Mechanisms Matter: Transportability of Cellular Perturbation Effects

Anonymous Authors[1]

Abstract

Predicting cellular responses to genetic or chemical perturbations across biological contexts is central to drug development and disease understanding. Despite increases in data and model scale, deep learning models have not consistently outperformed simple baselines. Leveraging causal transportability theory, we show that cross-context generalization is governed by shared causal mechanisms, not merely distributional similarity. To enable controlled evaluation, we develop a causal simulator that generates realistic semi-synthetic Perturb-seq datasets with tunable mechanistic divergence, providing benchmarks with known ground-truth causal structure. Further, we adapt the Vendi diversity score to the perturbation setting as a diagnostic for mode collapse, a failure mode invisible to standard per-perturbation metrics. Extensive experiments across four deep learning models and six simple baselines on semi-synthetic and real Perturb-seq datasets reveal a cross-context generalization gap: performance under cross-context splits drops substantially, often to simple baseline levels. Notably, even on synthetic data with fully specified causal structure, no model generalized across contexts with different causal mechanisms. These results underscore the need for cross-context evaluation, diversity-aware metrics, and mechanistically grounded inductive biases.

1. Introduction

A central goal of computational drug discovery is to predict how cells respond to genetic or chemical perturbations across diverse biological contexts, including cell types, tissues, and genetic backgrounds (Dixit et al., 2016; Hetzel et al., 2022). Accurate cross-context perturbation response prediction could transform target identification and drug development by replacing costly exhaustive experimentation with principled in-silico prioritization (Bock et al., 2022; Roohani et al., 2025).

Large-scale perturbation atlases generated by Perturb-seq (Norman et al., 2019; Replogle et al., 2022; Zhu et al., 2025) and related technologies now catalog single-cell transcriptomic responses to hundreds of perturbations. Yet the combinatorial space of perturbations ($\sim 10^{60}$ chemical compounds, $\sim$20,000 genetic targets), cell types, donors, and disease states remains vastly under-sampled, limiting model generalization. Filling this space experimentally is infeasible: perturbation responses depend on cell type, differentiation stage, culture conditions, and genetic background, so measurements in one context do not automatically transfer to another. The Virtual Cell initiative (Roohani et al., 2025) seeks to bridge this gap through models trained on pooled multi-context data, but under what conditions such transfer is theoretically justified remains unclear.

A growing body of evidence indicates that increasing dataset size and model capacity has not translated into meaningful gains over simple baselines for perturbation response prediction (Bendidi et al., 2024; Ahlmann-Eltze et al., 2025; Csendes et al., 2025; Viñas Torné et al., 2025; Wenteler et al., 2025). Two complementary explanations have emerged. First, standard evaluation metrics may themselves be unreliable: whole-transcriptome Pearson correlation correlates strongly with the cosine similarity between each perturbation effect and the population mean (Viñas Torné et al., 2025; Csendes et al., 2025), rewarding models that merely reproduce the average response. This problem is compounded by mode collapse, in which a model predicts near-identical responses across distinct perturbations (Wu et al., 2025; Mejia et al., 2025), since per-perturbation metrics are structurally blind to this failure: a model that outputs the population-average shift can achieve moderate Pearson correlation and low MAE while capturing no perturbation-specific biology. Second, models may lack the inductive biases needed to capture the causal mechanisms that determine perturbation effects (Lorch et al., 2026; Dibaeinia et al., 2026; Wenteler et al., 2025). Disentangling these failure modes, metric distortion, mode collapse, and model mis-specification, is a prerequisite for meaningful progress.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Submitted to the 2026 Workshop on Generative and Agentic AI for Biology (ICML 2026). Do not distribute.

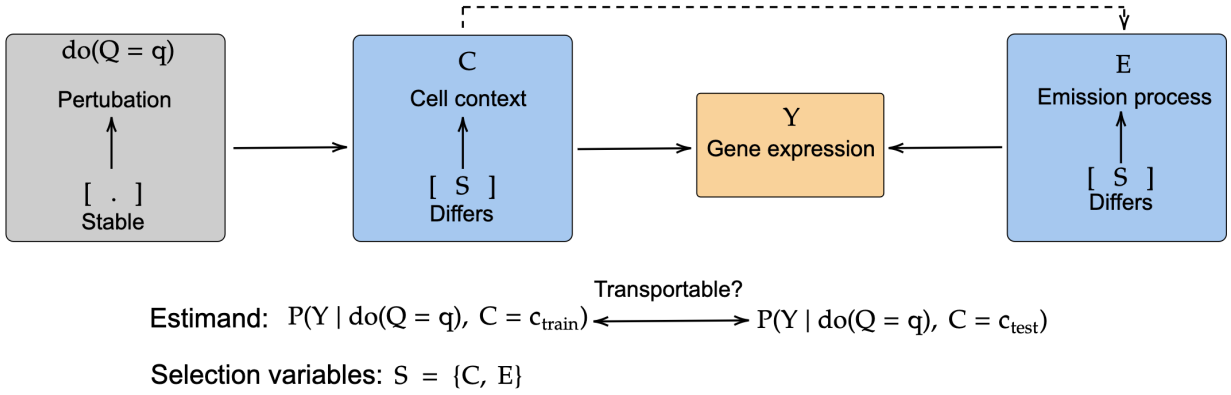

Figure 1. Overview of the transportability problem in cellular perturbation prediction. The goal is to predict $P\big(Y \mid \mathrm{do}(Q=q),\ C=c_{\mathrm{test}}\big)$, an *unseen (perturbation, context) combination*, from training data $\{(q,\ c_{\mathrm{train}}),\ (q',\ c_{\mathrm{test}}) : q' \neq q\}$. The selection variables $S = \{C, E\}$, biological context C and emission process E, may differ between training and test environments, where observed components of E (*e.g.*, library size, batch) can be adjusted for, while latent components (*e.g.*, dropout rate, technical noise) remain sources of domain shift.

Even when evaluation is corrected and models are made more expressive, formal guarantees on cross-context generalization remain absent. Scaling data and architecture (Cui et al., 2024; Hao et al., 2024), encoding perturbation structure (Roohani et al., 2024; Lorch et al., 2026), disentangling perturbation effects (Lotfollahi et al., 2023; Lopez et al., 2018), and matching distributions (Lotfollahi et al., 2019; Bunne et al., 2023; Adduri et al., 2025) have each shown promise within individual contexts, but lack formal conditions for when predictions transfer (Dibaeinia et al., 2026) nor has the cross-context generalization gap itself been systematically quantified. Causal transportability (Bareinboim & Pearl, 2013) offers a principled framework: *a perturbation effect generalizes from one context to another if and only if the causal mechanisms mediating its downstream effect are invariant across both*. As the causal cascades mediating perturbation effects are rarely fully characterized in biology (Amit et al., 2009), this criterion serves as a guiding framework rather than a verifiable condition, but it underscores that distributional similarity alone does not guarantee transportability (Figure 1). To address these challenges, we make the following contributions:

- We formally connect perturbation prediction to causal transportability (Bareinboim & Pearl, 2013), showing that cross-context generalization is governed by shared causal mechanisms, not distributional similarity.
- Extending Lorch et al. (2026) to the cross-context setting, we develop a causal simulator that generates semi-synthetic Perturb-seq data with tunable mechanistic divergence and known ground-truth causal structure, enabling controlled evaluation of generalization as a function of the degree of mechanistic difference between contexts.
- We adapt the Vendi diversity score (Dan Friedman & Dieng, 2023) to the perturbation setting as a reference-free diagnostic for prediction diversity. Coupled with the perturbation discrimination score (PDS (Wu et al., 2025)), which measures whether predictions are correctly matched to their targets, these metrics jointly diagnose mode collapse and quantify discrimination capacity.
- Through extensive evaluation of four deep learning models and six simple baselines (three context-agnostic and their context-specific counterparts) on semi-synthetic and real Perturb-seq data, we demonstrate that progress in perturbation prediction requires closing the cross-context generalization gap with mechanistically grounded models and benchmarks.

2. Perturbation transportability as a causal problem

We formally frame the virtual cell challenge (Roohani et al., 2025) as a causal transportability problem: under what conditions can cellular perturbation effects observed in one biological context be validly transferred to another? Leveraging causal transportability theory (Bareinboim & Pearl, 2013) and a semi-synthetic data-generating process, we derive necessary and sufficient conditions for such transfer within the linear regime and show that when specific *causal mechanisms* differ across contexts, systematic bias is irrecoverable regardless of sample size.

2.1. Problem formulation

Let $Y \in \mathbb{Z}^G$ denote gene-expression counts, $Q \in \{0, 1, \ldots, P\}$ the perturbation index (Q=0 for control), and C the biological context (*e.g.*, cell type, tissue, or experimental condition). The goal of perturbation prediction is to infer

$$P\big(Y \mid \mathrm{do}(Q{=}q),\ C = c_{\mathrm{test}}\big) \tag{1}$$

for an *unseen (perturbation, context) combination*, given training interventional data $\{(q, c_{\text{train}}),\ (q', c_{\text{test}}) : q' \neq q\}$ and control observations $P(Y \mid \text{do}(Q{=}0),\ C{=}c_{\text{test}})$ in the target context. For pseudo-bulk prediction, the estimand reduces to the conditional mean $\mathbb{E}\big[Y \mid \text{do}(Q{=}q),\ C = c_{\text{test}}\big]$. We illustrate the assumed selection diagram for cellular perturbation transportability in Figure 1. The emission process E governs the mapping from the latent state to observed scRNA-seq counts and may itself vary across context, yielding selection variables $S = \{C, E\}$. We assume two key unobserved mechanisms that may differ across cellular contexts:

(i) the causal influence matrix $A_C \in \mathbb{R}^{G\times G}$, a linear approximation of the active gene regulatory network (GRN) in context C: entry $[A_C]_{ij} \neq 0$ indicates that gene j directly regulates gene i, with positive values denoting activation and negative values repression;

(ii) the basal expression vector $B_C \in \mathbb{R}^G$, encoding the context-specific baseline expression program, reflecting the combined influence of transcription rates, epigenetic state, chromatin accessibility, and lineage-specific regulatory activity (Huang et al., 2005; Klemm et al., 2019).

Transportability depends on *which* of these differ across contexts and whether the differences are identifiable from available data. In the trivial case $c_{\text{train}} = c_{\text{test}}$, the problem reduces to interpolation over seen (perturbations, context) combinations, and transportability holds by construction. Throughout, we assume observations $\mathcal{D} = \{(\boldsymbol{y}_i, c_i, q_i)\}_{i=1}^N$.

2.2. Causal simulator with tunable transportability

We generate semi-synthetic scRNA-seq data from a continuous-time linear stochastic differential equation (SDE) in a latent gene-expression space, followed by a negative binomial observation model. We term this simulator CAUSALDGP, see Figure 2, for an illustration of the assumed selection diagram. Following Lorch et al. (2026), we adopt the Ornstein–Uhlenbeck (OU) process (Uhlenbeck & Ornstein, 1930), which is *causally structured, stationary, and analytically tractable*: these properties allow us to derive closed-form transportability conditions while generating realistic gene expression counts through a negative binomial likelihood.

Specifically, we model the latent expression state $X(t) \in \mathbb{R}^G$ of a cell with G genes as a multivariate OU process (Lorch et al., 2026):

$$dX(t) = \big(A_C\, X(t) + B_C + \Gamma_Q\big)\, dt + \sigma\, dW(t)\,, \quad (2)$$

where $A_C \in \mathbb{R}^{G\times G}$ is a sparse causal influence matrix encoding the GRN under biological context C, $B_C \in \mathbb{R}^G$ is a context-specific baseline drift bias, $Q \in \{0, 1, \ldots, P\}$ indexes the perturbation condition, and $\Gamma_Q \in \mathbb{R}^G$ is a sparse shift vector encoding the direct effect of perturbation Q (with $\Gamma_Q = \mathbf{0}$ for the unperturbed control $Q = 0$), $dW(t)$ denotes G-dimensional Brownian motion, and $\sigma > 0$ is a diffusion coefficient encoding *intrinsic biological stochasticity*. We refer to Lorch et al. (2026) for detailed construction of A_C. Provided A_C is Hurwitz (all eigenvalues with strictly negative real part), the process admits a unique stationary Gaussian distribution

$$X_\infty^{(q)} \sim \mathcal{N}\big(\boldsymbol{\mu}_q^C,\ \Sigma_C\big)\,, \qquad \boldsymbol{\mu}_q^C = -A_C^{-1}(B_C + \Gamma_q)\,, \quad (3)$$

where the covariance Σ_C is the unique positive-definite solution of the continuous Lyapunov equation ($A_C\,\Sigma_C + \Sigma_C\,A_C^\top = -\sigma^2\, I_G$).

Emission process In practice, the latent states $X(t)$ are *unobserved*. Given a stationary sample $\boldsymbol{x} \sim X_\infty^{(q)}$, observed counts are generated independently per gene through an emission process E, modeled as a negative binomial (NB) likelihood, capturing both *technical measurement noise* and *intrinsic transcriptional variability*:

$$Y_g \sim \text{NB}\big(\mu_g,\ \theta_g\big)\,, \qquad \mu_g \coloneqq \text{softplus}(x_g)\,, \quad (4)$$

where μ_g is the mean parameter and θ_g is a gene-specific inverse-dispersion. The softplus link $\text{softplus}(z) = \log(1 + e^z)$ maps real-valued latent states to non-negative intensities. The negative binomial captures intrinsic transcriptional noise arising from bursty gene expression (Amrhein et al., 2019), and has become the standard observation model for scRNA-seq (Lopez et al., 2018).

Collectively, the emission process E is fully specified by the gene-wise dispersion parameters $\{\theta_g\}_{g=1}^G$ and the softplus link, defines the mapping from latent state to observed counts. As illustrated in the selection diagram (Figure 2), we assume E is invariant across contexts ($E_C = E$ for all C), so all non-transportability in the semi-synthetic data arises from the latent dynamics alone. In practice, context shifts in library size, capture efficiency, or dispersion would introduce additional measurement-level non-transportability not modeled here. Following Mejia et al. (2025), we set $\theta_g = \hat{\theta}_g$ where $\hat{\theta}_g$ is estimated from the Norman et al. (2019) dataset via maximum likelihood, ensuring realistic count statistics while retaining full control over the latent causal structure. Complete details are provided in Appendix B.2.

