Today’s webinar will begin shortly

optibrium

Intellegens

Using Deep Learning to Impute Protein Activity

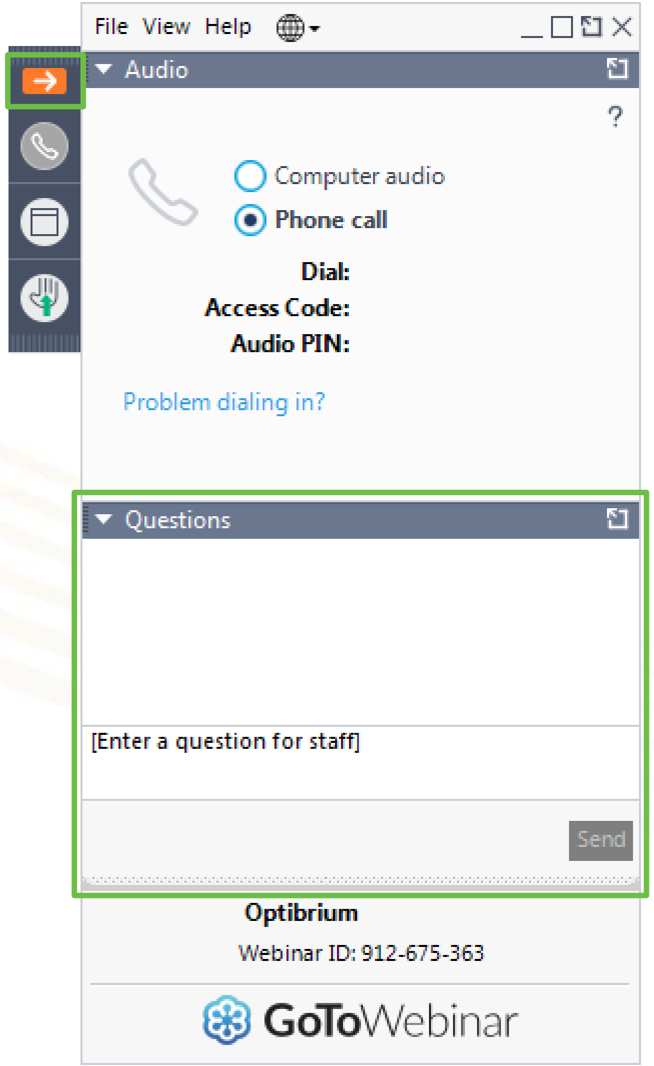
Webinar: 7th February 2019

Matt Segall – matt@optibrium.com Tom Whitehead – tom@intellegens.ai

Before We Begin…

- Thank you for joining us today
- Please feel free to ask questions at any time using the GoToWebinar "Questions" control panel
  - Questions will be answered at the end of the presentation
- You can minimise the control panel if you wish
- The presentation is being recorded and will be made available on the Optibrium Community website:

  www.optibrium.com/community



Today’s Host Speaker

Matt Segall, CEO Optibrium

Matt has a Master of Science in computation from the University of Oxford and a PhD in theoretical physics from the University of Cambridge. As Associate Director at Camitro (UK), ArQule Inc. and then Inpharmatica, he led a team developing predictive ADME models and state-of-the-art intuitive decision-support and visualization tools for drug discovery.

In January 2006, he became responsible for management of Inpharmatica's ADME business, including experimental ADME services and the StarDrop software platform. Following acquisition of Inpharmatica, Matt became Senior Director responsible for BioFocus DPI's ADMET division and in 2009 led a management buyout of the StarDrop business to found Optibrium.

Quantitative Structure-Activity Relationships

Predicting compound properties to guide design and selection

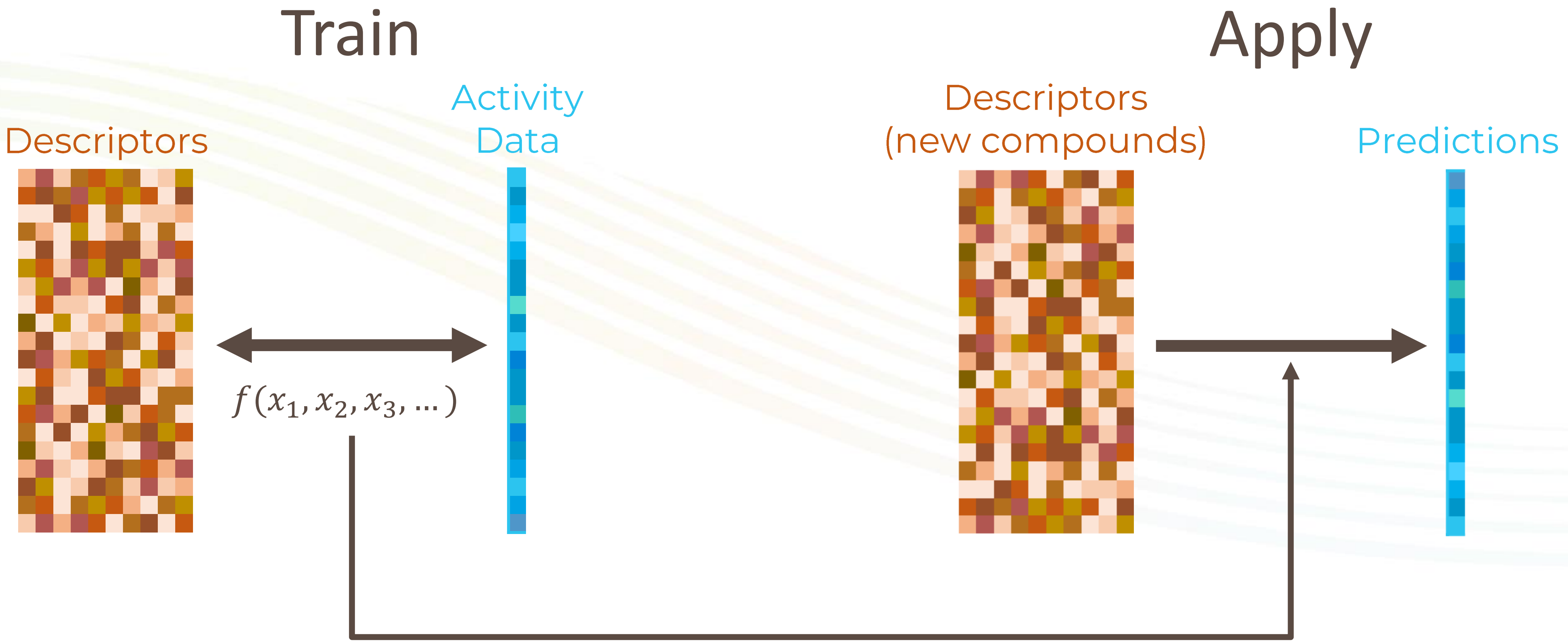
$$y = f(x_1, x_2, x_3, \dots) \pm \varepsilon$$

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…

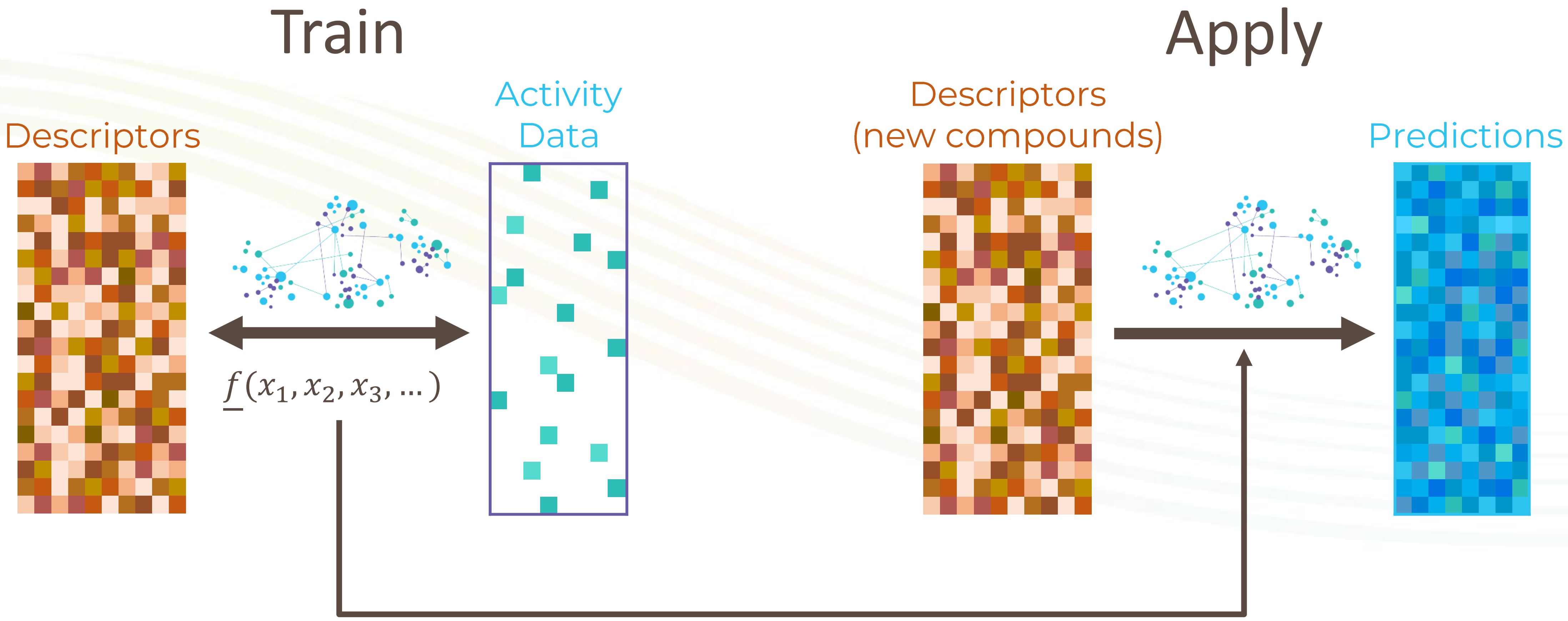
*Waldman et al. JCAMD (2015) DOI: 10.1007/s10822-015-9865-0

Quantitative Structure-Activity Relationships



Multi-Target Prediction

E.g. Deep learning

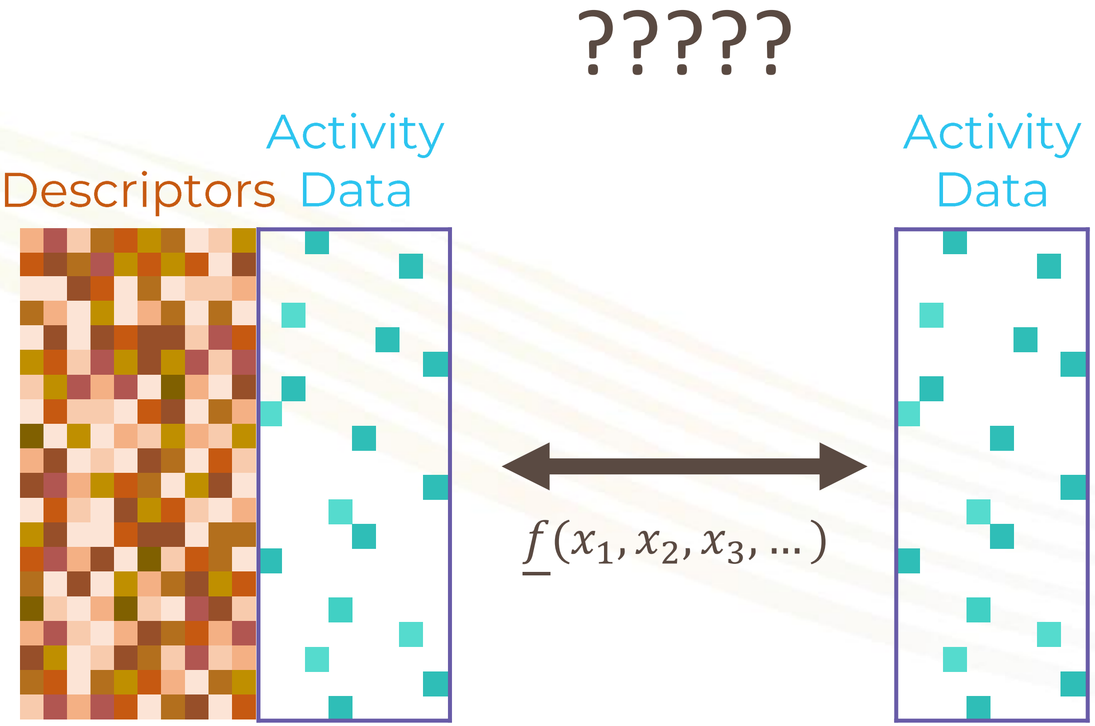


The Challenges of Applying Deep Learning

- 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

Learning From Sparse Data?



