LATENT CAUSAL PROPAGATION FOR OUT-OF-DISTRIBUTION SINGLE-CELL PERTURBATION PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting single-cell responses to unseen genetic or chemical perturbations requires generalization beyond the perturbations observed during training. Yet the conditions under which latent representations support such generalization remain insufficiently understood. We study this problem through latent causal propagation: perturbations intervene on a basal cell state, propagate through a shared latent structural causal model (SCM), and produce transcriptional responses through a domain-invariant conditional generation kernel. Under explicit assumptions, we establish identifiability results for perturbation-induced latent distributions and, under additional restrictions, the latent structural matrix. We further characterize the cross-domain optimality of an oracle Bayes rule conditioned on the causal latent state and bound target-domain generation error in terms of latent representation and conditional generator approximation errors. Guided by this framework, we introduce scCausalDiff, which combines a learned latent SCM with conditional diffusion generation. Experiments on genetic and chemical perturbation benchmarks under strict OOD splits demonstrate competitive predictive performance across mean-response and distributional metrics.

1 INTRODUCTION

Predicting how single cells respond to previously unseen genetic or chemical perturbations is a central challenge in computational biology, with implications for target identification, mechanistic studies, and drug development (Srivatsan et al., 2020; Replogle et al., 2022; Dixit et al., 2016). Large-scale perturbation screens provide increasingly rich measurements of transcriptional responses, yet the space of interventions across genes, compounds, doses, and cellular contexts remains far beyond experimental coverage. A central computational challenge is therefore to predict responses in perturbation domains absent from training (Hetzel et al., 2022; Wei et al., 2026). This out-of-distribution (OOD) setting requires assumptions that connect observed and unobserved interventions; accurate fitting of the observed perturbation–response pairs alone does not establish such a connection.

Existing approaches include compositional and disentangled autoencoders (Lotfollahi et al., 2023; Hetzel et al., 2022; Piran et al., 2024), graph-based predictors (Roohani et al., 2024), foundation-model-based methods (Cui et al., 2024; Hao et al., 2024), and diffusion-based generators (Chi et al., 2026; Liang et al., 2025). These methods incorporate diverse inductive biases, including compositional representations, biological knowledge graphs, and pretrained embeddings, and have enabled substantial empirical progress. However, their predictive success does not by itself characterize the conditions under which a learned representation supports unseen perturbation prediction. In particular, the relationship between latent perturbation modeling, cross-domain invariance, and target-domain prediction error remains insufficiently understood.

A growing line of work explicitly introduces causal structure into single-cell perturbation modeling (Liu et al., 2026; Lopez et al., 2023; Mao et al., 2024). Lorch et al. (2026) model expression through a stationary diffusion process observed under measurement noise, combining prediction of unseen perturbation combinations with analysis of regulatory effects. Jiang et al. (2026) introduce a latent dynamical causal model and establish conditions under which latent causal variables are identifiable up to an equivalence class. These advances motivate a distinction between whether latent

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.

structure can be recovered and under what conditions its estimation supports prediction in unseen perturbation domains. Identifiability alone does not quantify the prediction error of a learned model under an unseen intervention. We investigate this connection by relating errors in the perturbation-induced latent distribution and conditional response generator to target-domain generation error.

Our starting point is to model perturbations as interventions on a basal latent cell state. Each perturbation induces an additive latent shift, which propagates through a shared structural causal model (SCM), and the resulting state generates transcriptional responses through a domain-invariant conditional generation kernel. Under shared exogenous noise and a fixed linear SCM, these interventions induce a latent location family: perturbations change the mean while preserving the centered distribution and covariance. An unseen perturbation thus corresponds to a new shift within a shared latent mechanism. Predicting that shift remains a learning problem; the structural formulation specifies how it propagates and how its estimation error affects the generated response.

Within this framework, we distinguish three theoretical questions. First, under an injectivity condition on the observation operator induced by the fixed conditional kernel, we establish identifiability of perturbation-induced latent distributions from the corresponding observed response distributions. Identifiability of the structural matrix requires additional assumptions and is a separate result, rather than a prerequisite for all subsequent prediction guarantees. Second, invariance of the domain-invariant conditional generation kernel supports an oracle Bayes rule that is pointwise optimal across perturbation domains (Peters et al., 2016; Rojas-Carulla et al., 2018). This oracle statement does not imply that a predictor fitted only on source-domain data is automatically accurate in unseen domains. Third, for learned representations and generators, we derive a target-domain generation error bound that separates their respective approximation errors. Figure 1 summarizes these components; formal statements and proofs appear in Section 4 and Appendix B.

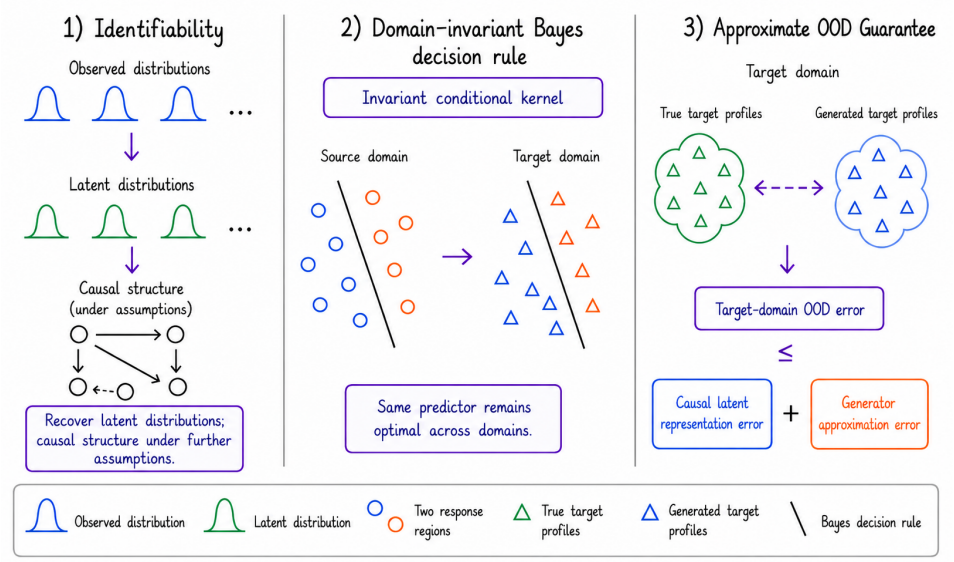

Figure 1: Overview of the latent causal framework underlying scCausalDiff. Under the stated assumptions, the analysis establishes latent-distribution identifiability, characterizes the cross-domain optimality of an oracle conditional Bayes rule, and bounds target-domain generation error through latent representation and conditional generator approximation errors.

Guided by this framework, we introduce **scCausalDiff**. An encoder maps basal cell expression to a latent state, a perturbation encoder predicts an additive intervention shift, and a learned SCM propagates the shifted state. Acyclicity of the structural matrix is enforced through a differentiable constraint optimized jointly with the model parameters (Zheng et al., 2018). The propagated latent state then conditions a diffusion generator that produces post-perturbation expression profiles.

The main contributions of our work are as follows:

- **A theoretical framework for OOD perturbation prediction.** We establish identifiability results for latent distributions and structural parameters, characterize oracle cross-domain optimality, and derive a target-domain generation error bound that separates representation and generator approximation errors.
- **A model with explicit latent causal propagation.** scCausalDiff combines a learned latent intervention shift, propagation through a shared SCM, and conditional diffusion generation within a unified prediction framework.
- **Evaluation under strict OOD perturbation splits.** We assess mean-response accuracy and distributional fidelity on genetic and chemical perturbation benchmarks, obtaining competitive performance across these complementary criteria. Structural replacements, decoder comparisons, and latent-dimension analyses examine the model's empirical behavior, while biological analyses investigate associations between latent response directions and known functional programs.

2 Related Work

2.1 Single-cell perturbation prediction

Single-cell perturbation prediction has been studied through compositional, graph-based, and generative approaches. CPA (Lotfollahi et al., 2023) and ChemCPA (Hetzel et al., 2022) combine basal cell states with latent perturbation shifts, while BioLord (Piran et al., 2024) learns disentangled representations of cell states and experimental attributes. scLAMBDA (Wang et al., 2024a) combines pretrained gene embeddings with a variational generative model that separates basal and perturbation-related representations. Graph-based methods incorporate biological relationships: GEARS (Roohani et al., 2024) uses gene knowledge graphs, and TxPert (Wenkel et al., 2026) integrates multiple biological knowledge graphs to encode perturbations. Foundation models such as scGPT (Cui et al., 2024) and scFoundation (Hao et al., 2024) leverage pretrained single-cell representations. Diffusion-based approaches include scPPDM (Liang et al., 2025) and Doloris (Chi et al., 2026); the latter learns control and perturbed distributions with dual conditional diffusion models and implicitly aligns them through a shared Gaussian space. Our work combines conditional diffusion generation with an explicit latent SCM and studies the conditions under which this decomposition supports OOD perturbation prediction.

2.2 Domain generalization and invariant prediction

OOD perturbation prediction can be formulated as domain generalization, with each perturbation condition defining a domain. Invariant prediction connects transfer across domains to stable conditional mechanisms (Peters et al., 2016; Ahuja et al., 2020). Related work studies mechanism-preserving transformations (Wang et al., 2022), invariant representation learning (Chen et al., 2022; Wang et al., 2024b; Miao et al., 2025), and prediction under interventions (Rojas-Carulla et al., 2018; Magliacane et al., 2018). Conditional invariance can support a common oracle prediction rule, but does not by itself guarantee that a predictor estimated from source data is accurate in unseen domains. We instantiate this distinction for single-cell perturbations: our analysis establishes domain-invariant optimality of an oracle predictor under the stated assumptions, and separately bounds target-domain generation error through latent representation and conditional generator approximation errors.

2.3 Causal representation learning

Causal representation learning studies latent mechanisms underlying observations (Schölkopf et al., 2021). Nonlinear ICA establishes identifiability using auxiliary variables (Hyvarinen et al., 2019; Khemakhem et al., 2020). Subsequent work exploits interventions (Ahuja et al., 2023; von Kügelgen et al., 2023), mechanism sparsity (Lachapelle et al., 2022), and paired observations or augmentations (von Kügelgen et al., 2021). Complementary approaches formulate DAG learning as continuous optimization (Zheng et al., 2018; Ng et al., 2020).

For biological data, SENA-discrepancy-VAE (SENA) (de la Fuente Cedeño et al., 2025) incorporates gene-to-biological-process priors and represents causal latent factors through learned combinations of biological processes. In contrast, scCausalDiff learns structural relations among latent coordinates

without predefined biological assignments and examines their functional associations through post-hoc analysis. Our theoretical analysis focuses on perturbation-induced latent distributions under a fixed conditional generation kernel, distinguishing their identifiability from recovery of the structural matrix. Beyond identifiability, we characterize cross-domain optimality of an oracle prediction rule and derive an OOD generation bound in terms of latent representation and conditional generator approximation errors. This connects the analysis of latent causal structure to the prediction of unseen perturbation responses.

3 Model Instantiation and Learning

Guided by the decomposition of OOD generation error into latent representation and conditional generator approximation terms, scCausalDiff instantiates our framework through control and perturbation encoding, shared causal propagation, and conditional diffusion generation (Figure 2).

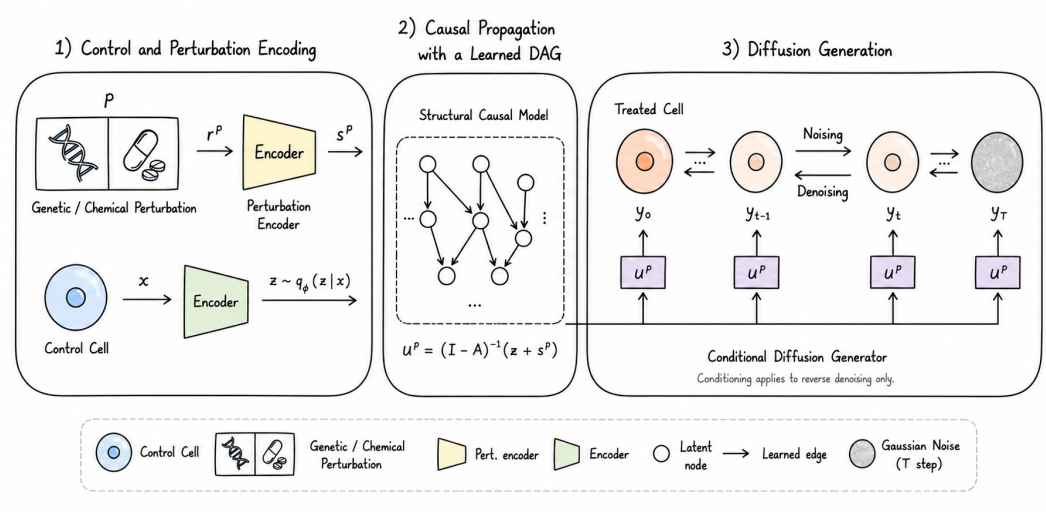

Figure 2: Overview of scCausalDiff. Control expression x and perturbation representation r^p are encoded into a basal state z and an intervention shift s^p, respectively. Their sum is propagated through a shared learned SCM to obtain u^p, which conditions reverse diffusion. Forward noising is used during training; generation starts from Gaussian noise.

Control and perturbation encoding. A variational encoder maps control-cell expression x to a diagonal Gaussian distribution $q_\phi(z \mid x)$, regularized toward a standard normal prior. We sample the basal state z using the reparameterization trick. A perturbation encoder maps the input representation r^p to a latent intervention shift s^p. Their sum $z + s^p$ implements the separation of basal cellular variation and perturbation-dependent intervention in our theoretical framework.

Causal propagation with a learned latent SCM. We compute $u^p = (I - A)^{-1}(z + s^p)$ using a learnable zero-diagonal adjacency matrix A shared across perturbation conditions. Without prescribing a topological order, we optimize the NOTEARS constraint $h(A) = \mathrm{tr}\big(e^{A \odot A}\big) - d = 0$ through an augmented Lagrangian scheme (Zheng et al., 2018), together with sparsity regularization. This implements the shared structural transformation in Section 4.1. Graph nodes denote latent coordinates rather than predefined genes or pathways, and structural identifiability remains conditional on the additional assumptions of Theorem 4.5.