2.3. Transportability conditions

Consider two biological contexts (*e.g.*, cell types) C and C' with potentially distinct causal influence matrices $A_C, A_{C'}$ and baseline biases $B_C, B_{C'}$, but a shared perturbation Γ_q. From (3), the perturbation effect in each context is

$$\boldsymbol{\tau}_C^{(q)} \coloneqq \boldsymbol{\mu}_q^C - \boldsymbol{\mu}_0^C = -A_C^{-1}\,\Gamma_q, \qquad \boldsymbol{\tau}_{C'}^{(q)} = -A_{C'}^{-1}\,\Gamma_q. \quad (5)$$

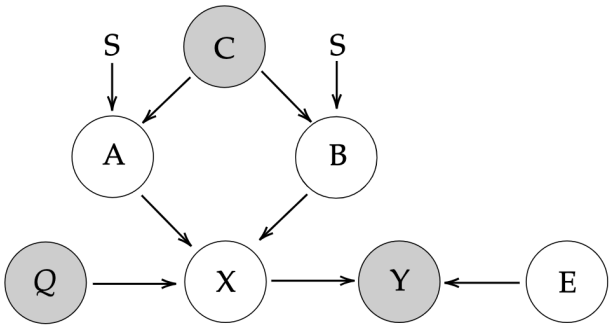

Figure 2. Selection diagram of the assumed CAUSALDGP, with unobserved mechanistic, context-specific selection variables $S = \{A, B\}$. Cell context C, perturbation Q, and gene expression counts Y are observed. The unobserved emission process E, mapping the latent state X to observed counts Y, is invariant across contexts.

Proposition 1 (Transportability of perturbation effects)**.** *A perturbation effect is* transportable *from C to C' if and only if $\boldsymbol{\tau}_C^{(q)} = \boldsymbol{\tau}_{C'}^{(q)}$, which holds whenever*

$$A_C^{-1}\,\Gamma_q = A_{C'}^{-1}\,\Gamma_q. \tag{6}$$

Crucially, this condition depends on *which causal mechanisms differ* between contexts, specifically, the columns of A_C^{-1} and $A_{C'}^{-1}$ indexed by the support of Γ_q, and is neither implied nor precluded by divergence between the marginal distributions $P_C(Y)$ and $P_{C'}(Y)$. Marginal differences can arise from variation in B, σ, or the emission process that leave transportability intact, or from changes in A that do not affect the causal pathways downstream of the perturbed genes. This formalizes the claim that *transportability is governed by shared causal mechanisms, not by distributional similarity*.

Recovering A_C^{-1} from perturbation data requires $P = G$ linearly independent perturbations. When $P < G$, only the restriction of A_C^{-1} to the column space spanned by the observed $\{\Gamma_q\}$ is identifiable, and transportability can be assessed only along those directions. Current Perturb-seq screens operate in the $P \ll G$ regime, making full recovery infeasible. Even when $P = G$, practical identification of A_C^{-1} remains confounded by measurement noise, variable knockdown efficiency, and nonlinearity in the true regulatory dynamics, yielding at best a noisy approximation of the causal influence matrix.

3. Metrics

In practice, most models do not infer the causal matrix A_C (5), and the ground-truth perturbation effect is typically unknown. Deep learning models approximate the full conditional distribution $P_\phi\big(Y \mid \mathrm{do}(Q{=}q),\, C\big)$, from which predicted means are obtained, while simple baselines directly estimate $\mathbb{E}\big[Y \mid \mathrm{do}(Q{=}q),\, C\big]$ by averaging observed training samples. In both cases, we define the predicted perturbation effect as

$$\hat{\boldsymbol{\delta}}_q = \hat{\mathbb{E}}\big[Y \mid \mathrm{do}(Q{=}q),\, C\big] - \hat{\mathbb{E}}\big[Y \mid \mathrm{do}(Q{=}0),\, C\big]\,, \tag{7}$$

where $\hat{\mathbb{E}}$ denotes the model's predicted mean expression, *a.k.a* pseudo-bulk, whether derived from a learned distribution or from empirical averaging. This formulation requires only predicted means for both the perturbed and control conditions, making downstream metrics applicable to all models. Below we describe the proposed Vendi score for model diagnosis. Additional metrics are detailed in the Appendix, spanning three families: *reconstruction metrics* including distributional (MMD, FD, JSD) and point-wise (MSE, MAE) measures (Appendix D.3); *perturbation effect metrics* including Pearson $\boldsymbol{\delta}$ (14), the perturbation effect discrimination score (PDS) (13), and coefficient of determination $R^2_{\boldsymbol{\delta}}$ (15) (Appendix D.1); and *differentially expressed gene (DEG) set metrics* including Jaccard index, recall, and precision (Appendix D.2). Because per-perturbation metrics are sensitive to control bias and perturbation-gene sparsity (Appendix E.4), we report DEG-restricted variants throughout.

Algorithm 1 Cellular Perturbation Response Vendi Score

Require: Count matrices Y_0 (control, $N_0 \times G$), Y_q (perturbation q, $N_q \times G$) for $q = 1, \ldots, P$; PCA dimension d; kernel bandwidth γ; library-size target ℓ
Ensure: Vendi score VS $\in [1,\, P]$
Log-normalize: $\tilde{\boldsymbol{y}}_i \leftarrow \log\big(1 + \ell \cdot \boldsymbol{y}_i / \sum_g y_{ig}\big)$ for all cells i
Fit PCA on controls: $\mathcal{P} \leftarrow \mathrm{PCA}_d(\tilde{Y}_0)$
Embed perturbed cells: $\boldsymbol{z}_i \leftarrow \mathcal{P}(\tilde{\boldsymbol{y}}_i) \in \mathbb{R}^d$ for all non-control cells i
for $q = 1$ **to** P **do**
  $\mathcal{Z}_q \leftarrow \{\boldsymbol{z}_i : q_i = q\}$
end for
Build similarity matrix:
if pseudo-bulk then
  $D_{qq'} \leftarrow \|\bar{\boldsymbol{z}}_q - \bar{\boldsymbol{z}}_{q'}\|_2^2$ {squared Euclidean}
else
  $D_{qq'} \leftarrow \mathrm{MMD}^2_\gamma(\mathcal{Z}_q,\, \mathcal{Z}_{q'})$ {MMD^2 with RBF kernel}
end if
$\sigma_{\mathrm{sim}} \leftarrow \mathrm{median}\big(\{\sqrt{D_{qq'}} : q < q',\ D_{qq'} > 0\}\big)$
$K_{qq'} \leftarrow \exp\big(-D_{qq'}/2\sigma^2_{\mathrm{sim}}\big)$ for $q, q' = 1, \ldots, P$
Compute eigenvalues $\lambda_1, \ldots, \lambda_P$ of K
$\tilde{\lambda}_i \leftarrow \lambda_i / \sum_j \lambda_j$ for $i = 1, \ldots, P$
VS $\leftarrow \exp\big(-\sum_{i=1}^{P} \tilde{\lambda}_i \log \tilde{\lambda}_i\big)$
return VS

Table 1. Perturbation prediction models and their transport assumptions for estimating $P(Y \mid \mathrm{do}(Q{=}q), C{=}c_{\mathrm{test}})$. Full transport formulae are given in Table 3 (Appendix A).

Model	Transport mechanism	Context-aware	GRN prior	Granularity
Control mean	Identity (no perturbation signal)	✗	✗	Pseudo-bulk
Control mean (ctx)	Identity (no perturbation signal)	✓	✗	Pseudo-bulk
Pert. mean	Perturbation average	✗	✗	Pseudo-bulk
Pert. mean (ctx)	Perturbation average	✓	✗	Pseudo-bulk
linearPCA	Additive in gene space	✗	✗	Pseudo-bulk
linearPCA (ctx)	Additive in gene space	✓	✗	Pseudo-bulk
scVI	Additive in latent space	✓	✗	Single-cell
CPA	Additive in latent space	✓	✗	Single-cell
GEARS	Additive residual in gene space	✗	✓	Single-cell
STATE	Nonlinear (attention)	✓	✗	Single-cell

3.1. Perturbation response diversity: Vendi score

Per-perturbation accuracy metrics, such as Pearson $\boldsymbol{\delta}$, evaluate each predicted effect $\hat{\boldsymbol{\delta}}_q$ (7) independently against its ground truth $\boldsymbol{\delta}_q$. These metrics are susceptible to *mode collapse* (Wu et al., 2025): a model that predicts near-identical responses across distinct perturbations achieves moderate per-perturbation accuracy (by staying close to the population mean) while failing to capture perturbation-specific biology. Moreover, these metrics are unreliable when perturbation effects are sparse, reflecting signal strength rather than model quality (Appendix E.4).

To directly quantify this failure, we measure the *diversity* of predicted perturbation responses using a *perturbation Vendi score* adapted from Dan Friedman & Dieng (2023), computed according to Algorithm 1. To mitigate data sparsity, high dimensionality, and technical noise, we first project log-normalized counts into a low-dimensional subspace via PCA fit on control cells. We then construct a pairwise similarity matrix between perturbation populations: (i) for single-cell models that approximate the full conditional $P_\phi\big(Y \mid \mathrm{do}(Q{=}q), C\big)$, we use the non-parametric maximum mean discrepancy (MMD) (Gretton et al., 2012) with an RBF kernel; (ii) for pseudo-bulk baselines, we use Euclidean distance.

Fitting PCA on control cells ensures the embedding captures biologically meaningful variation without being dominated by the perturbation signal itself. The resulting pairwise similarity matrix is robust to cell-to-cell variability within each condition. We compute the Vendi score for both ground-truth and predicted perturbation responses and report the *Vendi ratio*,

$$\mathrm{VR} = \frac{\mathrm{VS}(\hat{Y}_1, \ldots, \hat{Y}_P)}{\mathrm{VS}(Y_1, \ldots, Y_P)}, \tag{8}$$

where VS denotes the Vendi score and Y_q, $\hat{Y}_q$ are the ground-truth and predicted perturbation responses, respectively. Since $\mathrm{VS}(Y_1, \ldots, Y_P) \in [1, P]$ by construction, a ratio near 1 indicates that the model preserves the diversity of perturbation-specific signals; a ratio near $1/P$ diagnoses mode collapse. This failure mode is invisible to per-perturbation metrics: predicting the mean effect $\bar{\boldsymbol{\delta}}$ for every perturbation can correlate moderately with each individual $\boldsymbol{\delta}_q$ while erasing all perturbation-specific signal.

4. Experiments

All code is implemented in Python and will be publicly available on GitHub upon publication.

4.1. Baselines

We consider four deep learning models and six simple baselines (three global and three context-specific variants). The deep learning models include graph-based (GEARS (Roohani et al., 2024)), latent-additive (CPA (Lotfollahi et al., 2023), scVI (Lopez et al., 2018)), and set-transformer-based (STATE (Adduri et al., 2025)) architectures. The simple baselines, namely, control mean, perturbation mean (Csendes et al., 2025; Kernfeld et al., 2023), and linear PCA (Ahlmann-Eltze et al., 2025) are extended to context-specific variants by computing per-context pseudo-bulk means. We adapt scVI to perturbation prediction following the proposed transport formula. Given our focus on transportability of perturbation effects to *unseen (perturbation, context) combinations*, we use the state-transition variant of the STATE model with raw counts as inputs rather than the pretrained state-embedding model. GEARS leverages an external gene ontology database and co-expression patterns derived from training data as inputs to its graph neural network. For synthetic data, we use only the co-expression graph, as no external gene ontology is available. See Table 1 for a summary and Table 3 (Appendix) for detailed transport formulae.

Table 2. Summary of the cellular perturbation datasets after preprocessing. HVG: Highly variable genes.

Statistic	ZHU25	REPLOGLE22	NORMAN19	CAUSALDGP	DIRECTDGP
Cell type	Primary CD4+ T cells	K562 and RPE1 (cell lines)	K562 (cell line)	-	-
Perturbation type	CRISPRi knockdown	CRISPRi knockdown	CRISPRi over-expression	synthetic	synthetic
Control cells N_0	393,216	22,176	11,855	1,024	128–1,024
Perturbed cells N_P	203,870	266,087	71,742	1,024×10	(128–512)×(20–50)
HVGs G	3,463	7,226	8,192	128	1,024–8,192
Perturbations P	91	828	157	128	20–50
Biological replicates	4 (donors)	0	0	-	-
Conditions	3 (Rest, Stim8hr, Stim48hr)	1	1	-	-
Total contexts C	12 (donor × condition)	2	1	2	1

4.2. Datasets

We focus on Perturb-seq datasets because they directly intervene on individual genes, providing a clean experimental analog of the $\mathrm{do}(\cdot)$ operator for probing causal gene-function relationships at single-cell resolution. See Table 2 for summary statistics.

Semi-synthetic benchmarks We use two data generating approaches. DIRECTDGP, adopted from Mejia et al. (2025), simulates scRNA-seq perturbation data within a single context using a negative binomial model with parameters estimated from the NORMAN19 dataset. We use it to study metric distortion in the standard within-context setting. CAUSALDGP (Section 2.1) extends this to two contexts with distinct latent causal mechanisms (A_C, B_C). By varying which mechanisms differ across contexts (`None`, `A`, `B`, `Both`), it enables control over whether perturbation effects are transportable by construction. Refer to Appendix B for complete simulation details.

Real-world datasets We evaluate on three Perturb-seq datasets spanning progressively more challenging transportability settings:

- NORMAN19 (Norman et al., 2019): single- and double-gene perturbations in K562 cells (single context). Used to expose metric distortion on a canonical within-context benchmark.
- REPLOGLE22 (Replogle et al., 2022): genome-scale perturbations across K562 and RPE1 cell lines (two contexts). These lines represent markedly different cellular programs (hematopoietic vs. epithelial), providing a test bed for cross-context transportability.
- ZHU25 (Zhu et al., 2025): large-scale primary human CD4+ T-cell Perturb-seq across 12 contexts (4 donors × 3 activation states). Unlike the immortalized lines in REPLOGLE22, this dataset captures biologically relevant heterogeneity from both activation state and inter-individual variation (Brodin et al., 2015; Chapfuwa et al., 2025).

We provide complete pre-processing details and dataset statistics in Appendix C.1.

