Imputation of Protein Activity Data Using Deep Learning

Matthew Segall – matt@optibrium.com
Tom Whitehead, Intellegens – tom@intellegens.co.uk

Overview

- Prediction of compound activities in drug discovery
  - Quantitative structure-activity relationships
  - New ‘deep learning’ methods
  - Challenges of deep learning in drug discovery
- Intellegens’ Alchemite™ technology for deep learning – Tom Whitehead
  - Learning from sparse, noisy data
  - Example application to compound activity prediction
- Summary

Quantitative Structure-Activity Relationships

Predicting compound properties to guide design and selection

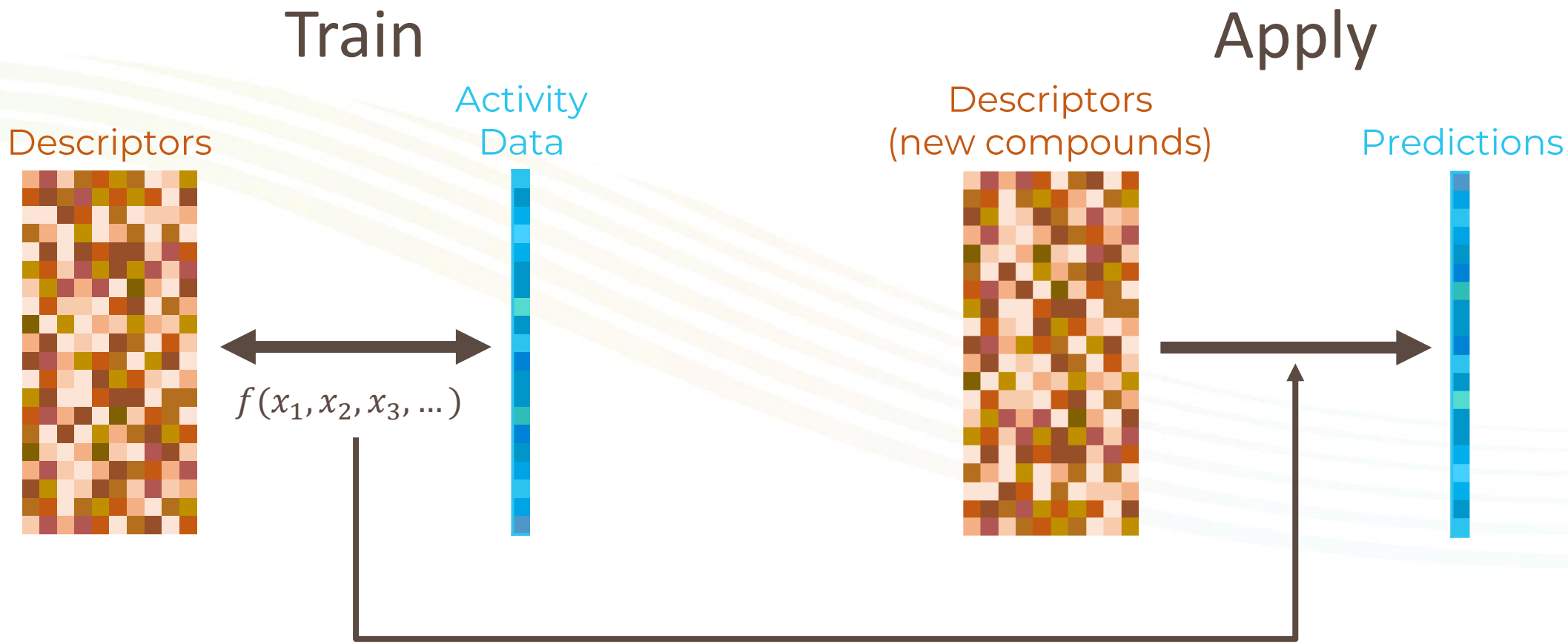
$$y = f(x_1, x_2, x_3, \dots) \pm \varepsilon$$

Statistical uncertainty

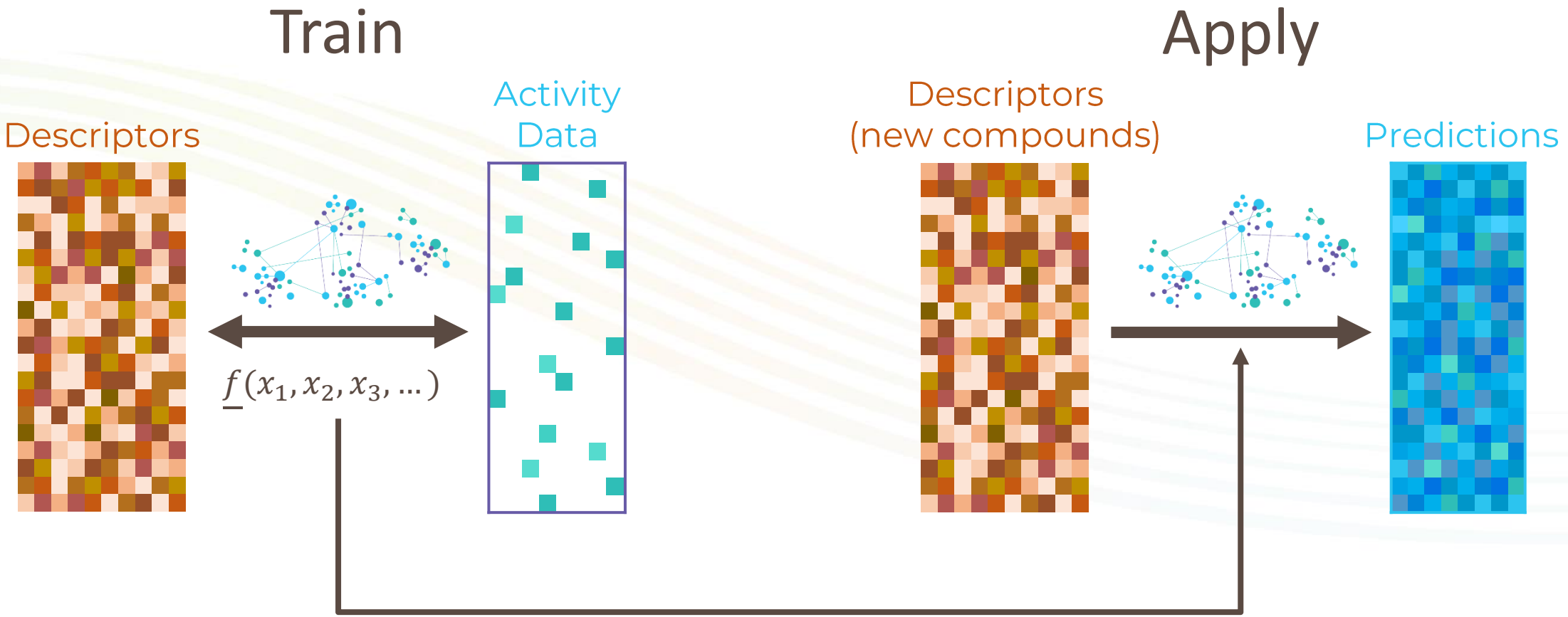
- Data
  - Quality data is essential
  - Public data need very careful curation* (and may not be good enough)
- Descriptors, e.g.
  - Whole molecule properties, e.g. logP, MW, PSA…
  - Structural descriptors, SMARTS, fingerprints…
- Machine learning method, e.g.
  - Artificial neural networks, support vector machines, random forest, Gaussian processes…

