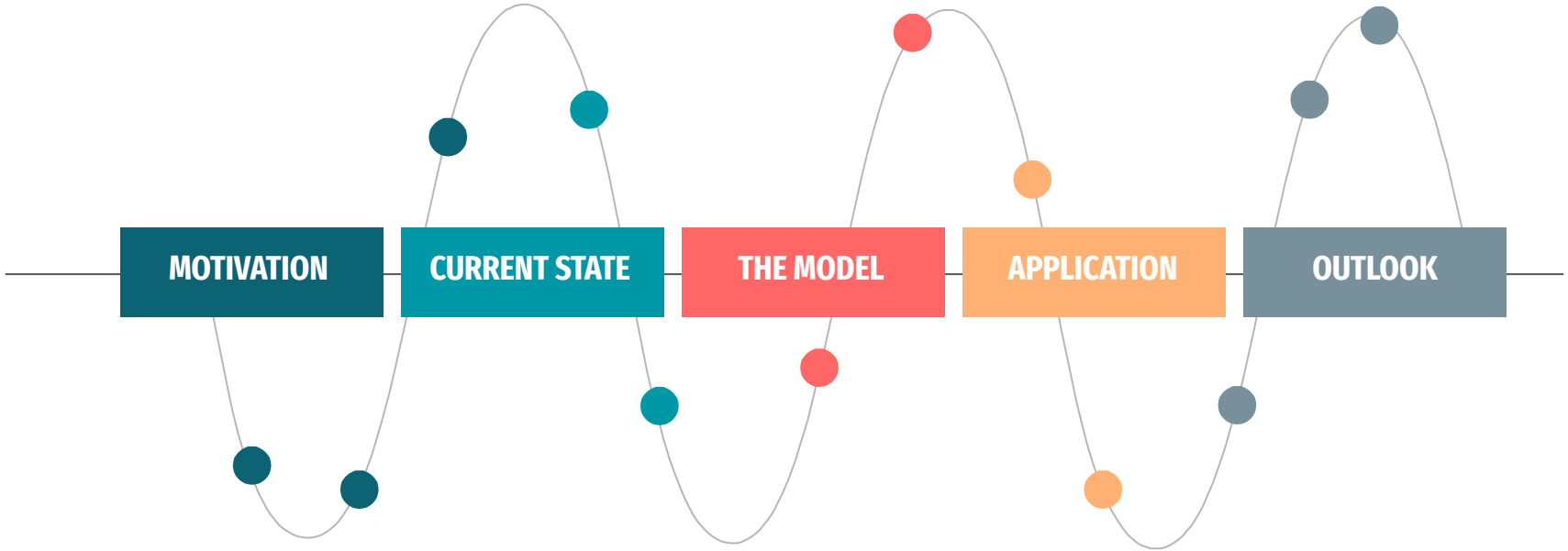
GRN Inference

Master's Thesis by Lorena Méndez

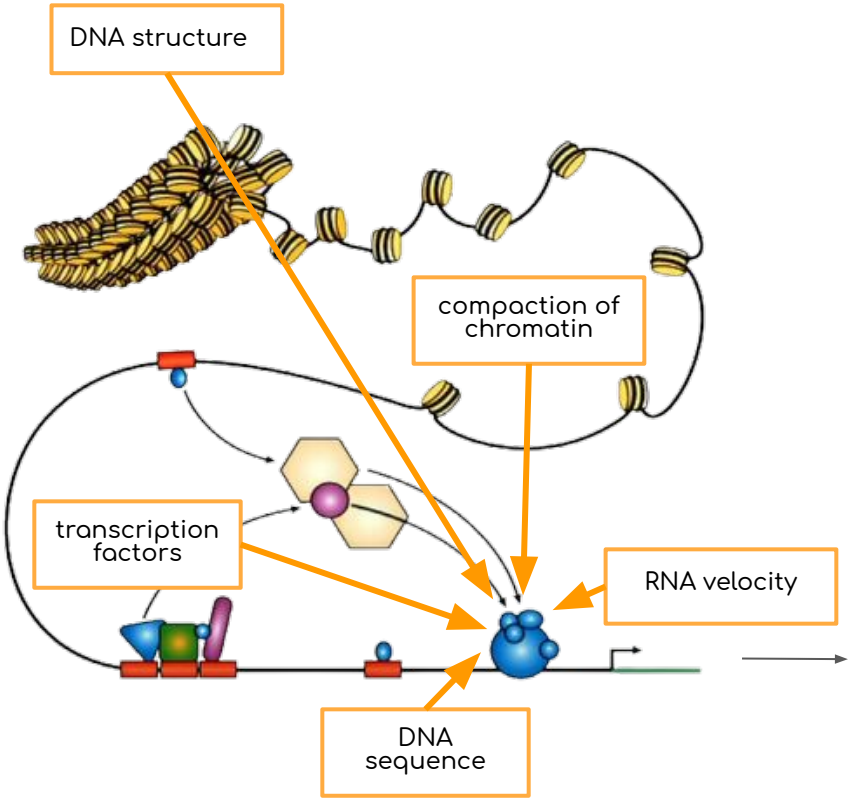
Volker Bergen

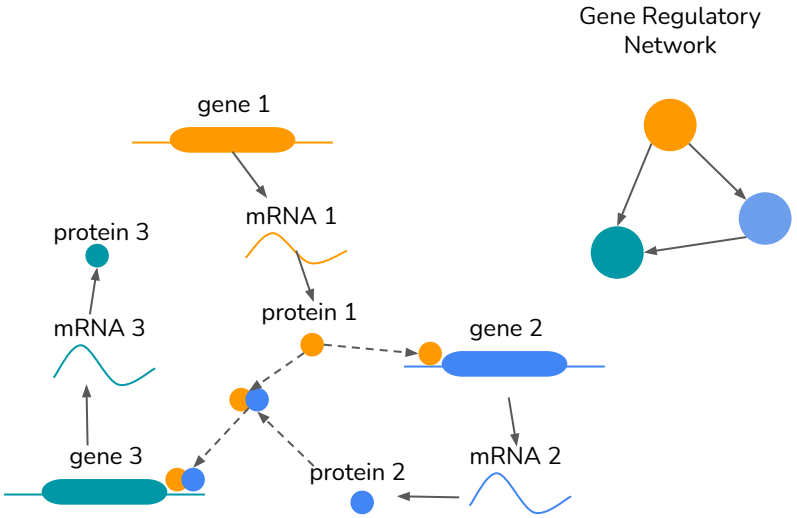
Fabian Theis

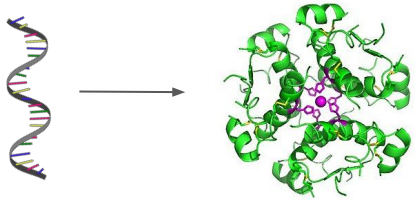
10.08.2021


MOTIVATION
CURRENT STATE
THE MODEL
APPLICATION
OUTLOOK

Transcriptional Regulation

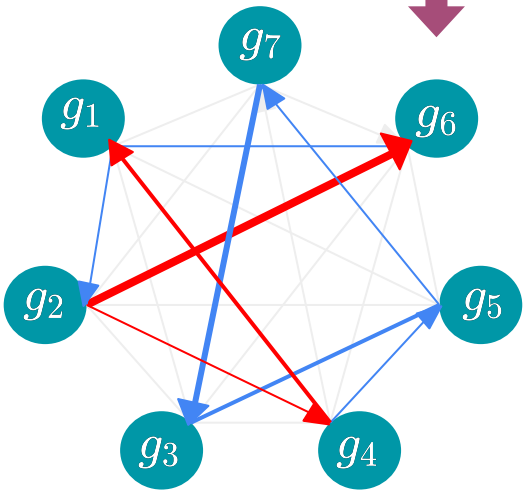




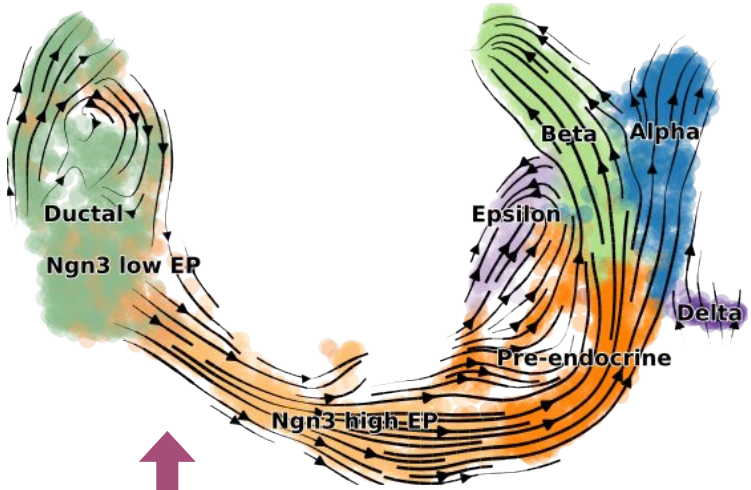


It has to do with curiosity...