?????

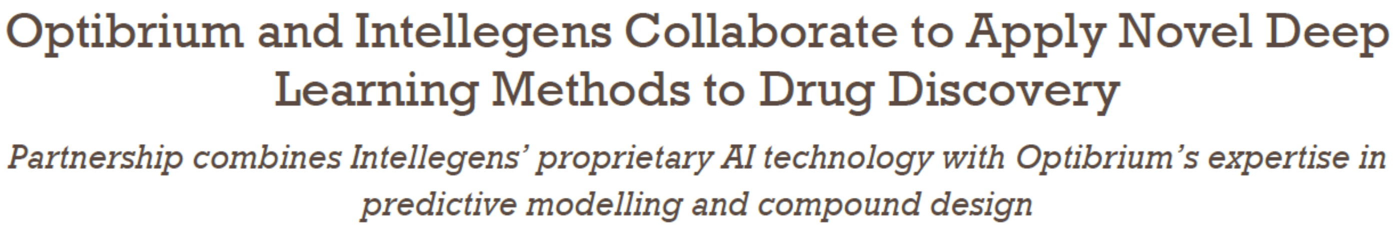
Activity Data

Descriptors

Activity Data

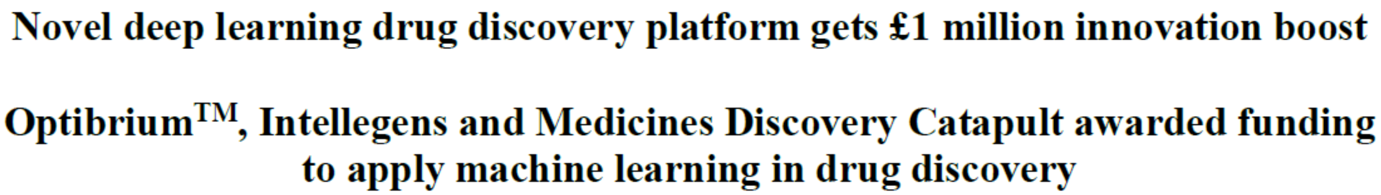
$\underline{f}(x_1, x_2, x_3, \ldots)$

Collaboration with 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



Novel deep learning drug discovery platform gets £1 million innovation boost

Optibrium^{TM}, Intellegens and Medicines Discovery Catapult awarded funding to apply machine learning in drug discovery

Today's Guest Speaker

Tom Whitehead, Head of Machine Learning, Intellegens

Dr Tom Whitehead is head of machine learning at Intellegens, a deep learning startup company based in Cambridge, UK. Intellegens focusses on handling sparse, noisy, experimental data, and Tom is leading the application of Intellegens' unique tools to drug discovery.

Tom has a PhD in theoretical physics from the University of Cambridge, and now focusses on the development and utilisation of deep learning methods for difficult, high-value data problems.

Using deep learning to impute protein activity

Intellegens

Dr Tom Whitehead

Intellegens

Novel **deep learning** architecture for handling **sparse and noisy** data

Dr Tom Whitehead

Head of machine learning, leading the application of deep learning technology to **drug discovery**

intellegens.ai

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

Inputs

Outputs

Alchemite™ deep learning

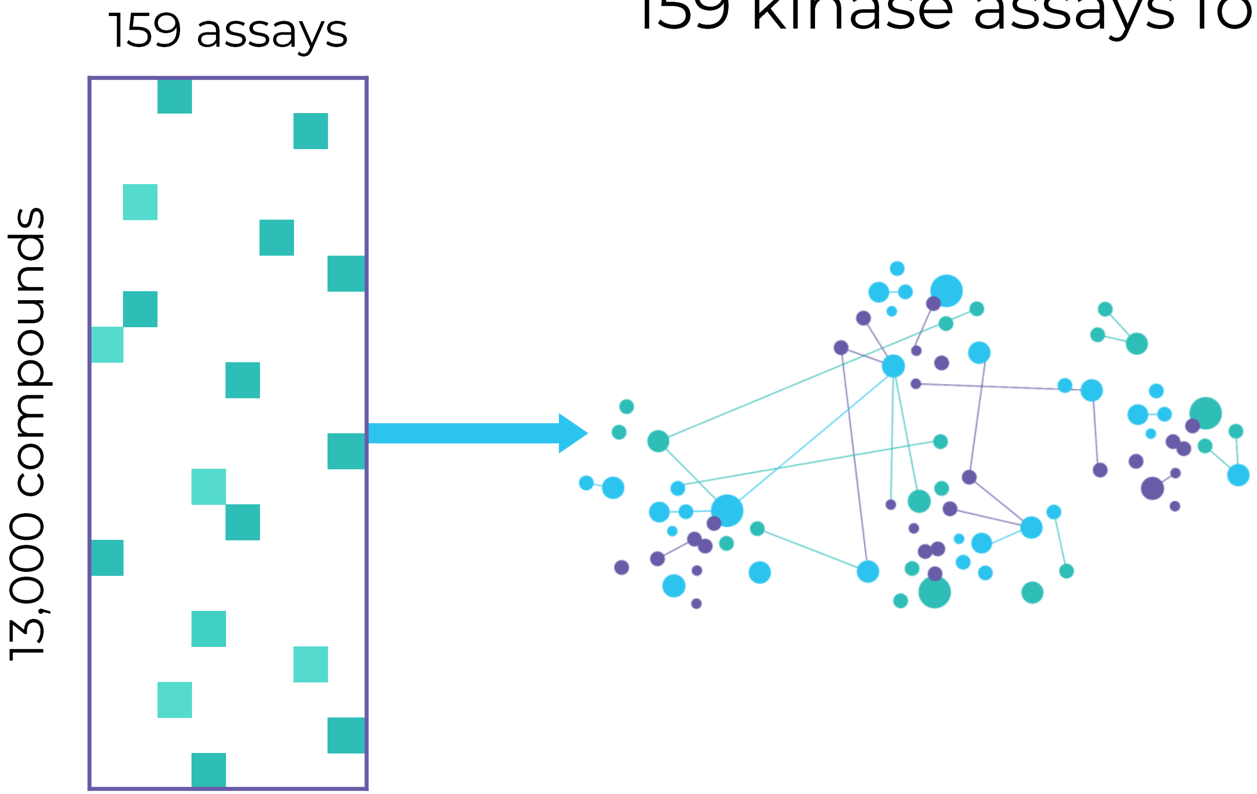
Inputs

Outputs

Novartis dataset to benchmark machine learning

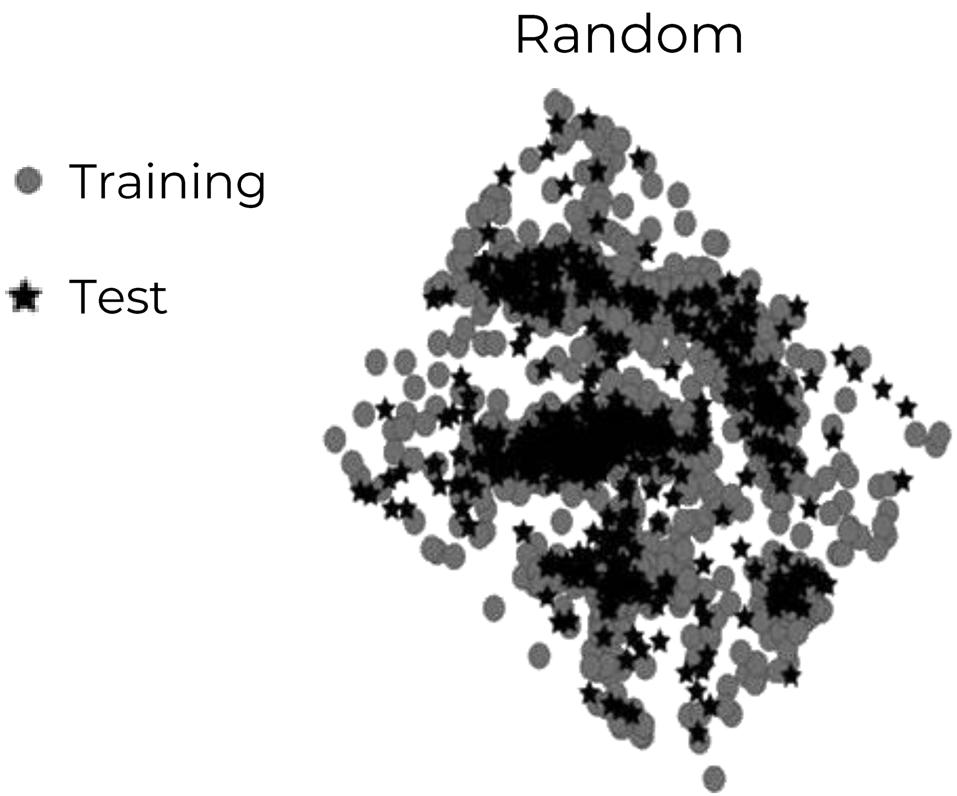
159 kinase assays for 13,000 compounds

Data **5%** complete



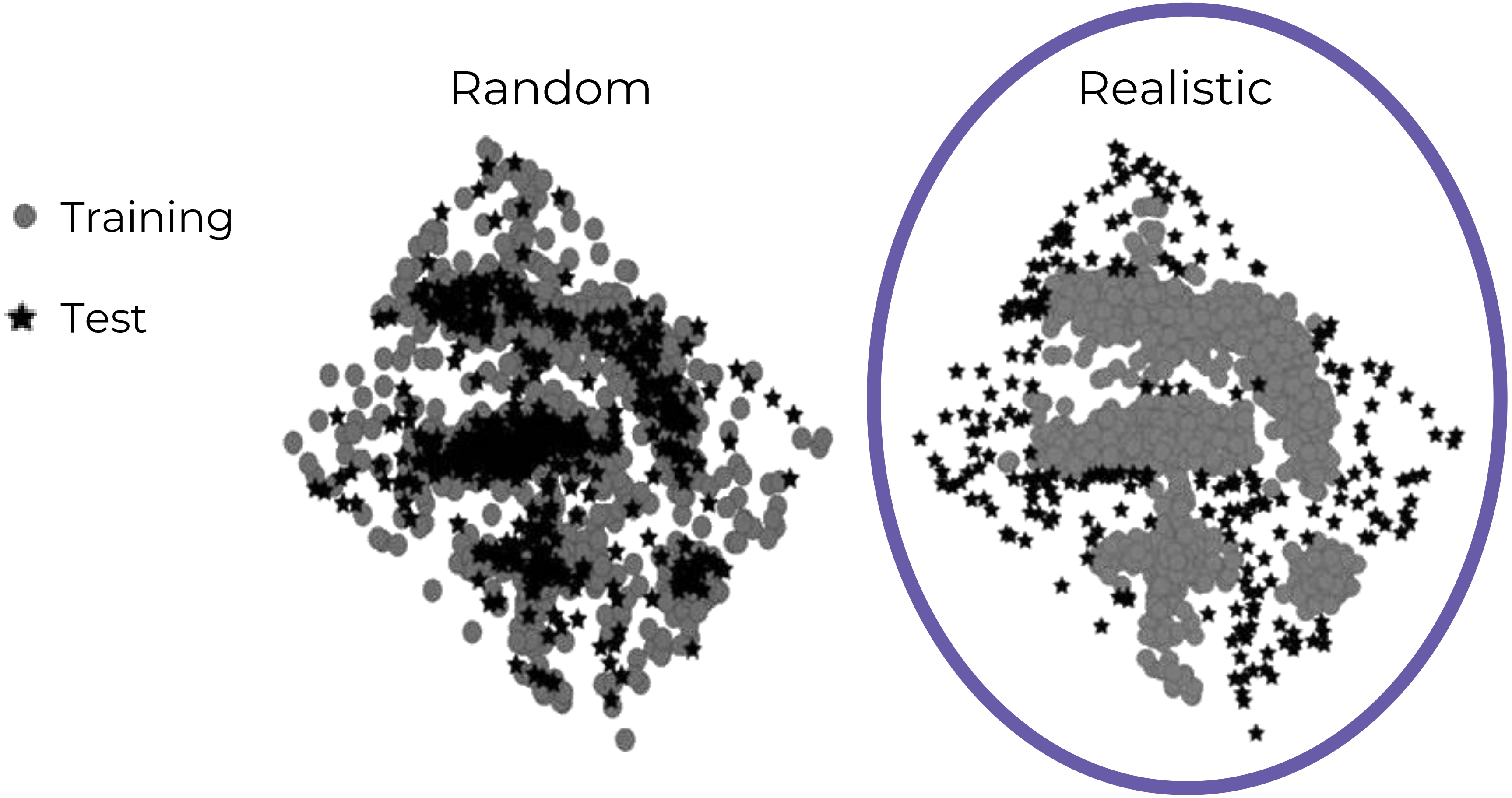
Data from ChEMBL
Martin, Polyakov, Tian, and Perez, J. Chem. Inf. Model. 57, 2077 (2017)

