GIDER: GUIDE-INFORMED DUAL-STAGE MODELING OF CRISPR–CAS9 EDITING RESPONSES IN SINGLE CELLS

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting single-cell transcriptional responses to specific CRISPR–Cas9 editing designs can support the assessment of intervention outcomes before experimentation. However, target-gene identity alone cannot distinguish different single-guide RNAs (sgRNAs) targeting the same gene under identical cellular backgrounds, limiting design-specific response prediction. Sequence information provides finer-grained conditions, but how to organize sgRNA–site relationships and their corresponding DNA contexts into predictive conditions for unseen editing designs remains underexplored. We therefore introduce GIDER, a two-stage framework inspired by gene-editing mechanisms that models targeting relationships and the corresponding site contexts as interdependent conditioning information. First, GIDER incorporates target-gene priors to encode sgRNA relationships with intended target sites and computationally predicted candidate off-target sites. It then uses each site's query to read its own long-range DNA context and reuses the first-stage site weights for aggregation. Auxiliary supervision is provided by observed differences in mean expression between perturbed and background-matched control populations in the training data. Conditioned on these representations and control expression, GIDER combines discrete change-state prediction with continuous magnitude generation to learn perturbed expression distributions from unpaired data. Across unseen-sgRNA and unseen-target-gene tasks on Frangieh and Papalexi, GIDER achieves the lowest all-gene EMD, defined as the mean gene-wise Wasserstein-1 distance, in all four settings, with reductions of 6.00%–19.62% relative to the best baseline in each setting. On Papalexi, it leads all six error metrics and the perturbation discrimination score (PDS cosine) in both tasks. Controlled ablations assess the contributions of editing information, conditioning structure, and response supervision. These results suggest that structured editing representations improve prediction of the transcriptional consequences of specific editing designs,providing a computational basis for virtual gene editing to inform experimental design.

1 INTRODUCTION

Gene editing enables targeted changes to genomic sequences, providing an important tool for investigating gene function and modulating cellular states (Cong et al., 2013). Advances in single-cell sequencing allow researchers to profile transcriptomes under different editing conditions, providing a basis for systematically studying the cellular consequences of genetic interventions (Frangieh et al., 2021). However, combinations of target genes, single-guide RNAs (sgRNAs), and cellular contexts define an experimental space that is impractical to measure exhaustively. This challenge motivates single-cell genetic perturbation prediction: inferring transcriptional response distributions under unmeasured interventions from intervention specifications and control cell populations (Chi et al., 2026b). Such predictions offer a concrete step toward virtual gene editing by estimating the transcriptional consequences of proposed editing designs before experiments, thereby informing experimental design. Existing methods represent perturbations at different levels of granularity. Gene-level approaches use target-gene indicators, gene embeddings, or functional relationships to describe interventions, as exemplified by scGPT (Cui et al., 2024) and GEARS (Roohani et al., 2024). How-

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

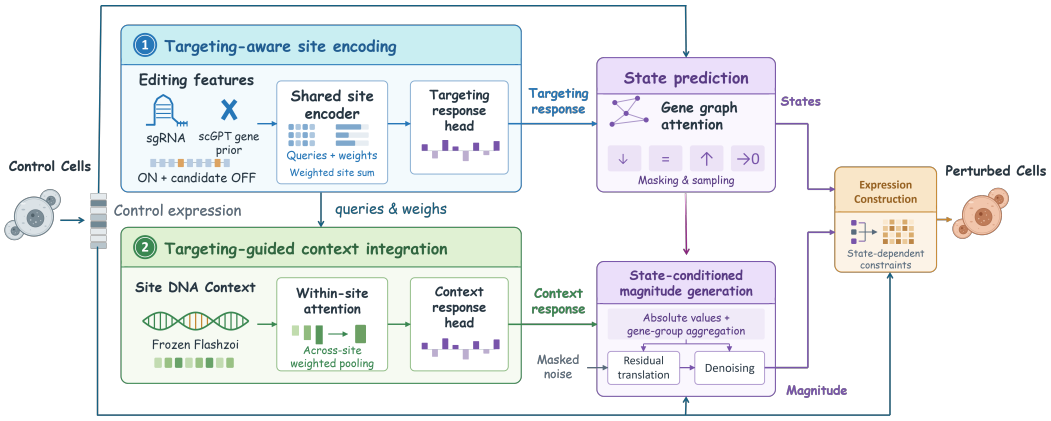

Figure 1: Overview of GIDER. Stage 1 models sgRNA–site relationships to generate queries and weights that guide DNA-context integration in Stage 2. The resulting responses condition discrete change-state prediction and state-conditioned magnitude generation. Sampled states and generated magnitudes are combined with control expression to reconstruct perturbed expression samples. Both response heads receive auxiliary supervision from observed population mean-expression differences.

ever, gene-level descriptions have limitations when evaluating specific editing designs. Single-cell screens have shown that sgRNAs targeting the same gene can differ in the fraction of cells exhibiting detectable transcriptional perturbations (Papalexi et al., 2021). When the perturbation condition consists solely of target-gene identity, a model cannot distinguish different sgRNAs targeting that gene under identical control inputs and experimental backgrounds. This limits its ability to evaluate alternative editing designs for the same target gene. Sequence information enables more detailed perturbation representations. STRAND uses queries derived from targeted regions to attend to extended DNA contexts (Fu et al., 2026), while PerturbNet uses protein sequences corresponding to specified variants in its coding-variant prediction setting (Yu et al., 2025). These methods incorporate information about genomic loci or variants, whereas a specific Cas9 editing design also involves targeting relationships between an sgRNA and its potential sites of action. Sequence compatibility between sgRNA and DNA provides information relevant to targeting specificity (Hsu et al., 2013), while the DNA context of each site supplies information about its surrounding sequence. Linking these sources of information could make context retrieval and aggregation depend on the targeting relationships of a specific sgRNA, providing more detailed conditions for predicting the response to an editing design. Therefore,*Can relationships between an sgRNA, its intended target site, and computationally predicted candidate off-target sites guide the integration of their corresponding DNA contexts to improve transcriptional response prediction under unmeasured editing conditions?*

To address this question, we introduce **GIDER**, a two-stage framework for single-cell perturbation prediction inspired by gene-editing mechanisms. Motivated by sgRNA-guided target recognition, the first stage incorporates a target-gene prior and encodes relationships between the sgRNA and its intended and candidate off-target sites, producing site queries and predictive relevance weights. The second stage uses these queries to read the extended DNA context of each corresponding site and reuses the first-stage weights for aggregation, progressively integrating targeting relationships with their associated genomic contexts. Separate response heads map the outputs of the two stages to gene-wise response space. Both heads receive auxiliary supervision from observed differences in mean expression between training perturbed populations and training control populations under the same experimental background. The downstream generation module predicts expression change states using the targeting response and control expression. It then generates change magnitudes conditioned on these states, a condition derived from the context response, and control expression through a process inspired by decoupled residual denoising diffusion (DRDD) (Lin et al., 2026). Together, these components learn perturbed single-cell expression distributions from unpaired observations. Across the unseen-sgRNA and unseen-target-gene tasks on the Frangieh and Papalexi datasets, GIDER achieves the lowest EMD among the evaluated methods on both all modeled genes and the top 20 genes in the differential-expression ranking in all four settings. Controlled ablations

further support the predictive contributions of editing information, the dependency between the two stages, and auxiliary response supervision.

Our main contributions are as follows:

- **Perturbation representations for specific CRISPR–Cas9 editing designs.** We jointly represent sgRNAs, intended target genes, intended target sites, and computationally predicted candidate off-target sites, explicitly linking sgRNA–site sequence relationships to the DNA context of each site to construct conditions for response prediction.
- **Structured two-stage conditioning.** We preserve the correspondence between sgRNA–site features and DNA contexts, derive site-specific queries from targeting relationships to guide context retrieval, and reuse site weights for cross-site aggregation. Population-response supervision maps both stages to gene-wise conditions for expression generation.
- **Evaluation under unseen editing conditions.** We evaluate unseen-sgRNA and unseen-target-gene prediction on two datasets, with controlled ablations on Papalexi SG assessing the predictive value of conditioning information and its organization.

