Imputing Compound Activities Based on Sparse and Noisy Data

Thomas Whitehead*, Benedict Irwin[†], Peter Hunt[†], Matthew Segall[†], Gareth Conduit*

*Intellegens [†]Optibrium. Email: matt@optibrium.com

© 2019 Optibrium Ltd.

Optibrium™, StarDrop™, Card View™, Nova™, Auto-Modeller™, WhichP450™, Glowing Molecule™ and MPO Explorer™ are trademarks of Optibrium Ltd.

The Challenges of Applying Deep Learning to Drug Discovery Data

- 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

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, …

Whitehead *et al.* J. Chem. Inf. Model (2019) **59**(3), pp 1197–1204

Deep learning

Inputs

Outputs

Alchemite™ deep learning

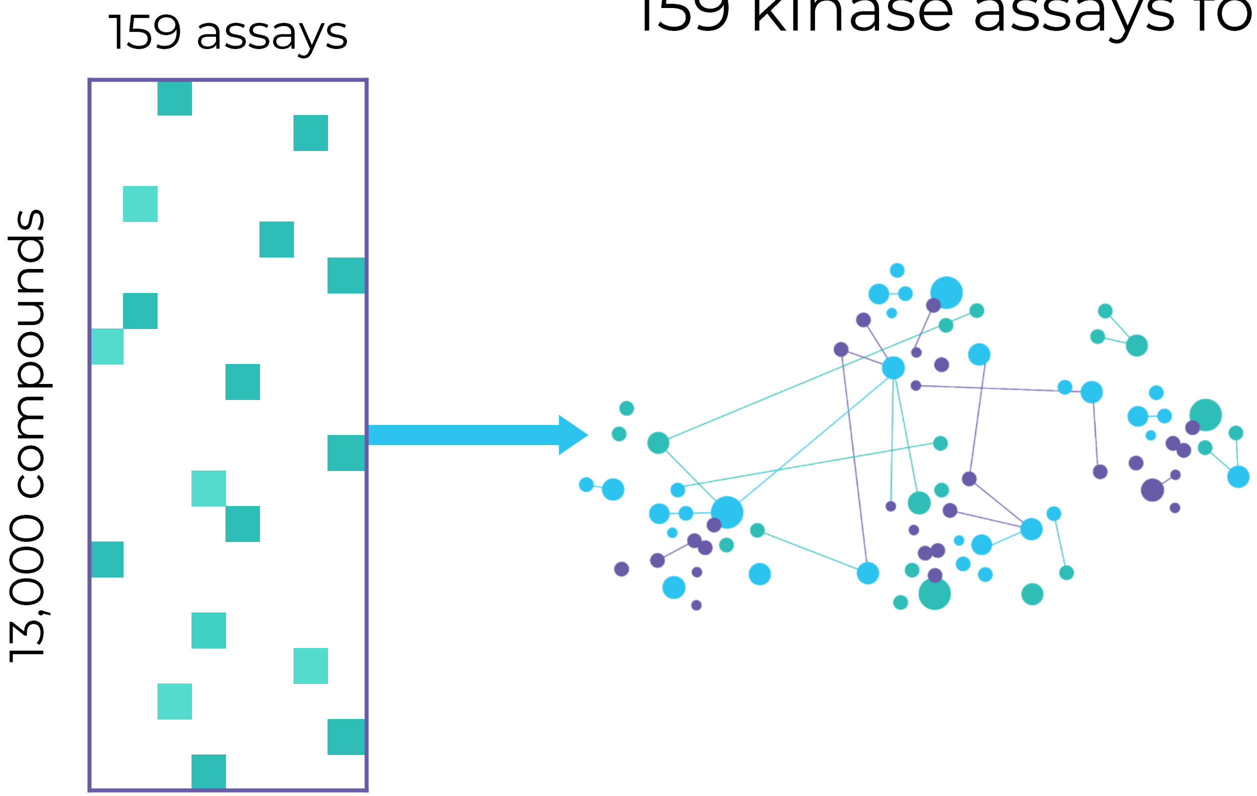
Inputs
Outputs

Whitehead *et al.* J. Chem. Inf. Model (2019) **59**(3), pp 1197–1204

Novartis dataset to benchmark machine learning

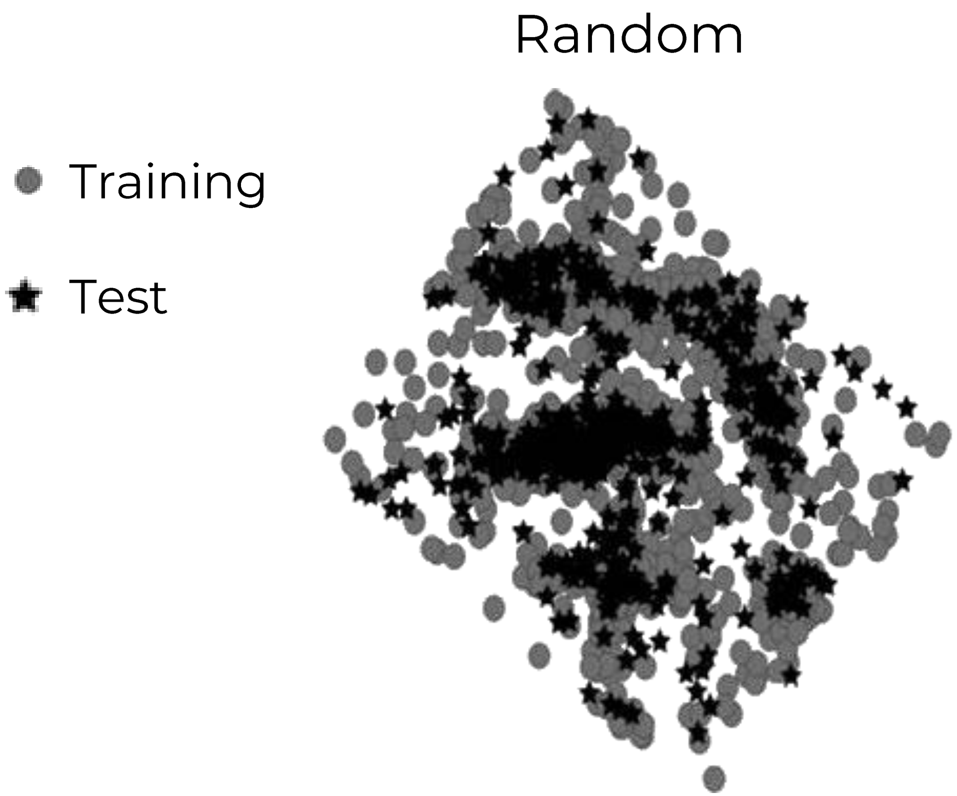
159 kinase assays for 13,000 compounds

Data **5%** complete



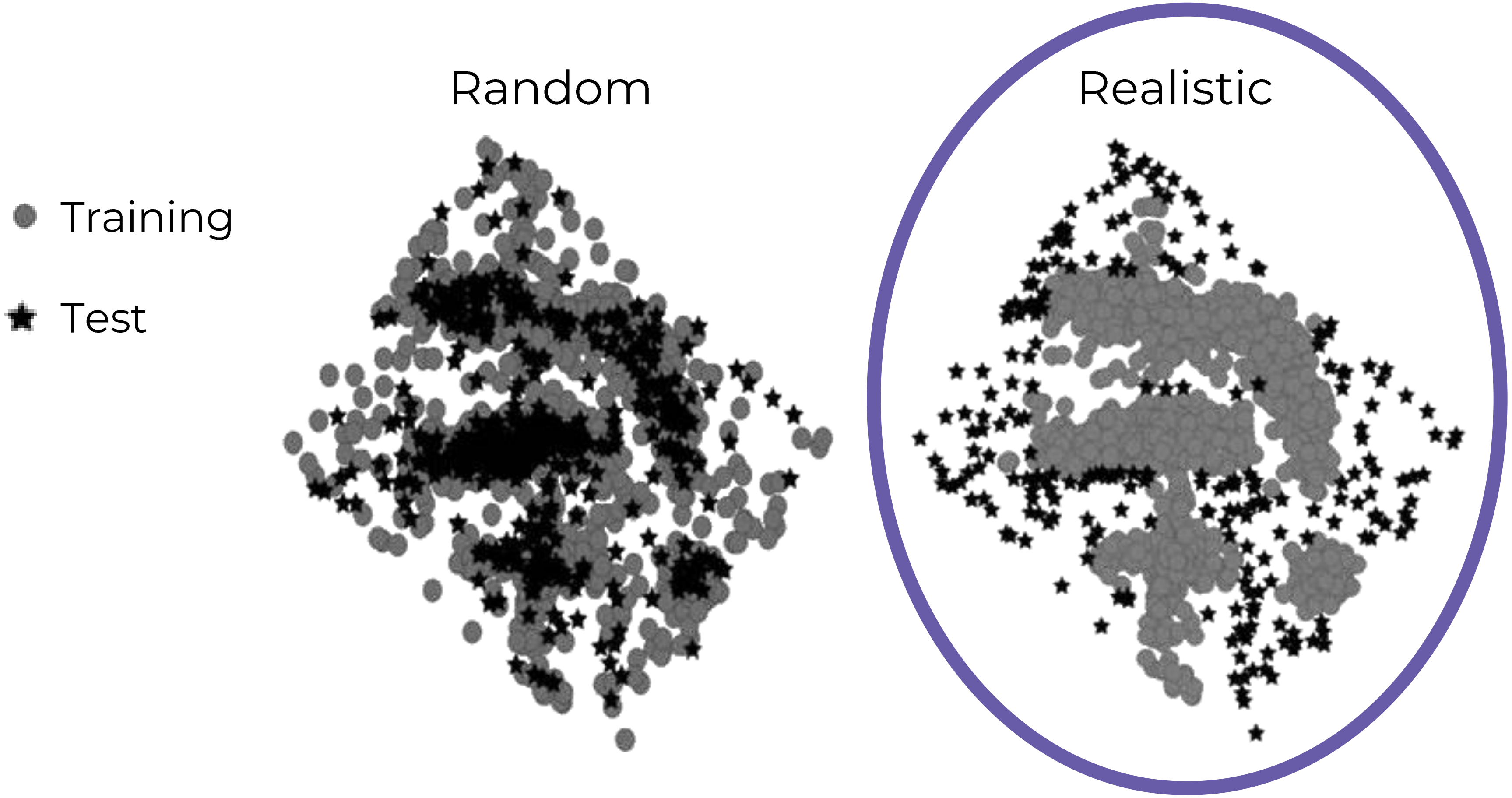
Data from ChEMBL
Martin, Polyakov, Tian, and Perez, J. Chem. Inf. Model. 57, 2077 (2017)

