D3i4AD

Diagnostic and Drug Discovery Initiative for Alzheimer's Disease

MARIE CURIE ACTIONS

# Modelling phenotypic assays using machine learning

Antonio de la Vega de León

ACS Fall 2018 Meeting

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

The University Of Sheffield.

# Introduction

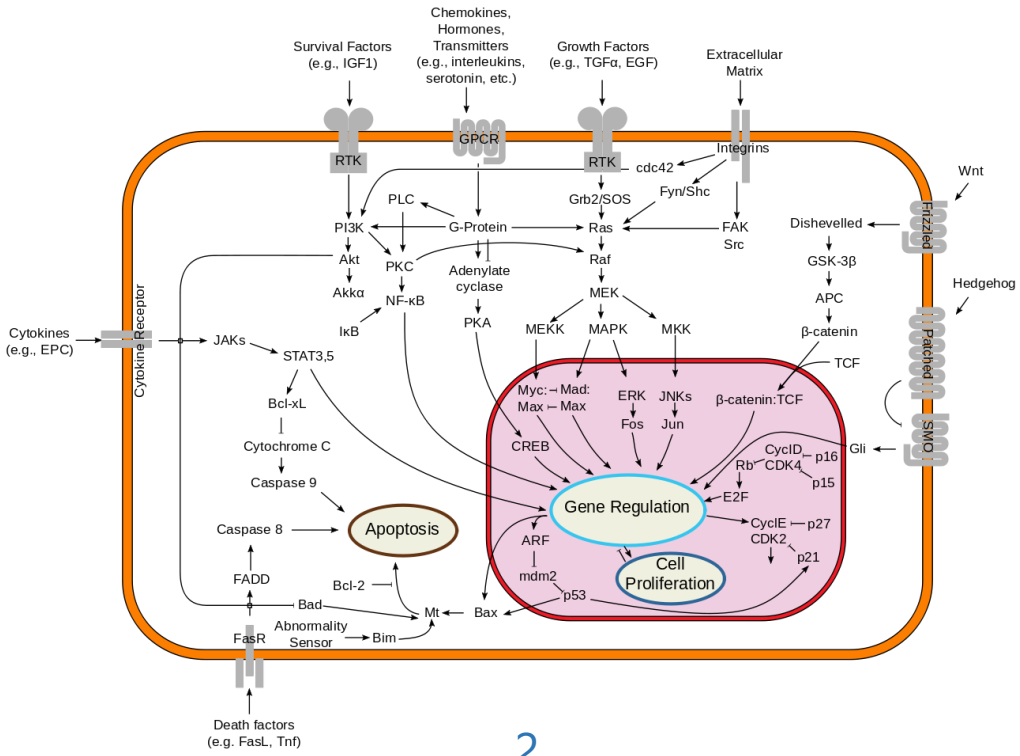
Data collection

Results

# Introduction

Phenotypic screening is a valuable tool to test compounds in an experimental model close to the disease condition

Main drawback: it is unknown what the molecule is doing inside the cell, what point in the pathway it is affecting






https://commons.wikimedia.org/wiki/File:Signal_transduction_pathways.svg

# Introduction

My goal with the project was to generate a machine learning model that, given a molecule, would say:

- If that molecule would be active in the phenotypic screen
- What targets in the pathway it would be hitting

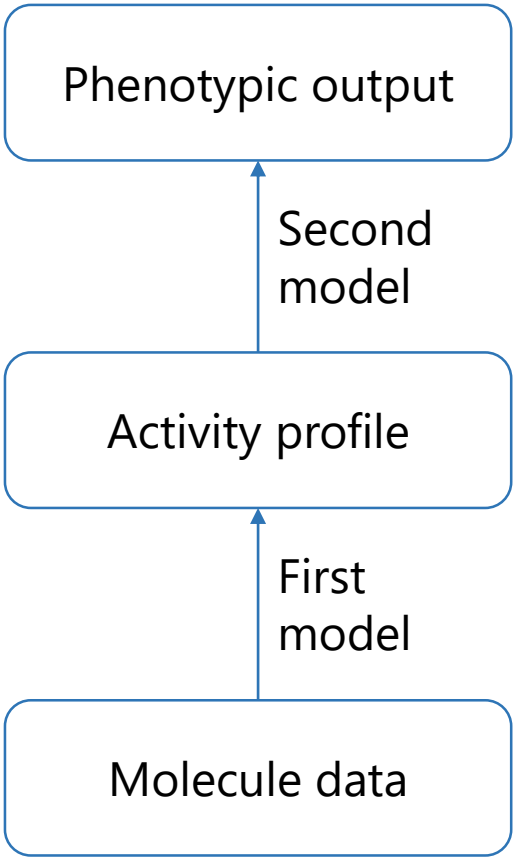
Similar work has been done by Al-Ali et al. (2015) and Drakakis et al. (2014) but both focused on a very broad target coverage

