Using deep neural networks with heterogeneous chemical data to support phenotypic assay campaigns

Antonio de la Vega de León

Information School
The University of Sheffield
www.sheffield.ac.uk/is
Member of the iSchools network
iSchools

The
University
Of
Sheffield.

Deep neural networks: challenges and opportunities

Antonio de la Vega de León

Information School
The University of Sheffield
www.sheffield.ac.uk/is
Member of the iSchools network
iSchools

The University Of Sheffield.

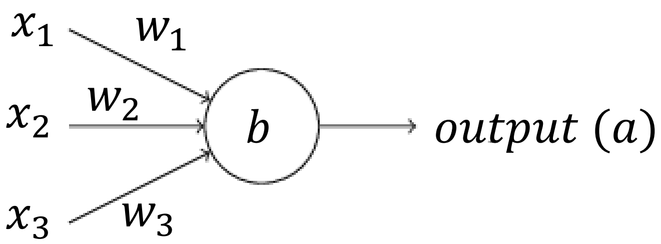
Deep neural networks

Hyperparameter selection

Effect of missing data

Deep neural networks

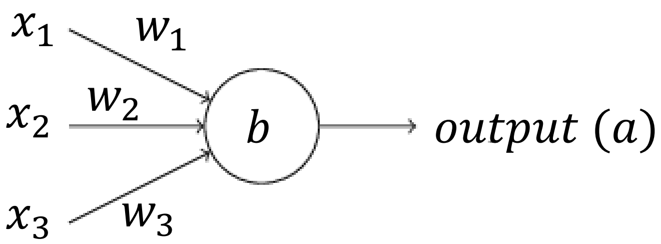
Deep neural networks (DNNs) are machine learning models based on simple, nonlinear units (neurons)



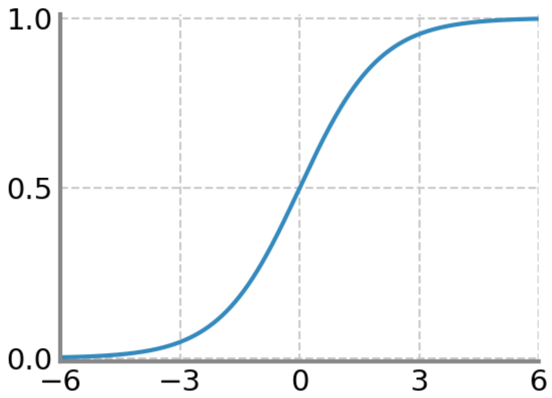
$$z = w \cdot x + b$$
$$a = f(z)$$

Deep neural networks

Deep neural networks (DNNs) are machine learning models based on simple, nonlinear units (neurons)



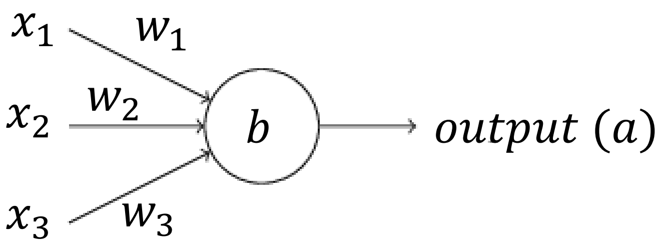
$$z = w \cdot x + b$$
$$a = f(z)$$



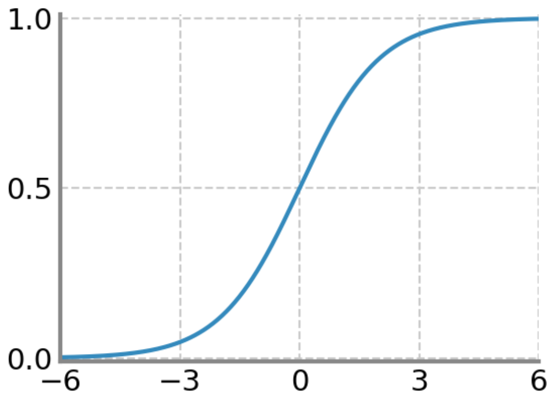
$$a = sigmoid(z) = \frac{1}{1+e^{-z}}$$

Deep neural networks

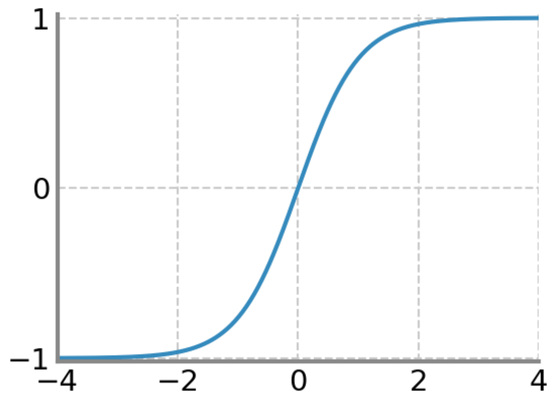
Deep neural networks (DNNs) are machine learning models based on simple, nonlinear units (neurons)



$$z = w \cdot x + b$$
$$a = f(z)$$



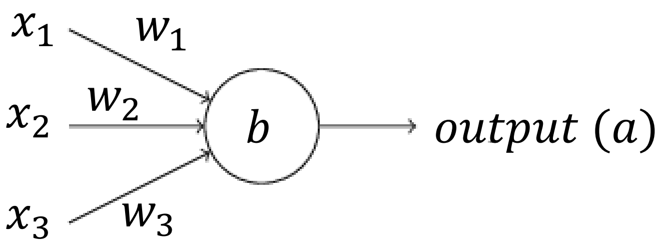
$$a = sigmoid(z) = \frac{1}{1+e^{-z}}$$



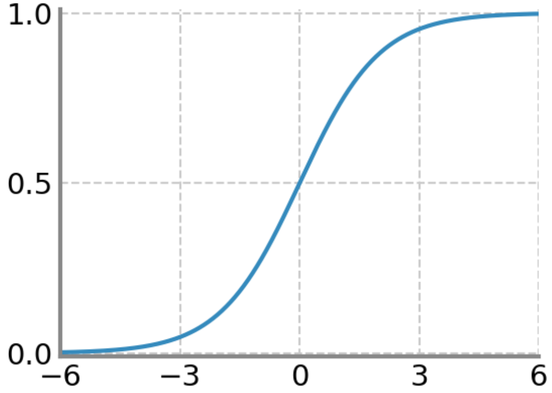
$$a = tanh(z) = \frac{e^{z}-e^{-z}}{e^{z}+e^{-z}}$$

Deep neural networks

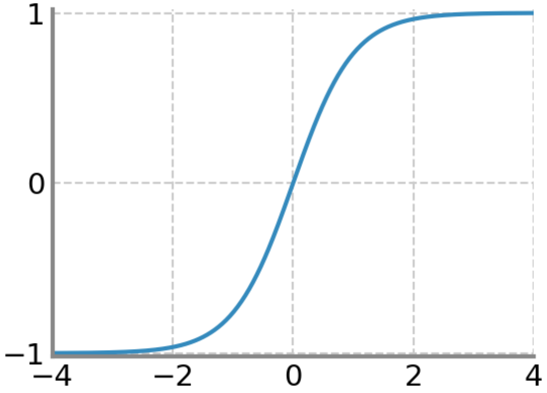
Deep neural networks (DNNs) are machine learning models based on simple, nonlinear units (neurons)



