Asymmetric predictability

An information theoretic approach to causal inference

Soumik Purkayastha,
Department of Biostatistics, University of Michigan.

2023 Joint Statistical Meetings,
Toronto, Canada.


Peter X. K. Song, PhD

Dept. of Biostatistics, University of Michigan


Karen E. Peterson, DSc

Dept. of Nutritional Sciences, University of Michigan


Jackie Goodrich, PhD

Dept. of Environmental Health Sciences, University of Michigan

Acknowledgements

“Asymmetry mirrors underlying causal structures*”

“Asymmetry mirrors underlying causal structures*”

*Postulates and assumptions apply

Does $X \to Y$ or $Y \to X$?

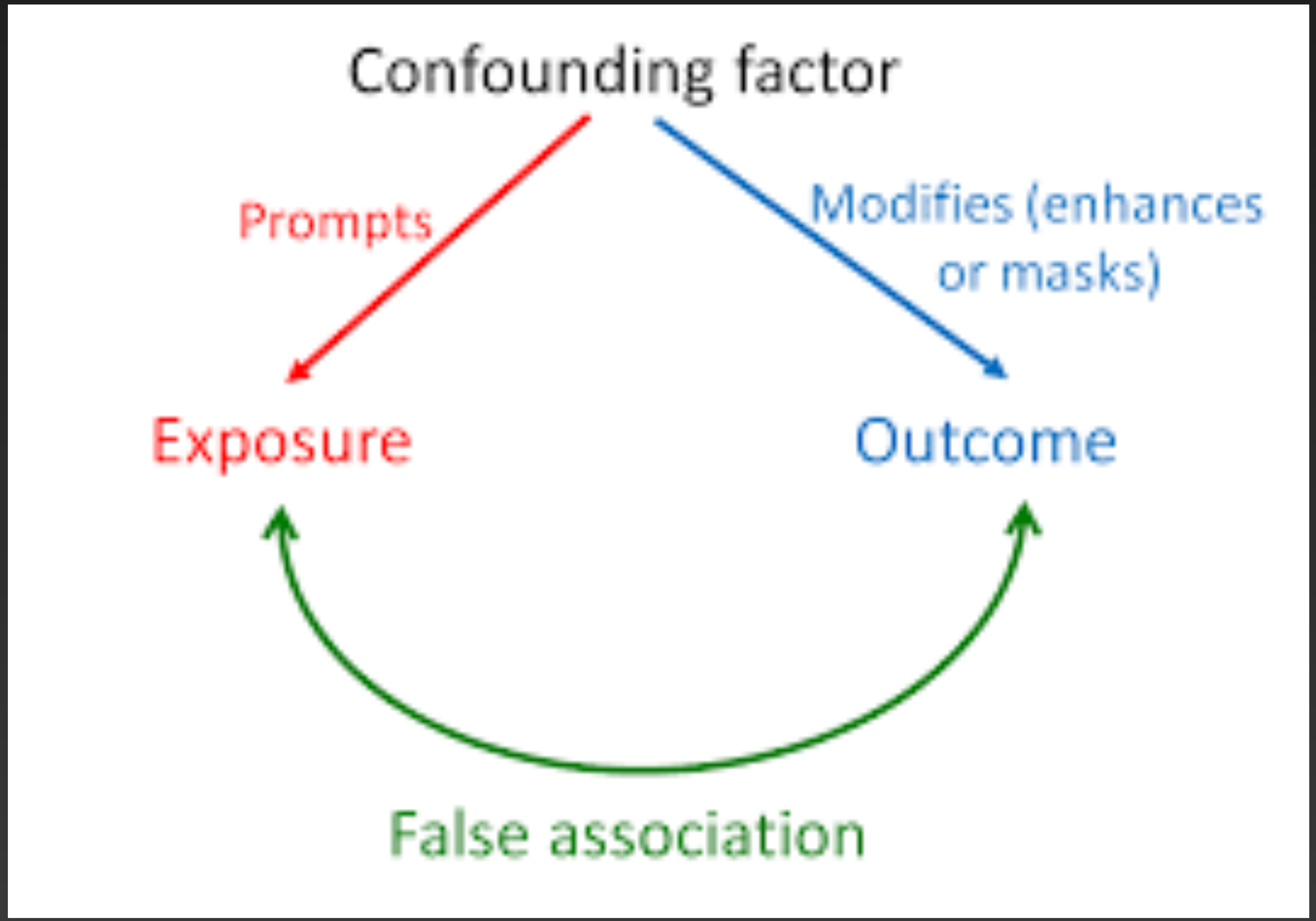
Introduction

- Inferring $X \to Y$ or $Y \to X$ from observations $(x_1, y_1), \ldots, (x_n, y_n)$ from f_{XY} is challenging.

Does $X \rightarrow Y$ or $Y \rightarrow X$?

Introduction

- Inferring $X \rightarrow Y$ or $Y \rightarrow X$ from observations $(x_1, y_1), \ldots, (x_n, y_n)$ from f_{XY} is challenging.
- Simplifying assumptions made:
  - Common confounders:

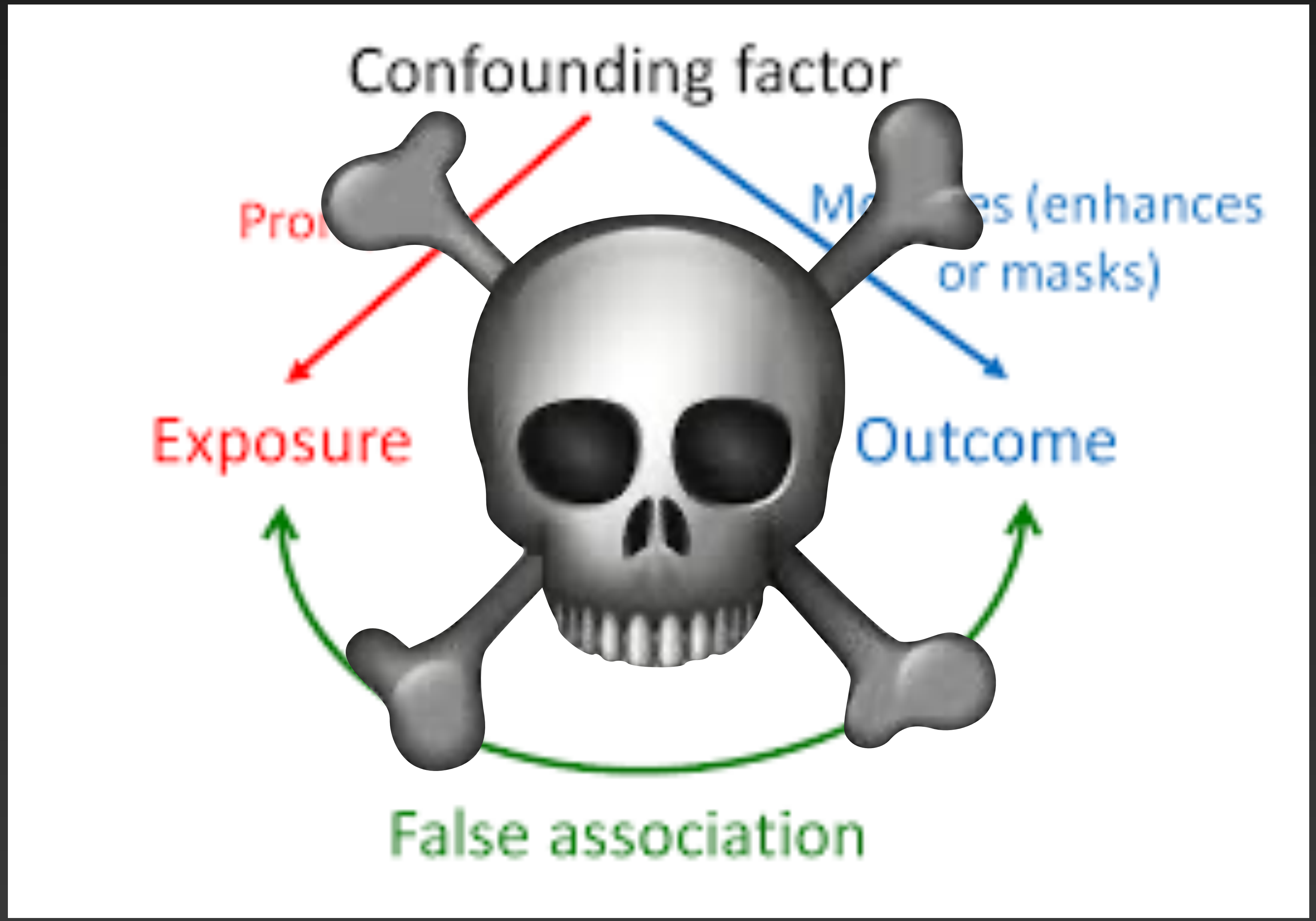


Does $X \rightarrow Y$ or $Y \rightarrow X$?

Introduction

- Inferring $X \rightarrow Y$ or $Y \rightarrow X$ from observations $(x_1, y_1), \ldots, (x_n, y_n)$ from f_{XY} is challenging.