Conditional diffusion generation. During training, forward diffusion corrupts observed perturbed expression y_0 into noisy states y_t. A shared denoising network predicts y_0 from y_t, the timestep t, and the propagated representation u^p. At generation time, u^p conditions iterative reverse denoising from Gaussian noise to produce post-perturbation expression profiles. The shared generator implements the domain-invariant conditional generation kernel in our framework, with perturbation-specific information supplied through u^p.

Optimization. We train the model by minimizing $\mathcal{L} = \mathcal{L}_{\text{rec}} + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}} + \alpha h(A) + \frac{\rho}{2}h(A)^2 + \lambda_1\mathcal{R}_1(A)$, where $\mathcal{L}_{\text{rec}}$ measures denoising reconstruction error and $\mathcal{L}_{\text{KL}}$ regularizes the basal posterior toward a Gaussian prior. The remaining terms encourage acyclicity and sparsity. Full definitions and augmented Lagrangian updates are provided in Appendix A.4.

4 Theoretical Guarantees on Identifiability and OOD Generalization

This section develops a population-level analysis of the latent structural framework underlying the model in Section 3. We first establish identifiability of perturbation-induced latent distributions and, under additional structural assumptions, of the latent structural matrix. We then characterize the cross-domain optimality of an oracle Bayes decision rule under a shared conditional generation kernel. Finally, we derive a target-domain generation error bound that separates the contributions of latent representation and conditional generator approximation errors. Together, these results connect latent identifiability, oracle prediction, and approximate OOD generation within a unified framework. All proofs are deferred to Appendix B.

4.1 Model setup

Let $\mathcal{P}$ be a finite or countable index set of perturbation domains, including the control domain and genetic or chemical perturbation conditions. For each domain $p \in \mathcal{P}$, we observe data $(X, Y^p) \in \mathcal{X} \times \mathcal{Y}$, where X denotes the basal or control-state expression profile and Y^p denotes the post-perturbation response under domain p. We consider a generalization setting in which data are available from a subset of source perturbation domains $\mathcal{P}_{\text{src}} \subseteq \mathcal{P}$, and the goal is to generalize to unseen target perturbation domains $\mathcal{P}_{\text{tgt}} \subseteq \mathcal{P}$.

We assume that the domain-specific response Y^p is generated through a causal latent representation. Let $\mathcal{U} := \mathbb{R}^d$ denote the latent representation space. At the population level, let $Z \in \mathbb{R}^d$ be a latent exogenous variable inferred from the basal state, whose distribution is shared across perturbation domains. Each domain p is associated with a deterministic perturbation vector $s^p \in \mathbb{R}^d$, which represents a perturbation-induced additive shift in latent space.

We consider the following population-level structural conditional generation model. There exists a shared structural matrix $A \in \mathbb{R}^{d\times d}$ that is DAG-compatible. Hence $I - A$ is invertible, and we write $B := (I - A)^{-1}$. For each perturbation domain $p \in \mathcal{P}$, the perturbation-shifted latent state is propagated through the latent SCM as $U^p = B(Z + s^p) = (I - A)^{-1}(Z + s^p)$. We refer to U^p as the causal latent representation under perturbation domain p. Moreover, the observed response is generated from U^p through a domain-invariant conditional generation kernel: there exists a conditional kernel $K(\cdot \mid u)$, representing the population-level counterpart of the conditional diffusion generator, such that, for every $p \in \mathcal{P}$, $Y^p \mid (U^p = u, X = x) \sim K(\cdot \mid u)$. Thus, the perturbation domain affects Y^p only through the causal latent representation U^p, while both the structural transformation B and the conditional generation kernel K are shared across domains.

4.2 Identifiability of U^p and A

We first establish conditions under which observed response distributions identify perturbation-induced latent distributions. We then study recovery of the structural matrix A under additional triangularity and independence assumptions.

Definition 4.1 (Identifiability of $\{\mathcal{F}(U^p)\}_{p\in\mathcal{P}}$)**.** *Suppose the model in Section 4.1 holds, with $U^p = (I - A)^{-1}(Z + s^p)$ and $Y^p \mid U^p = u \sim K(\cdot \mid u)$ for $p \in \mathcal{P}$. We say that the family $\{\mathcal{F}(U^p)\}_{p\in\mathcal{P}}$ is identifiable from $\{\mathcal{F}(Y^p)\}_{p\in\mathcal{P}}$ if the following implication holds.*

For any alternative parameterization $(\widetilde{A}, \widetilde{Z}, \widetilde{s}^p)$ satisfying the same structural model, with a domain-shared exogenous distribution $\widetilde{Z} \sim \mathcal{N}(\widetilde{\mu}_Z, \widetilde{\Sigma}_Z)$, $\widetilde{\Sigma}_Z \succ 0$, and the same conditional kernel K and latent coordinate space, and inducing

$$\widetilde{U}^p = (I - \widetilde{A})^{-1}(\widetilde{Z} + \widetilde{s}^p), \qquad \widetilde{Y}^p \mid \widetilde{U}^p = u \sim K(\cdot \mid u),$$

the condition $\mathcal{F}(\widetilde{Y}^p) = \mathcal{F}(Y^p), \forall p \in \mathcal{P}$, implies $\mathcal{F}(\widetilde{U}^p) = \mathcal{F}(U^p), \forall p \in \mathcal{P}$.

Assumption 4.2 (Shared latent exogenous distribution)**.** *The latent exogenous variable $Z \in \mathbb{R}^d$ satisfies $Z \sim \mathcal{N}(\mu_Z, \Sigma_Z)$, with $\mu_Z \in \mathbb{R}^d$ and $\Sigma_Z \succ 0$. Its distribution is shared across perturbation domains.*

Assumption 4.3 (Injectivity of the observation operator on Gaussian measures)**.** *Define the observation operator $\Phi_Q(\cdot) := \int K(\cdot \mid u)\, dQ(u)$.*

Assume that Φ is injective on the class of nondegenerate Gaussian measures on $\mathbb{R}^d$, i.e., if $Q, Q' \in \mathcal{G} := \{\mathcal{N}(m, \Sigma) : m \in \mathbb{R}^d, \Sigma \succ 0\}$ and $\Phi_Q(\cdot) = \Phi_{Q'}(\cdot)$, then necessarily $Q = Q'$.

The injectivity assumption ensures that distinct Gaussian latent distributions remain distinguishable under the shared conditional generation kernel.

Theorem 4.4 (Identifiability of $\mathcal{F}(U^p)$ without identifiability of A)**.** *Suppose the model in Section 4.1 holds. Under Assumptions 4.2 and 4.3, the family of perturbation-induced causal latent distributions $\{\mathcal{F}(U^p) : p \in \mathcal{P}\}$ is identifiable from the family of observed response distributions $\{\mathcal{F}(Y^p) : p \in \mathcal{P}\}$. Moreover, for each $p \in \mathcal{P}$, $\mathcal{F}(U^p) = \mathcal{N}(m_p, \Sigma_U)$, where*

$$m_p = B(\mu_Z + s^p), \qquad \Sigma_U = B\Sigma_Z B^\top.$$

In particular, the covariance Σ_U is invariant across perturbation domains, while the perturbation-specific variation is entirely encoded by the mean shift

$$m_p - m_{p'} = B(s^p - s^{p'}), \qquad p, p' \in \mathcal{P}.$$

Theorem 4.4 establishes that, under the injectivity of the observation operator on the Gaussian latent class, the family of perturbation-induced causal latent distributions $\{\mathcal{F}(U^p)\}_{p\in\mathcal{P}}$ is uniquely determined by the observed response distributions (proof in Appendix B.1). This is a distributional identifiability result: it identifies the domain-specific causal latent laws, but it does not by itself imply that the underlying structural matrix A is identifiable.

To move from distributional identifiability to structural identifiability, one needs additional restrictions on the latent structural equation. In particular, if the coordinates of the latent exogenous noise are mutually independent and the structural matrix A is strictly lower triangular, then the covariance Σ_U admits a unique triangular structural factorization. The following result shows that, under these additional assumptions, the identifiable covariance Σ_U determines $B = (I - A)^{-1}$ and hence A.

Theorem 4.5 (Identifiability of A from the identifiable distribution of U)**.** *Let $U = (I - A)^{-1}(Z + s^p) = B(Z + s^p), B := (I - A)^{-1}$, where $A \in \mathbb{R}^{d\times d}$ and $Z, s^p \in \mathbb{R}^d$. Assume that:*

1. *A is strictly lower triangular under a fixed topological order, i.e., $A_{ii} = 0$ for all i and $A_{ij} = 0$ whenever $i \le j$;*
2. *the coordinates of Z are mutually independent with mean zero and unknown variances, so that $\Sigma_Z = \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d), \lambda_i > 0$;*
3. *the distribution of U is identifiable.*

Then A is identifiable. More precisely, both Λ and A are uniquely determined by Σ_U.

Remark 4.6. *Every DAG admits a permutation matrix P such that $A_P = PAP^\top$ is strictly lower triangular. Writing $\Lambda_P = P\Lambda P^\top$, the covariance transforms as*

$$P\Sigma_U P^\top = (I - A_P)^{-1}\Lambda_P(I - A_P)^{-\top}.$$

For a fixed ordering P, the triangular factorization uniquely determines A_P and Λ_P, and hence $A = P^\top A_P P$. When the ordering is unknown, however, different orderings may yield distinct structural matrices compatible with the same covariance. Fixed-order uniqueness therefore does not imply uniqueness up to node relabeling without additional restrictions. The subsequent OOD results do not require structural identifiability.

Remark 4.7. *If the coordinates of Z are mutually independent and identically distributed as $\mathcal{N}(0, \sigma^2)$, i.e. $\Sigma_Z = \Lambda = \sigma^2 I$, then A is identifiable regardless of whether it is strictly lower triangular or not (Peters & Bühlmann, 2014).*

The proof of Theorem 4.5 and a discussion of Remark 4.6 are provided in Appendix B.2.

4.3 DOMAIN-INVARIANT BAYES DECISION RULE

The preceding results concern identifiability of the causal latent distribution $\mathcal{F}(U^p)$ and the latent structural matrix A. We next turn to a decision-theoretic implication of the invariant conditional generation assumption. Since the conditional law of Y^p given U^p is shared across domains, the latent conditional risk induced by the source domains is not specific to the source domains themselves. Instead, it defines a domain-invariant Bayes decision rule on the causal latent representation. The same source-induced rule can therefore be evaluated in unseen target perturbation domains and remains Bayes-optimal there at the oracle level.

Definition 4.8 (Source-induced latent conditional risk)**.** *Let $L : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$ be a measurable loss. Define the source-induced latent conditional risk by $R_{\mathrm{src}}(u; w) := \int_{\mathcal{Y}} L(w, y) K(dy \mid u)$, for $(u, w) \in \mathcal{U} \times \mathcal{Y}$. Suppose the model in Section 4.1 holds, for every source domain $p \in \mathcal{P}_{\mathrm{src}}$, this risk coincides with the conditional risk*

$$R_{\mathrm{src}}(u; w) = \mathbb{E}[L(w, Y^p) \mid U^p = u],$$

whenever the conditional expectation is well defined. Assume that the pointwise minimizer admits a measurable selection on $\mathcal{U}$. A measurable function $h^\star_{\mathrm{src}} : \mathcal{U} \to \mathcal{Y}$ is called a source-induced oracle Bayes decision rule if

$$h^\star_{\mathrm{src}}(u) \in \arg\min_{w \in \mathcal{Y}} R_{\mathrm{src}}(u; w), \qquad \text{for every } u \in \mathcal{U}.$$

We assume that this oracle rule has finite expected loss in every domain under consideration.

Theorem 4.9 (Domain-invariant Bayes decision rule)**.** *Suppose the model in Section 4.1 holds, and let $h^\star_{\mathrm{src}}$ be a source-induced oracle Bayes decision rule in the sense of Definition 4.8. Then, for every target domain $p \in \mathcal{P}_{\mathrm{tgt}}$ and every measurable decision rule $h : \mathcal{U} \to \mathcal{Y}$,*

$$\mathbb{E}[L(h^\star_{\mathrm{src}}(U^p), Y^p)] \le \mathbb{E}[L(h(U^p), Y^p)].$$

Theorem 4.9 formalizes a source-to-target consequence of the domain-invariant conditional generation kernel. The rule $h^\star_{\mathrm{src}}$ is defined through the shared latent conditional risk by minimizing the expected prediction loss at each latent state; however, because this risk is determined by the domain-shared kernel $K(\cdot \mid u)$, the same rule remains Bayes-optimal in unseen target perturbation domains. Thus, the causal latent representation does not merely provide a domain-specific predictor; it supports a domain-invariant Bayes decision rule that can be transferred from source perturbation domains to target perturbation domains at the oracle level. The proof is given in Appendix B.3.

4.4 APPROXIMATE OOD GUARANTEE UNDER LEARNED REPRESENTATIONS

We now consider the practically relevant setting in which the oracle causal latent representation U^p and the invariant conditional kernel $K(\cdot \mid u)$ are replaced by learned approximations. Let $\hat{U}^p \in \mathcal{U}$ denote the learned approximation of U^p in domain p, and let $Q_{\mathrm{gen}}(\cdot \mid \hat{u})$ denote the learned conditional generator given the learned representation $\hat{u} \in \mathcal{U}$.

For a target domain $p \in \mathcal{P}_{\mathrm{tgt}}$, the true conditional distribution of Y^p given $X = x$ is $P_{Y^p|X}(\cdot \mid x) = \int_{\mathcal{U}} K(\cdot \mid u)\, P_{U^p|X}(du \mid x)$, whereas the learned conditional distribution induced by the generator is $P^{Q_{\mathrm{gen}}}_{Y^p|X}(\cdot \mid x) = \int_{\mathcal{U}} Q_{\mathrm{gen}}(\cdot \mid \hat{u})\, P_{\hat{U}^p|X}(d\hat{u} \mid x)$. Let dist be a discrepancy between probability measures on $\mathcal{Y}$. The following result shows that, under Assumption B.1, the target-domain conditional generation error is controlled by the learned representation error and the learned generator error.

