DeMixPert: Decomposed Response Modeling with Gaussian Mixtures for OOD Single-Cell Perturbation Prediction

Jiawen Liu[1], Xuechenxiao Cao[1], Yutong Li[2], Bing Liu[1], Jiaming Liang[1], Tinghe Zhang[2], Xiaoqi Sheng[2], Hongmin Cai[1,2]*

[1]School of Computer Science and Engineering, South China University of Technology, Guangzhou, China
[2]School of Future Technology, South China University of Technology, Guangzhou, China
*hmcai@scut.edu.cn

Abstract

Predicting transcriptome-wide responses to unseen genetic perturbations remains a major computational challenge because accurate prediction requires recovering both perturbation-specific transcriptional shifts and heterogeneous cellular responses. Existing methods often entangle deterministic response structure with stochastic population-level variation, causing dominant shared patterns to mask weaker perturbation-specific signals and impair distributional modeling. To address these challenges, we propose **DeMixPert**, an approach for **De**composed response Modeling with Gaussian **Mix**tures for Out-Of-Distribution (OOD) single-cell **Pert**urbation prediction. DeMixPert decomposes perturbation-induced changes into a basal-state-dependent systematic response, a perturbation-specific response, and population-level variation. The systematic component is derived from the basal state encoded from control-cell expression, whereas the perturbation-specific component is inferred from pretrained target embeddings for unseen-target generalization. DeMixPert models population-level variation using a Gaussian prototype Invertible Network and adaptively combines reusable Gaussian prototypes according to the basal state and perturbation condition. The resulting mixture is mapped to a condition-specific variation distribution. Sampled variations are integrated with the systematic and perturbation-specific components, followed by joint decoding with the basal state to reconstruct perturbed-cell gene expression. Experimental results show that DeMixPert effectively captures heterogeneous single-cell perturbation responses and achieves superior performance across unseen-perturbation settings. The source code is made publicly available upon publication.

Introduction

Single-cell genetic perturbation profiling couples controlled genetic interventions with transcriptome-wide readouts, enabling cell-resolved characterization of transcriptional responses and the regulatory programs underlying them (Adamson et al. 2016; Dixit et al. 2016). Despite its value, the scalability of such profiling is limited by the combinatorial growth of experimental conditions across perturbation targets, cellular states, and biological contexts (Cheng et al. 2026), making exhaustive measurement infeasible. Accordingly, *in silico* perturbation-response prediction offers a scalable means of estimating cellular responses under experimentally profiled conditions.

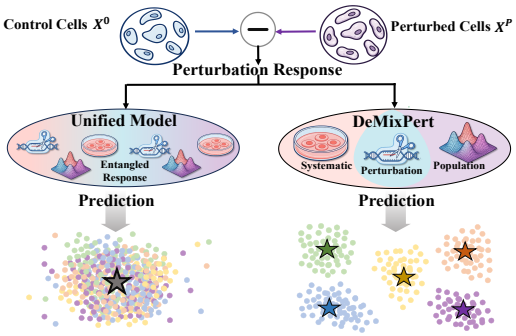

Figure 1: Comparison of unified and decomposed response modeling. By separating entangled response components, DeMixPert better preserves perturbation-specific signals and condition-dependent population structure.

However, predicting responses under experimentally unobserved conditions involves more than estimating an average perturbation-induced expression shift (Yu et al. 2025). These challenges can be understood through three intertwined components of perturbation responses, as illustrated in Fig. 1. First, the basal cellular state contributes a systematic component that shapes responses across perturbation conditions (Dong et al. 2023; Song et al. 2025). This shared structure should be learned from observed conditions and transferred to unseen ones. Second, each perturbation condition induces a perturbation-specific transcriptional effect that distinguishes its response from those of other interventions (Norman et al. 2019; Replogle et al. 2020). For an unseen perturbation, this effect must be inferred from biological relationships with targets observed during training. Third, cells exposed to the same perturbation may occupy distinct response subpopulations, whose internal structure and relative abundance vary with cellular state and environmental context (Adamson et al. 2016; Frangieh et al. 2021). Consequently, Out-Of-Distribution (OOD) prediction imposes three distinct requirements involving systematic-structure transfer, unseen-target effect inference, and condition-specific population-distribution estimation without direct observations.

*Corresponding author.

Existing methods improve perturbation-response prediction from different perspectives. Latent-variable approaches represent perturbation responses through state transformations or factorized perturbation components (Lotfollahi et al. 2023). In parallel, Knowledge-guided methods exploit gene representations or biological priors to generalize from limited perturbation observations (Cui et al. 2024). From another perspective, population-distribution approaches learn transitions from control to perturbed populations using optimal transport, set-level distributional objectives, or flow matching (Bunne et al. 2023; Adduri et al. 2025; Yu et al. 2026). Despite substantial advances in response transfer and distributional modeling, most approaches still formulate the full perturbation response as a monolithic prediction target, leaving shared systematic response, perturbation-specific response, and condition-dependent population-level variation entangled. This limitation is consequential because Systema shows that shared systematic variation can dominate standard evaluation metrics (Viñas Torné et al. 2025). Accordingly, accurate OOD prediction requires disentangling basal-state-dependent systematic responses, perturbation-specific responses, and population-level variation.

In this paper, we propose **DeMixPert**, a **De**composed response framework with Gaussian **Mix**tures for out-of-distribution single-cell **Pert**urbation prediction. By decomposing perturbation responses, DeMixPert assigns shared systematic response, perturbation-specific response, and population variation to separate modules. The basal-state-dependent systematic response is integrated with the perturbation-specific response to define a deterministic response center. This separation preserves transferable response structure while preventing shared variation from masking perturbation-specific signals. Subsequently, a Gaussian prototype Invertible Network models population-level variation around the response center by adaptively weighting reusable Gaussian prototypes according to the basal state and perturbation embedding. The resulting mixture is transformed into a condition-specific variation distribution through an invertible mapping. Finally, sampled variations are added to the response center and decoded together with the basal state to reconstruct post-perturbation gene expression. This design improves response recovery for unseen perturbation targets while preserving perturbation discriminability and distributional fidelity. Our main contributions are summarized as follows:

- We propose DeMixPert, a decomposed response model for OOD single-cell perturbation prediction. DeMixPert disentangles deterministic response structure from stochastic cellular heterogeneity and captures condition-specific response distributions using adaptive Gaussian mixtures.
- We introduce a Gaussian prototype Invertible Network for population-level variation modeling. The invertible network uses context-adaptive weights to combine reusable Gaussian mixture prototypes, thereby estimating condition-specific population-variation distributions.
- Quantitative analyses demonstrate that DeMixPert improves OOD perturbation prediction by accurately recovering perturbation-specific responses while preserving population-level distributional fidelity.

Related Work

Genetic Perturbation Prediction

Existing approaches formulate cellular responses as latent transformations: scGen applies an average shift, whereas CPA factorizes treatment, dose, and covariate contributions (Lotfollahi, Wolf, and Theis 2019; Lotfollahi et al. 2023). For unseen targets, GEARS leverages gene-relation graphs, while GenePert draws on external gene embeddings (Roohani, Huang, and Leskovec 2024; Chen and Zou 2024). To capture heterogeneity in unpaired data, STATE predicts post-intervention cell sets, whereas scDFM learns full conditional distributions via flow matching (Adduri et al. 2025; Yu et al. 2026). Systema further reveals that shared systematic variation can dominate standard metrics and mask target-specific signals (Viñas Torné et al. 2025). Nevertheless, most prior work does not jointly disentangle basal-state-dependent systematic responses, perturbation-specific responses, and condition-dependent population-level variation.

Gaussian Mixtures and Distribution Modeling

Gaussian mixture (GM) models provide a flexible representation of complex population structure (McLachlan and Peel 2000), while normalizing flows transform tractable source distributions through invertible mappings (Dinh, Krueger, and Bengio 2014). PerturbNet combines perturbation embeddings with a conditional invertible network to generate cell-state distributions (Yu et al. 2025). More recently, MixFlow replaces the conventional unimodal Gaussian source with a descriptor-conditioned Gaussian mixture and transports it using conditional flow matching (Rubbi et al. 2026). Despite improved distributional flexibility, existing approaches still model perturbation responses monolithically, conflating systematic and perturbation-specific effects with residual variation. This entanglement limits component-wise generalization and mechanistic interpretation.

Methodology

Preliminaries

In the context of OOD single-cell perturbation prediction, we decompose responses into systematic, perturbation-specific, and population-level components to capture transferable patterns and condition-specific heterogeneity.

Let $\mathcal{D} = \left\{X^0, \{X^p\}_{p\in\mathcal{P}}\right\}$ denote a single-cell genetic perturbation dataset, where $\mathcal{P}$ is the set of perturbation conditions, each involving one or more target genes. The control population is defined by $X^0 = \{x_j^0 \in \mathbb{R}^G\}_{j=1}^{N_0}$. Here, N_0 is the number of control cells and G denotes the number of selected genes. For each perturbation condition $p \in \mathcal{P}$, the corresponding perturbed population is defined by $X^p = \{x_i^p\}_{i=1}^{N_p}$, where N_p is the number of cells observed under perturbation p.

Since the control and perturbed populations are unpaired, for each perturbed cell x_i^p, we randomly sample a control cell x_j^0 from X^0 as its reference and define the per-

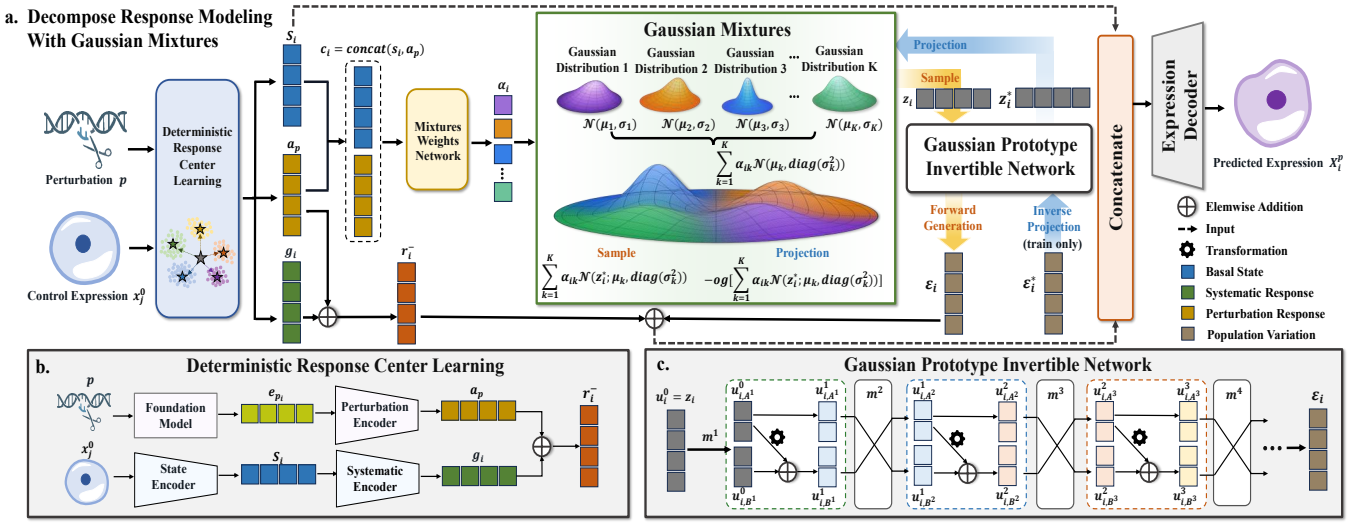

Figure 2: Overview of DeMixPert. **(a)** The overall workflow of DeMixPert. **(b)** The architecture of the Deterministic Response Center Learning module. **(c)** The architecture of the Gaussian prototype Invertible Network.

turbation response as $\Delta x_i^p = x_i^p - x_j^0$. To facilitate subsequent response modeling, the expression-space response Δx_i^p is projected into a latent space using an encoder E_r: $r_i^* = E_r(\Delta x_i^p) \in \mathbb{R}^{d_r}$. The objective of DeMixPert is to predict a perturbation response $\hat{r}_i$ that approximates the target response r_i^*. In DeMixPert, $\hat{r}_i$ is modeled as three components: $\hat{r}_i = g_i + a_p + \epsilon_i$. Here, g_i, a_p, and ϵ_i denote the systematic response, perturbation-specific response, and population-level response variation, respectively.

