# Computational medicine in action

Ewa Szczurek
Faculty of Mathematics, Informatics and Mechanics, University of Warsaw
szczurek@mimuw.edu.pl
https://www.mimuw.edu.pl/~szczurek/

UNIVERSITY OF WARSAW

WYDZIAŁ MATEMATYKI, INFORMATYKI I MECHANIKI

# Research in our lab

- Method development
  - Machine learning
    - Probabilistic graphical models
    - Deep learning
  - Statistical data analysis

- Always at the service of an important medical cause
  - Cancer
  - Antimicrobial peptides
  - COVID-19

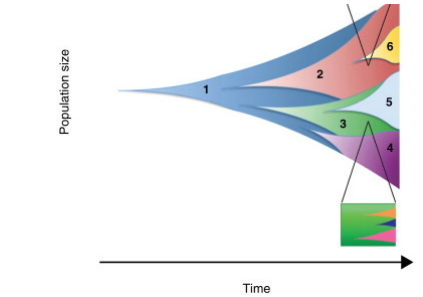
- Part 1: overview
- Part 2: more detail on drug sensitivity prediction

- https://www.mimuw.edu.pl/~szczurek/

# Part 1. Overview

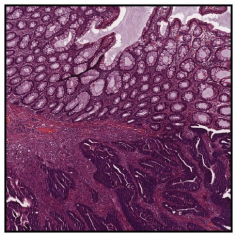
# Applications: how to model and understand…

## Tumor evolution & microenvironment



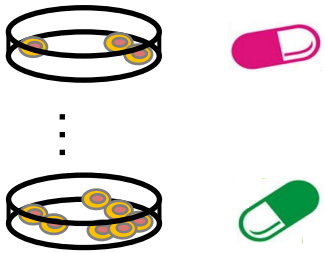

Darvish Shafighi *et al., Genome Med,* 2021
Markowska, Cąkała *et al.*, *BioRxiv*, 2021
Lähnemann *et al.*, *Genome Biol*, 2020
Szczurek *et al., PloS CB,* 2020
Geras *et al., in preparation*
Kang *et al.*, *in preparation*

## Medical images



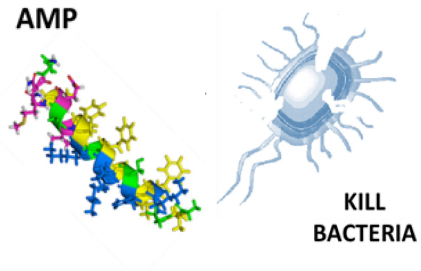
Rączkowski *et al., Sci Rep,* 2019
Rączkowski *et al., BiorXiv,* 2021

## Sensitivity of cancer cells to drugs



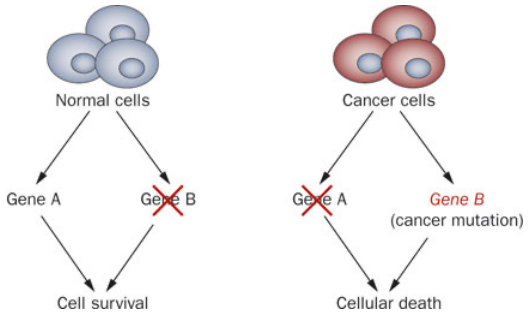
Koras *et al., Sci Rep,* 2020
Koras *et al., Sci Rep,* 2021
Koras *et al., in preparation*

## Generation of anti-microbial peptides



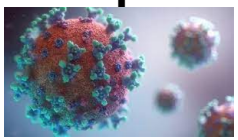

Szymczak, Możejko, *et al., in preparation*

## Synthetic lethality, genetic interactions




Markowska *et al.*, *in preparation*
Elmes *et al., Plos One,* 2021
Tiuryn, Szczurek, *Bioinformatics*, 2019

## COVID-19 epidemiology

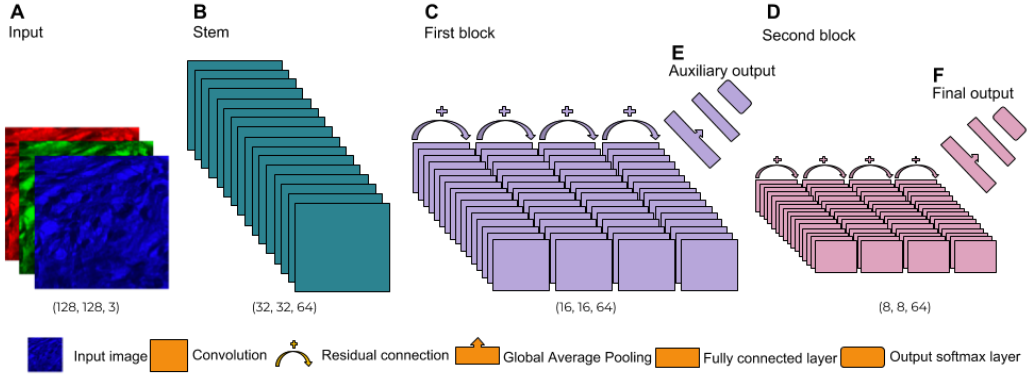


Priesemann *et al., The Lancet,* 2021 (2x)
Priesemann *et al., The Lancet,* 2020
Ifthekar *et al.*, *The Lancet Regional Health*, 2021
Krueger, Gogolewski, Bodych *et al., medRxiv,* 2021
Adamik *et al., medRxiv,* 2020
Bock *et al., medRxiv,* 2020

# Method development: deep learning

## ARA-CNN

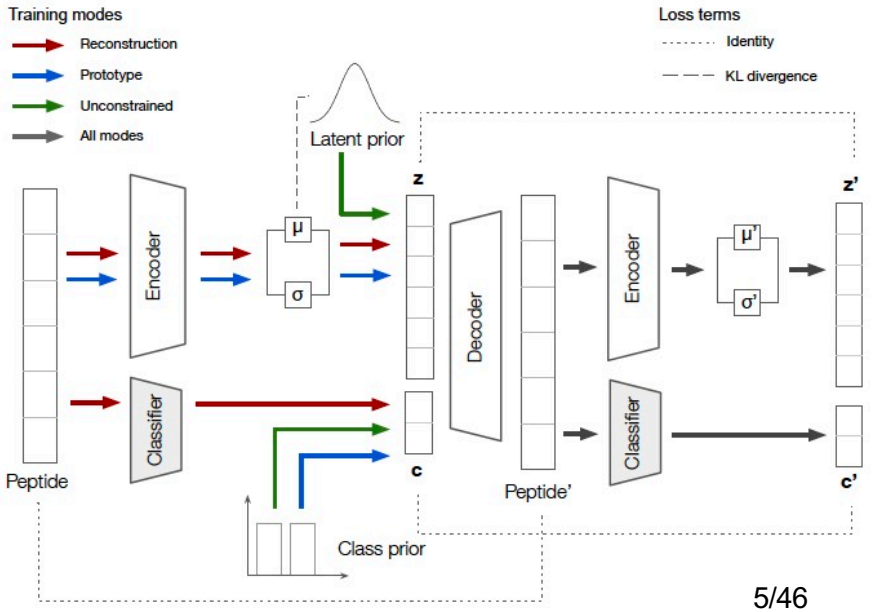
CNNs, uncertainty, active learning (Rączkowski *et al., Sci Rep,* 2019, *BioRxiv* 2021)




## HydrAMP

Conditional variational autoencoders

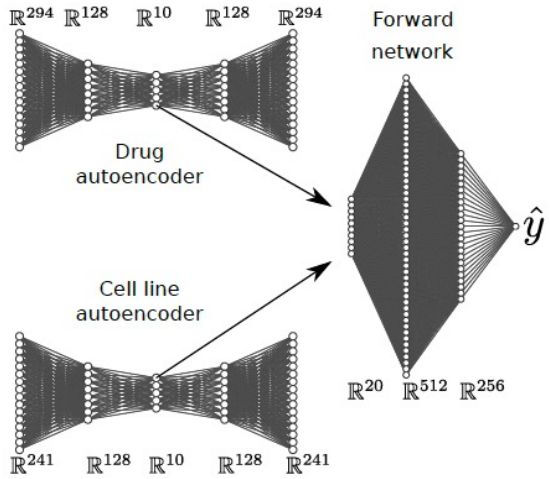
(Szymczak, Możejko *et al., in preparation*)




## DEERS

Autoencoders, recommendation systems

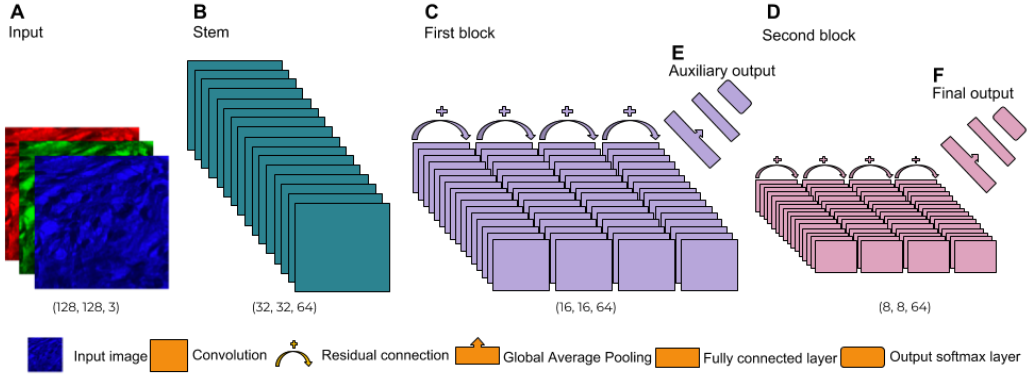
(Koras *et al., Sci Rep, 2021*)



# Method development: deep learning

**ARA-CNN**

CNNs, uncertainty, active learning (Rączkowski *et al., Sci Re*

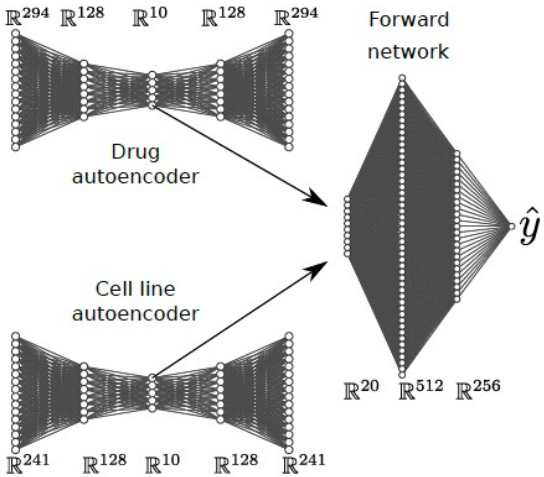



- Highly flexible models
- Highly predictive
- Large numbers of parameters
- Difficult to interpret – custom approach to interpretability needed

**DEERS**

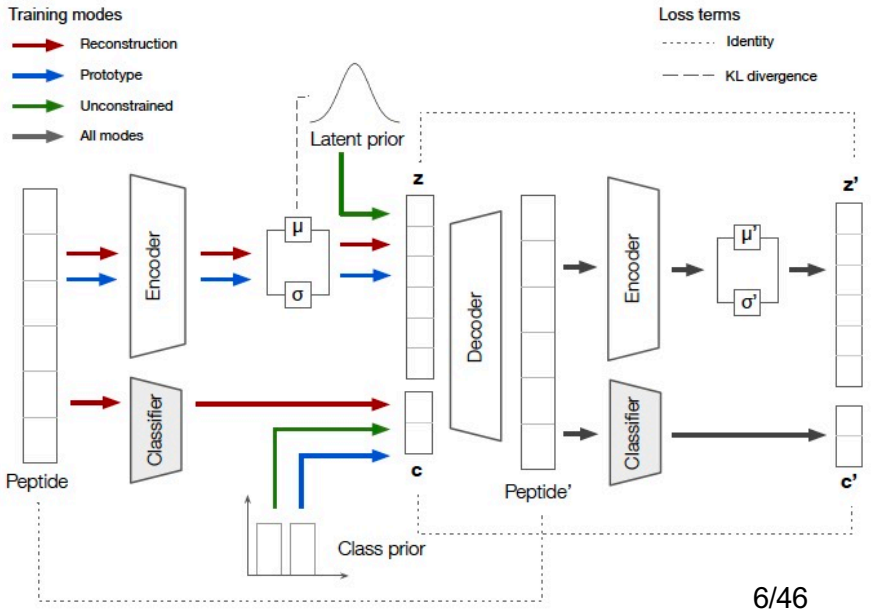
Autoencoders, recommendation systems

(Koras *et al., Sci Rep, 2021*)



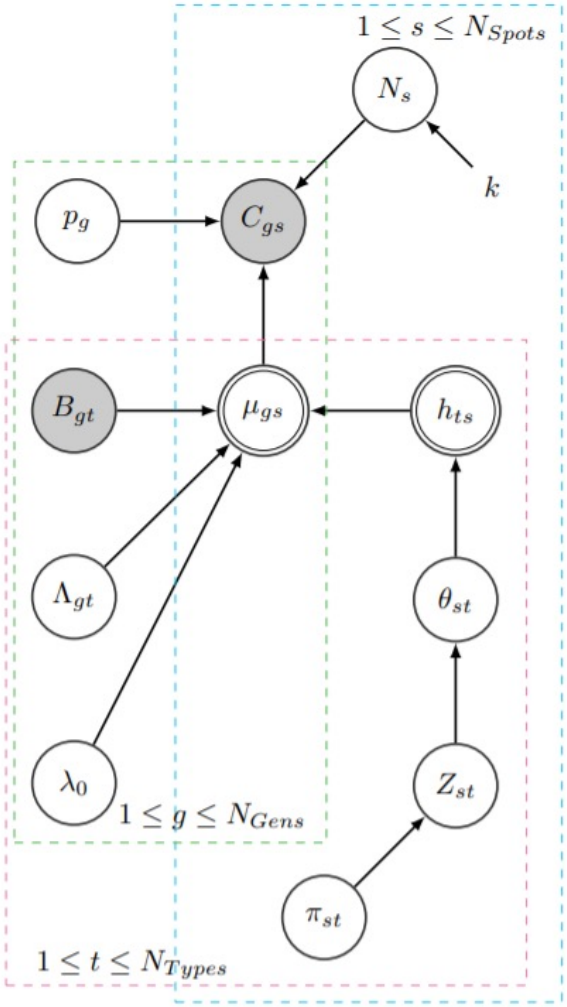

Conditional variational autoencoders

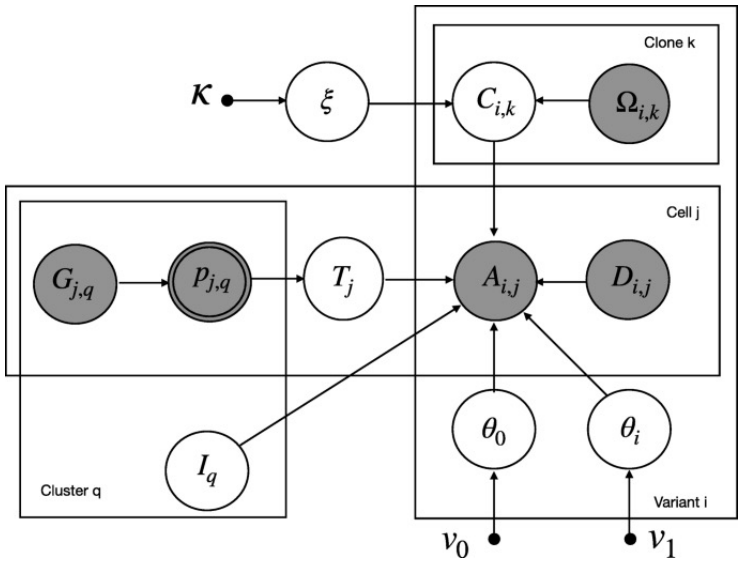
(Szymczak, Możejko *et al., in preparation*)



# Method development: probabilistic graphical models

**Celloscope** (Geras *et al.,* in preparation)
- Metropolis Hastings within Gibbs



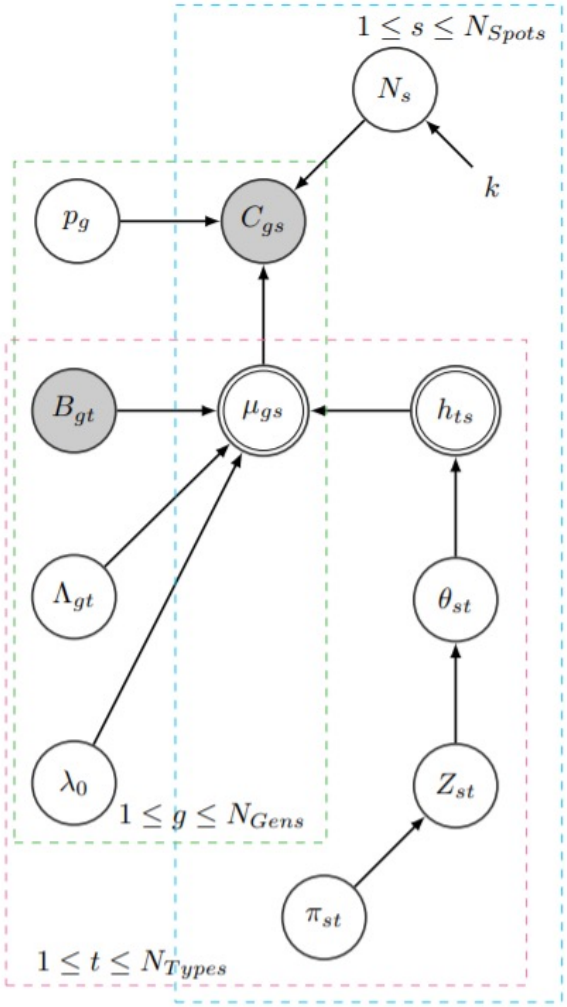




**CACTUS** (Darvish Shafighi *et al., Genome Med,* 2021)
- Gibbs sampler

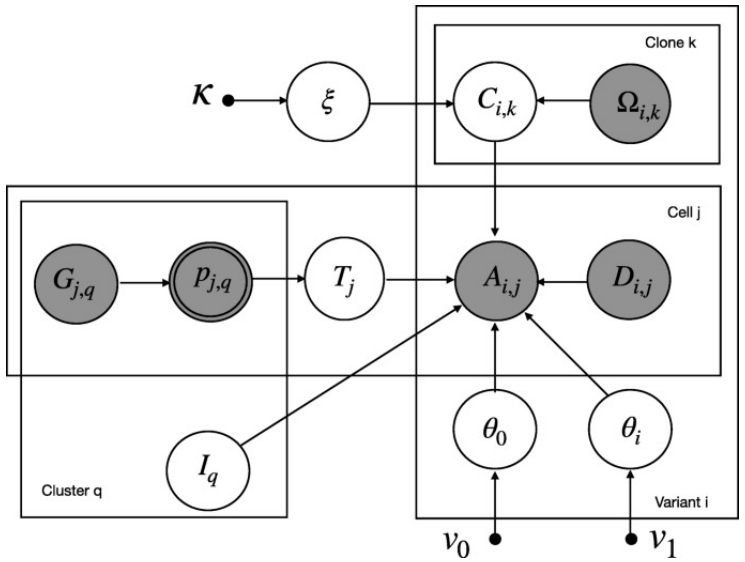
# Method development: probabilistic graphical models