- Simplifying assumptions made:
  - Common confounders: **not allowed!**

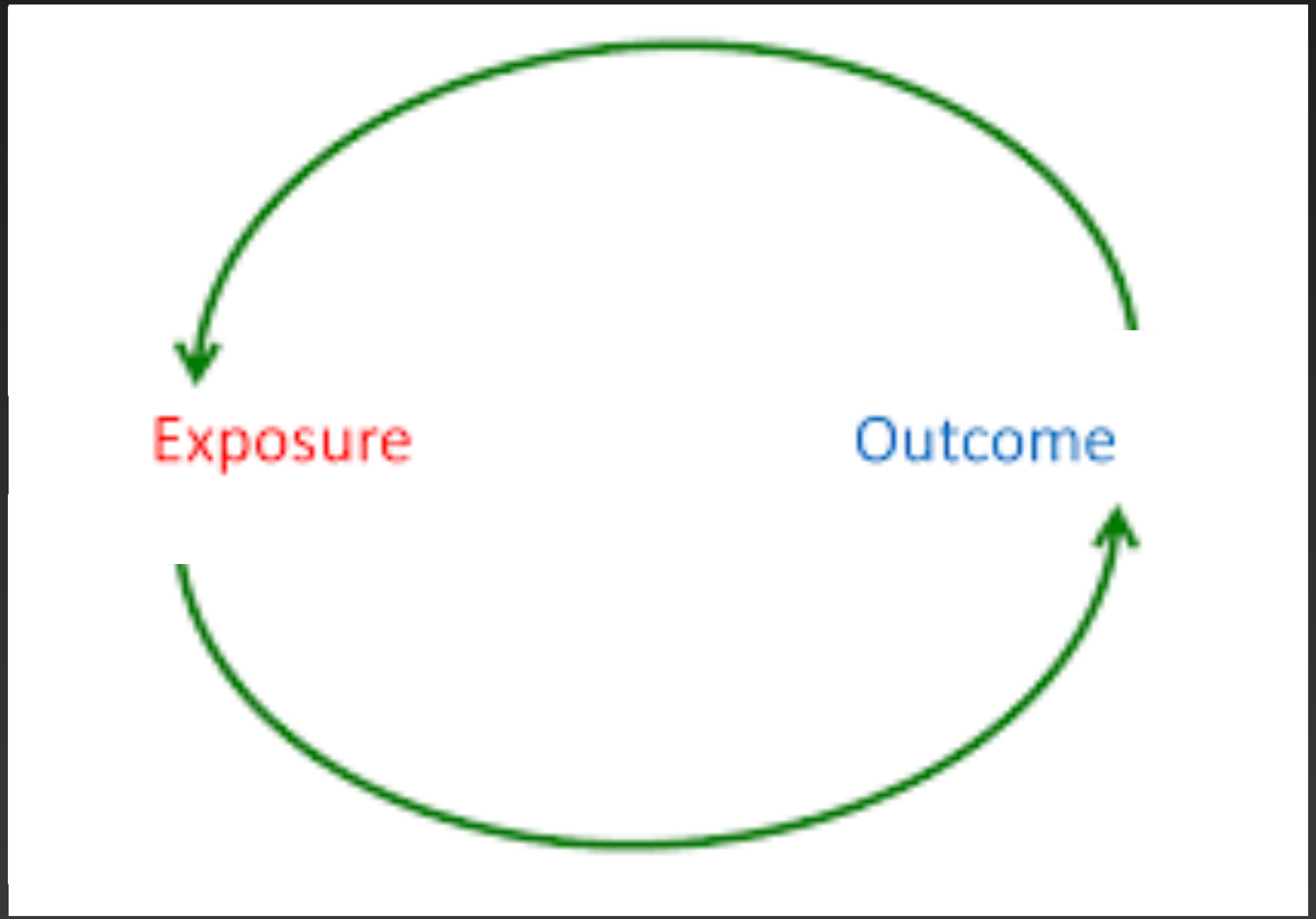


Does $X \rightarrow Y$ or $Y \rightarrow X$?

Introduction

- Inferring $X \rightarrow Y$ or $Y \rightarrow X$ from observations $(x_1, y_1), \dots, (x_n, y_n)$ from f_{XY} is challenging.

- Simplifying assumptions made:
  - Common confounders: **not allowed!**
  - Bidirectional influence:

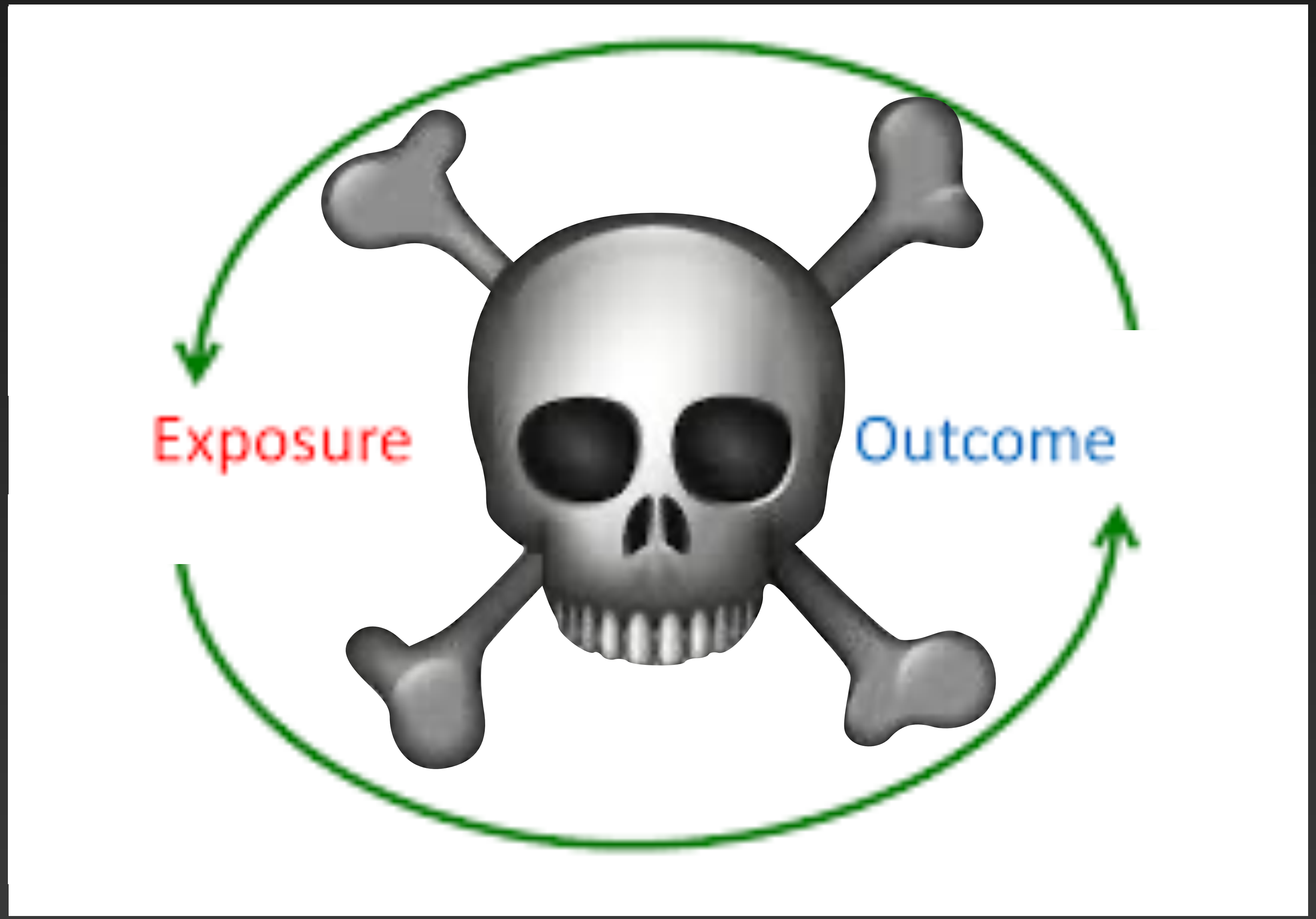


Does $X \to Y$ or $Y \to X$?

Introduction

- Inferring $X \to Y$ or $Y \to X$ from observations $(x_1, y_1), \ldots, (x_n, y_n)$ from f_{XY} is challenging.

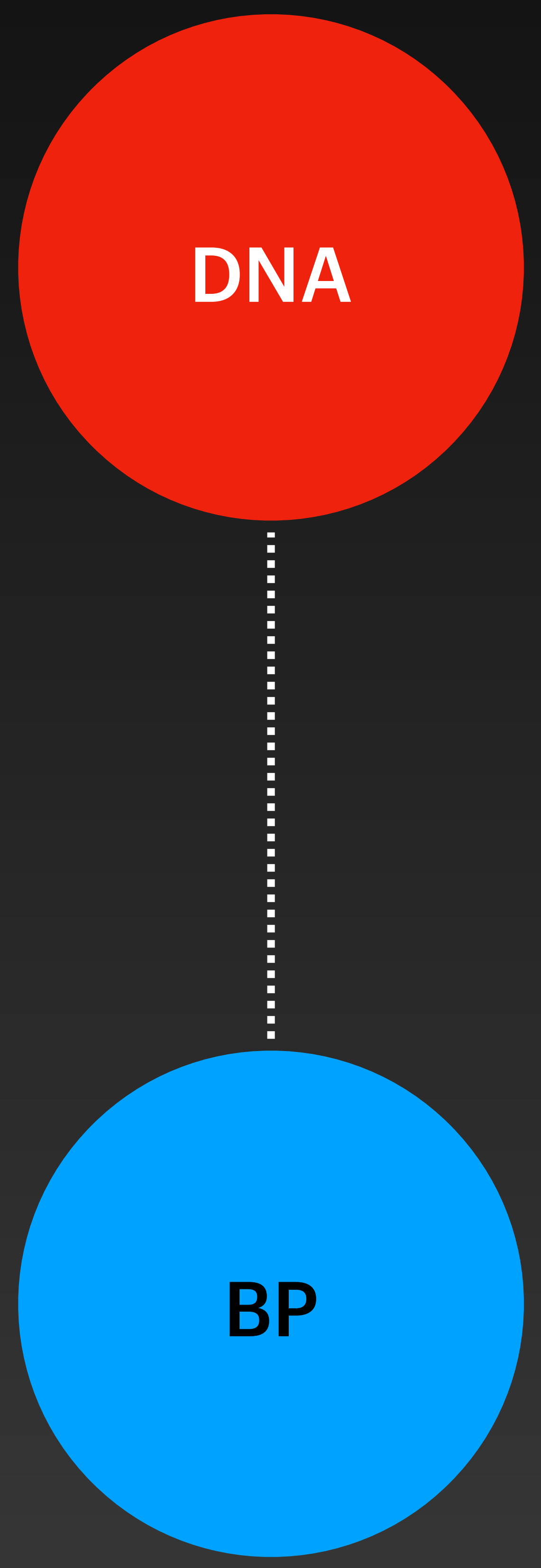
- Simplifying assumptions made:
  - Common confounders: **not allowed!**
  - Bidirectional influence: **not allowed!**



Methodology motivated by an epigenetic question

Does *DNAm* → *BP* or *BP* → *DNAm*?