Novartis dataset distribution



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

Novartis dataset is realistically distributed



Random

Realistic

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

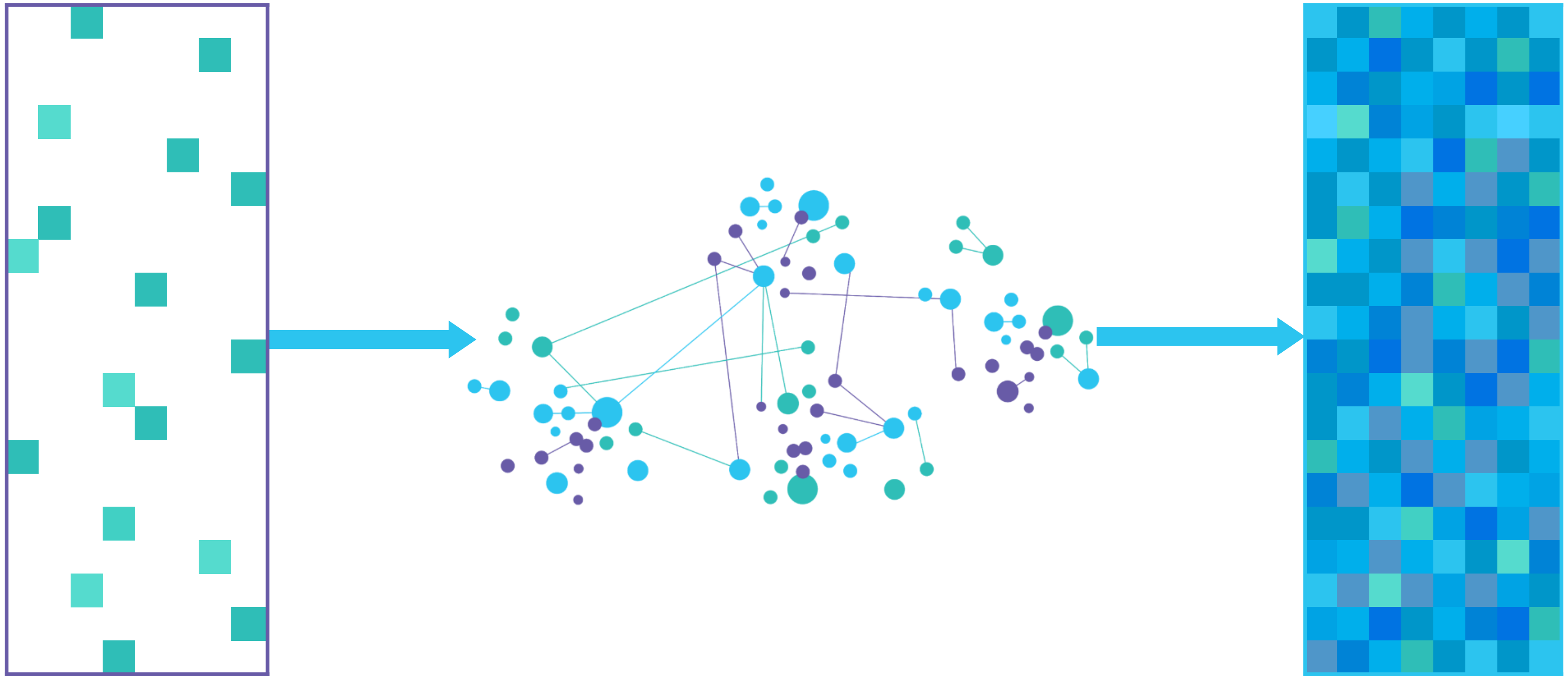
Accuracy metrics

Coefficient of determination, R^2

Root Mean Square Error, RMSE

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

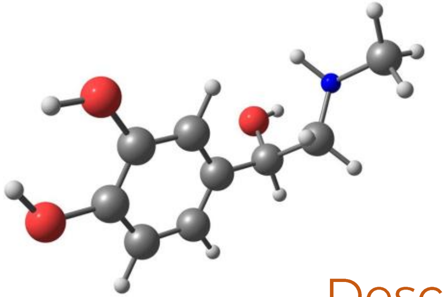
Aim: impute missing assay values

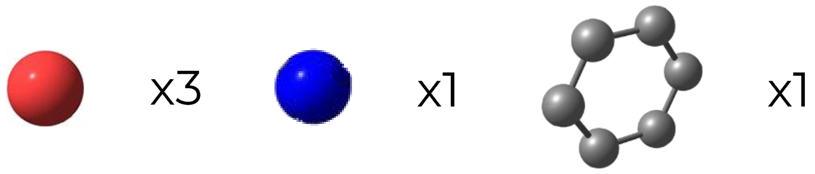


