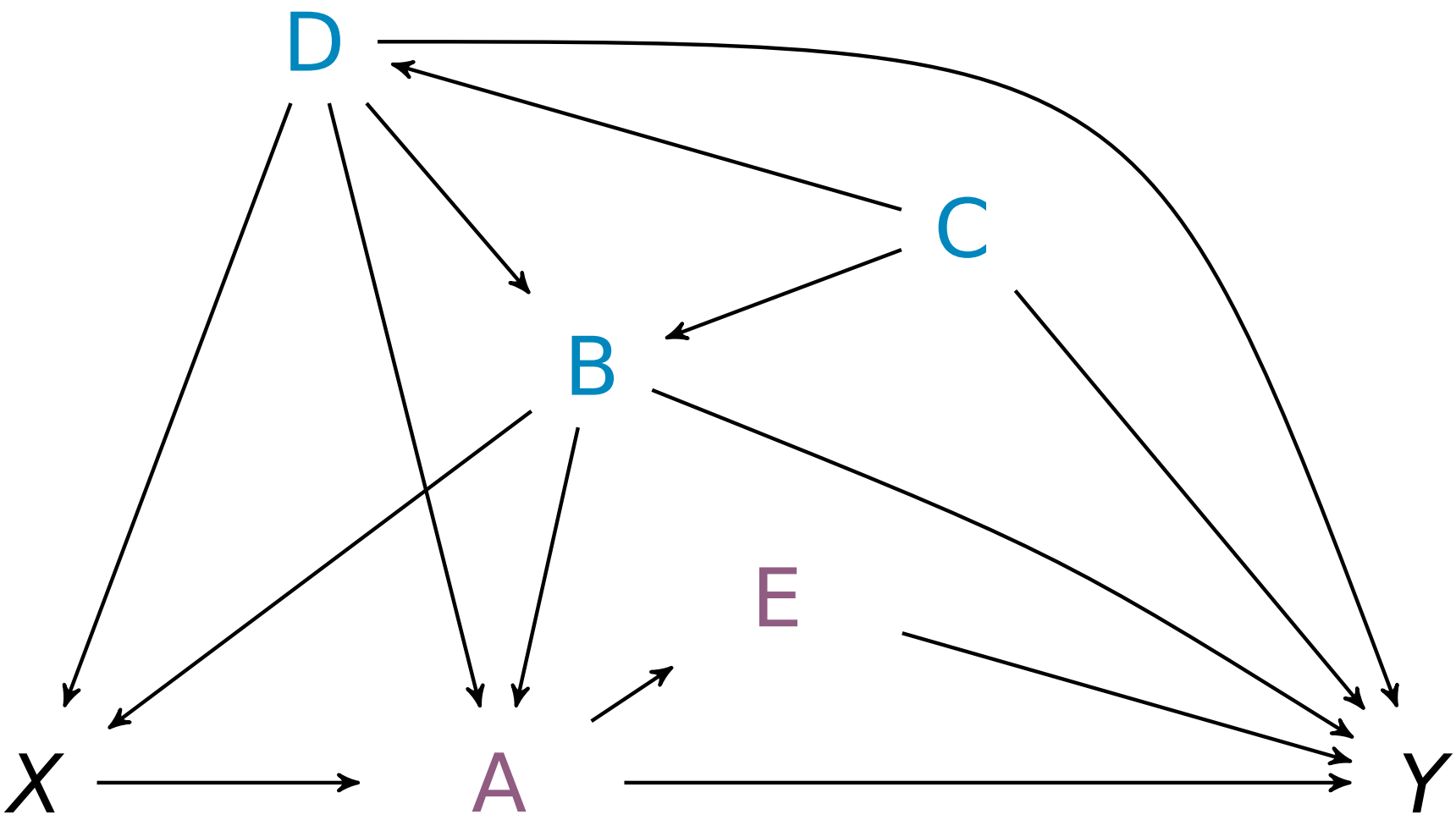
Graphical Criteria for Efficient Total Effect Estimation in Causal Linear Models

Emilija Perković

joint work with F. Richard Guo, Leonard Henckel,
Markus Kalisch, and Marloes Maathuis

Department of Statistics, University of Washington

Observational Causal DAG



Causal Directed Acyclic Graph (DAG) $\mathcal{D}$.

Goal

- Estimate the total causal effect of X on Y

  from observational data.

Observational data

Randomized control studies

Goal

- Estimate the total causal effect of X on Y
  - the change in Y due to $do(x)$-
  from observational data.

- $do(x)$: an intervention that sets variables X to x.

Observational data

Randomized control studies

Goal

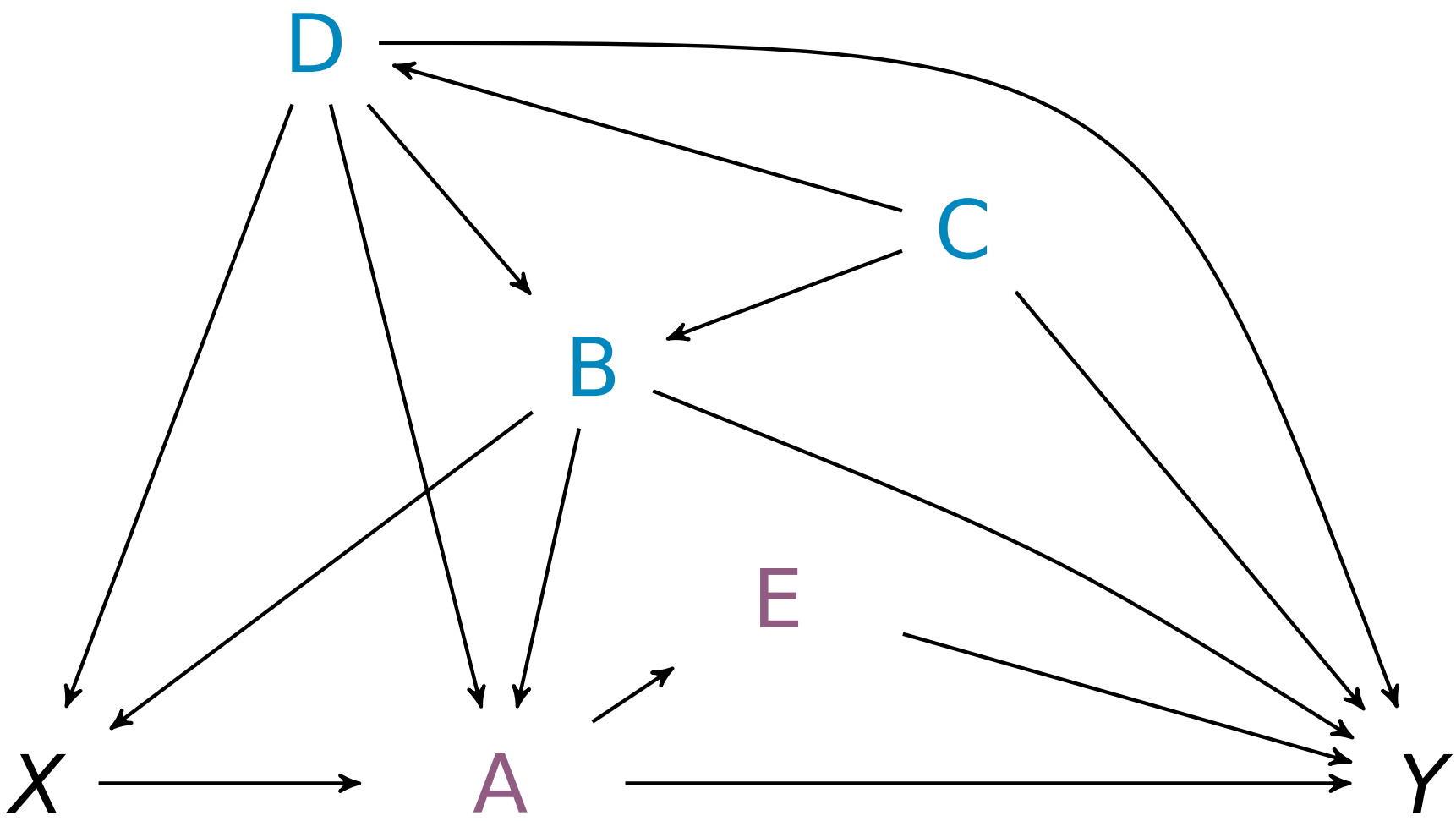
- Estimate the total causal effect of X on Y
  - the change in Y due to $do(x)$-
  from observational data.

- $do(x)$: an intervention that sets variables X to x.
  $f(y|do(x)) \neq f(y|x)$.