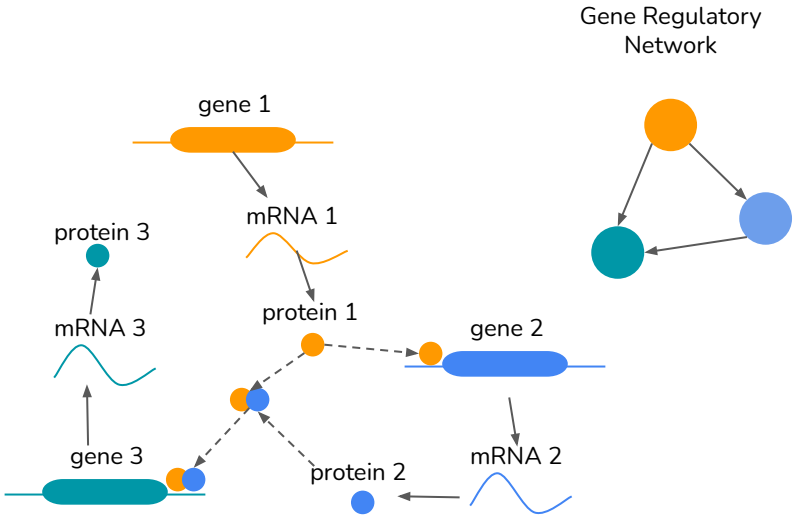
Gene Regulatory Network



RNA velocity



Gene regulatory networks



TF-gene interactions.

<u>nodes:</u> genes

<u>edges:</u> regulatory relationships

- directed (TF → gene)
- signed (+: activation, -: repression)
- weighted (strength)

Image source: [3]

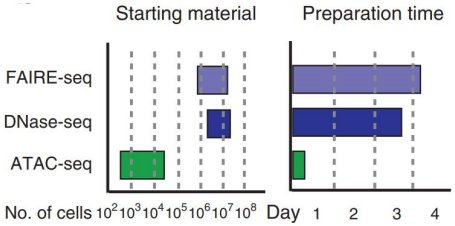
A problem over two decades...

WHY?

- causality inference
- noise
- complex interactions
- delayed response
- non-linearities
- TF combinations
- autoregulations

WHAT CAN WE DO?

- try new and different algorithms/approaches.
- incorporate complementary information (scATAC-seq and RNA velocity).



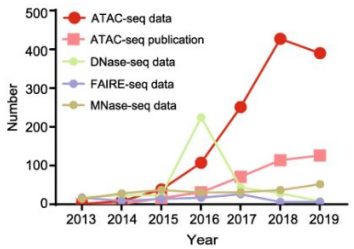
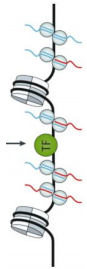


Image sources: [1] [2]

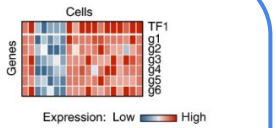
Inspiration

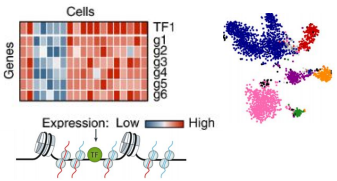


SCENIC

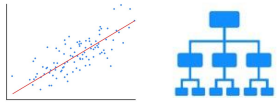
CellOracle

Input

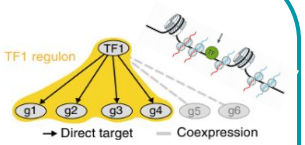


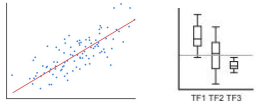


Method

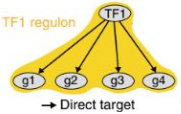


GENIE3/GRNBoost



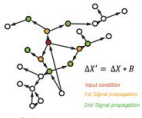


Bagging Ridge Regression



Beyond

Clustering



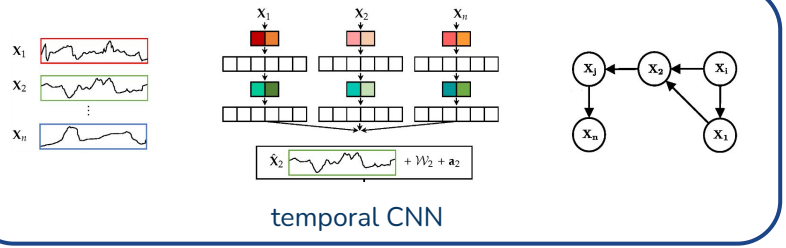
Simulation

Ideas

- non-linear relationships
- additional info for pruning
- one network per system
- Problem: missing temporal connection, own ATAC-seq.

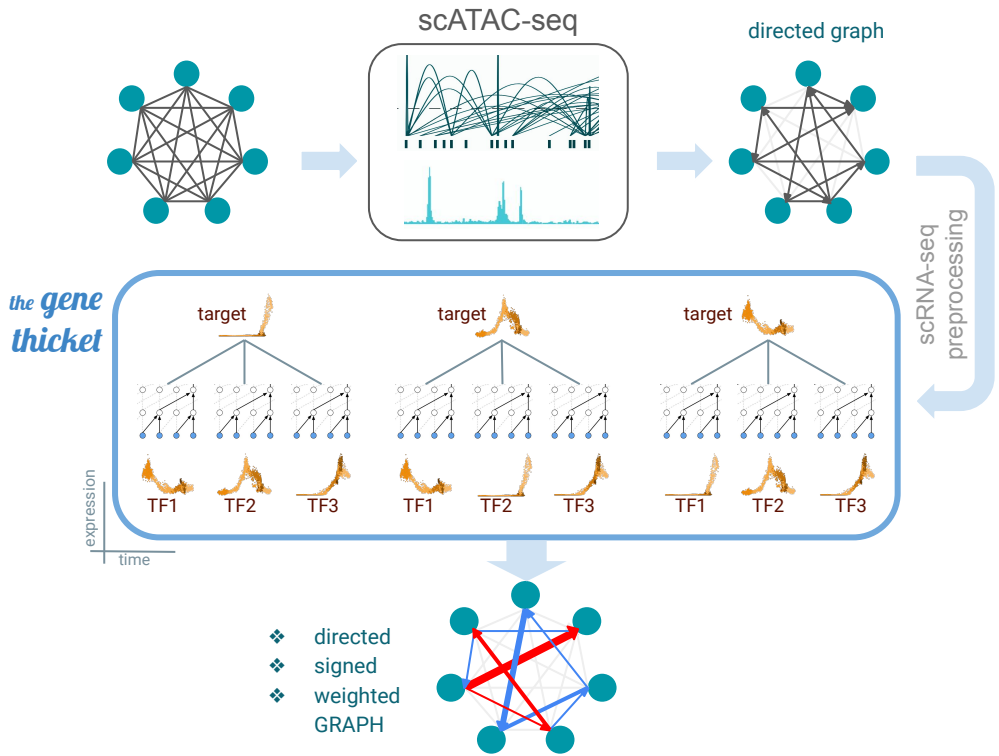
- ATAC-seq preprocessing
- ATAC data as prior
- Simulation. Problem: only measures the effect of the change in TF expression (need to know t0, t1). **TIME SERIES**
- Problem: linear relationships, network per cluster, missing temporal connection.

TCDF



- non-linear relationships
- directed graph through attention interpretation
- delay discovery
- causal validation
- Problem: not for genes, no ATAC-seq, no weights, no signs, depends on present and past.

