Imputation of assay activity data using deep learning

Tom Whitehead, Peter Hunt, Matt Segall, Gareth Conduit

Neural network algorithm to

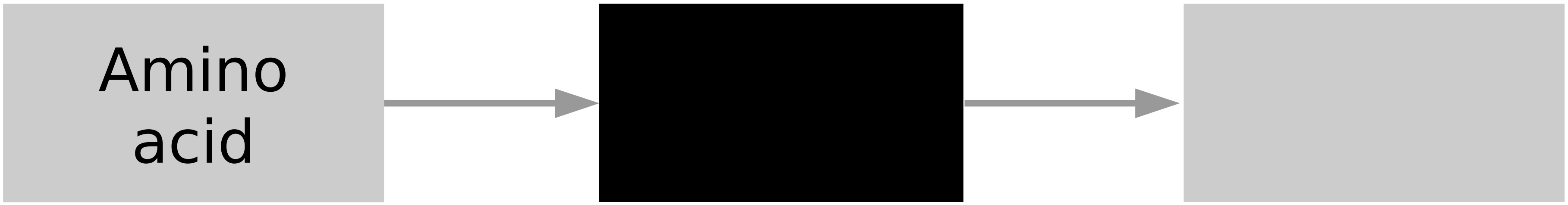
Utilise chemical descriptors, assay bioactivity, and simulations in combination

Impute assay bioactivity levels from sparse data

Reduce the need for experiments and accelerate drug discovery

Generic with proven applications in drug design and materials discovery

A black box



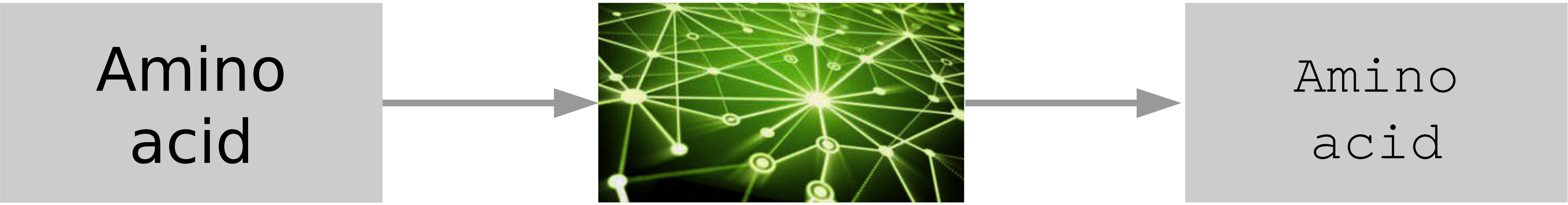
Amino acid

Train with complete data

Assay bioactivity

Assay bioactivity

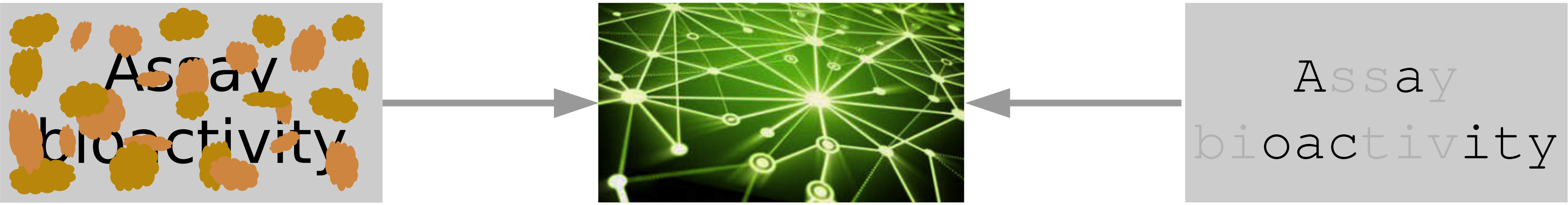
Predict with complete data



Amino acid

Amino acid

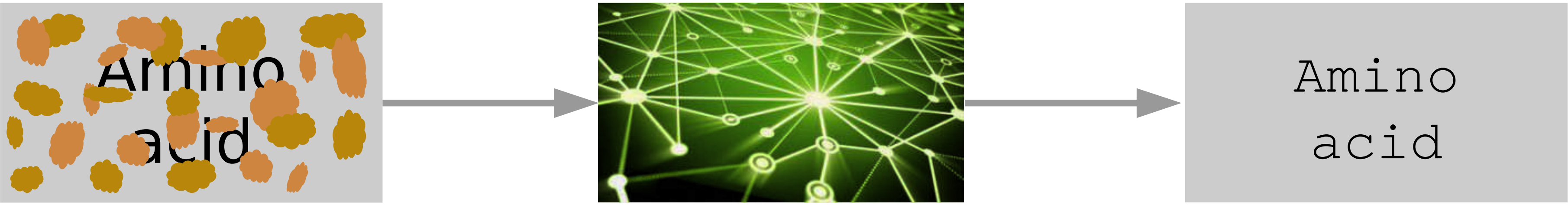
Train with fragmented data



Assay bioactivity

Assay bioactivity

Predict with fragmented data

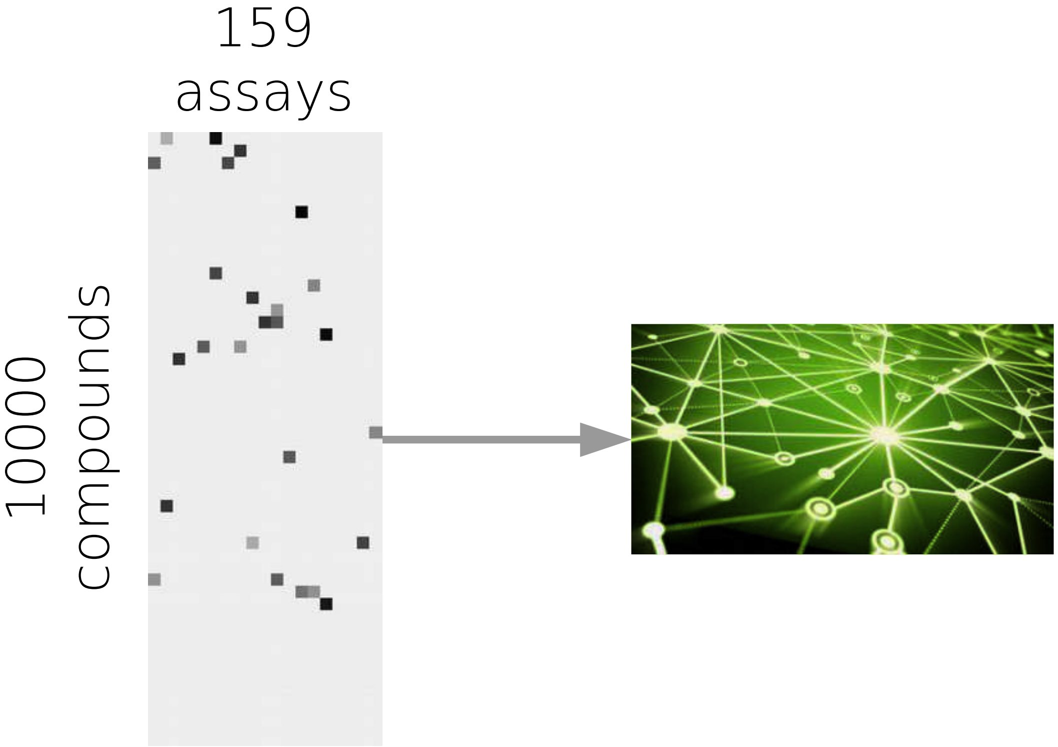


Amino acid

Amino acid

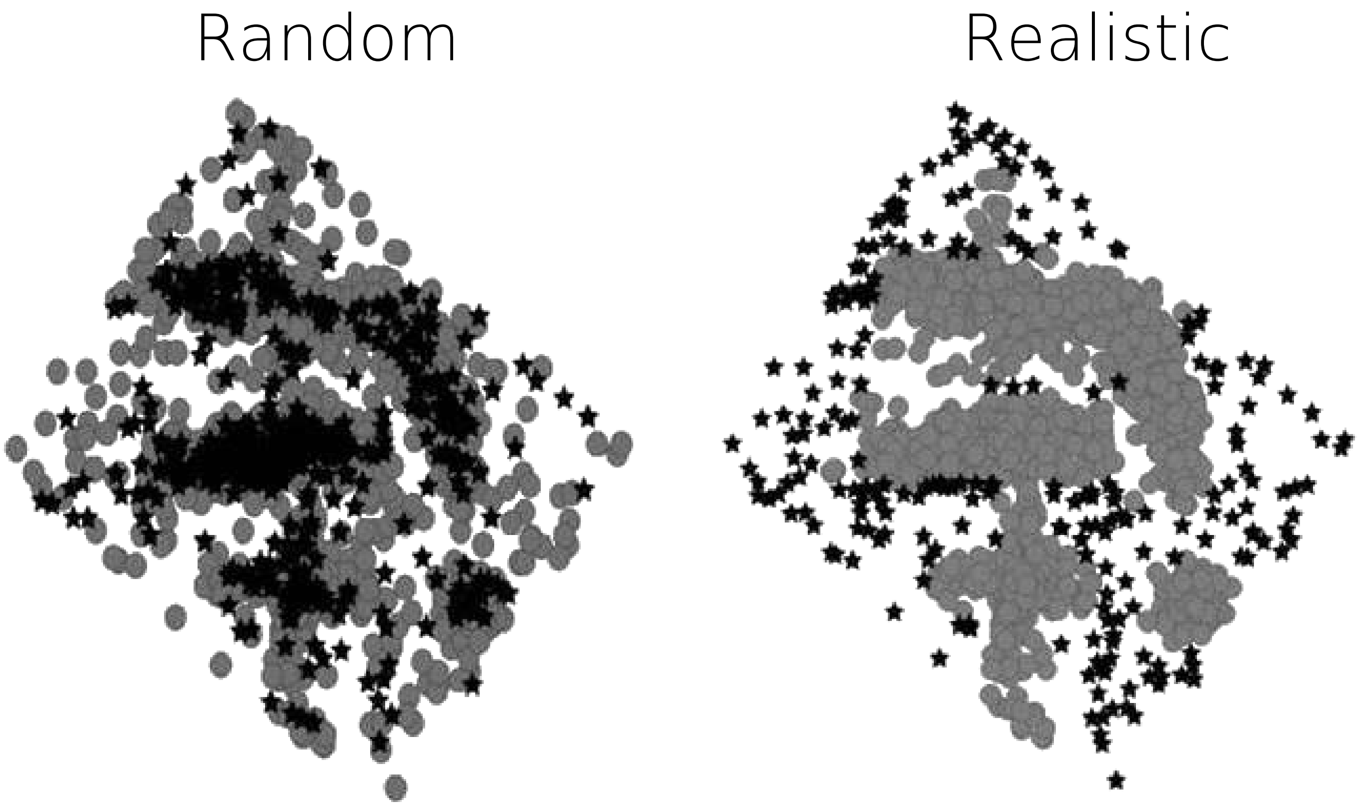
Novartis dataset to benchmark machine learning

