Learning the Shape of Causality: Manifold-Aware Network Trajectory Analysis (MANTRA)

Peter Driscoll

Abstract

We study program-mediated prediction of red blood cell traits from CRISPRi perturbations through a unified objective that couples a GRN prior, manifold-aware filtering, and SMR/TWAS-based trait readouts. Given a dose regulated input, we map GRN-predicted gene-level effects through an Energy-Guided, geometry-aware filter, project onto cNMF programs, and predict trait changes via learned program→trait weights. Our training objective combines (i) laplacian smoothing on the learned manifold, (ii) weighted least squares fit to SMR effects, and (iii) a dose monotonicity penalty across sgRNA-UMI quartiles. We evaluate on K562 and will extend to HCT116 in the final report, with preregistered metrics, e.g., R^2, dose response, and portability against baselines including Ota β-regression. We hypothesize that the constraints induced by manifold-realism and GRN-causality will enrich trait prediction and dose consistency.

Preprint. Under review.

1 Introduction

Question. Do *GRN-informed, manifold-constrained* program projections yield stronger concordance with dose-stratified KD responses and RBC trait directions than program-only baselines (*Ota et al.*)?

Context & related work. To respect biological geometry in single cells, we will learn a manifold on *unperturbed* K562 and HCT116 cell states using diffusion-based embeddings [2] or deep latent models such as scVI [5]. Building on the EGGFM framework from the Knowles lab (Zweig, Zhang, Azizi, and Knowles) [8], which distills an energy/score model into a *Riemannian* metric tensor $G(x)$, we construct a geometry-aware Laplacian for manifold smoothing and geodesic interpolation. Program discovery uses NMF/cNMF to obtain interpretable modules [1, 4]; trait priors come from SMR/TWAS summary data [7]; and regulator priors may leverage GWPS [1] [3, 6].

Planned approach overview. (i) We will initialize gene-level effects with GWPS-constructed GRN priors; (ii) enforce manifold realism learned on *unperturbed* cells; (iii) map gene-space effects to program coordinates via NMF; (iv) read out trait deltas via program weights derived from SMR/TWAS summary statistics.

2 Data and EDA

We constructed an AnnData object for unperturbed cells, computed standard QC covariates (UMIs/cell, genes/cell, mitochondrial content), and selected highly variable genes (HVGs) using a Seurat/Seurat-v3 flavor conditional on integer-likeness of counts. Table 1 summarizes the QC distributions; Figures 1 and 2 show HVG diagnostics and QC violins, respectively.

Table 1: Summary statistics for per-cell QC covariates

Statistic	Total mRNA UMIs / cell	Genes detected / cell	Mitochondrial UMI (%)
count	247,914	247,914	247,914
mean	13,420	3,498.45	6.12
std	6,907.36	858.52	1.60
min	1,457	595	0.00
25%	9,128	2,879	5.07
50%	11,266	3,459	6.14
75%	15,222	4,042	7.18
max	147,863	7,965	11.00

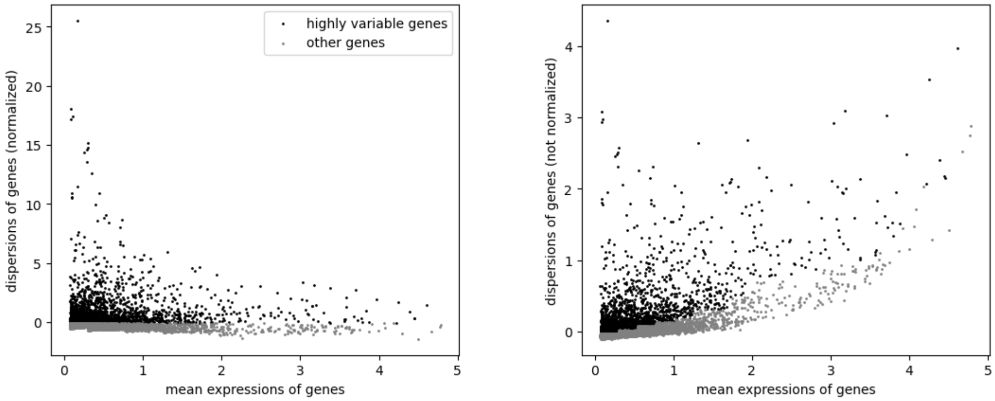

Figure 1: HVG dispersion with respect to mean expression.

