Collaborative Targeted Learning in Genomic
When Linkage Disequilibrium Challenges Causal Inference

Juliet Asantewaa Sarpong

Maxwell Institute Graduate School · University of Edinburgh | Supervisors: Dr. S. Beentjes · Prof. C. Ponting | Advisor: Dr. A. Khamseh

YEAR 2 - PHD STATISTICS

DEPARTMENT OF STATISTICS

SCHOOL OF MATHEMATICS

BIOINFORMATICS

When thousands of correlated genetic variants surround a single DNA variant of interest, can we still reliably estimate its **causal effect** on disease? This work benchmarks **six non/semi-parametric estimators** under realistic genomic confounding; testing whether theoretical advantages hold empirically across **different** linkage disequilibrium environments.

R · Julia · Python | Genomics · Variant analysis | TMLE.jl (open source) | HPC · Eddie cluster

1 MOTIVATION

Estimating causal effects from **observational healthcare data** requires adjusting for confounding. In high-dimensional settings such as genomic data, standard TMLE faces two critical challenges:

CHALLENGE 1 - MODEL MISSPECIFICATION

Parametric models impose assumptions that rarely hold in complex clinical data, leading to biased ATE estimates even when model fit appears adequate.

CHALLENGE 2 - POSITIVITY VIOLATIONS

When some covariate strata have near-zero treatment probability, inverse probability weights become extreme, inflating variance and destabilising inference. Especially severe in genomic data due to *linkage disequilibrium*.

Collaborative TMLE (C-TMLE) addresses both by embedding **outcome-guided propensity score selection** into the targeting step, coupling nuisance estimation with the causal parameter of interest.

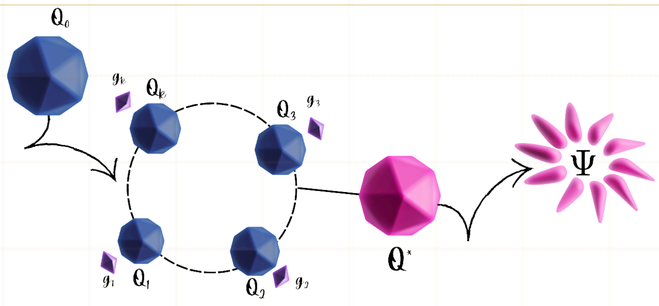


Fig. 1: C-TMLE flow - iterative collaborative targeting from Q_0 through propensity score sequence $g_1 \ldots g_k$ to targeted Q^* and Ψ

2 SIMULATION DESIGN

A synthetic genotype matrix is generated with **exponential-decay LD structure** matching the northern & western European ancestry reference population. Three treatment variants are selected for different LD Categories, enabling direct comparison of estimator performance as the positivity challenge varies by design.

DATA-GENERATING MECHANISM

n = 10,000 subjects · p = 1,000 variants over ~4,875 kb · MAF ~ Beta(0.5, 0.5) rescaled to [0.05, 0.50] · Recombination hotspots every 75kb · Hardy–Weinberg equilibrium verified (>95% variants) · FVE = 0.4 · True ATE = 1 · 200 replicates per scenario · Executed on Edinburgh HPC (Eddie) as parallel array jobs

TWO DATA-GENERATING MODELS COMPARED

Linear DGM: $\mu_Y = \sum \beta \cdot X_{jk}$ (5 causal main effects - consistent with GWAS assumption)

Interaction DGM: $\mu_Y = \sum \beta \cdot \prod_{l \in Sk} X_l$ (5 higher-order interactions, order drawn from {1,2,3,4} - reflecting epistasis)

All estimators target the **Average Treatment Effect (ATE)** of a binarised DNA variant A on a continuous phenotype Y, adjusting for the surrounding variants W as confounders.

$$\Psi(P_0) = E_W[E[Y|A=1,W] - E[Y|A=0,W]]$$

Marginal ATE - causal effect of carrying the minor allele

3 METHODS - SIX ESTIMATORS

TMLE

Standard baseline. Propensity score estimated independently via LASSO. Susceptible to variance inflation when LD induces near-zero treatment probabilities.