**Celloscope** (Geras *et al.,* in preparation)

- Metropolis Hastings within Gibbs




- Relatively smaller
- Each random variable corresponds to some entity in the system
- Conditional probability distributions describe relations between variables
- More precise description



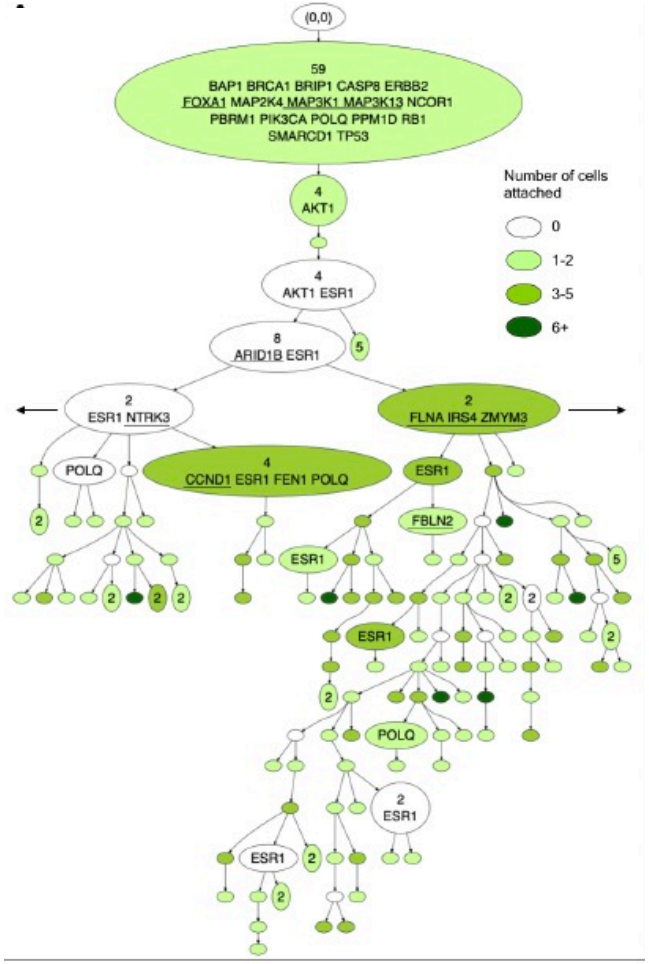

**CACTUS** (Darvish Shafighi *et al., Genome Med,* 2021)

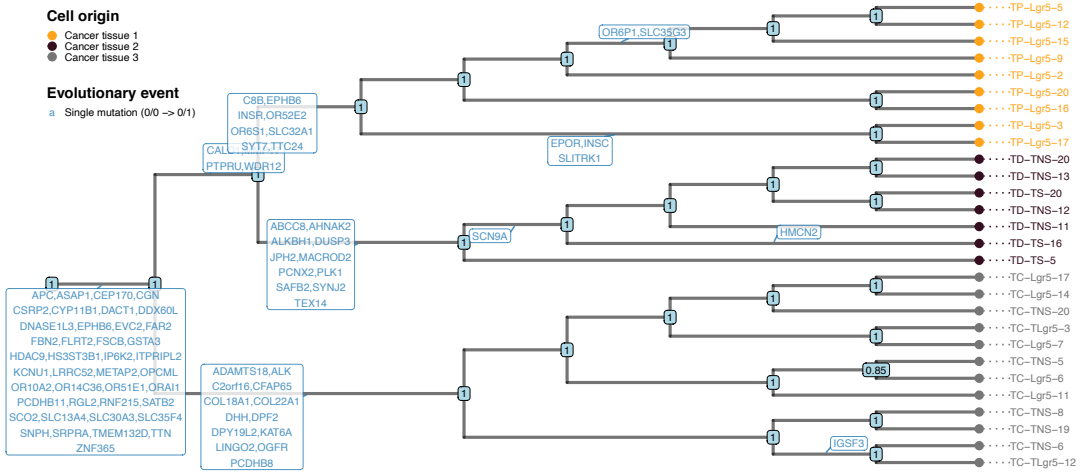
- Gibbs sampler

# Method development: probabilistic graphical models with tree structure model learning

**CONET** (Markowska, Cąkała *et al., BioRxiv* 2021)

- MCMC sampler



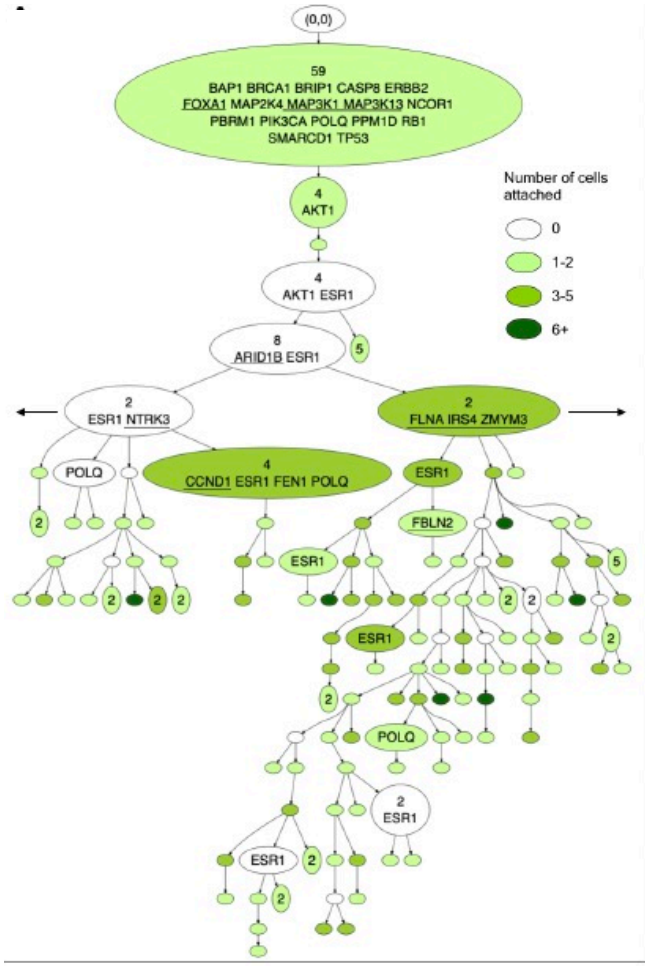




**SIEVE** (Kang *et al., in preparation*)

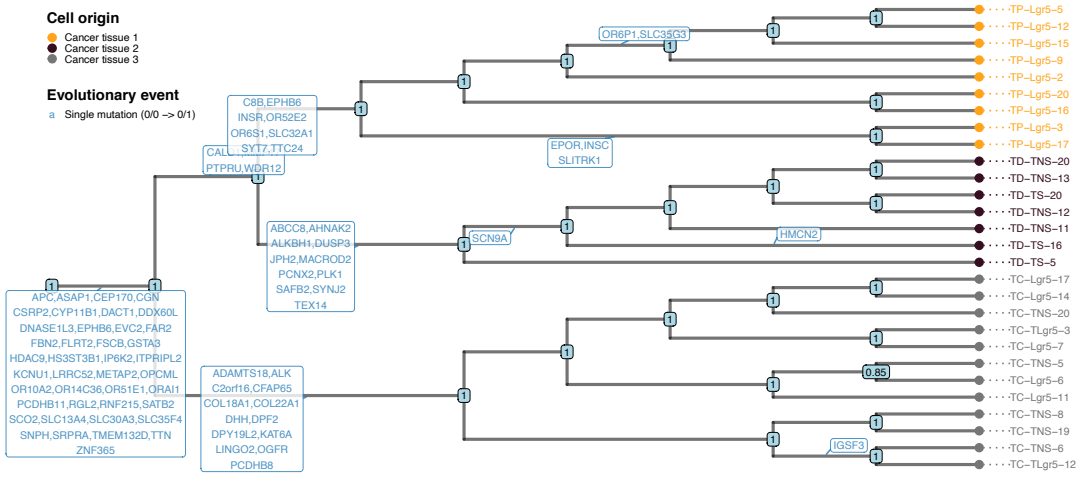
- MCMC sampler

# Method development: probabilistic graphical models with tree structure model learning

**CONET** (Markowska, Cąkała *et al., BioRxiv* 202
- MCMC sampler




- The tree describes the evolutionary history of the tumor
- A probabilistic graphical model of the data conditional on the tree
- Difficulty: learning the probabilistic model of the data and learning the tree structure at the same time




**SIEVE** (Kang *et al., in preparation*)
- MCMC sampler

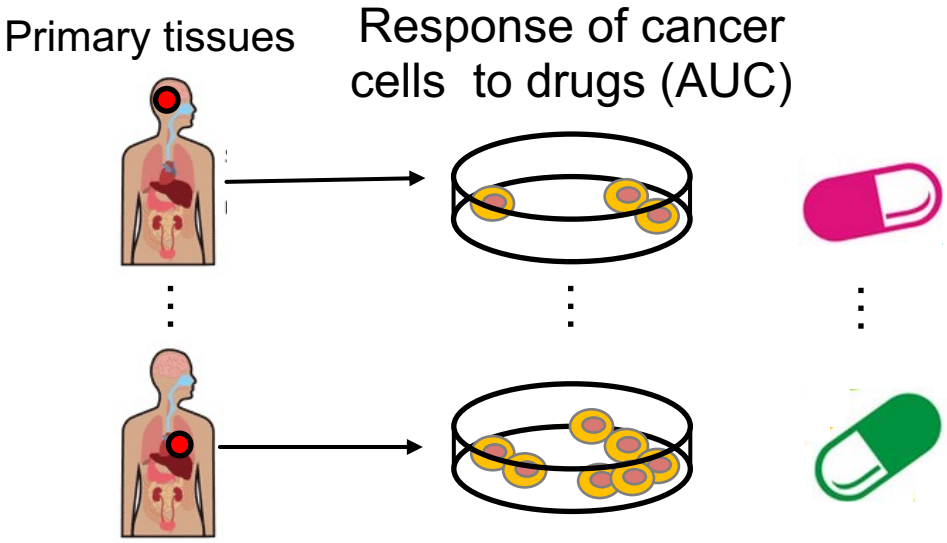
# Others working on similar topics in Poland

- **Probabilisttic graphical models of biological phenomena:** Anna Gambin, Błażej Miasojedow (MIM UW)
- **Deep learning applied to biological phenomena**: Bartek Wilczyński, (MIM UW)
- **Feature selection**: Witold Rudnicki (Białystok University)
- **Interpretability**: Przemysław Biecek (MIM UW and Warsaw University of Technology)
- **Single cell sequencing in tumors**: Bożena Kamińska (Nencki Institute), Marcin Tabaka (International Centre for Translational Eye Research)
- **Medical image analysis, machine learning**: Tomasz Trzciński (Warsaw University of Technology)
- **Gaussian mixture autoencoders**: Marek Smieja, Jacek Tabor (Jagiellonian University)

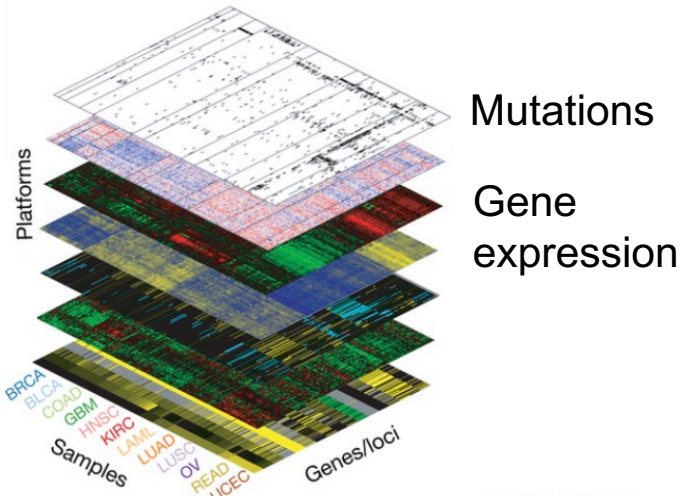
# Part 2. Modeling sensitivity of cancer cell lines to drugs

Krzysztof Koras

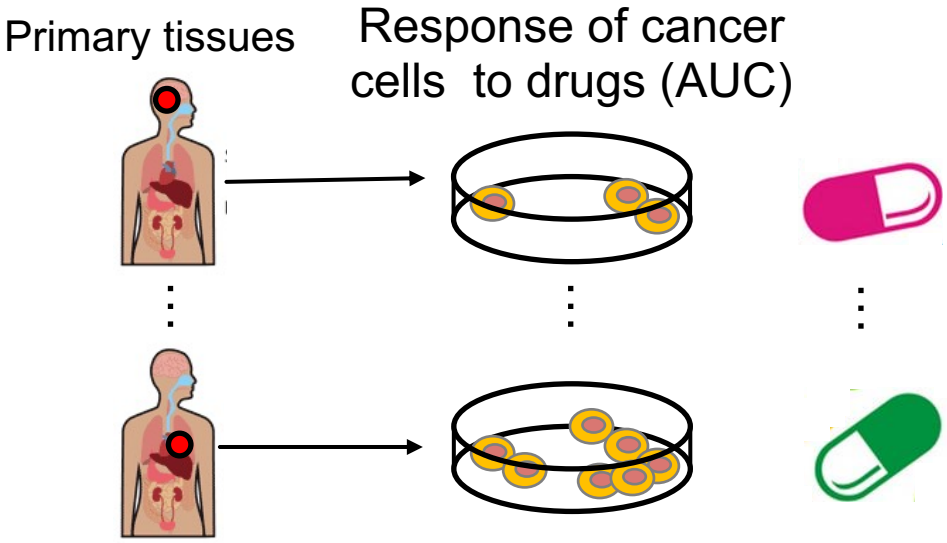
# Measuring sensitivity of cancer cell lines to drugs



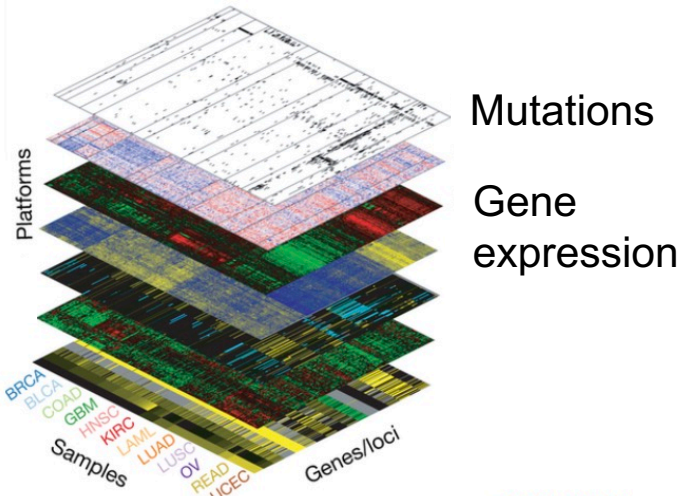

Molecular features of cancer cells



# Understanding sensitivity of cancer cell lines to drugs




Molecular features of cancer cells

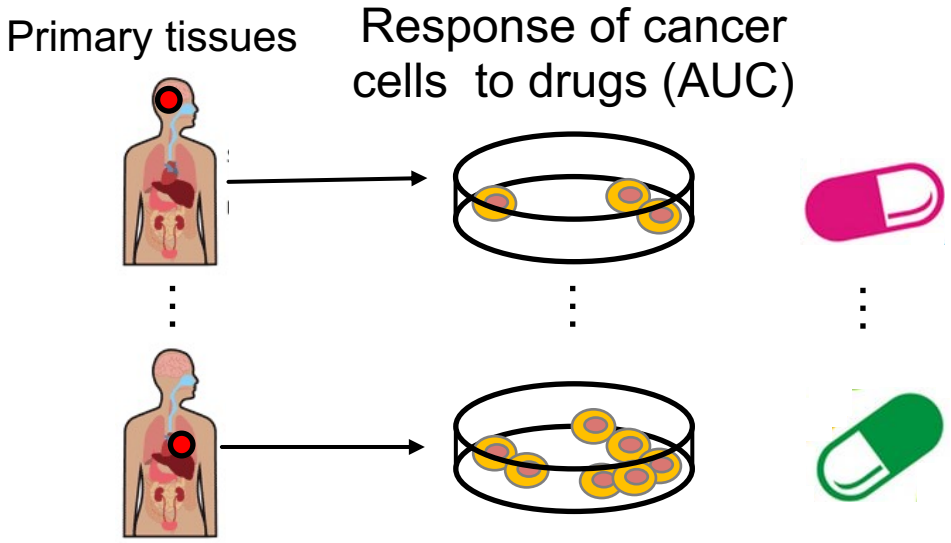


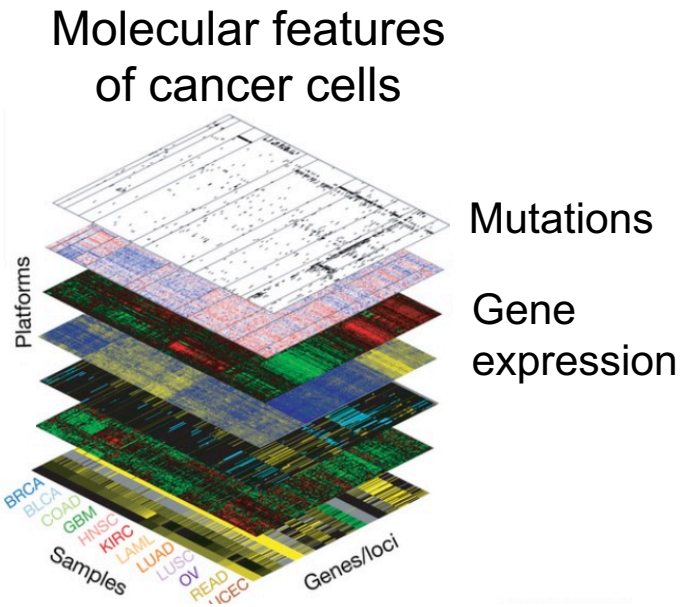

**Machine learning task:**

- **Given**: features of a drug and molecular features of a cancer cell line
- **Predict** the response (AUC value) of the cell line to the drug.

