GeneBench-Pro Case Study: LD-aware cis-MVMR with Winner's Curse and Residual Pleiotropy

GeneBench-Pro

June 26, 2026

1 Overview

cis-Mendelian randomization for drug-target validation often uses nearby, correlated genetic instruments and summary-level molecular-association data rather than randomized perturbation data. This case study asks whether an analyst can recover direct disease log-odds effects for two nearby proteins from protein and disease association summaries, a locus correlation reference, and protein measurement records. The practical bottleneck is not a single MR fit: the recoverable target depends on assay-scale calibration, allele harmonization, winner's-curse avoidance, LD-aware joint modeling, and residual diagnostics.

Because the prompt asks for effects per one SD increase in $\log_{10}$ concentration, the target is the finite-sample estimate rather than the nominal simulation parameter, with the realized PROTA and PROTB values reported in the answer-field table below. Analyses that ignore one part of the data structure can still appear scientifically plausible, but they target different quantities. The correct inferential arc is: (1) estimate usable protein SDs from batch-structured measurements, excluding a degraded assay batch; (2) align protein effects to the disease effect allele; (3) use the holdout protein-association phase rather than winner's-cursed screening effects; and (4) fit an LD-aware joint protein model, then remove residual clusters produced by a latent third protein before refitting. Each step has a visible data signature and a numerical consequence.

2 Released Prompt and Files

Prompt

```
You are given association summary statistics and metadata for two nearby proteins (PROTA
    and PROTB), a binary disease outcome, a locus correlation reference, and protein
    measurement records.
Goal: estimate the direct log-odds effect of each protein on the disease outcome per +1
    SD increase in log10 concentration, conditional on the other protein.
Interpretation: theta_PROTA and theta_PROTB use the same log-odds per-SD scale defined in
     the goal.
These data came from a real experiment; you will be graded not just on numerical
    correctness but the quality of analytical reasoning you exhibit; do not attempt to
    take any shortcuts.
Return your final answer as exactly one JSON object.
Do not wrap the JSON in markdown.
Do not add prose before or after the JSON.
Do not omit any keys shown in the example.
Return the JSON object in your final answer:
{
  "answer": {
    "theta_PROTA": <float>,
    "theta_PROTB": <float>
  },
  "reasoning": "<description of method and QC>"
}
```

Released data files

File	Format	Contents
`protein_PROTA_phase1.tsv.gz`	.tsv.gz	Screening-phase PROTA association summary statistics.
`protein_PROTA_phase2.tsv.gz`	.tsv.gz	Holdout-phase PROTA association summary statistics.
`protein_PROTB_phase1.tsv.gz`	.tsv.gz	Screening-phase PROTB association summary statistics.
`protein_PROTB_phase2.tsv.gz`	.tsv.gz	Holdout-phase PROTB association summary statistics.
`disease_association.tsv.gz`	.tsv.gz	Binary-disease association summary statistics with case/control counts.
`dataset_metadata.tsv.gz`	.tsv.gz	Dataset roles, sample sizes, ancestry, and measurement units.
`variant_metadata.tsv.gz`	.tsv.gz	SNP positions, alleles, and minor allele frequencies.
`protein_measurements.tsv.gz`	.tsv.gz	Individual protein measurements by batch, with repeated batch-level `control_cv`.
`locus_correlation_EUR.npz`	.npz	SNP correlation matrix and SNP order for the European-ancestry locus reference.

3 Answer Fields and Tolerances

The output schema maps directly to the two fitted coefficients:

$$\texttt{theta\_PROTA} = \widehat{\theta}_A, \qquad \texttt{theta\_PROTB} = \widehat{\theta}_B. \tag{1}$$

Answer field	Ground truth	Tolerance / matching rule	Interpretation
`theta_PROTA`	0.293155	Absolute error ≤ 0.025; valid range $[-1, 1]$	Direct effect of PROTA conditional on PROTB after harmonization, LD-aware GLS, winner's-curse scaling, and residual-cluster pruning.
`theta_PROTB`	0.222073	Absolute error ≤ 0.025; valid range $[-1, 1]$	Direct effect of PROTB conditional on PROTA under the same fitted multivariable MR system.