Validate against realistically-split holdout set

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

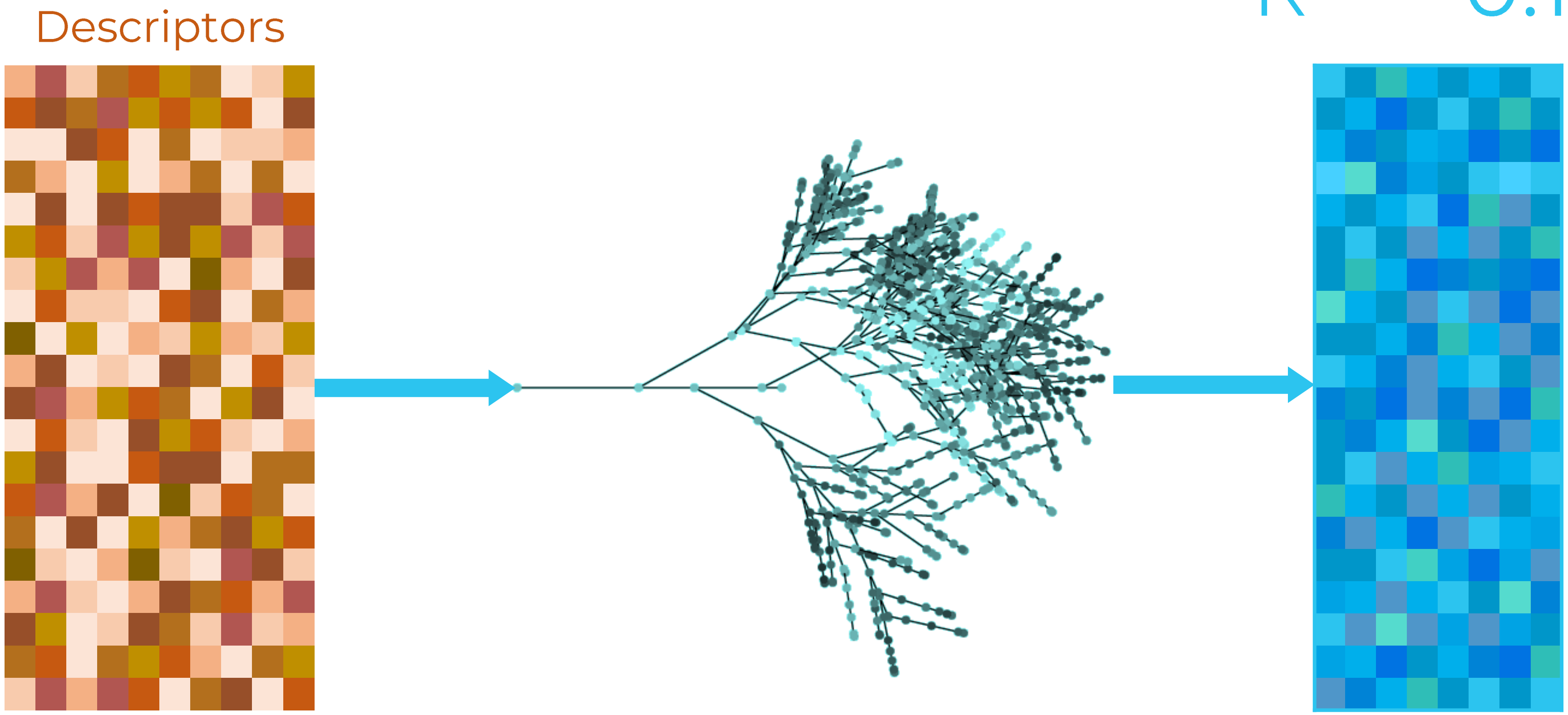
Random forest regression



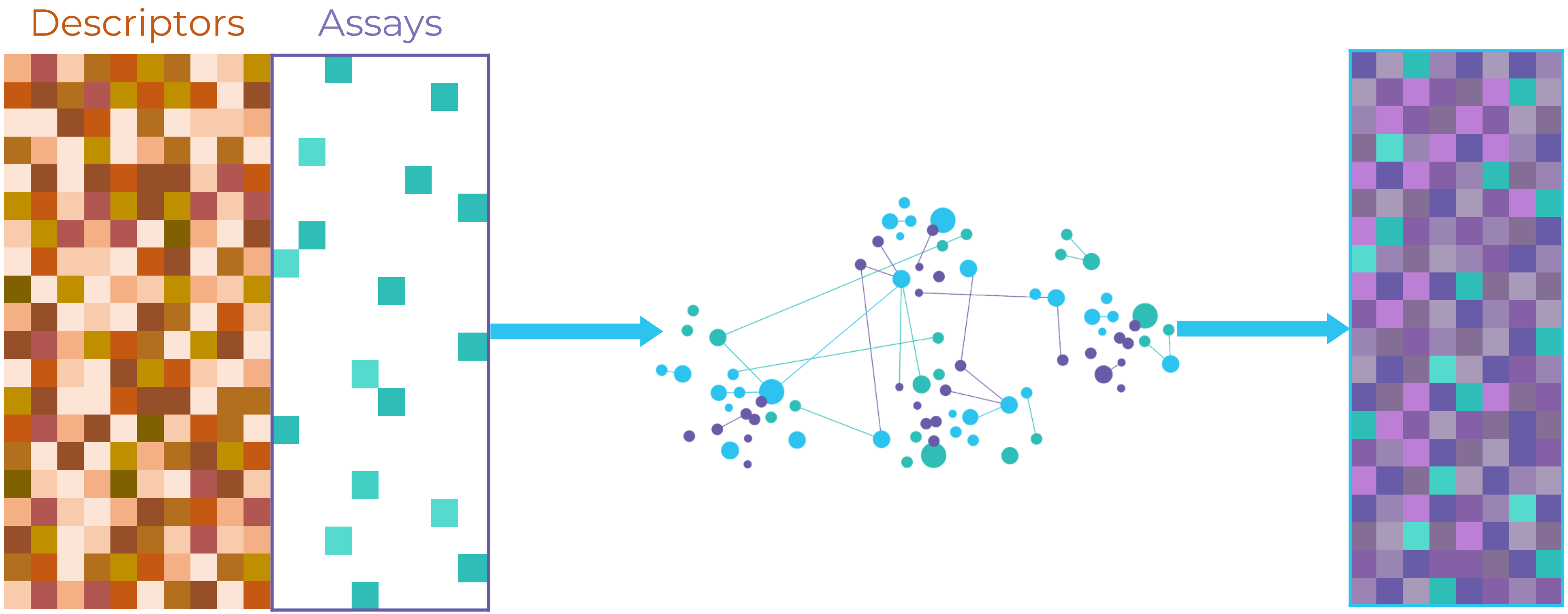


Molecular weight = 183 Da

$R^2 = -0.19$



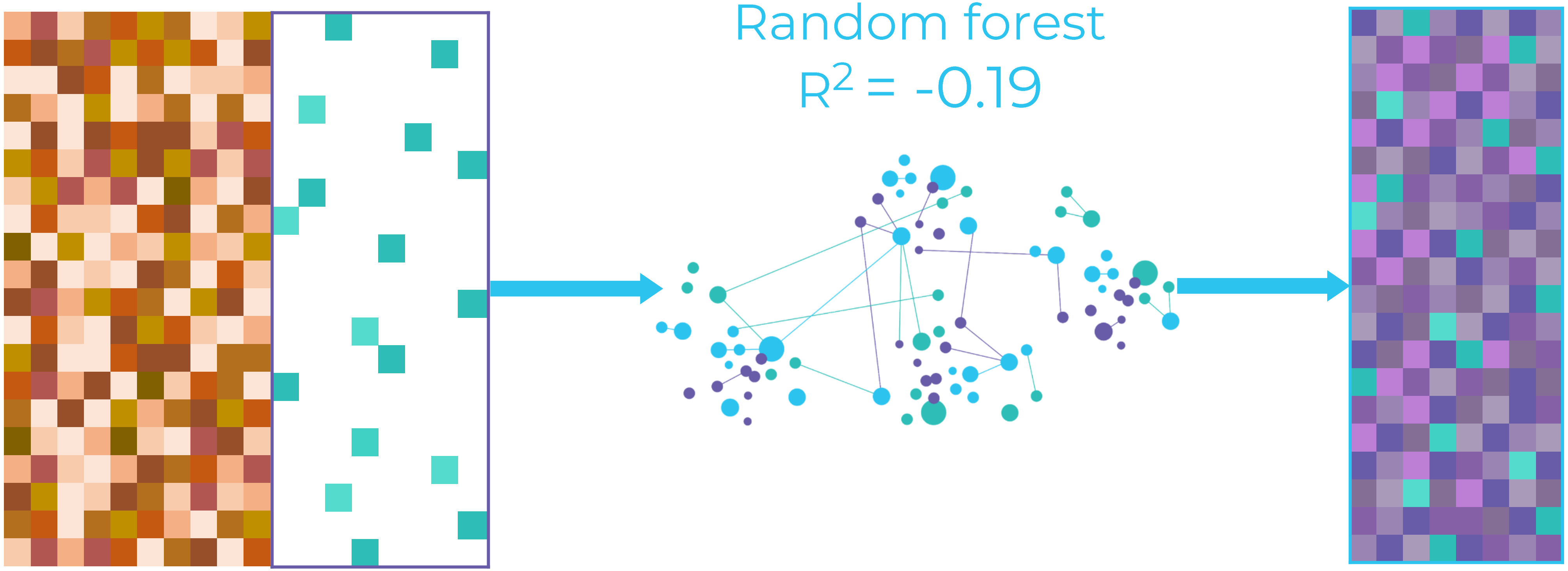
Descriptors and bioactivity values



Whitehead *et al.* J. Chem. Inf. Model (2019) **59**(3), pp 1197–1204

Deep learning predictions

$R^2 = 0.46$



Random forest

$R^2 = -0.19$

