Bayesian optimal experimental design for inferring causal structure

Jeff Miller

Joint work with Michele Zemplenyi

Harvard T.H. Chan School of Public Health
Department of Biostatistics

ISBA World Meeting || June 30, 2022 ||
Session on "Bayesian experimental design for causal inference"

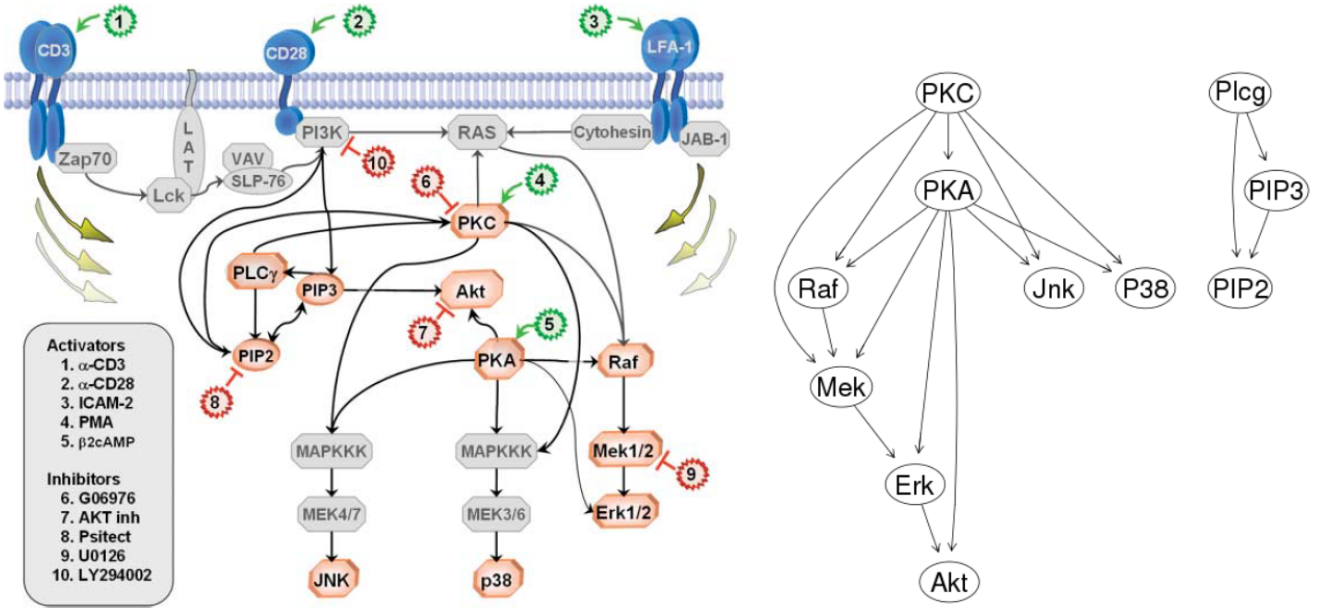
Preprint: https://arxiv.org/abs/2103.15229

Outline

Outline

Motivation

- Inferring causal structure is key to understanding many systems.
- Many methods exist for inferring causality from observational data.
- However, observational data only determines the structure of a causal network up to the Markov equivalence class.
- Without strong assumptions, interventional data are needed to fully resolve networks.



Motivation

- Different intervention experiments yield different amounts of information.
- Since experiments are often expensive and time-consuming, it is advantageous to select interventions that provide the most information.
- Optimal experimental design, a.k.a. active learning, attempts to optimize this experiment selection process.

Motivation

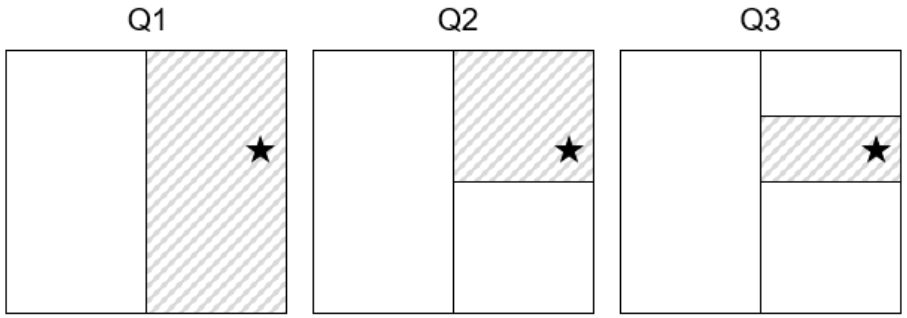
- From the Bayesian perspective, a brute-force approach would be as follows: for each candidate experiment,
  1. generate hypothetical datasets from the posterior predictive,
  2. perform posterior inference on each dataset, and
  3. compute a function of the posterior summarizing the information gain.
- Averaging over many hypothetical datasets would yield an estimate of the posterior expected amount of information gain for each candidate experiment.
- However, this brute-force approach would involve an inordinate amount of computation.

This talk

- We develop a novel Bayesian experiment selection criterion that is principled and computationally tractable.
- Roughly, we consider the asymptotic information gain that each experiment would yield in the limit of infinitely many replicates, as a proxy for the expected gain from finitely many replicates.
- It turns out that the reduction in entropy can be easily computed using samples from the current posterior, without generating or performing inference on any hypothetical datasets.
- This leads to a vast reduction in the computational burden required to select experiments.

Outline

Intuition: Twenty questions game



- Two players: Questioner and Answerer.
- Answerer thinks of an object. Questioner has a prior over objects that the Answerer might select.
- A question such as, "Is the object living?" partitions the objects into two parts: living and non-living.
- Given the answer, the questioner can eliminate objects in one part, and update their beliefs.
- Questioner chooses questions to focus the posterior as fast as possible.

