MECHANISMS MATTER: TRANSPORTABILITY OF CELLULAR PERTURBATION EFFECTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting cellular responses to perturbations across biological contexts is central to drug development, yet deep learning models often fail to outperform 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 score to the perturbation setting to quantify effective response diversity and diagnose mode collapse, a failure mode invisible to standard per-perturbation metrics. Extensive experiments across six deep learning models and five 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 generalizes across contexts with different causal mechanisms. Across real and semi-synthetic datasets, lower cross-context concordance of transcriptome-wide impact tracks larger transfer gaps, providing an empirical proxy for perturbation-relevant mechanistic divergence.

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; Nadig 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 undersampled, 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 per-perturbation metrics can be misleading. Whole-transcriptome Pearson correlation, for example, rewards alignment with the population-mean effect (Viñas Torné et al., 2025; Csendes et al., 2025). More generally, scoring perturbations independently can miss mode collapse, in which models predict nearly identical responses across perturbations (Wu et al., 2025; Mejia et al., 2025). The perturbation discrimination score (PDS) (Wu et al., 2025) detects this failure mode, but its ℓ_1 and ℓ_2 variants are scale-sensitive (Liu et al., 2025). Second, models may lack the inductive biases needed to capture the causal mechanisms governing perturbation effects (Lorch et al., 2026; Dibaeinia et al., 2026; Wenteler et al., 2025). Distinguishing

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

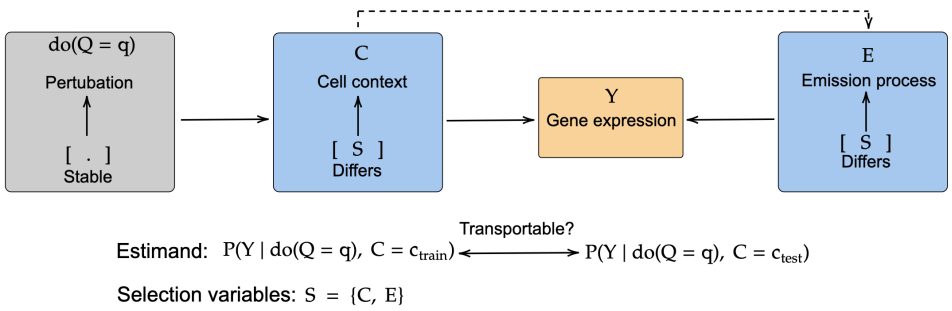


Figure 1: Overview of the transportability problem in cellular perturbation prediction. Given training pairs $\{(q, c_{\text{train}}), (q', c_{\text{test}}) : q' \neq q\}$, the goal is to predict $P(Y \mid \text{do}(Q = q), C = c_{\text{test}})$ for the *unseen (perturbation, context) pair* (q, c_{test}), where Y denotes the cellular response and $\text{do}(Q = q)$ assigns perturbation q (Pearl, 2009). The selection variables $S = \{C, E\}$ denote biological context and emission process, respectively. Observed components of E (*e.g.*, library size and batch) can be adjusted for, whereas latent components (*e.g.*, dropout rate and technical noise) induce domain shift.

among metric distortion, mode collapse, and model mis-specification is a prerequisite for meaningful progress.

Even when evaluation is corrected and models are made more expressive, formal guarantees of cross-context generalization remain absent (Dibaeinia et al., 2026). Scaling data and architecture (Theodoris et al., 2023; Cui et al., 2024; Hao et al., 2024), disentangling perturbation effects (Lotfollahi et al., 2023; Piran et al., 2024; Zhang et al., 2024), matching distributions (Lotfollahi et al., 2019; Bunne et al., 2023; Dong et al., 2023; Adduri et al., 2026; Lorch et al., 2026; Klein et al., 2025; Palla et al., 2026), and incorporating biological priors (Roohani et al., 2024; He et al., 2025) have each shown promise within individual contexts. Yet the cross-context generalization gap itself has not been systematically quantified. Causal transportability (Bareinboim and Pearl, 2013) offers a principled framework for determining whether and how perturbation effects generalize across contexts, including when mechanism shifts are identifiable and adjustable through a valid transport formula. Because causal cascades mediating perturbation effects are rarely fully characterized in biology (Amit et al., 2009), we treat mechanism invariance as a sufficient but not necessary guiding principle, emphasizing 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 and Pearl, 2013), showing that cross-context generalization is governed by shared causal mechanisms, not distributional similarity.
- We extend the framework of Lorch et al. (2026) into a cross-context causal simulator with realistic semi-synthetic data, known causal structure, and tunable mechanistic divergence, enabling controlled evaluation of generalization.
- We adapt the Vendi score (Friedman and Dieng, 2023) as a diagnostic of effective response diversity. Together, the Vendi score and PDS (Wu et al., 2025) characterize perturbation-response geometry in terms of effective diversity and discrimination.
- Across six deep learning models and five simple baselines on semi-synthetic and real Perturb-seq datasets, we reveal a cross-context generalization gap that widens as transcriptome-wide impact (TI; (Nadig et al., 2025)) concordance decreases, supporting TI concordance as an empirical proxy for mechanistic divergence across contexts.

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 and 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 or model capacity.

2.1 PROBLEM FORMULATION

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

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

where $\mathrm{do}(Q{=}q)$ denotes an intervention that assigns perturbation q (Pearl, 2009). We estimate this distribution for an *unseen (perturbation, context) combination*, using interventional samples from the observed combinations $\{(q, c_{\mathrm{train}})\} \cup \{(q', c_{\mathrm{test}}) : q' \neq q\}$ and target-context control samples drawn from $P(Y \mid \mathrm{do}(Q{=}0), C{=}c_{\mathrm{test}})$. For pseudobulk prediction, the estimand reduces to the interventional mean $\mathbb{E}\big[Y \mid \mathrm{do}(Q{=}q),\ C = c_{\mathrm{test}}\big]$. 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 (positive: activation, negative: 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).

A context shift alters A_C, B_C, or both: A_C rewires the regulatory program and thus the perturbation response (cell-type and lineage differences), while B_C shifts the basal state (batch and activation-state variation), leaving the latent response mechanism intact but potentially altering measured effects under nonlinear emission. Transportability depends on *which* of these differ across contexts and whether the differences are identifiable from the available data. In the trivial case $c_{\mathrm{train}} = c_{\mathrm{test}}$, the problem reduces to interpolation over seen (perturbations, context) combinations, and transportability holds by construction. Throughout, let $\mathcal{D} = \{(\boldsymbol{y}_i, c_i, q_i)\}_{i=1}^N$ denote a dataset of N observations.

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 and 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):

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

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

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

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,

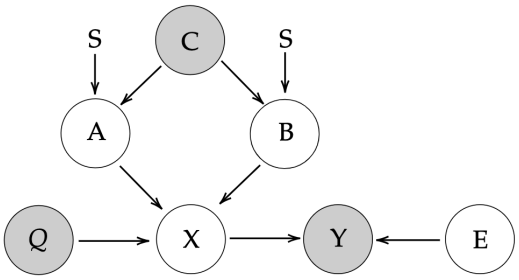


Figure 2: Selection diagram of the assumed CAUSALDGP, with unobserved mechanistic, context-specific selection variables $S = \{A, B\}$. Cell context C, perturbation Q, and gene expression counts Y are observed. The unobserved emission process E, mapping the latent state X to observed counts Y, is invariant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

modeled as a negative binomial (NB) likelihood, capturing both *technical measurement noise* and *intrinsic transcriptional variability*:

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

where μ_g is the mean parameter and θ_g is a gene-specific inverse-dispersion. The softplus link $\mathrm{softplus}(z) = \log(1 + e^z)$ 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, which 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

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

Proposition 1 (Direct transportability of perturbation effects) *From (5) the effect of perturbation q is directly transportable from C to C' without context-specific adjustment if and only if*

$$\boldsymbol{\tau}_C^{(q)} = \boldsymbol{\tau}_{C'}^{(q)} \quad \Longleftrightarrow \quad \big(A_C^{-1} - A_{C'}^{-1}\big)\,\Gamma_q = 0. \tag{6}$$

Crucially, this condition is *perturbation-specific*: for perturbation q, it involves only the columns of A_C^{-1} and $A_{C'}^{-1}$ indexed by the support of Γ_q. Thus, a pair of contexts may permit direct latent transport for one perturbation but not another. Marginal divergence between $P_C(Y)$ and $P_{C'}(Y)$ neither establishes nor precludes direct latent transport: even when $\boldsymbol{\tau}_C^{(q)} = \boldsymbol{\tau}_{C'}^{(q)}$, nonlinear emission can yield different measured perturbation effects across baselines B_C and $B_{C'}$, while differences in σ and irrelevant pathways in A may further shift observed distributions. Thus, perturbation-relevant mechanism invariance, unlike overall distributional similarity, is sufficient for direct latent transport.

More generally, effect equality is unnecessary if a context-specific transport map is identifiable. When Γ_q is shared across contexts, then $\boldsymbol{\tau}_{C'}^{(q)} = A_{C'}^{-1} A_C \boldsymbol{\tau}_C^{(q)}$. Thus, unequal effects are transportable when $A_{C'}^{-1} A_C$ is identifiable; otherwise, the held-out target effect cannot generally be recovered. Typical Perturb-seq screens ($P \ll G$) can identify A_C^{-1} only on $\mathrm{span}\{\Gamma_q\}$; full recovery requires G independent perturbations. Even then, noise, variable perturbation efficiency, and nonlinear dynamics make single-cell GRN inference challenging and approximate (Pratapa et al., 2020).

3 METRICS

Algorithm 1 Null-Calibrated Pseudobulk Vendi Score

Require: Log-normalized empirical control-cell vectors $\{\tilde{\boldsymbol{y}}_{0i}^{\mathrm{emp}}\}_{i=1}^{N_0}$; response pseudobulks $\{\bar{\tilde{\boldsymbol{y}}}_q^r\}_{q=1}^P$, $r \in \{\mathrm{emp}, \mathrm{pred}\}$; PCA dimension d; null-calibration parameters $(n_{\mathrm{null}}, S, \rho, \kappa)$
Ensure: Pseudobulk Vendi score VS $\in [1,\ P]$
1: Fit $\mathcal{P} \leftarrow \mathrm{PCA}_d(\{\bar{\tilde{\boldsymbol{y}}}_q^{\mathrm{emp}}\}_{q=1}^P)$ and embed response $\boldsymbol{z}_q^r \leftarrow \mathcal{P}(\bar{\tilde{\boldsymbol{y}}}_q^r)$ for $q = 1, \ldots, P$
2: **for** $s = 1$ **to** S **do**
3: Sample disjoint control-cell index sets $\mathcal{I}_s, \mathcal{J}_s \subseteq \{1, \ldots, N_0\}$ such that $|\mathcal{I}_s| = |\mathcal{J}_s| = n_{\mathrm{null}}$.
4: Empirical-control pseudobulks: $\bar{\tilde{\boldsymbol{y}}}_{\mathcal{U}}^{\mathrm{emp}} \leftarrow \frac{1}{n_{\mathrm{null}}} \sum_{i \in \mathcal{U}} \tilde{\boldsymbol{y}}_{0i}^{\mathrm{emp}}$, where $\mathcal{U} \in \{\mathcal{I}_s, \mathcal{J}_s\}$.
5: Null squared PCA distance: $\delta_s \leftarrow \left\| \mathcal{P}(\bar{\tilde{\boldsymbol{y}}}_{\mathcal{I}_s}^{\mathrm{emp}}) - \mathcal{P}(\bar{\tilde{\boldsymbol{y}}}_{\mathcal{J}_s}^{\mathrm{emp}}) \right\|_2^2$
6: **end for**
7: Null-distance ρ-quantile: $\delta_\rho \leftarrow \mathrm{Quantile}_\rho(\delta_1, \ldots, \delta_S)$
8: Calibrate the RBF variance to map δ_ρ to similarity κ: $\sigma_{\mathrm{null}}^2 \leftarrow -\frac{\delta_\rho}{2 \log \kappa}$
9: Response similarity matrix: $K_{qq'} \leftarrow \exp\!\left(-\|\boldsymbol{z}_q^r - \boldsymbol{z}_{q'}^r\|_2^2 / (2\sigma_{\mathrm{null}}^2)\right)$ for $q, q' = 1, \ldots, P$
10: Let $\lambda_1, \ldots, \lambda_P$ be the eigenvalues of K and $p_i \leftarrow \lambda_i / \sum_j \lambda_j$ for $i = 1, \ldots, P$
11: $\mathrm{VS} \leftarrow \exp\left(-\sum_{i=1}^{P} p_i \log p_i \right)$, with $0 \log 0 \coloneqq 0$
12: **return** VS

In practice, most models do not infer the causal matrix A_C in (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. Our focus is on quantifying the *cross-context generalization gap*: the degradation in perturbation-effect prediction when models transfer from observed to unseen biological contexts. We therefore use perturbation-effect metrics that capture complementary aspects of this gap. 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* pseudobulk, 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 quantifying perturbation-response diversity. 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.4); *perturbation effect metrics* including Pearson $\boldsymbol{\delta}$ (14), the perturbation effect discrimination score (PDS) (13), coefficient of determination $R^2_{\boldsymbol{\delta}}$ (15), and transcriptome-wide impact (TI) (19)(Appendix D.2); and *differentially expressed gene (DEG) set metrics* including Jaccard index, recall, and precision (Appendix D.3). Because per-perturbation metrics are sensitive to control bias and perturbation-gene sparsity (Appendix E.6), we report DEG-restricted variants throughout.

3.1 EFFECTIVE PERTURBATION RESPONSE DIVERSITY: VENDI SCORE

Per-perturbation metrics such as Pearson $\boldsymbol{\delta}$ evaluate predictions independently, potentially masking mode collapse (Wu et al., 2025) or reflecting signal strength rather than model quality for sparse effects (Appendix E.6). We therefore quantify effective response diversity using a *null-calibrated* pseudobulk Vendi score adapted from Friedman and Dieng (2023), which converts pairwise distances in an empirical-reference PCA space into RBF similarities (Algorithm 1).

We calibrate the RBF scale using pairs of disjoint pseudobulks sampled from empirical control cells, each containing n_{null} cells. We choose the scale such that the ρ-quantile of their squared PCA distances corresponds to a target similarity κ (Table 7, Appendix). This calibration captures finite-sample variability at the reference size n_{null} but does not explicitly adjust for perturbation-specific sample sizes N_q. We limit extreme mismatch by requiring $N_q \geq 64$. The Vendi score is

the exponential spectral entropy of the resulting perturbation-similarity matrix and represents the effective number of distinct perturbation responses. We compute empirical and predicted scores using the same empirical PCA transformation and null calibration and report the *Vendi Ratio*:

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

Here, VS denotes the Vendi score, and Y_q and $\hat{Y}_q$ denote the empirical and predicted perturbation responses, respectively. The raw predicted Vendi score is unnormalized, whereas the Vendi Ratio normalizes predicted diversity by empirical diversity. A ratio near one indicates preserved empirical diversity, although finite-sample noise may cause an optimal deterministic predictor to yield $\mathrm{VR} < 1$. Complete collapse of the predicted responses gives $\mathrm{VS}(\hat{Y}) = 1$ and hence $\mathrm{VR} = 1/\mathrm{VS}(Y)$, which equals $1/P$ only when the empirical responses attain maximal diversity.