2 Related Work

2.1 Representations of Perturbation Conditions

Existing methods represent perturbation conditions through intervention identities, gene relationships, and sequence-derived priors. CPA (Lotfollahi et al., 2023) disentangles basal cell state, perturbations, and experimental covariates in latent space. GEARS (Roohani et al., 2024) and TxPert (Wenkel et al., 2026) use gene relationship graphs to predict responses to unseen genetic perturbations, while scGPT (Cui et al., 2024) adapts pretrained single-cell representations to perturbation prediction. Sequence information further enriches these conditions: PerturbNet (Yu et al., 2025) uses protein sequence representations to predict responses to coding variants; CisTransCell (Zhang et al., 2026) incorporates regulatory DNA and transcript sequence priors; and STRAND (Fu et al., 2026) uses queries derived from targeted genomic regions to attend to long-range DNA context, aligning the resulting representations with population-level transcriptional changes in an RNA latent space. Our approach represents an editing design through the sgRNA, intended target gene, intended on-target site, and computationally predicted candidate off-target sites. Site-specific queries derived from sgRNA–site relationships guide the integration of the corresponding DNA contexts, with site weights shared across the two modeling stages.

2.2 Response Distribution Modeling from Unpaired Populations

Methods for learning response distributions without paired observations of the same cell before and after intervention can be grouped by their primary modeling mechanism. *Optimal-transport mapping*, exemplified by CellOT (Bunne et al., 2023), learns mappings from control to perturbed populations. *Flow-matching methods*, including CellFlow (Klein et al., 2025) and scDFM (Yu et al., 2026), generate responses through conditional velocity fields, with scDFM additionally incorporating a distribution-matching objective. *Diffusion and bridge models* include Squidiff (He et al., 2026), which uses conditional denoising; Doloris (Chi et al., 2026b), which connects separately learned control and perturbed distributions through a shared Gaussian latent space; and Departures (Chi et al., 2026a), which jointly models zero/nonzero expression states and continuous expression values. Our method uses within-condition optimal transport (OT) coupling to construct statistical training pairs and parameterizes responses relative to control expression through four discrete change states and continuous change magnitudes. Generation is conditioned on mechanism-inspired representations of sgRNA targeting relationships and the corresponding genomic contexts.

3 Method

3.1 Problem formulation and framework overview

Problem formulation. We predict single-cell transcriptional response distributions from an editing design and control expression. For each training condition $c = (g, e)$, we observe cells subjected

to single-guide RNA (sgRNA) g and control cells under the same experimental background e, such as culture conditions or exogenous treatments. The data do not contain paired pre- and post-editing measurements of the same cell. Let $x, y \in \mathbb{R}^G_{\geq 0}$ denote control and perturbed expression over G modeled genes, respectively, after $\log(1 + \mathrm{CP10K})$ normalization and division by a fixed scale $\kappa > 0$. Given an editing design $\mathcal{E}_g$ and control distribution $P_{0,e}$, we learn a conditional generative model whose population prediction

$$\widehat{P}_{\Theta,g,e}(\mathrm{d}y) = \int p_\Theta(\mathrm{d}y \mid x, \mathcal{E}_g, e)\, P_{0,e}(\mathrm{d}x) \tag{1}$$

approximates the observed perturbed distribution $P_{g,e}$, where Θ collects all trainable parameters. The objective is population distribution prediction; a generated sample is not interpreted as its input control cell's true post-editing state.

Framework overview. As shown in Figure 1, GIDER first encodes sgRNA–site relationships and then uses the resulting queries and weights to integrate each site's corresponding DNA context. Under observed population-response supervision, two response heads map these representations to gene-wise conditions: the targeting response conditions state prediction, while the context response, states, and control expression condition magnitude generation. States and magnitudes are combined with control expression to produce perturbed response.

3.2 Mechanism-inspired two-stage condition modeling

Editing-condition representation. For sgRNA g, the editing design comprises the intended target gene $\tau(g)$, one intended on-target site (ON), and up to K candidate off-target sites (OFF): $\mathcal{E}_g = (\tau(g), \{(p_{gs}, \rho_{gs}, m_{gs}, H_{gs})\}_{s=1}^S)$, with $S = K+1$. Here, p_{gs} contains short sgRNA–site sequence, mismatch, and positional features; ρ_{gs} denotes the ON/OFF role; and $m_{gs} \in \{0, 1\}$ masks missing candidates. The ON site is located using the experimental sgRNA, reference genome, and target-gene annotations. Candidate OFF sites are retrieved with Cas-OFFinder (Bae et al., 2014) and ranked and selected with CCLMoff (Du et al., 2025). A frozen Flashzoi encoder (Hingerl et al., 2025) extracts features from extended reference DNA context at each site; regional pooling yields $H_{gs} \in \mathbb{R}^{B \times d_H}$, with B regions and feature dimension d_H. Index s preserves the correspondence between short features and DNA context. Candidate OFF sites are computational predictions, not verified off-target events; feature construction details are provided in the Appendix C.

Stage 1: targeting-aware site representations. Inspired by sgRNA-guided sequence recognition, we combine short sgRNA–site features, a target-gene prior, and site roles. The prior is a shared projection of a frozen scGPT GeneEncoder feature (Cui et al., 2024), $u_{\tau(g)} = W_{\mathrm{gene}} z^{\mathrm{scGPT}}_{\tau(g)} + b_{\mathrm{gene}}$. Site queries, weights, and the targeting representation are

$$\begin{aligned} q_{gs} &= m_{gs} \tanh\big(W_q[f_p(p_{gs}) + u_{\tau(g)} + v_{\rho_{gs}}] + b_q\big)\,, \\ w_{gs} &= m_{gs}\, \mathrm{sigmoid}(a_q^\top q_{gs} + b_w), \qquad h_T = \sum_{s=1}^S w_{gs} q_{gs}. \end{aligned} \tag{2}$$

Here, f_p maps the short features, and $v_{\rho_{gs}}$ is a role embedding. ON and OFF sites share computational modules; no trainable identity embeddings specific to individual sgRNAs or target genes are used. Independent gates allow multiple sites to contribute; their weights encode predictive relevance, not Cas9 cleavage probabilities. This stage does not access extended DNA context and passes its queries and weights to Stage 2.

Stage 2: targeting-guided context integration. To account for the dependence of transcriptional consequences on site context, each query q_{gs} reads only its own site's DNA context. With $v_{gsb} = W_H\, \mathrm{LN}_0(H_{gsb})$, we compute

$$\begin{aligned} \alpha_{gsb} &= \mathrm{softmax}_b\left(\frac{q_{gs}^\top (v_{gsb} + r_b)}{\sqrt{d}}\right), \\ \ell_{gs} &= f_H\left(\sum_{b=1}^B \alpha_{gsb} v_{gsb}\right), \qquad h_C = \sum_{s=1}^S w_{gs} \ell_{gs}. \end{aligned} \tag{3}$$

Here, LN_0 denotes layer normalization without affine parameters, r_b encodes regional position, d is the query/projection dimension, and f_H maps the aggregated context. Queries select regions within sites; shared weights aggregate across sites. The value pathway preserves zero inputs, ensuring $h_C = 0$ when all context features are zero (Appendix D).

3.3 Supervision from observed population RNA responses

To connect editing representations with transcriptional responses, two independently parameterized response heads map the targeting and context representations to a gene-indexed response space:

$$\widehat{d}_{j,c} = D_j(h_j(g), b_e, e) \in \mathbb{R}^G, \qquad j \in \{T, C\}, \quad c = (g, e).$$

Here, b_e is a fixed mean-expression reference computed from training control cells under background e and retained at inference; its construction is specified in Appendix A.4. Each output coordinate corresponds to one modeled gene. Both heads are supervised by the same observed population response,

$$r_c^{\mathrm{obs}} = \frac{1}{|\mathcal{Y}_c^{\mathrm{tr}}|} \sum_{y \in \mathcal{Y}_c^{\mathrm{tr}}} y - \frac{1}{|\mathcal{X}_e^{\mathrm{tr}}|} \sum_{x \in \mathcal{X}_e^{\mathrm{tr}}} x, \tag{4}$$

where $\mathcal{Y}_c^{\mathrm{tr}}$ contains the training perturbed cells under condition c, and $\mathcal{X}_e^{\mathrm{tr}}$ contains all training control cells under background e. The target is computed directly from observed expression in the model's expression space and retains both positive and negative changes. For the set $\mathcal{U}$ of distinct conditions in a minibatch, the auxiliary loss is

$$\mathcal{L}_{\mathrm{obs}} = \frac{1}{2|\mathcal{U}|G} \sum_{c \in \mathcal{U}} \sum_{j \in \{T,C\}} \sum_{k=1}^{G} \left(\frac{\widehat{d}_{j,c,k} - r_{c,k}^{\mathrm{obs}}}{s_k} \right)^2, \tag{5}$$

where s_k is the root mean square of gene k's responses across training conditions, floored at 0.01. The heads predict complete responses, not additive effect components. For generation, the signed targeting response $\widehat{d}_T$ conditions state prediction; gene-wise absolute values of $\widehat{d}_C$ are aggregated over fixed gene-graph groups to form

$$c_A = \frac{1}{P} \sum_{p=1}^{P} \left(\frac{1}{|\mathcal{G}_p|} \sum_{k \in \mathcal{G}_p} |\widehat{d}_{C,k}| \right) a_p, \tag{6}$$

where $\mathcal{G}_p$ is a gene group and a_p a learnable group embedding. The magnitude networks receive c_A as a global additive condition. This compression discards response signs and within-group gene-specific magnitude patterns; gene-specific information remains available through control expression and discrete states. During both training and inference, the generator uses predicted response conditions; observed population-response targets enter only the auxiliary loss.