General setup

- More generally, suppose:
  - θ is a parameter of interest,
  - ν is a nuisance parameter,
  - $\pi(\theta, \nu)$ is the prior,
  - $f(\theta)$ is the answer to a research question f (for example, is a certain hypothesis true), and
  - $X_{1:N} = (X_1, \ldots, X_N)$ are data from N replicates of an experiment performed to obtain information about $f(\theta)$.

- By analogy with Twenty Questions, θ is the object to be inferred, f is the question, and $X_{1:N}$ is a noisy "answer".

Conditions

- Suppose:

$$(\theta, \nu) \sim \pi$$
$$X_1, \ldots, X_N | \theta, \nu \sim P_{\theta,\nu} \text{ i.i.d.}$$

and $f(\theta)$ is a function of θ with the following three properties:

1. $\theta \perp\!\!\!\perp X_{1:N} \mid f(\theta)$,
2. $f(\theta)$ is identifiable, in the sense that there is a function g such that $g(P_{\theta,\nu}) = f(\theta)$ almost surely, and
3. $f(\theta)$ can only take one of finitely many values.

- Condition 1 is that if we know the true answer to question f, then the experiment provides no additional information about θ.
- Condition 2 is that the answer $f(\theta)$ is uniquely determined by the distribution of X_n.

General criterion

- We seek experiments that minimize the posterior entropy $H(\theta \mid X_{1:N})$.

- **Theorem**: *Under the conditions above,*

$$H(\theta \mid X_{1:N}) \xrightarrow[N \to \infty]{} H(\theta) - H(f(\theta)).$$

- The information gain, in terms of reduction of entropy, is

$$H(\theta) - H(\theta \mid X_{1:N}).$$

- Thus, when N is large, the information gain is approximately $H(f(\theta))$, the entropy of the answer $f(\theta)$ *under the prior*.

- Thus, to approximate the information to be gained by a particular question f, we need only work with the prior — not the posterior $\theta|X_{1:N}$ for a yet unobserved dataset $X_{1:N}$.

Sequentially selecting experiments

- Now, suppose that instead of $\pi(\theta, \nu)$ being the prior, $\pi(\theta, \nu)$ is the current posterior given all the data from any previous experiments.

- For each candidate experiment e, let $X_1^e, \ldots, X_N^e | \theta, \nu \sim P_{\theta,\nu}^e$ i.i.d. be hypothetical random data from N replicates of experiment e.

- Suppose $f_e(\theta)$ satisfies our assumed conditions above.

- Our proposed method of choosing the next experiment e is:
  1. Generate samples $\theta_1, \ldots, \theta_T$ from the current posterior π.
  2. Compute $\hat{p}_e(y) = \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\big(f_e(\theta_t) = y\big)$ for each candidate experiment e.
  3. Select the experiment e with the largest value of the entropy

$$\widehat{H}(f_e(\theta)) := -\sum_y \hat{p}_e(y) \log \hat{p}_e(y).$$

Features of the proposed approach

- All it requires is posterior samples given the current data — no need to simulate or do inference on hypothetical future data.
- While the justification is asymptotic, our criterion accounts for finite sample uncertainty in a natural way.
  - The posterior π quantifies uncertainty in θ based on finitely many previous experiments and finitely many replicates of each experiment.
- A further advantage is that $f_e(\theta)$ often takes a small number of values, and thus, $H(f_e(\theta))$ is often much easier to approximate than $H(\theta \mid X^e_{1:N})$ or even $H(\theta)$.

Outline

Selecting experiments for causal networks

- Causal network models specify the joint distribution of the data when variables are manipulated (or unmanipulated).

- It is common to specify causal networks via:
  1. a directed acyclic graph G and
  2. the conditional probability distribution (CPD) of each node given the values of its parent nodes.

- In general, an intervention partitions the set of graphs into equivalence classes such that
  - the graphs in each class are indistinguishable with respect to this intervention (corresponding to our Condition 1), and
  - graphs in different classes are distinguishable (corresponding to Condition 2).

- Thus, in our notation, θ is the graph G, and the "question" f corresponds to this partition.

Selecting experiments for causal networks

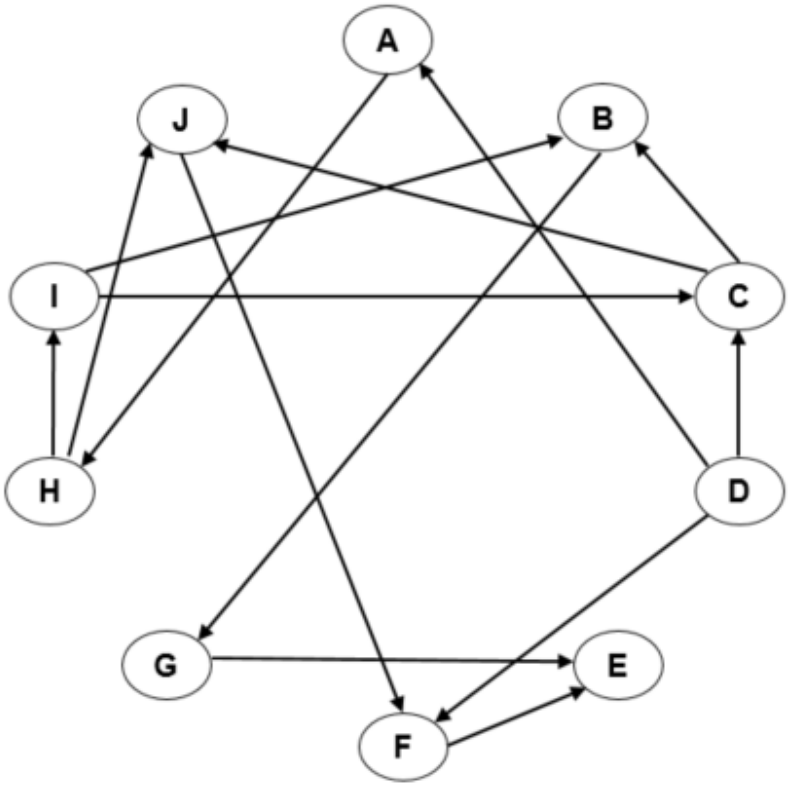
- The Markov equivalence class (MEC) is a natural choice of function f to use in our criterion.
- However, it may be possible to obtain good performance using other partition schemes as well.
- To this end, we considered the following array of partition schemes:
  1. MEC: $f_e(G) =$ Markov equivalence class when intervening on node e,
  2. Child Set (CS): $f_e(G) =$ set of children of node e,
  3. Descendant Set (DS): $f_e(G) =$ set of descendants of node e,
  4. Parent Set (PS): $f_e(G) =$ set of parents of node e.