Novartis dataset distribution



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

Novartis dataset is realistically distributed



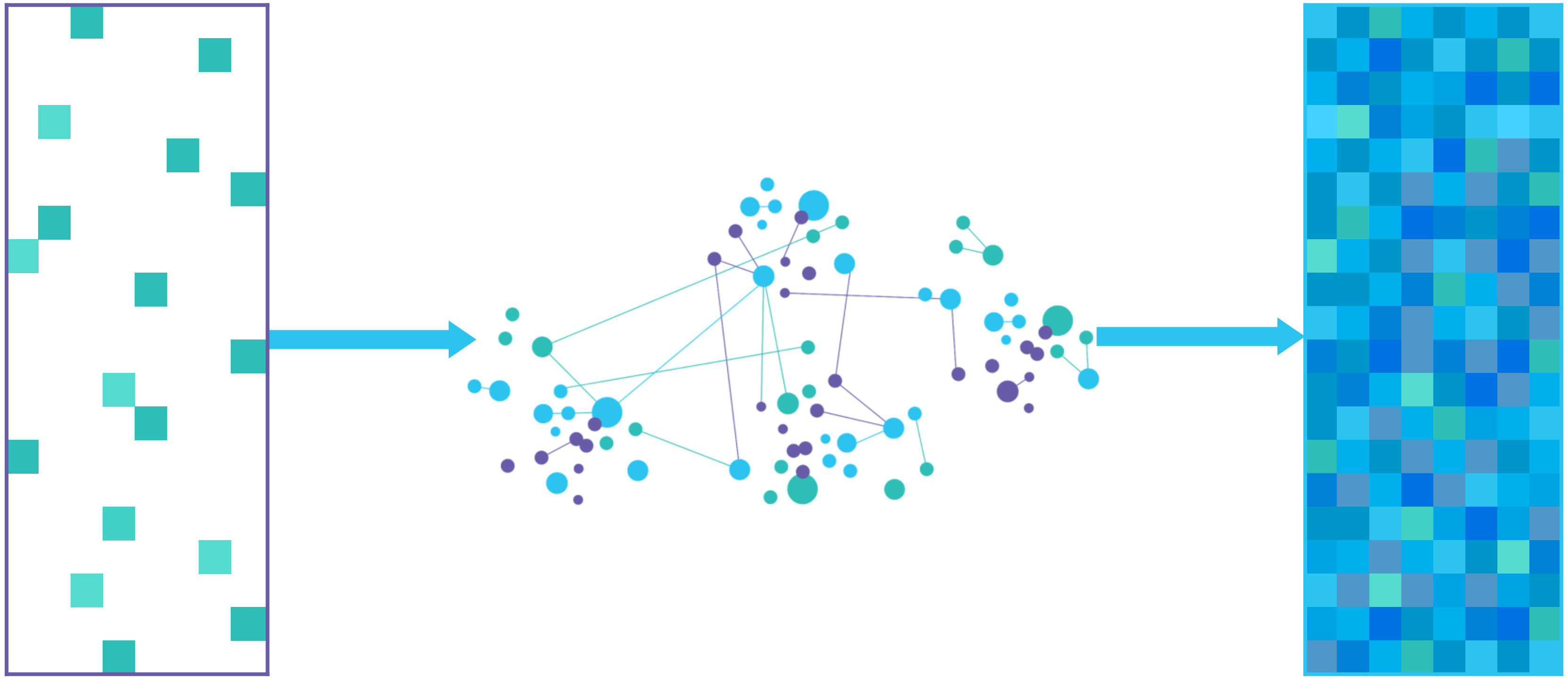
Data from ChEMBL
Martin, Polyakov, Tian, and Perez, J. Chem. Inf. Model. 57, 2077 (2017)

