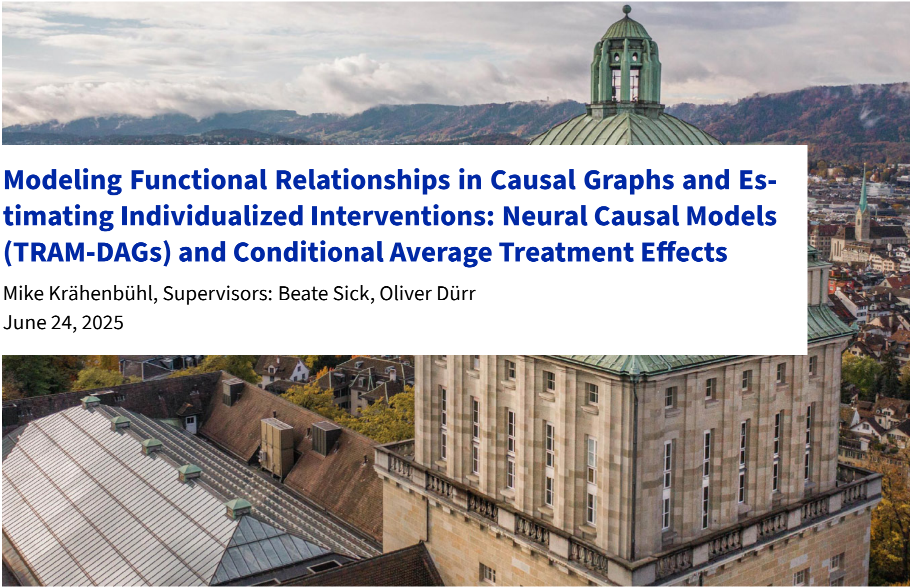


Modeling Functional Relationships in Causal Graphs and Estimating Individualized Interventions: Neural Causal Models (TRAM-DAGs) and Conditional Average Treatment Effects

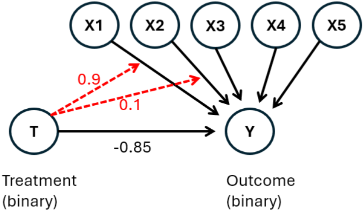
Mike Krähenbühl, Supervisors: Beate Sick, Oliver Dürr

June 24, 2025

Simulation Case 1: Fully Observed

Setup:

- $n = 20{,}000$
- $T \sim \text{Bernoulli}(0.5)$
- $\mathbf{X} = (X_1, \ldots, X_5)^\top \sim \mathcal{N}(\mathbf{0}, \Sigma)$
- $\mathbf{X_{TX}} = (X_1, X_2)^\top$ interaction

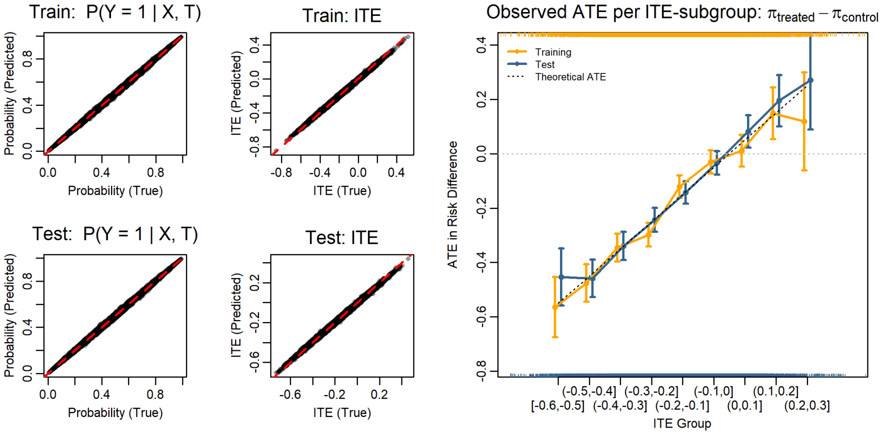


Outcome model:

$$\mathbb{P}(Y = 1 \mid \mathbf{X}, T) = \text{logit}^{-1}\left(\beta_0 + \beta_T T + \boldsymbol{\beta}_X^\top \mathbf{X} + T \cdot \boldsymbol{\beta}_{TX}^\top \mathbf{X_{TX}}\right)$$

