REGINA: REGULARIZED ENCODER WITH LATENT CYCLE-GAN FOR IN-VITRO NEURAL CELL PERTURBATION APPROXIMATION.

Anonymous authors
Paper under double-blind review

ABSTRACT

Counterfactual inference from unpaired observational data is a fundamental challenge in machine learning. A prominent example arises in single-cell biology, where predicting cellular responses to genetic perturbations is inherently unpaired, as each cell is destroyed during measurement. Existing approaches address this challenge using heuristic assumptions, such as restricting the input space to highly variable genes or incorporating external biological priors. However, these methods frequently optimize for mean point-estimates, resulting in catastrophic output variance collapse and a failure to capture biological heterogeneity. We introduce REGINA, a fully data-driven framework that formulates perturbation prediction as an unpaired distribution matching problem in a learned latent space. By combining structured latent representation learning with adversarial, cycle-consistent mappings between control and perturbed populations, REGINA captures non-linear, population-level transformations without requiring paired observations. A prompt-based latent conditioning mechanism further enables robust compositional generalization to novel perturbation combinations. Empirically, REGINA maintains competitive local prediction accuracy on key differentially expressed genes while generating biologically realistic single-cell distributions (evaluated via MMD and Wasserstein distance).

1 INTRODUCTION

Counterfactual inference from unpaired observational data is a fundamental challenge in machine learning Park & Kellis (2021). The goal is to predict the effect of an intervention on an individual unit when only population-level data from control and target groups are available, without any paired samples. This setting arises naturally in many scientific domains, from economics to personalized medicine. A compelling example is single-cell biology, where researchers aim to predict how a cell's gene expression profile – its transcriptional state – changes after a genetic perturbation (e.g., a gene knockout). Because the measurement process destroys the cell, it is impossible to observe the same cell both before and after intervention. Consequently, we only have unpaired control and perturbed cell populations. Despite the importance of this problem, existing machine learning approaches rely on restrictive heuristics. Some models like GEARS Roohani et al. (2024) incorporate biological knowledge graphs, introducing dependencies on incomplete databases. Methods like scGen Lotfollahi et al. (2019) and CPA model Lotfollahi et al. (2023) perturbations as additive shifts in a latent space, which cannot capture complex, non-linear interactions (e.g., epistasis).

We propose **REGINA** (**R**egularized **E**ncoder with Latent Cycle-**G**AN for **I**n-vitro **N**eural Cell Perturbation **A**pproximation), a fully data-driven framework that directly addresses the core challenge: **unpaired distribution matching in a learned latent space**. REGINA learns structured latent representations, in which control and perturbed populations can be related through non-linear, bidirectional transformations. This enables the model to capture population-level effects and complex regulatory responses without requiring paired observations. Furthermore, by conditioning these transformations on latent perturbation signals, the framework achieves robust compositional generalization to entirely unseen interventions.

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

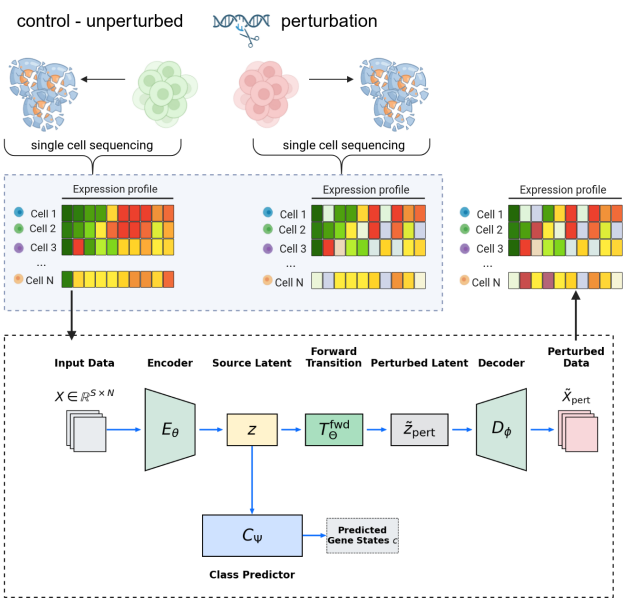

Figure 1: REGINA workflow:Unpaired control and perturbed cells are encoded into a latent space regularized by a classifier and center loss. A prompt-conditioned Cycle-GAN learns bidirectional transitions, and the decoder reconstructs the counterfactual expression.

We evaluate REGINA on genome-scale Replogle et al. (2022) and combinatorial Norman et al. (2019) Perturb-seq benchmarks. Our results show that REGINA achieves strong local prediction accuracy (Pearson correlation) while improving global distributional fidelity (Maximum Mean Discrepancy) compared to GEARS and CPA. Accurate modeling of global distributions is essential in virtual-cell perturbation settings, where interventions alter not only mean expression but also population heterogeneity and subpopulation structure, which are not captured by pointwise prediction metrics. Moreover, we identify an inherent trade-off between local precision and global variance preservation.

2 Problem Formulation

Let $\mathcal{X} \subset \mathbb{R}^d$ be the space of cell states (gene expression vectors). We observe two unpaired empirical distributions, a control distribution $\mathbb{P}_{\text{ctrl}}$ over $\mathcal{X}$ and for each perturbation $p \in \mathcal{P}$, a $\mathbb{P}^{(p)}_{\text{pert}}$ over $\mathcal{X}$.

We never have a pair $(X_{\text{ctrl}}, X^{(p)}_{\text{pert}})$ from the same cell, therefore it's a classic unpaired distribution matching problem.

Our goal is to learn a counterfactual transition operator $T_p : \mathcal{X} \to \mathcal{X}$ such that

$$T_p(\mathbb{P}_{\text{ctrl}}) = \mathbb{P}^{(p)}_{\text{pert}},$$

yields the target distribution by pushing the control distribution forward. For unseen perturbations $p' \notin \mathcal{P}_{\text{train}}$, we require the model to infer the resulting cellular state given only a descriptor of p'.

This is an instance of unpaired distribution matching. The challenge is to learn T_p without any paired examples and without heuristic input pruning.

3 Related Work

The cycle-consistent adversarial framework (CycleGAN) Zhu et al. (2017a) introduced the idea of learning bidirectional mappings with a cycle-consistency loss, enabling unpaired image-to-image translation. Subsequent extensions have applied this paradigm to other domains, including cross-modal retrieval Hiasa et al. (2018) and domain adaptation Mütze et al. (2022). While effective for

visual and structured data, these methods typically operate directly in the input or feature space and do not address the specific challenges of high-dimensional, sparse genomic data.