4 Structure Diagram

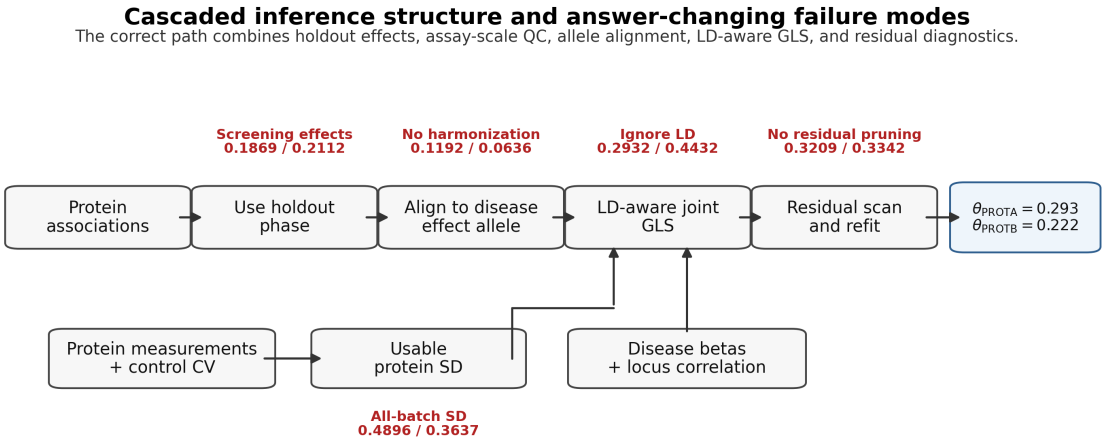

Figure 1: Cascaded inference structure. Boxes are analysis inputs or decisions and arrows show the dependency order. Red annotations above or below a step give the wrong PROTA/PROTB estimates produced by skipping that step; the blue box at right gives the released target estimates. PROTA and PROTB are the two protein exposures, `control_cv` is a batch-level control-sample coefficient of variation, LD is linkage disequilibrium/correlation among SNP instruments, and GLS is generalized least squares.

5 Variables and Assumptions

- $m = 80$: SNPs in a single 800 kb locus, indexed by j.
- x_j: physical position of SNP j from 50,000,000 to 50,800,000 bp.

- p_j: minor allele frequency drawn from Uniform(0.05, 0.5).
- $\mathbf{R}$: SNP correlation matrix supplied as `locus_correlation_EUR.npz`.
- $\boldsymbol{\alpha}_A$: per-SD SNP effects on PROTA, nonzero at indices 20, 24, and 55 with effects 0.20, 0.25, and 0.15.
- $\boldsymbol{\alpha}_B$: per-SD SNP effects on PROTB, nonzero at indices 25, 47, and 60 with effects 0.25, 0.28, and 0.18.
- $\theta_A = 0.30$, $\theta_B = 0.20$: simulation-reference direct effects on disease log-odds.
- $\sigma_A = 0.25$, $\sigma_B = 0.20$: true usable within-batch SDs of $\log_{10}$ concentration.
- $N_1 = 12000$, $N_2 = 8000$: screening and holdout protein-association sample sizes.
- $N_{\text{case}} = 15000$, $N_{\text{control}} = 35000$: disease-association sample sizes.
- $\lambda_{\text{WC}} = 1.60$: screening-phase winner's-curse multiplier for genome-wide-significant protein associations.
- Batch 3: degraded measurement batch with SD 0.625 for PROTA and 0.500 for PROTB; its batch-level control-sample CV is 0.314 versus 0.082 and 0.089 for batches 1 and 2.

6 Data-Generating Process

The LD/correlation structure is

$$R_{jk} = \exp\left(-\frac{|x_j - x_k|}{35000}\right) + 10^{-6}\mathbf{1}_{j=k}. \tag{2}$$

This creates a realistic local correlation pattern, making independent-instrument IVW inappropriate.

For a protein association study with sample size N and per-SD SNP effects $\boldsymbol{\alpha}$, z-scores are generated as

$$\mathbf{Z} \sim \mathcal{N}\left(\sqrt{N}\mathbf{R}\boldsymbol{\alpha},\ \mathbf{R}\right), \tag{3}$$

with standard error

$$\text{SE}_j = [N \cdot 2p_j(1-p_j)]^{-1/2}, \qquad \hat{\beta}_j = Z_j \text{SE}_j. \tag{4}$$

The disease study uses effective sample size

$$N_{\text{eff}} = \frac{4N_{\text{case}}N_{\text{control}}}{N_{\text{case}} + N_{\text{control}}} \ = \ 42000 \tag{5}$$

and disease genetic effects

$$\boldsymbol{\alpha}_Y = 0.30\boldsymbol{\alpha}_A + 0.20\boldsymbol{\alpha}_B. \tag{6}$$

