Elucidating the Design Space of Generative Models for Single-Cell Perturbation Prediction

Anonymous Authors[1]

Abstract

We introduce `ExpressionVAE`, the first discrete-latent perturbation model for single-cell data: a vector-quantized variational autoencoder paired with a perturbation-conditioned discrete prior. On Replogle and Parse 1M it achieves state-of-the-art on every distributional and cell-eval state metric we evaluate, with order-of-magnitude gaps on Fréchet distance and MMD^2 over the strongest continuous-latent baseline. We test two prior families (autoregressive and masked discrete diffusion) and find they achieve effectively identical numbers, isolating the gain to the discrete latent. A controlled output-head ablation further reveals a single design axis governing decoder-head choice, the richness of the inference-time sampling distribution, with standard evaluation metrics partitioning into three groups whose rankings flip along it. Finally, on a held-out CRISPRi reversion benchmark of 1,732 perturbations under inflammatory cytokine stress, the frozen encoder effectively matches scGPT model (trained on $10\times$ larger dataset) on biological selectivity.

1. Introduction

A cell can be viewed as an information-processing system that maps environmental inputs to molecular responses through an internal regulatory program (Fitch, 2021). While the structure of the regulatory network itself is not directly accessible, single-cell RNA sequencing provides a high-dimensional partial observation of its activity (Macosko et al., 2015; Gulati et al., 2020; Zeng & Dai, 2019). Crucially, this observation is discrete: we do not measure continuous expression levels, but count individual molecular events. Across many cells and conditions, these observations implicitly encode the statistical dependencies induced by the underlying regulatory structure. The objective of a virtual cell model is to learn these dependencies well enough to generate realistic cellular states and to predict how modifications to the regulatory program, such as genetic or chemical perturbations, alter the resulting transcriptome (Dixit et al., 2016; Adamson et al., 2016). If reliable, virtual cells stand to enable a new frontier of in-silico drug discovery.

The central object of observed gene expression is the count matrix: a sparse, zero-inflated, high-dimensional integer-valued tensor recording how many mRNA molecules of each gene were captured from each cell. These properties make direct generative modeling on raw counts remarkably difficult for most models. Autoregressive factorizations impose a sequential ordering on genes, an inductive bias misaligned with their natural exchangeability. Flow matching tends to collapse to the mean under the sparsity of count data. Masked diffusion is the only family that recovers reasonable structure on raw counts (Zhang et al., 2026; Bhattacharya et al., 2026).

These difficulties on raw counts motivate a latent-space approach. Autoregressive models have demonstrated predictable scaling laws on language data (Hoffmann et al., 2022), an appealing property as virtual cell models scale up, and they operate naturally over discrete tokens. This motivates a discrete latent representation.

We introduce `ExpressionVAE`, a vector-quantized variational autoencoder that encodes each cell as a fixed-length sequence of discrete codes, paired with a perturbation-conditioned prior over those codes. Vector quantization commits the encoder to a finite codebook of cell-state representations that any standard discrete prior can model exactly. To control for the choice of prior, we train two from different generative-model families, autoregressive (AR) and masked discrete diffusion (MDLM); any gain that survives the swap is attributable to the latent representation rather than to a particular prior architecture.

Beyond the latent and prior, two further design questions recur in the single-cell perturbation literature, and we treat each as a controlled experiment. First, single-cell models vary substantially in their decoder output head, and published rankings across cross-entropy, MSE, hurdle, and negative-binomial heads are inconsistent in ways that often

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Submitted to the 2026 Workshop on Generative and Agentic AI for Biology (ICML 2026). Do not distribute.

track the choice of evaluation metric rather than any property of the heads themselves. We run a within-architecture head ablation on the same trained backbone to expose the underlying axis. Second, we evaluate the frozen encoder on an out-of-distribution CRISPRi reversion benchmark to test whether the resulting design choices reflect a generalizable inductive bias rather than over-fitting our controlled metric suite.

Concretely, our contributions are:

- **Discrete-latent perturbation modeling.** We introduce the first discrete-latent representation learning approach for single-cell perturbation modeling, and show that pairing `ExpressionVAE` with either an autoregressive or a masked-diffusion prior achieves state-of-the-art on the distributional and cell-eval state metrics we evaluate, on both Replogle and Parse 1M, often by an order of magnitude on MMD^2 and Fréchet distance over the strongest continuous-latent baseline. The two priors achieve essentially identical performance, isolating the gain to the discrete latent representation rather than to the prior architecture (Section 4.1).
- **A measurement-theoretic account of decoder-head choice.** We show that the apparent conflicts between published evaluation metrics resolve into a clean three-group structure (distribution and DE-discovery, mean-ranking, and a hybrid) once a single design axis, the richness of the inference-time sampling distribution, is made explicit. A within-architecture experiment using a parametric negative-binomial head with two inference rules (sample vs. mean) that share trained weights isolates this axis from confounds in capacity, parameterization, and training objective, and yields a metric-to-head mapping for downstream practitioners (Section 4.2).
- **Out-of-distribution generalization.** On a held-out CRISPRi reversion benchmark of 1,732 perturbations under inflammatory cytokine stress (Wong et al., 2026), the hurdle decoder identified by our controlled ablation is the strongest pairing on biological selectivity for both training datasets, and the frozen encoder effectively matches a $10\times$-larger foundation model (scGPT) on this OOD perturbation-ranking task, evidence that the discrete-latent perturbation prior captures inflammation-axis structure efficiently rather than memorizing the in-distribution metric set (Section 4.3).

2. Related Work

We draw from two literatures: single-cell generative and perturbation-response models, and the latent modeling toolkit from vision and language.

Generative and perturbation-response models. scVI (Lopez et al., 2018a;b) established the canonical Gaussian-latent, Negative Binomial-likelihood VAE recipe, elaborated by the conditional-VAE thread of scGen, trVAE, and CPA (Lotfollahi et al., 2019; 2020; 2023) and the latent-diffusion thread of scDiffusion (Luo et al., 2024), CFGen (Palma et al., 2024), and scLDM (Palla et al., 2025) with flow-matching priors (Lipman et al., 2022; Peebles & Xie, 2023); the Hurdle head we evaluate traces to MAST (Finak et al., 2015). Beyond this lineage, perturbation-response models pretrained on tens to hundreds of millions of cells (Geneformer (Theodoris et al., 2023), scGPT (Cui et al., 2024), scFoundation (Hao et al., 2024), STATE (Adduri et al., 2025), Tahoe-x1 (Gandhi et al., 2025)) increasingly condition on ESM-2 embeddings (Lin et al., 2023). All embed raw counts into continuous vectors. A concurrent line generates directly in the discrete count domain, via Score Entropy Discrete Diffusion (DCM (Bhattacharya et al., 2026; Lou et al., 2023)), absorbing-state masked diffusion in the D3PM (Austin et al., 2021) lineage (Lingshu-Cell (Zhang et al., 2026), X-Cell (Wang et al., 2026a), scDiVa (Wang et al., 2026b)), or Stack's in-context approach (Dong et al., 2026). We evaluate on Replogle Perturb-seq (Replogle et al., 2022) and the Parse cytokine atlas (Oesinghaus et al., 2025), with phenotypic reversion (Wong et al., 2026) as our out-of-distribution NFκB experiment.

Latent generative modeling: tokenizers and priors. On the tokenizer side, VQ-VAE (van den Oord et al., 2017; Razavi et al., 2019) and MaskGIT (Chang et al., 2022) introduced learnable discrete codes and masked parallel decoding, Latent Diffusion (Rombach et al., 2022) established the continuous-latent plus diffusion-prior recipe, and FSQ (Mentzer et al., 2023) eliminated the codebook collapse documented in single-cell VQ (Liu et al., 2025). On the prior side, the absorbing-state masked discrete diffusion family has matured rapidly from D3PM (Austin et al., 2021) through SEDD (Lou et al., 2023), MDLM (Sahoo et al., 2024), and MD4 (Shi et al., 2024) to RADD (Ou et al., 2025), which unifies it with any-order autoregression. We combine learned discrete latents with flexible generative priors, a combination no existing model explores.

