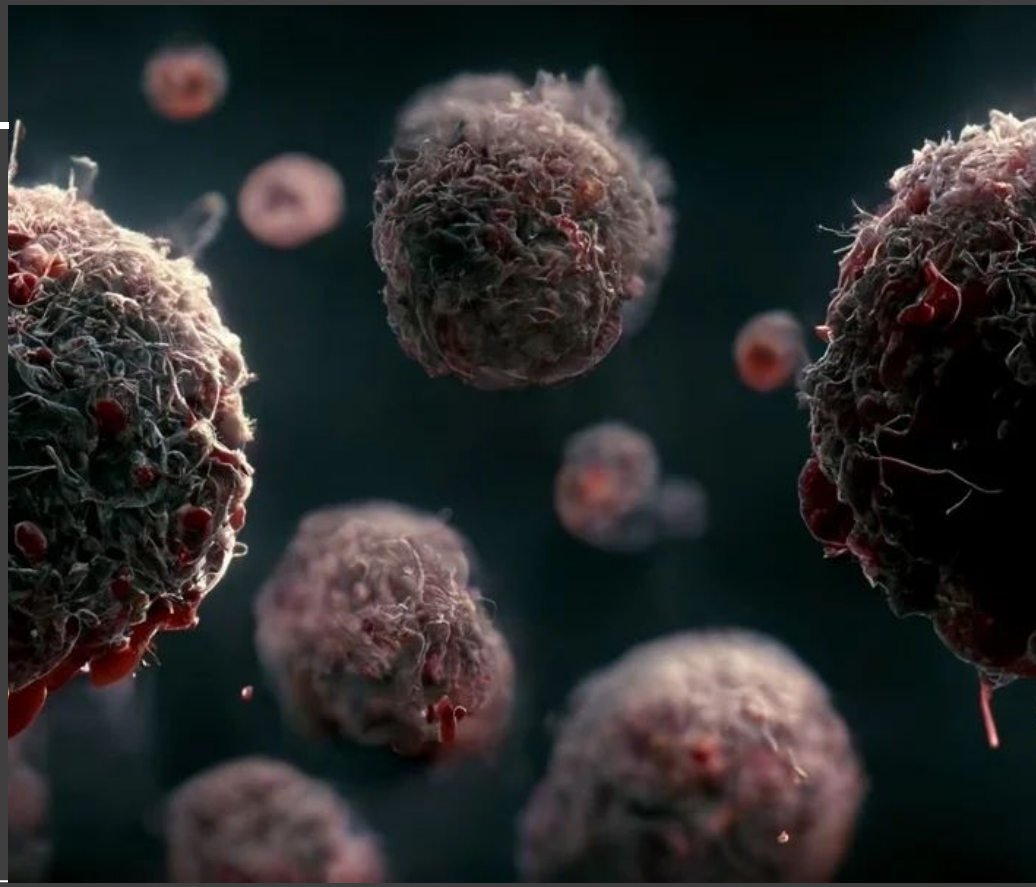
The Alan Turing Institute

Causal Discovery For Cancer

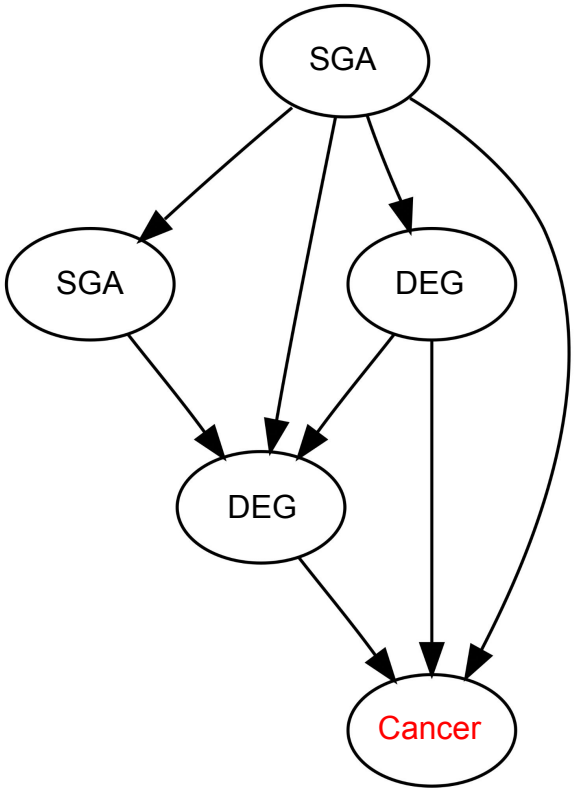
Joseph Boyle, Jonathan Erskine, Ellen Visscher, Pingfan Song, Behrad Koohy, Christian Cabrera, Jindong Gu