3.4 State-conditioned transcriptional response generation

We parameterize responses by discrete change states Q and nonnegative magnitudes A:

$$p_\Theta(Q, \mathrm{d}A \mid x, \mathcal{E}_g, e) = p_\psi(Q \mid x, \widehat{d}_T, e)\, p_\theta(\mathrm{d}A \mid Q, x, c_A, e). \tag{7}$$

Here, ψ and θ parameterize the state and magnitude networks, respectively; A denotes magnitudes after terminal constraints, including deterministic values for some states. To obtain supervision from unpaired data, we couple training control and perturbed cells within each condition using entropy-regularized optimal transport (OT), with uniform marginals and normalized squared expression distances as costs. Sampled endpoints $(x, y) \sim \pi_c$ provide state and magnitude targets, not measured single-cell editing trajectories. The OT construction and implementation settings are provided in Appendix E.

Discrete change states. For a training endpoint pair, define

$$Q_k^\star = \begin{cases} 0, & 0 < y_k < x_k, \\ 1, & y_k = x_k, \\ 2, & y_k > x_k, \\ 3, & y_k = 0 < x_k. \end{cases} \tag{8}$$

These states denote a decrease to a positive value, exact equality, an increase, and a decrease to zero; exact equality is not a statistical criterion for absence of a response. A graph-attention network jointly computes gene-wise logits, followed by gene-wise softmax probabilities $p_{\psi,k}$, with $p_\psi(Q \mid x, \widehat{d}_T, e) = \prod_{k=1}^{G} p_{\psi,k}(Q_k \mid x, \widehat{d}_T, e)$. States are sampled independently conditional on these inputs. When $x_k = 0$, infeasible states 0 and 3 are masked and probabilities renormalized during both training and inference. The cross-entropy loss is

$$\mathcal{L}_Q = \mathbb{E}_c \mathbb{E}_{(x,y)\sim\pi_c}\left[-\frac{1}{G}\sum_{k=1}^{G}\log p_{\psi,k}(Q_k^\star \mid x, \widehat{d}_T, e)\right], \tag{9}$$

where the outer expectation follows the training condition sampler.

Continuous change magnitudes. States 1 and 3 have deterministic magnitudes 0 and x_k; continuous generation applies to states 0 and 2. Define $M_k = \mathbf{1}\{Q_k^\star \in \{0,2\}\}$ and $a^\star = M \odot |y - x|/\gamma$, where $\gamma > 0$ is a fixed scale estimated from training endpoints. Inspired by the separation of translation and denoising in decoupled residual denoising diffusion (DRDD) (Lin et al., 2026), we construct two path segments conditioned on $\chi = (x, Q^\star, M, c_A, e)$:

$$\begin{aligned} z_t &= t\,a^\star + \sigma_A M \odot \epsilon, \\ v_u &= a^\star + \sigma_A(1-u)M \odot \xi, \end{aligned} \qquad \epsilon, \xi \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I). \tag{10}$$

The times $t, u \in [0,1]$ and the two Gaussian noises are sampled independently, and $\sigma_A > 0$ is the noise scale. The first segment introduces the target magnitude at a fixed noise level; the second removes noise. Their connecting training endpoints satisfy $\mathrm{Law}(z_1 \mid \chi) = \mathrm{Law}(v_0 \mid \chi)$ in conditional distribution, without requiring independently sampled endpoints to coincide. Separate residual and noise networks regress $-a^\star$ and ξ:

$$\begin{aligned} \mathcal{L}_R &= \mathbb{E}[\ell_M(R_{\theta_R}(z_t, t; \chi), -a^\star)], \\ \mathcal{L}_N &= \mathbb{E}[\ell_M(N_{\theta_N}(v_u, u; \chi), \xi)]. \end{aligned} \tag{11}$$

Here,

$$\ell_M(r, s) = \frac{\sum_k M_k (r_k - s_k)^2}{\max(1, \sum_k M_k)}$$

normalizes each cell's error by its number of active coordinates before averaging over samples; both losses vanish when no coordinates are active. The expectations include training endpoints, times, and Gaussian noises. Under squared-error regression, the networks estimate the conditional means of their respective targets on active coordinates. Since the pathwise derivatives are $a^\star$ and $-\sigma_A M \odot \xi$, the corresponding conditional-mean velocity estimates are $-M \odot R_{\theta_R}$ and $-\sigma_A M \odot N_{\theta_N}$, motivating the sampling dynamics in Section 3.5. The networks use independently parameterized graph-attention architectures and receive the current continuous state, time, and condition χ. Control expression remains fixed throughout both segments. Training uses endpoint-derived states $Q^\star$, whereas inference uses states $\widehat{Q}$ sampled from the state predictor.

3.5 Joint training and expression reconstruction

The joint objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_Q + \mathcal{L}_R + \sigma_A^2 \mathcal{L}_N + \lambda_{\text{obs}} \mathcal{L}_{\text{obs}}, \tag{12}$$

where λ_{obs} weights response supervision and σ_A^2 converts noise-prediction error to the squared-error scale of the corresponding path velocity. Response targets and normalization scales are computed using training data only. The scGPT and Flashzoi feature extractors remain frozen; all other modules are jointly trained.

At inference, we compute response conditions from the editing design and experimental background, and sample $\widehat{Q}$ conditional on control expression x. Set $\widehat{M}_k = \mathbf{1}\{\widehat{Q}_k \in \{0,2\}\}$ and $\widehat{\chi} = (x, \widehat{Q}, \widehat{M}, c_A, e)$. Starting from $z_0 = \sigma_A \widehat{M} \odot \epsilon$, $\epsilon \sim \mathcal{N}(0, I)$, we sequentially integrate

$$\begin{aligned} \frac{\mathrm{d}z_t}{\mathrm{d}t} &= -\widehat{M} \odot R_{\theta_R}(z_t, t; \widehat{\chi}), \qquad v_0 = z_1, \\ \frac{\mathrm{d}v_u}{\mathrm{d}u} &= -\sigma_A \widehat{M} \odot N_{\theta_N}(v_u, u; \widehat{\chi}). \end{aligned} \tag{13}$$

Each segment uses 20 explicit Euler steps. Inactive coordinates remain zero, while intermediate states may be negative. We restore the magnitude scale only at the endpoint, $a = \gamma v_1$, and reconstruct expression as

$$\widehat{y}_k = \begin{cases} x_k - \mathrm{clip}(a_k, 0, x_k), & \widehat{Q}_k = 0, \\ x_k, & \widehat{Q}_k = 1, \\ x_k + \max(0, a_k), & \widehat{Q}_k = 2, \\ 0, & \widehat{Q}_k = 3. \end{cases} \tag{14}$$

This guarantees nonnegative outputs and enforces exact equality and decrease-to-zero states analytically. Terminal projection permits state 0 to reach the equality or zero boundary and state 2 to reach equality. Repeated sampling of control cells, change states, and initial noise yields the predicted perturbed population. We recover the original normalized expression space as $\widehat{y}^{\mathrm{obs}} = \kappa \widehat{y}$. Inference does not access perturbed cells or observed response targets for the target condition.

4 EXPERIMENTS AND RESULTS

4.1 EXPERIMENTAL SETUP

Datasets. We evaluate GIDER on the Frangieh (Frangieh et al., 2021) and Papalexi (Papalexi et al., 2021) single-cell CRISPR–Cas9 datasets, which investigate melanoma immune evasion and immune-checkpoint regulation in THP-1 cells, respectively. Both provide sgRNA annotations and single-cell transcriptomes; we use RNA expression as the prediction target. Dataset statistics, preprocessing, and task splits are detailed in Appendix B.