Theorem 4.10 (Approximate OOD conditional generation)**.** *Suppose the model in Section 4.1 and Assumption B.1 hold. Then for every target domain $p \in \mathcal{P}_{\mathrm{tgt}}$,*

$$\mathbb{E}_X\Big[\mathrm{dist}\Big(P_{Y^p|X}(\cdot \mid X), P^{Q_{\mathrm{gen}}}_{Y^p|X}(\cdot \mid X)\Big)\Big] \le \omega(\delta_U) + \delta_Q.$$

In particular, if $\omega(r) = Kr$ for some $K > 0$, then

$$\mathbb{E}_X\Big[\mathrm{dist}\Big(P_{Y^p|X}(\cdot \mid X), P^{Q_{\mathrm{gen}}}_{Y^p|X}(\cdot \mid X)\Big)\Big] \le K\delta_U + \delta_Q.$$

Theorem 4.10 bounds target-domain conditional generation error by $\omega(\delta_U) + \delta_Q$, where the two terms capture latent representation error and conditional generator approximation error, respectively. The bound thus separates their contributions to the generation error guarantee. The proof is provided in Appendix B.4.

Under the stated assumptions, these population-level results establish identifiability of perturbation-induced latent distributions and, with additional structural restrictions, of the latent SCM. They also establish cross-domain optimality of an oracle Bayes rule and quantify how learned approximations affect target-domain generation.

5 Experiments

5.1 Experimental Setup

Datasets. We evaluate scCausalDiff on Replogle (Replogle et al., 2022) for single-gene perturbations, Norman (Norman et al., 2019) for two-gene combinatorial perturbations, and sci-Plex3 (Srivatsan et al., 2020) for chemical perturbations. Dataset descriptions, preprocessing, and OOD split protocols are provided in Appendix C.1.

Baselines. We compare scCausalDiff with representative perturbation prediction methods, including SCCVAE (Liu et al., 2026), GEARS, TxPert, scLAMBDA, SENA-discrepancy-VAE, Doloris, BioLord, and ChemCPA, using task-appropriate baselines for each dataset. Implementation details and task-specific adaptations are provided in Appendix C.3.

Evaluation. We evaluate three complementary aspects of prediction: mean-response accuracy (Δ-P and Δ-S), perturbation discrimination (PDS-cos and PDS-L1), and distributional fidelity (MMD and energy distance). We report means and standard deviations over five folds. Metric definitions and evaluation procedures are provided in Appendix C.4.

5.2 Results

OOD perturbation prediction. Table 1 shows that scCausalDiff achieves the best mean score on five of six metrics on Replogle and four of six on sci-Plex3. It combines strong mean-response accuracy with perturbation discrimination and distributional fidelity, while the relative performance of competing methods varies across these criteria. Results on Norman further demonstrate its distributional performance under combinatorial perturbations (Appendix D.1). Additional metric analyses are provided in Appendix D.2.

Structural and decoder ablations. Replacing the learned latent structure with alternative transformations reduces mean-response accuracy, indicating the fitted model's dependence on its learned propagation map. The diffusion-based variant achieves better predictive accuracy and distributional fidelity than the tested VAE-based variant. Detailed ablations and latent-dimension analyses are provided in Appendix D.3.

Biological associations of latent responses. Propagated perturbation representations exhibit associations with curated functional classes (Figure 3). Complementary output-gene analyses link latent directions to programs including translation, RNA processing, and stress response. Propagation concentrates perturbation-response variation, while a held-out *WDR3* case illustrates correspondence with observed mean responses. These analyses characterize functional associations rather than establish gene-regulatory interactions. Details are provided in Appendix D.4.

Empirical diagnostics of theoretical assumptions. Diagnostics of the learned basal representation show approximate marginal Gaussianity across most coordinates and weak pairwise dependence. Restricted Jacobian analyses further indicate local denoiser sensitivity to latent variations. Details are provided in Appendix D.5.

Table 1: OOD perturbation prediction on Replogle and sci-Plex3. Mean$_{\pm\text{std}}$ over five folds; best in bold. Metrics assess mean-response accuracy, perturbation discrimination, and distributional fidelity. All methods use the same control reference within each fold.

Model	Mean shift		Discrimination		Distribution	
	Δ-P ↑	Δ-S ↑	PDS-cos ↑	PDS-L1 ↑	MMD ↓	E-dist ↓
Gene perturbation prediction (Replogle, 5-fold OOD)						
SENA	$0.3594_{\pm 0.037}$	$0.3545_{\pm 0.039}$	$0.4951_{\pm 0.013}$	$0.5033_{\pm 0.005}$	$0.0616_{\pm 0.003}$	$1.7758_{\pm 0.097}$
Doloris	$0.3984_{\pm 0.018}$	$0.3767_{\pm 0.015}$	$0.5241_{\pm 0.014}$	$0.5100_{\pm 0.009}$	$0.0430_{\pm 0.006}$	$1.3176_{\pm 0.173}$
SCCVAE	$0.4485_{\pm 0.023}$	$0.4271_{\pm 0.015}$	$0.5189_{\pm 0.022}$	$0.5104_{\pm 0.016}$	$0.0371_{\pm 0.002}$	$1.1665_{\pm 0.062}$
GEARS	$0.5904_{\pm 0.098}$	$0.5605_{\pm 0.093}$	$0.7543_{\pm 0.137}$	$0.7337_{\pm 0.083}$	$0.0336_{\pm 0.031}$	$1.1118_{\pm 0.875}$
TxPert	$0.6732_{\pm 0.023}$	$0.6541_{\pm 0.020}$	$0.7691_{\pm 0.029}$	$0.7254_{\pm 0.016}$	$0.0397_{\pm 0.003}$	$2.7793_{\pm 0.127}$
scLAMBDA	$0.6827_{\pm 0.036}$	$0.6779_{\pm 0.029}$	$0.7686_{\pm 0.043}$	$0.7354_{\pm 0.021}$	$\mathbf{0.0223}_{\pm 0.003}$	$0.7846_{\pm 0.085}$
Ours	$\mathbf{0.6893}_{\pm 0.037}$	$\mathbf{0.6814}_{\pm 0.031}$	$\mathbf{0.7834}_{\pm 0.039}$	$\mathbf{0.7689}_{\pm 0.025}$	$0.0232_{\pm 0.003}$	$\mathbf{0.7606}_{\pm 0.079}$
Drug perturbation prediction (sci-Plex3, 5-fold OOD)						
Doloris	$0.2927_{\pm 0.127}$	$0.2350_{\pm 0.076}$	$0.5298_{\pm 0.020}$	$0.5207_{\pm 0.015}$	$0.0610_{\pm 0.046}$	$1.7185_{\pm 1.417}$
SCCVAE	$0.3953_{\pm 0.018}$	$0.2597_{\pm 0.006}$	$0.5055_{\pm 0.007}$	$0.5055_{\pm 0.005}$	$0.0264_{\pm 0.004}$	$1.4789_{\pm 0.181}$
ChemCPA	$0.5240_{\pm 0.014}$	$0.3671_{\pm 0.008}$	$0.5276_{\pm 0.016}$	$0.5049_{\pm 0.002}$	$0.0242_{\pm 0.002}$	$1.4487_{\pm 0.095}$
scLAMBDA	$0.5904_{\pm 0.008}$	$\mathbf{0.5364}_{\pm 0.003}$	$0.5711_{\pm 0.017}$	$\mathbf{0.5300}_{\pm 0.012}$	$0.0163_{\pm 0.001}$	$0.5873_{\pm 0.033}$
BioLord	$0.5932_{\pm 0.007}$	$0.4757_{\pm 0.011}$	$0.5405_{\pm 0.020}$	$0.5074_{\pm 0.004}$	$0.0391_{\pm 0.002}$	$4.2969_{\pm 0.065}$
Ours	$\mathbf{0.6270}_{\pm 0.017}$	$0.5187_{\pm 0.006}$	$\mathbf{0.5718}_{\pm 0.014}$	$0.5292_{\pm 0.010}$	$\mathbf{0.0110}_{\pm 0.002}$	$\mathbf{0.3587}_{\pm 0.050}$

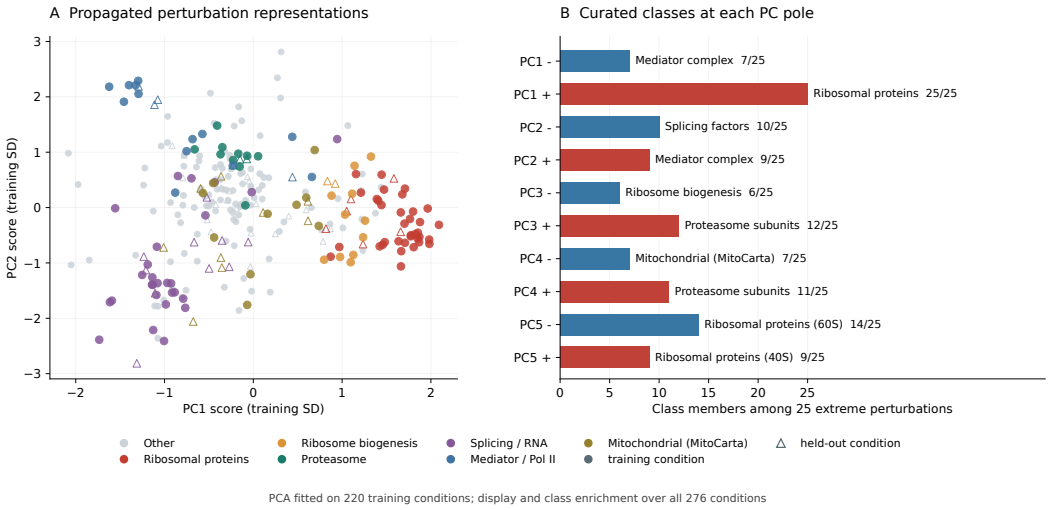

Figure 3: Biological associations of propagated representations (Replogle, fold 1, $d = 1024$). (A) Scores on the first two propagated PCs for training conditions (filled circles) and held-out conditions (open triangles), coloured by functional class. (B) The most enriched class among the 25 most extreme conditions at each pole of the first five PCs, with class-member counts. Enrichment is assessed using one-sided Fisher tests with Benjamini–Hochberg correction within each pole. PCA is fitted on 220 training conditions; display and class enrichment use all 276 conditions.

6 Conclusion

We introduced scCausalDiff, combining latent interventions, shared structural propagation, and conditional diffusion for unseen perturbation prediction. Under explicit assumptions, our theory distinguishes identifiability from oracle cross-domain optimality and bounds generation error through representation and generator approximation errors. Across three benchmarks, scCausalDiff combines competitive mean-response prediction with strong distributional performance on chemical and combinatorial perturbations. These findings motivate joint evaluation of mean effects and cell-population distributions, while biological analyses associate propagated responses with functional programs. Future work will incorporate biological priors into structure learning and develop cell-state-dependent interventions to improve generalization across cellular contexts.

AI USE STATEMENT

In this work, we used generative AI tools, including ChatGPT, to assist with three types of tasks: language editing and polishing of the manuscript; literature retrieval and discovery of potentially relevant prior work; and partial mathematical derivation, proof development, and proof checking.

All AI-assisted work was reviewed and verified by the authors. Mathematical arguments and theoretical claims were independently checked by the authors, and bibliographic information and citation-related claims were checked against the relevant source materials. AI-assisted language was reviewed and revised by the authors before inclusion in the manuscript. Synthetic datasets used in the experiments were generated by the specified data-generating processes and computational code, rather than by generative AI.

Generative AI tools did not make final scientific decisions or replace author verification of mathematical claims, experimental results, or cited sources. We take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

The model architecture, diffusion process, and training objective are described in Appendix A. The theoretical assumptions and proofs are provided in Appendix B. Dataset preprocessing and perturbation splits, implementation settings, baseline configurations, and evaluation metrics are documented in Appendices C.1, C.2, C.3, and C.4, respectively. Additional ablations, biological analyses, and empirical diagnostics are reported in Appendix D.

REFERENCES

Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar. Invariant risk minimization games. In *International Conference on Machine Learning*, pp. 145–155. PMLR, 2020.

Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *International Conference on Machine Learning*, pp. 372–407. PMLR, 2023.

Michael Ashburner et al. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1): 25–29, 2000. doi: 10.1038/75556.

Yiqun T. Chen and James Zou. Simple and effective embedding model for single-cell biology built from ChatGPT. *Nature Biomedical Engineering*, 9(4):483–493, 2025. doi: 10.1038/s41551-024-01284-6.

Yongqiang Chen, Yonggang Zhang, Yatao Bian, Han Yang, Kaili Ma, Binghui Xie, Tongliang Liu, Bo Han, and James Cheng. Learning causally invariant representations for out-of-distribution generalization on graphs. *Advances in Neural Information Processing Systems*, 35:22131–22148, 2022.

Changxi Chi, Jun Xia, Yufei Huang, Zhuoli Ouyang, Cheng Tan, Yunfan Liu, Jingbo Zhou, Chang Yu, Liangyu Yuan, Siyuan Li, Zelin Zang, and Stan Z. Li. Doloris: Dual conditional diffusion implicit bridges with sparsity masking strategy for unpaired single-cell perturbation estimation. In *International Conference on Learning Representations*, 2026.

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*, 21(8):1470–1480, 2024.

Jesus de la Fuente Cedeño, Robert Lehmann, Carlos Ruiz-Arenas, Jan Voges, Irene Marín-Goñi, Xabier Martinez de Morentin, David Gomez-Cabrero, Idoia Ochoa, Jesper Tegnér, Vincenzo Lagani, and Mikel Hernaez. Interpretable causal representation learning for biological data in the pathway space. In *The Thirteenth International Conference on Learning Representations*, 2025.