Overview

Fig. 2 outlines the DeMixPert framework. As shown in Fig. 2 (a), the Deterministic Response Center Learning module takes the perturbation condition and control-cell expression as inputs. The module separates the response into systematic and perturbation-specific components, whose combination forms the deterministic response center. Meanwhile, the basal-state and perturbation representations are concatenated to predict mixture weights, adapting a set of shared Gaussian prototypes to the current cell state and perturbation condition. The adapted Gaussian mixture is connected to the Gaussian prototype Invertible Network for population-variation modeling. The forward transformation generates population-level variations from source samples. Conversely, the training-only inverse transformation projects observed target variations into the Gaussian-mixture space. The corresponding conditional likelihood provides joint supervision for mixture-weight estimation and invertible transformation learning. Generated population-level variations are subsequently added to the deterministic response center. Thereafter, the resulting response representation is concatenated with the basal-state representation for expression decoding.

Deterministic Response Center Learning

To prevent systematic variation from obscuring perturbation-specific effects, DeMixPert constructs a deterministic response center by combining separately learned systematic and perturbation-specific responses. The three encoders E_s, E_{sys}, and E_{pert} produce the basal-state representation, systematic response, and perturbation-specific response, respectively: $s_i = E_s(x_j^0) \in \mathbb{R}^{d_s}$, $g_i = E_{\text{sys}}(s_i) \in \mathbb{R}^{d_r}$, and $a_p = E_{\text{pert}}(e_p) \in \mathbb{R}^{d_r}$. Here, s_i summarizes the unperturbed cellular context and is retained for subsequent variation modeling. The systematic response g_i captures basal-state-dependent patterns shared across perturbation conditions, whereas a_p captures the perturbation-specific response derived from the perturbation embedding e_p. We derive e_p from pretrained scGPT representations (Cui et al. 2024) for OOD generalization. The e_p provides prior knowledge of biological relationships among genes for inferring perturbation-specific responses without direct observations.

Finally, the resulting systematic and perturbation-specific responses are combined to form the deterministic response center: $\bar{r}_i = g_i + a_p \in \mathbb{R}^{d_r}$. The $\bar{r}_i$ denotes the deterministic response pattern jointly specified by the basal cellular state s_i and perturbation target e_p in the response latent space $\mathbb{R}^{d_r}$.

Gaussian prototype Invertible Network

The deterministic response center captures systematic and perturbation-specific responses but not the conditional population-level variation distribution. Accordingly, DeMixPert models the remaining population-level variation using a Gaussian prototype Invertible Network. Using the target response r_i^*, the target variation from the deterministic response center is given by $\epsilon_i^* = r_i^* - \bar{r}_i \in \mathbb{R}^{d_r}$.

For each perturbation condition, the target variations of its observed cells form samples from the corresponding population-variation distribution. To capture potential variation patterns shared across perturbations, the target variations ϵ^* pooled from all training conditions are modeled using a

K-component diagonal Gaussian mixtures:

$$q_{\mathrm{GM}}(\epsilon) = \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\epsilon; \mu_k, \mathrm{diag}(\sigma_k^2)\right), \tag{1}$$

where π_k, μ_k, and $\mathrm{diag}(\sigma_k^2)$ denote the weight, mean, and diagonal variance of the k-th Gaussian prototype, respectively. These parameters are periodically re-estimated from the target variations using the expectation-maximization algorithm (Dempster, Laird, and Rubin 1977). Further details are provided in Supplementary Appendix F.

The global Gaussian Mixtures provide reusable Gaussian prototypes shared across perturbations. However, the contribution of each prototype depends jointly on the basal state and perturbation condition. Accordingly, s_i and a_p are concatenated to predicts Gaussian prototype weights:

$$\alpha_i = \mathrm{softmax}\left(\mathrm{MLP}(\mathrm{concat}(s_i, a_p))\right) \in \mathbb{R}^K, \tag{2}$$

where $\mathrm{concat}(\cdot,\cdot)$ denotes concatenation, $\mathrm{MLP}(\cdot)$ denotes a multilayer perceptron, and $\mathrm{softmax}(\cdot)$ normalizes the predicted prototype weights. The conditional Gaussian mixture is then defined as:

$$P(z \mid \alpha_i) = \sum_{k=1}^{K} \alpha_{ik} \mathcal{N}\left(z; \mu_k, \mathrm{diag}(\sigma_k^2)\right). \tag{3}$$

The prototype parameters are shared globally, while their contributions are adapted to the condition representation c_i. Then, a source variable $z_i \sim P(z \mid \alpha_i)$ is sampled from the conditional mixture and passed through the invertible network T_θ, producing a sample ϵ_i from the final population-variation distribution: $\epsilon_i = T_\theta(z_i; c_i)$.

Let $u_i^{(0)} = z_i$ and $u_i^l = F_l\left(u_i^{(l-1)}; c_i\right) \in \mathbb{R}^{d_r}$ denote the output of the l-th coupling layer. A binary mask $m_l \in \{0,1\}^{d_r}$ partitions the feature dimensions into two complementary index sets, A^l and B^l, corresponding to mask values of 1 and 0, respectively, such that $A^\ell \cap B^\ell = \emptyset$, $A^\ell \cup B^\ell = \{1, \ldots, d_r\}$. Accordingly, the two input subvectors are defined as $u_{i,A^l}^{(l-1)} := \left(u_{ij}^{(l-1)}\right)_{\in A^l} \in \mathbb{R}^{|A^l|}$, $u_{i,B^l}^{(l-1)} := \left(u_{ij}^{(l-1)}\right)_{j \in B^l} \in \mathbb{R}^{|B^l|}$. The $u_i^{(l-1)}$ can be recovered by restoring the two subvectors to the positions specified by m_l: $u_i^{(l-1)} = \mathrm{merge}_{m_l}\left(u_{i,A^l}^{(l-1)}, u_{i,B^l}^{(l-1)}\right)$. Here, merge_{m^l} restores the two subsets to their original feature positions specified by m^l. The mask is fixed within each layer but varies across layers, allowing all feature dimensions to be updated.

Within the l-th coupling layer, $u_{i,A^l}^{(l-1)}$ remain unchanged and is used to determine the transformation applied to $u_{i,B^l}^{(l-1)}$. DeMixPert first measures how closely the unchanged features match each Gaussian prototype. The normalized matching weight associated with prototype k is

$$\gamma_{ik}^l = \frac{\pi_k \mathcal{N}\left(u_{i,A^l}^{(l-1)}; \mu_{k,A^l}, \mathrm{diag}\left(\sigma_{k,A^l}^2\right)\right)}{\sum_{q=1}^{K} \pi_q \mathcal{N}\left(u_{i,A^l}^{(l-1)}; \mu_{q,A^l}, \mathrm{diag}\left(\sigma_{q,A^l}^2\right)\right)}. \tag{4}$$

In parallel, a layer-specific MLP h^l produces prototype modulation coefficients from c_i: $b_i^l = h^l(c_i) \in \mathbb{R}^K$. Here, γ_i^l captures feature-level prototype matching, whereas b_i^l captures the contribution of each prototype under the current basal-state and perturbation context. The two coefficients are combined to generate the additive shift applied to B^l:

$$f^l(u_{i,A^l}, c_i) = W^l\left(\gamma_i^l \odot b_i^l\right) \in \mathbb{R}^{|B^l|}, \tag{5}$$

where $W^l \in \mathbb{R}^{|B^l| \times K}$ is a learnable projection matrix and $\odot$ denotes element-wise multiplication. The additive coupling transformation is defined as follows: $u_{i,A^l}^l = u_{i,A^l}^{l-1}$, $u_{i,B^l}^l = u_{i,B^l}^{l-1} + f_l\left(u_{i,A^l}^{l-1}, c_i\right)$. The two subsets are then reassembled to obtain the complete output of the l-th coupling layer:

$$F_l\left(u_i^{l-1}; c_i\right) = u_i^l = \mathrm{merge}_{m^l}\left(u_{i,A^l}^l, u_{i,B^l}^l\right). \tag{6}$$

Since the A^l subset remains unchanged, the same shift function can be evaluated during inversion (Dinh, Sohl-Dickstein, and Bengio 2016). The inverse transformation is defined by $u_{i,A^l}^{l-1} = u_{i,A^l}^l$ and $u_{i,B^l}^{l-1} = u_{i,B^l}^l - f_l\left(u_{i,A^l}^l, c_i\right)$.

The recovered subsets are reassembled to obtain the complete output of the inverse layer:

$$F_l^{-1}\left(u_i^l; c_i\right) = u_i^{l-1} = \mathrm{merge}_{m^l}\left(u_{i,A^l}^{l-1}, u_{i,B^l}^{l-1}\right). \tag{7}$$

Each coupling layer is bijective and has a triangular Jacobian with unit determinant (Dinh, Krueger, and Bengio 2014). This invertibility maps each target variation back to the Gaussian mixtures. Its likelihood under the source distribution supervises the context-dependent prototype weights and the invertible transformation. The corresponding likelihood objective is introduced in the training loss.

After L coupling layers, the generated variation sample is

$$\epsilon_i = T_\theta(z_i; c_i) = F_L \circ \cdots \circ F_1(z_i; c_i) \in \mathbb{R}^{d_r}. \tag{8}$$

For perturbation condition p, the collection $\{\epsilon_i : p_i = p\}$ forms samples from its predicted population-level variation distribution. Each sample perturbs the deterministic response center to yield the response latent: $\hat{r}_i = \bar{r}_i + \epsilon_i$.

Expression Decoder

The post-perturbation expression is jointly determined by the original cellular context and the perturbation-induced response. The basal-state representation s_i captures the cellular context, whereas the predicted response latent $\hat{r}_i$ encodes the perturbation effect. The concatenated representation is then decoded to reconstruct post-perturbation gene expression: $\hat{x}_i^p = D_\psi\left(\mathrm{concat}(s_i, \hat{r}_i)\right)$. Therein, D_ψ denotes the expression decoder.