To characterize complementary aspects of *perturbation-response geometry*, we report the Vendi score alongside PDS. PDS measures response distinguishability, whereas the Vendi score quantifies effective diversity by accounting for redundancy. Together, they assess whether responses are both distinct and nonredundant.

4 Experiments

4.1 Baselines

We evaluate six deep learning models spanning transport-based (STATE (Adduri et al., 2026), scLDM (Palla et al., 2026)), graph-based (GEARS (Roohani et al., 2024)), disentanglement-based (CPA (Lotfollahi et al., 2023), scVI (Lopez et al., 2018)), and foundation-model approaches (STATE-Geneformer, combining Geneformer (Theodoris et al., 2023) embeddings with a STATE decoder). We also evaluate five simple baselines: a global control mean, plus both global and context-specific variants of perturbation mean (Csendes et al., 2025; Kernfeld et al., 2023) and linearPCA (Ahlmann-Eltze et al., 2025). We adapt scVI to perturbation prediction using our transport formula. To predict *unseen (perturbation, context) combinations*, we use STATE's raw-count state-transition variant rather than its pretrained embedding model. Geneformer's pretraining predates ZHU25 and excludes immortalized cell lines, making our benchmark profiles unseen despite including primary immune cells. GEARS uses gene ontology and training-derived co-expression graphs; for synthetic data, which lack an external ontology, we use only the latter. See Tables 2 and 3 in Appendix A for a summary of the considered baselines and their detailed transport formulae, respectively.

4.2 Datasets

We focus on Perturb-seq because its targeted gene interventions provide an experimental analog of Pearl's $\mathrm{do}(\cdot)$ operator (Pearl, 2009) for studying causal gene function at single-cell resolution. Extending the framework to cytokines and chemical perturbations (Klein et al., 2025) remains an important direction for future work. See Table 1 for summary statistics and Appendix C.1 for complete details.

Table 1: Summary of the cellular perturbation datasets after preprocessing. HVGs: Highly variable genes.

Statistic	ZHU25	REPLOGLE22/NADIG25	NORMAN19	CAUSALDGP	DIRECTDGP
Cell type	Primary CD4+ T cells	RPE1–K562 / HepG2–K562 / Jurkat–K562	K562	-	-
Perturbation type	CRISPRi knockdown	CRISPRi knockdown	CRISPRa over-expression	synthetic	synthetic
Control cells N_0	393,216	22,176 / 15,667 / 22,704	11,855	1,024	128–1,024
Perturbed cells N_P	203,870	266,087 / 144,160 / 300,954	71,742	1,024×128	(128–512)×(20–50)
HVGs G	3,463	7,226 / 7,595 / 7,563	8,192	128	1,024–8,192
Perturbations P	91	827 / 479 / 983	156	128	20–50
Total contexts C	12 (4 donors × 3 conditions)	2 / 2 / 2	1	2	1

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

behavior, including metric distortion and Vendi score validation, in controlled within-context settings. CAUSALDGP (Section 2.2) 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.

Validating CAUSALDGP Figure 6 (Appendix) illustrates cell-level heterogeneity and context-dependent shifts in expression space. Beyond this qualitative visualization, we evaluate marginal fidelity, as emphasized in prior simulator studies (Zappia et al., 2017; Sun et al., 2021), and downstream task behavior. We compare CAUSALDGP with real Perturb-seq data within each cellular context using marginal, co-expression, perturbation-effect, effective-diversity, and discriminability metrics (Table 5, Appendix). Given the heterogeneity of real Perturb-seq data, CAUSALDGP falls within observed ranges for dispersion (CoV), sparsity (Zero), co-expression (Frac$|r|>0.3$), and effect size ($\sqrt{\text{TI}}$). Across scenarios, Vendi reaches 20.1–27.2% of its theoretical maximum, $P = 128$, indicating substantial effective diversity. Given the simulator design, high split-half PDS is expected by construction and serves as a sanity check that sampling noise does not obscure the simulated perturbation effects.

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 (one context), used to study metric distortion and Vendi-score sensitivity.
- REPLOGLE22/NADIG25 (Replogle et al., 2022; Nadig et al., 2025): single-gene perturbations across K562, RPE1, Jurkat, and HepG2 cell lines (four contexts). K562 and RPE1 are from Replogle et al. (2022), while Jurkat and HepG2 are from Nadig et al. (2025). These lines contrast hematopoietic (K562, Jurkat) with epithelial (RPE1, HepG2) and cancer with non-transformed (RPE1), providing a test bed across genuine biological divergence. We treat the partner line as the source context and K562 as the target (evaluation), defining one transportability task per pair.
- ZHU25 (Zhu et al., 2025): large-scale primary human CD4+ T-cell Perturb-seq across 12 contexts (4 donors $\times$ 3 activation states: resting, 8hrs post-stimulation, and 48hrs post-stimulation). Unlike the immortalized lines in REPLOGLE22/NADIG25, this dataset captures biologically relevant heterogeneity from both activation state and inter-individual variation (Brodin et al., 2015; Chapfuwa et al., 2025).

Context-aware splitting strategy 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 7, Appendix). Both splits are derived from a shared random seed and planning step to ensure identical randomization decisions, see Appendix C.2 for details.

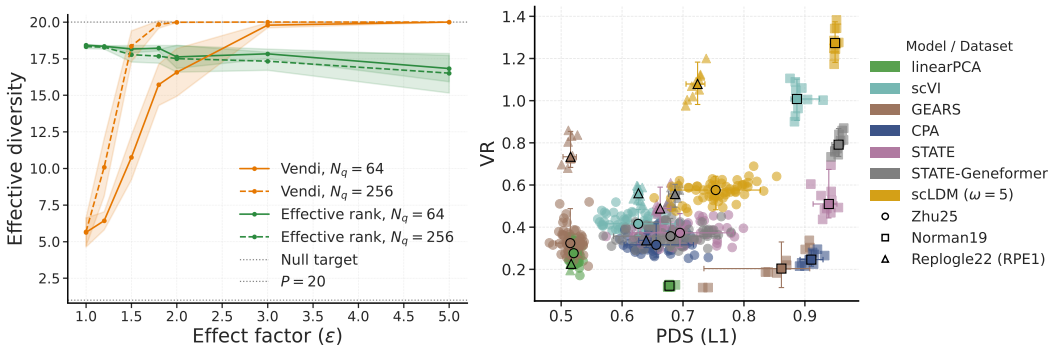
4.3 RESULTS

4.3.1 VALIDATION AND ROBUSTNESS OF THE VENDI SCORE

We compare Vendi score with covariance effective rank (Roy and Vetterli, 2007), the exponential entropy of the normalized covariance eigenvalues, to assess the value of Vendi's null-calibrated distance kernel. Across 100 synthetic DIRECTDGP datasets, both measures increase with the number of perturbations P (Figure 8, Appendix E.2). However, Vendi tracks increasing effect strength from the population null, whereas effective rank remains high under weak effects (Figure 3a). Moreover, sample-size and deterministic-prediction biases are largest at low effect factors and diminish with increasing effect strength and larger N_q (Figure 10, Appendix). In Gaussian-noise experiments on NORMAN19 (Figure 9, Appendix E.2), Vendi and PDS (ℓ_1) remain stable under low-to-moderate noise and decline under severe noise, whereas effective rank increases with noise.

4.3.2 PERTURBATION RESPONSE GEOMETRY: DIVERSITY AND DISCRIMINATION

We characterize perturbation-response geometry through relative diversity (VR; (8)) and discrimination (PDS; (13); Figure 3b). On NORMAN19, most models achieve high PDS but low VR,



(a) Vendi score and covariance effective rank. (b) Vendi ratio (VR) vs. PDS under within-context evaluation.

Figure 3: (a) Effective diversity in synthetic DIRECTDGP datasets ($P = 20$; $p_{\text{effect}} = 0.1$; $\beta = 0$). Lines show means over five replicates (shading: ± 1 SD); solid/dashed lines indicate $N_q = 64/256$. (b) VR vs. PDS (ℓ_1)across models and datasets. Color and shape indicate model and dataset; enlarged markers and error bars show medians and 95% percentile intervals per model–dataset pair.

suggesting partial diversity collapse; scVI is well calibrated, whereas scLDM overestimates diversity. Across datasets, scLDM performs among the strongest models, leading on REPLOGLE22/NADIG25, while STATE-Geneformer attains the highest PDS on NORMAN19. All models underestimate diversity on ZHU25, where GEARS also exhibits near-random PDS. On REPLOGLE22/NADIG25 (RPE1–K562), the largest perturbation set, PDS is lower than on NORMAN19, although scLDM retains strong discrimination and near-unity VR. Notably, under cross-context evaluation, PDS broadly declines, approaching random performance on ZHU25 (Figure 11; Appendix). Thus, unseen perturbation–context combinations consistently impair discrimination but affect diversity in a model- and dataset-dependent manner.

4.3.3 CROSS-CONTEXT GENERALIZATION GAP

Figure 5 summarizes the cross-context generalization gap on CAUSALDGP and ZHU25 (full results in Figures 12–15, Appendix). On CAUSALDGP, scLDM and CPA outperform the baselines within context. Cross-context performance remains high when only the baseline drift B_C changes, consistent with Proposition 1, under which latent-effect transportability is independent of B_C. Performance falls to baseline levels when perturbation-relevant mechanisms in A_C change, alone or together with B_C, consistent with a loss of direct latent transportability.

On ZHU25, performance similarly declines across the held-out donors and activation states. While CPA, STATE, scLDM, and STATE-Geneformer outperform simple baselines within-context evaluation, their advantage largely vanishes in 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/NADIG25 (Figure 4; full results in Figure 13–14, Appendix E.4), where models are trained on RPE1 and evaluated on held-out K562 perturbations, the cross-context gap is notably smaller: CPA, STATE, and STATE-Geneformer remain competitive with simple baselines even under cross-context evaluation. De-

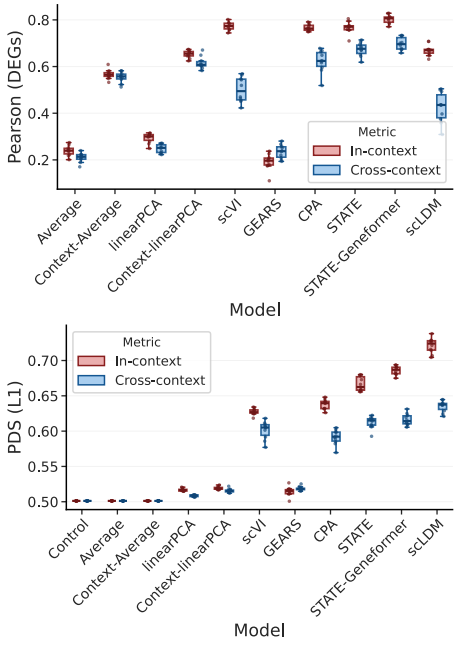


Figure 4: Cross-context generalization gap on REPLOGLE22/NADIG25 for Pearson $\boldsymbol{\delta}$ (DEG-restricted) and PDS (ℓ_1). Models are trained on RPE1 and evaluated on held-out K562 perturbations. scLDM uses $\omega = 5$.

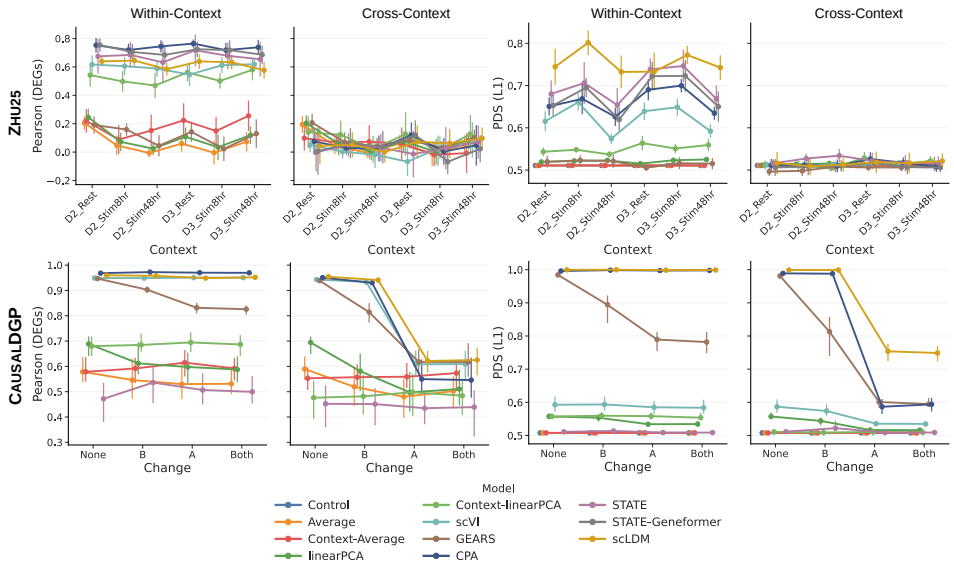


Figure 5: Cross-context gaps on ZHU25 (top) and CAUSALDGP (bottom) for DEG-restricted Pearson $\boldsymbol{\delta}$ and PDS (ℓ_1). The x-axis shows six ZHU25 test contexts (2 donors $\times$ 3 activation states) and four CAUSALDGP mechanism modes (`None`, `A`, `B`, `Both`). We plot medians and interquartile ranges. scLDM is evaluated with classifier-free guidance weight $\omega = 5$.

spite weak overall correlation between these cell lines (Nadig et al., 2025), conserved effects for some perturbations may support transport of the transferable signal.

Alignment between theory and experiments We test the prediction in (6) that direct-transport performance declines as *perturbation-relevant causal mechanisms* diverge across contexts, using both real Perturb-seq datasets and CAUSALDGP. As a proxy for mechanistic similarity, we use the Pearson correlation between source- and target-context noise-adjusted TI values across matched perturbations; TI is defined in (19) (Nadig et al., 2025). Across three REPLOGLE22/NADIG25 source-to-K562 tasks, we evaluate the four best-performing models: CPA, STATE, STATE-Geneformer, and scLDM, each trained separately on RPE1, Jurkat, and HepG2. Jurkat and HepG2 exhibit similarly high TI concordance with K562, whereas RPE1 exhibits lower concordance, which is biologically plausible given RPE1's non-cancer origin (Table 9, Appendix). Consistent with this difference, RPE1 $\rightarrow$ K562 yields the largest transfer gaps across all four models for both DEG-restricted Pearson $\boldsymbol{\delta}$ and PDS (ℓ_1) (Figure 14, Appendix). Moreover, varying A in CAUSALDGP reproduces the observed TI–transfer-gap relationship across the REPLOGLE22/NADIG25 TI range. Cross-donor TI correlations in ZHU25 are substantially lower and coincide with broad generalization failures (Figure 5). Although TI correlation is only a proxy and does not identify the underlying causal mechanisms, consistency across models and across real and synthetic Perturb-seq datasets supports the predicted association between greater mechanistic divergence and reduced direct-transport performance.