Here, we modelled with high precision that targets of the affected pathway and focus the prediction on them

# Introduction

Similar to Drakakis *et al.*, the first idea is to first predict activity for proteins involved in the phenotypic screen and use this information in the prediction of phenotypic response

Predicting an activity profile can be modelled as a multitask machine learning problem, where several targets are predicted at the same time



# Introduction

Deep neural networks (DNN) have become popular in chemoinformatics since they won the Merck Kaggle competition in 2012

They have been used to predict activity, aqueous solubility, drug induced liver injury, and more

Why use DNN?

- State-of-the-art performance
- Inherently multi-target, able to deal with missing labels and can produce structured output

Introduction

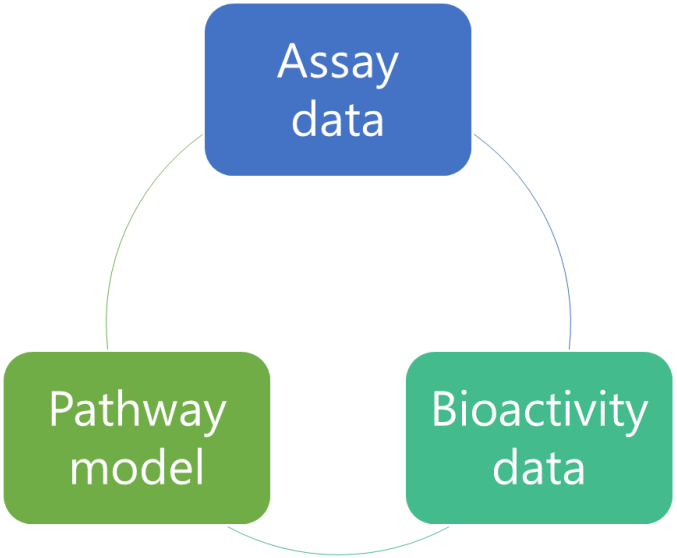
# Data collection

Results

# Creating a proof-of-concept data set

For our aim, we needed to have three sets of related data:

- Activity on a phenotypic assay
- Pathway that is affected in the assay
- Bioactivity information against targets in the pathway

