Scalable Causal Structure Learning via Amortized Conditional Independence Testing

James
Leiner

Carnegie Mellon University

Brian
Manzo

University of Michigan

Aaditya
Ramdas

Carnegie Mellon University

Wesley
Tansey

Memorial Sloan Kettering Cancer Center

James Leiner
May 8, 2025

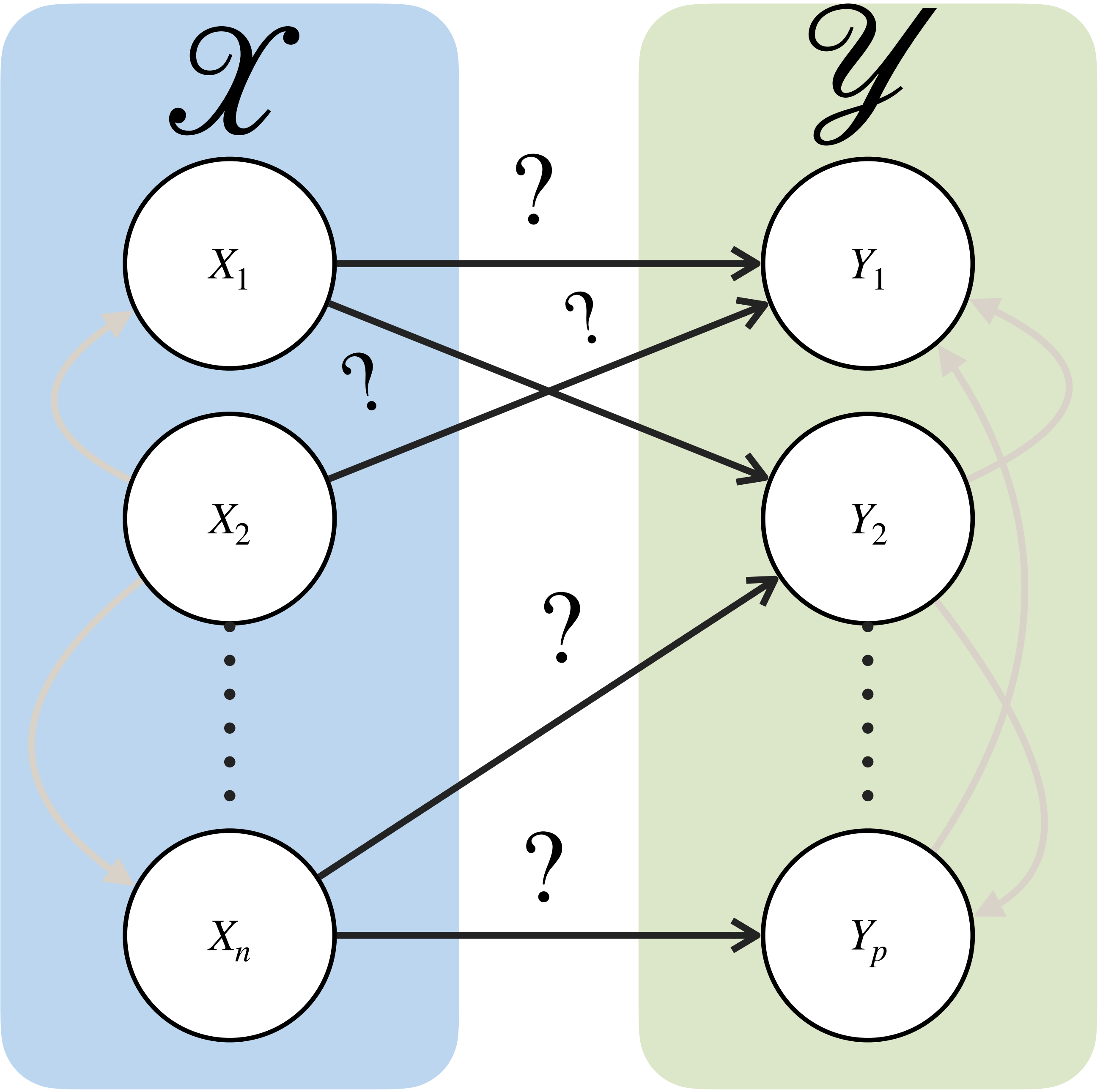
Consider a causal graph with two sets of nodes, $\mathscr{X}$ and $\mathscr{Y}$

Assume that $\mathscr{X}$ predates $\mathscr{Y}$

Key Question: Which edges exist between $\mathscr{X}$ and $\mathscr{Y}$?