[1] GWPS: genome-wide pooled CRISPR screens; Perturb-seq: single-cell transcriptomic CRISPR maps

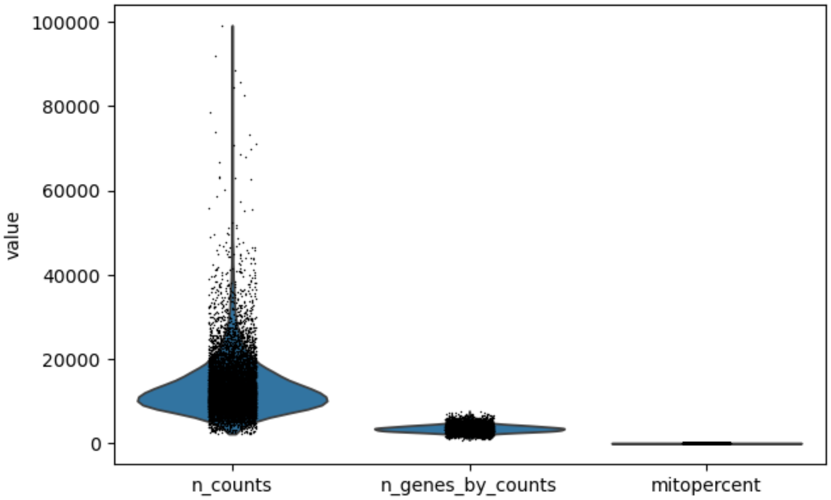

Figure 2: QC distributions across unperturbed cells.

3 Methods

Model summary (notation). We will use a GRN prior β to form gene-level deltas for a dose regulated input u, filter them by manifold geometry, project onto programs W, and read out trait change via program weights $\theta^{(t)}$:

$$\widetilde{\Delta E}=\beta\, u, \qquad \widehat{\Delta E}=(I+\lambda\, L_{\mathcal{M}})^{-1}\widetilde{\Delta E}, \qquad \Delta a=W^{\top}\widehat{\Delta E}, \qquad \widehat{\Delta \mathrm{Trait}}^{(t)}=\theta^{(t)\top}\Delta a. \quad (1)$$

3.1 GRN priors

We initialize gene-level effect predictions with a GWPS GRN matrix β; for a dose regulated input u we set $\widetilde{\Delta E} = \beta\, u$. When multiple regulators are combined, we sum the corresponding columns in β before geometry filtering. See Appendix B.1 for inclusion criteria and QC.

3.2 Manifold realism via EGGFM metric tensor

We will learn the cell-state manifold $(\mathcal{M}, G)$ on *unperturbed* cells (e.g., diffusion maps or scVI) and adopt the *EGGFM* view that distills an energy/score model into a Riemannian metric tensor $G(x)$ [8]. The induced geometry defines geodesic distance $d_{\mathcal{M}}$ and a geometry-aware Laplacian $L_{\mathcal{M}}$. To encourage on-manifold realism in predicted perturbation effects, we apply a Tikhonov/graph-Laplacian smoother in this learned geometry:

$$\widehat{\Delta E} \;=\; (I+\lambda\, L_{\mathcal{M}})^{-1}\widetilde{\Delta E},$$

with $\lambda > 0$ selected on a small grid. Realism diagnostics include kNN-overlap and geodesic displacement (*pre* vs. *post*) on the unperturbed manifold. Our manifold realism is directly inspired by the EGGFM-derived metric tensor $G(x)$ (Knowles lab), which induces the geometry-aware Laplacian $L_{\mathcal{M}}$ used for smoothing.

3.3 Program discovery (cNMF)

We learn program loadings W on unperturbed cells using cNMF; a K-sweep (e.g., $K \in \{8, 12, 16, 20\}$) and split-half stability (Jaccard of top genes) guide the choice of K. Programs provide the map $\Delta a = W^{\top}\widehat{\Delta E}$ used in prediction. See Appendix B.2.

