# BEYOND MEAN SHIFTS: PREDICTING DISTRIBUTIONAL RESPONSES TO UNSEEN GENETIC PERTURBATIONS

**Kalyan Ramakrishnan,*[*]   Jonathan G. Hedley,*[*]   Sisi Qu,**
**Puneet K. Dokania ,   Philip H. S. Torr**
University of Oxford
jonathan.hedley@eng.ox.ac.uk

**Cesar A. Prada-Medina,   Julien Fauqueur,   Kaspar Märtens**
Novo Nordisk
KQTM@novonordisk.com

## ABSTRACT

We introduce a simple, histogram-based approach for predicting distributional responses in gene expression following genetic perturbations. This is an essential task in early-stage drug discovery, where such responses can offer insights into gene function and inform target identification. Existing methods only optimize for changes in the mean expression, overlooking stochasticity inherent in single-cell data. We instead model per-gene expression distributions, predicting histograms conditioned on perturbations. This captures higher-order statistics (variance, skewness, kurtosis), where our method outperforms baselines at a fraction of the training cost. To generalize to unseen perturbations, we incorporate prior knowledge via gene embeddings from large language models (LLMs). While modeling a richer output space, the method remains competitive in predicting mean expression changes. This work demonstrates that explicitly modeling distributional responses yields richer biological insights while remaining practical and efficient.

## 1 INTRODUCTION

Predicting transcriptomic responses to genetic perturbations is a core problem in functional genomics and early-stage drug discovery. Given a perturbed gene, the goal is to predict how expression changes across the genome. Because single-cell expression is inherently stochastic, reflecting transcriptional noise, regulatory variability, and measurement uncertainty (Paulsson, 2005; Raj & van Oudenaarden, 2008), a perturbation induces not only mean shifts but also changes in distributional shape across cells. These distributional effects are often mechanistically meaningful: variance changes can reflect altered regulation or cell-cycle desynchronization, and multimodality can indicate state transitions or emerging subpopulations. Recent Perturb-seq assays (Dixit et al., 2016; Norman et al., 2019; Replogle et al., 2022) make this explicit by profiling thousands of single cells per perturbation, revealing changes in variance, skewness, zero inflation, and modality alongside mean shifts (Fig. 1). However, exhaustive experimental profiling across perturbations and contexts is infeasible, motivating predictive models that generalize to unseen perturbations.

Most existing perturbation predictors either model only mean responses or rely on restrictive output likelihoods. Latent-variable generative models such as CPA (Lotfollahi et al., 2023) and sVAE+(Lopez et al., 2023) use probabilistic decoders but typically assume fixed parametric forms that struggle to capture the diversity of single-cell expression patterns (de Torrenté et al., 2020). Methods such as STATE (Adduri et al., 2025) target population-level set prediction, while graph- and ontology-based approaches like GEARS (Roohani et al., 2024) incorporate biological structure to improve generalization. Large foundation models (e.g., scGPT (Cui et al., 2024), scFoundation (Hao

[*]Equal Contribution

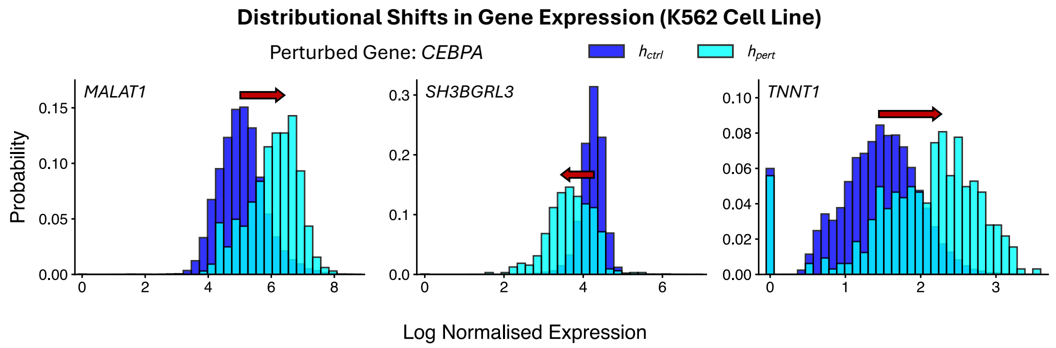

Figure 1: Gene expression distributional shifts following a CEBPA perturbation in the K562 cell line for three example genes. In each case, the control (dark blue) and perturbed (cyan) histograms differ in the mean *and* overall shape, including variance, skewness, zero inflation, and modality.

et al., 2024)) learn broad representations from observational atlases, but commonly focus on mean shifts; moreover, strong evaluations show they may not outperform simple linear baselines on unseen perturbations (Ahlmann-Eltze et al., 2024). A complementary route is to encode prior knowledge directly in the perturbation representation itself: LLM-informed gene embeddings can capture functional similarity (Rives et al., 2021; Theodoris et al., 2023; Chen & Zou, 2024), improving generalization in mean prediction (Märtens et al., 2024), and have been used by scLAMBDA (Wang et al., 2024a) to condition a generative model for heterogeneous responses.

Although generative models are often motivated as learning the joint response across genes, this claim is rarely validated in an interpretable way. Any joint model implies per-gene marginals; if marginal distributions are inaccurate, the implied joint response is unreliable. Yet existing evaluations typically emphasize population-level or mean-based criteria without directly assessing per-gene marginal shape.