Accuracy metric

Coefficient of determination, R^2

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

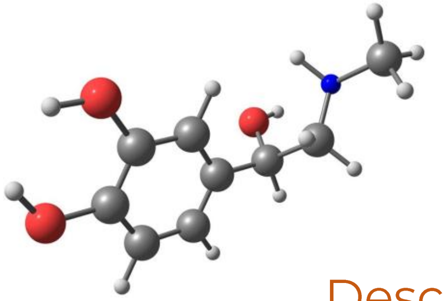
Aim: impute missing assay values

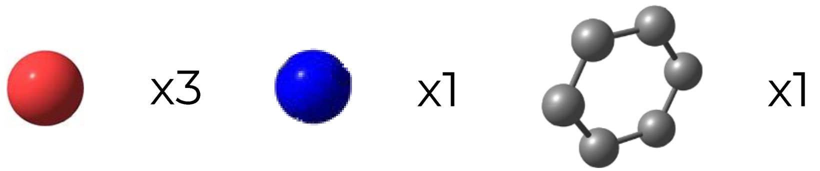


Validate against
realistically-split holdout set

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

Random forest regression

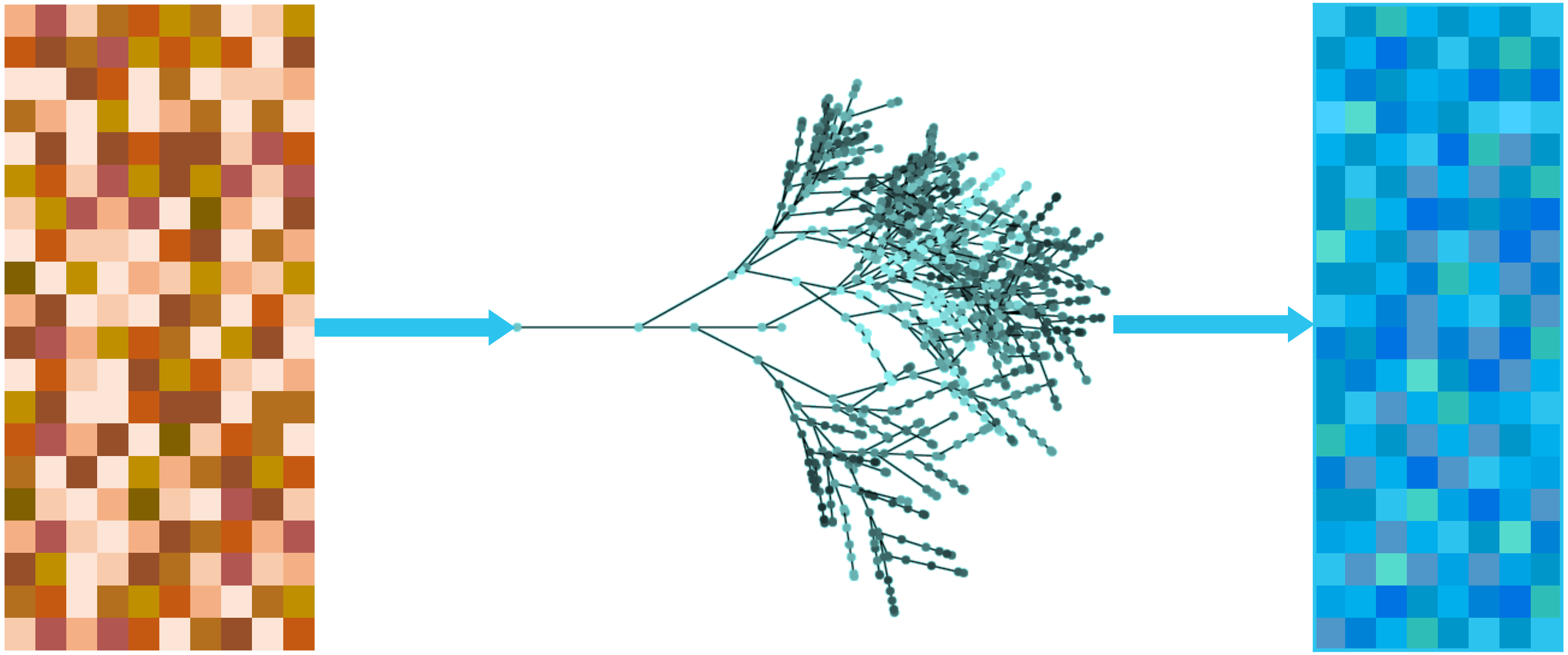




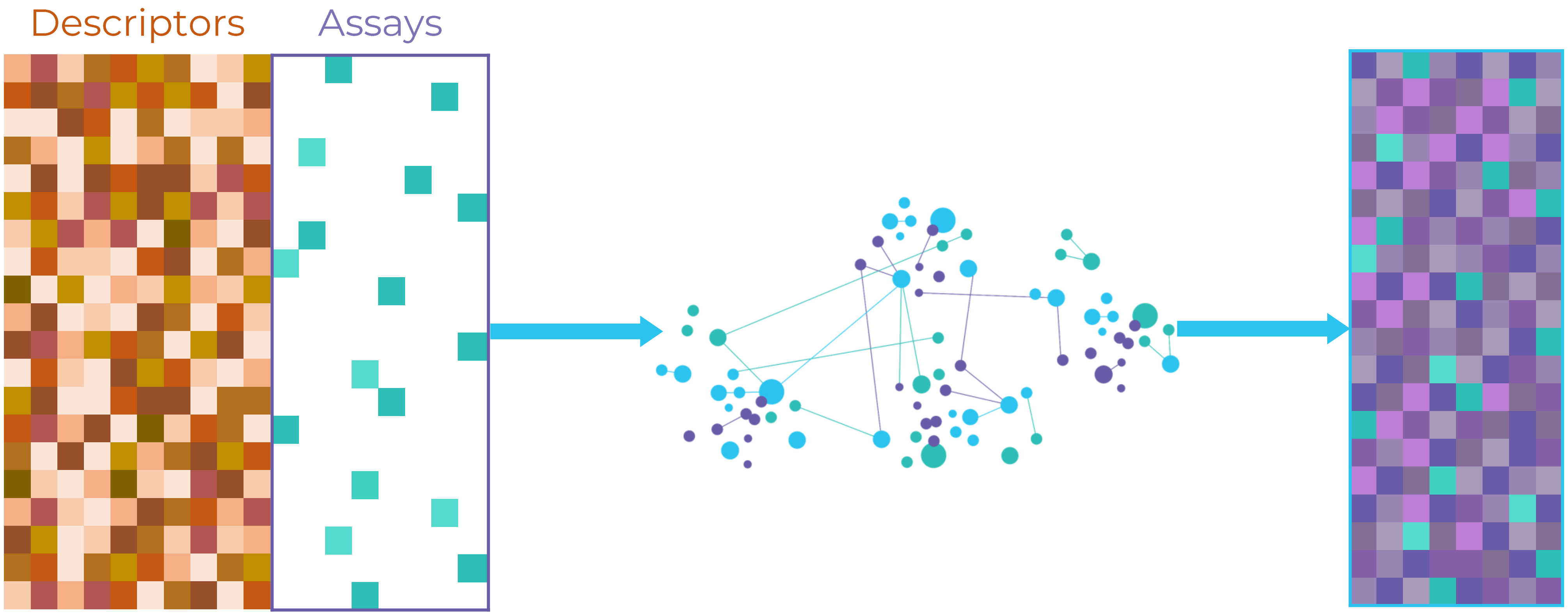
Molecular weight = 183 Da

$R^2 = -0.19$