The arrow of time implies that…

- No edge can be directed from $\mathscr{X}$ to $\mathscr{Y}$
- Edges between nodes in the same set can be oriented in any direction



A first step is to reduce the question to one of conditional independence relations

Assume the graph…

- satisfies the global directed Markov property
- is d-separation faithful
- does not contain latent confounders

$X_j \to Y_k$ is absent $\Longleftrightarrow$ $\exists S \subseteq Y_{-k}$ such that $X_j \perp Y_k \,|\, S, X_{-j}$

$X_j \to Y_k$ is present $\Longleftrightarrow$ $X_j \not\perp Y_k \,|\, S, X_{-j}$ for all $S \subseteq Y_{-k}$

Our goal is to learn edge-specific p-values for the graph

Key Inequality $p_{X_j \to Y_k} \leq \max_{S \subseteq Y_{-k}} p_{X_j \perp Y_k | S, X_{-j}}$

Exhaustive querying of all CI relationships is valid but may not be computationally feasible for even moderately sized graphs…

Prior work on causal discovery either…

- Searches for a graph (e.g. by maximizing a score function) but does not produce p-values with frequentist guarantees
- Outputs edge-specific p-values but only under the assumption of zero Type II error (i.e. no erroneous edge deletions) [Strobl et al., 2019]

We tackle this problem in two steps

1. **Find a function $T_{X_j,Y_k}(\,\cdot\,)$ that takes in S as an input and outputs a statistic for the hypothesis** $X_j \perp Y_k \,|\, S, X_{-j}$
2. **Use discrete optimization to find** $\hat{S} :=$ **arg** $\min_{S \subseteq Y_{-k}} T_{X_j,Y_k}(S)$

Shah and Peters [2018]

Generalized Covariance Measure

Target Estimand: $\mathbb{E}\left[\mathbb{E}[X_jY_k \,|\, S, X_{-j}] - \mathbb{E}[X_j \,|\, S, X_{-j}]\mathbb{E}[Y_k \,|\, S, X_{-j}]\right]$

(expected conditional covariance)

Inputs: Flexible ML estimates

$\hat{X}_j$ **targeting** $\mathbb{E}\left[X_j \,|\, S, X_{-j}\right]$

$\hat{Y}_k$ **targeting** $\mathbb{E}\left[Y_k \,|\, S, X_{-j}\right]$

Test statistic

Let $R_i = \left(X_j^i - \widehat{X}_j^i\right)\left(Y_k^i - \hat{Y}_k^i\right)$

If ML estimates converge sufficiently fast, then under the null (and appropriate regularity conditions),

$$T_{X_j,Y_k}^{(n)} := \frac{\sqrt{n} \cdot \frac{1}{n}\sum_{i=1}^{n} R_i}{\left(\frac{1}{n}\sum_{i=1}^{n} R_i^2 - \left(\frac{1}{n}\sum_{r=1}^{n} R_r\right)^2\right)^{1/2}} \approx N(0,1)$$

(won't have power against alternatives that are dependent but with 0 expected conditional covariance)

