Causality for Transcriptomics?

P. Bertin, S. Francis, J. Viviano, Y. Bengio, J. P. Cohen

Mila, Université de Montréal

February 12, 2020

Outline

Not published yet

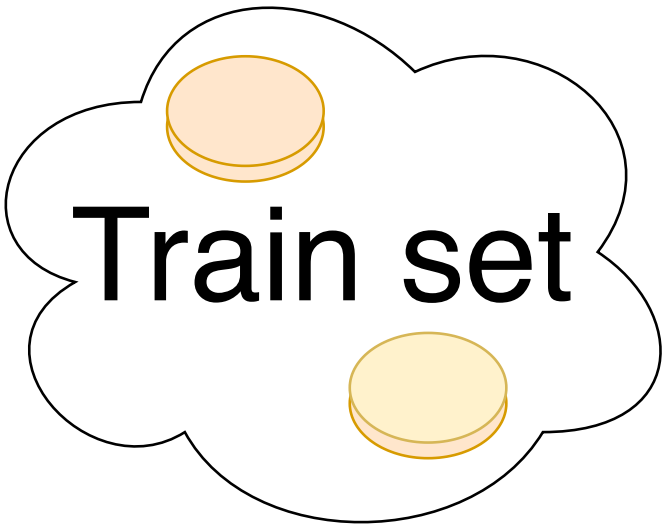
- We will present the main directions of our ongoing project
- No results yet
- But we would be happy to get your expert insight!

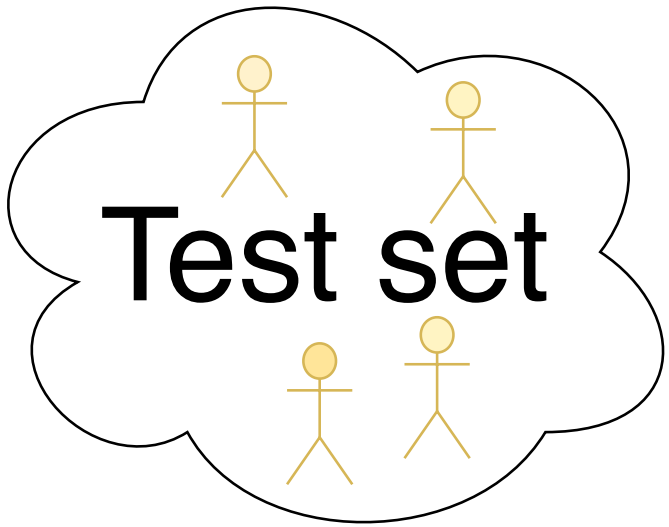
ML for medicine

Some challenges

Let us present two challenges of ML for medicine

Robustness



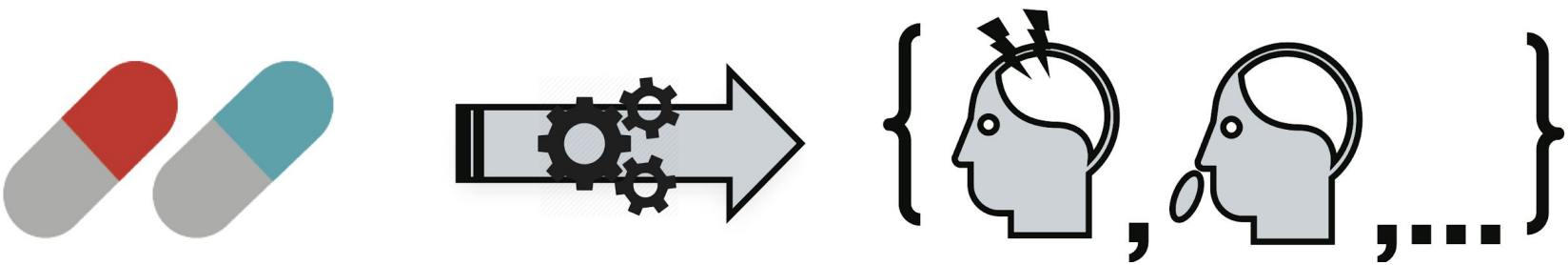


Lab experiments

Patients

- At test time the setting is most of the time different
- We need **robustness**

Drug effect prediction



- We would like to **predict the effect** of a given drug/compound (at the cell level, living organism level...)