159 kinase assays for 10000 compounds, data 5% complete



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

Novartis dataset is realistically distributed



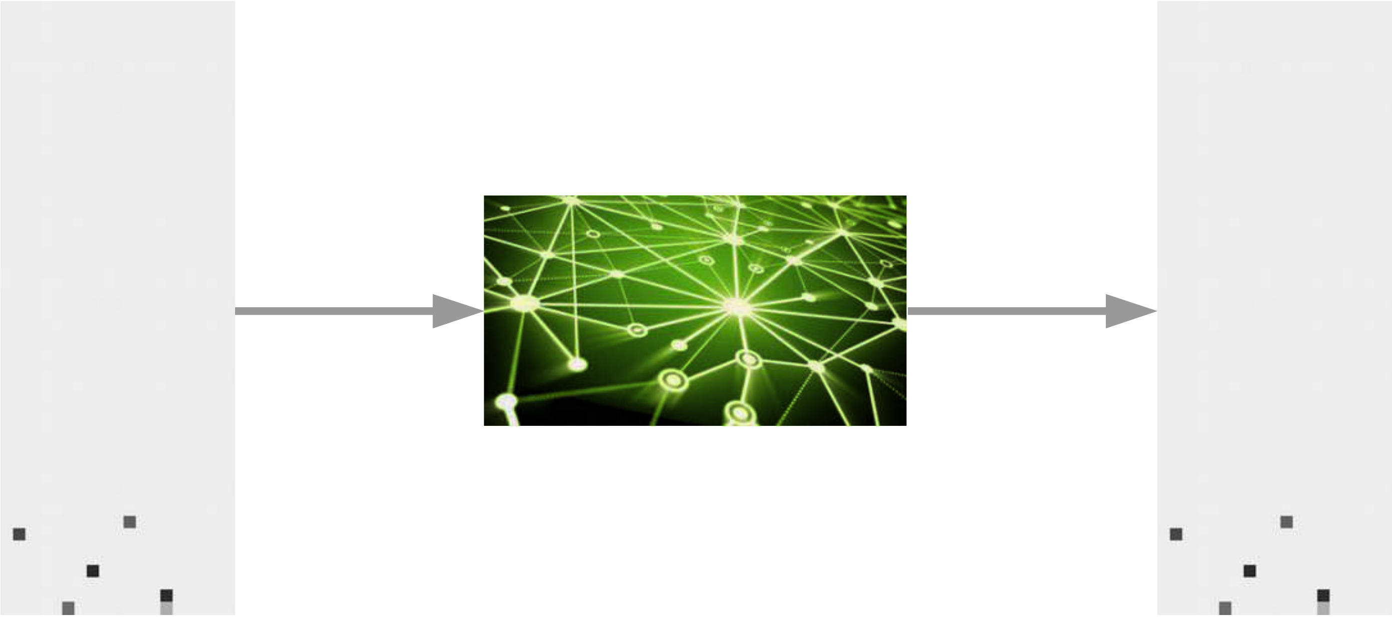
Random

Realistic

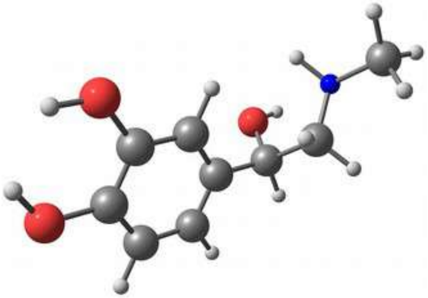
Data from ChEMBL
Martin, Polyakov, Tian, and Perez,
J. Chem. Inf. Model. 57, 2077 (2017)

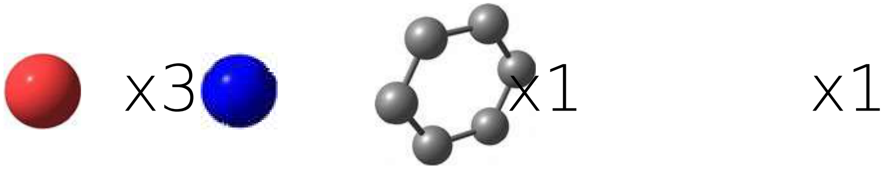
Want to impute missing entries

Validate using a realistically split holdout data set, extrapolate to new chemical space