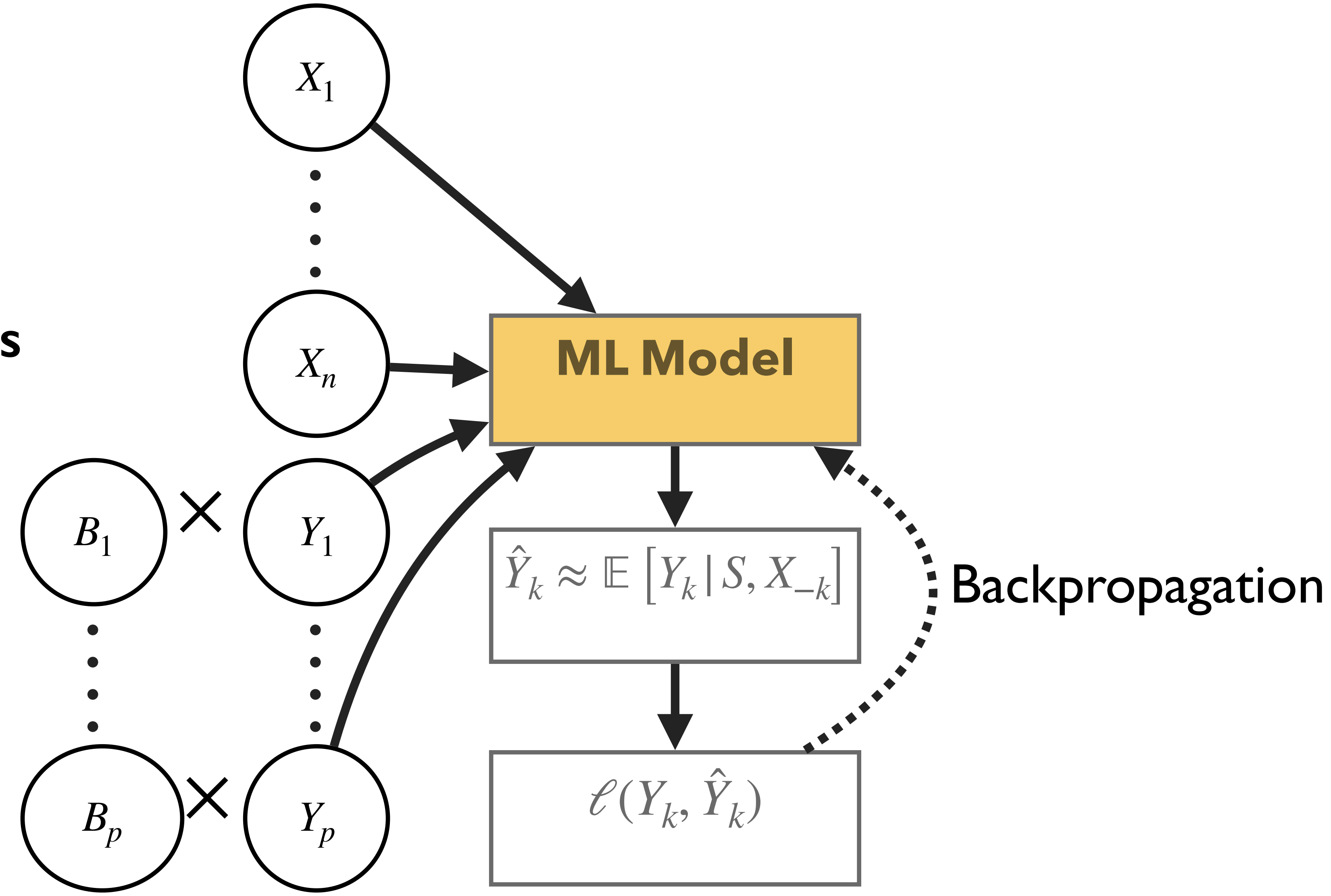
Using the GCM converts the CI testing problem to one of conditional mean estimation

Desiderata: train models $\hat{X}_j(\cdot)$ and $\hat{Y}_k(\cdot)$ that target $\mathbb{E}\left[X_j \mid S, X_{-j}\right]$ **and** $\mathbb{E}\left[Y_k \mid S, X_{-j}\right]$

Intuitively, we need to "hide" some pieces of information during training to mask out $Y_k \notin S$

During training, sample masks $B_k \sim$ Ber(p)

$S := \{Y_k$ **s.t.** $B_k = 1\}$



When using model, manually let $B_k = 1$ for all $Y_k \in S$ (given arbitrary choice of S)

Training process mimics process of an end user arbitrarily evaluating different conditioning subsets