Screening protein betas with $p < 5 \times 10^{-8}$ are multiplied by 1.60. Protein betas are then stored per $\log_{10}$ concentration unit by multiplying the per-SD effects by σ_A or σ_B:

$$\hat{\gamma}_{Aj}^{\log 10} = \sigma_A \hat{\gamma}_{Aj}^{\text{SD}}, \qquad \hat{\gamma}_{Bj}^{\log 10} = \sigma_B \hat{\gamma}_{Bj}^{\text{SD}}. \tag{7}$$

The protein measurement file contains three batches per protein. Batches 1 and 2 have different means but the true usable SD; batch 3 has the same general concentration range but 2.5 times larger within-batch SD:

$$M_{pb\ell} \sim \mathcal{N}(\mu_{pb}, \sigma^2_{pb}). \tag{8}$$

For PROTA, the batch means are 2.00, 2.50, and 2.25; for PROTB they are 1.80, 2.20, and 2.00. The released measurement file also repeats a neutral batch-level assay-QC value, `control_cv`, on each measurement record: batches 1 and 2 have values 0.082 and 0.089, while batch 3 has 0.314. This exposes degraded assay precision without directly labeling the batch as excluded.

Finally, a latent third protein introduces outcome-only residual signal at 18 core SNPs:

$$\hat{\beta}^{\text{obs}}_{Yj} = \hat{\beta}_{Yj} + \delta_j, \tag{9}$$

where $\delta_j > 0$ for indices 24–30, 47–52, and 66–70. The first two regions include protein-causal edge sentinels, so lead-sentinel shortcuts keep invalid instruments; the third region catches all-SNP analyses that only prune the first two visible disease-only regions. LD causes the residual diagnostic to flag a wider local halo than the directly perturbed selected instruments.

7 Analyst Walkthrough

Step 1: infer the usable protein scale

A simple pooled SD from all protein measurements returns 0.4699 for PROTA and 0.3728 for PROTB, almost twice the usable within-batch scale. Even centering within all batches is still wrong because batch 3 is a high-variance assay run, giving 0.4218 and 0.3303. The per-batch SDs are 0.2524, 0.2527, and 0.6373 for PROTA, and 0.2050, 0.1983, and 0.4959 for PROTB; the corresponding `control_cv` values are 0.082, 0.089, and 0.314. The batch-distribution and per-batch SD figures below show the mean shifts and the visible assay-QC separation that isolates the degraded batch. This is an assay-QC decision, not an implementation of ComBat, but it uses the same principle that non-biological batch variation should be diagnosed rather than pooled blindly [6].

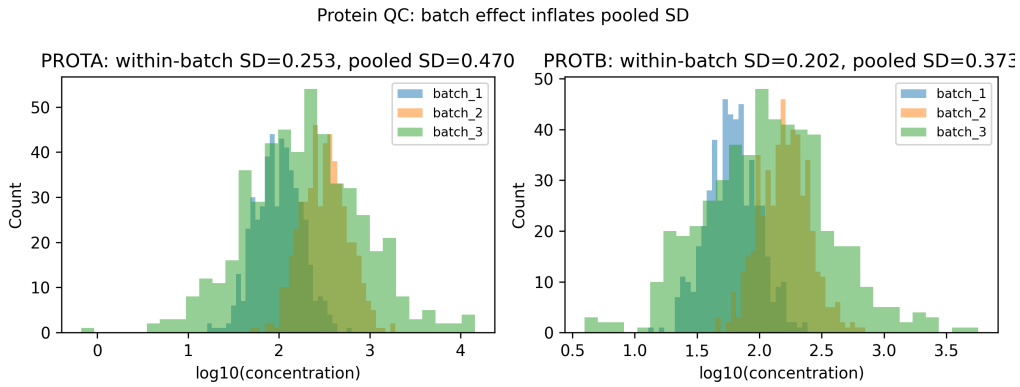

Figure 2: Protein measurement distributions by batch. Each panel is one protein, with $\log_{10}$ concentration on the x-axis and record count on the y-axis. Blue, orange, and green histograms are batches 1, 2, and 3. The panel titles report the usable within-batch SD after excluding the degraded batch and the naive pooled SD across all records. Mean shifts make pooled SD inappropriate, while the broader green batch-3 distribution shows degraded assay precision.