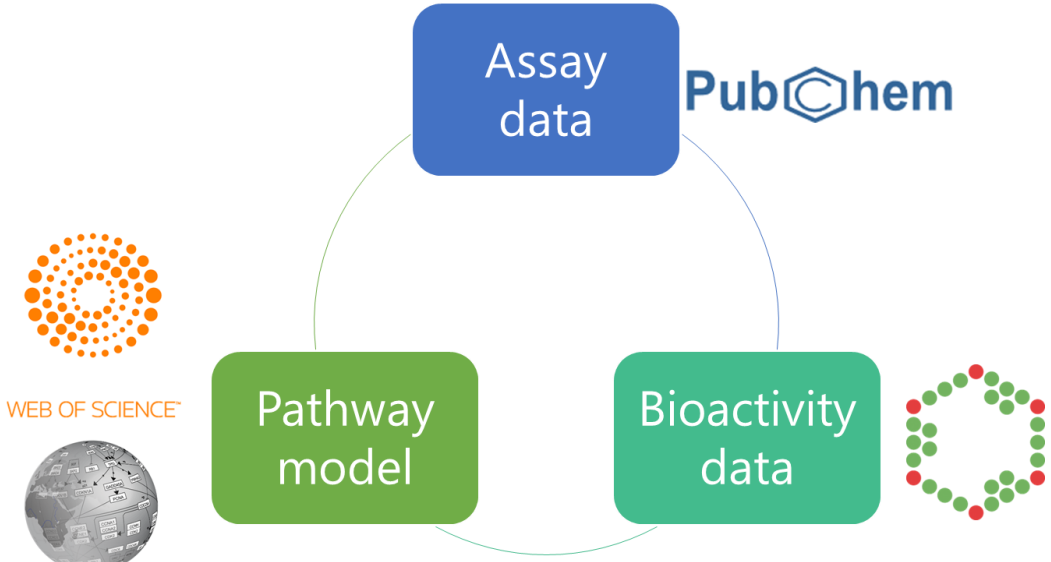


# Data search

PubChem search focused on cell-based “summary” assays that point to primary, confirmatory and counter-screening information of a project

At the same time, pathway information for the screening should be available in Wikipathways

Bioactivity information for proteins in the pathway is searched in ChEMBL

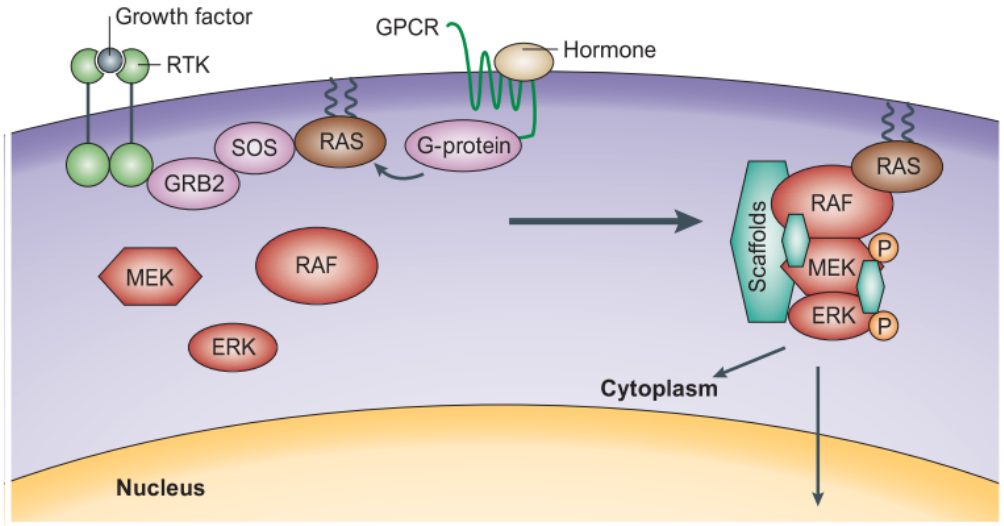


# ERK signalling pathway

The model follows a classical MAP kinase cascade

Numbers are active and inactive compounds for each assay

| | | Assay type | |
|---|---|---|---|
| | | Primary | Confirmatory |
| Activator | EGF | 528 / 124,388 | 33 / 14 |
| | Vasopressin | 692 / 64,632 | 40 / 27 |