*Waldman et al. JCAMD (2015) DOI: 10.1007/s10822-015-9865-0

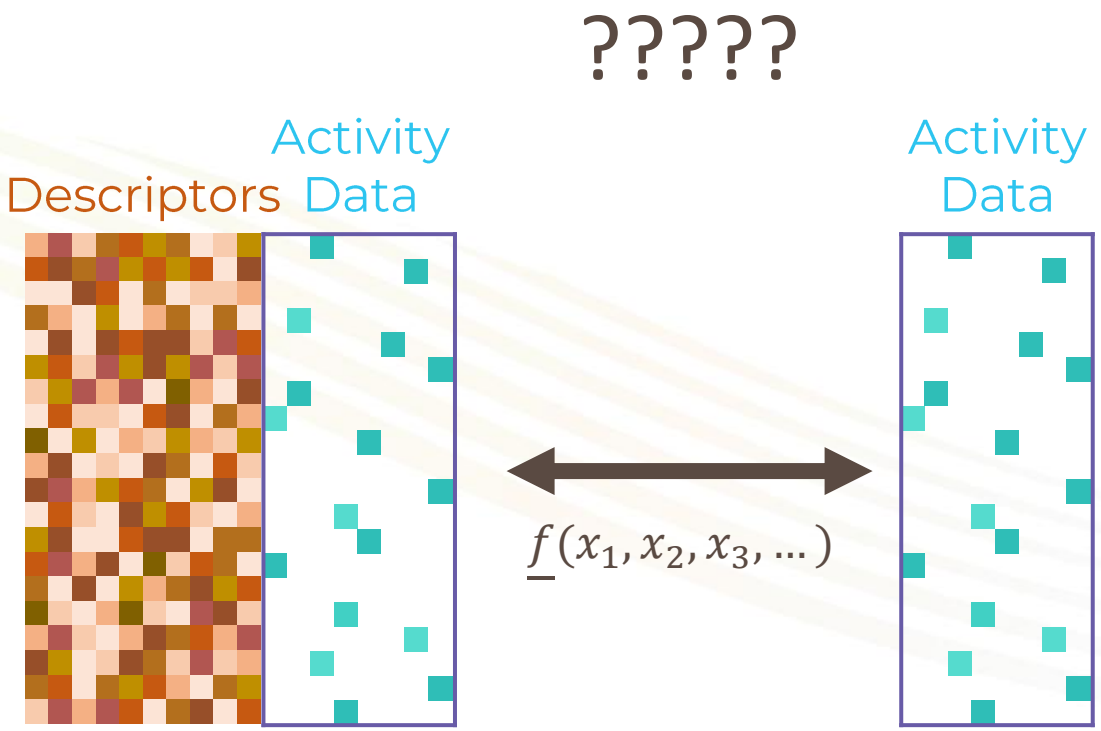
Quantitative Structure Activity Relationships



Multi-Target Prediction
E.g. Deep learning



Learning From Sparse Data?



The Challenges of Applying Deep Learning

- Application of conventional deep learning to traditional QSAR modelling offers little advantage
  - Robert Sheridan (Merck) reported an average improvement in R^2 of 0.04 over random forests across 30 representative QSAR data sets*
- Challenges
  - Compound bioactivity/property data is very sparse
  - 'Big data' in pharma is not very big! O(10^6) compounds and O(10^7) experimental data points
  - Biological data is noisy. ~0.3-0.5 log unit experimental variability
- How can we learn from these experimental data to make better predictions for compound bioactivities and properties?

*AI in Chemical Research, Switzerland, Sept.9 2018

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

Imputation of assay activity data using deep learning

Intellegens

optibrium

Tom Whitehead
Matt Segall

Unique deep learning algorithm

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

Impute assay bioactivity levels from sparse data

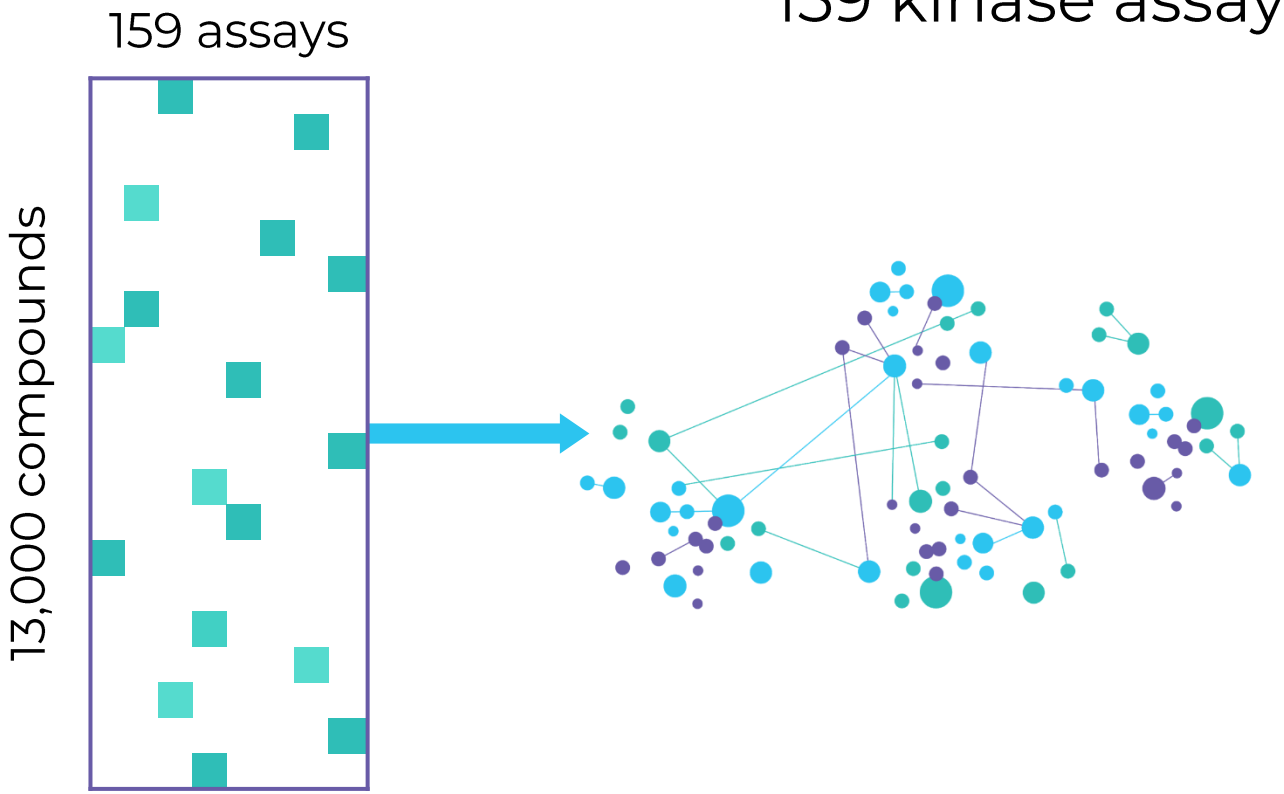
Understand and exploit **uncertainties** and noise to improve confidence in predictions