3.4 Trait readout (SMR/TWAS-informed)

Let $s^{(t)} \in \mathbb{R}^G$ denote gene-level SMR/TWAS effects for trait t (HEIDI-filtered when available) with precision Σ^{-1}. We fit program weights by weighted least squares,

$$\hat{\theta}^{(t)} \; = \; \arg\min_{\theta} \; \left\| s^{(t)} - W\,\theta \right\|_{\Sigma^{-1}}^{2},$$

predict $\widehat{\Delta \mathrm{Trait}}^{(t)} = \theta^{(t)\top}\,\Delta a$.

Baselines. (B1) SMR/TWAS-only (no programs/manifold); (B2) program-mean readout; (B3) linear readout $\langle \theta,\, W^\top \Delta E \rangle$ without manifold filtering; (B4) Ota β-regression: $\widehat{\Delta \mathrm{Trait}}^{(t)}_{\text{Ota-}\beta}(r) = s^{(t)\top}\,\beta_{:,r}$.

3.5 Regulator selection and dose stratification

We predefine inclusion criteria for targets: confident dual-sgRNA constructs with adequate coverage and baseline expression, passing CRISPRi QC and sign-consistency checks; $Q4$ denotes the top 25% by per-cell sgRNA-UMI (pooled $Q1$–$Q3$ baseline). In HCT116 we stratify by sgRNA-UMI quartiles ($Q1 \ldots Q4$); $Q4$ is the top-dose slice for main analyses, and the monotonicity penalty is defined in the overall loss (Appendix B.3).

3.6 Loss and ablations

The training objective combines geometry smoothing, program readout fitting, and dose monotonicity:

$$\mathcal{L} = \alpha \left\| \widehat{\Delta E} - \widetilde{\Delta E} \right\|_2^2 + \lambda\, \widehat{\Delta E}^{\top} L_{\mathcal{M}}\, \widehat{\Delta E} + \sum_t \left\| s^{(t)} - W\,\theta^{(t)} \right\|_{\Sigma^{-1}}^2 + \sum_{q=1}^{4} [m_q]_- \; . \tag{2}$$

Here m_q denotes the (signed) monotonicity margin for quartile q (larger dose should not reduce $|\Delta \mathrm{Trait}^{(t)}|$); $[\cdot]_-$ is the negative-part penalty defined in the preamble.

We will report a 2 × 2 ablation grid isolating the value of the GRN prior and manifold filter: {No/No, GRN/No, No/Manifold, GRN/Manifold}.

4 Preliminary results

As shown in Fig. 3, PCA resolves coherent structure; baseline performance and calibration/sign accuracy appear in Figs. 4 and 5.

Structure and QC. QC covariates vary smoothly across structure; no batch-driven axes observed.

HVGs. Expected mean–dispersion tradeoff observed; HVGs concentrate in biologically informative ranges.

Program discovery. We defer the cNMF K-sweep and stability selection to the final; interim baselines are reported without a finalized program basis.