Generalization tasks. For *unseen sgRNAs* (SG), we hold out all cells associated with each test sgRNA across all experimental backgrounds, while retaining other sgRNAs targeting the same gene in training. For *unseen target genes* (UG), we hold out all cells associated with all sgRNAs targeting each test gene across all backgrounds. Within each dataset and task, all methods use identical splits and test conditions. Predicted populations are generated from background-matched controls and compared with held-out perturbed populations.

Baselines. We compare GIDER with GEARS (Roohani et al., 2024), scGPT (Cui et al., 2024), biolord (Piran et al., 2024), DOLORIS (Chi et al., 2026b), Departures (Chi et al., 2026a), and scDFM (Yu et al., 2026). We also include a ridge-regression baseline that predicts population means from editing-condition features and adds centered control-expression deviations. For the SG tasks on both datasets, SameTarget-shift transfers the average training response of sgRNAs targeting the same gene under the same background to evaluation control cells. External methods are adapted from author-released implementations to common data splits, an evaluation gene panel, and a scoring protocol. Method-specific conditioning and output construction are described in the Appendix H.

Evaluation metrics. Each sgRNA–experimental-background combination is an evaluation unit. On all modeled genes (All) and the top 20 genes in the differential-expression ranking (DE20), we report RMSE between population means, energy distance (E-distance), and the mean per-gene one-dimensional Wasserstein-1 distance (EMD), yielding six error metrics. We also report PDS cosine, which measures perturbation distinguishability through the rank of the matching condition under cosine distance. Metric definitions, differential-expression ranking, and score aggregation are detailed in the Appendix F.

4.2 MAIN RESULTS

Unseen sgRNAs. GIDER achieves the lowest All E-distance, All EMD, and DE20 EMD on both datasets, as well as the lowest values across all six error metrics and the highest PDS cosine on Papalexi (Table 1). Compared with SameTarget-shift, GIDER reduces All EMD by approximately 10.5% on Frangieh and 11.8% on Papalexi. These results support improved transcriptome-wide mean and distributional fit under unseen sgRNA conditions relative to directly transferring same-target, background-matched training responses. On Frangieh, however, SameTarget-shift performs better on DE20 RMSE, DE20 E-distance, and PDS cosine, while scGPT achieves the lowest All and DE20 RMSE; the gains therefore do not extend to every evaluation dimension.

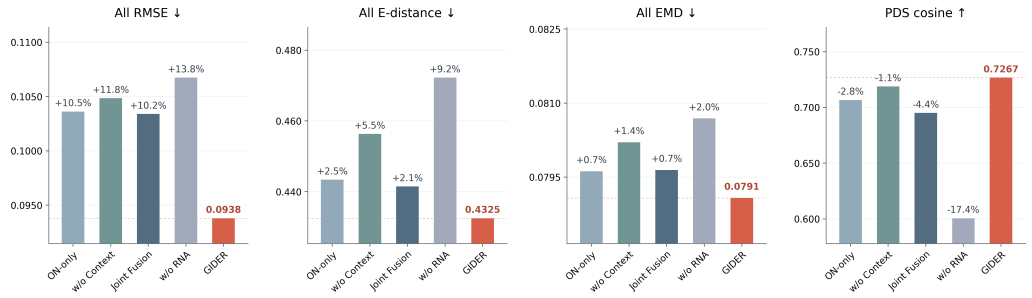

Figure 2: Ablation study on Papalexi in the unseen-sgRNA (SG) setting. We compare the full GIDER model with four variants: ON-only, removal of long-range DNA context, Joint Fusion, and removal of RNA auxiliary supervision. All variants use the same data split and evaluation metrics.

Table 1: Unseen sgRNA (SG) prediction on Frangieh and Papalexi. All and DE20 denote all modeled genes and the top 20 genes in the differential-expression ranking, respectively. E-dist. denotes energy distance; PDS cosine is the cosine-based perturbation discrimination score. SameTarget-shift transfers the average training response of other sgRNAs targeting the same gene under the same background. Bold indicates the best reported value within each dataset and metric.

Dataset	**Model**	**All ↓**			**DE20 ↓**			**PDS ↑**
		RMSE	E-dist.	EMD	RMSE	E-dist.	EMD	cosine
Frangieh	Ridge	0.1417	1.0889	0.0905	0.1481	0.6934	0.1171	0.5437
	GEARS	0.0873	0.3266	0.0726	0.1476	0.6891	0.1177	0.5533
	scGPT	**0.0722**	2.9224	0.2036	**0.1138**	0.8630	0.1148	0.5433
	biolord	0.0775	3.0848	0.2098	0.1310	0.9349	0.1187	0.5902
	DOLORIS	0.1030	0.8085	0.1031	0.1601	0.6765	0.1220	0.4941
	Departures	0.1423	1.3301	0.1249	0.1841	0.7623	0.1477	0.5015
	scDFM	0.0949	0.4967	0.0834	0.1489	0.6874	0.1218	0.5124
	SameTarget-shift	0.0921	0.4271	0.0721	0.1293	**0.6175**	0.1075	**0.6041**
	GIDER (ours)	0.0839	**0.3228**	**0.0645**	0.1301	0.6360	**0.0977**	0.5839
Papalexi	Ridge	0.1212	0.6085	0.0887	0.3629	0.9855	0.3261	0.5000
	GEARS	0.1167	0.5917	0.1000	0.2801	0.8524	0.2613	0.7183
	scGPT	0.1062	2.8807	0.2108	0.1993	1.1647	0.2969	0.6550
	biolord	0.1012	3.2288	0.2226	0.2279	1.2970	0.3275	0.6711
	DOLORIS	0.1268	0.8904	0.1153	0.3360	0.9433	0.3059	0.6322
	Departures	0.1637	1.3242	0.1333	0.4422	1.1523	0.3957	0.5022
	scDFM	0.1264	0.7279	0.1069	0.2901	0.8261	0.2539	0.5589
	SameTarget-shift	0.1105	0.5160	0.0897	0.2482	0.7851	0.2318	0.7056
	GIDER (ours)	**0.0938**	**0.4325**	**0.0791**	**0.1988**	**0.6887**	**0.1837**	**0.7267**

Unseen target genes. On Frangieh, GIDER achieves the lowest All EMD, DE20 E-distance, and DE20 EMD, together with the highest PDS cosine (Table 2). Its All EMD is approximately 6.00% below the best baseline for this metric. The advantage is not uniform: scGPT achieves the lowest All and DE20 RMSE, while GEARS leads on All E-distance. On Papalexi, GIDER achieves the lowest values across all six error metrics and the highest PDS cosine, with All EMD approximately 19.62% below the best baseline. Its PDS cosine of 0.5141 exceeds biolord's 0.5079 but remains close to the random-ranking expectation of 0.5.

Overall, GIDER achieves the lowest All EMD and DE20 EMD in all four settings, supporting improved gene-wise marginal distribution prediction across both evaluation gene sets. Gains in population-mean prediction, multivariate distributional fit, and condition discrimination remain dependent on the dataset and task.

Table 2: Unseen target gene (UG) prediction on Frangieh and Papalexi. All and DE20 denote all modeled genes and the top 20 genes in the differential-expression ranking, respectively. E-dist. denotes energy distance; PDS cosine is the cosine-based perturbation discrimination score. SameTarget-shift is not applicable because the UG split excludes all training perturbations of each test target gene. Bold indicates the best reported value within each dataset and metric.