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*, 167(7): 1853–1866, 2016.

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*, 21(8):1481–1491, 2024.

Leon Hetzel, Simon Boehm, Niki Kilbertus, Stephan Günnemann, Fabian Theis, et al. Predicting cellular responses to novel drug perturbations at a single-cell resolution. *Advances in Neural Information Processing Systems*, 35:26711–26722, 2022.

Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 859–868. PMLR, 2019.

Wenkang Jiang, Yuhang Liu, Erdun Gao, Ehsan Abbasnejad, Lina Yao, and Javen Qinfeng Shi. Learning latent dynamical causal processes for single-cell perturbation prediction. *arXiv preprint arXiv:2605.25581*, 2026.

Minoru Kanehisa, Miho Furumichi, Yoko Sato, Yuriko Matsuura, and Mari Ishiguro-Watanabe. KEGG: biological systems database as a model of the real world. *Nucleic Acids Research*, 53(D1): D672–D677, 2025. doi: 10.1093/nar/gkae909.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pp. 2207–2217. PMLR, 2020.

Liis Kolberg, Uku Raudvere, Ivan Kuzmin, Priit Adler, Jaak Vilo, and Hedi Peterson. g:Profiler: interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). *Nucleic Acids Research*, 51(W1):W207–W212, 2023. doi: 10.1093/nar/gkad347.

Sébastien Lachapelle, Pau Rodriguez, Yash Sharma, Katie E Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ICA. In *Conference on Causal Learning and Reasoning*, pp. 428–484. PMLR, 2022.

Zhaokang Liang, Shuyang Zhuang, Xiaoran Jiao, Weian Mao, Hao Chen, and Chunhua Shen. scPPDM: A diffusion model for single-cell drug-response prediction. *arXiv preprint arXiv:2510.11726*, 2025.

Emily Liu, Jiaqi Zhang, and Caroline Uhler. Learning genetic perturbation effects with variational causal inference. *PLOS Computational Biology*, 22(2):e1013194, 2026. doi: 10.1371/journal.pcbi.1013194.

Romain Lopez, Natasa Tagasovska, Stephen Ra, Kyunghyun Cho, Jonathan Pritchard, and Aviv Regev. Learning causal representations of single cells via sparse mechanism shift modeling. In *Conference on Causal Learning and Reasoning*, pp. 662–691. PMLR, 2023.

Lars Lorch, Jiaqi Zhang, Charlotte Bunne, Andreas Krause, Bernhard Schölkopf, and Caroline Uhler. Latent causal diffusions for single-cell perturbation modeling. *arXiv preprint arXiv:2601.15341*, 2026.

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*, 19(6):e11517, 2023.

Sara Magliacane, Thijs Van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, and Joris M Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. *Advances in Neural Information Processing Systems*, 31, 2018.

Haiyi Mao, Romain Lopez, Kai Liu, Jan-Christian Hütter, David Richmond, Panayiotis V Benos, and Lin Qiu. Learning identifiable factorized causal representations of cellular responses. *Advances in Neural Information Processing Systems*, 37:121630–121669, 2024.

Qiaowei Miao, Yawei Luo, and Yi Yang. DICS: Find domain-invariant and class-specific features for out-of-distribution generalization. In *ICASSP 2025 – 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Ignavier Ng, AmirEmad Ghassami, and Kun Zhang. On the role of sparsity and DAG constraints for learning linear DAGs. *Advances in Neural Information Processing Systems*, 33:17943–17954, 2020.

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*, 365(6455):786–793, 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 *International Conference on Machine Learning*, 2026.

J. Peters and P. Bühlmann. Identifiability of gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2014.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):947–1012, 2016.

Zoe Piran, Niv Cohen, Yedid Hoshen, and Mor Nitzan. Disentanglement of single-cell data with biolord. *Nature Biotechnology*, 42(11):1678–1683, 2024.

Eliot Ragueneau, Chuqiao Gong, Pierre Sinquin, et al. The Reactome knowledgebase 2026. *Nucleic Acids Research*, 54(D1):D673–D681, 2026. doi: 10.1093/nar/gkaf1223.

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*, 185(14):2559–2575, 2022.

Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *Journal of Machine Learning Research*, 19(36):1–34, 2018.

Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nature Biotechnology*, 42(6):927–935, 2024.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Sanjay R Srivatsan, José L McFaline-Figueroa, Vijay Ramani, Lauren Saunders, Junyue Cao, Jonathan Packer, Hannah A Pliner, Dana L Jackson, Riza M Daza, Lena Christiansen, et al. Massively multiplex chemical transcriptomics at single-cell resolution. *Science*, 367(6473):45–51, 2020.

The Gene Ontology Consortium. The Gene Ontology knowledgebase in 2026. *Nucleic Acids Research*, 54(D1):D1779–D1792, 2026. doi: 10.1093/nar/gkaf1292.

Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. *Advances in Neural Information Processing Systems*, 34:16451–16467, 2021.

Julius von Kügelgen, Michel Besserve, Liang Wendong, Luigi Gresele, Armin Kekić, Elias Bareinboim, David Blei, and Bernhard Schölkopf. Nonparametric identifiability of causal representations from unknown interventions. *Advances in Neural Information Processing Systems*, 36:48603–48638, 2023.

Gefei Wang, Tianyu Liu, Jia Zhao, Youshu Cheng, and Hongyu Zhao. Modeling and predicting single-cell multi-gene perturbation responses with scLAMBDA. *bioRxiv*, 2024a. doi: 10.1101/2024.12.04.626878.

Ruoyu Wang, Mingyang Yi, Zhitang Chen, and Shengyu Zhu. Out-of-distribution generalization with causal invariant transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 375–385, 2022.

Yujie Wang, Kui Yu, Guodu Xiang, Fuyuan Cao, and Jiye Liang. Discovering causally invariant features for out-of-distribution generalization. *Pattern Recognition*, 150:110338, 2024b.

Zhiting Wei, Yiheng Wang, Yicheng Gao, Shuguang Wang, Ping Li, Duanmiao Si, Yuli Gao, Siqi Wu, Danlu Li, Kejing Dong, et al. Benchmarking algorithms for generalizable single-cell perturbation response prediction. *Nature Methods*, 23(2):451–464, 2026. doi: 10.1038/s41592-025-02980-0.

Frederik Wenkel, Wilson Tu, Cassandra Masschelein, Hamed Shirzad, Liam Hodgson, Ihab Bendidi, Cian Eastwood, Shawn T. Whitfield, Craig Russell, Yassir El Mesbahi, Jiarui Ding, Marta M. Fay, Berton Earnshaw, Emmanuel Noutahi, and Alisandra K. Denton. TxPert: using multiple knowledge graphs for prediction of transcriptomic perturbation effects. *Nature Biotechnology*, 2026. doi: 10.1038/s41587-026-03113-4.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. DAGs with no tears: Continuous optimization for structure learning. *Advances in Neural Information Processing Systems*, 31, 2018.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-Mol: A universal 3d molecular representation learning framework. In *International Conference on Learning Representations*, 2023.

A Additional Method Details

This section provides additional details of the scCausalDiff architecture, following the three modules shown in Figure 2: control and perturbation encoding, causal propagation with a learned DAG, and diffusion generation. Optimization settings, loss weights, and training hyperparameters are provided in Appendix C.2.

A.1 Control and Perturbation Encoding

Given a control cell expression profile $x \in \mathbb{R}^G$, scCausalDiff uses a variational encoder to infer a control-cell latent distribution:

$$q_\phi(z \mid x) = \mathcal{N}\big(\mu_\phi(x), \operatorname{diag}\left(\sigma_\phi^2(x)\right)\big), \tag{1}$$

where $\mu_\phi(x) \in \mathbb{R}^d$ and $\sigma_\phi^2(x) \in \mathbb{R}^d$ are the encoder-predicted mean and diagonal variance. A control latent state is sampled from this posterior:

$$z \sim q_\phi(z \mid x). \tag{2}$$

During training, the sampling step is implemented using the reparameterization trick:

$$z = \mu_\phi(x) + \sigma_\phi(x) \odot \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I). \tag{3}$$

The sampled latent state z represents the basal state of the control cell before perturbation.

For a perturbation condition p, let $r^p \in \mathbb{R}^D$ denote its observed representation, such as a genetic perturbation embedding or a chemical perturbation representation. A perturbation embedding module maps r^p to a latent intervention shift:

$$s^p = \gamma g_\psi(r^p), \qquad s^p \in \mathbb{R}^d, \tag{4}$$

where g_ψ is the perturbation encoder and γ controls the magnitude of the perturbation effect. The control latent state and perturbation shift are combined additively:

$$\tilde{z}^p = z + s^p. \tag{5}$$

The resulting $\tilde{z}^p$ serves as the perturbation-shifted exogenous input to the causal propagation module.

This encoding formulation is shared across genetic and chemical perturbation prediction. For genetic perturbations, r^p is a fixed representation of the targeted gene. For chemical perturbations, r^p contains molecular and dosage information. In both cases, the perturbation encoder transforms the observed perturbation representation into a latent intervention shift in the same space as the control-cell latent state.

A.2 Causal Propagation with a Learned Latent SCM

The perturbation-shifted latent state $\tilde{z}^p$ is propagated through the learned SCM. The causal latent representation $u^p \in \mathbb{R}^d$ satisfies the linear structural equation

$$u^p = Au^p + \tilde{z}^p, \tag{6}$$

where $A \in \mathbb{R}^{d\times d}$ is a learnable weighted adjacency matrix over latent dimensions. Assuming $I - A$ is invertible, Eq. 6 admits the closed-form solution

$$u^p = (I - A)^{-1}\tilde{z}^p = (I - A)^{-1}(z + s^p). \tag{7}$$

Thus, the perturbation first modifies the exogenous latent input and then propagates through the learned latent graph before conditioning the expression generator.

To constrain A to a directed acyclic graph we use the differentiable NOTEARS formulation

$$h(A) = \mathrm{tr}\big(e^{A\odot A}\big) - d, \tag{8}$$

where $\odot$ denotes element-wise multiplication and $h(A) = 0$ if and only if A is acyclic, optimized with an augmented Lagrangian scheme. We initialize A at the empty graph, so that $(I - A)^{-1}$ starts at the identity and the intervention shift initially reaches the generator unpropagated. An ℓ_1 penalty on A is applied throughout training to discourage unnecessary edges. The propagated representation u^p then conditions the diffusion generator.

A.3 Diffusion Generation

The causal latent representation u^p conditions a diffusion generator that models the post-perturbation expression distribution. For a fixed perturbation condition, we write y instead of y^p for notational simplicity.

The forward diffusion process gradually corrupts the clean expression profile y_0 by adding Gaussian noise:

$$q(y_t \mid y_0) = \mathcal{N}\left(y_t; \sqrt{\bar{\alpha}_t}y_0, (1 - \bar{\alpha}_t)I\right), \tag{9}$$

or equivalently,

$$y_t = \sqrt{\bar{\alpha}_t}y_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \qquad \epsilon \sim \mathcal{N}(0, I). \tag{10}$$

Here $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$ and $\alpha_t = 1 - \beta_t$, where $\{\beta_t\}_{t=1}^{T}$ is the variance schedule.

The denoising network predicts the clean expression profile from the noisy profile, timestep, and causal latent representation:

$$\hat{y}_0 = D_\theta(y_t, t, u^p). \tag{11}$$

In our implementation the timestep is encoded with a sinusoidal embedding followed by a multilayer projection. Since the model width is set equal to the latent dimension d, the causal latent representation u^p conditions the denoiser by direct addition to the block input, without a learned projection.

At inference time, generation starts from Gaussian noise $y_T \sim \mathcal{N}(0, I)$ and proceeds by iterative reverse diffusion. Given the prediction $\hat{y}_0 = D_\theta(y_t, t, u^p)$, the posterior mean is computed as

$$\mu_\theta(y_t, t, u^p) = c_{1,t}\hat{y}_0 + c_{2,t}y_t, \tag{12}$$

where $c_{1,t}$ and $c_{2,t}$ are determined by the diffusion noise schedule. The reverse update is

$$y_{t-1} = \mu_\theta(y_t, t, u^p) + \sigma_t\eta, \qquad \eta \sim \mathcal{N}(0, I), \tag{13}$$

with the noise term omitted at the final step. Repeating this process from $t = T$ to $t = 1$ yields the generated post-perturbation expression profile.

A.4 Training Objective

We optimize the encoder, perturbation module, latent SCM, and conditional denoiser using the following objective.

Expression reconstruction. For each training example, we sample $z \sim q_\phi(z \mid x)$, $t \sim \mathrm{Uniform}\{1, \ldots, T\}$, and $\epsilon \sim \mathcal{N}(0, I)$. We construct $u^p = (I - A)^{-1}(z + s^p)$ and $y_t = \sqrt{\bar{\alpha}_t} y_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon$. The denoiser directly predicts clean expression, with loss

$$\mathcal{L}_{\mathrm{rec}} = \mathbb{E}\left[\frac{1}{G}\left\|D_\theta(y_t, t, u^p) - y_0\right\|_2^2\right], \tag{14}$$

where G is the number of expression features and the expectation covers training examples and latent, timestep, and noise sampling.

Basal-posterior regularization. We regularize the diagonal Gaussian posterior toward the standard Gaussian prior:

$$\begin{aligned}\mathcal{L}_{\mathrm{KL}} &= \mathbb{E}_x[D_{\mathrm{KL}}(q_\phi(z \mid x) \,\|\, \mathcal{N}(0, I))] \\ &= \frac{1}{2}\mathbb{E}_x\left[\sum_{j=1}^{d}\left(\mu_{\phi,j}(x)^2 + \sigma_{\phi,j}(x)^2 - \log \sigma_{\phi,j}(x)^2 - 1\right)\right].\end{aligned} \tag{15}$$

Structural regularization. The adjacency matrix has zero diagonal. We encourage acyclicity using $h(A) = \mathrm{tr}(\exp(A \odot A)) - d$ and sparsity using $\mathcal{R}_1(A) = \frac{1}{d(d-1)}\sum_{i \neq j}|A_{ij}|$. The complete objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{rec}} + \lambda_{\mathrm{KL}}\mathcal{L}_{\mathrm{KL}} + \alpha h(A) + \frac{\rho}{2}h(A)^2 + \lambda_1 \mathcal{R}_1(A). \tag{16}$$

We update α and ρ through an augmented Lagrangian schedule (Zheng et al., 2018). At scheduled updates, if $h(A)$ exceeds the tolerance and is greater than 0.25 times its previous value, we set $\rho \leftarrow \min(10\rho, \rho_{\max})$. Whenever $h(A)$ exceeds the tolerance, we then update $\alpha \leftarrow \alpha + \rho h(A)$. Loss weights, initialization, and optimization settings are specified in Appendix C.2.