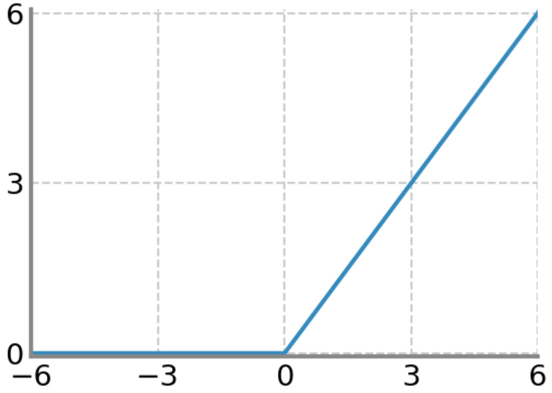
$$z = w \cdot x + b$$
$$a = f(z)$$



$$a = sigmoid(z) = \frac{1}{1+e^{-z}}$$



$$a = tanh(z) = \frac{e^{z}-e^{-z}}{e^{z}+e^{-z}}$$



$$a = relu(z) = \max(0, z)$$

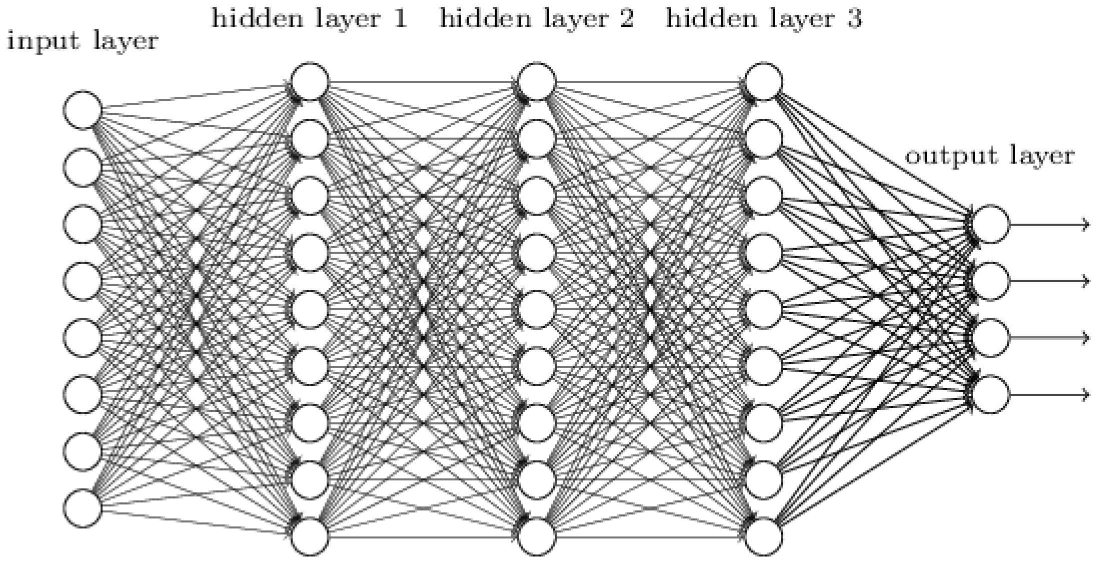
Information School
The University of Sheffield
www.sheffield.ac.uk/is
Member of the iSchools network
iSchools

The University Of Sheffield.

Deep neural networks

Neurons are organized into layers, where neurons in one layer are connected to all neurons in the previous and the next layer

The input layer represents the data descriptors while the output layer has a neuron per prediction task

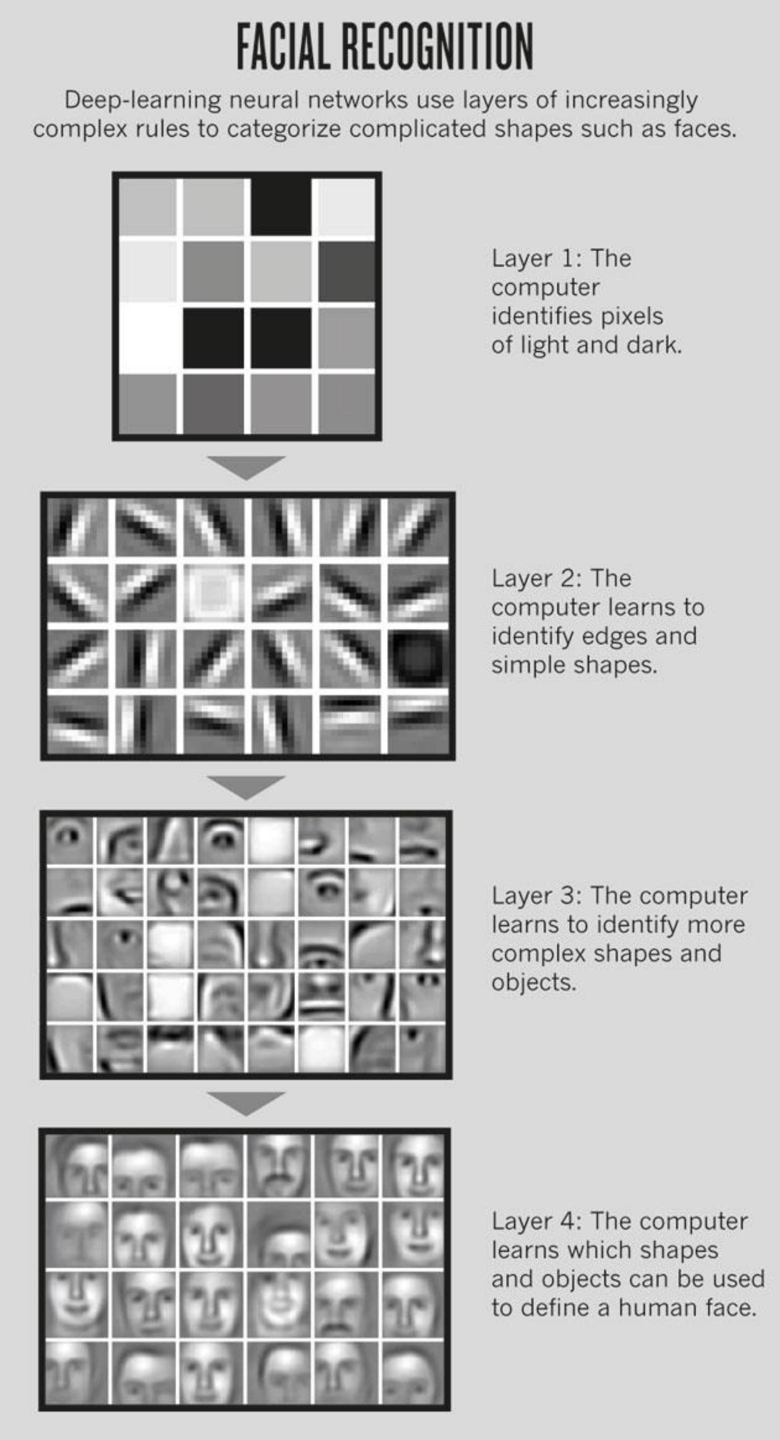


Deep learning

The idea behind deep learning is that of a hierarchical learning process

Early layers identify simple patterns that latter layers use to learn complex patterns

This reduces the need for feature selection or complex feature engineering



FACIAL RECOGNITION

Deep-learning neural networks use layers of increasingly complex rules to categorize complicated shapes such as faces.