C-TMLE$_0$

Extends **TMLE** with an additional clever covariate $\tilde{H}$ stabilising the EIF with respect to the regularisation parameter λ, without full variable selection.

C-TMLE$_1$ (LASSO)

Collaborative propensity model selection via LASSO over a grid of λ values; CV minimises joint causal loss rather than propensity fit alone.

Scalable C-TMLE

Computationally efficient C-TMLE ordering candidate confounders by partial correlation with TMLE residuals; forward selection accordingly.

Wang-TMLE

TMLE under a semi-parametric linear outcome model $E_0(Y|A,W) = \beta_0 A + \theta(W)$; clever covariate $H(A,W) = A - \hat{g}_n(W)$ designed for genetic association settings.

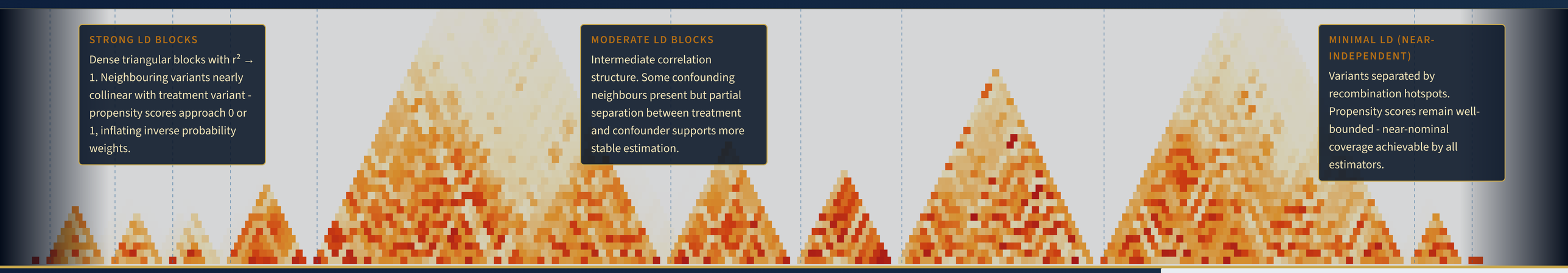
Wang-CTMLE

Collaborative Wang-TMLE selecting nearby variants via pairwise correlation cut-off $|r| > 0.7$ within 1Mb of the treatment variant. Best performer under strong LD.

WHY LD MATTERS

Linkage Disequilibrium Creates *Near-Positivity Violations:* The Main Challenge

CEU population LD matrix (r^2) - each triangular block represents a genomic region where variants are highly correlated, making it hard to separate their individual effects.



STRONG LD BLOCKS

Dense triangular blocks with $r^2 \to 1$. Neighbouring variants nearly collinear with treatment variant - propensity scores approach 0 or 1, inflating inverse probability weights.