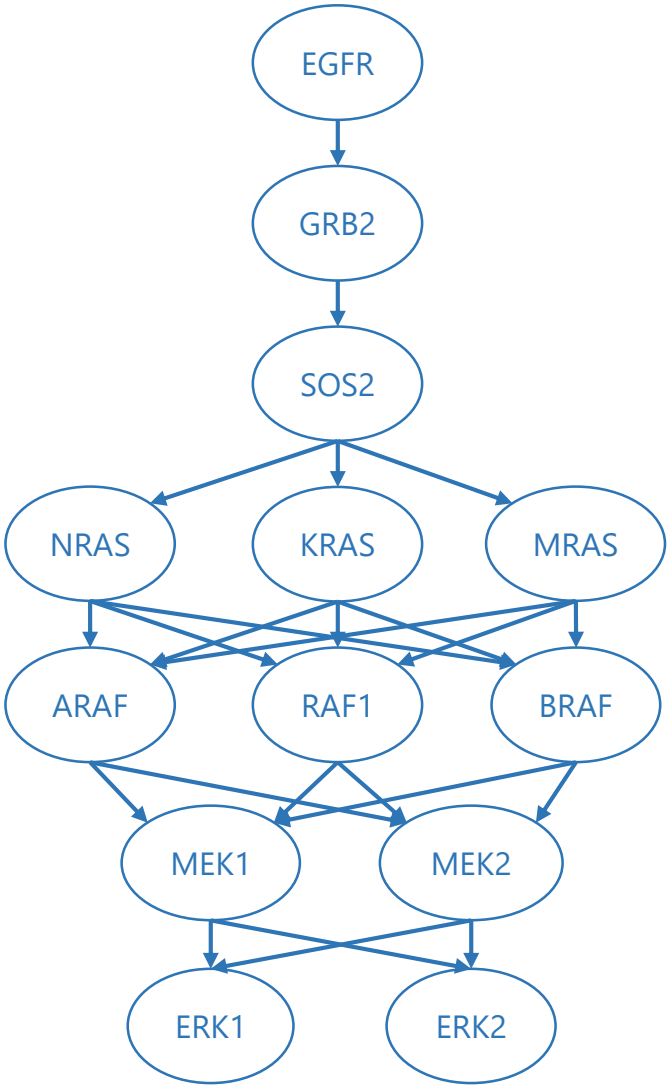



Wellbrock *et al.* Nat. Rev. Mol. Cell Biol. 2016

# ERK signalling pathway

Pathway modelled was extracted from WikiPathways (WP3284) and refined using the literature

Different human isoforms of RAS, RAF, MEK and ERK are included



# ERK signalling pathway

## Activity classification:

- For molar measures a threshold of 1 μM used
- Measures with no numerical value but activity annotations "Active", "Not active" and "Inactive" also considered

Targets in red were considered to have too little information for predictive modelling

| Targets | #Compounds | #Actives | #Inactives |
|---|---|---|---|
| EGFR | 7,220 | 3,088 | 4,132 |
| GRB2 | 322 | 154 | 168 |
| SOS2 | 0 | 0 | 0 |
| KRAS | 14 | 0 | 13 |
| NRAS | 1 | 1 | 0 |
| MRAS | 0 | 0 | 0 |
| BRAF | 1,986 | 1,571 | 415 |
| RAF1 | 788 | 432 | 356 |
| ARAF | 4 | 3 | 1 |
| MEK1 | 1,141 | 573 | 568 |
| MEK2 | 392 | 68 | 324 |
| ERK1 | 1,070 | 77 | 993 |
| ERK2 | 16,274 | 2,026 | 14,248 |
| TOTAL | 28,032 | 7,619 | 20,413 |

Introduction

Data collection

Results

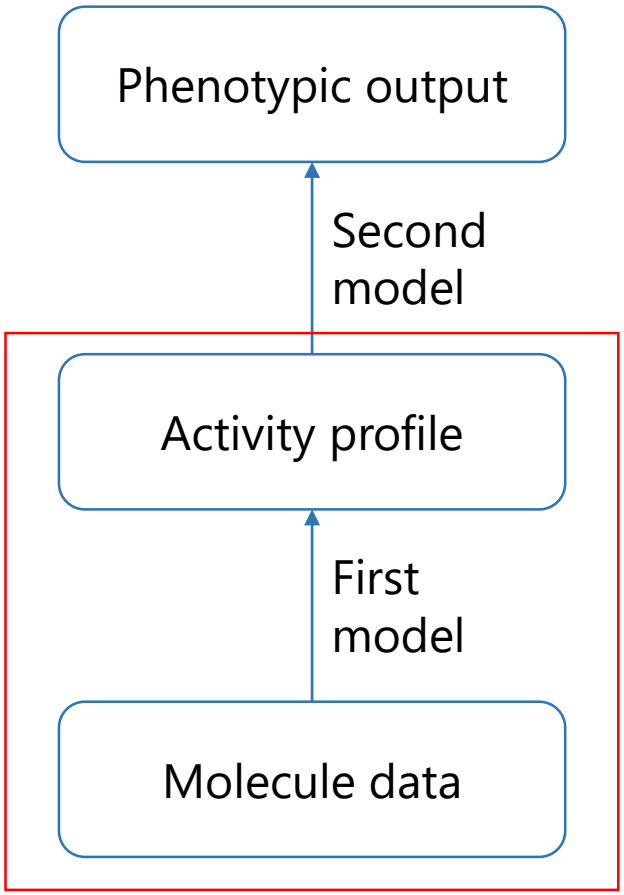
# Selecting the first model

Input will be the bioactivity data set with hashed ECFP4 fingerprint as molecular representation