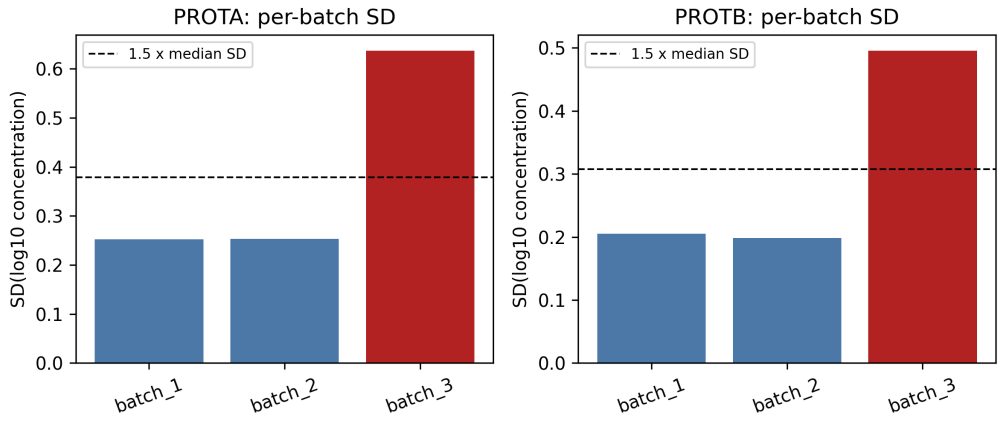

Figure 3: Per-batch SD diagnostic. Each panel is one protein; bars show per-batch SD of $\log_{10}$ concentration. Blue bars are retained high-quality batches, the red bar is the excluded degraded batch, and the horizontal dashed line is the 1.5-fold median-SD cutoff used to flag a high-variance batch. The diagnostic is not a hidden biological threshold: the two high-quality batches cluster together, batch 3 has both larger SD and higher `control_cv`, and the resulting usable SDs are 0.2526 for PROTA and 0.2017 for PROTB.

```
batch_sds = qc.groupby(["protein_id", "batch"]).log10_concentration.std()
cutoff = 1.5 * batch_sds.groupby(level=0).median()
good = batch_sds <= cutoff.reindex(batch_sds.index, level=0)
```

Step 2: align effect alleles

An analyst who reads only the beta columns will combine effects measured on opposite alleles. The allele-harmonization figure below shows that a subset of holdout protein betas is the sign-reversed representation of the same SNP effect. Aligning to the disease effect allele is a standard MR harmonization step; skipping it gives substantially attenuated direct effects.

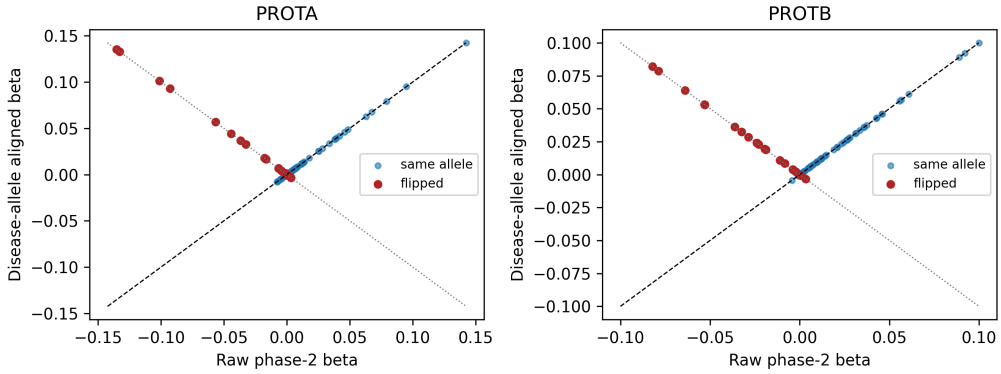

Figure 4: Allele harmonization for holdout protein effects. Each panel is one protein. The x-axis is the raw phase-2 protein beta as stored in the protein association file, and the y-axis is the same beta after aligning the protein effect allele to the disease-association effect allele. Blue points already use the same allele orientation as the disease file; red points use the opposite allele and are sign-flipped. The black dashed identity line marks unchanged betas, while the light gray negative-slope line marks the expected location of sign-flipped betas.

```
flip = (protein.effect_allele == disease.other_allele) & \
       (protein.other_allele == disease.effect_allele)
protein.loc[flip, "beta"] *= -1
```

Step 3: avoid winner's-cursed screening effects

The screening phase contains inflated protein betas at genome-wide-significant SNPs. This is visible by comparing phase-1 and phase-2 effects in the phase-comparison figure below. Using screening effects in the final model shrinks the inferred disease effects to 0.1869 and 0.2112.