4.3. Context-aware splitting strategy

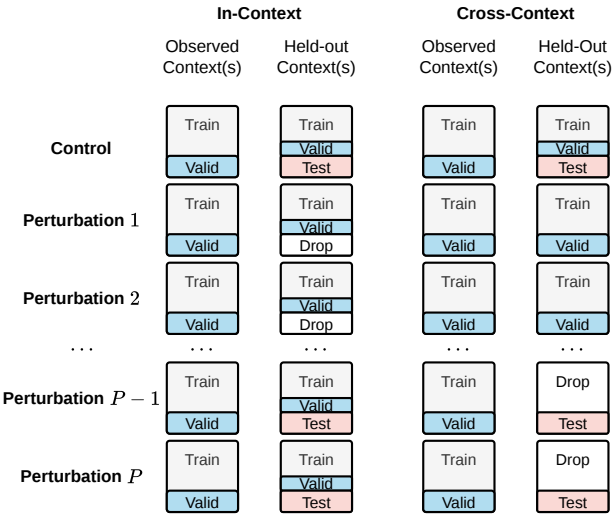

Figure 3. Within-context vs. cross-context splitting strategy. Cells marked "Drop" are removed during fairness alignment.

To fairly compare *within-context* and *cross-context* generalization, we design a splitting strategy that aligns the two evaluation regimes on three axes: held-out perturbations, test-set size, and training-set size (Figure 3). Both splits are derived from a shared random seed and planning step to ensure identical randomization decisions, see Appendix C.2 for details.

Within-context split For each (perturbation, context) combination (q, c), cells are split $50/20/30\%$ into train, validation, and test sets. The model sees every pair during training but is evaluated on held-out cells.

Cross-context split Given a target context c_{test}, we randomly hold out 50% of its non-control perturbations denoted $\mathcal{Q}_{\text{held}}$, from training, creating unseen (perturbation, context) combinations (q, c_{test}) for $q \in \mathcal{Q}_{\text{held}}$. The remaining perturbations in $\{(q', c_{\text{test}}) : q' \notin \mathcal{Q}_{\text{held}}\}$ and all perturbations in other contexts, including $\mathcal{Q}_{\text{held}}$ itself, form the training pool. Control cells in c_{test} are split $50/20/30\%$ as in the

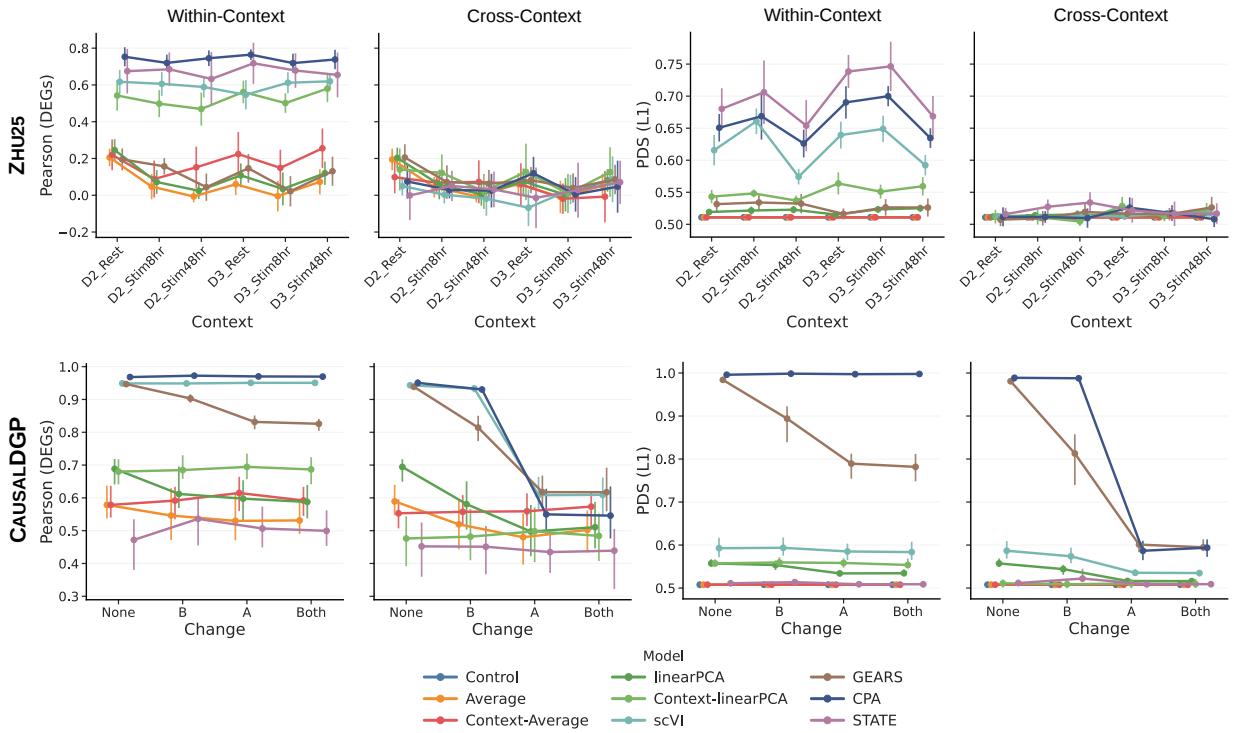

Figure 4. Cross-context generalization gap on ZHU25 (top) and CAUSALDGP (bottom) under within-context and cross-context evaluation for Pearson δ (DEG-restricted) and PDS (ℓ_1). For ZHU25, the x-axis denotes the six test contexts (2 held-out donors $\times$ 3 activation states); for CAUSALDGP, it denotes the four cross-context mechanism modes (`None`, `A`, `B`, `Both`). We plot median with 50% confidence interval (CI). Full results are shown in Figures 9–11 (Appendix E.3).

within-context regime, preserving held-out controls for evaluation.

Fairness alignment To ensure the two regimes are directly comparable, we apply three post-hoc alignment steps (marked "Drop" in Figure 3):

(i) *Perturbation alignment*: the within-context test set in c_{test} is restricted to the same held-out perturbations $\mathcal{Q}_{\text{held}}$, so that both regimes are evaluated on identical (q, c_{test}) combinations.

(ii) *Test-size alignment*: cross-context test sets are downsampled to match the within-context test-set size per perturbation.

(iii) *Train-size alignment*: the larger training set is downsampled to match the smaller.

4.4. Results

4.4.1. MODEL DIAGNOSIS: DIVERSITY VS. DISCRIMINATION

The Vendi ratio (VR) (8) and perturbation discrimination score (PDS) (13) jointly diagnose model performance along the perturbation diversity and discrimination axes (Figure 5). On NORMAN19, all models achieve high PDS, indicating correct perturbation matching, but scVI and CPA exhibit

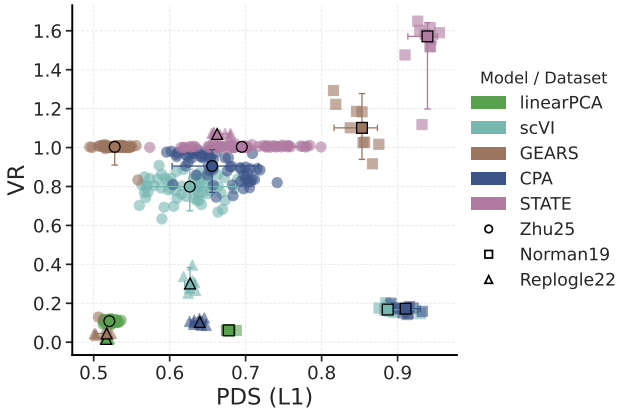

Figure 5. Vendi ratio (VR) vs. PDS across models and datasets under within-context evaluation. Color encodes model and marker shape encodes dataset. Bold markers and error bars show the median and 95% CI per model–dataset pair. Higher is better on both axes.

lower VR, suggesting partial mode collapse despite accurate per-perturbation predictions. STATE consistently attains high PDS and VR across all three datasets. LinearPCA consistently shows low VR, confirming that its additive gene-space transport collapses perturbation diversity. On ZHU25, all models except linearPCA achieve high VR, but GEARS

shows lower PDS, indicating diverse yet poorly matched predictions. On REPLOGLE22, which has the largest number of perturbations, both VR and PDS drop for all models except STATE, consistent with the increased difficulty of maintaining diversity and discrimination as the number of perturbation grows. Under cross-context evaluation (Figure 8 in Appendix), PDS drops across all models and datasets, indicating that discrimination capacity degrades when predicting unseen (perturbation, context) combinations.

4.4.2. CROSS-CONTEXT GENERALIZATION GAP

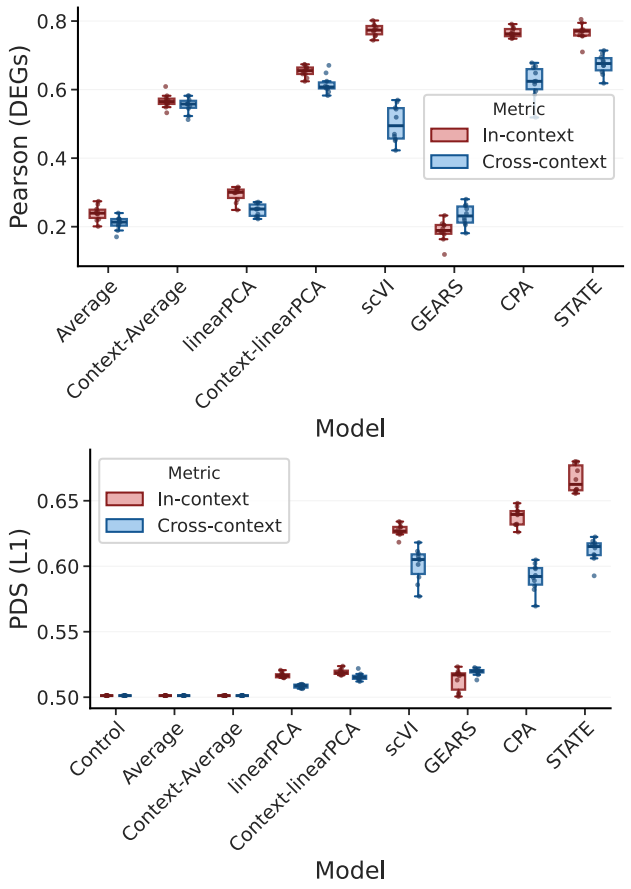

Figure 6. Cross-context generalization gap on REPLOGLE22: Pearson $\boldsymbol{\delta}$ (DEG-restricted) and PDS (ℓ_1) under within-context vs. cross-context evaluation. Models are trained on RPE1 and evaluated on held-out K562 perturbations. Full results in Figure 10 (Appendix E.3).

Figure 4 summarizes the cross-context generalization gap on CAUSALDGP and ZHU25 (full results in Figures 11–9, Appendix). On CAUSALDGP, results are consistent with the transportability condition (6): GEARS and CPA outperform baselines under within-context evaluation, but performance drops when the baseline drift B_C differs across contexts and collapses to baseline levels when the causal influence matrix A_C or both (A_C and B_C) differs from the training context. On ZHU25, we observe a similar pattern: performance drops consistently across the held-out donors and activation states. While CPA and STATE outperform simple baselines under within-context evaluation, their advantage largely vanishes under cross-context splits. This suggests donor-specific variation in the causal mechanisms governing CD4+ T-cell perturbation responses, potentially driven by environmental factors, pathogen exposure history, or inter-individual genetic variation (Brodin et al., 2015; Chapfuwa et al., 2025).

On REPLOGLE22 (Figure 6; full results in Figure 10, Appendix), where models are trained on RPE1 and evaluated on held-out K562 perturbations, the cross-context gap is notably smaller: CPA and STATE remain competitive with simple baselines even under cross-context evaluation. Although Replogle et al. (2022) report that K562 and RPE1 transcriptional phenotypes are only weakly correlated overall, a subset of perturbations nonetheless exhibits conserved effects across these cell lines, and these shared responses may suffice for models that learn to transport the transferable component of the perturbation signal.

Limitations Our theoretical analysis relies on a linear SDE (6), whereas real perturbation responses involve nonlinear, multi-layered regulatory cascades (Amit et al., 2009). Closing this gap requires extending the framework to nonlinear dynamics. No existing benchmark provides ground-truth causal structure; semi-synthetic simulators like CAUSALDGP partially fill this gap but remain constrained by linearity and fixed emission assumptions. Developing benchmarks that combine known causal structure with realistic biological complexity remains an important open challenge.

5. Conclusion

Cross-context generalization in perturbation prediction depends on shared causal mechanisms, not merely distributional similarity. Across four deep learning models and six baselines, performance drops substantially under cross-context evaluation, often matching simple baselines. A joint diversity–discrimination diagnostic reveals complementary failure modes: simple baselines collapse to undifferentiated predictions, while deep learning models exhibit declining diversity as the perturbation space grows and reduced discrimination when evaluated across contexts. Even on synthetic data with known causal structure, no model recovered transportable perturbation effects across contexts under mechanistic divergence, suggesting the bottleneck is architectural. Together, these findings motivate cross-context evaluation as a standard benchmark practice, mechanistic inductive biases as a modeling priority, and multi-context CRISPR epistasis screens as a concrete experimental next step.

References

Adduri, A. K., Gautam, D., Bevilacqua, B., Imran, A., Shah, R., Naghipourfar, M., Teyssier, N., Ilango, R., Nagaraj, S., Dong, M., et al. Predicting cellular responses to perturbation across diverse contexts with state. *bioRxiv*, 2025.

Ahlmann-Eltze, C., Huber, W., and Anders, S. Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines. *Nature Methods*, 2025.

Amit, I., Garber, M., Chevrier, N., Leite, A. P., Donner, Y., Eisenhaure, T., Guttman, M., Grenier, J. K., Li, W., Zuk, O., et al. Unbiased reconstruction of a mammalian transcriptional network mediating pathogen responses. *Science*, 2009.

Amrhein, L., Harsha, K., and Fuchs, C. A mechanistic model for the negative binomial distribution of single-cell mrna counts. *bioRxiv*, 2019.

Bareinboim, E. and Pearl, J. A general algorithm for deciding transportability of experimental results. *Journal of causal Inference*, 2013.

Bendidi, I., Whitfield, S., Kenyon-Dean, K., Yedder, H. B., Mesbahi, Y. E., Noutahi, E., and Denton, A. K. Benchmarking transcriptomics foundation models for perturbation analysis: one pca still rules them all. *arXiv*, 2024.

Benjamini, Y. and Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 1995.