Training Objective

DeMixPert is trained end-to-end to jointly learn the deterministic response center and condition-specific population-level variation. The total training objective consists of four loss

components that supervise response alignment, population-variation likelihood, gene-expression reconstruction, and population-distribution matching.

Response Alignment Loss. A mean squared error loss is used to align the predicted response latent with its target: $\mathcal{L}_{\text{align}} = \frac{1}{N}\sum_{i=1}^{N} \|\hat{r}_i - r_i^*\|_2^2$, where N denotes the number of cells in the training batch.

Population-Variation Likelihood Loss. To learn the conditional population-level variation, each target variation is first mapped back to the Gaussian mixtures: $z_i^* = T_\theta^{-1}(\epsilon_i^*; c_i)$. We then minimize its negative log-likelihood under the conditional Gaussian mixtures:

$$\mathcal{L}_{\text{gm}} = -\frac{1}{N}\sum_{i=1}^{N} \log\left[\sum_{k=1}^{K} \alpha_{ik}\mathcal{N}\left(z_i^*; \mu_k, \text{diag}(\sigma_k^2)\right)\right], \quad (9)$$

where, the $\mathcal{L}_{\text{gm}}$ jointly supervises the context-dependent prototype weights and the invertible transformation.

Gene-Expression Reconstruction Loss. The reconstruction loss preserves gene expression information by reconstructing the perturbed expression from both the predicted and target responses, while recovering the control expression from a zero-response vector $\mathbf{0} \in \mathbb{R}^{d_r}$:

$$\begin{aligned}\mathcal{L}_{\text{rec}} = \mathcal{L}_{mse}(D_\psi([s_i, \hat{r}_i]), x_i^{p_i}) + \mathcal{L}_{mse}(D_\psi([s_i, r_i^*]), x_i^{p_i}) \\ + \mathcal{L}_{mse}\big(D_\psi([s_i, \mathbf{0}]), x_j^0\big)\,,\end{aligned} \quad (10)$$

where $\mathcal{L}_{mse}$ denotes the mean square error loss.

Distribution Loss. At the population level, we use energy distance to align the generated and observed expression distributions under each perturbation condition:

$$\mathcal{L}_{\text{dist}} = \frac{1}{|\mathcal{P}|}\sum_{p\in\mathcal{P}} \text{ED}\left(\{\hat{x}_i^p\}_{i:p_i=p}, \{x_i^p\}_{i:p_i=p}\right), \quad (11)$$

where $\mathcal{P}$ denotes the set of perturbation conditions and $\text{ED}(\cdot)$ is the energy distance (Székely and Rizzo 2013). The overall training objective is $\mathcal{L} = \mathcal{L}_{\text{align}} + \lambda_{\text{gm}}\mathcal{L}_{\text{gm}} + \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{dist}}$. The λ_{gm} balances the numerical scale of the negative log-likelihood against the other loss components. All trainable parameters are jointly optimized by minimizing $\mathcal{L}$.

Experiments

Experimental Setup

Datasets and preprocessing. We benchmark DeMixPert on four widely used single-cell genetic perturbation datasets. Adamson (Adamson et al. 2016) and Papalexi (Papalexi et al. 2021) contain single-gene perturbations, whereas Norman (Norman et al. 2019) and Replogle (Replogle et al. 2020) involve combinatorial perturbations. All datasets are processed following a standard single-cell RNA-seq preprocessing pipeline. For each dataset, we select 2,048 highly variable genes (HVGs) and additionally retain all perturbation-target genes to define the input gene space. Detailed dataset descriptions and statistics are provided in Appendix C.

Compared methods. We compare DeMixPert against five representative perturbation-response prediction methods: GEARS (Roohani, Huang, and Leskovec 2024), scGPT (Cui et al. 2024), GenePert (Chen and Zou 2024), scDFM (Yu et al. 2026), and STATE (Adduri et al. 2025). These methods span graph learning, pretrained foundation models, ridge regression, conditional flow matching, and set-based Transformers. Accordingly, the comparative experiments enable DeMixPert to be evaluated against methods with substantially different assumptions and predictive mechanisms.

Metrics and implementation. Model performance is evaluated at both the top-100 differentially expressed genes (DEGs) level and the all-gene level. At the top-100 DEG level, we report Common Differentially Expressed Genes (C-DEGs), Energy Distance (E-Dist), Wasserstein Distance (W-Dist), and Mean Squared Error (MSE). At the all-gene level, we report the Differential Expression Score (DES), MSE, Centroid Accuracy (Centroid Acc), and the Perturbation Discrimination Score (PDS) (Wei et al. 2026; Viñas Torné et al. 2025; Roohani et al. 2025). Together, these metrics assess differential-expression recovery, distributional agreement, expression error, and perturbation discrimination.

Condition-disjoint splits evaluate unseen perturbations. Single-gene datasets use 70%/10%/20% training/validation/test splits. For combinatorial datasets, training includes all single-gene conditions and half of the combinatorial conditions, while the remainder is divided equally between validation and test sets. Validation selects checkpoints, and results are reported as mean $\pm$ standard variation across random seeds. DeMixPert uses PyTorch and AdamW. Default and dataset-specific configurations are provided in AppendiX B and H, respectively.

Quantitative Comparison

Table 1 summarizes the results for unseen single-gene and combinatorial perturbations. On the single-gene datasets Papalexi and Adamson, DeMixPert substantially improves perturbation-specific and distributional recovery, achieving C-DEGs scores of 22.000 and 10.733, PDS values of 0.904 and 0.947, and the lowest E-Dist values of 0.171 and 0.448, respectively. Although ScDFM obtains the highest DES on Papalexi, its lower C-DEGs and PDS suggest that matching response magnitude alone is insufficient to recover perturbation-specific effects. From another direction, performance on unseen combinatorial perturbations evaluates the ability to preserve interaction-specific effects and generalize to novel gene combinations. On Norman and Replogle, DeMixPert achieves the highest PDS (0.981 and 0.938) and the lowest E-Dist (0.374 and 0.155). Its C-DEGs score reaches 31.650 on Norman, surpassing STATE's 21.006, and 12.378 on Replogle, close to the best baseline result of 12.711. Moreover, scGPT's high Centroid Acc of 0.940 but worst E-Dist of 3.174 on Norman demonstrates that accurate centroid prediction does not guarantee distributional recovery. Overall, these results show that DeMixPert generalizes effectively across both perturbation regimes while preserving perturbation specificity and population-level heterogeneity.

Dataset	Method	Top 100 DEGs Level				All Gene Level			
		C-DEGs↑	E-Dist↓	W-Dist↓	MSE↓	DES↑	MSE↓	Centroid Acc↑	PDS↑
Papalexi	GEARS	2.280 ± 0.460	0.617 ± 0.218	12.720 ± 0.263	0.065 ± 0.032	0.194 ± 0.115	0.024 ± 0.015	0.240 ± 0.089	0.640 ± 0.049
	scGPT	1.040 ± 0.862	0.208 ± 0.327	**3.468 ± 1.159**	0.056 ± 0.059	0.077 ± 0.045	0.010 ± 0.011	0.500 ± 0.000	0.623 ± 0.024
	GenePert	5.427 ± 0.889	0.337 ± 0.136	7.011 ± 0.190	0.027 ± 0.012	0.127 ± 0.067	0.004 ± 0.002	0.500 ± 0.000	0.600 ± 0.025
	scDFM	5.560 ± 1.506	1.971 ± 0.582	14.053 ± 0.324	0.274 ± 0.104	**0.212 ± 0.099**	0.099 ± 0.037	0.200 ± 0.141	0.568 ± 0.095
	STATE	4.760 ± 1.499	0.760 ± 0.214	13.364 ± 0.222	0.072 ± 0.033	0.198 ± 0.107	0.025 ± 0.014	0.160 ± 0.089	0.600 ± 0.000
	DeMixPert	**22.000 ± 4.854**	**0.171 ± 0.088**	6.369 ± 0.403	**0.008 ± 0.006**	0.200 ± 0.129	**0.002 ± 0.001**	**0.720 ± 0.110**	**0.904 ± 0.046**
Adamson	GEARS	2.280 ± 0.145	0.563 ± 0.125	9.654 ± 0.217	0.046 ± 0.015	0.313 ± 0.082	0.012 ± 0.003	0.133 ± 0.067	0.605 ± 0.026
	scGPT	4.151 ± 0.131	2.997 ± 0.310	**4.936 ± 0.326**	0.037 ± 0.014	**0.378 ± 0.078**	0.005 ± 0.002	**0.586 ± 0.079**	0.567 ± 0.037
	GenePert	0.063 ± 0.010	0.468 ± 0.174	4.938 ± 0.301	0.031 ± 0.013	0.105 ± 0.038	**0.002 ± 0.001**	0.533 ± 0.014	0.570 ± 0.025
	scDFM	3.400 ± 1.381	1.355 ± 0.216	10.771 ± 0.628	0.132 ± 0.027	0.278 ± 0.082	0.048 ± 0.007	0.067 ± 0.047	0.539 ± 0.040
	STATE	2.947 ± 0.357	0.820 ± 0.138	10.989 ± 0.169	0.052 ± 0.017	0.280 ± 0.091	0.015 ± 0.003	0.080 ± 0.030	0.534 ± 0.004
	DeMixPert	**10.733 ± 5.198**	**0.448 ± 0.052**	5.481 ± 0.113	**0.021 ± 0.003**	0.153 ± 0.078	0.006 ± 0.001	0.560 ± 0.076	**0.947 ± 0.020**
Norman	GEARS	3.850 ± 0.557	0.855 ± 0.220	6.380 ± 0.234	0.049 ± 0.019	0.334 ± 0.029	0.008 ± 0.002	0.237 ± 0.112	0.802 ± 0.090
	scGPT	6.644 ± 0.627	3.174 ± 0.157	**5.627 ± 0.239**	0.028 ± 0.007	**0.540 ± 0.031**	0.004 ± 0.001	**0.940 ± 0.013**	0.886 ± 0.025
	GenePert	19.181 ± 0.658	1.113 ± 0.154	6.305 ± 0.256	0.078 ± 0.013	0.404 ± 0.022	0.008 ± 0.011	0.639 ± 0.034	0.646 ± 0.033
	scDFM	15.350 ± 3.548	0.711 ± 0.148	7.607 ± 0.464	0.046 ± 0.011	0.389 ± 0.029	0.009 ± 0.003	0.463 ± 0.116	0.890 ± 0.037
	STATE	21.006 ± 5.872	1.341 ± 0.217	7.175 ± 0.177	0.105 ± 0.021	0.354 ± 0.013	0.011 ± 0.002	0.031 ± 0.000	0.524 ± 0.014
	DeMixPert	**31.650 ± 11.739**	**0.374 ± 0.056**	6.606 ± 0.269	**0.003 ± 0.0002**	0.506 ± 0.063	**0.003 ± 0.0002**	0.750 ± 0.163	**0.981 ± 0.005**
Replogle	GEARS	2.111 ± 0.176	0.448 ± 0.060	12.743 ± 0.114	0.038 ± 0.009	0.189 ± 0.047	0.012 ± 0.002	0.267 ± 0.127	0.669 ± 0.058
	scGPT	6.089 ± 0.755	2.263 ± 0.280	**4.302 ± 0.483**	0.011 ± 0.004	**0.269 ± 0.060**	0.002 ± 0.001	**0.778 ± 0.125**	0.654 ± 0.087
	GenePert	12.533 ± 2.881	0.200 ± 0.047	4.600 ± 0.550	**0.002 ± 0.001**	0.200 ± 0.047	0.002 ± 0.001	0.558 ± 0.048	0.580 ± 0.008
	scDFM	3.422 ± 1.630	0.802 ± 0.091	13.234 ± 0.272	0.092 ± 0.015	0.202 ± 0.053	0.026 ± 0.004	0.178 ± 0.061	0.588 ± 0.052
	STATE	**12.711 ± 3.278**	0.510 ± 0.075	13.008 ± 0.107	0.040 ± 0.012	0.207 ± 0.041	0.011 ± 0.002	0.111 ± 0.000	0.548 ± 0.014
	DeMixPert	12.378 ± 4.071	**0.155 ± 0.040**	5.614 ± 0.156	0.005 ± 0.001	0.227 ± 0.066	**0.002 ± 0.000**	0.778 ± 0.193	**0.938 ± 0.039**

Table 1: Comprehensive evaluation results on four genetic perturbation datasets, reported as mean ± standard variation. ↑ (↓) indicates that higher (lower) values are better. The best result for each metric is marked in **bold**.

