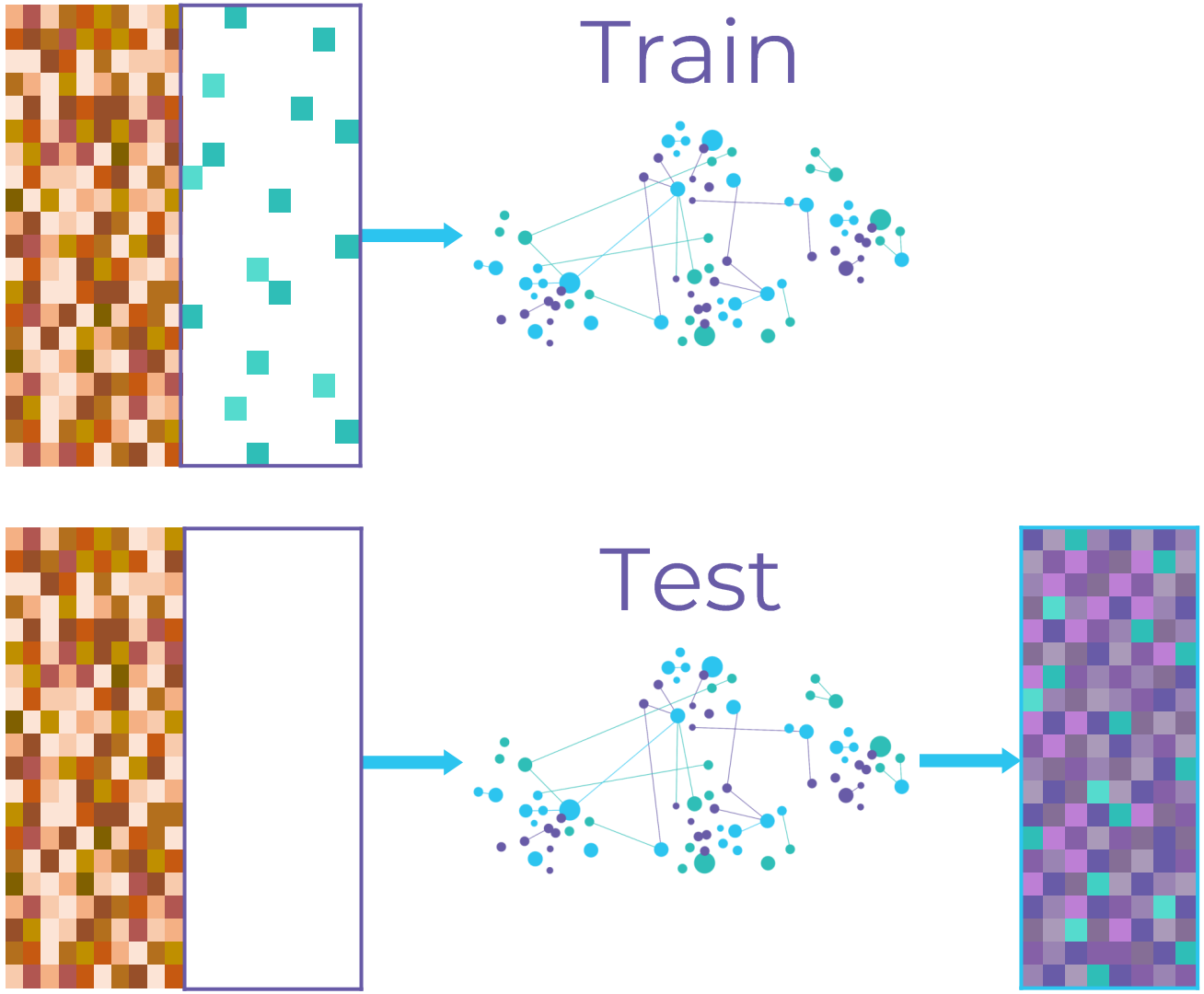
Accuracy on assays



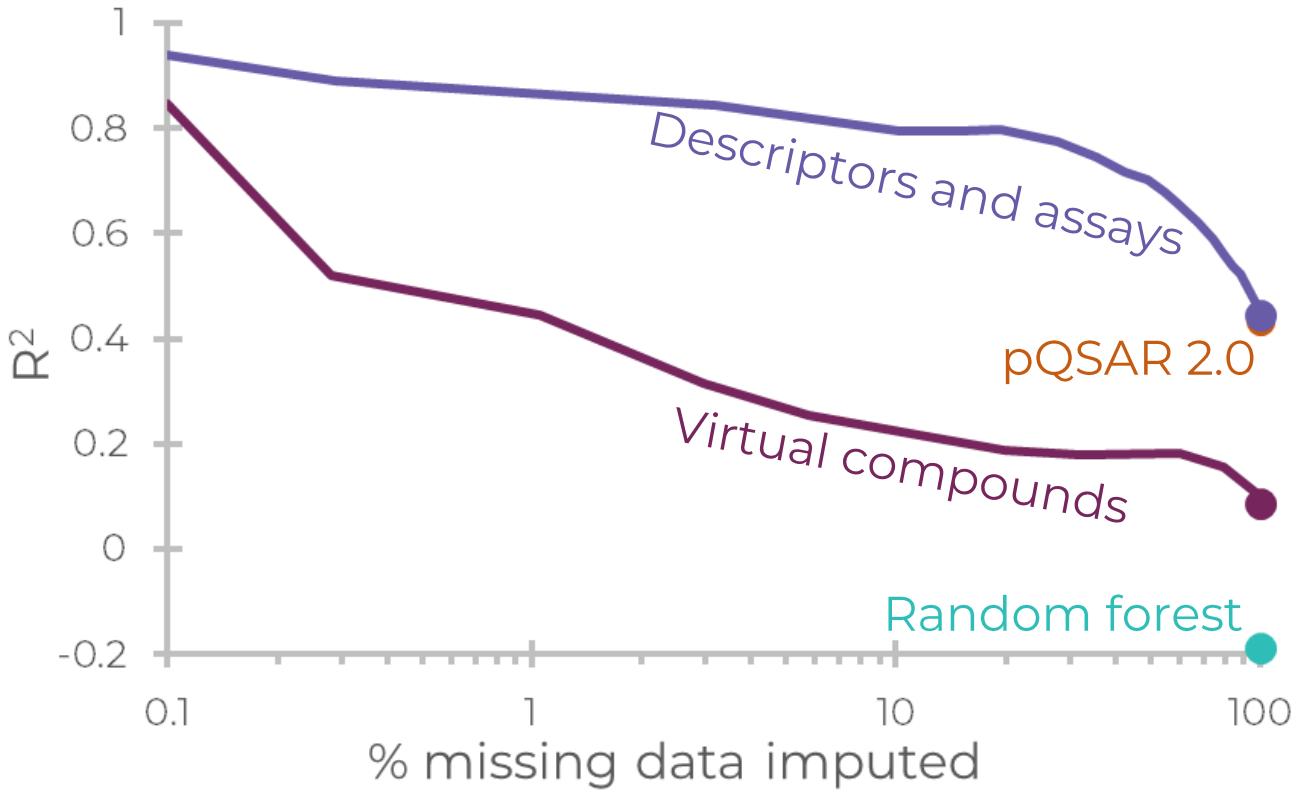
Accuracy vs level of data



Virtual compounds



Virtual compounds



Increasing confidence

intellegens.ai