3. Method

We propose a two-stage pipeline (Fig. 1). Stage 1 trains an `ExpressionVAE` that compresses each cell's gene-expression vector into a fixed-length sequence of discrete codes via a finite scalar quantization (FSQ) bottleneck (Mentzer et al., 2023). Stage 2 freezes the VAE and trains a perturbation-conditioned generative prior over the code sequence, conditioned on an ESM2-3B protein-language-model embedding of the targeted gene and an additive cell-type signal. At inference, the prior is queried with the em-

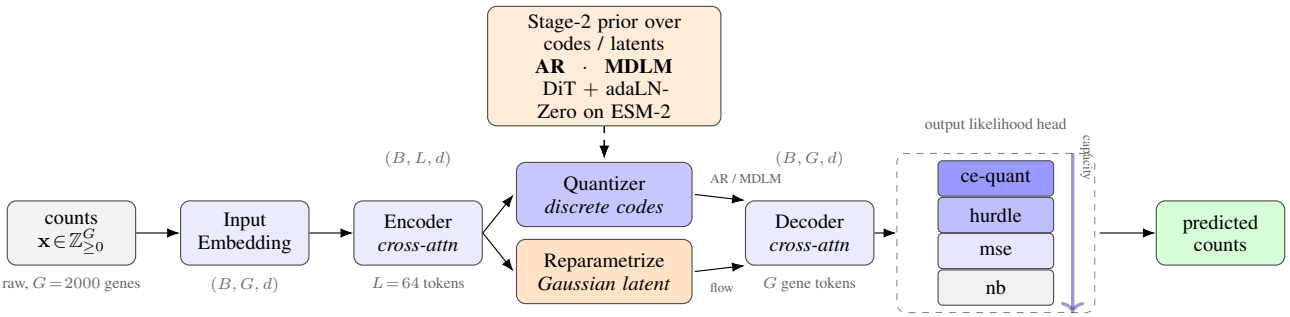

Figure 1. A two-stage pipeline. Stage 1 (the row of boxes): a cross-attention transformer encoder maps raw count vectors $\mathbf{x} \in \mathbb{Z}^G_{\geq 0}$ to L=64 latent tokens, which are either FSQ quantised into discrete codes (consumed by AR or MDLM priors) or reparametrised as continuous Gaussian latents (consumed by flow). A symmetric cross-attention decoder maps these back to per-gene representations, which one of four output likelihood heads converts to predictions. Stage 2 (sidebar): a DiT prior over the frozen codes / latents, conditioned on the ESM-2 perturbation embedding and optionally cell-type embedding via adaLN-Zero.

bedding of an unseen perturbation and the resulting latent is decoded into a predicted gene-expression profile. The two design axes varied across the experiments are the VAE *output head* (Section 4.2) and the choice of *prior* (Section 4.1).

3.1. Stage 1: ExpressionVAE

Encoder and decoder. The encoder maps a count vector to a set of L latent tokens via a cross-attention transformer. Each gene g is first embedded as $\mathbf{e}_g = f_{\text{in}}(x_g, \mathbf{v}_g)$, where f_{in} is an MLP applied to the log1p-normalized scalar count x_g and modulated by a learned gene-identity token $\mathbf{v}_g \in \mathbb{R}^d$. L learnable latent query tokens cross-attend over the G gene embeddings through a stack of transformer layers, producing a per-cell latent matrix $\mathbf{Z} \in \mathbb{R}^{L \times d}$. A symmetric cross-attention transformer decodes the quantized codes back to gene space: G learnable gene query tokens cross-attend over the projected code embeddings, yielding a per-gene representation $\mathbf{H} \in \mathbb{R}^{G \times d}$.

FSQ bottleneck. The encoder output $\mathbf{Z}$ is quantized per-dimension by FSQ (Mentzer et al., 2023), which replaces the learned VQ codebook with a parameter-free per-dimension rounding scheme. Given a level vector $\boldsymbol{\ell} = (\ell_1, \ldots, \ell_d) \in \mathbb{Z}^d_{>0}$, each dimension is bounded via $\tanh$ and rounded to the nearest integer on a uniform grid:

$$\hat{z}_k = \text{round}\left(\frac{\ell_k - 1}{2} \cdot \tanh(z_k)\right), \qquad k = 1, \ldots, d.$$

The result is a per-cell sequence of L discrete codes $\mathbf{c} = (c_1, \ldots, c_L)$ drawn from a vocabulary of size $\prod_k \ell_k$, eliminating the codebook collapse documented in single-cell VQ (Liu et al., 2025). We use $L = 64$ tokens with codebook size $K = 512$. A Gaussian-latent variant used in Section 4.3 also uses $L = 64$ latent tokens with $\beta_{\text{KL}} = 2 \times 10^{-6}$.

Output heads. An output head maps the per-gene representation $\mathbf{H}$ to a conditional distribution $p_\theta(\mathbf{x} \mid \mathbf{H})$. We evaluate four head families, corresponding to standard choices in the single-cell modeling literature, and refer to them by the labels used in Section 4.2.

`ce-quant` places a categorical over $[0, c_{\max}]$ per gene with logits produced by a linear projection of $\mathbf{h}_g$, trained against quantile-binned integer targets and sampled at inference (Section 3.1 below).

`hurdle` (Finak et al., 2015) parameterises a Bernoulli zero-gate $z_g \sim \text{Bern}(1 - p_{0,g})$ together with a regressed positive magnitude μ_g; the loss combines BCE on the gate with MSE on non-zero entries, and inference samples the gate while emitting μ_g deterministically.

`mse` regresses log1p-normalised counts directly, with $p(x_g \mid \mathbf{h}_g) = \mathcal{N}(\mu_g, 1)$, and emits μ_g with no stochastic step.

The negative binomial (NB) family parameterises $\text{NB}(\mu_g, \theta_g)$ per gene with library-size factorisation, following (Lopez et al., 2018a; Palla et al., 2025): $\mu_g = \ell_n \cdot \rho_g$, where $\boldsymbol{\rho} = \text{softmax}_g(\mathbf{w}^\top \mathbf{H})$ is a probability simplex over genes, $\ell_n = \sum_g x_g$ is the per-cell library size, and θ_g is a learned cell-independent dispersion. This family admits two distinct `predict()` rules at inference time without changing trained weights: `nb-sample` draws from $\text{NB}(\mu, \theta)$, and `nb-deter` returns μ directly. We treat these as two separate evaluation points throughout, since the choice of inference rule materially shifts which evaluation metrics the head wins (Section 4.2).

Quantile binning for the categorical head. The `ce-quant` head requires integer targets, so we quantile-bin the raw counts into a vocabulary of 20 bins chosen from the empirical training count distribution: 8 singleton bins for counts 0–7 capture the dominant low-count regime exactly, 7 bins covering 8–63 handle moderate counts at progressively coarser resolution, and 5 bins (64–95, 96–127, 128–191, 192–255, ≥ 256) cover the high-expression tail. Tokenisation maps each raw count to its bin index; detokenisation

maps each predicted bin back to the empirical conditional mean of counts within that bin range, not the arithmetic midpoint, since the within-bin distribution is heavy-tailed and the bias matters most for the unbounded ≥ 256 bucket. This scheme captures 100% of top-DEG per-perturbation means on Parse and 99.69% on Replogle within bounded bins.

3.2. Stage 2: Discrete prior over latent codes

Stage 2 freezes the VAE and trains a generative prior over the code sequence $\mathbf{c} = (c_1, \ldots, c_L)$, conditioned on a projected ESM2-3B perturbation embedding $\mathbf{p}$ and an additive cell-type signal. Both prior families share a DiT backbone with adaLN-Zero conditioning on $\mathbf{p}$ and the cell-type vector. The autoregressive (AR) prior factorises as $p(\mathbf{c} \mid \mathbf{p}) = \prod_{l=1}^{L} p(c_l \mid c_{<l}, \mathbf{p})$ via a causally-masked DiT, requiring L sequential decoding steps at inference. MDLM (Sahoo et al., 2024) trains a bidirectional DiT denoiser under an absorbing-state masking schedule $\alpha_t = 1 - t$, with the standard masked-token loss

$$\mathcal{L}_{\text{MDLM}} = \mathbb{E}_{t,\mathbf{z}_t}\left[\frac{1}{t}\sum_{i:\, z_t^i = \texttt{[M]}} -\log p_\theta(c^i \mid \mathbf{z}_t, \mathbf{p})\right],$$

and denoises all L tokens in a single forward pass. We use the same backbone, parameter count, and training schedule for both, so any gap in Section 4.1 reflects the prior family rather than capacity.

3.3. Inference

Given an unseen perturbation, the corresponding ESM2-3B embedding $\mathbf{p}^*$ is passed to the trained prior, which samples a latent sequence $\hat{\mathbf{c}} \sim p_\theta(\cdot \mid \mathbf{p}^*)$ via L autoregressive steps for AR or a single masked-denoising pass for MDLM. The frozen VAE decoder maps $\hat{\mathbf{c}}$ back to gene space, and the chosen output head emits a predicted profile $\hat{\mathbf{x}} \in \mathbb{R}^G$ following its `predict()` rule. The same frozen encoder is reused without any fine-tuning for the OOD evaluation in Section 4.3.

4. Experiments

Datasets. We evaluate on two large-scale single-cell perturbation screens. The train and holdout splits are matched to scLDM (Palla et al., 2025). Parse 1M consists of cytokine perturbations across approximately one million cells, with a CD4 Naive holdout of 27 cytokines. The Replogle essential-genome CRISPRi screen (Replogle et al., 2022) provides a HepG2 holdout of 372 knock-outs.