Problem statement

Somatic genomic alterations (SGAs) lead to differential expression of genes (DEGs) which are known to cause cancer by altering the genes that control cell growth and multiplication.

SGAs that cause DEGs are potential drivers of cancer.



Goal

Find a causal link between SGAs and DEGs in breast cancer patients.

Explanation of Data

Curated dataset – Mini-TCGA Cancer Dataset:

- contains data of Somatic Genomic Alteration (SGA) and Differentially Expressed Gene (DEG).
- **831 patients**, with samples from cancerous and non cancerous breast tissue
- **6 SGAs: genomic locations** known to be associated with breast cancer
  - 1 if mutation at location (in cancerous sample), 0 otherwise
- **60 DEGs: genes** that can be overly-expressed.
  - 1 if gene is differentially expressed (> 2 std) between cancerous and non-cancerous sample

Problem setup:

- Based on clinical assumptions, our causal graph is bipartite.
  - SGAs → DEGs
  - DEGs cannot cause SGAs
- 6 SGAs that are well-known drivers in breast cancer as inputs
- 60 DEGs which are regulated by these 6 SGAs.

Aim: predict the presence of a DEG given the SGA's in the patient.

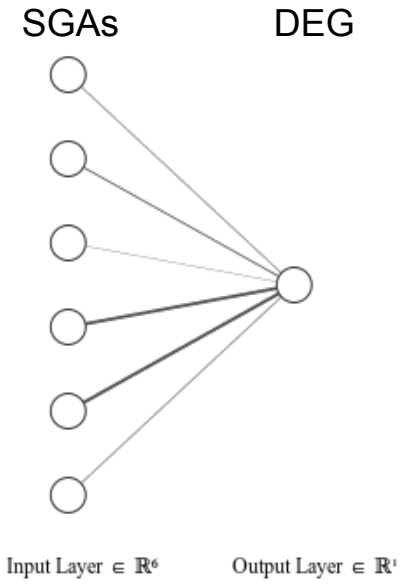


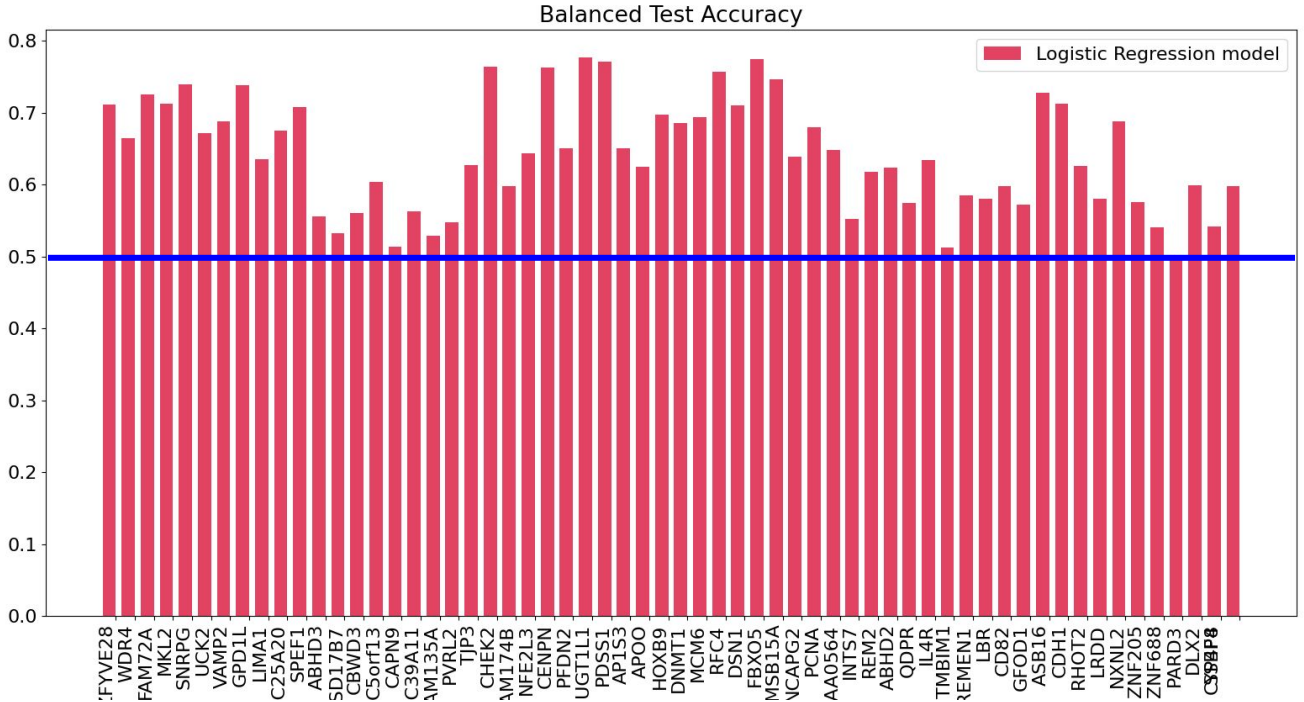
Figure: an example graph of causal links between SGAs and DEGs

Our Approach

Logistic Regression model performance - Test set Accuracy

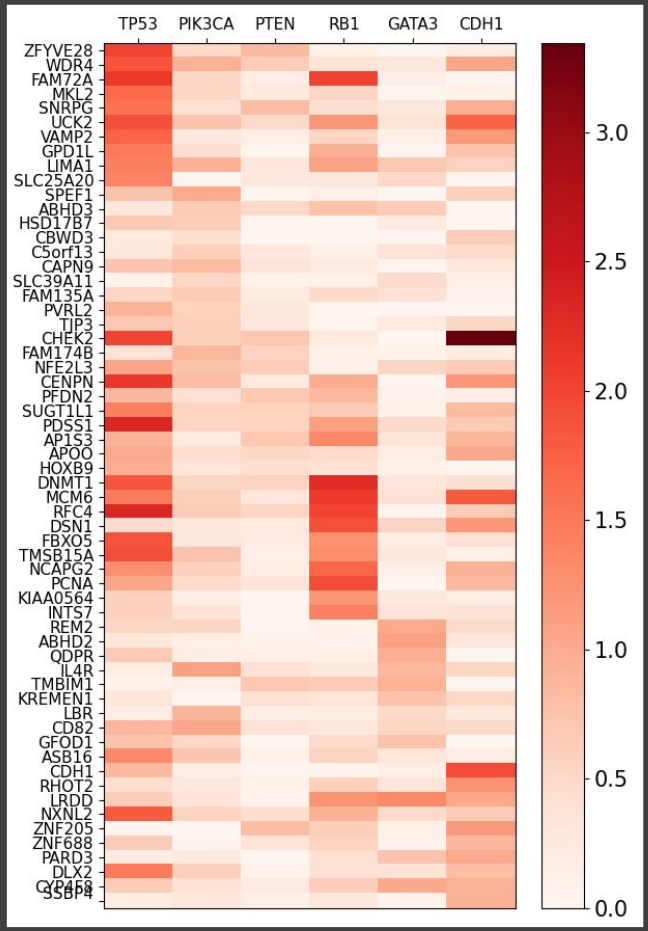
For each DEG we train a Logistic Regression classifier and evaluate it on our artificially balanced test set.

We show a random baseline in blue.