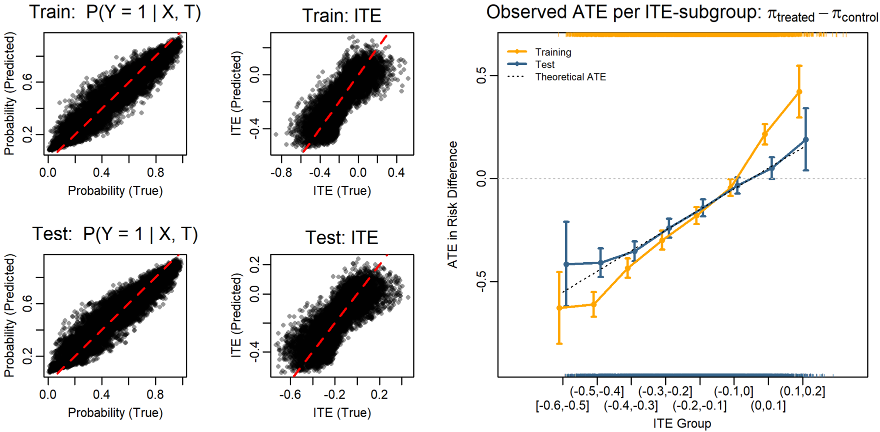
Simulation Case 1: Fully Observed

Results with T-learner logistic regression (glm):



Simulation Case 1: Fully Observed

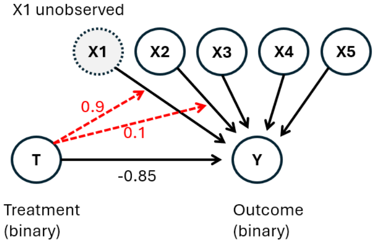
Results with T-learner Random Forest (comets package):



Simulation Case 2: Unobserved Interaction

Setup:

- $n = 20{,}000$
- $T \sim \text{Bernoulli}(0.5)$
- $\mathbf{X} = (X_1, \ldots, X_5)^\top \sim \mathcal{N}(\mathbf{0}, \Sigma)$
- $\mathbf{X_{TX}} = (X_1, X_2)^\top$ interaction



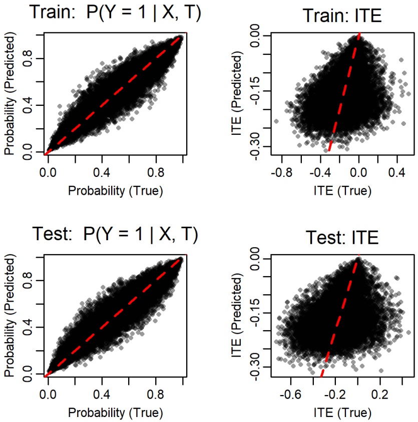
Outcome model:

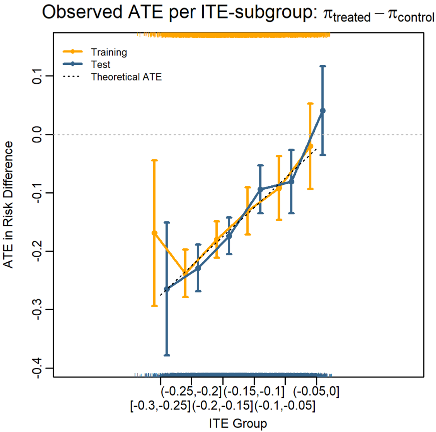
$$\mathbb{P}(Y = 1 \mid \mathbf{X}, T) = \text{logit}^{-1}\left(\beta_0 + \beta_T T + \boldsymbol{\beta}_X^\top \mathbf{X} + T \cdot \boldsymbol{\beta}_{TX}^\top \mathbf{X_{TX}}\right)$$

Note: Same DGP, but X_1 is not observed!