Mimics **precision medicine** application in the clinic.

# How can ML help to understand the drug action on cancer cell lines?




- **Feature selection**: which cell line features are predictive of the response?[1]
- **Multi-task learning**: capturing the action of multiple drugs on multiple cancer cell lines in a single model[2]
- **Interpretability** (explainability): what are the mechanisms behind the drug action on the cell?[2]
- **Representation learning** finding low-dimensional representations of drugs, cell lines[2,3]

[1]Koras *et al.*, *Sci Rep,* 2020
[2]Koras *et al., Sci Rep,* 2021
[3] Koras et al., *in preparation*

www.nature.com/scientificreports

scientific reports

Check for updates

OPEN

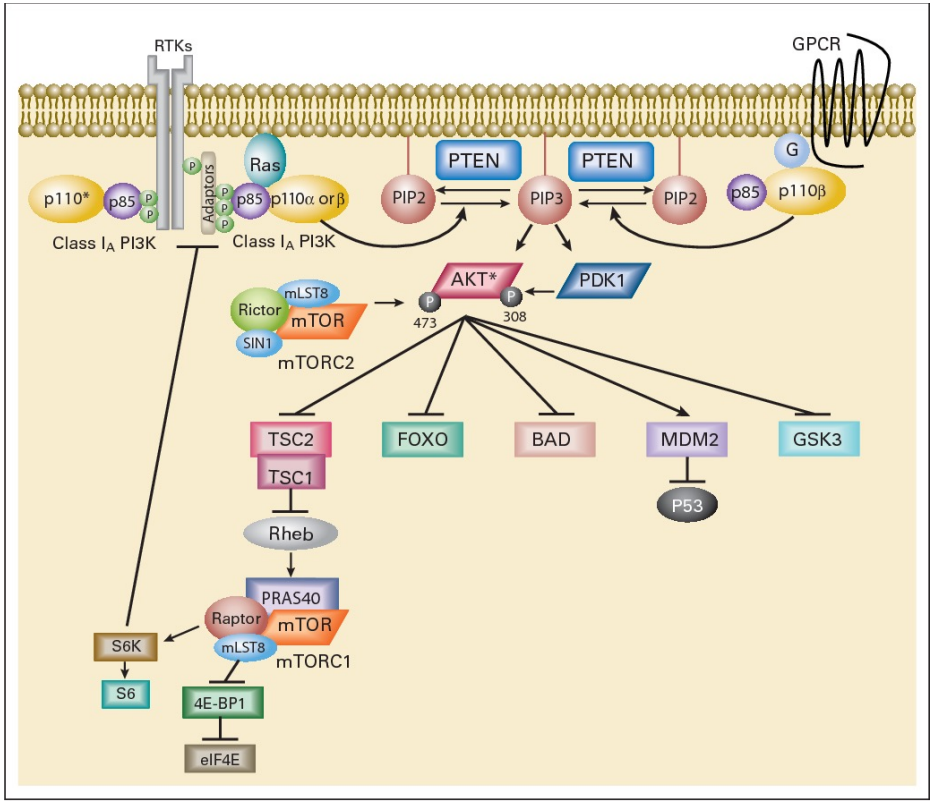
# Interpretable deep recommender system model for prediction of kinase inhibitor efficacy across cancer cell lines

Krzysztof Koras[1], Ewa Kizling[1], Dilafruz Juraeva[2], Eike Staub[2] & Ewa Szczurek[1]✉

Computational models for drug sensitivity prediction have the potential to significantly improve personalized cancer medicine. Drug sensitivity assays, combined with profiling of cancer cell lines and drugs become increasingly available for training such models. Multiple methods were proposed for predicting drug sensitivity from cancer cell line features, some in a multi-task fashion. So far, no such model leveraged drug inhibition profiles. Importantly, multi-task models require a tailored approach to model interpretability. In this work, we develop DEERS, a neural network recommender system for kinase inhibitor sensitivity prediction. The model utilizes molecular features of the cancer cell lines and kinase inhibition profiles of the drugs. DEERS incorporates two autoencoders to project cell line and drug features into 10-dimensional hidden representations and a feed-forward neural network to combine them into response prediction. We propose a novel interpretability approach, which in addition to the set of modeled features considers also the genes and processes outside of this set. Our approach outperforms simpler matrix factorization models, achieving $R = 0.82$ correlation between true and predicted response for the unseen cell lines. The interpretability analysis identifies 67 biological processes that drive the cell line sensitivity to particular compounds. Detailed case studies are shown for PHA-793887, XMD14-99 and Dabrafenib.

# What do we know about how the kinase inhibitors work?

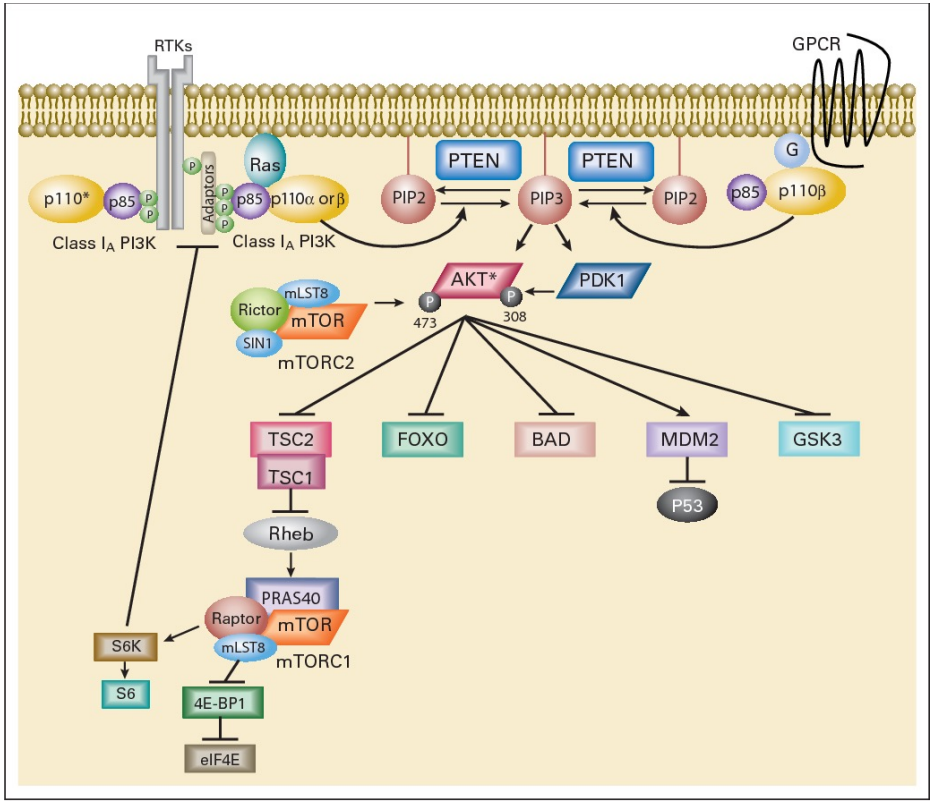
- Kinase inhibitors: drugs which target *kinases* (proteins)
- These kinases are usually part of some *biological process*




Courtney et al, Journal of clinical oncology, 2010

# What do we know about how the kinase inhibitors work?

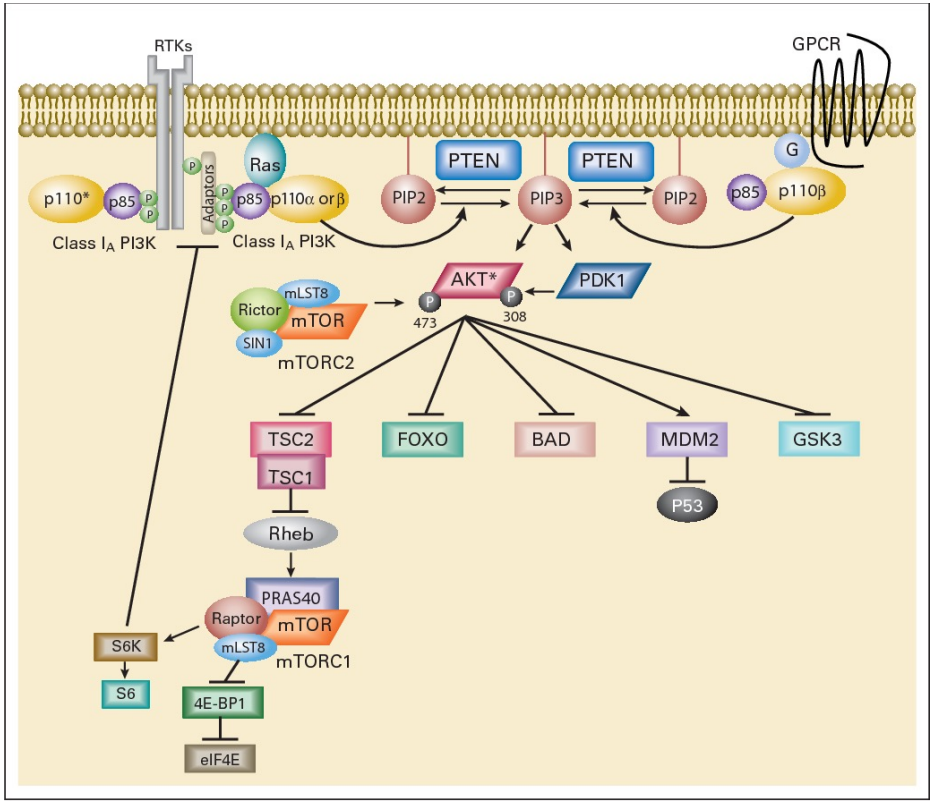
- Kinase inhibitors: drugs which target *kinases* (proteins)
- These kinases are usually part of some *biological process*
- This biological process can be important for cancer progression
- When the target kinase is inhibited, the process is perturbed




Courtney et al, Journal of clinical oncology, 2010

# What do we know about how the kinase inhibitors work?

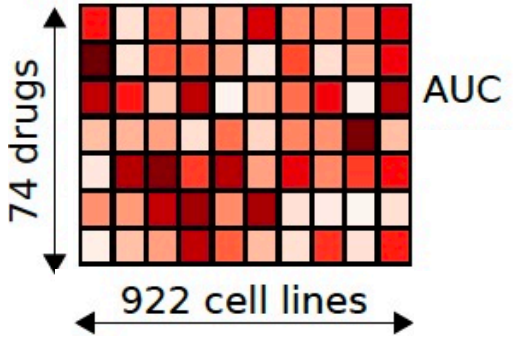
- Kinase inhibitors: drugs which target *kinases* (proteins)
- These kinases are usually part of some *biological process*
- This biological process can be important for cancer progression
- When the target kinase is inhibited, the process is perturbed
- Kinase inhibitors have their *off-targets*
- Their inhibition strengths on targets and and off-targets is measured by *inhibition profiles*




Courtney et al, Journal of clinical oncology, 2010

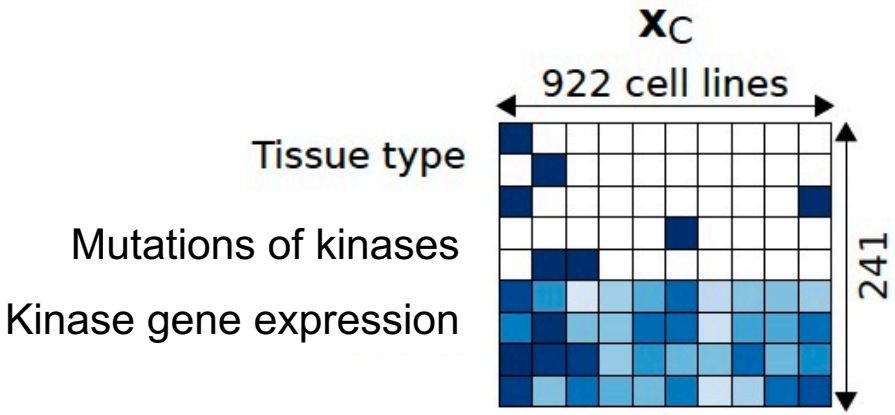
# Multi-task modeling approach: recommender system

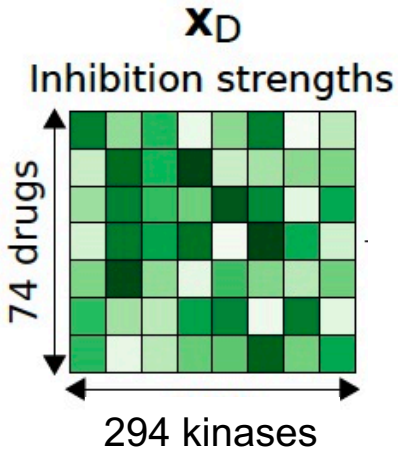
**Data:** 922 cell lines, 74 drugs, 52730 drug-cell line pairs in total

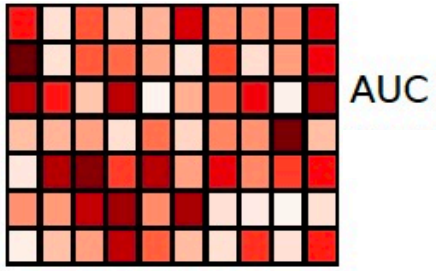


# Multi-task modeling approach: recommender system

**Data:** 922 cell lines, 74 drugs, 52730 drug-cell line pairs in total

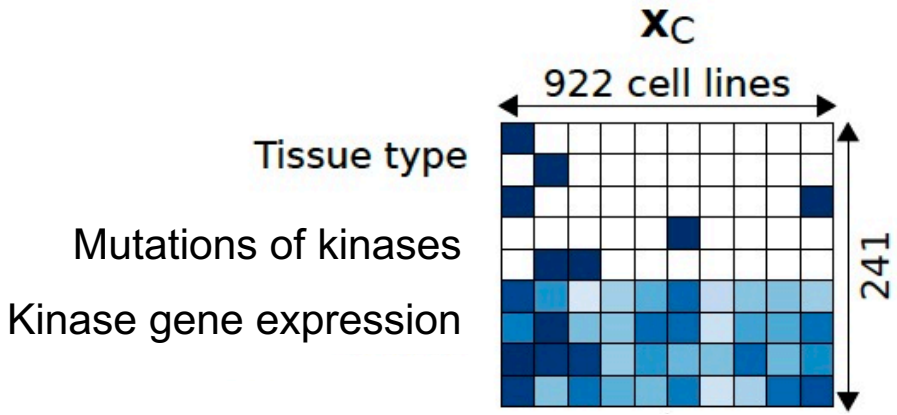


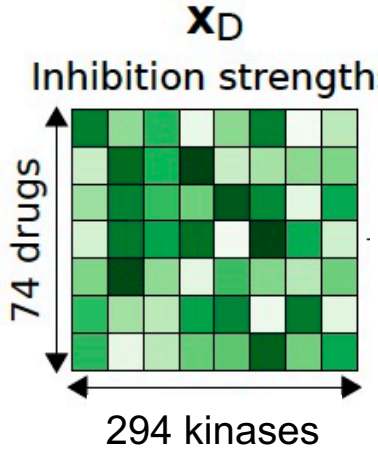



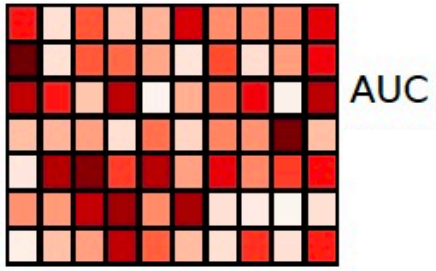



# Multi-task modeling approach: recommender system

**Data:** 922 cell lines, 74 drugs, 52730 drug-cell line pairs in total
**Recommender system:** recommending movies to users