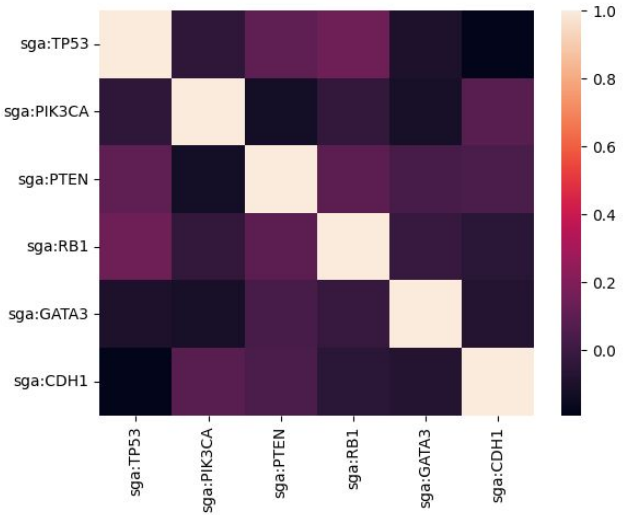


Inferring causality

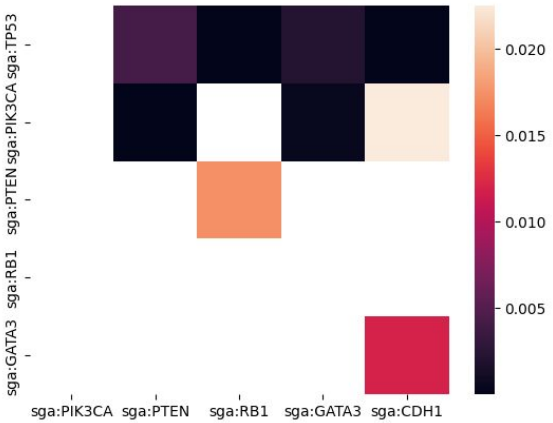
We can analyse the absolute feature weights to understand how input features contribute to our model predictions.



Testing Assumptions: Are SGAs independent?



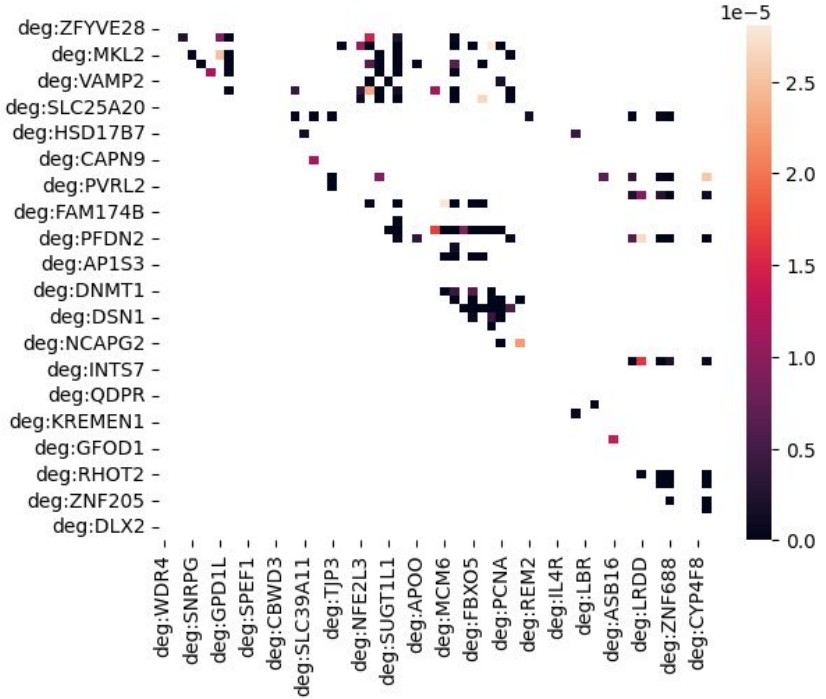
MI shows that MI is low across SGAs



G-test (similar to chi^2)- Bonferroni corrected significant values, lower = more sig (likely to be dependent)

Conclusion:
Not independent but relationship may be weak leading to low MI score

Testing Assumptions: Are DEGs conditionally independent given SGAs?