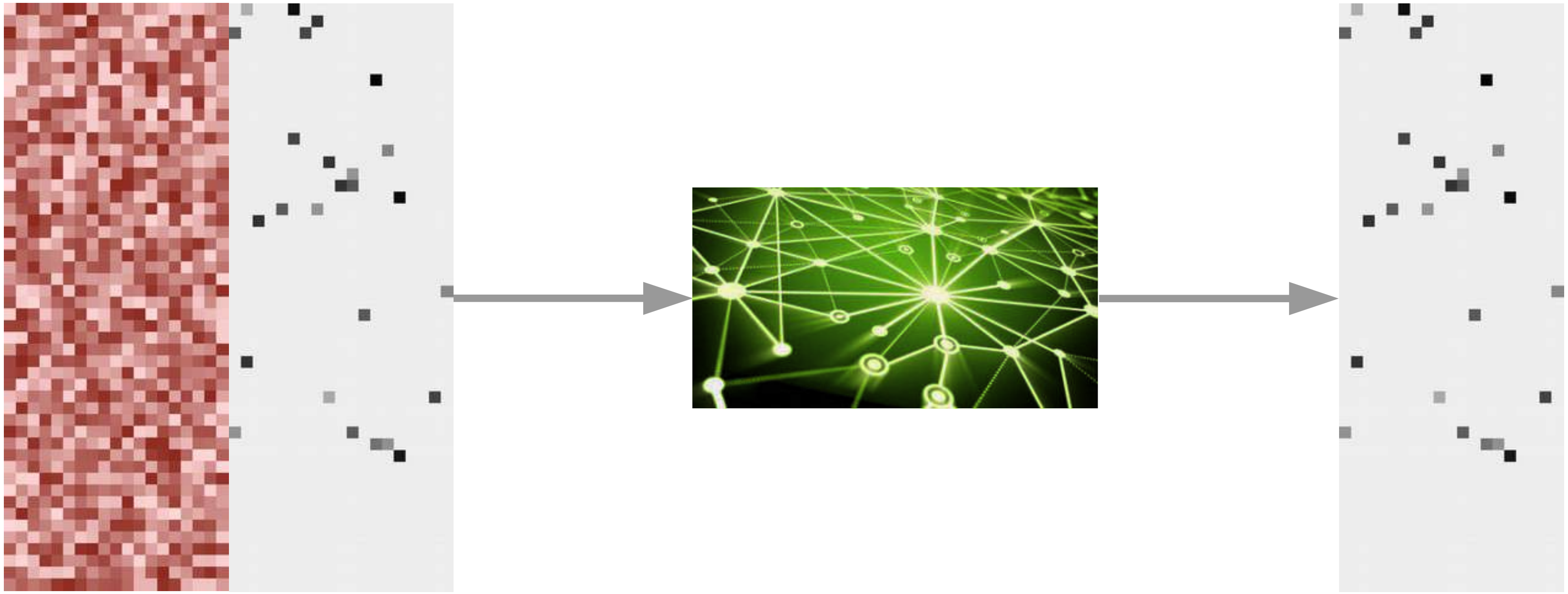


QSAR: quantitative structure-activity relationships

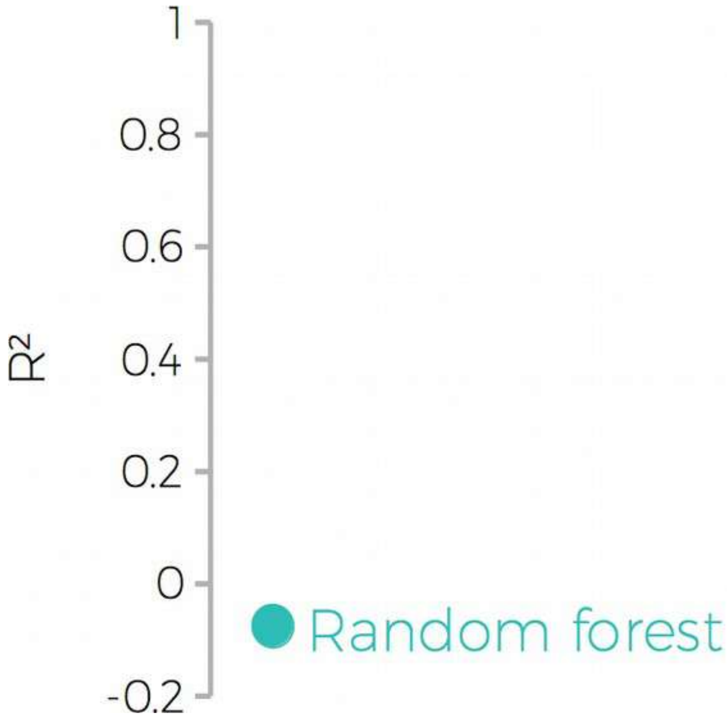




Molecular weight=183 Da

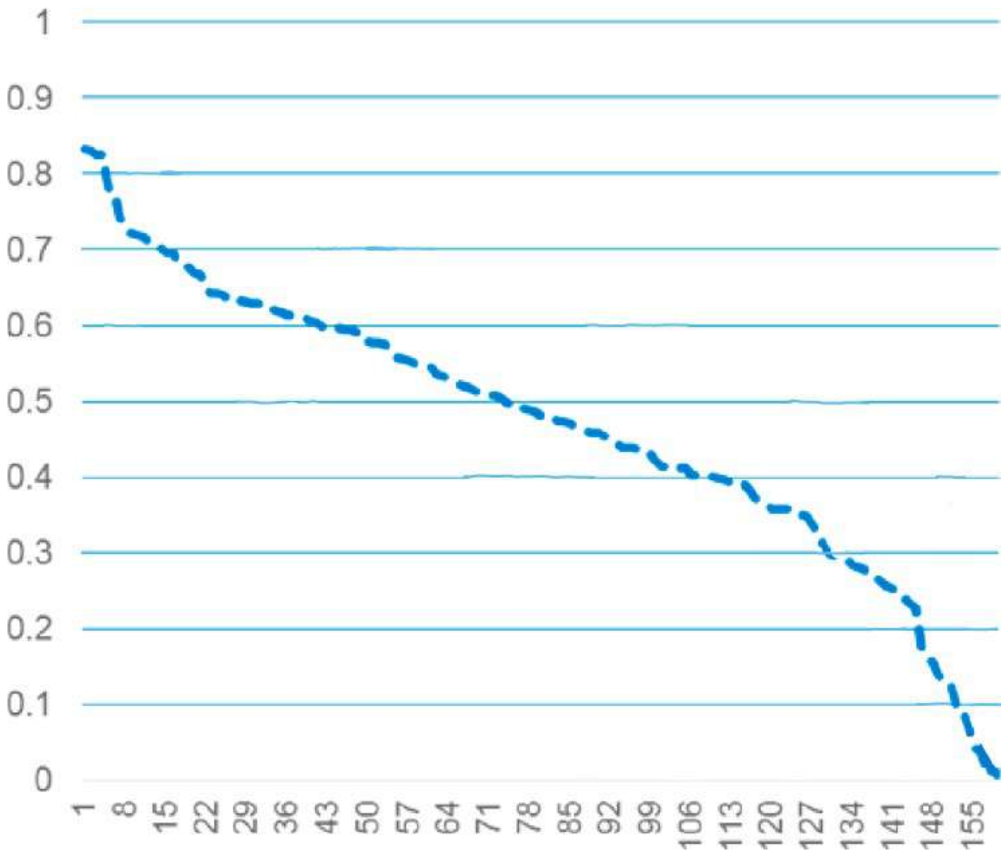


Random forest



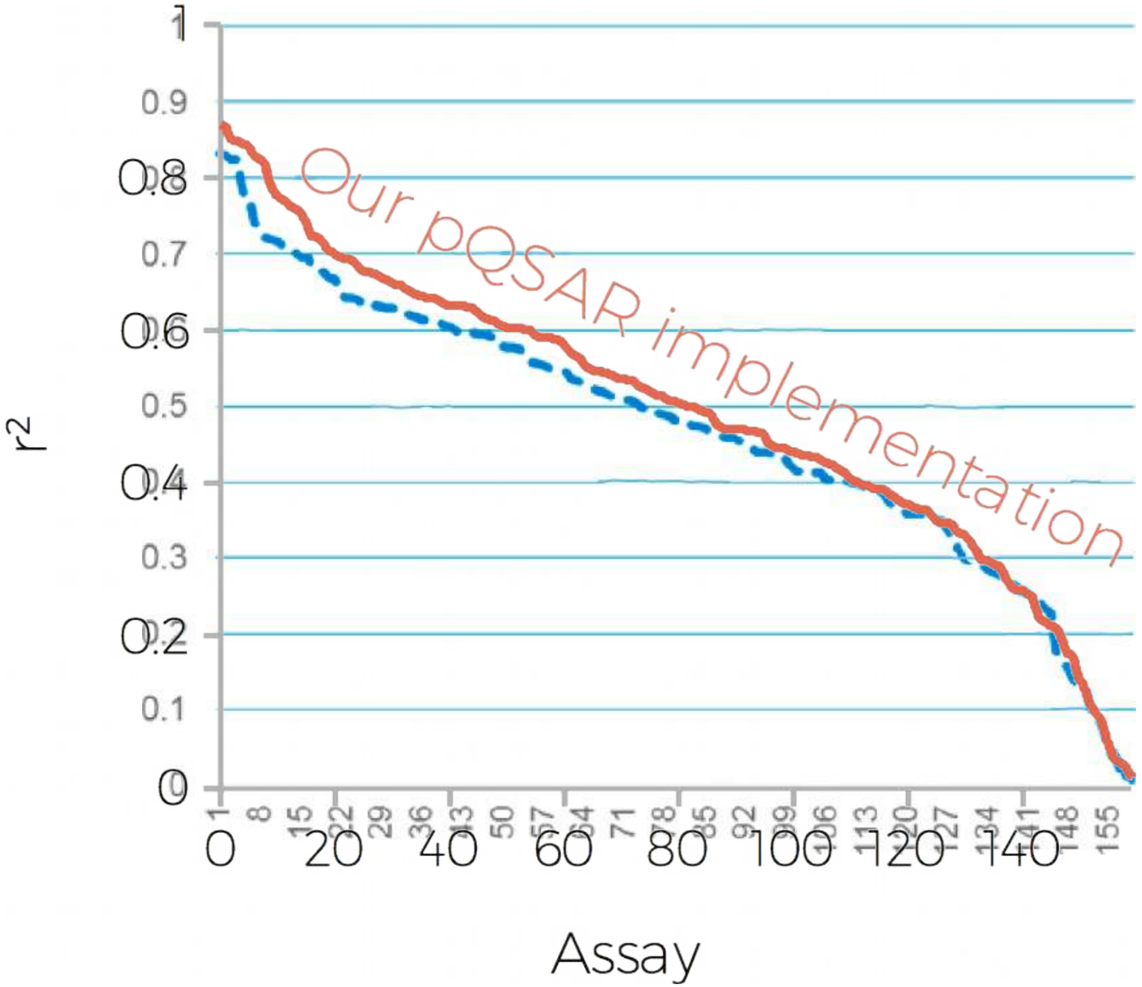
pQSAR: baseline results

pQSAR takes random forest models to impute activities as input to a partial least squares model

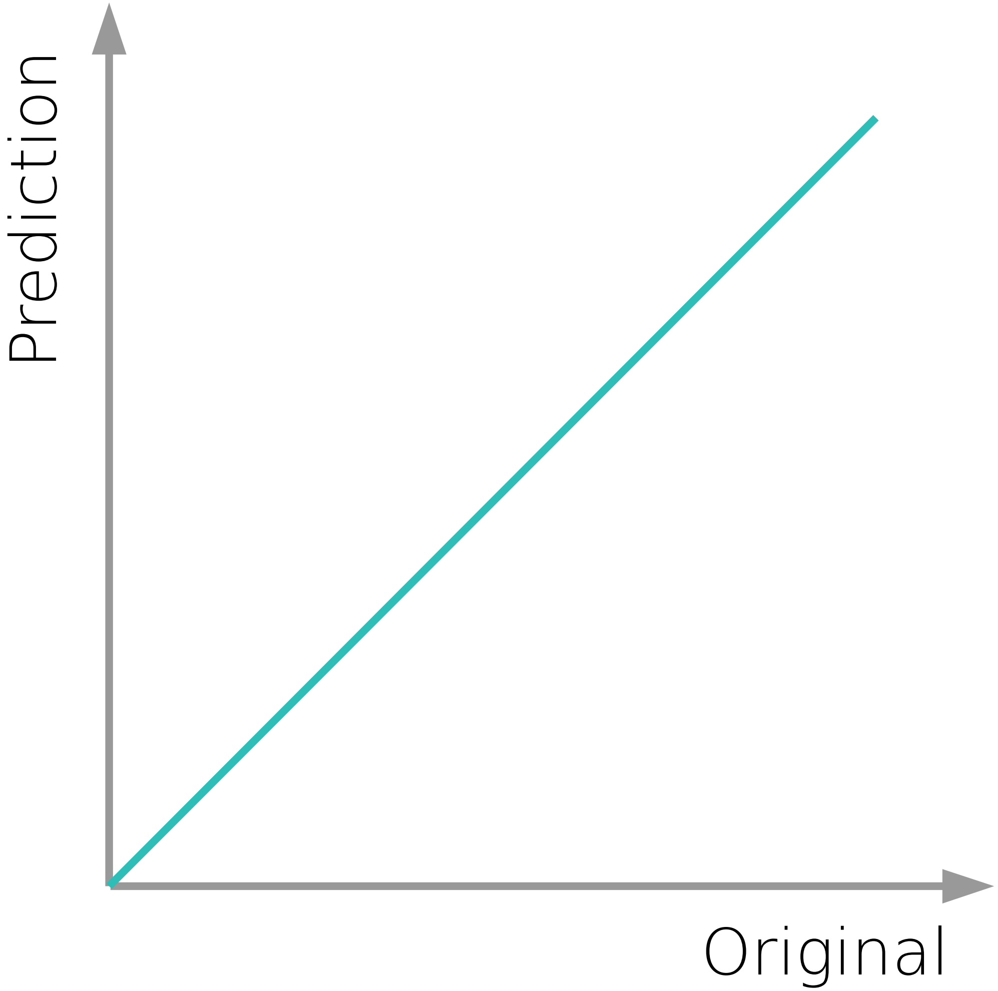


Martin, Polyakov, Tian, and Perez,
J. Chem. Inf. Model. 57, 2077 (2017)