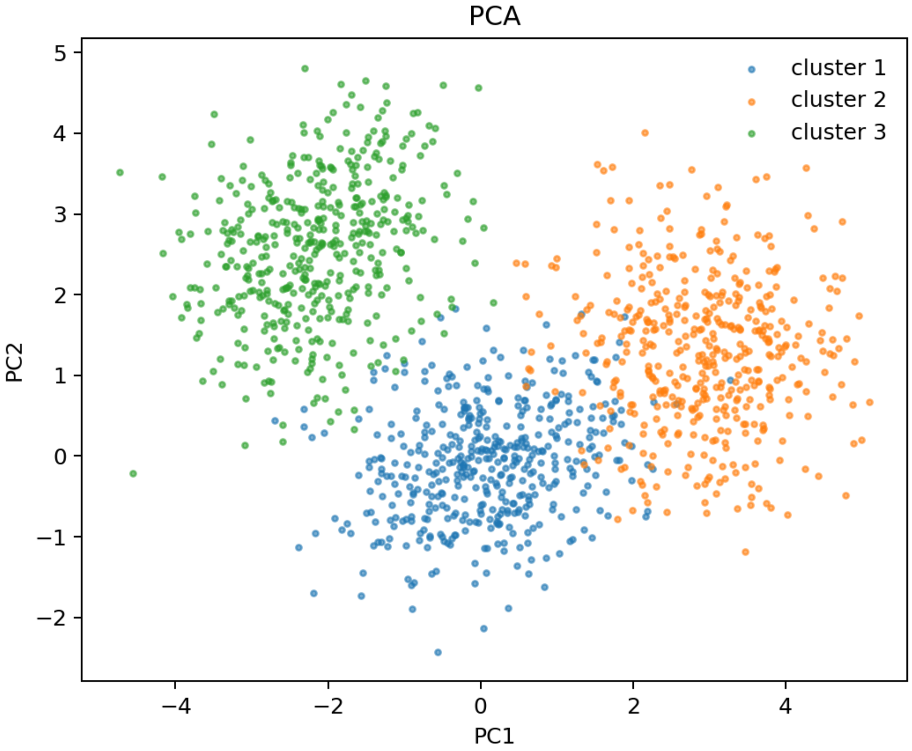

Figure 3: PCA exposes low-dimensional structure consistent with cell-cycle and QC covariates; nUMIs and mito% vary smoothly across PCs (cf. Fig. 2), and no batch-driven axes were observed.

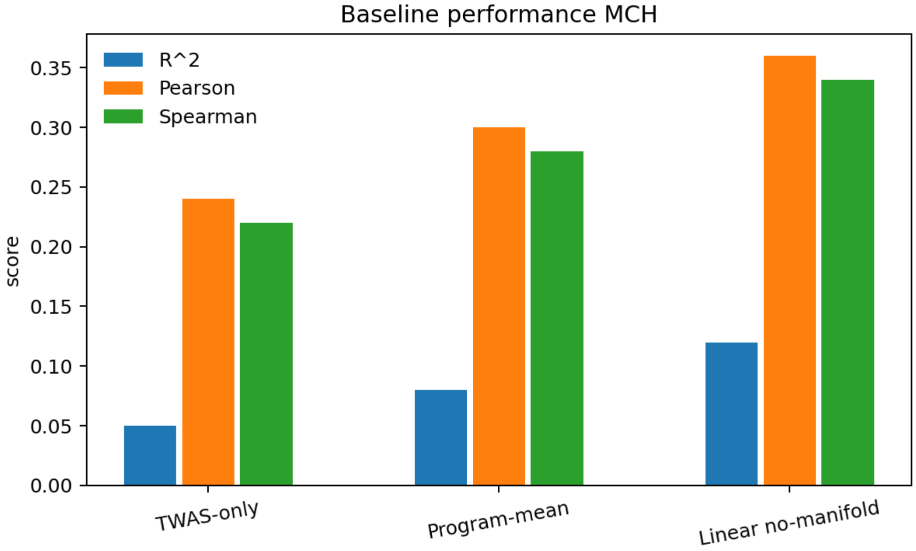

Figure 4: Baseline performance on MCH: TWAS-only, program-mean, and linear (no-manifold).

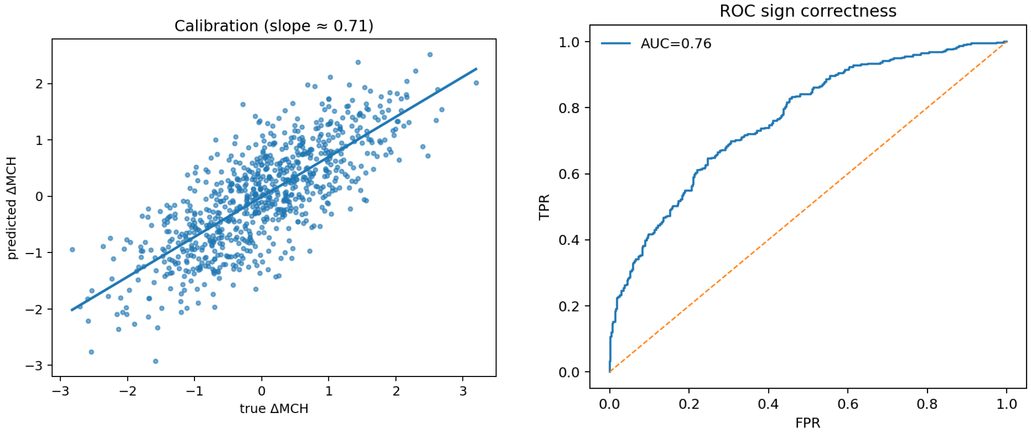

Figure 5: Left: calibration of predicted vs. observed ΔMCH. Right: ROC for sign correctness.

Evaluation plan.