Outline

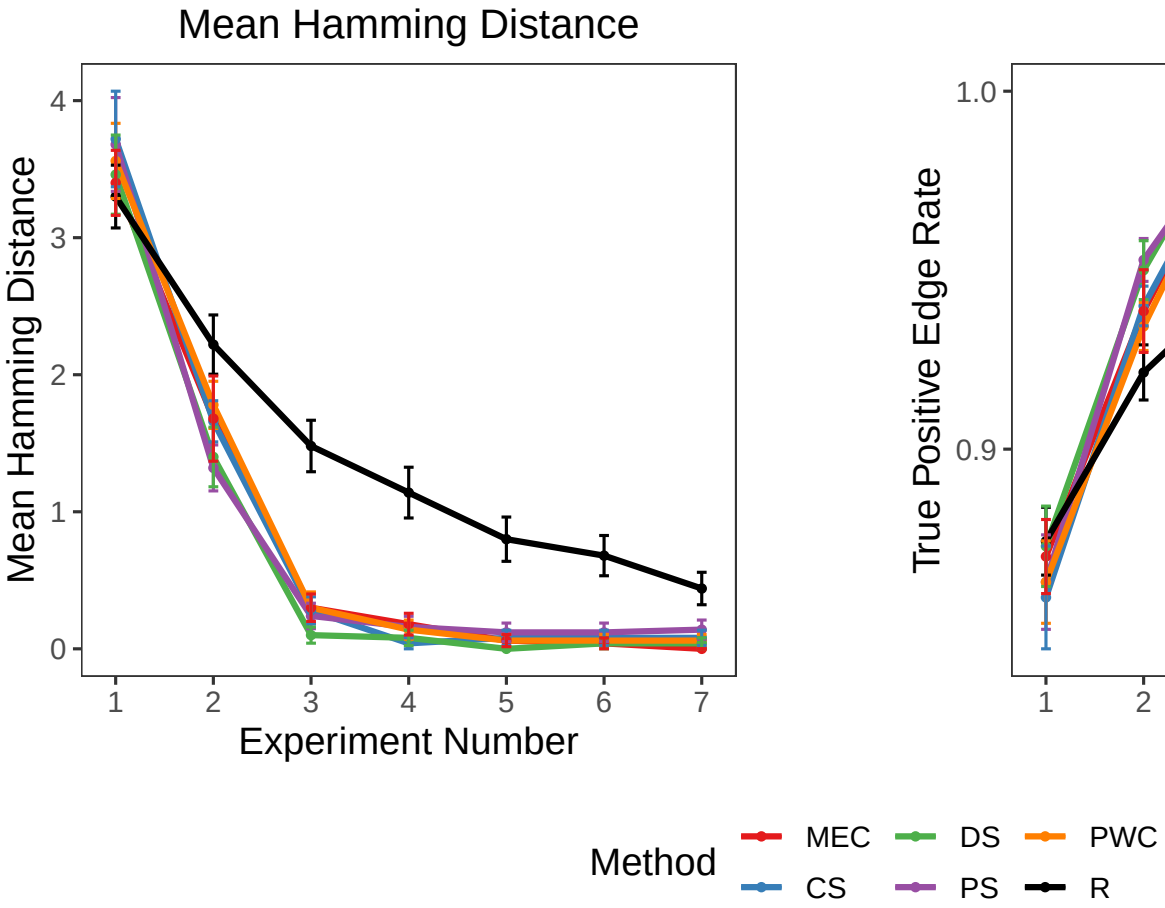
Simulations: Comparing partition schemes