Dataset	Variant	Top 100 DEGs		All Genes	
		C-DEGs ↑	W-Dist ↓	Centroid Acc ↑	PDS ↑
Adamson	w/o Decomp	10.520	5.927	0.240	0.733
	w/o Gene	8.773	5.772	0.267	0.787
	w/o Sys	10.400	5.746	0.360	0.846
	w/o GM	**11.107**	5.781	0.280	0.797
	DeMixPert	10.733	**5.481**	**0.560**	**0.947**
Replogle	w/o Decomp	12.022	5.905	0.244	0.736
	w/o Gene	9.733	5.852	0.333	0.760
	w/o Sys	11.556	5.832	0.333	0.800
	w/o GM	**14.356**	5.833	0.289	0.773
	DeMixPert	12.378	**5.614**	**0.778**	**0.938**

Table 2: The results of the ablation study.

Ablation Study

We conduct ablation experiments on Adamson and Replogle by comparing the complete model with four ablated variants. The w/o Decomp variant replaces decomposed response modeling with unified modeling. The w/o Gene and w/o Sys variants remove the embedding-derived perturbation-specific effect and the basal-state-dependent systematic response, respectively. For brevity, w/o GM denotes removal of the entire Gaussian prototype Invertible Network, including both the GM and the invertible network. The complete results are reported in Table 2.

Table 2 shows that DeMixPert leads overall with the lowest W-Dist and highest Centroid Acc/PDS. Removing decomposition lowers Centroid Acc/PDS to 0.240/0.244 and 0.733/0.736 on Adamson/Replogle; removing the systematic component also degrades both, confirming complementary roles in perturbation specificity and transferable basal-state patterns. Removing the embedding-derived perturbation-specific effect reduces C-DEGs from 10.733/12.378 to 8.773/9.733, confirming its importance for unseen perturbations. Although w/o GM raises C-DEGs to 11.107/14.356, it worsens all other metrics, showing that the deterministic response center alone cannot recover population distributions.

Sensitivity Analysis

Fig. 3 shows performance changes under different Gaussian prototype numbers K and GM-loss weights λ_{gm} relative to the reference setting ($K = 8$, $\lambda_{\mathrm{gm}} = 0.01$). Deviating from $K = 8$ generally degrades DEG recovery, distribution modeling, and expression reconstruction, whereas larger K improves perturbation identification, indicating a fidelity–separability trade-off. Similarly, increasing λ_{gm} improves identification but slightly affects other objectives. Overall, the reference setting achieves the best balance.

Interpretability Analysis

Two interpretability analyses are conducted to validate two central design principles of DeMixPert: explicit response decomposition and condition-dependent composition of shared Gaussian prototypes.

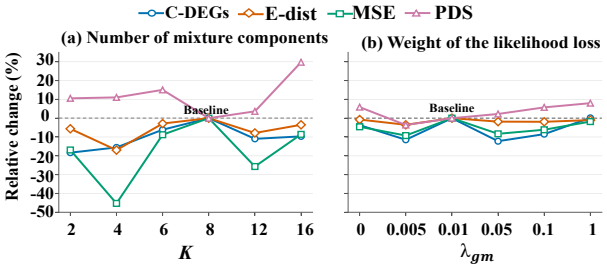

Figure 3: The result of the sensitivity analysis. Positive values indicate improvement and negative values degradation.

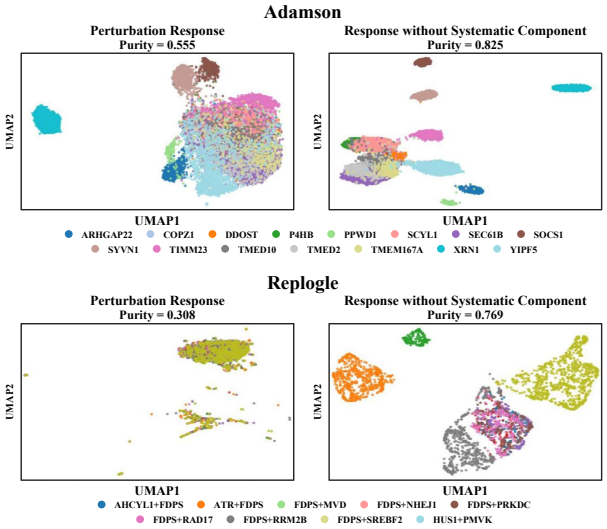

Figure 4: UMAP visualizations of the perturbation response r_i^* (left) and the response after removing the basal-state-dependent systematic component, $r_i^* - g_i$ (right), on the Adamson and Replogle datasets. Points are colored by perturbation condition, and cluster purity scores quantify the separation among perturbations.

Fig. 4 compares representations of predicted response before and after removing the systematic component g_i. Cluster purity is applied to assess the separation of basal-state variation from perturbation-associated structure. It investigates whether the basal-state-dependent systematic response interferes with perturbation-associated structures in the predicted response space. After removing g_i, the responses on both datasets form more compact and clearly separated clusters. Accordingly, cluster purity increases from 0.308 to 0.769 on Replogle and from 0.555 to 0.825 on Adamson. Since g_i is computed exclusively from the basal-state representation and is the only component removed in this comparison, these improvements indicate that it introduces variation that is not aligned with perturbation identity. This validates the response decomposition in DeMixPert and improves the discrimination of unseen perturbations.

Fig. 5 illustrates the relationship between Gaussian prototype weights and cellular distributions across different perturbation conditions. For each pair of perturbation conditions, D_α denotes the Jensen–Shannon (JS) divergence (Lin 2002) between their Gaussian prototype weight vectors α, whereas D_ϵ denotes the energy distance between their population distributions ϵ. Spearman correlation is then used to assess the overall rank association between the two distance matrices (Spearman 1961), and its statistical significance is evaluated using a Mantel-style permutation test (Dietz 1983). Across all pairs of perturbation conditions, significant overall positive rank associations are observed in both the Adamson ($\rho = 0.462$) and Replogle ($\rho = 0.512$) datasets, with Mantel $p < 0.001$ in both cases. The consistency of this association across the single-gene and combinatorial perturbation datasets indicates that differences in the learned prototype weights are closely associated with variation in perturbation-specific population distributions. These results provide empirical support for the effectiveness and interpretability of the Gaussian prototype composition mechanism.

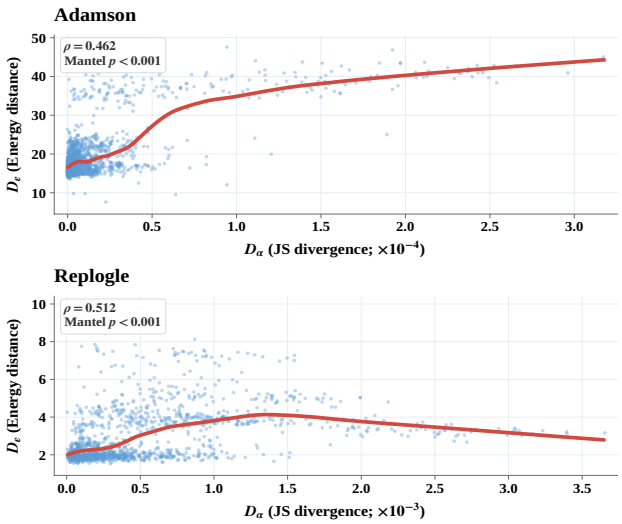

Figure 5: The correlation between Prototype-weight divergence and perturbation-specific population differences. Pairwise prototype-weight divergence D_α is compared with population-level distance D_ϵ in the Adamson and Replogle datasets. The red curves show the LOWESS-smoothed relationships between D_α and D_ϵ.

Conclusion

In this study, we present DeMixPert, a Gaussian-mixture-based response decomposition framework for OOD single-cell perturbation prediction. DeMixPert disentangles perturbation responses into systematic, perturbation-specific, and population-level components, each modeled through a dedicated mechanism. A Gaussian prototype Invertible Network further characterizes cellular heterogeneity through context-adaptive prototype composition. Extensive experiments demonstrate accurate recovery of perturbation-specific responses under OOD settings. The adaptive prototype composition also captures condition-specific population structures, offering a structured interpretation of heterogeneous cellular responses.

Acknowledgements

This work was supported in part by the National Key Research and Development Program of China (2024YFF1206600); in part by Guangdong S&T Programme (2025B0101130001); in part by the National Natural Science Foundation of China (62325204, 62502161, 62502163); in part by the China Postdoctoral Science Foundation (2025M782912).

References

Adamson, B.; Norman, T. M.; Jost, M.; Cho, M. Y.; Nuñez, J. K.; Chen, Y.; Villalta, J. E.; Gilbert, L. A.; Horlbeck, M. A.; Hein, M. Y.; et al. 2016. A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell*, 167(7): 1867–1882.

Adduri, A. K.; Gautam, D.; Bevilacqua, B.; Imran, A.; Shah, R.; Naghipourfar, M.; Teyssier, N.; Ilango, R.; Nagaraj, S.; Dong, M.; et al. 2025. Predicting cellular responses to perturbation across diverse contexts with State. *BioRxiv*, 2025–06.

Bunne, C.; Stark, S. G.; Gut, G.; Del Castillo, J. S.; Levesque, M.; Lehmann, K.-V.; Pelkmans, L.; Krause, A.; and Rätsch, G. 2023. Learning single-cell perturbation responses using neural optimal transport. *Nature methods*, 20(11): 1759–1768.

Chen, Y.; and Zou, J. 2024. Genepert: Leveraging genept embeddings for gene perturbation prediction. *bioRxiv*, 2024–10.

