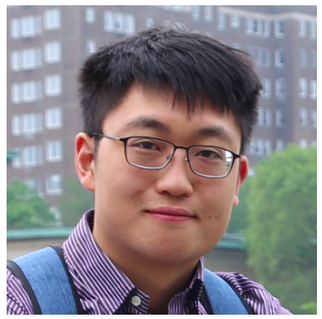
Active multiple testing w/ proxy p-value and e-values

DeGroot Workshop @ CMU 2025

Joint work with:



Neil Xu

Catherine Wang

Larry Wasserman

Kathryn Roeder

Aaditya Ramdas

Carnegie Mellon University

Hypothesis testing under resource constraints

Hypothesis testing under resource constraints

Hypothesis testing under resource constraints

Large scale hypothesis testing can meet resource constraints…

Hypothesis testing under resource constraints

Large scale hypothesis testing can meet resource constraints…

- *Computational constraints:* Computing the statistic for every hypothesis is slow

Hypothesis testing under resource constraints

Large scale hypothesis testing can meet resource constraints…

- *Computational constraints:* Computing the statistic for every hypothesis is slow
  - Cross-fitting + sampling.

Hypothesis testing under resource constraints

Large scale hypothesis testing can meet resource constraints…

- *Computational constraints:* Computing the statistic for every hypothesis is slow
  - Cross-fitting + sampling.
  - Nonconvex optimization.

Hypothesis testing under resource constraints

Large scale hypothesis testing can meet resource constraints…

- *Computational constraints:* Computing the statistic for every hypothesis is slow
  - Cross-fitting + sampling.
  - Nonconvex optimization.
- *Cost constraints:* obtaining the data is expensive

Hypothesis testing under resource constraints

Large scale hypothesis testing can meet resource constraints…

- *Computational constraints:* Computing the statistic for every hypothesis is slow
  - Cross-fitting + sampling.
  - Nonconvex optimization.
- *Cost constraints:* obtaining the data is expensive
  - treating patients

Hypothesis testing under resource constraints

Large scale hypothesis testing can meet resource constraints…

- *Computational constraints:* Computing the statistic for every hypothesis is slow
  - Cross-fitting + sampling.
  - Nonconvex optimization.
- *Cost constraints:* obtaining the data is expensive
  - treating patients
  - sending out coupons/promotions.

Hypothesis testing under resource constraints

Large scale hypothesis testing can meet resource constraints…

- *Computational constraints:* Computing the statistic for every hypothesis is slow
  - Cross-fitting + sampling.
  - Nonconvex optimization.
- *Cost constraints:* obtaining the data is expensive
  - treating patients
  - sending out coupons/promotions.
- *Time constraints:* true outcome is in the future.

Hypothesis testing under resource constraints

Large scale hypothesis testing can meet resource constraints…

- *Computational constraints:* Computing the statistic for every hypothesis is slow
  - Cross-fitting + sampling.
  - Nonconvex optimization.
- *Cost constraints:* obtaining the data is expensive
  - treating patients
  - sending out coupons/promotions.
- *Time constraints:* true outcome is in the future.
  - panel/longitudinal experiments

Hypothesis testing under resource constraints

Large scale hypothesis testing can meet resource constraints…

- *Computational constraints:* Computing the statistic for every hypothesis is slow
  - Cross-fitting + sampling.
  - Nonconvex optimization.
- *Cost constraints:* obtaining the data is expensive
  - treating patients
  - sending out coupons/promotions.
- *Time constraints:* true outcome is in the future.
  - panel/longitudinal experiments

Hypothesis testing under resource constraints

Large scale hypothesis testing can meet resource constraints…

- *Computational constraints:* Computing the statistic for every hypothesis is slow
  - Cross-fitting + sampling.
  - Nonconvex optimization.
- *Cost constraints:* obtaining the data is expensive
  - treating patients
  - sending out coupons/promotions.
- *Time constraints:* true outcome is in the future.
  - panel/longitudinal experiments

…but we often have “cheap” but inaccurate proxy statistics

Hypothesis testing under resource constraints

Large scale hypothesis testing can meet resource constraints…

- *Computational constraints:* Computing the statistic for every hypothesis is slow
  - Cross-fitting + sampling.
  - Nonconvex optimization.
- *Cost constraints:* obtaining the data is expensive
  - treating patients
  - sending out coupons/promotions.
- *Time constraints:* true outcome is in the future.
  - panel/longitudinal experiments

…but we often have “cheap” but inaccurate proxy statistics

➡ *Approximations*

Hypothesis testing under resource constraints

Large scale hypothesis testing can meet resource constraints…

- *Computational constraints:* Computing the statistic for every hypothesis is slow
  - Cross-fitting + sampling.
  - Nonconvex optimization.
- *Cost constraints:* obtaining the data is expensive
  - treating patients
  - sending out coupons/promotions.
- *Time constraints:* true outcome is in the future.
  - panel/longitudinal experiments

…but we often have “cheap” but inaccurate proxy statistics

➡ *Approximations*

- In-sample fitting once

Hypothesis testing under resource constraints

Large scale hypothesis testing can meet resource constraints…

- *Computational constraints:* Computing the statistic for every hypothesis is slow
  - Cross-fitting + sampling.
  - Nonconvex optimization.
- *Cost constraints:* obtaining the data is expensive
  - treating patients
  - sending out coupons/promotions.
- *Time constraints:* true outcome is in the future.
  - panel/longitudinal experiments

…but we often have “cheap” but inaccurate proxy statistics

➡ *Approximations*

- In-sample fitting once
- Neural networks

Hypothesis testing under resource constraints

Large scale hypothesis testing can meet resource constraints…

- *Computational constraints:* Computing the statistic for every hypothesis is slow
  - Cross-fitting + sampling.
  - Nonconvex optimization.
- *Cost constraints:* obtaining the data is expensive
  - treating patients
  - sending out coupons/promotions.
- *Time constraints:* true outcome is in the future.
  - panel/longitudinal experiments

…but we often have “cheap” but inaccurate proxy statistics

➡ *Approximations*

- In-sample fitting once
- Neural networks

Hypothesis testing under resource constraints

Large scale hypothesis testing can meet resource constraints…

- *Computational constraints:* Computing the statistic for every hypothesis is slow
  - Cross-fitting + sampling.
  - Nonconvex optimization.
- *Cost constraints:* obtaining the data is expensive
  - treating patients
  - sending out coupons/promotions.
- *Time constraints:* true outcome is in the future.
  - panel/longitudinal experiments

…but we often have “cheap” but inaccurate proxy statistics

➡ *Approximations*

- In-sample fitting once
- Neural networks

➡ *Outcome predictions*

Hypothesis testing under resource constraints

Large scale hypothesis testing can meet resource constraints…

- *Computational constraints:* Computing the statistic for every hypothesis is slow
  - Cross-fitting + sampling.
  - Nonconvex optimization.
- *Cost constraints:* obtaining the data is expensive
  - treating patients
  - sending out coupons/promotions.
- *Time constraints:* true outcome is in the future.
  - panel/longitudinal experiments

…but we often have “cheap” but inaccurate proxy statistics

➡ *Approximations*

- In-sample fitting once
- Neural networks

➡ *Outcome predictions*

- Machine learning predictors

Hypothesis testing under resource constraints

Large scale hypothesis testing can meet resource constraints…

- *Computational constraints:* Computing the statistic for every hypothesis is slow
  - Cross-fitting + sampling.
  - Nonconvex optimization.
- *Cost constraints:* obtaining the data is expensive
  - treating patients
  - sending out coupons/promotions.
- *Time constraints:* true outcome is in the future.
  - panel/longitudinal experiments

…but we often have “cheap” but inaccurate proxy statistics

➡ *Approximations*

- In-sample fitting once
- Neural networks

➡ *Outcome predictions*

- Machine learning predictors

Hypothesis testing under resource constraints

Large scale hypothesis testing can meet resource constraints…

- *Computational constraints:* Computing the statistic for every hypothesis is slow
  - Cross-fitting + sampling.
  - Nonconvex optimization.
- *Cost constraints:* obtaining the data is expensive
  - treating patients
  - sending out coupons/promotions.
- *Time constraints:* true outcome is in the future.
  - panel/longitudinal experiments

…but we often have “cheap” but inaccurate proxy statistics

➡ *Approximations*

- In-sample fitting once
- Neural networks

➡ *Outcome predictions*

- Machine learning predictors

➡ *Mediators*

Hypothesis testing under resource constraints

Large scale hypothesis testing can meet resource constraints…

- *Computational constraints:* Computing the statistic for every hypothesis is slow
  - Cross-fitting + sampling.
  - Nonconvex optimization.
- *Cost constraints:* obtaining the data is expensive
  - treating patients
  - sending out coupons/promotions.
- *Time constraints:* true outcome is in the future.
  - panel/longitudinal experiments

…but we often have “cheap” but inaccurate proxy statistics

➡ *Approximations*

- In-sample fitting once
- Neural networks

➡ *Outcome predictions*

- Machine learning predictors

➡ *Mediators*

- Preliminary outcomes

Hypothesis testing under resource constraints

Large scale hypothesis testing can meet resource constraints…

- *Computational constraints:* Computing the statistic for every hypothesis is slow
  - Cross-fitting + sampling.
  - Nonconvex optimization.
- *Cost constraints:* obtaining the data is expensive
  - treating patients
  - sending out coupons/promotions.
- *Time constraints:* true outcome is in the future.
  - panel/longitudinal experiments

…but we often have “cheap” but inaccurate proxy statistics

➡ *Approximations*

- In-sample fitting once
- Neural networks

➡ *Outcome predictions*

- Machine learning predictors

➡ *Mediators*

- Preliminary outcomes

Goal: Use proxy statistics to derive

Hypothesis testing under resource constraints

Large scale hypothesis testing can meet resource constraints…

- *Computational constraints:* Computing the statistic for every hypothesis is slow
  - Cross-fitting + sampling.
  - Nonconvex optimization.
- *Cost constraints:* obtaining the data is expensive
  - treating patients
  - sending out coupons/promotions.
- *Time constraints:* true outcome is in the future.
  - panel/longitudinal experiments

…but we often have “cheap” but inaccurate proxy statistics

➡ *Approximations*

- In-sample fitting once
- Neural networks

➡ *Outcome predictions*

- Machine learning predictors

➡ *Mediators*

- Preliminary outcomes

Goal: Use proxy statistics to derive

1. valid statistics for every hypothesis (correctness)