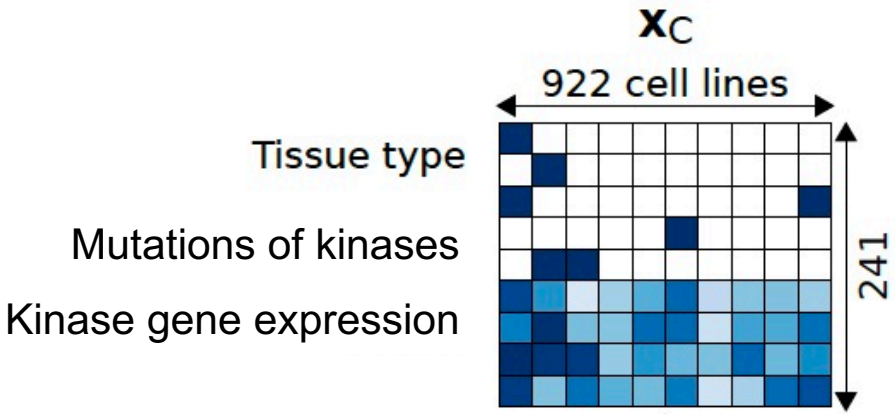



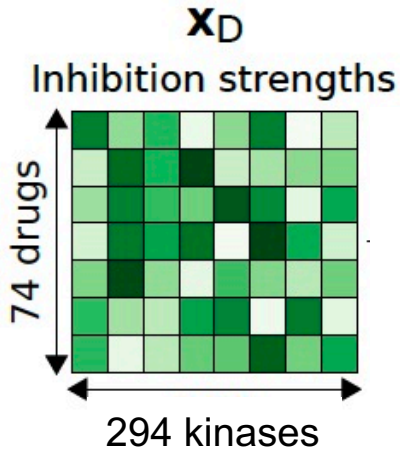
# Multi-task modeling approach: recommender system

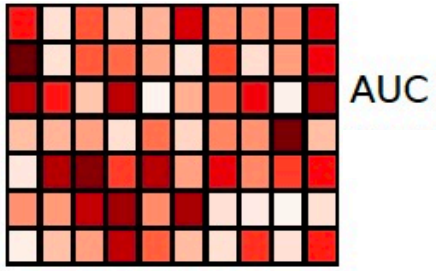
**Data:** 922 cell lines, 74 drugs, 52730 drug-cell line pairs in total

**Recommender system:** recommending ~~movies~~ to ~~users~~

drugs cancer cell lines



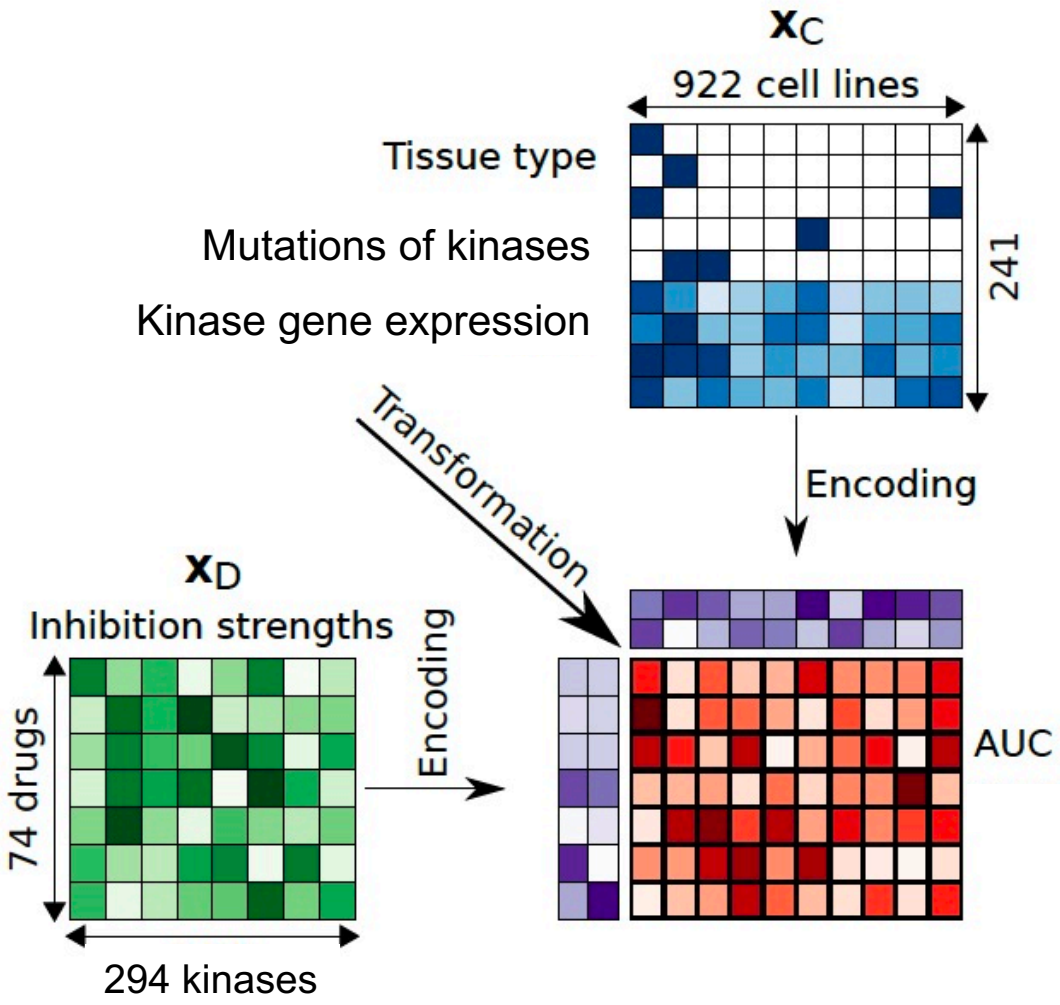



# Multi-task modeling approach: recommender system

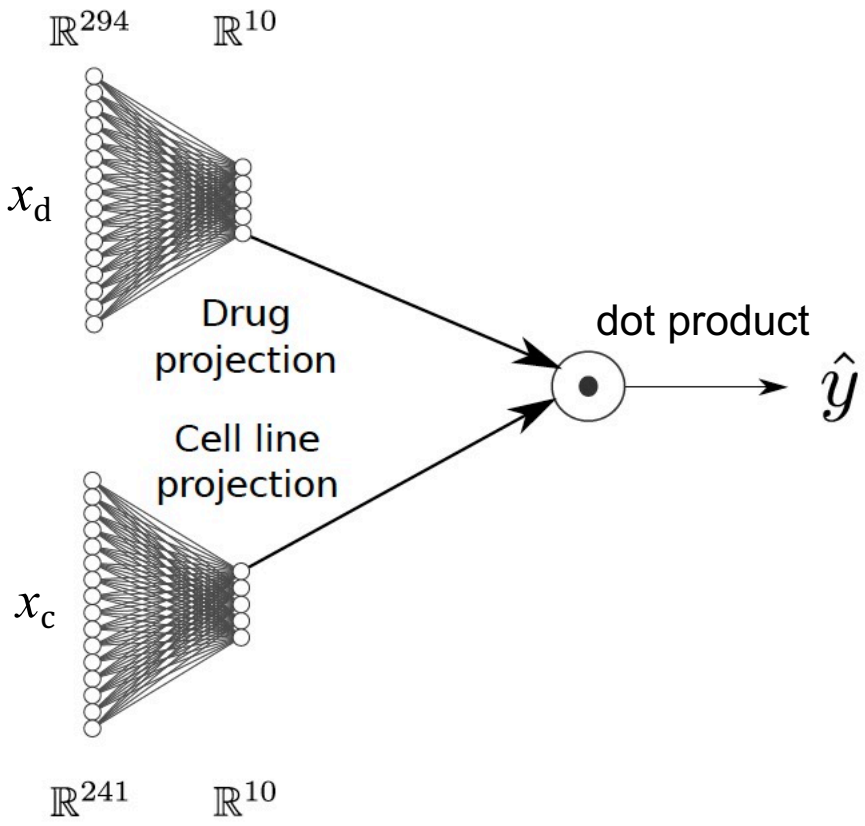
**Data:** 922 cell lines, 74 drugs, 52730 drug-cell line pairs in total

**Recommender system:** recommending ~~movies~~ to ~~users~~

drugs cancer cell lines



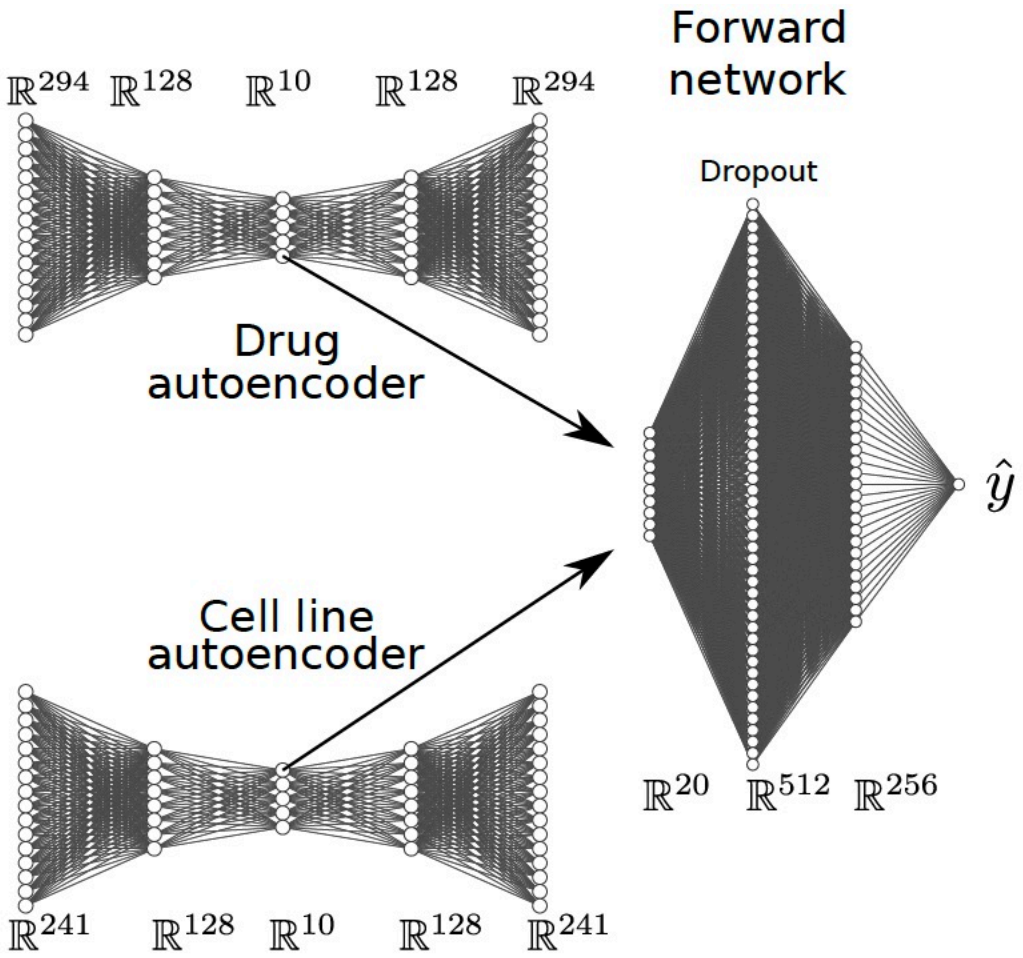
# Simple model: Matrix factorization with side information (MF)




Linear model

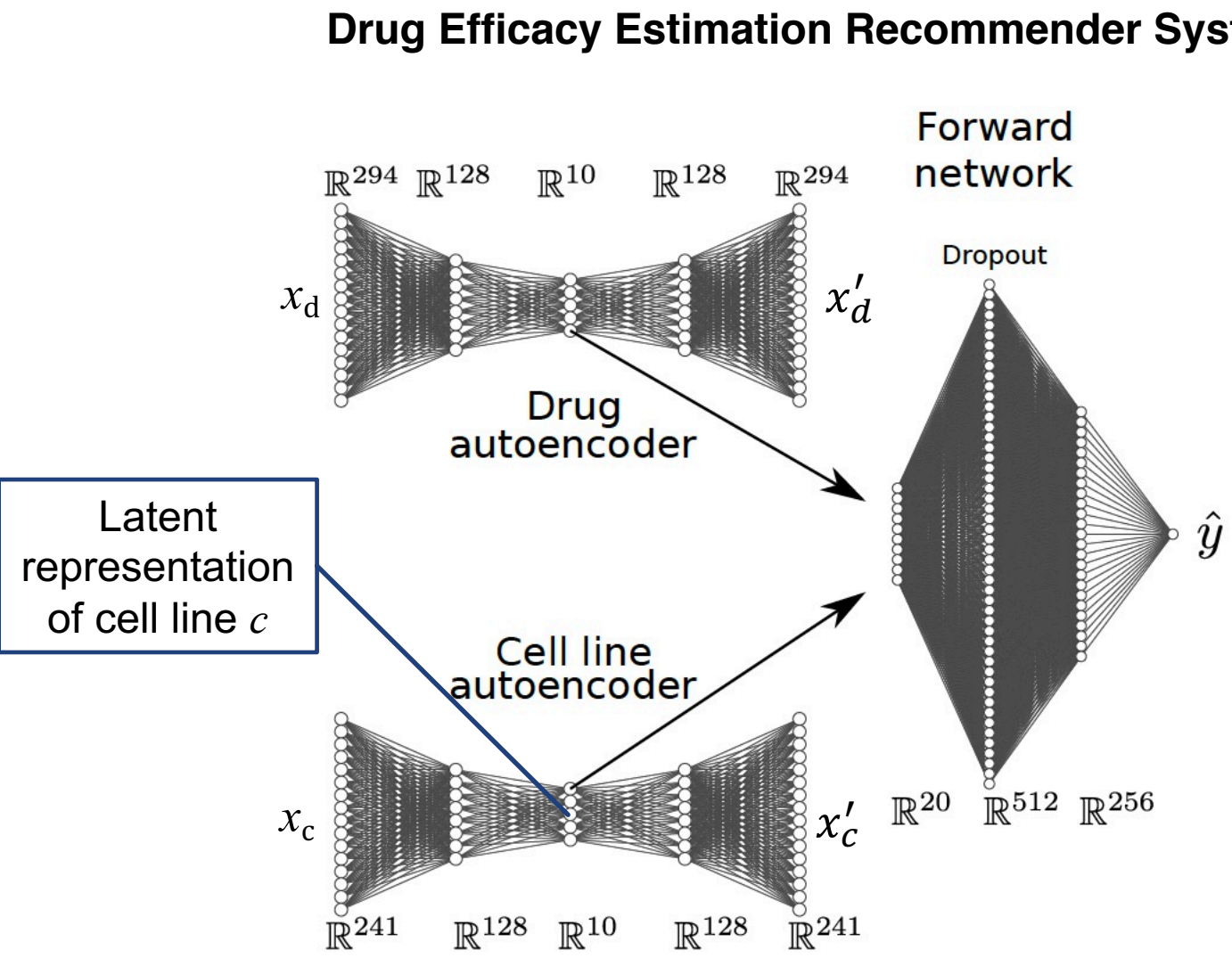
# DEERS: autoencoders for embedding and a feed forward network for the transformation

**Drug Efficacy Estimation Recommender System**



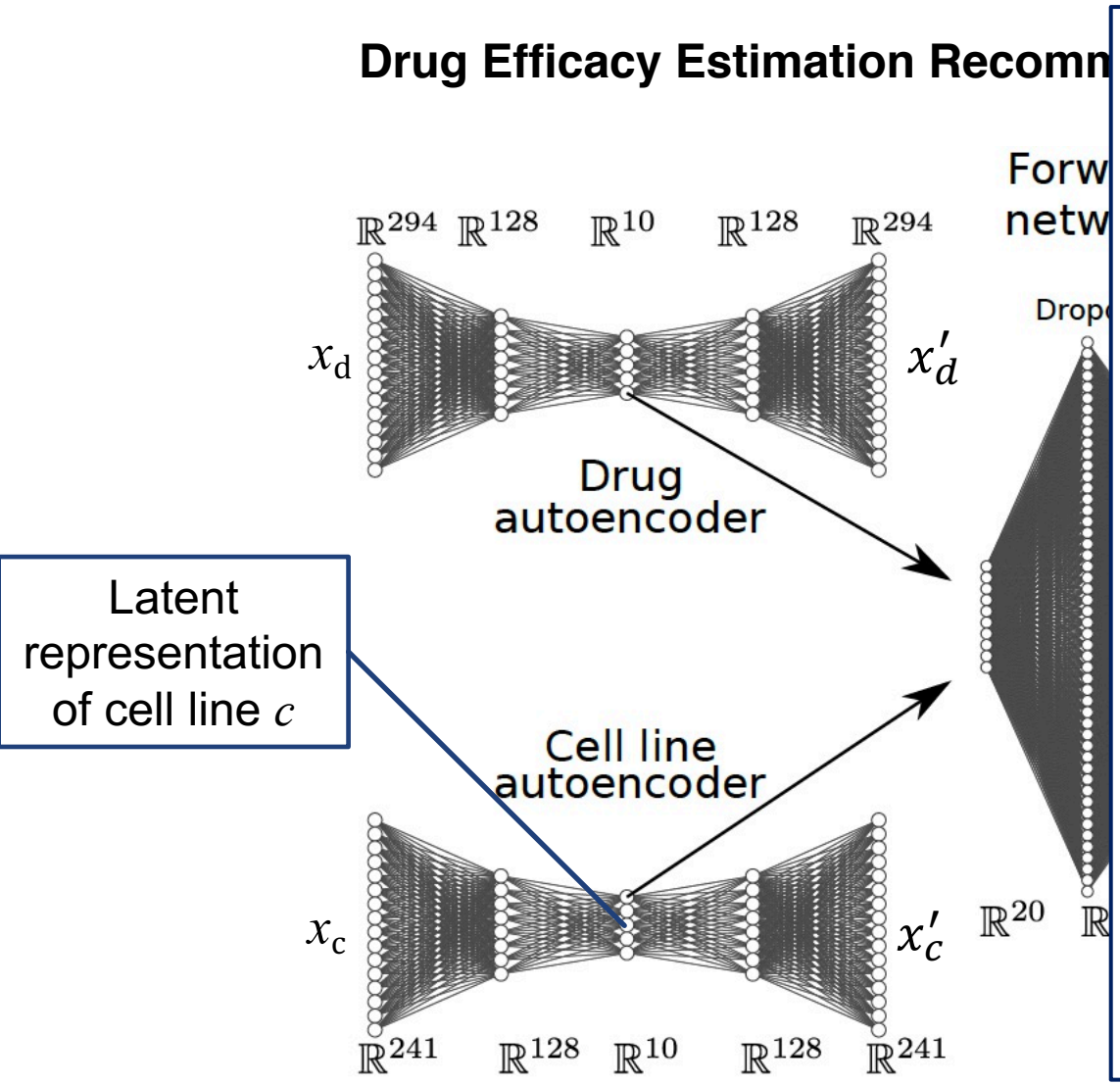

Non-linear model

# DEERS: autoencoders for embedding and a feed forward network for the transformation




Non-linear model with latent representations

# DEERS: autoencoders for embedding and a feed forward network for the transformation




Us: The model performs great

**Pharma colleagues: OK, but why?**

Us: Latent dimensions – good representations of drug and cell line data

**Pharma colleagues: OK, but what biology has this model learned?**

Us: Off-the-shelf approaches to interpretability

- Tell what in the input is associated with the response
- Are not enough.