Cheng, J.; Chi, C.; Zhou, J.; Xin, H.; and Xia, J. 2026. PRESCRIBE: Predicting Single-Cell Responses with Bayesian Estimation. *Advances in Neural Information Processing Systems*, 38: 145457–145489.

Cui, H.; Wang, C.; Maan, H.; Pang, K.; Luo, F.; Duan, N.; and Wang, B. 2024. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature methods*, 21(8): 1470–1480.

Dempster, A. P.; Laird, N. M.; and Rubin, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1): 1–22.

Dietz, E. J. 1983. Permutation tests for association between two distance matrices. *Systematic Biology*, 32(1): 21–26.

Dinh, L.; Krueger, D.; and Bengio, Y. 2014. Nice: Nonlinear independent components estimation. *arXiv preprint arXiv:1410.8516*.

Dinh, L.; Sohl-Dickstein, J.; and Bengio, S. 2016. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*.

Dixit, A.; Parnas, O.; Li, B.; Chen, J.; Fulco, C. P.; Jerby-Arnon, L.; Marjanovic, N. D.; Dionne, D.; Burks, T.; Raychowdhury, R.; et al. 2016. Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *cell*, 167(7): 1853–1866.

Dong, M.; Wang, B.; Wei, J.; de O. Fonseca, A. H.; Perry, C. J.; Frey, A.; Ouerghi, F.; Foxman, E. F.; Ishizuka, J. J.; Dhodapkar, R. M.; et al. 2023. Causal identification of single-cell experimental perturbation effects with CINEMA-OT. *Nature methods*, 20(11): 1769–1779.

Frangieh, C. J.; Melms, J. C.; Thakore, P. I.; Geiger-Schuller, K. R.; Ho, P.; Luoma, A. M.; Cleary, B.; Jerby-Arnon, L.; Malu, S.; Cuoco, M. S.; et al. 2021. Multimodal pooled Perturb-CITE-seq screens in patient models define mechanisms of cancer immune evasion. *Nature genetics*, 53(3): 332–341.

Lin, J. 2002. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information theory*, 37(1): 145–151.

Lotfollahi, M.; Klimovskaia Susmelj, A.; De Donno, C.; Hetzel, L.; Ji, Y.; Ibarra, I. L.; Srivatsan, S. R.; Naghipourfar, M.; Daza, R. M.; Martin, B.; et al. 2023. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular systems biology*, 19(6): MSB202211517.

Lotfollahi, M.; Wolf, F. A.; and Theis, F. J. 2019. scGen predicts single-cell perturbation responses. *Nature methods*, 16(8): 715–721.

McLachlan, G. J.; and Peel, D. 2000. *Finite mixture models*. John Wiley & Sons.

Norman, T. M.; Horlbeck, M. A.; Replogle, J. M.; Ge, A. Y.; Xu, A.; Jost, M.; Gilbert, L. A.; and Weissman, J. S. 2019. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science*, 365(6455): 786–793.

Papalexi, E.; Mimitou, E. P.; Butler, A. W.; Foster, S.; Bracken, B.; Mauck III, W. M.; Wessels, H.-H.; Hao, Y.; Yeung, B. Z.; Smibert, P.; et al. 2021. Characterizing the molecular regulation of inhibitory immune checkpoints with multimodal single-cell screens. *Nature genetics*, 53(3): 322–331.

Replogle, J. M.; Norman, T. M.; Xu, A.; Hussmann, J. A.; Chen, J.; Cogan, J. Z.; Meer, E. J.; Terry, J. M.; Riordan, D. P.; Srinivas, N.; et al. 2020. Combinatorial single-cell CRISPR screens by direct guide RNA capture and targeted sequencing. *Nature biotechnology*, 38(8): 954–961.

Roohani, Y.; Huang, K.; and Leskovec, J. 2024. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nature Biotechnology*, 42(6): 927–935.

Roohani, Y. H.; Hua, T. J.; Tung, P.-Y.; Bounds, L. R.; Yu, F. B.; Dobin, A.; Teyssier, N.; Adduri, A.; Woodrow, A.; Plosky, B. S.; et al. 2025. Virtual Cell Challenge: Toward a Turing test for the virtual cell. *Cell*, 188(13): 3370–3374.

Rubbi, A.; Akbarnejad, A.; Sanian, M. V.; Parast, A. Y.; Asadollahzadeh, H.; Amani, A.; Akhtar, N.; Cooper, S.; Bassett, A.; Liò, P.; et al. 2026. MixFlow: Mixture-Conditioned Flow Matching for Out-of-Distribution Generalization. *arXiv preprint arXiv:2601.11827*.

Song, B.; Liu, D.; Dai, W.; Mcmyn, N. F.; Wang, Q.; Yang, D.; Krejci, A.; Vasilyev, A.; Untermoser, N.; Loregger, A.; et al. 2025. Decoding heterogeneous single-cell perturbation responses. *Nature cell biology*, 27(3): 493–504.

Spearman, C. 1961. The proof and measurement of association between two things.

Székely, G. J.; and Rizzo, M. L. 2013. Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference*, 143(8): 1249–1272.

Viñas Torné, R.; Wiatrak, M.; Piran, Z.; Fan, S.; Jiang, L.; Teichmann, S. A.; Nitzan, M.; and Brbić, M. 2025. Systema: a framework for evaluating genetic perturbation response prediction beyond systematic variation. *Nature Biotechnology*, 1–10.

Wei, Z.; Wang, Y.; Gao, Y.; Wang, S.; Li, P.; Si, D.; Gao, Y.; Wu, S.; Li, D.; Dong, K.; et al. 2026. Benchmarking algorithms for generalizable single-cell perturbation response prediction. *Nature Methods*, 23(2): 451–464.

Yu, C.; Wang, C.; Liao, B.; and Wu, T. 2026. scdfm: Distributional flow matching model for robust single-cell perturbation prediction. *arXiv preprint arXiv:2602.07103*.

Yu, H.; Qian, W.; Song, Y.; and Welch, J. D. 2025. Perturbnet predicts single-cell responses to unseen chemical and genetic perturbations. *Molecular Systems Biology*, 21(8): 960.

Supplementary Appendix

Perturbation Embedding Construction

As stated in the main paper, the perturbation embedding is constructed from pretrained scGPT gene representations (Cui et al. 2024). Let $\mathcal{M}_p$ denote the set of target genes named by perturbation condition p, and let $\mathbf{e}_m \in \mathbb{R}^{512}$ be the pretrained scGPT embedding of target gene m. We define

$$e_p = \frac{1}{|\mathcal{M}_p|} \sum_{m \in \mathcal{M}_p} \mathbf{e}_m \in \mathbb{R}^{512}. \tag{12}$$

For a single-gene perturbation, this reduces to the embedding of that gene. For a combinatorial perturbation, the target-gene embeddings are averaged with equal weight. Conditions whose named target is unavailable in the frozen embedding table are removed before split generation.

Default Model Configuration

This section reports the default architecture and optimization configuration shared across all datasets. A summary of the model architecture is provided in Table 4, and dataset-specific training settings are reported in Supplementary Appendix H.

Hyperparameter	**Value**
State latent dimension	256
Response latent dimension	128
Perturbation embedding dimension	512
Gaussian components K	8
Coupling layers L	4
Condition-network hidden width	256
Prototype temperature	1.0
Optimizer	AdamW
Learning rate	5×10^{-5}
Weight decay	10^{-6}
Gradient clipping	global norm 5.0
Learning-rate scheduler	none (constant learning rate)

Default configuration shared across datasets. Dataset-specific batch sizes, evaluation intervals, Gaussian-mixture refresh settings, and training schedules are listed in Supplementary Appendix H.

Dataset Descriptions

We evaluate DeMixPert on four single-cell genetic perturbation datasets, comprising two single-gene perturbation datasets, Adamson and Papalexi, and two combinatorial perturbation datasets, Norman and Replogle. These datasets cover different perturbation technologies, cellular contexts, and prediction settings. Dataset-specific preprocessing and feature construction are described in Supplementary Appendix D, while the perturbation-level splitting protocol is described in Supplementary Appendix E.

Adamson. The Adamson dataset was generated using Perturb-seq in K562 cells to characterize transcriptional responses associated with the unfolded protein response (Adamson et al. 2016). Each non-control condition corresponds to a single targeted genetic perturbation.

Papalexi. The Papalexi dataset was generated using ECCITE-seq in THP-1 cells to measure the transcriptional consequences of CRISPR perturbations targeting immune-regulatory genes (Papalexi et al. 2021). The dataset contains single-gene perturbation conditions.

Norman. The Norman dataset contains CRISPR activation measurements in K562 cells and was designed to study transcriptional responses and genetic interactions induced by single-gene and combinatorial perturbations (Norman et al. 2019). The formal input uses one shared gene space for all cells and perturbation conditions.

Replogle. The Replogle dataset contains both single-gene and dual-gene perturbations measured using direct guide-RNA capture technology (Replogle et al. 2020).

Data Preprocessing

This section describes data quality control, normalization, gene selection, perturbation embedding construction, and the response reference used in the main-text formulation.

All models and metrics operate on the frozen `logNor` expression layer. The feature space is the union of the 2,048 genes selected by expression variability and the perturbation target genes that are available in the dataset. The resulting gene counts are reported in Table 5. Gene selection is performed once when constructing each frozen input, before the condition-level split is generated; it is therefore shared by all five seeds and all compared methods.

Perturbation embeddings are constructed as defined in Supplementary Appendix A.

Following the main text, a control cell x_j^0 is randomly sampled from the control population for each perturbed cell x_i^p, and the response is defined as

$$\Delta x_i^p = x_i^p - x_j^0. \tag{13}$$

Out-of-Distribution Data Splitting Protocol

We perform perturbation-level splitting to evaluate generalization to unseen perturbation conditions. The non-control perturbations assigned to the training, validation, and test sets are mutually disjoint. Control cells are not treated as prediction targets in the split; instead, they are retained as a shared reference population for constructing perturbation responses and evaluating differential expression.

Single-Gene Perturbation Datasets

For Adamson and Papalexi, all unique non-control perturbation conditions are first sorted and then randomly permuted using a seed-specific NumPy random number generator. Given N non-control conditions, the numbers assigned to the training and validation sets are computed as

$$N_{\text{train}} = \text{round}(0.7N), \qquad N_{\text{val}} = \text{round}(0.1N), \tag{14}$$

and the remaining conditions are assigned to the test set. For datasets with at least three non-control conditions, the implementation ensures that the validation and test sets are non-empty. The resulting conditions within each split are stored in lexicographic order.