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

Deep learning

Alchemite™ deep learning

Novartis dataset to benchmark machine learning



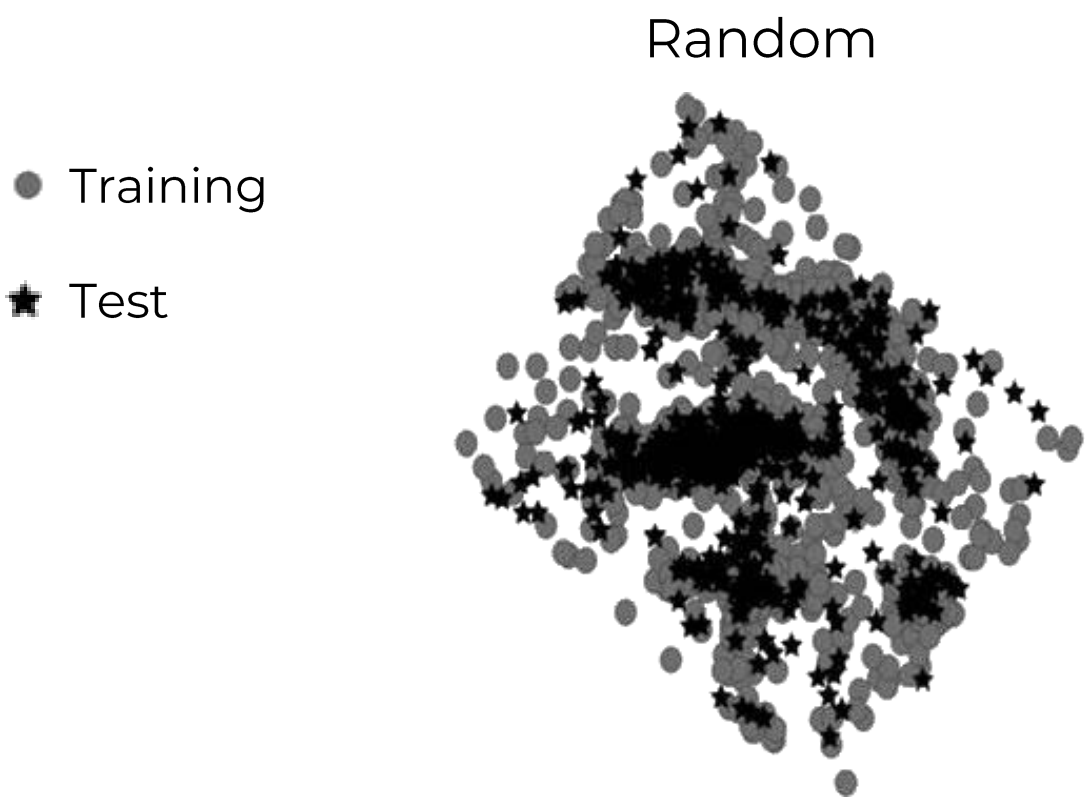
159 kinase assays for 13,000 compounds

Data **5%** complete

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

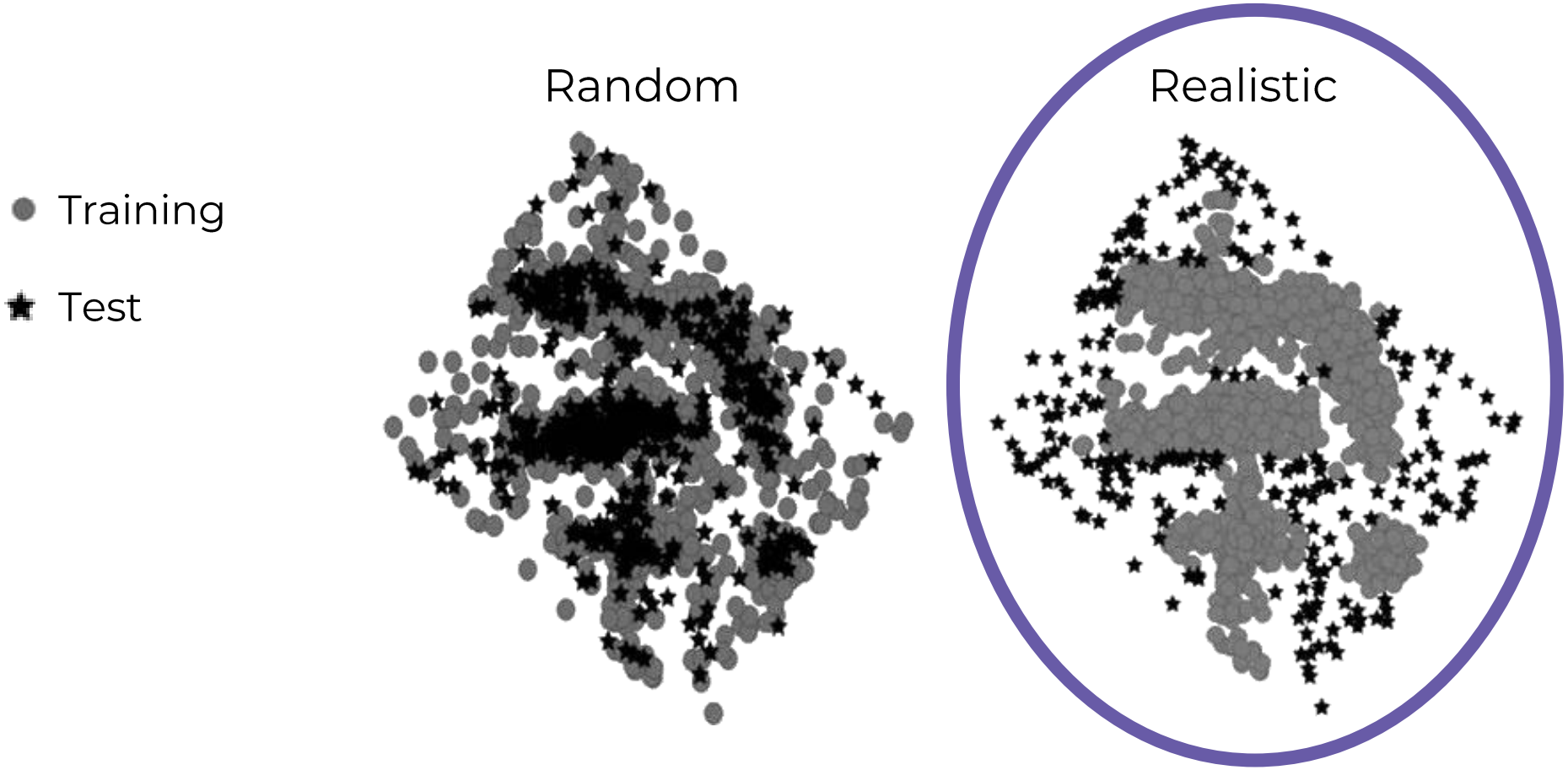
intellegens.ai