Metrics. We evaluate at two complementary levels. *Distribution* metrics (W_2, MMD^2 with an RBF kernel, FD Fréchet Distance) following (Palla et al., 2025) compare generated against true cells at the population level, and are sensitive to per-gene marginals and joint dispersion of the predicted cell cloud. The *cell-eval* per-perturbation metrics from (Adduri et al., 2025) fall into three functional groups by what they measure: (i) *DE-discovery* (*disc_l1*, *PR-AUC*, *Spearman sig*) compare FDR-based DE-test calls against ground truth and depend on the predicted mean *and* variance across cells in a perturbation; (ii) *mean-ranking* (*Spearman LFC*, *overlap@N*) rank genes by predicted mean log2 fold change and depend only on the predicted mean; and (iii) *Pearson* Δ, which correlates the full per-gene Δ vector and is sensitive to both. All numbers report mean $\pm$ SE across four seeds.

4.1. Discrete latents outperform continuous baselines

Existing single-cell perturbation models build on continuous latent representations (scVI, CPA, scLDM) or operate directly in expression space (scGPT, STATE). Discrete-latent representation learning, despite being the foundation of the strongest generative models in language and vision, has not previously been applied to perturbation modeling. We close this gap. The `ExpressionVAE` encodes each cell as a fixed-length sequence of discrete codes drawn from a vector-quantized bottleneck, and a perturbation-conditioned prior, either autoregressive (AR) or masked discrete diffusion (MDLM), models the joint distribution over those codes. We compare against six baselines spanning continuous latent VAEs, conditional perturbation models, expression-space foundation models, and a non-parametric mean baseline: scLDM (Palla et al., 2025), scVI (Lopez et al., 2018a), CPA (Lotfollahi et al., 2023), scGPT (Cui et al., 2024), STATE (Adduri et al., 2025), and PerturbMean. Results on Replogle and Parse 1M are reported in Table 1.

For each prior we report two rows, corresponding to the NB-sample and MSE output heads. These sit at the sampling and deterministic ends of the head spectrum analyzed in Section 4.2 and trade off across metric families; the practical takeaway here is that our model achieves the strongest result in *each* family. Three observations stand out.

Order-of-magnitude gains on distributional metrics. The headline result is on W_2, MMD^2, and FD, where the discrete-latent model improves over the strongest continuous-latent baseline (scLDM) by roughly an order of magnitude on MMD^2 and FD, on both datasets. On Replogle, MMD^2 drops from 0.200 to 0.010 and FD from 53.75 to 4.67; on Parse 1M, MMD^2 drops from 0.027 to 0.008 and FD from 18.14 to 3.11. Expression-space methods (scGPT, STATE) and continuous latent baselines (scVI, CPA) lag the discrete-latent model by one to three orders of magnitude on these metrics.

Table 1. **Discrete-latent perturbation modeling against six baselines on Replogle and Parse 1M.** For each prior (AR, MDLM) we report two rows, corresponding to the NB-sample and MSE output heads (Section 4.2). Bold = best in column for that dataset

Dataset	Prior	$W_2 \downarrow$	$\text{MMD}^2 \downarrow$	FD $\downarrow$	PDS $\uparrow$	PR-AUC $\uparrow$	Sp-sig $\uparrow$	Sp-LFC $\uparrow$	ovl@N $\uparrow$	P-Δ $\uparrow$
Replogle	AR	**7.969**	**0.010**	**4.67**	**0.803**	**0.298**	0.765	0.500	0.181	0.951
		8.989	0.098	34.40	0.751	0.233	**0.774**	0.584	**0.288**	**0.965**
	MDLM	8.105	0.013	5.20	0.779	0.292	0.747	0.508	0.188	0.950
		9.111	0.105	35.80	0.738	0.221	0.761	**0.585**	0.281	0.964
	STATE	20.580	0.730	366.64	0.641	0.210	0.721	0.357	0.212	0.420
	ScVI	17.359	0.453	284.47	0.517	0.190	0.598	0.257	0.085	0.350
	CPA	11.510	0.532	126.81	0.599	0.190	0.370	0.430	0.196	0.380
	ScGPT	34.166	3.087	1247.68	0.501	0.100	0.550	0.002	0.126	0.385
	ScLDM ($\omega = 1$)	11.292	0.200	53.75	0.730	–	–	0.413	0.163	–
	PerturbMean	–	–	–	0.616	0.190	0.190	0.246	0.162	–
Parse 1M	AR	**11.949**	**0.008**	**3.11**	0.920	0.493	**0.962**	0.793	0.400	**0.986**
		22.270	0.653	383.36	0.771	0.153	0.913	0.805	**0.480**	0.776
	MDLM	11.966	0.008	3.31	0.931	**0.495**	0.959	**0.812**	0.411	0.985
		22.314	0.660	384.68	0.770	0.150	0.928	0.786	0.463	0.776
	STATE	19.111	0.714	312.34	0.972	–	–	0.565	0.188	–
	ScVI	35.508	1.372	1233.11	0.519	–	–	0.010	0.044	–
	CPA	13.534	1.117	181.32	0.567	–	–	0.570	0.197	–
	ScGPT	22.870	2.203	523.93	0.520	–	–	0.057	0.077	–
	ScLDM ($\omega = 1$)	12.457	0.027	18.14	0.962	–	–	0.620	0.303	–
	PerturbMean	–	–	–	**0.973**	–	–	0.467	0.118	–

Two priors, the same numbers. AR and MDLM, drawn from different generative-model families and commonly viewed as alternatives, achieve effectively identical performance across all metrics on both datasets. This is direct evidence that the gain is from the discrete-latent representation rather than from a particular prior architecture, since the two priors share only the encoder and the codebook.

Top of both metric families. The model achieves the strongest result on every distributional and DE-discovery metric (sampling-rule row) *and* on every mean-ranking metric (deterministic-rule row), on both datasets. No baseline reaches the top of both families. Existing methods are competitive on distributional metrics or on ranking metrics, but never both: scLDM is the strongest baseline on distributional metrics yet trails on *Sp-LFC* and *overlap@N*, while PerturbMean and STATE are competitive on PDS and mean-ranking metrics but collapse on W_2, MMD^2, and FD. The two-rows-per-prior structure that produces this pattern is not incidental; Section 4.2 shows it reflects a single underlying axis governed by inference-time stochasticity.

4.2. Sampling-richness axis explains decoder-head behavior

A practical question for any single-cell perturbation model is which decoder head to use: cross-entropy, hurdle, MSE, negative binomial, or sampled negative binomial. The literature offers no consistent answer. Different papers report different metrics, and different metrics rank these heads differently. The two rows per prior in Table 1 make the conflict visible within a single backbone: the same trained model produces noticeably different metric profiles depending only on the head. We show here that this is not noise. The standard evaluation metrics partition into three groups, and within each group the head ordering is monotonic in a single underlying axis: the richness of the inference-time sampling distribution.

Five heads on a sampling-richness gradient. We evaluate five heads on the same trained `ExpressionVAE` backbone. `ce-quant` parameterizes a categorical over $[0, c_{\max}]$ per gene and samples from the softmax. `nb-sample` parameterizes a negative binomial $\text{NB}(\mu, \theta)$ per gene and emits a draw from it. `hurdle` parameterizes a Bernoulli zero-gate combined with a deterministic positive magnitude μ, sampling only the gate. `mse` emits a deterministic real-valued μ everywhere. `nb-deter` shares trained parameters with `nb-sample` and differs only in that `predict()` returns μ rather than a sample, which makes the prediction structurally anti-sparse since the softmax over genes leaks small mass to every gene. Read in the order `ce-quant` $\rightarrow$ `nb-sample` $\rightarrow$ `hurdle` $\rightarrow$ `mse` $\rightarrow$ `nb-deter`, the heads form a monotone gradient in the richness of the inference-time predictive distribution: full count sampling, parametric count sampling, sampled zero-gate over deterministic magnitude, fully deterministic, and fully deterministic plus structurally anti-sparse. Crucially, `nb-sample` and `nb-deter` share trained weights; the

only difference is the `predict()` rule.

Three groups, three trends. Figure 2 reports all five heads under both priors as per-metric heatmaps, with columns ordered to reflect the three-group partition. The structure is immediate from the color gradient. Group 1 (distribution) and Group 2 (DE-discovery) columns darken toward the sampling end of the head ordering, the top of each prior block; Group 3 (mean-ranking) columns darken toward the deterministic end, the bottom; *Pearson-*Δ flips winner between the two panels, with sampling heads winning on Parse 1M and deterministic heads winning on Replogle, exactly as a hybrid metric should behave when the dataset's noise structure changes. The Group 1 collapse is most dramatic on Parse 1M, where moving the head from sampling to deterministic shifts FD by more than two orders of magnitude (e.g., $1.84 \rightarrow 454.1$ along the AR block); on Replogle, the same trend holds in direction but is compressed in magnitude, consistent with that dataset's lower count dispersion.