Simulation Case 2: Unobserved Interaction

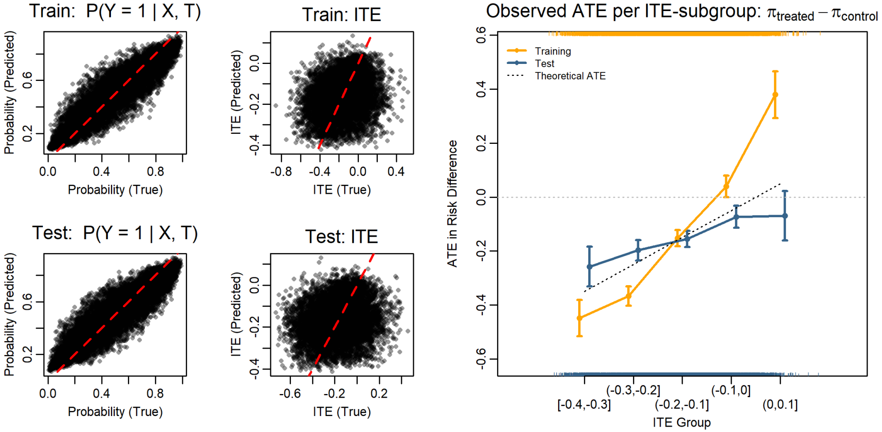
Results with T-learner logistic regression (glm):





Simulation Case 2: Unobserved Interaction

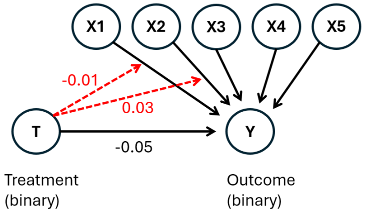
Results with T-learner Random Forest (comets):



Simulation Case 3: Fully Observed, Small Effects

Setup:

- $n = 20{,}000$
- $T \sim \text{Bernoulli}(0.5)$
- $\mathbf{X} = (X_1, \ldots, X_5)^\top \sim \mathcal{N}(\mathbf{0}, \Sigma)$
- $\mathbf{X_{TX}} = (X_1, X_2)^\top$ interaction



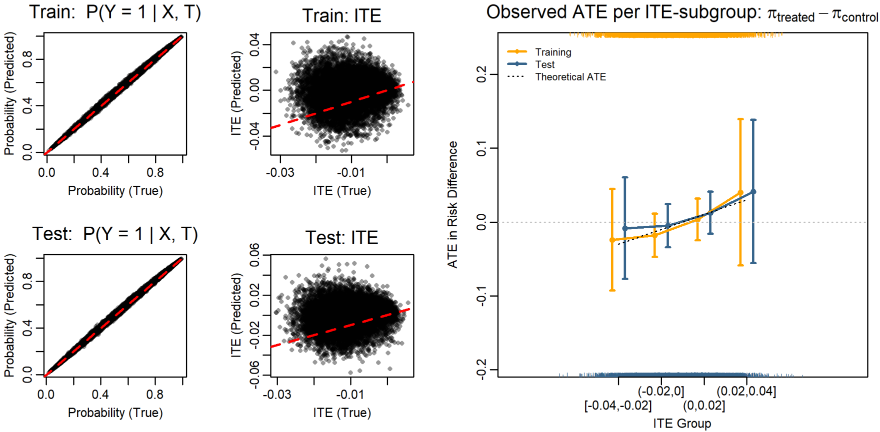
Outcome model:

$$\mathbb{P}(Y = 1 \mid \mathbf{X}, T) = \text{logit}^{-1}\left(\beta_0 + \beta_T T + \boldsymbol{\beta}_X^\top \mathbf{X} + T \cdot \boldsymbol{\beta}_{TX}^\top \mathbf{X_{TX}}\right)$$

Note: Same DGP, but weak treatment effects!

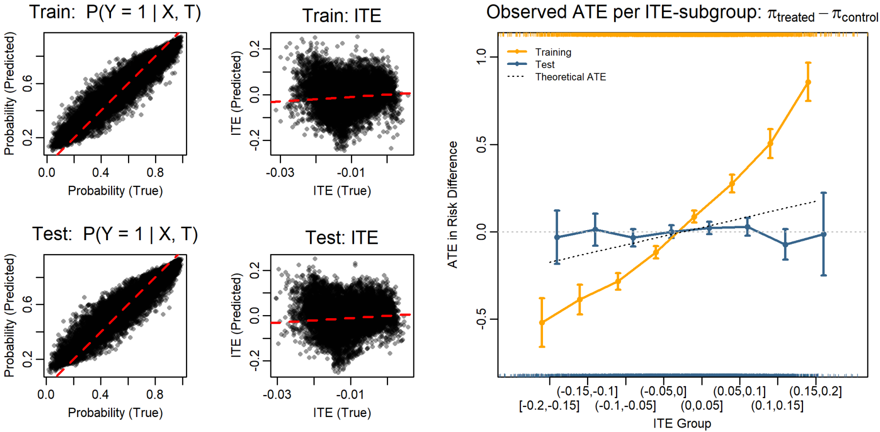
Simulation Case 3: Fully Observed, Small Effects