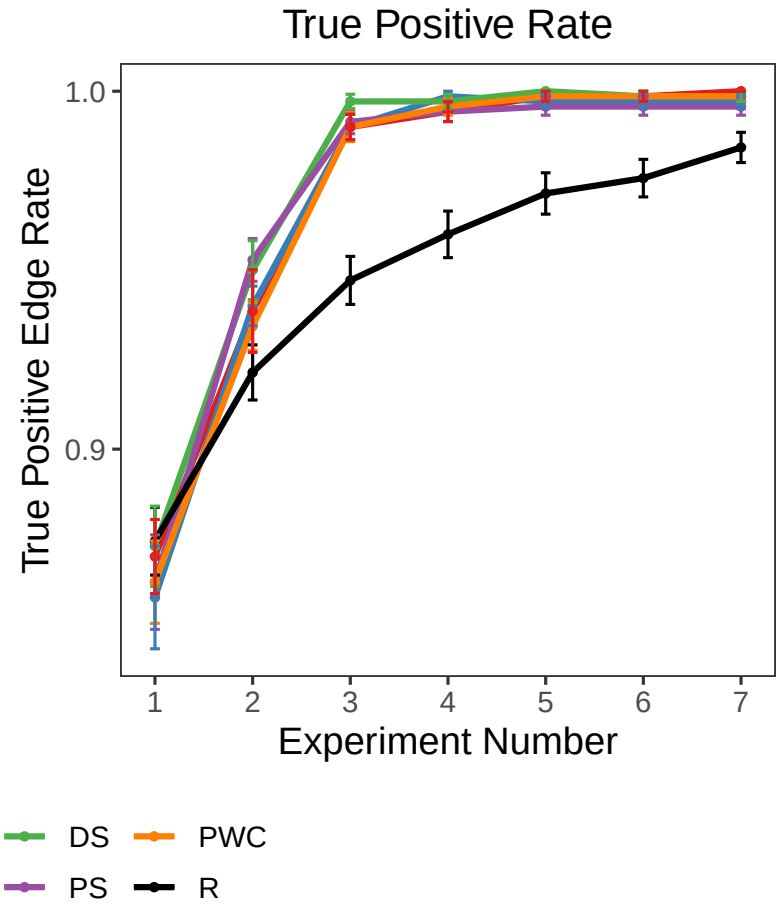
- First, we compare various partition schemes that could be used in our method.

- We simulate binary data using the following 10-node network as ground truth:



Simulations: Comparing partition schemes

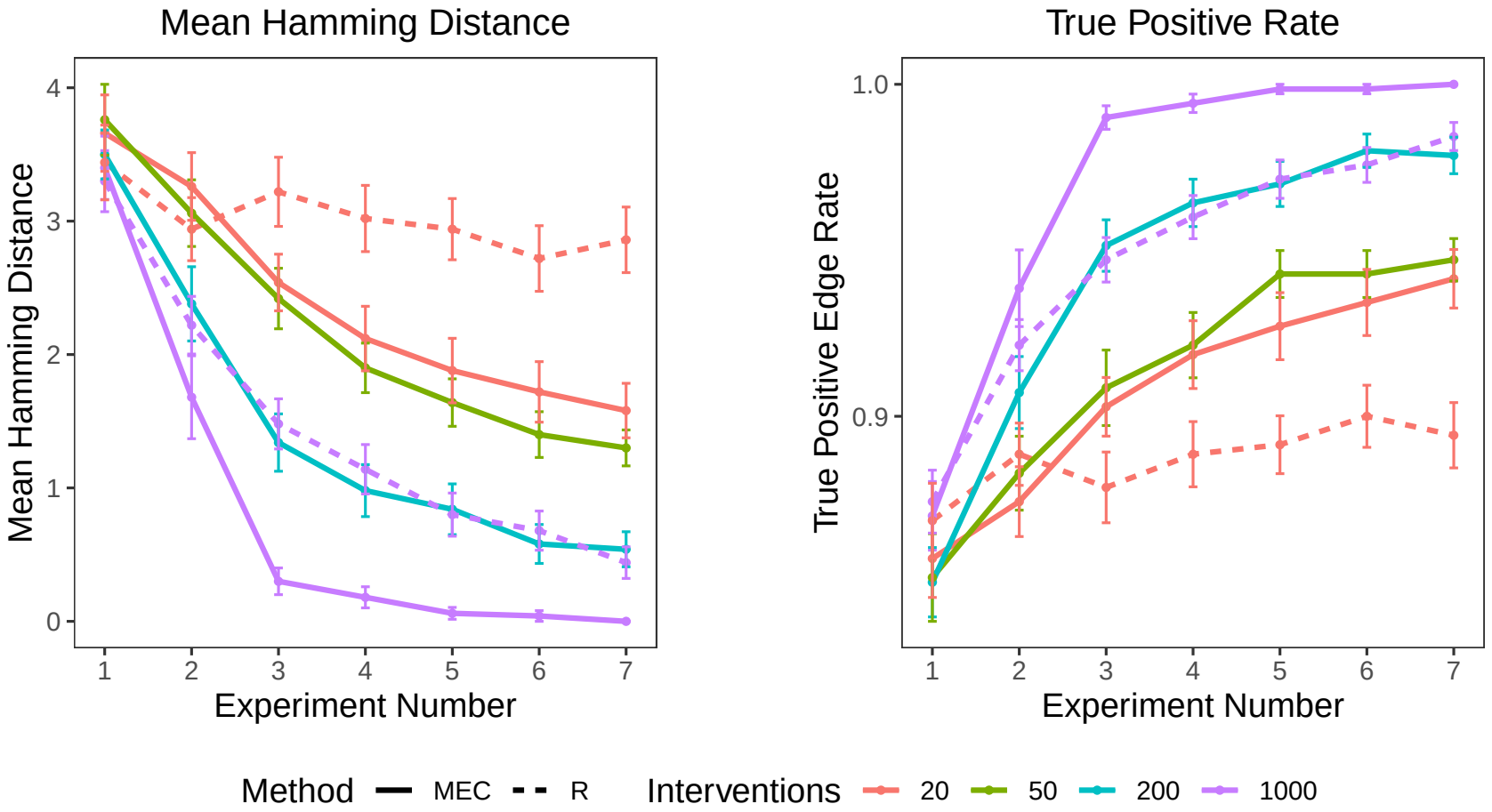




- Settings: $n_{sim} = 50$ simuation runs, $n_{exp} = 7$ interventional experiments per run, $n_{obs} = 1000$ observational data points, $n_{intv} = 1000$ data points for each interventional experiment.
- "R" = randomly selecting a node to intervene on.