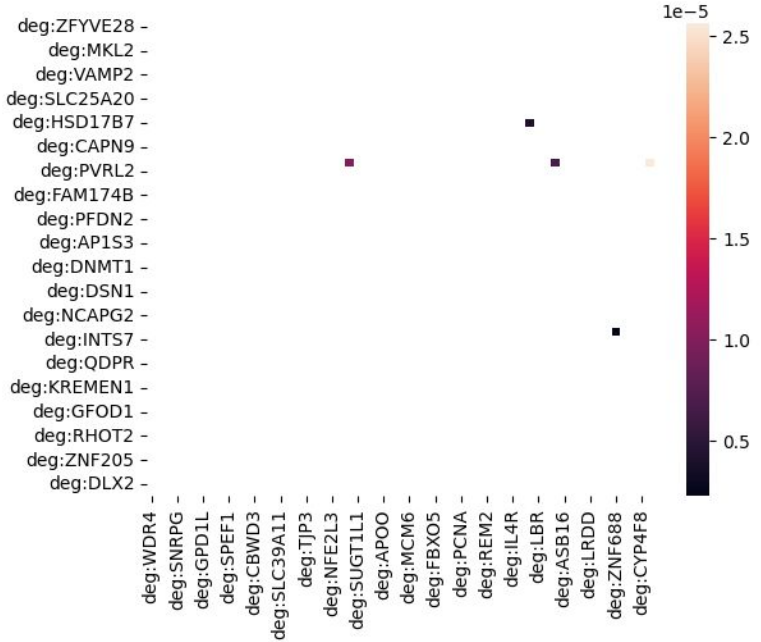
Conclusion:
Even when conditioning on SGAs some DEGs still fail independence test.

Can we determine underlying reason?

1. DEGs have causal effect on each other.
2. DEGs have causal effect on SGAs.
3. DEGs may have other confounders/causes which are not included in the current 6 SGAs.

G-test (similar to chi^2)- Bonferroni corrected significant values, lower = more sig (likely to be dependent)

Do some DEGs become dependent after conditioning on SGAs?



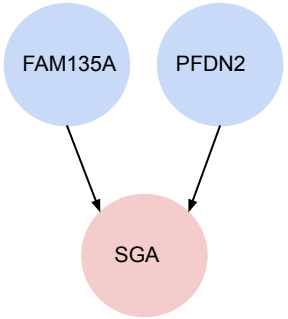
DEGs that fail independence test only after conditioning on SGAs

Conclusion:
Some DEGs becomes dependent after conditioning on SGAs.

E.g FAM135A (lipid metabolism) becomes dependent on:

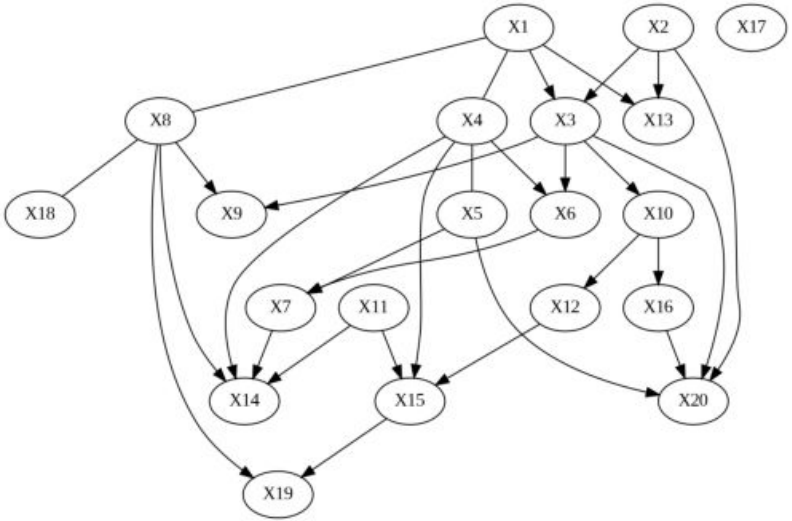
- PFDN2 (correct protein folding)
- GFOD1 (nucleotide binding)