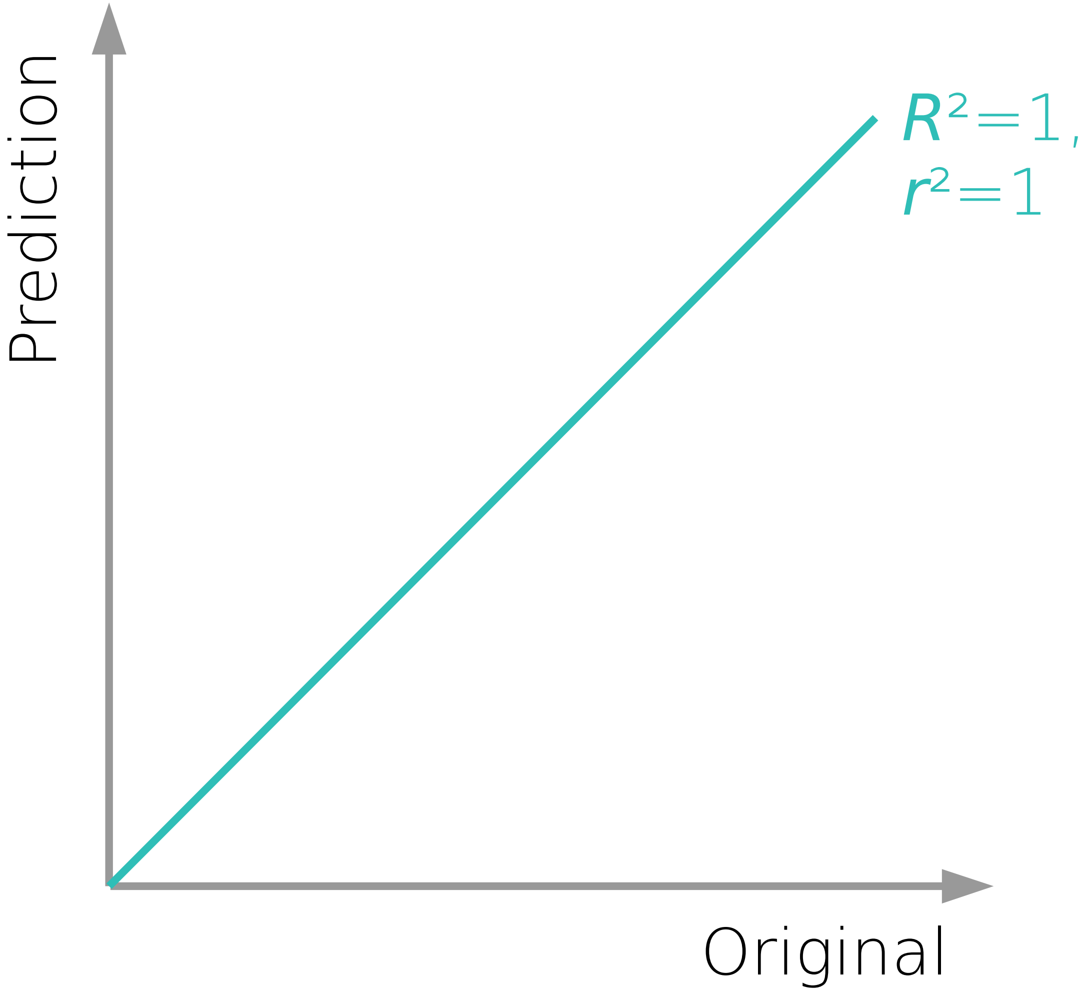
pQSAR: baseline results



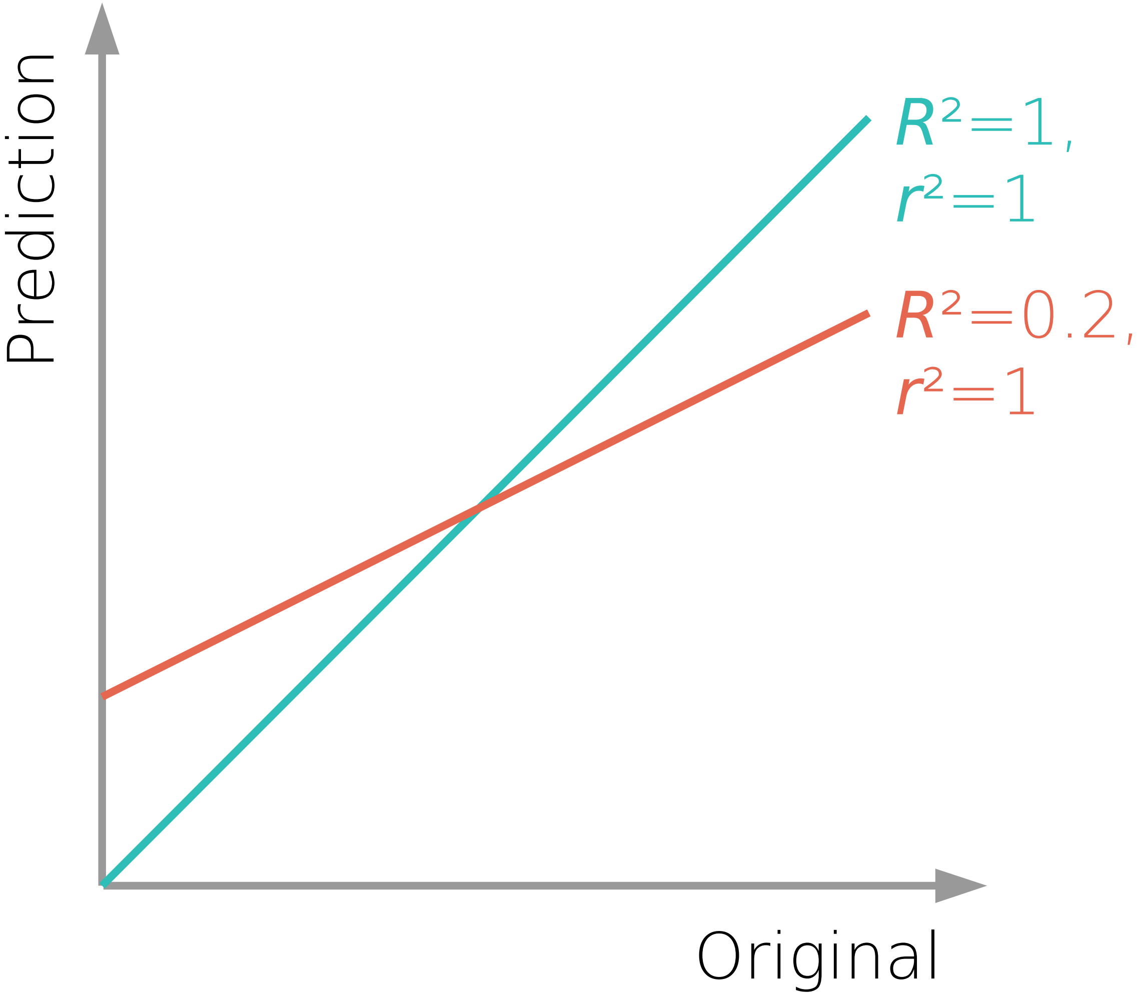
Benefits of the coefficient of determination, R^2



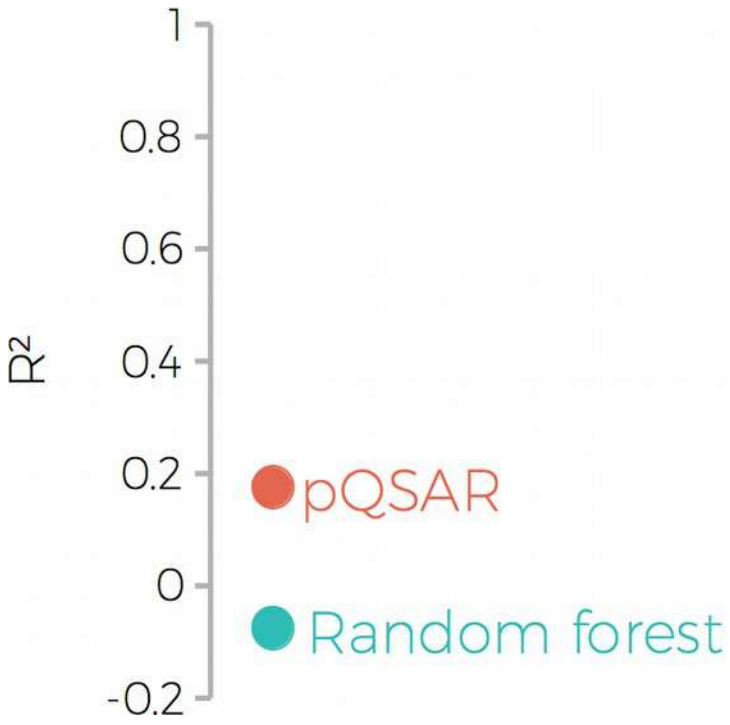
Benefits of the coefficient of determination, $\boldsymbol{R^2}$



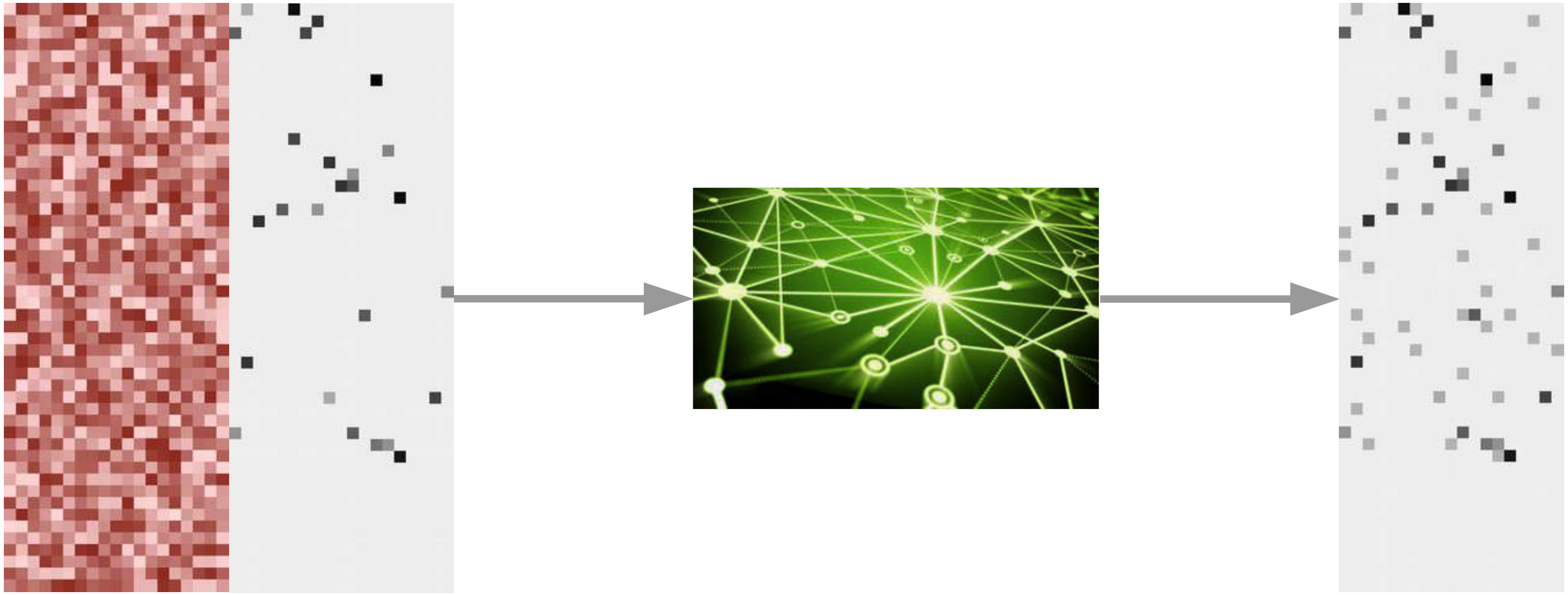
Benefits of the coefficient of determination, R^2