Symbol	Definition
Problem formulation	
$\mathcal{D}$	A single-cell genetic perturbation dataset.
X^0	The population of unperturbed control cells.
X^p	The observed cell population under perturbation condition p.
x_j^0, x_i^p	Gene-expression profiles of a control cell and a cell under perturbation p, respectively.
$\mathcal{P}$, p	The set of perturbation conditions and an individual perturbation condition, respectively.
N_0, N_p	The numbers of control cells and cells under perturbation p.
G	The number of selected genes in the expression space.
π_0	The distribution of unperturbed control cells.
$\pi(\cdot \mid p)$	The condition-specific distribution of cells under perturbation p.
f_θ	The stochastic conditional generator parameterized by θ.
Δx_i^p	The observed expression change defined in the main text, $\Delta x_i^p = x_i^p - x_j^0$, where x_j^0 is a randomly sampled control cell.
Response decomposition	
E_s, E_r	The basal-state encoder and response encoder, respectively.
$s_i \in \mathbb{R}^{d_s}$	The basal cellular-state representation of cell i.
$r_i^* \in \mathbb{R}^{d_r}$	The target response latent encoded from the observed expression change Δx_i^p.
d_s, d_r	The dimensions of the basal-state latent and response latent spaces.
g_i	The basal-state-dependent systematic response component.
$e_{p_i} \in \mathbb{R}^{d_e}$	The pretrained embedding of perturbation condition p_i.
d_e	The dimension of the pretrained perturbation embedding.
a_{p_i}	The perturbation-level average response effect derived from e_{p_i}.
$\bar{r}_i$	The deterministic response center, $\bar{r}_i = g_i + a_{p_i}$.
ϵ_i^*	The target cell-specific deviation, $\epsilon_i^* = r_i^* - \bar{r}_i$.
Gaussian-prototype residual modeling	
K	The number of diagonal Gaussian response prototypes.
π_k, μ_k, σ_k^2	The global mixture weight, mean, and diagonal variance of the k-th Gaussian prototype.
c_i	The joint condition representation, $c_i = \mathrm{concat}(s_i, a_{p_i})$.
α_{ik}	The condition-dependent mixture weight of prototype k for cell i.
z_i	A source variable sampled from the condition-dependent Gaussian mixture.
T_θ	The conditional invertible transformation that maps z_i to a cell-specific deviation.
L	The number of additive coupling layers in T_θ.
$\gamma_{ik}^{(\ell)}$	The prototype-matching responsibility of prototype k at coupling layer ℓ.
ϵ_i	The generated stochastic cell-specific deviation.
$\hat{r}_i$	The predicted response latent, $\hat{r}_i = \bar{r}_i + \epsilon_i$.
D_ψ	The expression decoder parameterized by ψ.
$\hat{x}_i^{p_i}$	The predicted post-perturbation expression profile.
Training objectives	
$\mathcal{L}_{\mathrm{align}}$	The response-latent alignment loss.
$\mathcal{L}_{\mathrm{gm}}$	The Gaussian-mixture negative log-likelihood loss.
$\mathcal{L}_{\mathrm{rec}}$	The expression reconstruction loss.
$\mathcal{L}_{\mathrm{dist}}$	The perturbation-condition-specific distribution matching loss.
$\mathcal{L}$	The overall training objective.

Table 3: Summary of the notation used in DeMixPert.

Combinatorial Perturbation Datasets

For Norman and Replogle, we follow the combinatorial perturbation protocol of (Wei et al. 2026). All single-gene perturbation conditions are included in the training set. The dual-gene conditions are randomly permuted, and

$$N_{\mathrm{dual,train}} = \mathrm{round}\left(0.5 N_{\mathrm{dual}}\right) \tag{15}$$

dual-gene conditions are assigned to training. If N_{remain} denotes the number of remaining dual-gene conditions, the validation set receives

$$N_{\mathrm{dual,val}} = \left\lfloor \frac{N_{\mathrm{remain}}}{2} \right\rfloor, \tag{16}$$

and all other remaining dual-gene conditions are assigned to the test set. Therefore, when the number of remaining dual-gene conditions is odd, the test set contains one more condition than the validation set.

Random Seeds and Split Validation

All methods are evaluated using the same five data-splitting seeds:

$$\mathcal{S} = \{17, 23, 29, 31, 37\}. \tag{17}$$

For each dataset and random seed, the generated split is saved before model training and reused by DeMixPert and all comparative methods. The implementation explicitly verifies

Module	Layer dimensions	Activation	Dropout	Norm.
State encoder	G–1024–512–256	SiLU after hidden layers	0.05	LayerNorm after hidden linear layers
Response encoder	G–1024–512–128	SiLU after hidden layers	0.05	LayerNorm after hidden linear layers
Systematic predictor	256–512; 3×(512–1024–512); 512–128	SiLU in blocks	0.05	LayerNorm at each block and output head
Anchor predictor	512–512; 3×(512–1024–512); 512–128	SiLU in blocks	0.05	LayerNorm at each block and output head
Condition network	384–256–256; heads 8 and 32	SiLU	0	None
Coupling flow	$4\times$ additive half-coupling; $8 \to 64$ shift map	None	0	None
Expression decoder	$384 \to G$	MLP-based mapping	–	–

Table 4: Main-text-aligned DeMixPert component summary. Hyphens denote successive linear-layer widths; residual-block repetition is written explicitly.

Dataset	Source	Species	Cell line	Sequencing	Perturbation	Cells	Control cells	Genes	Single	Dual
Adamson	(Adamson et al. 2016)	Human	K562	Perturb-seq	CRISPRi	56,998	7,629	2,061	76	0
Papalexi	(Papalexi et al. 2021)	Human	THP-1	ECCITE-seq	CRISPR knockout	19,340	2,000	2,050	24	0
Norman	(Norman et al. 2019)	Human	K562	Perturb-seq	CRISPRa	96,994	2,000	2,096	101	125
Replogle	(Replogle et al. 2020)	Human	K562	direct-capture Perturb-seq	CRISPRi	26,777	2,000	2,050	33	35

Table 5: Dataset sources and post-filter statistics used in the experiments. "Single" and "Dual" denote non-control perturbation conditions; every dataset additionally contains one shared control condition.

$$\mathcal{P}_{\text{train}} \cap \mathcal{P}_{\text{val}} = \mathcal{P}_{\text{train}} \cap \mathcal{P}_{\text{test}} = \mathcal{P}_{\text{val}} \cap \mathcal{P}_{\text{test}} = \emptyset. \quad (18)$$

This condition-level separation ensures that the reported performance measures generalization to unseen perturbation conditions rather than memorization of cells from perturbations observed during training.

Gaussian-Prototype EM Updates

This section describes the initialization, expectation-maximization procedure, and periodic refresh strategy of the Gaussian response prototypes.

Residual population. The Gaussian mixture model is fitted to training residuals only. With model parameters fixed in evaluation mode, the implementation computes

$$\epsilon_i^* = E_r(\Delta x_i^p) - \bar{r}_i \quad (19)$$

for every non-control training cell. At most the dataset-specific number of residuals in Table 7 is sampled uniformly without replacement. Residual extraction and mixture fitting use `no_grad`; consequently, the EM updates are not differentiated through.

Initialization. For $K = 8$, a seeded random permutation of the N residuals is generated and the first K residual vectors initialize the component means. The mixture weights are initialized uniformly, $\pi_k = 1/K$. Every component receives the same feature-wise population variance, with $\epsilon_{\text{cov}} = 10^{-5}$:

$$\sigma_{kj}^{2(0)} = \max\left\{ \frac{1}{N} \sum_{i=1}^{N} \left(\epsilon_{ij}^* - \bar{\epsilon}_j^*\right)^2, \epsilon_{\text{cov}} \right\}. \quad (20)$$

The covariance is diagonal; no full-covariance or tied-covariance parameter is estimated.

E step. At EM iteration t, let $v_{kj}^{(t)} = \max(\sigma_{kj}^{2(t)}, \epsilon_{\text{cov}})$. Responsibilities are computed in log space:

$$\begin{aligned} \ell_{ik}^{(t)} &= \log \max(\pi_k^{(t)}, 10^{-12}) \\ &\quad - \frac{1}{2} \sum_{j=1}^{d_r} \frac{(\epsilon_{ij}^* - \mu_{kj}^{(t)})^2}{v_{kj}^{(t)}} \\ &\quad - \frac{1}{2} \sum_{j=1}^{d_r} \left[\log v_{kj}^{(t)} + \log(2\pi) \right], \end{aligned} \quad (21)$$

$$\gamma_{ik}^{(t)} = \exp\left[\ell_{ik}^{(t)} - \text{LSE}_{m=1}^{K} \ell_{im}^{(t)} \right], \quad (22)$$

where LSE denotes log-sum-exp.

M step. Let $N_k^{(t)} = \max(\sum_i \gamma_{ik}^{(t)}, 10^{-8})$. The update is

$$\pi_k^{(t+1)} = \frac{N_k^{(t)}}{N}, \tag{23}$$

$$\mu_k^{(t+1)} = \frac{1}{N_k^{(t)}} \sum_i \gamma_{ik}^{(t)} \epsilon_i^*, \tag{24}$$

$$\sigma_{kj}^{2(t+1)} = \max\left\{ \frac{1}{N_k^{(t)}} \sum_i \gamma_{ik}^{(t)} \left(\epsilon_{ij}^* - \mu_{kj}^{(t+1)}\right)^2, \epsilon_{\text{cov}} \right\}. \tag{25}$$

The variance floor is the covariance regularizer; no additional Wishart, entropy, or Dirichlet penalty is used in EM.

Iteration and convergence. The mean log-likelihood computed in the E step,

$$\mathcal{Q}^{(t)} = \frac{1}{N} \sum_i \text{LSE}_{k=1}^K \ell_{ik}^{(t)} \tag{26}$$

is monitored once per EM iteration. EM stops when $|\mathcal{Q}^{(t)} - \mathcal{Q}^{(t-1)}| < 10^{-4}$ or after the dataset-specific maximum number of iterations in Table 7, whichever occurs first.

Empty and near-empty components. Responsibilities are soft, and the effective count is lower-bounded by 10^{-8}. Therefore, a near-empty component remains finite and is not deleted, merged, or randomly reinitialized. If fewer than K residual samples are available, the refresh routine leaves the previous mixture parameters unchanged rather than silently reducing K.

Refresh schedule and numerical stability. The Gaussian mixture model is initialized once at epoch 0 and refreshed after every dataset-specific number of training epochs using newly recomputed residuals. The refresh seed is the run seed at epoch 0 and `run_seed+epoch` thereafter. EM is performed in float32 on CPU and the resulting weights, means, and variances replace non-trainable model buffers. Besides the variance and effective-count floors, all mixture normalization uses log-sum-exp, Gaussian log-probabilities use a minimum variance of 10^{-6} at training/inference, and sampling applies the same 10^{-6} floor before the square root. Conditional mixture weights are produced with softmax; fixed global weights are clamped to 10^{-8} before taking logarithms.

Baseline References and Code

Table 6 lists the paper and official code repository for each comparison method used in the main paper. We omit repository-specific implementation details here and refer readers to the corresponding public releases.

Dataset-Specific Hyperparameters

This section reports the dataset-specific training, validation, and Gaussian-prototype refresh settings used in the experiments.

Method	Paper	Official code
GEARS	https://www.nature.com/articles/s41587-023-01905-6	https://github.com/snap-stanford/GEARS
scGPT	https://www.nature.com/articles/s41592-024-02201-0	https://github.com/bowang-lab/scGPT
GenePert	https://www.biorxiv.org/content/10.1101/2024.10.27.620513v1	https://github.com/zou-group/GenePert
scDFM	https://openreview.net/forum?id=QSGanMEcUV	https://github.com/AI4Science-WestlakeU/scDFM
STATE	https://www.biorxiv.org/content/10.1101/2025.06.26.661135v1	https://github.com/ArcInstitute/state

Table 6: Paper and official code links for the five comparison methods used in the main experiments.