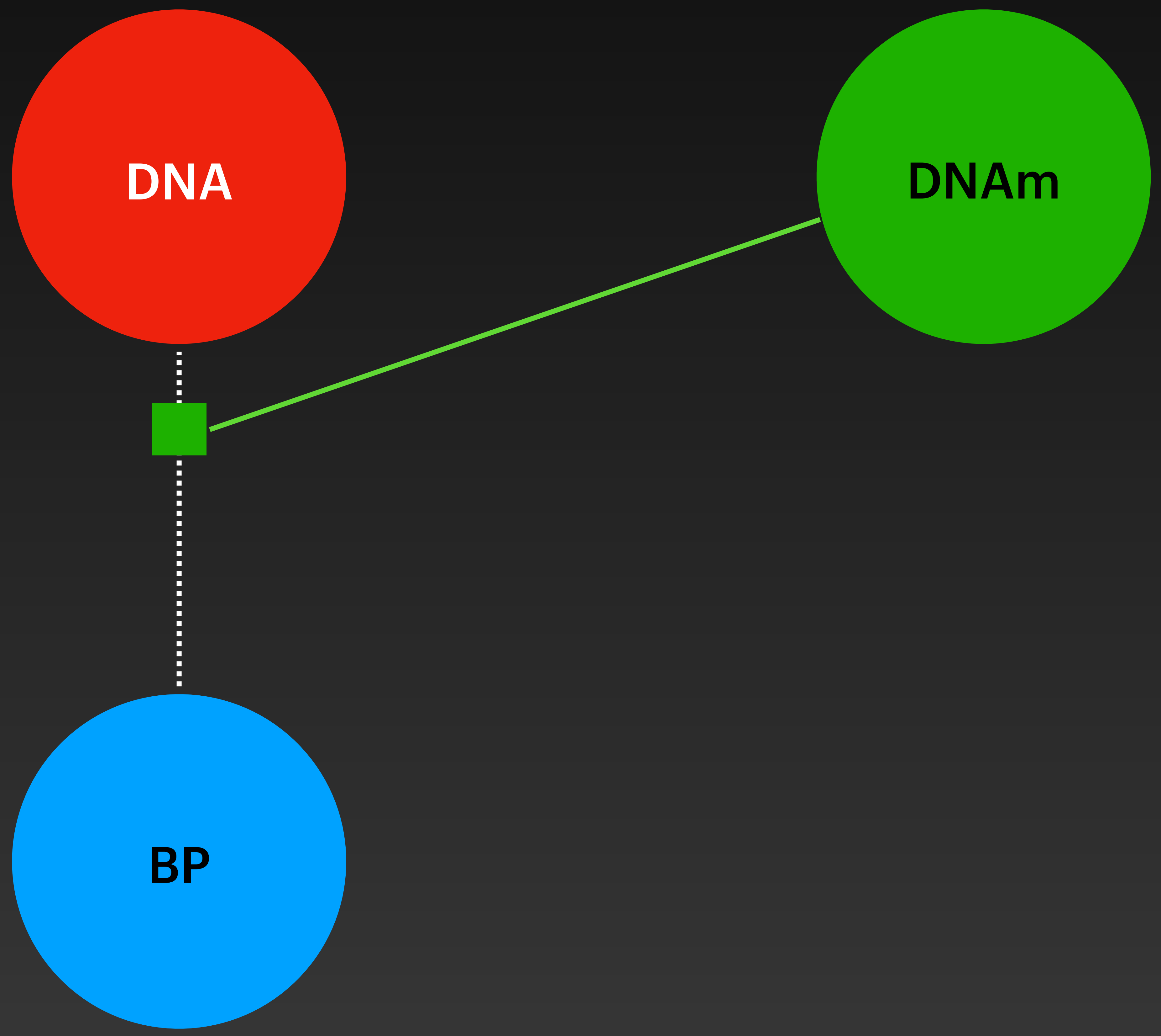
Early Life Exposures in Mexico to Environmental Toxicants (*ELEMENT*) cohort study



- Genes associated with blood pressure: ***ATP2B1***, ***FGF5***, and ***PRDM8***.

Does *DNAm* → *BP* or *BP* → *DNAm*?

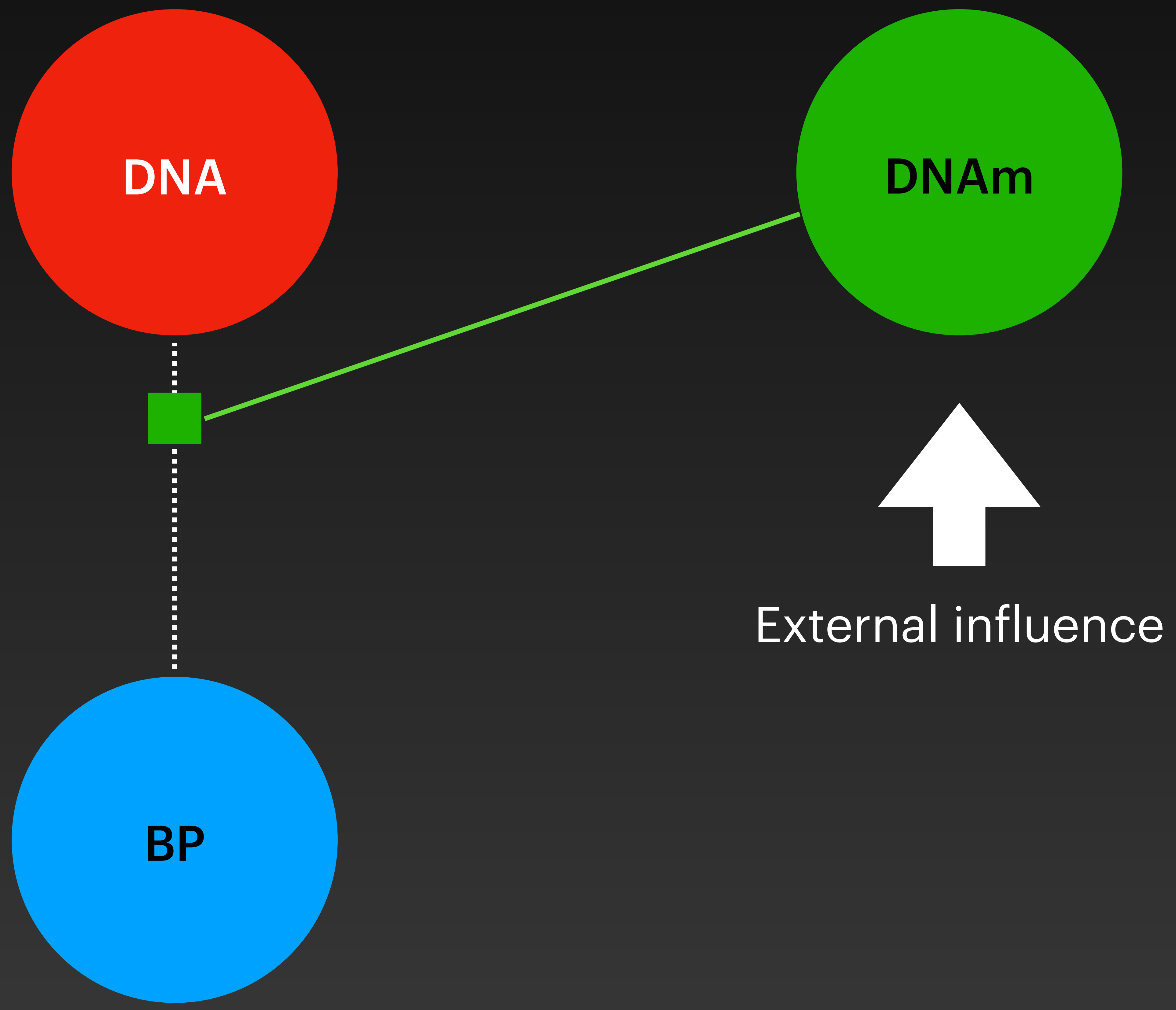
Early Life Exposures in Mexico to Environmental Toxicants (*ELEMENT*) cohort study



- Genes associated with blood pressure: ***ATP2B1***, ***FGF5***, and ***PRDM8***.
- Gene expression controlled by methylation.