Novartis dataset distribution



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

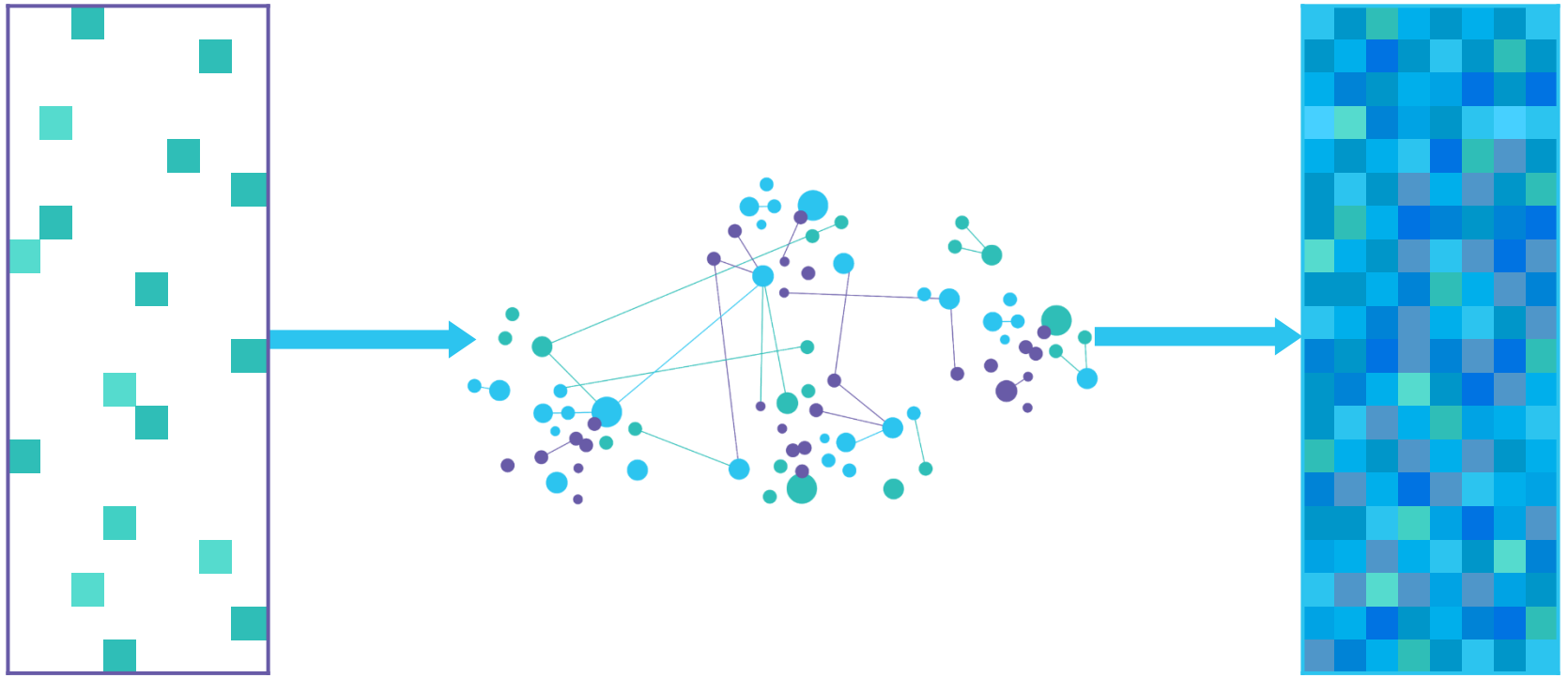
Novartis dataset is realistically distributed



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



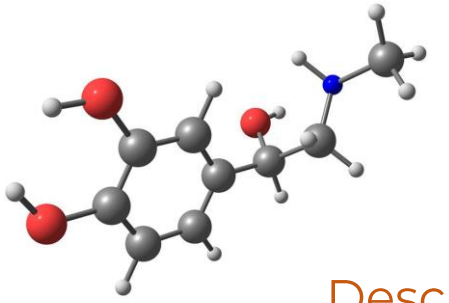
Data from ChEMBL
Martin, Polyakov, Tian, and Perez, J. Chem. Inf. Model. 57, 2077 (2017)