5 CONCLUSION

Cross-context generalization in perturbation prediction depends on shared causal mechanisms, not merely distributional similarity. Across six deep learning models and five simple baselines, performance drops substantially under cross-context evaluation, often matching simple baselines. A joint Vendi–PDS analysis characterizes perturbation-response geometry along effective-diversity and discrimination axes, revealing complementary failure modes: simple baselines produce low-diversity, poorly differentiated responses, whereas deep models generally achieve stronger in-context discrimination but often underestimate diversity and consistently lose discrimination under cross-context transfer. Even on synthetic data with known causal structure, no model generalizes across contexts with divergent mechanisms, exposing limitations of current architectures. Consistent with theory, transfer gaps widen as cross-context transcriptome-wide impact (TI) concordance declines, supporting low TI concordance as an empirical proxy for mechanistic divergence.

AI USE STATEMENT

In this work, we did not use large language models (LLMs) for tasks requiring disclosure under the ICLR 2027 AI Policy; these tasks were performed entirely by the authors or were not applicable. We used LLMs to create and edit software code, including plotting code, and to improve the readability of author-written manuscript text and checklist responses. All LLM-assisted code and text were reviewed by the authors; code was tested and its outputs inspected, while language edits were checked to preserve the intended scientific meaning. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of LLMs.

REPRODUCIBILITY STATEMENT

The paper provides a dataset summary (Table 1); preprocessing and filtering details (Appendix C.1 and Table 6); descriptions of the within-context and cross-context split construction (Section 4.2 and Appendix C.2); model definitions and transport formulae (Tables 2 and 3, Appendix A); metric definitions (Section 3 and Appendix D); and complete specifications for the semi-synthetic CAUSALDGP and DIRECTDGP simulators (Section 4.2 and Appendix B). Results supporting the main claims appear in Section 4.3, with complete results in Appendix E. The supplementary material contains the code and instructions needed to reproduce the experiments, including instructions in the code repository for accessing the data. All code is implemented in Python and will be made publicly available on GitHub upon publication.

REFERENCES

Abhinav K Adduri, Dhruv Gautam, Beatrice Bevilacqua, Mohsen Naghipourfar, Alishba Imran, Rohan Shah, Noam Teyssier, Rishi Verma, Christopher Carpenter, Basak Eraslan, et al. Predicting cellular responses to perturbation across diverse contexts with state. *Cell*, 2026.

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

Ido Amit, Manuel Garber, Nicolas Chevrier, Ana Paula Leite, Yoni Donner, Thomas Eisenhaure, Mitchell Guttman, Jennifer K Grenier, Weibo Li, Or Zuk, et al. Unbiased reconstruction of a mammalian transcriptional network mediating pathogen responses. *Science*, 2009.

Lisa Amrhein, Kumar Harsha, and Christiane Fuchs. A mechanistic model for the negative binomial distribution of single-cell mrna counts. *bioRxiv*, 2019.

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

Ihab Bendidi, Shawn Whitfield, Kian Kenyon-Dean, Hanene Ben Yedder, Yassir El Mesbahi, Emmanuel Noutahi, and Alisandra K Denton. Benchmarking transcriptomics foundation models for perturbation analysis: one pca still rules them all. *arXiv*, 2024.

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

Christoph Bock, Paul Datlinger, Florence Chardon, Matthew A Coelho, Matthew B Dong, Keith A Lawson, Tian Lu, Laetitia Maroc, Thomas M Norman, Bicna Song, et al. High-content crispr screening. *Nature Reviews Methods Primers*, 2022.

Petter Brodin, Vladimir Jojic, Tianxiang Gao, Sanchita Bhattacharya, Cesar J Lopez Angel, David Furman, Shai Shen-Orr, Cornelia L Dekker, Gary E Swan, Atul J Butte, et al. Variation in the human immune system is largely driven by non-heritable influences. *Cell*, 2015.

Charlotte Bunne, Stefan G Stark, Gabriele Gut, Jacobo Sarabia Del Castillo, Mitch Levesque, Kjong-Van Lehmann, Lucas Pelkmans, Andreas Krause, and Gunnar Rätsch. Learning single-cell perturbation responses using neural optimal transport. *Nature methods*, 2023.

Paidamoyo Chapfuwa, Ilker Demirel, Lorenzo Pisani, Javier Zazo, Elon Portugaly, H Jabran Zahid, and Julia Greissl. Scalable universal t-cell receptor embeddings from adaptive immune repertoires. In *ICLR*, 2025.

Gerold Csendes, Gema Sanz, Kristóf Z Szalay, and Bence Szalai. Benchmarking foundation cell models for post-perturbation rna-seq prediction. *BMC genomics*, 2025.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature methods*, 2024.

Payam Dibaeinia, Sudarshan Babu, Mei Knudson, Ali ElSheikh, Yibo Wen, Han Liu, Jason Perera, and Aly A Khan. Position: Virtual cells need context, not just scale. In *ICML Position Paper Track*, 2026.

Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P Fulco, Livnat Jerby-Arnon, Nemanja D Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, et al. Perturb-seq: dissecting molecular circuits with scalable single-cell rna profiling of pooled genetic screens. *Cell*, 2016.

Mingze Dong, Bao Wang, Jessica Wei, Antonio H de O. Fonseca, Curtis J Perry, Alexander Frey, Feriel Ouerghi, Ellen F Foxman, Jeffrey J Ishizuka, Rahul M Dhodapkar, et al. Causal identification of single-cell experimental perturbation effects with cinema-ot. *Nature methods*, 2023.

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

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *JMLR*, 2012.

Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, Xuegong Zhang, and Le Song. Large-scale foundation model on single-cell transcriptomics. *Nature methods*, 2024.

Chujun He, Jiaqi Zhang, Munther Dahleh, and Caroline Uhler. Morph predicts the single-cell outcome of genetic perturbations across conditions and data modalities. *bioRxiv*, 2025.

Leon Hetzel, Simon Boehm, Niki Kilbertus, Stephan Günnemann, Fabian Theis, et al. Predicting cellular responses to novel drug perturbations at a single-cell resolution. *NeurIPS*, 2022.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 2017.

Sui Huang, Gabriel Eichler, Yaneer Bar-Yam, and Donald E Ingber. Cell fates as high-dimensional attractor states of a complex gene regulatory network. *Physical review letters*, 2005.

Eric Kernfeld, Yunxiao Yang, Joshua S Weinstock, Alexis Battle, and Patrick Cahan. A systematic comparison of computational methods for expression forecasting. *bioRxiv*, 2023.

Dominik Klein, Jonas Simon Fleck, Daniil Bobrovskiy, Lea Zimmermann, Sören Becker, Alessandro Palma, Leander Dony, Alejandro Tejada-Lapuerta, Guillaume Huguet, Hsiu-Chuan Lin, et al. Cellflow enables generative single-cell phenotype modeling with flow matching. *bioRxiv*, 2025.

Sandy L Klemm, Zohar Shipony, and William J Greenleaf. Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics*, 2019.

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

Lawrence I-Kuei Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 1989.

Qiyuan Liu, Qirui Zhang, Jinhong Du, Siming Zhao, and Jingshu Wang. Effects of distance metrics and scaling on the perturbation discrimination score. *arXiv*, 2025.

Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 2018.

Lars Lorch, Jiaqi Zhang, Charlotte Bunne, Andreas Krause, Bernhard Schölkopf, and Caroline Uhler. Latent causal diffusions for single-cell perturbation modeling. *arXiv*, 2026.

Mohammad Lotfollahi, F Alexander Wolf, and Fabian J Theis. scgen predicts single-cell perturbation responses. *Nature methods*, 2019.

Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Leon Hetzel, Yuge Ji, Ignacio L Ibarra, Sanjay R Srivatsan, Mohsen Naghipourfar, Riza M Daza, Beth Martin, et al. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology*, 2023.

Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 2014.

Gabriel M Mejia, Henry E Miller, Francis JA Leblanc, Bo Wang, Brendan Swain, and Lucas Paulo de Lima Camillo. Diversity by design: Addressing mode collapse improves scrna-seq perturbation modeling on well-calibrated metrics. *arXiv*, 2025.

Ajay Nadig, Joseph M Replogle, Angela N Pogson, Mukundh Murthy, Steven A McCarroll, Jonathan S Weissman, Elise B Robinson, and Luke J O'Connor. Transcriptome-wide analysis of differential expression in perturbation atlases. *Nature Genetics*, 2025.

Thomas M Norman, Max A Horlbeck, Joseph M Replogle, Alex Y Ge, Albert Xu, Marco Jost, Luke A Gilbert, and Jonathan S Weissman. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science*, 2019.

Giovanni Palla, Sudarshan Babu, Payam Dibaeinia, James D Pearce, Donghui Li, Aly A Khan, Theofanis Karaletsos, and Jakub M. Tomczak. Scalable single-cell gene expression generation with latent diffusion models. In *ICML*, 2026.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009.

Zoe Piran, Niv Cohen, Yedid Hoshen, and Mor Nitzan. Disentanglement of single-cell data with biolord. *Nature Biotechnology*, 2024.

Aditya Pratapa, Amogh P Jalihal, Jeffrey N Law, Aditya Bharadwaj, and TM Murali. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature methods*, 2020.

Joseph M Replogle, Reuben A Saunders, Angela N Pogson, Jeffrey A Hussmann, Alexander Lenail, Alina Guna, Lauren Mascibroda, Eric J Wagner, Karen Adelman, Gila Lithwick-Yanai, et al. Mapping information-rich genotype-phenotype landscapes with genome-scale perturb-seq. *Cell*, 2022.

Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with gears. *Nature Biotechnology*, 2024.

Yusuf H Roohani, Tony J Hua, Po-Yuan Tung, Lexi R Bounds, Feiqiao B Yu, Alexander Dobin, Noam Teyssier, Abhinav Adduri, Alden Woodrow, Brian S Plosky, et al. Virtual cell challenge: Toward a turing test for the virtual cell. *Cell*, 2025.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *2007 15th European signal processing conference*. IEEE, 2007.

Matthew Stephens. False discovery rates: a new deal. *Biostatistics*, 2017.

Tianyi Sun, Dongyuan Song, Wei Vivian Li, and Jingyi Jessica Li. scdesign2: a transparent simulator that generates high-fidelity single-cell gene expression count data with gene correlations captured. *Genome biology*, 2021.

Christina V Theodoris, Ling Xiao, Anant Chopra, Mark D Chaffin, Zeina R Al Sayed, Matthew C Hill, Helene Mantineo, Elizabeth M Brydon, Zexian Zeng, X Shirley Liu, et al. Transfer learning enables predictions in network biology. *Nature*, 2023.

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

Cédric Villani. *Optimal transport: old and new*. Springer, 2009.

Ramon Viñas Torné, Maciej Wiatrak, Zoe Piran, Shuyang Fan, Liangze Jiang, Sarah A Teichmann, Mor Nitzan, and Maria Brbić. Systema: a framework for evaluating genetic perturbation response prediction beyond systematic variation. *Nature Biotechnology*, 2025.

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

Aaron Wenteler, Martina Occhetta, Nikhil Branson, Victor Curean, Magdalena Huebner, William Dee, William Connell, Siu Pui Chung, Alex Hawkins-Hooker, Yasha Ektefaie, César Miguel Valdez Córdova, and Amaya Gallagher-Syed. Perteval-scFM: Benchmarking single-cell foundation models for perturbation effect prediction. In *ICML*, 2025.

Yan Wu, Esther Wershof, Sebastian Schmon, Marcel Nassar, Błażej Osiński, Ridvan Eksi, Zichao Yan, Rory Stark, Kun Zhang, and Thore Graepel. Perturbench: Benchmarking machine learning models for cellular perturbation analysis. In *NeurIPS Datasets and Benchmarks Track*, 2025.

Luke Zappia, Belinda Phipson, and Alicia Oshlack. Splatter: simulation of single-cell rna sequencing data. *Genome biology*, 2017.

Ziqi Zhang, Xinye Zhao, Mehak Bindra, Peng Qiu, and Xiuwei Zhang. scdisinfact: disentangled learning for integration and prediction of multi-batch multi-condition single-cell rna-sequencing data. *Nature Communications*, 2024.

Ronghui Zhu, Emma Dann, Jun Yan, Justine Reyes Retana, Ryunosuke Goto, Reese C Guitche, Lillian Brixi, Mineto Ota, Austin Hartman, Theodore L Roth, et al. Genome-scale perturb-seq in primary human cd4+ t cells maps context-specific regulators of t cell programs and human immune traits. *Cell*, 2025.

A BASELINES

Table 2 summarizes the considered baselines. Control predicts the mean expression of all training control cells, whereas Average predicts the cell-weighted mean across all non-control training cells; both return the same vector for every perturbation and context. Table 3 presents detailed transport formulae for the remaining models. The CAUSALDGP simulator shares inductive biases with these methods but realizes them differently; we therefore group methods by their primary assumptions rather than their architectures. *Transport-based* methods learn a map from control to perturbed distributions: CAUSALDGP uses a linear SDE, scLDM learns a neural velocity field through conditional flow matching, and STATE uses optimal-transport-based matching between control and perturbed populations. STATE-Geneformer retains the STATE transport mechanism but represents basal cell states using pretrained Geneformer embeddings. *Disentanglement-based* methods (CPA/scVI) decompose latent variation into perturbation, context, and cell-state components. Finally, GEARS encodes explicit gene–gene structure using Gene Ontology and basal co-expression graphs.

Table 2: 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**	**Biological prior**	**Granularity**
Control	Control mean, identity (no perturbation signal)	✗	✗	Pseudobulk
Average	Perturbation average	✗	✗	Pseudobulk
Context-Average	Context-specific perturbation average	✓	✗	Pseudobulk
linearPCA	Additive in gene space	✗	✗	Pseudobulk
Context-linearPCA	Context-specific, additive in gene space	✓	✗	Pseudobulk
scVI	Additive in latent space	✓	✗	Single-cell
CPA	Additive in latent space	✓	✗	Single-cell
GEARS	Additive residual in gene space	✗	Gene Ontology and co-expression graphs	Single-cell
STATE	Nonlinear (attention)	✓	✗	Single-cell
STATE-Geneformer	Nonlinear (attention)	✓	Pretrained Geneformer embedding	Single-cell
scLDM	Conditional flow matching in VAE latent space	✓	✗	Single-cell

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
Context-Average	$\hat{\boldsymbol{y}}(q, c_{\mathrm{test}}) = \frac{1}{\lvert\mathcal{I}^{\mathrm{tr}}_{c_{\mathrm{test}}}\rvert} \sum_{i \in \mathcal{I}^{\mathrm{tr}}_{c_{\mathrm{test}}}} \boldsymbol{y}_i$, where $\mathcal{I}^{\mathrm{tr}}_{c_{\mathrm{test}}} = \{i \in \mathcal{D}_{\mathrm{train}} : C_i = c_{\mathrm{test}},\ Q_i \neq \mathrm{control}\}$. *Context-specific and perturbation-agnostic*: predicts the cell-weighted mean of seen non-control perturbations in the target context.
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 pseudobulks, $\bar{\boldsymbol{p}}_{\mathcal{T}(q)} = \frac{1}{\lvert\mathcal{T}(q)\rvert} \sum_{t\in\mathcal{T}(q)} U_t$ 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 pseudobulks. *Additive and context-free*: no conditioning on C. Operates on pseudobulk perturbation means, predicts single expression vector per perturbation. Assumes perturbation effects are fully determined by target gene identity.
Context-linearPCA	$\hat{\boldsymbol{y}}(q, c_{\mathrm{test}}) = U^{(c_{\mathrm{test}})} W^{(c_{\mathrm{test}})} \bar{\boldsymbol{p}}^{(c_{\mathrm{test}})}_{\mathcal{T}(q)} + \boldsymbol{b}^{(c_{\mathrm{test}})}$. Uses the same procedure as linearPCA, but refits U, W, and $\boldsymbol{b}$ using only seen non-control perturbation pseudobulks from the target-context training partition.
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*.
STATE-Geneformer	$\hat{\boldsymbol{y}}_i(q, c_{\mathrm{test}}) = f_{\mathrm{STATE}}\big(E_{\mathrm{GF}}(\boldsymbol{y}_{\mathrm{ctrl},i}),\ q,\ c_{\mathrm{test}},\ \{E_{\mathrm{GF}}(\boldsymbol{y}_{\mathrm{ctrl},j})\}_{j\neq i}\big)$, where E_{GF} denotes the pretrained Geneformer encoder. *Nonlinear in a pretrained latent representation*: replaces STATE's basal gene-expression input with a precomputed Geneformer embedding while retaining the STATE architecture and gene-space output.
scLDM	$\boldsymbol{z}_i(0) \sim \mathcal{N}(0, I)$, $\frac{\mathrm{d}\boldsymbol{z}_i(t)}{\mathrm{d}t} = v_\theta\big(\boldsymbol{z}_i(t), t \mid q, c_{\mathrm{test}}\big)$, $\hat{\boldsymbol{y}}_i(q, c_{\mathrm{test}}) \sim p_{\mathrm{Dec}}\big(\cdot \mid \boldsymbol{z}_i(1)\big)$, where v_θ is the DiT velocity field learned by conditional flow matching and p_{Dec} is the frozen VAE decoder. *Conditional flow matching in VAE latent space*: an ODE transports Gaussian noise to perturbation- and context-conditioned cell latents, which are decoded into gene-expression profiles.