The mechanism is straightforward. Group 1 metrics reward matching the full population cell cloud, including per-gene marginals and joint dispersion; Group 2 metrics reward DE-test calls that depend on both predicted mean and within-perturbation variance. Deterministic heads collapse the predicted cell cloud to a single point per condition, so within-perturbation variance is degenerate; DE tests applied to the predicted cells then operate on a vanishing σ and produce uninformative p-values, which propagates into PR-AUC, *disc_l1*, and *Sp-sig*. Group 3 metrics ask only about the per-gene predicted mean. Sampling injects between-cell noise that contaminates finite-sample mean estimates, so the noise that helps Groups 1 and 2 hurts Group 3.

The NB pair isolates the sampling axis. A natural confound is that `ce-quant` differs from `mse` on multiple axes simultaneously: discreteness, parametric form, output dimension, and training loss. To isolate sampling versus deterministic prediction from these confounds, we exploit that `nb-sample` and `nb-deter` share trained weights and differ only in `predict()`. Switching from μ to a draw from $\mathrm{NB}(\mu, \theta)$ moves NB from the bottom tier to the top tier on the distribution and DE-discovery groups (e.g., on Parse, $\mathrm{FD} : 454.06 \rightarrow 3.11$ and $\mathrm{PR\text{-}AUC} : 0.129 \rightarrow 0.493$ under AR), and reverses the direction on the mean-ranking group (e.g., on Replogle, $\mathrm{overlap@N} : 0.281 \rightarrow 0.181$ under AR). Because no other component of the model changes, the entire effect is localized to the sampling step at inference. This rules out parametric-form, capacity, discrete-vs-continuous, and training-objective confounds, and pins the cross-group structure on the sampling axis.

What metric measures what, and a recommendation. Group 1 and Group 2 metrics are both sampling-axis metrics: they reward recovery of within-perturbation cell variance, expressed once at the cell-distribution level (Group 1) and once filtered through DE testing (Group 2). Group 3 metrics are mean-axis metrics that reward clean per-gene mean estimates and penalize sampling noise. Pearson-Δ is a hybrid whose winner tracks dataset noise structure. The practical implication is that head choice is governed by what the practitioner intends to measure. We recommend reporting at least one metric from each group, and selecting the head to match the downstream task: a sampling head (`ce-quant` or sampled NB) for use cases that consume predicted cell distributions (synthetic-cohort DE, virtual screens, power calculations), and a deterministic-mean head (`mse` or NB-mean) for gene-list ranking against a reference. Reporting any single metric in isolation is misleading.

4.3. Usefulness of latent space for phenotypic reversion

Sections 4.1 and 4.2 establish that a discrete-latent perturbation model with the right decoder head dominates within-distribution metrics, and that head choice is governed by an inference-time sampling axis with a clean metric-to-head map. To test whether these design choices reflect a generalizable inductive bias rather than over-fitting the controlled metric set, we evaluate the *frozen* encoder on an out-of-distribution biological discovery task. No information about the OOD dataset enters training; the encoder is used only to embed cells, and metrics are computed directly on the resulting latents.

Benchmark. The Pfizer phenotypic-reversion benchmark (Wong et al., 2026) is a CRISPRi screen of 1,732 perturbations across 864,115 cells under IL-1β and TNF-α inflammation. The task is to rank perturbations capable of reverting the inflammatory state, with the canonical NF-κB activators (TNFRSF1A, TRADD, JUNB, JUND, NFKB1, NFKB2) as positive controls. We report three metrics on the frozen latents: per-perturbation Calinski-Harabasz scores among treated cells (Gene CH (T)) and untreated cells (Gene CH (U)), measuring latent separability; and Enrichment AUC, the rank-AUC of the NF-κB positive controls within the cosine-distance reversion ordering of the 1,732 perturbation centroids (random baseline 0.5).

Configurations. We evaluate the $2 \times 2 \times 4$ grid of training datasets (Parse 1M, Replogle), encoder bottlenecks (FSQ, Gaussian), and decoder heads (Quantile, Hurdle, MSE, NB), shown in Figure 3. The four heads correspond to the four head families from Section 4.2; the encoder bottleneck is the additional axis that the controlled ablation did not vary. Although the decoder is used only during training (the OOD evaluation reads the encoder output directly), its choice in-

	W2 ↓	MMD² ↓	FD ↓	Disc-L1 ↑	PR-AUC ↑	Spearman sig ↑	Spearman L2FC ↑	DEG Overlap ↑	Pearson Δ ↑
AR ce-quant	7.868 ±0.010	0.009 ±0.0003	4.640 ±0.080	0.787 ±0.0040	0.283 ±0.0010	0.769 ±0.0050	0.534 ±0.012	0.171 ±0.0030	0.950
AR nb-sample	7.969 ±0.012	0.010 ±0.0003	4.670 ±0.080	0.803 ±0.0020	0.298 ±0.0010	0.765 ±0.0010	0.500 ±0.015	0.181 ±0.0050	0.951
AR hurdle	7.899 ±0.010	0.013 ±0.0004	5.580 ±0.110	0.780 ±0.0020	0.294 ±0.0010	0.781 ±0.0030	0.529 ±0.0060	0.221 ±0.0010	0.953
AR mse	8.989 ±0.017	0.098 ±0.0007	34.40 ±0.240	0.751 ±0.0030	0.233 ±0.0020	0.774 ±0.0040	0.584 ±0.0090	0.288 ±0.0020	0.965
AR nb (deter)	11.94 ±0.013	0.303 ±0.0014	100.2 ±0.380	0.778 ±0.0010	0.223 ±0.0010	0.759 ±0.0050	0.558 ±0.0090	0.281 ±0.0020	0.957
mdlm ce-quant	8.045 ±0.020	0.015 ±0.0004	5.820 ±0.150	0.751 ±0.0030	0.275 ±0.0020	0.753 ±0.0050	0.501 ±0.0040	0.163 ±0.0020	0.949
mdlm nb-sample	8.105 ±0.0060	0.013 ±0.0001	5.200 ±0.060	0.779	0.292 ±0.0020	0.747	0.508 ±0.0040	0.188 ±0.0010	0.950
mdlm hurdle	8.009 ±0.0080	0.017 ±0.0005	6.460 ±0.070	0.755 ±0.0010	0.286 ±0.0010	0.754 ±0.0040	0.522 ±0.0090	0.217 ±0.0010	0.952
mdlm mse	9.111 ±0.021	0.105 ±0.0011	35.80 ±0.380	0.738 ±0.0040	0.221 ±0.0010	0.761 ±0.0070	0.585 ±0.0070	0.281 ±0.0010	0.964
mdlm nb (deter)	12.09 ±0.010	0.316 ±0.0008	102.4 ±0.200	0.751 ±0.0050	0.216 ±0.0010	0.733 ±0.0030	0.562 ±0.0080	0.276 ±0.0020	0.956

	W2 ↓	MMD² ↓	FD ↓	Disc-L1 ↑	PR-AUC ↑	Spearman sig ↑	Spearman L2FC ↑	DEG Overlap ↑	Pearson Δ ↑
AR ce-quant	11.97 ±0.011	0.004 ±0.0001	1.840 ±0.020	0.978 ±0.0030	0.497 ±0.0020	0.952 ±0.0070	0.741 ±0.015	0.409 ±0.0050	0.989
AR nb-sample	11.95 ±0.0080	0.007	3.110	0.920 ±0.010	0.493 ±0.0010	0.962 ±0.0050	0.793 ±0.0080	0.400 ±0.0030	0.986
AR hurdle	12.04 ±0.0070	0.011 ±0.0003	4.900 ±0.100	0.907 ±0.0020	0.433 ±0.0050	0.936 ±0.0030	0.766 ±0.012	0.302 ±0.0040	0.983
AR mse	22.27 ±0.108	0.653 ±0.0047	383.4* ±4.54	0.771 ±0.0040	0.153 ±0.0020	0.913 ±0.0070	0.805 ±0.0030	0.480 ±0.0050	0.776
AR nb (deter)	23.09 ±0.038	0.855 ±0.0017	454.1* ±1.74	0.754 ±0.0050	0.129 ±0.0020	0.785 ±0.0060	0.755 ±0.0040	0.462 ±0.0020	0.808
mdlm ce-quant	11.99 ±0.0040	0.005 ±0.0001	2.120 ±0.040	0.978 ±0.0020	0.500 ±0.0020	0.955 ±0.0070	0.766 ±0.0070	0.420 ±0.0050	0.989
mdlm nb-sample	11.97 ±0.0060	0.008 ±0.0002	3.310 ±0.010	0.931 ±0.0030	0.495 ±0.0040	0.959 ±0.0070	0.812 ±0.0030	0.411 ±0.0050	0.985
mdlm hurdle	12.07 ±0.010	0.013 ±0.0002	5.480 ±0.080	0.904 ±0.0040	0.423 ±0.0050	0.933 ±0.0050	0.749 ±0.021	0.302 ±0.0040	0.982
mdlm mse	22.31 ±0.073	0.660 ±0.0031	384.7* ±2.81	0.770 ±0.0030	0.150 ±0.0030	0.928 ±0.0060	0.786 ±0.0080	0.463 ±0.0050	0.776
mdlm nb (deter)	23.07 ±0.047	0.859 ±0.0023	453.1* ±2.09	0.755 ±0.0050	0.117 ±0.0010	0.759 ±0.012	0.754 ±0.0060	0.453 ±0.0020	0.809

Figure 2. **Output-head ablation on Replogle (top) and Parse 1M (bottom).** Rows are five output heads under AR and MDLM priors on the same trained backbone; columns are grouped by metric family. Cells shaded per column, darker indicates better; mean over four seeds, SE in ± notation.