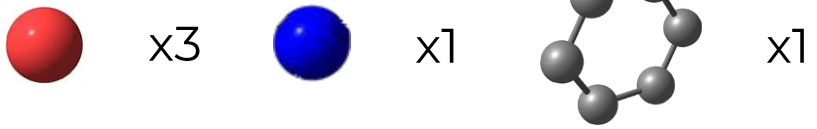
Accuracy metric

Coefficient of determination, R^2

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

Random forest regression

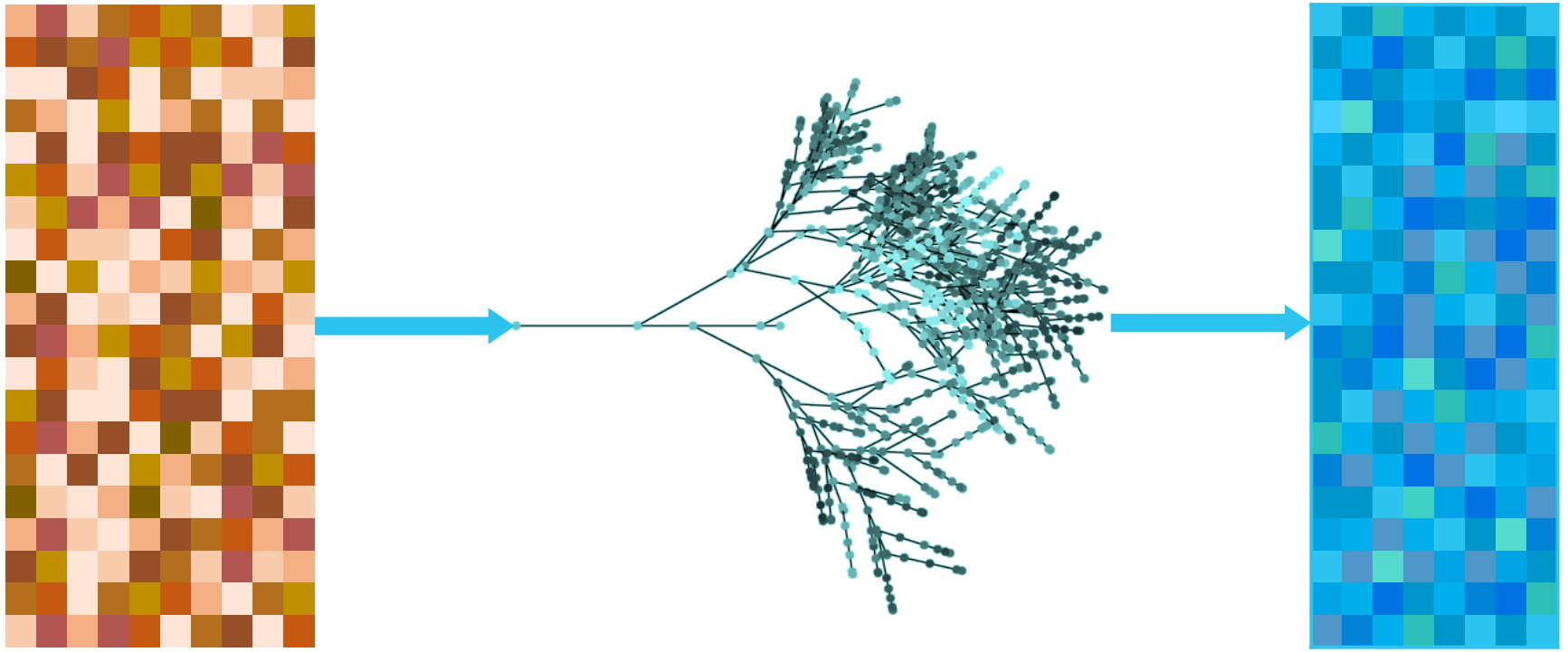




Molecular weight = 183 Da

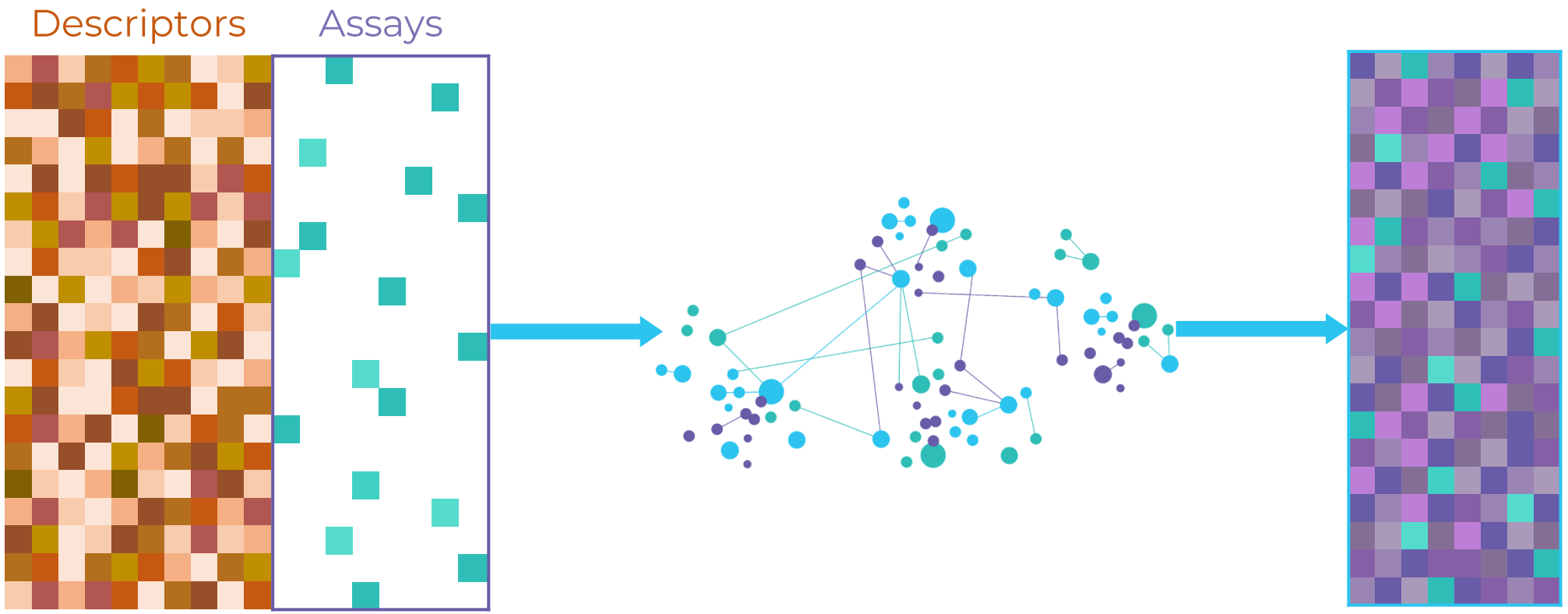