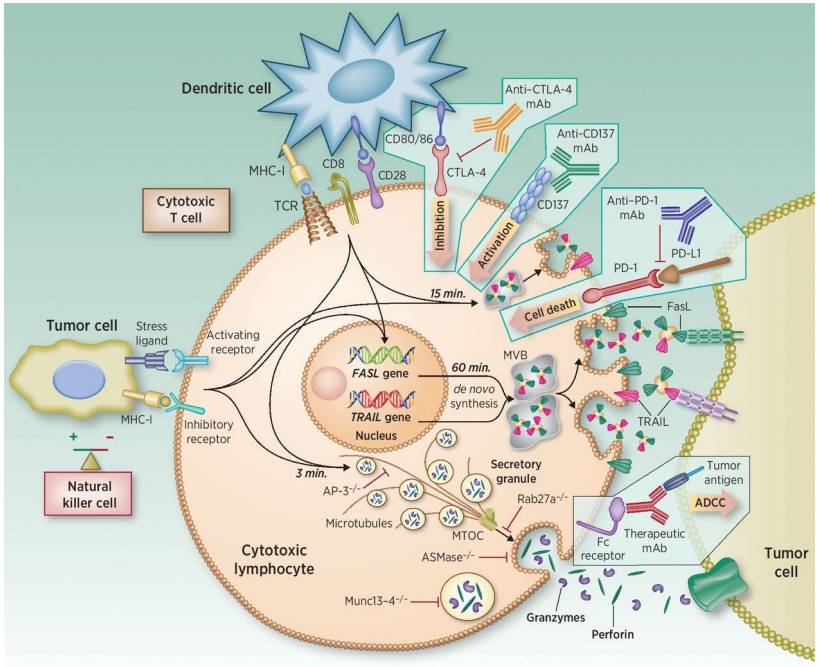
Proposed Approach

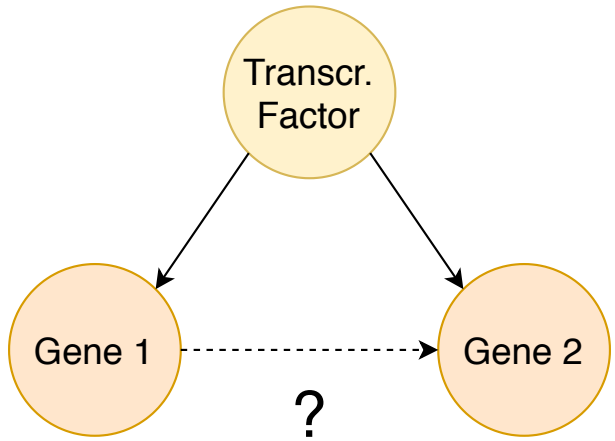
Proposed Approach

Better **model the mechanisms of the cell** from gene expression data in order to:

- Generalize well in different contexts (**robustness**)
- Predict the effect of a new compound (**perturbation effect prediction**)
- Predict the effect of new combinations/concentrations of compounds

How to understand the mechanisms of the cell?