Jang et al. [2017]

Gumbel-Softmax Optimization

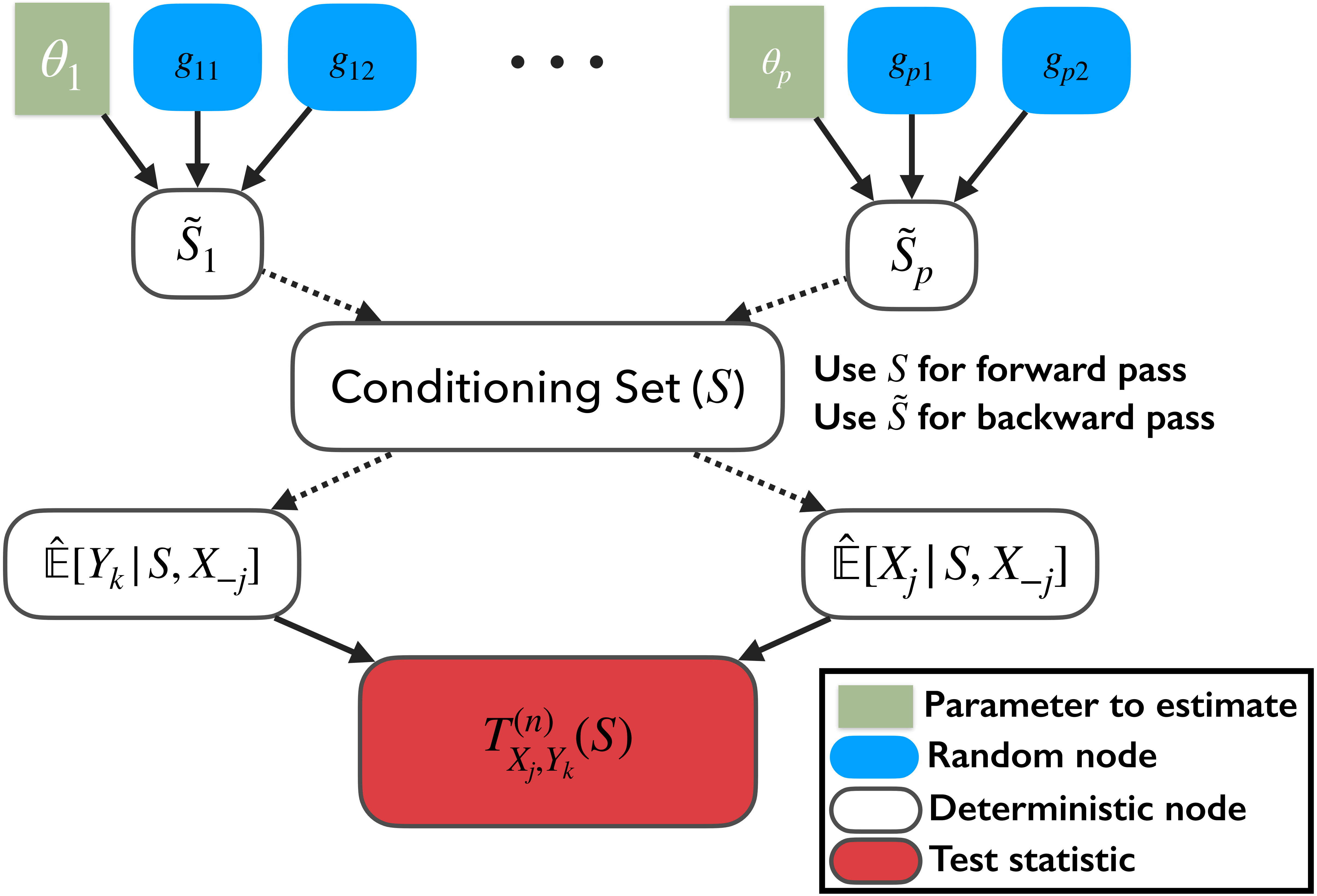
Desiderata: Learn $\arg \min_{\theta_1,\ldots,\theta_p} \mathbb{E}\left[T_n(S)\right]$ **where** $1_{Y_k \in S} \sim \text{Ber}(\theta_k)$