Dataset	**Model**	**All** ↓			**DE20** ↓			**PDS** ↑
		RMSE	E-dist.	EMD	RMSE	E-dist.	EMD	cosine
Frangieh	Ridge	0.0977	0.3399	0.0767	0.1532	0.7075	0.1260	0.5057
	GEARS	0.0960	**0.2916**	0.0779	0.1569	0.7238	0.1307	0.5074
	scGPT	**0.0854**	2.9298	0.2035	**0.1383**	0.9451	0.1335	0.5025
	biolord	0.0922	3.0848	0.2093	0.1445	0.9880	0.1342	0.5122
	DOLORIS	0.1065	0.7159	0.1058	0.1591	0.6980	0.1283	0.4982
	Departures	0.1614	1.4124	0.1359	0.1940	0.8180	0.1618	0.5007
	scDFM	0.1092	0.5804	0.0946	0.1547	0.7284	0.1323	0.4888
	GIDER (ours)	0.0909	0.3259	**0.0721**	0.1425	**0.6835**	**0.1150**	**0.5219**
Papalexi	Ridge	0.1106	0.7507	0.0795	0.1936	0.6181	0.1617	0.4807
	GEARS	0.0942	0.5167	0.0785	0.1895	0.6088	0.1627	0.4754
	scGPT	0.1011	2.6908	0.1969	0.1534	0.9272	0.2045	0.4860
	biolord	0.1133	3.3836	0.2314	0.1475	1.0843	0.2280	0.5079
	DOLORIS	0.1036	0.7939	0.0945	0.1774	0.5732	0.1460	0.5000
	Departures	0.1803	1.5961	0.1340	0.2638	0.8270	0.2257	0.4991
	scDFM	0.1276	0.9764	0.1061	0.1537	0.5124	0.1243	0.4895
	GIDER (ours)	**0.0794**	**0.3678**	**0.0631**	**0.1356**	**0.4660**	**0.1098**	**0.5141**

Ablation results. On Papalexi SG, the full model outperforms all ablated variants on the four reported metrics, supporting the predictive contributions of candidate-site information, long-range DNA context, and population-response auxiliary supervision in this setting. Replacing two-stage conditioning with Joint Fusion increases All RMSE by 10.2% and decreases PDS cosine by 4.4%, supporting the predictive value of query-guided reading and gated aggregation as an integrated design; removing RNA auxiliary supervision produces the largest degradation, increasing All RMSE and All E-distance by 13.8% and 9.2%, respectively, and decreasing PDS cosine by 17.4%.

5 CONCLUSION

We presented GIDER, a framework for predicting perturbation-induced expression distributions from unpaired single-cell data. Its two-stage conditioning, inspired by gene-editing mechanisms, explicitly links sgRNA–site relationships to the corresponding DNA contexts and combines supervision from observed population responses with discrete state prediction and conditional magnitude generation. Evaluations on unseen-sgRNA and unseen-target-gene tasks across two datasets demonstrate advantages in expression distribution prediction, while controlled ablations on Papalexi further support the value of editing information and how it is organized within the model. These findings provide empirical support for using the dependence between targeting relationships and site contexts as a structural prior for prediction. However,the urrent approach uses predicted candidate sites and reference sequences without explicitly modeling actual DNA editing outcomes. Future work could incorporate experimentally measured editing outcomes and cell-specific regulatory information to assess the site associations used by the model and its applicability across a broader range of cellular backgrounds, further supporting the evaluation of editing designs before experimentation.

ETHICS STATEMENT

This study conducts computational analyses of publicly available single-cell perturbation data to support the investigation of transcriptomic responses to CRISPR–Cas9 perturbations. Candidate off-target sites used by the model are identified computationally and do not represent experimentally confirmed off-target events. The predictions are intended for scientific research and the exploration of biological hypotheses; they do not establish the safety or clinical efficacy of genome editing.

Downstream biomedical applications require independent experimental validation and compliance with applicable ethical review requirements.

Reproducibility statement

Section 3 describes the model architecture, training objectives, and inference procedure. Appendix A.1 documents the datasets, preprocessing, and task splits; Appendices A.2–A.4 provide editing-feature construction, implementation details, and training and sampling settings. Evaluation metrics and aggregation procedures are specified in Appendix A.5, ablation configurations in Appendix A.6, and baseline adaptations in Appendix A.7. These descriptions specify the experimental protocol used to obtain the reported results.We will make the source code publicly available upon publication.

References

Arc Institute. Cell-Eval2: Gpu-accelerated evaluation of perturbation prediction models. GitHub repository, 2026. URL `https://github.com/ArcInstitute/cell-eval2`. Accessed September 25, 2026.

Sangsu Bae, Jeongbin Park, and Jin-Soo Kim. Cas-offinder: a fast and versatile algorithm that searches for potential off-target sites of cas9 rna-guided endonucleases. *Bioinformatics*, 30(10): 1473–1475, 2014.

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.

Changxi Chi, Yufei Huang, Jun Xia, Jiangbin Zheng, Yunfan Liu, Zelin Zang, and Stan Z Li. Departures: Distributional transport for single-cell perturbation prediction with neural schrödinger bridges. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 20544–20552, 2026a.

Changxi Chi, Jun Xia, Yufei Huang, Zhuoli Ouyang, Cheng Tan, Yunfan Liu, Jingbo Zhou, Chang Yu, Liangyu Yuan, Siyuan Li, et al. Doloris: Dual conditional diffusion implicit bridges with sparsity masking strategy for unpaired single-cell perturbation estimation. In *International Conference on Learning Representations*, volume 2026, pp. 52808–52823, 2026b.

Le Cong, F Ann Ran, David Cox, Shuailiang Lin, Robert Barretto, Naomi Habib, Patrick D Hsu, Xuebing Wu, Wenyan Jiang, Luciano A Marraffini, et al. Multiplex genome engineering using crispr/cas systems. *Science*, 339(6121):819–823, 2013.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature methods*, 21(8):1470–1480, 2024.

Weian Du, Liang Zhao, Kaichuan Diao, Yangyang Zheng, Qianyong Yang, Zhenzhen Zhu, Xiangxing Zhu, and Dongsheng Tang. A versatile crispr/cas9 system off-target prediction tool using language model. *Communications Biology*, 8(1):882, 2025.

Chris J Frangieh, Johannes C Melms, Pratiksha I Thakore, Kathryn R Geiger-Schuller, Patricia Ho, Adrienne M Luoma, Brian Cleary, Livnat Jerby-Arnon, Shruti Malu, Michael S Cuoco, et al. Multimodal pooled perturb-cite-seq screens in patient models define mechanisms of cancer immune evasion. *Nature genetics*, 53(3):332–341, 2021.

Boyang Fu, George Dasoulas, Sameer Gabbita, Xiang Lin, Shanghua Gao, Xiaorui Su, Soumya Ghosh, and Marinka Zitnik. Strand: Sequence-conditioned transport for single-cell perturbations. *arXiv preprint arXiv:2602.10156*, 2026.

Siyu He, Yuefei Zhu, Daniel Naveed Tavakol, Haotian Ye, Yeh-Hsing Lao, Zixian Zhu, Cong Xu, Shradha Chauhan, Guy Garty, Raju Tomer, et al. Squidiff: predicting cellular development and responses to perturbations using a diffusion model. *Nature methods*, 23:65–77, 2026.

Johannes C Hingerl, Alexander Karollus, and Julien Gagneur. Flashzoi: an enhanced borzoi for accelerated genomic analysis. *Bioinformatics*, 41(9):btaf467, 2025.

Patrick D Hsu, David A Scott, Joshua A Weinstein, F Ann Ran, Silvana Konermann, Vineeta Agarwala, Yinqing Li, Eli J Fine, Xuebing Wu, Ophir Shalem, et al. Dna targeting specificity of rna-guided cas9 nucleases. *Nature biotechnology*, 31(9):827–832, 2013.

Dominik Klein, Jonas Simon Fleck, Daniil Bobrovskiy, Lea Zimmermann, Sören Becker, Alessandro Palma, Leander Dony, Alejandro Tejada-Lapuerta, Guillaume Huguet, Hsiu-Chuan Lin, et al. Cellflow enables generative single-cell phenotype modeling with flow matching. *bioRxiv*, pp. 2025–04, 2025.

Ziyue Lin, Jiahe Hou, Hongyu Xia, Xinrui Xie, Feifei Wang, Yuyin Zhou, Wei Wang, Jiawei Liu, and Liangqiong Qu. Decoupled residual denoising diffusion models for unified and data efficient image-to-image translation. *2026 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Leon Hetzel, Yuge Ji, Ignacio L Ibarra, Sanjay R Srivatsan, Mohsen Naghipourfar, Riza M Daza, Beth Martin, et al. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular systems biology*, 19(6):MSB202211517, 2023.

Efthymia Papalexi, Eleni P Mimitou, Andrew W Butler, Samantha Foster, Bernadette Bracken, William M Mauck III, Hans-Hermann Wessels, Yuhan Hao, Bertrand Z Yeung, Peter Smibert, et al. Characterizing the molecular regulation of inhibitory immune checkpoints with multimodal single-cell screens. *Nature genetics*, 53(3):322–331, 2021.

Stefan Peidli, Tessa D Green, Ciyue Shen, Torsten Gross, Joseph Min, Samuele Garda, Bo Yuan, Linus J Schumacher, Jake P Taylor-King, Debora S Marks, et al. scperturb: harmonized single-cell perturbation data. *Nature Methods*, 21(3):531–540, 2024.