Observational data

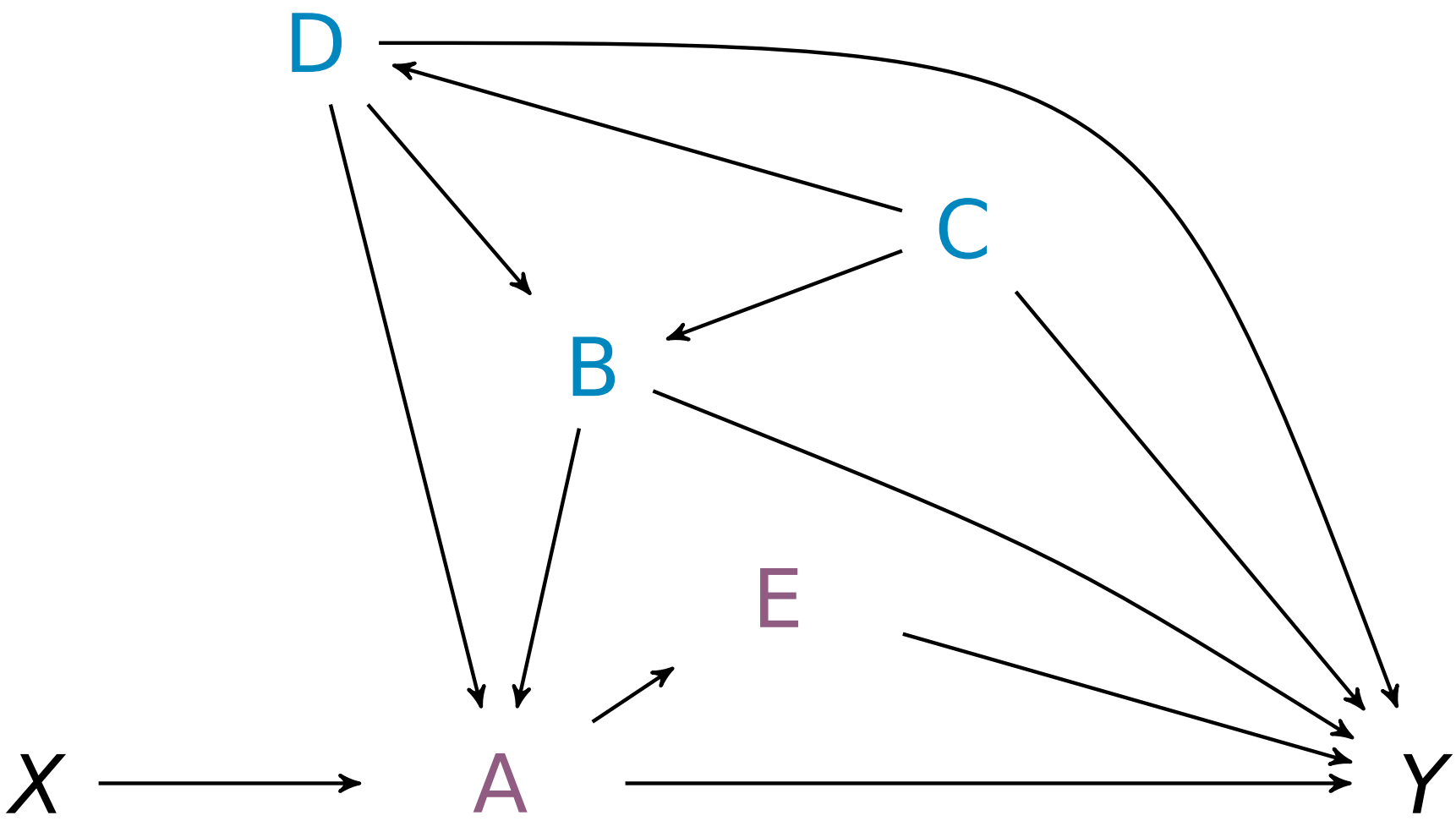
Randomized control studies

Observational Causal DAG



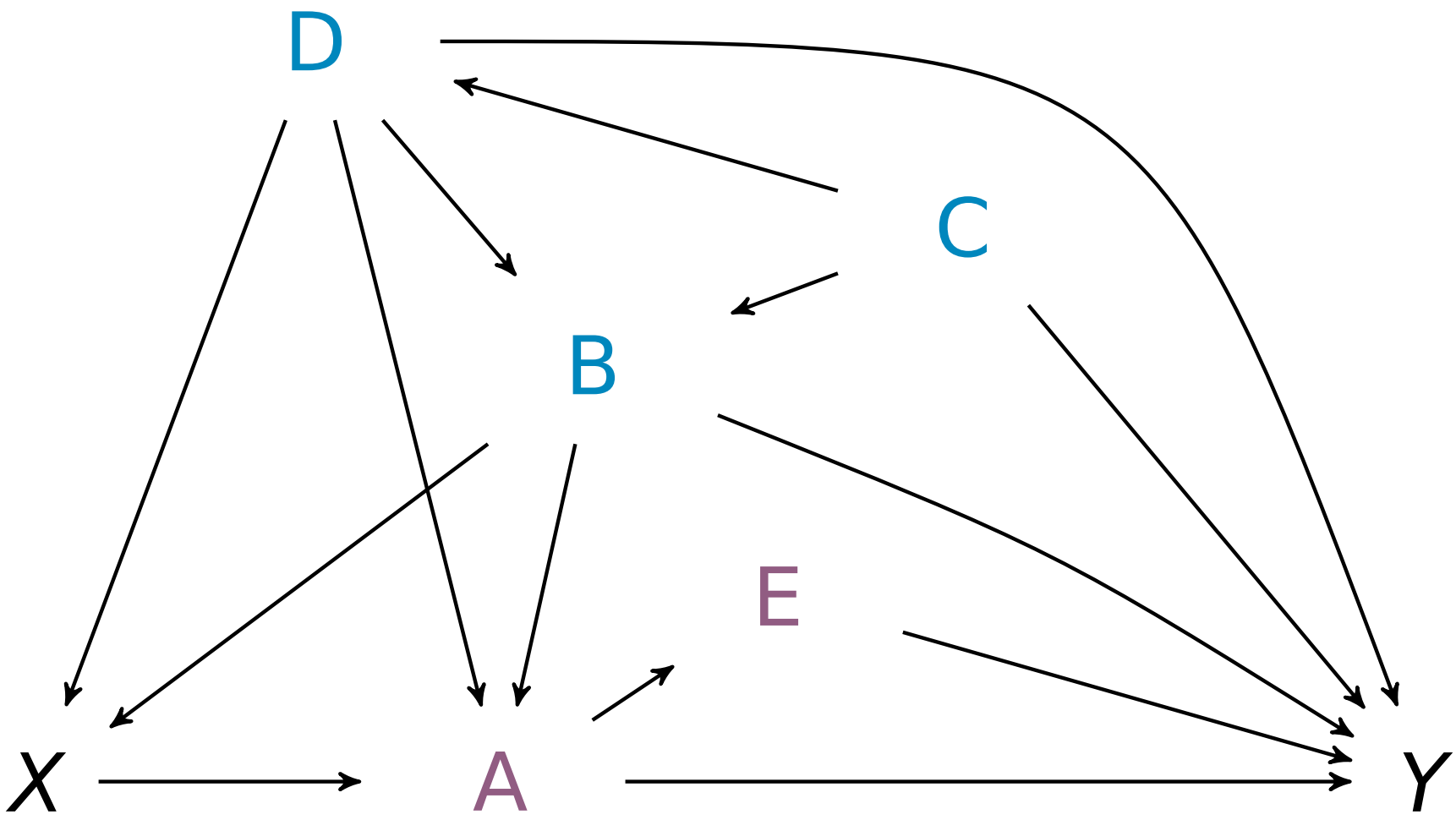
Causal Directed Acyclic Graph (DAG) $\mathcal{D}$.

Interventional Causal DAG



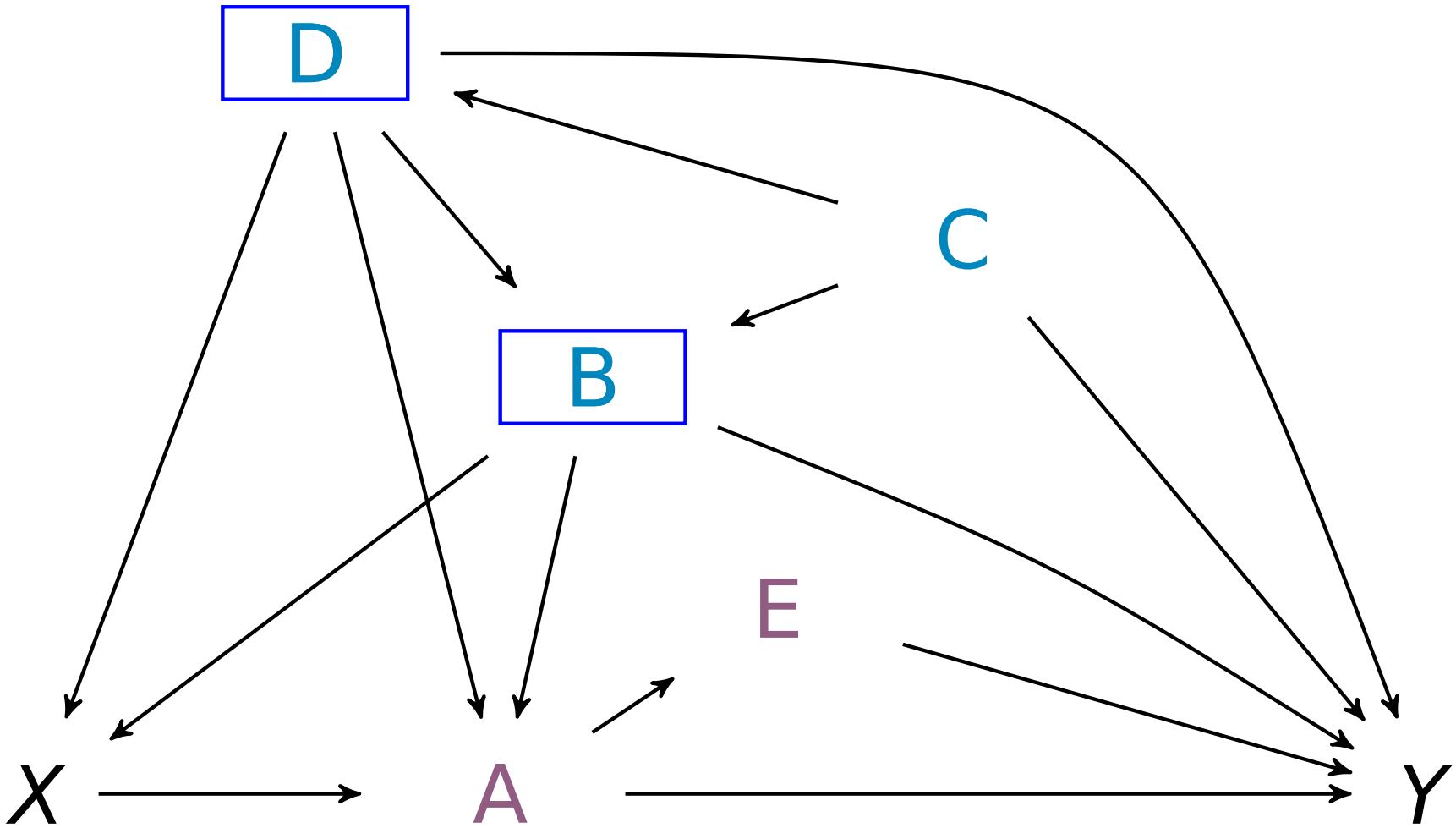
Causal DAG $\mathcal{D}$ **after a "do"-intervention on** X**.**

Observational Causal DAG



Use Pearl's Back-door Criterion

Observational Causal DAG



Use Pearl's Back-door Criterion

Overview of graphical criteria for identification

Graphical criterion	DAG
Back-door Criterion (Pearl '93)	$\Rightarrow$
G-formula (Robins '86)	$\Leftrightarrow$

$\Leftrightarrow$ - $\Rightarrow$ - sufficient for identification,
necessary and sufficient for identification

Back-door Adjustment: If Z is an back-door set, then

$$f(y|do(x)) = \int f(y|x,z)f(z)dz$$

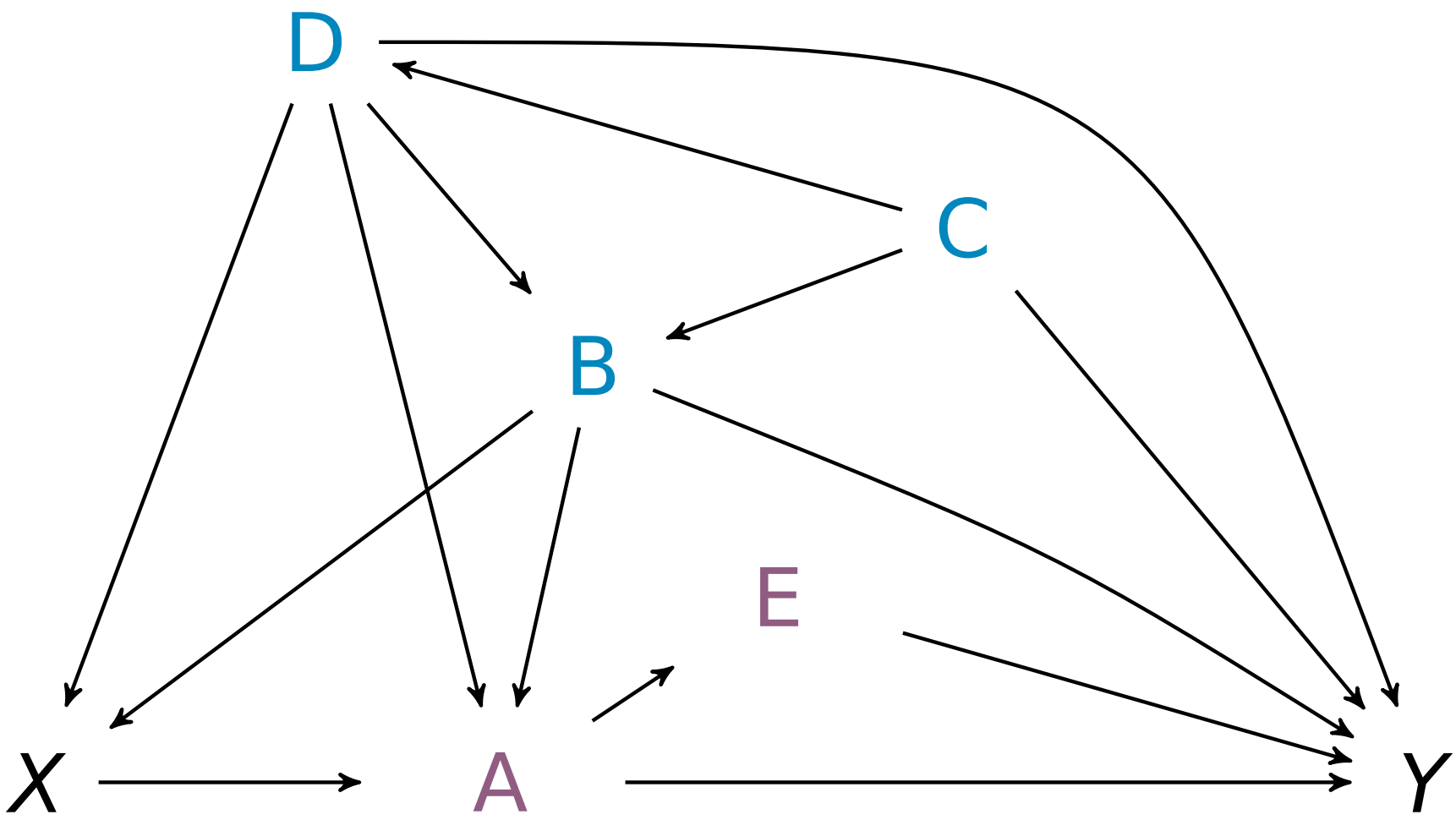
In the linear case, estimate using an adjusted regression.