The GeneThicket

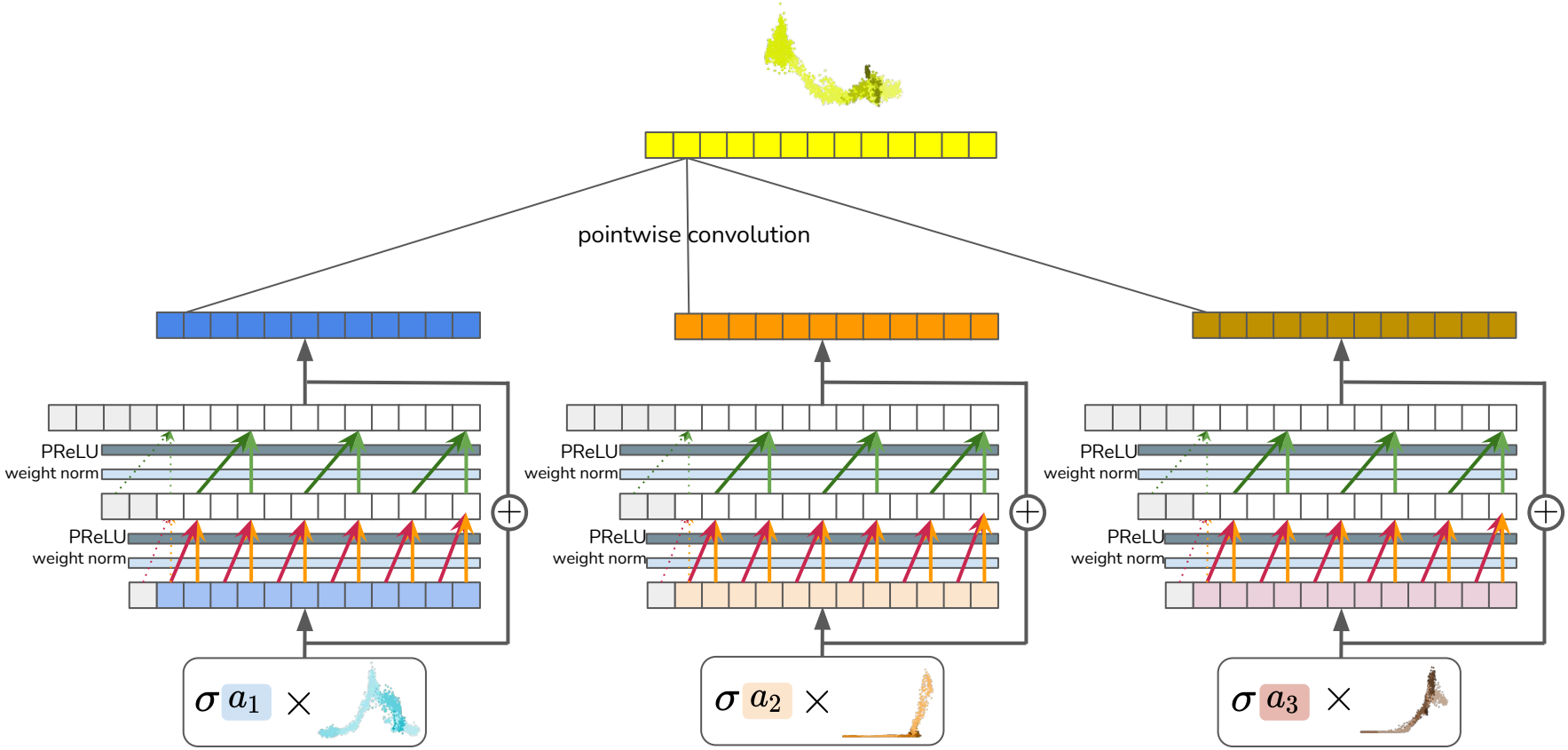


Novelties

GRN inference method that predicts the future of each cell

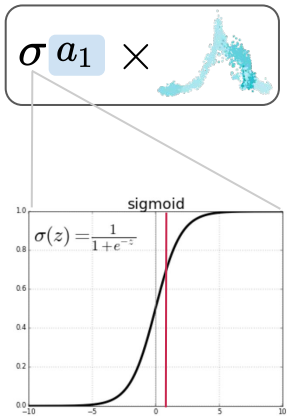
- ATAC-seq info as prior
- CNN: architecture, prediction using only the past, early stopping, Xavier initialization, weight normalization
- added signs
- modified filters of attention interpretation

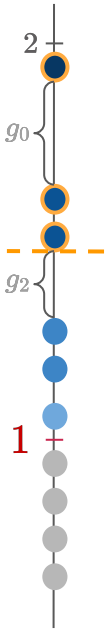
The network


pointwise convolution
PReLU
weight norm
PReLU
weight norm
PReLU
weight norm
PReLU
weight norm
PReLU
weight norm
PReLU
weight norm
$\sigma\ a_1\ \times$
$\sigma\ a_2\ \times$
$\sigma\ a_3\ \times$

Interpretability

Attention scores



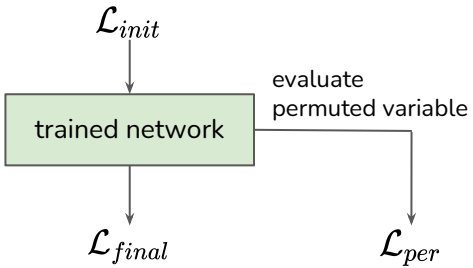


Causal Validation

- temporal precedence

  given by the architecture

- physical influence

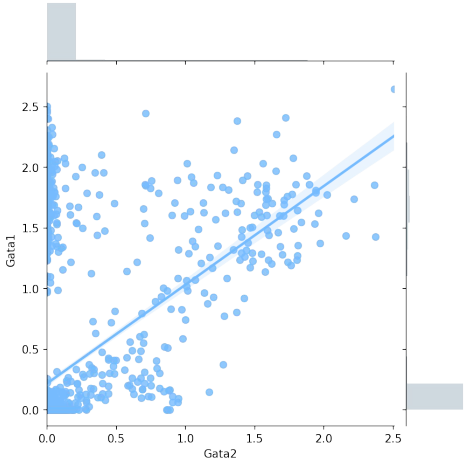


true cause if:

$$(\mathcal{L}_{final} - \mathcal{L}_{per}) \leq (\mathcal{L}_{final} - \mathcal{L}_{init}) \times s$$