Does *DNAm* → *BP* or *BP* → *DNAm*?

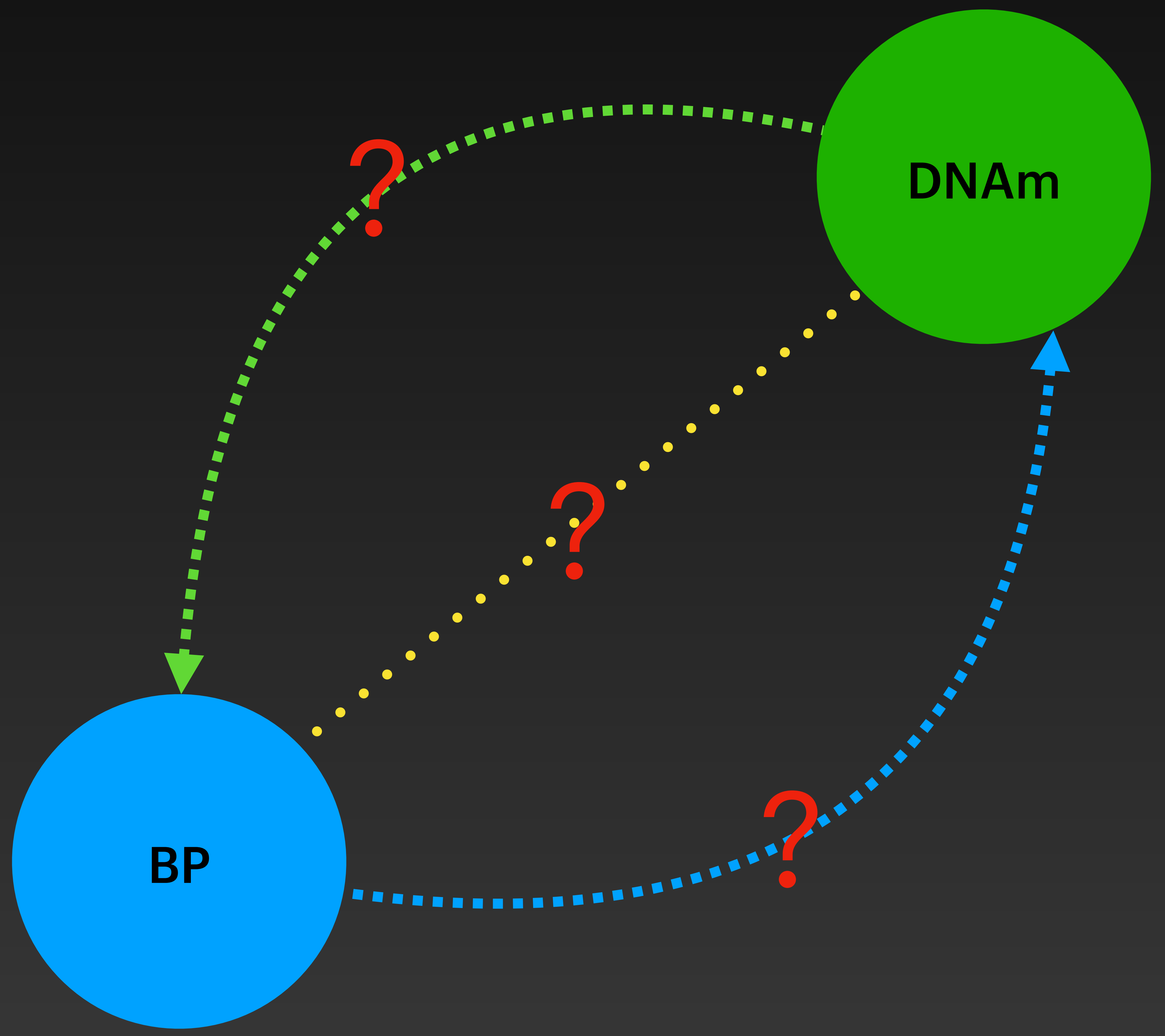
Early Life Exposures in Mexico to Environmental Toxicants (*ELEMENT*) cohort study



- Genes associated with blood pressure: ***ATP2B1***, ***FGF5***, and ***PRDM8***.
- Gene expression controlled by methylation.
- Methylation influenced by external features.

Does *DNAm* → *BP* or *BP* → *DNAm*?

Early Life Exposures in Mexico to Environmental Toxicants (*ELEMENT*) cohort study



- Genes associated with blood pressure: ***ATP2B1***, ***FGF5***, and ***PRDM8***.
- Gene expression controlled by methylation.
- Methylation influenced by external features.
- ***Questions:***

1. ***EpiGWAS for BP?***
2. ***DNAm → BP or BP → DNAm?***

Deliverables

Does $X \to Y$ or $Y \to X$?

Directed Mutual Information (DMI)

1. **Asymmetric predictability**: well-justified framework for studying statistical asymmetries between cause and effect.

Does $X \rightarrow Y$ or $Y \rightarrow X$?

Directed Mutual Information (DMI)

1. **Asymmetric predictability**: well-justified framework for studying statistical asymmetries between cause and effect.

2. New information theory-based measure: **Directed Mutual Information** (DMI).

   A. DMI can test for independence.

   B. DMI can quantify and estimate “asymmetries” between cause and effect.

Asymmetric predictability

Some information theoretic concepts

Entropy decomposition equation:

$$H(X, Y) = MI(X, Y) + H(X \mid Y) + H(Y \mid X)$$

Some information theoretic concepts

Entropy decomposition equation:

$$H(X,Y) = MI(X,Y) + H(X\,|\,Y) + H(Y\,|\,X)$$

Mutual information:

$$MI(X,Y) = E_{XY}\left[\log c_{XY}\right]$$

$$MI = 0 \iff X \perp Y$$

Symmetric!

Some information theoretic concepts

Entropy decomposition equation:

$$H(X,Y) = MI(X,Y) + H(X|Y) + H(Y|X)$$

Study independence with MI.

Depends only on copula c_{XY}

Mutual information:

$$MI(X,Y) = E_{XY}\left[\log c_{XY}\right]$$

$$MI = 0 \iff X \perp Y$$

Symmetric!

Some information theoretic concepts

Entropy decomposition equation:

$$H(X, Y) = MI(X, Y) + H(X \mid Y) + H(Y \mid X)$$

Joint entropy:

$$H(X, Y) = E_{XY}\left[-\log f_{XY}\right]$$

Marginal entropies $H(X)$ and $H(Y)$

Symmetric!

Some information theoretic concepts

Entropy decomposition equation:

$$H(X,Y) = MI(X,Y) + H(X\,|\,Y) + H(Y\,|\,X)$$

Total information
in (X, Y)

Joint entropy:

$$H(X,Y) = E_{XY}\left[-\log f_{XY}\right]$$

Marginal entropies $H(X)$ and $H(Y)$

Symmetric!

Some information theoretic concepts

Entropy decomposition equation:

$$H(X, Y) = MI(X, Y) + H(X \mid Y) + H(Y \mid X)$$

Conditional entropy:

$$H(X \mid Y) = E_{XY}\left[-\log \frac{f_{XY}}{f_X}\right]$$

Asymmetric!

Some information theoretic concepts

Entropy decomposition equation:

$$H(X, Y) = MI(X, Y) + H(X \mid Y) + H(Y \mid X)$$

Information needed to predict X if we know Y

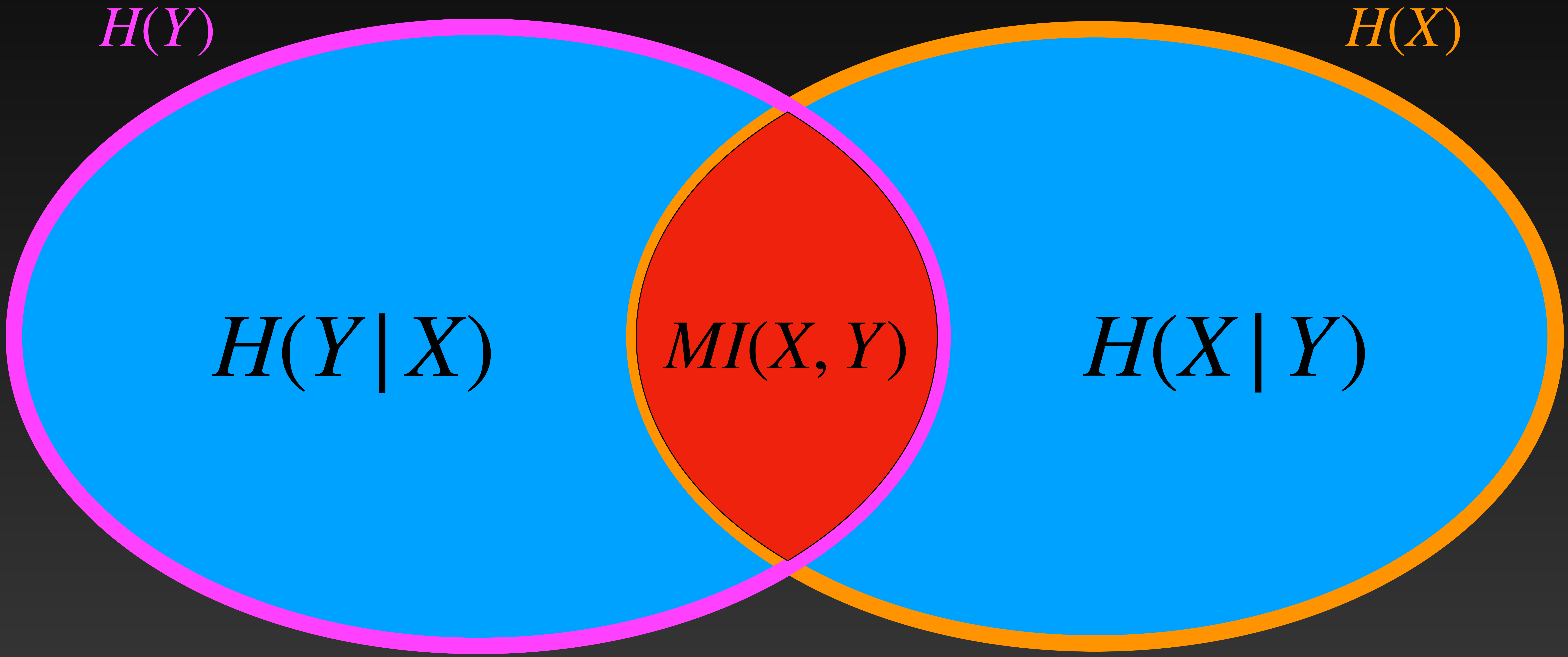
Conditional entropy:

$$H(X \mid Y) = E_{XY}\left[-\log \frac{f_{XY}}{f_X}\right]$$

Asymmetric!