G-formula: Let $V' = V \setminus \{X \cup Y\}$, then

$$f(y|do(x)) = \int \prod_{V_i \in V \setminus X} f(v_i|pa(v_i, \mathcal{D}))dv'.$$

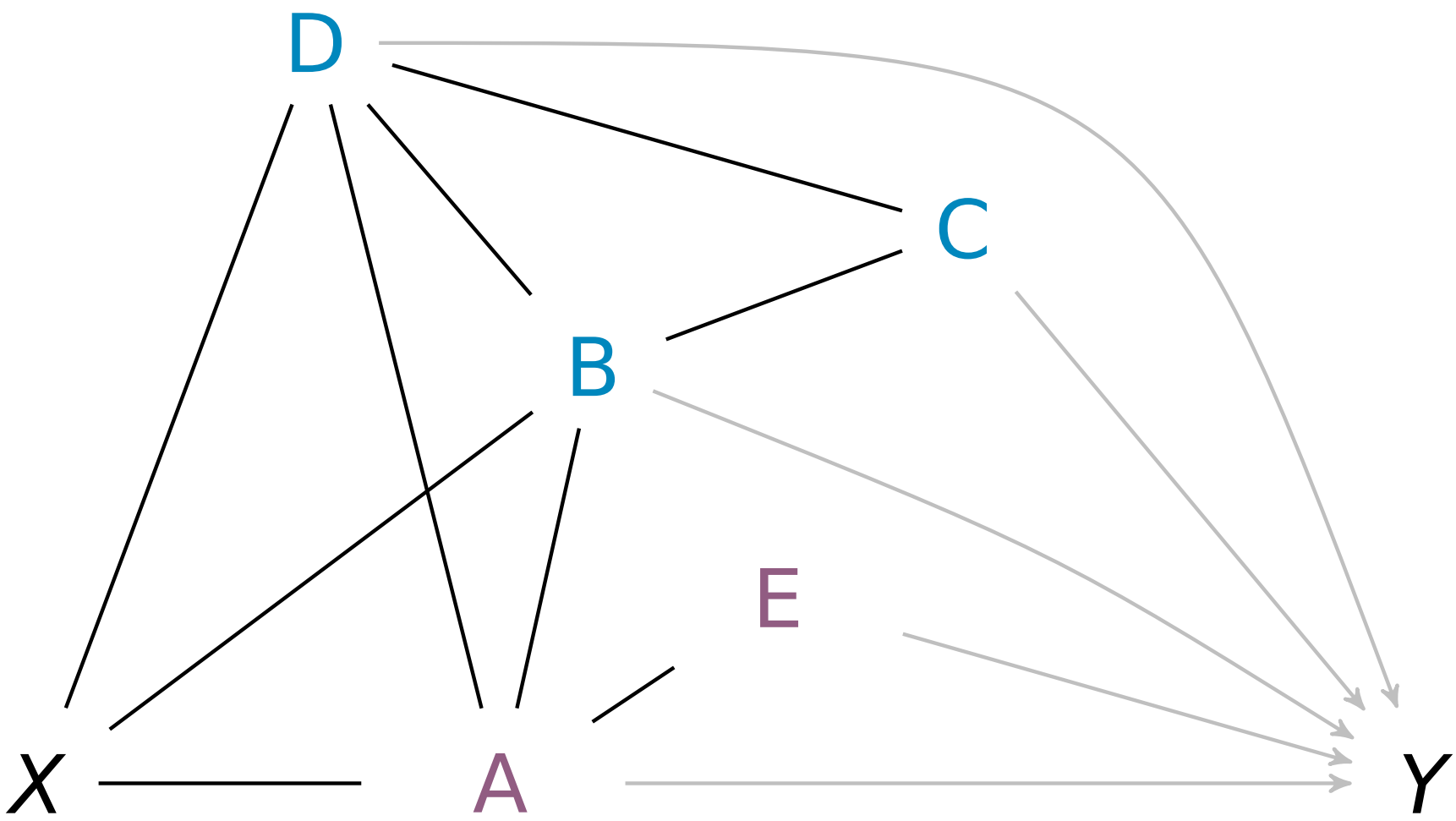
In the linear case, estimate using sequential adjusted regressions.

Problem solved?



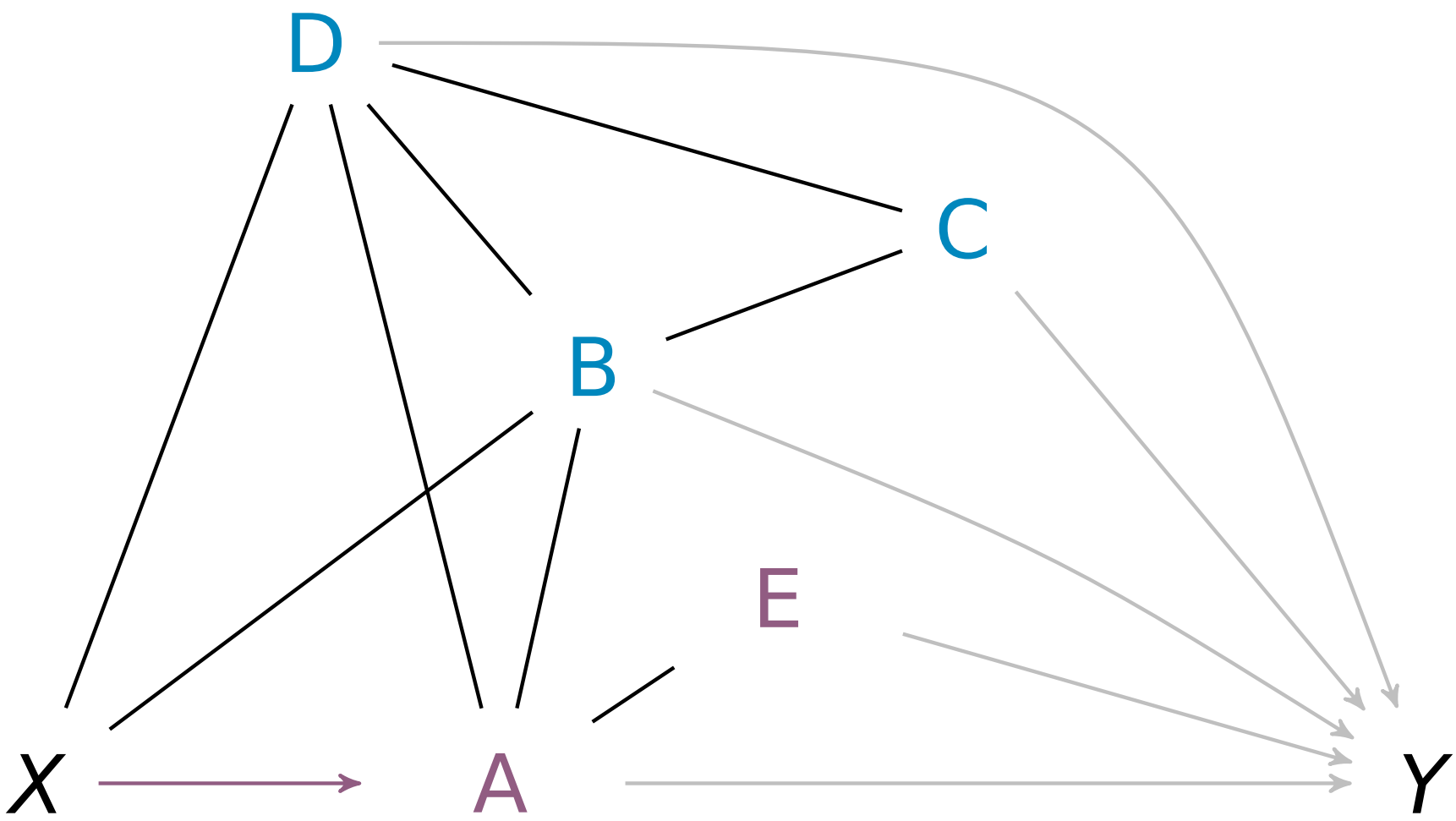
DAG $\mathcal{D}$.

Problem solved?



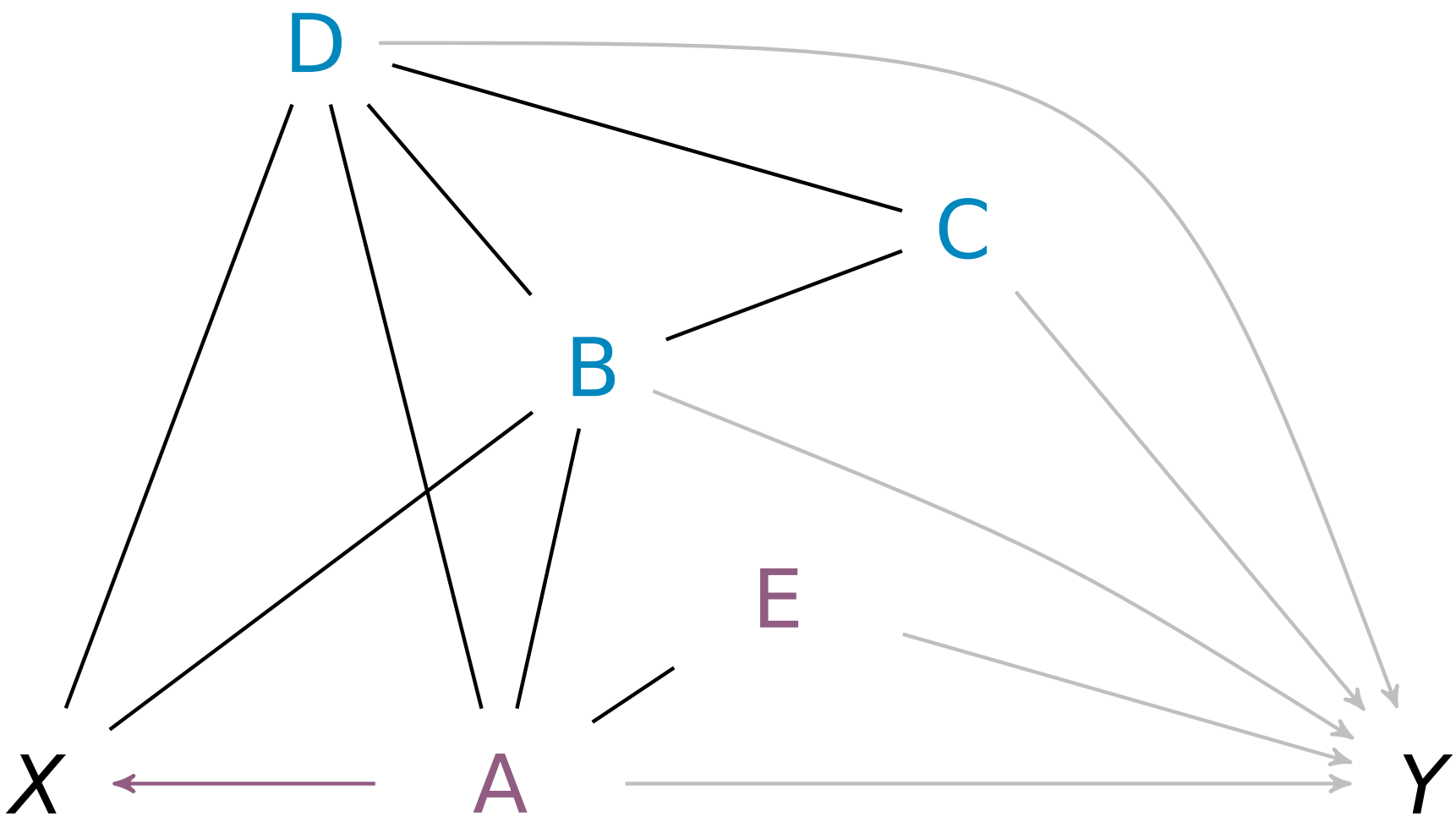
Completed Partially Directed Acyclic Graph (CPDAG) $\mathcal{C}$.