- We would like to identify the effect of a given gene on another gene
- **Correlation is not Causation**
- Lots of confounders

Figure taken from *https://clincancerres.aacrjournals.org/content/21/22/5047*

Module Networks

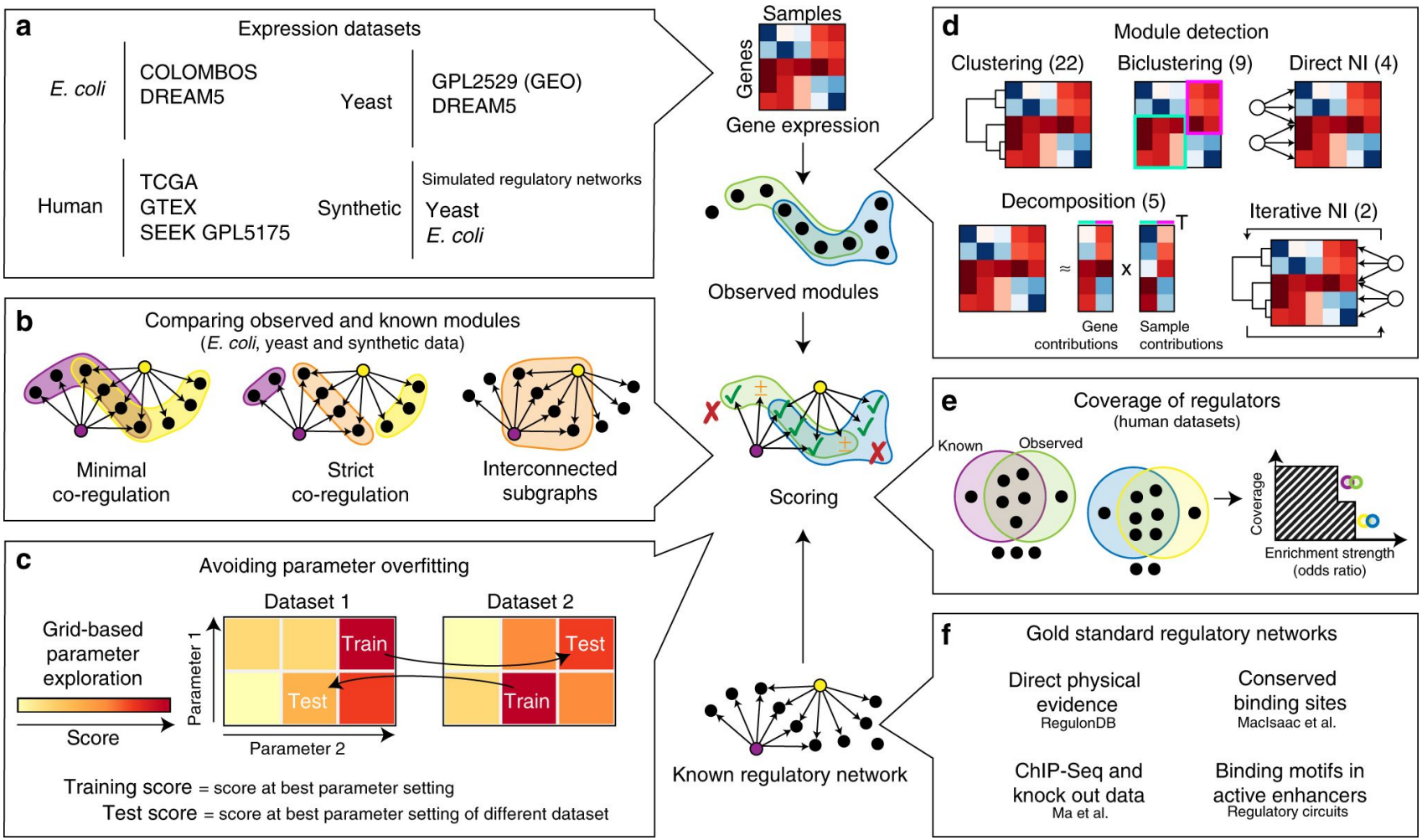


Figure taken from *A comprehensive evaluation of module detection methods for gene expression data*

What about curated graphs?

What about curated graphs?