Comparing conditional entropies

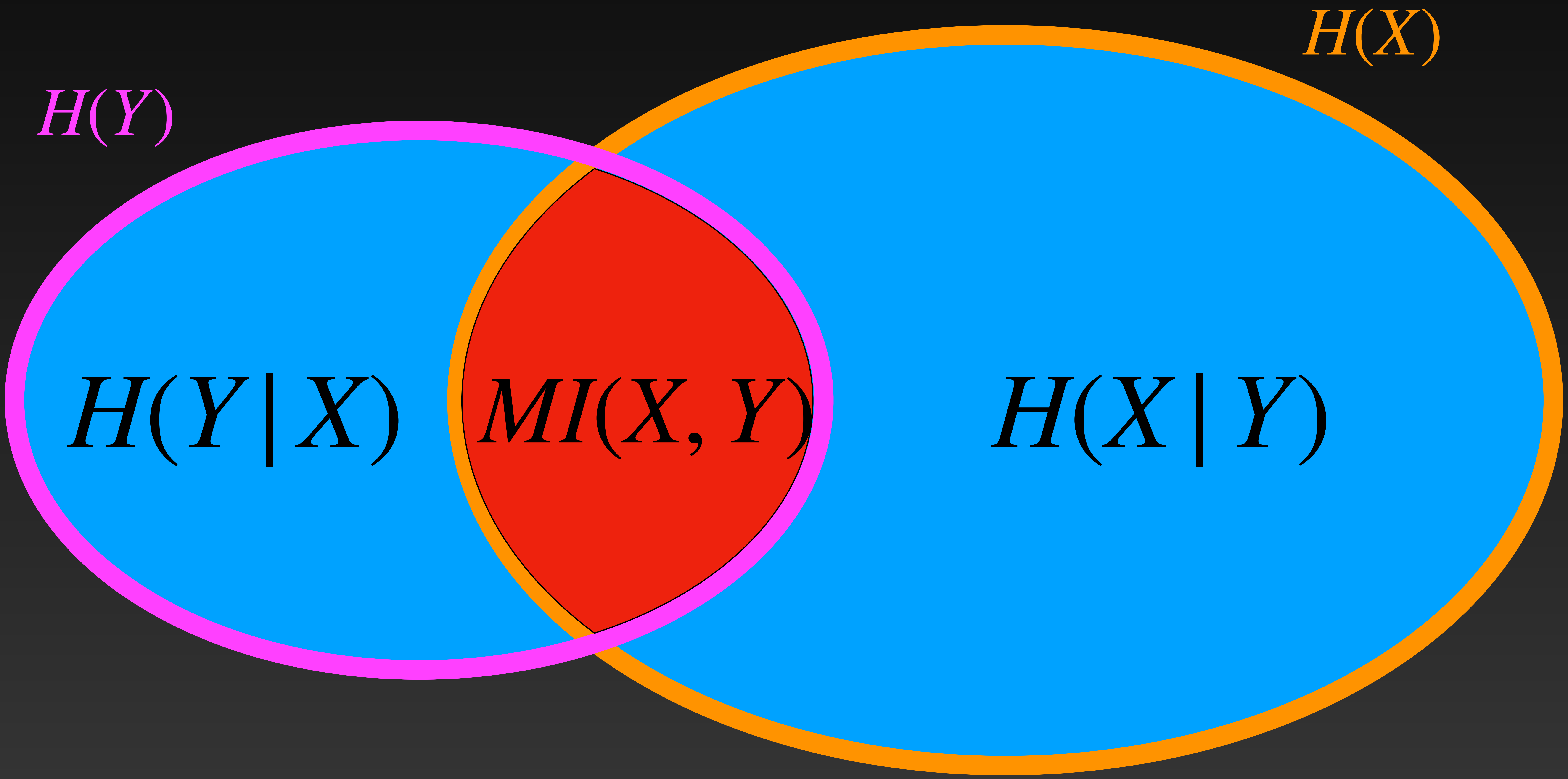
Symmetric entropy decomposition



Symmetric/balanced

Comparing conditional entropies

Asymmetric entropy decomposition



Asymmetry: “Less information needed to predict Y given X”

Asymmetric predictability

Causality in information space

Step 1: Consider $Y = g(X) + \epsilon$

Postulate: If $X \to Y$, the density f_X and the function g are "independent".

Asymmetric predictability

Causality in information space

Step 1: Consider $Y = g(X) + \epsilon$

Postulate: If $X \rightarrow Y$, the density f_X and the function g are “independent”.

Step 2: “peaks of f_X do not coincide with regions of large slope of g”

$$\text{Cov}\left[\log\left(g'\right), f_X\right] = 0$$

Asymmetric predictability

Causality in information space

Step 1: Consider $Y = g(X) + \epsilon$

Postulate: If $X \to Y$, the density f_X and the function g are "independent".

Step 2: "peaks of f_X do not coincide with regions of large slope of g"

$$\mathrm{Cov}\left[\log\left(g'\right), f_X\right] = 0$$

Step 3: after some more math...

If $X \to Y$, we have

$H(X \mid Y) = H(Y \mid X) +$ non-negative value

Asymmetric predictability

Causality in information space

Step 1: Consider $Y = g(X) + \epsilon$

Postulate: If $X \to Y$, the density f_X and the function g are "independent".

Step 2: "peaks of f_X do not coincide with regions of large slope of g"

$$\mathrm{Cov}\left[\log\left(g'\right), f_X\right] = 0$$

Step 3: after some more math...

If $X \to Y$, we have

$H(X \mid Y) = H(Y \mid X) +$ non-negative value

References:

Janzing, D., et al. (2012). ***Information-geometric approach to inferring causal directions.***
Daniusis, P., et al. (2012). ***Inferring deterministic causal relations.***

Towards defining DMI

Entropy ratio

Definition and properties

Entropy ratio compares $H(X|Y)$ and $H(Y|X)$:

$$ER(X|Y) = \frac{\exp\{H(X|Y)\}}{\exp\{H(X|Y)\} + \exp\{H(Y|X)\}}\,.$$

Entropy ratio

Definition and properties

Entropy ratio compares $H(X\,|\,Y)$ and $H(Y\,|\,X)$:

$$ER(X\,|\,Y) = \frac{\exp\{H(X\,|\,Y)\}}{\exp\{H(X\,|\,Y)\} + \exp\{H(Y\,|\,X)\}}.$$

$ER > 0$

Entropy ratio

Definition and properties

Entropy ratio compares $H(X\,|\,Y)$ and $H(Y\,|\,X)$:

$$ER(X\,|\,Y) = \frac{\exp\{H(X\,|\,Y)\}}{\exp\{H(X\,|\,Y)\} + \exp\{H(Y\,|\,X)\}}\,.$$

$ER > 0$

$$ER(X\,|\,Y) > ER(Y\,|\,X) \iff H(X\,|\,Y) > H(Y\,|\,X)$$

Directed Mutual Information (DMI)

Definition and properties

$$DMI(X \mid Y) = MI(X, Y) \times ER(X \mid Y)$$

Directed Mutual Information (DMI)

Definition and properties

Test for independence

$$DMI(X|Y) = MI(X,Y) \times ER(X|Y)$$

Properties:

1. $DMI = 0 \iff MI = 0 \iff X \perp Y$
2. $\Delta = DMI(X|Y) - DMI(Y|X) > 0 \iff H(X|Y) > H(Y|X) \iff X \to Y$

Directed Mutual Information (DMI)

Definition and properties

Detect asymmetry

$$DMI(X \mid Y) = MI(X, Y) \times ER(X \mid Y)$$

Properties:

1. $DMI = 0 \iff MI = 0 \iff X \perp Y$
2. $\Delta = DMI(X \mid Y) - DMI(Y \mid X) > 0 \iff H(X \mid Y) > H(Y \mid X) \iff X \to Y$

Directed Mutual Information (DMI)

Definition and properties

Test for independence

Detect asymmetry

$$DMI(X|Y) = MI(X,Y) \times ER(X|Y)$$

Properties:

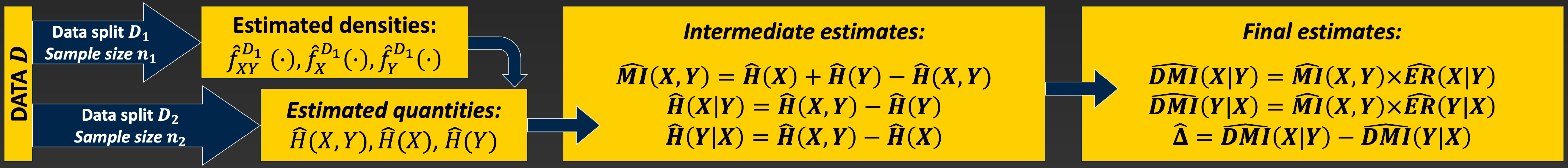
1. $DMI = 0 \iff MI = 0 \iff X \perp Y$
2. $\Delta = DMI(X|Y) - DMI(Y|X) > 0 \iff H(X|Y) > H(Y|X) \iff X \rightarrow Y$

Estimation and inference using DMI

Directed Mutual Information (DMI)

Estimation and inference

1. Estimate density functions (“nuisance parameter”) using one split
2. Evaluate entropy and mutual information using other split



DATA D

Data split D_1
Sample size n_1

Data split D_2
Sample size n_2

Estimated densities:
$\hat{f}_{XY}^{D_1}(\cdot), \hat{f}_{X}^{D_1}(\cdot), \hat{f}_{Y}^{D_1}(\cdot)$

Estimated quantities:
$\widehat{H}(X,Y), \widehat{H}(X), \widehat{H}(Y)$

Intermediate estimates:

$$\widehat{\boldsymbol{MI}}(\boldsymbol{X},\boldsymbol{Y}) = \widehat{\boldsymbol{H}}(\boldsymbol{X}) + \widehat{\boldsymbol{H}}(\boldsymbol{Y}) - \widehat{\boldsymbol{H}}(\boldsymbol{X},\boldsymbol{Y})$$
$$\widehat{\boldsymbol{H}}(\boldsymbol{X}|\boldsymbol{Y}) = \widehat{\boldsymbol{H}}(\boldsymbol{X},\boldsymbol{Y}) - \widehat{\boldsymbol{H}}(\boldsymbol{Y})$$
$$\widehat{\boldsymbol{H}}(\boldsymbol{Y}|\boldsymbol{X}) = \widehat{\boldsymbol{H}}(\boldsymbol{X},\boldsymbol{Y}) - \widehat{\boldsymbol{H}}(\boldsymbol{X})$$

Final estimates:

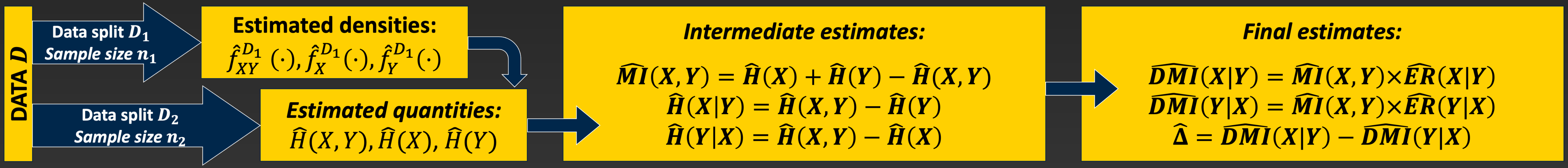
$$\widehat{\boldsymbol{DMI}}(\boldsymbol{X}|\boldsymbol{Y}) = \widehat{\boldsymbol{MI}}(\boldsymbol{X},\boldsymbol{Y}) \times \widehat{\boldsymbol{ER}}(\boldsymbol{X}|\boldsymbol{Y})$$
$$\widehat{\boldsymbol{DMI}}(\boldsymbol{Y}|\boldsymbol{X}) = \widehat{\boldsymbol{MI}}(\boldsymbol{X},\boldsymbol{Y}) \times \widehat{\boldsymbol{ER}}(\boldsymbol{Y}|\boldsymbol{X})$$
$$\widehat{\boldsymbol{\Delta}} = \widehat{\boldsymbol{DMI}}(\boldsymbol{X}|\boldsymbol{Y}) - \widehat{\boldsymbol{DMI}}(\boldsymbol{Y}|\boldsymbol{X})$$

Directed Mutual Information (DMI)

Estimation and inference

1. Estimate density functions (“nuisance parameter”) using one split
2. Evaluate entropy and mutual information using other split

Sample splitting eliminates bias due to nuisance parameter estimation.



DATA D

Data split D_1
Sample size n_1

Data split D_2
Sample size n_2

Estimated densities:
$\hat{f}_{XY}^{D_1}(\cdot), \hat{f}_{X}^{D_1}(\cdot), \hat{f}_{Y}^{D_1}(\cdot)$

Estimated quantities:
$\hat{H}(X,Y), \hat{H}(X), \hat{H}(Y)$

Intermediate estimates:

$$\widehat{\boldsymbol{MI}}(\boldsymbol{X},\boldsymbol{Y}) = \hat{\boldsymbol{H}}(\boldsymbol{X}) + \hat{\boldsymbol{H}}(\boldsymbol{Y}) - \hat{\boldsymbol{H}}(\boldsymbol{X},\boldsymbol{Y})$$
$$\hat{\boldsymbol{H}}(\boldsymbol{X}|\boldsymbol{Y}) = \hat{\boldsymbol{H}}(\boldsymbol{X},\boldsymbol{Y}) - \hat{\boldsymbol{H}}(\boldsymbol{Y})$$
$$\hat{\boldsymbol{H}}(\boldsymbol{Y}|\boldsymbol{X}) = \hat{\boldsymbol{H}}(\boldsymbol{X},\boldsymbol{Y}) - \hat{\boldsymbol{H}}(\boldsymbol{X})$$

Final estimates:

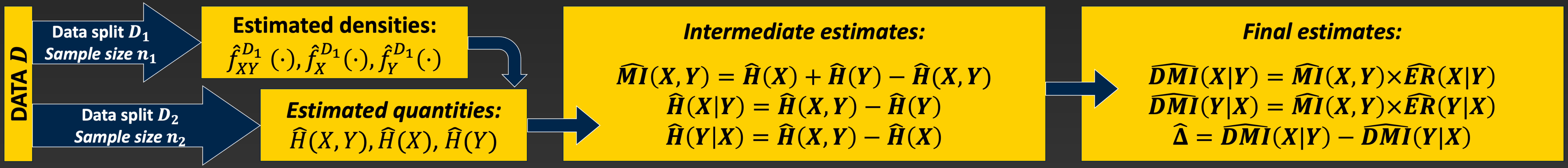
$$\widehat{\boldsymbol{DMI}}(\boldsymbol{X}|\boldsymbol{Y}) = \widehat{\boldsymbol{MI}}(\boldsymbol{X},\boldsymbol{Y})\times\widehat{\boldsymbol{ER}}(\boldsymbol{X}|\boldsymbol{Y})$$
$$\widehat{\boldsymbol{DMI}}(\boldsymbol{Y}|\boldsymbol{X}) = \widehat{\boldsymbol{MI}}(\boldsymbol{X},\boldsymbol{Y})\times\widehat{\boldsymbol{ER}}(\boldsymbol{Y}|\boldsymbol{X})$$
$$\hat{\boldsymbol{\Delta}} = \widehat{\boldsymbol{DMI}}(\boldsymbol{X}|\boldsymbol{Y}) - \widehat{\boldsymbol{DMI}}(\boldsymbol{Y}|\boldsymbol{X})$$

Directed Mutual Information (DMI)

Estimation and inference

1. Estimate density functions (“nuisance parameter”) using one split
2. Evaluate entropy and mutual information using other split

Cross-fitting yields improved empirical performance.



DATA D

Data split D_1
Sample size n_1

Data split D_2
Sample size n_2

Estimated densities:
$\hat{f}_{XY}^{D_1}(\cdot), \hat{f}_X^{D_1}(\cdot), \hat{f}_Y^{D_1}(\cdot)$

Estimated quantities:
$\hat{H}(X,Y), \hat{H}(X), \hat{H}(Y)$

Intermediate estimates:

$$\widehat{MI}(X,Y) = \hat{H}(X) + \hat{H}(Y) - \hat{H}(X,Y)$$
$$\hat{H}(X|Y) = \hat{H}(X,Y) - \hat{H}(Y)$$
$$\hat{H}(Y|X) = \hat{H}(X,Y) - \hat{H}(X)$$

Final estimates:

$$\widehat{DMI}(X|Y) = \widehat{MI}(X,Y) \times \widehat{ER}(X|Y)$$
$$\widehat{DMI}(Y|X) = \widehat{MI}(X,Y) \times \widehat{ER}(Y|X)$$
$$\hat{\Delta} = \widehat{DMI}(X|Y) - \widehat{DMI}(Y|X)$$

Directed Mutual Information (DMI)

Theoretical guarantees

1. Assuming the density functions are bounded, when $\min(n_1, n_2) \to \infty$, we have $D\hat{M}I(X \mid Y) \xrightarrow{p} DMI(X \mid Y)$.

2. Assuming $MI \neq 0$, $\min(n_1, n_2) \to \infty$, we have $\sqrt{n_2}\left(\hat{\Delta} - \Delta\right) \xrightarrow{D} N(0,\sigma_{\Delta}^2)$.

Directed Mutual Information (DMI)

Theoretical guarantees

Consistent estimates of DMI permit test of independence using permutation.