Hypothesis testing under resource constraints

Large scale hypothesis testing can meet resource constraints…

- *Computational constraints:* Computing the statistic for every hypothesis is slow
  - Cross-fitting + sampling.
  - Nonconvex optimization.
- *Cost constraints:* obtaining the data is expensive
  - treating patients
  - sending out coupons/promotions.
- *Time constraints:* true outcome is in the future.
  - panel/longitudinal experiments

…but we often have “cheap” but inaccurate proxy statistics

➡ *Approximations*
- In-sample fitting once
- Neural networks

➡ *Outcome predictions*
- Machine learning predictors

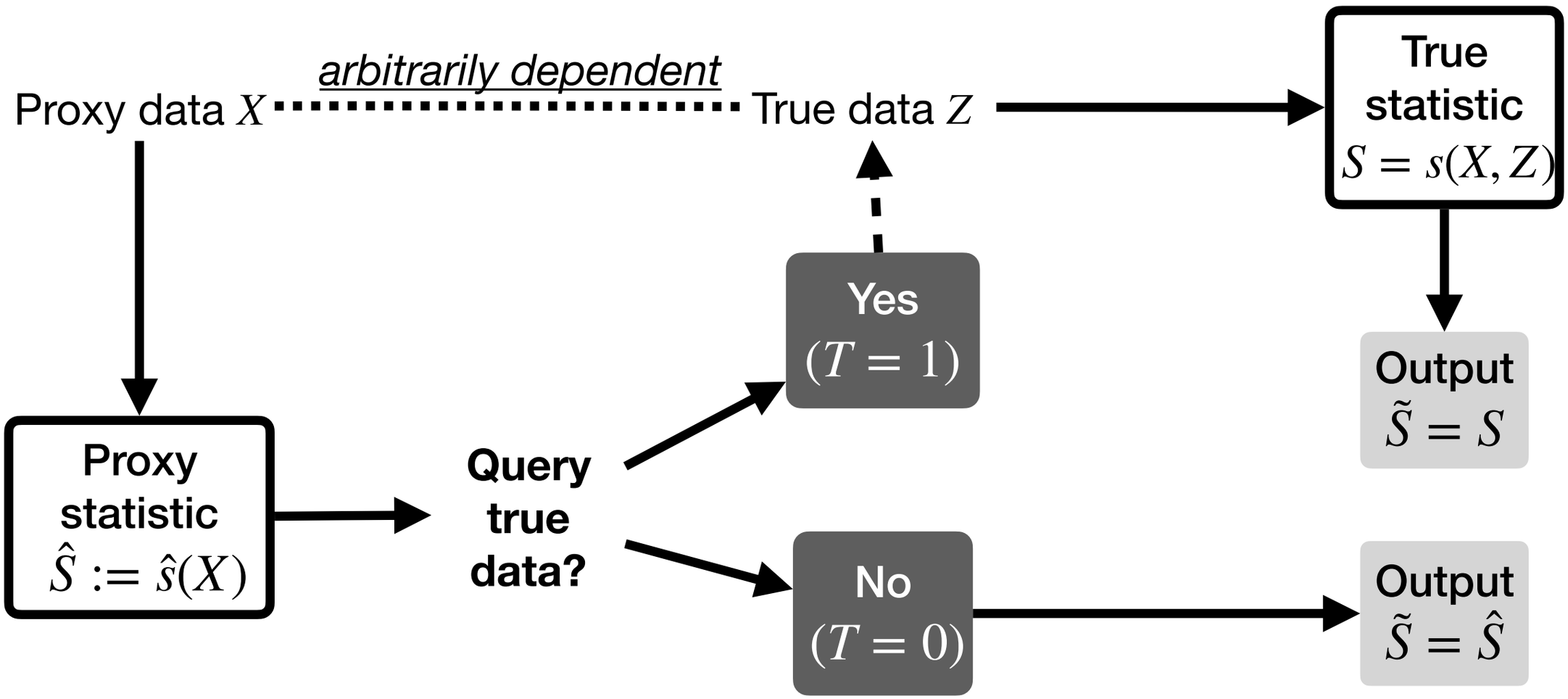
➡ *Mediators*
- Preliminary outcomes

Goal: Use proxy statistics to derive
1. valid statistics for every hypothesis (correctness)
2. while selectively computing few true statistics (efficiency)

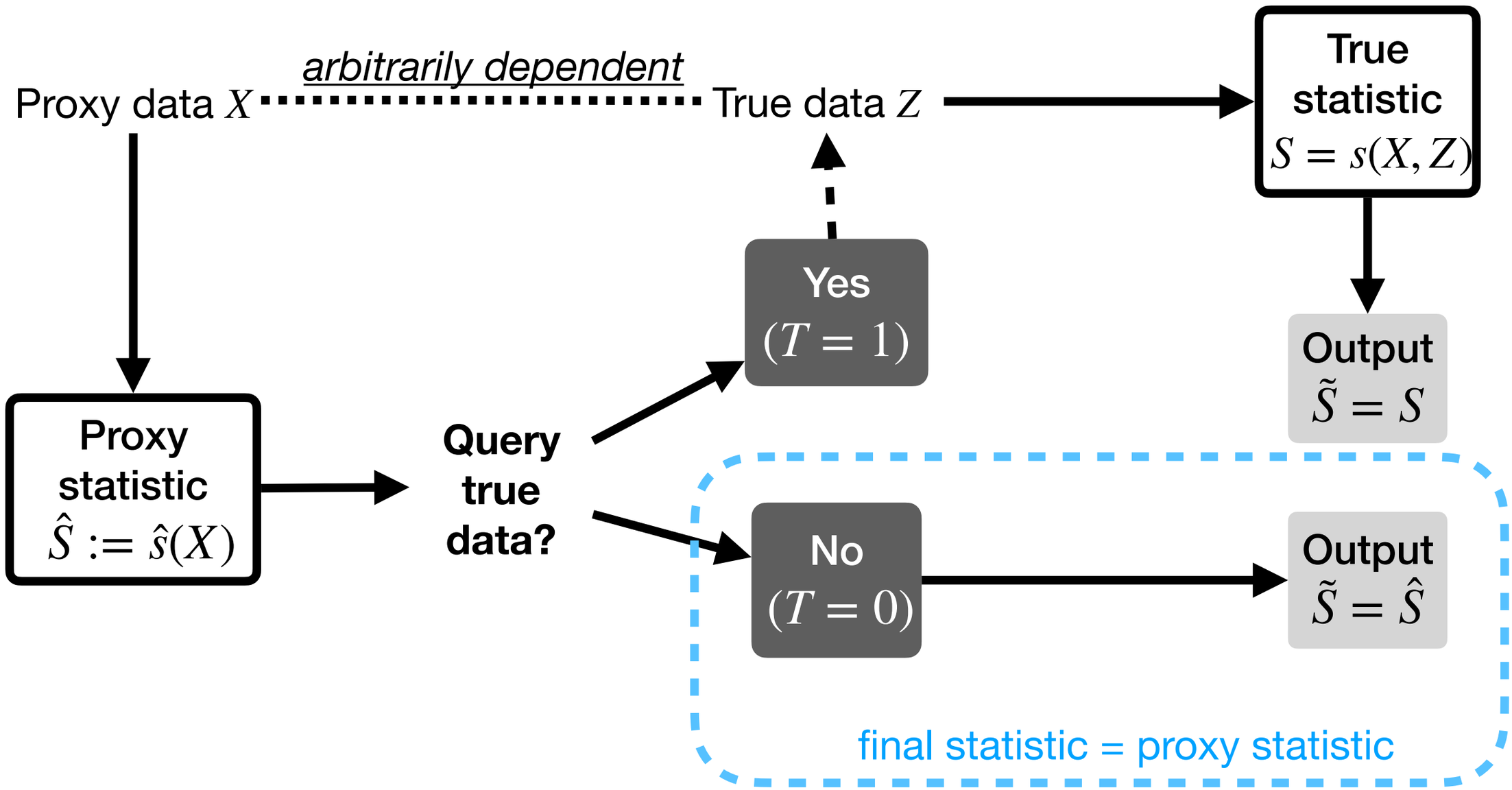
Outline

1. The active hypothesis testing framework
    1.1. Active p-values
2. Application: scCRISPR screening via proximal causal inference
    2.1. Proximal causal inference
    2.2. Two stage least squares
    2.3. Experimental results
3. Multiple testing w/ FDR control: active BH
4. Conclusion

Active hypothesis testing framework



Active hypothesis testing framework



Active p-values

Active p-values

- Q is a random variable in [0, 1] **(proxy p-value).**
- P is a bona-fide p-value: $\mathbb{P}(P \leq s) \leq s$ for all $s \in [0,1]$ **(true p-value).**

Active p-values

- Q is a random variable in [0, 1] **(proxy p-value).**
- P is a bona-fide p-value: $\mathbb{P}(P \leq s) \leq s$ for all $s \in [0,1]$ **(true p-value).**

Two kinds of active p-values:

Active p-values

- Q is a random variable in [0, 1] **(proxy p-value).**
- P is a bona-fide p-value: $\mathbb{P}(P \leq s) \leq s$ for all $s \in [0,1]$ **(true p-value).**

Two kinds of active p-values:

If Q and P are <u>independent under the null</u>

$$T \mid Q \sim \text{Bern}(1 - L/f(Q))$$

Final p-value $\tilde{P}^{\text{ind}} := (1 - T)Q + T \cdot P$

Active p-values

- Q is a random variable in [0, 1] **(proxy p-value).**
- P is a bona-fide p-value: $\mathbb{P}(P \leq s) \leq s$ for all $s \in [0,1]$ **(true p-value).**

Two kinds of active p-values:

If Q and P are <u>independent under the null</u>

$$T \mid Q \sim \text{Bern}(1 - L/f(Q))$$

Final p-value $\tilde{P}^{\text{ind}} := (1 - T)Q + T \cdot P$

f : density of Q under the null

$L \leq f(q)$ for all q

(1-d estimation problem)

Active p-values

- Q is a random variable in [0, 1] **(proxy p-value).**
- P is a bona-fide p-value: $\mathbb{P}(P \leq s) \leq s$ for all $s \in [0,1]$ **(true p-value).**

Two kinds of active p-values:

If Q and P are <u>independent under the null</u>

$$T \mid Q \sim \text{Bern}(1 - L/f(Q))$$

Final p-value $\tilde{P}^{\text{ind}} := (1-T)Q + T \cdot P$

If Q and P are <u>arbitrarily dependent</u> let

$$T \mid Q \sim \text{Bern}(1 - Q/2)$$

Final p-value $\tilde{P}^{\text{arb-dep}} := (1-T)Q + T \cdot 2P$

f : density of Q under the null

$L \leq f(q)$ for all q

(1-d estimation problem)