Problem solved?



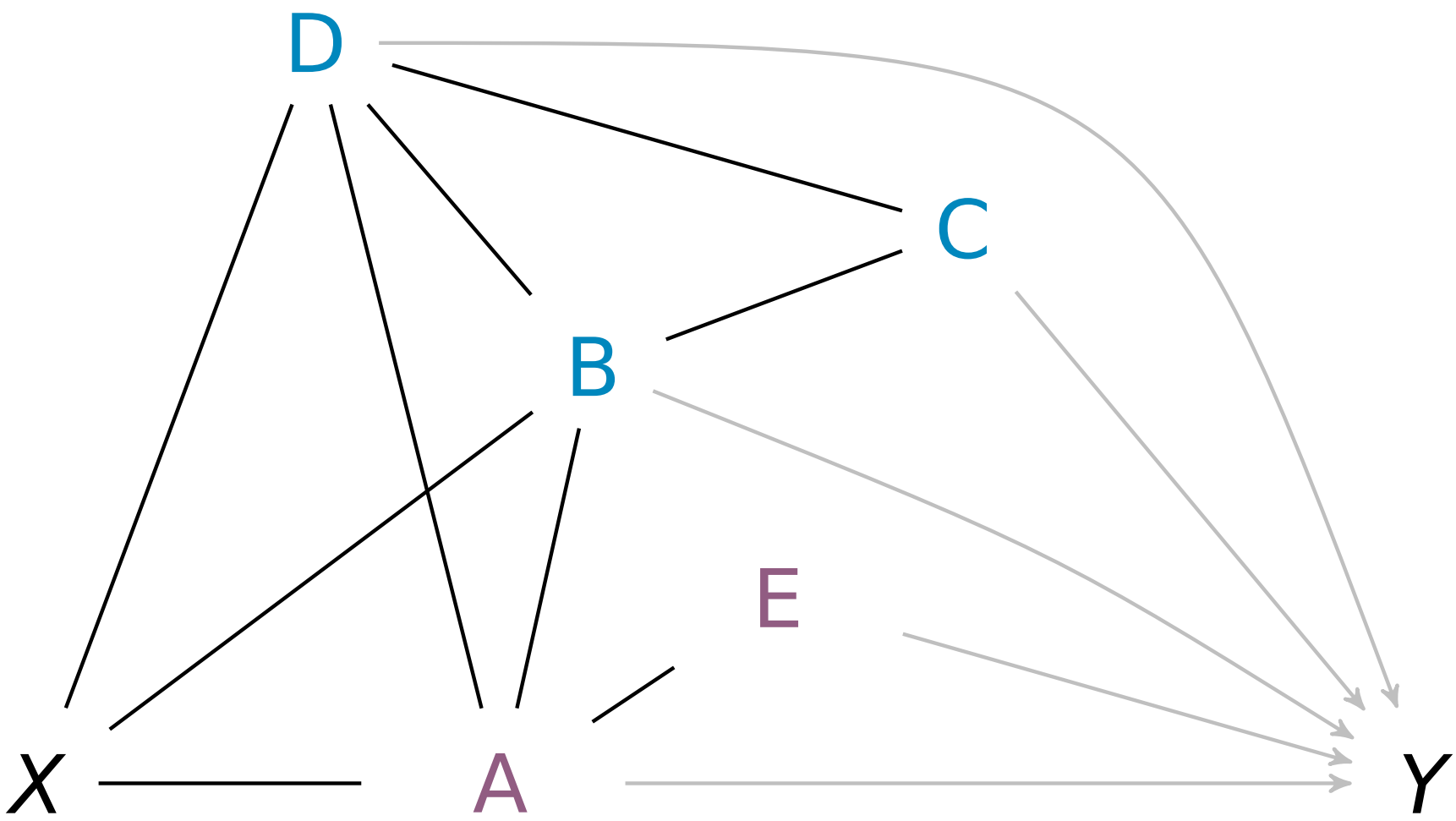
Completed Partially Directed Acyclic Graph (CPDAG) $\mathcal{C}$.

Problem solved?



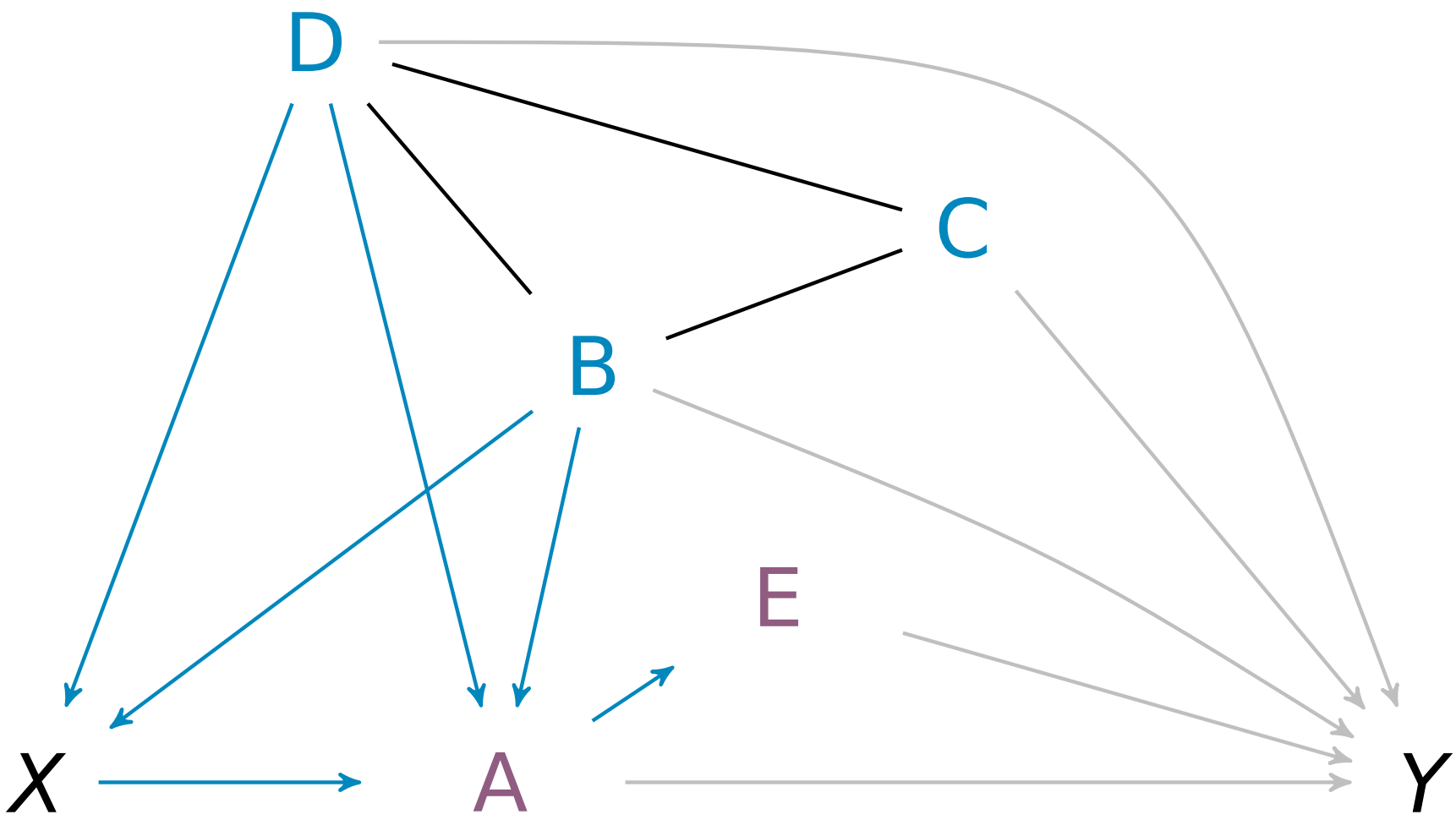
Completed Partially Directed Acyclic Graph (CPDAG) $\mathcal{C}$.

Problem solved?



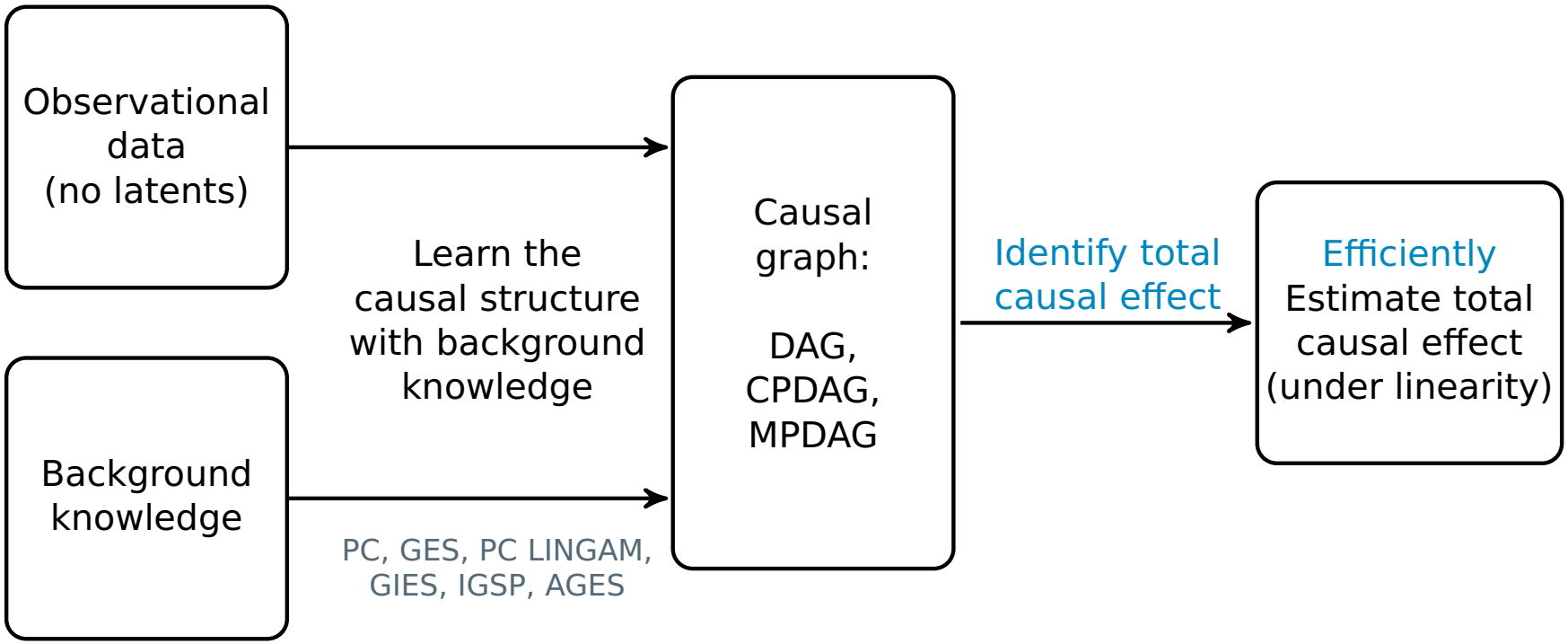
Completed Partially Directed Acyclic Graph (CPDAG) $\mathcal{C}$.

Problem solved?



Maximally oriented Partially Directed Acyclic Graph (MPDAG) $\mathcal{G}$.

Framework



- PC (Spirtes et al, 1993), GES (Chickering, 2002)
- Adding background knowledge (Meek, 1995; TETRAD, Scheines et al., 1998), PC LINGAM (Hoyer et al., 2008), GIES (Hauser and Bühlmann, 2012), IGSP (Wang et al., 2017), etc.