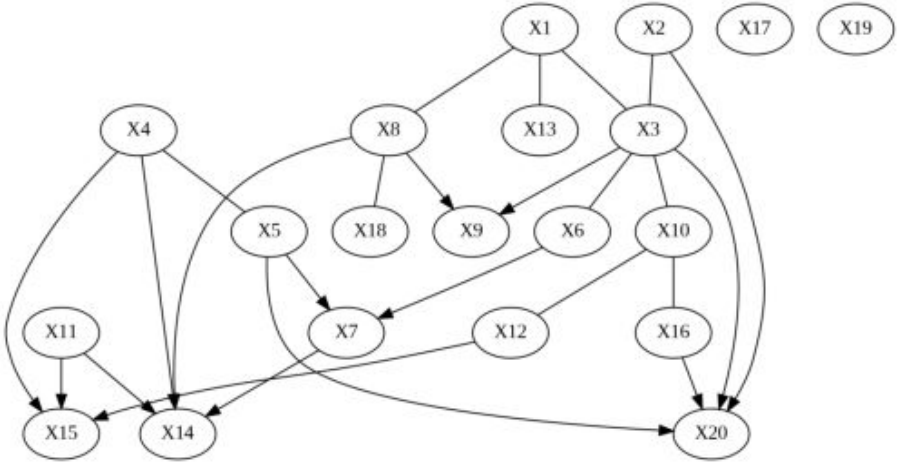
Is this an example of collider bias? That is:



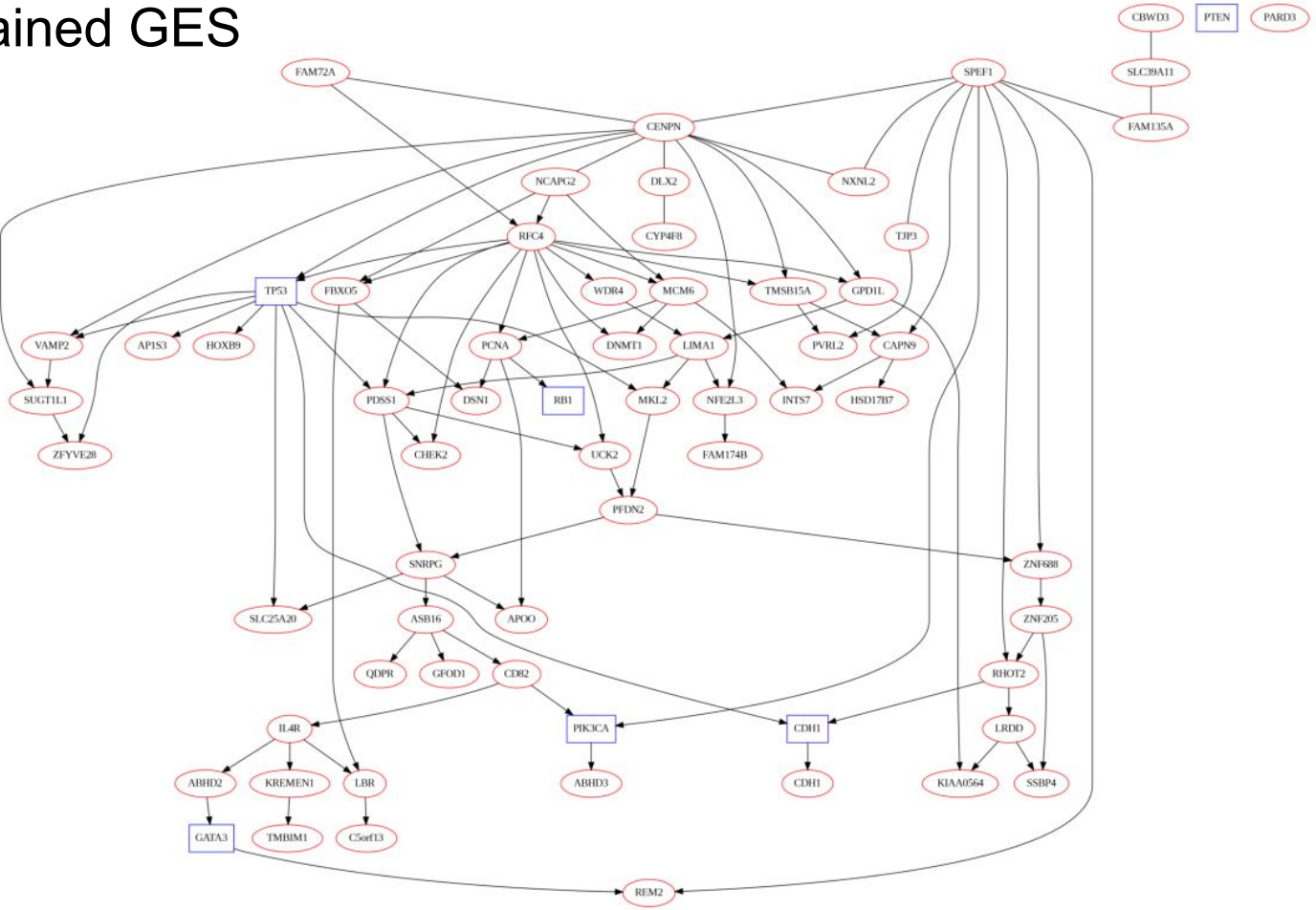
Unconstrained causal graph somewhat supports