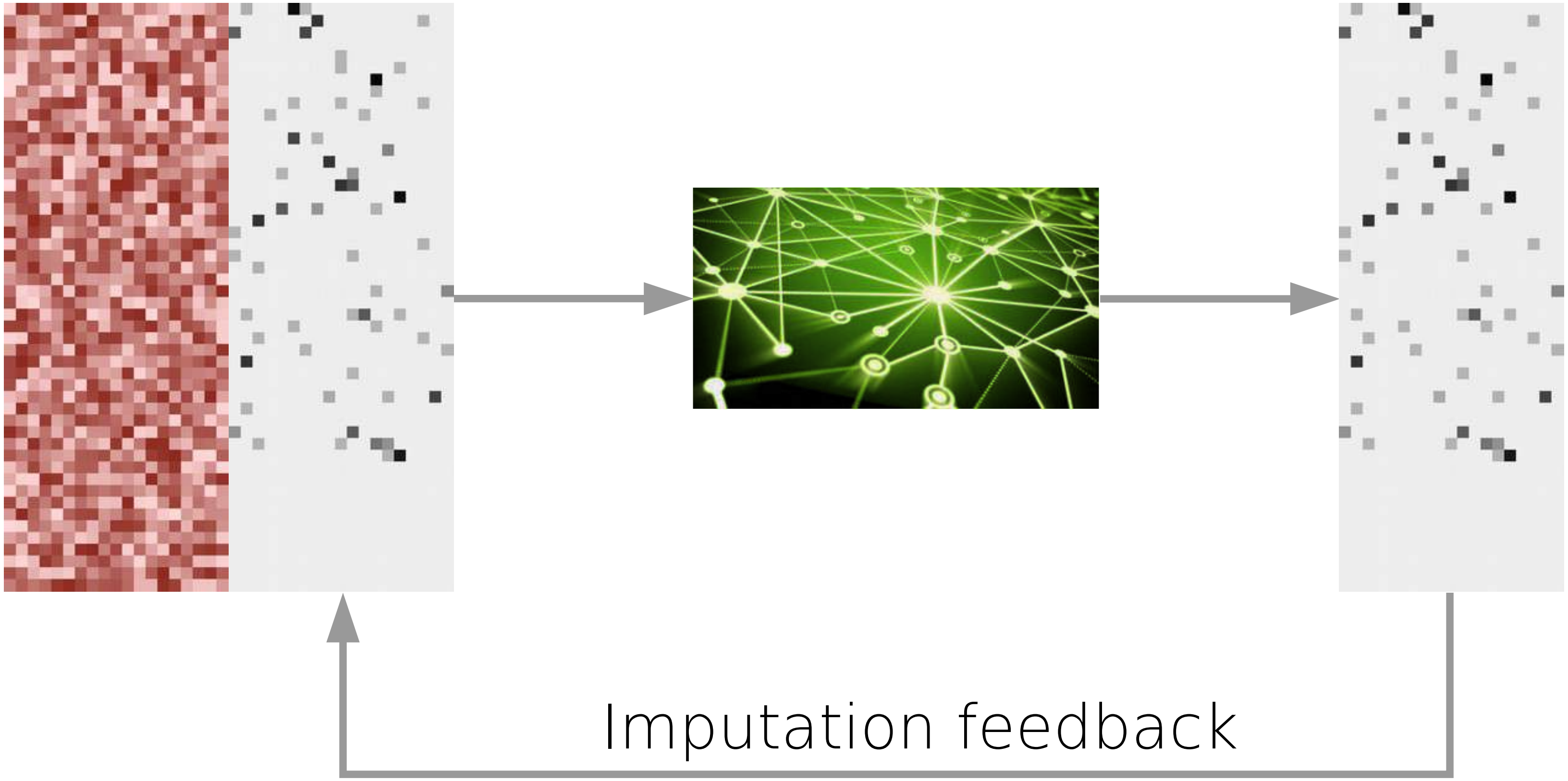
Predictions from pQSAR



QSAR: neural network can impute new activities

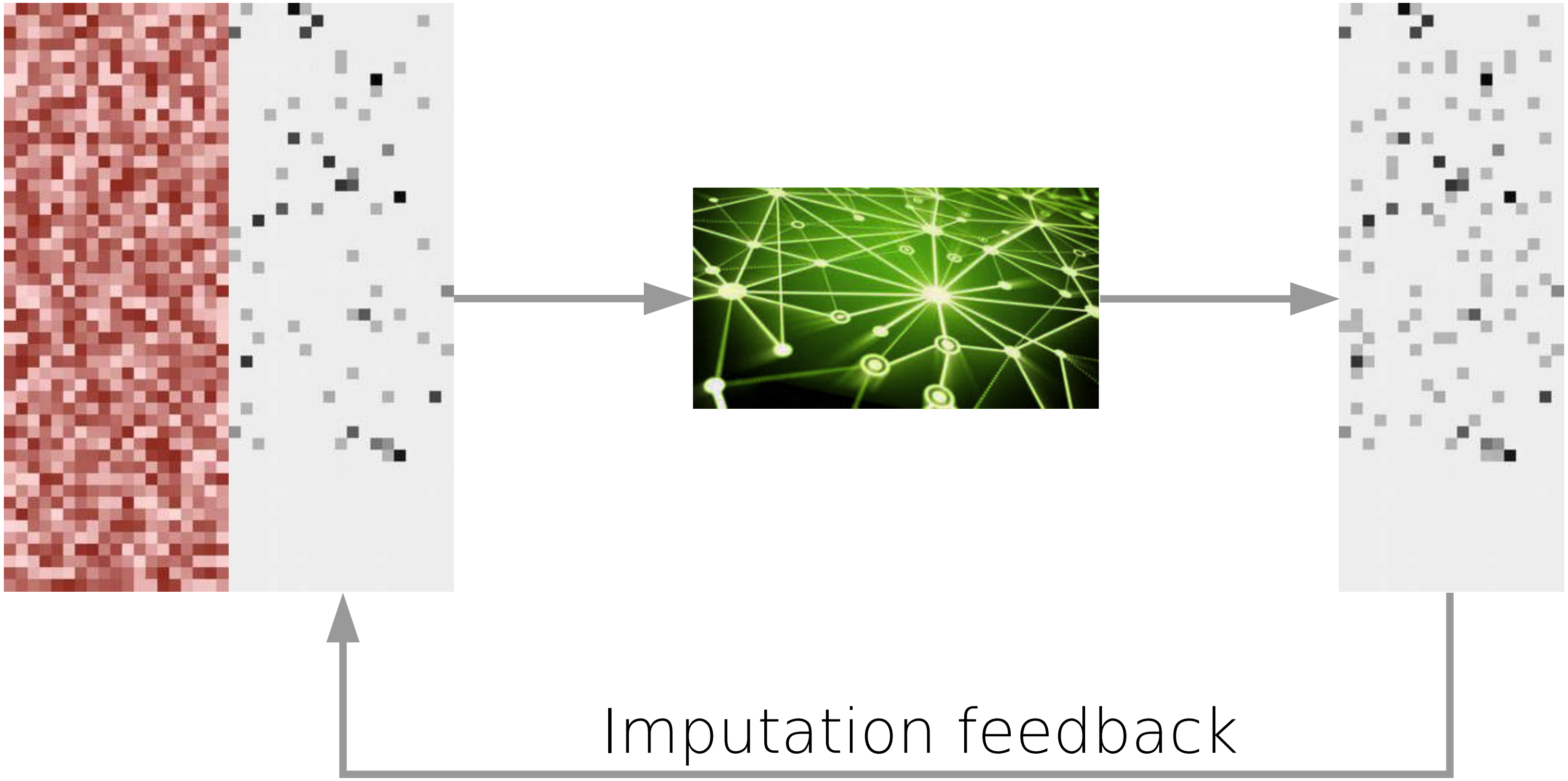


QSAR: neural network feedback loop



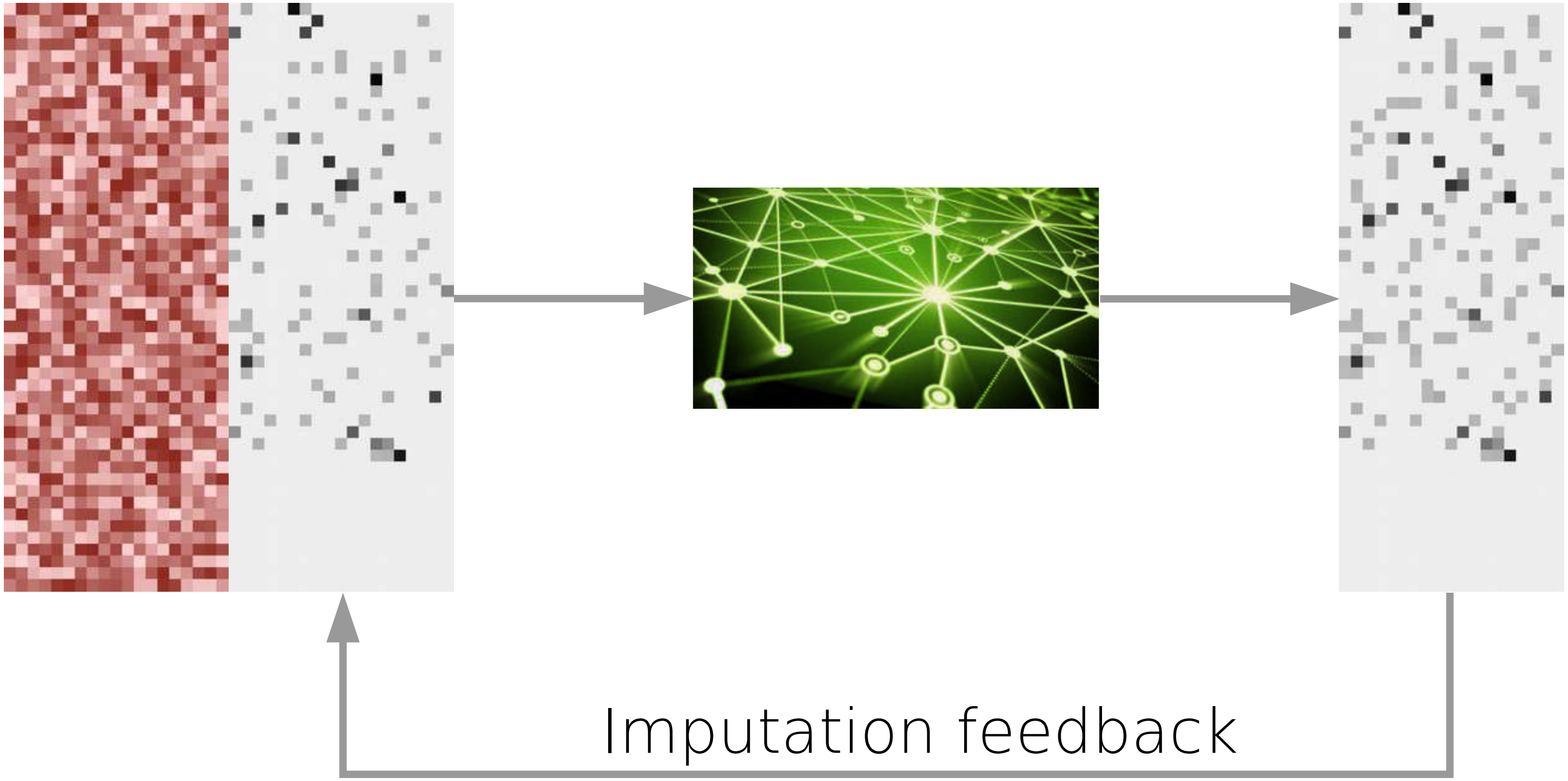
Imputation feedback

QSAR: neural network feedback loop