MODERATE LD BLOCKS

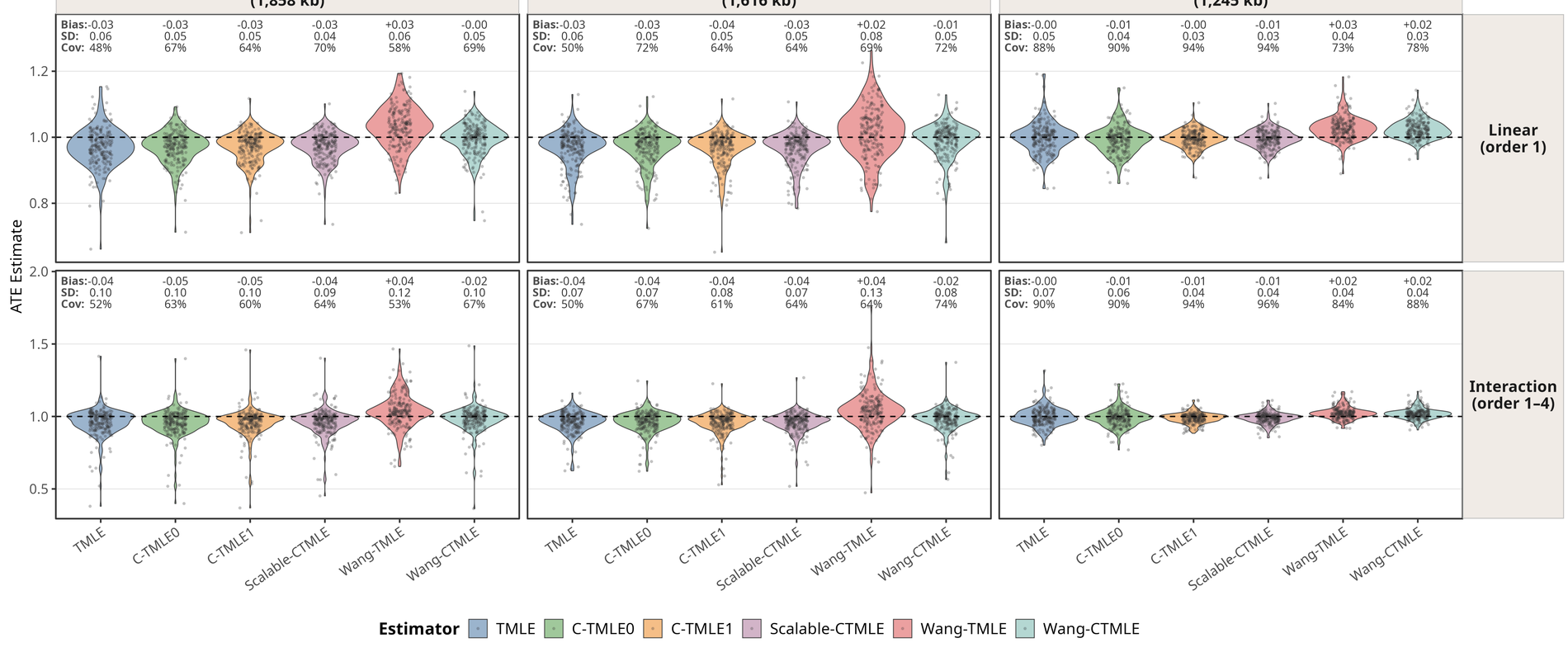
Intermediate correlation structure. Some confounding neighbours present but partial separation between treatment and confounder supports more stable estimation.

MINIMAL LD (NEAR-INDEPENDENT)

Variants separated by recombination hotspots. Propensity scores remain well-bounded - near-nominal coverage achievable by all estimators.

Results - ATE Estimates Across Estimators, LD Categories & DGMs

Rows: Linear DGM (additive, order 1) · Interaction DGM (epistatic, orders 1–4) · Cols: Strong · Moderate · Minimal LD



*Coverage is driven by LD category, not DGM. **Wang-CTMLE** leads under strong and moderate LD; **C-TMLE$_1$** and **Scalable-CTMLE** are competitive alternatives. Standard **TMLE** underperforms throughout.*

CONCLUSION

- **Wang-CTMLE** is the strongest performer under strong and moderate LD; its $|r| > 0.7$ pruning directly tackles positivity violations.
- **C-TMLE$_1$ and Scalable-CTMLE** are strong broadly-applicable alternatives.
- **TMLE consistently underperforms**; low coverage is driven by **variance underestimation**, not bias.
- Minimal LD yields near-nominal coverage for all estimators; confirming LD as the core challenge.

NEXT STEPS

⚡ TMLE.jl Integration | 🧬 UK Biobank | 📦 JOSS Paper

IMPLEMENTATION

R simulation pipeline executed as parallel array jobs on **Edinburgh HPC (Eddie)**.

References · van der Laan & Rose (2011, 2018) · van der Laan & Gruber (2010) · Ju et al. (2019) · Hines et al. (2022)

SUPERVISORS
Dr. S. Beentjes · Prof. C. Ponting

ADVISOR
Dr. A. Khamseh

NETWORK
CHAI Hub · University of Edinburgh