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 this paper, we use DIRECTDGP primarily as a controlled within-context benchmark for studying metric behavior, including metric distortion and Vendi score validation.

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_q	Number of cells under perturbation q	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	-
σ_ℓ	standard deviation of log-library size	0.1714	-
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 $\mathrm{Unif}([-3,3])$
Γ_q	Perturbation shift vector	-	One-sparse intervention
s_q	Perturbation magnitude	-	$\mathrm{Unif}([-6\log G, -2\log G] \cup [2\log G, 6\log G])$
ρ_{swap}	Row-wise rewiring fraction	-	0.5
σ	Diffusion coefficient in latent OU process	-	$\sqrt{2}$
β_{sp}	Softplus parameter	-	3

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-cell counts $\{N_q\}_{q=1}^{P}$, empirical parameters $\{\mu_g^c, \theta_g, \lambda_g\}_{g=1}^{G}$ derived from NORMAN19, and simulation parameters $p_{\text{effect}}, \beta, \epsilon, \mu_\ell, \sigma_\ell$

for $j = 1, \ldots, N_0$ **do**
  Sample $\ell_j^{(0)} \sim \text{LogNormal}(\mu_\ell, \sigma_\ell^2)$
  for $g = 1, \ldots, G$ **do**
    Set $\mu_{g,j}^{(0)} \leftarrow \ell_j^{(0)} \mu_g^c$
    Draw $Y_{g,j}^{(0)} \sim \text{NB}(\mu_{g,j}^{(0)}, \theta_g)$
  end for
end for
for $q = 1, \ldots, P$ **do**
  for $g = 1, \ldots, G$ **do**
    Sample $\alpha_{q,g} \in \{1, \epsilon, 1/\epsilon\}$ with probabilities $(1 - p_{\text{effect}}, p_{\text{effect}}/2, p_{\text{effect}}/2)$
  end for
  for $j = 1, \ldots, N_q$ **do**
    Sample $\ell_j^{(q)} \sim \text{LogNormal}(\mu_\ell, \sigma_\ell^2)$
    for $g = 1, \ldots, G$ **do**
      Set $\mu_{g,j}^{(q)} \leftarrow \ell_j^{(q)} \alpha_{q,g} \max\big(\mu_g^c + \beta\lambda_g, 0\big)$
      Draw $Y_{g,j}^{(q)} \sim \text{NB}(\mu_{g,j}^{(q)}, \theta_g)$
    end for
  end for
end for
return Concatenated count matrix $\{Y^{(q)}\}_{q=0}^{P}$ and perturbation metadata Q

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 $\text{Unif}\big([-3, -1] \cup [1, 3]\big)$. We use this fixed-in-degree construction to control network density and local regulatory complexity, thereby isolating context-specific mechanism shifts from variation in in-degree. This choice may underrepresent heterogeneity in perturbation propagation arising from highly connected regulators in biological networks. 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 it by a multiple of 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:

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

For each context–perturbation pair, we simulate up to 64 parallel trajectories, independently initialized from $\mathcal{N}(0, I_G)$ and driven by independent Gaussian increments. We use $\Delta t = 10^{-3}$, burn in each trajectory for 12,000 steps, and collect one state per trajectory per sampling round, advancing each trajectory by 10 steps between rounds. Conditional on the simulator parameters, different trajectories are independent. 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,

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

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

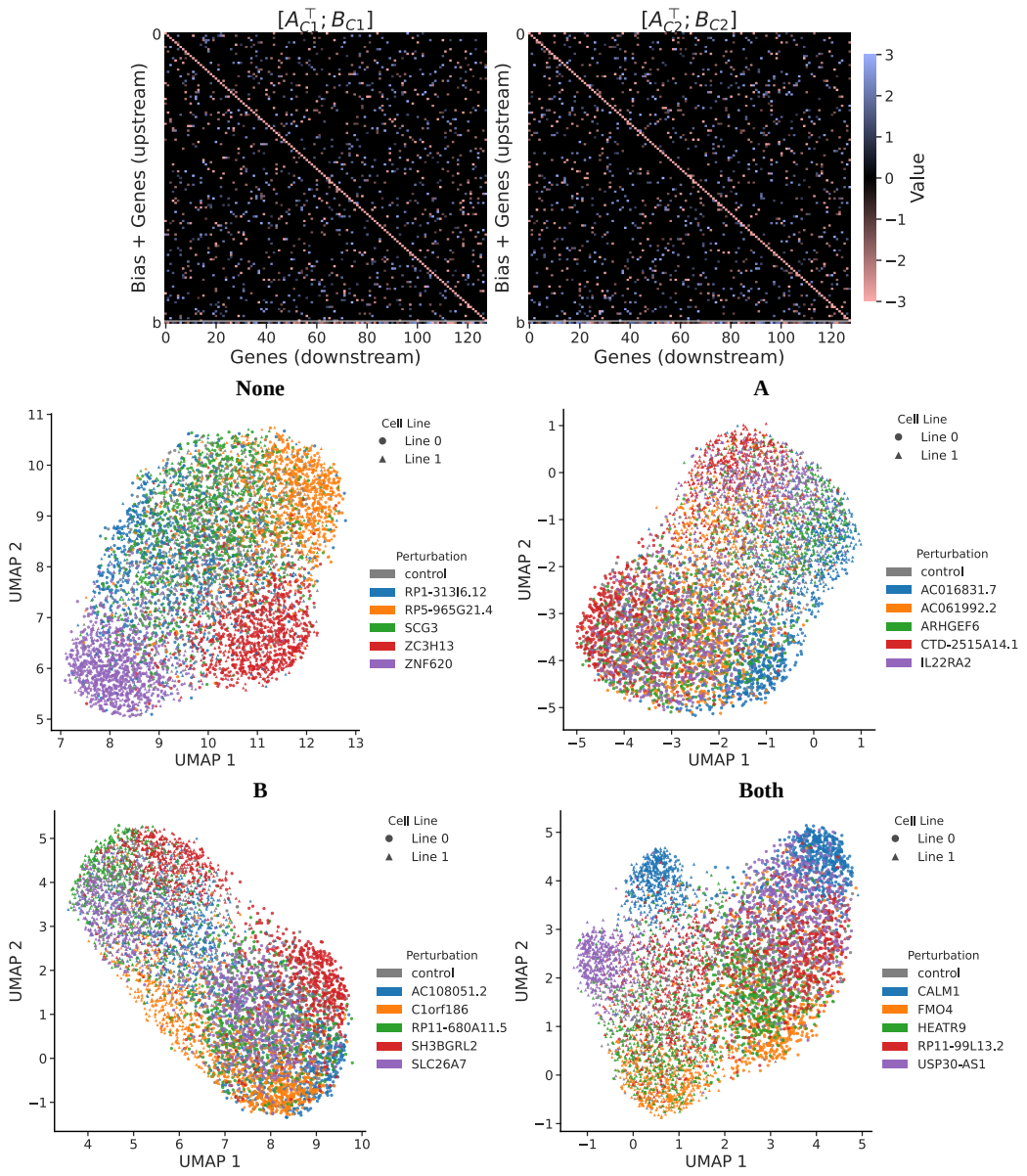


Figure 6: Qualitative visualization of CAUSALDGP under the four mechanism-diversity modes. **Top row:** context-specific latent mechanisms for the two simulated contexts, shown as concatenated parameter blocks $[A_C^\top; B_C]$, where A_C is the sparse signed interaction matrix and B_C is the baseline 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 $C \in \{0, 1\}$.

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

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

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 $Y_{ig} \sim \text{NB}(\mu_g(X_i), \theta_g)$ for each gene g (12)
  end for
end for
return 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.

B.2.1 VALIDATING CAUSALDGP

Qualitative cell-level heterogeneity Figure 6 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.

Benchmarking CAUSALDGP with real Perturb-seq data We contrast CAUSALDGP with real Perturb-seq data (per cell context) along marginal, co-expression, perturbation-effect, effective diversity, and discriminability metrics (Table 5). Notably, CAUSALDGP falls within observed ranges for dispersion (CoV), sparsity (Zero), co-expression (Frac$|r|>0.3$), and effect size ($\sqrt{\text{TI}}$).

Remarks Our theoretical analysis assumes a linear SDE (2), whereas real perturbation responses involve nonlinear, multilayered regulatory cascades (Amit et al., 2009). Linearity is confined to the latent drift; although the emission model uses a gene-wise nonlinear softplus link with a negative binomial observation model, CAUSALDGP does not explicitly represent nonlinear multi-gene syner-

Table 5: Perturbation effect, effective diversity, discriminability, and marginal statistics per cell-line context. **G**: number of genes; **P**: number of perturbations; **CoV**: gene-wise CoV (dispersion); **Zero**: gene-wise zero proportion (sparsity); **Frac**$|r|>0.3$: fraction of gene pairs with absolute Pearson correlation > 0.3 (metacell-aggregated); n_{DEG}: DEGs per perturbation; $\sqrt{\textbf{TI}}$: square root of *noise adjusted* transcriptome-wide impact (Nadig et al., 2025), reported as median [95% CI]; **Vendi**: pseudobulk Vendi diversity score; **PDS** (ℓ_1): perturbation discrimination score. Marginal statistics/effect values are bootstrap medians; Vendi and PDS (ℓ_1) (estimated by split-half self-consistency) are means over 5 seeds.

Dataset	Context	G	P	CoV	Zero	Frac$\|r\|>0.3$	n_{DEG}	$\sqrt{\mathrm{TI}}$ (median [95% CI])	Vendi	PDS (ℓ_1)
CAUSALDGP (A)	0	128	128	2.561	0.803	0.0337	65	1.121 [1.042, 1.241]	31.84	1.000
	1	128	128	2.479	0.811	0.0332	62	1.106 [1.005, 1.225]	33.39	1.000
CAUSALDGP (B)	0	128	128	2.611	0.824	0.0379	64	1.166 [1.079, 1.240]	30.56	1.000
	1	128	128	2.556	0.812	0.0333	61	1.110 [1.006, 1.213]	34.87	1.000
CAUSALDGP (A+B)	0	128	128	2.510	0.797	0.0378	65	1.077 [0.970, 1.165]	34.86	1.000
	1	128	128	2.570	0.819	0.0379	64	1.044 [0.972, 1.114]	30.76	1.000
CAUSALDGP (None)	0	128	128	2.561	0.806	0.0470	60	1.024 [0.945, 1.102]	25.78	1.000
	1	128	128	2.633	0.814	0.0282	64	1.066 [0.999, 1.123]	30.44	1.000
NORMAN19	K562	8,192	156	9.43	0.997	0.006	974	1.71 [1.69, 1.74]	97.63	0.992
REPLOGLE22/NADIG25	K562	7,226	827	1.424	0.590	0.041	48	0.086 [0.080, 0.090]	9.30	0.821
	RPE1	7,226	827	1.589	0.643	0.178	285	0.166 [0.149, 0.184]	9.45	0.913
ZHU25	D1_Stim48hr	3,463	91	2.31	0.861	0.0098	22	1.92 [1.87, 1.95]	2.94	0.914
	D2_Stim48hr	3,463	91	2.25	0.891	0.0121	184	2.36 [2.21, 2.41]	5.06	0.947
	D3_Stim48hr	3,463	91	2.19	0.864	0.0079	151	2.28 [2.22, 2.31]	6.09	0.973
	D4_Stim48hr	3,463	91	2.17	0.849	0.0070	6	1.79 [1.76, 1.80]	3.51	0.883

gies. Context-specific mechanisms are encoded by A_C and B_C, while the intervention is represented by Γ_q. This tractable formulation allows the effects of causal mechanism shifts to be precisely characterized, and the simulated cross-context generalization gaps are consistent with those observed in real Perturb-seq data. Context-dependent intervention effects could be incorporated by allowing Γ_q to depend on C, while nonlinear regulatory interactions would require replacing the linear drift with a nonlinear transition function. Existing real-data benchmarks lack ground-truth causal structure, while semi-synthetic simulators such as CAUSALDGP provide it, albeit under simplified dynamics and observation models. Importantly, the central conclusion remains unchanged: transportability depends on whether the causal mechanisms governing perturbation responses are invariant across contexts.

C DATASETS

C.1 PERTURB-SEQ DATASET PRE-PROCESSING

Table 6 summarizes the pre-processing pipeline. We follow Dibaeinia et al. (2026) for ZHU25 and Mejia et al. (2025) for NORMAN19 and REPLOGLE22/NADIG25.

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.

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 $\mathcal{Q}_{\mathrm{held}}$, from training, creating unseen (perturbation, context) combinations (q, c_{test}) for $q \in \mathcal{Q}_{\mathrm{held}}$. The remaining perturbations in $\{(q', c_{\mathrm{test}}) : q' \notin \mathcal{Q}_{\mathrm{held}}\}$ and all perturba-

Table 6: Dataset-specific pre-processing and filtering criteria for real-world Perturb-seq datasets.