Whitehead *et al.* J. Chem. Inf. Model (2019) **59**(3), pp 1197–1204

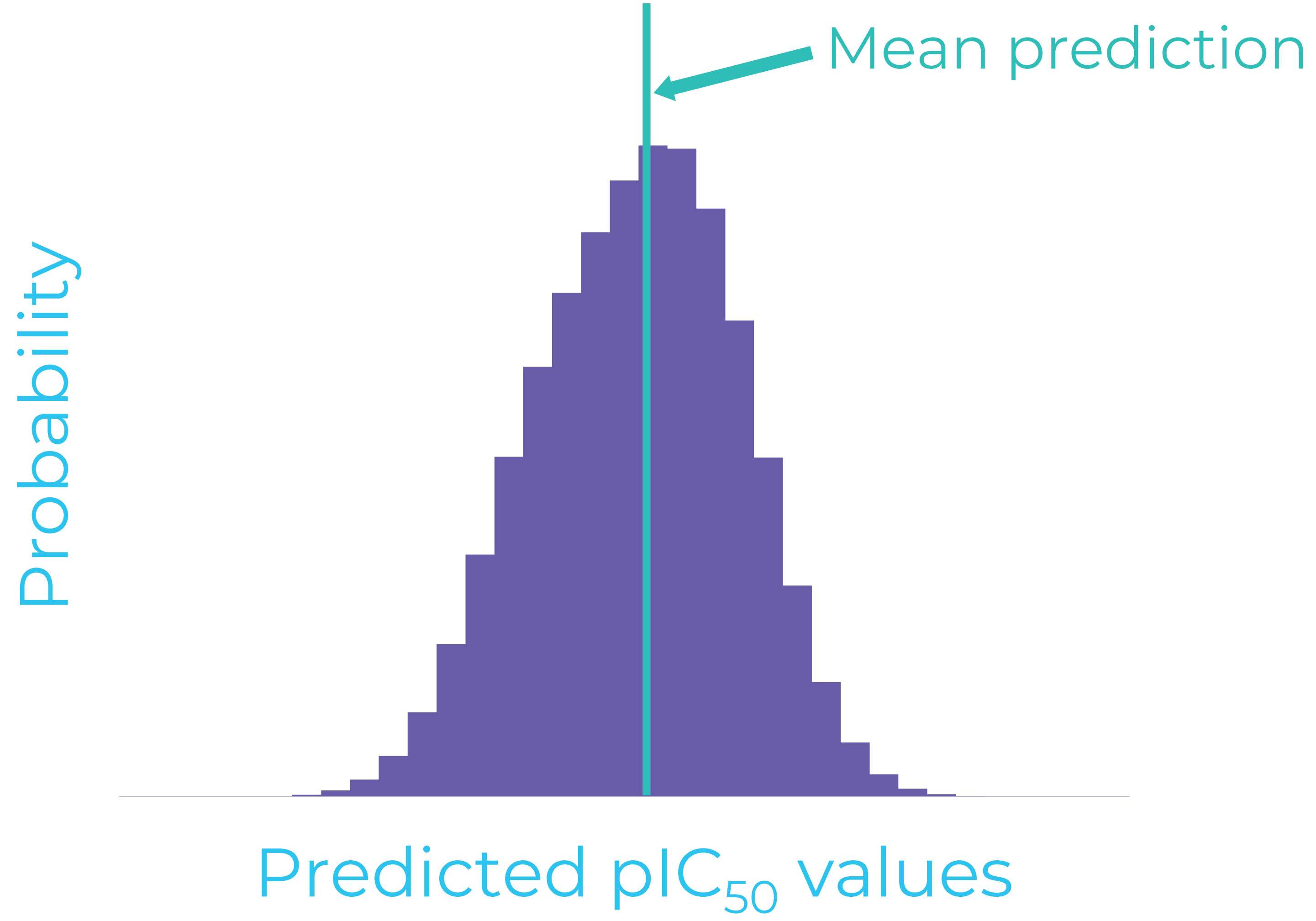
Comparison with other methods

Method	R^2	RMSE
Alchemite	**0.46***	**0.59**
Profile QSAR 2.0	0.43	0.61
Multi-target deep neural network (tensor-flow)	0.11	0.77
Collective matrix factorisation	-0.11	0.87
Random forest	-0.19	0.89

* N.B. Improved over published results (R^2=0.44). Also 1000× faster to build model!

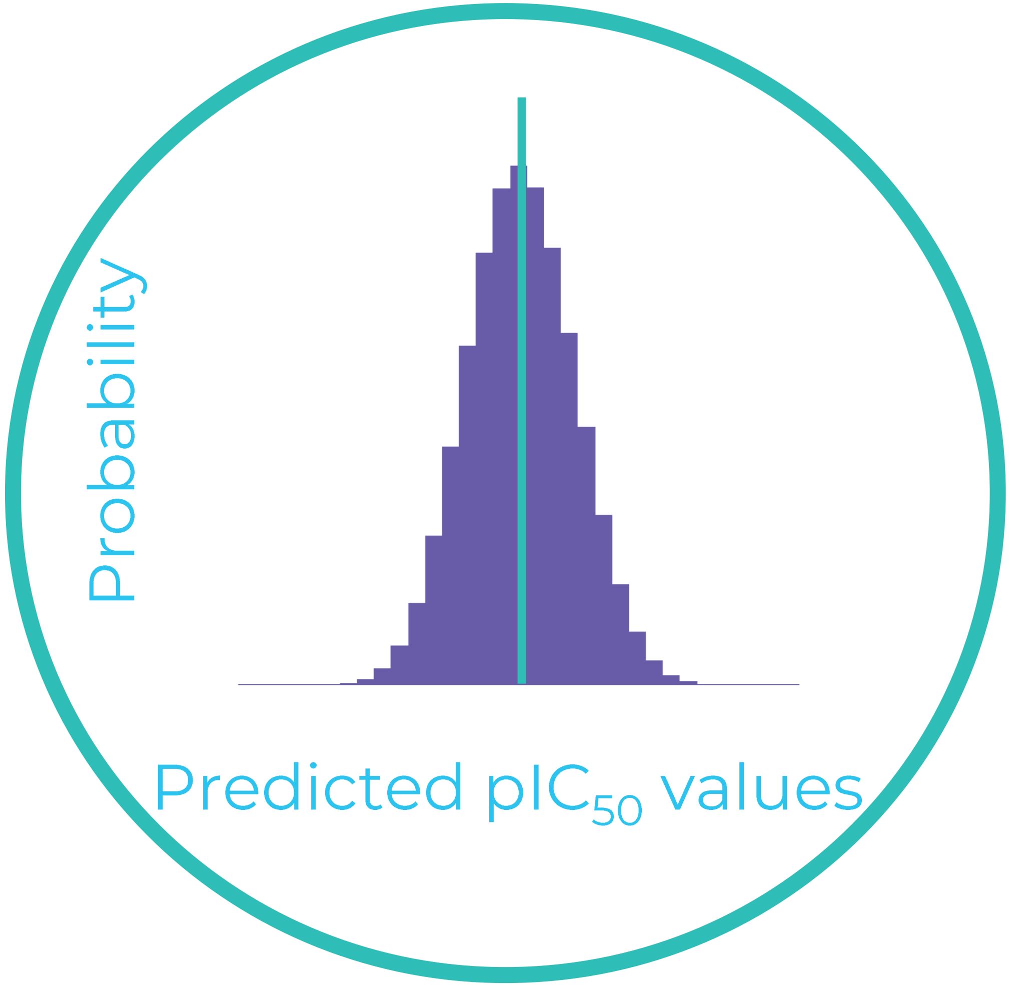
Whitehead *et al.* J. Chem. Inf. Model (2019) **59**(3), pp 1197–1204