Non-linear model with latent representations

# Custom interpretability analysis of the cell line autoencoder

- $\boldsymbol{Z}_{\mathrm{C}}$ – 10 x 922 matrix of latent dimension values for the cell lines
- $\boldsymbol{Z}_{\mathrm{C}}[:, c]$ – vector of 922 values for dimension $c$

# Custom interpretability analysis of the cell line autoencoder

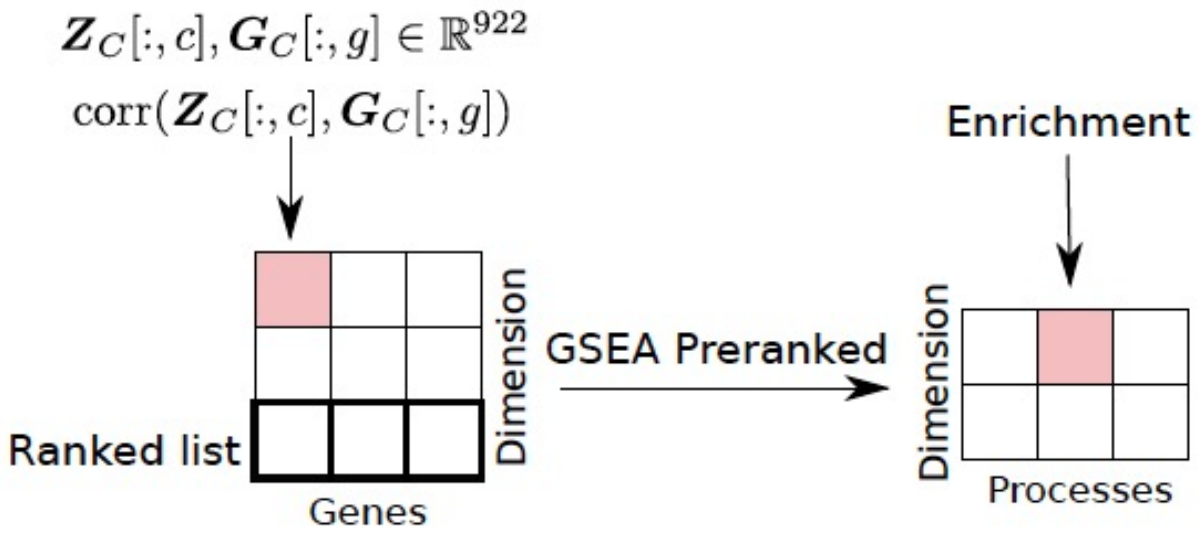
- $\boldsymbol{Z}_{\mathrm{C}}$ – 10 x 922 matrix of latent dimension values for the cell lines
- $\boldsymbol{Z}_{\mathrm{C}}[:, c]$ – vector of 922 values for dimension $c$

- $\boldsymbol{G}_{\mathrm{C}}$ – matrix of gene expression values for the cell lines **– for ~17,000 genes not seen by the model**
- $\boldsymbol{G}_{\mathrm{C}}[:, g]$ – vector of 922 values of expression of gene $g$

# Custom interpretability analysis of the cell line autoencoder

- $\boldsymbol{Z}_{\mathrm{C}}$ – 10 x 922 matrix of latent dimension values for the cell lines
- $\boldsymbol{Z}_{\mathrm{C}}[:, c]$ – vector of 922 values for dimension $c$

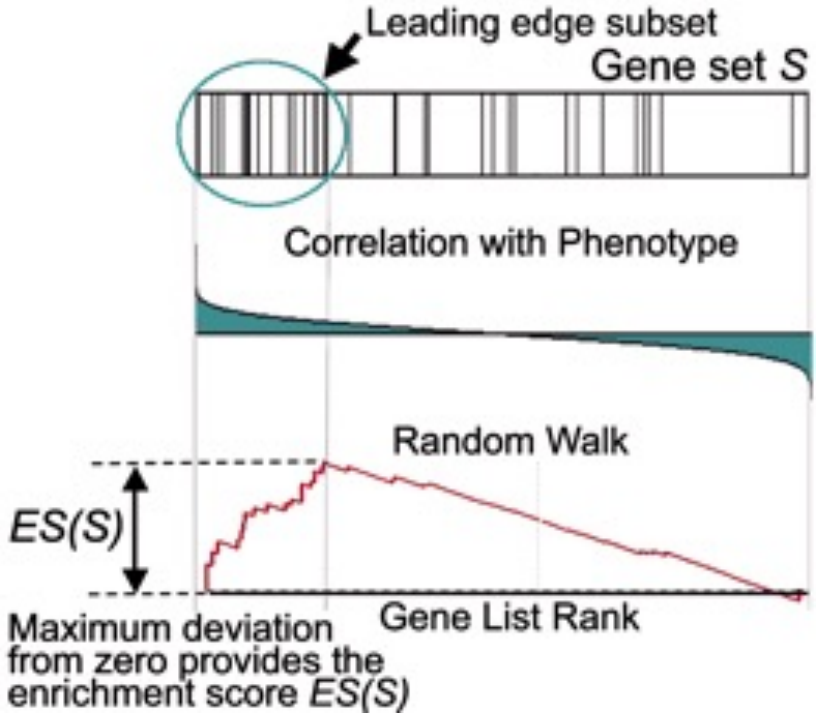
- $\boldsymbol{G}_{\mathrm{C}}$ – matrix of gene expression values for the cell lines **– for ~17,000 genes not seen by the model**
- $\boldsymbol{G}_{\mathrm{C}}[:, g]$ – vector of 922 values of expression of gene $g$




Genes are ranked by their correlation with the latent dimension

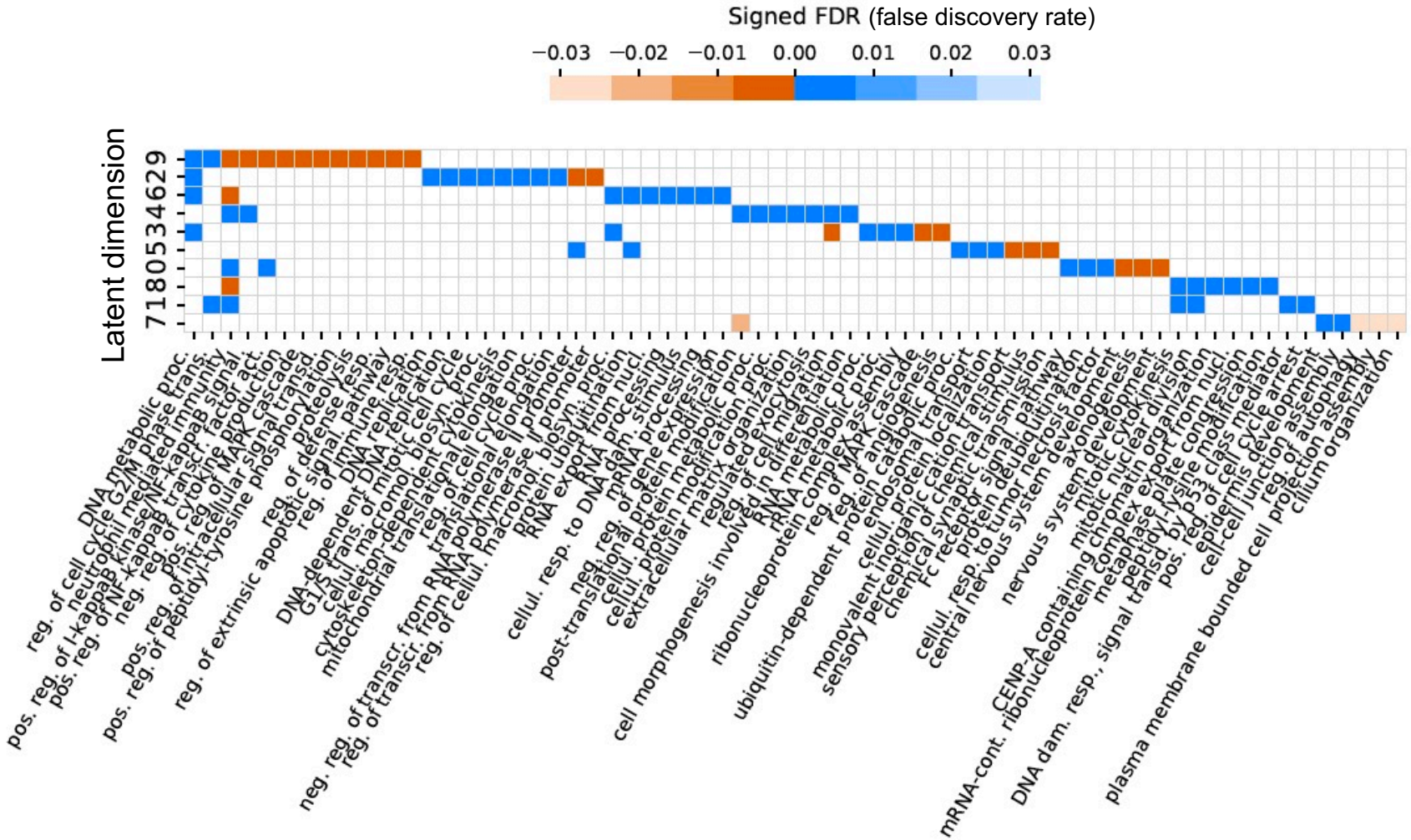
# Custom interpretability analysis of the cell line autoencoder

- GSEA pre-ranked – what is it?
- Computing enrichment of a set of genes *S* on the top of a ranked list



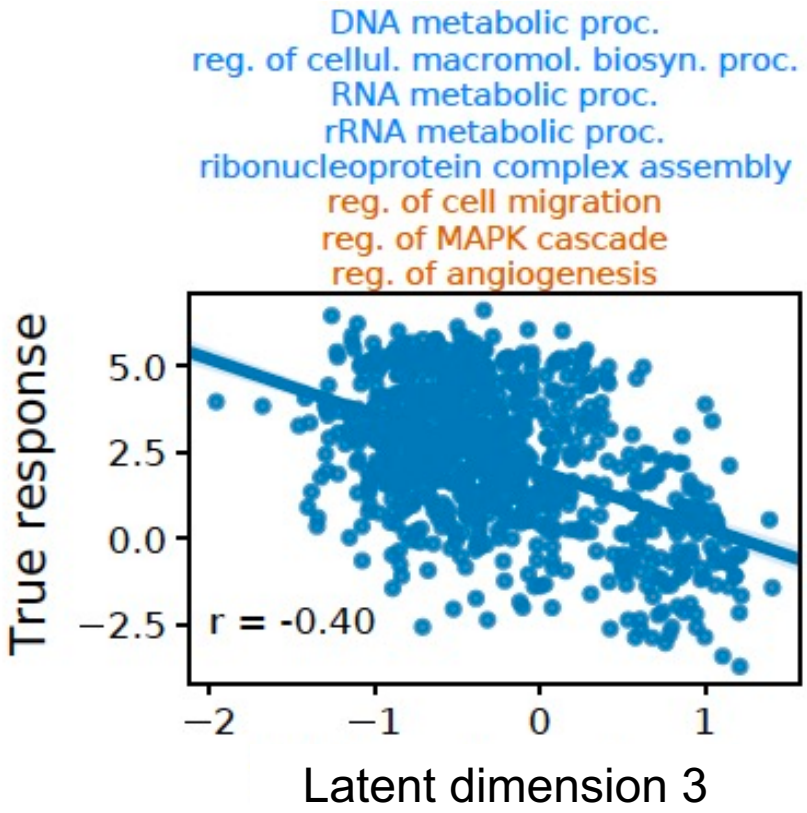
# Custom interpretability analysis of the cell line autoencoder

- Enrichment of the gene lists ranked by correlation with latent dimensions in known gene sets related to cellular processes

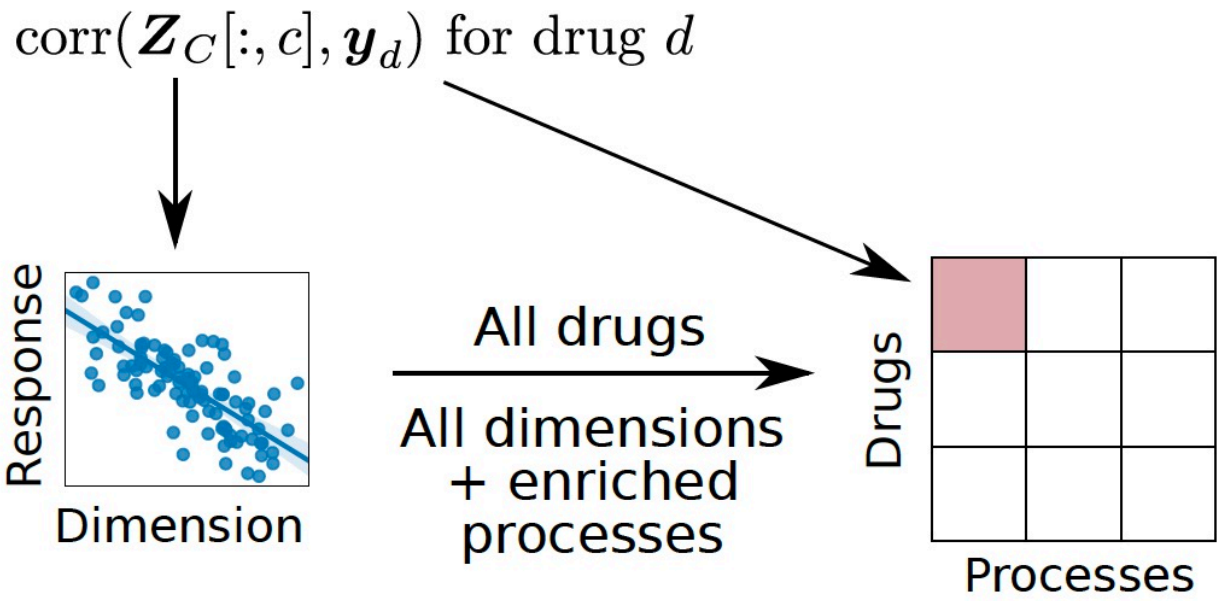


# Custom interpretability analysis of the drugs

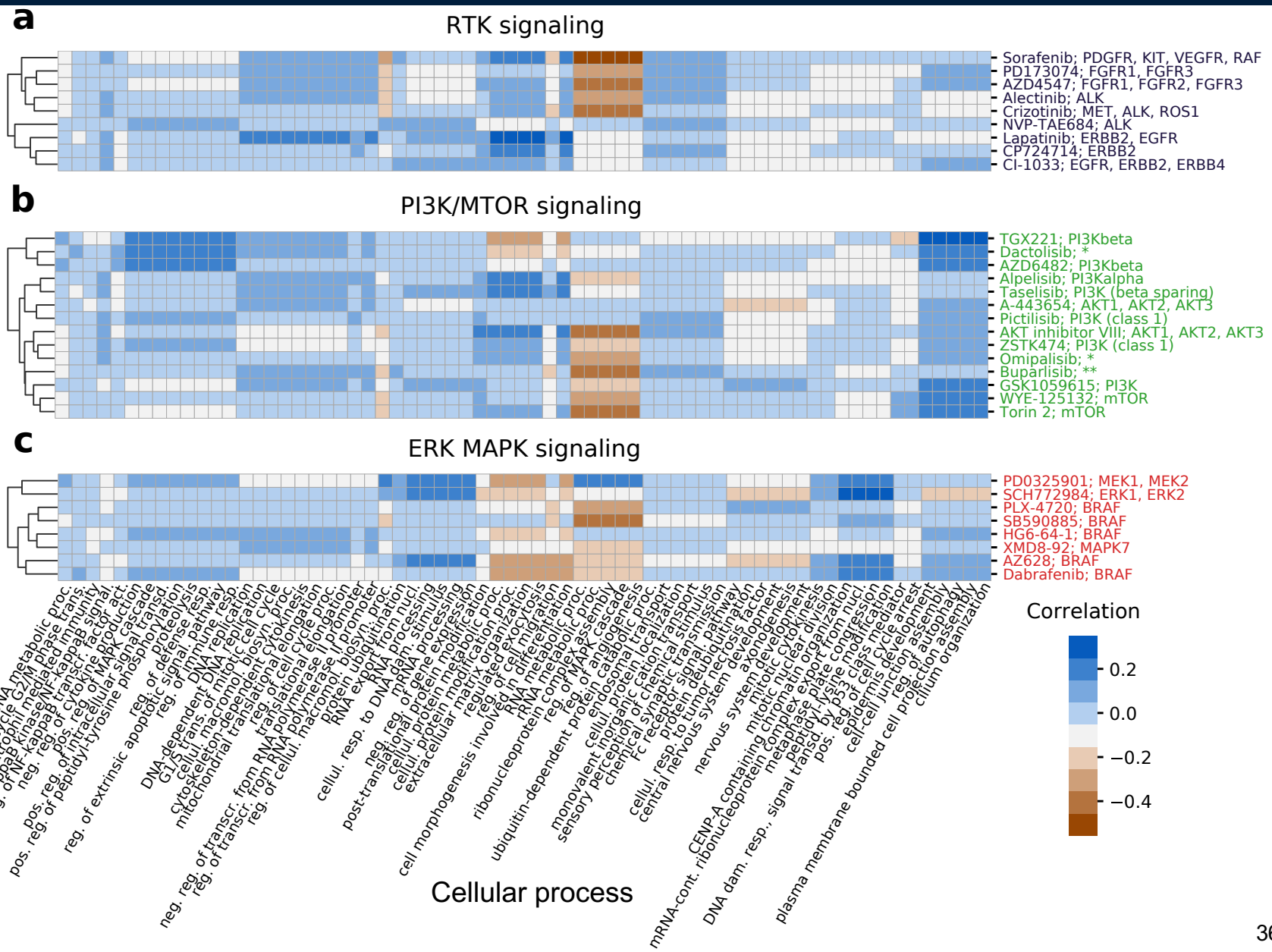
**PHA-793887**
CDK inhibitor
Used to treat leukemia