Modeling cellular responses to genetic perturbations from single-cell transcriptomic data has progressed from linear effect models to deep generative approaches Dixit et al. (2016). Despite strong empirical performance, existing methods rely on structural assumptions that limit their ability to capture complex, population-level perturbation dynamics.

Variational autoencoders (VAEs), such as scVI Lopez et al. (2018), provide a standard framework for learning low-dimensional representations of high-dimensional single-cell data. scGen Lotfollahi et al. (2019) models perturbations as vector arithmetic in latent space, implicitly assuming that perturbation effects correspond to global linear shifts.

Compositional Perturbation Autoencoders (CPA) Lotfollahi et al. (2021) extend this framework by disentangling basal cell states from perturbation effects using adversarial objectives and enabling compositional generalization. However, perturbations are still represented as additive latent components, limiting the ability to model complex interactions such as epistasis. Methods such as GEARS Roohani et al. (2024) incorporate external biological priors to improve generalization by leveraging gene ontology graphs within a graph neural network to predict unseen perturbations. While effective, this paradigm introduces dependencies on the completeness and accuracy of biological databases. Furthermore, to reduce computational complexity, such methods typically restrict the input space to highly variable genes (HVGs) Zhao et al. (2025b), discarding a substantial portion of the transcriptomic landscape and potentially removing subtle signals and rare subpopulation structure Yip et al. (2019). Recent studies Atti & Subramaniam (2025); Csendes et al. (2025) suggest that foundation models do not consistently outperform task-specific approaches in perturbation prediction, and their evaluation often requires different assumptions and training setups. Accordingly, we restrict our comparison to established task-specific baselines.

Because single-cell measurements are destructiveJost et al. (2026); Chen et al. (2022), perturbation modeling is inherently unpaired, requiring comparisons between control and perturbed populationsPark & Kellis (2021); Roohani et al. (2024); Lotfollahi et al. (2019). Prior approaches address this either through conditional generative modelsShankar (2026) or latent transformationsWang et al. (2026), but generally rely on implicit alignment assumptions or supervised signals derived from paired or pseudo-paired data Bunne et al. (2023).

4 METHOD

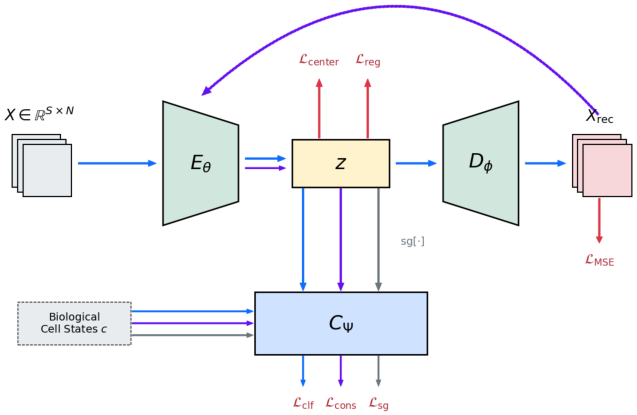

Figure 2: Training pipeline of the regularized autoencoder. Red text indicates the losses computed during training. Arrow colors correspond to the gradient flow of different variables.

4.1 PHASE ONE: REGULARIZED LATENT AUTOENCODER

Our approach follows the general paradigm of latent diffusion models Rombach et al. (2022), in which high-dimensional observations are first mapped into a lower-dimensional latent space via an autoencoder.

Given the high dimensionality of gene expression profiles, processing the full vector with standard linear layers is computationally heavy. To address this, we tokenize the input vector into a sequence of lower-dimensional segments. Our tokenization is a naive fixed-length split of the expression vector into consecutive segments. These segments are then processed by a Transformer encoder Vaswani et al. (2017), utilizing bidirectional self-attention Devlin et al. (2018) to capture complex, non-local gene dependencies.

Let $X \in \mathbb{R}^N$ denote a gene expression vector sampled from either the control or a perturbed population during training. An encoder E_θ projects X into a latent representation z, which is subsequently reconstructed as X_{rec} by a decoder D_ϕ:

$$X_{\text{rec}} = D_\phi(E_\theta(X)). \tag{1}$$

A standard autoencoder does not explicitly enforce preservation of biologically meaningful class information, such as the cellular state. To be able to study perturbation effects in the latent space, we augment the model with a latent classifier C_Ψ that predicts the auxiliary cell state directly from the latent representation, i.e. $\hat{c}_i = C_\Psi(z_i)$. Crucially, this classifier is not designed to predict the perturbation itself. Instead, it captures intrinsic biological properties relevant to virtual cell analysis, such as the cell cycle phase, the population of origin, or whether the cell is under stress. We decided not to place it in the reconstruction space due to the high dimensionality of virtual cell modeling.

Latent Space Regularization. The classifier is trained using a stop-gradient operation on the encoder output, preventing its updates from affecting the encoder:

$$\mathcal{L}_{\text{sg}} = \text{CE}\big(C_\Psi(\text{sg}[E_\theta(X)]), c\big), \tag{2}$$

where $\text{sg}[\cdot]$ denotes the stop-gradient operator, and CE is the cross-entropy loss. This loss is used exclusively for the classifier's training step.

Training a vanilla autoencoder typically results in an unregularized latent space that is unsuitable for downstream perturbation modeling Kingma & Welling (2013). To address this, we apply a *center loss* ($\mathcal{L}_{\text{CE}}$) Wen et al. (2016), which explicitly minimizes the intra-class variance by penalizing the Euclidean distance between each latent representation and the learned center of its biological state. This encourages the latent space to form compact, tightly grouped clusters. We further employ Kullback-Leibler divergence as an auxiliary regularization term ($\mathcal{L}_{\text{reg}}$)between the learned latent distribution and a standard normal prior.

To additionally regularize the encoder, we include a second classification term without the stop-gradient:

$$\mathcal{L}_{\text{CLF}} = \text{CE}\big(C_\Psi(E_\theta(X)),\, c\big). \tag{3}$$

This term encourages the encoder to produce latents that are themselves predictive of the cell state, while the classifier receives gradients from both terms (which only improves its accuracy and does not harm training dynamics). While the stop-gradient objective isolates the updates to the classifier, this second term enables active gradient flow back to the encoder, explicitly forcing it to learn class-separable representations.

The full autoencoder objective is then:

$$\mathcal{L}_{\text{AE}} = \lambda_{\text{MSE}}\mathcal{L}_{\text{MSE}} + \lambda_{\text{CE}}\mathcal{L}_{\text{CE}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}} + \lambda_{\text{CLF}}\mathcal{L}_{\text{CLF}}, \tag{4}$$

where $\mathcal{L}_{\text{MSE}}$ is the reconstruction loss, $\mathcal{L}_{\text{CE}}$ the center loss, $\mathcal{L}_{\text{reg}}$ combines auxiliary regularization terms, and the λ coefficients are scalar hyperparameters.