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

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

Frederik Wenkel, Wilson Tu, Cassandra Masschelein, Hamed Shirzad, Liam Hodgson, Ihab Bendidi, Cian Eastwood, Shawn T Whitfield, Craig Russell, Yassir El Mesbahi, et al. Txpert: using multiple knowledge graphs for prediction of transcriptomic perturbation effects. *Nature Biotechnology*, pp. 1–8, 2026.

Chenglei Yu, Chuanrui Wang, Bangyan Liao, and Tailin Wu. scdfm: Distributional flow matching model for robust single-cell perturbation prediction. In *International Conference on Learning Representations*, volume 2026, pp. 72511–72535, 2026.

Hengshi Yu, Weizhou Qian, Yuxuan Song, and Joshua D Welch. Perturbnet predicts single-cell responses to unseen chemical and genetic perturbations. *Molecular systems biology*, 21(8):960–982, 2025.

Wei Zhang, Xun Jiang, Yuesi Xi, and Ming Tang. Cistranscell: Single-cell perturbation prediction via gene function, regulatory control, and cellular context. *arXiv preprint arXiv:2606.13713*, 2026.

A Appendix

AI use statement

We used generative AI tools for language editing of the manuscript and programming assistance. The authors take responsibility for the final manuscript, implementation, experimental results, and scientific claims, including content developed with AI assistance.

B DATASETS, PREPROCESSING, AND TASK SPLITS

Data sources and filtering. We use the Frangieh and Papalexi single-cell CRISPR–Cas9 perturbation datasets, which investigate immune evasion in melanoma cells (Frangieh et al., 2021) and immune-checkpoint regulation in THP-1 cells (Papalexi et al., 2021), respectively. RNA data are obtained from a fixed release of scPerturb (Peidli et al., 2024), and editing conditions are cross-checked against the experimental cell metadata and sgRNA sequence tables. We predict RNA expression only; protein measurements are used neither as model inputs nor as prediction targets.

For Frangieh, we retain cells with an unambiguous assignment to a single targeting or non-targeting sgRNA and a total RNA count of at least 500 across the complete source gene set. We exclude perturbation records for which the intended on-target site cannot be uniquely identified or the intended target gene lacks an available expression coordinate. For Papalexi, we use the authors' single-sgRNA assignments and exclude cells assigned to CUL3g2, whose target-gene annotation is inconsistent with the sequence table. We do not further filter cells using Mixscape response labels. Table 3 summarizes the retained data.

Table 3: Dataset statistics after filtering. The sgRNA counts include targeting sgRNAs only.

Dataset	Total cells	Perturbed cells	Control cells	Targeting sgRNAs	Target genes
Frangieh	115,818	106,863	8,955	719	240
Papalexi	20,602	18,216	2,386	97	25

The Frangieh experimental backgrounds are Control, IFNγ, and Co-culture. Here, Control is the name of an experimental background and is distinct from a cell's status as a non-targeting control. The Papalexi background labels are `rep1|tx`, `rep2|tx`, and `rep3|tx`, denoting three experimental replicates under the same treatment.

Expression preprocessing. For cell i and gene k, we compute

$$u_{ik} = \log\left(1 + 10^4 \frac{n_{ik}}{L_i}\right), \tag{15}$$

where n_{ik} is the raw RNA count and L_i is the cell's library size over the complete source gene set. Library sizes are not recomputed after selecting the modeled genes.

For each task, highly variable genes are selected using only training control and training perturbed cells. We also retain expression coordinates for known, measurable target genes, yielding a fixed panel of $G = 2{,}000$ genes. The identities of held-out target genes may be used to define output coordinates, but their test perturbation expression profiles are used neither for highly variable gene selection nor for training. A separate gene panel is constructed for each task.

Task splits. In the unseen-sgRNA (SG) task, all perturbed cells associated with a held-out sgRNA are excluded from training across every experimental background, while other sgRNAs targeting the same gene remain in training. In the unseen-target-gene (UG) task, all sgRNAs targeting a test gene, together with their cells across all backgrounds, are held out. UG therefore withholds the perturbation responses associated with the target gene; expression of that gene can still be observed in other training cells. Table 4 reports the split sizes. A condition denotes an observed sgRNA–background combination. All methods use the same data splits and test conditions within a task.

Table 4: Training/test split sizes. Each entry reports training / test counts.

Setting	Perturbed cells	sgRNAs	Target genes	Conditions
Frangieh SG	70,889 / 35,974	479 / 240	240 / 240	1,426 / 708
Frangieh UG	86,777 / 20,086	573 / 146	191 / 49	1,705 / 429
Papalexi SG	14,111 / 4,105	72 / 25	25 / 25	214 / 75
Papalexi UG	14,208 / 4,008	77 / 20	20 / 5	229 / 60

Evaluation inputs. Training and evaluation controls are disjoint: 7,163 and 1,792 cells for Frangieh, and 1,908 and 478 for Papalexi, respectively. For methods requiring control-expression inputs, we use the same fixed set of 100 distinct evaluation controls within each background. Each method exports 100 expression vectors per test condition, which are compared with all observed perturbed cells for that condition. Methods that predict only population means repeat those estimates for distributional evaluation; these repeated vectors are not independent generated cells. Method-specific output construction is described in Appendix H.

C CONSTRUCTION OF EDITING-CONDITION FEATURES

Intended on-target and candidate off-target sites. Editing conditions are constructed from the experimentally specified 20-nt spacer sequence and intended target gene. We identify the intended on-target site (ON) using the GRCh38 reference genome and gene annotations, search for candidate off-target sites (OFF) with Cas-OFFinder (Bae et al., 2014), and rank the candidates with CCLMoff (Du et al., 2025). The search uses an NGG PAM, permits at most six mismatches, and excludes insertion or deletion bulges. Only the confirmed ON site is removed from the candidate list; perfect matches at other genomic locations remain eligible as candidate OFF sites.

We retain at most $K = 3$ OFF sites per sgRNA, giving one ON site and up to three OFF sites. Missing candidates are excluded using validity masks. CCLMoff scores are used only for ranking and are not supplied to the prediction model. Neither the candidate sites nor their scores are interpreted as observed editing events or cleavage probabilities in the experiment.

Short site features. Each site is represented by a 184-dimensional feature vector comprising:

- an 80-dimensional one-hot encoding of the 20-nt spacer;
- a 92-dimensional one-hot encoding of the 23-nt site DNA sequence, including the PAM; and
- 12 auxiliary sequence and positional features: an ON indicator; mismatch proportions over the full spacer and its PAM-proximal and PAM-distal regions; strand orientation; chromosome and reference cleavage-site position; chromosome and distance relationships to the ON site; the relationship between the reference cleavage site and the annotated interval of the intended target gene; and the GC fractions of the spacer and site sequences.

The reference cleavage site is inferred from sequence coordinates and is not an experimentally measured editing position. The validity mask indicates only whether a usable candidate exists at the corresponding input position.

Long-range DNA context. For each ON/OFF site, we extract a 524,288-bp reference DNA window centered on the reference cleavage site. Windows on the negative strand are reverse-complemented, and positions beyond chromosome boundaries are padded with N. In our implementation, frozen Flashzoi (Hingerl et al., 2025) provides an internal representation of size $6{,}144 \times 1{,}536$. We average-pool its tokens within eight contiguous regions and additionally average the central three tokens, yielding nine regional vectors of dimension 1,536. The central representation overlaps with the corresponding regional representations. These features are derived solely from reference DNA and contain neither measured chromatin states of the target cells nor post-perturbation RNA measurements.

Target-gene prior. The intended target gene is represented by a static 512-dimensional vector obtained from the frozen scGPT GeneEncoder, including its pretrained layer normalization (Cui et al., 2024). A trainable linear projection shared across all target genes maps this vector to 64 dimensions. This procedure does not read test-cell expression or learn a separate conditioning identity vector for each target gene. Different sgRNAs targeting the same gene share this prior but retain their own sequence, site, and context features.

D NETWORK IMPLEMENTATION

Two-stage condition modeling. The hidden dimension is 64. A two-layer multilayer perceptron (MLP) encodes the short site features. The resulting representation is added to the target-gene projection and site-role embedding to construct each site query. Independent sigmoid gates allow multiple sites to contribute simultaneously.

In the second stage, each query reads only the DNA context of its corresponding site, and the first-stage site weights are reused for aggregation. The context pathway uses layer normalization without affine parameters, a bias-free feature projection, and a bias-free local MLP. Regional positional embeddings contribute only to the attention keys.