Simulations: Effect of sample size

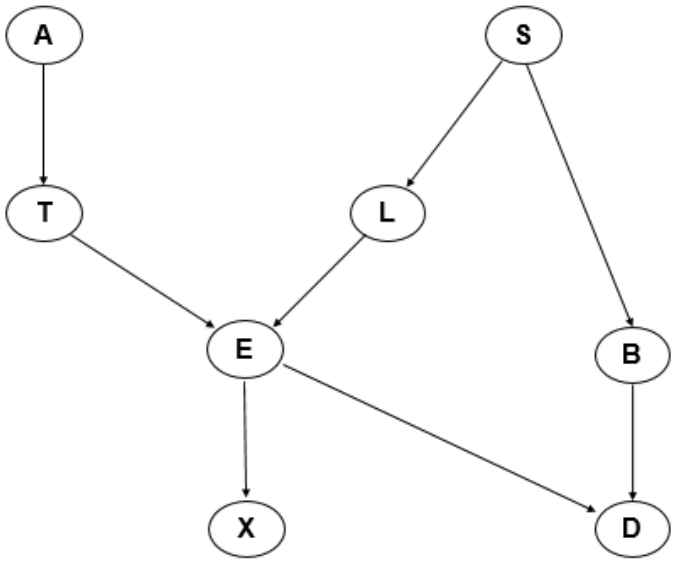
- We see the same pattern for smaller numbers of data points as well.
- Settings: Same as previous, but with $n_{intv} \in \{20, 50, 200, 1000\}$.

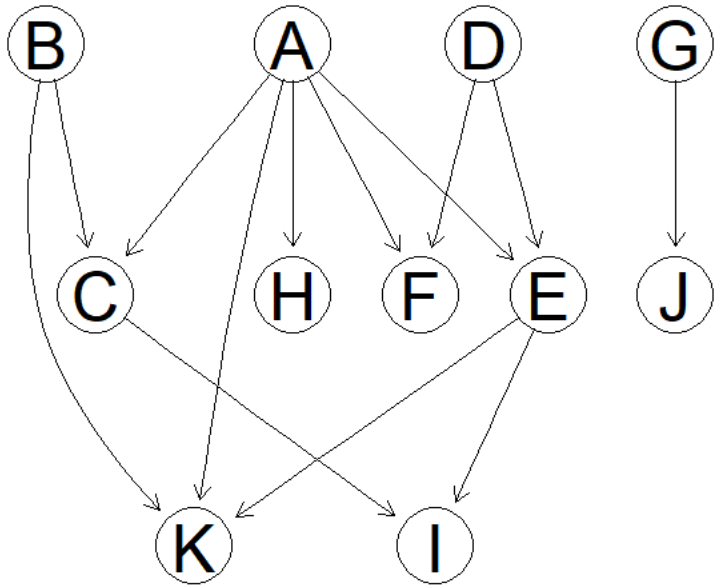


Simulations: Comparing with other methods

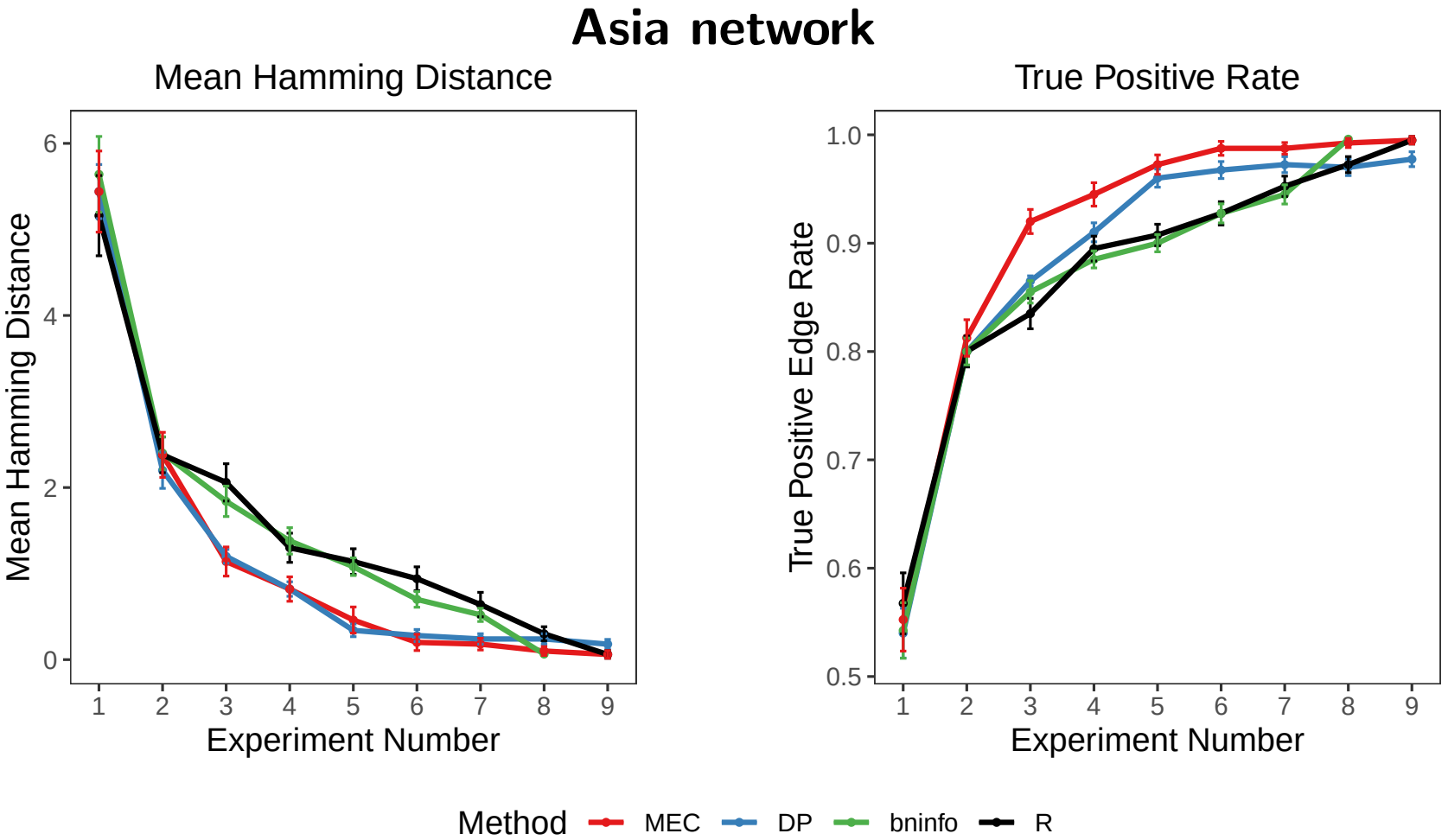
- We compared our algorithm to several other methods:
  - "bninfo" method of Ness et al. (2018)
  - dynamic programming (DP) method of Li and Leong (2009)
  - greedy equivalence search (GES) method of Chickering (2002)
  - greedy interventional equivalence search (GIES) method of Hauser and Bühlmann (2012)