Criterion	NORMAN19	REPLOGLE22/NADIG25	ZHU25
Cell-level QC			
Minimum detected genes per cell	200	200	100
Total UMI filter	Source-study QC[a]	Source-study QC[a]	$[\max\{1{,}400, m - 9\,\mathrm{MAD}\},\ m + 9\,\mathrm{MAD}]$
Gene-level QC			
Minimum gene abundance	≥ 3 cells	≥ 3 cells per pair	≥ 100 cells and ≥ 100 total UMIs
Feature selection			
Selection scope	All retained cells	Each K562–anchored pair	Each time point; donors pooled
HVG	8,192	7,226 (RPE1) / 7,595 (HepG2) / 7,563 (Jurkat)	3,463
Perturbation filtering			
Minimum cells per perturbation	256	64	64
Knockdown-ratio filter	N/A	Source-study QC[a]	< 0.5 at guide and cell levels[b]
Minimum donors per time point	N/A	N/A	≥ 2 at every time point
Dataset summary			
Analysis contexts	1 (K562)	4 cell lines in 3 K562–anchored pairs	12 donor–timepoint combinations (4×3)
Perturbation design	Single- and dual-gene (CRISPRa)	Single-gene (CRISPRi)	Single-gene (CRISPRi)

[a] No additional filter of this type was applied; source-study QC was retained.
[b] Ratios compare target-gene expression with the matched donor–timepoint control mean.
Final matrices were normalized to 10^4 counts per cell and `log1p`-transformed. HVG, highly variable gene; MAD, median absolute deviation; UMI, unique molecular identifier.

tions in other contexts, including $\mathcal{Q}_{\mathrm{held}}$ itself, form the training pool. Control cells in c_{test} are split $50/20/30\%$ as in the 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 7):

(i) *Perturbation alignment*: the within-context test set in c_{test} is restricted to the same held-out perturbations $\mathcal{Q}_{\mathrm{held}}$, 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.

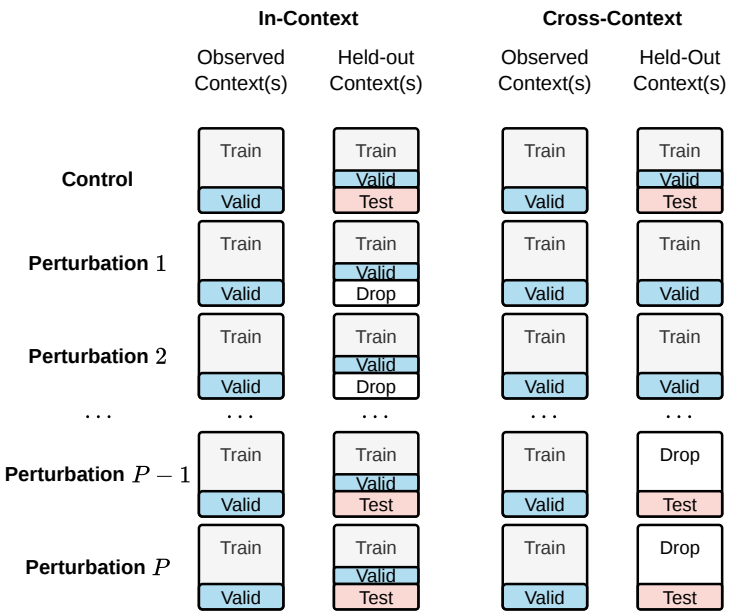


Figure 7: Within-context vs. cross-context splitting strategy. Cells marked "Drop" are removed during fairness alignment.

D METRICS

D.1 EFFECTIVE PERTURBATION RESPONSE DIVERSITY: VENDI SCORE

Table 7 details the null-calibration hyperparameters and how they are determined. Calibration is performed separately within each evaluation context.

Table 7: Null-calibration hyperparameters for the pseudobulk Vendi score.

Symbol	Parameter	Value	Summary
n_{null}	Control split size	$\min\left\{ \begin{array}{l} \lfloor \text{median}_q(N_q) \rfloor, \\ \lfloor N_0/2 \rfloor \end{array} \right\}$	Approximates the typical empirical perturbation sample size. The second term ensures that two disjoint control subsets can be formed.
S	Number of null splits	200	Number of independently generated pairs of disjoint control subsets. Cells may be reused across different splits.
ρ	Null-distance quantile	0.95	Uses the 95th percentile of the control split distances as a conservative estimate of finite-sample null variation.
κ	Target null similarity	0.95	Sets the RBF scale so that a distance equal to δ_ρ has similarity 0.95.
σ^2_{null}	Derived RBF variance	$-\dfrac{\delta_\rho}{2\log\kappa}$	Ensures $\exp[-\delta_\rho/(2\sigma^2_{\text{null}})] = \kappa$.

Here, N_q denotes the number of empirical cells for perturbation $q \in \{1, \ldots, P\}$, subject to the dataset-specific minimum thresholds reported in Appendix Table 6, and N_0 denotes the number of empirical control cells (see Table 1). Calibration is performed separately within each evaluation context. We set $d = \min(50, P, G)$ PCA components by default.

D.2 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 score 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 and 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).

Transcriptome-wide impact (TI) Following the transcriptome-wide impact (TI) framework proposed by Nadig et al. (2025), we combine `DESeq2` (Love et al., 2014) with adaptive shrinkage (`ash`) (Stephens, 2017) to estimate the noise-adjusted distribution of perturbation effects. Counts are summed within each condition and replicate to form pseudobulks. For gene g and pseudobulk i, DESeq2 fits

$$K_{gi} \sim \mathrm{NB}(\mu_{gi}, \alpha_g), \qquad \log \mu_{gi} = \log s_i + \phi_{g0} + \phi_{gq} x_{iq} + \sum_j \gamma_{gj} b_{ij}\,, \tag{16}$$

where s_i is size factor estimated by median-of-ratios by default, x_{iq} indicates perturbation q, b_{ij} indicates batch j, and α_g is the gene-specific dispersion. The perturbation-versus-control contrast

yields a $\log_2$ fold-change estimate $\hat{\beta}_{gq} = \hat{\phi}_{gq}/\ln 2$ and standard error $\hat{\sigma}_{gq} = \mathrm{SE}(\hat{\beta}_{gq})$, with sampling model

$$\hat{\beta}_{gq} \mid \beta_{gq}, \hat{\sigma}_{gq} \sim \mathcal{N}(\beta_{gq}, \hat{\sigma}_{gq}^2)\,. \tag{17}$$

To separate biological effects from estimation noise, `ash` models the centered true effects across genes using a zero-centered normal mixture,

$$g_q(\beta) = \pi_{q0}\delta_0 + \sum_{k=1}^{K} \pi_{qk}\mathcal{N}(0, \omega_k^2), \qquad \sum_{k=0}^{K} \pi_{qk} = 1\,, \tag{18}$$

and estimates the mixture weights by marginal maximum likelihood. TI is the variance of the fitted effect-size distribution,

$$\mathrm{TI}_q = \mathrm{Var}_{\beta\sim g_q}(\beta) = \sum_{k=1}^{K} \hat{\pi}_{qk}\omega_k^2\,. \tag{19}$$

Thus, TI is measured in squared $\log_2$-fold-change units, whereas $\sqrt{\mathrm{TI}_q}$ is in $\log_2$-fold-change units.

D.3 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 $\mathcal{D}_q = \{g : p_{q,g}^{\mathrm{adj}} < \alpha_{\mathrm{FDR}}\}$ denote the set of observed DEGs, where $p_{q,g}^{\mathrm{adj}}$ 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). Because DEG-set metrics require predicted single-cell profiles to estimate cell-to-cell variance, we compute them only for single-cell predictors. 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.4 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:

$$\mathrm{MAE}_q = \frac{1}{G}\sum_{g=1}^{G} |\hat{\mu}_{q,g} - \bar{\mu}_{q,g}|\,, \qquad \mathrm{MSE}_q = \frac{1}{G}\sum_{g=1}^{G} (\hat{\mu}_{q,g} - \bar{\mu}_{q,g})^2\,. \tag{20}$$

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 pseudobulk 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. Single-cell distributional distances (MMD, FD, and JSD) are applicable only to predictors that produce single-cell outputs.

Parametric (raw counts) Because scRNA-seq counts are sparse, overdispersed, and often heavy-tailed, we fit separate negative-binomial distributions to the observed and predicted count vectors for each gene and quantify their discrepancy using the Jensen–Shannon divergence (JSD) (Lin, 2002). Because the JSD between two negative-binomial distributions has no closed-form expression, we approximate it by summing their probability mass functions over a truncated discrete support, with the upper bound determined from the ($1 - 10^{-8}$) quantiles of the fitted distributions and the maximum empirical count. The truncated probability masses are then renormalized before computing the JSD. This deterministic procedure avoids Monte Carlo error and respects the discrete, overdispersed nature of the data, but depends on the assumed distributional family and remains 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 using 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, 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 the non-parametric distances in PCA space. For both MMD and FD, we fit a single shared PCA projection to the pooled observed control cells within each evaluation context. We then use this fixed projection to transform the observed and predicted cells for every perturbation condition. The distance is computed separately for each perturbation in this common PCA space and subsequently aggregated across perturbations.

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.

The linearPCA baseline is a bilinear model of the form $\hat{\boldsymbol{y}} = UW\bar{\boldsymbol{p}} + \boldsymbol{b}$, where U contains the PCA gene embeddings, $\bar{\boldsymbol{p}}$ is the mean embedding of the perturbation's target genes, and W is estimated using two-sided ridge regularization with $\lambda = 0.1$ (Table 3). The gene embeddings are obtained by applying full-SVD PCA with up to 10 components to the transpose of the matrix of training perturbation-level pseudobulks, treating genes as observations and perturbations as features.

Table 8 summarizes the architectural and training hyperparameters used by the deep learning models.

E.2 Validation and robustness of the Vendi score

Across 100 randomly simulated DirectDGP datasets (Figure 8) both Vendi and covariance effective rank increased with the number of perturbations, with similar OLS slopes ($\beta_1 = 0.870$ and 0.891, respectively). Effective rank showed a substantially stronger Pearson correlation with P ($r = 0.984$) than Vendi ($r = 0.660$) because it remained tightly coupled to its maximum value of $P - 1$. As expected, Vendi exhibited downward deviations when perturbation effects were weak or sparse. Thus, the higher correlation for effective rank likely reflects a near-mechanical dependence on the number of perturbation groups, whereas Vendi is more sensitive to whether those perturbations produce distinguishable responses. Moreover, in Gaussian-noise experiments on Norman19 (Figure 9), Vendi and PDS (ℓ_1) remain stable under low-to-moderate noise and decline under severe noise, whereas effective rank increases with noise.

Figure 10 evaluates the sensitivity of the Vendi score to perturbation sample size and finite-sample noise. Under the population null ($\epsilon = 1$), severe sample-size heterogeneity increased mean Vendi relative to balanced sampling. This inflation was attenuated at $\epsilon = 1.5$ and negligible at $\epsilon \geq 2$. The

Table 8: Architecture and training hyperparameters for deep learning models.

Model	Architecture	Training
scVI	Three-layer, width-128 MLP encoder and decoder; latent dimension 10; ReLU and batch normalization; encoder dropout 0.1; ZINB likelihood with gene-wise dispersion.	Adam (lr $= 10^{-3}$, wd $= 10^{-6}$); batch 512; up to 800 epochs; KL warmup 400 epochs; early-stopping patience 45.
GEARS	One co-expression and one perturbation-graph SGConv layer, using GO when available and co-expression otherwise; hidden width 64; shared recovery MLP $64 \to 128 \to 64$ for residual gene-space prediction; ReLU and batch normalization; uncertainty head predicting gene-wise log variance.	Adam (lr $= 10^{-3}$, wd $= 5 \times 10^{-4}$); batch 64; 20 epochs; StepLR ($\gamma = 0.5$ per epoch); gradient-value clipping at 1.
CPA	Variational latent dimension 128; hidden width 256; encoder/decoder configured with three layers; perturbation and context embeddings configured with two and one layers, respectively; ReLU; dropout 0.1; Gaussian/MSE reconstruction.	Two Adam optimizers (lr $= 10^{-3}$, wd $= 10^{-8}$); batch 128; up to 100 epochs; early-stopping patience 10; six warmup epochs; generator-to-adversary update ratio 1:6.
STATE	Two-layer basal and perturbation MLPs with width 64; width-64 bidirectional GPT-2 with two blocks, four heads, and FFN width 256; GELU; dropout 0.1; residual gene-space prediction.	Adam (lr $= 3 \times 10^{-4}$, wd $= 0$); batch of four sets of S cells; up to 100 epochs; early-stopping patience 10; energy-distance loss plus 0.1 context-classification and 10^{-5} residual L1 penalties.
scLDM VAE	Eight encoder and eight decoder blocks; width 32; eight self-attention and four cross-attention heads; 16 latent tokens of width 16; SwiGLU; dropout 0; NB decoder.	AdamWLegacy (lr $= 10^{-3}$, $\boldsymbol{\beta} = (0.9, 0.95)$, wd $= 0$); batch 128; $100\lfloor N_{\text{cells}}/128 \rfloor$ steps; 10% warmup and square-root decay over the final 10%.
scLDM DiT	Eight DiT blocks; width 256; eight heads; FFN width 684; 16 latent tokens; adaptive LayerNorm; dropout 0; joint conditioning with CFG dropout 0.8; linear flow matching with velocity prediction.	AdamW (lr $= 5 \times 10^{-4}$, wd $= 0$); same step budget as the VAE; 10% warmup followed by cosine decay; EMA decay 0.9999; adaptive Dopri5 sampling at 50 requested output times, with guidance weight $\omega \in \{1, 5, 10\}$

S is the minimum number of cells per perturbation–context group across the training and validation splits. N_{cells} is the number of scLDM training cells; the stated step budget is applied separately to the VAE and DiT stages.

large-pool mean oracle also yielded a lower Vendi Ratio than the sampling-matched oracle at low effect factors, but this difference diminished with increasing effect factor and sample size, becoming negligible by $N_q = 256$ at $\epsilon = 2$ and at all evaluated sample sizes for $\epsilon = 5$. Thus, fixed-size null calibration and deterministic prediction can introduce bias, but their practical impact is concentrated under null or low-effect conditions and is small for moderate-to-strong effects or larger sample sizes.

E.3 Perturbation response geometry: diversity and discrimination

Figure 11 provides the cross-context counterpart to Figure 3b for the two multi-context datasets, Zhu25 and Replogle22/Nadig25. Relative to within-context evaluation, the dominant change is a leftward shift in PDS, indicating weaker discrimination when models predict unseen perturbation–context combinations. On Zhu25 (circles), all models approach chance-level discrimination (PDS ≈ 0.5) and substantially underestimate response diversity (VR < 1). Thus, cross-context predictions on Zhu25 fail to preserve both perturbation identity and observed diversity.