- Baselines: (B1)–(B4) as in Section 3.4.
- Ablations (2×2 over $\{$GRN prior, manifold filter$\}$): (No, No), (GRN, No), (No, Manifold), (GRN, Manifold).
- Metrics: R^2, Pearson/Spearman, sign AUROC/AUPRC, calibration slope, kNN overlap pre/post, geodesic vs. Euclidean displacement.

5 Next steps (one-week plan)

1. Learn the manifold and fit diffusion maps/scVI on unperturbed K562, construct $L_{\mathcal{M}}$, and integrate $(I + \lambda L_{\mathcal{M}})^{-1}$ smoothing.
2. Build the GRN prior to derive β from GWPS/Perturb-seq and wire $\widehat{\Delta E} = \beta\, u$.
3. Finalize K via elbow + split-half program stability.
4. Fit program-level trait weights θ from SMR/TWAS, emit program loading barplot.
5. We will verify directional concordance with Ota K562 LoF/KD signs for representative regulators and report sign AUROC/AUPRC alongside the calibration slope.
6. Run baselines and $2{\times}2$ ablations; produce metrics CSV and summary figure.
7. Dose-aware validation using sgRNA-UMI quartiles; test monotone trends.

6 Reproducibility

All artifacts are generated by `01_qc_eda.py` and `02_prelim_figs.py`. We log commit hash, key thresholds, and counts in `manifest_qc.json`. Figures in this report are loaded from `./figures/`.

Acknowledgments

We build on manifold-aware interpolation and geometry for single-cell modeling [8]. Thanks to the course staff and collaborators for feedback.

References

[1] Jean-Philippe Brunet, Pablo Tamayo, Todd R. Golub, and Jill P. Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences*, 101(12):4164–4169, 2004. doi: 10.1073/pnas.0308531101.

[2] Ronald R. Coifman, Stéphane Lafon, Ann B. Lee, Mauro Maggioni, Boaz Nadler, Frederick Warner, and Steven W. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition on data: Diffusion maps. *Proceedings of the National Academy of Sciences*, 102(21): 7426–7431, 2005. doi: 10.1073/pnas.0500334102.

[3] Ota et al. Integrating perturb-seq with genetic association data to map causal regulatory programs in k562, 2025. Course resource / internal PDF shared in ML4FG; K562 GWPS anchor for regulator→program effects.

[4] Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999. doi: 10.1038/44565.

[5] Romain Lopez, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15:1053–1058, 2018. doi: 10.1038/s41592-018-0229-2.

[6] Joseph M. Replogle, Thomas M. Norman, and et al. Mapping information flow in mammalian cells with pooled single-cell crispr screens. *Cell*, 185(2):281–299.e19, 2022. doi: 10.1016/j.cell.2021.12.017.

[7] Z. Zhu, F. Zhang, H. Hu, A. Bakshi, M. R. Robinson, J. E. Powell, and et al. Integration of summary data from gwas and eqtl studies predicts complex trait gene targets. *Nature Genetics*, 48(5):481–487, 2016. doi: 10.1038/ng.3538.

[8] A. Zweig, M. Zhang, E. Azizi, and D. A. Knowles. Energy-guided geometric flow matching. *arXiv preprint arXiv:2509.25230*, 2025. URL `https://arxiv.org/abs/2509.25230`.

A Expanded Pipeline

Order of operations.

(i) **Program discovery (unperturbed):** run cNMF on unperturbed counts to obtain loadings W; select K via stability/coherence; annotate programs.

(ii) **Manifold learning (unperturbed):** fit a geometry/energy to induce a Riemannian metric $G(x)$; build a kNN graph and the geometry-aware Laplacian $L_{\mathcal{M}}$.

(iii) **Trait readout (SMR/TWAS):** estimate program→trait coefficients $\theta^{(t)}$ by WLS on SMR effects $s^{(t)}$ with precision Σ^{-1}.

(iv) **GRN prior:** form $\widetilde{\Delta E} = \beta\, u$ for a dose regulated input u (dose-stratified if available).

(v) **Geometry filtering:** obtain $\widehat{\Delta E} = (I + \lambda\, L_{\mathcal{M}})^{-1}\widetilde{\Delta E}$ (Tikhonov/graph smoothing in the learned geometry).

(vi) **Program/trait mapping:** compute $\Delta a = W^{\top}\widehat{\Delta E}$ and $\widehat{\Delta \mathrm{Trait}}^{(t)} = \langle \theta^{(t)}, \Delta a \rangle$.