Descriptors

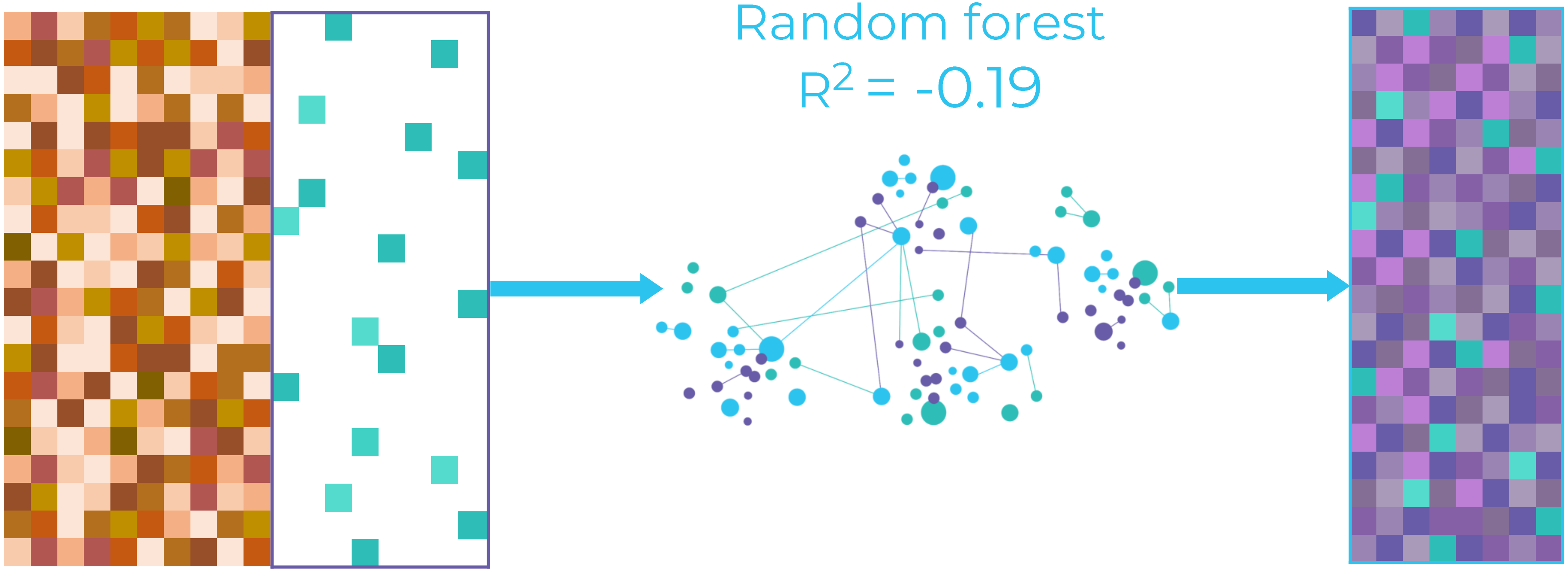


Descriptors and bioactivity values



Deep learning predictions

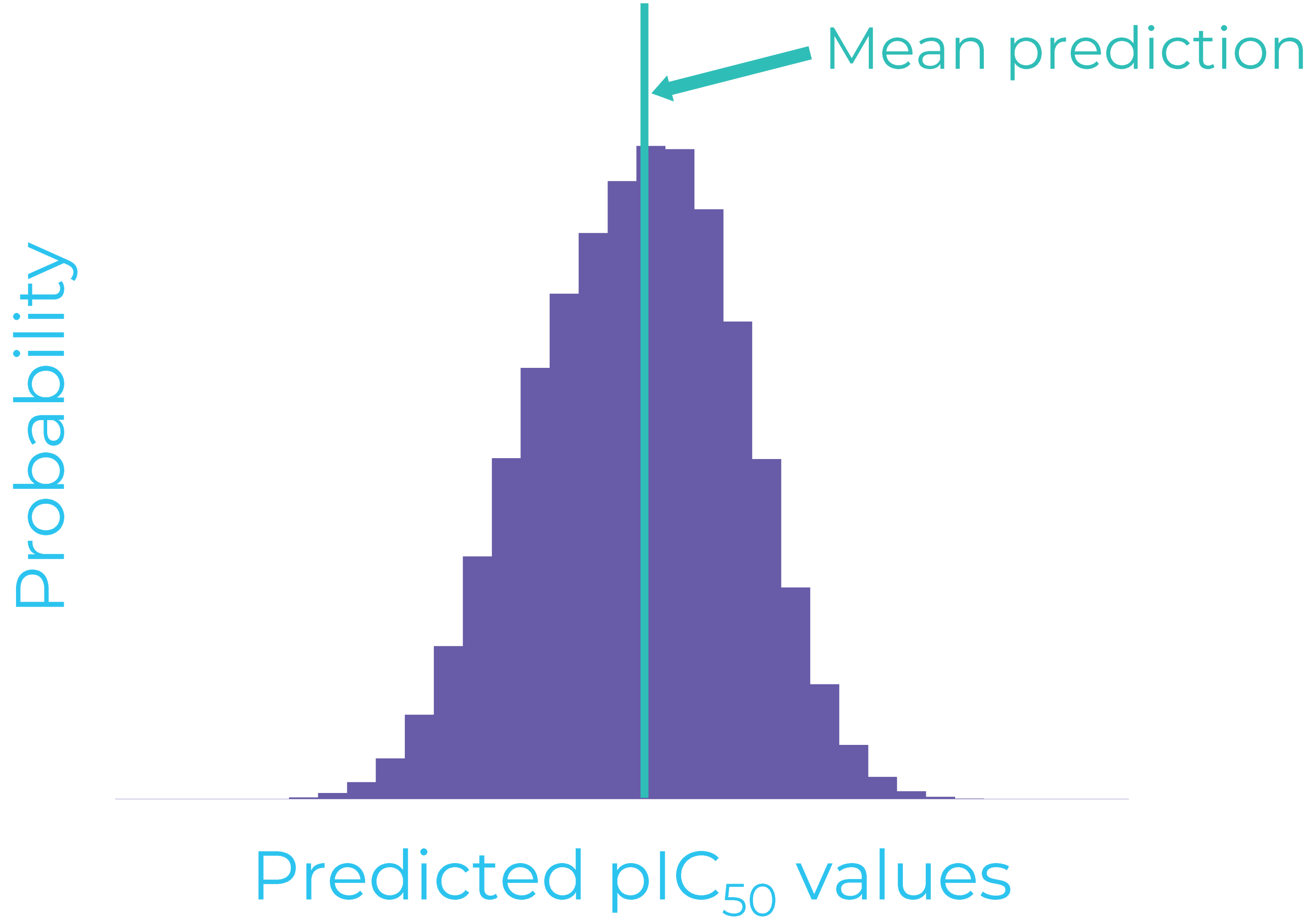
$R^2 = 0.44$



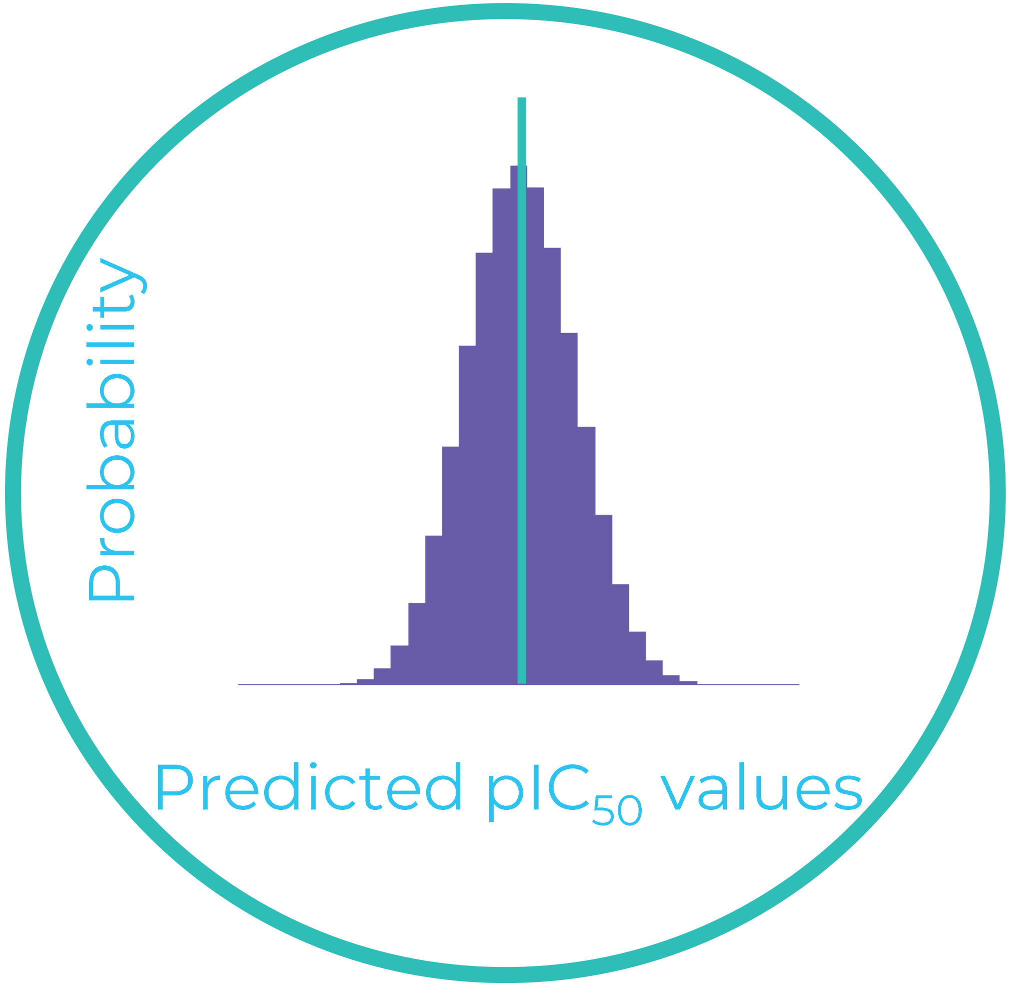
Random forest

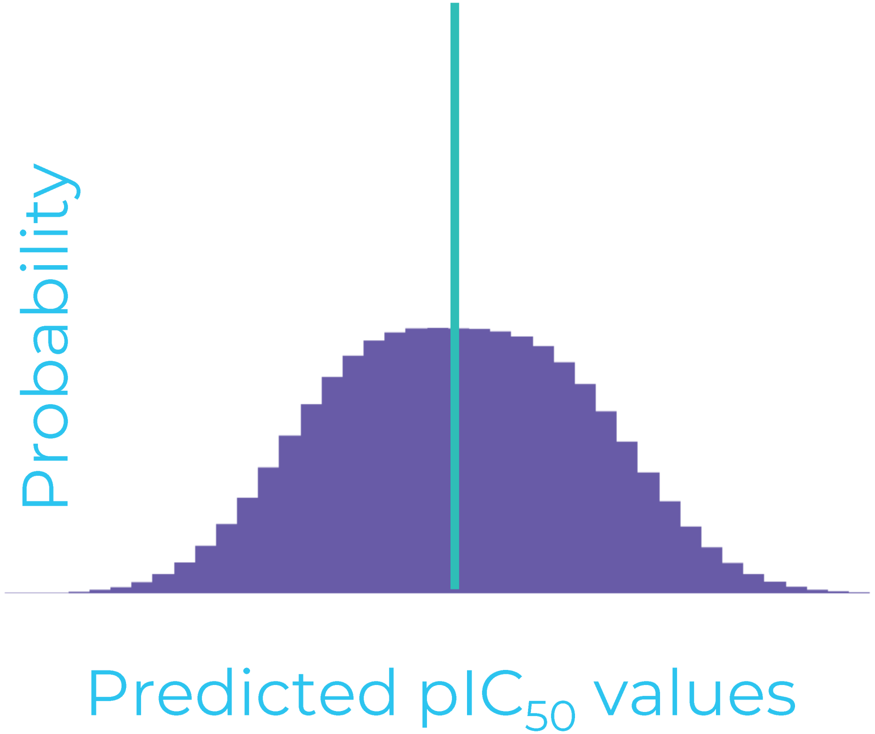
$R^2 = -0.19$

Calculate probability distribution

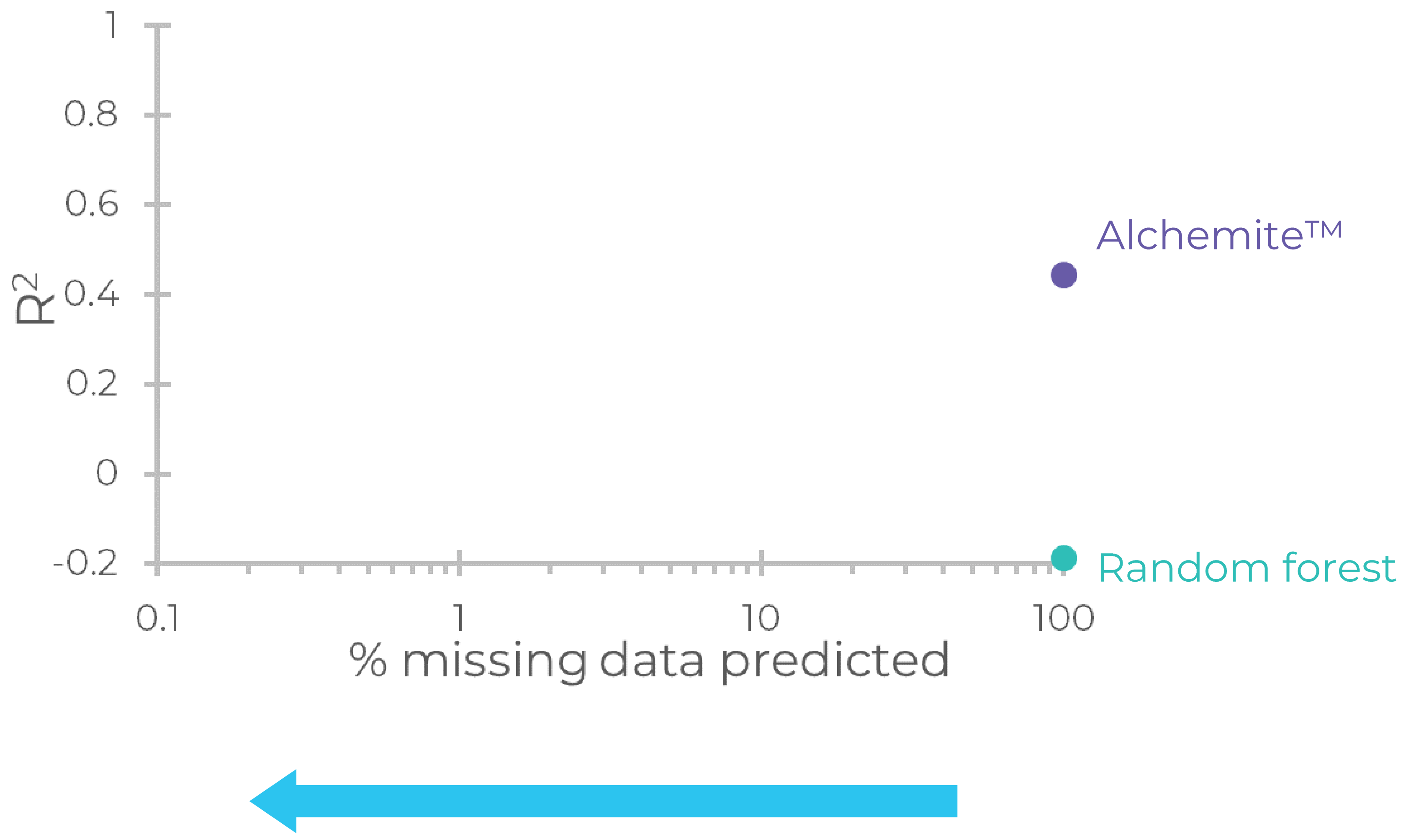


Focus on most confident predictions





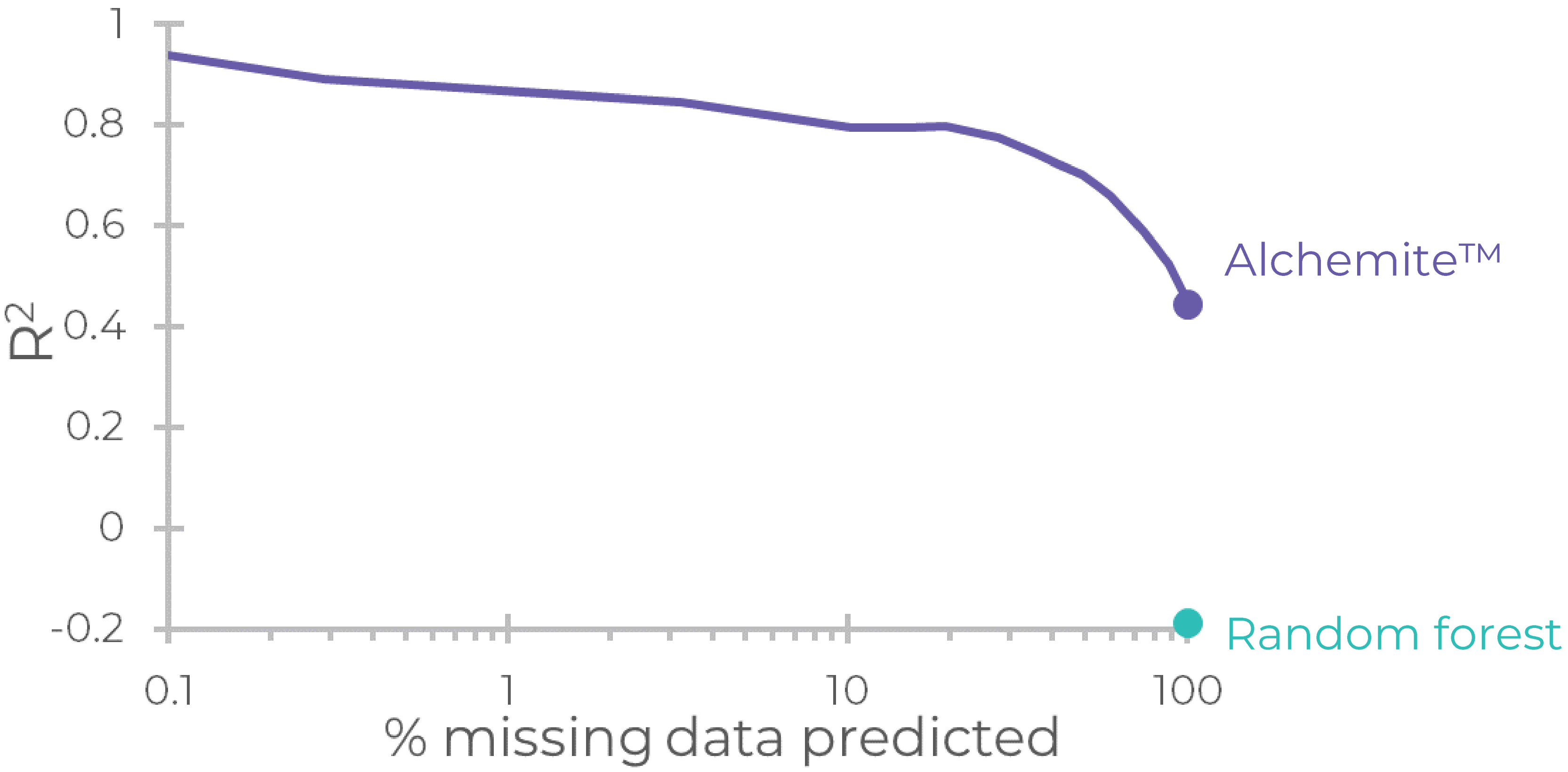
Reporting only most confident predictions