Curated graphs (StringDB...) do not seem to be well suited to provide prior knowledge for gene expression data

---

Analysis of Gene Interaction Graphs as Prior Knowledge for Machine Learning Models

---

Paul Bertin
Mila, Université de Montréal
Montréal, Canada

Mohammad Hashir
Mila, Université de Montréal
Montréal, Canada

Martin Weiss
Mila, Université de Montréal
Montréal, Canada

Vincent Frappier
Mila, Université de Montréal
Montréal, Canada

Theodore J. Perkins
Ottawa Hospital Research Institute
University of Ottawa
Ottawa, Canada

Geneviève Boucher
Institute for Research in Immunology and Cancer
Université de Montréal
Montréal, Canada

Joseph Paul Cohen
Mila, Université de Montréal
Montréal, Canada

Can we do better?

Can we do better?

We would like to use machine learning to **learn the mechanisms of the cell** directly from gene expression data

Outline

Causal versus Statistical Inference

Let us explore in more details the difference between Causal and Statistical Inference

Statistical Inference: Observation

- You try to describe a system by **looking** at it
- You have access to observational data
- You model $P(X)$ from the observations

Causal Inference: Action

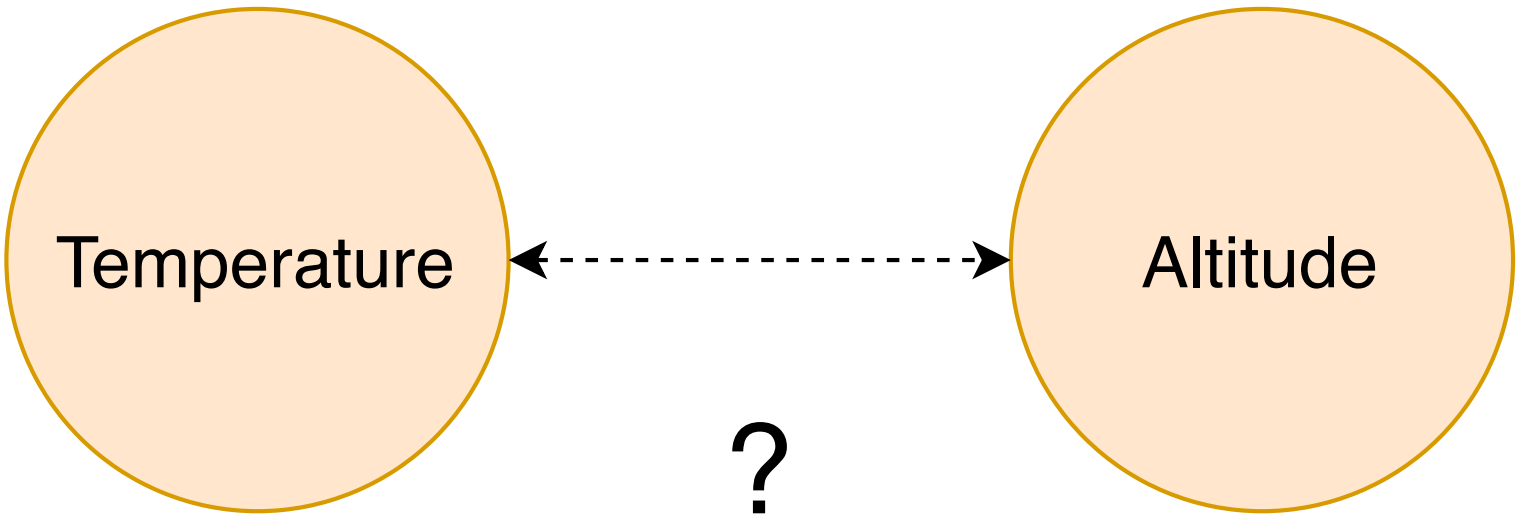
- You try to describe a system by **looking** and **acting** on it
- You have access to observational data **and** data under **intervention**
- You model $P(X)$ **and** the way $P(X)$ would change under intervention

An example

An example

Given a list of cities, can we understand the relation between **altitude** and **temperature**?