Results with T-learner logistic regression (glm):



Simulation Case 3: Fully Observed, Small Effects

Results with T-learner Random Forest (comets package):



ITE simulation: Interpretation

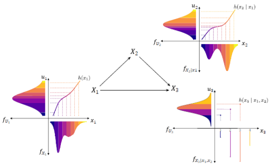
My interpretation:

- When a high predicted treatment effect (ITE) corresponds to a high observed effect in the train set (strong discrimination), but not in the test set, it might be due to **unobserved interaction variables** or **weak treatment effects**.
- This is more likely to occur with complex models, as they tend to overfit when the interaction is not observed.

TRAM-DAGs for ITE Estimation

Paper *"Interpretable Neural Causal Models with TRAM-DAGs"* (Sick and Dürr, 2025):

- Framework to model causal relationships
- Based on transformation models
- Rely on (deep) neural networks
- Compromise between interpretability and flexibility



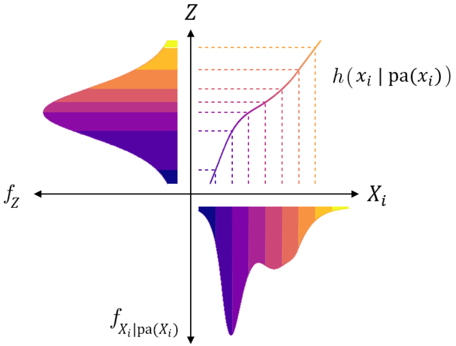
Our Claim: We can use TRAM-DAGs for ITE estimation, as long as the DAG is known and fully observed!

TRAM-DAGs: Structural Equations

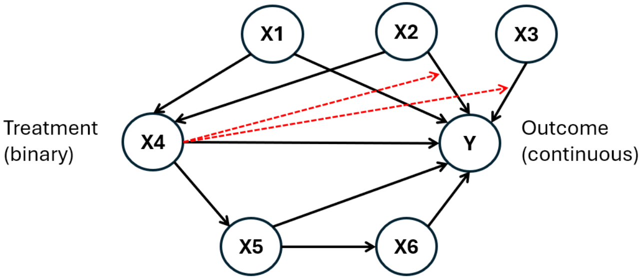
TRAM-DAGs estimate the structural equations with transformation functions h_i:

$$Z_i = h_i(X_i \mid \mathrm{pa}(X_i))$$
$$X_i = h_i^{-1}(Z_i, \mathrm{pa}(X_i)) = f_i(Z_i, \mathrm{pa}(X_i))$$

- $\mathrm{pa}(X_i)$: causal parents of X_i
- Z_i: noise distribution (e.g. standard logistic)



TRAM-DAGs: Example for ITE estimation

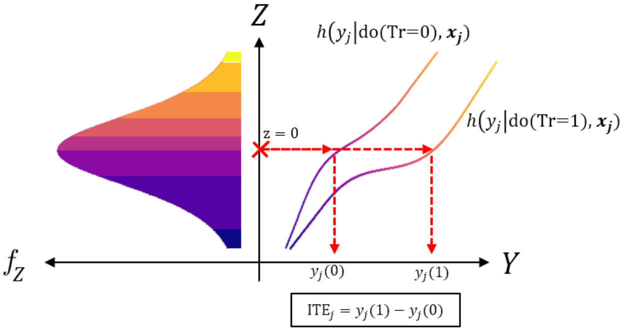


DGP:

- $X5 = h_5^{-1}(\epsilon - 0.8\,X4)$ $\rightarrow$ (depends on treatment)
- $X6 = h_6^{-1}(\epsilon + 0.5\,X5)$ $\rightarrow$ (depends on treatment through X5)
- $Y = h_7^{-1}(\epsilon - \beta_1 X1 - \beta_2 X2 - \beta_3 X3 - \beta_4 X4 - \beta_5 X5 - \beta_6 X6 - Tr \cdot (\beta_{2Tr} X2 + \beta_{3Tr} X3))$