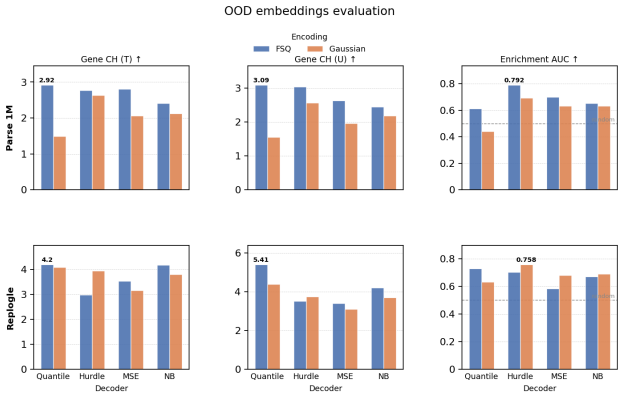

Figure 3. **Out-of-distribution evaluation on the Pfizer phenotypic-reversion benchmark.** Frozen `ExpressionVAE` encoders trained on Parse 1M (top) and Replogle (bottom), across encoder bottlenecks (FSQ, Gaussian) and decoder heads. Dashed line on the right shows the random baseline (0.5); annotated values mark the best configuration per panel.

fluences the frozen embedding by determining which signal the encoder must capture.

The decoder transfers; the encoder pairing is dataset-dependent. The Hurdle decoder identified by the controlled ablations is the strongest pairing on Enrichment AUC for both training datasets (0.792 on Parse 1M, 0.758 on Replogle); the decoder choice transfers cleanly without retuning. The optimal encoder pairing varies between datasets, however: FSQ + Hurdle wins on Parse 1M, while Gaussian + Hurdle wins on Replogle (FSQ + Hurdle reaches 0.704). Both axes contribute measurably to OOD performance, confirming that the decoder is not just a likelihood-modeling decision but shapes the representation itself.

Discrete encoders dominate separability. On the per-perturbation Calinski-Harabasz scores, FSQ + Quantile is the strongest configuration on both Gene CH (T) and Gene CH (U) on *both* datasets (2.92, 3.09 on Parse 1M; 4.20, 5.41 on Replogle), and FSQ encoders dominate Gaussian encoders across decoder pairings more generally. Hard quantization commits the encoder to a finite codebook of phenotypic states, producing sharper per-perturbation cluster structure in the latent space. This is consistent with the distributional-metric gap reported in Section 4.1: discrete latents allocate representational capacity in a way that preserves cell-state distinctions both within and outside the training distribution.

Comparison with scGPT. On the same benchmark, scGPT (Cui et al., 2024) achieves Enrichment AUC $= 0.81$. Our strongest Parse-trained configuration reaches 0.79, within typical evaluation noise of scGPT despite scGPT consuming roughly $10\times$ more pre-training data. The point is not a SOTA claim on the Pfizer benchmark, but a representation-efficiency claim: a discrete-latent perturbation prior trained on a single 1M-cell dataset effectively matches a foundation model trained on tens of millions of cells, on a biologically meaningful out-of-distribution ranking task. The explicit zero-gate decoder identified by our ablation transfers cleanly off the training distribution; hard quantization improves separability across both datasets and matches Gaussian on selectivity for Parse-trained encoders.

5. Conclusion

We introduced `ExpressionVAE`, the first discrete-latent perturbation model for single-cell data, and showed that pairing a vector-quantized bottleneck with either an autoregressive or a masked-diffusion prior achieves state-of-the-art across distributional and cell-eval metrics on Replogle and Parse 1M. Both priors achieve effectively identical performance, isolating the gain to the discrete latent rather than to a particular prior. A controlled output-head ablation further showed that decoder-head choice in this setting is governed by a single inference-time design axis, sampling versus deterministic prediction, with standard evaluation metrics partitioning into three groups along it; this resolves the apparent conflicts between published metric rankings into a clean three-way mapping between heads and downstream tasks. The hurdle decoder identified by the ablation transfers as the strongest selectivity pairing on a held-out CRISPRi reversion benchmark, where the frozen encoder effectively matches a $10\times$ larger foundation model at a fraction of the pretraining data. Taken together, these results suggest that a compact discrete latent captures perturbation-axis structure efficiently, and that further progress in virtual cell modeling may come more from improvements to the latent representation than from scaling generic foundation models directly on raw counts.

References

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

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

Austin, J., Johnson, D. D., Ho, J., Tarlow, D., and Van Den Berg, R. Structured denoising diffusion models in

discrete state-spaces. *Advances in neural information processing systems*, 34:17981–17993, 2021.

Bhattacharya, S., Gensbigler, C., Karim, S., and Lees, J. Discrete diffusion for single-cell gene expression modeling. *bioRxiv*, pp. 2026–02, 2026.

Chang, H., Zhang, H., Jiang, L., Liu, C., and Freeman, W. T. MaskGIT: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11315–11325, 2022.

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

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

Dong, M., Adduri, A., Gautam, D., Carpenter, C., Shah, R., Ricci-Tam, C., Kluger, Y., Burke, D. P., and Roohani, Y. H. Stack: In-context learning of single-cell biology. *bioRxiv*, pp. 2026–01, 2026.

Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K., Slichter, C. K., Miller, H. W., McElrath, M. J., Prlic, M., et al. Mast: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell rna sequencing data. *Genome biology*, 16(1):278, 2015.

Fitch, W. T. Information and the single cell. *Current Opinion in Neurobiology*, 71:150–157, 2021.

Gandhi, S., Javadi, F., Svensson, V., Khan, U., Jones, M. G., Yu, J., Merico, D., Goodarzi, H., and Alidoust, N. Tahoe-x1: Scaling perturbation-trained single-cell foundation models to 3 billion parameters. *bioRxiv*, pp. 2025–10, 2025.

Gulati, G. S., Sikandar, S. S., Wesche, D. J., Manjunath, A., Bharadwaj, A., Berger, M. J., Ilagan, F., Kuo, A. H., Hsieh, R. W., Cai, S., et al. Single-cell transcriptional diversity is a hallmark of developmental potential. *Science*, 367(6476):405–411, 2020.

Hao, M., Gong, J., Zeng, X., Liu, C., Guo, Y., Cheng, X., Wang, T., Ma, J., Zhang, X., and Song, L. Large-scale foundation model on single-cell transcriptomics. *Nature methods*, 21(8):1481–1491, 2024.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Rae, J. W., Vinyals, O., and Sifre, L. Training compute-optimal large language models. *ArXiv*, abs/2203.15556, 2022. URL https://api.semanticscholar.org/CorpusID:247778764.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637): 1123–1130, 2023.

Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M., and Le, M. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Liu, Y. et al. CellTok: Early-fusion multimodal large language model for single-cell transcriptomics via tokenization. *bioRxiv*, pp. 2025–10, 2025.

Lopez, R., Regier, J., Cole, M., Jordan, M., and Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15, 12 2018a. doi: 10.1038/s41592-018-0229-2.

Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., and Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058, 2018b.

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

Lotfollahi, M., Naghipourfar, M., Theis, F. J., and Wolf, F. A. Conditional out-of-distribution generation for unpaired data using transfer vae. *Bioinformatics*, 36 (Supplement_2):i610–i617, 2020.

Lotfollahi, M., Klimovskaia Susmelj, A., De Donno, C., Hetzel, L., Ji, Y., Ibarra, I. L., Srivatsan, S. R., Naghipourfar, M., Daza, R. M., Martin, B., et al. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular systems biology*, 19(6):e11517, 2023.