Layer 1: The computer identifies pixels of light and dark.

Layer 2: The computer learns to identify edges and simple shapes.

Layer 3: The computer learns to identify more complex shapes and objects.

Layer 4: The computer learns which shapes and objects can be used to define a human face.

Jones N. *Nature* 505, 146-148

Deep neural networks

Hyperparameter selection

Effect of missing data

Hyperparameter selection

DNNs usually require a large number of hyperparameters to be set in advance:

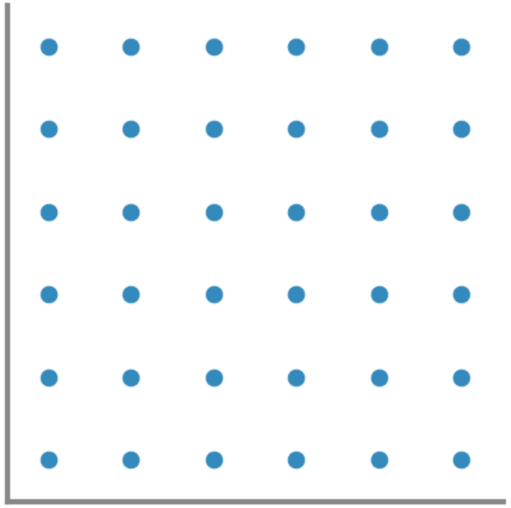
- Optimizer functions
- Learning rate
- Activation function
- Number of neurons
- Number of hidden layers
- Dropout amount
- Batch size
- Number of training steps
- Weight initialization

How best to find the combination that gives highest performance?

Search strategies

Grid search

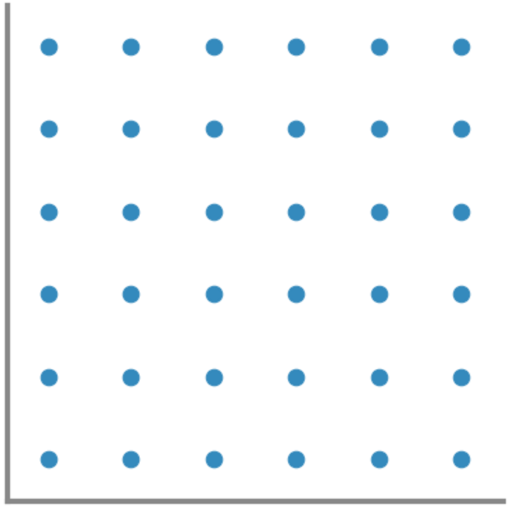
- For each parameter, a set of values is fixed and all possible parameter combinations are tested
- Impractical for large number of parameters



Search strategies

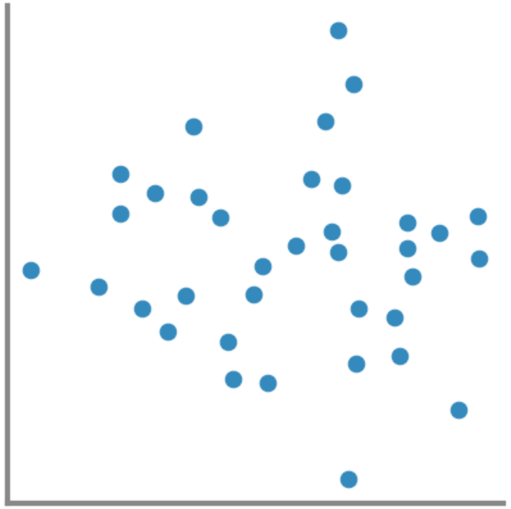
Grid search

- For each parameter, a set of values is fixed and all possible parameter combinations are tested
- Impractical for large number of parameters



Random search

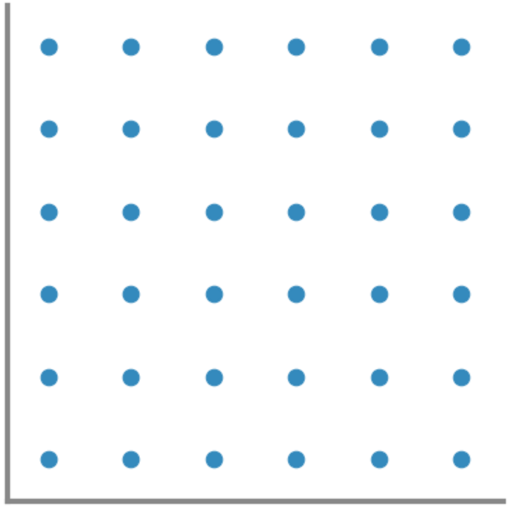
- For each parameter, a value range is fixed and at each iterations random numbers are used
- Can require large number of iterations to cover parameter space adequately



Search strategies

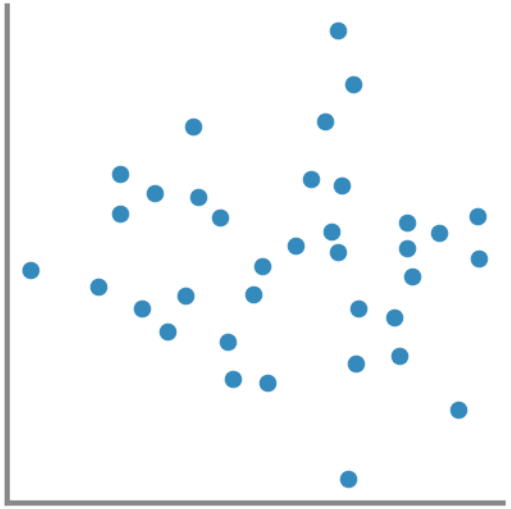
Grid search

- For each parameter, a set of values is fixed and all possible parameter combinations are tested
- Impractical for large number of parameters



Random search

- For each parameter, a value range is fixed and at each iterations random numbers are used
- Can require large number of iterations to cover parameter space adequately