Increasing confidence

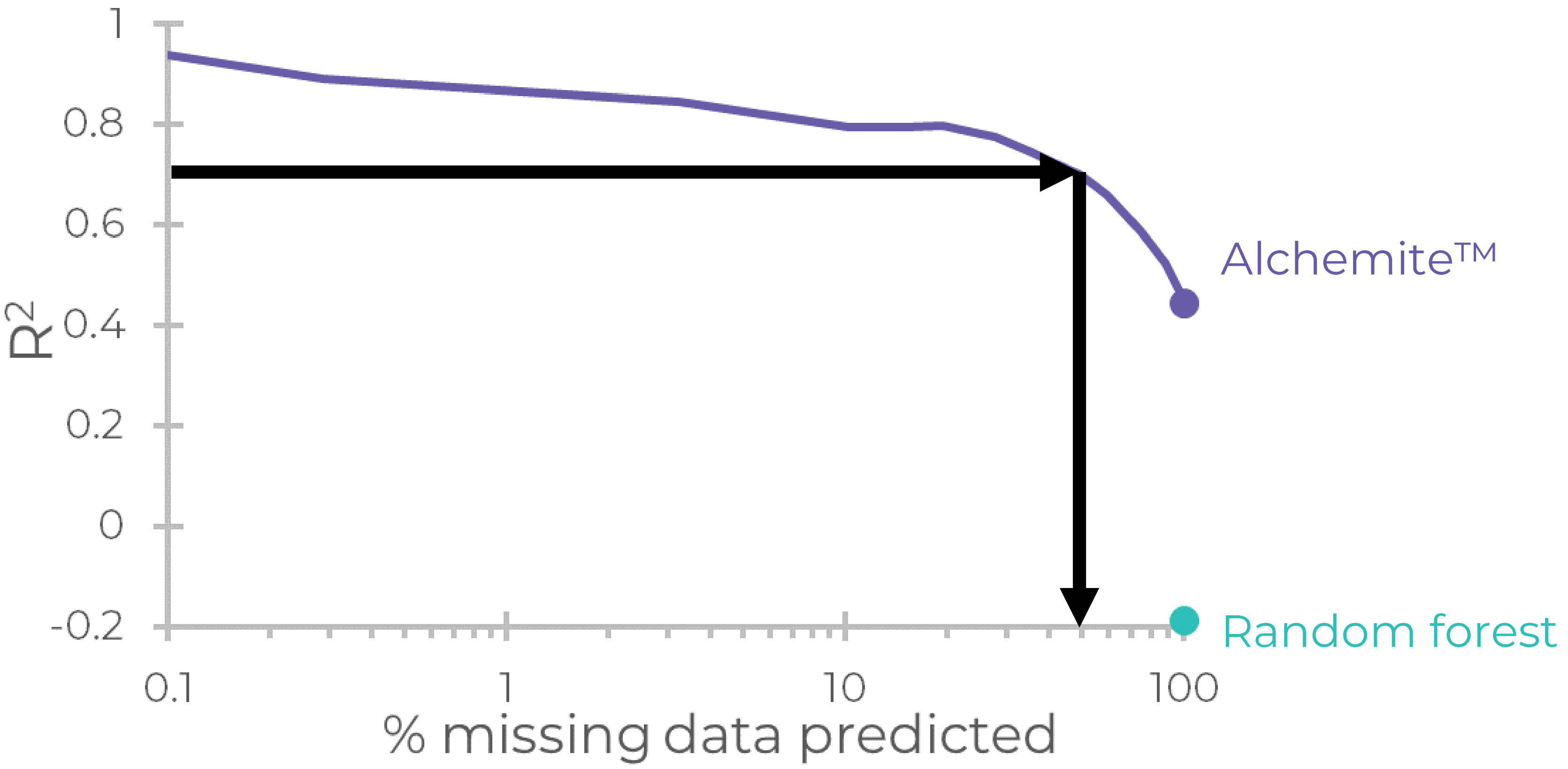
intellegens.ai

Reporting only most confident predictions



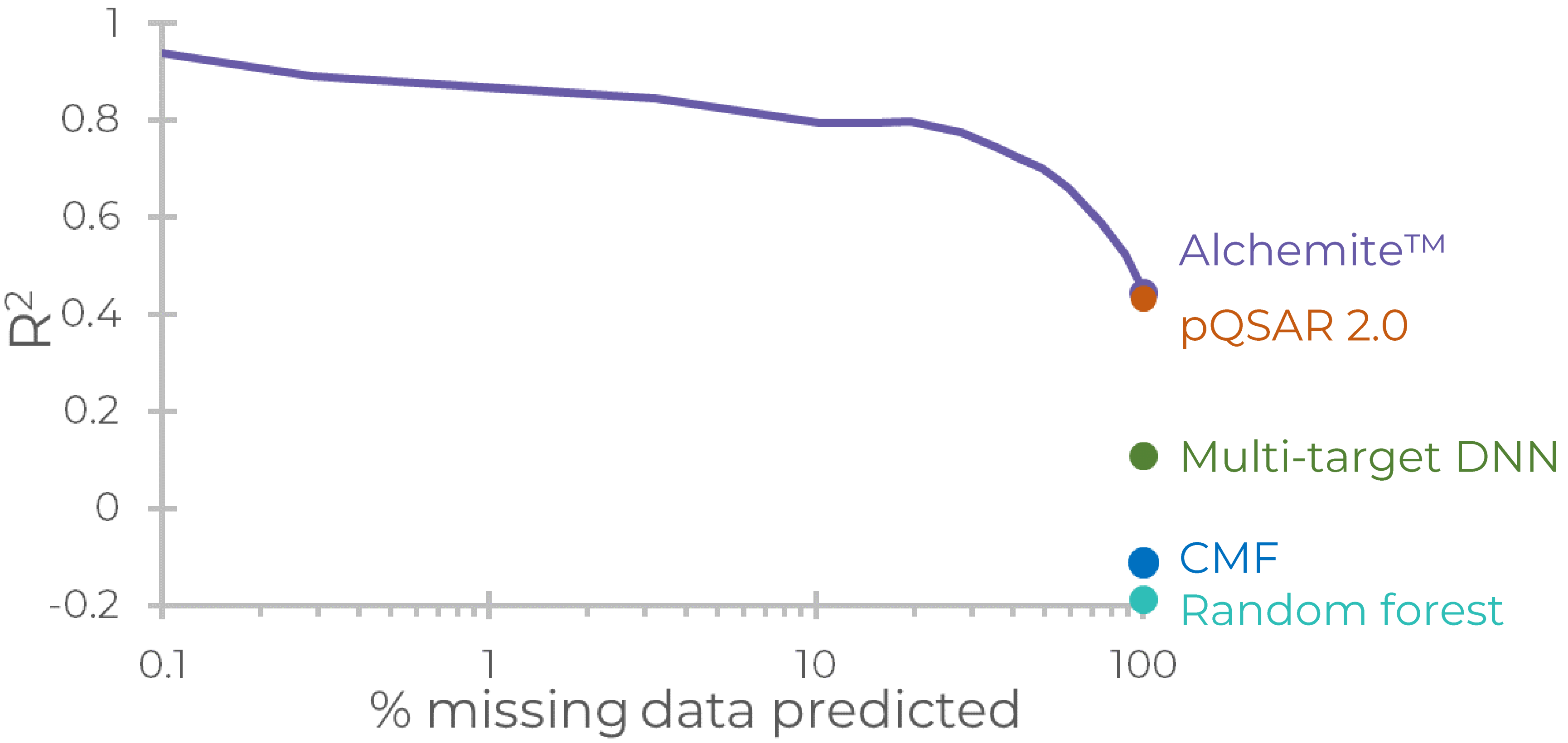
Increasing confidence

Reporting only most confident predictions



intellegens.ai

Comparison to other methods



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

Using Deep Learning to Impute Protein Activity

Poll …

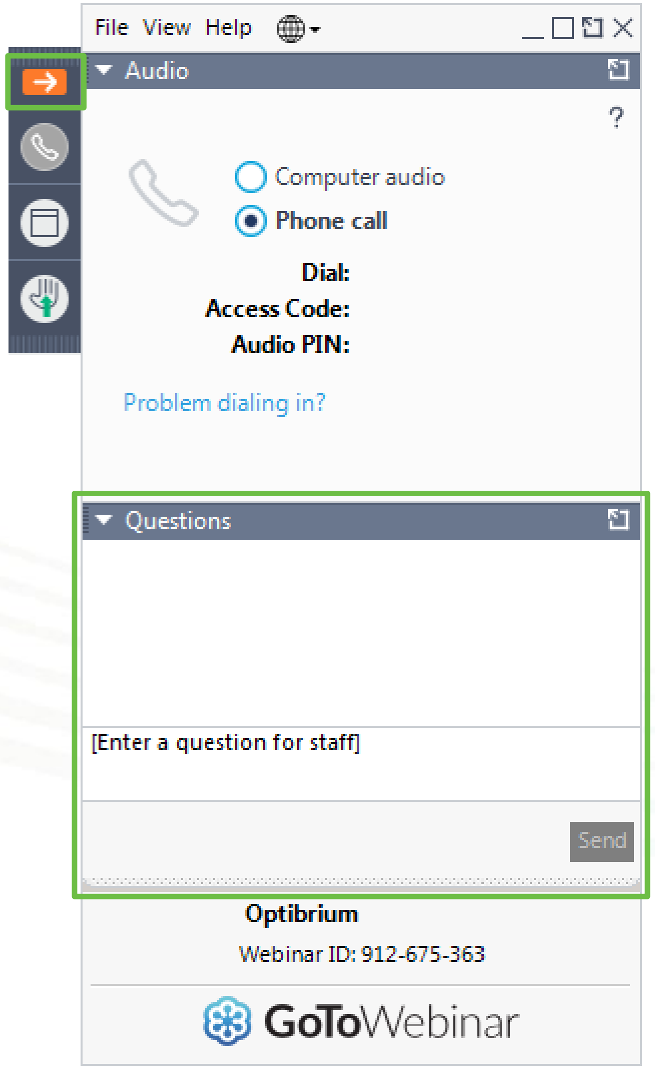
(tba)

Using Deep Learning to Impute Protein Activity

Questions

We will now respond to questions posted during the webinar

You may still ask questions using the GoToWebinar “Questions” control panel



○ Computer audio
● Phone call

Dial:
Access Code:
Audio PIN:

[Enter a question for staff]

Optibrium
Webinar ID: 912-675-363

Using Deep Learning to Impute Protein Activity

Thank you for attending today’s webinar

For further information please contact:

info@optibrium.com

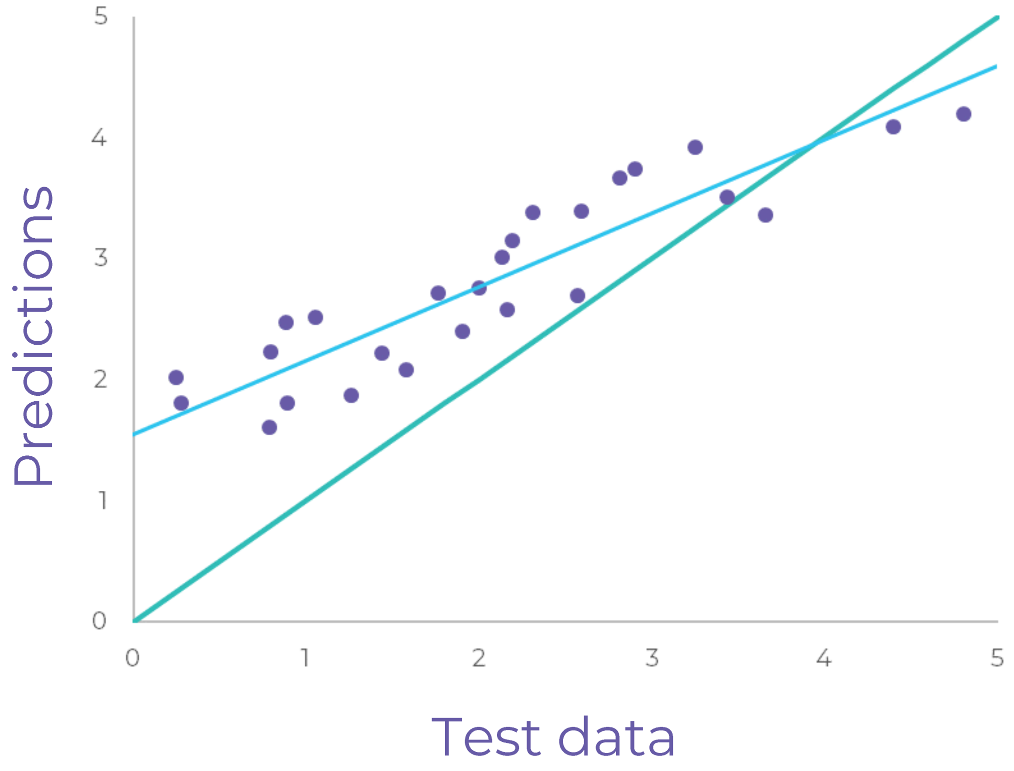
A recording of the presentation will be made available on the Optibrium Community website:

www.optibrium.com/community

optibrium

Intellegens

Coefficient of determination, R^2



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

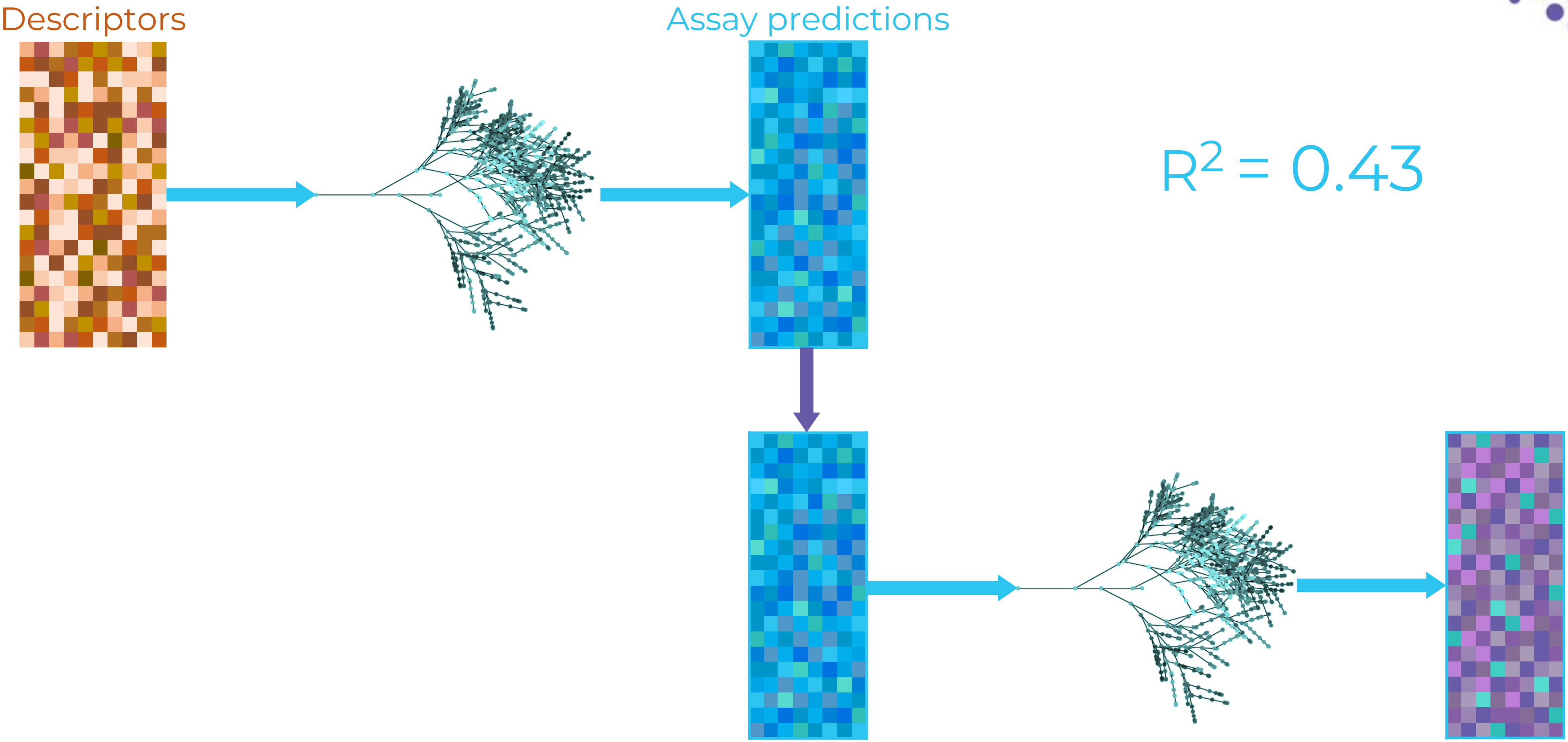
$r^2 = 0.94$

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

$R^2 = 0.77$

intellegens.ai

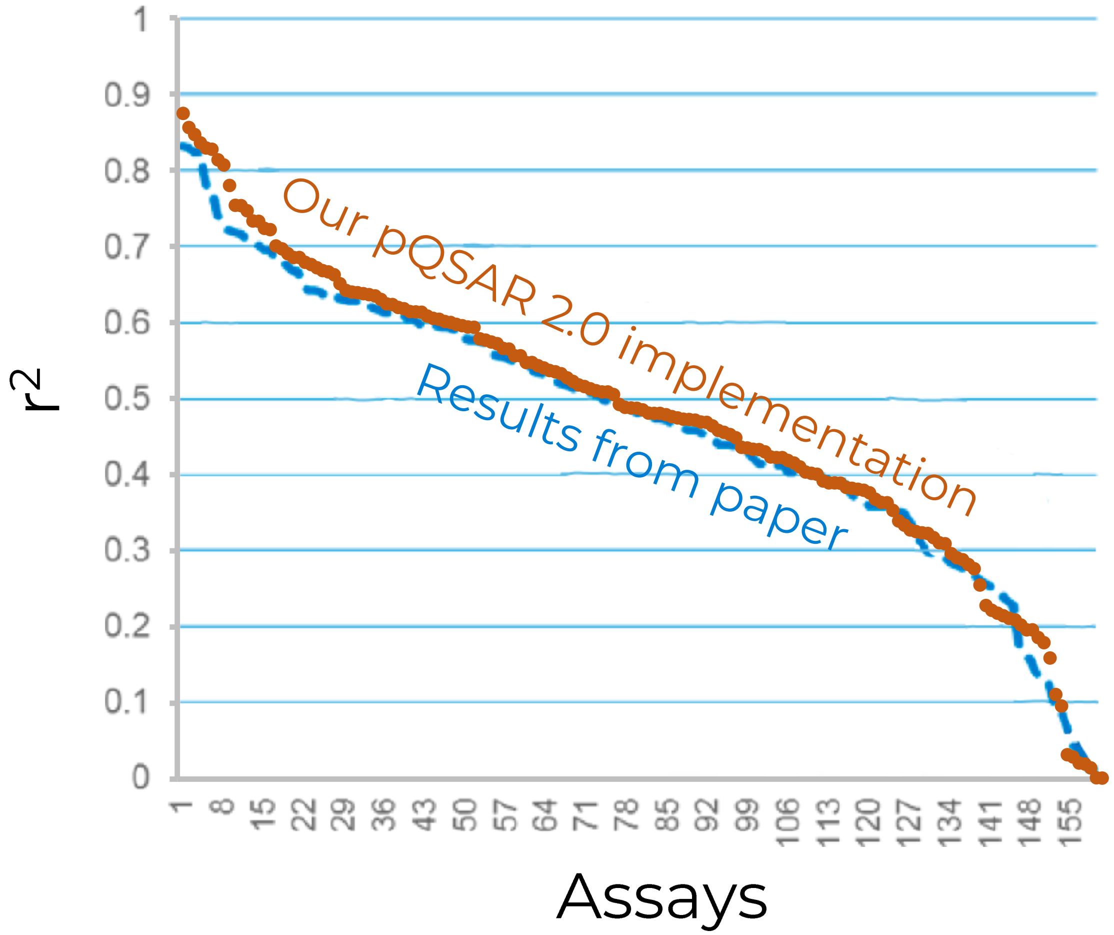
pQSAR 2.0 method



$R^2 = 0.43$

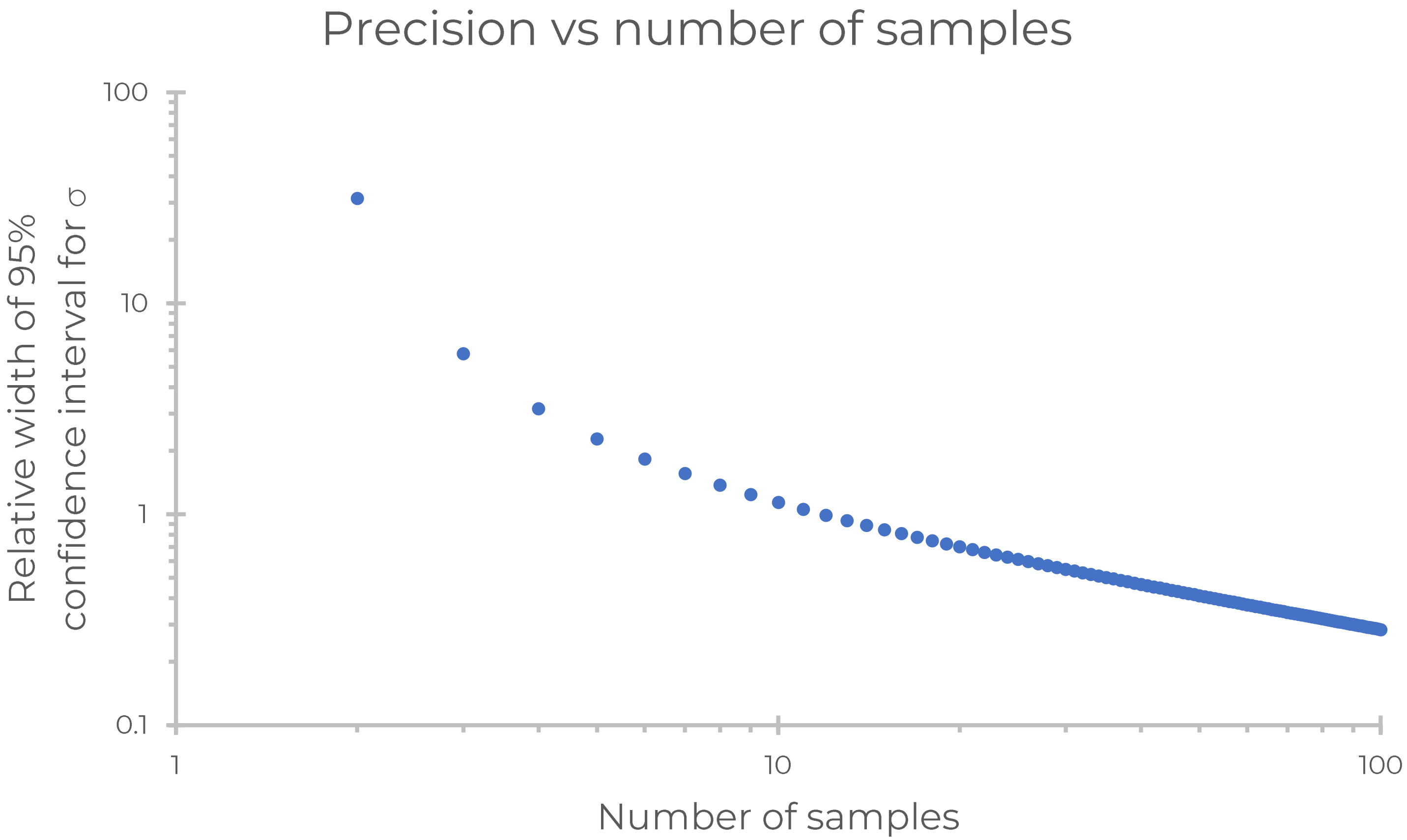
Martin, Polyakov, Tian, and Perez, J. Chem. Inf. Model. 57, 2077 (2017)