To enable back propagation, we replace $\frac{\partial T_n}{\partial S} \approx \frac{\partial T_n}{\partial \tilde{S}}$ **where** $\tilde{S}$ **is a continuous relaxation of** S

$$\tilde{S}_i = \frac{\exp\left((\log \theta_i + g_{i1})/\tau\right)}{\exp\left((\log \theta_i + g_{i1})/\tau\right) + \exp\left(\left(\log(1-\theta_i) + g_{i2}\right)/\tau\right)} \qquad g_{i1}, g_{i2} \sim \text{Gumbel}(0,1)$$

$\tau \to 0$ approximates a discrete distribution

We can now learn the conditioning set with gradient descent



Use S for forward pass
Use $\tilde{S}$ for backward pass

🟩 Parameter to estimate
🟦 Random node
□ Deterministic node
🟥 Test statistic

Results

We consider a cancer dataset as a motivating example

Dataset [Nguyen et al., 2022] $n = 22{,}352$ patients where

- $\mathscr{X}$ contains binary variables indicating whether certain mutations are contained in the primary tumor site
- $\mathscr{Y}$ contains binary variables indicating whether metastases have developed in secondary locations

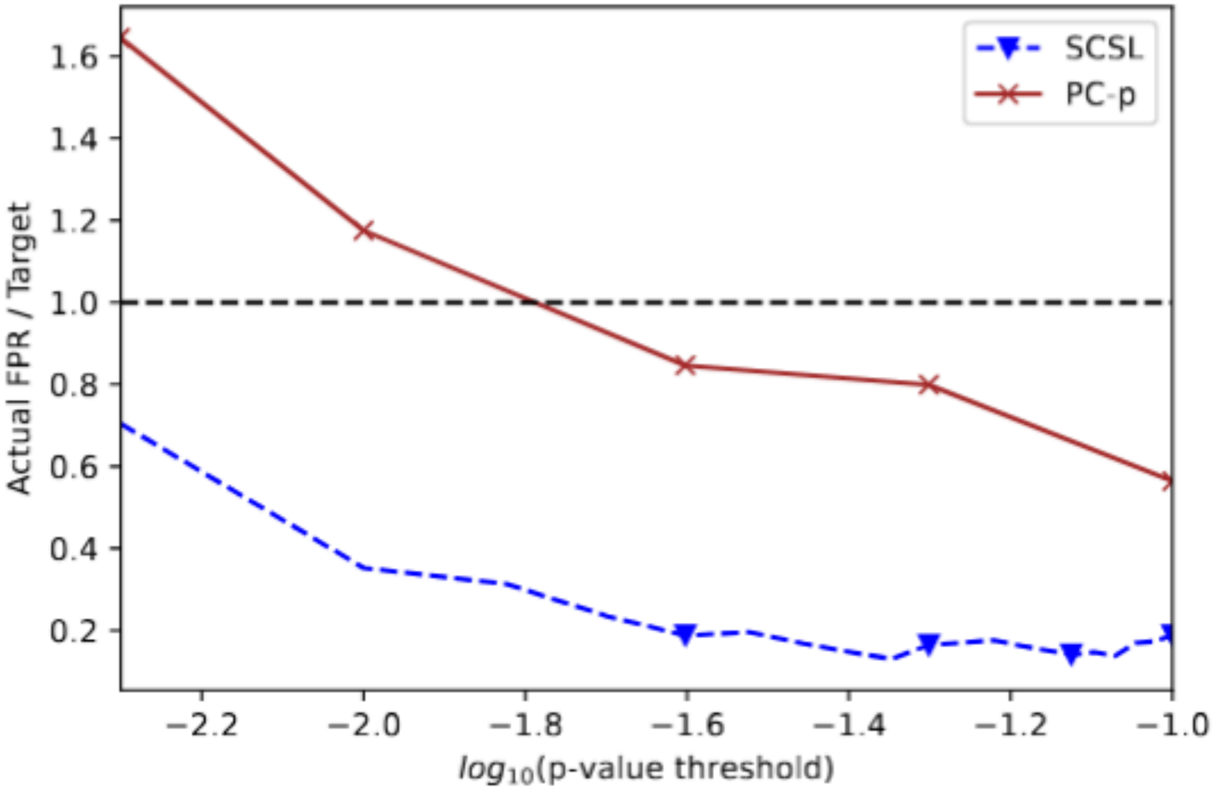
Discovering connections of the form $X_j \rightarrow Y_k$ allow us to proactively screen at-risk patients and better understand the progression of the disease.