1. Assuming the density functions are bounded, when $\min(n_1, n_2) \to \infty$, we have $D\hat{M}I(X \mid Y) \xrightarrow{p} DMI(X \mid Y)$.

2. Assuming $MI \neq 0$, $\min(n_1, n_2) \to \infty$, we have $\sqrt{n_2}\left(\hat{\Delta} - \Delta\right) \xrightarrow{D} N(0, \sigma_\Delta^2)$.

Directed Mutual Information (DMI)

Theoretical guarantees

1. Assuming the density functions are bounded, when $\min(n_1, n_2) \to \infty$, we have $D\hat{M}I(X|Y) \xrightarrow{p} DMI(X|Y)$.

Sign of $\hat{\Delta}$ informs $X \to Y$

(95%) **CI allows for calibration**

2. Assuming $MI \neq 0$, $\min(n_1, n_2) \to \infty$, we have $\sqrt{n_2}\left(\hat{\Delta} - \Delta\right) \xrightarrow{D} N(0,\sigma_{\Delta}^2)$.

New epigenetic insights using DMI

Does *DNAm* → BP or BP → *DNAm*

Application of DMI to methylation studies

- Cohort of 525 children of age 10 - 18 years in the ELEMENT cohort.

- 3 candidate genes: ***ATP2B1, FGF5,*** and ***PRDM8.***
  - Mildly correlated methylation sites: 21 for ***ATP2B1***, 21 for ***FGF5***, and 51 sites for ***PRDM8***.

Does *DNAm* → BP or BP → *DNAm*

Application of DMI to methylation studies

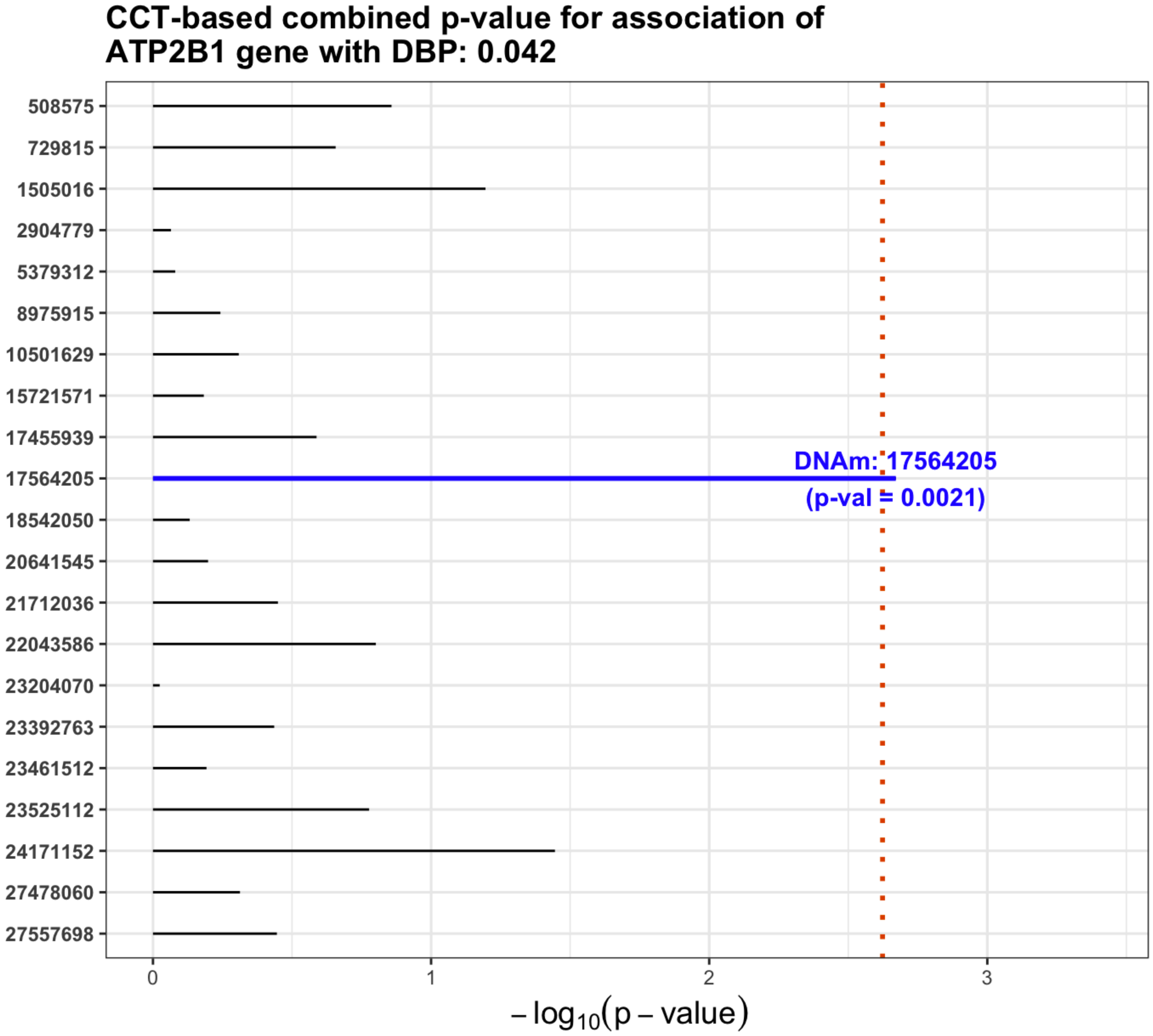
- Cohort of 525 children of age 10 - 18 years in the ELEMENT cohort.

- 3 candidate genes: ***ATP2B1, FGF5,*** and ***PRDM8.***
    - Mildly correlated methylation sites: 21 for ***ATP2B1***, 21 for ***FGF5***, and 51 sites for ***PRDM8***.

1. Test for association between *DNAm* site and BP for a given gene.
2. Aggregate all *DNAm*-findings using Cauchy combination test for each gene.
3. In *DNAm* sites associated with BP, check if *DNAm* → BP or BP → *DNAm*.

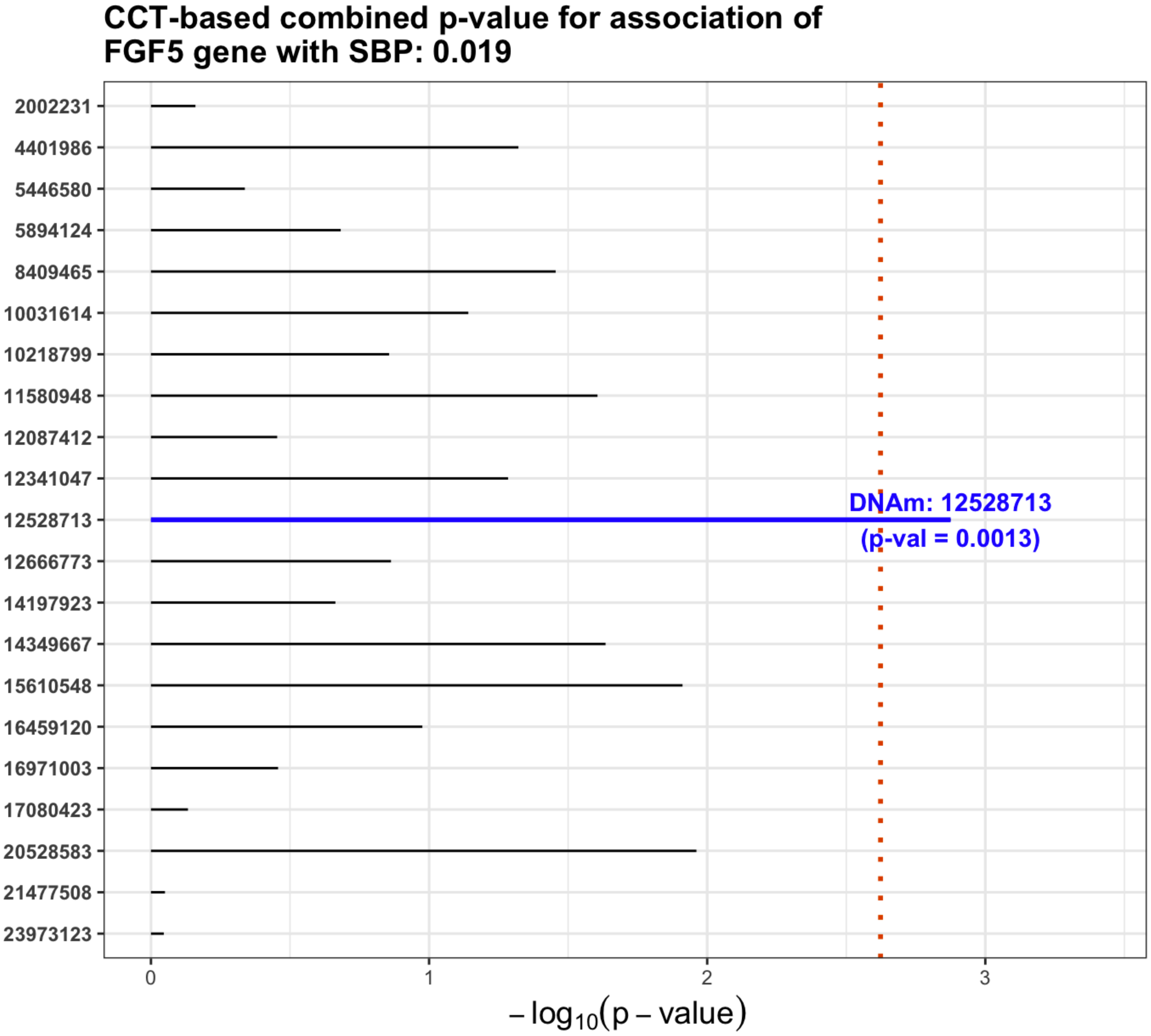
Finding 1: DNAm site #17564205 in *ATP2B1*