Predict all 8 targets at the same time

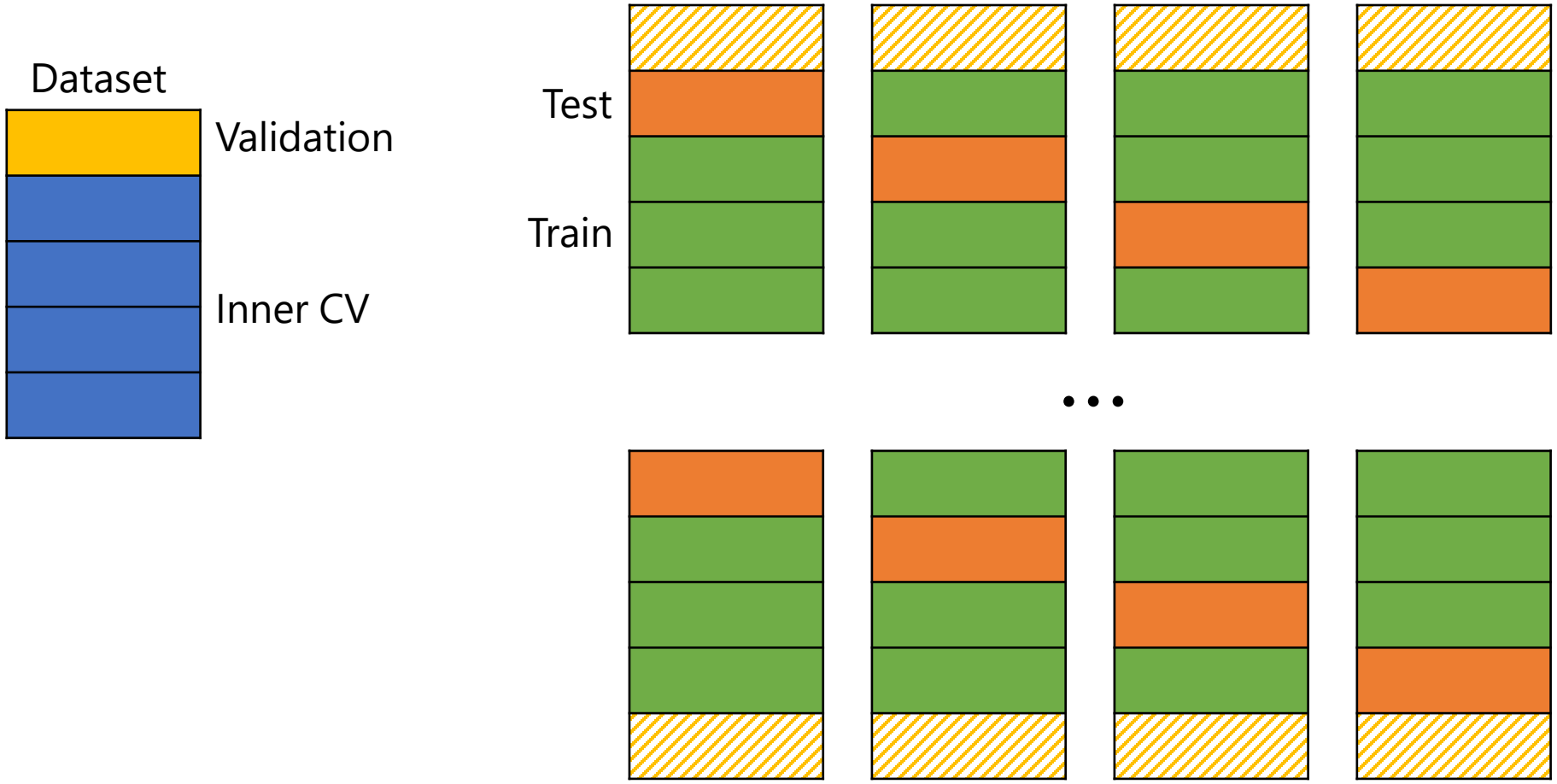
Measure MCC for each and average

Problem: how to choose best model?