Dataset	LR	Cell batch	Max epochs	Eval int.	EM int.	EM samples	EM iter.	K
Adamson	5×10^{-5}	1024	5000	1000	1000	5000	15	8
Papalexi	5×10^{-5}	4096	5000	100	100	50,000	50	8
Norman	5×10^{-5}	16,384	5000	200	200	20,000	20	8
Replogle	5×10^{-5}	1024	5000	200	100	30,000	50	8

Table 7: Dataset-specific optimization and Gaussian-mixture refresh configuration. "Cell batch" is the number of non-control cells in one optimizer update; gradient accumulation is not used. "Int." is measured in training epochs.

For every dataset, $\lambda_{\text{align}} = \lambda_{\text{rec}} = \lambda_{\text{dist}} = 1$ and $\lambda_{\text{gm}} = 0.01$. The Gaussian mixture model is also refreshed at epoch 0, so an EM interval of 100 denotes updates at epochs $0, 100, 200, \ldots$.

Additional Experimental Results

This section reports the sensitivity analysis of DeMixPert. The complete ablation results are already reported in the main paper; the module-level definitions and forward-path changes are provided in Supplementary Appendix J and are not duplicated as a second figure here.

Ablation Details and Evaluation Metrics

Ablation Implementation Details

The main paper defines four ablated variants:

- **w/o Decomp** replaces the explicit three-part decomposition $\hat{r}_i = g(s_i)+a(e_{p_i})+\epsilon_i$ with a unified response mapping. The response is no longer represented by separately parameterized systematic, target-specific, and population components.
- **w/o Gene** removes the embedding-derived target-specific response by setting $a(e_{p_i}) = \mathbf{0}$. Hence $\bar{r}_i = g(s_i)$; the pretrained embedding-derived deterministic response branch is absent.
- **w/o Sys** removes the basal-state-dependent systematic response by setting $g(s_i) = \mathbf{0}$, yielding $\bar{r}_i = a(e_{p_i})$. The state encoder remains active because s_i is still used by the condition network and expression decoder.
- **w/o GM** removes the entire Gaussian-Prototype Invertible Network, including both the Gaussian mixture and the invertible network. No GMM is fitted, no EM refresh is performed, no coupling transform is applied, and $\epsilon_i = \mathbf{0}$. The prediction therefore reduces to the deterministic response center, $\hat{r}_i = g(s_i) + a(e_{p_i})$. The inapplicable likelihood term $\lambda_{\text{gm}}\mathcal{L}_{\text{gm}}$ is disabled.

The ablation comparison is conducted on Adamson and Replogle. Each variant is independently initialized and retrained rather than produced by post-hoc masking of a trained complete model. Architecture-specific operations and their associated losses are disabled only when inapplicable to the corresponding variant.

Evaluation Protocol and Metric Definitions

We evaluate each test perturbation condition independently and then macro-average the resulting scores across test conditions. Let $\mathbf{X}^p \in \mathbb{R}^{N_p\times G}$ and $\widehat{\mathbf{X}}^p \in \mathbb{R}^{\widehat{N}_p\times G}$ denote the observed and predicted expression matrices under perturbation condition p, respectively. Their population-level mean expression vectors are

$$\boldsymbol{\mu}^p = \frac{1}{N_p}\sum_{i=1}^{N_p}\mathbf{x}_i^p, \qquad \widehat{\boldsymbol{\mu}}^p = \frac{1}{\widehat{N}_p}\sum_{i=1}^{\widehat{N}_p}\widehat{\mathbf{x}}_i^p. \tag{27}$$

We denote the mean expression vector of the observed control population by $\boldsymbol{\mu}^0$.

Top-100 DEG Metrics

Top-100 DEG selection. For each perturbation condition p, genes are ranked by comparing the observed perturbed cells with the observed control cells using the Wilcoxon rank-sum test. Let $\mathcal{S}_p^{\text{true}}$ denote the 100 highest-ranked genes. The same procedure is applied to the predicted perturbed cells and the observed control cells to obtain $\mathcal{S}_p^{\text{pred}}$. The observed set $\mathcal{S}_p^{\text{true}}$ is used as the evaluation gene space for MSE, E-Dist, and W-Dist at the top-100 DEG level.

Common differentially expressed genes (C-DEGs). C-DEGs measures the number of overlapping genes between the top-100 DEG sets obtained from the observed and predicted populations:

$$\text{C-DEGs}_p = \left|\mathcal{S}_p^{\text{true}} \cap \mathcal{S}_p^{\text{pred}}\right|. \tag{28}$$

The score ranges from 0 to 100, with a higher value indicating better recovery of the perturbation-responsive genes. The values reported in the main table may be non-integers because the condition-level counts are macro-averaged.

Mean squared error (MSE). For a gene set $\mathcal{S}$, the population-level MSE is defined as

$$\text{MSE}_p(\mathcal{S}) = \frac{1}{|\mathcal{S}|} \sum_{g \in \mathcal{S}} \left(\widehat{\mu}_g^p - \mu_g^p\right)^2. \tag{29}$$

At the top-100 DEG level, we use $\mathcal{S} = \mathcal{S}_p^{\text{true}}$. A lower value indicates more accurate recovery of the mean perturbation response.

Energy distance (E-Dist). Energy distance evaluates the discrepancy between the complete predicted and observed cell populations. Let $\widehat{\mathbf{x}}, \widehat{\mathbf{x}}'$ be independent samples from the predicted distribution and $\mathbf{x}, \mathbf{x}'$ be independent samples from the observed distribution. Energy distance is defined as

$$\begin{aligned}\text{E-Dist}_p = 2\mathbb{E}\left[\|\widehat{\mathbf{x}} - \mathbf{x}\|_2\right] - \mathbb{E}\left[\|\widehat{\mathbf{x}} - \widehat{\mathbf{x}}'\|_2\right] \\ - \mathbb{E}\left[\|\mathbf{x} - \mathbf{x}'\|_2\right].\end{aligned} \tag{30}$$

For the top-100 DEG result, all expression vectors are restricted to $\mathcal{S}_p^{\text{true}}$. Lower E-Dist indicates better agreement between the predicted and observed distributions. The code uses the empirical V-statistic

$$2\overline{d}(\widehat{X}^p, X^p) - \overline{d}(\widehat{X}^p, \widehat{X}^p) - \overline{d}(X^p, X^p), \tag{31}$$

where each bar is the mean of the complete pairwise Euclidean-distance matrix, including its zero diagonal for within-population terms. The result is clamped below at zero. Each population is subsampled without replacement to at most 2,000 cells with a fixed condition-index seed.

Wasserstein distance (W-Dist). W-Dist measures the optimal-transport cost between the empirical predicted and observed cell distributions. Using the squared Euclidean ground cost, it is calculated as

$$\text{W-Dist}_p = \left[\sum_{i,j} \gamma_{ij}^\star \left\|\widehat{\mathbf{x}}_i^p - \mathbf{x}_j^p\right\|_2^2\right]^{1/2}, \tag{32}$$

where $\gamma^\star$ denotes the transport coupling returned by the shared optimal-transport implementation. W-Dist is evaluated on $\mathcal{S}_p^{\text{true}}$, and a lower value indicates greater distributional similarity. The primary backend is `Pertpy Distance` with the Wasserstein metric; no additional cell subsampling is applied. If Pertpy is unavailable, the released fallback computes an entropically regularized Sinkhorn cost with $\varepsilon = 0.08$ and 35 iterations and reports the square root of the non-negative cost.

All-Gene Metrics

Differential Expression Score (DES). For each perturbation p, significant DEGs are identified separately from the observed and predicted populations relative to the same observed control population. We use the Wilcoxon rank-sum test with Benjamini–Hochberg correction and an adjusted p-value threshold of 0.05. Let $\mathcal{D}_p^{\text{true}}$ and $\mathcal{D}_p^{\text{pred}}$ denote the resulting significant gene sets.

When the predicted set contains more genes than the observed set, it is truncated to the $|\mathcal{D}_p^{\text{true}}|$ genes with the largest absolute log-fold changes. Denoting the resulting predicted set by $\widetilde{\mathcal{D}}_p^{\text{pred}}$, DES is

$$\text{DES}_p = \frac{\left|\mathcal{D}_p^{\text{true}} \cap \widetilde{\mathcal{D}}_p^{\text{pred}}\right|}{\left|\mathcal{D}_p^{\text{true}}\right|}. \tag{33}$$

Conditions with no significant observed DEGs are excluded from the DES average. A higher DES indicates more accurate recovery of statistically significant perturbation-responsive genes.

All-gene MSE. All-gene MSE uses the same definition as above, but is calculated over the complete frozen gene space rather than $\mathcal{S}_p^{\text{true}}$.

Centroid accuracy. For every test perturbation, we compare its predicted population centroid with the observed centroids of all test perturbations:

$$\text{CA}_p = \mathbf{1}\left[\arg\min_{q \in \mathcal{T}} \|\widehat{\boldsymbol{\mu}}^p - \boldsymbol{\mu}^q\|_2 = p\right], \tag{34}$$

where $\mathcal{T}$ is the set of test perturbation conditions. The final centroid accuracy is the mean of CA_p across conditions. A higher value means that predicted populations preserve condition-specific identities more accurately.

Perturbation Discrimination Score (PDS). The main comparison table denotes this metric as PDS. The L1-based implementation below evaluates whether the predicted response of condition p is most similar to the corresponding observed response rather than to another test perturbation. We first define the predicted and observed perturbation-effect vectors as

$$\widehat{\boldsymbol{\delta}}^p = \widehat{\boldsymbol{\mu}}^p - \boldsymbol{\mu}^0, \qquad \boldsymbol{\delta}^q = \boldsymbol{\mu}^q - \boldsymbol{\mu}^0. \tag{35}$$

For every pair (p, q), the directly perturbed target genes of both conditions are excluded, producing the comparison gene set $\mathcal{G}_{p,q}$. The L1 distance is

$$d_{p,q} = \sum_{g \in \mathcal{G}_{p,q}} \left|\widehat{\delta}_g^p - \delta_g^q\right|. \tag{36}$$

Let r_p be the rank of $d_{p,p}$ among $\{d_{p,q} : q \in \mathcal{T}\}$ in ascending order. The condition-level score is

$$\text{PDS}_p = 1 - \frac{r_p - 1}{|\mathcal{T}|}. \tag{37}$$

The final score is averaged across all test perturbations. A higher PDS indicates better discrimination between different

perturbation responses. Ties are resolved deterministically by the lexicographic condition name. If target-gene removal would eliminate the entire feature space, the implementation falls back to all genes. This L1-based quantity may be named `PDS-L1` in evaluator artifacts, but it is reported as PDS in the main paper.

Aggregation protocol. Across the five seeds defined in Supplementary Appendix E, the main table reports the arithmetic mean and sample standard deviation ($\mathrm{ddof} = 1$) across these five trials. DES conditions with no significant observed genes are represented by NaN and omitted from the DES macro-average. The per-condition evaluation caps are 64, 512, 256, and 256 cells for Adamson, Papalexi, Norman, and Replogle, respectively. All methods are evaluated using the same frozen evaluation implementation.