# Custom interpretability analysis of the drugs



# Custom interpretability analysis of the drugs

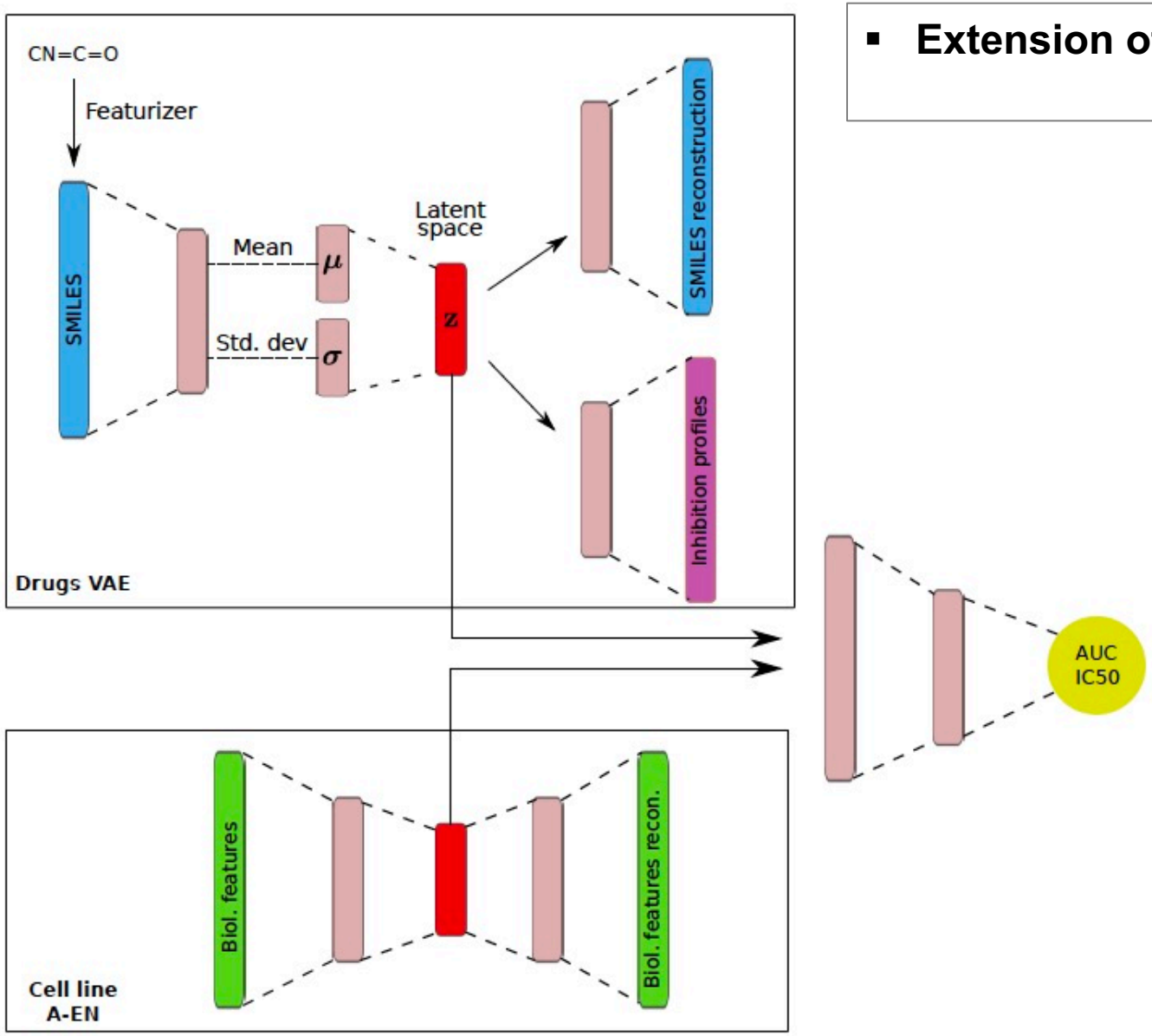


# Summary of the DEERS model

- Deep recommender system approach to predicting response of cancer cell lines to drugs based on drug and cell line features
- Custom approach to model interpretability
- Revealing general mechanisms of drug action

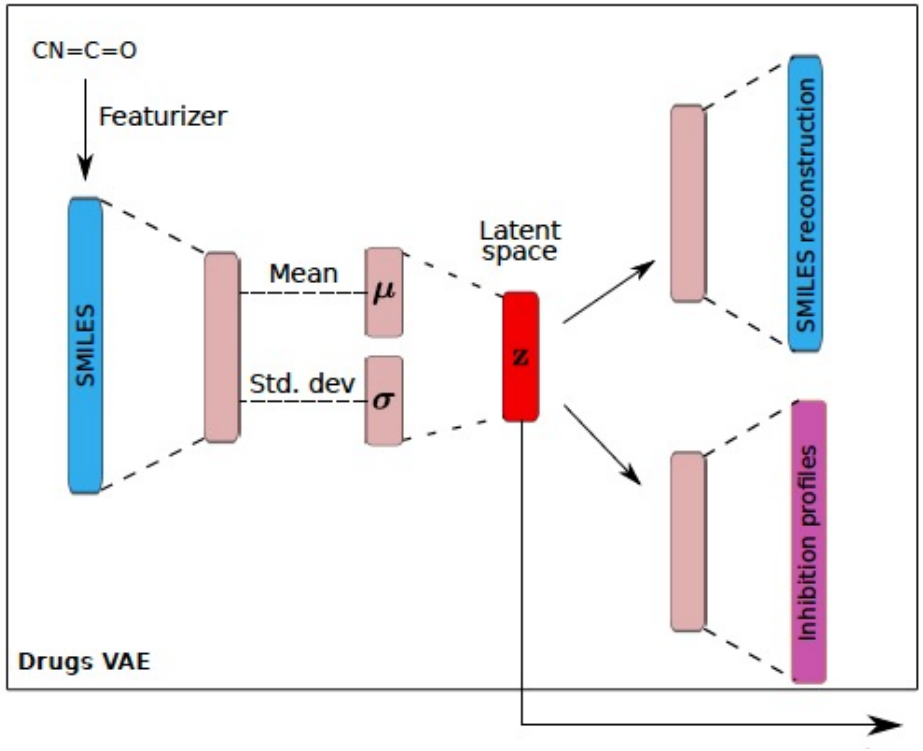
# Drug variational autoencoders with latent space clustering

# Drug variational autoencoders with latent space clustering



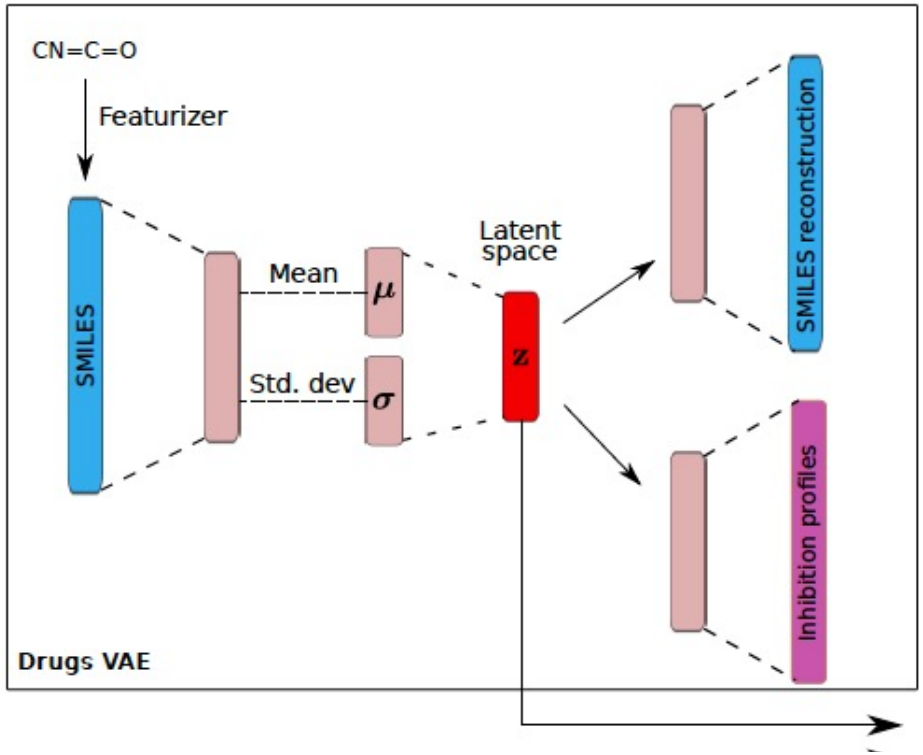

- **Extension of DEERs**

# Drug variational autoencoders with latent space clustering




- **Extension of DEERs**
- **The Drugs VAE**
  - Generative model for drugs
  - Takes a drug representation as input
  - Outputs an inhibition profile

# Drug variational autoencoders with latent space clustering

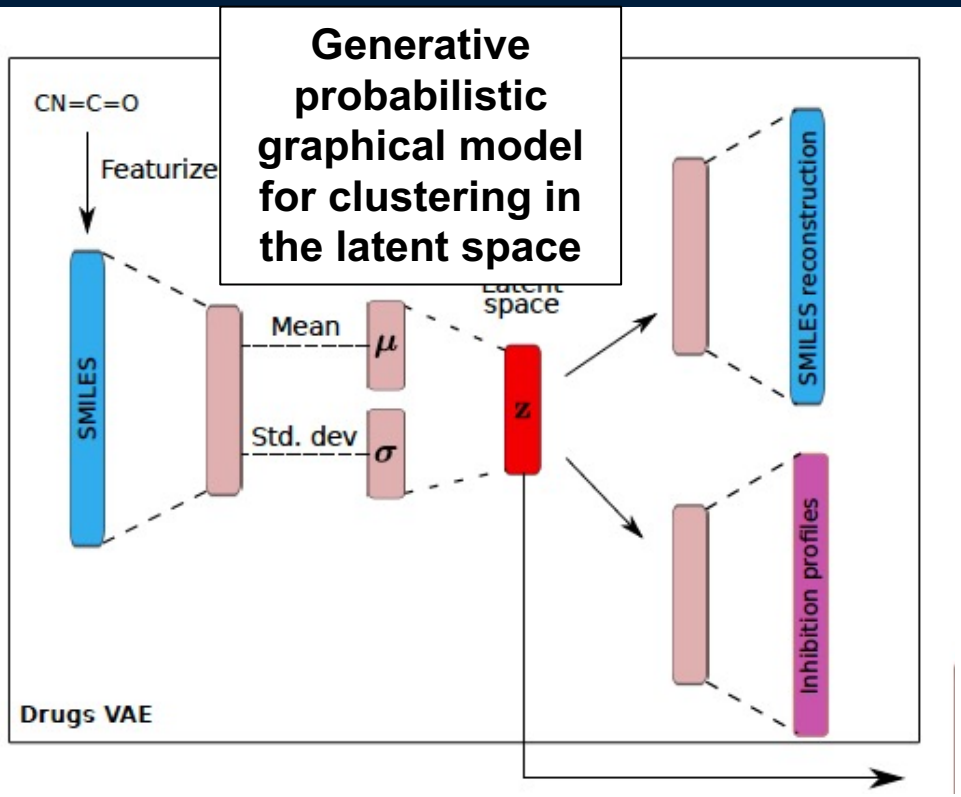



- **Main assumptions:**

- Drugs cluster by their inhibition profiles (*guiding data*)

- Drugs with similar inhibition profiles should also cluster in the latent space

# Drug variational autoencoders with latent space clustering

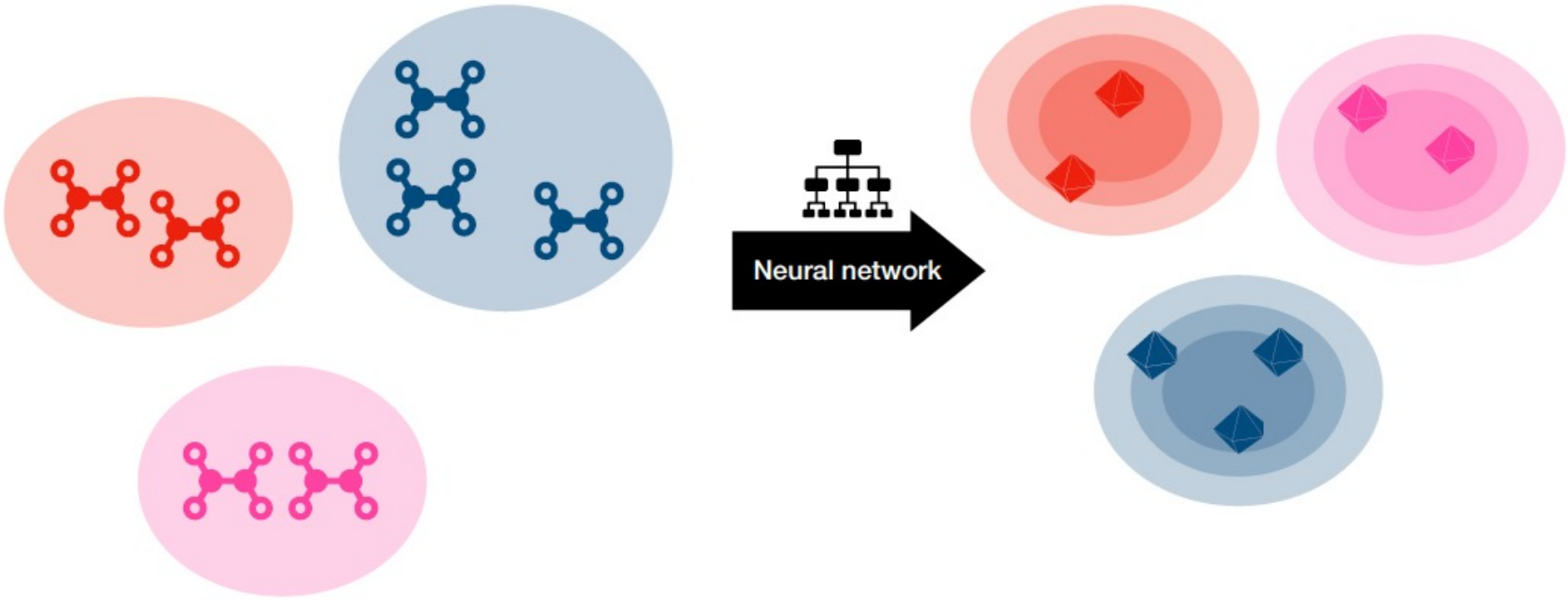



- **Main assumptions:**

- Drugs cluster by their inhibition profiles (*guiding data*)