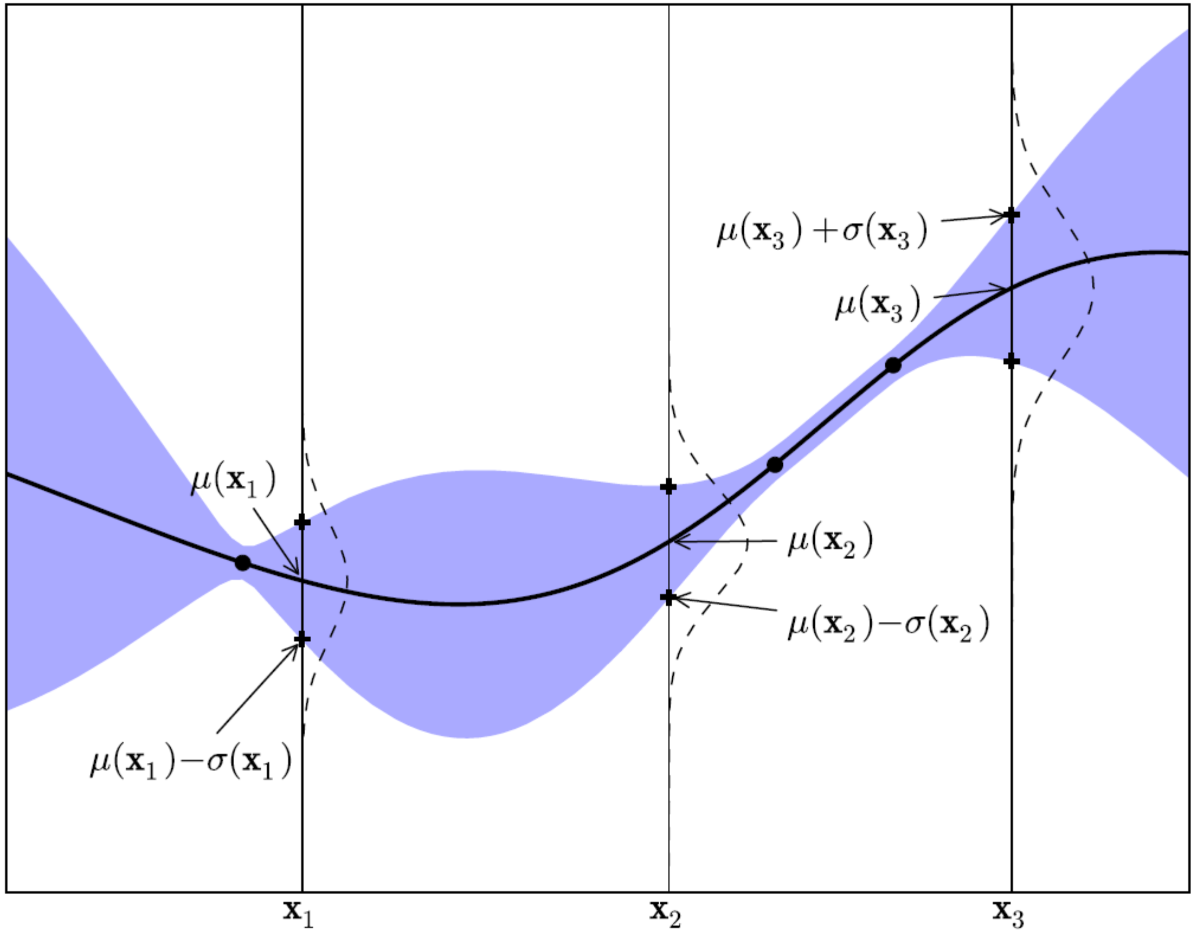
Bayesian optimization

Bayesian optimization

Methodology to find the optimum of a black-box function over a range of parameter values

Based on previous evaluations, it builds a probabilistic representation of the function (Gaussian Process or GP)

Based on the GP, a utility function determines the best point in parameter space to test in the next iteration



$\mu(\mathbf{x}_3)+\sigma(\mathbf{x}_3)$
$\mu(\mathbf{x}_3)$
$\mu(\mathbf{x}_1)$
$\mu(\mathbf{x}_2)$
$\mu(\mathbf{x}_2)-\sigma(\mathbf{x}_2)$
$\mu(\mathbf{x}_1)-\sigma(\mathbf{x}_1)$
$\mathbf{x}_1$
$\mathbf{x}_2$
$\mathbf{x}_3$

Brochu E, Cora VM & de Freitas N. *arXiv* 1012.2599

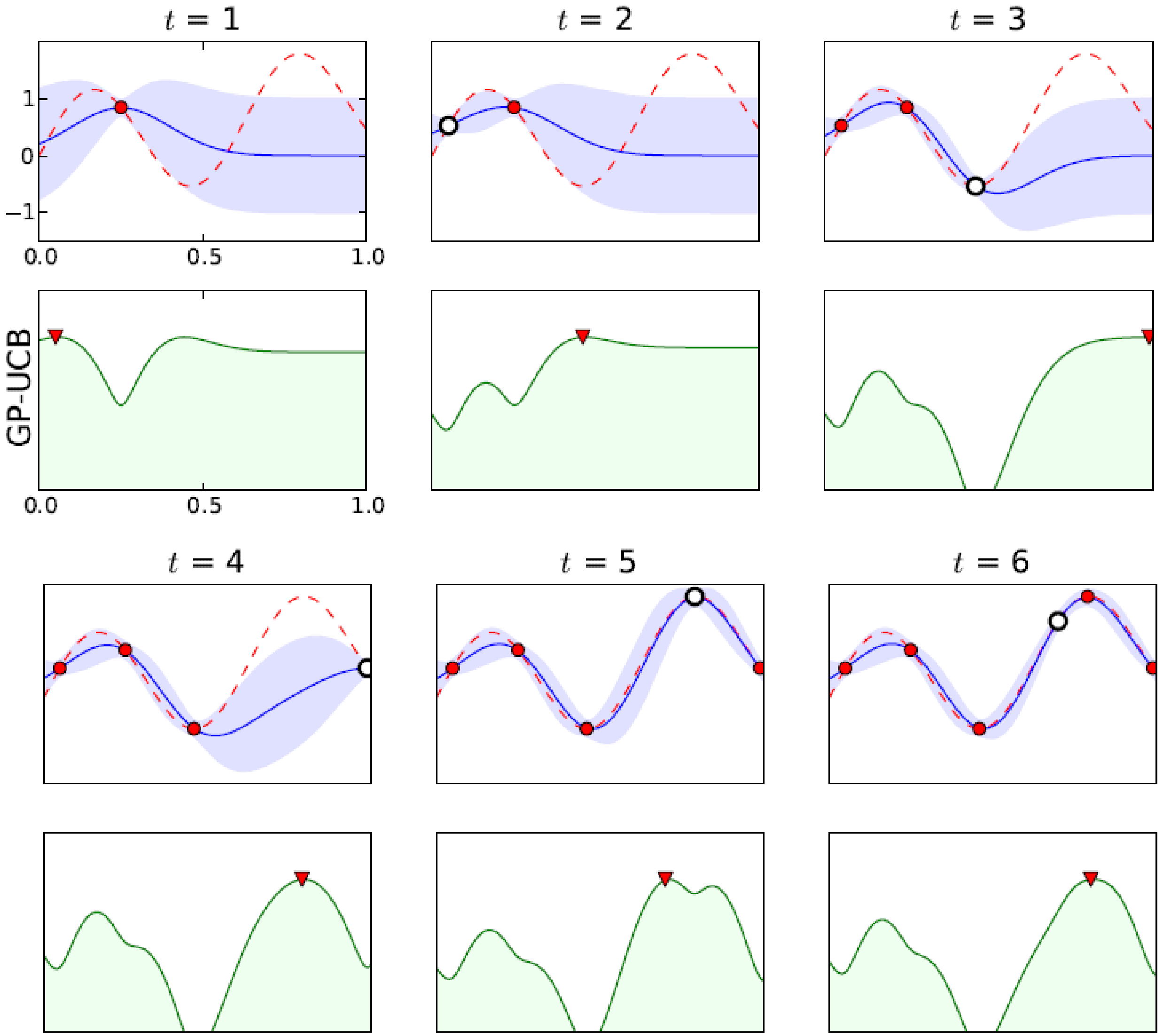
Bayesian optimization

Utility functions guide the search for the optimum:

- Probability of Improvement
- Expected Improvement
- GP-Upper Confidence Bounds

They all have a parameter that balances exploration/exploitation

At each step, the GP becomes better at representing the real formula



Brochu E, Cora VM & de Freitas N. *arXiv* 1012.2599

Bayesian optimization test

Tox21 data: collection of 12 toxicity assays with active/inactive labels

Random parameter selection vs Bayesian optimization with the 3 functions (POI, EI, and UCB) at values: 10^{-7}, 10^{-5}, 10^{-3}, 10^{-1}, or 10^{1}

200 DNN models per run were trained where the parameters varied:

- Optimizer algorithm: Adagrad, Adam, Ftrl, RMSProp, or SGD
- Learning rate: 10^{-7}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10^{1}, or 10^{3}

Performance is measured as area under the ROC curve (AUC)

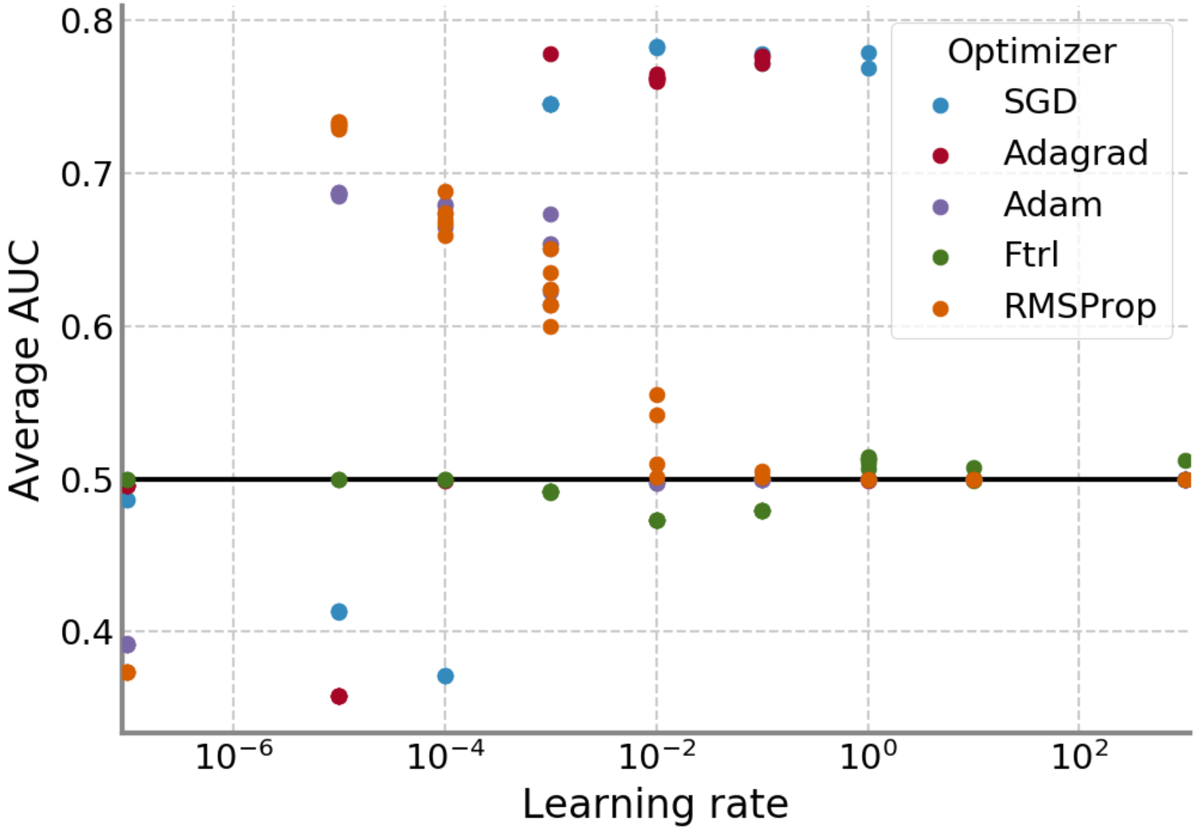
https://tripod.nih.gov/tox21/challenge/

Bayesian optimization test

The effect of both of these parameters are heavily related

Many combinations of optimizer and learning rate lead to models with random or worse performance