Reproducibility Details

Software environment. DeMixPert is implemented in Python and PyTorch. Data processing and evaluation use AnnData, Scanpy, NumPy, pandas, and SciPy.

Random seeds. The five split seeds defined in Supplementary Appendix E are also used for model initialization. For every run, the Python, NumPy, PyTorch, and CUDA random-number generators are initialized using the corresponding seed.

Model selection. Model selection is performed exclusively on the validation set. Specifically, the checkpoint with the lowest macro-averaged top-100 DEG MSE on the validation perturbation conditions is selected. The test set is evaluated only after checkpoint selection and is never used for model selection or hyperparameter tuning. The same checkpoint-selection rule is applied to every random seed.

Details of the Interpretability Experiments

Separability of the Decomposed Response Representation

We first examined whether the response decomposition isolates perturbation-specific information from variation driven by the basal cell state. For a perturbed cell i, the response representation was obtained from the expression difference between the perturbed cell x_i and its paired control cell c_i:

$$r_i^* = \mathrm{Enc}_{\mathrm{response}}(x_i - c_i). \tag{38}$$

The model decomposes this representation as

$$r_i^* \approx g_i + a_p + \epsilon_i, \tag{39}$$

where g_i denotes the systematic response associated with the basal cell state, a_p is the perturbation-specific anchor for perturbation p, and ϵ_i captures cell-level residual variation. The systematic component is predicted from the latent state of the paired control cell:

$$s_i = \mathrm{Enc}_{\mathrm{state}}(c_i), \qquad g_i = f_{\mathrm{sys}}(s_i). \tag{40}$$

To assess the effect of removing state-driven variation, we compared the following two response representations:

$$r_i^{\mathrm{raw}} = r_i^*, \tag{41}$$

$$r_i^{\mathrm{specific}} = r_i^* - g_i. \tag{42}$$

The second representation removes the systematic component predicted from the basal cell state while retaining the perturbation-specific anchor and cell-level residual variation.

Experimental protocol. All analyses used the most recent baseline checkpoint selected according to validation performance (`best.pt`). Checkpoints selected using test performance (`best_test.pt`), as well as checkpoints from sensitivity or ablation experiments, were not used.

We evaluated the Adamson and Replogle datasets using five random seeds:

$$\mathcal{S} = \{17, 23, 29, 31, 37\}\,. \tag{43}$$

For each dataset–seed combination, the analysis was restricted to perturbations in the test split. To prevent perturbations with larger cell populations from disproportionately influencing the clustering results, we randomly sampled an equal number of cells from every test perturbation.

For every sampled cell, we computed the raw response latent r_i^*, the control-state latent s_i, and the corresponding systematic response g_i. The specific response was subsequently obtained as $r_i^* - g_i$.

Within each dataset–seed combination, the raw and specific response representations were concatenated when fitting a `StandardScaler`. The resulting common scaling transformation was then applied to both representations. This ensured that the raw and specific responses were evaluated using the same feature scaling.

K-means clustering was performed separately on the standardized raw and specific representations. The number of clusters was set equal to the number of held-out perturbations:

$$K = \#\{\text{test perturbations}\}. \tag{44}$$

The resulting cluster assignments were evaluated against the ground-truth perturbation labels using the following metrics:

- **Adjusted Rand Index (ARI):** agreement between the inferred clustering and the true perturbation partition, adjusted for chance.
- **Normalized Mutual Information (NMI):** normalized information shared by the inferred cluster assignments and the ground-truth perturbation labels.
- **Purity:** the proportion of correctly assigned cells when each cluster is labeled according to its most frequent ground-truth perturbation.
- **Silhouette score:** the compactness and separation of the inferred clusters in the response latent space.

All quantitative metrics were computed directly in the standardized, high-dimensional response latent space. UMAP projections were used only for visualization and were not used for clustering or metric calculation.

Higher ARI, NMI, purity, and silhouette scores for r_i^{specific} than for r_i^{raw} indicate that subtracting g_i removes basal-state-driven variation and makes the perturbation-specific response structure more separable.

Association Between Prototype Weights and Residual Population Differences

We next investigated whether the Gaussian prototype weights learned by the model reflect genuine differences between perturbation-specific residual cell populations. Specifically, we tested whether perturbation pairs with more dissimilar prototype-weight profiles also exhibit more dissimilar residual-response distributions.

For perturbation p under control-cell state s_i, the mixture weights over the eight Gaussian prototypes are given by

$$\alpha_{i,p} = \text{softmax}\left(h_{\text{mix}}\left([s_i, a_p]\right)\right), \tag{45}$$

where a_p is the perturbation embedding and $\alpha_{i,p} \in \mathbb{R}^8$ contains the corresponding prototype weights.

For each perturbation, we computed its average prototype-weight profile:

$$\bar{\alpha}_p = \frac{1}{M}\sum_{i=1}^{M} \alpha_{i,p} = \frac{1}{M}\sum_{i=1}^{M} \text{softmax}\left(h_{\text{mix}}\left([s_i, a_p]\right)\right), \tag{46}$$

where M is the number of control cells used to calculate the profile.

Importantly, the same set of M control cells was used for every perturbation within a dataset–seed combination. This control ensures that differences between the profiles $\bar{\alpha}_p$ are attributable to the perturbations rather than to differences in basal cell-state composition.

Experimental protocol. We used the validation-selected baseline checkpoint (`best.pt`). The main visualization was generated using the training split for Adamson seed 23 and Replogle seed 29.

For each pair of perturbations (p, q), prototype-weight dissimilarity was quantified using Jensen–Shannon divergence:

$$D_\alpha(p, q) = \text{JS}\left(\bar{\alpha}_p, \bar{\alpha}_q\right). \tag{47}$$

We then derived the observed residual response for every perturbed cell. The encoded response was

$$r_i^* = \text{Enc}_{\text{response}}(x_i - c_i), \tag{48}$$

and the residual response was calculated by subtracting the systematic response and perturbation anchor:

$$\epsilon_i^* = r_i^* - (g_i + a_p). \tag{49}$$

Let E_p denote the collection of residual-response vectors associated with perturbation p. To prevent differences in cell counts from biasing the population-distance estimates, we sampled the same number of residual cells for every perturbation. Specifically, we used 30 cells per perturbation for Adamson seed 23 and 51 cells per perturbation for Replogle seed 29.

The balanced sampling procedure was independently repeated 20 times. For each repetition t, the distance between the residual populations of perturbations p and q was quantified using Energy Distance:

$$\text{ED}\left(E_p^{(t)}, E_q^{(t)}\right). \tag{50}$$

The final residual-population distance was calculated by averaging over the 20 sampling repetitions:

$$D_\epsilon(p, q) = \frac{1}{20}\sum_{t=1}^{20} \text{ED}\left(E_p^{(t)}, E_q^{(t)}\right). \tag{51}$$

The Adamson training split contained 53 perturbations and therefore yielded

$$\binom{53}{2} = \frac{53 \times 52}{2} = 1{,}378 \tag{52}$$

unique perturbation pairs. The Replogle training split contained 51 perturbations and yielded

$$\binom{51}{2} = \frac{51 \times 50}{2} = 1{,}275 \tag{53}$$

unique pairs. Each perturbation pair contributed one observation of the form

$$\left(D_\alpha(p, q), D_\epsilon(p, q)\right). \tag{54}$$

We quantified the association between the two distance matrices by computing Spearman's rank correlation over their upper-triangular elements:

$$\rho = \text{Spearman}\left(\text{upper}(D_\alpha), \text{upper}(D_\epsilon)\right). \tag{55}$$

Because pairwise distances are not statistically independent, significance was assessed using a Mantel-style permutation test rather than the conventional p-value returned by a standard Spearman correlation test. In each permutation, the perturbation labels of D_α were randomly reordered by jointly permuting its rows and columns, while D_ϵ was kept fixed. Spearman's ρ was then recomputed. This procedure was repeated 5,000 times, and the resulting permutation distribution was used to calculate the empirical p-value.

The resulting associations were

$$\text{Adamson, seed 23:} \quad \rho = 0.462, \quad p_{\text{Mantel}} < 0.001, \tag{56}$$

$$\text{Replogle, seed 29:} \quad \rho = 0.512, \quad p_{\text{Mantel}} < 0.001. \tag{57}$$

Visualization. For each dataset, the prototype-weight heatmap displays perturbations as rows and the eight Gaussian prototypes as columns. Each heatmap entry represents

$$\bar{\alpha}_{p,k}, \tag{58}$$

namely the average weight assigned by perturbation p to Gaussian prototype k. Perturbations were ordered by hierarchical clustering according to the similarity of their prototype-weight profiles.

The distance-association panel displays $D_\alpha(p, q)$ against $D_\epsilon(p, q)$ for all perturbation pairs. Because of the large number of overlapping observations, the pairs were visualized using a hexagonal-binning density plot. A LOWESS curve was added to show the overall trend. This curve was used only as a visual guide and did not contribute to the statistical significance test. The figure reports only Spearman's ρ and the Mantel permutation p-value.

The significant positive associations demonstrate that perturbations with more dissimilar prototype-weight profiles also tend to have more dissimilar residual cell populations. Thus, the learned Gaussian mixture weights are not arbitrary internal coefficients but encode interpretable, perturbation-specific population structure.

Construction of Sensitivity Scores

To summarize metrics with different scales and optimization directions in the sensitivity analysis presented in the main paper, we convert each metric into a direction-aligned percentage change relative to the selected configuration, $K = 8$ and $\lambda_{\text{gm}} = 0.01$. Let $x_m(p)$ denote the five-seed mean of metric m under parameter setting p, and let x_m^{ref} denote its value under the selected configuration. The transformed score is computed as

$$\Delta_m(p) = \begin{cases} 100\left(\dfrac{x_m(p)}{x_m^{\text{ref}}} - 1\right), & \text{if metric } m \text{ is higher-is-better,} \\ 100\left(1 - \dfrac{x_m(p)}{x_m^{\text{ref}}}\right), & \text{if metric } m \text{ is lower-is-better.} \end{cases} \tag{59}$$

The higher-is-better metrics are C-DEGs, DES, Centroid Accuracy, and PDS-L_1, whereas E-Dist, W-Dist, and MSE are lower-is-better. We then obtain each curve by taking the unweighted arithmetic mean of its constituent transformed metrics. Specifically, DEG recovery averages C-DEGs and DES; Distribution averages top-100 E-Dist, top-100 W-Dist, and all-gene E-Dist; Expression error averages top-100 MSE and all-gene MSE; and Perturbation ID averages Centroid Accuracy and PDS-L_1. All-gene E-Dist is used only as an auxiliary distributional measure in the sensitivity analysis and is not included in the primary benchmark table. It is computed using the same energy-distance definition over the complete input gene space. Consequently, positive values indicate improvement over the selected configuration, negative values indicate degradation, and the selected configuration corresponds to 0%. During each sensitivity analysis, one hyperparameter is varied while all remaining settings are held fixed.