Latent Consistency. To ensure training stabilityGulrajani et al. (2017); Arjovsky & Bottou (2017) and a self consistent encoder and decoder we applied consistency monitoring. A latent sample z_i is decoded and re-encoded as $z_{\text{rec}} = E_\theta(D_\phi(z_i))$, and both representations are required to produce consistent class predictions:

$$\mathcal{L}_{\text{cons}} = \text{CE}\big(C_\Psi(z_i),\, C_\Psi(z_{\text{rec}})\big), \tag{5}$$

where $\mathrm{CE}(\cdot)$ denotes the cross-entropy loss. The training is stopped if this error starts to constantly increase.

4.2 Second Phase: Latent Transition Modeling

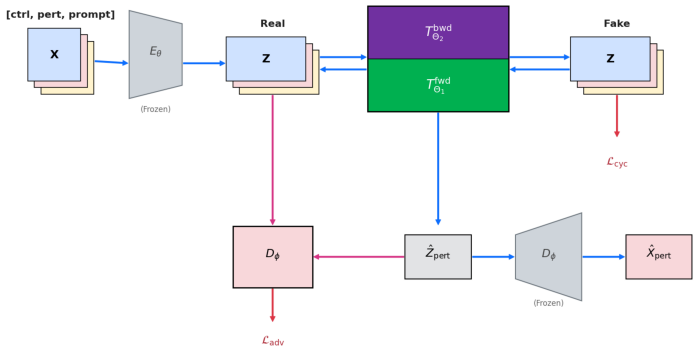

Figure 3: Training pipeline of the second phase of training. Red text indicates losses. Arrow colors correspond to gradient flow. Frozen modules and their labels are gray.

Due to the absence of paired control–perturbation samples, a supervised latent transition model cannot be trained directly. We therefore adopt a *latent cycle GAN* framework Zhu et al. (2017b); Nigam et al. (2025) that learns bidirectional mappings between control latent distributions Z_{ctrl} and perturbed latent distributions Z_{pert}.

A *forward transition block* $T^{\text{fwd}}_{\Theta_1}$ maps control latents to perturbed latents, while a *backward transition block* $T^{\text{bwd}}_{\Theta_2}$ performs the inverse mapping:

$$Z_{\text{pert}} = T^{\text{fwd}}_{\Theta_1}(Z_{\text{ctrl}}), \qquad Z_{\text{ctrl}} = T^{\text{bwd}}_{\Theta_2}(Z_{\text{pert}})$$

Since the true inverse of a biological perturbation is unknown, the backward transition model is tasked with learning an implicit inversion corresponding to the perturbation that generated the observed perturbed state. Converting perturbed state to control state has no biological meaning, it is a mathematical constraint for cycle-consistency to regularize the latent manifold, rather than a biological claim. To ensure consistency between these transformations, we apply a *cycle-consistency loss* $\mathcal{L}_{\text{cycle}}$ that enforces the reconstruction of latent vectors after forward–backward and backward–forward passes. This loss utilizes MSE to minimize the distance between the original latents and their cycled counterparts:

$$\mathcal{L}_{\text{cycle}} = \text{MSE}\left(Z_{\text{ctrl}}, T^{\text{bwd}}_{\Theta_2}(T^{\text{fwd}}_{\Theta_1}(Z_{\text{ctrl}}))\right) + \text{MSE}\left(Z_{\text{pert}}, T^{\text{fwd}}_{\Theta_1}(T^{\text{bwd}}_{\Theta_2}(Z_{\text{pert}}))\right)$$

4.2.1 Adversarial Training

Without additional constraints, the transition model may collapse to a trivial identity mapping, i.e., $T(Z) = Z$. To prevent this degeneration, we apply an adversarial discriminator D_ϕ operating in latent space Goodfellow et al. (2014), which distinguishes real latent vectors Z_{real} from transformed vectors $Z_{fake} = T(Z)$. To ensure that the learned transformations are perturbation-specific, the discriminator is explicitly conditioned on the perturbation index p, which is provided as a one-hot-encoded vector corresponding to the perturbed gene's position. Following Mao et al. Mao et al. (2016), we optimize D_ϕ using a Mean Squared Error (MSE) objective $\mathcal{L}_D$ to stabilize the adversarial training process.

$$\mathcal{L}_D = \text{MSE}(D_\phi(Z_{\text{real}}, p), 1) + \text{MSE}(D_\phi(T(Z_{\text{fake}}), p), 0)\,, \tag{6}$$

where $T(\cdot)$ denotes either the forward or backward transformation, Z_{real} is the latent vectors of real gene expression vectors and Z_{fake} is the latent vectors generated by our model. The transition model is then trained using the corresponding generator loss $\mathcal{L}_G$, which encourages the model to produce transformed latents that the discriminator classifies as "real" (target value of 1):

$$\mathcal{L}_G = \text{MSE}\left(D_\phi(T(Z_{\text{real}}), p), 1\right).$$

4.2.2 PROMPT-BASED CONDITIONING FOR UNSEEN PERTURBATIONS

To address unseen perturbations during training, we applied a *latent prompting* Li & Liang (2021); Lester et al. (2021) that explicitly encodes perturbation information. Given a perturbation index p, which denotes the index of the perturbed gene in the target gene expression vector, a synthetic gene expression vector X_{fake} is constructed by copying the control expression and assigning a fixed value of -1 to the perturbed gene at index p. This value is chosen because it does not appear in real expression vectors, serving as a clear signal that a perturbation has been applied. If multiple gene are perturbed the assigned value is also going to be multiplied with that amount. The corresponding latent embedding $Z_{\text{fake}} = E_\theta(X_{\text{fake}})$ is used to define a perturbation prompt $Z_{\text{prompt}} = Z_{\text{ctrl}} - Z_{\text{fake}}$. We employ this method to guide the generators of our model, since the effect of the perturbation is not deterministic even on the target gene. The transition model is conditioned on this prompt as

$$Z_{\text{pert}}^{\text{fwd}} = T_{\Theta_1}^{\text{fwd}}(Z_{\text{ctrl}}, Z_{\text{prompt}}), \qquad Z_{\text{ctrl}}^{\text{bwd}} = T_{\Theta_2}^{\text{bwd}}(Z_{\text{pert}}, Z_{\text{prompt}}). \tag{7}$$

The same prompt is used for both forward and backward transformations.