SHD: 5 of 30

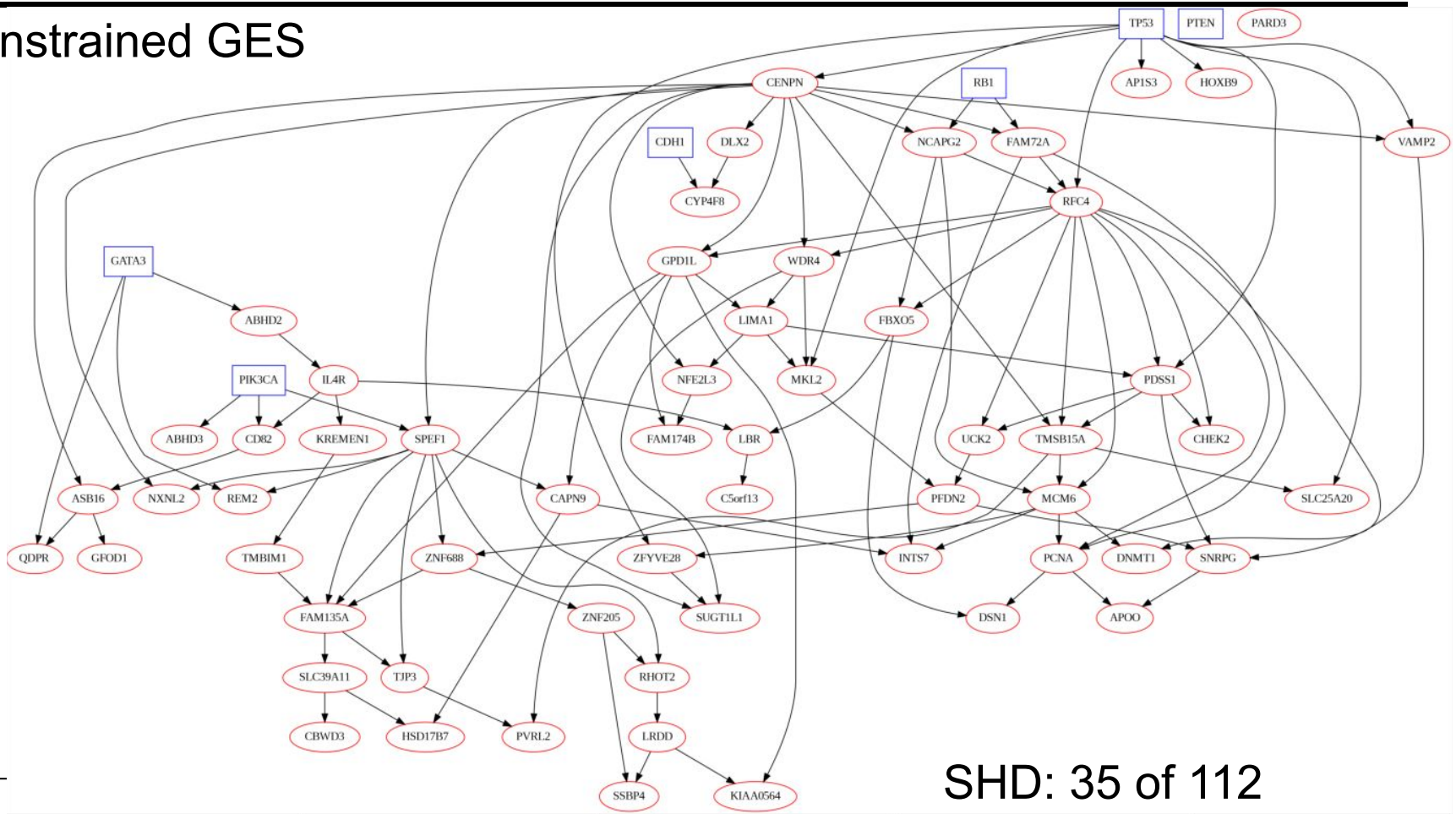

X1
X2
X17
X8
X4
X3
X13
X18
X9
X5
X6
X10
X7
X11
X12
X16
X14
X15
X20
X19


X1
X2
X17
X19
X4
X8
X13
X3
X5
X18
X9
X6
X10
X11
X7
X12
X16
X15
X14
X20

Unconstrained GES



Constrained GES



SHD: 35 of 112

More info about SGAs and DEGs

- sga:TP53: TP53 gene encodes the **tumor suppressor protein p53**; most commonly mutated genes in cancer; involved in regulating cell cycle progression, DNA repair.
- sga:PIK3CA: PIK3CA gene is important for **cell growth and survival**.
- sga:PTEN: PTEN is a **tumor suppressor** that regulates PI3K signaling.
- sga:RB1: **tumor suppressor** that inhibit the activity of the E2F transcription factors. mutation → uncontrolled cell proliferation.
- sga:GATA3: GATA3 gene encodes a **transcription factor** that is important for mammary gland development and differentiation.
- sga:CDH1: CDH1 gene encodes E-cadherin, a protein that is important for **cell-cell adhesion.** Mutation can lead to the dissociation of cells and promote cancer invasion and metastasis.

- deg:CHEK2: encodes the checkpoint kinase 2 protein, which is involved in DNA damage response and cell cycle checkpoint control. linked to an increased risk of breast, prostate, and other cancers.
- deg:DNMT1: encodes an enzyme that plays a role in DNA methylation and gene regulation.
- deg:FBXO5: encodes F-box protein 5 involved in protein degradation via the ubiquitin-proteasome pathway.