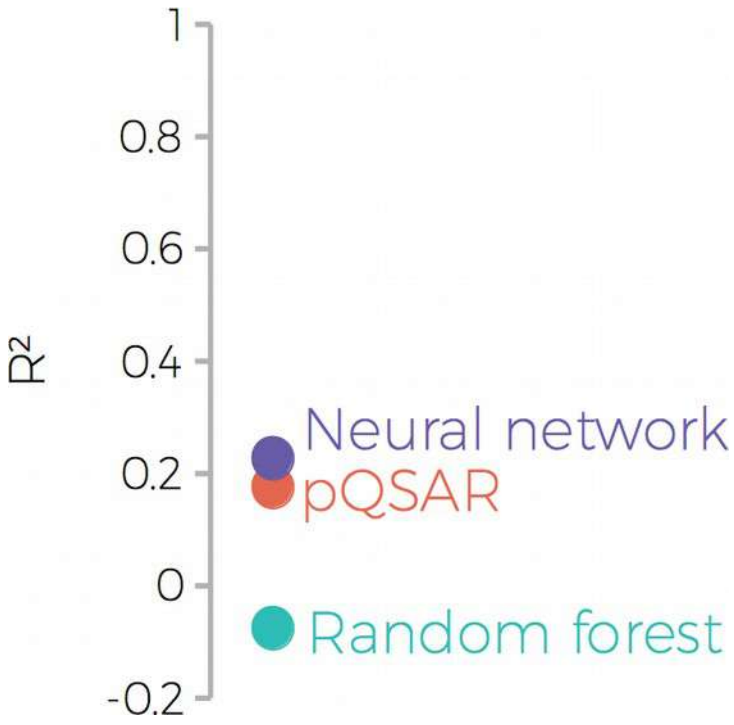
Imputation feedback

QSAR: neural network feedback loop

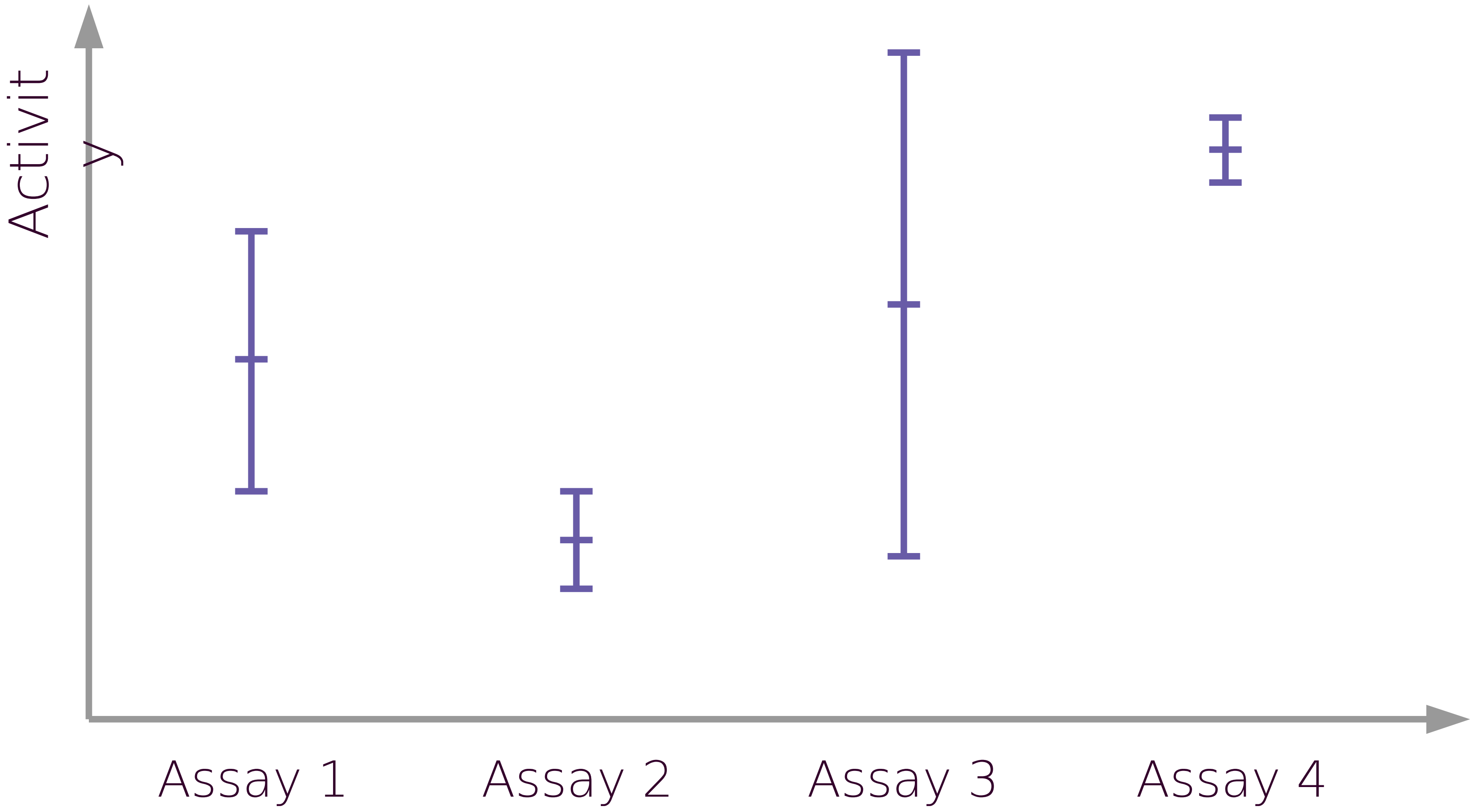


Imputation feedback

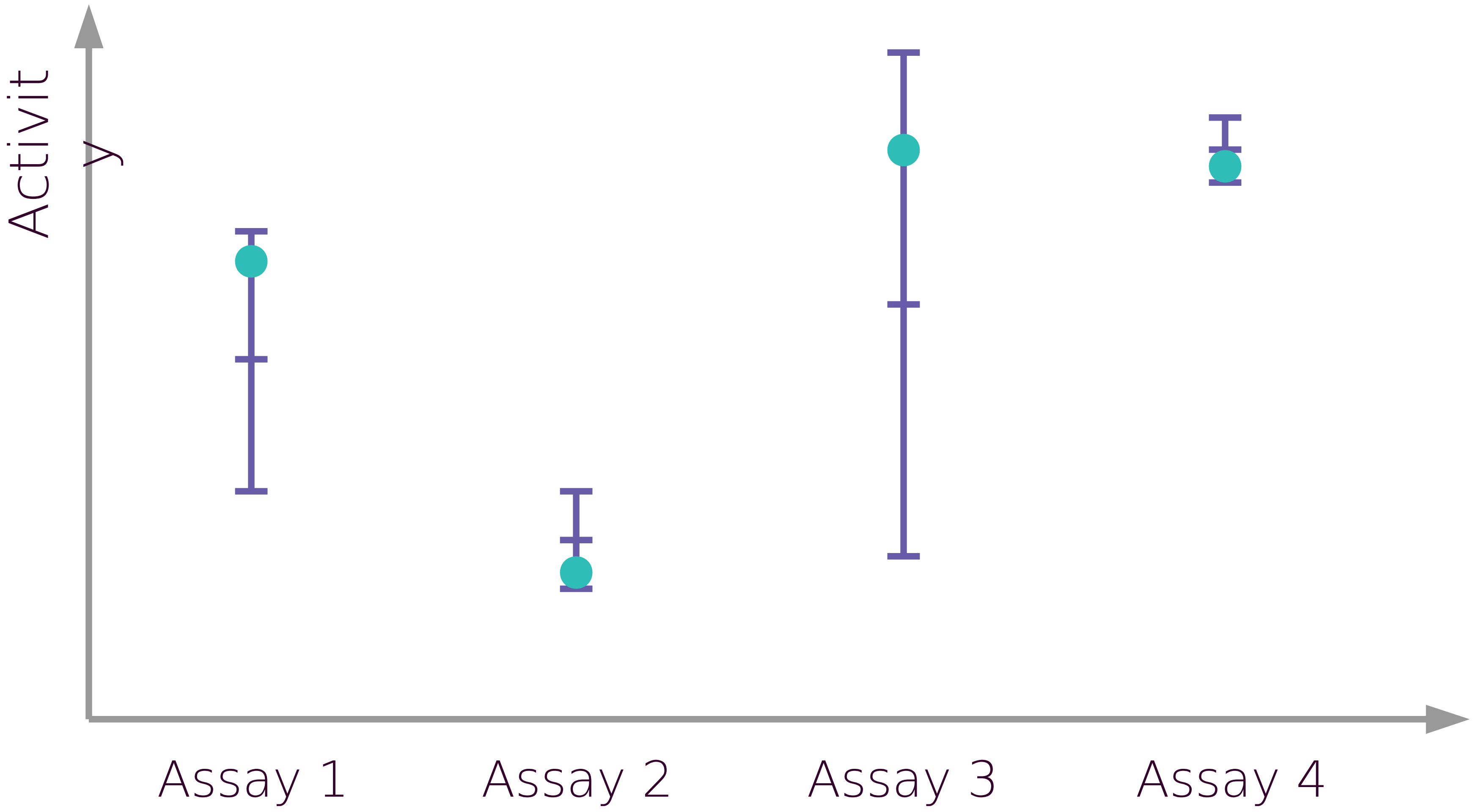
Predictions by the neural network



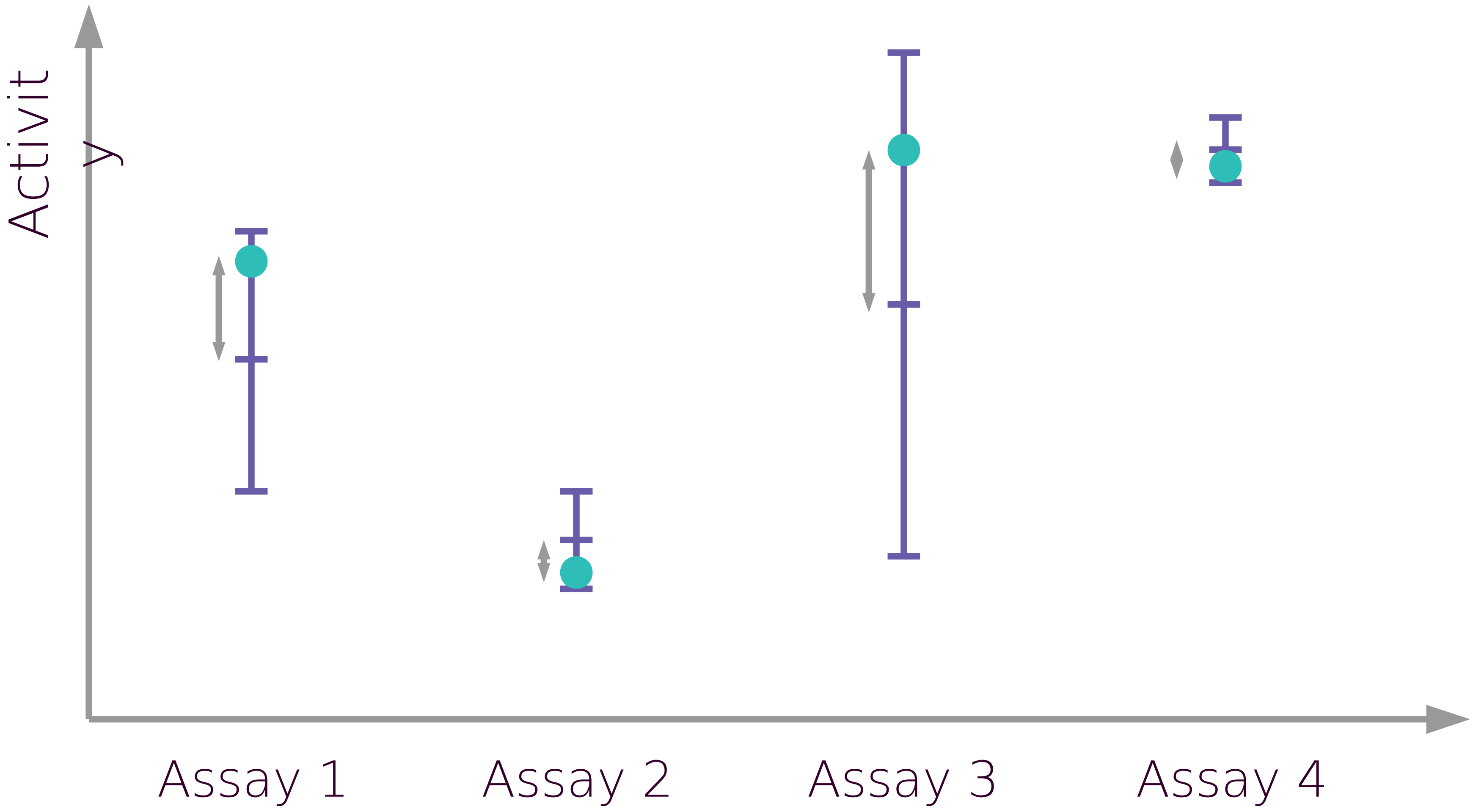
Predicted activities have an uncertainty