Active p-values

- Q is a random variable in [0, 1] **(proxy p-value).**
- P is a bona-fide p-value: $\mathbb{P}(P \leq s) \leq s$ for all $s \in [0,1]$ **(true p-value).**

Two kinds of active p-values:

If Q and P are <u>independent under the null</u>

$$T \mid Q \sim \text{Bern}(1 - L/f(Q))$$

Final p-value $\tilde{P}^{\text{ind}} := (1-T)Q + T \cdot P$

If Q and P are <u>arbitrarily dependent</u> let

$$T \mid Q \sim \text{Bern}(1 - Q/2)$$

Final p-value $\tilde{P}^{\text{arb-dep}} := (1-T)Q + T \cdot 2P$

f : density of Q under the null

$L \leq f(q)$ for all q

(1-d estimation problem)

not likely to sample nulls
(if Q is predictive of P)

Active p-values

- Q is a random variable in [0, 1] **(proxy p-value).**
- P is a bona-fide p-value: $\mathbb{P}(P \leq s) \leq s$ for all $s \in [0,1]$ **(true p-value).**

Two kinds of active p-values:

If Q and P are <u>independent under the null</u>

$$T \mid Q \sim \text{Bern}(1 - L/f(Q))$$

Final p-value $\tilde{P}^{\text{ind}} := (1-T)Q + T \cdot P$

f : density of Q under the null

$L \leq f(q)$ for all q

(1-d estimation problem)

If Q and P are <u>arbitrarily dependent</u> let

$$T \mid Q \sim \text{Bern}(1 - Q/2)$$

Final p-value $\tilde{P}^{\text{arb-dep}} := (1-T)Q + T \cdot 2P$

not likely to sample nulls (if Q is predictive of P)

Pays a union bound cost on the true p-value

Active p-values

- Q is a random variable in [0, 1] **(proxy p-value).**
- P is a bona-fide p-value: $\mathbb{P}(P \leq s) \leq s$ for all $s \in [0,1]$ **(true p-value).**

Two kinds of active p-values:

If Q and P are <u>independent under the null</u>

$$T \mid Q \sim \text{Bern}(1 - L/f(Q))$$

Final p-value $\tilde{P}^{\text{ind}} := (1-T)Q + T \cdot P$

f : density of Q under the null

$L \leq f(q)$ for all q

(1-d estimation problem)

If Q and P are <u>arbitrarily dependent</u> let

$$T \mid Q \sim \text{Bern}(1 - Q/2)$$

Final p-value $\tilde{P}^{\text{arb-dep}} := (1-T)Q + T \cdot 2P$

not likely to sample nulls (if Q is predictive of P)

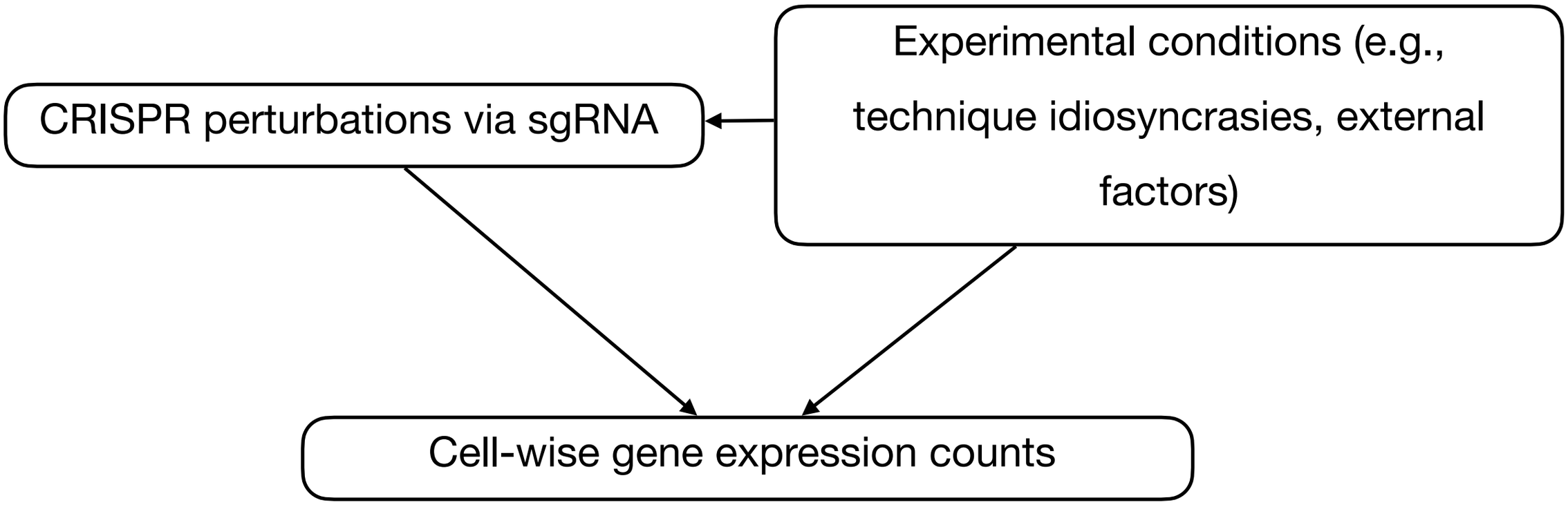
Pays a union bound cost on the true p-value

Theorem (ours): Active p-values $\tilde{P}^{\text{arb-dep}}$ and $\tilde{P}^{\text{ind}}$ are bona-fide p-values.

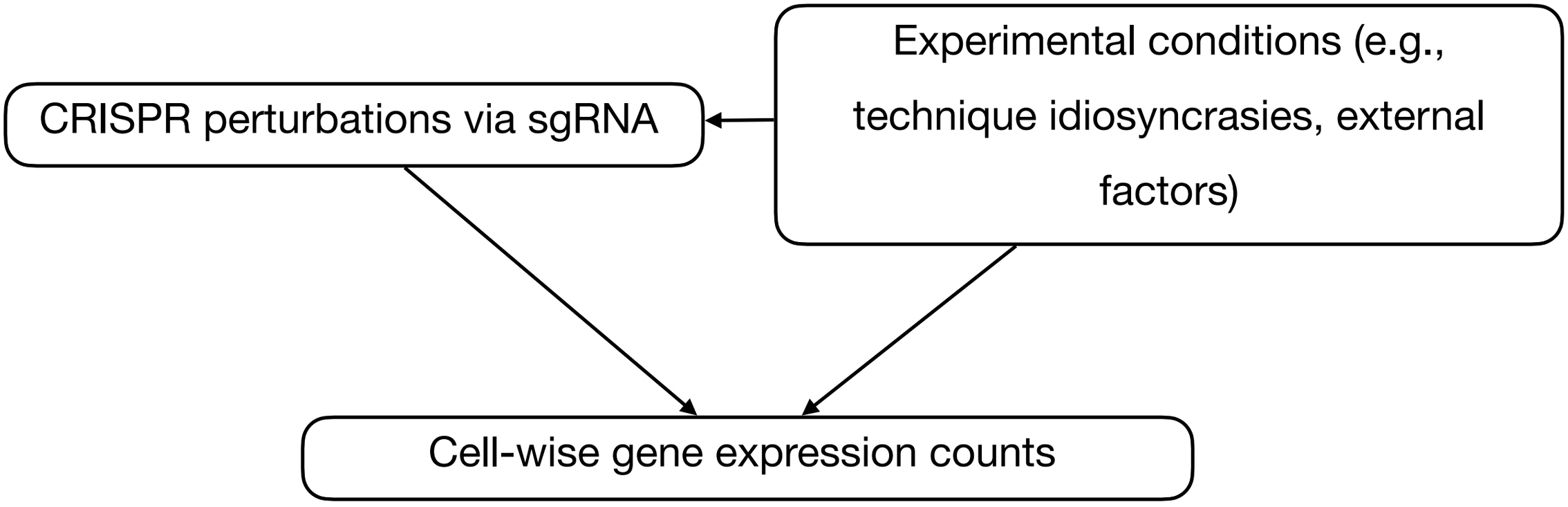
Outline

1. The active hypothesis testing framework
    1.1. Active p-values
2. Application: scCRISPR screening via proximal causal inference
    2.1. Proximal causal inference
    2.2. Two stage least squares
    2.3. Experimental results
3. Multiple testing w/ FDR control: active BH
4. Conclusion

scCRISPR w/ gene perturbations



scCRISPR w/ gene perturbations

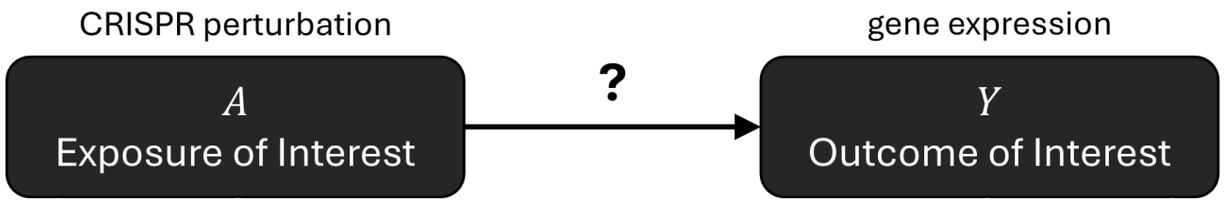


Goal: Test for causal effect of gene perturbation on cell-wise expression counts.