$R^2 = -0.19$

Descriptors



intellegens.ai

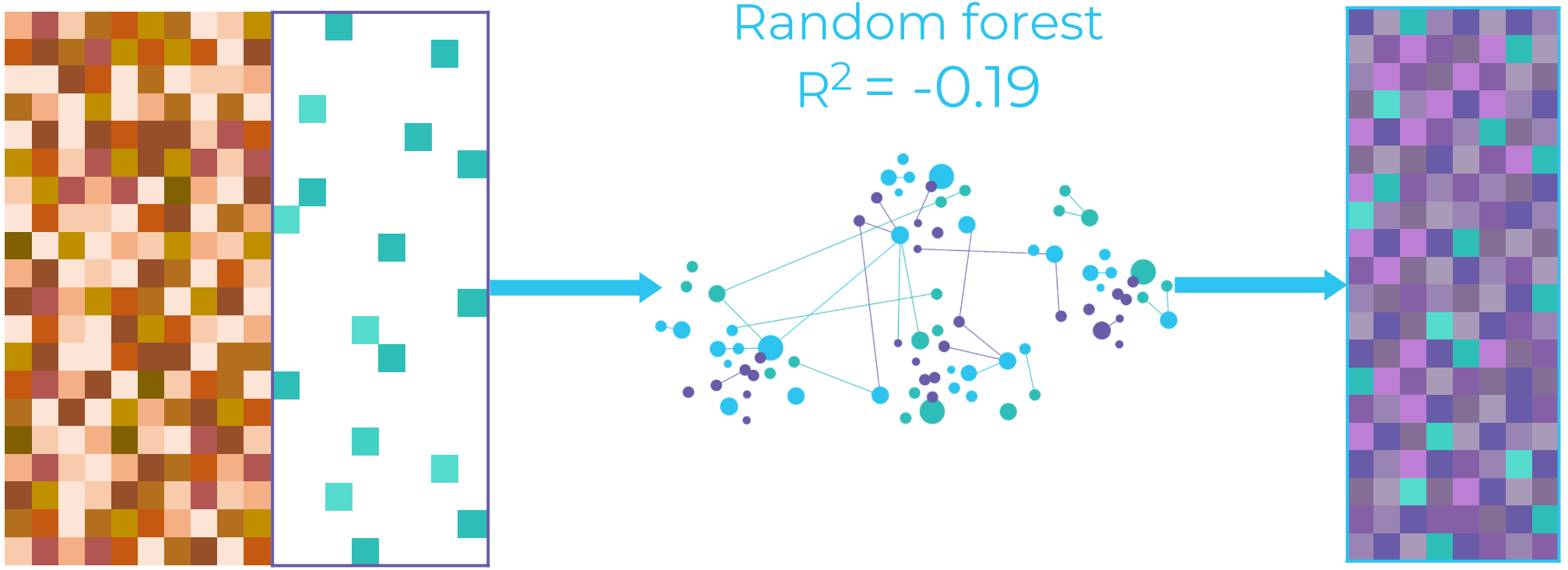
Descriptors and bioactivity values



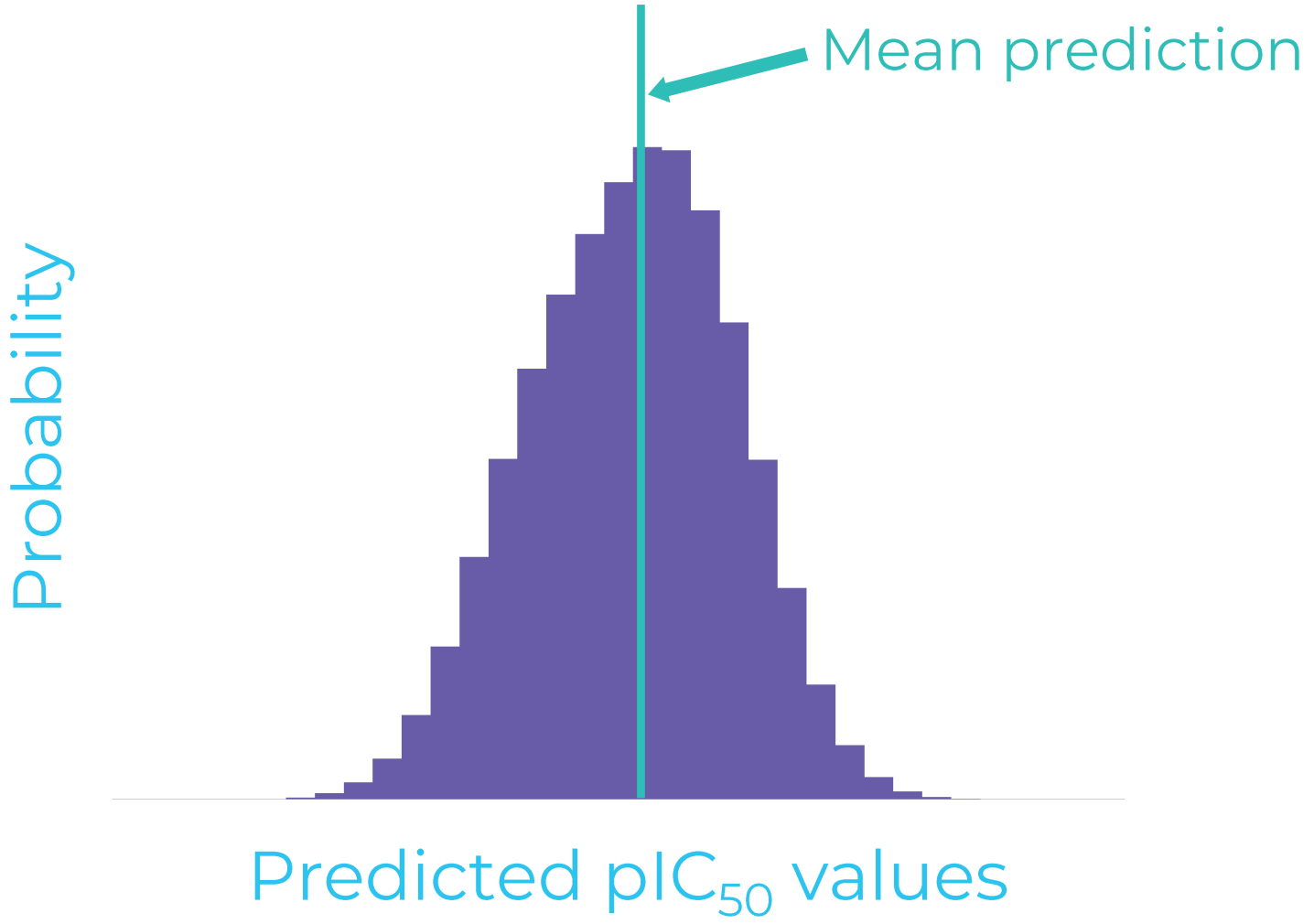
intellegens.ai

Deep learning predictions

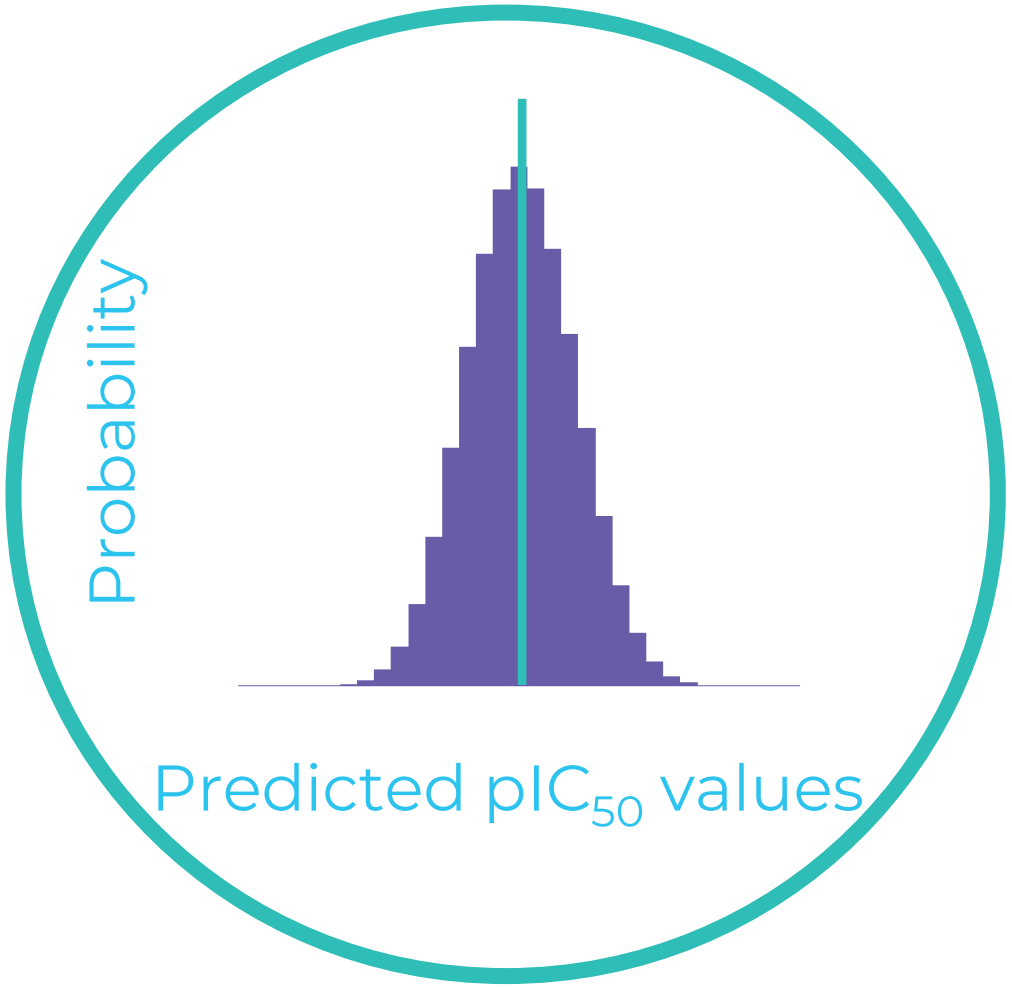
$R^2 = 0.44$