Overview of graphical criteria for identification

Graphical criterion	DAG	CPDAG	MPDAG
Back-door Criterion (Pearl 1993)	$\Rightarrow$		
Generalized adjustment (Shpitser et al '10, Perković et al '17, '18)	$\Rightarrow$	$\Rightarrow$	$\Rightarrow$
G-formula (Robins '86)	$\Leftrightarrow$		
Causal identification formula (Perković '20)	$\Leftrightarrow$	$\Leftrightarrow$	$\Leftrightarrow$

$\Rightarrow$ - sufficient for identification,
$\Leftrightarrow$ - necessary and sufficient for identification

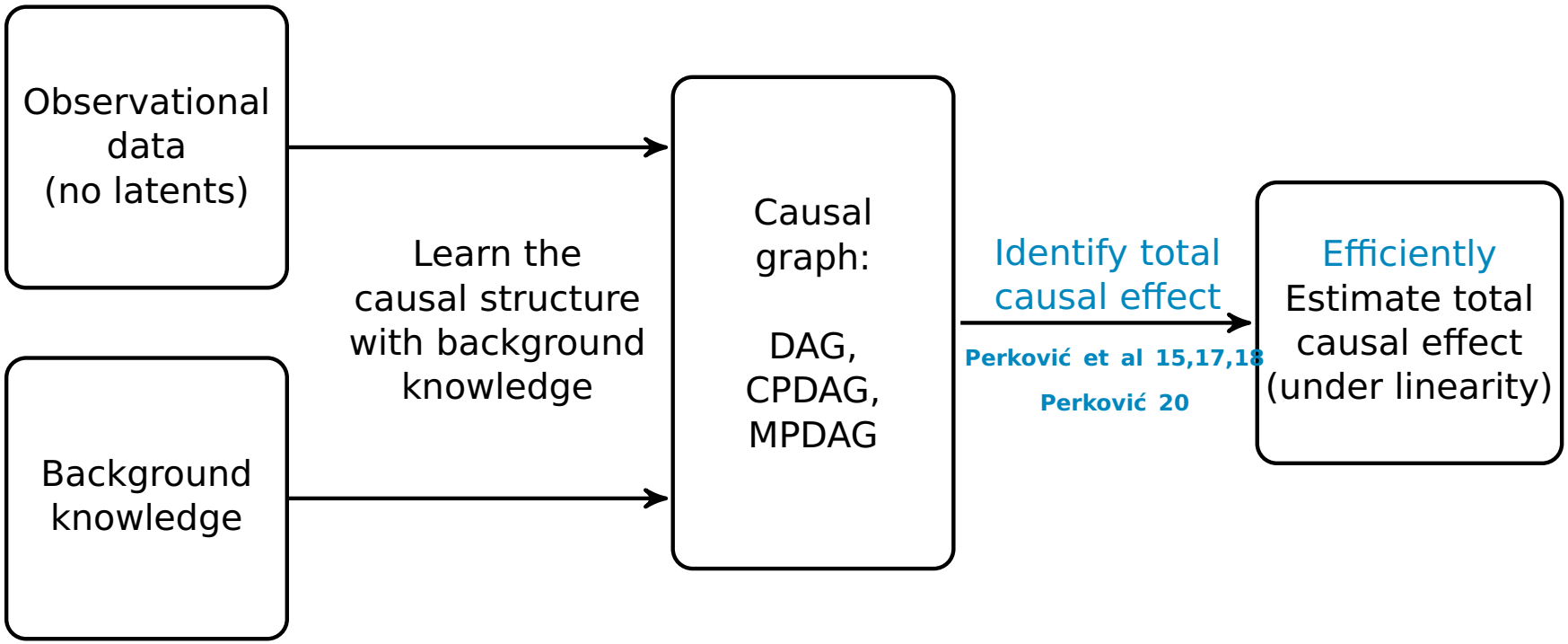
Adjustment: Z is an adjustment set if

$$f(y|do(x)) = \int f(y|x,z)f(z)dz$$

Causal Identification Formula: Let $S = an(Y, \mathcal{G}_{V\setminus X}) \setminus Y$ and let $(S_1, \ldots, S_k)$ be the partition of $S \cup Y$ into undirected components

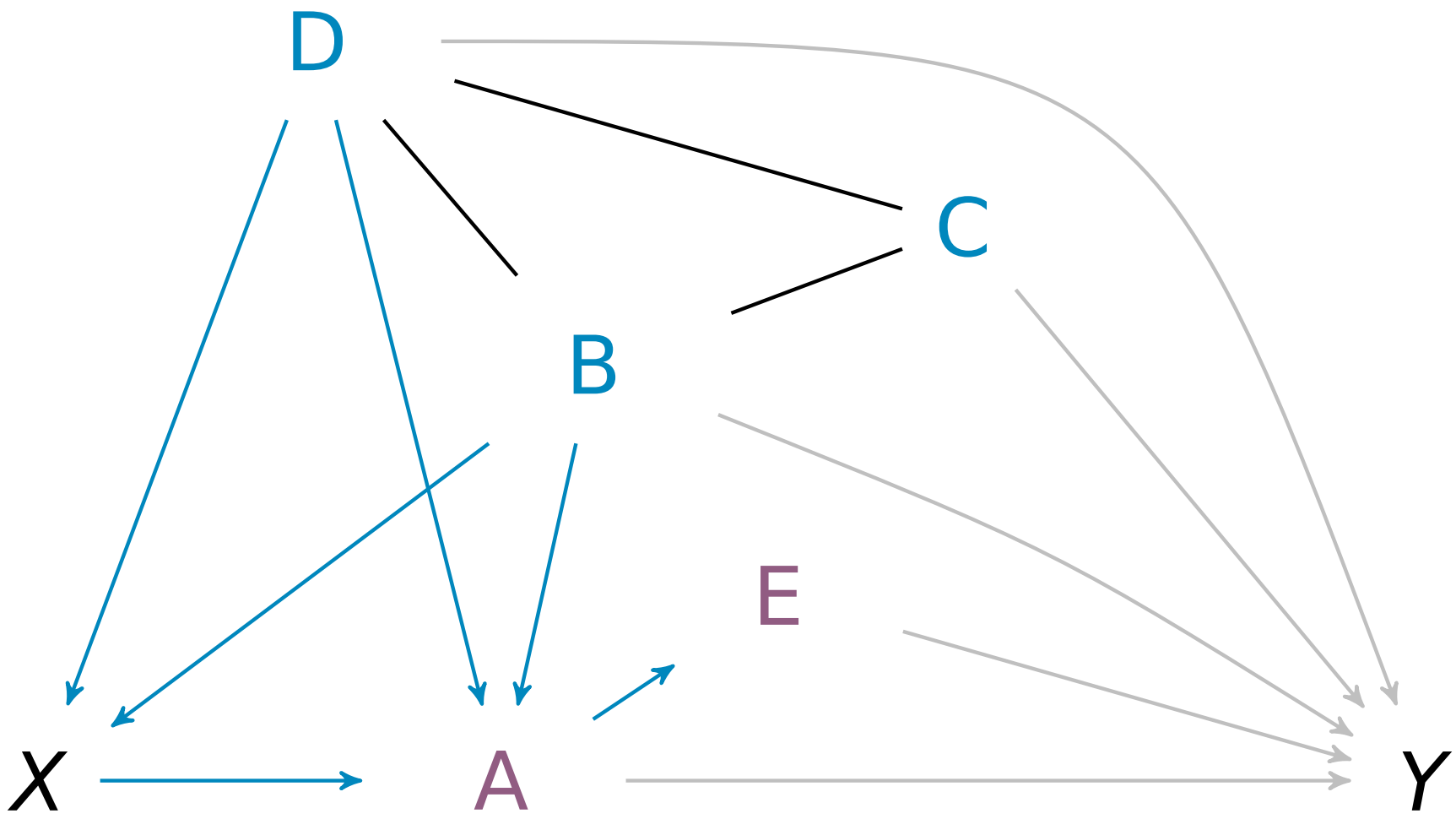
$$f(y|do(x)) = \int \prod_{i=1}^{k} f(s_i|\text{pa}(s_i, \mathcal{G}))ds$$

Framework



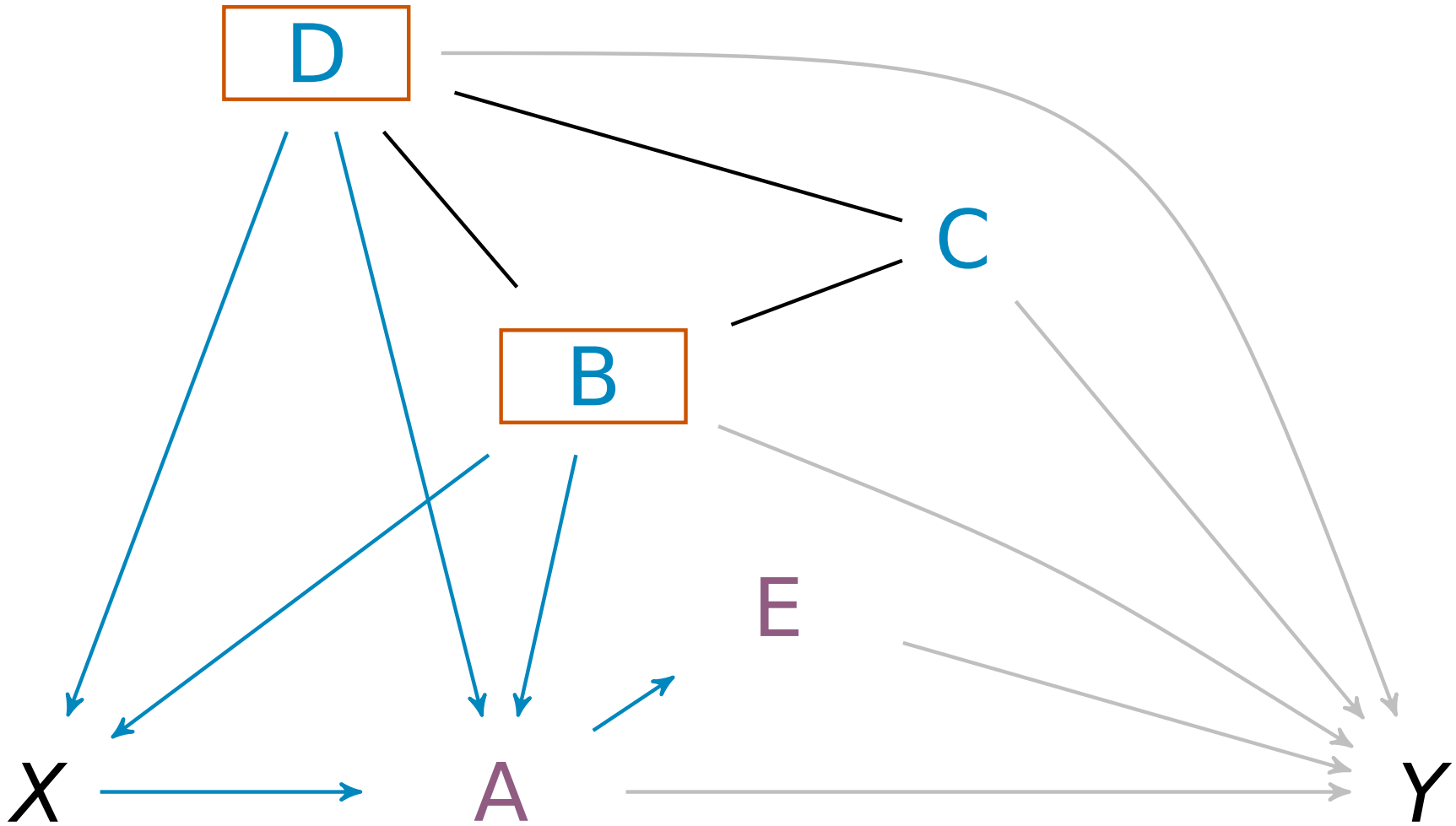
- Generalized adjustment criterion (Perković et al 2015, 2017, 2018). Extends to hidden variable settings.
- Causal identification formula (generalized g-formula, Perković 2020).