(vii) **Evaluation and ablations:** report trait fit (R^2, Pearson/Spearman, calibration), LoF/KD sign agreement, dose monotonicity across quartiles, manifold realism via kNN overlap and geodesic displacement, and K562→HCT116 portability; run ablations, i.e., no-manifold, no-GRN, etc..

B Expanded Methods

B.1 GRN priors from Perturb-seq (K562)

From the Ota K562 dataset, estimate an empirical matrix $\beta_{:,r}$ relating regulator r to gene-level expression deltas. For multi-regulator inputs, combine the corresponding columns of β before geometry filtering. Polarity and magnitude are cross-checked against LoF/KD directions reported by *Ota et al.* where applicable.

B.2 Program space via cNMF

Discovery. Grid K in a modest range consistent with the interim ($K \in \{8, 12, 16, 20\}$). For each K run multiple NMF initializations (e.g., NNDSVD start; HALS or multiplicative updates, `max_iter`=1000, `tol`=10^{-4}). Select K by: (i) split-half stability (Jaccard of top genes), (ii) relative reconstruction error, and (iii) biological coherence (GO/MSigDB enrichments). Normalize columns of W; drop weak/duplicated programs; annotate remaining.

B.3 Dose-stratified estimation (planned for final)

When dose proxies are available, stratify cells by sgRNA-UMI quartiles ($Q1 \ldots Q4$). For each regulator r, estimate $\beta^{(q)}_{:,r}$ and use *top-dose* $Q4$ for main analyses; confirm monotone trends by per-gene Spearman across quartiles. Compare pooled vs. $Q4$ through the $\Delta E \mapsto \Delta a \mapsto \widehat{\Delta \mathrm{Trait}}^{(t)}$ pipeline.

B.4 Regulator selection

Include confident dual-sgRNA targets with adequate coverage and baseline expression; define $Q4$ as the top 25% by per-cell sgRNA-UMI (pooled $Q1$–$Q3$ baseline). Exclude targets with inconsistent signs or failing CRISPRi QC; thresholds are defined a priori and logged.

B.5 Manifold constraint (EGGFM-derived metric)

Learn a *Riemannian* cell-state manifold $(\mathcal{M}, G)$ on *unperturbed* cells using energy/score models distilled to a metric tensor $G(x)$ (EGGFM) [8]. Use G to define geodesic distance $d_{\mathcal{M}}$ and build a G-aware Laplacian $L_{\mathcal{M}}$ (neighbors/weights under $d_{\mathcal{M}}$).

1. **Geometry-induced graph.** Construct a kNN graph with weights $w_{ij} = \exp(-d_{\mathcal{M}}(x_i, x_j)^2/\varepsilon)$ and $L_{\mathcal{M}} = I - D^{-1/2} W D^{-1/2}$.
2. **Graph smoothing.** Apply $(I + \lambda L_{\mathcal{M}})^{-1}$ to obtain $\widehat{\Delta E}$ from $\widetilde{\Delta E}$, selecting λ on realism criteria (e.g., kNN-overlap, geodesic displacement).
3. **Program mapping.** Map to programs/traits: $\Delta a = W^\top \widehat{\Delta E}$, $\widehat{\Delta \mathrm{Trait}}^{(t)} = \langle \theta^{(t)}, \Delta a \rangle$.

B.6 SMR linkage to blood traits

Using SMR [7] with HEIDI filtering, regress gene-level SMR effects $s^{(t)}$ onto W to obtain program coefficients:

$$s^{(t)} = W\,\theta^{(t)} + \varepsilon, \qquad \hat{\theta}^{(t)} = (W^\top \Sigma^{-1} W)^{-1} W^\top \Sigma^{-1} s^{(t)}.$$

Predicted trait change for a perturbation:

$$\widehat{\Delta \mathrm{Trait}}^{(t)} \;=\; \langle \theta^{(t)},\, \Delta a \rangle \;=\; \sum_m \theta_m^{(t)}\, \Delta a_m.$$

LD safeguards. Beyond HEIDI: (i) use fine-mapped eQTLs and harmonized alleles for $s^{(t)}$; (ii) when available, require minimal regional colocalization support; (iii) report retained proportions after each filter.