# Selecting the first model

To select the best DNN model, a nested cross-validation was performed



# Selecting the first model

Nested cross-validation with 5 outer and 4 inner folds

Leave one fold as validation and perform cross-validation in the other folds (inner cross-validation)

Best model for each inner fold is tested in all outer folds

100 models with random hyperparameter values were tested in each run of the inner cross-validation

# Results of nested cross-validation

Values are average MCC over folds in each CV

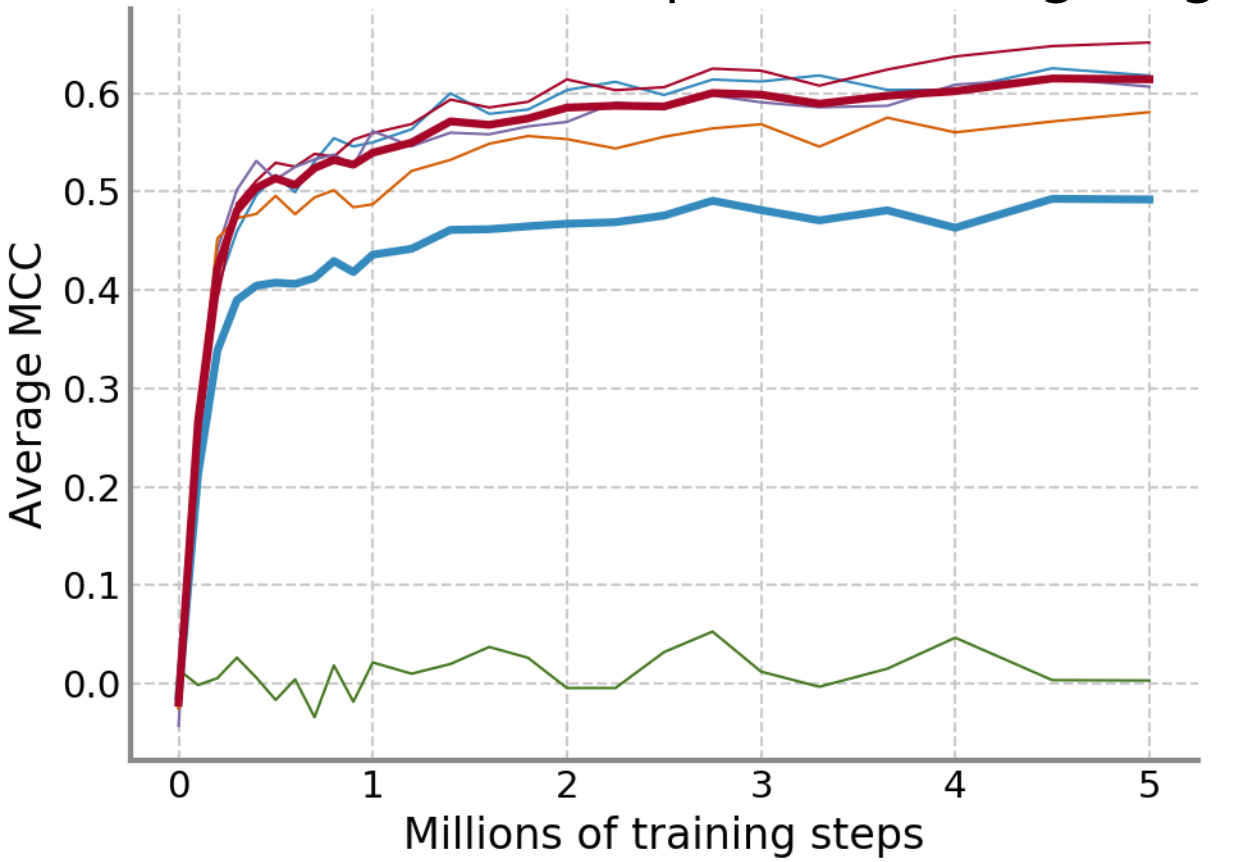
Model 49 was the best one

| Outer Fold | Best Model | Inner CV | Outer CV |
|---|---|---|---|
| 1 | **49** | 0.241 | **0.234** |
| 2 | 92 | 0.187 | 0.221 |
| 3 | 13 | 0.217 | 0.185 |
| 4 | 69 | 0.030 | 0.157 |
| 5 | **49** | 0.230 | **0.235** |