Lou, A., Meng, C., and Ermon, S. Discrete diffusion modeling by estimating the ratios of the data distribution. In *International Conference on Machine Learning*, 2023. URL https://api.semanticscholar.org/CorpusID:264451832.

Luo, E., Hao, M., Wei, L., and Zhang, X. scdiffusion: conditional generation of high-quality single-cell data using diffusion model. *Bioinformatics*, 40, 2024. URL https://api.semanticscholar.org/CorpusID:266844972.

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

Mentzer, F., Minnen, D. C., Agustsson, E., and Tschannen, M. Finite scalar quantization: Vq-vae made simple. *ArXiv*, abs/2309.15505, 2023. URL https://api.semanticscholar.org/CorpusID:263153393.

Oesinghaus, L., Becker, S., Vornholz, L., Papalexi, E., Pangallo, J., Moinfar, A. A., Liu, J., La Fleur, A., Shulman, M., Marrujo, S., et al. A single-cell cytokine dictionary of human peripheral blood. *bioRxiv*, 2025.

Ou, J., Nie, S., Xue, K., Zhu, F., Sun, J., Li, Z., and Li, C. Your absorbing discrete diffusion secretly models the conditional distributions of clean data. In *International Conference on Learning Representations (ICLR)*, 2025.

Palla, G., Babu, S., Dibaeinia, P., Pearce, J. D., Li, D., Khan, A. A., Karaletsos, T., and Tomczak, J. M. Scalable single-cell gene expression generation with latent diffusion models. *ArXiv*, abs/2511.02986, 2025. URL https://api.semanticscholar.org/CorpusID:282758604.

Palma, A., Richter, T., Zhang, H., Lubetzki, M., Tong, A., Dittadi, A., and Theis, F. J. Multi-modal and multi-attribute generation of single cells with cfgen. In *International Conference on Learning Representations*, 2024. URL https://api.semanticscholar.org/CorpusID:271218151.

Peebles, W. and Xie, S. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Razavi, A., van den Oord, A., and Vinyals, O. Generating diverse high-fidelity images with VQ-VAE-2. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Replogle, J. M., Saunders, R. A., Pogson, A. N., Hussmann, J. A., Lenail, A., Guna, A., Mascibroda, L., Wagner, E. J., Adelman, K., Lithwick-Yanai, G., et al. Mapping information-rich genotype-phenotype landscapes with genome-scale perturb-seq. *Cell*, 185(14):2559–2575, 2022.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. 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.

Sahoo, S. S., Arriola, M., Schiff, Y., Gokaslan, A., Marroquin, E., Chiu, J. T., Rush, A., and Kuleshov, V. Simple and effective masked diffusion language models. *ArXiv*, abs/2406.07524, 2024. URL https://api.semanticscholar.org/CorpusID:270380319.

Shi, J., Han, K., Wang, Z., Doucet, A., and Titsias, M. Simplified and generalized masked diffusion for discrete data. *Advances in neural information processing systems*, 37:103131–103167, 2024.

Theodoris, C. V., Xiao, L., Chopra, A., Chaffin, M. D., Al Sayed, Z. R., Hill, M. C., Mantineo, H., Brydon, E. M., Zeng, Z., Liu, X. S., et al. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, 2023.

van den Oord, A., Vinyals, O., and Kavukcuoglu, K. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Wang, C., Karimzadeh, M., Ravindra, N. G., Bounds, L. R., Alerasool, N., Huang, A. C., Ma, S., Gulbranson, D. R., Cui, H., Lee, Y., et al. X-cell: Scaling causal perturbation prediction across diverse cellular contexts via diffusion language models. *bioRxiv*, pp. 2026–03, 2026a.

Wang, M., Chen, C., Jiang, G., Ren, Z., Zhao, C., Shi, L., and Ma, Y. Scdiva: Masked discrete diffusion for joint modeling of single-cell identity and expression. *arXiv preprint arXiv:2602.03477*, 2026b.

Wong, D. R., Piper, M., Qiao, J., Russo, M., Jean, P., Clevert, D.-A., Arroyo, J. D., and Pashos, E. Phenotypic reversion and target prioritization for cellular inflammation via representation learning with foundation models. *bioRxiv*, pp. 2026–03, 2026.

Zeng, T. and Dai, H. Single-cell rna sequencing-based computational analysis to describe disease heterogeneity. *Frontiers in Genetics*, 10:629, 2019.

Zhang, H., Yuan, G.-H., Yuan, C., Xu, T., Bian, T., Cheng, H., Huang, W., Zhao, D., and Rong, Y. Lingshu-cell: A generative cellular world model for transcriptome modeling toward virtual cells. *arXiv preprint arXiv:2603.25240*, 2026.

A. Hyperparameters and training configuration

Tables 2 and 3 list the hyperparameters used for the main-paper VAE backbone and prior models. The same configuration is shared across Parse 1M and Replogle, with two dataset-specific differences: (i) the perturbation conditioning source — ESM2-3B protein-language-model embeddings keyed by the knocked-out gene on Replogle, and a learnable per-cytokine table on Parse 1M (cytokine ligands have no direct protein-sequence analog) — and (ii) the train/holdout split definitions, matched to scLDM in both cases. Both datasets use additive cell-type conditioning on the adaLN-Zero path.

Table 2. **Stage 1: `ExpressionVAE` hyperparameters.** Shared across both datasets and all four output heads. The Gaussian-latent variant referenced in Section 4.3 swaps the FSQ bottleneck for a Gaussian-posterior bottleneck with the parameters in the lower block; all other settings are unchanged.

Latent sequence length L	64
FSQ codebook size K	512
Encoder / decoder	Cross-attention transformer
Gene input embedding	MLP over $\log_1 p$ count + learned gene-identity tokens
Optimiser	AdamW
Learning rate	6×10^{-4}
Schedule	Cosine
Warm-up steps	2,500
Training iterations	75,000
Batch size	128
Gaussian-latent variant: per-token latent dim	8
Gaussian-latent variant: β_{KL}	2×10^{-6}

Table 3. **Stage 2: prior hyperparameters.** Shared across the autoregressive (AR) and masked-discrete-diffusion (MDLM) priors. The DiT backbone is identical (same depth, width, head count, parameter budget) so any gap in Section 4.1 reflects the prior family rather than capacity.

Backbone	DiT (Diffusion Transformer)
Conditioning	adaLN-Zero on perturbation + cell-type embedding
Model dimension d_{model}	384
Attention heads	8
Layers	6
Feed-forward dimension	1,536
Optimiser	AdamW
Learning rate	2×10^{-4}
Schedule	Cosine
Warm-up steps	3,000
Gradient clipping (max-norm)	1.0
Training iterations	75,000
Batch size	128
MDLM: masking schedule	absorbing, $\alpha_t = 1 - t$
MDLM: t_ϵ	5×10^{-3}
MDLM: inference denoising steps	128

B. Codebook size and latent-token-count ablation

We swept the FSQ latent sequence length $L \in \{64, 128\}$ against the codebook size $K \in \{512, 1024\}$ on the Replogle (HepG2) holdout, paired with the AR-FSQ-MSE and MDLM-FSQ-MSE configurations from the main paper. We additionally report AR-FSQ-NB, MDLM-FSQ-NB, and a Flow-Gaussian-MSE configuration as neighbouring anchor points in the design space. Results are in Table 4.

The numbers in Table 4 were obtained on a different gene-set than the main-paper results and are therefore not directly

comparable to Table 1; they are, however, internally consistent across rows. Based on this ablation we carried forward $L = 64$ and $K = 512$ to the main experiments.

Table 4. **Codebook size and latent-token-count ablation on the Replogle (HepG2) holdout.** Bold = best per column.

Run	PDS ↑	ovl@N ↑	PR-AUC ↑	Sp-sig ↑	W_2 ↓	MMD^2 ↓	FD ↓
AR-FSQ-MSE L=128, K=1024	0.7373	0.2800	0.2225	0.7519	**10.02**	**0.0706**	56.26
AR-FSQ-MSE L=128, K=512	0.7398	**0.2914**	**0.2301**	0.7499	10.17	0.0734	57.28
AR-FSQ-MSE L=64, K=1024	0.7474	0.2773	0.2196	0.7563	10.06	0.0720	56.70
AR-FSQ-MSE L=64, K=512	0.7510	0.2834	0.2177	**0.7856**	10.06	0.0716	57.19
AR-FSQ-NB	**0.7594**	0.2790	0.2224	0.7678	12.35	0.1514	116.30
MDLM-FSQ-MSE L=128, K=1024	0.7030	0.2750	0.2071	0.7265	10.18	0.0791	59.31
MDLM-FSQ-MSE L=128, K=512	0.7147	0.2790	0.2135	0.7440	10.25	0.0776	57.75
MDLM-FSQ-MSE L=64, K=1024	0.7126	0.2692	0.2064	0.7495	10.19	0.0771	57.80
MDLM-FSQ-MSE L=64, K=512	0.7169	0.2816	0.2161	0.7344	10.27	0.0777	60.17
MDLM-FSQ-NB	0.7259	0.2728	0.2112	0.7406	12.49	0.1580	118.57
Flow-Gaussian-MSE	0.7514	0.2820	0.2204	0.7290	10.17	0.0707	**55.10**

C. Additional results

Table 5. **Parse 1M**: full prior × head ablation on Distribution metrics. We add results from running 4-seed pooled mean ± SE runs.