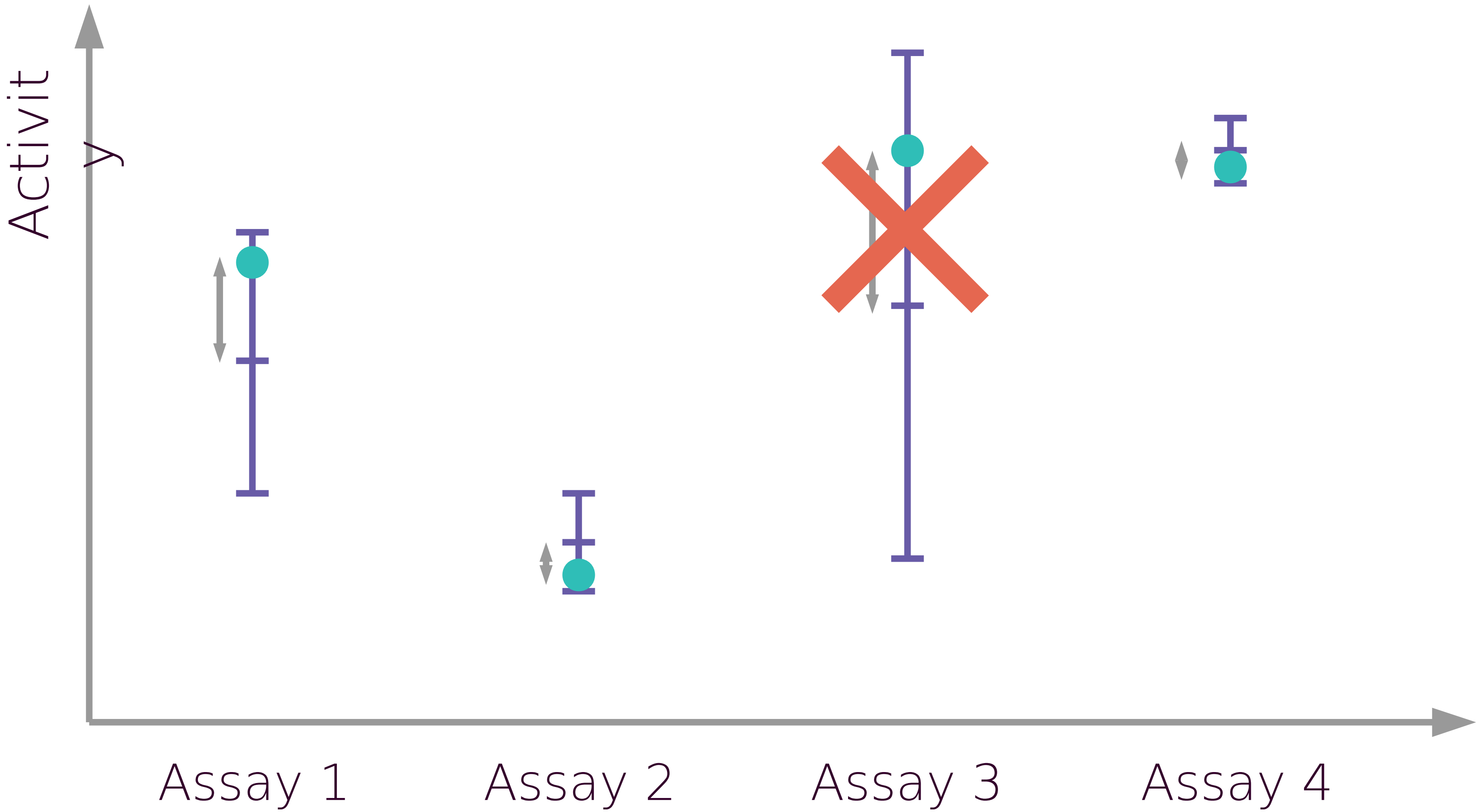
Validation data within one standard deviation



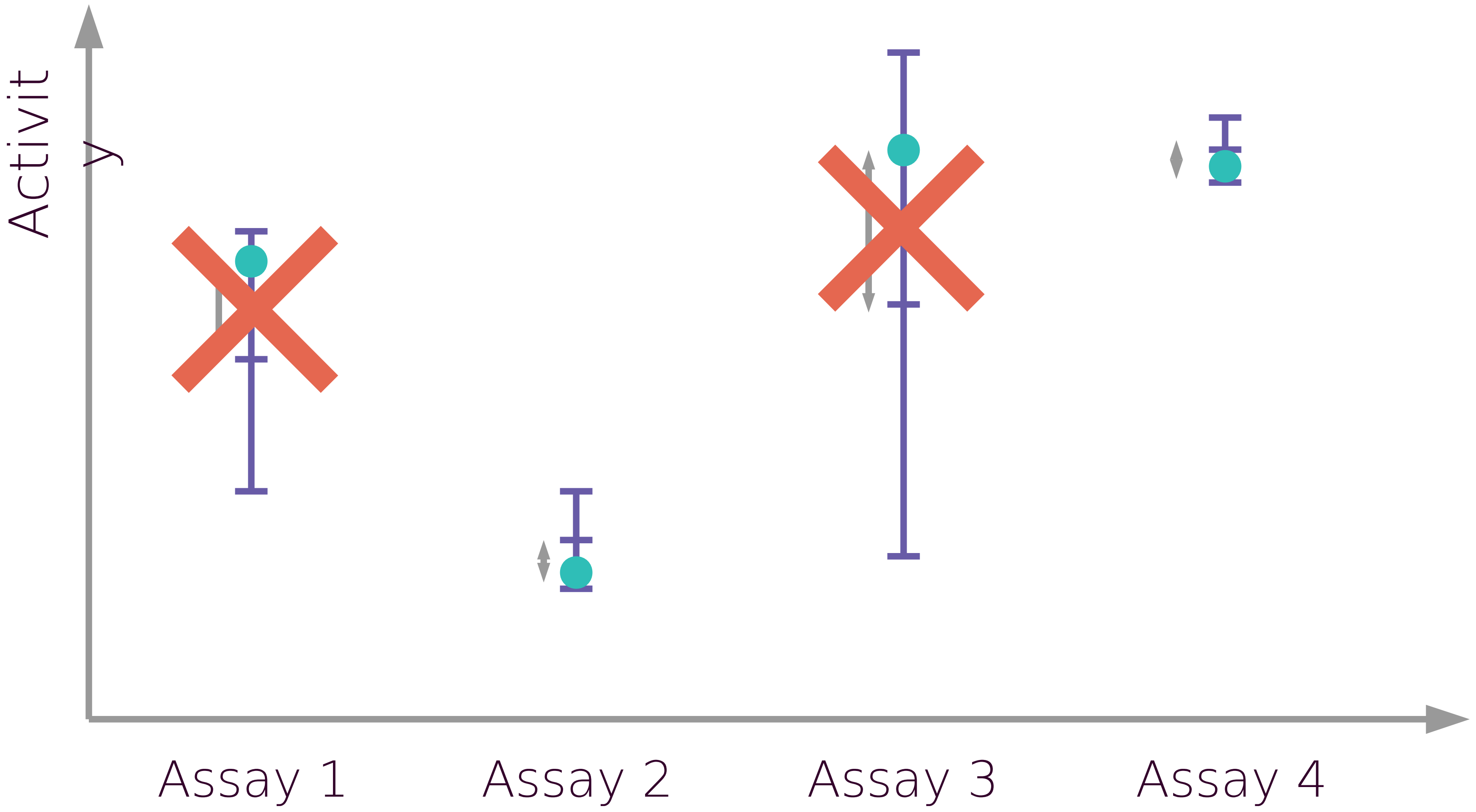
*R*² metric calculated with difference from mean



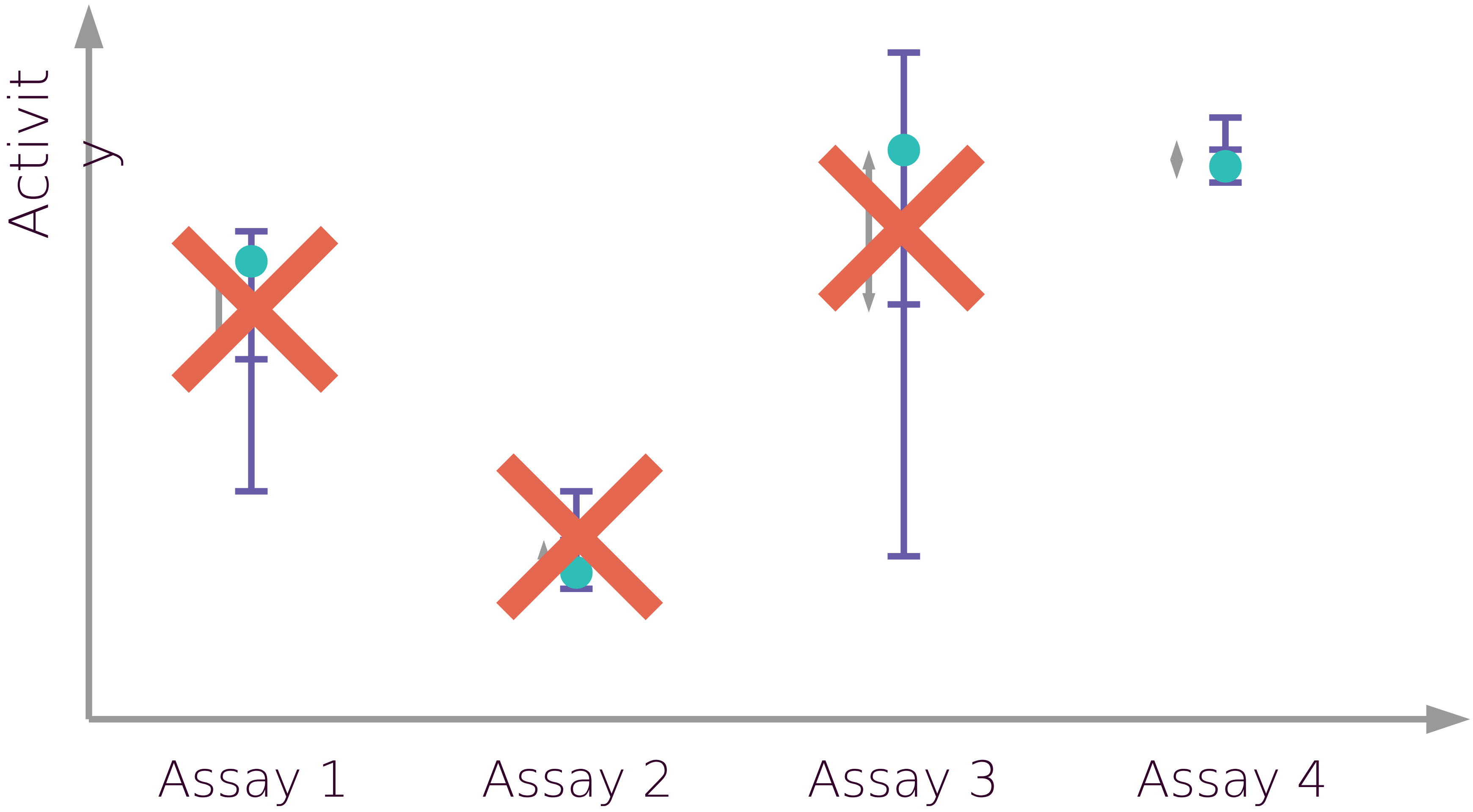
Impute 75% of data with smallest uncertainty