# Improving first model

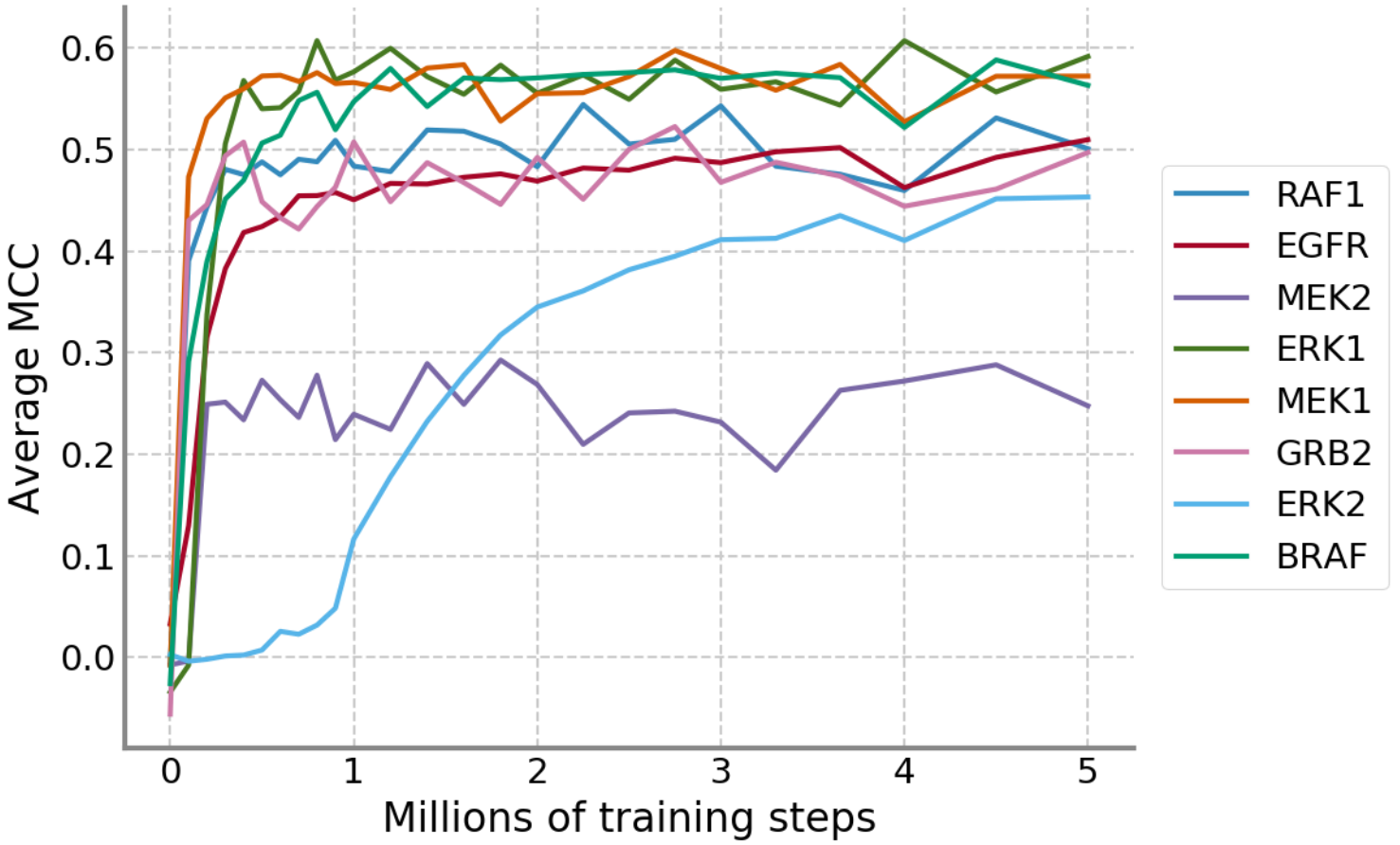
Nested cross-validation was done with low amount of training