We test using semi-synthetic data…

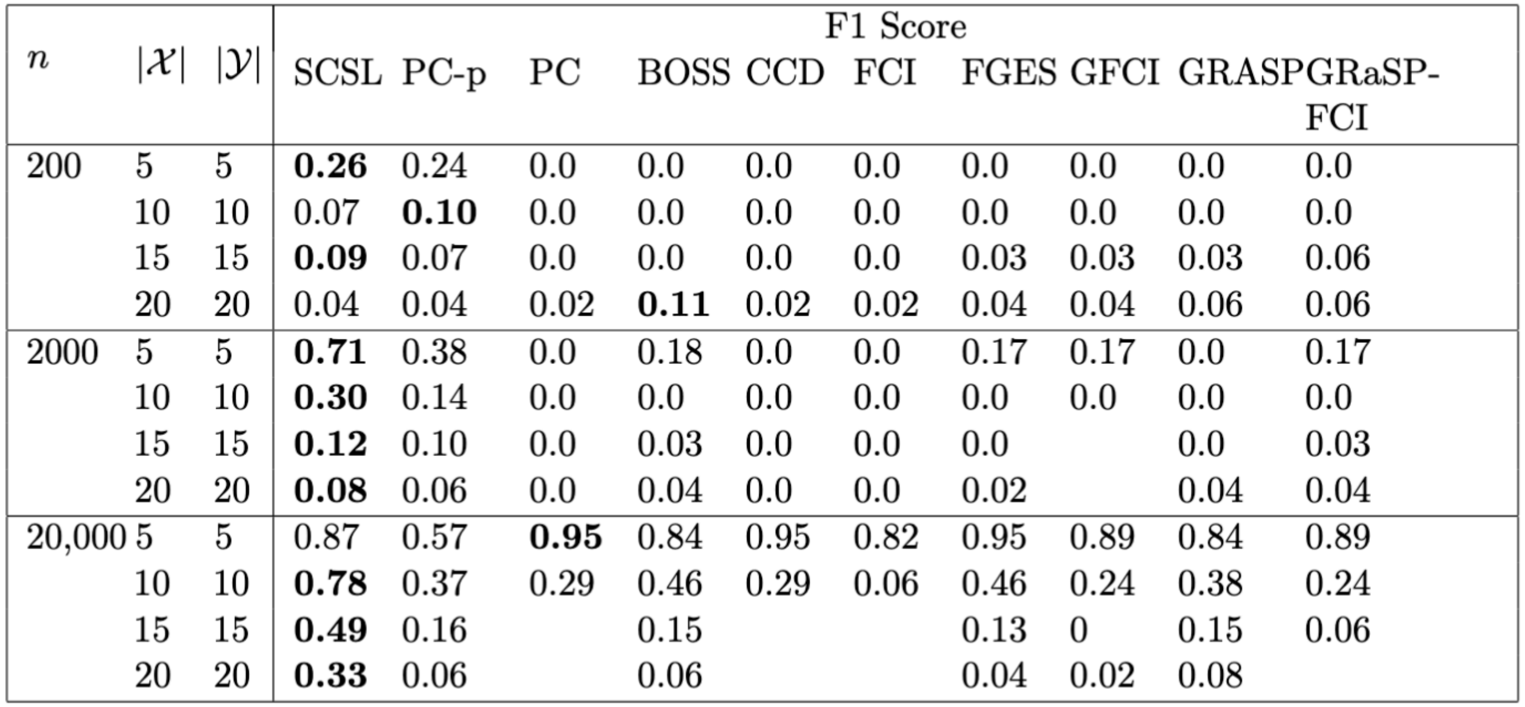
1. Posit a logistic model $\mathscr{P}$ relating $\mathscr{X}$ and $\mathscr{Y}$.
2. For reach patient, we calculate $\pi_i := \mathscr{P}(\mathscr{Y}_i \,|\, \mathscr{X}_i)$ as the likelihood of this row under the assumed model.
3. Construct new dataset by sampling Cat$(\pi_1, \ldots \pi_n)$.
4. This preserves marginal distributions of $\mathscr{X}$ and $\mathscr{Y}$ while providing ground truth knowledge of causal relationship