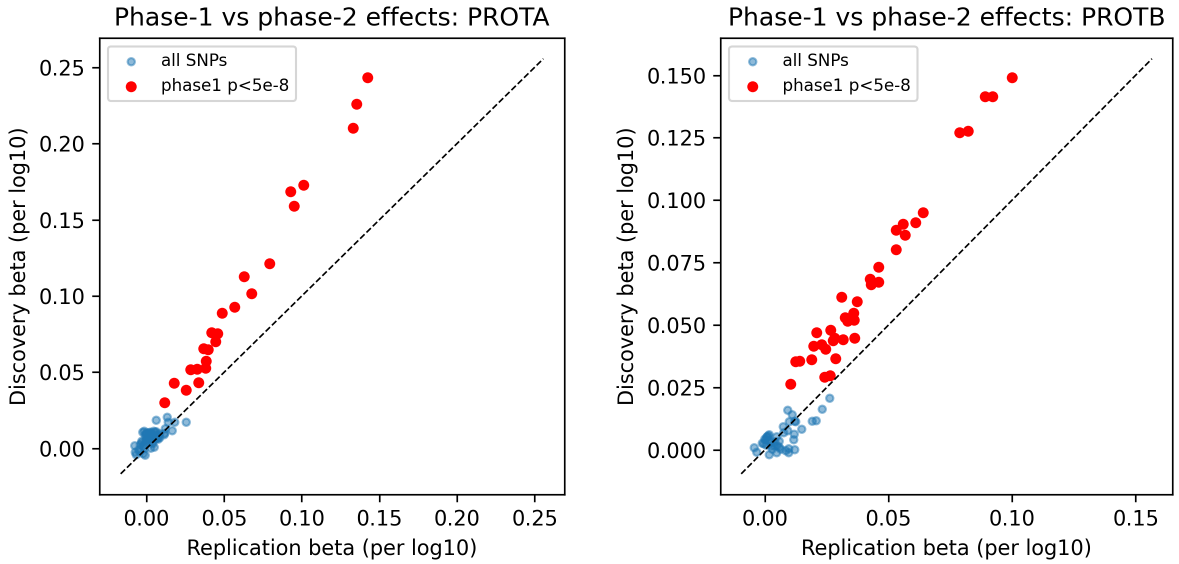

Figure 5: Screening versus holdout protein effects. The left panel is PROTA and the right panel is PROTB. The x-axis is the phase-2 holdout/replication beta per $\log_{10}$ concentration, and the y-axis is the phase-1 discovery/screening beta on the same scale. Blue points are all SNPs; red points are SNPs that were genome-wide significant in the phase-1 screen ($p < 5 \times 10^{-8}$). The black dashed diagonal is the no-inflation reference line $y = x$; red points lying above this line show winner's-curse inflation in the screening phase.

Step 4: fit the joint LD-aware model

The disease and protein signals occupy the same correlated locus. The selected instrument set is the phase-2 union of SNPs with $p_A < 5 \times 10^{-8}$ or $p_B < 5 \times 10^{-8}$, giving 40 instruments; no clumping is applied, and the LD submatrix is taken at those same indices. A univariable analysis gives 0.4486 and 0.4493 because each protein partially proxies the other. A model that ignores LD gives 0.2932 and 0.4432 because correlated instruments are misweighted. The correct covariance for the disease effects is

$$\boldsymbol{\Sigma}_Y = \mathrm{diag}(\mathbf{s}_Y)\mathbf{R}\mathrm{diag}(\mathbf{s}_Y). \tag{10}$$