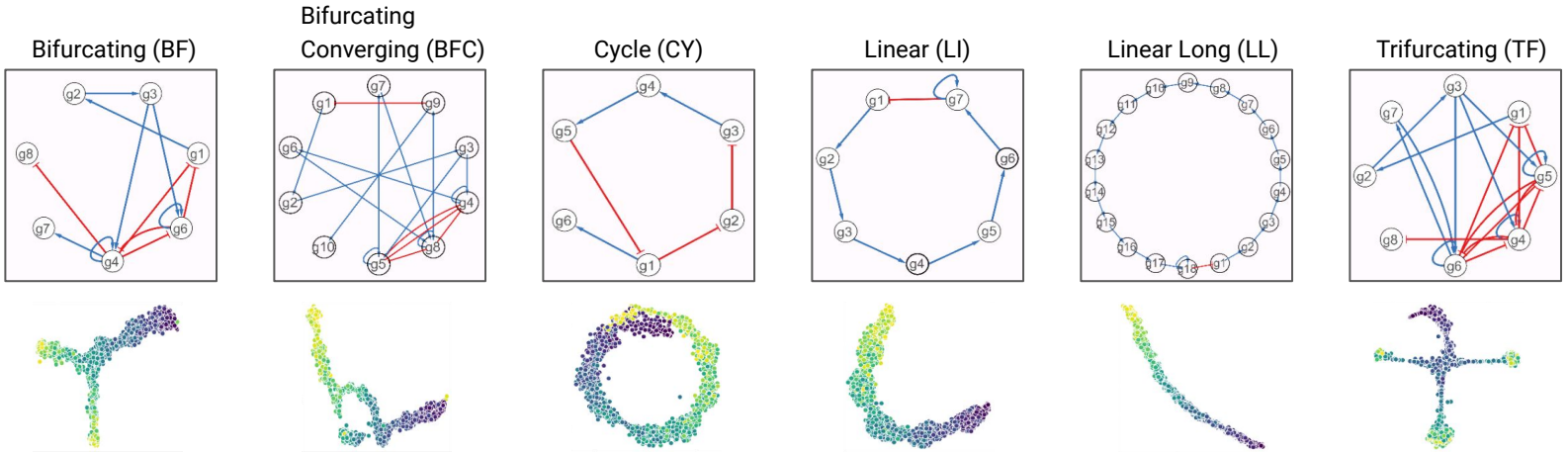
Signs

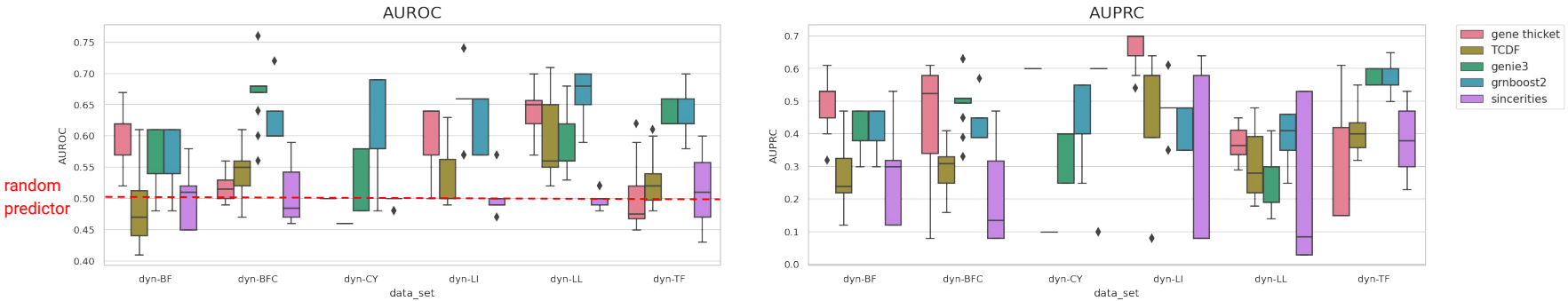
Pearson Correlation



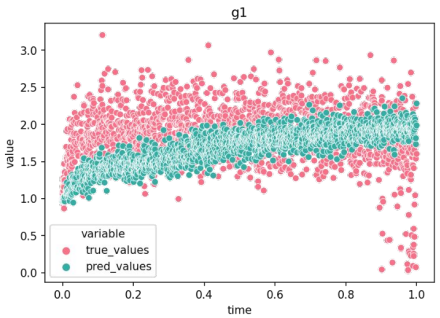
APPLICATION

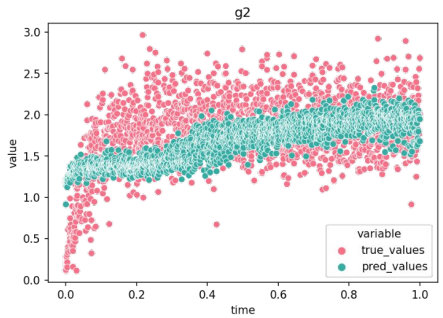
Synthetic Datasets

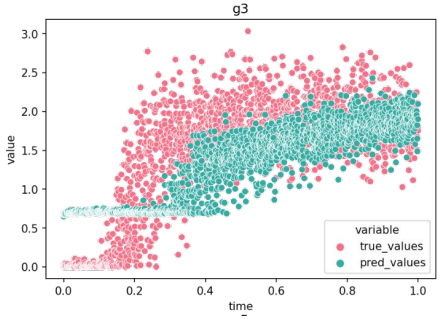


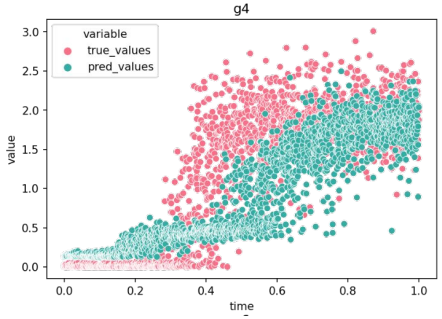


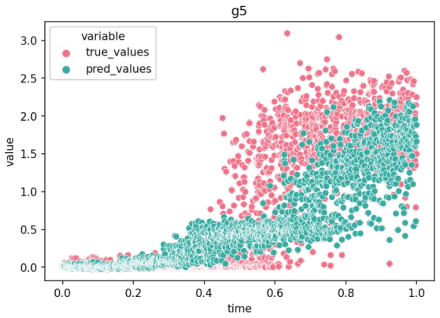
Gene Expression Predictions

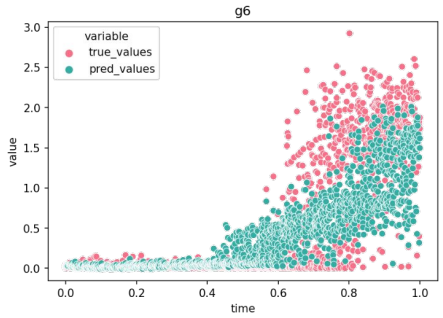


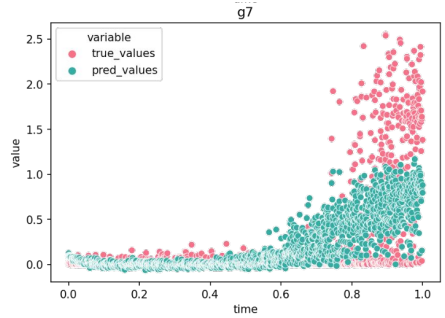


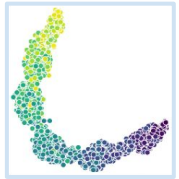


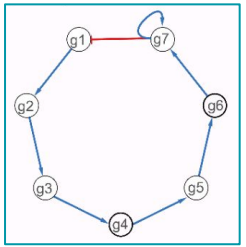






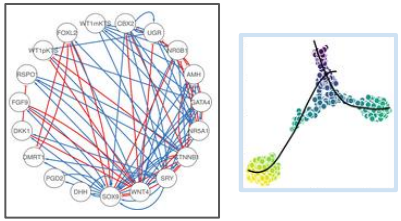




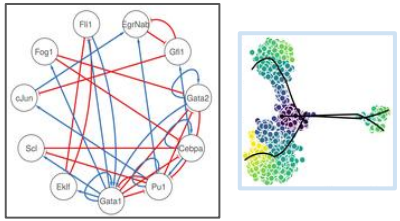


Curated Datasets

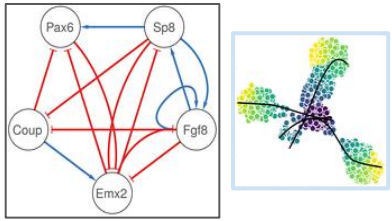
Gonadal Sex Determination (GSD)



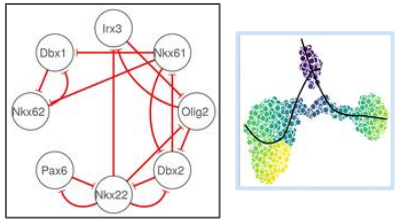
Hematopoietic Stem Cell Differentiation (HSC)



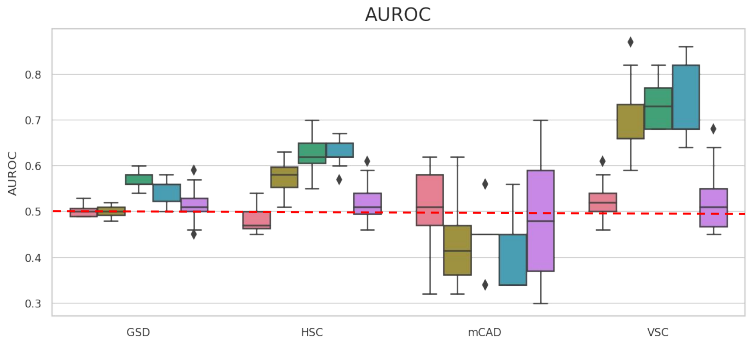
Mammalian Cortical Area Development (mCAD)