MPDAG



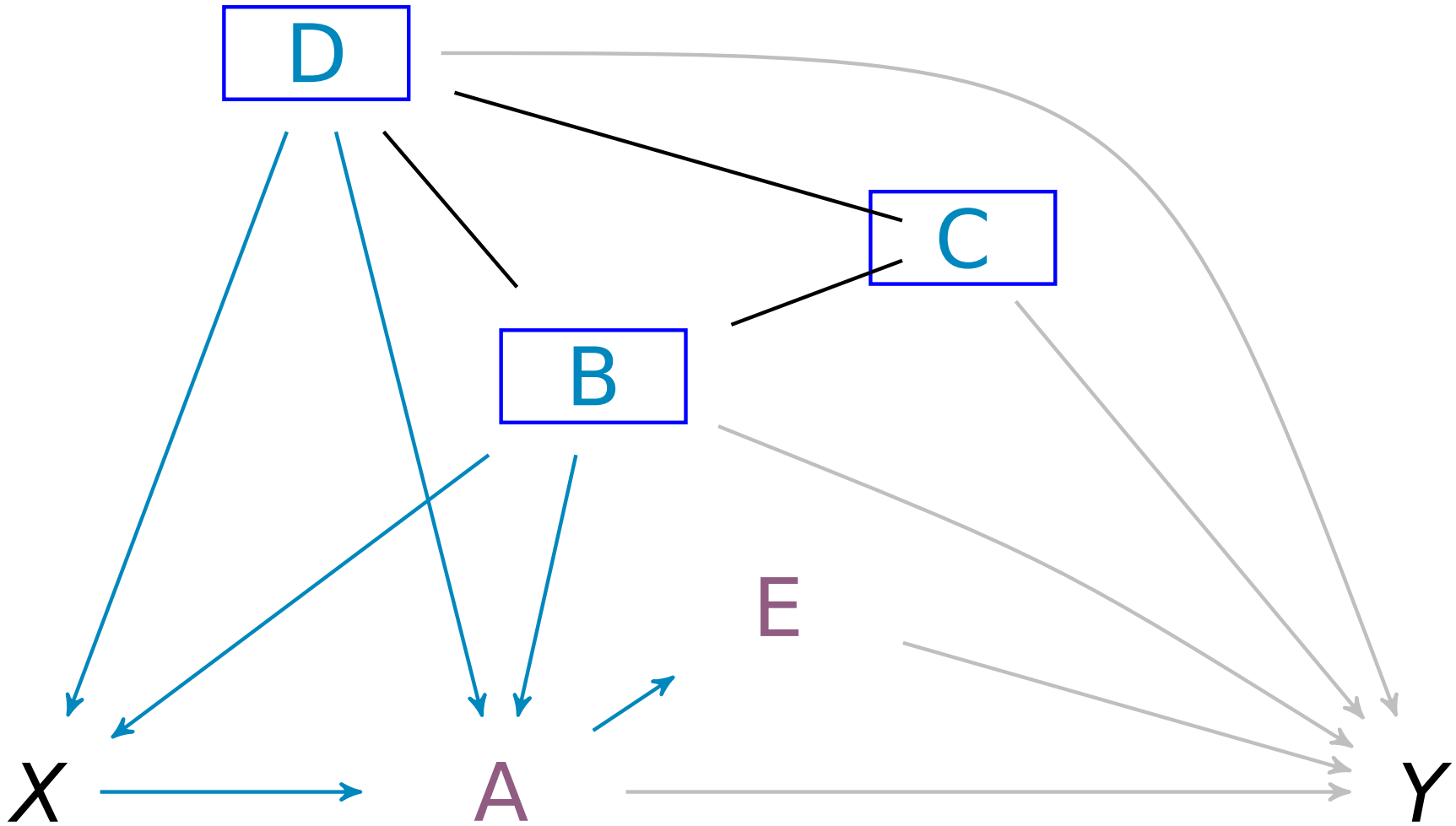
Which is more efficient? Adjustment set $\{B, D\}$, or $\{B, C, D\}$, or $\mathcal{G}$-regression ?

MPDAG



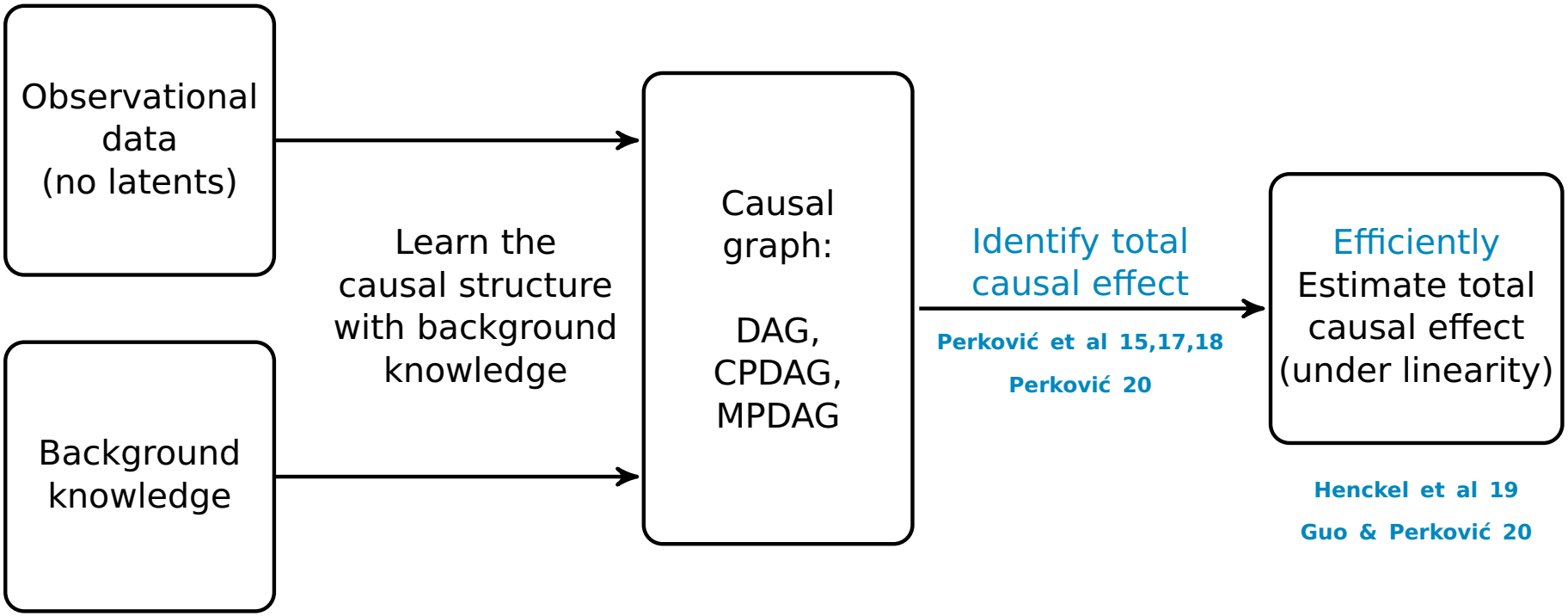
Which is more efficient? Adjustment set $\{B, D\}$, or $\{B, C, D\}$, or $\mathcal{G}$-regression ?

MPDAG



Which is more efficient? Adjustment set $\{B, D\}$, or $\{B, C, D\}$, or $\mathcal{G}$-regression ?

Framework



- Comparison of adjustment sets (Henckel et al 2019).
  Extends partially to hidden variable settings.
- Comparison of $\mathcal{G}$-regression with adjustment and other regression based methods (Guo & Perković 2020).

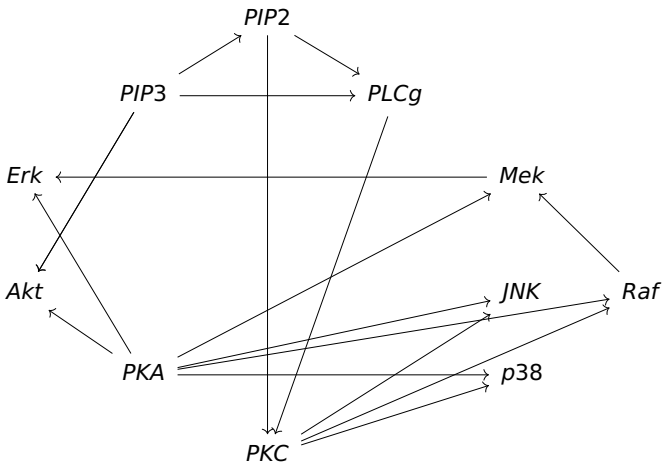
Results of Henckel et al 2019

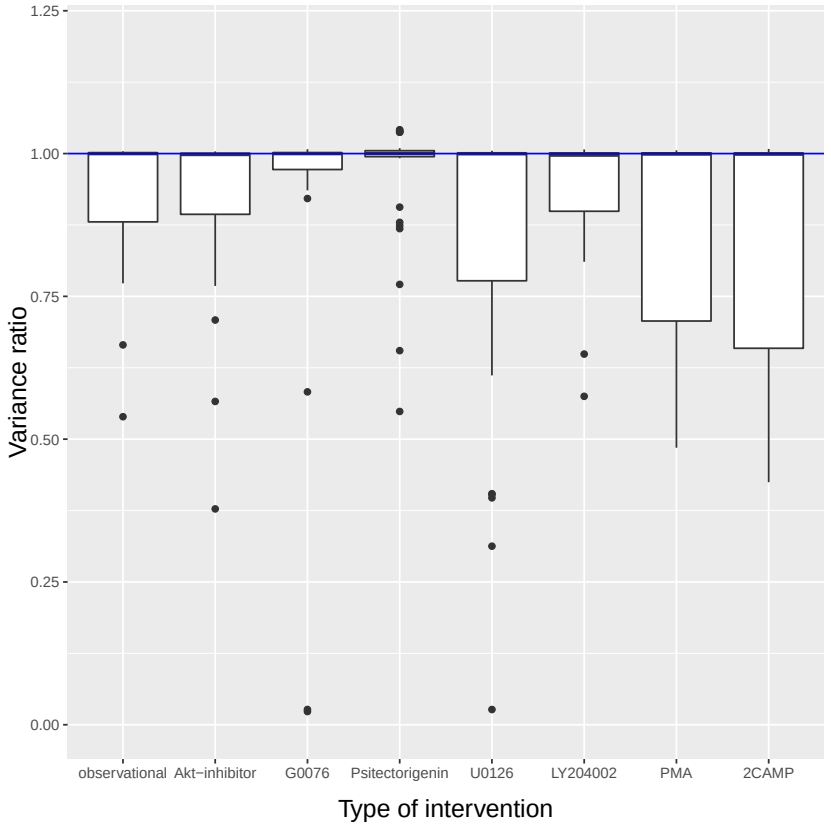
- A graphical criterion for pairwise adjustment set comparison in terms of asymptotic variance.
- A variance reducing pruning procedure for each adjustment set.
- A graphical characterization of an asymptotically optimal valid adjustment set $\mathbf{O}(X, Y, \mathcal{G})$ (does not hold in the hidden variable setting).
- Results implemented in `R` package `pcalg`.
- Key Takeaways: $pa(X, \mathcal{G})$ performs very poorly, a smaller set is generally not better than a larger set, including variables "close" to Y is beneficial.

Results of Guo & Perković 2020

- $\mathcal{G}$-regression consistently estimates any identified effect.
- $\mathcal{G}$-regression is asymptotically most efficient among all consistent and differentiable estimators based on the first two moments.
- This includes all regression based estimators e.g.
  1. covariate adjustment (Henckel et al, 2019, Witte et al, 2020),
  2. recursive regressions (Nandy et al, 2017, Gupta et al, 2020),
  3. modified Cholesky decomposition (Nandy et al, 2017).
- Results implemented in `R` package `eff2` at github.com/richardkwo/eff2

Comparing Adjustment Sets, Henckel et al 2019





- (left) Consensus protein signaling causal network in human T-Cells (Sach, et al ’05).
- (right) Boxplots of variance ratios comparing the causal effect estimate using the optimal set and the parent set in 8 experimental settings.
  Single-cell flow cytometry data of Sach et al 2005.