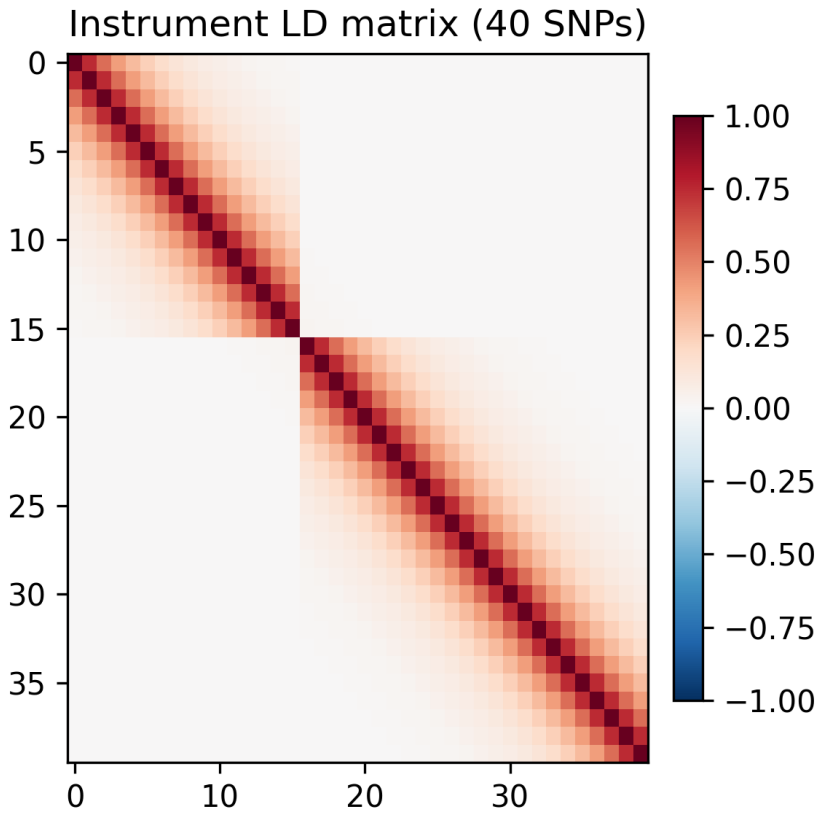

Figure 6: LD/correlation matrix among the 40 selected SNP instruments. Rows and columns index the same selected instruments in locus order. The color bar gives the pairwise SNP correlation: dark red is strong positive correlation, white is near zero, and blue would indicate negative correlation. The red off-diagonal blocks show correlated instrument clusters, so treating instruments as independent is not defensible.

```
Sigma = np.diag(se_y) @ R_instruments @ np.diag(se_y)
theta = solve(Gamma.T @ inv(Sigma) @ Gamma,
              Gamma.T @ inv(Sigma) @ beta_y)
```

Step 5: diagnose the residual cluster and refit

The selected-instrument LD-aware joint model before pruning gives 0.3209 and 0.3342. The residuals are structured rather than exchangeable: they cluster by position where the latent third protein affects the disease association, including protein-causal edge sentinels in the first two clusters. The residual-pruning figure below shows the initial residuals and the pruning threshold. Refitting after pruning gives the target. A fixed all-SNP shortcut that removes only the first two regions gives 0.2699 and 0.2839, because it misses the third residual cluster.

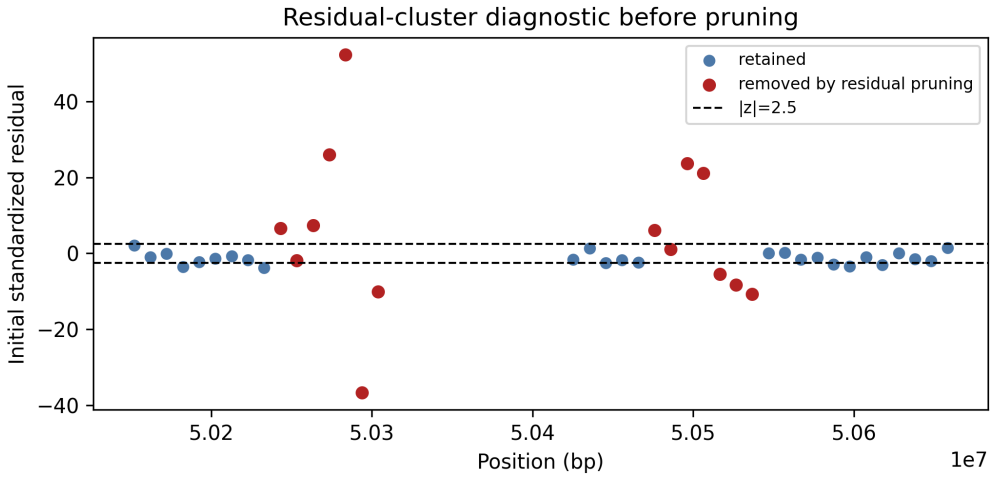

Figure 7: Residual-cluster diagnostic before pruning. Each point is a selected SNP instrument, plotted by genomic position on the x-axis and its initial standardized residual from the LD-aware joint model on the y-axis. Blue points are retained for the final refit; red points are the cumulative set removed by iterative residual pruning. The horizontal dashed lines mark the per-iteration residual threshold $|z| = 2.5$. Some red points fall within the initial dashed band because pruning removes the current worst residual, refits the model, and can make additional clustered instruments exceed the threshold in later iterations. Red points form positionally clustered outliers, including the directly perturbed core plus an LD-correlated halo.

```
while max(abs(std_resid)) > 2.5:
    remove_worst_residual()
    refit_ld_aware_joint_model()
```

Final fitted estimate. The final model uses holdout protein effects, allele alignment, SD conversion with $\hat{\sigma}_A = 0.2526$ and $\hat{\sigma}_B = 0.2017$, and LD-aware residual-pruned GLS. The realized estimates are listed in the answer-field table above. The ablation table summarizes the major wrong estimates against the target.