- Drugs with similar inhibition profiles should also cluster in the latent space

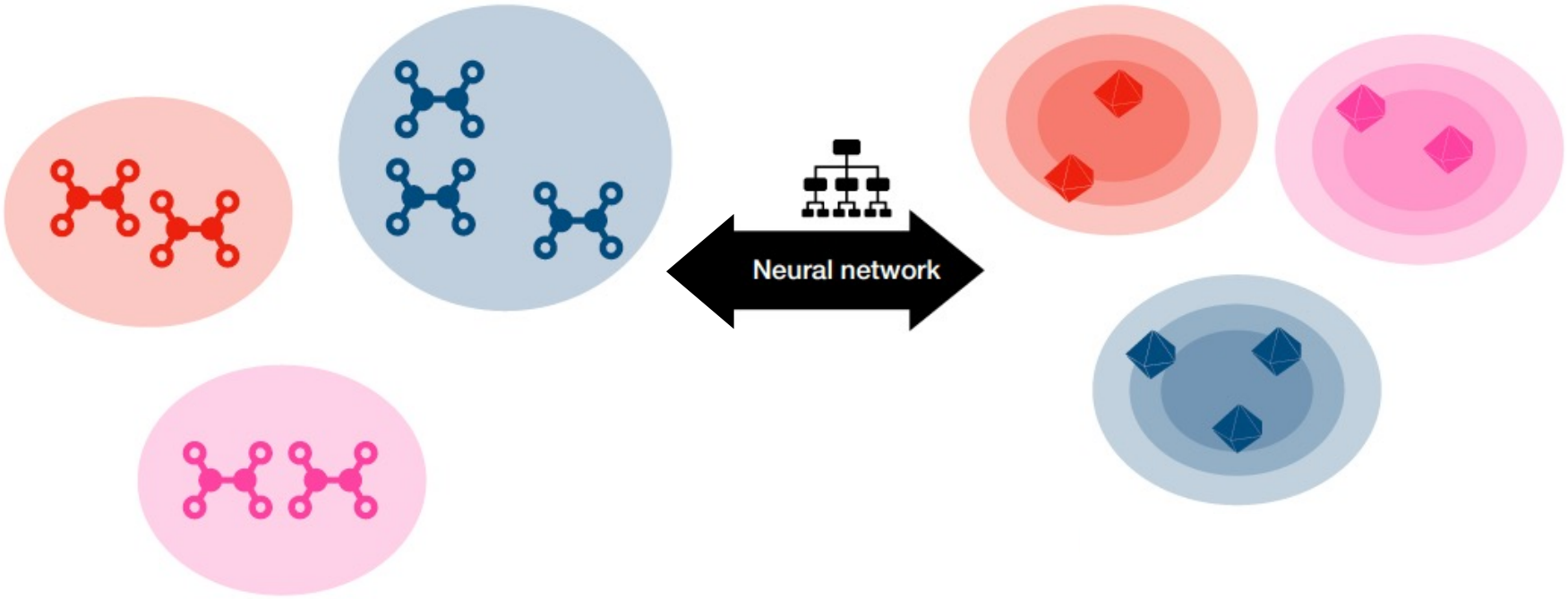
# Drug variational autoencoders with latent space clustering (1)




**Given clustering in the guiding data**

**Learning the clustering in the latent space**
Mixtures of Gaussians

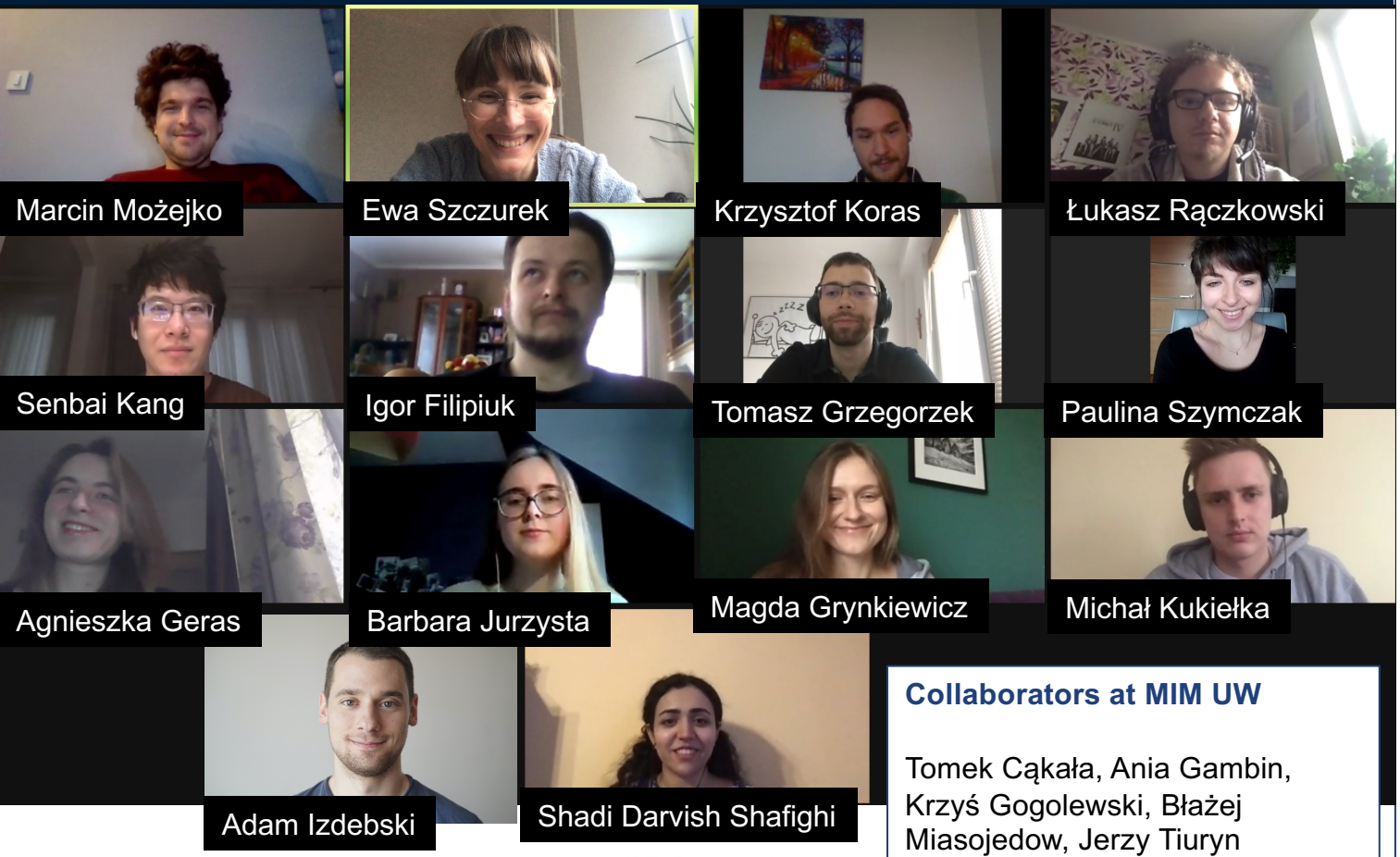
# Drug variational autoencoders with latent space clustering (2)




**Learning the clustering in the guiding data**
Mixtures of Gaussians

**Learning the clustering in the latent space**
Mixtures of Gaussians

# Acknowledgements




**Collaborators at MIM UW**

Tomek Cąkała, Ania Gambin, Krzyś Gogolewski, Błażej Miasojedow, Jerzy Tiuryn


**Funding**

UNIVERSITY OF WARSAW

HORIZON 2020

MERCK

NATIONAL SCIENCE CENTRE POLAND


**Collaborating labs**

**ETH Zurich**
Niko Beerenwinkel, Jack Kuipers

**Jagiellonian University**
Tomasz Kościółek

**Leiden Med Center, The Netherlands**
Kees van Bergen, Szymon Kiełbasa

**KTH, Sweden**
Nicola Crosetto, Jens Lagergren

**Medical University of Lublin, Poland**
Paweł Krawczyk

**Merck Biopharma, Germany**
Eike Staub

**Sorbonne, France**
Alessandra Carbone

**Uni Vigo, Spain**
David Posada

**Warsaw Medical Uni**
Dominika Nowis, Łukasz Koperski

**Uni Wrocław**
Tyll Krueger

# Positions, collaboration

- https://www.mimuw.edu.pl/~szczurek/positions.html
- Looking for 2 postdocs!
- Please apply at szczurek@mimuw.edu.pl

Thank you for your attention!
Questions?

www.nature.com/scientificreports

SCIENTIFIC REPORTS

natureresearch



OPEN

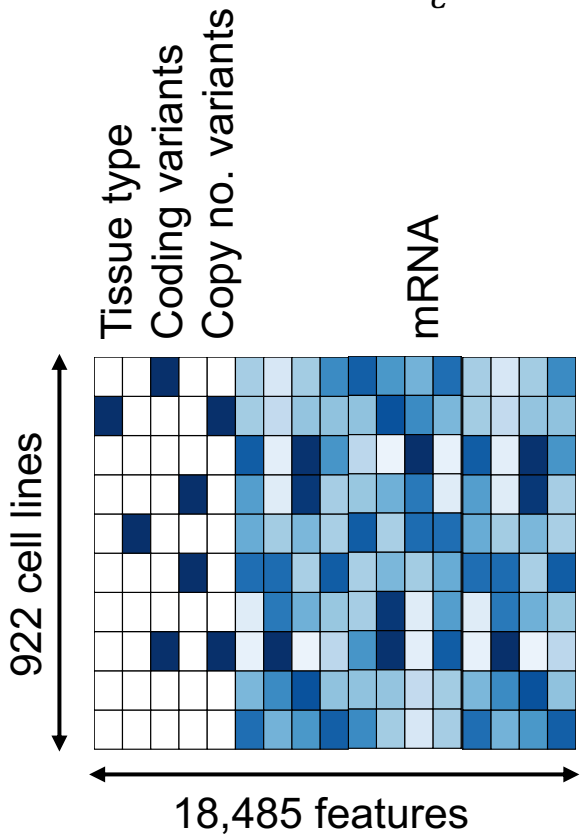
# Feature selection strategies for drug sensitivity prediction

Krzysztof Koras[1], Dilafruz Juraeva[2], Julian Kreis[2], Johanna Mazur[2], Eike Staub[2] & Ewa Szczurek[1]✉

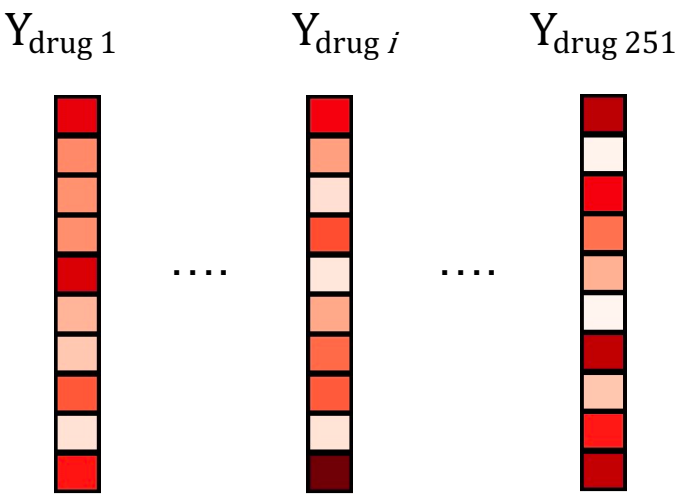
Drug sensitivity prediction constitutes one of the main challenges in personalized medicine. Critically, the sensitivity of cancer cells to treatment depends on an unknown subset of a large number of biological features. Here, we compare standard, data-driven feature selection approaches to feature selection driven by prior knowledge of drug targets, target pathways, and gene expression signatures. We asses these methodologies on Genomics of Drug Sensitivity in Cancer (GDSC) dataset, evaluating 2484 unique models. For 23 drugs, better predictive performance is achieved when the features are selected according to prior knowledge of drug targets and pathways. The best correlation of observed and predicted response using the test set is achieved for Linifanib ($r = 0.75$). Extending the drug-dependent features with gene expression signatures yields the most predictive models for 60 drugs, with the best performing example of Dabrafenib. For many compounds, even a very small subset of drug-related features is highly predictive of drug sensitivity. Small feature sets selected using prior knowledge are more predictive for drugs targeting specific genes and pathways, while models with wider feature sets perform better for drugs affecting general cellular mechanisms. Appropriate feature selection strategies facilitate the development of interpretable models that are indicative for therapy design.

# High-dimensional cell line data & drug response measurements

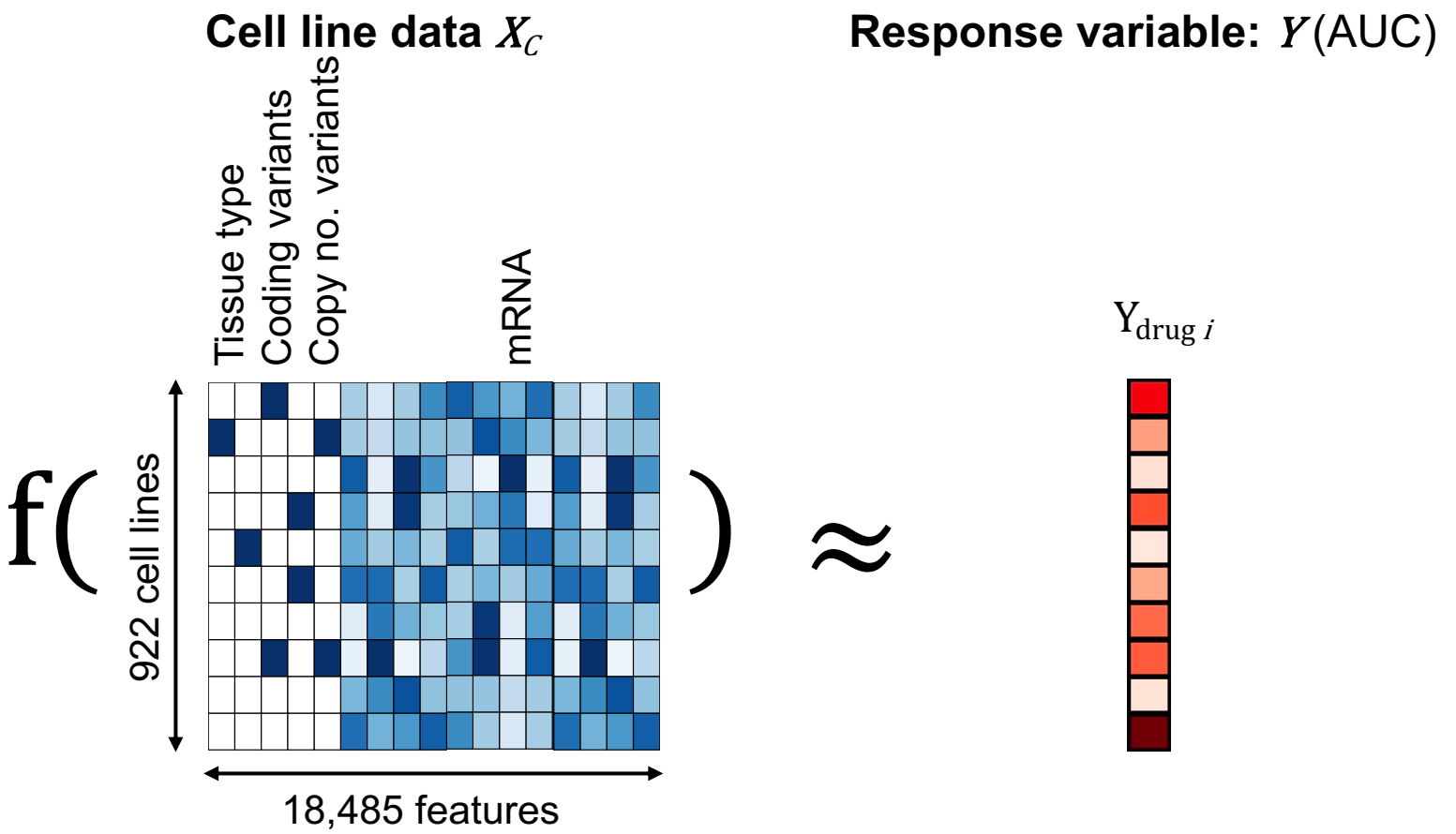
**Cell line data $X_C$**




**Response variable: $Y$(AUC)**

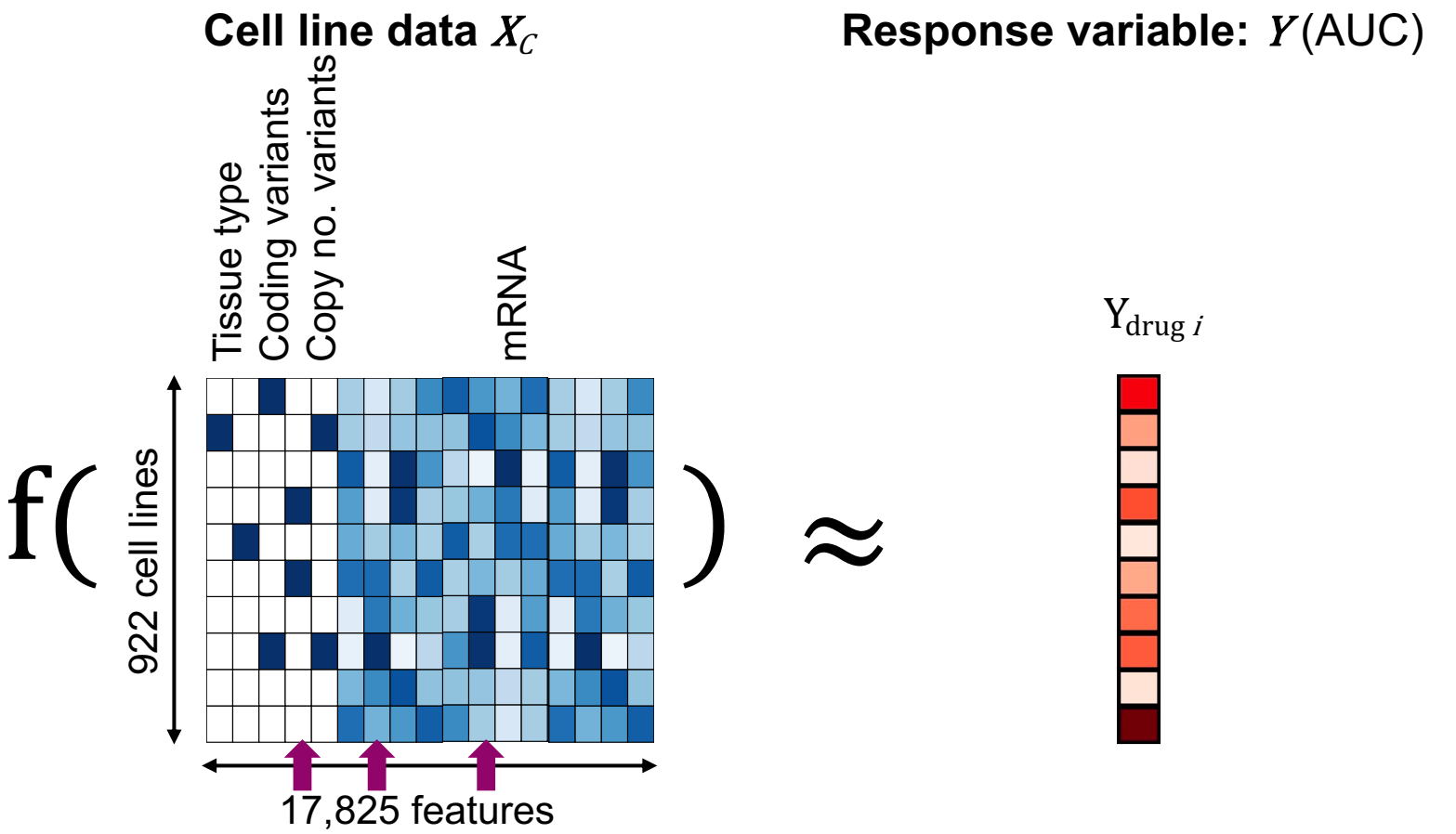


# Prediction




- Use $X_c$ and $Y_{drug\ i}$ as training data for a model f
- When a new observation $x$ comes, $f(x)$ should be *close to* the true $y$
- Our models: elastic net (linear), random forest (non-linear)
- Evaluation measure: Correlation, RMSE