Prior	Head	W_2 ↓	MMD^2 ↓	FD ↓
AR	ce-quantile	$11.975_{\pm 0.011}$	$\mathbf{0.0041}_{\pm 0.0001}$	$\mathbf{1.84}_{\pm 0.02}$
	hurdle	$12.039_{\pm 0.007}$	$0.0114_{\pm 0.0003}$	$4.90_{\pm 0.10}$
	mse	$22.270_{\pm 0.108}$	$0.6534_{\pm 0.0047}$	$383.36_{\pm 4.54}$
	nb	$23.088_{\pm 0.038}$	$0.8550_{\pm 0.0017}$	$454.06_{\pm 1.74}$
flow	ce-quantile	$12.218_{\pm 0.008}$	$0.0246_{\pm 0.0002}$	$10.90_{\pm 0.15}$
	hurdle	$11.918_{\pm 0.005}$	$0.0074_{\pm 0.0001}$	$2.95_{\pm 0.03}$
	mse	$\mathbf{11.890}_{\pm 0.005}$	$0.0108_{\pm 0.0003}$	$3.98_{\pm 0.10}$
	nb	$11.966_{\pm 0.010}$	$0.0127_{\pm 0.0002}$	$4.93_{\pm 0.07}$
MDLM	ce-quantile	$11.992_{\pm 0.004}$	$0.0049_{\pm 0.0001}$	$2.12_{\pm 0.04}$
	hurdle	$12.073_{\pm 0.010}$	$0.0128_{\pm 0.0002}$	$5.48_{\pm 0.08}$
	mse	$22.314_{\pm 0.073}$	$0.6601_{\pm 0.0031}$	$384.68_{\pm 2.81}$
	nb	$23.068_{\pm 0.047}$	$0.8585_{\pm 0.0023}$	$453.13_{\pm 2.09}$

Table 6. **Replogle (HepG2)**: full prior × head ablation on Distribution metrics.We add results from running 4-seed pooled mean ± SE runs.

Prior	Head	$W_2 \downarrow$	$MMD^2 \downarrow$	FD ↓
AR	ce-quantile	**7.868**$_{\pm 0.010}$	**0.0087**$_{\pm 0.0003}$	$4.64_{\pm 0.08}$
	hurdle	$7.899_{\pm 0.010}$	$0.0129_{\pm 0.0004}$	$5.58_{\pm 0.11}$
	mse	$8.989_{\pm 0.017}$	$0.0978_{\pm 0.0007}$	$34.40_{\pm 0.24}$
	nb	$11.939_{\pm 0.013}$	$0.3029_{\pm 0.0014}$	$100.25_{\pm 0.38}$
flow	ce-quantile	$8.050_{\pm 0.008}$	$0.0091_{\pm 0.0001}$	**4.53**$_{\pm 0.08}$
	hurdle	$7.908_{\pm 0.005}$	$0.0118_{\pm 0.0002}$	$5.17_{\pm 0.06}$
	mse	$8.973_{\pm 0.019}$	$0.0793_{\pm 0.0012}$	$27.45_{\pm 0.34}$
	nb	$10.884_{\pm 0.035}$	$0.2106_{\pm 0.0029}$	$68.15_{\pm 0.85}$
MDLM	ce-quantile	$8.045_{\pm 0.020}$	$0.0151_{\pm 0.0004}$	$5.82_{\pm 0.15}$
	hurdle	$8.009_{\pm 0.008}$	$0.0173_{\pm 0.0005}$	$6.46_{\pm 0.07}$
	mse	$9.111_{\pm 0.021}$	$0.1052_{\pm 0.0011}$	$35.80_{\pm 0.38}$
	nb	$12.086_{\pm 0.010}$	$0.3159_{\pm 0.0008}$	$102.37_{\pm 0.20}$

Table 7. **Parse 1M**: per-cell + per-gene moment fidelity. We add results from running 4-seed pooled mean ± SE runs.

Prior	Head	R^2 mean ↑	R^2 var ↑	PCC per-cell ↑	MSE per-cell ↓
AR	ce-quantile	**0.999**$_{\pm 0.000}$	**0.999**$_{\pm 0.000}$	$0.341_{\pm 0.000}$	$0.9972_{\pm 0.0005}$
	hurdle	$0.995_{\pm 0.000}$	$0.990_{\pm 0.000}$	$0.336_{\pm 0.000}$	$1.0320_{\pm 0.0001}$
	mse	$-0.308_{\pm 0.001}$	$0.472_{\pm 0.001}$	**0.460**$_{\pm 0.000}$	**0.9614**$_{\pm 0.0002}$
	nb	$-0.330_{\pm 0.000}$	$0.382_{\pm 0.000}$	$0.446_{\pm 0.000}$	$1.0674_{\pm 0.0003}$
flow	ce-quantile	$0.985_{\pm 0.000}$	$0.970_{\pm 0.000}$	$0.324_{\pm 0.000}$	$1.0476_{\pm 0.0002}$
	hurdle	$0.998_{\pm 0.000}$	$0.996_{\pm 0.000}$	$0.341_{\pm 0.000}$	$0.9991_{\pm 0.0004}$
	mse	$0.971_{\pm 0.000}$	$0.994_{\pm 0.000}$	$0.349_{\pm 0.000}$	$0.9810_{\pm 0.0003}$
	nb	$0.985_{\pm 0.000}$	$0.985_{\pm 0.000}$	$0.342_{\pm 0.000}$	$1.0001_{\pm 0.0004}$
MDLM	ce-quantile	$0.999_{\pm 0.000}$	$0.998_{\pm 0.000}$	$0.342_{\pm 0.000}$	$0.9971_{\pm 0.0005}$
	hurdle	$0.994_{\pm 0.000}$	$0.989_{\pm 0.000}$	$0.336_{\pm 0.000}$	$1.0341_{\pm 0.0004}$
	mse	$-0.315_{\pm 0.004}$	$0.469_{\pm 0.001}$	**0.460**$_{\pm 0.000}$	$0.9630_{\pm 0.0008}$
	nb	$-0.325_{\pm 0.003}$	$0.381_{\pm 0.001}$	$0.446_{\pm 0.000}$	$1.0674_{\pm 0.0012}$

Table 8. **Replogle (HepG2)**: per-cell + per-gene moment fidelity.We add results from running 4-seed pooled mean ± SE runs.

Prior	Head	R^2 mean ↑	R^2 var ↑	PCC per-cell ↑	MSE per-cell ↓
AR	ce-quantile	**1.000**$_{\pm 0.000}$	**0.998**$_{\pm 0.000}$	$0.601_{\pm 0.000}$	$0.8901_{\pm 0.0009}$
	hurdle	$0.999_{\pm 0.000}$	$0.909_{\pm 0.003}$	$0.615_{\pm 0.000}$	$0.8502_{\pm 0.0012}$
	mse	$0.982_{\pm 0.000}$	$-2.201_{\pm 0.008}$	$0.738_{\pm 0.000}$	$0.5274_{\pm 0.0005}$
	nb	$0.934_{\pm 0.000}$	$-2.369_{\pm 0.005}$	$0.738_{\pm 0.000}$	$0.5528_{\pm 0.0002}$
flow	ce-quantile	**1.000**$_{\pm 0.000}$	$0.996_{\pm 0.000}$	$0.603_{\pm 0.000}$	$0.8881_{\pm 0.0006}$
	hurdle	$0.999_{\pm 0.000}$	$0.933_{\pm 0.001}$	$0.615_{\pm 0.000}$	$0.8498_{\pm 0.0003}$
	mse	$0.984_{\pm 0.000}$	$-1.776_{\pm 0.003}$	$0.720_{\pm 0.000}$	$0.5615_{\pm 0.0005}$
	nb	$0.948_{\pm 0.001}$	$-1.511_{\pm 0.006}$	$0.697_{\pm 0.001}$	$0.6216_{\pm 0.0011}$
MDLM	ce-quantile	**1.000**$_{\pm 0.000}$	$0.993_{\pm 0.001}$	$0.603_{\pm 0.000}$	$0.8840_{\pm 0.0010}$
	hurdle	$0.998_{\pm 0.000}$	$0.878_{\pm 0.008}$	$0.617_{\pm 0.001}$	$0.8458_{\pm 0.0023}$
	mse	$0.981_{\pm 0.000}$	$-2.243_{\pm 0.009}$	$0.739_{\pm 0.000}$	**0.5248**$_{\pm 0.0002}$
	nb	$0.933_{\pm 0.000}$	$-2.428_{\pm 0.012}$	**0.740**$_{\pm 0.000}$	$0.5501_{\pm 0.0006}$