8 Estimand

The estimand is the pair of direct disease log-odds effects per one usable within-batch SD increase in $\log_{10}$ protein concentration:

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_A \\ \theta_B \end{bmatrix}, \tag{11}$$

where θ_A is the effect of PROTA conditional on PROTB, and θ_B is the effect of PROTB conditional on PROTA. The finite-sample estimator is preferred over the nominal DGP because association noise and residual-cluster pruning define the recoverable answer.

9 Estimator

Let $\mathbf{y}$ be the disease beta vector for the 40 selected phase-2 instruments after allele alignment, $\mathbf{s}$ its standard errors, $\mathbf{R}$ the matching 40-by-40 correlation submatrix, and $\boldsymbol{\Gamma}$ the $K \times 2$ matrix of

holdout protein betas converted to per-SD units. The LD-aware GLS estimator is

$$\widehat{\boldsymbol{\theta}} = (\boldsymbol{\Gamma}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\Gamma})^{-1} \boldsymbol{\Gamma}^T \boldsymbol{\Sigma}^{-1} \mathbf{y}, \qquad \boldsymbol{\Sigma} = \mathrm{diag}(\mathbf{s}) \mathbf{R} \mathrm{diag}(\mathbf{s}). \tag{12}$$

Before the final refit, standardized residuals are computed from the Cholesky factor of $\boldsymbol{\Sigma}$:

$$\mathbf{r} = \mathbf{L}^{-1}(\mathbf{y} - \boldsymbol{\Gamma}\widehat{\boldsymbol{\theta}}), \qquad \mathbf{L}\mathbf{L}^T = \boldsymbol{\Sigma}. \tag{13}$$

The reference implementation removes the worst residual while $\max_j |r_j| > 2.5$ and then refits. The threshold is used as a diagnostic implementation choice, not as a hidden target definition: thresholds from 2.5 through 5 recover both effects within tolerance in this dataset, while much looser thresholds leave residual-cluster contamination and fail the PROTB field. This GLS fit treats exposure betas as fixed relative to the outcome uncertainty, which is the usual summarized-data IVW approximation rather than a full errors-in-variables model. Standard multivariable MR supports direct-effect estimation with multiple exposures [1], and fine-mapped MR work supports summarized association/correlation inputs when candidate instruments are correlated [2]. The residual scan is a diagnostic for locally invalid instruments, not a claim to implement cisMR-cML or MR-PRESSO; robust cis-MR and MR-PRESSO literature instead motivate treating correlated cis instruments and horizontal pleiotropy as answer-changing risks [3,4]. Allele harmonization follows standard two-sample MR conventions [5], and the batch-scale step is grounded in visible assay-QC evidence while citing batch-effect literature only for the general principle that non-biological batch variation should be handled explicitly [6].

10 Decision-Point and Ablation Walkthrough

The table combines the full ablation outputs with the stage at which each incomplete analysis fails. The released answer contract accepts absolute error no larger than 0.025 for both `theta_PROTA` and `theta_PROTB`; rows labeled "no" violate at least one of those fields.