Gene-wise response mapping. The targeting and context response heads have independent parameters. For branch $j \in \{T, C\}$, output gene k, and condition $c = (g, e)$, the predicted response is

$$\widehat{d}_{j,c,k} = \frac{0.01}{\sqrt{64}} \left\langle f_j(h_j),\, E^{\text{gene}}_{j,k} + E^{\text{env}}_{j,e} + \text{Linear}_j(b_{e,k}) \right\rangle, \tag{16}$$

where $E^{\text{gene}}_{j,k}$ and $E^{\text{env}}_{j,e}$ are embeddings of the output gene and experimental background, respectively, and b_e is the fixed training-control reference. The map f_j is a bias-free $64 \to 64 \to 64$ MLP with an intermediate SiLU activation. Both heads predict signed expression changes in a fixed output gene coordinate system. The output-gene embeddings in these heads are distinct from the target-gene prior used to encode the editing condition.

For a fixed sgRNA, background effects in each response head are restricted to a gene-shared offset and a shared scalar multiplying the background-specific control reference. The generator additionally uses control expression and background conditioning.

With a zero-preserving projection of the aggregated group condition, the context pathway satisfies

$$H = 0 \implies h_C = 0 \implies \widehat{d}_C = 0 \implies c_A = 0. \tag{17}$$

This property constrains the context-conditioning pathway; the amplitude network can still use control expression, discrete states, and the experimental background.

Amplitude conditioning and generation networks. The context response is transformed by taking gene-wise absolute values and aggregating within fixed gene-graph groups. We use 16 groups of 125 genes, formed by partitioning a deterministic traversal order of the graph. These groups compress the response condition and are not assumed to represent biological pathways. The mean group condition is projected and added globally to the amplitude network.

The state, residual, and noise networks have independent parameters. Each consists of two graph-attention layers with two 32-dimensional attention heads per layer. The networks use pre-normalization, SiLU activations, residual connections, and output dropout of 0.05; dropout on attention coefficients is zero. The gene graph combines expression correlations in training controls with similarities between scGPT gene vectors. Nearest-neighbor edges from the two sources are merged and symmetrized, and self-loops are added. Each experiment uses a precomputed, fixed graph; all variants within an ablation study share the same graph and grouping.

The full model has 884,815 trainable parameters, excluding the frozen models used for feature precomputation.

E TRAINING AND INFERENCE SETTINGS

Expression scale and control references. Within each task, the model uses scaled expression $x = u/\kappa$, where u is the $\log(1+\text{CP10K})$ expression and κ is the maximum expression value among the training control and perturbed cells. Predictions are multiplied by κ before evaluation. For each background e, the response-head reference b_e is the mean scaled expression of a fixed subset of 100 training control cells. The auxiliary response target instead subtracts the mean of all training controls in that background from the corresponding training perturbed mean. The main experiment and its corresponding ablations share the frozen expression scale, control references, gene graph,

gene grouping, and magnitude scale. All such quantities are determined from the training split and retained at inference.

Population-response supervision. Each response head predicts the same complete, signed population response. The gene-wise normalization scale is the root mean square of the responses across training conditions, with a lower bound of 0.01. The auxiliary loss weights the distinct conditions represented in the current batch equally. Deduplication is performed within each batch only and does not imply uniform sampling of conditions over training. The response targets and normalization scales are computed before training and held fixed. At inference, only the model-predicted response conditions are used.

OT endpoint construction. At each training epoch, we reconstruct an entropically regularized optimal transport (OT) coupling within each sgRNA–background condition. For a condition containing n training perturbed cells, we sample n cells without replacement from the training-control pool in the same background. The transport cost is the squared Euclidean distance in the model expression space, normalized by the maximum entry of the cost matrix. Both marginal distributions are uniform, and the entropic regularization coefficient is 0.1. If all costs are zero, we use the independent uniform coupling.

We then sample n endpoint pairs with replacement according to the joint transport probabilities to construct state and amplitude supervision. All variants in a given ablation study reuse the same training endpoints for the corresponding epoch. The coupling uses training data only and provides statistical endpoint associations, not measured before-and-after trajectories of individual cells.

The amplitude scale γ is estimated as the root mean square of all nonzero expression differences in training OT endpoint pairs and is subsequently held fixed. The four states are defined by exact comparisons between endpoint expression values, without a threshold for small changes. Both decreasing states are masked when control expression is zero. Continuous losses are evaluated only at coordinates requiring amplitude generation, normalized by the number of active coordinates in each cell, and then averaged across the batch.

Optimization. GIDER and all ablation variants are trained for 50 epochs using AdamW, with a peak learning rate of 10^{-3}, weight decay of 10^{-2}, and batch size of 64. The full model uses a population-response auxiliary loss weight of 1. All ablation variants share the optimization settings and learning-rate schedule of the corresponding full-model run. Table 5 summarizes the principal settings. Baseline-specific conditioning representations and dataset adaptations are described in Appendix H.

Table 5: Optimization and inference settings for GIDER and its ablated variants.

Setting	Value
Optimizer	AdamW
Peak learning rate	10^{-3}
Weight decay	10^{-2}
Batch size	64
Learning-rate schedule	One epoch of linear warmup, followed by cosine decay
RNA auxiliary loss weight (full model)	1
Continuous-path time sampling	Independent samples from $\mathcal{U}(0.001, 0.999)$ for the two segments
Inference integration steps	20 for translation and 20 for denoising

Inference. The model first samples a discrete state vector once, conditioned on control expression and the targeting response. Gaussian noise is initialized on the active coordinates, followed by amplitude translation and denoising. Both segments use explicit Euler integration, with no additional noise injected at their junction. Control expression, editing conditions, background, discrete states, and active coordinates remain fixed throughout generation. Intermediate continuous states may be negative; state-specific amplitude constraints are applied only at the endpoint. Each input control cell produces one expression sample, giving 100 predicted cells per test condition.

F EVALUATION METRICS AND AGGREGATION

The main text reports seven metrics: RMSE, E-distance, and EMD on all modeled genes (All) and the top 20 genes in the differential-expression ranking (DE20), together with PDS cosine.

Evaluation gene sets. All contains the 2,000 modeled genes for the corresponding task. For DE20, genes are ranked using the absolute Welch-type statistic comparing observed test perturbed cells with all evaluation controls from the same background:

$$T_k = \frac{|\overline{y}_k - \overline{x}_k|}{\sqrt{s^2_{y,k}/n_y + s^2_{x,k}/n_x + 10^{-12}}}, \tag{18}$$

where $s^2_{y,k}$ and $s^2_{x,k}$ are sample variances, and n_y and n_x are the respective cell counts. For each condition, all methods use the same 20 highest-ranked genes. This set is used only for evaluation and does not constitute a set of statistically significant differentially expressed genes selected after multiple-testing correction.

Let $Y = \{y_i\}_{i=1}^n$ and $\widehat{Y} = \{\widehat{y}_j\}_{j=1}^m$ denote the observed and predicted populations, restricted to the d evaluation genes. The vectors in this subsection are in the unscaled $\log(1 + \mathrm{CP10K})$ space.

Population-mean error. RMSE compares the two population means:

$$\mathrm{RMSE}(Y, \widehat{Y}) = \sqrt{\frac{1}{d}\sum_{k=1}^{d}\left(\overline{Y}_k - \overline{\widehat{Y}}_k\right)^2}. \tag{19}$$

This metric does not require cell-wise pairing.

Multivariate distributional discrepancy. We compute E-distance as

$$\begin{aligned}\widehat{E}_U = \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\|y_i - \widehat{y}_j\|_2 \\ - \frac{1}{n(n-1)}\sum_{i\neq i'}\|y_i - y_{i'}\|_2 \\ - \frac{1}{m(m-1)}\sum_{j\neq j'}\|\widehat{y}_j - \widehat{y}_{j'}\|_2, \\ D_E(Y, \widehat{Y}) = \sqrt{\max(\widehat{E}_U, 0)}.\end{aligned} \tag{20}$$

The within-population sums exclude diagonal pairs. The internal U-statistic estimate is clipped at zero before taking the square root to obtain the reported value.

Marginal distributional discrepancy. EMD is the mean of the gene-wise one-dimensional Wasserstein-1 distances:

$$\mathrm{EMD}(Y, \widehat{Y}) = \frac{1}{d}\sum_{k=1}^{d} W_1(P_k, \widehat{P}_k), \tag{21}$$

where P_k and $\widehat{P}_k$ are the observed and predicted empirical distributions for gene k. This metric accommodates unequal population sizes and measures discrepancies between the marginal gene-expression distributions.