5 EXPERIMENTS

To rigorously assess the generalization capability of our models, we evaluate their ability to predict cellular responses to unseen perturbations. Rather than employing a naive random split across individual cells, which risks data leakage and overestimates model capacity, we split the dataset such that specific perturbations are entirely held out from training. The model is subsequently evaluated on its ability to infer the transcriptomic shifts induced by these novel conditions, thereby testing its capacity for robust compositional generalization. Furthermore, because global transcriptomic profiles are inherently noisy, it is standard practice to focus the evaluation on the genes most responsive to a given perturbation. Accordingly, we compute our gene-specific metrics primarily on the top 20 most differentially expressed genes (DEGs) relative to the unperturbed control state Wolf et al. (2017); Zhao et al. (2025a). For a fair comparison, we utilize the exact data splits provided by the GEARS pipeline to train and evaluate all models. All reported metrics represent the mean ± standard deviation aggregated across the held-out perturbations. Following standard benchmarks in virtual cell modeling Lotfollahi et al. (2019); Roohani et al. (2024), we quantify local prediction precision using the Normalized Mean Squared Error (NMSE), Pearson correlation, and directional error of the perturbation effect relative to the control baseline. To evaluate global distributional fidelity, we compute the Maximum Mean Discrepancy (MMD), Wasserstein distance, Energy distance, and the variance correlation between the predicted and ground-truth perturbed cell populations. Lower values indicate better performance for all metrics, with the exception of Pearson and variance correlations, where higher is better.

5.1 SINGLE GENE PERTURBATIONS

We evaluate REGINA on two standard Perturb-seq benchmarks for single-cell perturbation: the Adamson dataset Adamson et al. (2016) and the large-scale Replogle RPE1 screen Replogle et al. (2022). Adamson validates robust local prediction, while Replogle's massive scale (236,429 single-cell transcriptomes, 11,485 controls, and 8,749 measured genes) provides the high-dimensional complexity required to rigorously test full joint distribution matching.

5.2 MULTI-GENE PERTURBATIONS AND COMPOSITIONAL GENERALIZATION

To evaluate REGINA's capacity for compositional generalization, we extend our benchmarking to the Norman dataset Norman et al. (2019), which comprises both single and pairwise genetic perturbations. From a representation learning perspective, predicting these combinatorial effects presents a rigorous challenge for out-of-distribution (OOD) generalization, as the model must extrapolate beyond simple additive logic to capture complex, non-linear transcriptomic interactions. Specifically, we evaluate the model on holdout pairwise combinations where only the individual constituent perturbations were observed during training. Demonstrating high fidelity on this task proves that REGINA's unpaired matching approach naturally captures the complex, non-linear geometry of combinatorial perturbation spaces.

Table 1: Precision metrics on the Adamson and Replogle datasets. Best results are bolded, and second-best are underlined.

Dataset	Model	Pearson	NMSE	Dir. Error
Adamson (All Genes)	GEARS	$\underline{0.5854 \pm 0.3548}$	$\underline{0.7631 \pm 0.6413}$	$\underline{0.2341 \pm 0.1063}$
	PerturbNet	0.2166 ± 0.2742	4.7775 ± 1.4690	0.2627 ± 0.1039
	REGINA (Ours)	$\mathbf{0.6429 \pm 0.3948}$	$\mathbf{0.5960 \pm 0.6445}$	$\mathbf{0.1833 \pm 0.0948}$
Adamson (Top 20 DEGs)	GEARS	$\underline{0.6419 \pm 0.5984}$	$\underline{0.4880 \pm 0.5266}$	$\underline{0.1452 \pm 0.2768}$
	PerturbNet	0.4059 ± 0.3911	1.3000 ± 0.5007	0.4238 ± 0.1702
	REGINA (Ours)	$\mathbf{0.6510 \pm 0.5554}$	$\mathbf{0.4788 \pm 0.4726}$	$\mathbf{0.1286 \pm 0.2383}$
Replogle (All Genes)	GEARS	$\mathbf{0.589 \pm 0.212}$	1.145 ± 1.238	$\mathbf{0.297 \pm 0.058}$
	REGINA (Ours)	0.559 ± 0.106	$\mathbf{0.729 \pm 0.187}$	0.304 ± 0.033
Replogle (Top 20 DEGs)	GEARS	0.696 ± 0.317	$\mathbf{0.539 \pm 0.610}$	$\mathbf{0.110 \pm 0.157}$
	REGINA (Ours)	$\mathbf{0.697 \pm 0.238}$	0.674 ± 0.181	0.112 ± 0.128

Table 2: Distributional metrics on the Adamson and Replogle datasets. Best results are bolded, and second-best are underlined.

Dataset	Model	MMD	Wasserstein	Variance Correlation	Energy Loss
Adamson	GEARS	1.5859 ± 0.2048	0.2612 ± 0.0187	0.0000 ± 0.0000	15.0395 ± 1.7858
	PerturbNet	$\mathbf{0.5614 \pm 0.1140}$	$\mathbf{0.2101 \pm 0.0258}$	$\mathbf{0.8908 \pm 0.0277}$	$\mathbf{7.7913 \pm 2.0317}$
	REGINA (Ours)	$\underline{1.5287 \pm 0.1986}$	$\underline{0.2547 \pm 0.0171}$	$\underline{0.4281 \pm 0.0526}$	$\underline{14.1082 \pm 1.8789}$
Replogle	GEARS	1.509 ± 0.117	0.327 ± 0.019	0.000 ± 0.000	24.561 ± 2.487
	REGINA (Ours)	$\mathbf{1.319 \pm 0.171}$	$\mathbf{0.276 \pm 0.022}$	$\mathbf{0.442 \pm 0.086}$	$\mathbf{17.132 \pm 3.374}$

5.3 Ablations and stability study

We evaluate REGINA's core components by ablating them individually on the Norman dataset (Table 5). Removing the **latent classifier** or replacing the **latent prompting** mechanism with learnable projections causes catastrophic failure on unseen perturbations (e.g., Pearson collapses to 0.08 and -0.05, respectively), proving their critical roles in structuring the latent space and enabling compositional generalization. Ablating the **center loss** noticeably degrades both local precision and global fidelity, confirming that clustering by cell state improves overall representation integrity. Finally, removing the **discriminator** primarily harms distributional metrics (e.g., increased MMD and Wasserstein) rather than local precision, confirming that adversarial training is the specific driver behind REGINA's distribution matching capabilities.

To ensure our results are not artifacts of a specific data split, we conducted a rigorous stability analysis across five independent random splits (Table 6). The model demonstrates consistent performance across all configurations.

Furthermore, to evaluate REGINA's capacity for compositional generalization, we follow the evaluation protocol established by GEARS and stratify the model's performance based on the effective exposure of perturbations during training. As shown in the lower section of Table 6, REGINA maintains strong predictive capabilities even in the most challenging scenarios—such as inferring the

Table 3: Performance metrics on the Norman dataset evaluated across all genes and the top 20 differentially expressed genes. Best results are bolded, and second-best results are underlined.