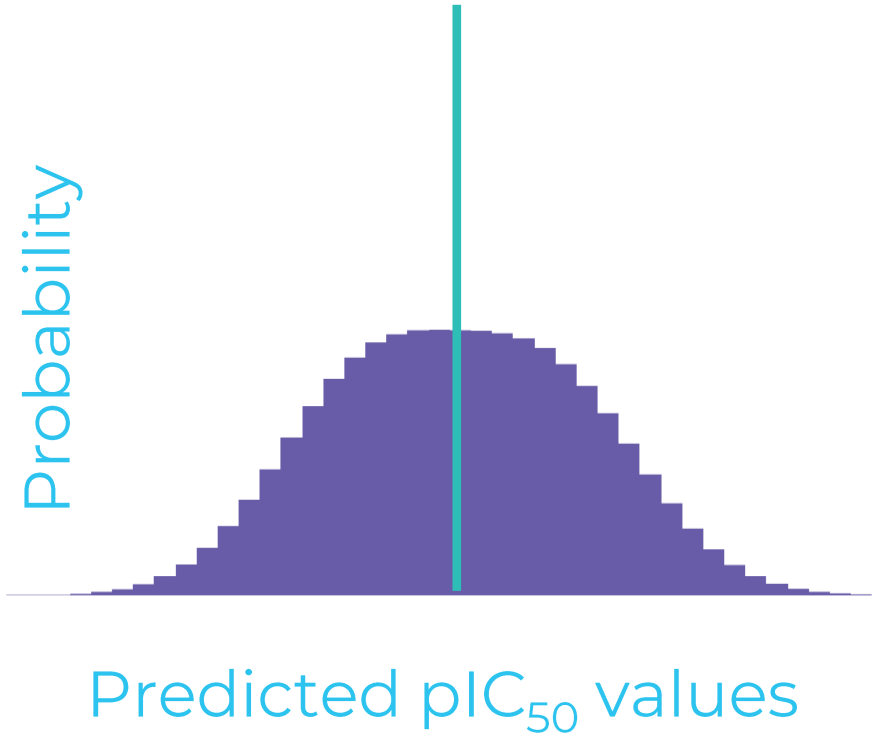


Calculate probability distribution



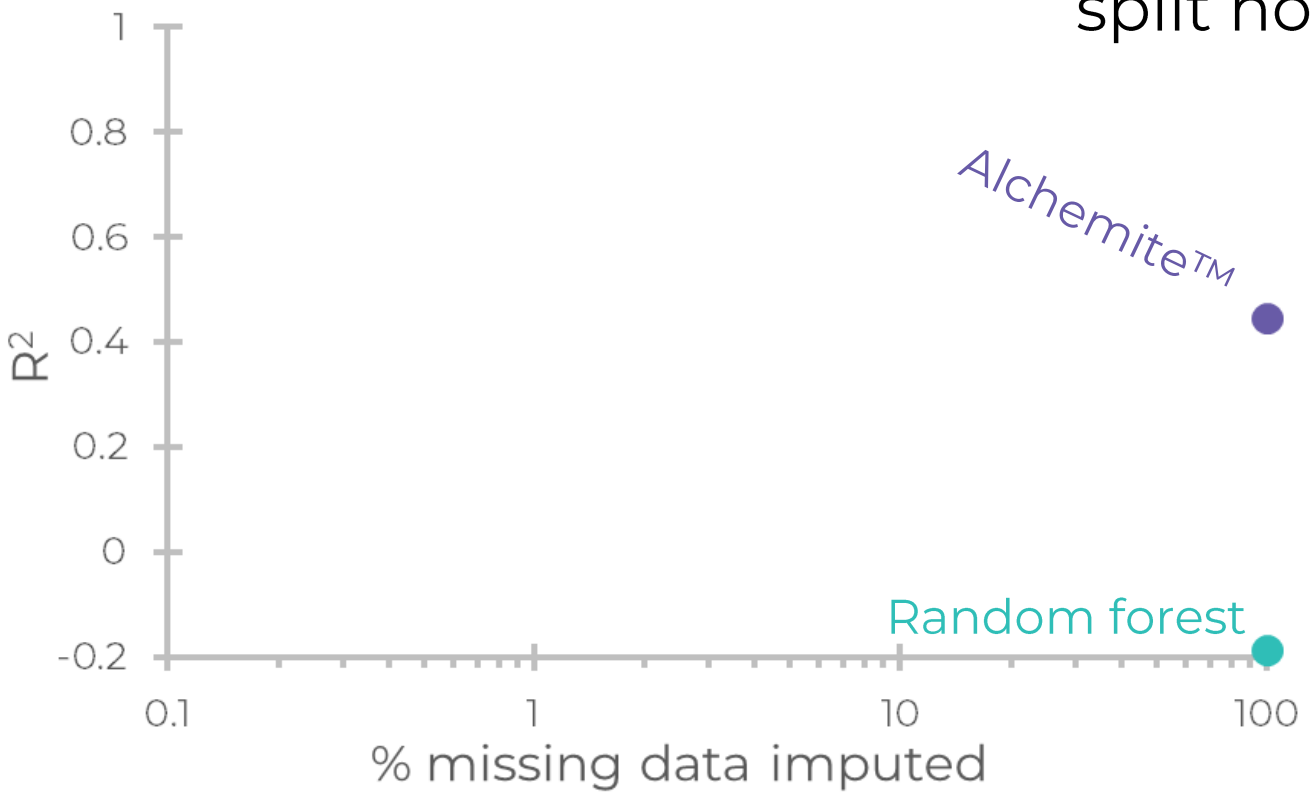
Focus on most confident predictions





Reporting only most confident predictions

Validating against realistically-split holdout set



Increasing confidence