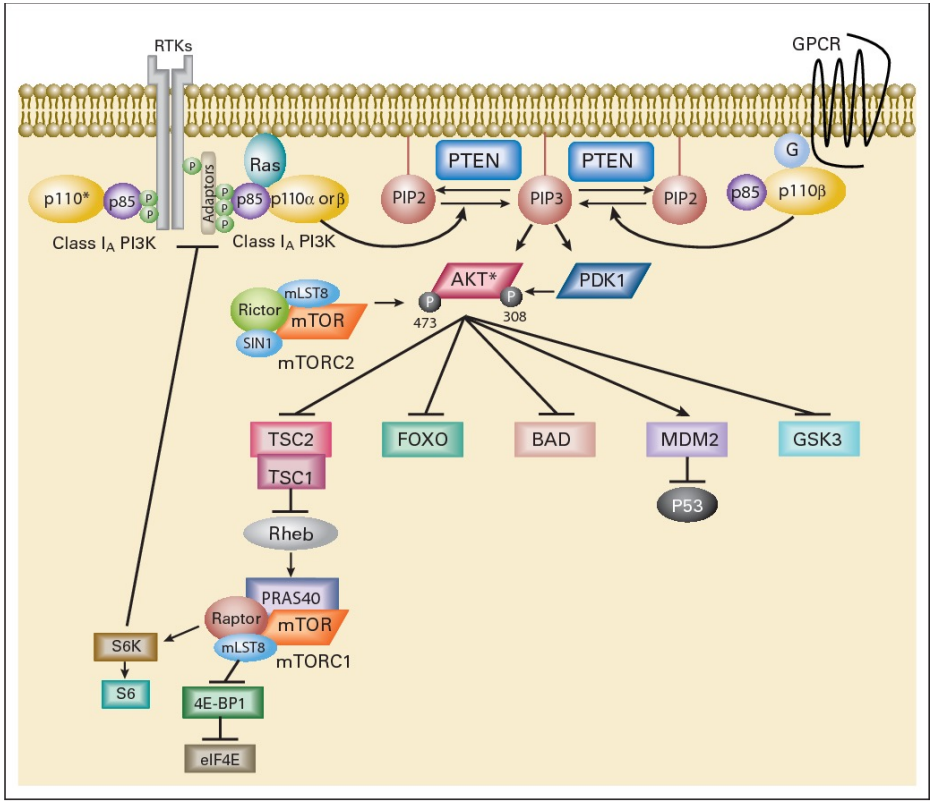
# Feature selection




- Given a model f, identify such a relatively small set of **features** that are most informative for that model's prediction.
- Elastic net and random forest offer that.
- Use prior knowledge

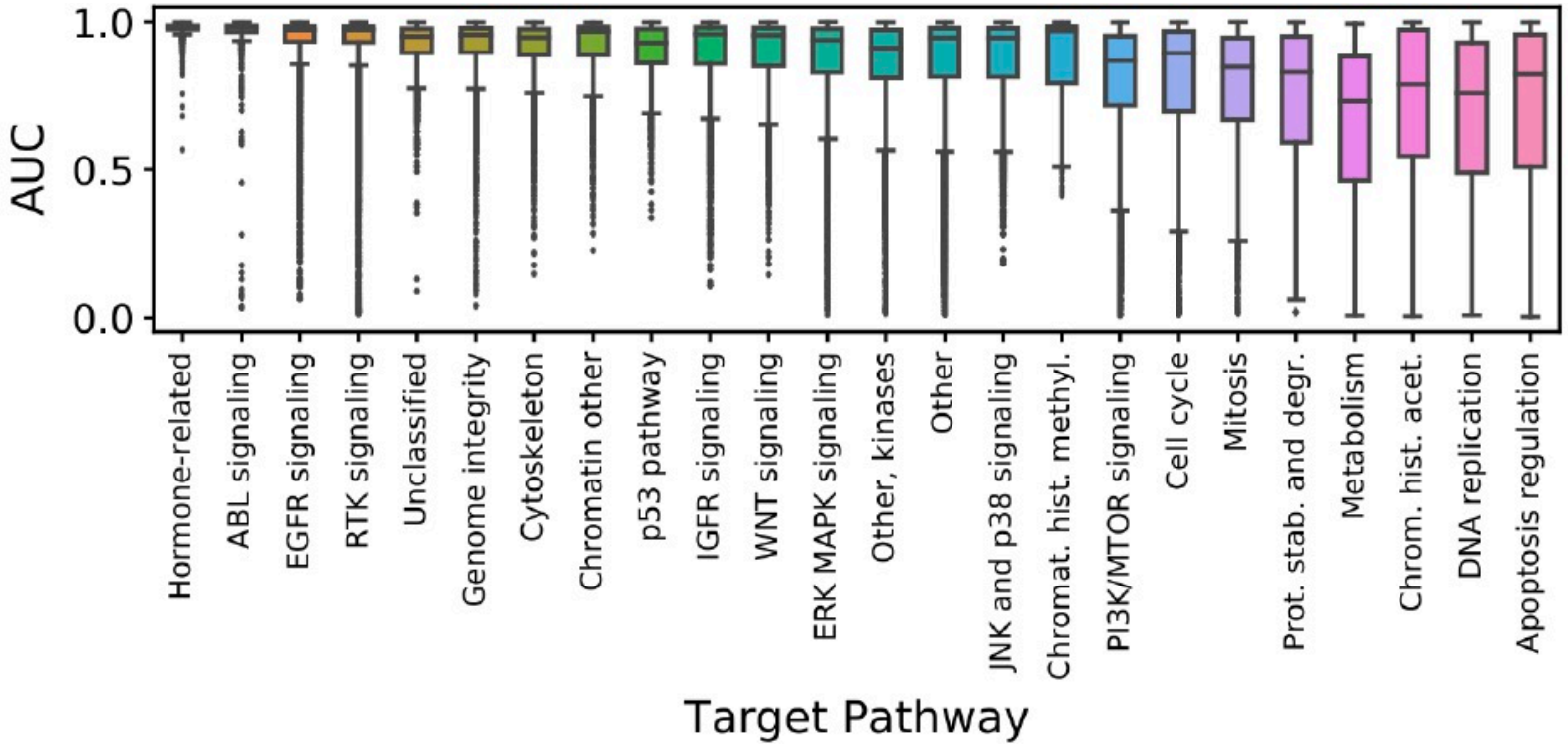
# What do we know about how the drugs work?

- Drugs have their *targets*
- Drugs have their *target pathways*
- Expression of many genes participating in a certain phenomenon can be summarized by a *gene expression signature* of a smaller set of genes

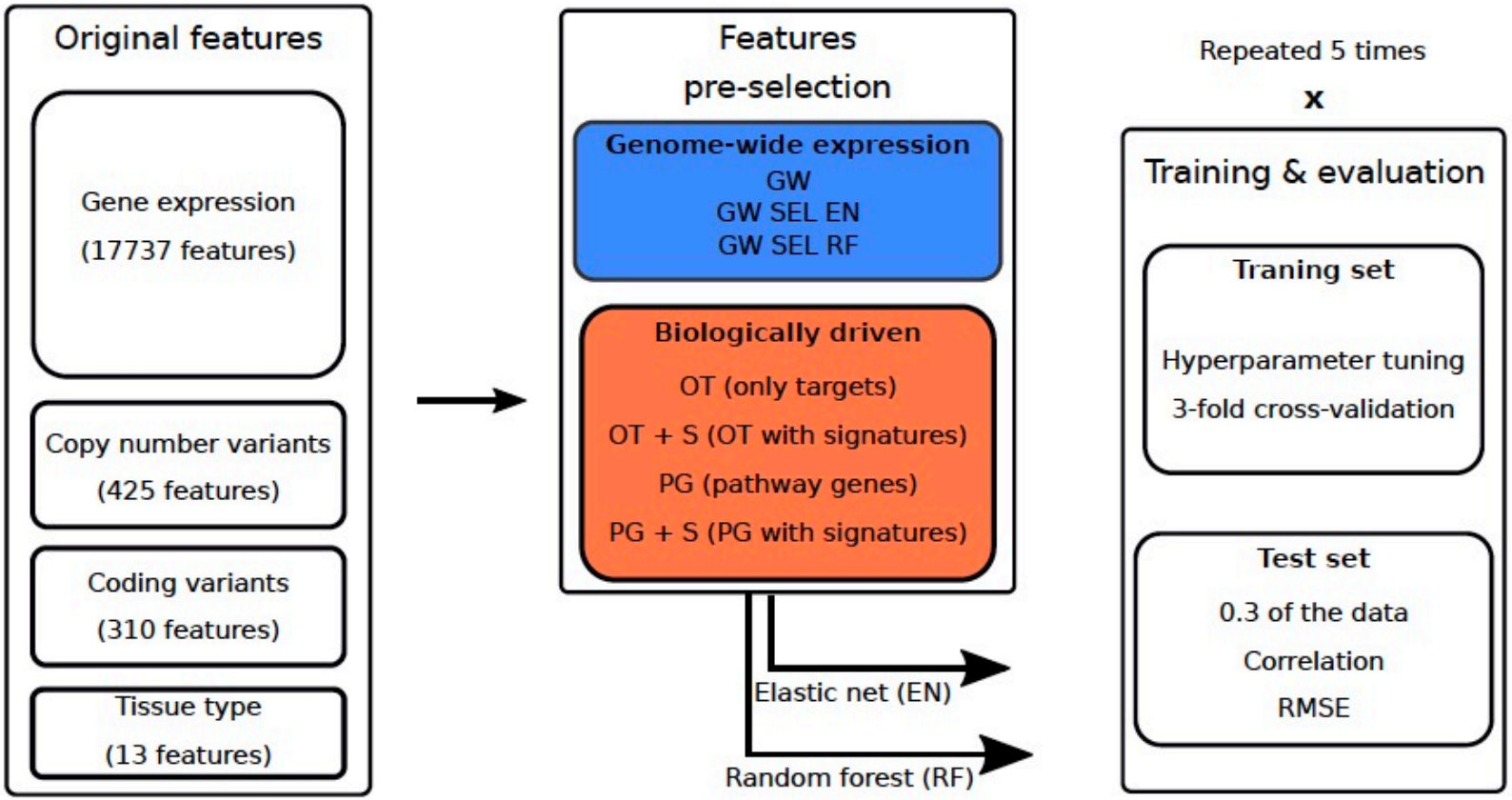



Courtney et al, Journal of clinical oncology, 2010

# AUC is different for different target pathways and is biased towards no response

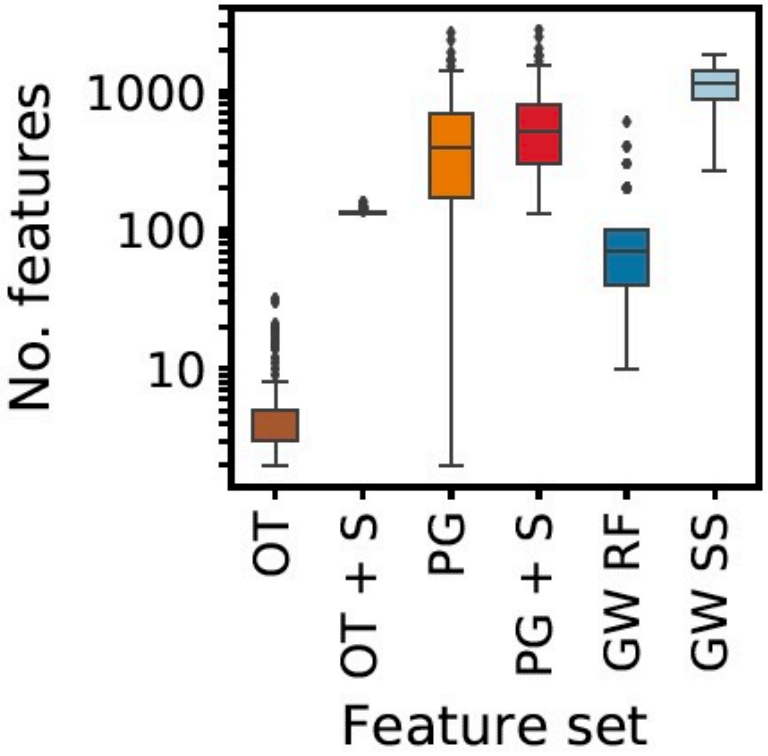


# Experimental setup

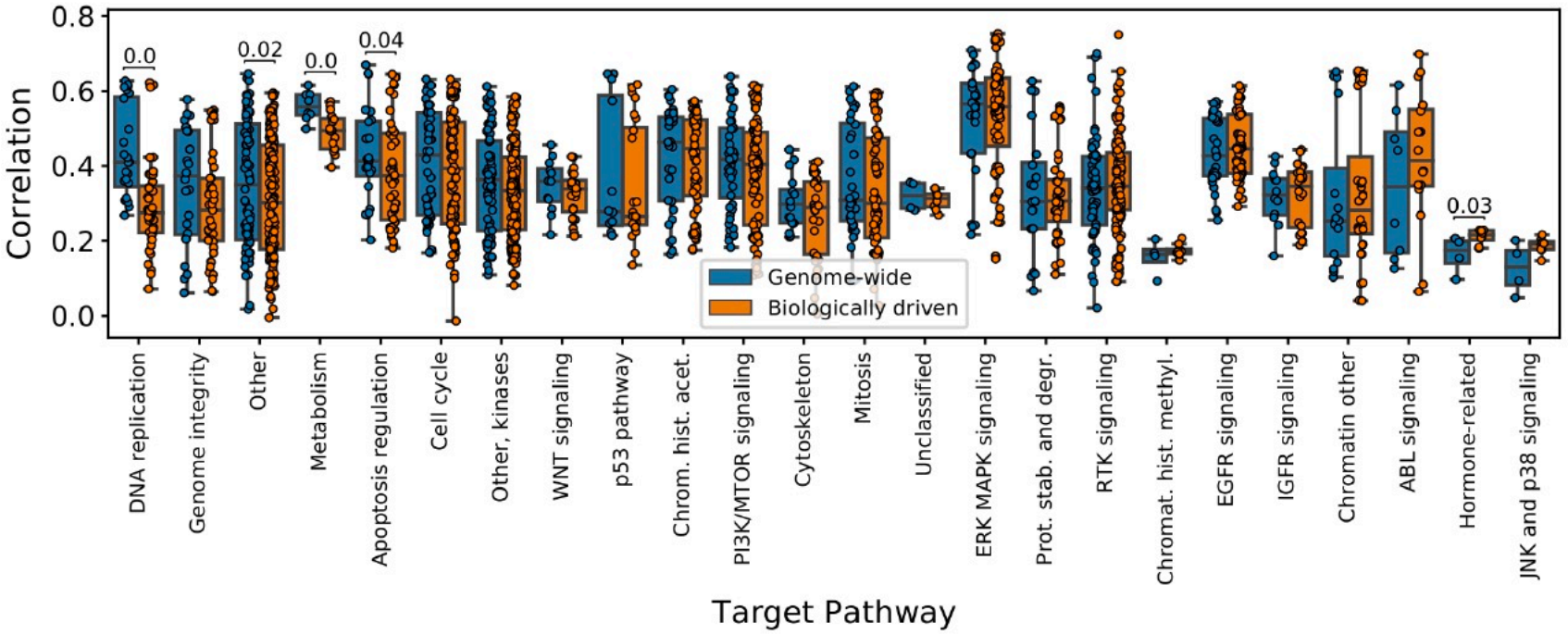



2484 models in total

# Feature selection based on target genes gives very few features



# Genome-wide and biologically driven selection perform similarly well



# For some drugs, the very few features based on targets give the best predictive performance