Perturbation discrimination. We compute PDS cosine using the ranking and target-gene exclusion conventions of Cell-Eval2 (Arc Institute, 2026), with response vectors defined as follows. Within each background, the candidate set consists of the test sgRNAs. Observed and predicted responses are obtained by subtracting the mean expression of the same 100 evaluation control cells from the respective population means. For all comparisons within a candidate set, we exclude the

union of measurable target-gene coordinates associated with its sgRNAs, so that every comparison uses the same gene coordinates.

For each predicted response, we rank all observed candidate responses by increasing cosine distance. If the correct sgRNA has zero-based average rank r_g among $N > 1$ candidates, its score is

$$\mathrm{PDS}_{\mathrm{cosine}}(g) = 1 - \frac{r_g}{N-1}. \tag{22}$$

Ties receive average ranks. The score measures the rank of the matching perturbation, rather than the cosine similarity of a single response pair or a classification accuracy. It equals 0.5 when all candidates tie and has expectation 0.5 under a uniformly random ranking. In the current SG splits, each target gene contributes one test sgRNA; PDS on this candidate set therefore evaluates retrieval across test perturbation conditions, rather than separately assessing discrimination among sgRNAs targeting the same gene.

Metric aggregation. For each metric, we first average equally over valid test conditions within each experimental background and then average equally over backgrounds. Conditions are not weighted by their numbers of perturbed cells. If an observed population contains fewer than two cells, E-distance and the DE20 evaluations requiring sample variances are undefined; these conditions are excluded from the corresponding averages rather than assigned zero. The same condition-validity rules are applied to every method.

Table 6: Controlled ablations on Papalexi SG. All variants retain $\mathcal{L}_{\mathrm{gen}}$. The final column lists the retained response auxiliary losses.

Variant	Modification	Auxiliary loss
ON-only	Mask OFF validity indicators, short features, and long DNA contexts; retain the ON site and the original computational pathways.	$\frac{1}{2}\mathcal{L}_T + \frac{1}{2}\mathcal{L}_C$
w/o Long-DNA	Set long DNA inputs to zero at every site; retain short features, site roles, and the target-gene prior.	$\frac{1}{2}\mathcal{L}_T$
Joint Fusion	Retain identical inputs and within-site correspondence; replace query-guided reading and gated aggregation with within-site joint MLP encoding, regional mean pooling, and summation over sites.	$\frac{1}{2}\mathcal{L}_T + \frac{1}{2}\mathcal{L}_C$
w/o RNA supervision	Remove response auxiliary losses while retaining both response heads and their generation-conditioning pathways.	None

G Ablation design for Papalexi SG

We examine candidate off-target information, long-range DNA context, the overall two-stage conditioning design, and population RNA-response supervision on Papalexi SG. All variants are retrained and share the data split, gene panel, training assets, generation-network architecture, and evaluation inputs. Each is compared with the full model trained under the same ablation configuration. Joint Fusion retains identical inputs and within-site correspondence but jointly replaces query-guided context retrieval and gated aggregation; this comparison assesses the overall conditioning design rather than the independent effects of its individual operations.

Let $\mathcal{L}_T$ and $\mathcal{L}_C$ denote the standardized mean squared errors of the targeting and context response heads, respectively. The full-model objective is

$$\mathcal{L}_{\mathrm{full}} = \mathcal{L}_{\mathrm{gen}} + \tfrac{1}{2}\mathcal{L}_T + \tfrac{1}{2}\mathcal{L}_C, \tag{23}$$

where

$$\mathcal{L}_{\mathrm{gen}} = \mathcal{L}_Q + \mathcal{L}_R + \sigma_A^2 \mathcal{L}_N \tag{24}$$

comprises the state-classification, amplitude-residual prediction, and noise-prediction losses. Table 6 specifies the changes and retained auxiliary losses for each variant. For *w/o Long-DNA*, the context response is structurally zero, and only the targeting response auxiliary loss is retained.

Table 7: Conditioning representations and principal adaptations of the evaluated external baselines.

Method	Conditioning and adaptation
GEARS (Roohani et al., 2024)	Uses a target-gene GO graph, a co-expression graph constructed from training expression, and background-matched controls. Five Frangieh targets lacking GO information are retained as nodes with self-loops.
scGPT (Cui et al., 2024)	Uses target-gene perturbation indicators. The gene panel is aligned with the pretrained vocabulary, with trainable tokens added for missing genes. Training samples 1,536 gene tokens; inference uses all 2,000 modeled genes.
biolord (Piran et al., 2024)	Uses GO-derived conditions and the genetic-perturbation mean-prediction configuration. Each predicted mean is repeated 100 times for evaluation.
DOLORIS (Chi et al., 2026b)	Retains the expression-generation and Mask branches, using whole-human scGPT features for the graph prior. When training conditions are insufficient, the PCA dimension is reduced and the resulting features are zero-padded to 32 dimensions.
Departures (Chi et al., 2026a)	Retains target-gene and background conditioning, OT coupling, and continuous/discrete generation. The graph prior uses whole-human scGPT features.
scDFM (Yu et al., 2026)	Generates perturbed expression distributions conditioned on control states using conditional flow matching and a maximum mean discrepancy (MMD) objective. Evaluated using the common data splits, gene panel, and scoring protocol.

H Baseline Implementations and Adaptations

Common protocol. External baselines are adapted from author-released implementations to the same data splits, a fixed 2,000-gene panel within each task, and a common evaluation protocol. Data-dependent statistics are estimated from training cells only. Predictions are restored to the common expression scale, and negative values are set to zero. Methods requiring control-expression inputs receive the same evaluation controls within each experimental background.

External methods. Table 7 summarizes the conditioning representations and principal adaptations used for each method. These baseline configurations use gene-level perturbation conditions and do not directly receive sgRNA sequences. For biolord, repeating the predicted mean yields a point-mass distribution without model-generated cell-to-cell variation.

Ridge-regression baseline. The regressor maps pooled editing features and target-gene indicators z_g to the absolute perturbed population mean, $\widehat{\mu}_g = Wz_g + b$. Training guide–background conditions receive equal weights, the ridge penalty is 1, and feature standardization uses training guides only. The regressor receives neither background labels nor control expression. In the common $\log(1 + \mathrm{CP10K})$ evaluation space, predictions are constructed as

$$\widehat{y}_{i,g,e} = \max\left(0, \widehat{\mu}_g + x_{i,e}^{\mathrm{ctrl}} - \bar{x}_e^{\mathrm{ctrl,tr}}\right), \tag{25}$$

where $\bar{x}_e^{\mathrm{ctrl,tr}}$ is the mean expression of all training controls in background e. Thus, centered control variation is introduced at prediction time without an additional learned distribution generator.

Same-target response transfer. We evaluate the parameter-free SameTarget-shift baseline on both Frangieh SG and Papalexi SG. For a test sgRNA g and background e, let $\mathcal{T}^{\mathrm{tr}}_{\tau(g),e}$ denote the training sgRNAs targeting the same gene under that background. We average their training perturbed population means with equal sgRNA weights:

$$\bar{\mu}_{\tau(g),e} = \frac{1}{|\mathcal{T}^{\mathrm{tr}}_{\tau(g),e}|} \sum_{g' \in \mathcal{T}^{\mathrm{tr}}_{\tau(g),e}} \bar{y}^{\mathrm{tr}}_{g',e}. \tag{26}$$

The transferred response is the difference between this mean and the corresponding training-control mean, giving

$$\widehat{y}^{\text{shift}}_{i,g,e} = \max\bigl(0,\, x^{\text{ctrl}}_{i,e} + \bar{\mu}_{\tau(g),e} - \bar{x}^{\text{ctrl,tr}}_{e}\bigr)\,. \tag{27}$$

Here, $x^{\text{ctrl}}_{i,e}$ is an evaluation control cell, and $\bar{x}^{\text{ctrl,tr}}_{e}$ is the mean of all training controls under background e. Each test condition produces 100 expression vectors without using test perturbed expression to construct predictions. SameTarget-shift is not applicable to UG because that split excludes all training perturbations of each test target gene.

Although STRAND (Fu et al., 2026) mentions pipeline code in its supplementary materials, we could not locate a publicly accessible author implementation as of September 26, 2026. We therefore do not include it in the quantitative comparison.