The winner of the Tox21 data challenge achieved an average AUC value of 0.84



Mayr A *et al. Front. Environ. Sci.* 3:80

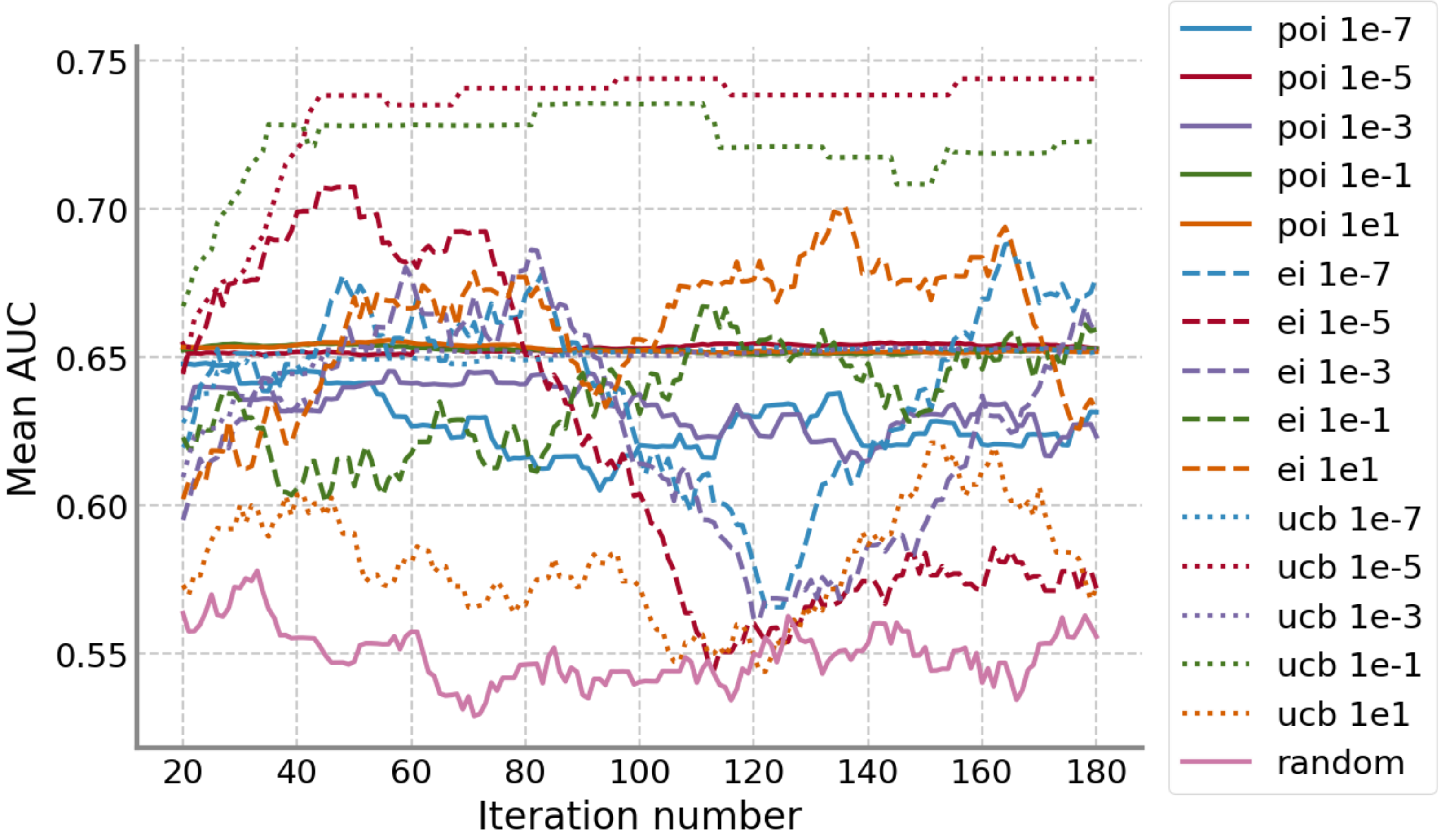
Bayesian optimization test

Random search has very low AUC values throughout

Results of Bayesian optimization are really mixed

Some runs have very stable performance, but far from the optimum

Most runs find good values but do not hold there



Deep neural networks

Hyperparameter selection

Effect of missing data

Incomplete data sets

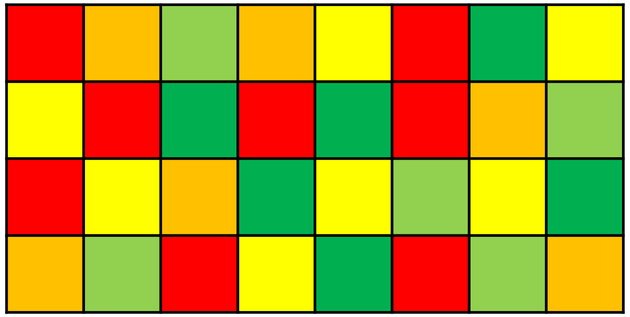
When combining activity records from several targets/assays in a pathway, the resulting activity matrix is usually incomplete

It has been stated that even small data sets can help in preventing bias and generalizing the networks internal representation

We were interested to test how performance deteriorates with increasing sparseness of activity labels

Ramsundar B *et al. arXiv* 1502.02072

Incomplete data sets



When combining activity records from several targets/assays in a pathway, the resulting activity matrix is usually incomplete

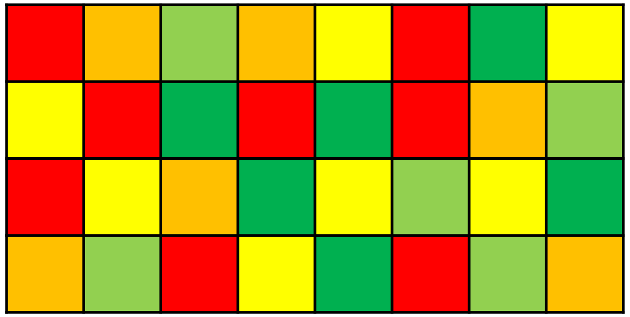
It has been stated that even small data sets can help in preventing bias and generalizing the networks internal representation

We were interested to test how performance deteriorates with increasing sparseness of activity labels

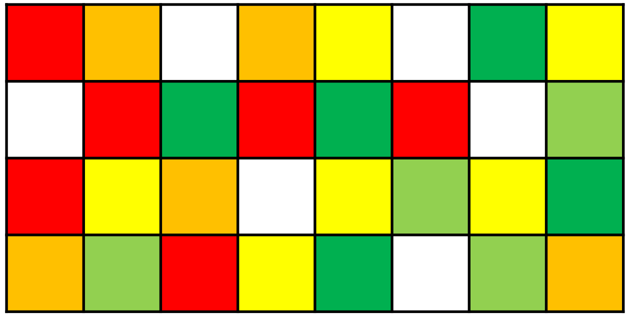
Ramsundar B *et al. arXiv* 1502.02072

Incomplete data sets

When combining activity records from several targets/assays in a pathway, the resulting activity matrix is usually incomplete



It has been stated that even small data sets can help in preventing bias and generalizing the networks internal representation



We were interested to test how performance deteriorates with increasing sparseness of activity labels

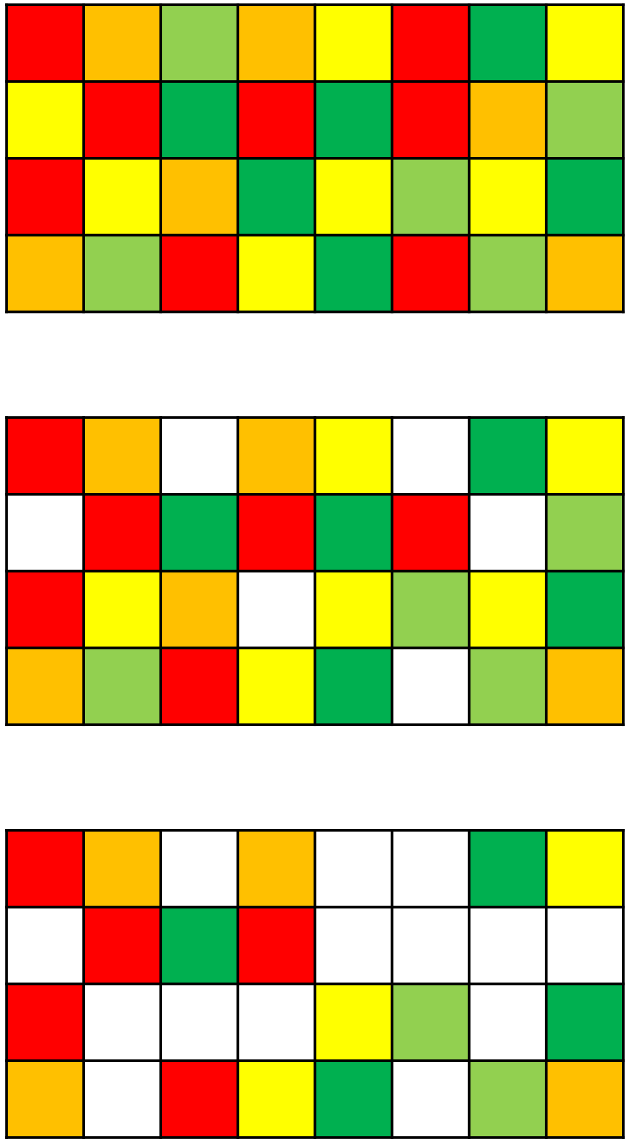
Ramsundar B *et al. arXiv* 1502.02072

Incomplete data sets

When combining activity records from several targets/assays in a pathway, the resulting activity matrix is usually incomplete

It has been stated that even small data sets can help in preventing bias and generalizing the networks internal representation

We were interested to test how performance deteriorates with increasing sparseness of activity labels



Ramsundar B *et al. arXiv* 1502.02072

Testing data sparseness

PKIS (GSK Published Kinase Inhibitor Set) data: percent inhibition of ~ 370 compounds in 454 kinase assays (for most kinases there are values at 1 and 0.1μM)