Dataset	Model	Pearson	NMSE	Dir. Error
Norman (All Genes)	CPA	0.3027 ± 0.0760	$\underline{0.9649 \pm 0.1088}$	$\mathbf{0.1011 \pm 0.0085}$
	GEARS	$\underline{0.5022 \pm 0.1377}$	1.3865 ± 1.1378	0.1307 ± 0.0212
	REGINA (Ours)	$\mathbf{0.5483 \pm 0.1153}$	$\mathbf{0.7502 \pm 0.1689}$	$\underline{0.1305 \pm 0.0080}$
Norman (Top 20 DEGs)	CPA	0.4466 ± 0.1926	0.9070 ± 0.0923	0.2813 ± 0.1221
	GEARS	$\underline{0.6699 \pm 0.3129}$	$\mathbf{0.5566 \pm 0.2387}$	$\mathbf{0.1107 \pm 0.1028}$
	REGINA (Ours)	$\mathbf{0.6844 \pm 0.2944}$	$\underline{0.6856 \pm 0.1696}$	$\underline{0.1276 \pm 0.1057}$

Table 4: Distributional metrics on the Norman dataset. Best results are bolded, and second-best results are underlined.

Dataset	Model	MMD	Wasserstein	Variance Correlation	Energy Loss
Norman	CPA	$\underline{1.5066 \pm 0.1217}$	$\mathbf{0.1511 \pm 0.0250}$	$\mathbf{0.7142 \pm 0.0354}$	$\mathbf{7.6613 \pm 1.8771}$
	GEARS	1.7089 ± 0.1461	0.1779 ± 0.0231	0.0000 ± 0.0000	10.7794 ± 1.5272
	REGINA (Ours)	$\mathbf{1.4776 \pm 0.1167}$	$\underline{0.1583 \pm 0.0257}$	$\underline{0.3270 \pm 0.0654}$	$\underline{8.3483 \pm 1.9317}$

Table 5: Ablation study of REGINA components on the Norman dataset. Performance is evaluated using Pearson correlation and Directional Error on the top 20 differentially expressed genes, alongside MMD and Wasserstein distance. Best results are bolded.

Model Configuration	Pearson ↑	Dir. Error ↓	MMD ↓	Wasserstein ↓
REGINA (Full Model)	$\mathbf{0.6844 \pm 0.2944}$	$\mathbf{0.1276 \pm 0.1057}$	$\mathbf{1.4776 \pm 0.1167}$	$\mathbf{0.1583 \pm 0.0257}$
w/o Center Loss	0.5599 ± 0.4169	0.2117 ± 0.1937	1.8696 ± 0.2696	0.1697 ± 0.0292
w/o Latent Classifier	0.0820 ± 0.3043	0.4322 ± 0.1619	4.1214 ± 0.6969	0.3157 ± 0.0555
w/o Adversarial Training	0.6617 ± 0.2292	0.1776 ± 0.1259	1.5742 ± 0.1247	0.1603 ± 0.0246
w/o Latent Prompting	-0.0547 ± 0.2016	0.5136 ± 0.1179	5.6422 ± 0.3384	7.8000 ± 3.4901

effects of a combined perturbation where neither individual gene was seen during training (0/2 seen), as well as on entirely unseen single perturbations.

Table 6: Stability and compositional generalization of REGINA. The top section demonstrates performance across five independent random splits, while the bottom section details generalization performance categorized by the effective exposure of combined perturbations during training.

Evaluation Setup	Pearson Top 5% ↑	Dir. Error ↓	MMD ↓	Wasserstein ↓
Stability Across Random Splits				
Split 1	0.634 ± 0.181	0.179 ± 0.117	1.519 ± 0.087	0.050 ± 0.002
Split 2	0.520 ± 0.317	0.210 ± 0.194	1.354 ± 0.079	0.048 ± 0.002
Split 3	0.597 ± 0.202	0.189 ± 0.117	1.498 ± 0.089	0.046 ± 0.002
Split 4	0.558 ± 0.287	0.208 ± 0.158	1.460 ± 0.083	0.042 ± 0.002
Split 5	0.543 ± 0.266	0.214 ± 0.167	1.527 ± 0.101	0.052 ± 0.002
Generalization by Effective Exposure				
0/2 Seen (Novel Combo)	0.626 ± 0.061	0.187 ± 0.049	2.215 ± 0.331	0.056 ± 0.013
1/2 Seen	0.575 ± 0.042	0.194 ± 0.014	1.624 ± 0.207	0.053 ± 0.004
2/2 Seen	0.657 ± 0.039	0.137 ± 0.027	1.919 ± 0.283	0.055 ± 0.011

5.4 Latent space

UMAP projections (Figure 4) confirm our regularization successfully structures the latent space. While raw gene expression exhibits heavily overlapping cell cycle phases (S, G1, G2M), REGINA's latent space enforces clear categorical separation. This demonstrates that the center and classifier losses effectively organize cells by biological state.

To further evaluate our latent space and the similarity of latent representations across different perturbations, we analyzed their cluster-distribution similarity (Figure 5). The results indicate that the vast majority of perturbation conditions exhibit strong grouping. This confirms that REGINA successfully isolates distinct transcriptomic responses into perturbation-specific latent distributions.

6 Discussion

In this work, we introduced REGINA, a fully data-driven representation learning framework that reformulates single-cell perturbation modeling as an unpaired distribution matching problem. By pairing a regularized Transformer autoencoder with a latent Cycle-GAN, REGINA successfully

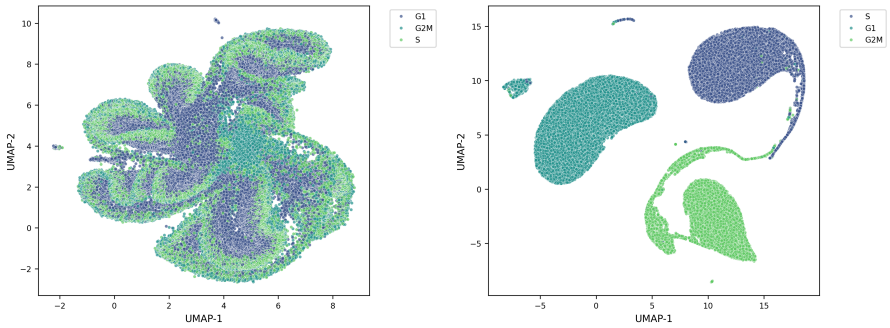

Figure 4: UMAP visualization of expression vectors on the Replogle dataset, colorized by cell cycle phase (left) and visualization of the same data from their latent embeddings (right).