- We considered the 8-node "Asia" binary network and an 11-node Gaussian network:



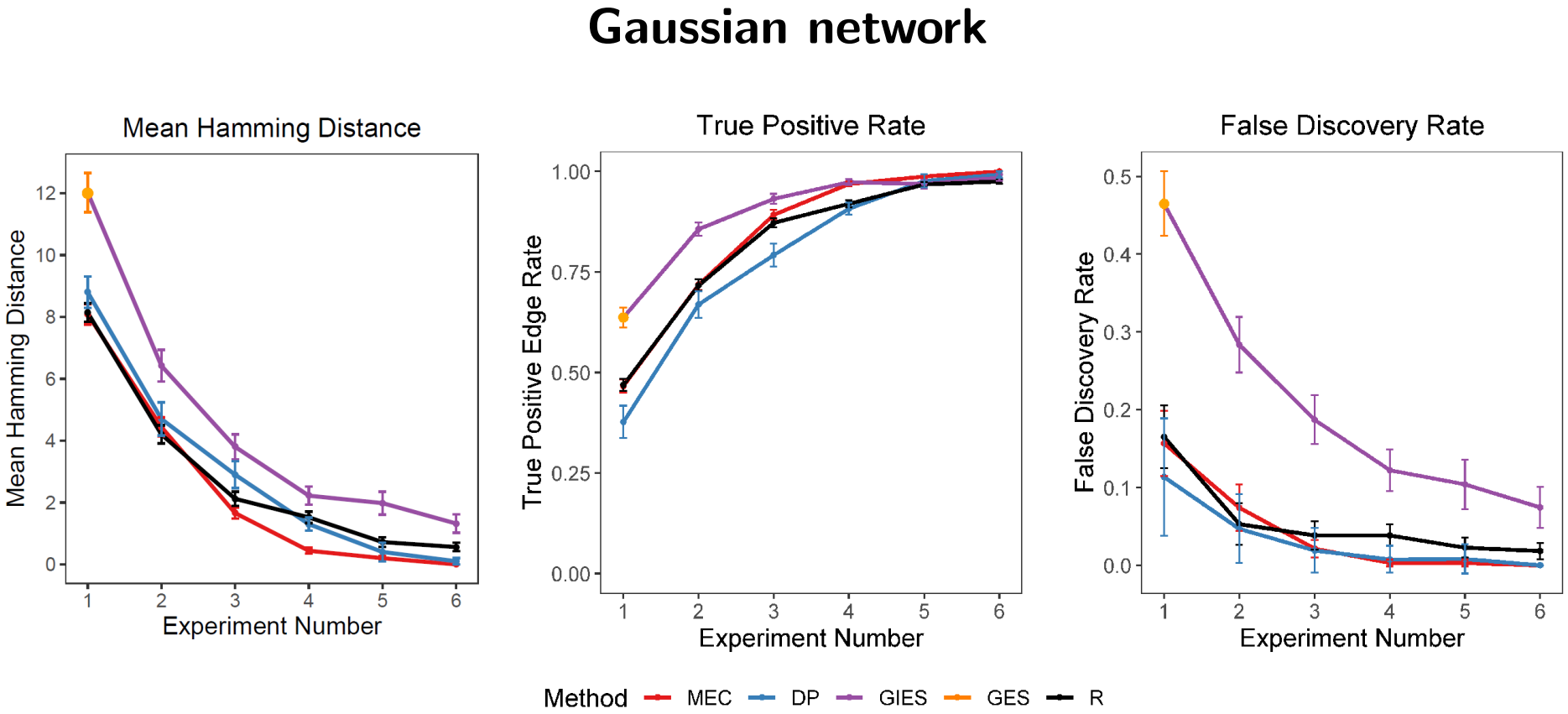


Simulations: Comparing with other methods



- Settings: $n_{sim} = 50$ simuation runs, $n_{exp} = 9$ interventional experiments per run, $n_{obs} = 300$ observational data points, $n_{intv} = 300$ data points for each interventional experiment.