Bock, C., Datlinger, P., Chardon, F., Coelho, M. A., Dong, M. B., Lawson, K. A., Lu, T., Maroc, L., Norman, T. M., Song, B., et al. High-content crispr screening. *Nature Reviews Methods Primers*, 2022.

Brodin, P., Jojic, V., Gao, T., Bhattacharya, S., Angel, C. J. L., Furman, D., Shen-Orr, S., Dekker, C. L., Swan, G. E., Butte, A. J., et al. Variation in the human immune system is largely driven by non-heritable influences. *Cell*, 2015.

Bunne, C., Stark, S. G., Gut, G., Del Castillo, J. S., Levesque, M., Lehmann, K.-V., Pelkmans, L., Krause, A., and Rätsch, G. Learning single-cell perturbation responses using neural optimal transport. *Nature methods*, 2023.

Chapfuwa, P., Demirel, I., Pisani, L., Zazo, J., Portugaly, E., Zahid, H. J., and Greissl, J. Scalable universal t-cell receptor embeddings from adaptive immune repertoires. In *ICLR*, 2025.

Csendes, G., Sanz, G., Szalay, K. Z., and Szalai, B. Benchmarking foundation cell models for post-perturbation rna-seq prediction. *BMC genomics*, 2025.

Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., Duan, N., and Wang, B. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature methods*, 2024.

Dan Friedman, D. and Dieng, A. B. The vendi score: A diversity evaluation metric for machine learning. *TMLR*, 2023.

Dibaeinia, P., Babu, S., Knudson, M., ElSheikh, A., Wen, Y., Liu, H., Perera, J., and Khan, A. A. Virtual cells need context, not just scale. *bioRxiv*, 2026.

Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C. P., Jerby-Arnon, L., Marjanovic, N. D., Dionne, D., Burks, T., Raychowdhury, R., et al. Perturb-seq: dissecting molecular circuits with scalable single-cell rna profiling of pooled genetic screens. *Cell*, 2016.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *JMLR*, 2012.

Hao, M., Gong, J., Zeng, X., Liu, C., Guo, Y., Cheng, X., Wang, T., Ma, J., Zhang, X., and Song, L. Large-scale foundation model on single-cell transcriptomics. *Nature methods*, 2024.

Hetzel, L., Boehm, S., Kilbertus, N., Günnemann, S., Theis, F., et al. Predicting cellular responses to novel drug perturbations at a single-cell resolution. *NeurIPS*, 2022.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 2017.

Huang, S., Eichler, G., Bar-Yam, Y., and Ingber, D. E. Cell fates as high-dimensional attractor states of a complex gene regulatory network. *Physical review letters*, 2005.

Kernfeld, E., Yang, Y., Weinstock, J. S., Battle, A., and Cahan, P. A systematic comparison of computational methods for expression forecasting. *bioRxiv*, 2023.

Klemm, S. L., Shipony, Z., and Greenleaf, W. J. Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics*, 2019.

Lin, J. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 2002.

Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., and Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 2018.

Lorch, L., Zhang, J., Bunne, C., Krause, A., Schölkopf, B., and Uhler, C. Latent causal diffusions for single-cell perturbation modeling. *arXiv*, 2026.

Lotfollahi, M., Wolf, F. A., and Theis, F. J. scgen predicts single-cell perturbation responses. *Nature methods*, 2019.

Lotfollahi, M., Klimovskaia Susmelj, A., De Donno, C., Hetzel, L., Ji, Y., Ibarra, I. L., Srivatsan, S. R., Naghipourfar, M., Daza, R. M., Martin, B., et al. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology*, 2023.

Mejia, G. M., Miller, H. E., Leblanc, F. J., Wang, B., Swain, B., and Camillo, L. P. d. L. Diversity by design: Addressing mode collapse improves scrna-seq perturbation modeling on well-calibrated metrics. *arXiv*, 2025.

Norman, T. M., Horlbeck, M. A., Replogle, J. M., Ge, A. Y., Xu, A., Jost, M., Gilbert, L. A., and Weissman, J. S. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science*, 2019.

Replogle, J. M., Saunders, R. A., Pogson, A. N., Hussmann, J. A., Lenail, A., Guna, A., Mascibroda, L., Wagner, E. J., Adelman, K., Lithwick-Yanai, G., et al. Mapping information-rich genotype-phenotype landscapes with genome-scale perturb-seq. *Cell*, 2022.

Roohani, Y., Huang, K., and Leskovec, J. Predicting transcriptional outcomes of novel multigene perturbations with gears. *Nature Biotechnology*, 2024.

Roohani, Y. H., Hua, T. J., Tung, P.-Y., Bounds, L. R., Yu, F. B., Dobin, A., Teyssier, N., Adduri, A., Woodrow, A., Plosky, B. S., et al. Virtual cell challenge: Toward a turing test for the virtual cell. *Cell*, 2025.

Uhlenbeck, G. E. and Ornstein, L. S. On the theory of the brownian motion. *Physical review*, 1930.

Villani, C. et al. *Optimal transport: old and new*. Springer, 2009.

Viñas Torné, R., Wiatrak, M., Piran, Z., Fan, S., Jiang, L., Teichmann, S. A., Nitzan, M., and Brbić, M. Systema: a framework for evaluating genetic perturbation response prediction beyond systematic variation. *Nature Biotechnology*, 2025.

Welch, B. L. The generalization of 'student's'problem when several different population varlances are involved. *Biometrika*, 1947.

Wenteler, A., Occhetta, M., Branson, N., Curean, V., Huebner, M., Dee, W., Connell, W., Chung, S. P., Hawkins-Hooker, A., Ektefaie, Y., Córdova, C. M. V., and Gallagher-Syed, A. Perteval-scFM: Benchmarking single-cell foundation models for perturbation effect prediction. In *ICML*, 2025.

Wu, Y., Wershof, E., Schmon, S. M., Nassar, M., Osiński, B., Eksi, R., Yan, Z., Stark, R., Zhang, K., and Graepel, T. Perturbench: Benchmarking machine learning models for cellular perturbation analysis. In *NeurIPS Datasets and Benchmarks Track*, 2025.

Zhu, R., Dann, E., Yan, J., Retana, J. R., Goto, R., Guitche, R. C., Petersen, L. K., Ota, M., Pritchard, J. K., and Marson, A. Genome-scale perturb-seq in primary human cd4+ t cells maps context-specific regulators of t cell programs and human immune traits. *bioRxiv*, 2025.

A. Baselines

Table 3. Perturbation prediction models and their assumed transport formulae for estimating $P(Y \mid \mathrm{do}(Q = q),\ C = c_{\mathrm{test}})$ from training data $\{(q,\ c_{\mathrm{train}}),\ (q',\ c_{\mathrm{test}}) : q' \neq q\}$.

Model	Assumed transport formula
linearPCA	$\hat{\boldsymbol{y}}(q) = \underbrace{U}_{\mathbb{R}^{G\times k}}\ \underbrace{W}_{\mathbb{R}^{k\times k}}\ \underbrace{\bar{\boldsymbol{p}}_{\mathcal{T}(q)}}_{\mathbb{R}^{k\times 1}} + \underbrace{\boldsymbol{b}}_{\mathbb{R}^{G\times 1}}$, where U are PCA gene embeddings with k components learned from training perturbation pseudo-bulks, $\bar{\boldsymbol{p}}_{\mathcal{T}(q)} = \frac{1}{\lvert\mathcal{T}(q)\rvert}\sum_{t\in\mathcal{T}(q)} U_t$ averages over targeted gene embeddings, $W \in \mathbb{R}^{k\times k}$ is fit via ridge regression, and $\boldsymbol{b}$ is mean expression across all training perturbation pseudo-bulks. *Additive and context-free*: no conditioning on C. Operates on pseudo-bulk perturbation means, predicts single expression vector per perturbation. Assumes perturbation effects are fully determined by target gene identity.
scVI	$\hat{Y}(q, c_{\mathrm{test}}) = \mathrm{Dec}\big(\bar{\boldsymbol{z}}_{\mathrm{ctrl}}(c_{\mathrm{test}}) + \overline{\boldsymbol{\delta z}}(q),\ \ q,\ \ c_{\mathrm{test}}\big)$, where $\boldsymbol{z} = \mathrm{Enc}(\boldsymbol{y})$, $\bar{\boldsymbol{z}}(q,c) = \frac{1}{\lvert\mathcal{N}_{q,c}\rvert}\sum_{n\in\mathcal{N}_{q,c}} \boldsymbol{z}_n$ is the empirical mean over training data cells with perturbation q in context c, and $\overline{\boldsymbol{\delta z}}(q) = \frac{1}{\lvert\mathcal{C}_q\rvert}\sum_{c\in\mathcal{C}_q}\big[\bar{\boldsymbol{z}}(q,c) - \bar{\boldsymbol{z}}_{\mathrm{ctrl}}(c)\big]$. *Additive in latent space*: decoder conditioned on both q and C provides partial modulation in gene space via the ZINB likelihood.
CPA	$\hat{\boldsymbol{y}}_i(q, c_{\mathrm{test}}) = \mathrm{Dec}\big(\underbrace{\mathrm{Enc}(\boldsymbol{y}_{\mathrm{ctrl},i})}_{\boldsymbol{z}_{\mathrm{basal}}} + \underbrace{\mathrm{MLP}_{\mathrm{pert}}(q)}_{\boldsymbol{z}_{\mathrm{pert}}} + \underbrace{\mathrm{MLP}_{\mathrm{cov}}(c_{\mathrm{test}})}_{\boldsymbol{z}_{\mathrm{cov}}}\big)$, where $\boldsymbol{y}_{\mathrm{ctrl},i}$ is a per-cell control expression input. *Additive in latent space*: learned perturbation and covariate embeddings compose linearly with the encoded basal state. Adversarial training encourages $\boldsymbol{z}_{\mathrm{basal}}$ to be perturbation- and covariate-invariant.
GEARS	$\hat{\boldsymbol{y}}_i(q, c_{\mathrm{test}}) = \boldsymbol{y}_{\mathrm{ctrl},i}(c_{\mathrm{test}}) + \mathrm{Dec}\big(\Phi_{\mathrm{gene}}(E_g, G_{\mathrm{coexp}}) + \mathbb{I}_{g\in\mathcal{T}(q)} \cdot \Psi(\Phi_{\mathrm{pert}}(E_q, G_{\mathrm{sim}}))\big)$ *Additive residual in gene space*: perturbation injected only at target gene positions, propagated to other genes via a cross-gene MLP. Shared graphs G_{coexp}, G_{sim} are context-invariant.
STATE	$\hat{\boldsymbol{y}}_i(q, c_{\mathrm{test}}) = \mathrm{Dec}\big(\mathrm{Transformer}\big(\mathrm{MLP}_{\mathrm{pert}}(q) + \mathrm{MLP}_{\mathrm{basal}}(\boldsymbol{y}_{\mathrm{ctrl},i}) + \mathrm{Emb}(c_{\mathrm{test}})\big) + \boldsymbol{z}_{\mathrm{basal},i}\big)$, equivalently $\hat{\boldsymbol{y}}_i = f\big(\boldsymbol{y}_{\mathrm{ctrl},i},\ q,\ c_{\mathrm{test}},\ \{\boldsymbol{y}_{\mathrm{ctrl},j}\}_{j\neq i}\big)$, where f is the full forward map over a set of S covariate-matched control cells and $\boldsymbol{z}_{\mathrm{basal},i} = \mathrm{MLP}_{\mathrm{basal}}(\boldsymbol{y}_{\mathrm{ctrl},i})$. *Additive at the transformer input*: the residual skip adds $\boldsymbol{z}_{\mathrm{basal},i}$ back before decoding, preserving per-cell identity. Nonlinear and non-separable after attention, *implicit transport formula defined by learned parameters*.

B. Data generation process

Here we provide additional implementation details for the two data-generating processes used in this paper. See Table 4 for a summary of the hyperparameters.

B.1. Additional details for DIRECTDGP

DIRECTDGP is adapted from the simulation framework of Mejia et al. (2025). We refer the reader to that work for the full derivation and motivation, and depict the process in Algorithm 2 for completeness. Briefly, DIRECTDGP generates single-cell perturbation data in a *single biological context*, with gene-wise negative-binomial counts calibrated to mimic key statistical properties of NORMAN19, including sparse counts, heterogeneous library size, and control-referenced bias. In

Table 4. Simulation parameters for the two semi-synthetic data-generating processes used in this work. DIRECTDGP, adapted from Mejia et al. (2025), is primarily used for within-context metric stress tests, whereas CAUSALDGP is the proposed two-context mechanistic simulator for studying transportability under controlled mechanism shift.

Parameter	Meaning	DIRECTDGP	CAUSALDGP
Shared parameters			
$\lvert\mathcal{C}\rvert$	Number of biological contexts	1	2
G	Number of genes	1024–8192	128
N_0	Number of control cells	128–1024	1024
N_k	Number of cells per perturbation	128–512	1024
P	Number of perturbations	20–50	128
θ_g	Gene-wise inverse-dispersion	Estimated from NORMAN19	Estimated from NORMAN19
DirectDGP-specific parameters			
p_{effect}	Probability that a gene is perturbed	0.001–0.1	-
ϵ	Multiplicative perturbation factor	1.2–5.0	-
β	Control-bias scale	0.0–2.0	-
μ_ℓ	Mean log-library size	0.2–5.0	-
μ_g^c	Baseline control mean	Estimated from NORMAN19	-
λ_g	Gene-specific control-bias term	Estimated from NORMAN19	-
CausalDGP-specific parameters			
A_c	Context-specific interaction matrix	-	Sparse, randomized, spectrally shifted to be Hurwitz
B_c	Context-specific baseline bias	-	Entries sampled from $\text{Unif}([-3,3])$
Γ_q	Perturbation shift vector	-	One-sparse intervention
s_q	Perturbation magnitude	-	$\text{Unif}\big([-6\log G, -2\log G] \cup [2\log G, 6\log G]\big)$
ρ_{swap}	Row-wise rewiring fraction	-	0.5
σ	Diffusion coefficient in latent OU process	-	$\sqrt{2}$
β_{sp}	Softplus parameter	-	3

this paper, we use DIRECTDGP primarily as a controlled within-context benchmark for studying metric distortion rather than transportability.