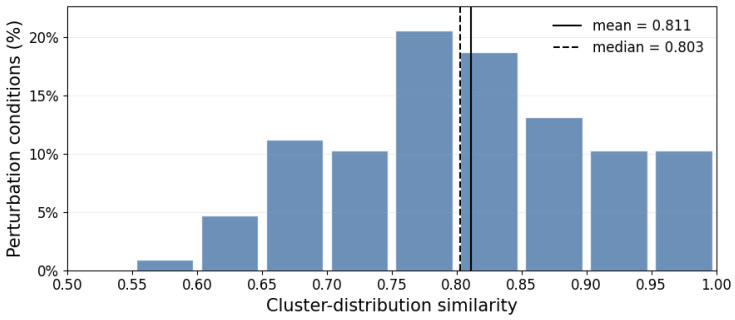

Figure 5: Cluster-distribution similarity of latent representations across different perturbations. The y-axis indicates the percentage of perturbation conditions achieving a given similarity score, demonstrating that the vast majority maintain high intra-cluster coherence

bypasses the heuristic dimensionality reduction methods common in current baselines. It achieves competitive local precision (Pearson) and superior MMD, while revealing an inherent trade-off between local accuracy and global variance preservation – complementing CPA which excels on variance. While the current framework establishes a strong foundation for unmasked perturbation modeling, we identify two primary avenues for future expansion. REGINA has the potential to significantly accelerate the early stages of drug discovery, although such computational predictions still require rigorous experimental validation prior to clinical application. Our current tokenization method does not follow any biological intuition, the splits are based on occurrence in the fixed gene expression vector. One possible improvement is, rather than utilizing arbitrary partitioning, tokens could be structurally defined by established gene regulatory networks or any nuanced meaningful strategy. A core advantage of REGINA is its ability to operate purely on raw data without relying on external knowledge graphs to mask the input space. The prompt-based conditioning mechanism (Z_{prompt}) utilized in our transition model provides a natural, non-destructive interface for injecting additional information. Future work will explore multimodal extension of the current model.

6.1 Limitations

Because REGINA relies on a fully data-driven representation learning paradigm, its performance is fundamentally bounded by the perturbation diversity of the training data. As observed in our experiments on the Dixit dataset (detailed in the Appendix 7.2.1 Dixit et al. (2016)), when the training regime contains a severely limited number of unique perturbations (e.g., $N = 24$), the model struggles to construct a latent manifold dense enough to support reliable inference for held-out conditions.

REPRODUCIBILITY STATEMENT

The source code for REGINA, including model implementations, training scripts, evaluation pipelines, and hyperparameter configurations, is available at: `https://anonymous.4open.science/r/single_cell-4F06`. All datasets used in this paper (Replogle, Norman, Adamson, Dixit) are publicly accessible via the GEARS pipeline.All experiments were run on an NVIDIA A100 GPU (40GB) cluster and it took approximately 2 hours to train each experiment.

AI USE STATEMENT

In this work, we used generative AI tools for writing polish, grammar checking, and rephrasing specific expressions to improve the clarity and readability of the manuscript. Additionally, we used generative AI tools for software development, specifically to assist in generating isolated code functions and for routine debugging of our experimental pipeline.

All AI-generated code functions and debugging solutions were manually reviewed, tested for correctness. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Britt Adamson, Thomas M Norman, Marco Jost, Min Y Cho, James K Nuñez, Yuwen Chen, Jacqueline E Villalta, Luke A Gilbert, Max A Horlbeck, Marco Y Hein, et al. A multiplexed single-cell crispr screening platform enables systematic dissection of the unfolded protein response. *Cell*, 167(7):1867–1882, 2016.

Martin Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks, 2017. URL https://arxiv.org/abs/1701.04862.

Srijan Atti and Shankar Subramaniam. Fundamental limitations of foundation models in single-cell transcriptomics. *bioRxiv*, 2025. doi: 10.1101/2025.06.26.661767. URL https://www.biorxiv.org/content/early/2025/06/28/2025.06.26.661767.

Charlotte Bunne, Stefan G Stark, Gabriele Gut, Jacobo Sarabia Del Castillo, Mitch Levesque, Kjong-Van Lehmann, Lucas Pelkmans, Andreas Krause, and Gunnar Rätsch. Learning single-cell perturbation responses using neural optimal transport. *Nature methods*, 20(11):1759–1768, 2023.

Wei Chen et al. Live-seq enables temporal transcriptomic recording of single cells. *Nature*, 608 (7924):733–740, 2022. doi: 10.1038/s41586-022-05046-9.

Gerold Csendes, Gema Sanz, Krist'of Z. Szalay, and Bence Szalai. Benchmarking foundation cell models for post-perturbation RNA-seq prediction. *BMC Genomics*, 26(1):393, 2025. doi: 10.1186/s12864-025-11600-2.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

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.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C. Courville. Improved training of wasserstein gans. *CoRR*, abs/1704.00028, 2017. URL http://arxiv.org/abs/1704.00028.

Yuta Hiasa, Yoshito Otake, Masaki Takao, Takumi Matsuoka, Kazuma Takashima, Jerry L. Prince, Nobuhiko Sugano, and Yoshinobu Sato. Cross-modality image synthesis from unpaired data using cyclegan: Effects of gradient consistency loss and training data size, 2018. URL https://arxiv.org/abs/1803.06629.

P. J. Jost, D. Weindl, K. Wunderling, C. Thiele, and J. Hasenauer. Suggested experimental design and computational modeling to infer single-cell lipid dynamics from a single destructive measurement. *iScience*, 29(4):115474, 2026. doi: 10.1016/j.isci.2026.115474.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *CoRR*, abs/2104.08691, 2021. URL https://arxiv.org/abs/2104.08691.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *CoRR*, abs/2101.00190, 2021. URL https://arxiv.org/abs/2101.00190.

Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058, 2018.

Mohammad Lotfollahi, F Alexander Wolf, and Fabian J Theis. scgen predicts single-cell perturbation responses. *Nature methods*, 16(8):715–721, 2019.

Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Yuge Ji, Ignacio L Ibarra, F Alexander Wolf, Nafissa Yakubova, Fabian J Theis, and David Lopez-Paz. Compositional perturbation autoencoder for single-cell response modeling. *BioRxiv*, 2021.

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):MSB202211517, 2023.

Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. *arXiv preprint arXiv:1611.04076*, 2016.

Annika Mütze, Matthias Rottmann, and Hanno Gottschalk. Semi-supervised domain adaptation with cyclegan guided by a downstream task loss, 2022. URL https://arxiv.org/abs/2208.08815.