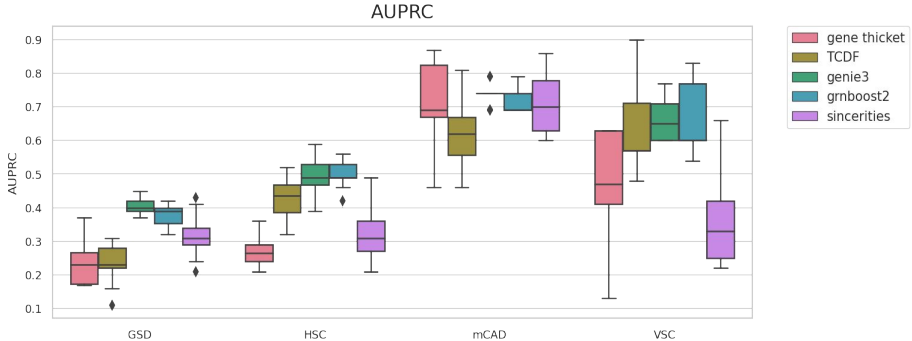


Ventral Spinal Cord Development (VSC)

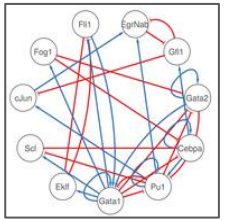


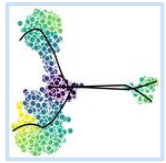
random predictor

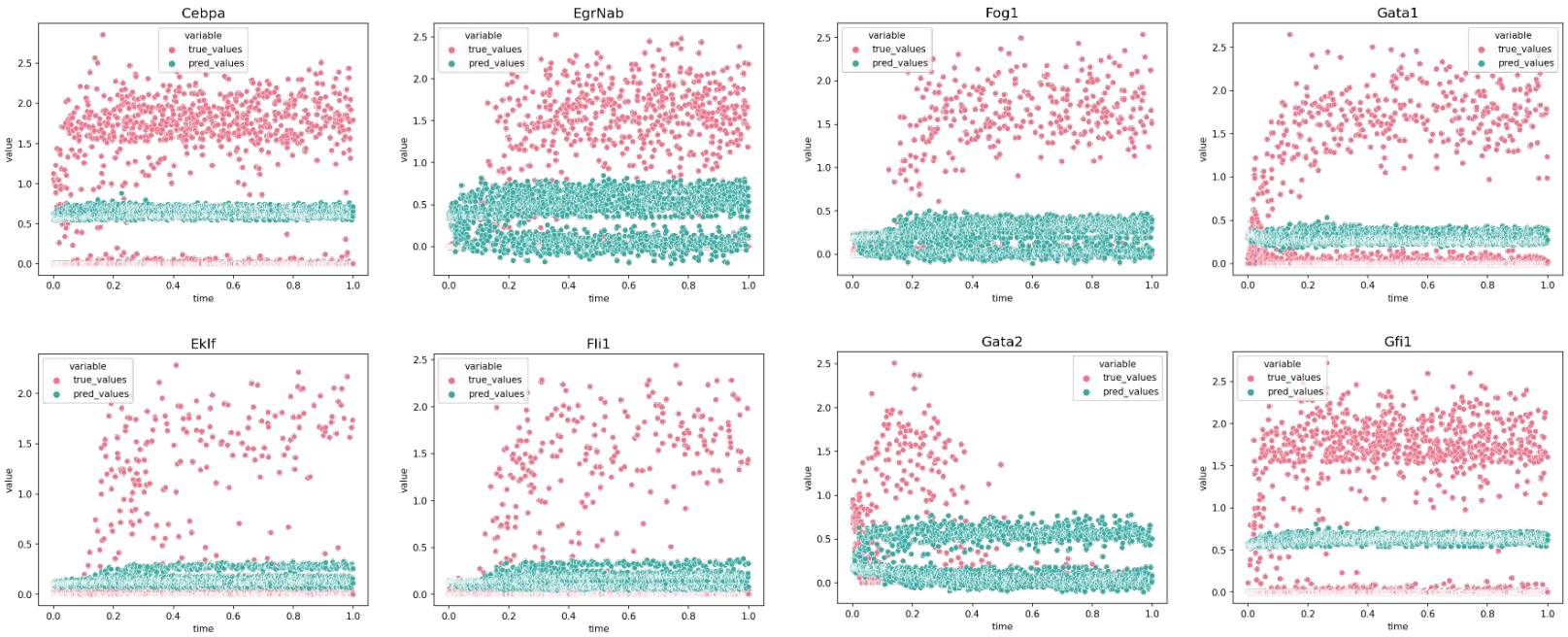




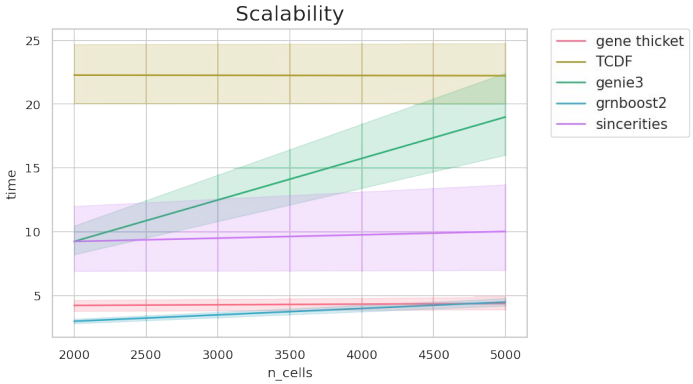
Gene Expression Predictions

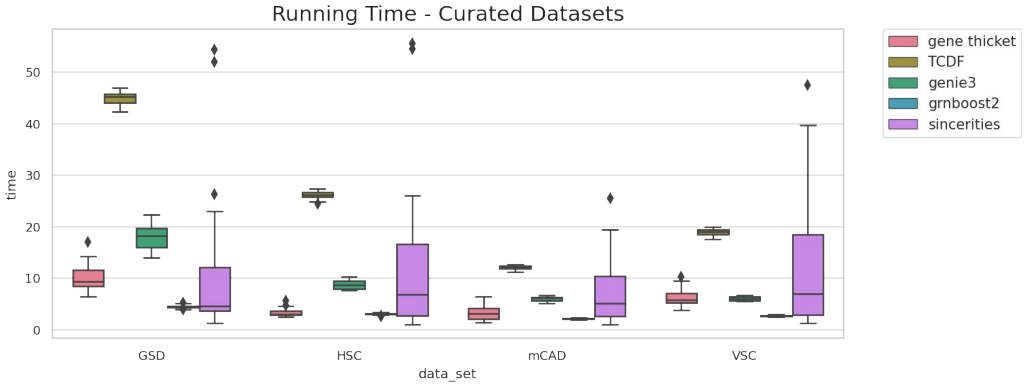


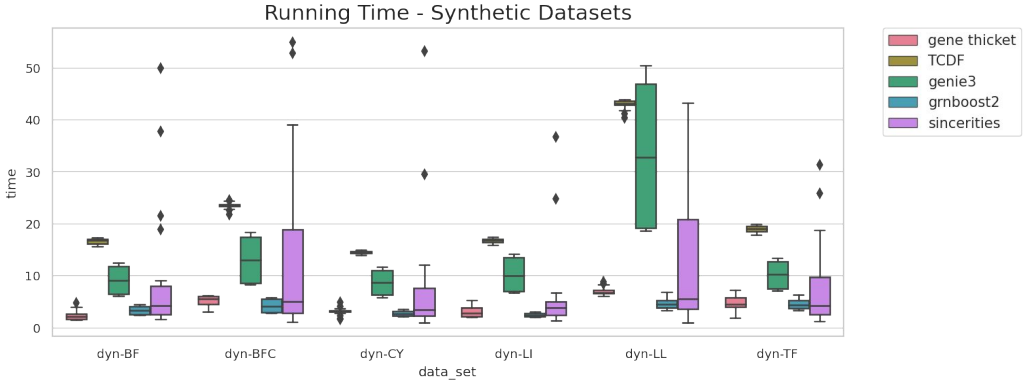




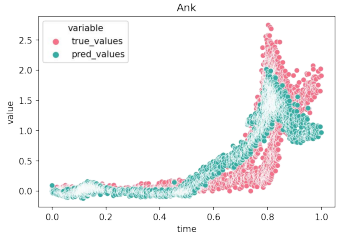
Scalability

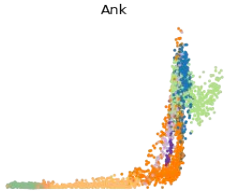


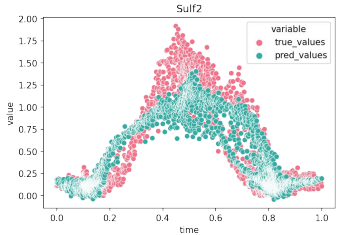


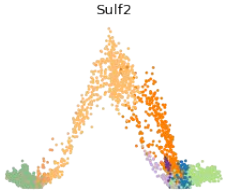


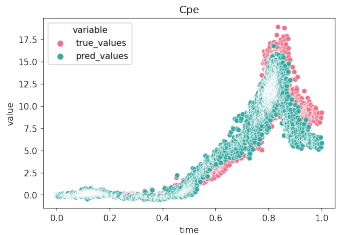
GRN - Pancreas Data

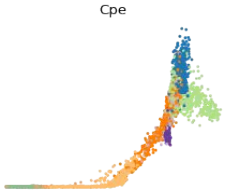


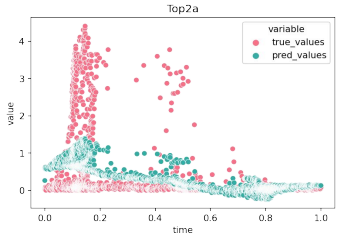


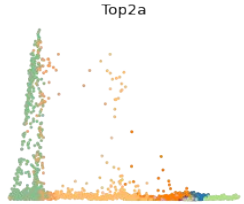


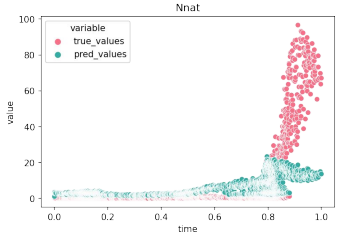