- Strongly associated with **diastolic** BP.
- Strong signal to drive gene-wide association with DBP.
- Next: check *DNAm* → BP or BP → *DNAm?*



Finding 2: DNAm site #125287 in *FGF5*

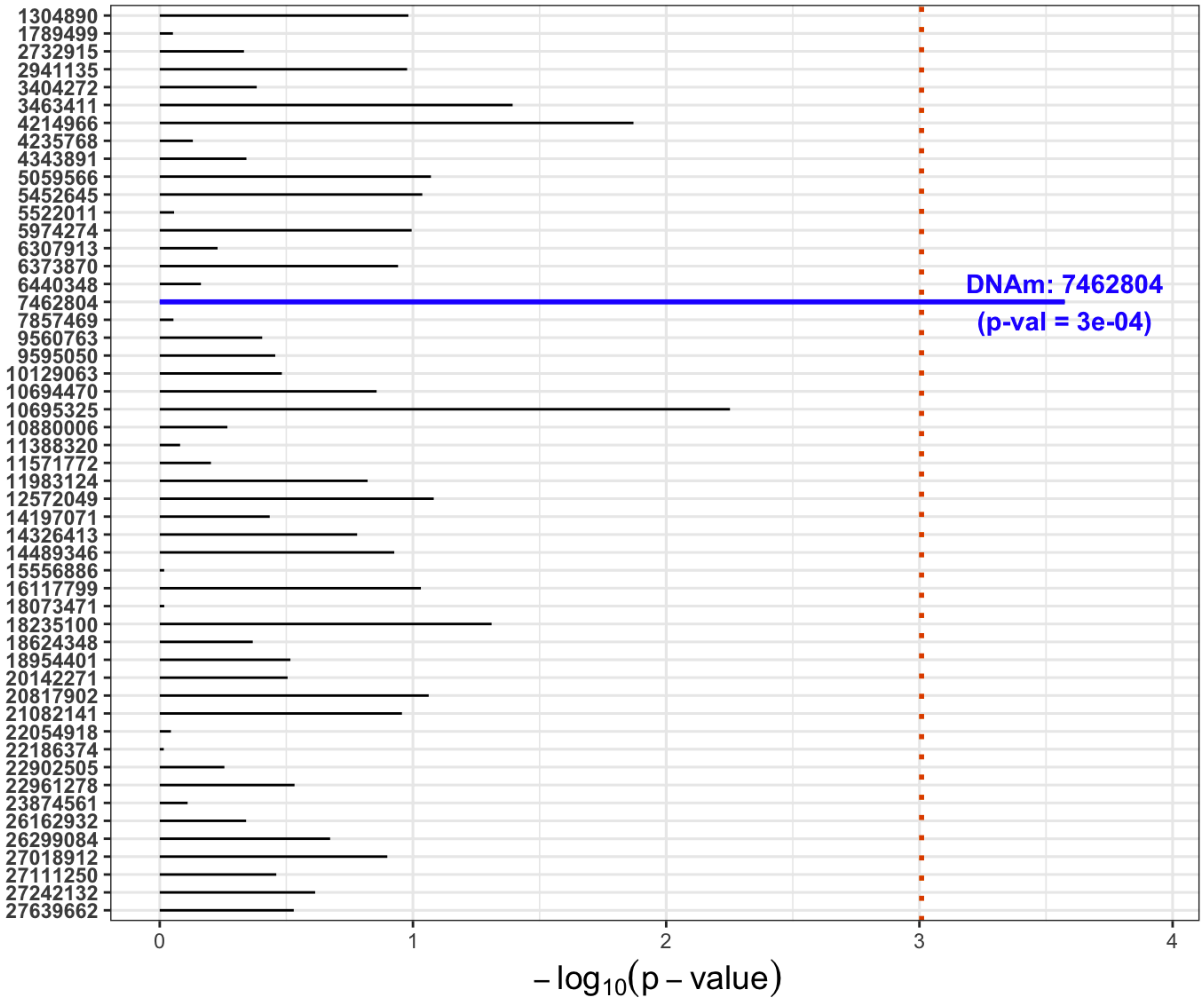
- Strongly associated with **systolic** BP.
- Strong signal to drive gene-wide association with SBP.
- Next: check *DNAm* → BP or BP → *DNAm?*



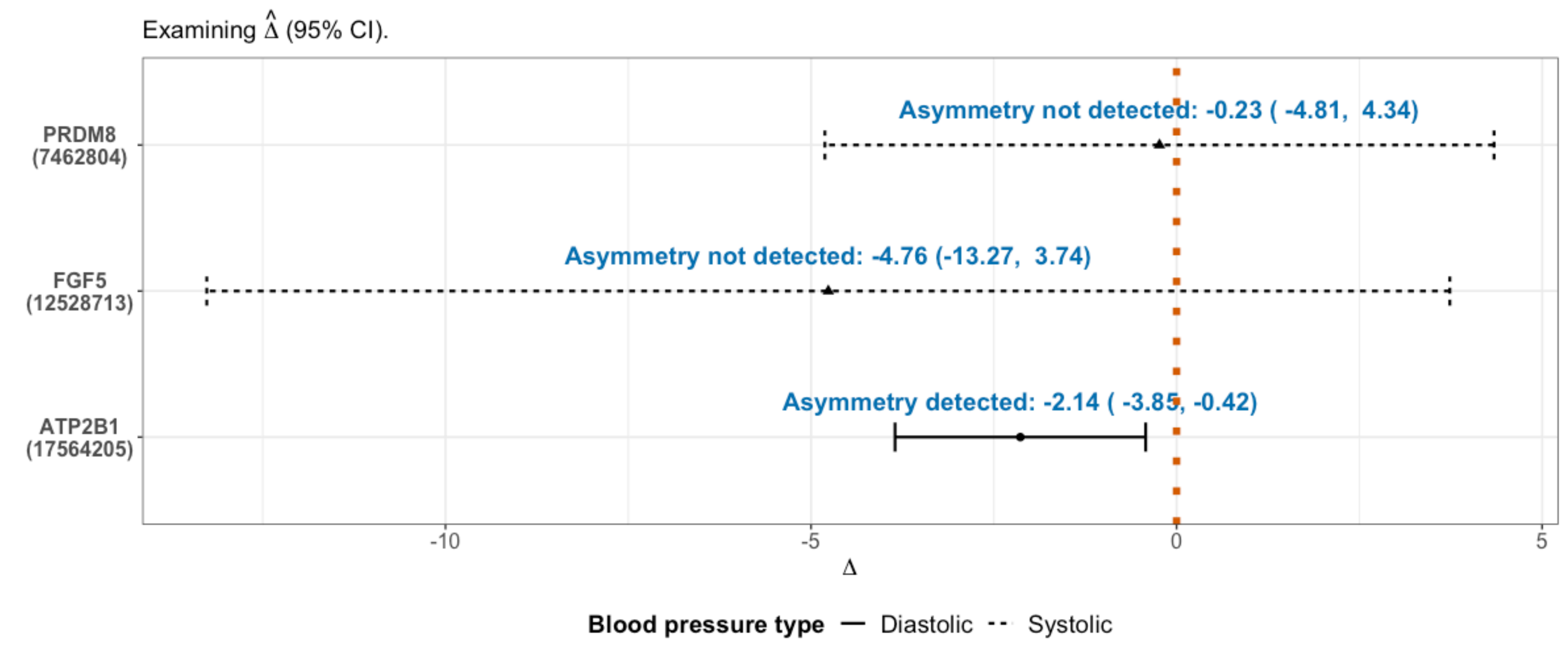
Finding 3: DNAm site #7462804 in *PRDM8*

- Strongly associated with **systolic** BP.
- Strong signal to drive gene-wide association with SBP.
- Next: check *DNAm* → BP or BP → *DNAm?*

CCT-based combined p-value for association of PRDM8 gene with SBP: 0.012



Finding 4: DBP → #17564205 in *ATP2B1*



1. No asymmetry detected: ***#7462804*** (PRDM8) and ***#12528713*** (FGF5)
2. Asymmetry detected: **Diastolic BP → #17564205** in ***ATP2B1.***

Thank you for your time!

soumikp@umich.edu

References

- **ELEMENT study reference**
  - Perng, Wei, et al. "Early life exposure in Mexico to environmental toxicants (ELEMENT) project." BMJ open 9.8 (2019): e030427.
- **Motivation of entropy ratio**
  - Varin, Cristiano, Manuela Cattelan, and David Firth. "Statistical modelling of citation exchange between statistics journals." Journal of the Royal Statistical Society. Series A,(Statistics in Society) 179.1 (2016): 1.
- **Information geometric causal inference**
  - Janzing, D., et al. (2012). Information-geometric approach to inferring causal directions. Artificial Intelligence, 182–183, 1–31. https://doi.org/10.1016/j.artint.2012.01.002
  - Daniusis, P., et al. (2012). Inferring deterministic causal relations. arXiv. https://doi.org/10.48550/arXiv.1203.3475
- **Sample splitting and cross-fitting**
  - Zivich, P. N., & Breskin, A. (2021). Machine learning for causal inference: on the use of cross-fit estimators. Epidemiology (Cambridge, Mass.), 32(3), 393.