¼ of the compounds are used as test set and the rest is used for training

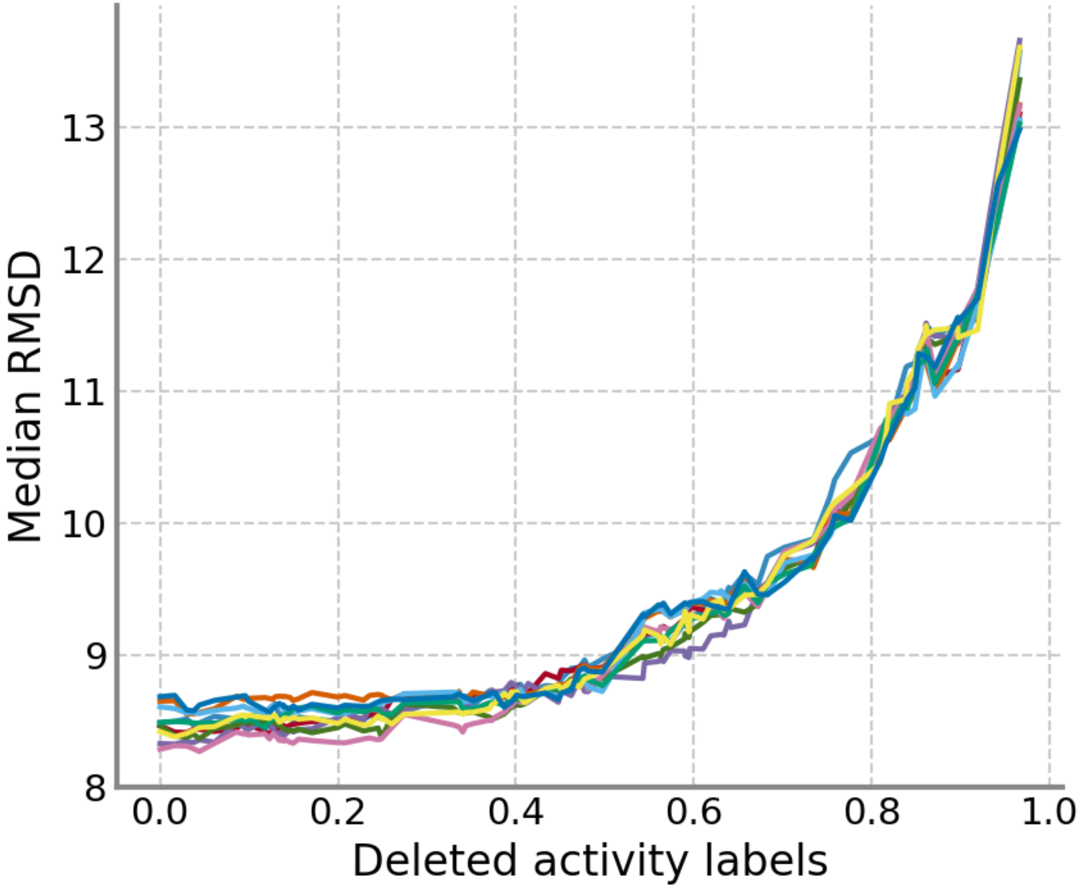
10 sets of DNN hyperparameter values were set, and for each set 100 DNN models were trained with different amounts of training labels removed

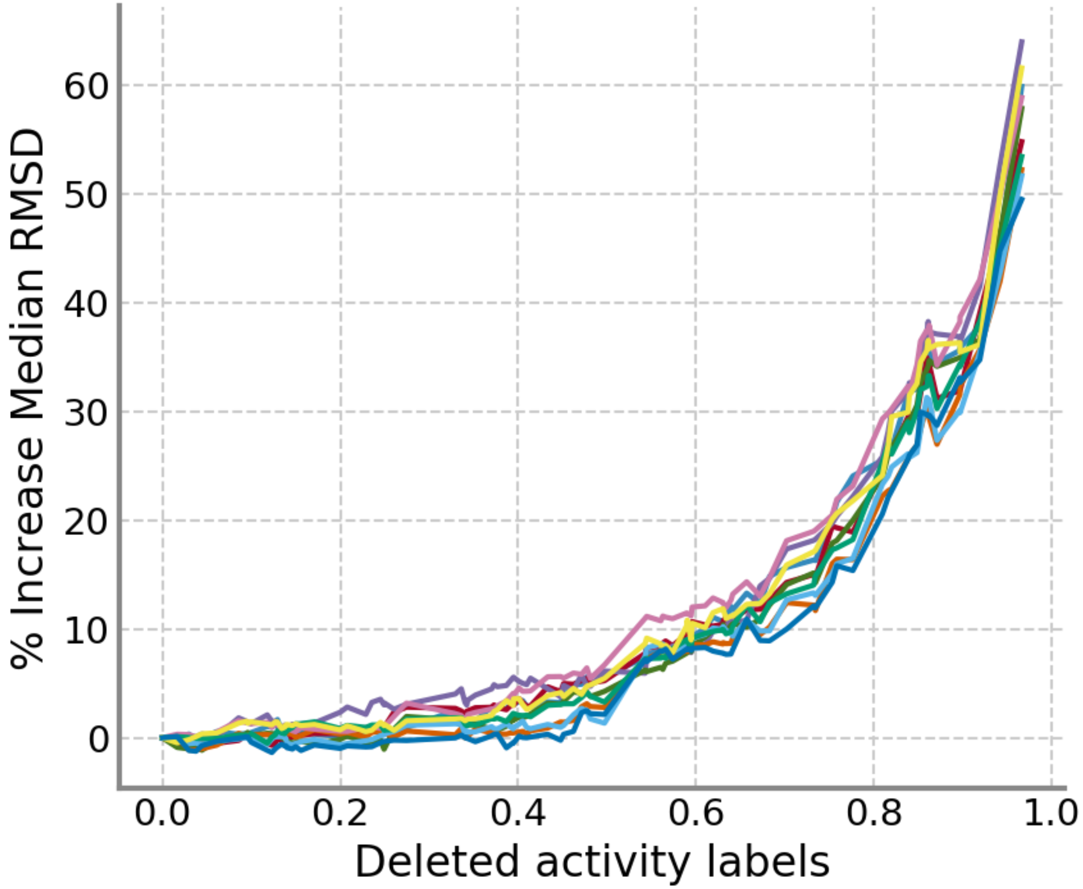
Performance was measured as RMSD $RMSD = \sqrt{\frac{\sum(y_i - \widehat{y_i})^2}{n}}$

https://www.ebi.ac.uk/chembldb/extra/PKIS/
Knapp S et al. *Nat. Chem. Biol.* 9, 3–6