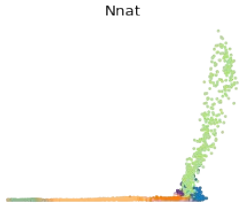


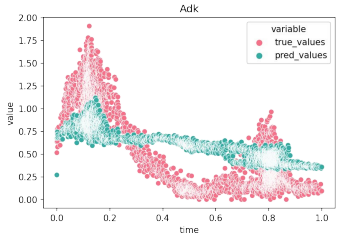


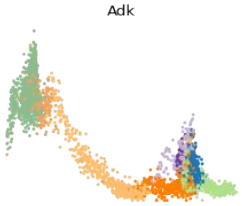




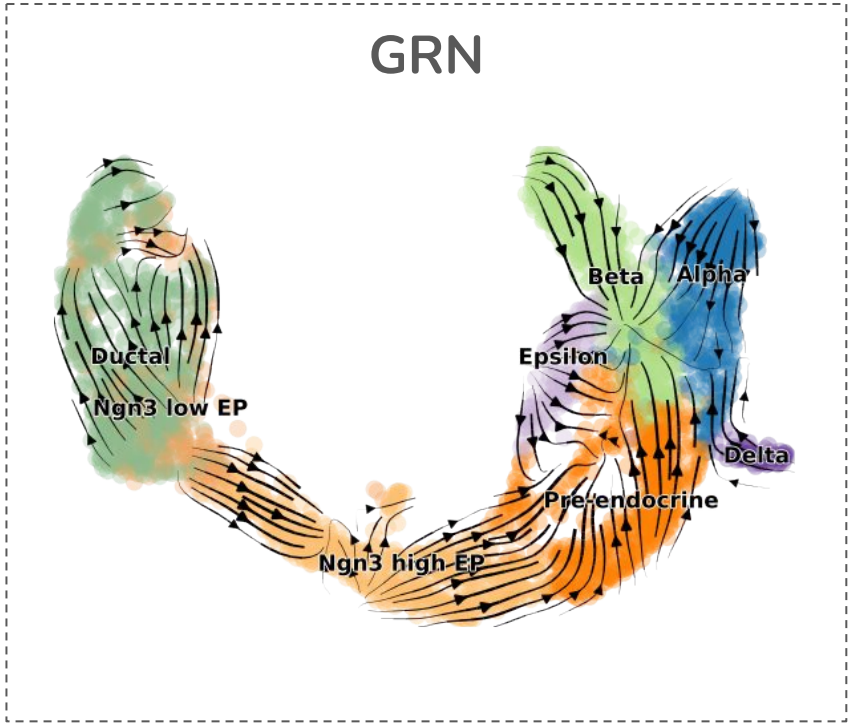


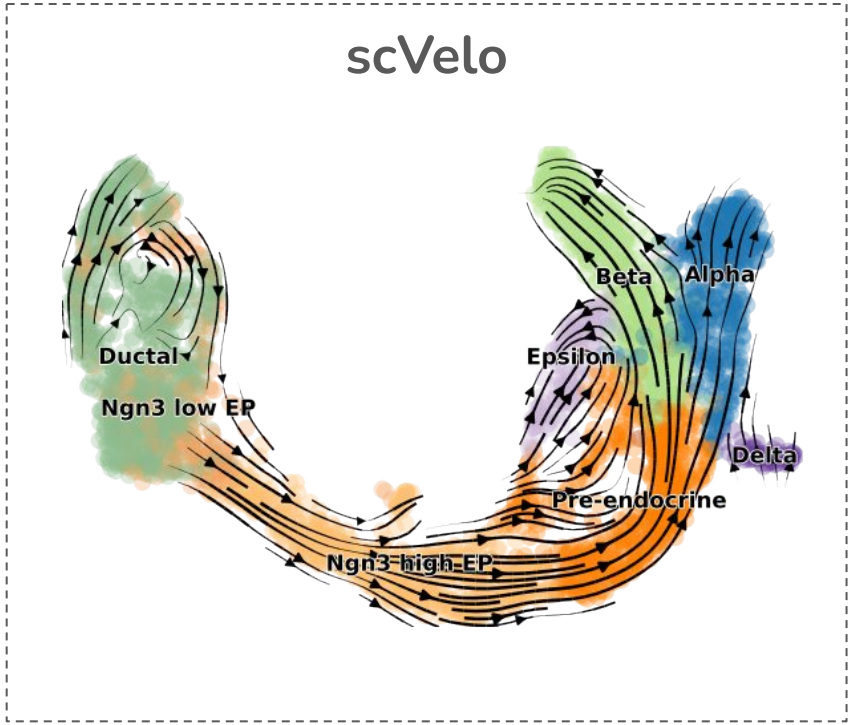




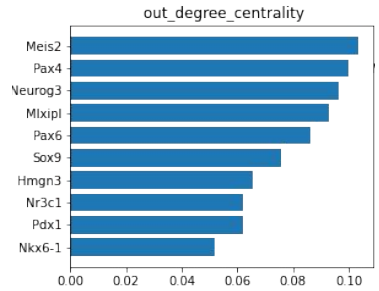


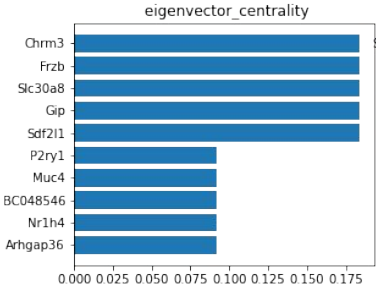
GRN velocities

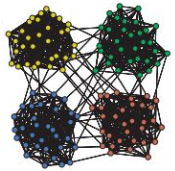




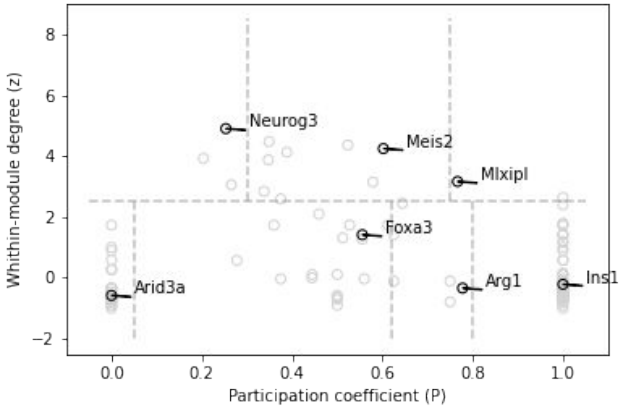
GRN - Pancreas Data

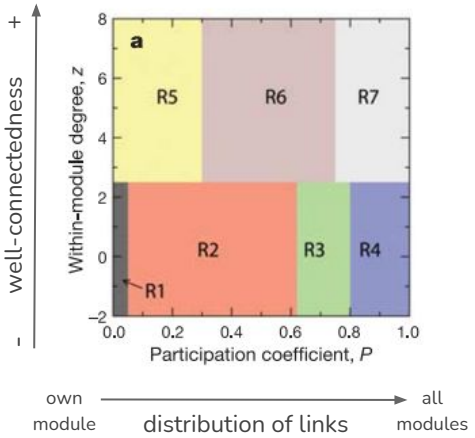






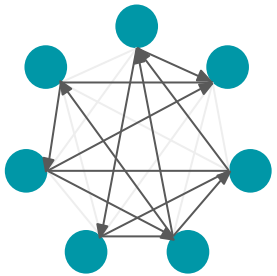
Gene Cartography Analysis





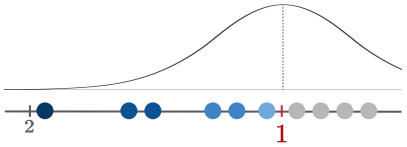
Outlook

include ATAC-seq data as a soft prior



Example with ATAC-seq data

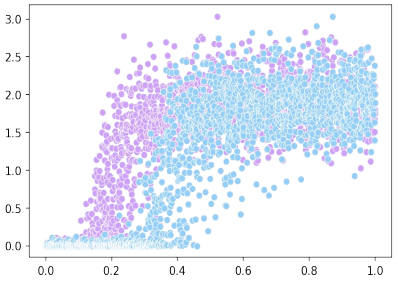
use a statistical method for: attention score selection and causal validation



$$(\mathcal{L}_{final} - \mathcal{L}_{per}) \leq (\mathcal{L}_{final} - \mathcal{L}_{init}) \times s$$