On Replogle22/Nadig25 (triangles), the degradation is milder and model performance separates more clearly. scLDM retains the strongest discrimination, whereas scVI best preserves diversity, with VR close to one. STATE and STATE-Geneformer retain moderate PDS but underestimate diversity, while GEARS and linearPCA remain near chance and exhibit low VR. These results

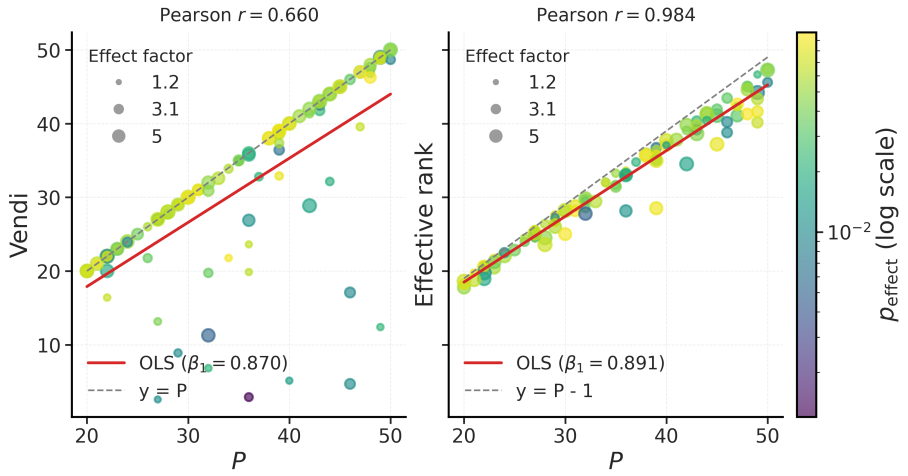
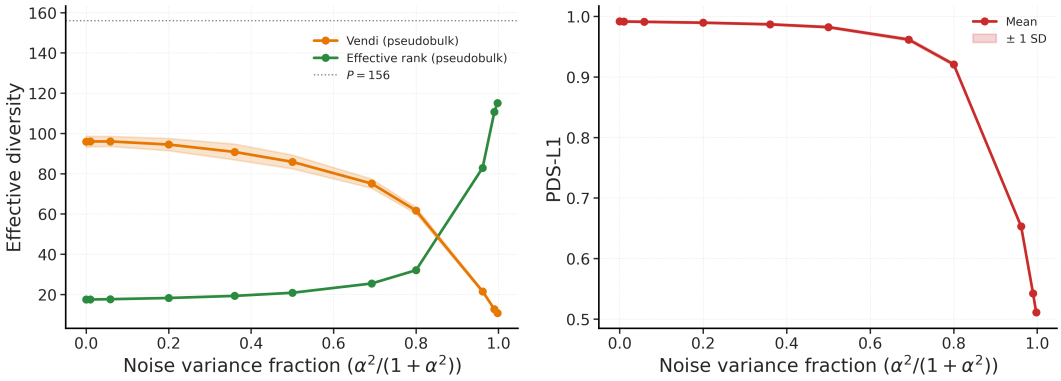


Figure 8: Raw empirical Vendi score (left) and covariance effective rank (right) versus the number of perturbations P across 100 randomly sampled synthetic DIRECTDGP datasets. Each circle is one dataset. Color encodes the perturbation-effect probability p_{effect} on a logarithmic scale, and marker size encodes the effect factor. Red lines show ordinary least-squares fits, with the estimated slopes reported in the legends; Pearson correlations are reported above each panel. Gray dashed lines denote the reference relationships $\mathrm{VS}(Y) = P$ and $\mathrm{erank}(Y) = P - 1$, respectively.



(a) Vendi score and covariance effective rank.

(b) PDS-L1.

Figure 9: Robustness to Gaussian noise on the NORMAN19 dataset. We inject Gaussian noise scaled to each gene's own variability, $\sigma_g = \alpha\,\mathrm{std}_g$, so α is a dimensionless noise-to-signal ratio and the noise variance fraction is $\alpha^2/(1+\alpha^2)$. (a) Vendi score and covariance effective rank as functions of the noise variance fraction. (b) PDS (ℓ_1) over the same noise sweep. Lines and markers show the mean across 10 noise realizations, and shaded bands indicate ± 1 standard deviation.

reveal two distinct, model-dependent failure modes under cross-context transfer: loss of perturbation discrimination and loss of response diversity. Strong within-context performance therefore does not necessarily imply robust generalization across biological contexts.

E.4 CROSS-CONTEXT GENERALIZATION GAP

Appendix Figures 12–15 provide the full per-context results underlying Section 4.3.3 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 relative effective diversity.

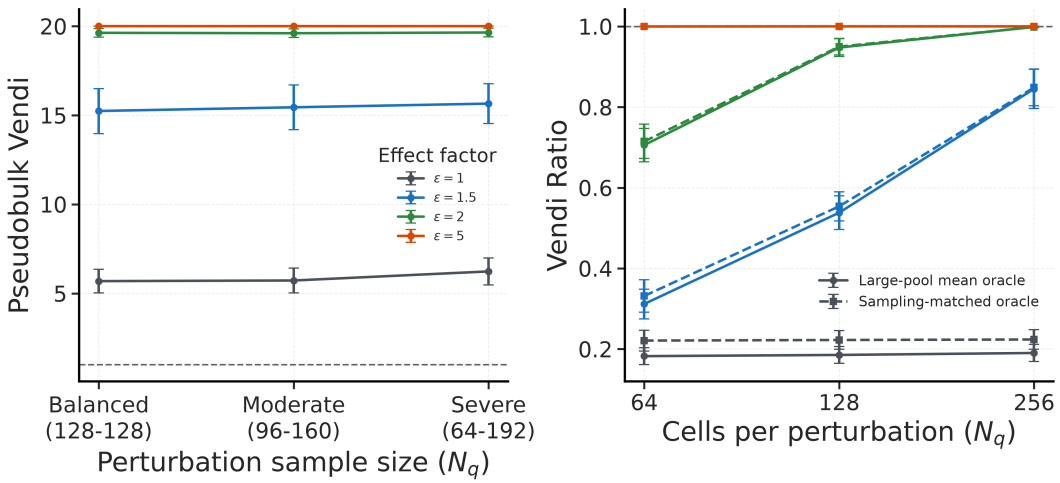


Figure 10: Sensitivity of Vendi score to perturbation sample size and finite-sample noise. (left) Mean Vendi score across balanced ($N_q = 128$), moderately heterogeneous ($N_q = 96$–160), and severely heterogeneous ($N_q = 64$–192) sampling approach. All approaches have the same median $N_q = 128$ and total number of perturbed cells. Colors denote DIRECTDGP effect factors $\epsilon \in \{1, 1.5, 2, 5\}$. (right) Vendi Ratios for a large-pool mean oracle, computed from 2,048 cells per perturbation, and an independent sampling-matched oracle containing the same N_q as the empirical target. Ratios are shown for $N_q \in \{64, 128, 256\}$; colors denote effect factor, while solid circles and dashed squares denote the large-pool and sampling-matched oracles, respectively. Points and error bars show mean $\pm$ one standard deviation across 10 paired simulation seeds. Simulations used $P = 20$, $G = 1024$, $N_0 = 512$, $p_{\text{effect}} = 0.1$, and $\beta = 0$. Horizontal dotted lines indicate a Vendi score or Vendi Ratio of one.

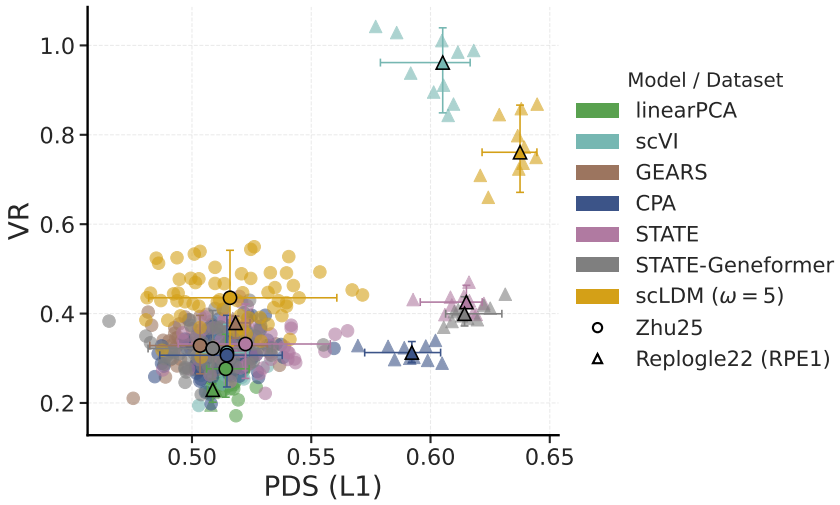


Figure 11: Vendi ratio (VR) vs. PDS (ℓ_1) across models and datasets under cross-context evaluation. Color encodes the model and marker shape encodes dataset. Bold markers and error bars show the median and 95% percentile intervals per model–dataset pair. Higher is better on both axes.

E.4.1 CROSS-CONTEXT GENERALIZATION GAP FOR ZHU25

Figure 12 shows a broad cross-context generalization gap on ZHU25 across all six held-out donor–activation combinations. Under within-context evaluation, STATE, STATE-Geneformer, and scLDM generally perform best across DEG-restricted R^2, DES recall, MAE, and MMD, while CPA remains competitive, particularly in R^2. Under cross-context evaluation, these advantages contract sharply. Across all six contexts, DEG-restricted R^2 decreases and MAE and MMD increase for all four

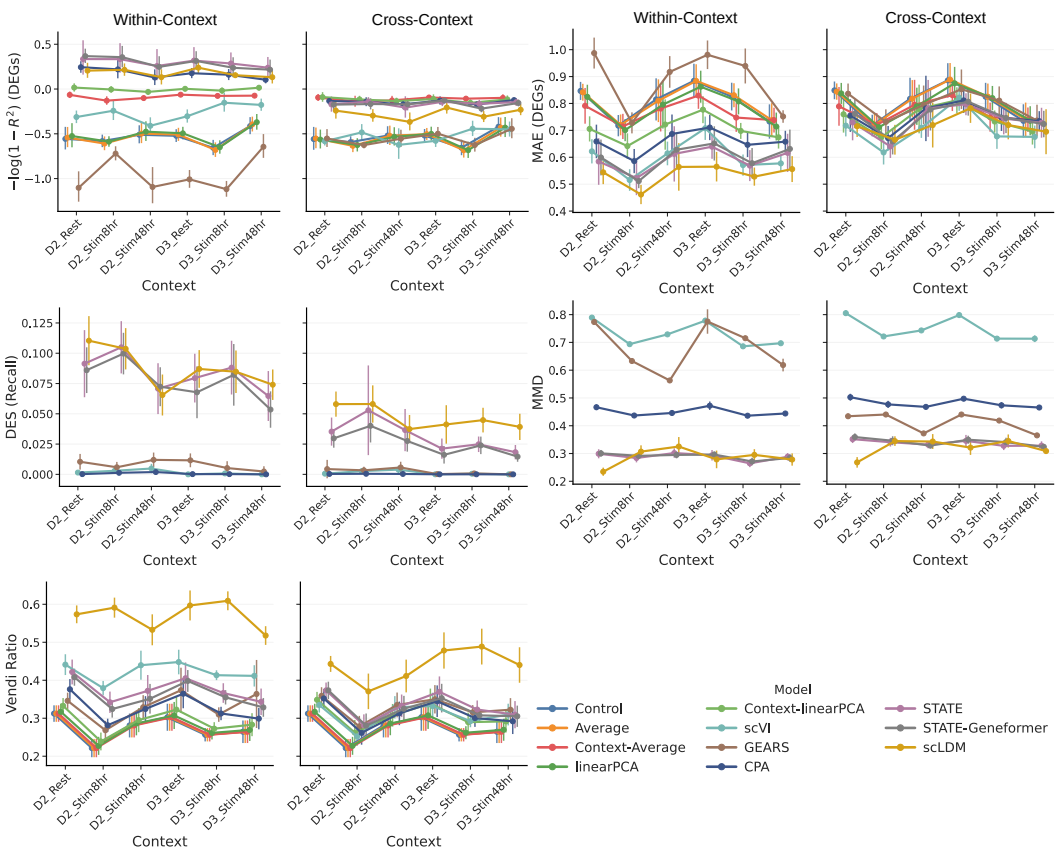


Figure 12: Additional results on the cross-context generalization gap in ZHU25 under within-context and cross-context evaluation. The x-axis denotes the six test contexts (2 held-out donors $\times$ 3 activation states). We plot medians and interquartile ranges. scLDM is evaluated with $\omega = 5$.

models; DES recall also declines consistently for STATE, STATE-Geneformer, and scLDM. Median R^2 falls below zero, indicating poor perturbation-specific agreement for unseen perturbation–context combinations.

Because these declines recur across all six contexts, they are not driven by a single donor or activation state. Relative performance nevertheless remains metric-dependent: CPA, STATE, and STATE-Geneformer retain the strongest R^2, whereas scLDM retains the highest DES recall and lowest MMD. Weaker methods exhibit less consistent changes on some metrics, partly because their within-context performance is already low.

VR requires a more nuanced interpretation. It changes little for CPA and only moderately for STATE and STATE-Geneformer, but declines substantially for scLDM. Moreover, cross-context VR remains well below one for every model, indicating systematic underestimation of response diversity. Thus, cross-context failure is not explained solely by additional diversity collapse: models also lose alignment between predicted responses and the correct perturbations in held-out biological contexts.

E.4.2 CROSS-CONTEXT GENERALIZATION GAP FOR REPLOGLE22/NADIG25

Figure 13 shows systematic cross-context generalization gaps for RPE1–K562. Although RPE1–K562 is the least concordant REPLOGLE22/NADIG25 pair, its TI correlation remains substantially higher than the cross-donor correlations in ZHU25 (Table 9). Cross-context evaluation therefore reduces performance without eliminating perturbation-specific signal. Performance is metric-dependent. Within context, STATE-Geneformer achieves the strongest DEG-restricted Pearson and R^2_{δ}, whereas scLDM achieves the strongest PDS (ℓ_1) and lowest MMD; STATE and CPA remain competitive. The context-aware linear baselines are also informative: Context-Average and Context-linearPCA retain moderate DEG-restricted Pearson under transfer, with Context-linearPCA approaching CPA. However,

their near-random PDS and negative R^2_δ indicate limited perturbation-specific discrimination. Thus, simple context-conditioned summaries capture a transportable component of the average response structure, but not the full perturbation-specific geometry.

Across RPE1–K562, HepG2–K562, and Jurkat–K562, CPA, STATE, STATE-Geneformer, and scLDM exhibit positive transfer gaps in DEG-restricted Pearson and PDS. RPE1–K562, which has the lowest TI concordance (Table 9), consistently exhibits the largest gaps in both metrics (Figure 14). Other metrics show less consistent relationships with TI concordance, with changes in R^2_δ, MAE, DES recall, MMD, and VR depending on both model and context pair. Thus, mechanistic similarity aligns most clearly with transfer in DEG-restricted Pearson and PDS.