Temperature and Altitude



Statistics on cities

- We are given a list of cities with altitude A and average temperature T
- We can model the **joint distribution** $P(A, T)$

Problem

This **does not give us any information on the actual mechanisms** of the system

- If our observations are *biased*, we could draw wrong conclusions.
- *e.g. low cities in Sweden vs high cities in the tropics*

Acting?

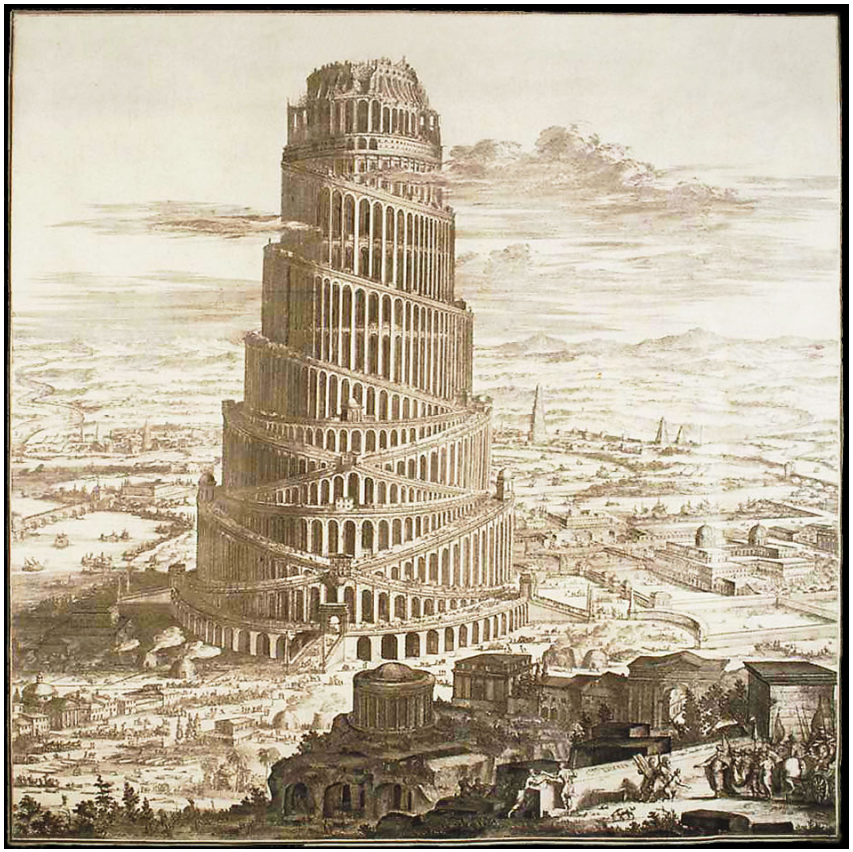
Acting?

We **have to act** on our system (cities) in order to **understand what actually happens**

Temperature and Altitude

- Setting the city on fire does not change its altitude!

Temperature and Altitude



- Moving the city to the top of a giant tower changes the temperature!