Proximal causal inference via negative controls

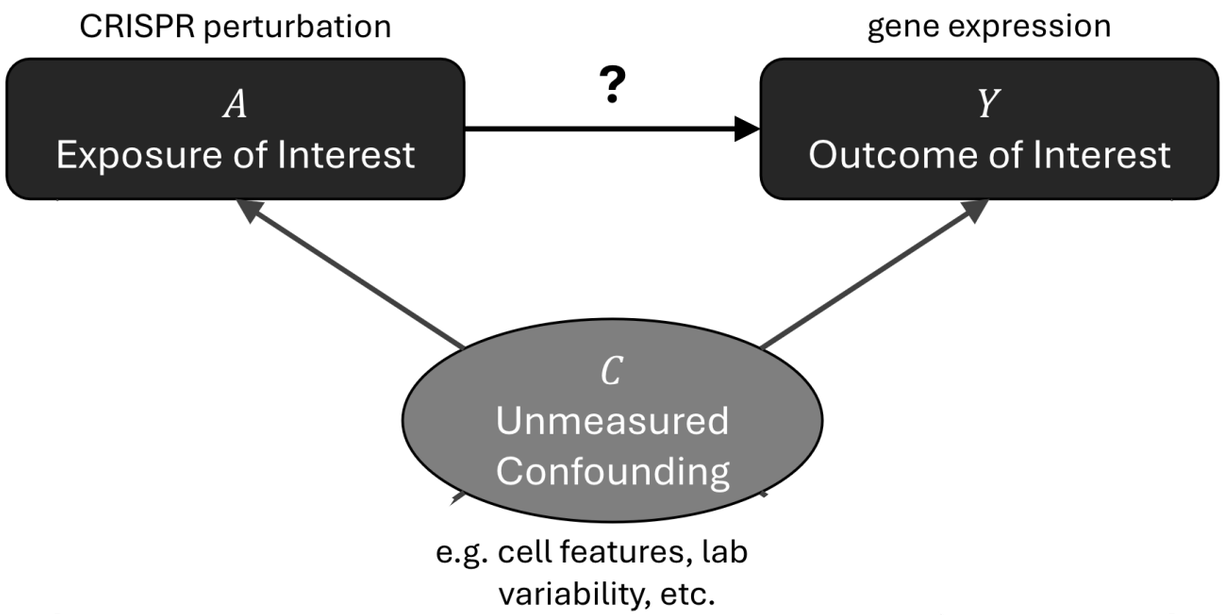
[1] Liu et al. Regression-Based Proximal Causal Inference. arXiv:2402.00335, 2024.

Proximal causal inference via negative controls



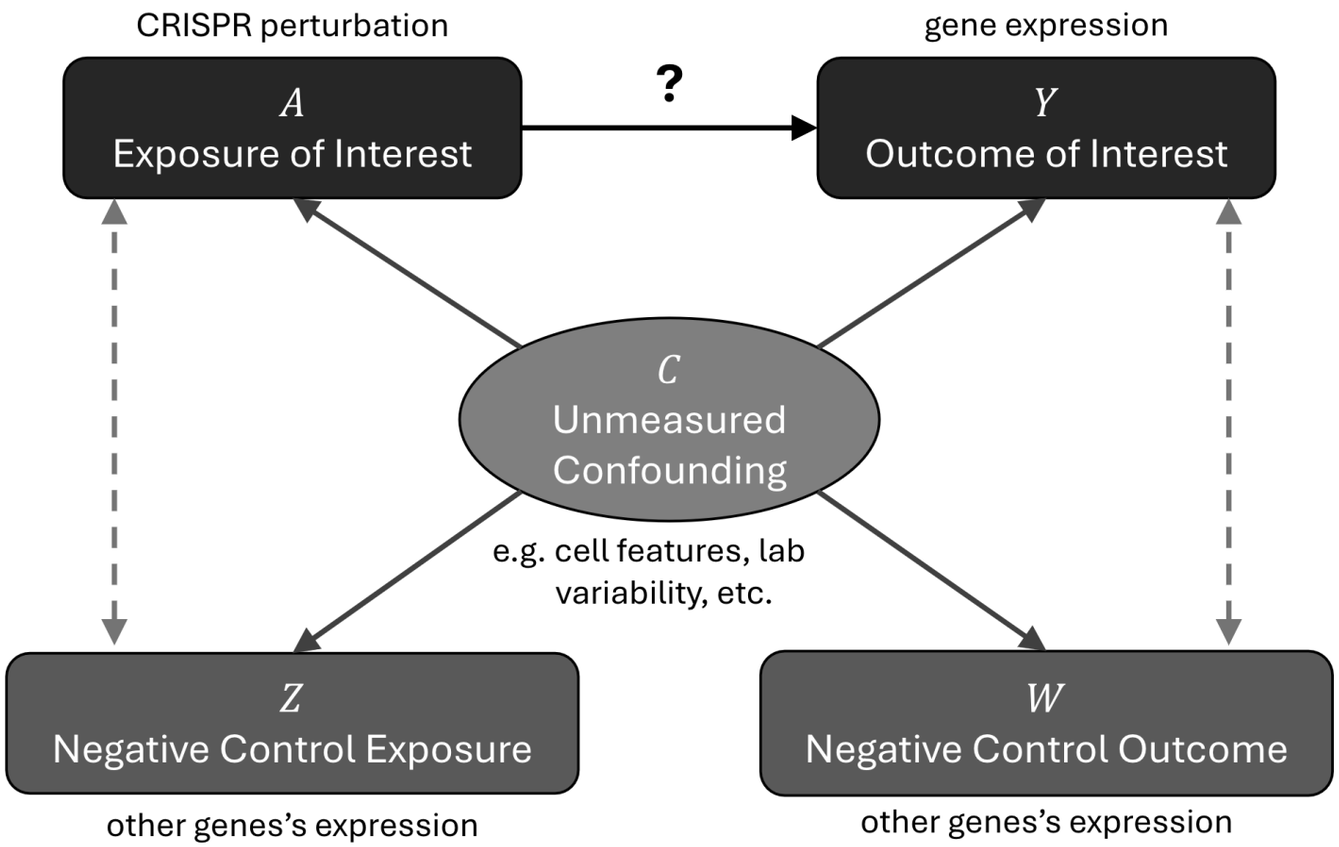
[1] Liu et al. Regression-Based Proximal Causal Inference. arXiv:2402.00335, 2024.

Proximal causal inference via negative controls



[1] Liu et al. Regression-Based Proximal Causal Inference. arXiv:2402.00335, 2024.

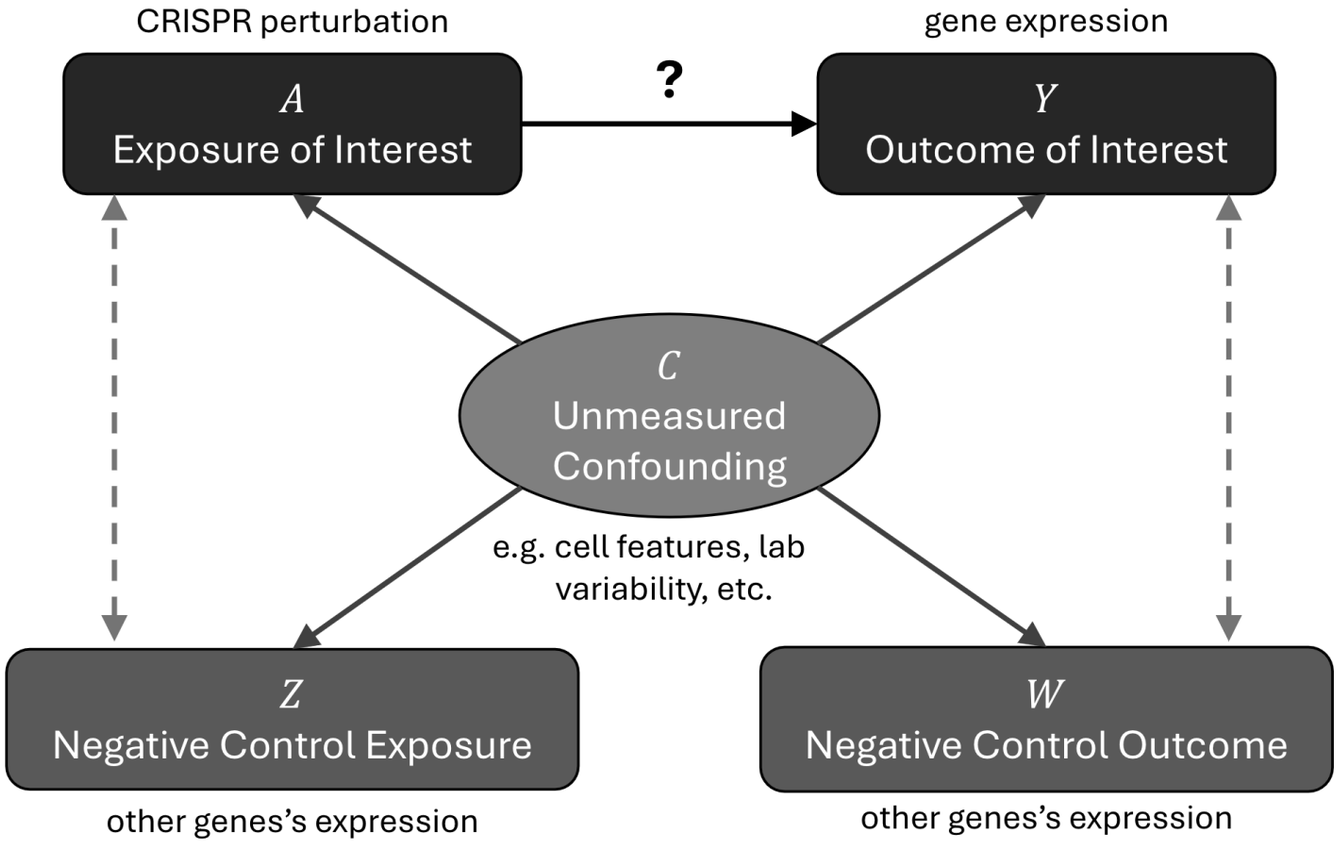
Proximal causal inference via negative controls



[1] Liu et al. Regression-Based Proximal Causal Inference. arXiv:2402.00335, 2024.