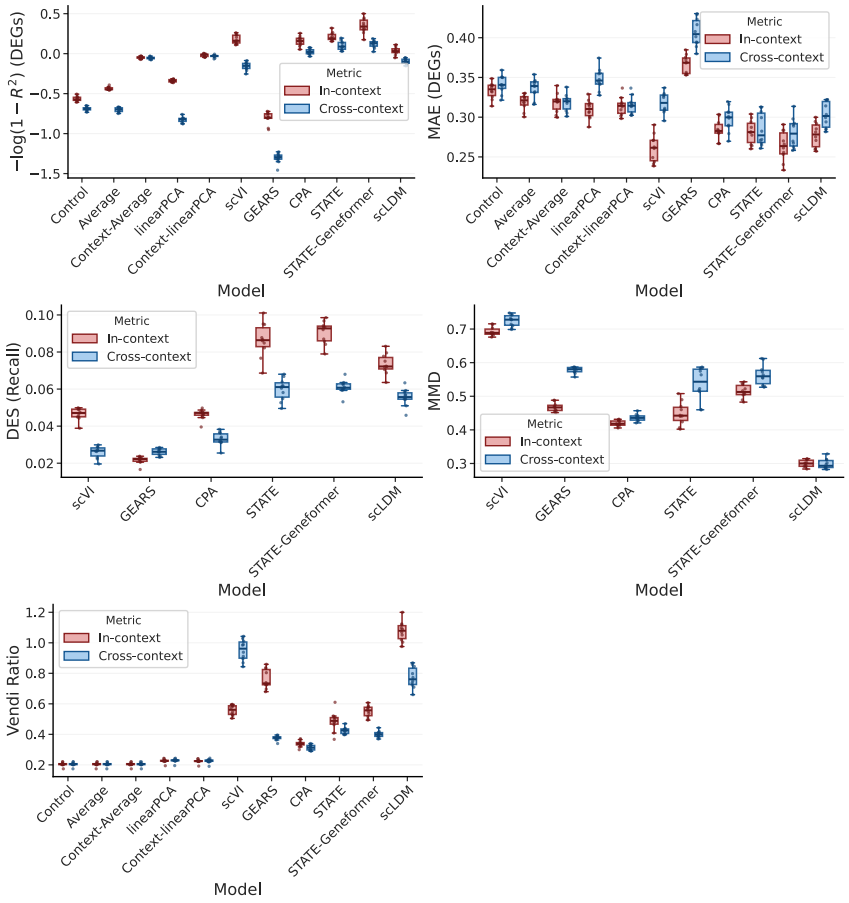


Figure 13: Additional results for the cross-context generalization gap on REPLOGLE22/NADIG25 (RPE1–K562) under within-context vs. cross-context evaluation. Models are trained on RPE1 and evaluated on held-out K562 perturbations. scLDM is evaluated with $\omega = 5$.

E.4.3 CROSS-CONTEXT GENERALIZATION GAP FOR CAUSALDGP

Figure 15 provides a controlled view of cross-context generalization because the source of context shift is known and tunable by construction. The results support the transportability condition in Section 2.3. When neither the baseline state nor the interaction mechanism changes (`None`), CPA, GEARS, and scLDM retain strong cross-context performance, with only small reductions in DEG-restricted Pearson, PDS, and DES recall. Changing only the baseline drift (`B`) produces moderate degradation for CPA and GEARS, whereas scLDM remains comparatively stable.

The pattern changes sharply when the interaction matrix A_C differs across contexts. Under `A` and `Both`, CPA, GEARS, and scLDM exhibit substantially larger transfer gaps: DEG-restricted R^2 and Pearson decline, MAE increases, and PDS and DES recall deteriorate. scLDM nevertheless retains the strongest PDS and DES recall under these shifts. STATE remains near chance in PDS across conditions, limiting its ability to reveal the same ordering because of a floor effect.

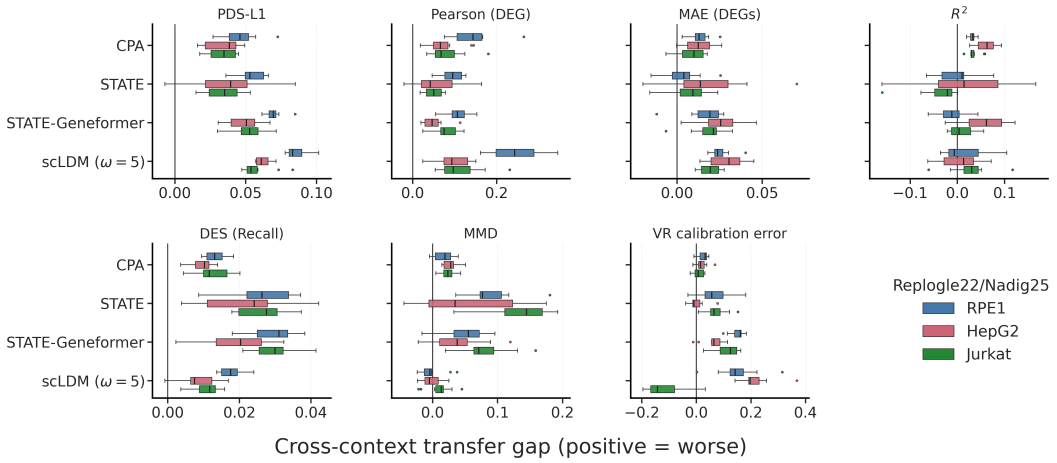


Figure 14: Cross-context transfer gaps on REPLOGLE22/NADIG25 for RPE1–K562, HepG2–K562, and Jurkat–K562. Models are trained in the indicated source context and evaluated on held-out K562 perturbations. Box plots summarize 10 paired trial-level gaps. Gaps are oriented so that positive values consistently indicate worse cross-context performance. For VR, the gap is the cross-context increase in absolute deviation from perfect calibration (VR $= 1$).

For scLDM, VR moves closer to one under `A` and `Both` even as perturbation-specific accuracy deteriorates. Thus, improved diversity calibration does not imply correct perturbation-response alignment, reinforcing that diversity and discrimination capture distinct failure modes. Overall, transfer gaps are smallest under `None`, larger under `B`, and largest under `A` and `Both` for the models that perform well in context. This supports the prediction that changes to the perturbation-response mechanism are more damaging than baseline-state shifts alone.

E.4.4 ALIGNMENT BETWEEN THEORY AND EXPERIMENTS

Table 9: Cross-context concordance of transcriptome-wide impact (TI). For each dataset and context pair, per-perturbation TI estimates in the source context are compared with those in the target context across the shared P perturbations. We report Pearson r, Spearman ρ, and Lin's concordance correlation coefficient (CCC) (Lin, 1989), which penalizes departures from the identity line $y = x$. Lower values across these metrics indicate greater context dependence in perturbation impact.

Dataset	Context	P	Pearson r	Spearman ρ	Lin's CCC
CAUSALDGP	`None` (0 → 1)	128	0.962	0.945	0.962
	`B` (0 → 1)	128	0.929	0.919	0.928
	`A` (0 → 1)	128	0.507	0.567	0.507
	`A + B` (0 → 1)	128	0.549	0.630	0.546
ZHU25 Stim48hr	D1 → D3	91	0.093	0.101	0.085
	D4 → D3	91	0.261	0.228	0.203
	D1 → D2	91	-0.072	-0.069	-0.057
	D4 → D2	91	-0.094	-0.093	-0.062
REPLOGLE22/NADIG25	RPE1 → K562	827	0.639	0.631	0.301
	HepG2 → K562	479	0.692	0.803	0.426
	Jurkat → K562	983	0.698	0.750	0.648

Table 9 grounds the divergence parameter in an empirically measurable quantity: the cross-context concordance of transcriptome-wide impact (TI) induced by the CAUSALDGP influence matrix, A, recapitulates the range observed across real biological contexts. For example, as the divergence parameter increases, CAUSALDGP reproduces the low cross-donor TI concordance observed in ZHU25 (Figure 5). This low concordance coincides with poor generalization across all models, supporting TI concordance as a real-data proxy for mechanistic divergence. A similar pattern

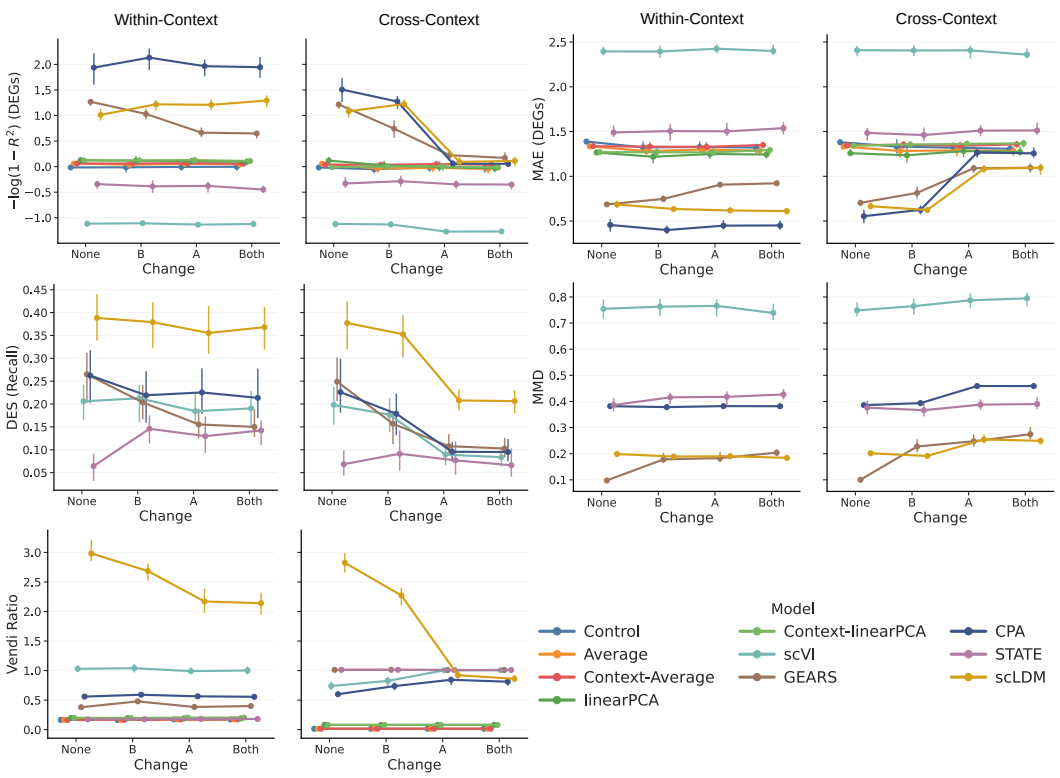


Figure 15: Additional results on the cross-context generalization gap in CAUSALDGP under within-context and cross-context evaluation. The x-axis denotes the four cross-context mechanism modes (`None`, `A`, `B`, `Both`). We plot medians and interquartile ranges. scLDM is evaluated with $\omega = 5$.

Table 10: STATE on REPLOGLE22/NADIG25 (RPE1 → K562), for the 532 perturbations held out in ≥ 5 of 10 seeds. We test whether each perturbation's mean transfer gap (PDS (ℓ_1)) differs from zero (variance-stabilized and BH FDR-controlled). No perturbation shows a corroborated *positive* gap.

Class	n	% of perts	Median gap	Mean gap	$\sum$ gap (% of total)
Consistent negative (transfer failure)	57	10.7%	−0.153	−0.239	−13.61 (55%)
No significant gap	475	89.3%	−0.011	−0.024	−11.18 (45%)
All analyzed	532	100%	−0.016	−0.047	−24.79 (100%)

emerges across cell lines: Jurkat and HepG2 exhibit comparably high TI concordance with K562, whereas RPE1 exhibits substantially lower concordance. Correspondingly, RPE1 → K562 produces the largest transfer gaps across all four models for both DEG-restricted Pearson $\boldsymbol{\delta}$ and PDS (ℓ_1) (Figure 14). Together, these results support the theoretical prediction in (6) that transfer gaps widen as TI concordance decreases.

A perturbation-level analysis of STATE on RPE1 → K562 further localizes this aggregate transfer penalty. Table 10 shows that the cross-context penalty is concentrated among a small, reproducible subset of perturbations. Moreover, perturbations with more divergent effects exhibit larger cross-context generalization gaps, reinforcing the relationship between effect divergence and the reduced TI concordance reported in Table 9.

E.5 scLDM CLASSIFIER-FREE GUIDANCE SENSITIVITY ANALYSIS

Figures 16–18 compare classifier-free guidance weight $\omega \in \{1, 5, 10\}$, revealing metric-dependent effects. On CAUSALDGP, stronger guidance improves cross-context PDS (ℓ_1) when A changes, but reduces cross-context Pearson $\boldsymbol{\delta}$. For RPE1–K562, PDS improves while cross-context Pearson $\boldsymbol{\delta}$

changes little. On ZHU25, guidance improves in-context Pearson $\boldsymbol{\delta}$ and PDS, but their cross-context values remain near zero and chance, respectively. Across datasets, stronger guidance increases both VR and MMD, indicating that greater response diversity does not necessarily improve distributional fidelity. Thus, tuning ω changes the trade-offs between metrics but does not eliminate the cross-context generalization gaps.

Table 11 also summarizes the sensitivity of scLDM to ω. Overall, PDS (ℓ_1) increases with guidance strength, whereas Pearson $\boldsymbol{\delta}$ (DEG) generally improves from $\omega = 1$ to $\omega = 5$ before plateauing or declining at $\omega = 10$. These results reveal a trade-off between perturbation discriminability and per-gene fidelity under stronger guidance. VR indicates that predicted effective diversity is lower than observed at $\omega = 1$, generally closest to observed at $\omega = 5$, and higher than observed at $\omega = 10$.

Accordingly, we use $\omega = 5$ throughout the remaining experiments because it provides the best overall balance among discriminability, per-gene fidelity, and diversity calibration.

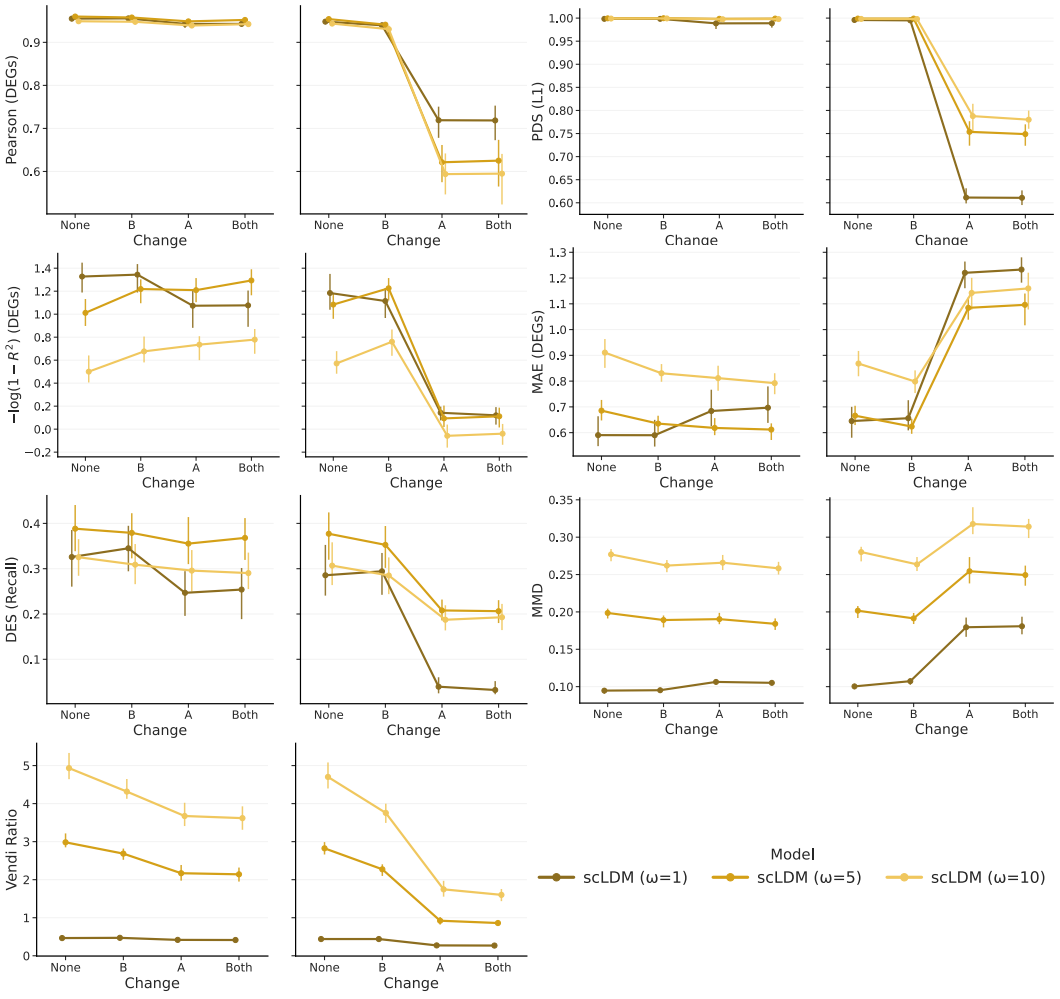


Figure 16: Sensitivity of scLDM to the classifier-free guidance weight ω on CAUSALDGP.