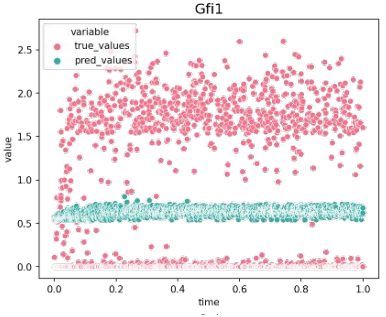
Uncertainty estimator

compute link sign based on temporal trends



activation (+)

adapt network to handle different trajectories

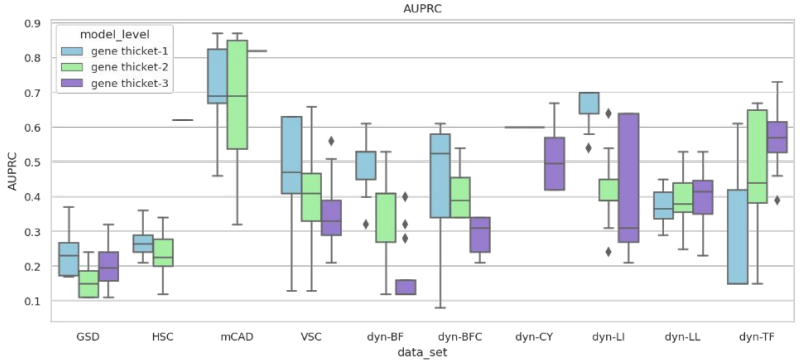


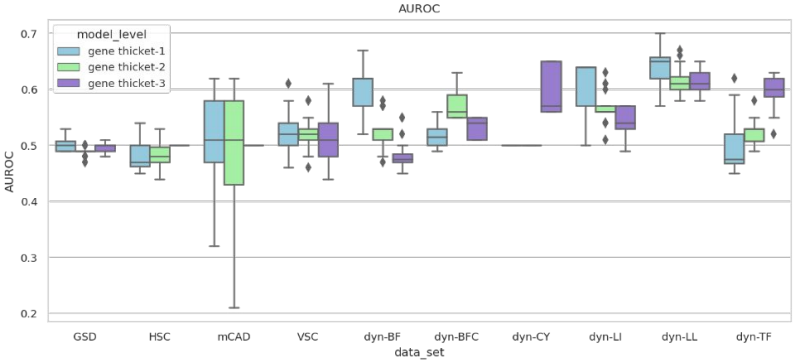
Conclusion

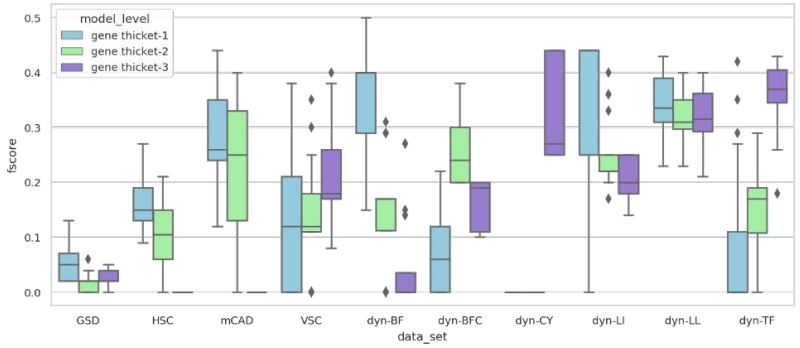
- This is a first step to iterative GRN-velocity approach.
- We can compute cell displacements using a GRN inference method.
- CNN's are flexible when computing GRNs.
  - able to look at many points in the past
  - approximate non-linear trends
  - scalable
  - can be interpretable (even discover delays)
- We need to consider:
  - multiple trajectories
  - abrupt changes in gene expression trends
  - uncertainty

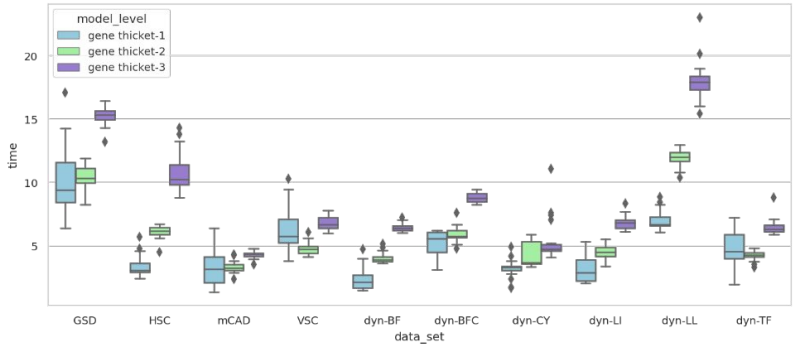
THANK YOU!