B.2. Additional details of CAUSALDGP

This section provides additional implementation details for CAUSALDGP. Recall from Section 2.2 that, conditional on biological context $C \in \{0,1\}$ and perturbation $Q \in \{0,1,\dots,P\}$, the latent state $X(t) \in \mathbb{R}^G$ evolves according to an affine OU process, and the observed counts are generated through a gene-wise negative binomial emission process. Here we specify how the simulator instantiates the causal interaction matrix A_C, the context-specific baseline drift B_C, the shift vector Γ_Q, and the sampling procedure.

Gene sub-sampling and empirical anchoring To anchor the observation model to realistic single-cell count statistics, we subsample G genes without replacement from a larger empirical pool of M genes whose dispersion parameters have been estimated from the NORMAN19 dataset:

$$I = \{i_1, \dots, i_G\} \subset \{1, \dots, M\}\,.$$

We retain the corresponding gene names and inverse-dispersion parameters,

$$\theta_g = \widehat{\theta}_{i_g}, \qquad g = 1, \dots, G\,.$$

Construction of the sparse causal interaction matrix As we described in Section 2.2, the entry convention for the causal interaction matrix A_C is

$$[A_C]_{ij} \neq 0 \qquad \Longleftrightarrow \qquad \text{gene } j \text{ directly regulates gene } i\,.$$

Algorithm 2 DIRECTDGP

Require: Number of genes G, perturbations P, control cells N_0, perturbed cells per perturbation N_k, empirical parameters $\{\mu_g^c, \theta_g, \lambda_g\}_{g=1}^G$ (estimated from NORMAN19), simulation parameters $\beta, \delta, \epsilon, \mu_\ell, \sigma_\ell^2$

for $j = 1, \ldots, N_0$ **do**
  Sample library size $\ell_j \sim \mathrm{LogNormal}(\mu_\ell, \sigma_\ell^2)$
  for $g = 1, \ldots, G$ **do**
    Set $\mu_{g,j}^{(0)} \leftarrow \ell_j \, \mu_g^c$
    Draw $Y_{g,j}^{(0)} \sim \mathrm{NB}(\mu_{g,j}^{(0)}, \theta_g)$
  end for
end for
for $q = 1, \ldots, P$ **do**
  for $j = 1, \ldots, N_k$ **do**
    Sample library size $\ell_j \sim \mathrm{LogNormal}(\mu_\ell, \sigma_\ell^2)$
    for $g = 1, \ldots, G$ **do**
      Sample $\alpha_{q,g} \in \{1, \epsilon, 1/\epsilon\}$ with probabilities $(1 - p_{\text{effect}},\ p_{\text{effect}}/2,\ p_{\text{effect}}/2)$
      Set $\mu_{g,j}^{(q)} \leftarrow \ell_j \, \alpha_{q,g} (\mu_g^c + \beta \lambda_g)$
      Draw $Y_{g,j}^{(q)} \sim \mathrm{NB}(\mu_{g,j}^{(q)}, \theta_g)$
    end for
  end for
end forreturn Count matrix $\{Y^{(q)}\}_{q=0}^P$ with metadata Q_j

Thus, the deterministic drift of gene i depends on its upstream regulators:

$$\frac{dX_i(t)}{dt} \;=\; \sum_{j=1}^{G} [A_C]_{ij} X_j(t) + [B_C]_i + [\Gamma_Q]_i \,.$$

For each target gene i, we sample exactly 10 distinct regulators from $\{1, \ldots, G\} \setminus \{i\}$ and assign each edge an independent signed weight from $\mathrm{Unif}\big([-3, -1] \cup [1, 3]\big)$. A randomly generated sparse matrix is not necessarily stable. To guarantee that (2) admits a unique stationary Gaussian distribution, we stabilize each raw matrix by shifting along the identity:

$$\widetilde{A}_C = A_C - (\widehat{\alpha}_C - \gamma) I \,, \tag{9}$$

where $\widehat{\alpha}_C$ estimates the spectra abscissa (largest real part among eigenvalues) of the raw A_C and $\gamma < 0$ is a user-chosen stability margin (set to $\gamma = -1$ throughout). If λ is an eigenvalue of A_C, the corresponding eigenvalue of $\widetilde{A}_C$ is

$$\widetilde{\lambda} = \lambda - (\widehat{\alpha}_C - \gamma) \,.$$

When the spectral–abscissa estimate is accurate,

$$\max_i \Re(\widetilde{\lambda}_i) = \gamma < 0 \,,$$

so $\widetilde{A}_C$ is Hurwitz. Throughout the paper we overload notation and write A_C for the stabilized matrix.

Mechanism-level context diversity The two contexts $C \in \{0, 1\}$ differ through their latent mechanisms (A_C, B_C), not through post-hoc output perturbations. We define four diversity modes:

$$\begin{aligned}
\texttt{None}: &\quad A_1 = A_0 \,, \quad B_1 = B_0 \,, \\
\texttt{A}: &\quad A_1 \neq A_0 \,, \quad B_1 = B_0 \,, \\
\texttt{B}: &\quad A_1 = A_0 \,, \quad B_1 \neq B_0 \,, \\
\texttt{Both}: &\quad A_1 \neq A_0 \,, \quad B_1 \neq B_0 \,.
\end{aligned}$$

When matrix diversity is enabled, A_1 is obtained by row-wise rewiring of A_0: in each row a fraction of existing nonzero entries is moved to previously zero positions, preserving both the row sparsity pattern and the multiset of edge weights while changing which regulators act on each target. Each context matrix is then stabilized independently via the identity shift in (9), since rewiring can alter the spectral abscissa. When bias diversity is enabled, B_1 is resampled coordinate-wise from $[B_1]_g \sim \mathrm{Unif}([-3, 3])$. By construction, cross-context non-transportability arises directly from mechanism shift.

Perturbation shift vector Each non-control perturbation targets exactly one gene. We sample P distinct targets

$$t_1, \ldots, t_P \in \{1, \ldots, G\}$$

without replacement and define

$$\Gamma_0 = 0, \qquad \Gamma_q = s_q e_{t_q}, \quad q = 1, \ldots, P,$$

where e_{t_q} is the t_q-th standard basis vector and the perturbation magnitudes s_q are drawn from

$$s_q \sim \mathrm{Unif}\Big([-6 \log G, -2 \log G] \cup [2 \log G, 6 \log G]\Big).$$

Each perturbation therefore acts directly on a single coordinate but propagates globally through the network dynamics.

Euler–Maruyama sampling Although the stationary distribution of (2) is Gaussian in latent space, we sample from it via Euler–Maruyama (EM) discretization with step size Δt:

$$X_{n+1} = X_n + \Delta t\,(A_C X_n + B_C + \Gamma_Q) + \sigma\sqrt{\Delta t}\,\varepsilon_n, \qquad \varepsilon_n \sim \mathcal{N}(0, I_G). \tag{10}$$

We set $\Delta t = 10^{-3}$ with a burn-in of 12,000 steps and thinning every 10 steps. Let

$$M_C = I + \Delta t\, A_C, \qquad d_{c,q} = B_c + \Gamma_q.$$

Then (10) becomes

$$X_{n+1} = M_C\, X_n + \Delta t\, d_{c,q} + \sigma\sqrt{\Delta t}\,\varepsilon_n.$$

Emission process and count parameterization Given a latent state $\boldsymbol{x} \in \mathbb{R}^G$, we map it to nonnegative gene-wise mean expression via the softplus link:

$$\mu_g(\boldsymbol{x}) = \mathrm{softplus}_{\beta_{\mathrm{sp}}}(x_g) = \frac{1}{\beta_{\mathrm{sp}}} \log\big(1 + e^{\beta_{\mathrm{sp}} x_g}\big). \tag{11}$$

The observed count for gene g is then drawn independently as

$$Y_g \mid X = \boldsymbol{x} \quad \sim \quad \mathrm{NB}\left(n = \theta_g,\ p_g = \frac{\theta_g}{\theta_g + \mu_g(\boldsymbol{x})}\right). \tag{12}$$

Under this parameterization,

$$\mathbb{E}[Y_g \mid X = \boldsymbol{x}] = \mu_g(\boldsymbol{x}),$$
$$\mathrm{Var}[Y_g \mid X = \boldsymbol{x}] = \mu_g(\boldsymbol{x}) + \frac{\mu_g(\boldsymbol{x})^2}{\theta_g},$$

recovering the standard negative-binomial mean–dispersion model used in scRNA-seq analysis.

Condition-wise sampling scheme Within each condition, context labels are assigned independently as

$$C_i \sim \mathrm{Bernoulli}(0.5),$$

so the realized number of cells per context is random rather than exactly balanced. Conditional on (C_i, Q_i), we instantiate the corresponding EM sampler, draw a latent state, transform it via the softplus link, and sample observed counts from the negative-binomial model. Algorithm 3 summarizes the full procedure for generating the CAUSALDGP dataset.

Algorithm 3 CAUSALDGP

Require: Number of genes G, perturbations P, control cells per condition N_0, perturbed cells per condition N_k, gene dispersions $\{\theta_g\}$ (estimated from NORMAN19), diversity mode $\in \{\texttt{None}, \texttt{A}, \texttt{B}, \texttt{Both}\}$

Sample G genes and corresponding dispersions from an empirical gene pool; randomly permute gene indices
Construct sparse signed interaction matrix A_0 and baseline drift B_0
Stabilize A_0 via spectral shifting (9) to ensure it is Hurwitz
Construct (A_1, B_1) from (A_0, B_0) according to the diversity mode; stabilize A_1 independently
Sample distinct perturbation targets $\{t_q\}_{q=1}^{P}$ and magnitudes $\{s_q\}_{q=1}^{P}$; set $\Gamma_0 \leftarrow 0$ and $\Gamma_q \leftarrow s_q e_{t_q}$ for $q \geq 1$
for $q = 0, \ldots, P$ **do**
  Set $N_q \leftarrow N_0$ if $q = 0$, else $N_q \leftarrow N_k$
  for $i = 1, \ldots, N_q$ **do**
    Sample context $C_i \sim \text{Bernoulli}(0.5)$
    Draw latent state X_i via Euler–Maruyama (10) with parameters $(A_{C_i}, B_{C_i}, \Gamma_q)$
    Compute mean $\mu_g(X_i)$ via softplus link in (11)
    Draw counts $Y_{ig} \sim \text{NB}(\theta_g, \mu_g(X_i))$ for each gene g (12)
  end for
end forreturn Count matrix $\{Y_i\}$ with metadata (Q_i, C_i) for all cells

Remarks Two implementation details are worth noting. First, because the emission process is shared across contexts, all non-transportability in CAUSALDGP arises from latent mechanism differences in (A_C, B_C), not from measurement-level shift. Second, because the affected-gene masks are defined from equilibrium mean shifts, they provide exact ground truth for evaluating whether a model recovers the downstream support of each perturbation effect.

Visualization Figure 7 illustrates how CAUSALDGP generates context variation at the mechanism level and how this variation manifests in the simulated single-cell expression manifold. The heatmaps in the top row display the latent parameters for the two contexts as stacked blocks $[A_C^\top; B_C]$, combining the sparse signed interaction matrix A_C with the context-specific baseline drift B_C. This highlights that the simulator does not create a new context by simply shifting observed expression values after sampling. Instead, the context enters upstream through the latent dynamics themselves, by altering the regulatory interactions, the baseline drift, or both.

The four UMAP panels in the bottom row then show how these mechanism choices translate into observable geometry. Each point is a simulated cell, color identifies perturbation, and marker shape identifies the context. Under `None` ($A_1 = A_0$, $B_1 = B_0$), the two contexts overlap almost completely, as expected when perturbation effects are fully transportable by construction. Under `A`, where only the interaction matrix differs across contexts, the global shape of the manifold is visibly warped, reflecting changes in how perturbation signals propagate through the regulatory network. Under `B`, where only the baseline drift changes, the manifold exhibits a more uniform displacement. Under `Both`, the geometry is the most strongly separated, indicating the largest departure from transportability.

C. Datasets

C.1. Perturb-seq dataset pre-processing

Table 5 summarizes the pre-processing pipeline. We follow Dibaeinia et al. (2026) for ZHU25 and Mejia et al. (2025) for NORMAN19 and REPLOGLE22.

C.2. Context-aware splitting strategy

Shared pre-split planner Before constructing either final split, we run a shared planning step. The planner constructs provisional splits for both strategies, selects the perturbations to hold out in the target context(s), records the within-context test-set size for each context-perturbation group, and computes the overall training-set size under both schemes. These quantities are then fed into the subsequent fairness-alignment stage. This shared planner ensures that the final within-context and cross-context splits are directly comparable, rather than reflecting different random choices or different effective data budgets.

Table 5. Data pre-processing pipeline for real-world Perturb-seq datasets. All datasets undergo library-size normalization to 10,000 counts per cell followed by $\log(1+y)$ transformation. HVG selection uses the SEURAT V3 method throughout.

Pre-processing step	NORMAN19	REPLOGLE22	ZHU25
Cell-level QC			
Min. genes per cell	200	200	100
UMI outlier filtering	—[a]	—[a]	MAD-based
Gene-level QC			
Min. cells per gene	3	3	100
Gene set	Per-dataset	Shared across cell lines	Shared across 12 contexts
Feature selection			
No. HVGs	7,226	8,192	3,463
Perturbation filtering			
Min. cells per perturbation	256	64	64
Knockdown quality filter	—[a]	—[a]	Ratio < 0.5
Min. donors per timepoint	—	—	2
Dataset summary			
Contexts	1 (K562)	2 (K562, RPE1)	12 (4 donors $\times$ 3 activation states)
Perturbation type	Single + double gene	Single gene	Single gene

[a] Applied by original authors prior to data release.

D. Metrics

D.1. Perturbation effect metrics

Perturbation effect discrimination score (PDS) To assess whether each predicted perturbation effect is closer to its own ground truth than to those of other perturbations, we adopt the PDS of Wu et al. (2025). For each perturbation q, r_q measures the fraction of competing predictions that the matched prediction $\hat{\boldsymbol{\delta}}_q$ outranks:

$$\mathrm{PDS} \coloneqq \frac{1}{Q}\sum_{q=1}^{Q} r_q\,, \qquad r_q \coloneqq \frac{1}{Q-1}\sum_{\substack{q'=1\\ q'\neq q}}^{Q} \mathbb{I}\big(d(\hat{\boldsymbol{\delta}}_q,\, \boldsymbol{\delta}_q) \;\le\; d(\hat{\boldsymbol{\delta}}_{q'},\, \boldsymbol{\delta}_q)\big)\,, \tag{13}$$

where $\hat{\boldsymbol{\delta}}_q$ is the predicted perturbation effect (7), $\boldsymbol{\delta}_q$ is the corresponding observed effect, and d is a distance function (we report ℓ_1, but ℓ_2 and cosine variants could also be considered). A perfect model places every prediction closest to its own target, yielding $\mathrm{PDS} = 1$; a mode-collapsed model that emits near-identical profiles across perturbations achieves $\mathrm{PDS} \approx 0.5$, the expected value under random assignment. PDS therefore complements per-perturbation metrics such as Pearson $\boldsymbol{\delta}$ and $R^2_{\boldsymbol{\delta}}$, which can remain high when a model captures the shared control-to-perturbation shift but fails to distinguish individual perturbation effects. Unlike the Vendi score, which requires no ground truth, PDS requires matched observations and is applicable only at evaluation time. Together, the Vendi ratio and PDS provide complementary diagnostics: the former measures whether predictions are *diverse*, the latter whether they are *correctly matched* to their targets.