B Assumptions and Proofs

B.1 Proof of Theorem 4.4

Theorem 4.4 [Identifiability of $\mathcal{F}(U^p)$ without identifiability of A]

Suppose the model in Section 4.1 holds, under Assumptions 4.2 and 4.3, the family of latent distributions

$$\{\mathcal{F}(U^p) : p \in \mathcal{P}\}$$

is identifiable from the family of observed distributions

$$\{\mathcal{F}(Y^p) : p \in \mathcal{P}\}.$$

Moreover, for each $p \in \mathcal{P}$,

$$\mathcal{F}(U^p) = \mathcal{N}(m_p, \Sigma_U),$$

where

$$m_p = B(\mu_Z + s^p), \qquad \Sigma_U = B\Sigma_Z B^\top.$$

In particular, the covariance Σ_U is invariant across perturbation domains, while the perturbation-specific variation is entirely encoded by the mean shift

$$m_p - m_{p'} = B(s^p - s^{p'}), \qquad p, p' \in \mathcal{P}.$$

Proof. By Assumption 4.2,

$$Z \sim \mathcal{N}(\mu_Z, \Sigma_Z),$$

and for each $p \in \mathcal{P}$,

$$U^p = B(Z + s^p).$$

Since s^p is deterministic, we have

$$Z + s^p \sim \mathcal{N}(\mu_Z + s^p, \Sigma_Z).$$

Applying the linear transformation B, it follows that

$$U^p \sim \mathcal{N}\big(B(\mu_Z + s^p),\, B\Sigma_Z B^\top\big).$$

Therefore,

$$\mathcal{F}(U^p) = \mathcal{N}(m_p, \Sigma_U),$$

with

$$m_p = B(\mu_Z + s^p), \qquad \Sigma_U = B\Sigma_Z B^\top.$$

This proves the explicit Gaussian form of $\mathcal{F}(U^p)$ and shows that the covariance Σ_U does not depend on p.

Next, the observed variable Y^p is generated through the shared conditional kernel $K(\cdot \mid u)$:

$$Y^p \mid U^p = u \sim K(\cdot \mid u).$$

Hence the marginal law of Y^p can be written as

$$\mathcal{F}(Y^p) = \int K(\cdot \mid u)\, d\mathcal{F}(U^p)(u) = \Phi\big(\mathcal{F}(U^p)\big),$$

where

$$\Phi(Q) := \int K(\cdot \mid u)\, dQ(u).$$

Now consider any alternative parameterization

$$(\widetilde{A}, \widetilde{Z}, \widetilde{s}^p),$$

which induces

$$\widetilde{U}^p = (I - \widetilde{A})^{-1}(\widetilde{Z} + \widetilde{s}^p), \qquad \widetilde{Y}^p \mid \widetilde{U}^p = u \sim K(\cdot \mid u).$$

Suppose that for all $p \in \mathcal{P}$,

$$\mathcal{F}(\widetilde{Y}^p) = \mathcal{F}(Y^p).$$

By Definition 4.1, the alternative exogenous variable satisfies $\widetilde{Z} \sim \mathcal{N}(\widetilde{\mu}_Z, \widetilde{\Sigma}_Z)$ with $\widetilde{\Sigma}_Z \succ 0$. Since $\widetilde{A}$ is DAG-compatible, $I - \widetilde{A}$ is invertible. Thus, for each p, the invertible affine transformation $\widetilde{U}^p = (I - \widetilde{A})^{-1}(\widetilde{Z} + \widetilde{s}^p)$ has a nondegenerate Gaussian distribution, and

$$\mathcal{F}(\widetilde{Y}^p) = \Phi\big(\mathcal{F}(\widetilde{U}^p)\big).$$

Therefore,

$$\Phi\big(\mathcal{F}(\widetilde{U}^p)\big) = \Phi\big(\mathcal{F}(U^p)\big), \qquad \forall p \in \mathcal{P}.$$

Since both $\mathcal{F}(\widetilde{U}^p)$ and $\mathcal{F}(U^p)$ belong to the Gaussian class $\mathcal{G}$, Assumption 4.3 implies that

$$\mathcal{F}(\widetilde{U}^p) = \mathcal{F}(U^p), \qquad \forall p \in \mathcal{P}.$$

By Definition 4.1, this proves that the family

$$\{\mathcal{F}(U^p) : p \in \mathcal{P}\}$$

is identifiable from

$$\{\mathcal{F}(Y^p) : p \in \mathcal{P}\}.$$

The theorem follows. □

B.2 Proof of Theorem 4.5

Theorem 4.5 [Identifiability of A from the identifiable distribution of U]

Let

$$U = (I - A)^{-1}(Z + s^p) = B(Z + s^p), \qquad B := (I - A)^{-1},$$

where $A \in \mathbb{R}^{d \times d}$ and $Z, s^p \in \mathbb{R}^d$. Assume that:

1. A is strictly lower triangular, i.e., $A_{ii} = 0$ for all i and $A_{ij} = 0$ whenever $i \le j$;

2. the coordinates of Z are mutually independent with mean zero and unknown variances, so that
$$\Sigma_Z = \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d), \qquad \lambda_i > 0;$$
3. the distribution of U is identifiable.

Then A is identifiable. More precisely, both Λ and A are uniquely determined by Σ_U.

Proof. Since
$$U = B(Z + s^p),$$
we have
$$\Sigma_U = B\Lambda B^\top.$$
Define
$$M := B\Lambda^{1/2}.$$
Then
$$\Sigma_U = MM^\top.$$
Because A is strictly lower triangular, $I - A$ is lower triangular with diagonal entries equal to 1, and hence
$$B = (I - A)^{-1}$$
is also lower triangular with diagonal entries equal to 1. Therefore $M = B\Lambda^{1/2}$ is lower triangular with positive diagonal entries
$$M_{ii} = \sqrt{\lambda_i} > 0.$$
Now Σ_U is positive definite, so its Cholesky factorization is unique. Since M is lower triangular with positive diagonal entries and satisfies
$$\Sigma_U = MM^\top,$$
it is uniquely determined by Σ_U. Hence $M = B\Lambda^{1/2}$ is identifiable.

Next, consider
$$M^{-1} = (B\Lambda^{1/2})^{-1} = \Lambda^{-1/2}B^{-1}.$$
Using $B^{-1} = I - A$, we obtain
$$M^{-1} = \Lambda^{-1/2}(I - A) = \Lambda^{-1/2} - \Lambda^{-1/2}A.$$
Since A is strictly lower triangular, its diagonal entries are zero. Therefore the diagonal entries of M^{-1} are exactly those of $\Lambda^{-1/2}$:
$$(M^{-1})_{ii} = \lambda_i^{-1/2}, \qquad i = 1, \ldots, d.$$
Thus $\Lambda^{-1/2}$, and hence Λ, are identifiable from the diagonal of M^{-1}.

Finally, since both M^{-1} and $\Lambda^{-1/2}$ are identifiable, the matrix
$$\Lambda^{-1/2}A = \Lambda^{-1/2} - M^{-1}$$
is identifiable. Multiplying on the left by $\Lambda^{1/2}$ yields
$$A = I - \Lambda^{1/2}M^{-1},$$
so A is identifiable.

Therefore both Λ and A are uniquely determined by Σ_U, and hence by the identifiable distribution of U. □

Discussion of Remark 4.6. For any fixed permutation matrix P, let

$$P\Sigma_U P^\top = L_P L_P^\top$$

be its unique Cholesky factorization, with L_P lower triangular and positive diagonal. Define

$$D_P = \mathrm{diag}\big((L_P)_{11}, \ldots, (L_P)_{dd}\big), \qquad B_P = L_P D_P^{-1},$$

and

$$\Lambda_P = D_P^2, \qquad A_P = I - B_P^{-1}.$$

Then B_P is unit lower triangular, A_P is strictly lower triangular, and

$$P\Sigma_U P^\top = (I - A_P)^{-1}\Lambda_P(I - A_P)^{-\top}.$$

Thus, the structural factorization is unique conditional on P. Applying the construction under different orderings can yield distinct DAGs after mapping back to the original coordinates. Consequently, uniqueness for each fixed ordering does not establish uniqueness across orderings.

B.3 Proof of Theorem 4.9

Theorem 4.9 [Domain-invariant Bayes decision rule]

Suppose the model in Section 4.1 holds and let $h^\star_{\mathrm{src}}$ be a source-induced oracle Bayes decision rule in the sense of Definition 4.8. Then, for every target domain $p \in \mathcal{P}_{\mathrm{tgt}}$ and every measurable decision rule $h : \mathcal{U} \to \mathcal{Y}$,

$$\mathbb{E}[L(h^\star_{\mathrm{src}}(U^p), Y^p)] \le \mathbb{E}[L(h(U^p), Y^p)]\,.$$

Proof. Fix any target domain $p \in \mathcal{P}_{\mathrm{tgt}}$ and any measurable decision rule $h : \mathcal{U} \to \mathcal{Y}$. By model specification, the conditional law of Y^p given $U^p = u$ is the domain-invariant kernel $K(\cdot \mid u)$. Hence,

$$\begin{aligned}\mathbb{E}[L(h(U^p), Y^p)|U^p] &= \int_{\mathcal{Y}} L(h(U^p), y)\, K(dy \mid U^p) \\ &= R_{\mathrm{src}}(U^p; h(U^p))\,,\end{aligned}$$

where the second equality follows from the definition of R_{src} and the invariance of $K(\cdot \mid u)$. Taking expectation gives

$$\mathbb{E}[L(h(U^p), Y^p)] = \mathbb{E}[R_{\mathrm{src}}(U^p; h(U^p))]\,.$$

The same argument applied to $h^\star_{\mathrm{src}}$ gives

$$\mathbb{E}[L(h^\star_{\mathrm{src}}(U^p), Y^p)] = \mathbb{E}[R_{\mathrm{src}}(U^p; h^\star_{\mathrm{src}}(U^p))]\,.$$

By Definition 4.8, $h^\star_{\mathrm{src}}(u) \in \arg\min_{w \in \mathcal{Y}} R_{\mathrm{src}}(u; w)$ for every $u \in \mathcal{U}$. Therefore,

$$R_{\mathrm{src}}(u; h^\star_{\mathrm{src}}(u)) \le R_{\mathrm{src}}(u; h(u))\,, \qquad \text{for every } u \in \mathcal{U}.$$

Substituting $u = U^p$ yields

$$R_{\mathrm{src}}(U^p; h^\star_{\mathrm{src}}(U^p)) \le R_{\mathrm{src}}(U^p; h(U^p)) \qquad \text{almost surely.}$$

Taking expectation, we obtain

$$\mathbb{E}[R_{\mathrm{src}}(U^p; h^\star_{\mathrm{src}}(U^p))] \le \mathbb{E}[R_{\mathrm{src}}(U^p; h(U^p))]\,.$$

Combining this inequality with the preceding equalities gives

$$\mathbb{E}[L(h^\star_{\mathrm{src}}(U^p), Y^p)] \le \mathbb{E}[L(h(U^p), Y^p)]\,.$$

This completes the proof. □

B.4 Assumption and Proof of Theorem 4.10

Assumption B.1 (Approximate representation and stable conditional generation)**.** *Assume that:*

(a) $(\mathcal{U}, d_{\mathcal{U}})$ *is a metric space, and* dist *is a nonnegative discrepancy between probability measures on* $\mathcal{Y}$ *satisfying the triangle inequality.*

(b) There exists a nondecreasing concave function $\omega : [0,\infty) \to [0,\infty)$ *with* $\omega(0) = 0$ *such that, for all* $u, u' \in \mathcal{U}$,

$$\mathrm{dist}(K(\cdot \mid u), K(\cdot \mid u')) \le \omega(d_{\mathcal{U}}(u, u'))\,.$$

(c) $\sup_{p \in \mathcal{P}_{\mathrm{src}} \cup \mathcal{P}_{\mathrm{tgt}}} \mathbb{E}\Big[d_{\mathcal{U}}(U^p, \hat{U}^p)\Big] \le \delta_U$.

(d) $\sup_{p \in \mathcal{P}_{\mathrm{src}} \cup \mathcal{P}_{\mathrm{tgt}}} \mathbb{E}\Big[\mathrm{dist}\Big(K(\cdot \mid \hat{U}^p), Q_{\mathrm{gen}}(\cdot \mid \hat{U}^p)\Big)\Big] \le \delta_Q$.

(e) For any two probability kernels $G_1(\cdot \mid v)$ *and* $G_2(\cdot \mid v)$ *on* $\mathcal{Y}$ *and any random variable* V *taking values in* $\mathcal{U}$,

$$\mathrm{dist}\Big(\int_{\mathcal{U}} G_1(\cdot \mid v)\, P_V(dv), \int_{\mathcal{U}} G_2(\cdot \mid v)\, P_V(dv)\Big) \le \mathbb{E}[\mathrm{dist}(G_1(\cdot \mid V), G_2(\cdot \mid V))]\,.$$

(f) For any coupled random variables (V, V') *taking values in* $\mathcal{U}$,

$$\mathrm{dist}\Big(\int_{\mathcal{U}} K(\cdot \mid v)\, P_V(dv), \int_{\mathcal{U}} K(\cdot \mid v')\, P_{V'}(dv')\Big) \le \mathbb{E}[\mathrm{dist}(K(\cdot \mid V), K(\cdot \mid V'))]\,.$$

Theorem 4.10[Approximate OOD conditional generation]

Suppose the model in Section 4.1 and Assumption B.1 hold. Then for every target domain $p \in \mathcal{P}_{\mathrm{tgt}}$,

$$\mathbb{E}_X\Big[\mathrm{dist}\Big(P_{Y^p|X}(\cdot \mid X), P^{Q_{\mathrm{gen}}}_{Y^p|X}(\cdot \mid X)\Big)\Big] \le \omega(\delta_U) + \delta_Q.$$

In particular, if $\omega(r) = Kr$ for some $K > 0$, then

$$\mathbb{E}_X\Big[\mathrm{dist}\Big(P_{Y^p|X}(\cdot \mid X), P^{Q_{\mathrm{gen}}}_{Y^p|X}(\cdot \mid X)\Big)\Big] \le K\delta_U + \delta_Q.$$

Proof. Fix any target domain $p \in \mathcal{P}_{\mathrm{tgt}}$.