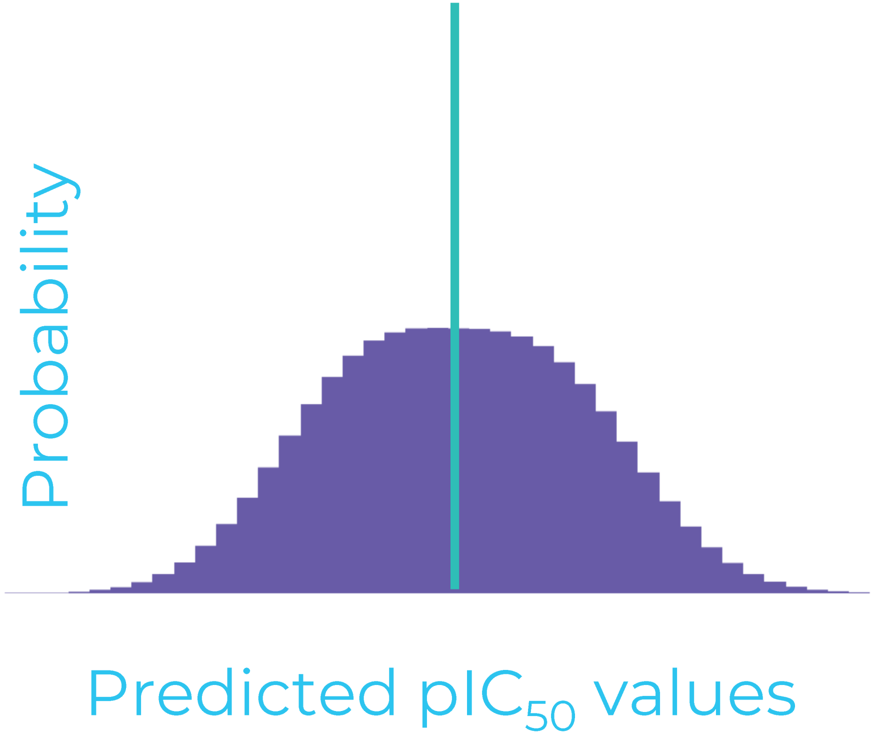
Calculate probability distribution



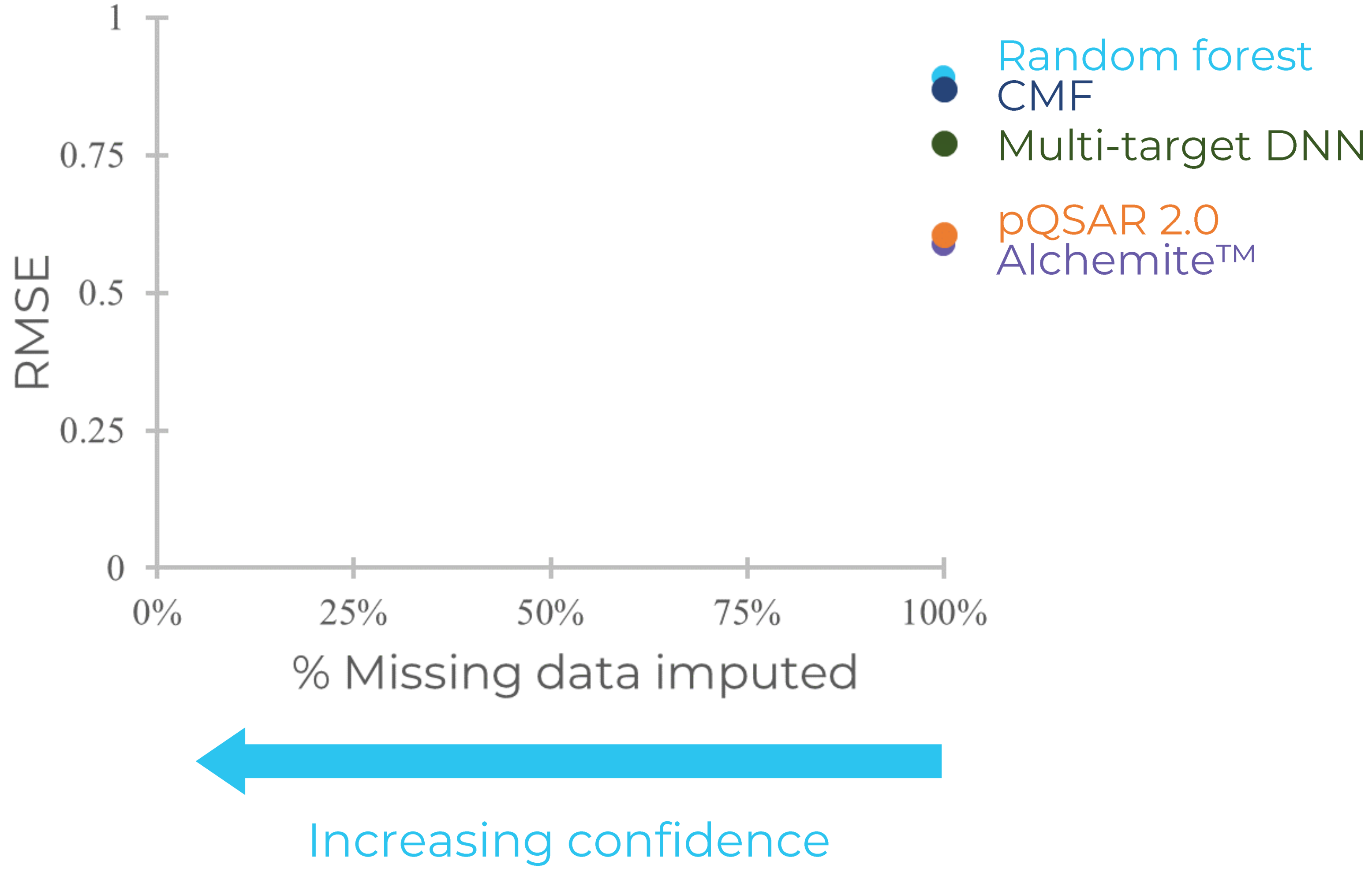
Whitehead *et al.* J. Chem. Inf. Model (2019) **59**(3), pp 1197–1204