Shivangi Nigam, Adarsh Prasad Behera, Shekhar Verma, and P. Nagabhushan. Learning latent representations for image translation using frequency distributed cyclegan, 2025. URL https://arxiv.org/abs/2508.03415.

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.

Yongjin P Park and Manolis Kellis. Cocoa-diff: counterfactual inference for single-cell gene expression analysis. *Genome Biology*, 22(1):228, 2021.

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.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

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

Shiv Shankar. Celltarnet: Single-cell perturbation prediction using transformer based normalizing flow. ICLR 2026 Workshop on Language Models for Robotics and Manipulation (LMRL), 2026. URL https://openreview.net/forum?id=fd2dt4zOwa.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, volume 30, 2017.

Yuchen Wang, Xingjian Chen, Xiangtao Li, and Ka-Chun Wong. screpa: Predicting single-cell perturbation responses with cycle-consistent representation alignment. *Computational Biology and Chemistry*, 120(Part 2):108712, 2026. doi: 10.1016/j.compbiolchem.2025.108712.

Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. *European Conference on Computer Vision*, pp. 499–515, 2016.

F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. Scanpy for analysis of large-scale single-cell gene expression data. *bioRxiv*, 2017. doi: 10.1101/174029. URL https://www.biorxiv.org/content/early/2017/08/09/174029.

Shun H Yip, Pak Chung Sham, and Junwen Wang. Evaluation of tools for highly variable gene discovery from single-cell rna-seq data. *Briefings in bioinformatics*, 20(4):1583–1589, 2019.

Ruzhang Zhao, Jiuyao Lu, Yuzi Li, Weiqiang Zhou, Ni Zhao, and Hongkai Ji. A systematic evaluation of highly variable gene selection methods for single-cell rna-sequencing. *Genome Biology*, 26(1): 424, 2025a. doi: 10.1186/s13059-025-03887-x.

Ruzhang Zhao, Jiuyao Lu, Yuzi Li, Weiqiang Zhou, Ni Zhao, and Hongkai Ji. A systematic evaluation of highly variable gene selection methods for single-cell rna-sequencing. *Genome Biology*, 26(1): 424, 2025b.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *CoRR*, abs/1703.10593, 2017a. URL `http://arxiv.org/abs/1703.10593`.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017b.

7 Appendix

7.1 Prompting motivation

As illustrated in Figure 6, genetic perturbations rarely result in a complete ablation of the target gene's expression. Consequently, employing a naive masking strategy—such as artificially zeroing out the targeted gene's value—is biologically inaccurate and destroys residual transcriptomic signal. This empirical observation directly motivated our development of the latent prompting mechanism, which explicitly conditions the network on the interventional state without corrupting the raw, unmasked input vector.

7.2 Biological context

In the wet lab, a population of cells is subjected to various genetic perturbations, with each cell tagged by a unique cell barcode and a guide RNA (sgRNA) identifying the targeted gene. Each captured mRNA molecule is tagged with a Unique Molecular Identifier (UMI) to distinguish biological transcripts from PCR duplicates. After lysis, mRNA is reverse-transcribed into cDNA and sequenced, yielding a Gene Expression Matrix $X \in \mathbb{R}^{S \times N}$, where S is the number of cells and N $\mathcal{O}$(10,000) is the

7.2.1 Smaller-Scale Benchmarks: Adamson and Dixit

To evaluate REGINA's performance across varying degrees of dataset complexity, we benchmarked the model on the Adamson and Dixit datasets. While genome-scale datasets like Replogle test scalability, these classical benchmarks contain a significantly lower number of unique genetic perturbations, allowing us to probe the model's data efficiency. The results, detailed in Tables 7 and 8, reveal contrast in performance dictated by the scale of the training distribution.

Dixit: Limitations in the Data-Scarce Regime. Conversely, the Dixit dataset represents an extreme low-resource environment, containing only 24 unique single-gene perturbations. In this regime, REGINA experiences severe performance degradation on local metrics, failing to accurately predict differential expression shifts (evidenced by a negative Pearson correlation on the Top 20 DEGs in Table 7). This failure is fully expected, because REGINA utilizes an unpaired distribution matching paradigm (Cycle-GAN) and operates without the assistance of external biological knowledge graphs. When perturbations are too sparse, the latent generator lacks the context necessary to infer unseen compositional changes, whereas baseline models like GEARS are partially insulated against this sparsity due to their reliance on explicitly hardcoded Gene Ontology (GO) graphs. Interestingly, REGINA still outperforms GEARS on several global metrics (Wasserstein and Energy Loss in Table 8), indicating that while it correctly models the overall shape of the cell population, the data-driven transition model cannot pinpoint specific gene targets without sufficient pretraining diversity. This confirms that REGINA's architecture is uniquely optimized for high-throughput, genome-scale modeling where purely empirical representation learning overtakes heuristic knowledge graphs.