To optimize the model, we find optimal training length



# Improving first model

Except for MEK2, there is no big target difference



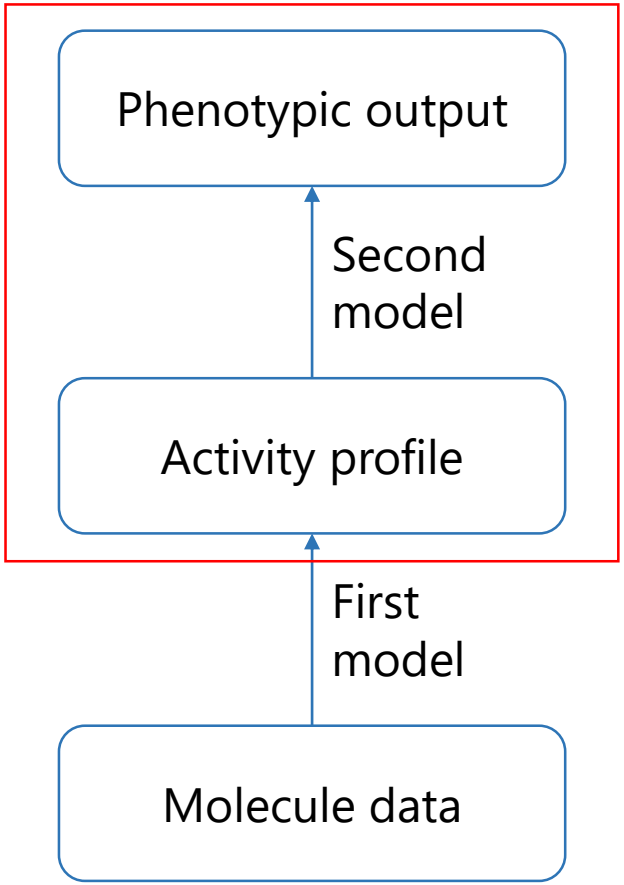
# Training the second model

Input is phenotypic data set with activity prediction from first model as molecular representation

Our first tests are done on the primary assays: activated by either EGF (AID1414) or vasopressin (AID995)

Problem: highly unbalanced with large number of inactive molecules

Work is still in progress



# Training the second model

Using binary activity prediction representation with AID1414 didn't lead to a good model

We analysed how predicted activity related to assay activity

We looked at calculating the odds ratio of the compound being active in the screen depending on if its active for a target

$$OR = \frac{\frac{\#P_A and\ C_A}{\#P_I and\ C_A}}{\frac{\#P_A and\ C_I}{\#P_I and\ C_I}}\ P: PubChem\ C: ChEMBL$$

# Odds ratio analysis

In all targets, predicted active compounds are enriched in phenotypically active compounds

Enrichment strength varies from weak (1.2) to strong (19)

| Target | $P_AC_A$ | $P_IC_A$ | $P_AC_I$ | $P_IC_I$ | OR |
|---|---|---|---|---|---|
| EGFR | 44 | 976 | 484 | 123412 | 11.495 |
| GRB2 | 90 | 18337 | 438 | 106051 | 1.188 |
| BRAF | 27 | 3472 | 501 | 120916 | 1.877 |
| RAF1 | 21 | 4159 | 507 | 120229 | 1.197 |
| MEK1 | 64 | 6742 | 464 | 117646 | 2.407 |
| MEK2 | 69 | 8953 | 459 | 115435 | 1.938 |
| ERK1 | 6 | 170 | 522 | 124218 | 8.399 |
| ERK2 | 7 | 88 | 521 | 124300 | 18.978 |

# Odds ratio analysis

There are 210 compounds that are active in the screen and also active against one of the targets

This means there are 318 compounds that are active in the screen but not active against any targets

Using binary active/inactive descriptor might not give enough information to predict

Further work is ongoing to provide acceptable predictions

# Conclusion

We have established a framework to model phenotypic assays using public data

We have generated a proof-of-concept data set

We generated a fairly successful activity prediction model

Work is on going in getting the second model right

# 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

Funding for this research has been provided by the European commission under IAPP Grant No. 612347