Focus on most confident predictions

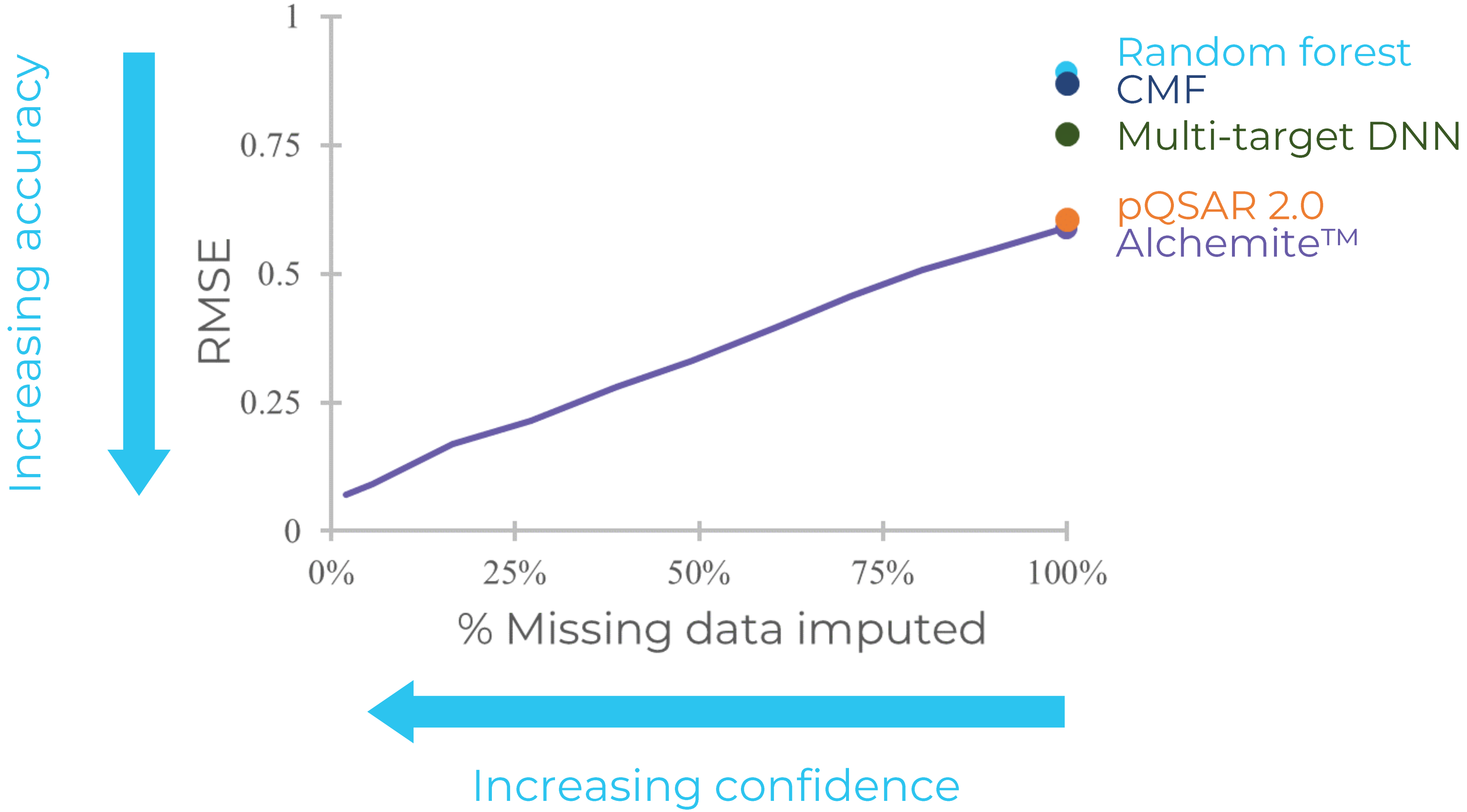




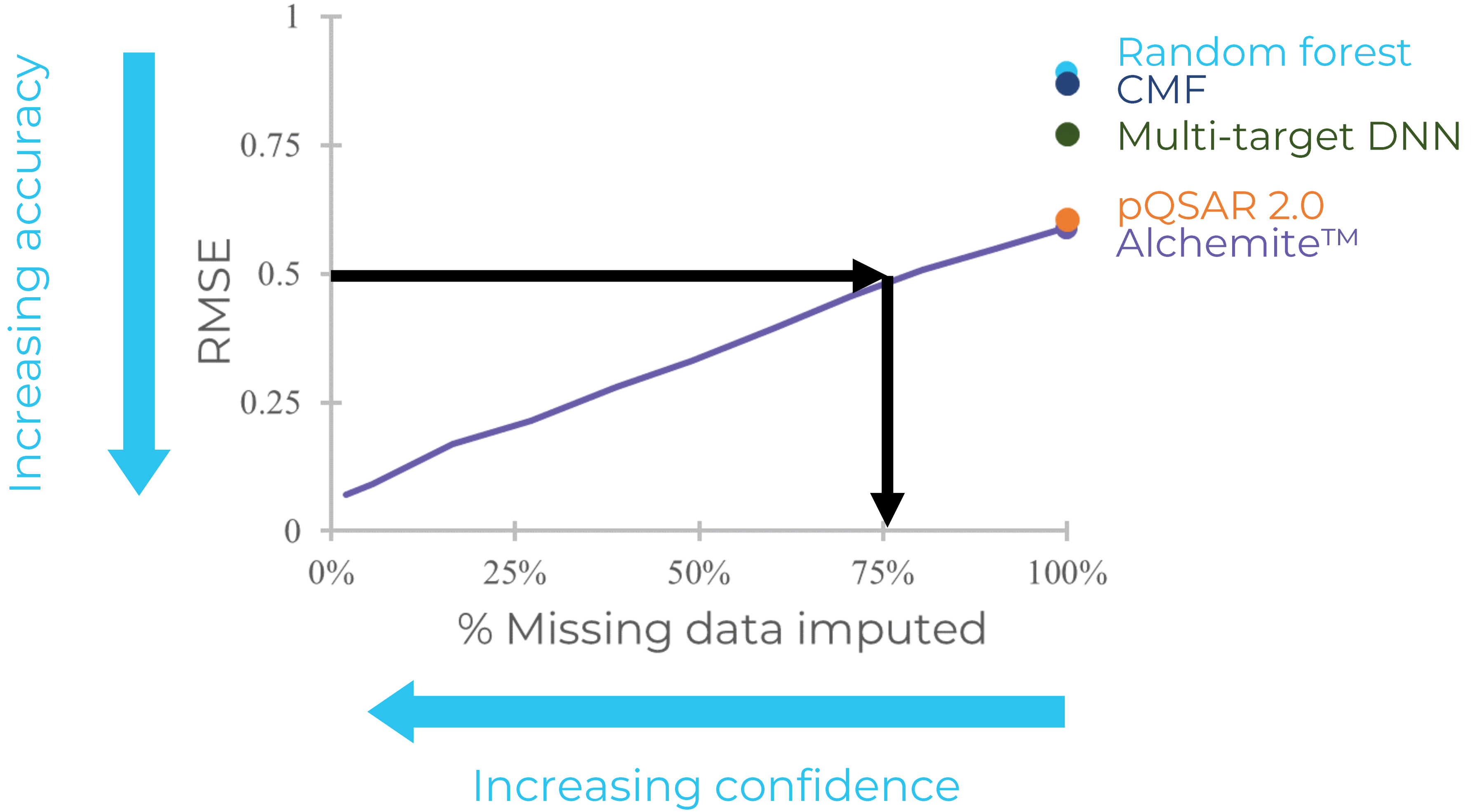
Reporting only most confident predictions