Decision point	Analysis / ablation	Quantitative output	Pass?	Failure point	Why the approach is wrong
Reference pipeline	Full LD-aware MVMR	PROTA 0.2932, PROTB 0.2221; errors 0.0000/0.0000	yes	none	Reference joint LD-aware MVMR after scale, allele, winner's-curse, and residual-cluster corrections.
Accepted sensitivity	Residual threshold 5.0	0.2946/0.2298; errors 0.0014/0.0077	yes	none	A moderately looser residual cutoff retains the same inferential target; much looser cutoffs retain contaminated instruments and fail PROTB.
Residual cluster	No residual outlier detection	0.3209/0.3342; errors 0.0278/0.1121	no	Step 5	Leaves a latent third-protein residual cluster in the instrument set.
Joint model	Univariable MR	0.4486/0.4493; errors 0.1554/0.2272	no	Step 4	Estimates each protein separately even though the local instruments proxy both proteins through LD.
Winner's curse	Discovery-screen effects	0.1869/0.2112; errors 0.1062/0.0109	no	Step 3	Uses discovery-screen effects as instruments after selection inflated those betas.
Scale conversion	No unit conversion	1.2705/1.6570; errors 0.9774/1.4349	no	Step 1	Leaves protein effects on incompatible assay units rather than converting to the usable replicate scale.
Allele harmonization	No allele harmonization	0.1192/0.0636; errors 0.1740/0.1585	no	Step 2	Fails to sign-align protein betas to the disease effect allele.
Scale conversion	Pooled SD across batches	0.5970/0.6177; errors 0.3039/0.3957	no	Step 1	Lets the degraded batch inflate the exposure standard deviation.
Scale conversion	All-batch SD after pruning	0.4896/0.3637; errors 0.1964/0.1416	no	Step 1	Applies residual pruning but still estimates the unit conversion on all batches, including the degraded batch.
LD-aware weighting	Independent-IV MVMR	0.2932/0.4432; errors 0.0001/0.2211	no	Step 4	Treats correlated instruments as independent and overweights clustered information.
Residual cluster	All 80 instruments	0.3092/0.3562; errors 0.0161/0.1341	no	Step 5	Uses the broad selected set without removing disease-only residual clusters.
Instrument selection	F-statistic threshold only	0.3013/0.3640; errors 0.0081/0.1419	no	Steps 3–5	A strength-threshold filter keeps invalid clustered instruments.
Instrument selection	Lead variant only	0.2980/0.3540; errors 0.0048/0.1319	no	Steps 4–5	Lead variants remain contaminated and do not replace residual diagnostics.
Instrument selection	Phase-1 clumped sentinels	0.3219/0.3417; errors 0.0288/0.1196	no	Steps 3–5	Uses winner's-cursed phase-1 sentinels and bypasses the residual scan.

Decision point	Analysis / ablation	Quantitative output	Pass?	Failure point	Why the approach is wrong
Residual pruning	Two-region pruning, no LD	0.2699/0.2839; errors 0.0233/0.0618	no	Step 5	Removes the first two visible residual blocks but misses a third disease-only cluster.
Residual pruning	Two-region pruning, LD-aware	0.2734/0.2656; errors 0.0197/0.0435	no	Step 5	Adds LD weighting to an incompletely pruned instrument set.
Scale conversion	Meta all-batch SD	0.3674/0.2919; errors 0.0743/0.0698	no	Step 1	Apparent averaging partly cancels errors but still uses the wrong scale.
Robust alternative	Egger-style sensitivity model	0.2404/0.1689; errors 0.0527/0.0532	no	Estimator parity	The intercept-augmented sensitivity model targets a different pleiotropy structure and does not recover the released direct-effect estimand.
Robust alternative	MR-PRESSO-style pruning	0.2503/0.3577; errors 0.0428/0.1356	no	Estimator parity	PRESSO-style outlier handling does not solve the joint LD-aware residual-cluster problem.

Table 2: Unified decision-point and ablation walkthrough for the cis-MVMR problem.

11 References

1. Sanderson E, Davey Smith G, Windmeijer F, Bowden J. An examination of multivariable Mendelian randomization in the single-sample and two-sample summary data settings. *International Journal of Epidemiology*. 2019;48:713–727. DOI: https://doi.org/10.1093/ije/dyy262.

2. Burgess S, Zuber V, Valdes-Marquez E, Sun BB, Hopewell JC. Mendelian randomization with fine-mapped genetic data: choosing from large numbers of correlated instrumental variables. *Genetic Epidemiology*. 2017;41:714–725. DOI: https://doi.org/10.1002/gepi.22077.

3. Lin Z, Pan W. A robust cis-Mendelian randomization method with application to drug target discovery. *Nature Communications*. 2024;15:6072. DOI: https://doi.org/10.1038/s41467-024-50385-y.

4. Verbanck M, Chen CY, Neale B, Do R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nature Genetics*. 2018;50:693–698. DOI: https://doi.org/10.1038/s41588-018-0099-7.

5. Hemani G, Haycock P, Zheng J, Gaunt T, Elsworth B, Palmer T. TwoSampleMR allele harmonisation documentation. https://mrcieu.github.io/TwoSampleMR/articles/harmonise.html.

6. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. 2007;8:118–127. DOI: https://doi.org/10.1093/biostatistics/kxj037.