SCSL controls Type I error and has high power

The only other causal discovery method that produces p-values has inflated type I error, while SCSL is conservative.



SCSL also often has improved performance even when compared to methods not designed for frequentist error control…

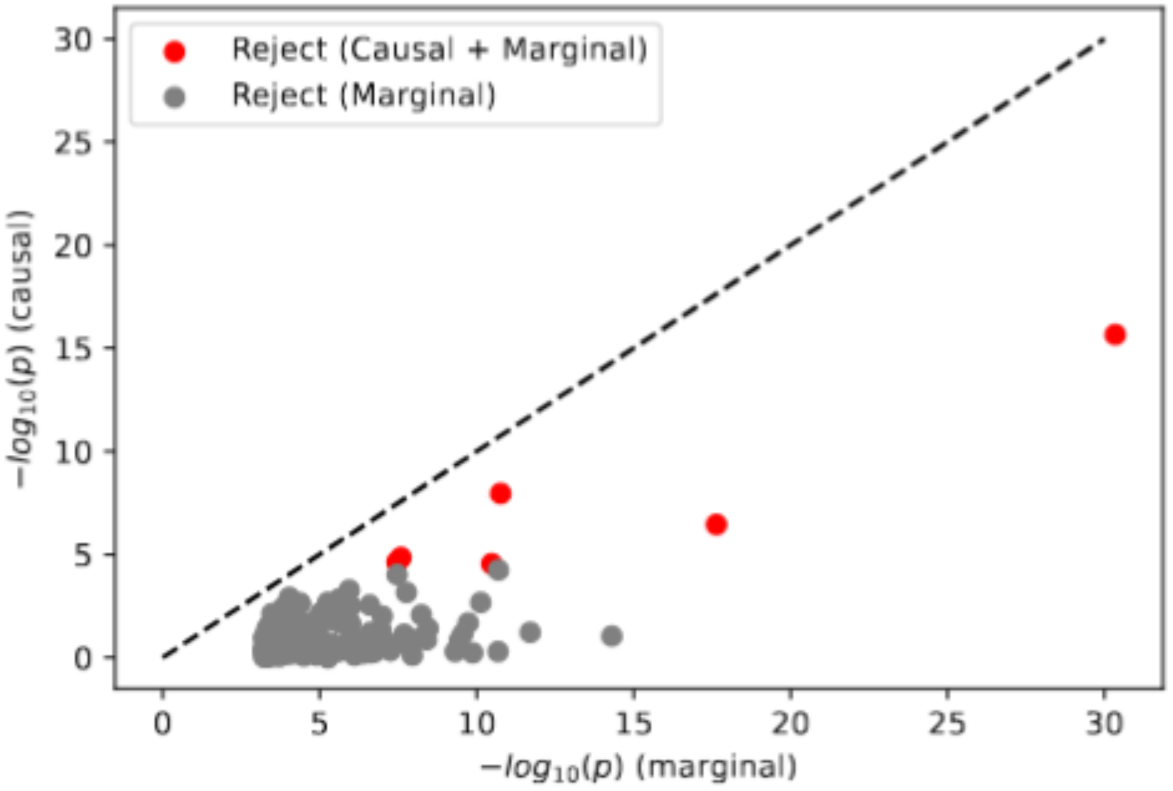


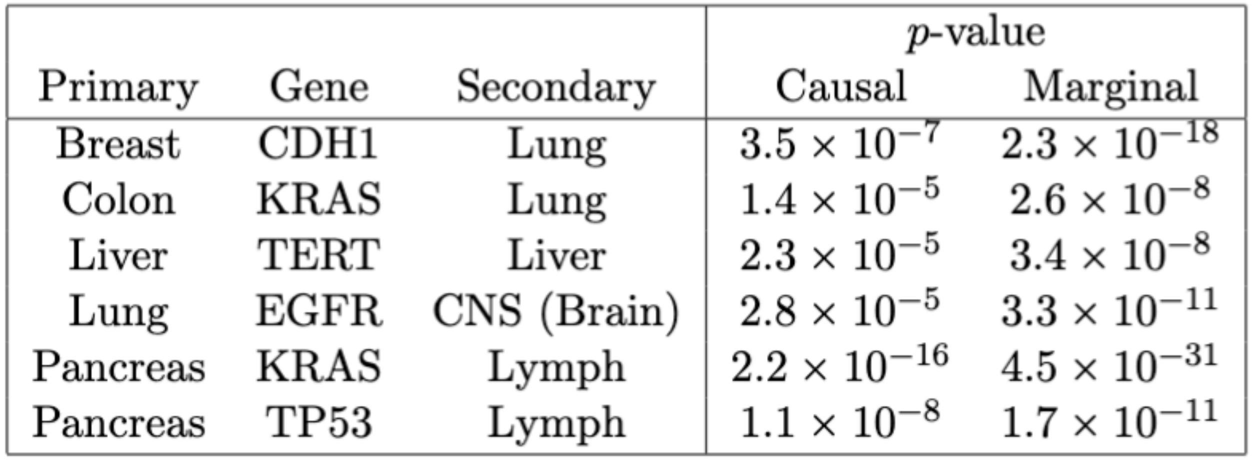
n	$\lvert\mathcal{X}\rvert$	$\lvert\mathcal{Y}\rvert$	F1 Score									
			SCSL	PC-p	PC	BOSS	CCD	FCI	FGES	GFCI	GRASP	GRaSP-FCI
200	5	5	**0.26**	0.24	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	10	10	0.07	**0.10**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	15	15	**0.09**	0.07	0.0	0.0	0.0	0.0	0.03	0.03	0.03	0.06
	20	20	0.04	0.04	0.02	**0.11**	0.02	0.02	0.04	0.04	0.06	0.06
2000	5	5	**0.71**	0.38	0.0	0.18	0.0	0.0	0.17	0.17	0.0	0.17
	10	10	**0.30**	0.14	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	15	15	**0.12**	0.10	0.0	0.03	0.0	0.0	0.0		0.0	0.03