Simulations: Comparing with other methods



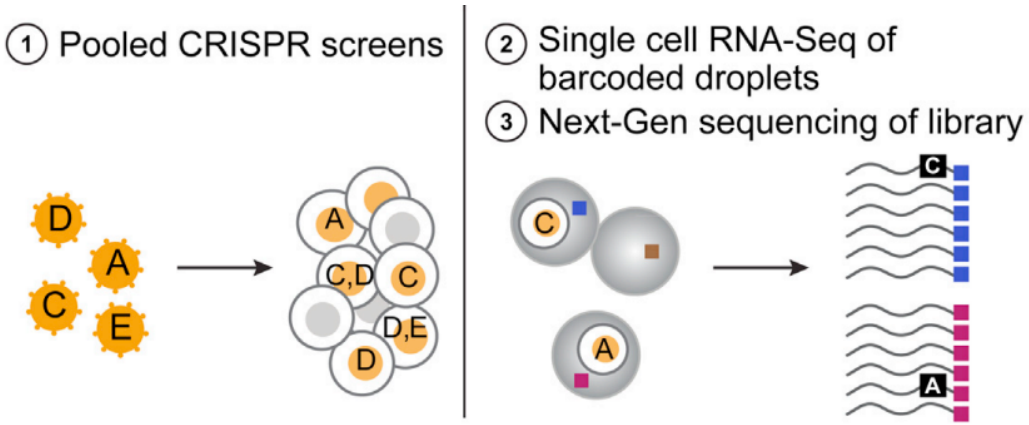
- Settings: $n_{sim} = 50$ simuation runs, $n_{exp} = 6$ interventional experiments per run, $n_{obs} = 50$ observational data points, $n_{intv} = 50$ data points for each interventional experiment.

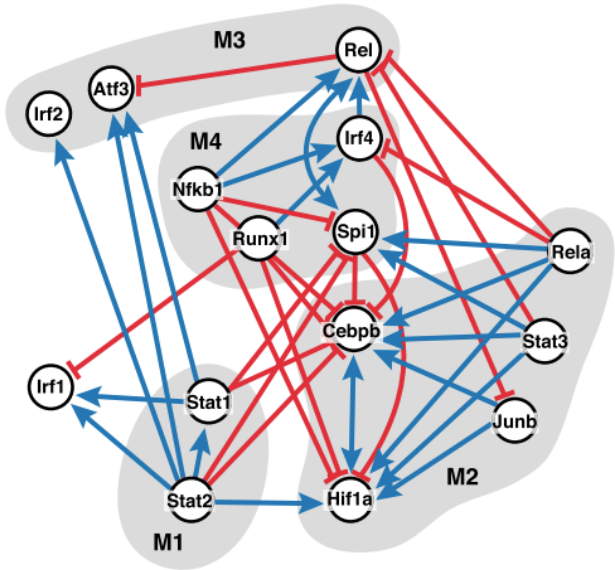
Outline

Application: Perturb-seq gene expression dataset

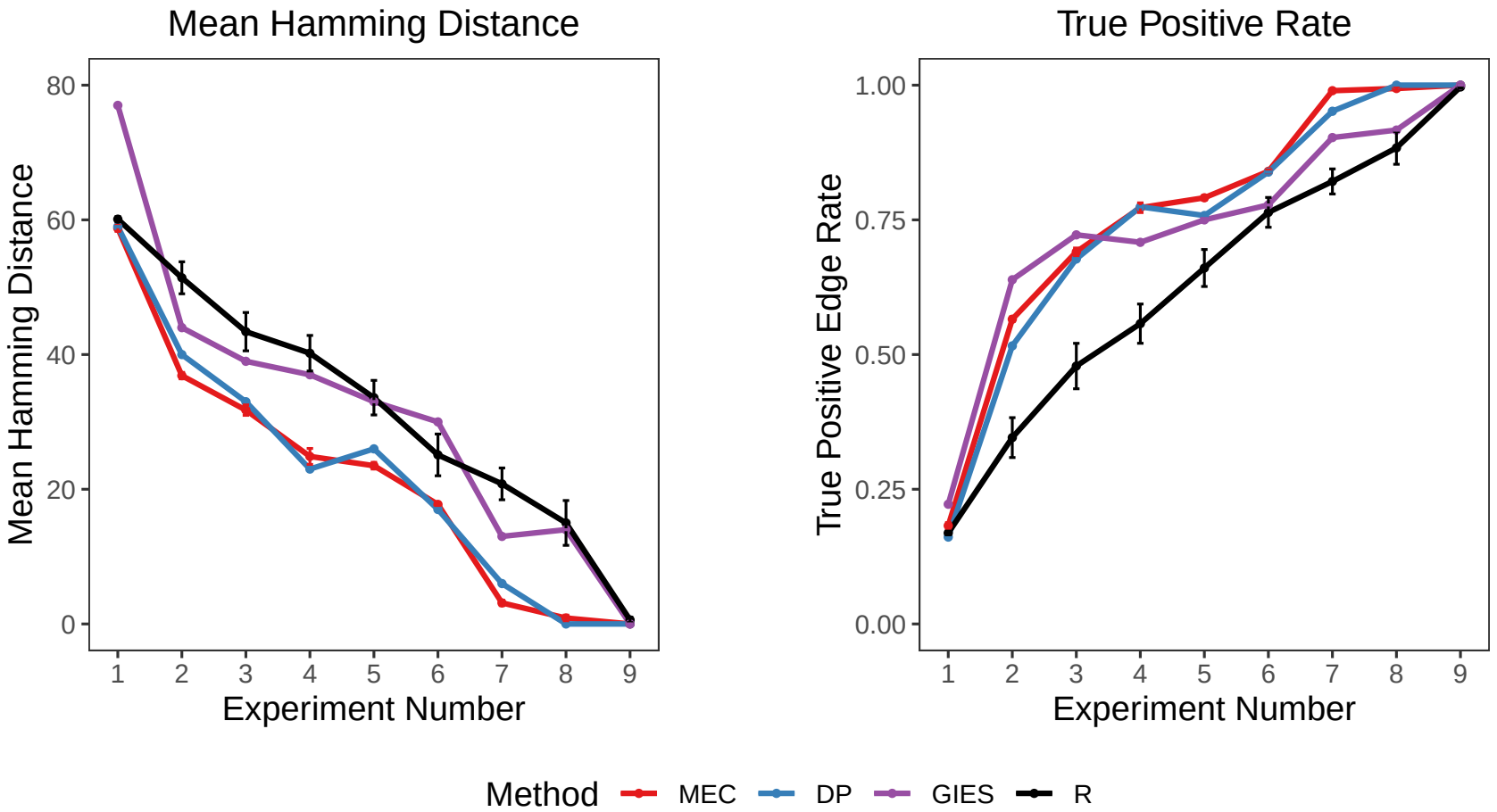
- We apply our method to the Perturb-seq data (Dixit et al., 2016), and compare with the interventional greedy sparsest permutation (IGSP) structure learning algorithm of Wang et al. (2017).