For P_X-almost every $x \in \mathcal{X}$, the true conditional distribution is

$$P_{Y^p|X}(\cdot \mid x) = \int_{\mathcal{U}} K(\cdot \mid u)\, P_{U^p|X}(du \mid x),$$

and the learned conditional distribution is

$$P^{Q_{\mathrm{gen}}}_{Y^p|X}(\cdot \mid x) = \int_{\mathcal{U}} Q_{\mathrm{gen}}(\cdot \mid \hat{u})\, P_{\hat{U}^p|X}(d\hat{u} \mid x).$$

Introduce the intermediate distribution

$$\bar{P}^K_{Y^p|X}(\cdot \mid x) := \int_{\mathcal{U}} K(\cdot \mid \hat{u})\, P_{\hat{U}^p|X}(d\hat{u} \mid x).$$

By the triangle inequality,

$$\begin{aligned}
&\mathrm{dist}\Big(P_{Y^p|X}(\cdot \mid x), P^{Q_{\mathrm{gen}}}_{Y^p|X}(\cdot \mid x)\Big) \\
&\le \mathrm{dist}\Big(P_{Y^p|X}(\cdot \mid x), \bar{P}^K_{Y^p|X}(\cdot \mid x)\Big) \\
&\quad + \mathrm{dist}\Big(\bar{P}^K_{Y^p|X}(\cdot \mid x), P^{Q_{\mathrm{gen}}}_{Y^p|X}(\cdot \mid x)\Big)\,.
\end{aligned}$$

First term. By Assumption B.1(f), applied conditionally on $X = x$ to the coupled pair $(U^p, \hat{U}^p)$,

$$\begin{aligned}&\mathrm{dist}\Big(P_{Y^p|X}(\cdot \mid x), \bar{P}^K_{Y^p|X}(\cdot \mid x)\Big)\\ &\le \mathbb{E}\Big[\mathrm{dist}\Big(K(\cdot \mid U^p), K(\cdot \mid \hat{U}^p)\Big)\Big| X = x\Big].\end{aligned}$$

By Assumption B.1(b),

$$\mathrm{dist}\Big(K(\cdot \mid U^p), K(\cdot \mid \hat{U}^p)\Big) \le \omega\Big(d_{\mathcal{U}}(U^p, \hat{U}^p)\Big).$$

Therefore,

$$\mathrm{dist}\Big(P_{Y^p|X}(\cdot \mid x), \bar{P}^K_{Y^p|X}(\cdot \mid x)\Big) \le \mathbb{E}\Big[\omega\Big(d_{\mathcal{U}}(U^p, \hat{U}^p)\Big)\Big| X = x\Big].$$

Taking expectation over X gives

$$\begin{aligned}&\mathbb{E}_X\Big[\mathrm{dist}\Big(P_{Y^p|X}(\cdot \mid X), \bar{P}^K_{Y^p|X}(\cdot \mid X)\Big)\Big]\\ &\le \mathbb{E}\Big[\omega\Big(d_{\mathcal{U}}(U^p, \hat{U}^p)\Big)\Big].\end{aligned}$$

Since ω is concave, Jensen's inequality gives

$$\mathbb{E}\Big[\omega\Big(d_{\mathcal{U}}(U^p, \hat{U}^p)\Big)\Big] \le \omega\Big(\mathbb{E}\Big[d_{\mathcal{U}}(U^p, \hat{U}^p)\Big]\Big).$$

By Assumption B.1(c) and the monotonicity of ω,

$$\mathbb{E}_X\Big[\mathrm{dist}\Big(P_{Y^p|X}(\cdot \mid X), \bar{P}^K_{Y^p|X}(\cdot \mid X)\Big)\Big] \le \omega(\delta_U).$$

Second term. By Assumption B.1(e), applied conditionally on $X = x$ with

$$G_1(\cdot \mid \hat{u}) = K(\cdot \mid \hat{u}), \qquad G_2(\cdot \mid \hat{u}) = Q_{\mathrm{gen}}(\cdot \mid \hat{u}),$$

we have

$$\begin{aligned}&\mathrm{dist}\Big(\bar{P}^K_{Y^p|X}(\cdot \mid x), P^{Q_{\mathrm{gen}}}_{Y^p|X}(\cdot \mid x)\Big)\\ &\le \mathbb{E}\Big[\mathrm{dist}\Big(K(\cdot \mid \hat{U}^p), Q_{\mathrm{gen}}(\cdot \mid \hat{U}^p)\Big)\Big| X = x\Big].\end{aligned}$$

Taking expectation over X yields

$$\begin{aligned}&\mathbb{E}_X\Big[\mathrm{dist}\Big(\bar{P}^K_{Y^p|X}(\cdot \mid X), P^{Q_{\mathrm{gen}}}_{Y^p|X}(\cdot \mid X)\Big)\Big]\\ &\le \mathbb{E}\Big[\mathrm{dist}\Big(K(\cdot \mid \hat{U}^p), Q_{\mathrm{gen}}(\cdot \mid \hat{U}^p)\Big)\Big] \le \delta_Q,\end{aligned}$$

where the last inequality follows from Assumption B.1(d).

Combining the two bounds with the triangle inequality gives

$$\mathbb{E}_X\Big[\mathrm{dist}\Big(P_{Y^p|X}(\cdot \mid X), P^{Q_{\mathrm{gen}}}_{Y^p|X}(\cdot \mid X)\Big)\Big] \le \omega(\delta_U) + \delta_Q.$$

The linear case $\omega(r) = Kr$ follows immediately. □

C Experimental Details

C.1 Datasets

Replogle (genetic). CRISPRi Perturb-seq in K562 (Replogle et al., 2022), preprocessed as in (Liu et al., 2026), comprising 101,930 cells over 8,563 genes. The library targets 285 genes and includes 10,691 non-targeting control cells; every condition retains at least 200 cells. Of the 285 targets, 276 have a GenePT protein embedding and are the conditions the model sees. We construct five out-of-distribution folds by holding out whole target genes: no cell of a held-out gene appears in training, so generalization is to perturbations never observed, not to unseen cells of a seen perturbation.

sci-Plex3 (chemical). Dose-resolved chemical screen (Srivatsan et al., 2020), preprocessed as in (Wei et al., 2026) and restricted to the MCF7 cell line, with 223,630 cells over 5,839 genes. A condition is a (compound, dose) pair: 187 compounds at four doses give 748 conditions, against 1,011 vehicle control cells. Five out-of-distribution folds are formed over conditions, giving roughly 155,000 training, 22,000 validation and 44,000 test cells per fold.

Norman (combinatorial). We use the Norman two-gene combinatorial perturbation dataset (Norman et al., 2019), as preprocessed and released by Doloris (Chi et al., 2026). It contains 91,205 cells across 5,045 genes, including 7,353 control cells, with 152 single-gene and 131 two-gene conditions. All single-gene conditions are assigned to training. Two-gene conditions are shuffled with seed 42 and divided into five folds, with 26–27 pairs held out per fold to evaluate generalization to unseen combinations.

C.2 Implementation Details

Perturbation features. Genetic perturbations use the 3072-dimensional GenePT protein embeddings (Chen & Zou, 2025), which encode a target gene through a language model description of its protein product. A multi-gene condition is represented by the sum of its members' vectors; because the perturbation encoder is a bias-free linear map, $\mathrm{enc}(a+b) = \mathrm{enc}(a)+\mathrm{enc}(b)$, so the combinatorial case inherits the additivity the single-gene case relies on and requires no separate mechanism.

Chemical perturbations use 512-dimensional Uni-Mol embeddings (Zhou et al., 2023) computed from each compound's 3D conformer with the `unimol_v1_84m` checkpoint. Since sci-Plex3 has four doses per compound but one embedding per compound, the dose is appended as $\log_{10}(\text{dose})$, giving a 513-dimensional feature; this keeps the encoder bias-free and leaves the injection path and the identifiability argument unchanged.

Architecture and optimization. The latent width is $d = 2048$ on Replogle, 1024 on Norman and 256 on sci-Plex3, chosen by the sweep in Appendix D.3. The denoising network is a stack of 4 residual blocks, each a pre-norm two-layer MLP with a $4\times$ hidden expansion and GELU activation. The diffusion process uses $T = 1000$ steps with a linear β schedule and is parameterized to predict y_0 directly. Models are trained with Adam at learning rate 2×10^{-5}, batch size 128, for 100 epochs, with a KL weight of 2.3×10^{-4} and weight decay 0 on the genetic tasks and 10^{-6} on sci-Plex3; the checkpoint with the lowest validation reconstruction loss is selected for evaluation. The acyclicity constraint is the NOTEARS penalty (Zheng et al., 2018), $h(A) = \mathrm{tr}\big(e^{A\odot A}\big) - d$, optimized with an augmented Lagrangian using $\rho_0 = 1$, $\alpha_0 = 0$, multiplier 10, $\rho_{\max} = 10^8$, tolerance $h(A) \le 10^{-8}$ and a progress threshold of 0.25, with A initialized at the empty graph. All runs use seed 42; the fold-to-fold standard deviations reported throughout are therefore across data splits, not across seeds. Training was performed on NVIDIA RTX A6000.

C.3 Baselines

All baselines are run from their authors' released code with their published hyperparameters, retrained on our folds, and scored by the same evaluation code against the same control sample. Where a baseline exposes a choice that our data forces, we record it: TxPert is run in `sample_cell` control mode on Norman, because that dataset carries no batch annotation and the default batch-matched mode is unavailable. Doloris and TxPert accept multi-gene conditions in their own forward pass and are used directly on Norman without modification.

C.4 Evaluation Metric Details

For each perturbation condition c we compute three condition-level mean expression profiles over genes: the control mean $\mu^{\mathrm{ctrl}} \in \mathbb{R}^G$, the ground-truth post-perturbation mean $\mu^{\mathrm{true},(c)}$, and the predicted post-perturbation mean $\mu^{\mathrm{pred},(c)}$.

Δ correlations. The condition-level post–control expression shift is

$$\boldsymbol{\Delta}^{(c)} = \mu^{\mathrm{true},(c)} - \mu^{\mathrm{ctrl}}, \qquad \widehat{\boldsymbol{\Delta}}^{(c)} = \mu^{\mathrm{pred},(c)} - \mu^{\mathrm{ctrl}}, \tag{17}$$

and Δ-P, Δ-S are the Pearson and Spearman correlations between $\mathbf{\Delta}^{(c)}$ and $\widehat{\mathbf{\Delta}}^{(c)}$. A DE-restricted variant restricts both vectors to the K genes with the largest absolute observed shift,

$$S_c^{(K)} = \mathrm{TopK}\Big(\Big|\mathbf{\Delta}^{(c)}\Big|\Big), \tag{18}$$

and is reported for several K in Appendix D.2.

Perturbation discrimination score (PDS). The Δ correlations judge each condition in isolation and do not test whether a prediction is specific to the condition it was made for. Let $\{c_1, \ldots, c_n\}$ be the held-out conditions of a fold. For the prediction $\widehat{\mathbf{\Delta}}^{(c_i)}$ we rank all n observed shifts by

$$d_{ij} = \begin{cases} \big\|\mathbf{\Delta}^{(c_j)} - \widehat{\mathbf{\Delta}}^{(c_i)}\big\|_1, & \text{PDS-L1}, \\ 1 - \cos\big(\mathbf{\Delta}^{(c_j)}, \widehat{\mathbf{\Delta}}^{(c_i)}\big), & \text{PDS-cos}, \end{cases} \tag{19}$$

and record the position $r_i \in \{0, \ldots, n-1\}$ of the matching condition. The score is the normalized rank $\frac{1}{n}\sum_i (1 - r_i/n)$, so 1 means every prediction retrieves its own condition first. Under uniformly random rankings, the expected score is $(n+1)/(2n)$, approximately 0.5. For genetic perturbations the expression of the targeted gene itself is excluded from Eq. 19: a model that merely reproduces the knockdown of the targeted gene could otherwise identify the condition without predicting any downstream response.

Distribution metrics. To complement condition-level mean metrics, we evaluate agreement between predicted and observed cell populations. Following the use of 30 principal components for distributional evaluation in Palla et al. (2026), we compute MMD and energy distance in a 30-dimensional PCA space. PCA is fitted on the control cells within each fold, with the same projection applied to observed and predicted cells for all methods. For each condition, we sample $N = 100$ observed and $N = 100$ predicted cells, denoted by $\{a_i\}_{i=1}^N$ and $\{b_j\}_{j=1}^N$ after projection. We estimate energy distance as

$$\mathrm{E} = \frac{2}{N^2}\sum_{i,j}\|a_i - b_j\|_2 - \frac{1}{N^2}\sum_{i,j}\|a_i - a_j\|_2 - \frac{1}{N^2}\sum_{i,j}\|b_i - b_j\|_2. \tag{20}$$

At the population level, energy distance is zero if and only if the projected distributions coincide, provided they have finite first moments.