pQSAR 2.0 results

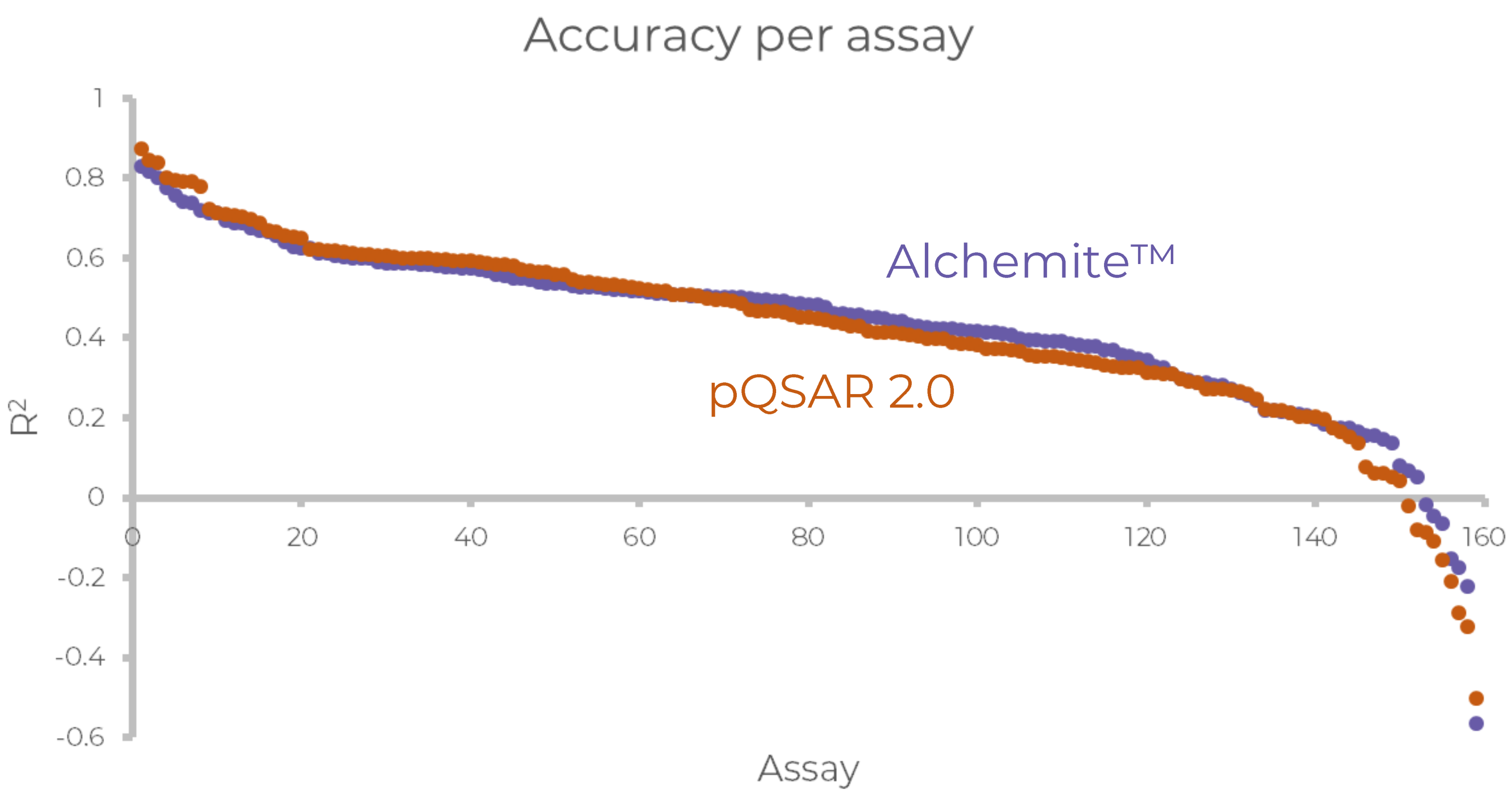


Martin, Polyakov, Tian, and Perez, J. Chem. Inf. Model. 57, 2077 (2017)

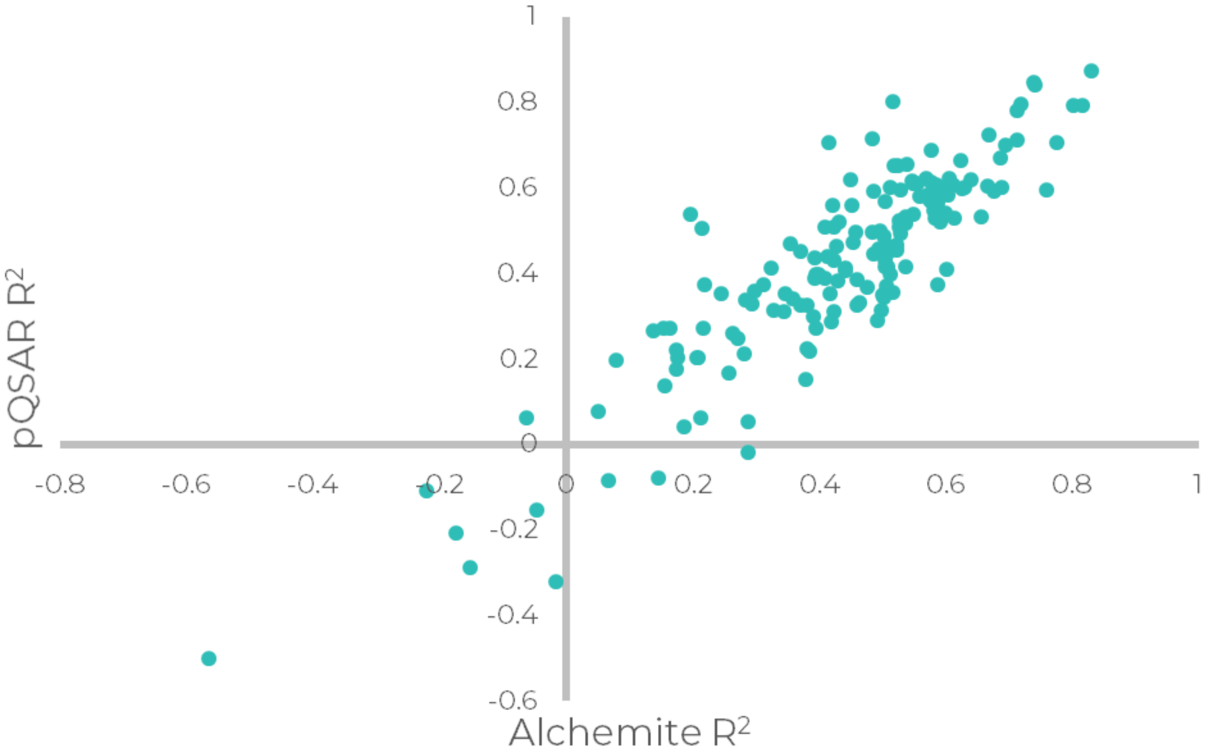
Samples from probability distribution



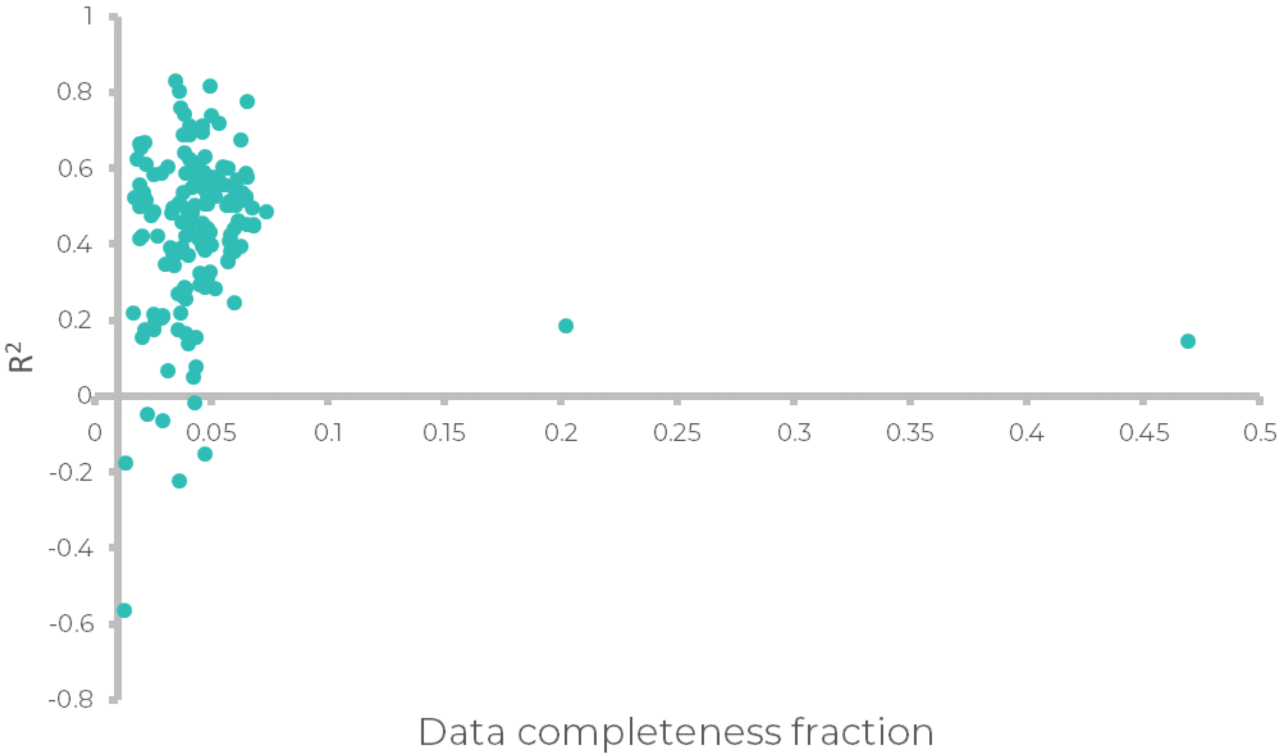
Accuracy per assay



Accuracy on assays

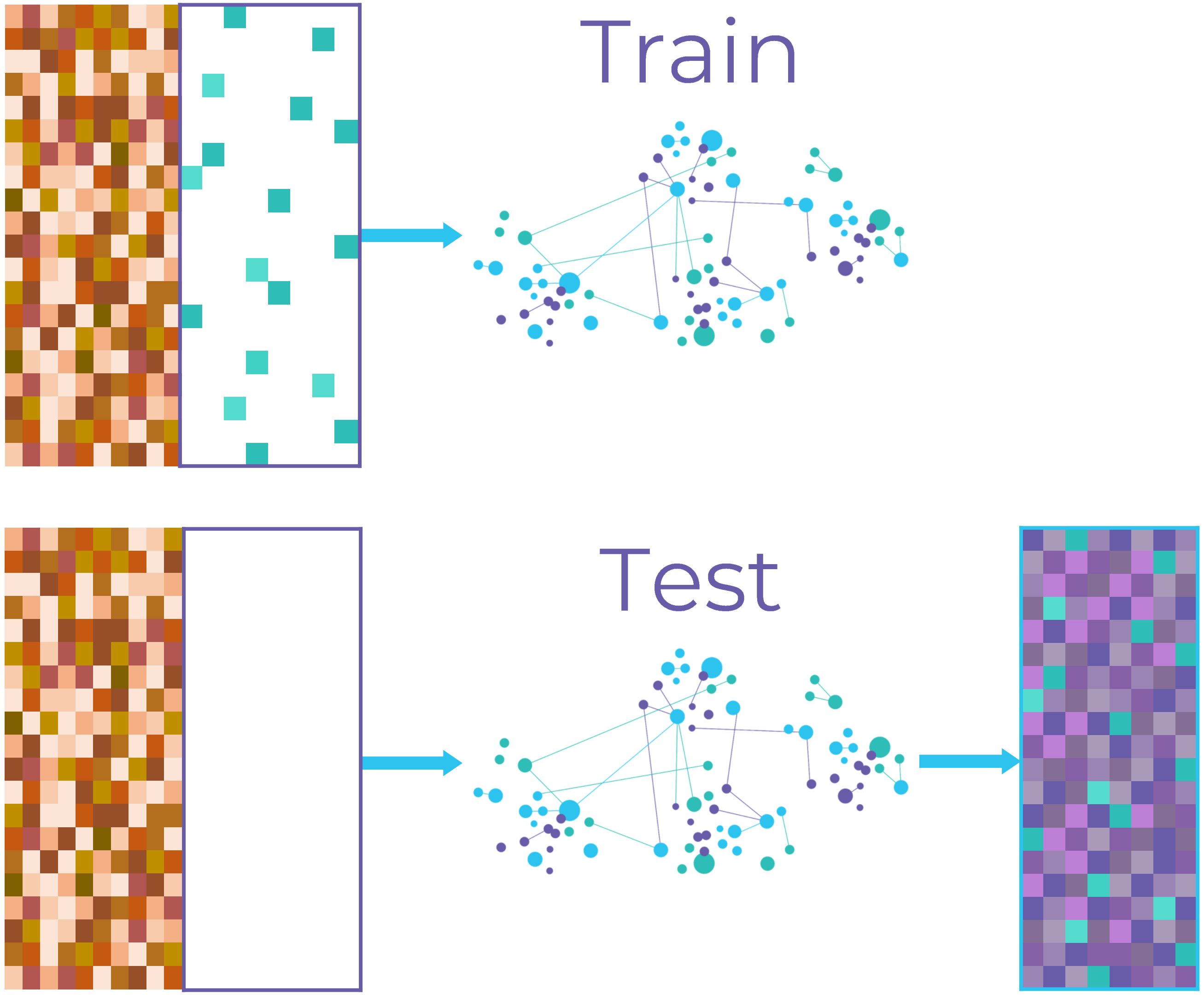


Accuracy vs level of data



intellegens.ai

Virtual compounds



intellegens.ai