Proximal causal inference via negative controls

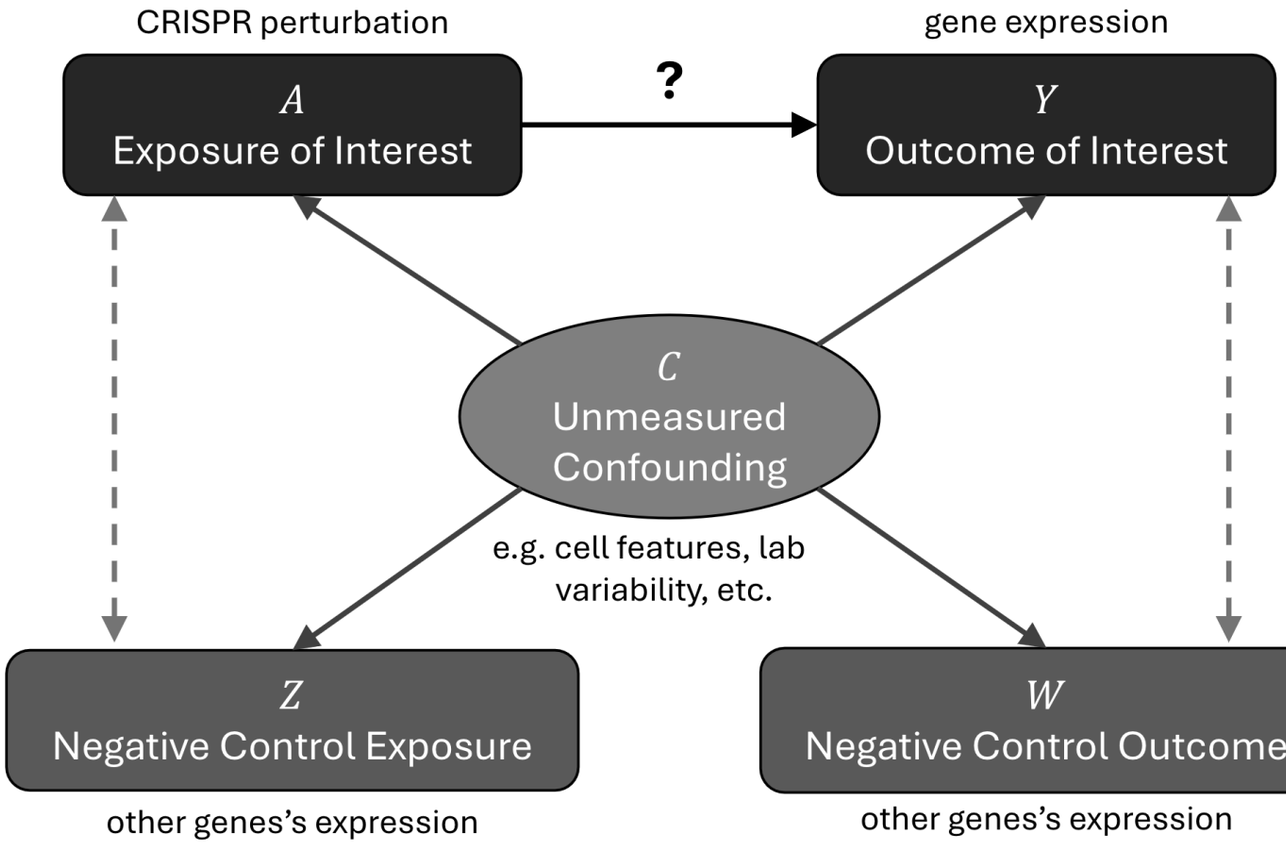
$$H_0 : \psi^{\mathsf{ATE}} = \mathbb{E}[Y^1 - Y^0] \neq 0$$



[1] Liu et al. Regression-Based Proximal Causal Inference. arXiv:2402.00335, 2024.

Proximal causal inference via negative controls

$$H_0 : \psi^{\mathsf{ATE}} = \mathbb{E}[Y^1 - Y^0] \neq 0$$



Linear proximal causal inference [1].

- $\mathbb{E}[Y \mid A, Z, C] = \beta_0 + \psi^{ATE} A + \beta_C^T C$
- $\mathbb{E}[W \mid A, Z, C] = \alpha_0 + \mathbf{M}^T C$

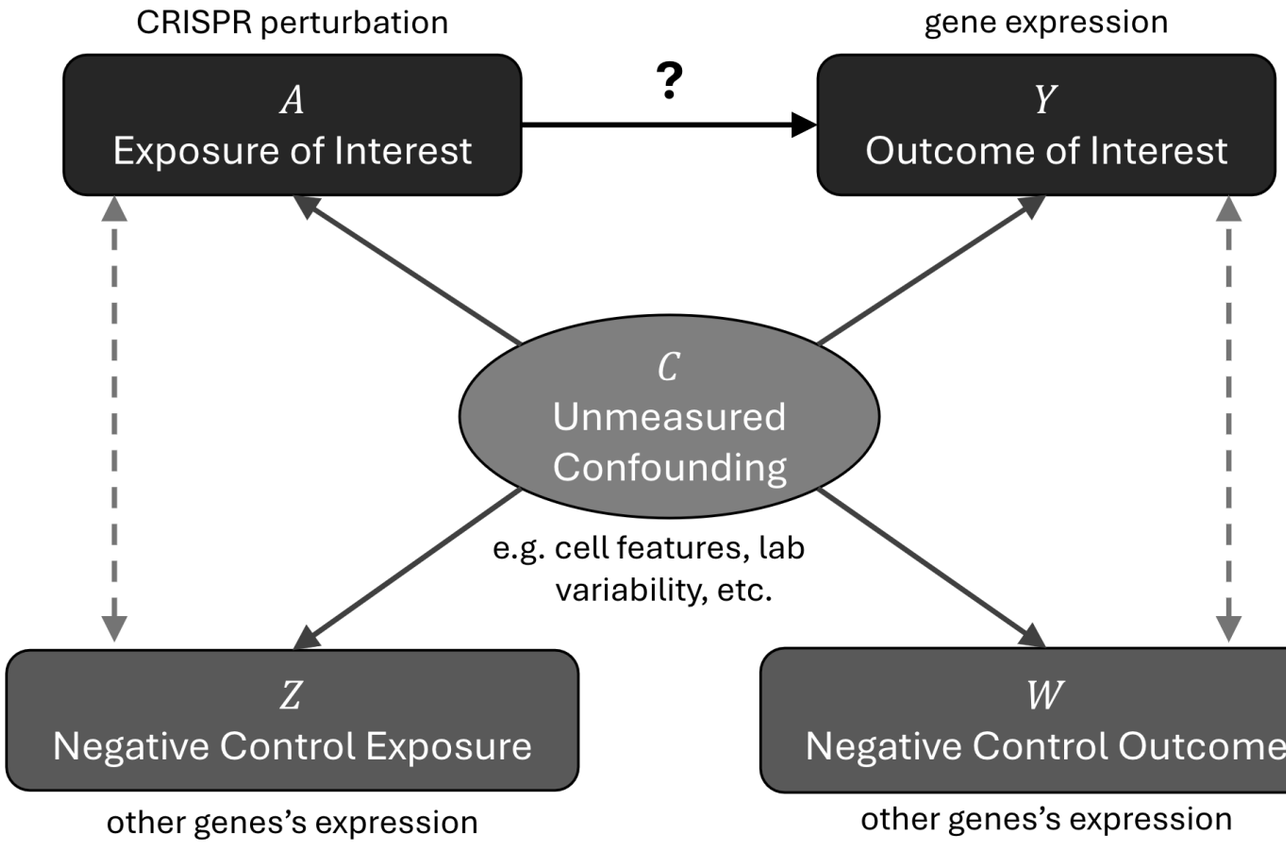
this implies…

- $\mathbb{E}[Y \mid A, Z] = \beta_0' + \psi^{\mathsf{ATE}} A + (\beta_w')^T \mathbb{E}[W \mid A, Z]$

[1] Liu et al. Regression-Based Proximal Causal Inference. arXiv:2402.00335, 2024.

Proximal causal inference via negative controls

$$H_0 : \psi^{\mathsf{ATE}} = \mathbb{E}[Y^1 - Y^0] \neq 0$$



Linear proximal causal inference [1].

- $\mathbb{E}[Y \mid A, Z, C] = \beta_0 + \psi^{ATE} A + \beta_C^T C$
- $\mathbb{E}[W \mid A, Z, C] = \alpha_0 + \mathbf{M}^T C$

this implies…

- $\mathbb{E}[Y \mid A, Z] = \beta_0' + \psi^{\mathsf{ATE}} A + (\beta_w')^T \mathbb{E}[W \mid A, Z]$

We can run two stage-least squares to approximate $\mathbb{E}[W \mid A, Z]$ and estimate ψ^{ATE}.

[1] Liu et al. Regression-Based Proximal Causal Inference. arXiv:2402.00335, 2024.

Computing the proxy and true p-value via least squares

Outcomes $Y \in \mathbb{R}^n$

Computing the proxy and true p-value via least squares

Outcomes $Y \in \mathbb{R}^n$

Treatments: $\bar{A} = [\mathbf{1}, A] \in \mathbb{R}^{n \times 2}$

Computing the proxy and true p-value via least squares

Outcomes $Y \in \mathbb{R}^n$

Treatments: $\bar{A} = [\mathbf{1}, A] \in \mathbb{R}^{n \times 2}$

$\hat{\psi}^{OLS} = (\bar{A}^T \bar{A})^{-1} \bar{A}^T Y$

Computing the proxy and true p-value via least squares

Outcomes $Y \in \mathbb{R}^n$

Treatments: $\bar{A} = [\mathbf{1}, A] \in \mathbb{R}^{n \times 2}$

$$\hat{\psi}^{OLS} = (\bar{A}^T \bar{A})^{-1} \bar{A}^T Y$$

$$\hat{\sigma}^{OLS} = \sqrt{(\bar{A}^T \bar{A})^{-1}_{1,1} \cdot ||Y - \bar{A}^T \hat{\psi}^{OLS}||_2^2 / (n-2)}$$

Computing the proxy and true p-value via least squares

Outcomes $Y \in \mathbb{R}^n$

Treatments: $\bar{A} = [\mathbf{1}, A] \in \mathbb{R}^{n \times 2}$

$$\hat{\psi}^{OLS} = (\bar{A}^T \bar{A})^{-1} \bar{A}^T Y$$

$$\hat{\sigma}^{OLS} = \sqrt{(\bar{A}^T \bar{A})^{-1}_{1,1} \cdot ||\, Y - \bar{A}^T \hat{\psi}^{OLS} \,||_2^2 / (n-2)}$$

Proxy (OLS estimator)**:** $Q = 2F_{n-2}(\, - |\hat{\psi}^{OLS}| / \hat{\sigma}^{OLS})$

Computing the proxy and true p-value via least squares

Outcomes $Y \in \mathbb{R}^n$

Treatments: $\bar{A} = [\mathbf{1}, A] \in \mathbb{R}^{n \times 2}$

$$\hat{\psi}^{OLS} = (\bar{A}^T \bar{A})^{-1} \bar{A}^T Y$$

$$\hat{\sigma}^{OLS} = \sqrt{(\bar{A}^T \bar{A})^{-1}_{1,1} \cdot ||Y - \bar{A}^T \hat{\psi}^{OLS}||_2^2 / (n-2)}$$

Proxy (OLS estimator)**:** $Q = 2F_{n-2}(-|\hat{\psi}^{OLS}|/\hat{\sigma}^{OLS})$