- deg:PDSS1: encodes the prenyl diphosphate synthase subunit 1 protein, which is involved in the biosynthesis of ubiquinone. associated with cancer development and progression, as well as mitochondrial dysfunction.
- deg:TJP3: This gene encodes the tight junction protein 3, which is involved in cell-cell adhesion and barrier function in epithelial and endothelial tissues. Dysregulation of TJP3 expression has been linked to various types of cancer, including lung and ovarian cancer.
- deg:TP53: this gene encodes the tumor suppressor protein p53, which plays a critical role in regulating cell cycle progression, DNA repair, and apoptosis.

Future outlook and work

- AI's huge potential for scientific discovery,

  e.g. cancer cause discovery.
- Human knowledge is also crucial, e.g. 40 hours VS 4 seconds
- Potential future work
  - Ask clinical experts to help interpret the results;
  - Use a larger TCGA dataset;
  - Inject more clinical knowledge into the causal discovery process;
  - Diverse effective metrics for evaluation of graph discovery results;
  - Improved algorithm stability;

Thanks for listening!

Validating Assumptions:

In simplest analysis, assume bipartite structure

1. SGAs are independent
2. DEGs are conditionally independent given the SGAs

If 2. fails this may be because:

a. DEGs have causal effect on each other
b. DEGs have causal effect on SGAs (for example differentially expressed genes may cause errors in DNA replication process leading to more SGAs)
c. DEGs may have other confounders which are not included in the current 6 SGAs.