Pearson $\boldsymbol{\delta}$ and $R^2_{\boldsymbol{\delta}}$ For each perturbation q, we compute gene-level effects $\boldsymbol{\delta}_q$ and $\hat{\boldsymbol{\delta}}_q$ from the observed and predicted mean profiles, respectively. The Pearson correlation between $\hat{\boldsymbol{\delta}}_q$ and $\boldsymbol{\delta}_q$ across genes measures whether the model recovers the correct *direction* of perturbation effects irrespective of their magnitude:

$$\begin{aligned}\mathrm{Pearson}_{\boldsymbol{\delta}}(\boldsymbol{\delta}_q,\, \hat{\boldsymbol{\delta}}_q) &= \frac{\sum_{g=1}^{G}\big(\delta_{q,g}-\bar{\delta}_q\big)\big(\hat{\delta}_{q,g}-\bar{\hat{\delta}}_q\big)}{\sqrt{\sum_{g=1}^{G}\big(\delta_{q,g}-\bar{\delta}_q\big)^2}\,\sqrt{\sum_{g=1}^{G}\big(\hat{\delta}_{q,g}-\bar{\hat{\delta}}_q\big)^2}} \\ &= \frac{\tilde{\boldsymbol{\delta}}_q \cdot \tilde{\hat{\boldsymbol{\delta}}}_q}{\|\tilde{\boldsymbol{\delta}}_q\|\,\|\tilde{\hat{\boldsymbol{\delta}}}_q\|}\,,\end{aligned} \tag{14}$$

where $\tilde{\boldsymbol{\delta}}_q = \boldsymbol{\delta}_q - \bar{\delta}_q \mathbf{1}$. The coefficient of determination $R^2_{\boldsymbol{\delta}}$ additionally penalizes miscalibrated effect sizes:

$$R^2_{\boldsymbol{\delta}}(\boldsymbol{\delta}_q, \hat{\boldsymbol{\delta}}_q) = 1 - \frac{\sum_{g=1}^{G} \big(\delta_{q,g} - \hat{\delta}_{q,g}\big)^2}{\sum_{g=1}^{G} \big(\delta_{q,g} - \bar{\delta}_q\big)^2}\,. \tag{15}$$

A model that predicts effects proportional to the truth ($\hat{\boldsymbol{\delta}}_q = \alpha\, \boldsymbol{\delta}_q$, $\alpha \neq 1$) achieves perfect Pearson $\boldsymbol{\delta}$ but $R^2_{\boldsymbol{\delta}} < 1$. Both metrics are computed across all G genes and, separately, restricted to differentially expressed genes (DEGs) identified by a Welch t-test with Benjamini–Hochberg correction (Welch, 1947; Benjamini & Hochberg, 1995) at significance level $\alpha_{\mathrm{FDR}} = 0.05$. We report the median over perturbations. Because both metrics are evaluated per perturbation and then aggregated, a mode-collapsed model that predicts the population-average effect can still attain moderate Pearson $\boldsymbol{\delta}$ whenever perturbation effects share a common direction, though $R^2_{\boldsymbol{\delta}}$ is more sensitive to this failure. This motivates the complementary use of PDS (13) and VR (8).

D.2. DEG-level set metrics

Beyond weighting gene-level scores by DEG membership, we treat differential expression as a binary classification problem. For each perturbation q, let $\mathcal{D}_q = \{g : p^{\mathrm{adj}}_{q,g} < \alpha_{\mathrm{FDR}}\}$ denote the set of observed DEGs, where $p^{\mathrm{adj}}_{q,g}$ is the Benjamini–Hochberg-adjusted p-value from a Welch t-test comparing single-cell expression of gene g between $P\big(Y_g \mid \mathrm{do}(Q{=}q),\, C\big)$ and $P\big(Y_g \mid \mathrm{do}(Q{=}0),\, C\big)$, with $\alpha_{\mathrm{FDR}} = 0.05$. Let $\hat{\mathcal{D}}_q$ denote the analogous set obtained by applying the same test to predicted single-cell profiles. Standard set-overlap metrics then quantify whether the model preserves the combinatorial identity of each perturbation:

- *Jaccard similarity:* $J_q = |\mathcal{D}_q \cap \hat{\mathcal{D}}_q| \,/\, |\mathcal{D}_q \cup \hat{\mathcal{D}}_q|$,
- *Precision:* $\mathrm{Prec}_q = |\mathcal{D}_q \cap \hat{\mathcal{D}}_q| \,/\, |\hat{\mathcal{D}}_q|$,
- *Recall:* $\mathrm{Rec}_q = |\mathcal{D}_q \cap \hat{\mathcal{D}}_q| \,/\, |\mathcal{D}_q|$.

Recall, also called *top-k DEG recall* when $\hat{\mathcal{D}}_q$ is defined by the k most significant predicted genes, is the variant adopted by Wu et al. (2025). These metrics complement the continuous scores above: a model can achieve high Pearson $\boldsymbol{\delta}$ by capturing large effects while missing weaker but biologically significant genes, a failure that low recall exposes. We report the median of each metric over perturbations.

D.3. Reconstruction metrics

Mean absolute and mean squared error We report point-wise reconstruction error on raw counts. Let $\bar{\mu}_{q,g}$ denote the empirical mean expression of gene g among cells receiving perturbation q, and let $\hat{\mu}_{q,g}$ denote the corresponding model prediction. For perturbation q:

$$\mathrm{MAE}_q = \frac{1}{G}\sum_{g=1}^{G} |\hat{\mu}_{q,g} - \bar{\mu}_{q,g}|\,, \qquad \mathrm{MSE}_q = \frac{1}{G}\sum_{g=1}^{G} (\hat{\mu}_{q,g} - \bar{\mu}_{q,g})^2\,. \tag{16}$$

Unlike Pearson $\boldsymbol{\delta}$ and $R^2_{\boldsymbol{\delta}}$, these errors are computed on absolute expression levels rather than shifts from control and are hence sensitive to baseline calibration, *i.e.*, unperturbed gene expression contributes to the error independently of perturbation-effect accuracy. As with the effect-based scores, both are evaluated across all G genes and restricted to DEGs, and we report the median MAE over perturbations.

Distributional distances The pseudo-bulk metrics above collapse each perturbation population to a single mean vector, discarding cell-to-cell heterogeneity. Distribution-level metrics retain this information and fall into two complementary classes.

Parametric (raw counts) Because scRNA-seq counts are sparse and often heavy-tailed, we fit a negative binomial to the observed and predicted count vectors for each gene and measure divergence between the fitted distributions, via Jensen–Shannon divergence (JSD) (Lin, 2002). This approach respects the discrete, over-dispersed nature of the data but requires choosing a likelihood family and is sensitive to library-size normalization.

Non-parametric (latent embedding) An alternative projects both observed and predicted cells into a lower-dimensional representation, such as PCA or a foundation model embedding (*e.g.*, scGPT (Cui et al., 2024)), and compares the resulting

point clouds with a distributional distance. Common choices include the maximum mean discrepancy $\mathrm{MMD}(\hat{P}_q, P_q)$ with an RBF kernel (Gretton et al., 2012), the p-Wasserstein distance (Villani et al., 2009), and the Fréchet distance (FD) computed from the first two moments of the embedding (Heusel et al., 2017). These metrics are agnostic to the count distribution but depend on the quality and dimensionality of the embedding. As with the Vendi score (Algorithm 1), we compute non-parametric distances in PCA space: (i) for MMD, the projection is fit to observed control cells; (ii) for FD, it is fit to observed perturbed cells. In both cases, predicted cells are projected through the same learned PCA before computing the distance.

E. Full experimental results

E.1. Reproducing the experiments

All experiments were conducted on a SLURM-managed high-performance computing cluster using GPU-enabled compute nodes. Data generation, loading, and pre-processing used multi-process CPU workers (Intel Xeon Gold 6448Y), while model training and inference used GPU acceleration (NVIDIA L40S, 48 GB). Each job was allocated sufficient system memory (64 GB for synthetic datasets and 256 GB for real datasets) to support parallel workers and GPU execution. Runtimes generally ranged from several hours to approximately 3 days for the largest configurations over 10 runs.

E.2. Model diagnosis: diversity vs. discrimination

Figure 8 provides the cross-context counterpart of Figure 5 on the two multi-context datasets, ZHU25 and REPLOGLE22. Relative to within-context evaluation, the most consistent change is a leftward shift in PDS, indicating that discrimination degrades once models must predict unseen (perturbation, context) combinations. On ZHU25 (circles), nearly all models cluster near PDS ≈ 0.5, close to the random-assignment regime, even when their VR remains close to 1. This suggests that several models still generate predictions with substantial diversity, but that diversity is no longer correctly aligned with the target perturbation. In particular, scVI, GEARS, and STATE all maintain high VR on ZHU25 yet achieve low PDS, whereas linearPCA fails on both axes with near-chance discrimination and strongly collapsed diversity. CPA occupies an intermediate position, retaining more diversity than linearPCA but still showing a clear drop in both metrics relative to its within-context performance. On REPLOGLE22 (squares), the degradation is milder and the models separate more clearly by failure mode. STATE remains in the most desirable upper-right region, achieving both high VR and the strongest PDS among the compared methods. By contrast, scVI and CPA attain moderate PDS but very low VR, indicating that they partially preserve perturbation discrimination while collapsing the diversity of predicted responses. GEARS shows the opposite pattern: its VR remains high, but its PDS stays near chance, implying diverse yet poorly discriminated predictions. Taken together, Figure 8 shows that cross-context transfer can fail in at least two distinct ways (loss of discrimination and loss of diversity) and that these failure modes are model-dependent. This complements the main-text within-context analysis by showing that strong in-context performance does not necessarily translate to robust generalization across biological contexts.

E.3. Cross-context generalization gap

Appendix Figures 9–11 provide the full per-context results underlying Section 4.4.2 across all three benchmarks. We report five complementary metrics: $R^2_{\boldsymbol{\delta}}$ (DEGs-restricted; plotted as $-\log(1 - R^2_{\boldsymbol{\delta}})$ to accommodate its wide range) and DES (recall), where larger values indicate better perturbation-specific agreement; MAE (DEGs-restricted) and MMD, where smaller values are better; and VR, which measures prediction diversity.

E.3.1. CROSS-CONTEXT GENERALIZATION GAP FOR ZHU25

Figure 9 shows that the cross-context generalization gap on ZHU25 is broad and systematic across all six held-out donor–activation combinations. Under within-context evaluation, CPA and STATE are generally the strongest models: they attain the highest $R^2_{\boldsymbol{\delta}}$ and DES (recall), competitive or lowest MAE, and favorable MMD, clearly outperforming the simple baselines. However, this advantage contracts sharply under cross-context evaluation. For nearly all methods, $R^2_{\boldsymbol{\delta}}$ drops toward or below zero, DES (recall) decreases substantially, and MAE increases across held-out contexts, indicating that perturbation-specific agreement becomes much harder once the target donor and activation state are unseen during training.

The degradation appears across all six contexts, suggesting that the gap reflects a genuine mechanism shift across donor–activation combinations rather than a failure specific to one context. STATE remains comparatively strong under cross-context

evaluation in terms of DES (recall) and MMD, but its margin over simpler approaches is much smaller than under within-context evaluation. CPA shows a similar loss of advantage: it remains competitive, yet the gap between CPA and the stronger baselines narrows substantially once the (perturbation, context) combinations are unseen.

An important nuance is that VR changes much less than the other metrics. This indicates that the cross-context failure on ZHU25 is not simply a collapse of the prediction manifold. Instead, many models continue to generate diverse outputs, but those outputs are no longer well aligned with the correct perturbation in the held-out biological context.

E.3.2. CROSS-CONTEXT GENERALIZATION GAP FOR REPLOGLE22

Figure 10 shows that the cross-context gap on REPLOGLE22 is real but substantially smaller than on ZHU25. This is consistent with the interpretation in Section 4.4.2: because K562 and RPE1 perturbation responses are weakly correlated (Replogle et al., 2022), transport across these two cell lines is more plausible than across donors and activation states in ZHU25. As a result, cross-context evaluation reduces performance but does not erase it. In particular, STATE remains the strongest and most stable model overall, while CPA also stays competitive under transfer from RPE1 to held-out K562 perturbations.

Specifically, for R^2_{δ} and MAE on DEGs, STATE and CPA degrade only moderately and remain favorable relative to the baselines, whereas scVI shows a larger drop. The context-aware linear baselines are also informative: Context-Average and Context-linearPCA remain comparatively strong on correlation-style metrics, suggesting that a substantial transportable component of the perturbation response can already be captured by simple context-conditioned summaries. This helps explain why the advantage of the more expressive models is smaller on REPLOGLE22 than on ZHU25.

E.3.3. CROSS-CONTEXT GENERALIZATION GAP FOR CAUSALDGP

Figure 11 provides the cleanest controlled view of the cross-context gap, because the source of context shift is known and tunable by construction.

The results are consistent with the transportability condition in Section 2.3. When there is no mechanism shift (`None`), the strongest models under within-context evaluation, particularly CPA and GEARS, remain relatively strong under cross-context evaluation as well, although some degradation is still visible. When only the baseline drift B_C changes (`B`), performance degrades moderately: CPA and GEARS lose some R^2_{δ} and DES (recall) but they still remain clearly above the simple baselines.