$O(n)$ runtime —- very fast. (does not account for neg. controls)

Computing the proxy and true p-value via least squares

Outcomes $Y \in \mathbb{R}^n$

Treatments: $\bar{A} = [\mathbf{1}, A] \in \mathbb{R}^{n \times 2}$

$$\hat{\psi}^{OLS} = (\bar{A}^T \bar{A})^{-1} \bar{A}^T Y$$

$$\hat{\sigma}^{OLS} = \sqrt{(\bar{A}^T \bar{A})^{-1}_{1,1} \cdot ||Y - \bar{A}^T \hat{\psi}^{OLS}||_2^2 / (n-2)}$$

Proxy (OLS estimator)**:** $Q = 2F_{n-2}(-|\hat{\psi}^{OLS}|/\hat{\sigma}^{OLS})$

$O(n)$ runtime —- very fast. (does not account for neg. controls)

Treatments + neg. contr. exposures: $\bar{Z} = [\mathbf{1}, A, Z] \in \mathbb{R}^{n \times (d+2)}$

Computing the proxy and true p-value via least squares

Outcomes $Y \in \mathbb{R}^n$

Treatments: $\bar{A} = [\mathbf{1}, A] \in \mathbb{R}^{n \times 2}$

$$\hat{\psi}^{OLS} = (\bar{A}^T \bar{A})^{-1} \bar{A}^T Y$$

$$\hat{\sigma}^{OLS} = \sqrt{(\bar{A}^T \bar{A})^{-1}_{1,1} \cdot ||Y - \bar{A}^T \hat{\psi}^{OLS}||_2^2 / (n-2)}$$

Proxy (OLS estimator)**:** $Q = 2F_{n-2}(-|\hat{\psi}^{OLS}|/\hat{\sigma}^{OLS})$

$O(n)$ runtime —- very fast. (does not account for neg. controls)

Treatments + neg. contr. exposures: $\bar{Z} = [\mathbf{1}, A, Z] \in \mathbb{R}^{n \times (d+2)}$

First stage: $\hat{W} = (\bar{Z}^T \bar{Z})^{-1} \bar{Z}^T W,$

Computing the proxy and true p-value via least squares

Outcomes $Y \in \mathbb{R}^n$

Treatments: $\bar{A} = [\mathbf{1}, A] \in \mathbb{R}^{n \times 2}$

$$\hat{\psi}^{OLS} = (\bar{A}^T \bar{A})^{-1} \bar{A}^T Y$$

$$\hat{\sigma}^{OLS} = \sqrt{(\bar{A}^T \bar{A})^{-1}_{1,1} \cdot ||Y - \bar{A}^T \hat{\psi}^{OLS}||_2^2 / (n-2)}$$

Proxy (OLS estimator)**:** $Q = 2F_{n-2}(-|\hat{\psi}^{OLS}|/\hat{\sigma}^{OLS})$

$O(n)$ runtime —- very fast. (does not account for neg. controls)

Treatments + neg. contr. exposures: $\bar{Z} = [\mathbf{1}, A, Z] \in \mathbb{R}^{n \times (d+2)}$

First stage: $\hat{W} = (\bar{Z}^T \bar{Z})^{-1} \bar{Z}^T W,$

Computing the proxy and true p-value via least squares

Outcomes $Y \in \mathbb{R}^n$

Treatments: $\bar{A} = [\mathbf{1}, A] \in \mathbb{R}^{n \times 2}$

$$\hat{\psi}^{OLS} = (\bar{A}^T \bar{A})^{-1} \bar{A}^T Y$$

$$\hat{\sigma}^{OLS} = \sqrt{(\bar{A}^T \bar{A})^{-1}_{1,1} \cdot ||Y - \bar{A}^T \hat{\psi}^{OLS}||_2^2 / (n-2)}$$

Proxy (OLS estimator)**:** $Q = 2F_{n-2}(-|\hat{\psi}^{OLS}|/\hat{\sigma}^{OLS})$

$O(n)$ runtime —- very fast. (does not account for neg. controls)

Treatments + neg. contr. exposures: $\bar{Z} = [\mathbf{1}, A, Z] \in \mathbb{R}^{n \times (d+2)}$

First stage: $\hat{W} = (\bar{Z}^T \bar{Z})^{-1} \bar{Z}^T W,$

Treatments + est. neg. contr. outcomes: $\bar{W} = [1, A, \hat{W}] \in \mathbb{R}^{n \times (d+2)}$

Computing the proxy and true p-value via least squares

Outcomes $Y \in \mathbb{R}^n$

Treatments: $\bar{A} = [\mathbf{1}, A] \in \mathbb{R}^{n \times 2}$

$$\hat{\psi}^{OLS} = (\bar{A}^T \bar{A})^{-1} \bar{A}^T Y$$

$$\hat{\sigma}^{OLS} = \sqrt{(\bar{A}^T \bar{A})^{-1}_{1,1} \cdot ||Y - \bar{A}^T \hat{\psi}^{OLS}||_2^2 / (n-2)}$$

Proxy (OLS estimator)**:** $Q = 2F_{n-2}(-|\hat{\psi}^{OLS}|/\hat{\sigma}^{OLS})$

$O(n)$ runtime —- very fast. (does not account for neg. controls)

Treatments + neg. contr. exposures: $\bar{Z} = [\mathbf{1}, A, Z] \in \mathbb{R}^{n \times (d+2)}$

First stage: $\hat{W} = (\bar{Z}^T \bar{Z})^{-1} \bar{Z}^T W,$

Treatments + est. neg. contr. outcomes: $\bar{W} = [1, A, \hat{W}] \in \mathbb{R}^{n \times (d+2)}$

Second stage: $\hat{\psi}^{2SLS} = (\bar{W}^T \bar{W}) \bar{W}^T Y$

Computing the proxy and true p-value via least squares

Outcomes $Y \in \mathbb{R}^n$

Treatments: $\bar{A} = [\mathbf{1}, A] \in \mathbb{R}^{n \times 2}$

$$\hat{\psi}^{OLS} = (\bar{A}^T \bar{A})^{-1} \bar{A}^T Y$$

$$\hat{\sigma}^{OLS} = \sqrt{(\bar{A}^T \bar{A})^{-1}_{1,1} \cdot ||Y - \bar{A}^T \hat{\psi}^{OLS}||_2^2 / (n-2)}$$

Proxy (OLS estimator)**:** $Q = 2F_{n-2}(-|\hat{\psi}^{OLS}|/\hat{\sigma}^{OLS})$

$O(n)$ runtime —- very fast. (does not account for neg. controls)

Treatments + neg. contr. exposures: $\bar{Z} = [\mathbf{1}, A, Z] \in \mathbb{R}^{n \times (d+2)}$

First stage: $\hat{W} = (\bar{Z}^T \bar{Z})^{-1} \bar{Z}^T W,$

Treatments + est. neg. contr. outcomes: $\bar{W} = [1, A, \hat{W}] \in \mathbb{R}^{n \times (d+2)}$

Second stage: $\hat{\psi}^{2SLS} = (\bar{W}^T \bar{W}) \bar{W}^T Y$

Computing the proxy and true p-value via least squares

Outcomes $Y \in \mathbb{R}^n$

Treatments: $\bar{A} = [\mathbf{1}, A] \in \mathbb{R}^{n \times 2}$

$\hat{\psi}^{OLS} = (\bar{A}^T \bar{A})^{-1} \bar{A}^T Y$

$\hat{\sigma}^{OLS} = \sqrt{(\bar{A}^T \bar{A})^{-1}_{1,1} \cdot ||Y - \bar{A}^T \hat{\psi}^{OLS}||_2^2 / (n-2)}$

Proxy (OLS estimator)**:** $Q = 2F_{n-2}(-|\hat{\psi}^{OLS}|/\hat{\sigma}^{OLS})$

$O(n)$ runtime —- very fast. (does not account for neg. controls)

Treatments + neg. contr. exposures: $\bar{Z} = [\mathbf{1}, A, Z] \in \mathbb{R}^{n \times (d+2)}$

First stage: $\hat{W} = (\bar{Z}^T \bar{Z})^{-1} \bar{Z}^T W,$

Treatments + est. neg. contr. outcomes: $\bar{W} = [1, A, \hat{W}] \in \mathbb{R}^{n \times (d+2)}$

Second stage: $\hat{\psi}^{2SLS} = (\bar{W}^T \bar{W}) \bar{W}^T Y$

$\hat{\sigma}^{2SLS} = \sqrt{\hat{A}^{-1} \hat{B} (\hat{A}^{-1})^T}$ where $\hat{A}^{-1}, \hat{B} \in \mathbb{R}^{O(d^2) \times O(d^2)}$

Computing the proxy and true p-value via least squares

Outcomes $Y \in \mathbb{R}^n$

Treatments: $\bar{A} = [\mathbf{1}, A] \in \mathbb{R}^{n \times 2}$

$\hat{\psi}^{OLS} = (\bar{A}^T \bar{A})^{-1} \bar{A}^T Y$

$\hat{\sigma}^{OLS} = \sqrt{(\bar{A}^T \bar{A})^{-1}_{1,1} \cdot ||Y - \bar{A}^T \hat{\psi}^{OLS}||_2^2 / (n-2)}$

Proxy (OLS estimator)**:** $Q = 2F_{n-2}(-|\hat{\psi}^{OLS}|/\hat{\sigma}^{OLS})$

$O(n)$ runtime —- very fast. (does not account for neg. controls)

Treatments + neg. contr. exposures: $\bar{Z} = [\mathbf{1}, A, Z] \in \mathbb{R}^{n \times (d+2)}$

First stage: $\hat{W} = (\bar{Z}^T \bar{Z})^{-1} \bar{Z}^T W,$

Treatments + est. neg. contr. outcomes: $\bar{W} = [1, A, \hat{W}] \in \mathbb{R}^{n \times (d+2)}$

Second stage: $\hat{\psi}^{2SLS} = (\bar{W}^T \bar{W}) \bar{W}^T Y$

$\hat{\sigma}^{2SLS} = \sqrt{\hat{A}^{-1} \hat{B} (\hat{A}^{-1})^T}$ where $\hat{A}^{-1}, \hat{B} \in \mathbb{R}^{O(d^2) \times O(d^2)}$