E.6 METRIC DISTORTION

Figure 19 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.

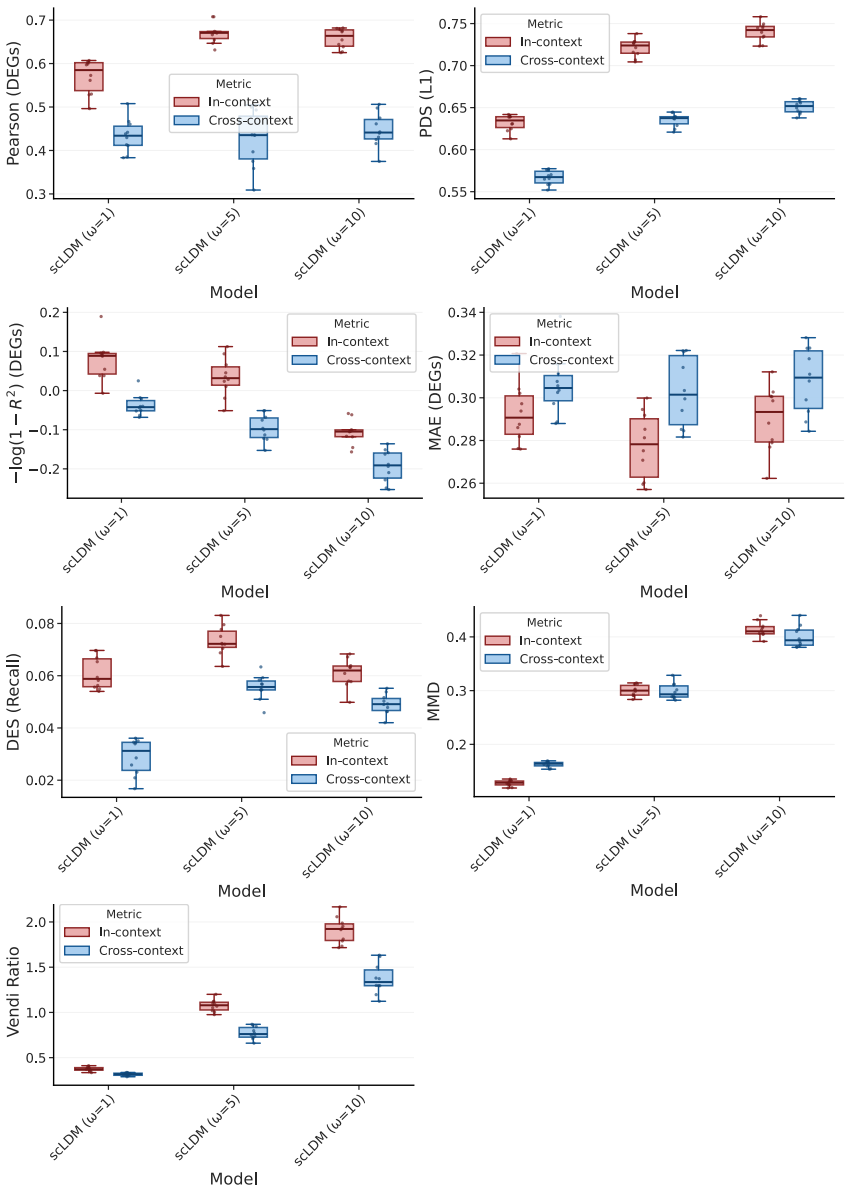


Figure 17: Sensitivity of scLDM to the classifier-free guidance weight ω onREPLOGLE22/NADIG25.

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 $\mu_{g,j}^{(q)} = \ell_j \alpha_{q,g}(\mu_g^c + \beta\lambda_g)$.

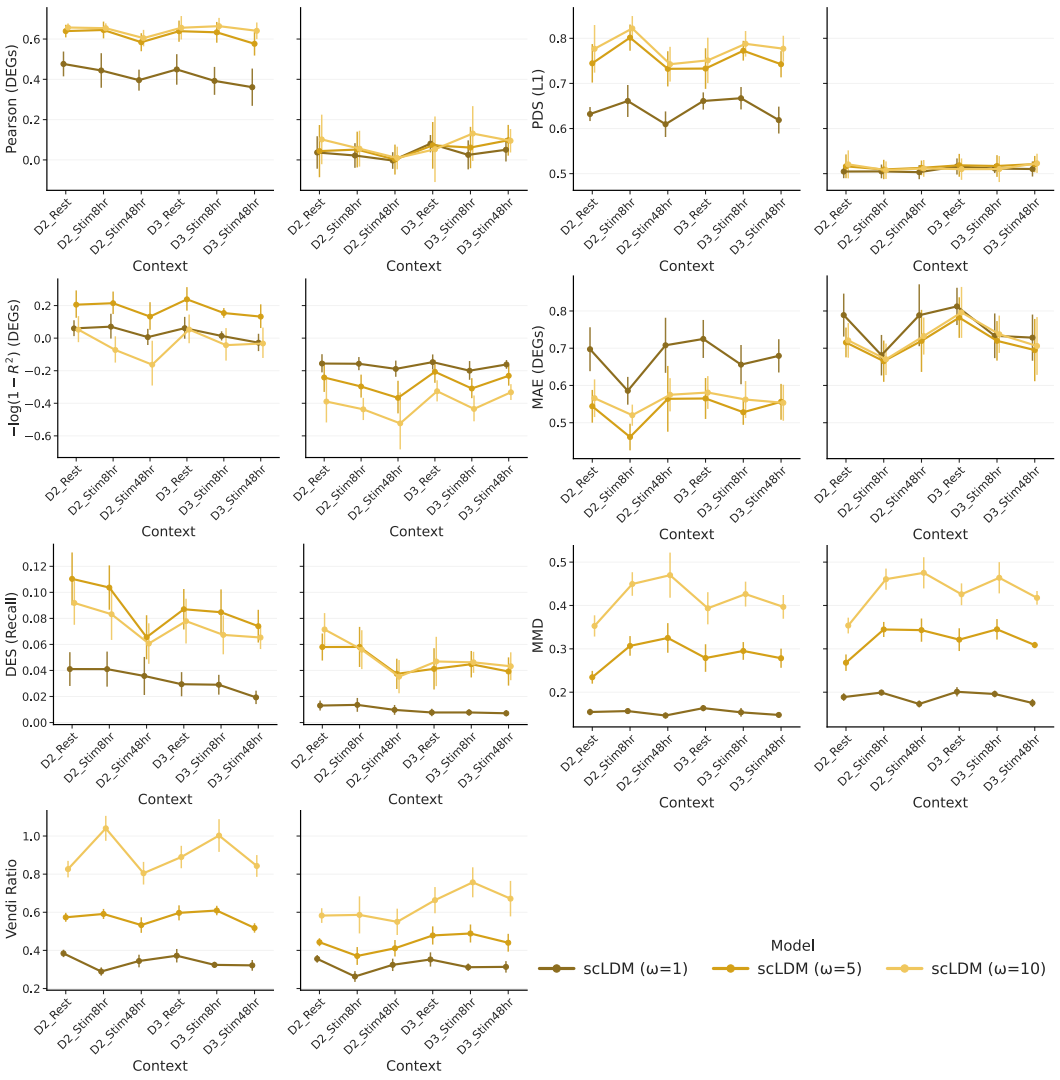


Figure 18: Sensitivity of scLDM to the classifier-free guidance weight ω on ZHU25.

Table 11: Sensitivity of scLDM to the classifier-free guidance weight ω across three K562-anchored REPLOGLE22/NADIG25 cell-line pairs. Values are reported as median [Q1, Q3] over 10 trials for PDS (ℓ_1), Pearson δ (DEG-restricted), Vendi ratio (VR). Cross-context performance measures generalization to K562 perturbations withheld from K562 training but observed in the paired source cell line.

Source	Model	PDS (ℓ_1)		Pearson δ (DEG)		VR	
		In-context	Cross-context	In-context	Cross-context	In-context	Cross-context
RPE1	scLDM ($\omega = 1$)	0.635 [0.626, 0.639]	0.567 [0.560, 0.574]	0.585 [0.538, 0.602]	0.434 [0.412, 0.456]	0.372 [0.363, 0.390]	0.318 [0.305, 0.328]
	scLDM ($\omega = 5$)	0.724 [0.715, 0.728]	0.638 [0.631, 0.639]	0.671 [0.658, 0.674]	0.435 [0.381, 0.479]	1.079 [1.027, 1.112]	0.761 [0.727, 0.833]
	scLDM ($\omega = 10$)	0.742 [0.734, 0.747]	0.652 [0.645, 0.657]	0.664 [0.640, 0.677]	0.441 [0.427, 0.471]	1.922 [1.795, 1.979]	1.336 [1.295, 1.470]
HepG2	scLDM ($\omega = 1$)	0.693 [0.684, 0.695]	0.629 [0.623, 0.631]	0.685 [0.667, 0.705]	0.584 [0.531, 0.611]	0.263 [0.250, 0.267]	0.217 [0.205, 0.228]
	scLDM ($\omega = 5$)	0.787 [0.782, 0.797]	0.727 [0.720, 0.736]	0.758 [0.748, 0.763]	0.666 [0.631, 0.700]	0.870 [0.784, 0.890]	0.614 [0.589, 0.670]
	scLDM ($\omega = 10$)	0.804 [0.794, 0.812]	0.739 [0.735, 0.745]	0.734 [0.727, 0.744]	0.677 [0.639, 0.694]	1.394 [1.278, 1.436]	0.994 [0.959, 1.105]
Jurkat	scLDM ($\omega = 1$)	0.654 [0.648, 0.656]	0.604 [0.598, 0.613]	0.630 [0.620, 0.649]	0.483 [0.465, 0.519]	0.405 [0.390, 0.416]	0.355 [0.341, 0.367]
	scLDM ($\omega = 5$)	0.751 [0.749, 0.756]	0.697 [0.690, 0.702]	0.729 [0.719, 0.744]	0.645 [0.591, 0.660]	1.227 [1.203, 1.272]	1.056 [0.846, 1.091]
	scLDM ($\omega = 10$)	0.766 [0.762, 0.773]	0.711 [0.704, 0.715]	0.704 [0.700, 0.720]	0.645 [0.609, 0.662]	2.221 [2.106, 2.332]	1.899 [1.462, 1.961]

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

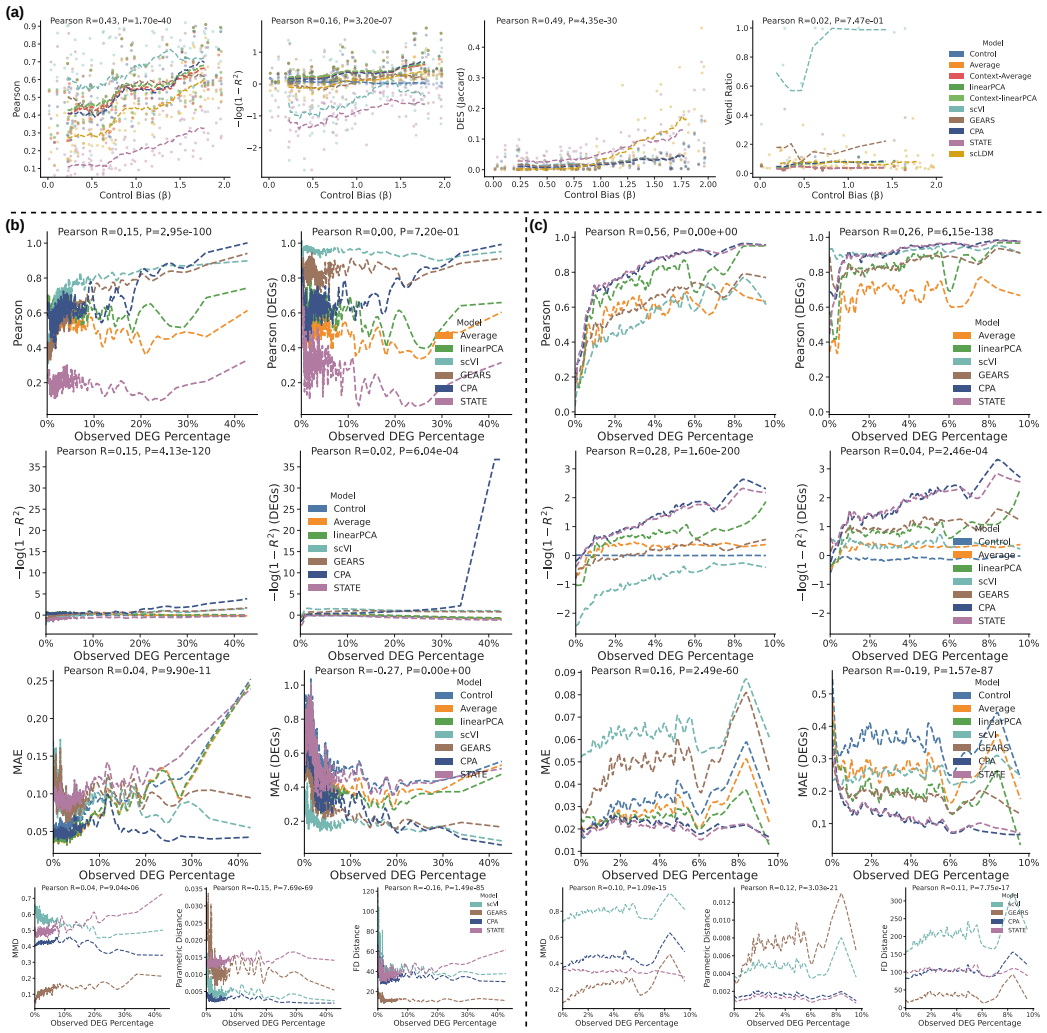


Figure 19: Metric distortion on DIRECTDGP and NORMAN19. Points show individual runs, coloured by model; dashed curves show per-model LOESS-smoothed trends. Top annotations report the Pearson correlation (R) and two-sided p-value (P) computed from all raw points in each panel. (a) On DIRECTDGP: Pearson $\boldsymbol{\delta}$, $R^2_{\boldsymbol{\delta}}$, 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.

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 19, Pearson $\boldsymbol{\delta}$, $R^2_{\boldsymbol{\delta}}$, and DES (Jaccard) all show noticeable dependence on control bias β, with correlations of $R = 0.43$, 0.16, and 0.49, respectively (all $P \leq 1 \times 10^{-7}$). 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 ($R = 0.02$, $P = 0.747$) 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 19 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.