G-regression, Guo & Perković 2020

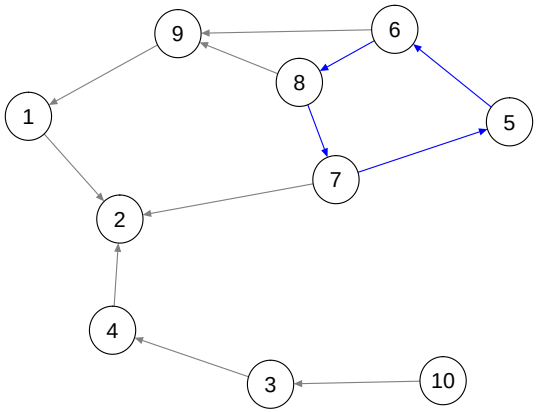
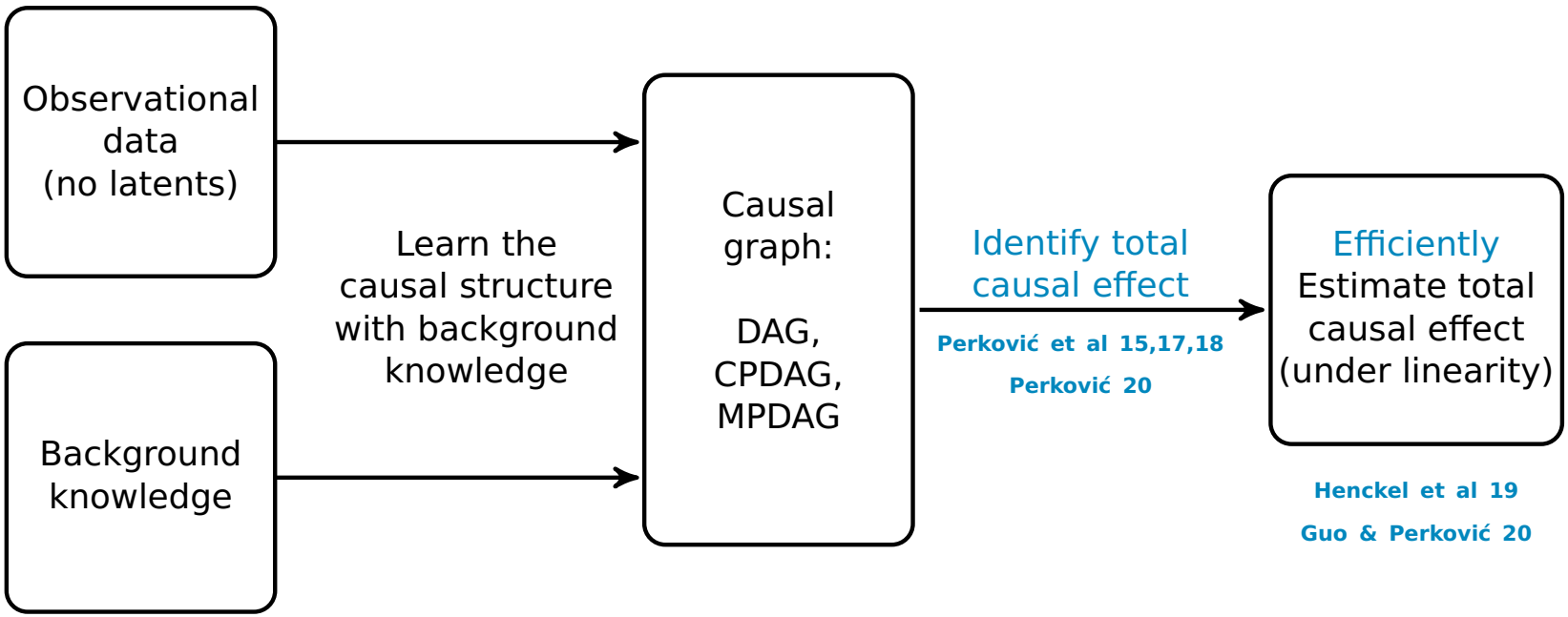


Table: Normalized sq. errors when predicting double knockouts

	$\mathcal{E}^*$			$\mathcal{E}$		
removed	$\nexists$ `adj.O`	`adj.O`	$\mathcal{G}$-`reg`	`IDA.R`	$\mathcal{G}$-`reg`	`baseline`
$5 \to 6$	36%	43%	**35%**	46%	**30%**	81%
$6 \to 8$	42%	**29%**	32%	33%	**26%**	81%
$8 \to 7$	60%	39%	**35%**	45%	**44%**	81%
$7 \to 5$	46%	40%	**33%**	45%	**34%**	81%

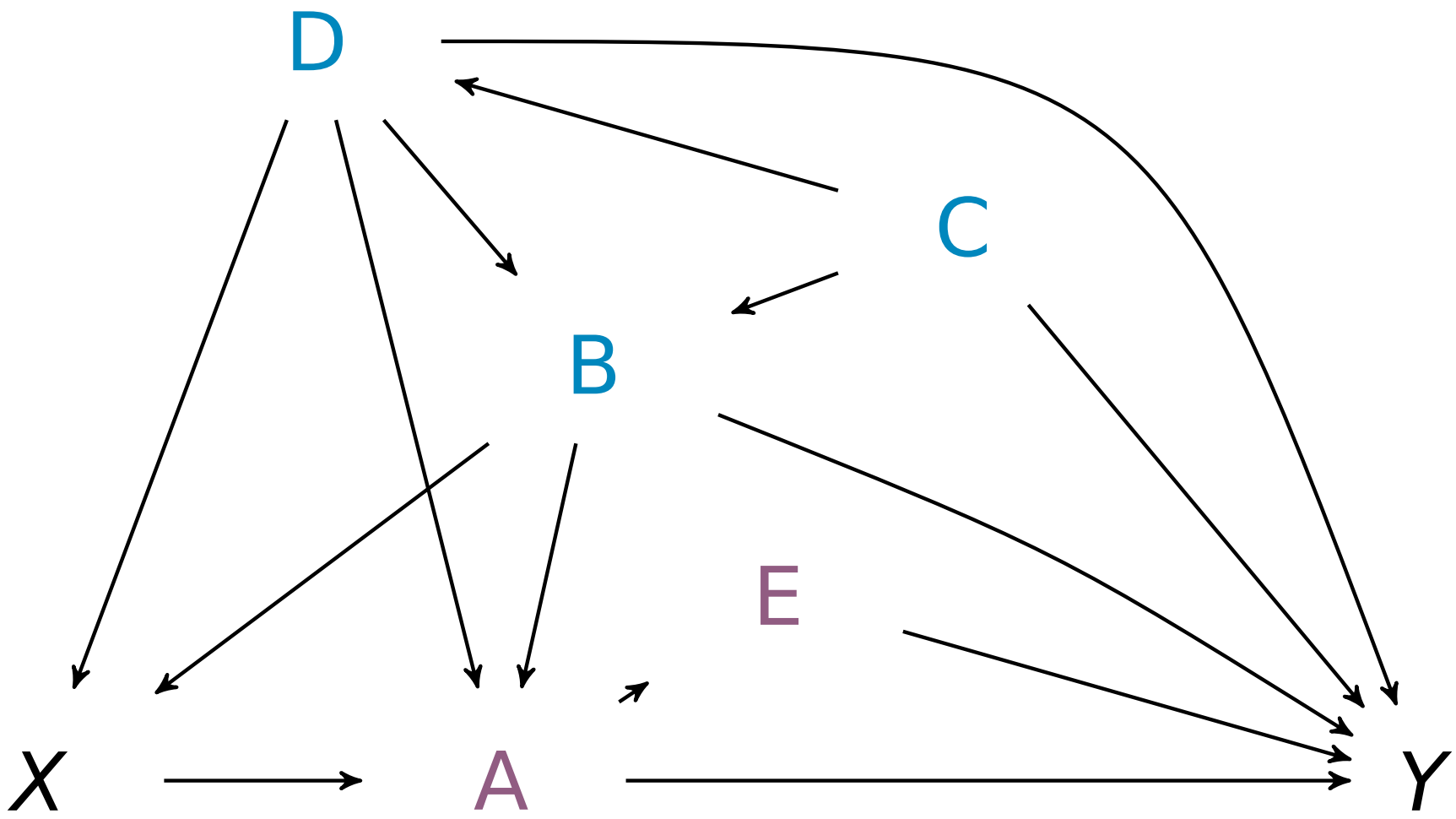
- (Left) Gene regulation network from DREAM4 challenge based on the 5th Size10 dataset (Marbach et al, 2009). An SDE model generated the data under wild type, perturbed steady state and knockout interventions. The true network contains a cycle.
- The performance is evaluated with normalized squared error using interventional knockout data on $\{(6,8),(7,8),(8,10),(8,5),(8,9)\}$.

Summary



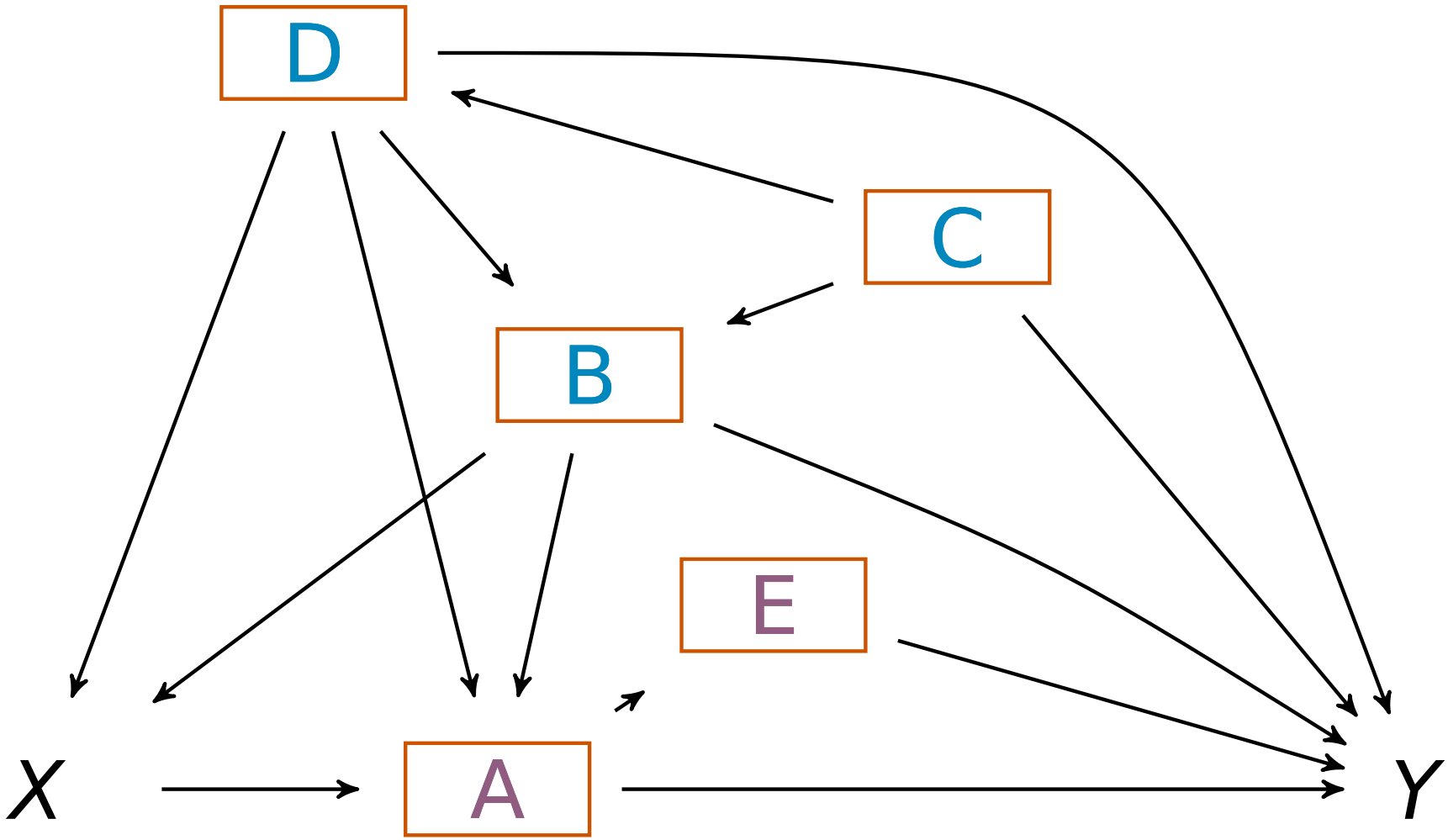
Thanks!