True (2SLS estimator)**:** $P = 2\Phi(-|\hat{\psi}^{2SLS}|/\hat{\sigma}^{2SLS})$

Computing the proxy and true p-value via least squares

Outcomes $Y \in \mathbb{R}^n$

Treatments: $\bar{A} = [\mathbf{1}, A] \in \mathbb{R}^{n \times 2}$

$\hat{\psi}^{OLS} = (\bar{A}^T \bar{A})^{-1} \bar{A}^T Y$

$\hat{\sigma}^{OLS} = \sqrt{(\bar{A}^T \bar{A})^{-1}_{1,1} \cdot ||Y - \bar{A}^T \hat{\psi}^{OLS}||_2^2 / (n-2)}$

Proxy (OLS estimator)**:** $Q = 2F_{n-2}(-|\hat{\psi}^{OLS}|/\hat{\sigma}^{OLS})$

$O(n)$ runtime —- very fast. (does not account for neg. controls)

Treatments + neg. contr. exposures: $\bar{Z} = [\mathbf{1}, A, Z] \in \mathbb{R}^{n \times (d+2)}$

First stage: $\hat{W} = (\bar{Z}^T \bar{Z})^{-1} \bar{Z}^T W,$

Treatments + est. neg. contr. outcomes: $\bar{W} = [1, A, \hat{W}] \in \mathbb{R}^{n \times (d+2)}$

Second stage: $\hat{\psi}^{2SLS} = (\bar{W}^T \bar{W}) \bar{W}^T Y$

$\hat{\sigma}^{2SLS} = \sqrt{\hat{A}^{-1} \hat{B} (\hat{A}^{-1})^T}$ where $\hat{A}^{-1}, \hat{B} \in \mathbb{R}^{O(d^2) \times O(d^2)}$

True (2SLS estimator)**:** $P = 2\Phi(-|\hat{\psi}^{2SLS}|/\hat{\sigma}^{2SLS})$

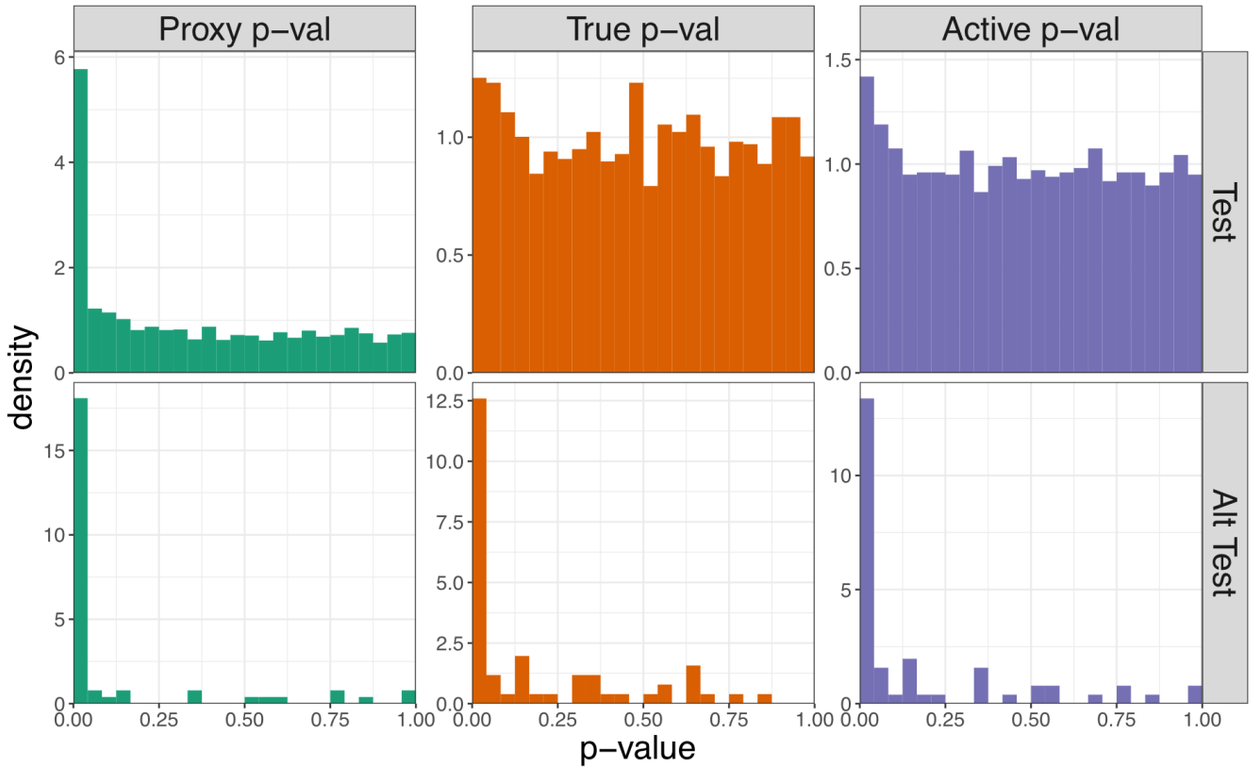
$O(nd^4)$ to compute $\hat{A}^{-1}$ and B.

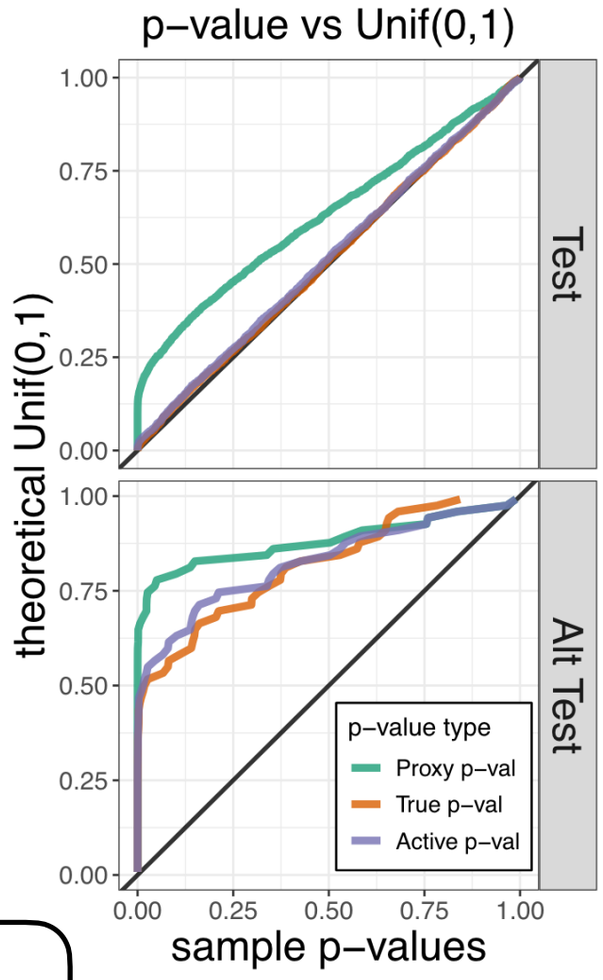
$O(d^6)$ to compute $\hat{A}^{-1} \hat{B} (\hat{A}^{-1})^T$.

Total complexity:

$O(nd^4 + d^6)$

Experimental results on scCRIPSR data

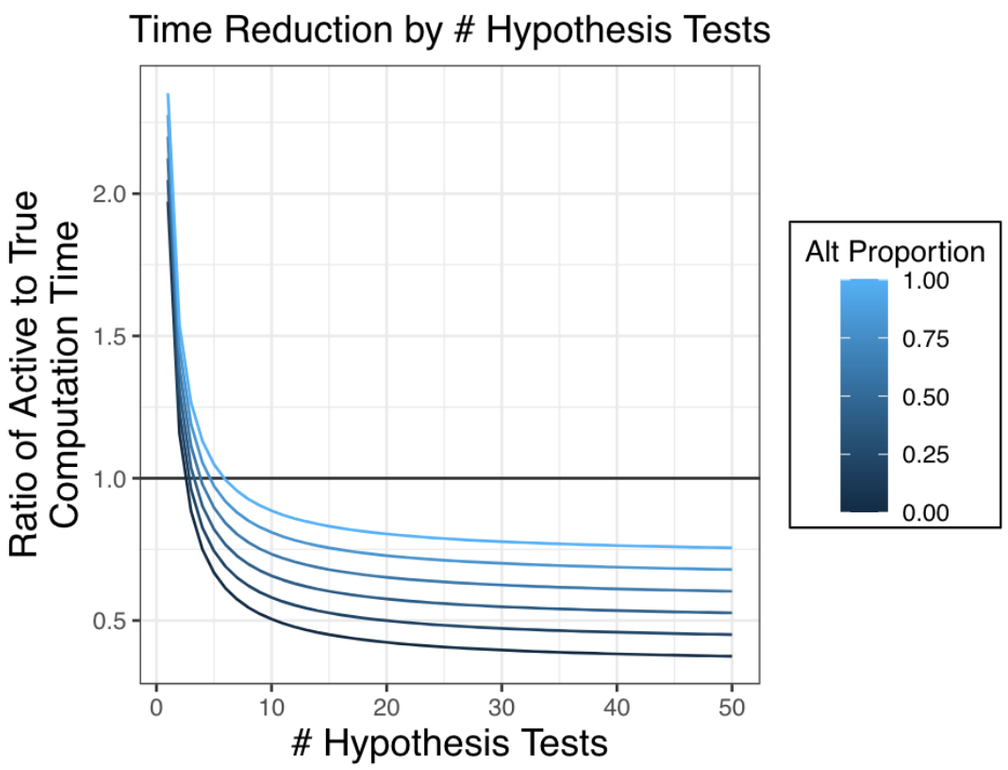


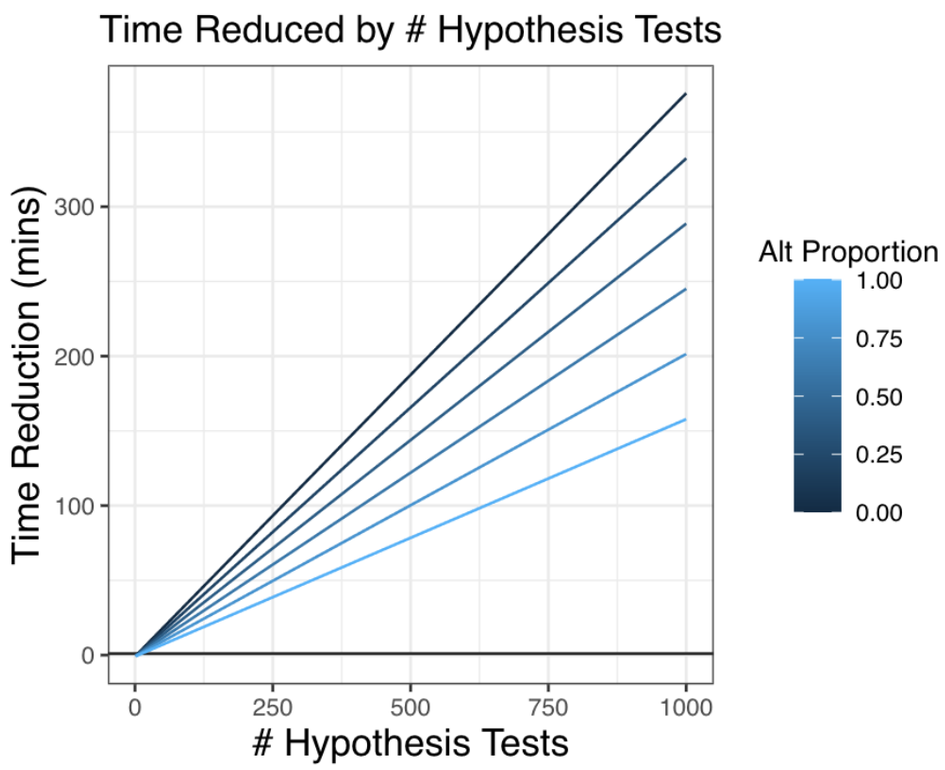


2000 genes pre-filtered [1] from original cancer marker scCRISPR dataset [2]. multimodal single-cell screens.