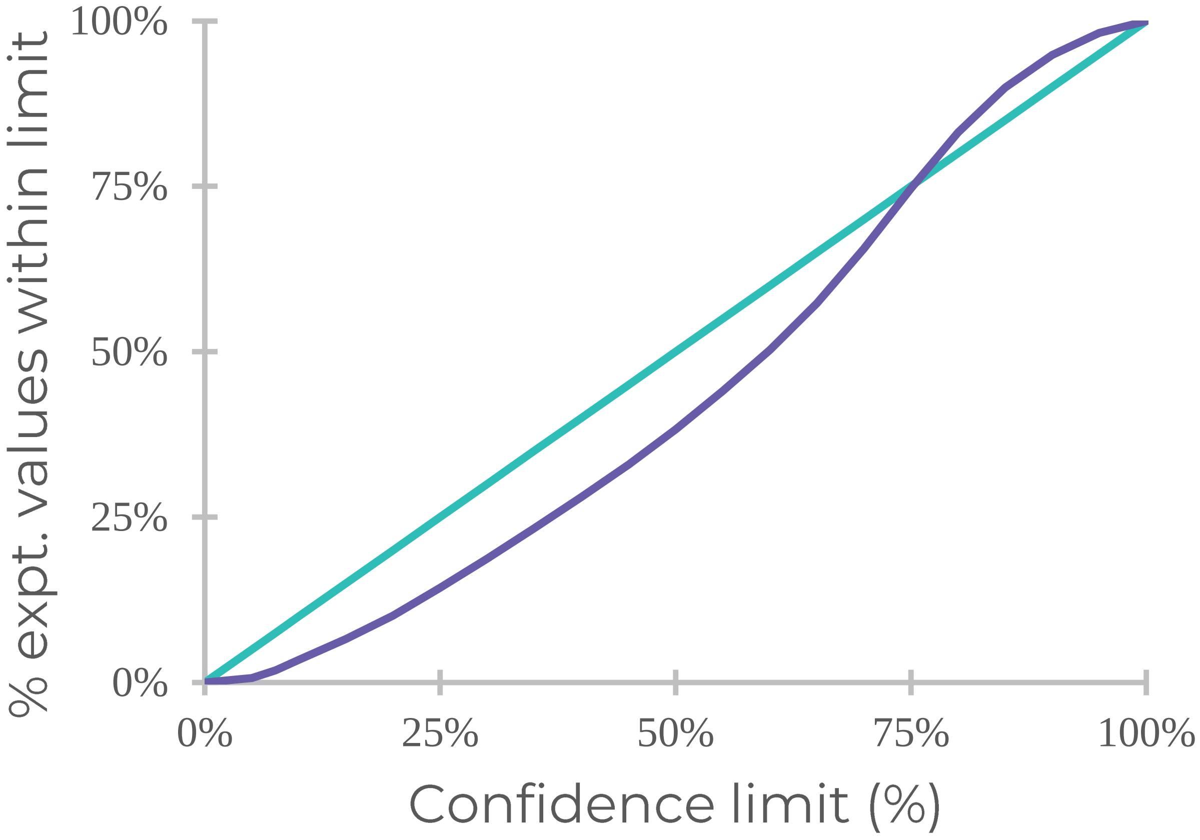
Reporting only most confident predictions