Observational Causal DAG



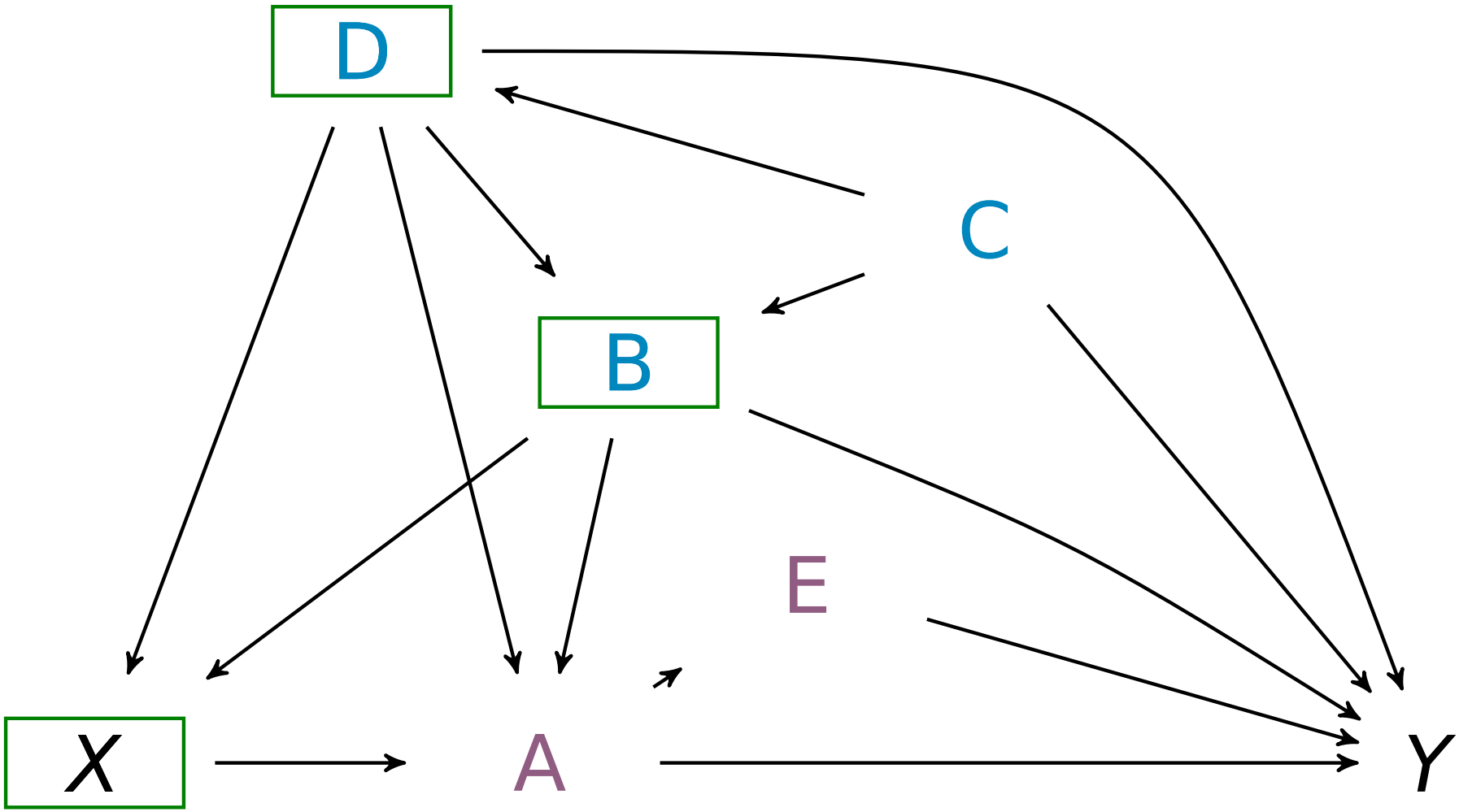
Use G-formula or the sequential backdoor criterion

Observational Causal DAG



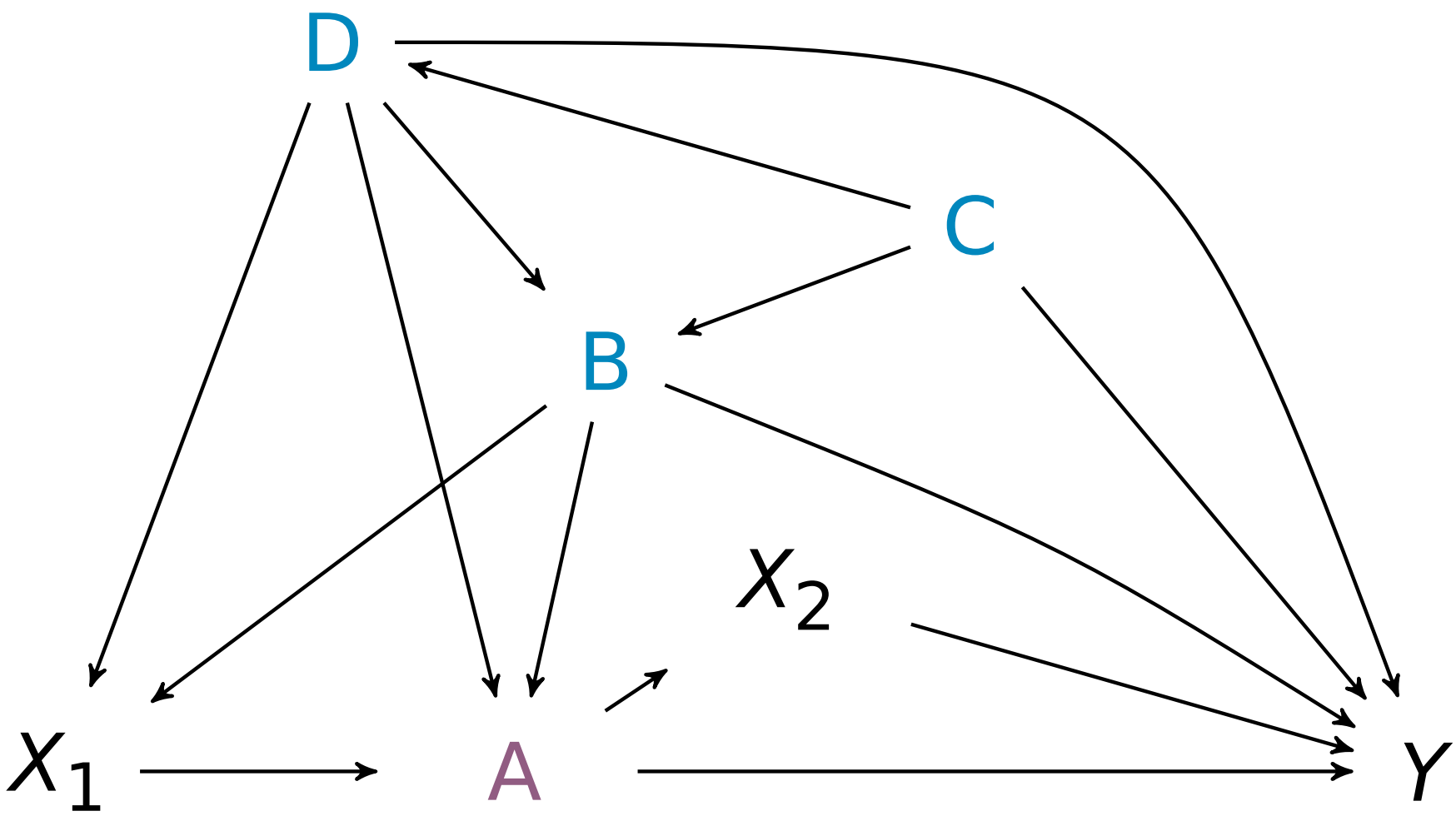
Use G-formula or the sequential backdoor criterion

Observational Causal DAG



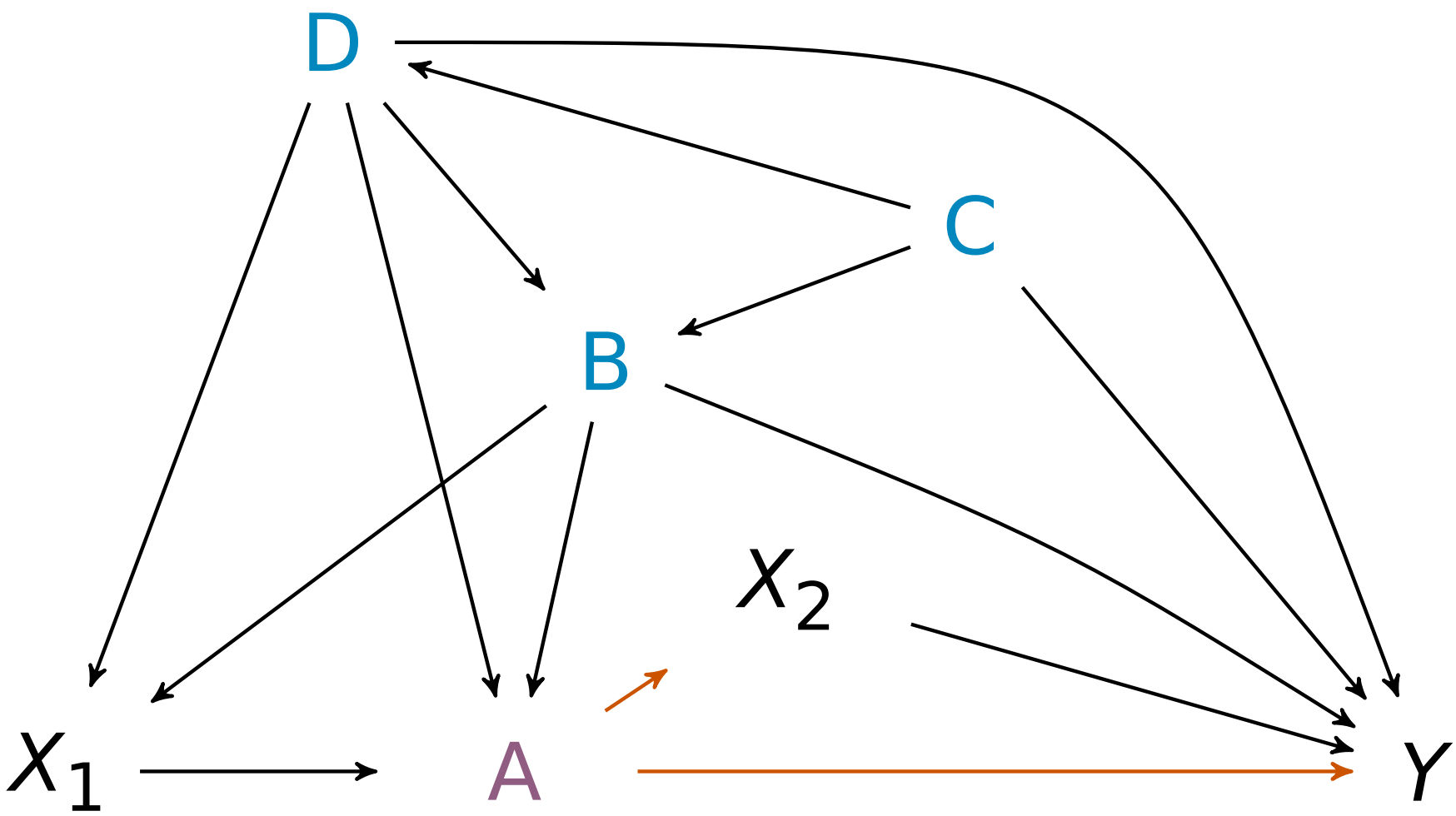
Use G-formula or the sequential backdoor criterion

Joint intervention



Use G-formula, or the sequential back-door criterion.

Joint intervention



Use G-formula, or the sequential back-door criterion.