Testing data sparseness

Median performance holds well even if half of the training data is removed

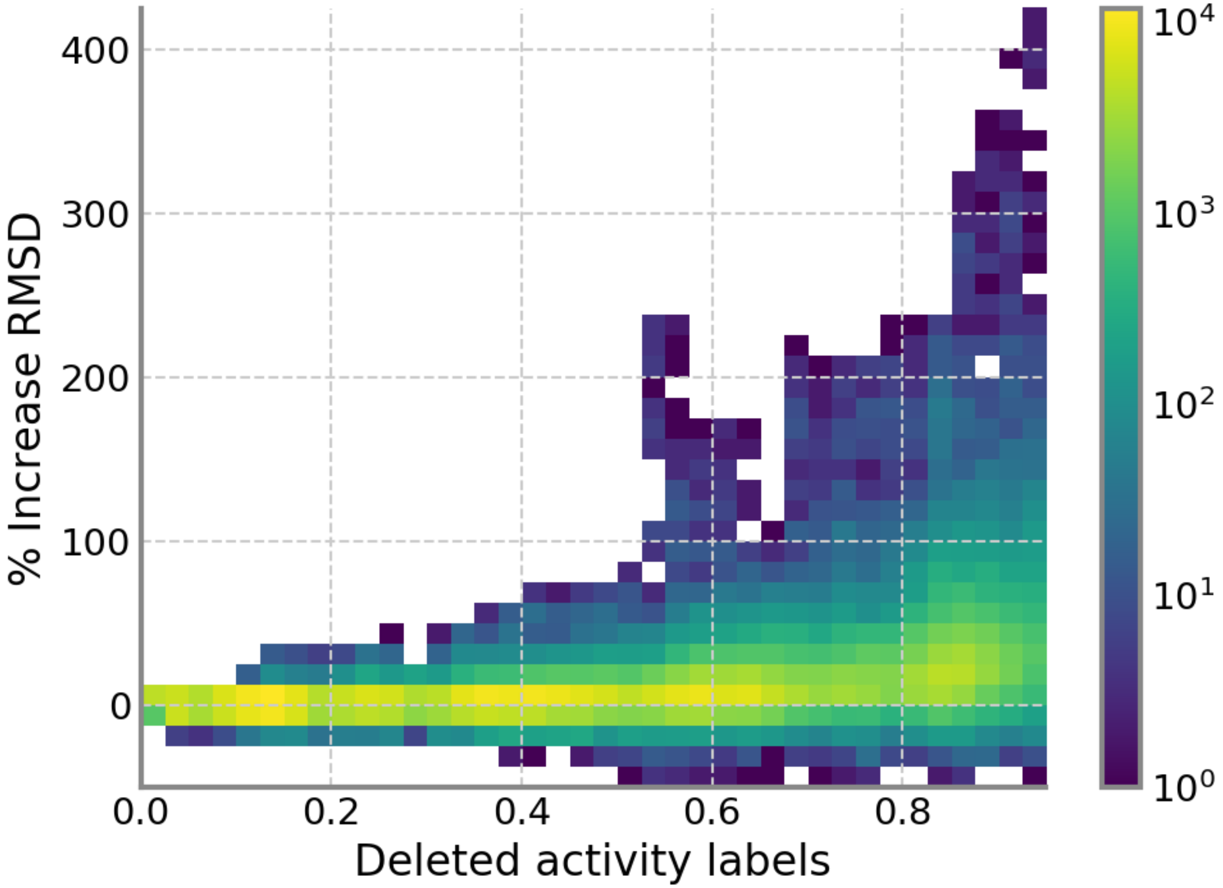




Testing data sparseness

Median performance masks large differences between individual prediction tasks

Most tasks see small differences even with large decreases of training data

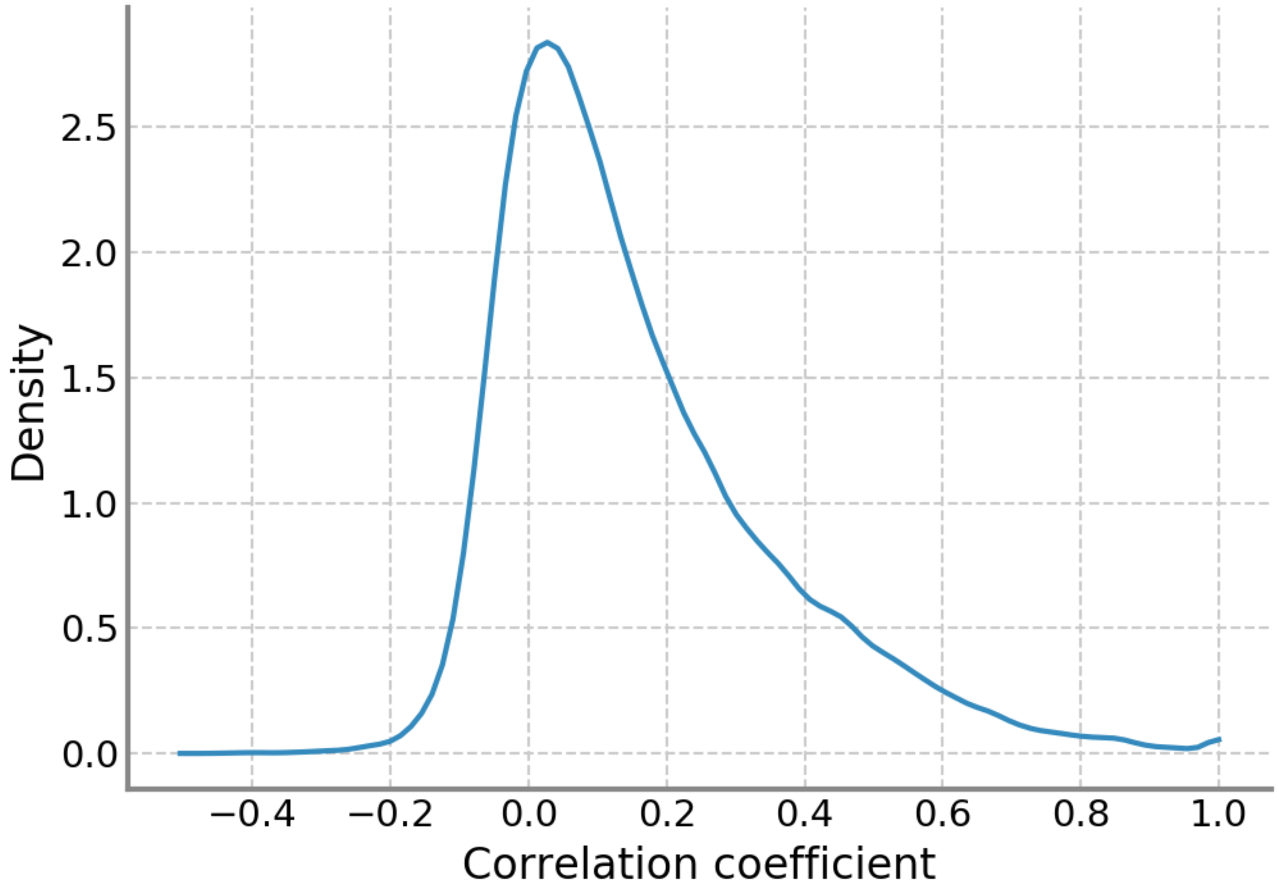


Testing data sparseness

Median performance masks large differences between individual prediction tasks

Most tasks see small differences even with large decreases of training data

Correlation between tasks is not large enough to explain the performance consistency



Conclusions

Deep neural networks are becoming increasingly popular in chemoinformatics

Bayesian optimization provided mixed results and might not be worth the computing cost compared to random parameter selection

Data sparseness is a frequent issue in multi-target data sets such as those that model biological pathways

The performance of deep neural networks is resilient in low to middle data sparseness scenarios

Acknowledgements

University of Sheffield:

- Prof. Dr. Val Gillet
- Prof. Dr. Beining Chen

Eli Lilly:

- Dr. David Evans
- Dr. Matthew Baumgartner

Colleagues from the Information School and the Chemistry Department at Sheffield

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n°612347

D3i4AD

Diagnostic and Drug Discovery Initiative for Alzheimer's Disease

MARIE CURIE ACTIONS

SEVENTH FRAMEWORK PROGRAMME