The picture changes dramatically when the latent interaction matrix A_C differs across contexts. Under `A` and `Both`, the performance of CPA and GEARS drops sharply: R^2_{δ} collapses toward baseline levels, MAE increases markedly, and DES (recall) declines substantially. This pattern is exactly what one would expect if successful cross-context prediction requires the underlying perturbation-response mechanism to be preserved. Changing only the baseline state makes transfer harder, but changing the regulatory mechanism that governs how perturbations propagate through the system is far more damaging.

Another striking feature is that VR does not mirror this deterioration. For several models, VR remains high or even increases under `A` and `Both`, despite the sharp decline in R^2_{δ} and DES (recall). This again underscores that diversity and discrimination are distinct axes: a model can continue to generate varied outputs under mechanism shift while failing to align those outputs with the correct perturbation-specific target.

In summary, these results directly support the interpretation that the cross-context generalization gap is smallest when contexts share the same underlying mechanism, larger when only the baseline state changes, and largest when the perturbation-response mechanism itself differs.

E.4. Metric distortion

Figure 12 evaluates *metric distortion*, i.e., whether an evaluation metric varies systematically with a nuisance property of the data rather than purely reflecting model quality. In panel (a), the nuisance factor is the control bias (β), while in panels (b) and (c), it is the observed DEG percentage. If a metric changes substantially as these nuisance factors vary, part of its value may reflect dataset characteristics such as bias strength or perturbation magnitude rather than genuine predictive performance.

In panel (a) each point represents one evaluation result for a given model under one sampled setting, and the dashed curve shows a model-specific locally estimated scatterplot smoothing (LOESS) trend that summarizes how the metric changes with

the nuisance factor. In the remaining panels, we omit individual points to avoid visual clutter and show only the smoothed trends.

The annotation at the top of each subplot reports the Pearson correlation (R) between the metric and the x-axis variable, computed over all raw points in that panel, together with its two-sided p-value. A robust metric should exhibit relatively flat LOESS curves and a small absolute correlation, indicating that it is not tracking the nuisance variable. Because these panels contain many observations, very small p-values can arise even for weak effects. We therefore place greater emphasis on the magnitude of the correlation and the visual slope of the smoothed curves than on statistical significance alone.

Metric distortion with respect to control bias on DIRECTDGP In DIRECTDGP (Algorithm 2), control bias is introduced by the additive term $\beta\lambda_g$ in the perturbed mean $\mu_{g,j}^{(q)} = \ell_j \alpha_{q,g}(\mu_g^c + \beta\lambda_g)$. Here λ_g is a gene-specific bias pattern estimated from the real dataset, and β controls its strength. Because this term is shared across all perturbations, it induces a common control-referenced shift that is not specific to any individual perturbation. Consequently, larger β can make simulated data appear easier to predict even without actually improving recovery of perturbation-specific effects. We therefore prefer metrics that are insensitive to control bias; otherwise, a metric may reward nuisance structure in the simulator rather than genuine model quality, leading to unfair comparisons.

In panel (a) of Figure 12, Pearson $\boldsymbol{\delta}$, $R^2_{\boldsymbol{\delta}}$, and DES (Jaccard) all show noticeable dependence on control bias β, with correlations of $R = 0.38$, 0.16, and 0.44, respectively (all $P \leq 1 \times 10^{-18}$). Their LOESS curves also drift visibly upward as β increases, even for simple baselines such as Control or Average. This indicates that these metrics improve systematically as control bias strengthens, meaning their values are partly driven by a nuisance property of the data-generation process. By contrast, VR is the only metric in panel (a) that remains largely insensitive to control bias, with near-zero and non-significant correlation ($R = -0.02$, $P = 0.681$) and comparatively flat trends across models. This suggests that VR is substantially more robust to control bias than the other metrics considered.

Metric distortion with respect to observed DEG percentage Panels (b) and (c) of Figure 12 examine distortion with respect to DEG percentage on DIRECTDGP and NORMAN19, respectively. A strong correlation with the observed DEG percentage is undesirable because it implies that the metric is partly measuring perturbation strength or task difficulty rather than prediction quality alone. Such dependence can confound comparisons across perturbations, since models may appear better simply because the underlying perturbation affects a larger fraction of genes. We therefore prefer metrics that are less sensitive to DEG percentage.

For both DIRECTDGP and NORMAN19, the DEG-restricted versions of Pearson $\boldsymbol{\delta}$ and $R^2_{\boldsymbol{\delta}}$ are consistently more robust than their all-gene counterparts. On DIRECTDGP, Pearson $\boldsymbol{\delta}$ computed on all genes has a positive correlation with observed DEG percentage ($R = 0.15$), whereas the DEG-only version is essentially uncorrelated ($R = 0.00$). A similar pattern holds for $R^2_{\boldsymbol{\delta}}$: the all-gene version has $R = 0.15$, while the DEG-only version is much less sensitive ($R = -0.02$). On NORMAN19 the same trend is even clearer. Pearson $\boldsymbol{\delta}$ on all genes shows strong dependence on DEG percentage ($R = 0.56$), whereas the DEG-only version reduces this substantially ($R = 0.26$). Likewise, $R^2_{\boldsymbol{\delta}}$ drops from $R = 0.28$ (all genes) to $R = 0.04$ (DEG-restricted). These results indicate that restricting evaluation to DEGs reduces metric distortion and yields a fairer assessment across perturbations with different response magnitudes.

In contrast, MAE is not robust in either formulation. On DIRECTDGP, both all-gene and DEG-only MAE vary systematically with DEG percentage, and on NORMAN19 both versions also remain sensitive to this nuisance factor. Unlike Pearson $\boldsymbol{\delta}$ and $R^2_{\boldsymbol{\delta}}$, switching from all genes to DEGs does not resolve the distortion problem for MAE. This suggests that MAE remains entangled with perturbation magnitude and is therefore less suitable when robust evaluation across heterogeneous perturbations is desired.

Finally, the distance-based metrics in the bottom row of panels (b) and (c) are generally less distorted than several of the gene-wise metrics. Among them, MMD appears the most robust, showing the weakest dependence on observed DEG percentage. Parametric distance and FD exhibit somewhat larger drift with DEG percentage in at least some settings. Taken together, these results suggest that MMD is the most insensitive of the distance-based metrics to variation in perturbation strength.

F. Broader impacts

This work provides a framework for identifying when perturbation predictions are unreliable across biological contexts, which can help prevent misallocation of experimental resources in drug development. We caution that inflated within-context metrics may mask cross-context failures. Hence model predictions should not guide therapeutic decisions without independent validation. We do not foresee direct negative societal consequences from this evaluation framework.

G. Declaration of LLM usage

We used a large language model only to assist with language editing and checklist wording, while all scientific content, experimental design, analysis, and conclusions were developed by the authors.

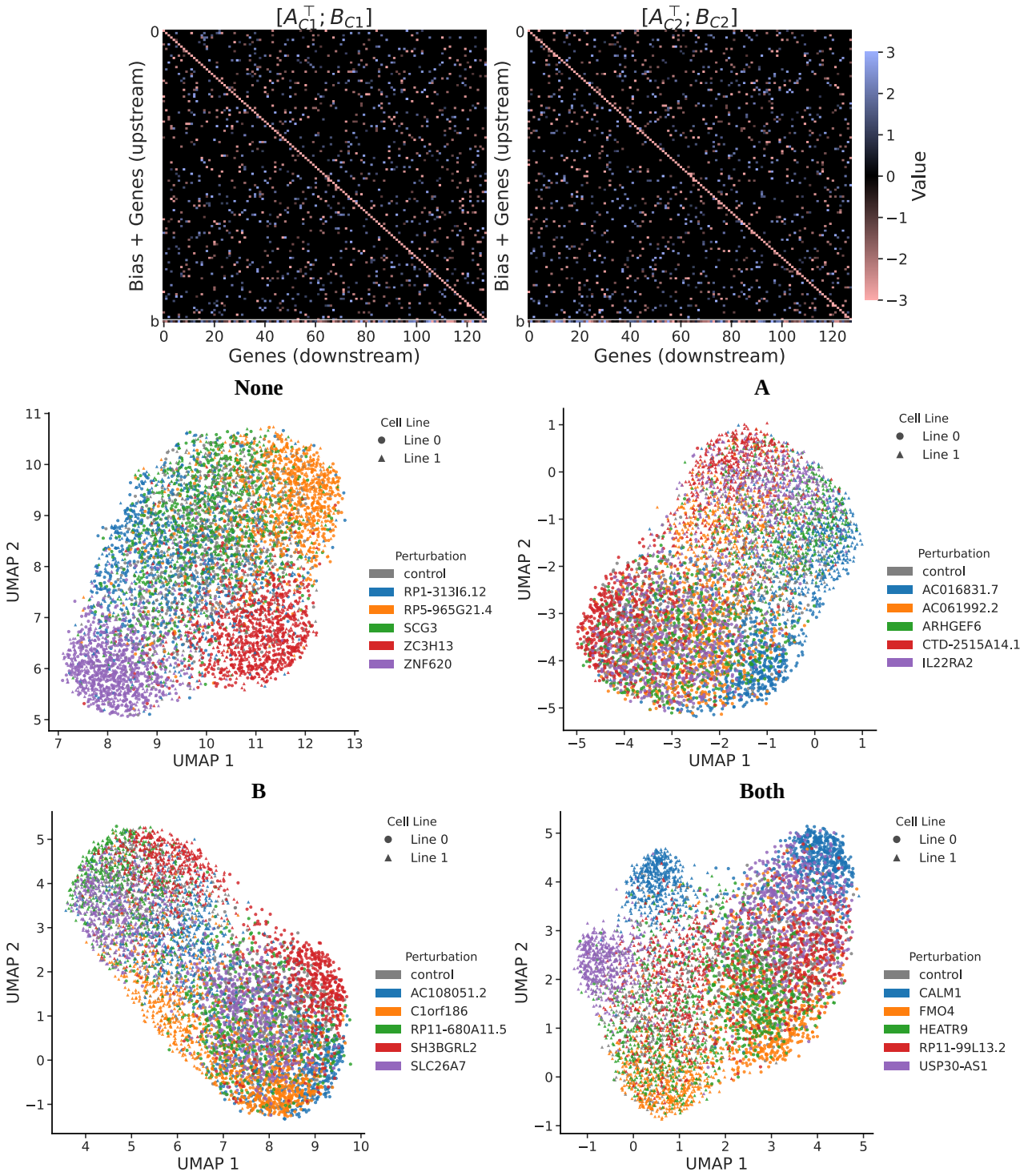

Figure 7. Qualitative visualization of CAUSALDGP under the four mechanism-diversity modes. **Top row:** context-specific latent mechanisms for the two simulated contexts, shown as concatenated parameter blocks $[A_C^\top;\ B_C]$, where A_C is the sparse signed interaction matrix and B_C is the baseline drift vector. **Bottom row:** UMAP projections of simulated cells under each diversity mode (`None`, `A`, `B`, `Both`), with 128 genes, 5 perturbations plus control, and 1,024 cells per condition. UMAP was computed on the top 50 principal components of the raw count matrix. Color denotes perturbation identity, and marker shape denotes context $C \in \{0, 1\}$.

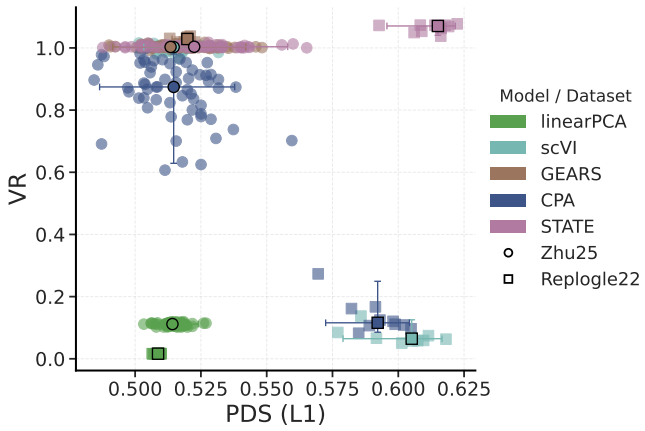

Figure 8. Vendi ratio (VR) vs. PDS across models and datasets under cross-context evaluation. Color encodes the model and marker shape encodes dataset. Bold markers and error bars show the median and 95% CI per model–dataset pair. Higher is better on both axes.

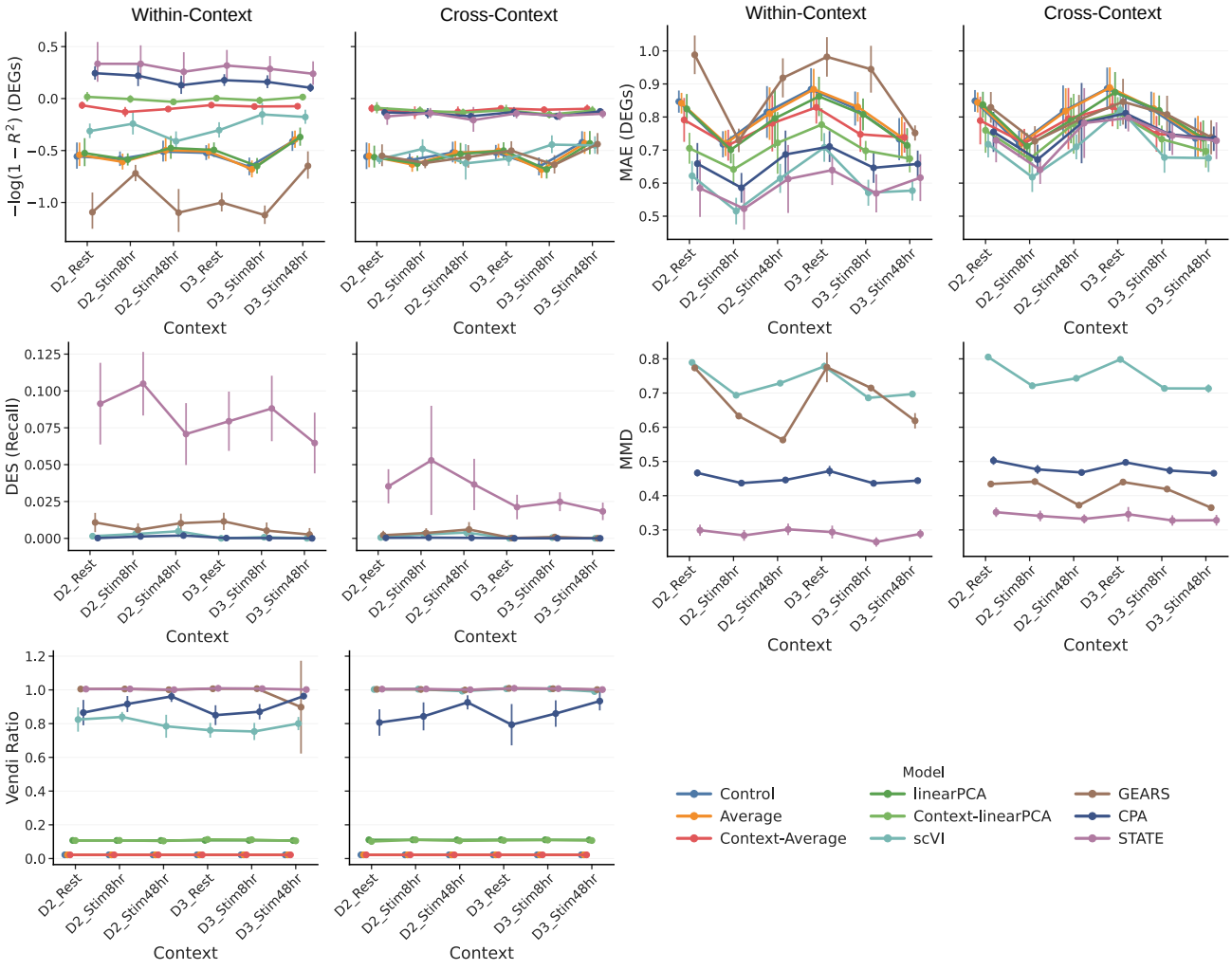

Figure 9. Additional results on the cross-context generalization gap in ZHU25 under within-context and cross-context evaluation. The x-axis denotes the six test contexts (2 held-out donors $\times$ 3 activation states). We plot median with 50% CI.