Choosing Number of Blocks

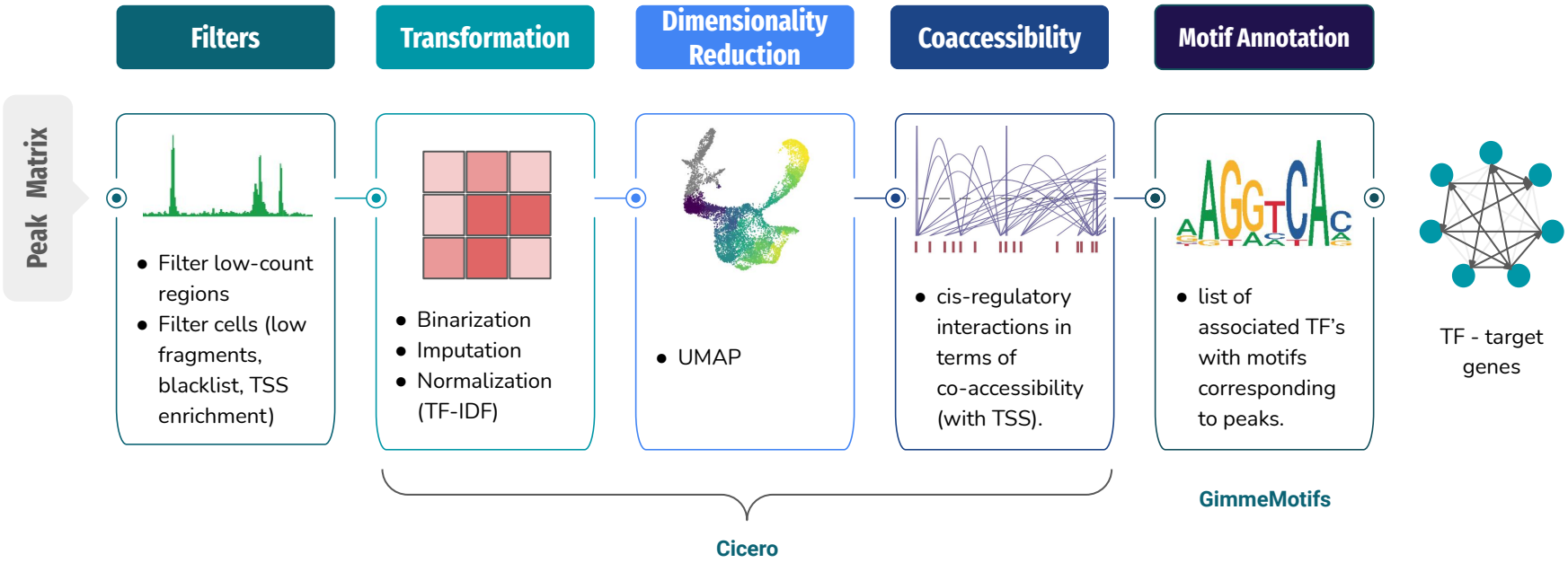








ATAC-seq preprocessing



As in CellOracle

Gene Cartography Analysis

1. Identify functional modules.

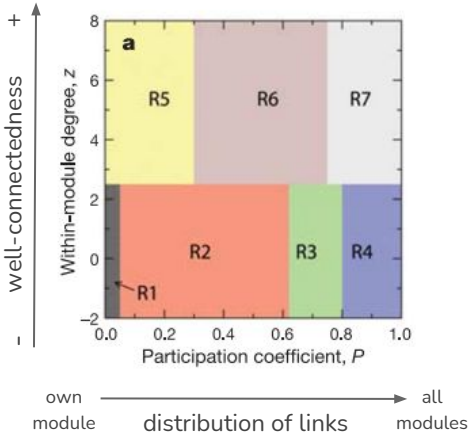
2. Compute **within-module degree (zi)** and **participation coefficient (pi)**.

- **zi** measures how “well connected” node i is to other nodes in the module. (higher = better connected)
- **pi** measures how “well-distributed” the links of a node are among different modules. (1 if links are uniformly distributed among all the modules, 0 if all its links are within its own module)

3. Identify different roles of nodes (TFs).

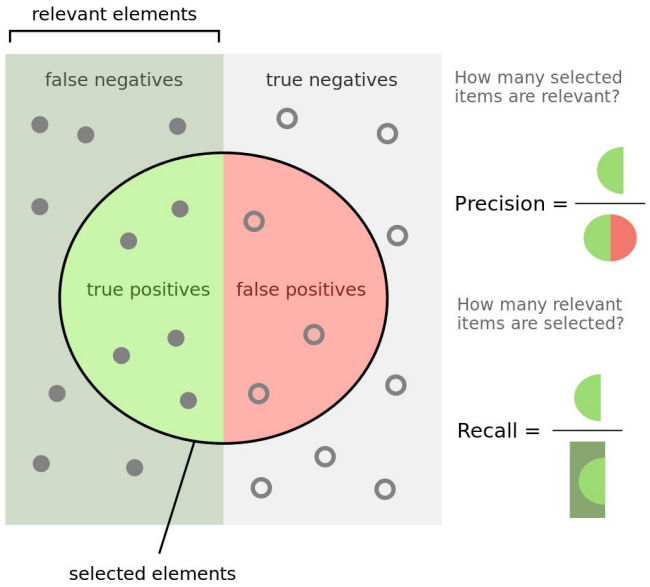
NON-HUBS (z < 2.5):

- **ultra-peripheral (R1):** all their links within their module.
- **peripheral (R2):** most links within their module.
- **non-hub connector (R3):** many links to other modules.
- **non-hub kinless (R4):** links homogeneously distributed among all modules.



HUBS (z ≥ 2.5):

- **provincial (R5):** vast majority of links within their module.
- **connector hub (R6):** many links to most of the other modules.
- **kinless hub (R7):** links homogeneously distributed among all modules.

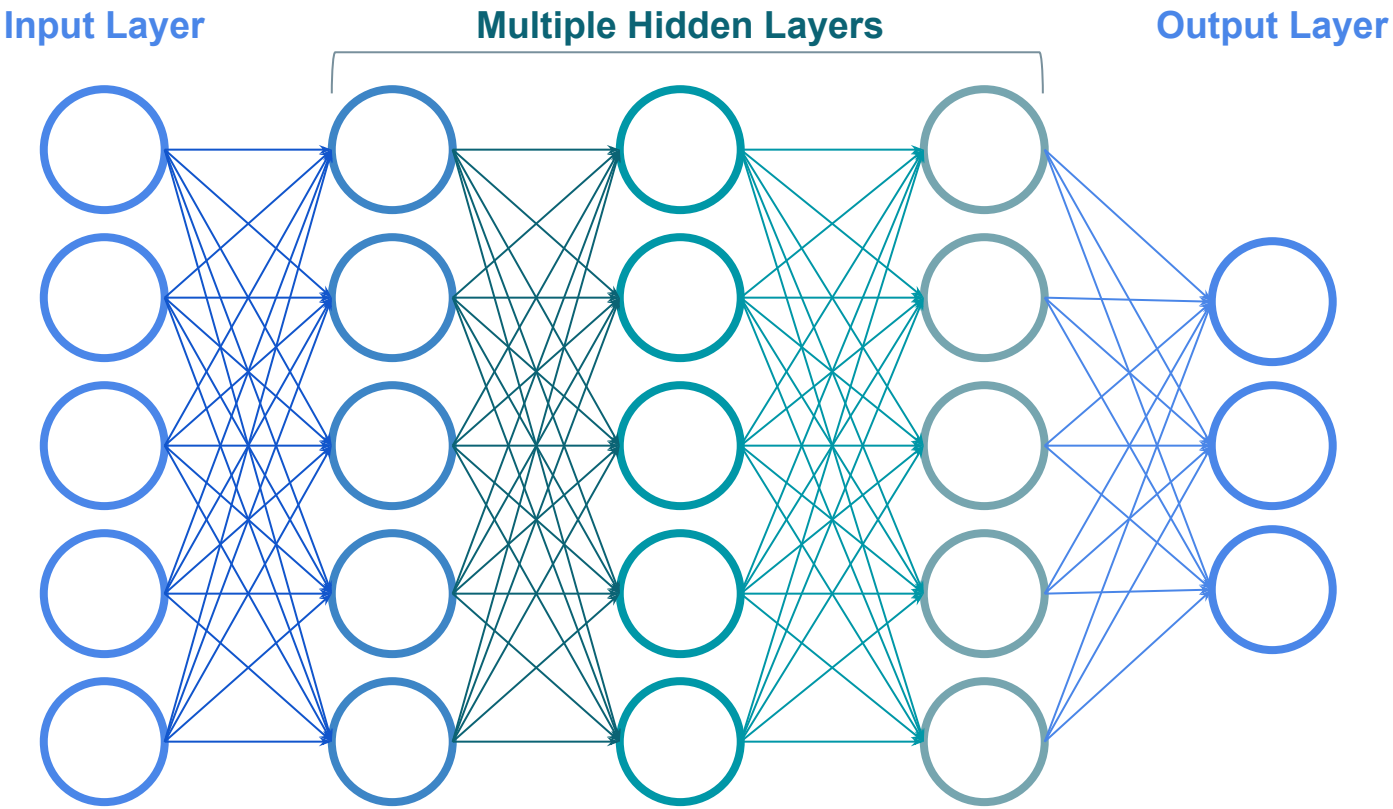


How many selected items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

How many relevant items are selected?

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$


Input Layer
Multiple Hidden Layers
Output Layer