Table 9. **All prior × head ablations across cell-eval metrics**

Dataset	Prior	Head	disc_l1 ↑	ovl@N ↑	Sp-LFC ↑	PR-AUC ↑	MAE-l2 ↓	P-△ ↑	MAE-pb ↓	Sp-sig ↑
Parse 1M	AR	ce-quantile	**0.978**$_{\pm 0.003}$	0.409$_{\pm 0.005}$	0.741$_{\pm 0.015}$	0.497$_{\pm 0.002}$	0.259$_{\pm 0.001}$	0.989$_{\pm 0.000}$	**0.023**$_{\pm 0.000}$	0.952$_{\pm 0.007}$
		hurdle	0.907$_{\pm 0.002}$	0.302$_{\pm 0.004}$	0.766$_{\pm 0.012}$	0.433$_{\pm 0.005}$	0.300$_{\pm 0.002}$	0.983$_{\pm 0.000}$	0.030$_{\pm 0.000}$	0.936$_{\pm 0.003}$
		mse	0.771$_{\pm 0.004}$	0.480$_{\pm 0.005}$	0.805$_{\pm 0.003}$	0.153$_{\pm 0.002}$	0.282$_{\pm 0.001}$	0.776$_{\pm 0.000}$	0.324$_{\pm 0.000}$	0.913$_{\pm 0.007}$
		nb	0.754$_{\pm 0.005}$	0.462$_{\pm 0.002}$	0.755$_{\pm 0.004}$	0.129$_{\pm 0.002}$	0.299$_{\pm 0.001}$	0.808$_{\pm 0.000}$	0.338$_{\pm 0.000}$	0.785$_{\pm 0.006}$
	flow	ce-quantile	0.681$_{\pm 0.006}$	0.241$_{\pm 0.005}$	0.719$_{\pm 0.005}$	0.355$_{\pm 0.005}$	0.355$_{\pm 0.002}$	0.965$_{\pm 0.000}$	0.040$_{\pm 0.000}$	0.946$_{\pm 0.004}$
		hurdle	**0.977**$_{\pm 0.001}$	0.459$_{\pm 0.004}$	0.895$_{\pm 0.003}$	**0.552**$_{\pm 0.004}$	0.196$_{\pm 0.001}$	**0.991**$_{\pm 0.000}$	0.025$_{\pm 0.000}$	**0.962**$_{\pm 0.004}$
		mse	0.975$_{\pm 0.001}$	**0.513**$_{\pm 0.004}$	**0.903**$_{\pm 0.002}$	0.132$_{\pm 0.002}$	**0.183**$_{\pm 0.001}$	0.976$_{\pm 0.000}$	0.054$_{\pm 0.000}$	0.850$_{\pm 0.005}$
		nb	0.961$_{\pm 0.002}$	0.477$_{\pm 0.003}$	0.873$_{\pm 0.005}$	0.155$_{\pm 0.004}$	0.227$_{\pm 0.002}$	0.978$_{\pm 0.000}$	0.032$_{\pm 0.000}$	0.847$_{\pm 0.017}$
	MDLM	ce-quantile	**0.978**$_{\pm 0.002}$	0.420$_{\pm 0.005}$	0.766$_{\pm 0.007}$	0.500$_{\pm 0.002}$	0.254$_{\pm 0.001}$	0.989$_{\pm 0.000}$	**0.023**$_{\pm 0.000}$	**0.955**$_{\pm 0.007}$
		hurdle	0.904$_{\pm 0.004}$	0.302$_{\pm 0.004}$	0.749$_{\pm 0.021}$	0.423$_{\pm 0.005}$	0.306$_{\pm 0.003}$	0.982$_{\pm 0.000}$	0.031$_{\pm 0.000}$	0.933$_{\pm 0.005}$
		mse	0.770$_{\pm 0.003}$	0.463$_{\pm 0.005}$	0.786$_{\pm 0.008}$	0.150$_{\pm 0.003}$	0.289$_{\pm 0.002}$	0.776$_{\pm 0.000}$	0.325$_{\pm 0.000}$	0.928$_{\pm 0.006}$
		nb	0.755$_{\pm 0.005}$	0.453$_{\pm 0.002}$	0.754$_{\pm 0.006}$	0.117$_{\pm 0.001}$	0.304$_{\pm 0.001}$	0.809$_{\pm 0.000}$	0.338$_{\pm 0.001}$	0.759$_{\pm 0.012}$
Replogle	AR	ce-quantile	0.787$_{\pm 0.004}$	0.171$_{\pm 0.003}$	0.534$_{\pm 0.012}$	0.283$_{\pm 0.001}$	0.496$_{\pm 0.003}$	0.950$_{\pm 0.000}$	0.109$_{\pm 0.000}$	0.769$_{\pm 0.005}$
		hurdle	0.780$_{\pm 0.002}$	0.221$_{\pm 0.001}$	0.529$_{\pm 0.006}$	**0.294**$_{\pm 0.001}$	0.493$_{\pm 0.002}$	0.953$_{\pm 0.000}$	0.108$_{\pm 0.000}$	**0.781**$_{\pm 0.003}$
		mse	0.751$_{\pm 0.003}$	**0.288**$_{\pm 0.002}$	0.584$_{\pm 0.009}$	0.233$_{\pm 0.002}$	0.484$_{\pm 0.002}$	**0.965**$_{\pm 0.000}$	0.129$_{\pm 0.000}$	0.774$_{\pm 0.004}$
		nb	0.778$_{\pm 0.001}$	0.281$_{\pm 0.002}$	0.558$_{\pm 0.009}$	0.223$_{\pm 0.001}$	0.493$_{\pm 0.002}$	0.957$_{\pm 0.000}$	0.207$_{\pm 0.000}$	0.759$_{\pm 0.005}$
	flow	ce-quantile	0.784$_{\pm 0.001}$	0.170$_{\pm 0.005}$	0.498$_{\pm 0.007}$	0.281$_{\pm 0.002}$	0.497$_{\pm 0.001}$	0.951$_{\pm 0.000}$	0.109$_{\pm 0.000}$	0.752$_{\pm 0.007}$
		hurdle	0.786$_{\pm 0.001}$	0.210$_{\pm 0.001}$	0.519$_{\pm 0.005}$	0.293$_{\pm 0.001}$	0.489$_{\pm 0.002}$	0.953$_{\pm 0.000}$	**0.107**$_{\pm 0.000}$	0.756$_{\pm 0.005}$
		mse	0.765$_{\pm 0.003}$	0.278$_{\pm 0.001}$	0.570$_{\pm 0.007}$	0.221$_{\pm 0.001}$	**0.481**$_{\pm 0.004}$	0.963$_{\pm 0.000}$	0.125$_{\pm 0.001}$	0.755$_{\pm 0.001}$
		nb	**0.795**$_{\pm 0.001}$	0.268$_{\pm 0.004}$	0.546$_{\pm 0.009}$	0.206$_{\pm 0.000}$	**0.482**$_{\pm 0.002}$	0.944$_{\pm 0.000}$	0.189$_{\pm 0.001}$	0.751$_{\pm 0.008}$
	MDLM	ce-quantile	0.751$_{\pm 0.003}$	0.163$_{\pm 0.002}$	0.501$_{\pm 0.004}$	0.275$_{\pm 0.002}$	0.504$_{\pm 0.003}$	0.949$_{\pm 0.000}$	0.110$_{\pm 0.000}$	0.753$_{\pm 0.005}$
		hurdle	0.755$_{\pm 0.001}$	0.217$_{\pm 0.001}$	0.522$_{\pm 0.009}$	0.286$_{\pm 0.001}$	0.496$_{\pm 0.002}$	0.952$_{\pm 0.000}$	0.110$_{\pm 0.000}$	0.754$_{\pm 0.004}$
		mse	0.738$_{\pm 0.004}$	0.281$_{\pm 0.001}$	**0.585**$_{\pm 0.007}$	0.221$_{\pm 0.001}$	**0.480**$_{\pm 0.002}$	0.964$_{\pm 0.000}$	0.129$_{\pm 0.000}$	0.761$_{\pm 0.007}$
		nb	0.751$_{\pm 0.005}$	0.276$_{\pm 0.002}$	0.562$_{\pm 0.008}$	0.216$_{\pm 0.001}$	0.494$_{\pm 0.003}$	0.956$_{\pm 0.000}$	0.209$_{\pm 0.001}$	0.733$_{\pm 0.003}$