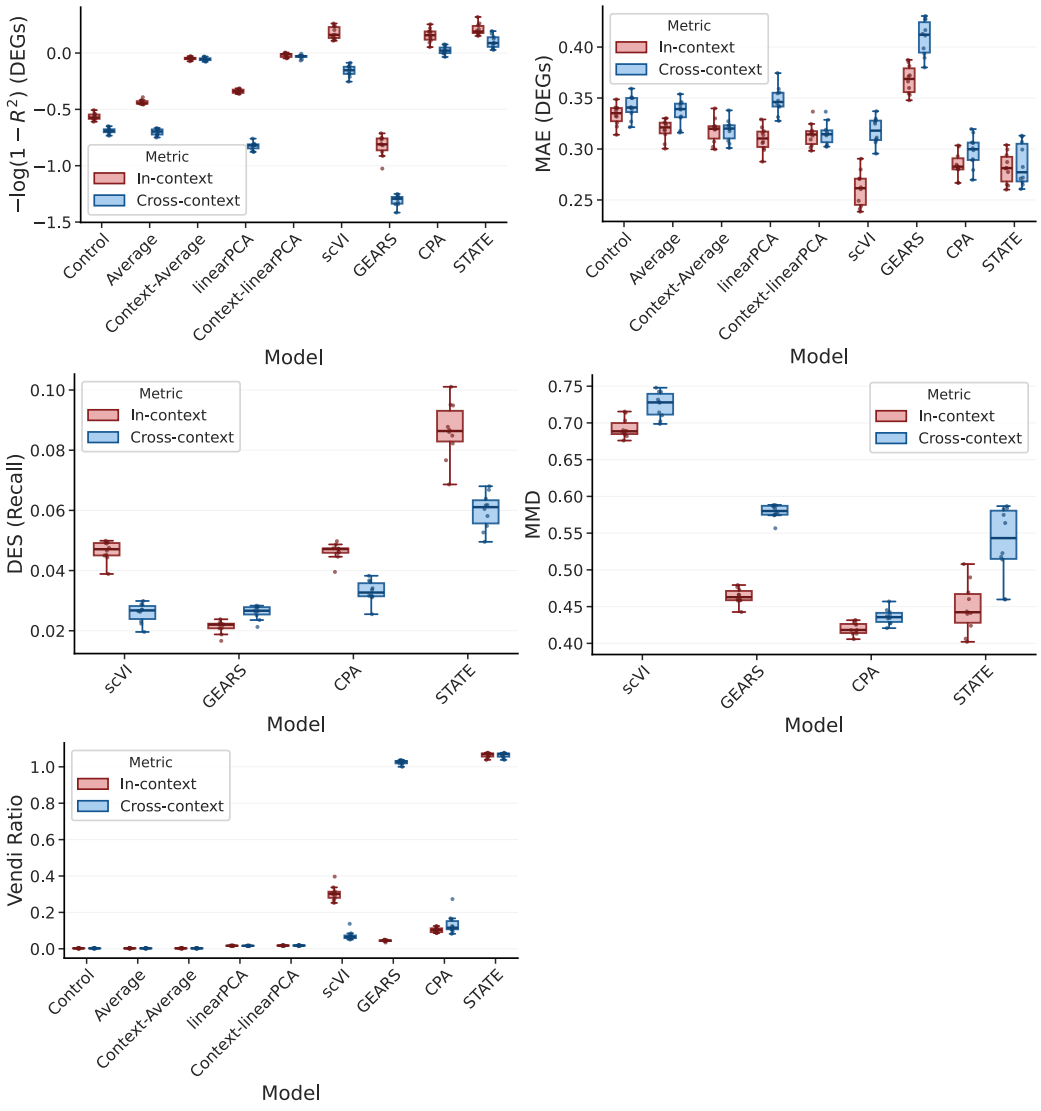

Figure 10. Additional results for the cross-context generalization gap on REPLOGLE22 under within-context vs. cross-context evaluation. Models are trained on RPE1 and evaluated on held-out K562 perturbations.

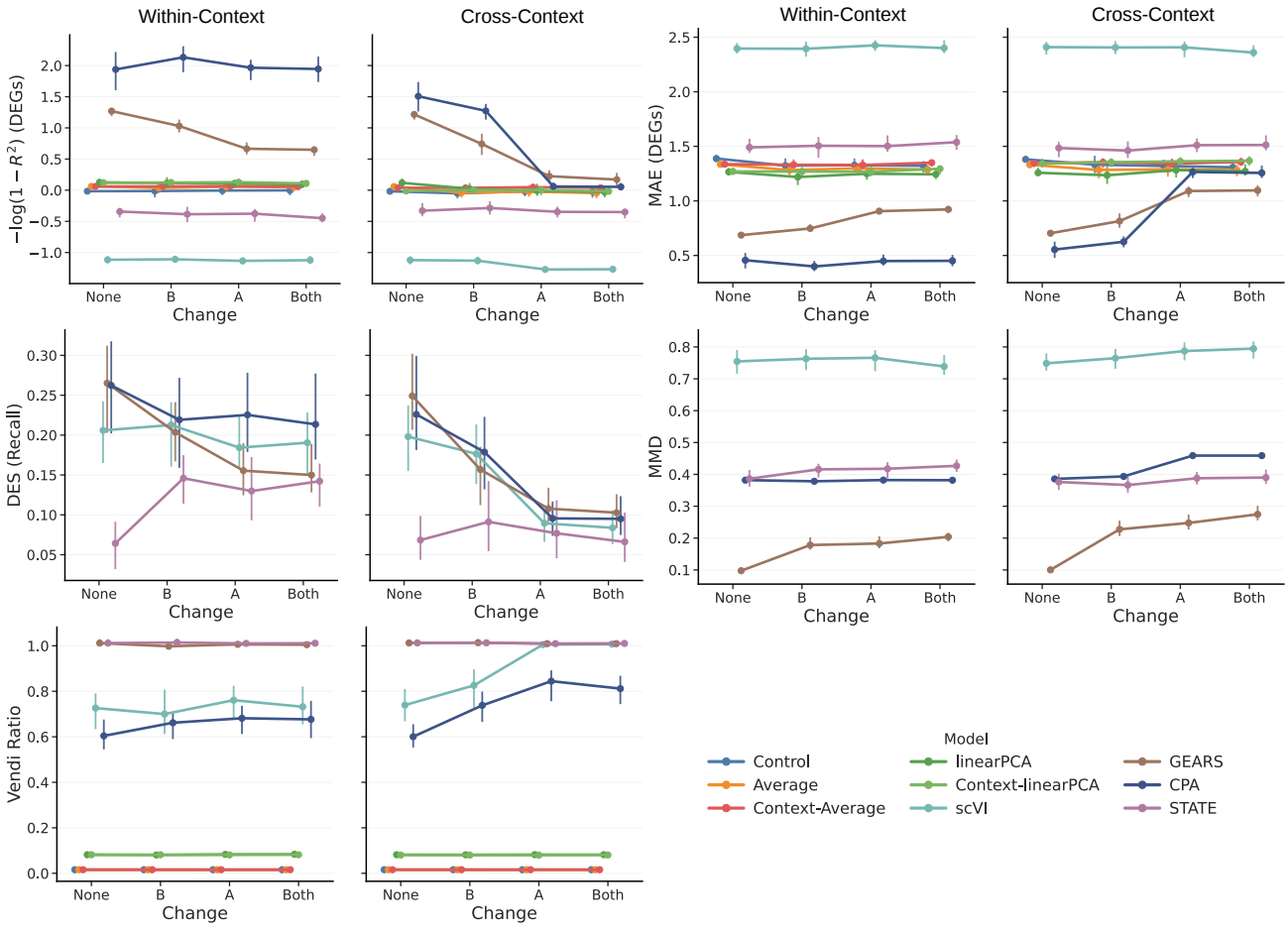

Figure 11. Additional results on the cross-context generalization gap in CAUSALDGP under within-context and cross-context evaluation. The x-axis denotes the four cross-context mechanism modes (`None`, `A`, `B`, `Both`). We plot median with 50% CI.

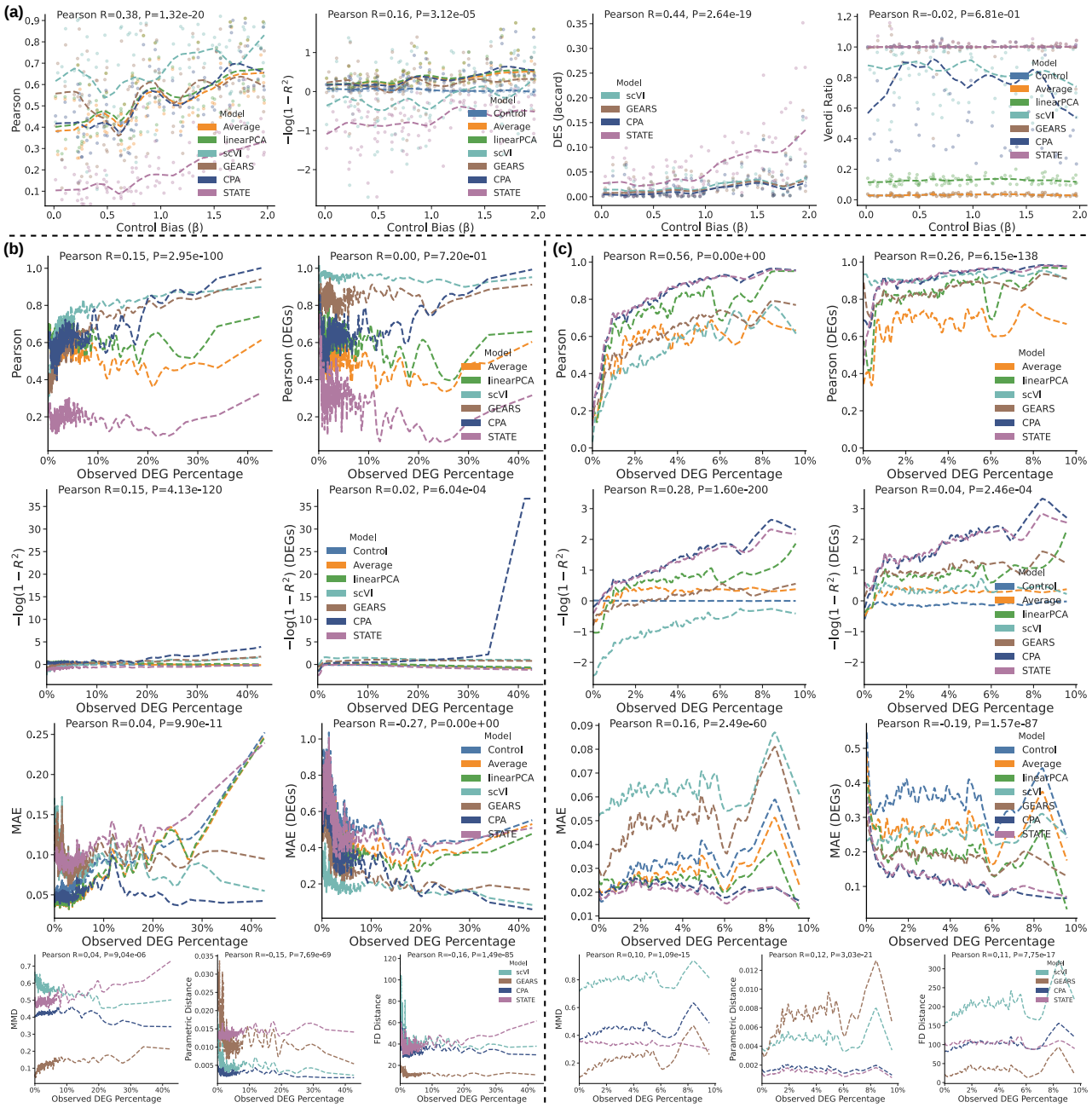

Figure 12. Metric distortion on DIRECTDGP and NORMAN19. (a) On DIRECTDGP: Pearson $\boldsymbol{\delta}$, $R^2_{\boldsymbol{\delta}}$, DES (Jaccard), and Vendi Ratio (VR) versus control bias (β) across 100 random parameter draws. (b) On DIRECTDGP: Pearson $\boldsymbol{\delta}$, $R^2_{\boldsymbol{\delta}}$, and MAE on all genes and on DEGs, together with MMD, parametric distance, and FD, versus observed DEG percentage across all perturbations and 100 random parameter draws. (c) Corresponding results on NORMAN19, aggregated across all perturbations over 10 repeated experiments. Points show individual runs, coloured by model; dashed curves show per-model LOESS-smoothed trends. Top annotations report the Pearson correlation (R) and two-sided p-value (P) computed from all raw points in each panel.