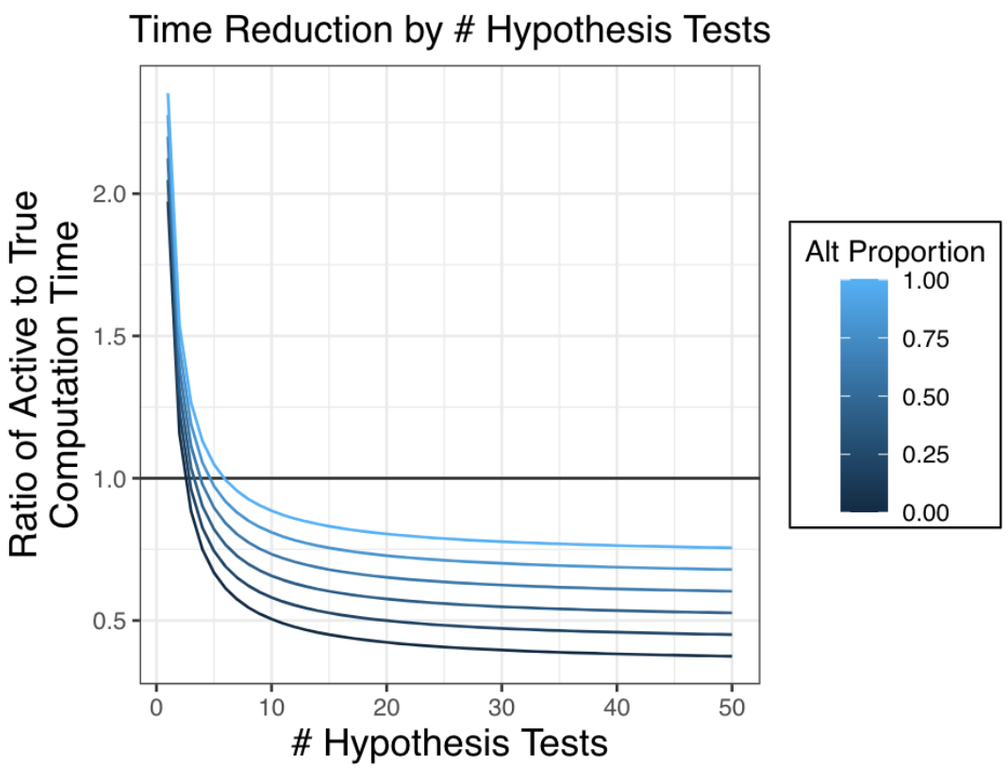
[1] Papalexi et al. Characterizing the molecular regulation of inhibitory immune checkpoints with
multimodal single-cell screens. *Nature Genetics,* 2021.
[2] Townes et al. Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model *Genome Biology,* 2-10.

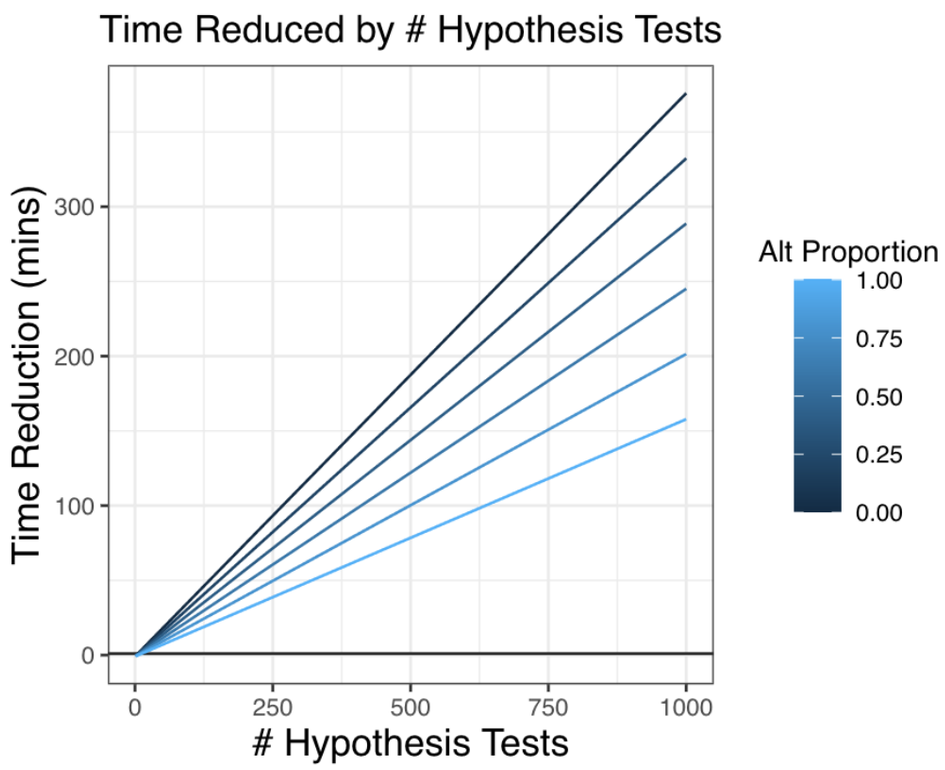
Experimental results on computation time





Experimental results on computation time





Significant reduction in computation time, while maintaining power.

Active BH for FDR control

Active BH for FDR control

Multiple testing for K hypotheses —- $I_0 \subseteq [K]$ are the true nulls. Output discovery set $R \subseteq [K]$ s.t. *false discovery rate (FDR)*

$$\text{FDR} = \mathbb{E}\left[\frac{|I_0 \cap R|}{|R| \vee 1}\right] \leq \alpha \text{ for fixed } \alpha \in [0,1]$$

Active BH for FDR control

Multiple testing for K hypotheses —- $I_0 \subseteq [K]$ are the true nulls. Output discovery set $R \subseteq [K]$ s.t. *false discovery rate (FDR)*

$$\text{FDR} = \mathbb{E}\left[\frac{|I_0 \cap R|}{|R| \vee 1}\right] \leq \alpha \text{ for fixed } \alpha \in [0,1]$$

Active BH procedure:

Access to $(Q_1, \ldots, Q_K)$ proxy p-values, and $(P_1, \ldots, P_K)$ are independent true p-values.

Apply Benjamini-Hochberg (BH) procedure to $(\tilde{P}_1, \ldots, \tilde{P}_K)$ (any active p-values), i.e.,

$$k^* = \max\left\{ k \in [K] : \sum_{i=1}^{K} \mathbf{1}\{\tilde{P}_i \leq \alpha k/K\} \geq k \right\} \text{ and } R = \{k \in [K] : P_k \leq \alpha k^*/K\}$$

Active BH for FDR control

Multiple testing for K hypotheses —- $I_0 \subseteq [K]$ are the true nulls. Output discovery set $R \subseteq [K]$ s.t. *false discovery rate (FDR)*

$$\text{FDR} = \mathbb{E}\left[\frac{|I_0 \cap R|}{|R| \vee 1}\right] \leq \alpha \text{ for fixed } \alpha \in [0,1]$$

Active BH procedure:

Access to $(Q_1, \ldots, Q_K)$ proxy p-values, and $(P_1, \ldots, P_K)$ are independent true p-values.

Apply Benjamini-Hochberg (BH) procedure to $(\tilde{P}_1, \ldots, \tilde{P}_K)$ (any active p-values), i.e.,

$$k^* = \max\left\{k \in [K] : \sum_{i=1}^{K} \mathbf{1}\{\tilde{P}_i \leq \alpha k/K\} \geq k\right\} \text{ and } R = \{k \in [K] : P_k \leq \alpha k^*/K\}$$

Theorem (ours): If $(P_1, \ldots, P_K)$ are independent, $(Q_1, \ldots, Q_K)$ are arbitrarily dependent, and (P_i, Q_i) satisfies active p-value dependence requirement, then FDR $\leq \alpha(1 + \log(1/\alpha))$.

Conclusion

Conclusion

- Active hypothesis testing framework + application for proximal causal inference in scCRISPR screening

Conclusion

- Active hypothesis testing framework + application for proximal causal inference in scCRISPR screening
- Extensions

Conclusion

- Active hypothesis testing framework + application for proximal causal inference in scCRISPR screening
- Extensions
  - E-value versions, i.e., active e-value and active e-BH

Conclusion

- Active hypothesis testing framework + application for proximal causal inference in scCRISPR screening
- Extensions
  - E-value versions, i.e., active e-value and active e-BH
  - Interactive + multilevel computation of proxies.

Conclusion

- Active hypothesis testing framework + application for proximal causal inference in scCRISPR screening
- Extensions
  - E-value versions, i.e., active e-value and active e-BH
  - Interactive + multilevel computation of proxies.
  - Joint density estimation of proxy + true p-value

Conclusion

- Active hypothesis testing framework + application for proximal causal inference in scCRISPR screening
- Extensions
  - E-value versions, i.e., active e-value and active e-BH
  - Interactive + multilevel computation of proxies.
  - Joint density estimation of proxy + true p-value

Thanks!

“Active multiple testing with proxy p-values and e-values”
arXiv:2502.05715