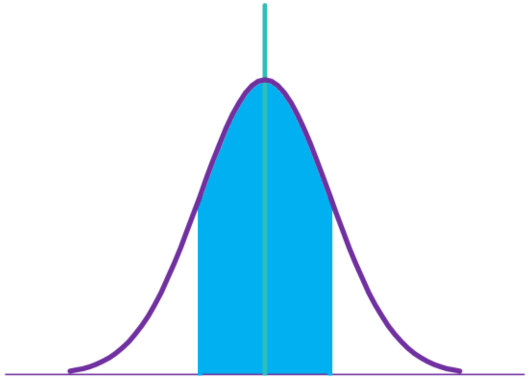


Reporting only most confident predictions



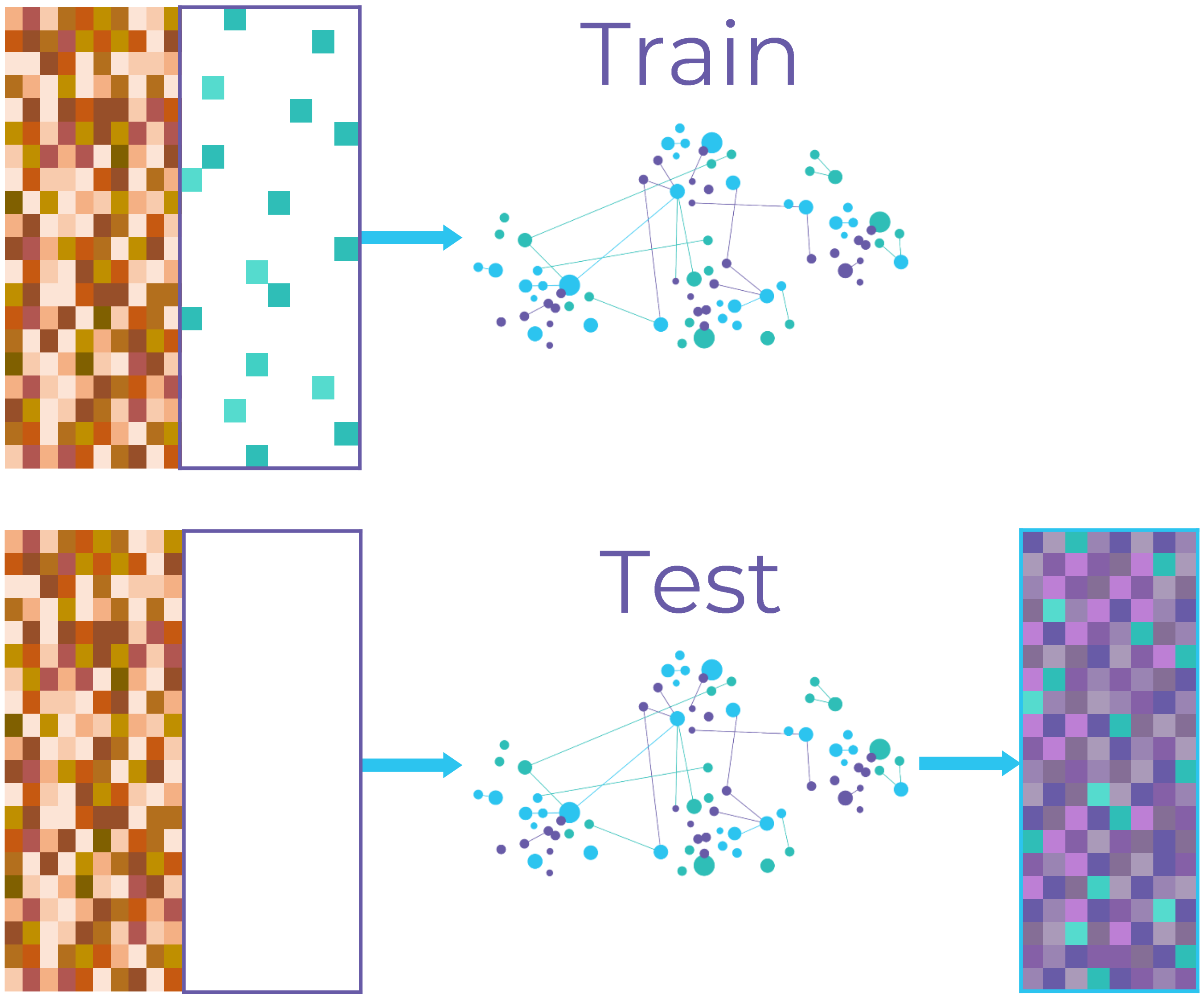
Absolute accuracy of uncertainties





N.B. Assumes normally distributed errors e.g. 62% of results within 1 SD

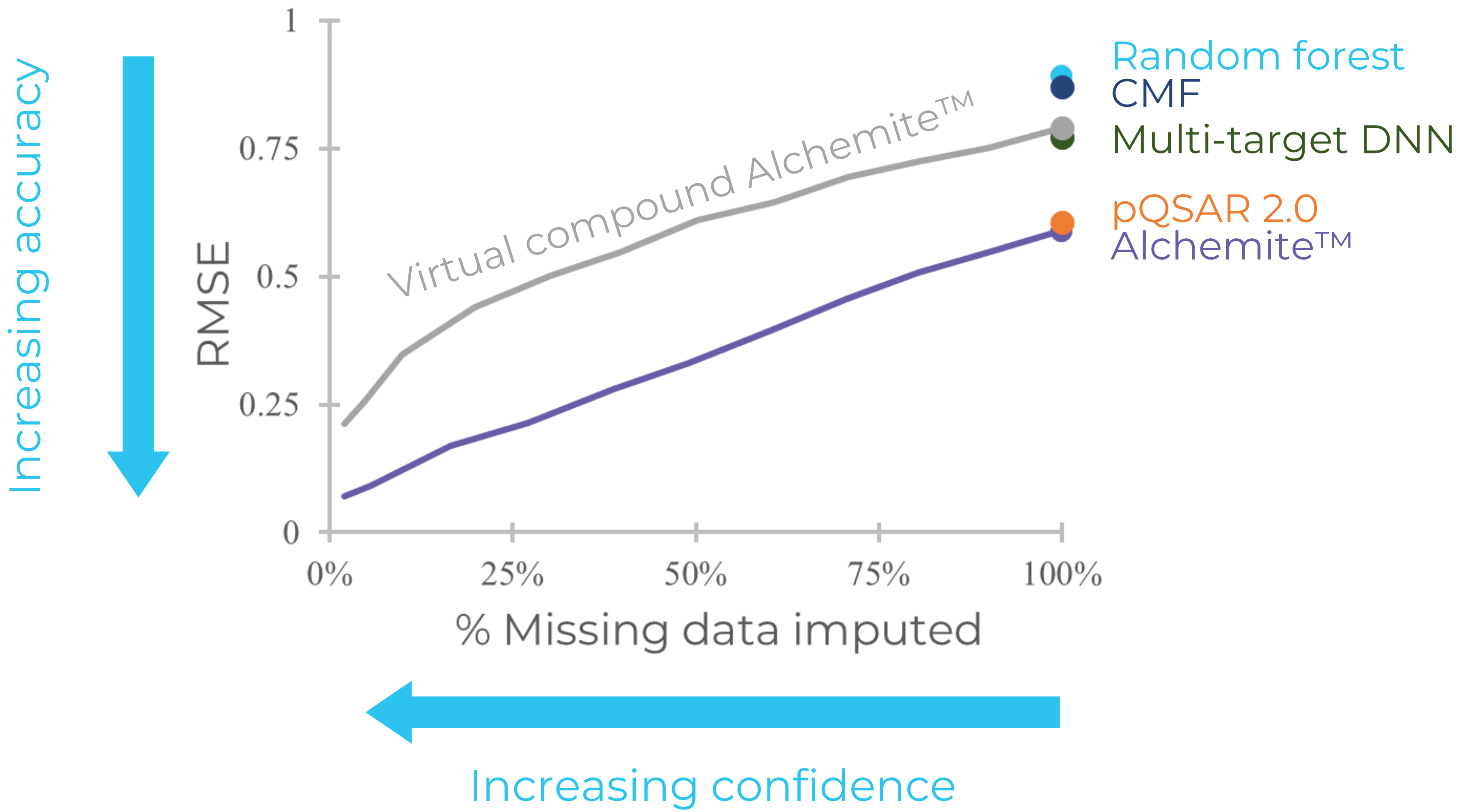
Application to virtual compounds



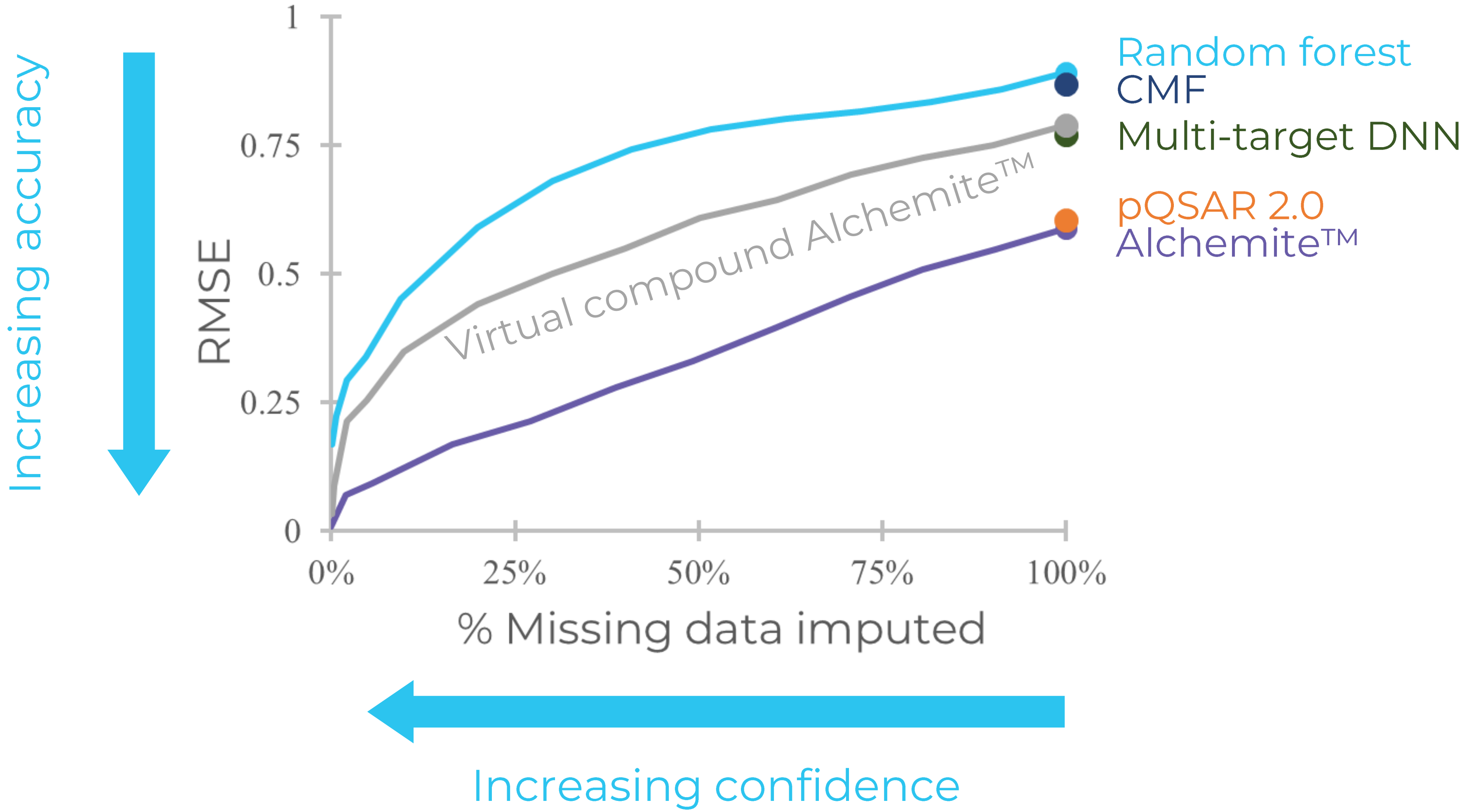
Train

Test

Application to virtual compounds



Random forest confidence predictions



Conclusions

- 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
- Impute results of missing assays to **high accuracy**
- **Broadly applicable** to other endpoints, e.g. physicochemical, ADME, tox…
- **Applicable** to pharma-scale data sets
- For more details: **Whitehead *et al.* J. Chem. Inf. Model (2019) 59(3), pp 1197–1204**
  - matt@optibrium.com

© 2019 Optibrium Ltd.