We also estimate squared maximum mean discrepancy, denoted by MMD below, using the Gaussian kernel $k_\sigma(a, b) = \exp(-\|a - b\|_2^2/(2\sigma^2))$:

$$\mathrm{MMD}_\sigma = \frac{1}{N^2}\sum_{i,j}k_\sigma(a_i, a_j) + \frac{1}{N^2}\sum_{i,j}k_\sigma(b_i, b_j) - \frac{2}{N^2}\sum_{i,j}k_\sigma(a_i, b_j). \tag{21}$$

The reference bandwidth σ_0^2 is determined by the median heuristic on the pooled observed cells of the condition and N control cells. We average the MMD estimates over $\sigma^2 \in \{0.5\sigma_0^2, \sigma_0^2, 2\sigma_0^2\}$. Both metrics are computed per condition and then averaged; lower values indicate better distributional agreement.

D Additional Results

This section presents additional results on combinatorial perturbation prediction (Appendix D.1), DE-restricted correlations (Appendix D.2), ablation studies (Appendix D.3), biological associations of latent response programs (Appendix D.4), and empirical diagnostics of theoretical assumptions (Appendix D.5).

D.1 Combinatorial two-gene perturbations

Table 2 reports the Norman benchmark. TxPert and scLAMBDA match or slightly exceed scCausalDiff on mean shift, with no consistent per-fold winner (Δ-P favours us in 2/5 folds against TxPert and 3/5 against scLAMBDA), while the distribution metrics separate the methods decisively and in the same direction on every fold. scCausalDiff is first on both distribution metrics in 5/5 folds against every baseline; TxPert, the strongest competitor on mean shift, has an energy distance $8.8\times$ larger than ours.

Table 2: Two-gene combinatorial perturbation prediction (Norman, 5-fold OOD). Mean$_{\pm\text{std}}$. Best in bold.

Model	Mean shift		Discrimination		Distribution	
	Δ-P ↑	Δ-S ↑	PDS-cos ↑	PDS-L1 ↑	MMD ↓	E-dist ↓
SCCVAE	$0.5621_{\pm 0.054}$	$0.2885_{\pm 0.021}$	$0.6374_{\pm 0.073}$	$0.6475_{\pm 0.059}$	$0.0682_{\pm 0.004}$	$1.1780_{\pm 0.084}$
SENA	$0.5977_{\pm 0.009}$	$0.2834_{\pm 0.011}$	$0.7119_{\pm 0.069}$	$0.6356_{\pm 0.037}$	$0.0521_{\pm 0.003}$	$0.8239_{\pm 0.041}$
Doloris	$0.6765_{\pm 0.027}$	$0.4081_{\pm 0.018}$	$0.8588_{\pm 0.047}$	$0.7237_{\pm 0.032}$	$0.0367_{\pm 0.005}$	$0.6110_{\pm 0.090}$
GEARS	$0.6994_{\pm 0.018}$	$0.4217_{\pm 0.017}$	$0.8640_{\pm 0.024}$	$0.9044_{\pm 0.017}$	$0.0368_{\pm 0.010}$	$0.6170_{\pm 0.147}$
scLAMBDA	$0.8499_{\pm 0.018}$	$\mathbf{0.5533}_{\pm 0.015}$	$0.9376_{\pm 0.023}$	$0.8888_{\pm 0.041}$	$0.0319_{\pm 0.009}$	$0.6379_{\pm 0.120}$
TxPert	$\mathbf{0.8610}_{\pm 0.018}$	$0.4808_{\pm 0.013}$	$\mathbf{0.9636}_{\pm 0.014}$	$\mathbf{0.9386}_{\pm 0.018}$	$0.0658_{\pm 0.008}$	$3.2029_{\pm 0.169}$
Ours	$0.8453_{\pm 0.024}$	$0.5271_{\pm 0.010}$	$0.9248_{\pm 0.051}$	$0.9318_{\pm 0.034}$	$\mathbf{0.0201}_{\pm 0.004}$	$\mathbf{0.3654}_{\pm 0.074}$

D.2 DE-RESTRICTED CORRELATIONS

Figure 4 evaluates Δ-P and Δ-S on the top-K genes ranked by absolute observed shift, ranging from 50 genes to the full transcriptome. On sci-Plex3, scCausalDiff leads on both correlations at the evaluated DE-restricted cut-offs, although scLAMBDA achieves a higher whole-transcriptome Δ-S. On Replogle, TxPert leads on DE-restricted Δ-P, whereas scCausalDiff achieves a higher whole-transcriptome correlation. Several curves initially increase before declining as more genes are included. These results show that the selected gene set affects both correlation values and method rankings, complementing the population-level comparisons in Table 1.

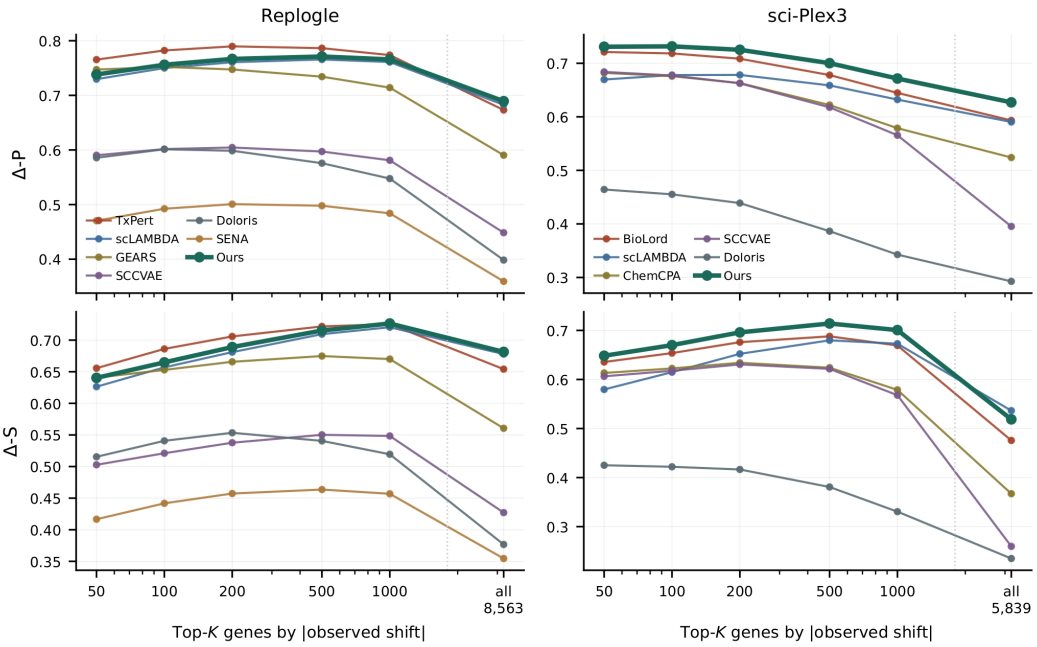

Figure 4: DE-restricted correlations against the number of genes retained, ranked by absolute observed shift. The rightmost point uses the full transcriptome and equals the value in Table 1. Five-fold means.

D.3 ABLATION STUDIES

We examine the learned propagation structure, the conditional generator, and the latent dimensionality. Decoder and latent-width comparisons use a single-stage training configuration. Structural replacements are evaluated on the models reported in Table 1.

Learned latent propagation. We compare the learned structure with three alternatives: no propagation ($A = 0$), a random latent structure with matched propagation-operator norm, and an orthogonal

transformation. The learned structure achieves higher mean-response correlations and lower distributional discrepancies on both Replogle and sci-Plex3.

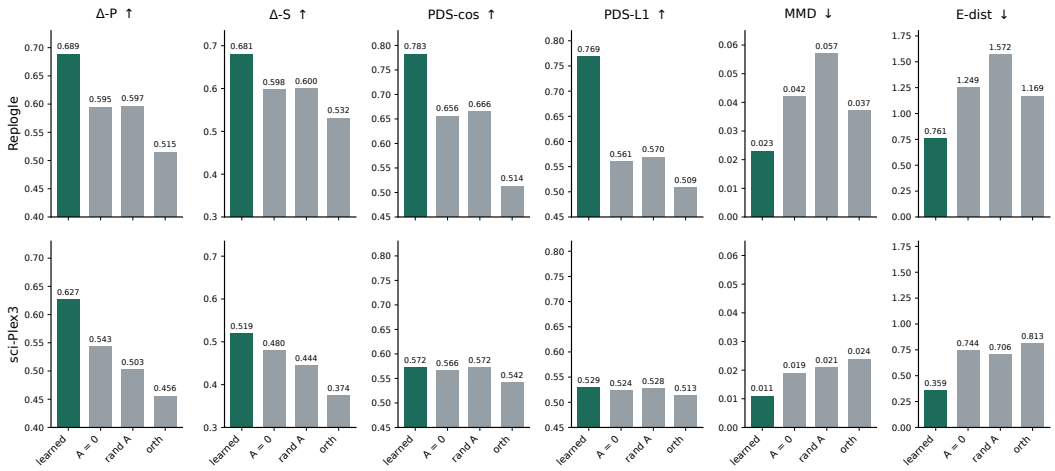

Figure 5: Replacing the learned latent structure at inference time, all other weights held fixed: $A = 0$, a random A rescaled to the original $\|(I - A)^{-1}\|_2$, and an orthogonal transform. Five-fold means; each metric shares one vertical range across the two rows. PDS-cos and PDS-L1 barely separate the arms on sci-Plex3, where those metrics have little dynamic range across all methods.

Conditional generation. Figure 6 compares diffusion-based and VAE-based generation under matched remaining training settings. Under the evaluated readout procedures, diffusion achieves higher mean-response correlations and lower distributional discrepancies on both datasets. PDS scores also improve on Replogle, while the two variants obtain similar PDS scores on sci-Plex3.

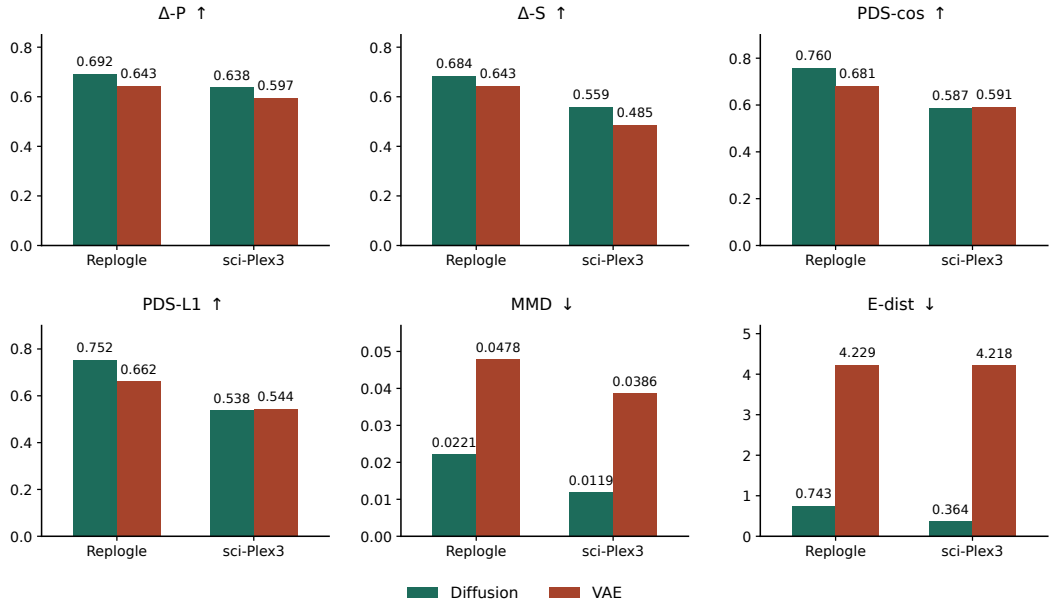

Figure 6: Comparison of diffusion-based and VAE-based generation under the same remaining training configuration. Five-fold means on Replogle and sci-Plex3.

Sensitivity to latent dimensionality. Figure 7 evaluates six metrics for $d \in \{256, 512, 1024, 2048\}$ on Replogle and sci-Plex3. On Replogle, increasing d improves mean-response correlations and

PDS scores while reducing distributional discrepancies, with smaller gains at larger dimensions. On sci-Plex3, Δ-P decreases slightly with increasing d, whereas PDS scores improve and Δ-S varies non-monotonically. Distributional discrepancies remain similar at $d = 256$ and 512 but increase at larger dimensions. These results indicate dataset-dependent effects of latent dimensionality, with no single width optimizing every metric.

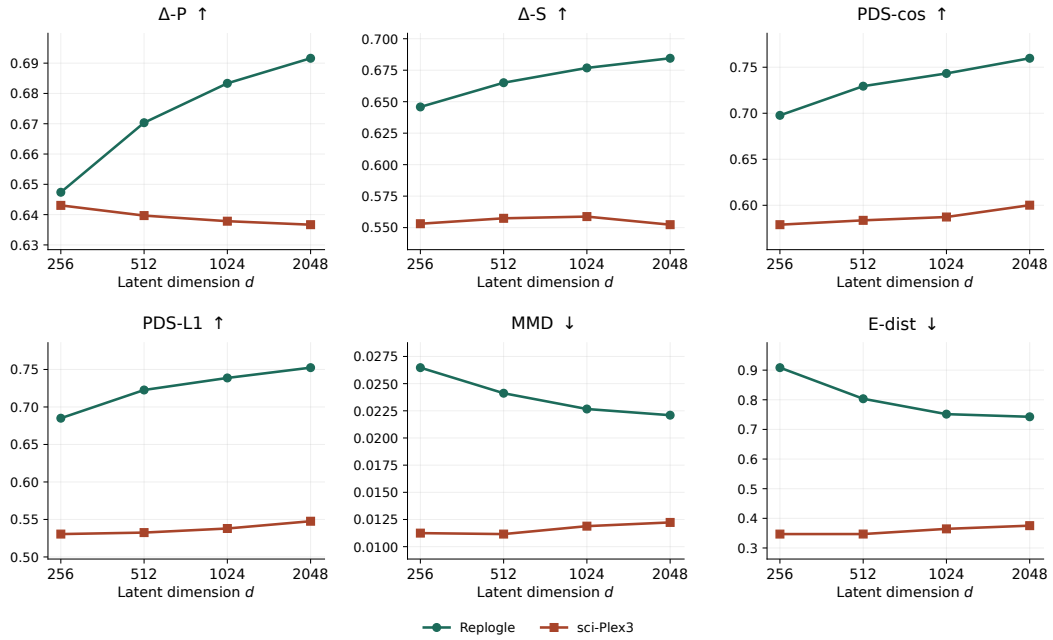

Figure 7: Sensitivity to latent dimensionality on Replogle and sci-Plex3 across six evaluation metrics. Curves show five-fold means for $d \in \{256, 512, 1024, 2048\}$. The denoiser width changes together with d under direct additive conditioning.