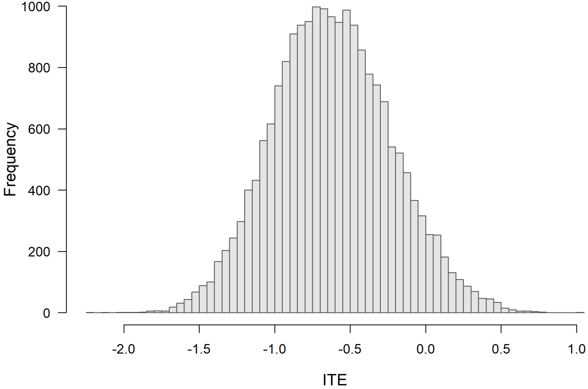
TRAM-DAGs: Example for ITE estimation

$$\text{ITE} = \text{median}(Y \mid \text{do}(T = 1), X) - \text{median}(Y \mid \text{do}(T = 0), X)$$



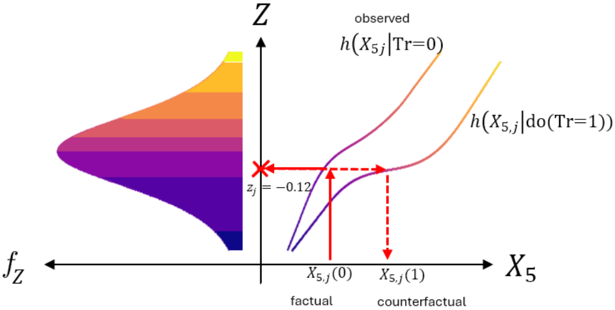
TRAM-DAGs: Example for ITE estimation

$$\text{ITE} = \text{median}(Y \mid \text{do}(T = 1), X) - \text{median}(Y \mid \text{do}(T = 0), X)$$



TRAM-DAGs: Estimate Potential Outcomes

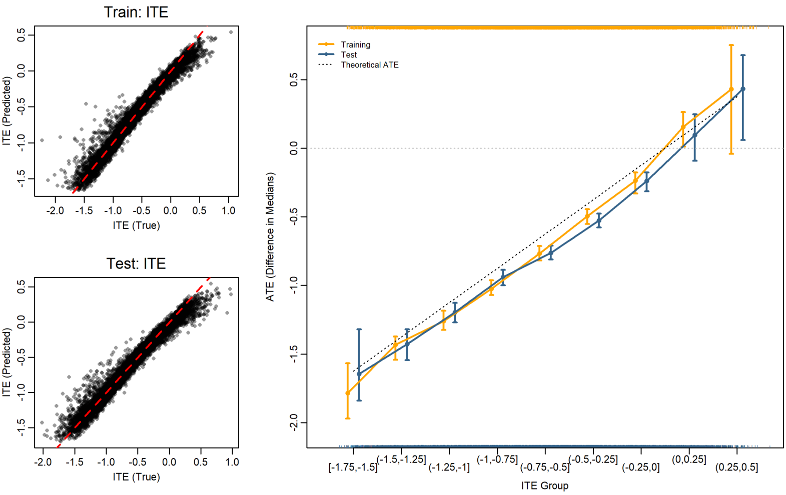
If we observe a $X5$ under $Tr = 0$, we can determine the counterfactual $X5$ under $Tr = 1$ with the observed latent value z_j:



TRAM-DAGs: Estimate Potential Outcomes

1. Estimate each $h_i(X_i \mid \mathrm{pa}(X_i))$ fully flexible (deep-NN / complex intercept)
2. Take the train set or a test set
3. $Z_i = h(X_i \mid pa(X_i))$ gives us the (observed) latent variable for each X_i
4. Determine counterfactuals for X5 and X6 with the (observed) latent variables Z_i
5. Determine medians of potential outcomes $Y(1)$ and $Y(0)$
6. $\mathrm{ITE} = \mathrm{median}(Y(1) \mid X_{tx}) - \mathrm{median}(Y(0) \mid X_{ct})$

TRAM-DAGs: Example for ITE estimation (Results)



TRAM-DAGs: Example for ITE estimation (Results)

ATE TRAM-DAG: estimated as $\text{mean}(\text{ITE}_{predicted})$:

-0.619 (-0.627 to -0.617)

ATE from RCT (randomized:) estimated as observed $\text{median}(Y \mid T = 1)$ - $\text{median}(Y \mid T = 0)$:

-0.637 (-0.662 to -0.610)

— confidence intervals obtained by bootstrapping

References

Sick, B. and Dürr, O. (2025). Interpretable neural causal models with tram-dags. Accepted at the CLeaR 2025 Conference.