Outline

L1000 dataset

L1000 dataset

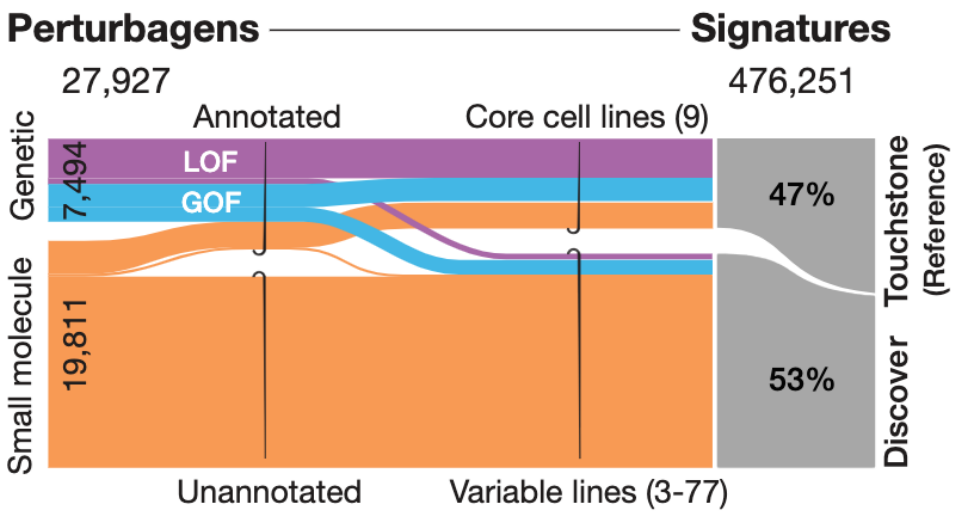
Gene expression data with

- Different cell lines
- Different growth conditions (perturbations by compounds)

L1000 dataset

The CMap dataset of cellular signatures catalogs transcriptional responses of human cells to chemical and genetic perturbation. Here you can find the 1.3M L1000 profiles and the tools for their analysis.

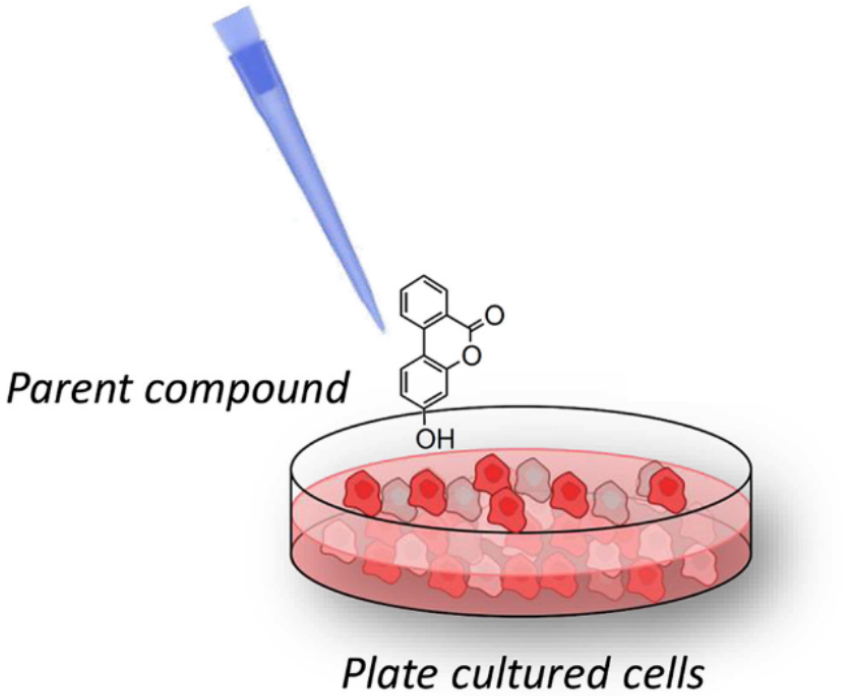
A total of 27,927 perturbagens have been profiled to produce 476,251 expression signatures. About half of those signatures make up the Touchstone (reference) dataset generated from testing well-annotated genetic and small-molecular perturbagens in a core panel of cell lines. The remainder make up the Discover dataset, generated from profiling uncharacterized small molecules in a variable number of cell lines.



ConnectivityMap

BROAD INSTITUTE

Link with Causality?



We can act on the cell

We have a system on which we can act, why not use Causal Inference?

Terminology

Terminology

We call **environment** a given cell line that has been perturbed by a given compound

- Other choices are possible!
- Question: Take into account the concentration/incubation time?