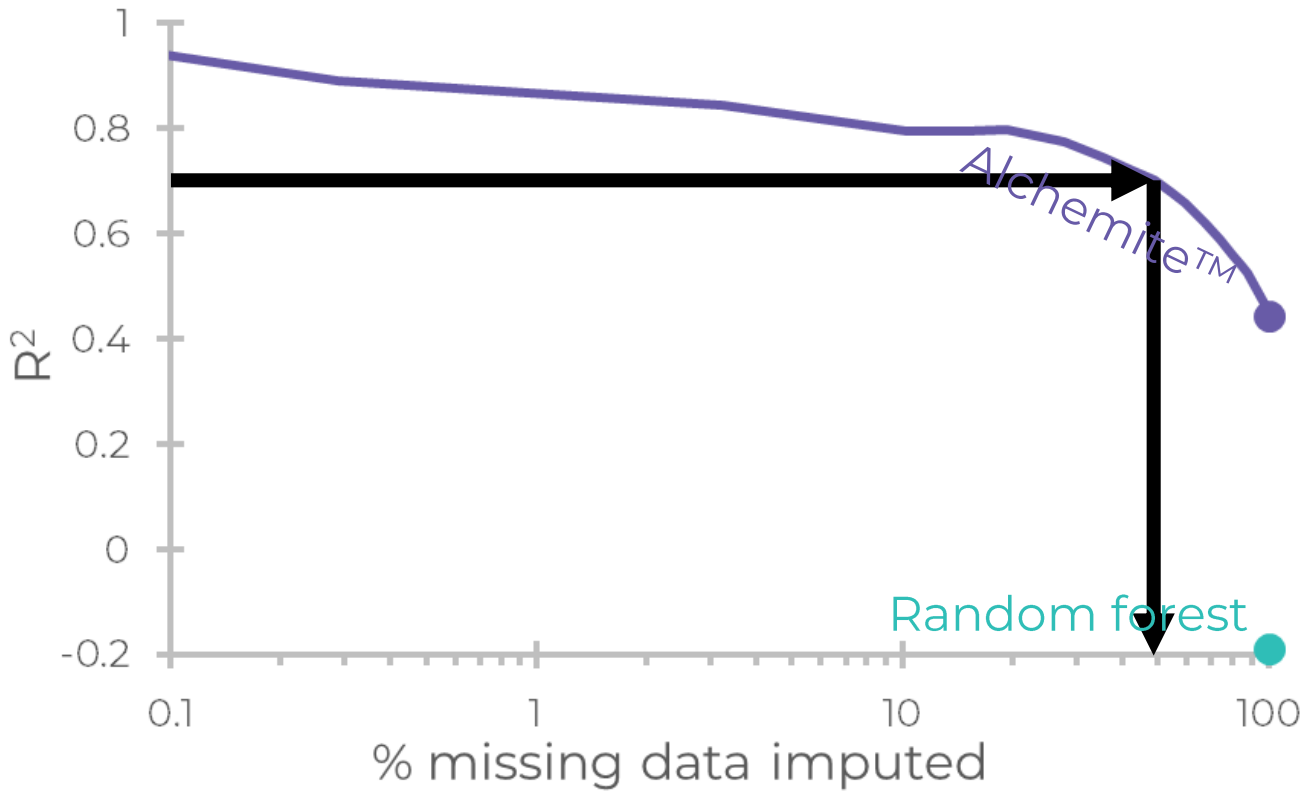
intellegens.ai

Reporting only most confident predictions

Increasing confidence

intellegens.ai

Reporting only most confident predictions



Increasing confidence

intellegens.ai

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