	20	20	**0.08**	0.06	0.0	0.04	0.0	0.0	0.02		0.04	0.04
20,000	5	5	0.87	0.57	**0.95**	0.84	0.95	0.82	0.95	0.89	0.84	0.89
	10	10	**0.78**	0.37	0.29	0.46	0.29	0.06	0.46	0.24	0.38	0.24
	15	15	**0.49**	0.16		0.15			0.13	0	0.15	0.06
	20	20	**0.33**	0.06		0.06			0.04	0.02	0.08	

On real data, the method reveals interesting connections between mutations and metastases

In the original study, 161 discoveries were identified using associative p-values with a Benjamini-Hochberg (BH) adjustment

Only 6 discoveries remain when substituting causal p-values with the same BH adjustment.





Primary	Gene	Secondary	p-value Causal	 Marginal
Breast	CDH1	Lung	3.5×10^{-7}	2.3×10^{-18}
Colon	KRAS	Lung	1.4×10^{-5}	2.6×10^{-8}
Liver	TERT	Liver	2.3×10^{-5}	3.4×10^{-8}
Lung	EGFR	CNS (Brain)	2.8×10^{-5}	3.3×10^{-11}
Pancreas	KRAS	Lymph	2.2×10^{-16}	4.5×10^{-31}
Pancreas	TP53	Lymph	1.1×10^{-8}	1.7×10^{-11}

Thank you!