- To facilitate comparison with IGSP, we follow Wang et al. (2017) in using 992 observational samples and 13,435 interventional samples under 8 gene interventions for 14 transcription factors of interest.





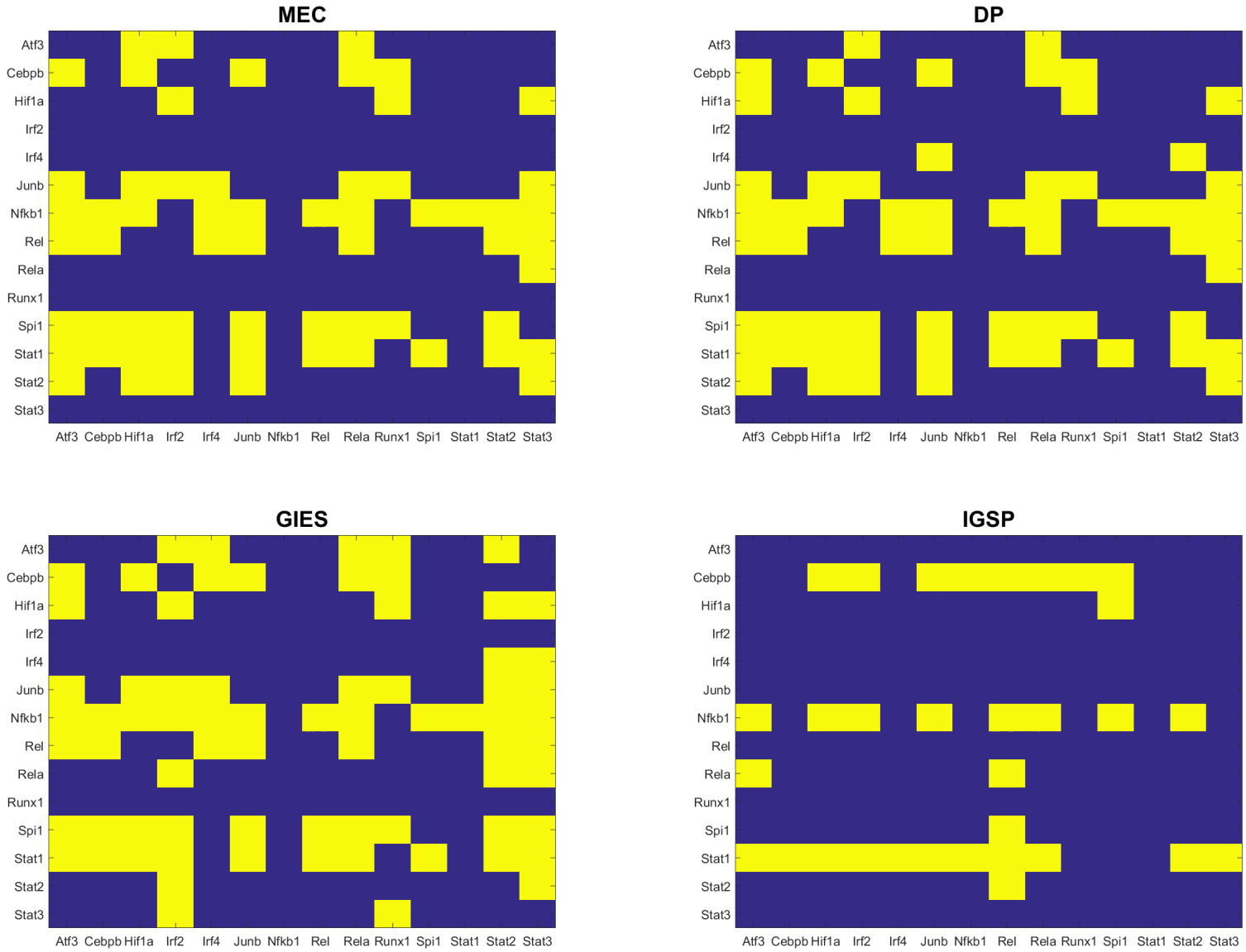
Application: Perturb-seq gene expression dataset



- Since ground truth is unknown, we self-benchmark each method by treating the final estimated network for each method as its ground truth.

Application: Perturb-seq gene expression dataset

Adjacency matrices of final estimated networks



Conclusion

- Our general criterion provides a computationally fast way to select experiments that optimize an approximation to the information gain.
- Empirically it works well for a range of natural partition schemes.
- Challenges and future work:
  - Existing posterior inference algorithms for causal networks are too slow to handle larger numbers of nodes.
  - Relaxing the acyclicity assumption is important for some real-world networks.
  - It would be interesting to try the method on other settings, beyond causal networks.

Bayesian optimal experimental design for inferring causal structure

Jeff Miller

Joint work with Michele Zemplenyi

Harvard T.H. Chan School of Public Health
Department of Biostatistics

ISBA World Meeting || June 30, 2022 ||
Session on "Bayesian experimental design for causal inference"

Preprint: https://arxiv.org/abs/2103.15229