D.4 Biological associations of latent response programs

Analysis setting. We examine how latent propagation organizes perturbation responses and relates them to biological functions. The analysis uses the Replogle model with $d = 1024$, whose latent structural matrix is learned with the NOTEARS acyclicity constraint. Unless otherwise stated, results concern fold 1, comprising 220 training and 56 held-out perturbations with valid input representations. We compare the direct shift s^p with the propagated shift $(I - A)^{-1}s^p$. PCA and expression-program regressions are fitted on training conditions; held-out conditions are projected without refitting.

Organization of response variation. We first fit PCA separately to the direct and propagated training representations. Explaining 90% of their variance requires 5–7 components before propagation and five components afterwards across all folds (Table 3). Propagation thus concentrates response variation into fewer dominant directions in four of the five folds. We use the first five propagated components of fold 1 for the subsequent biological annotations and case study.

Table 3: Components required for 90% of the perturbation-response variance, per fold.

Representation	fold 0	fold 1	fold 2	fold 3	fold 4
Direct shift	6	7	6	5	7
After propagation	5	5	5	5	5

Functional classes along latent directions. Figure 3 relates propagated PC scores to manually curated functional classes of perturbed genes. For each component, we select the 25 perturbations

at each pole and assess class enrichment using one-sided Fisher tests with Benjamini–Hochberg correction within each pole. The analysis includes all 276 valid conditions, which also define the enrichment background. The positive pole of PC1 contains 25 ribosomal-protein perturbations, out of 41 in the background. Other poles show associations with Mediator, splicing, and proteasome classes. These patterns connect the dominant latent directions to functional similarities among perturbations.

Downstream expression programs. To connect these directions to transcriptional responses, we jointly regress observed mean expression effects on the five standardized propagated-PC scores using training conditions. For each component, the 200 genes with the largest positive and negative regression coefficients are analyzed separately using g:Profiler (Kolberg et al., 2023), querying GO Biological Process (Ashburner et al., 2000; The Gene Ontology Consortium, 2026), Reactome (Ragueneau et al., 2026), and KEGG (Kanehisa et al., 2025). Mapped measured genes define the custom background. Figure 8 reports the two terms with the smallest service-adjusted p-values per coefficient-sign group. For example, PC3 combines a positive-pole association with proteasome perturbations and positive output-gene associations with amino-acid metabolism and stress response. The perturbation-side and expression-side analyses therefore provide complementary functional annotations. Table 4 further reports held-out expression-program associations for the direct and propagated representations, each using its own training-fitted PCA and regression.

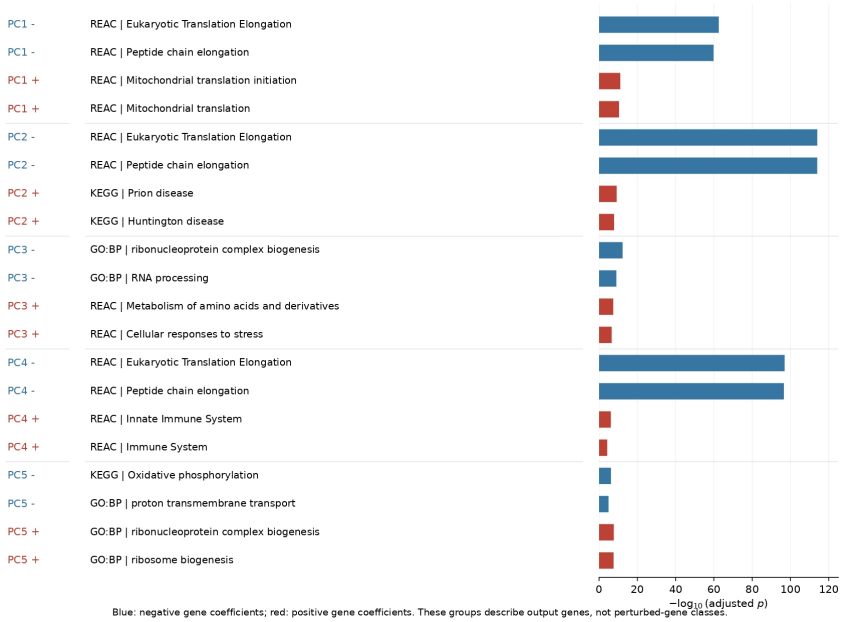

Figure 8: Pathway enrichment of the output genes with the largest positive (red) and negative (blue) regression coefficients on each propagated component, two most significant terms per group. The regression is fitted on training conditions. These terms describe downstream expression programs, not classes of perturbed genes.

Table 4: Held-out Spearman agreement between each component and its expression program, fitted on training conditions only ($n = 56$ held-out conditions). The two rows come from separately fitted PCAs and regressions, so components at the same index are not the same direction.

Representation	1	2	3	4	5
Direct shift	0.520	0.496	0.563	0.307	0.519
After propagation	0.510	0.597	0.595	0.393	0.613

A held-out perturbation case. Figure 9 examines *WDR3*, which is excluded from fold-1 training entirely. Both shifts are projected onto the same propagated-PC basis and evaluated with one expression probe fitted on propagated training representations. Propagation changes the allocation

across response directions and increases Pearson agreement with the observed top-200 gene effects from $r = 0.711$ to $r = 0.957$. The generated cells reach $r = 0.969$ on the 40 genes with the largest observed effect, while systematically underestimating effect magnitudes, visible as a slope below one in panel D. This held-out case illustrates how propagation reshapes latent response coordinates and how the resulting predictions align with observed expression effects. Together, these analyses support biological associations of the learned representations rather than identification of gene-regulatory interactions.

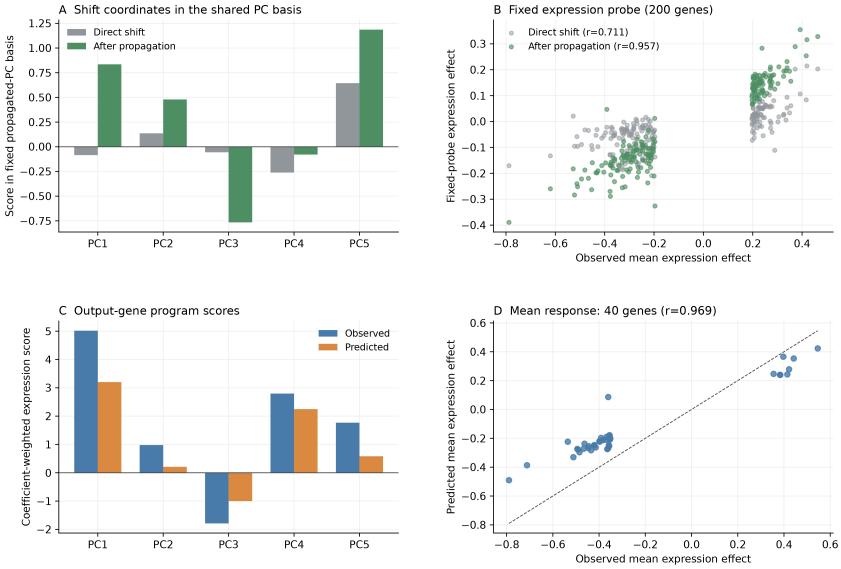

Figure 9: Strictly held-out *WDR3*. (A) shift coordinates in the shared propagated PC basis, before and after propagation. (B) a probe fitted on the propagated training representation and then held fixed, applied to both representations. (C) output-gene program scores, observed against predicted. (D) generated mean response on the 40 genes with the largest observed effect; correlations are Pearson.

D.5 Empirical Diagnostics of Theoretical Assumptions

Diagnostic setting. We examine the same Replogle model used in the biological analysis (fold 1, $d = 1024$), with the latent structural matrix learned using the NOTEARS acyclicity constraint. We sample 4,096 control cells and repeat posterior sampling three times to characterize the aggregate basal distribution $q(z) = \mathbb{E}_X[q_\phi(z \mid X)]$. These diagnostics assess the empirical behavior of the learned representation in relation to the population-level assumptions.

Gaussianity and coordinate dependence. Figure 10 summarizes marginal normality, coordinate correlations, and the allocation of basal-state variation. D'Agostino–Pearson tests with Benjamini–Hochberg correction at 0.05 reject normality for the same four of 1,024 coordinates in all three posterior draws; none of 64 tested random projections is rejected. We define the background-signal fraction of coordinate j as $\mathrm{Var}_X[\mu_j(X)]/\mathrm{Var}(Z_j)$. Four of the five coordinates with a background-signal fraction above 50% also exhibit non-Gaussianity. Aggregate correlations, computed using the law of total covariance, are small: the mean and maximum absolute off-diagonal correlations are 0.0011 and 0.0799. A complementary permutation-based distance-correlation test detects no significant dependence among 32 tested coordinate pairs after correction (Figure 11). These results indicate weak pairwise dependence and approximate marginal Gaussianity across most coordinates, with deviations concentrated in several high-background-signal coordinates.

Shared variation and observation-map sensitivity. Applying different shifts to common basal samples and inverting the SCM recovers the original samples with relative error below 1.5×10^{-7}, confirming the implementation of shared exogenous variation. This sharing is imposed by the model rather than independently established from observed perturbation domains. We additionally

evaluate local denoiser sensitivity along 32 random orthogonal latent directions at three conditional anchors and two diffusion timesteps. All six restricted Jacobians have numerical rank 32 at a relative singular-value threshold of 10^{-3} (Figure 11). This provides evidence of local sensitivity in the tested subspaces. Joint independence and injectivity of the conditional observation operator remain theoretical assumptions rather than conclusions of these diagnostics.

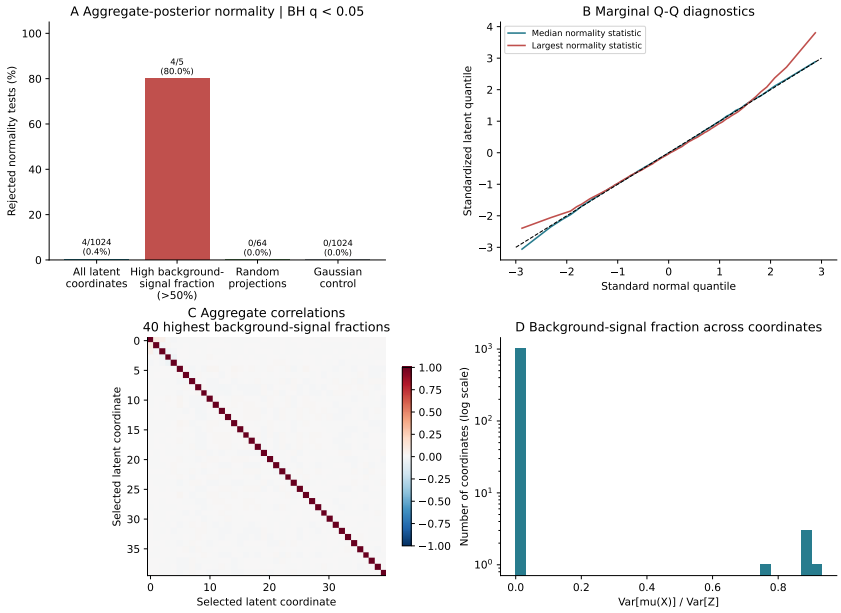

Figure 10: Aggregate basal-latent diagnostics. (A) Normality-test rejection fractions. (B) Representative marginal Q–Q plots. (C) Correlations among the 40 coordinates with the highest background-signal fractions, selected without dimensionality reduction. (D) Background-signal fractions across all 1,024 coordinates, defined as $\mathrm{Var}_X[\mu_j(X)]/\mathrm{Var}(Z_j)$.

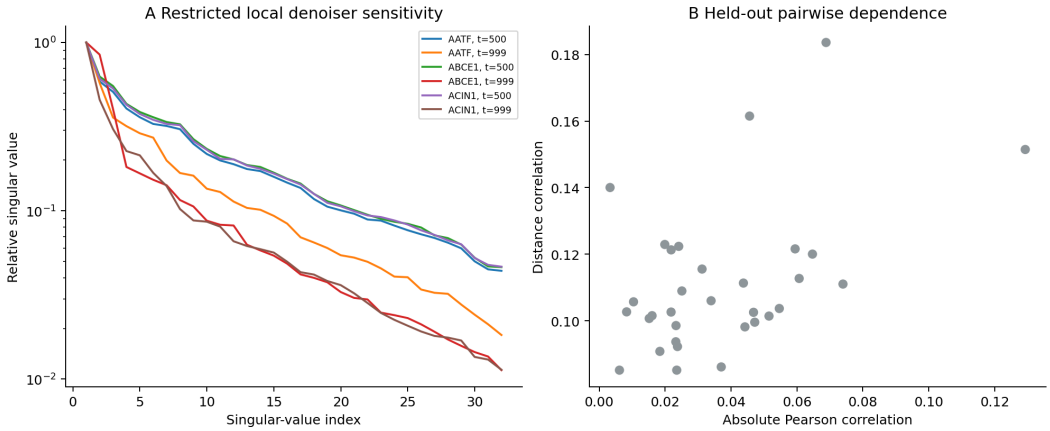

Figure 11: Local sensitivity and pairwise dependence. (A) Normalized singular-value spectra of denoiser Jacobians restricted to 32 latent directions. (B) Absolute Pearson correlation and distance correlation for 32 coordinate pairs, evaluated on cells disjoint from those used for pair selection. No pair is significant after multiple-testing correction.