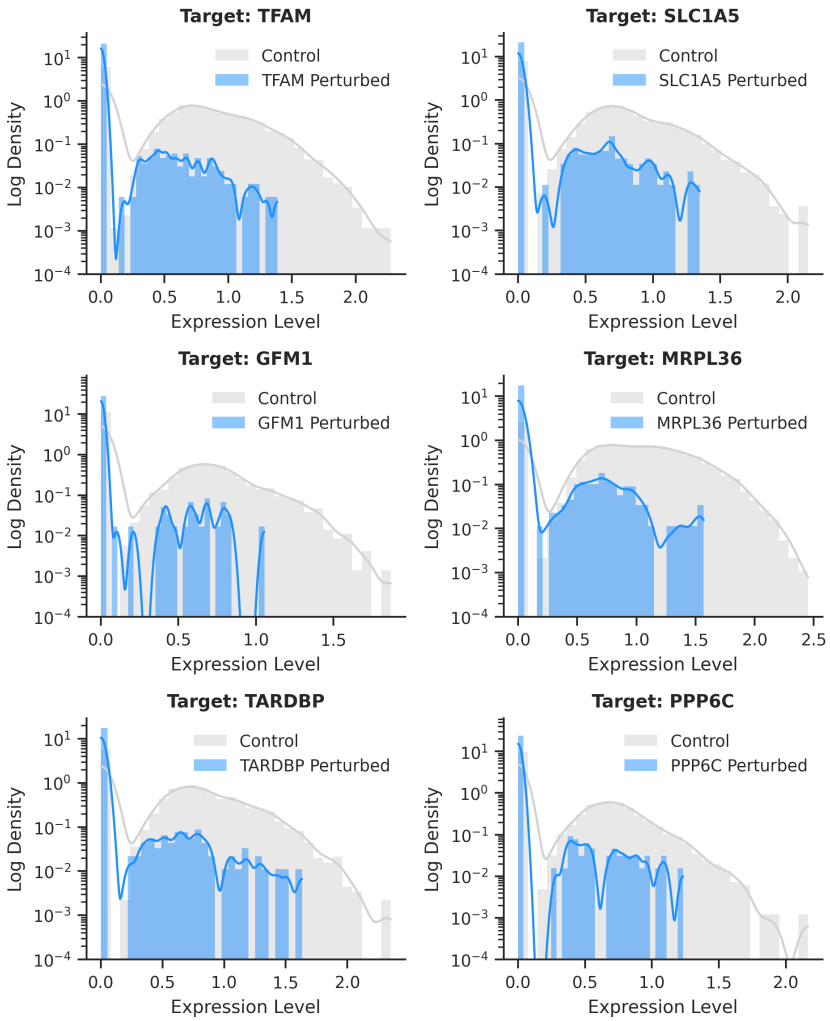

Figure 6: Gene expression levels of a specific gene, after it has been perturbed and and in the control state in the Replogle dataset.

Table 7: Performance metrics on the Dixit dataset evaluated across all genes and the top 20 differentially expressed genes (Top 20 DEGs). Best results are bolded.

Dataset	Model	Pearson	NMSE	Dir. Error
Dixit (All Genes)	GEARS	$\mathbf{0.3656 \pm 0.0377}$	$\mathbf{1.0721 \pm 0.1170}$	0.3038 ± 0.0181
	REGINA (Ours)	0.1930 ± 0.0498	1.3543 ± 0.1506	$\mathbf{0.2922 \pm 0.0130}$
Dixit (Top 20 DEGs)	GEARS	$\mathbf{0.1743 \pm 0.1606}$	$\mathbf{1.1443 \pm 0.2232}$	$\mathbf{0.4400 \pm 0.0583}$
	REGINA (Ours)	-0.3082 ± 0.1455	2.0018 ± 0.2309	0.5900 ± 0.0663

Table 8: Distributional metrics on the Dixit dataset. Best results are bolded.

Dataset	Model	MMD	Wasserstein	Variance Correlation	Energy Loss
Dixit	GEARS	$\mathbf{1.4740 \pm 0.0308}$	0.1961 ± 0.0035	0.0000 ± 0.0000	10.5429 ± 0.1160
	REGINA (Ours)	1.4811 ± 0.0168	$\mathbf{0.1874 \pm 0.0019}$	$\mathbf{0.2168 \pm 0.0811}$	$\mathbf{9.8388 \pm 0.1527}$

7.2.2 Extended Visualization of Replogle Performance

To supplement the aggregated tabular metrics, we provide the full performance distributions of REGINA and GEARS evaluated on the Replogle dataset. Figure 8 illustrates the density of local precision metrics (Pearson correlation, NMSE, and Directional Error), while Figure 7 visualizes the global distributional metrics (MMD, Wasserstein, Variance Correlation, and Energy Loss) across all held-out perturbation conditions. These distributions demonstrate that REGINA does not merely improve the mean performance, but consistently shifts the entire mass of the error distribution toward better predictive accuracy and higher distributional fidelity compared to the baseline.

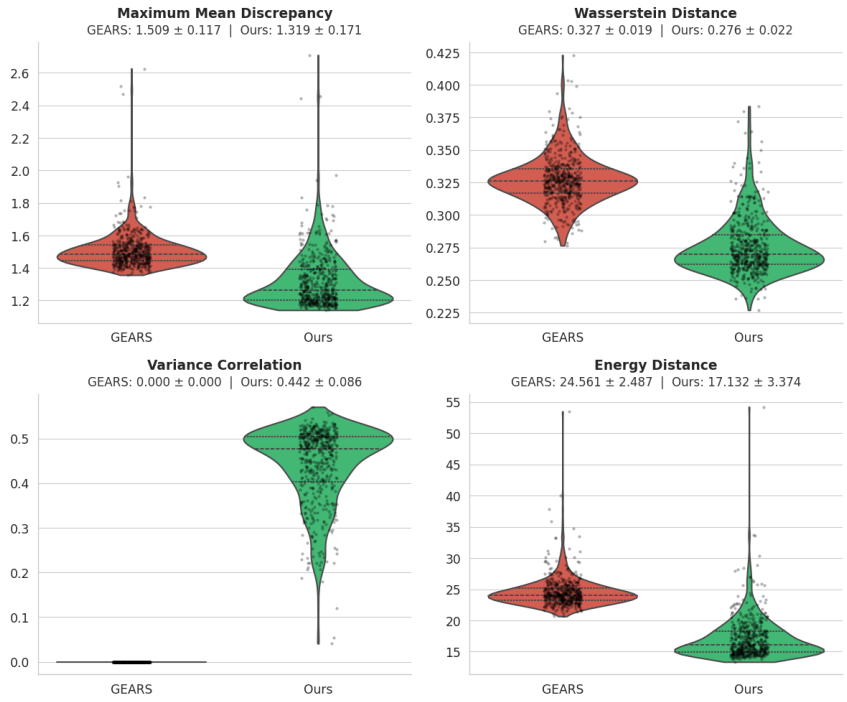

Figure 7: Distribution results on Replogle RPE1. REGINA (green) consistently achieves lower distributional divergence compared to GEARS across held-out perturbations.

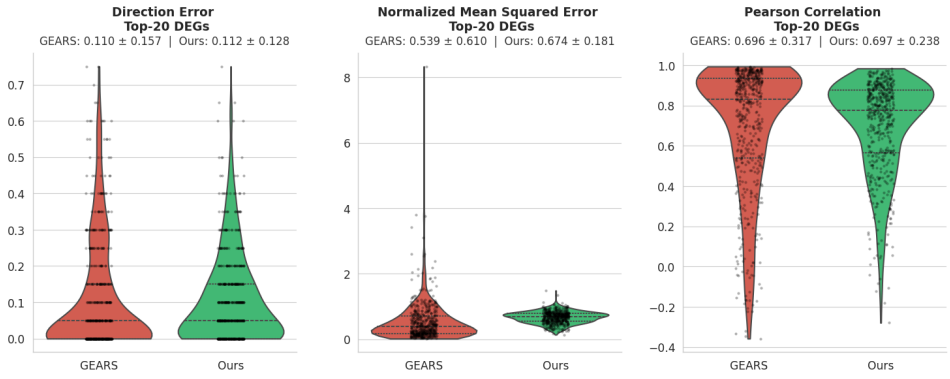

Figure 8: Precision results on Replogle RPE1. REGINA (green) consistently achieves similar precision compared to GEARS across held-out perturbations.