Outline

Main idea

Main idea

We hypothesize that the mechanisms that regulate gene expression data change in a **sparse** way from one environment to the other

Note

- Compared to conventions used for *Regulatory Networks*, we will represent links as arrows, even if the link is inhibitory

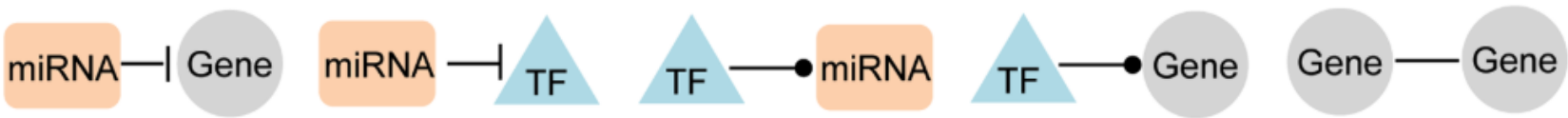
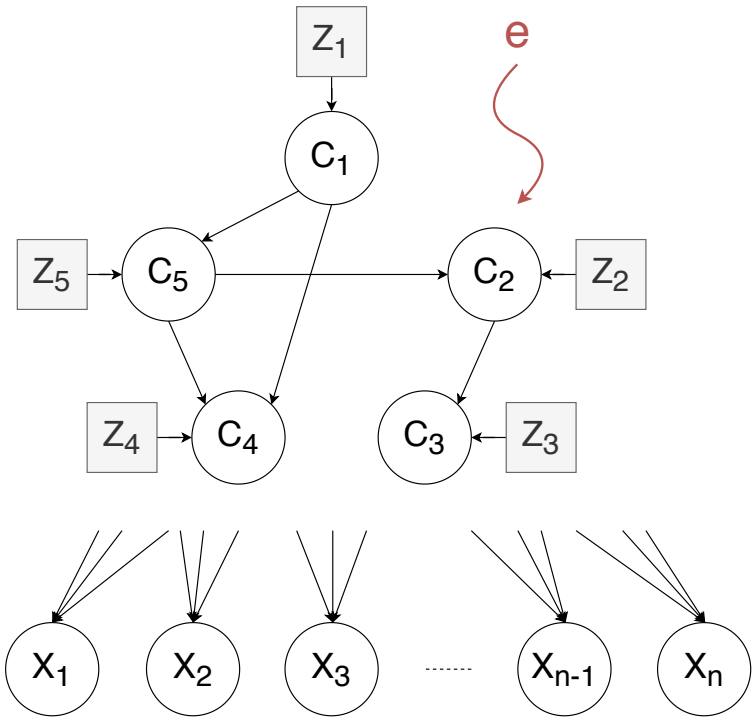


Figure taken from *Uncovering MicroRNA and Transcription Factor Mediated Regulatory Networks in Glioblastoma*

Our assumptions



- There exist a set of hidden variables C whose interactions are governed by fixed mechanisms
- The gene expressions X we observe are functions of the values of the hidden variables $X = f(C)$. f is invariant across environments.
- A given compound acts on only **one** of the hidden variables

Outline

Challenges

Challenges

There are several challenges that prevent us from using common causal discovery tools

- **High dimensionality**: $\mathcal{O}(2^{20k^2})$ possible gene interaction graphs
- **Hidden confounders**: not every variable of interest is observed (*e.g.* variations in the non coding genome)
- **Compounds can act on lots of things** at the same time (move towards CRISPR experiments with the **Perturbseq** dataset?)

Combining different datasets

Combining different datasets

- Use TCGA along with L1000?
- Problem: different preprocessing (environments might not be sparse anymore)

- We also want to experiment with single cell data under CRISPR perturbations (Perturbseq dataset)

Getting a fixed size embedding for compounds

Getting a fixed size embedding for compounds

- We need a fixed size representation of each environment
- Use SMILE representation of compounds?

- Which fingerprint to use?

How to evaluate the model?

Evaluation

- Measure performance in unseen environments (cell lines/compounds) : **gene expression reconstruction error**
- Use the *causal* embedding in **downstream tasks** (*e.g.* histological type) and see if it improves performance

- Other ideas?

Make sense of the embedding?

Interesting question

Would we be able to **relate** the variables of the *causal* embedding **to actual biological entities**?

Thanks!

Stay tuned, we hope to have results soon!