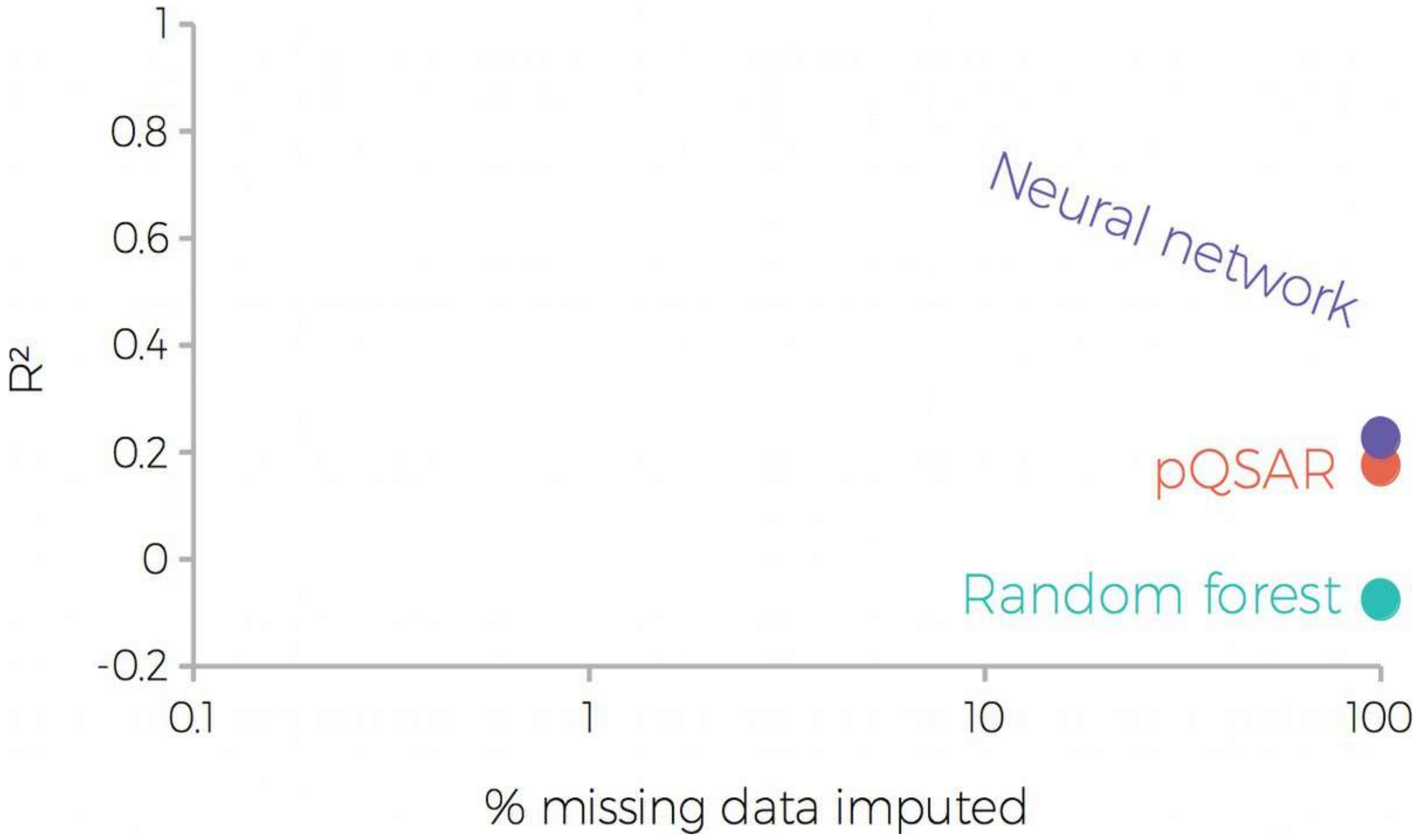
Impute 50% of data with smallest uncertainty



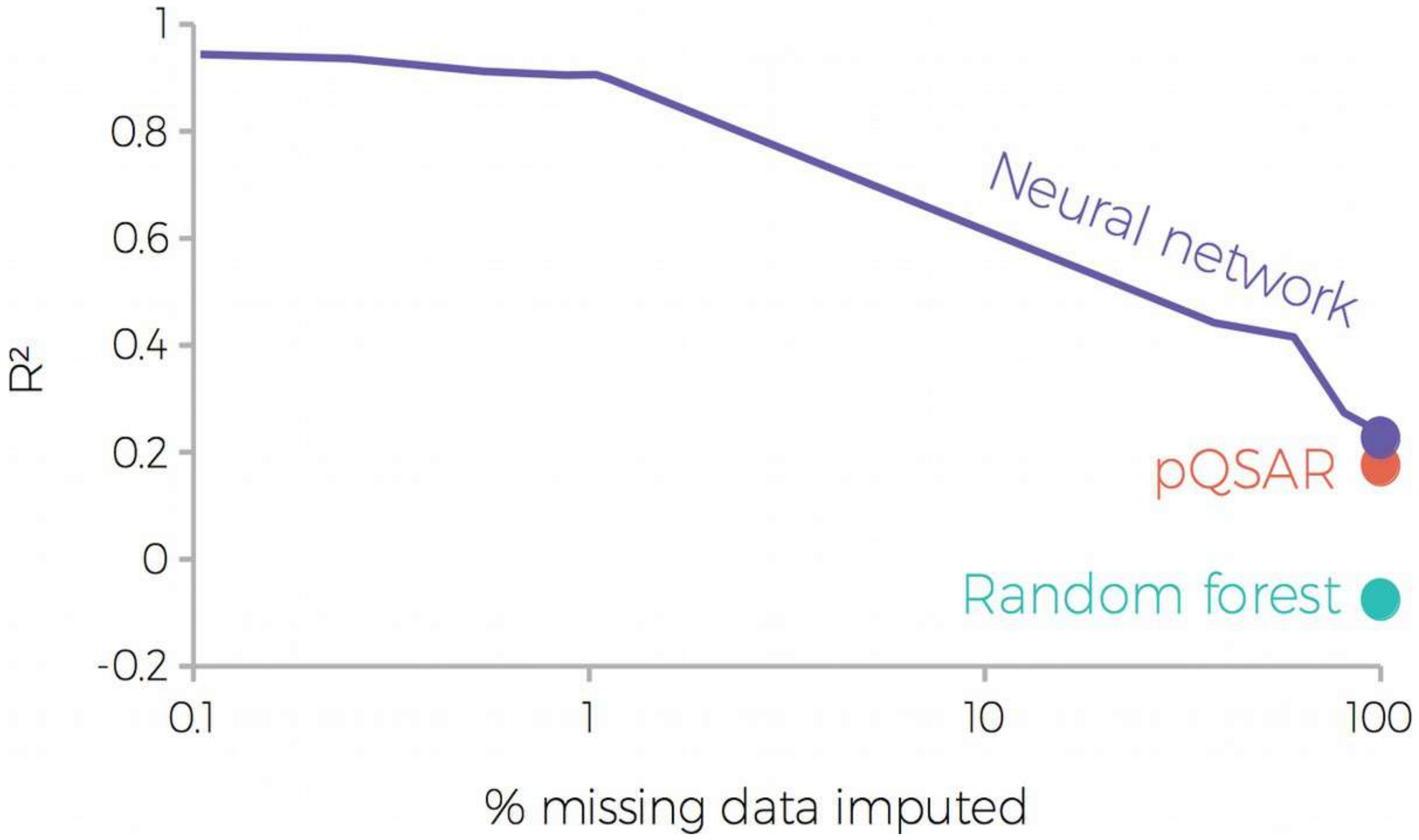
Impute 25% of data with smallest uncertainty



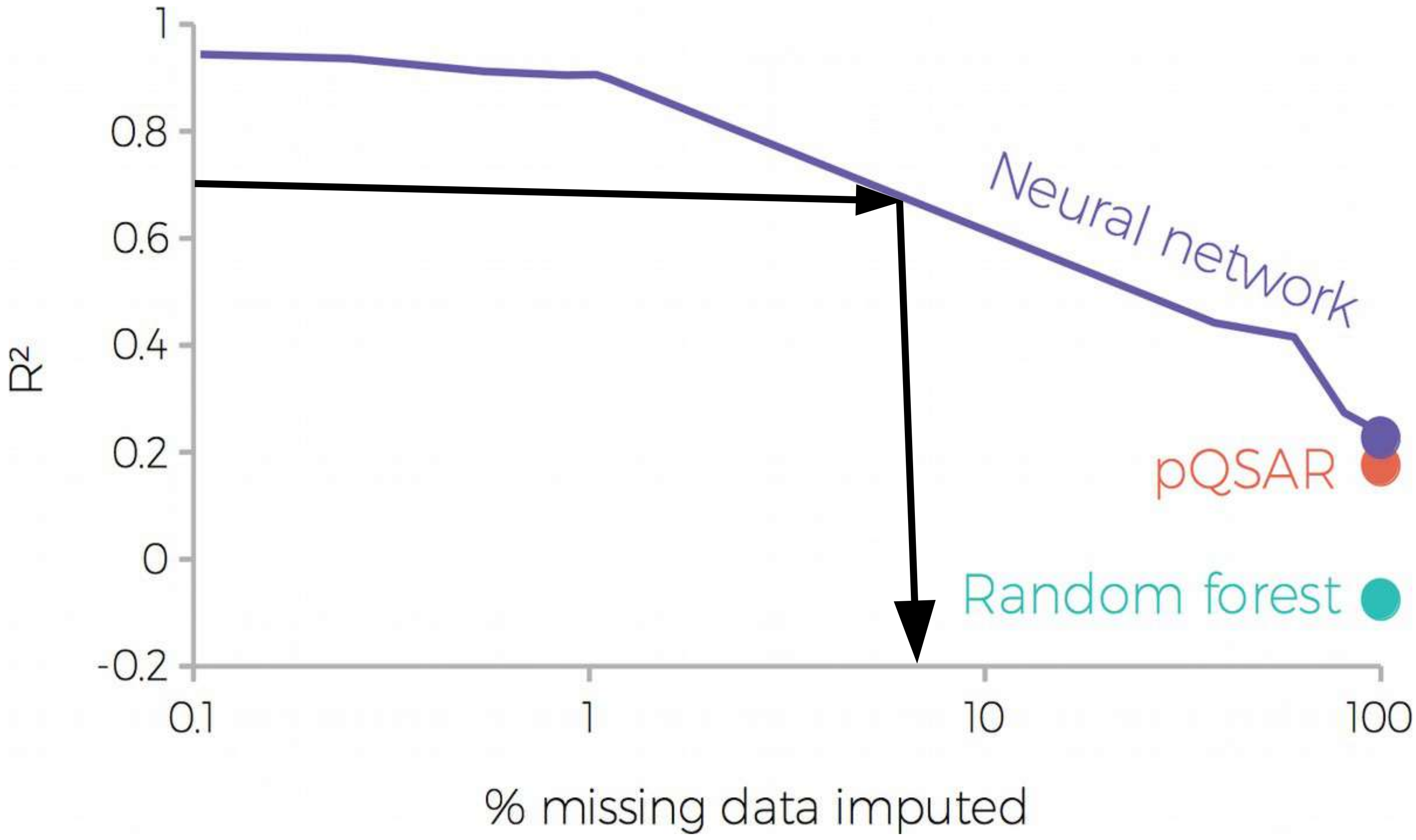
Improve performance by exploiting uncertainties



Improve performance by exploiting uncertainties



Improve performance by exploiting uncertainties



Summary

Train across all endpoints simultaneously to pull out activity-activity correlations

Impute values in sparse matrix to high accuracy, enables identification of new hits and activity profiling of compounds

Understand and exploit uncertainties to dial-in on most confident results

Combine all sources of information into a holistic imputation and design tool

Intellegens
gareth@intellegens.ai

optibrium
info@optibrium.com