We address both challenges: predicting per-gene expression distributions while generalizing to unseen perturbations. Our approach combines LLM-informed gene embeddings with a non-parametric histogram output that can represent rich distributional changes with a small number of bins. Across Perturb-seq benchmarks (Norman et al., 2019; Replogle et al., 2022), we show improved distributional accuracy over parametric baselines, competitive mean prediction, and substantially faster training. Beyond metrics, the model surfaces distributional signatures such as variance inflation and bimodality shifts, enabling mechanistic insight that mean-only predictors miss.

## 2 Related Work

**Generative models for perturbation effects.** Generative models are widely used to simulate transcriptional responses in single cells. scVI (Lopez et al., 2018) models gene expression counts with a negative binomial VAE (Kingma et al., 2013) to capture measurement noise and variability across cells. scGen (Lotfollahi et al., 2019) and CPA (Lotfollahi et al., 2023) learn latent-space shifts that encode perturbation effects, enabling interpolation across conditions. sVAE+ (Lopez et al., 2023) and SAMS-VAE (Bereket & Karaletsos, 2023) separate baseline and perturbation components to improve interpretability, while CellOT (Bunne et al., 2023) aligns control and perturbed populations via neural optimal transport. STATE (Adduri et al., 2025) models perturbations as population-level shifts by matching perturbed cell distributions. However, all remain limited to perturbations seen during training in their current implementation.

**Generalization to unseen perturbations.** Generalizing to unseen perturbations is critical for scalable profiling. GEARS (Roohani et al., 2024) addresses this by using gene ontology and co-expression graphs. Foundation models such as scGPT (Cui et al., 2024) and scFoundation (Hao et al., 2024) learn gene and cell embeddings from large-scale data using transformers. Despite their scale, these models predict only mean responses. Moreover, a recent benchmark (Ahlmann-Eltze et al., 2024) shows they struggle to outperform simple baselines on unseen perturbations.

**Gene embeddings from language models.** Pretrained language models have recently been adopted to encode biological priors. GenePT (Chen & Zou, 2024) generates embeddings from NCBI gene text descriptions, and shows that these embeddings encode underlying biology, e.g., gene function. Märtens et al. (2024) incorporate similar embeddings into perturbation models, demonstrating state-of-the-art generalization to unseen targets, but focusing on mean shifts. scLAMBDA Wang et al. (2024a) similarly conditions a generative VAE on LLM-derived gene embeddings and aims to capture heterogeneous single-cell responses.

# 3 METHOD

## 3.1 OVERVIEW

We aim to model responses to genetic perturbations by learning a conditional distribution over gene expression values given a perturbation $p \in \{0,1\}^P$, where $P$ is the number of perturbable genes. Let $\mathbf{x} \in \mathbb{R}^G_{\geq 0}$ be a random vector representing post-perturbation log-normalized expression across $G$ genes. The task is to approximate the conditional distribution $q(\mathbf{x}|p)$. In principle, this is a joint distribution over all genes, residing in a high-dimensional space that is challenging to model directly (partly due to the sparsity and noise inherent in single-cell data). We therefore focus on predicting gene-wise marginal distributions $q(x_g|p)$ for each gene $g$. While this ignores gene-gene correlations, it allows tractable distribution modeling and captures per-gene variability across cells.

## 3.2 HISTOGRAM CONSTRUCTION

To represent marginal expression distributions, we discretize each gene's expression range into a constant number of fixed-width bins and model the result as a histogram. Thus, we approximate $q(x_g|p) \approx h(x_g|p)$, where $h$ is a piecewise-uniform density over the expression range.

For each gene $g$, expression values in the training data, $x_g^{(\text{train})}$, (i.e. cells used to fit the model, aggregated across training perturbations) are binned into $B$ intervals of width $w_g$. All bin ranges and edges are computed from training cells only and then held fixed: held-out data are discretized using the same bin edges to prevent data leakage. Expression ranges are symmetrically extended by a constant $\epsilon_g$ to avoid truncation:

$$r_g = \left[\min(x_g^{(\text{train})}) - \epsilon_g, \max(x_g^{(\text{train})}) + \epsilon_g\right], \tag{1}$$

where $\epsilon_g = [\max(x_g^{(\text{train})}) - \min(x_g^{(\text{train})})]/2(B-1)$. This choice of $\epsilon_g$ ensures the first bin is centered exactly at the minimum, allowing the model to assign mass exactly at zero to better capture the sparsity in single-cell data. For genes with a constant expression (e.g., all zeros), we assign a small positive upper bound to avoid degenerate binning and ensure non-negative support. Histograms are then computed by assigning a probability mass to each bin based on the fraction of cells falling within it, yielding a valid discrete distribution.

## 3.3 LLM-INFORMED PERTURBATION ENCODING

We map each perturbation $p$ to a continuous vector $\mathbf{e}_p \in \mathbb{R}^d$ using a fixed function, where $d$ is the embedding dimension. Following the success of LLM-informed perturbation models Märtens et al. (2024), we construct this gene representation by combining two complementary sources of biological information: (i) text embeddings of NCBI **gene descriptions** using text-embedding-ada-002, as proposed by Chen & Zou (2024) in GenePT, and (ii) **protein-sequence** embeddings from the ProtT5 protein language model (Elnaggar et al., 2022). We concatenate these embeddings to obtain a single per-gene vector, which is kept frozen during training and serves as a biologically informed input that supports generalisation to unseen perturbations.

## 3.4 MODEL ARCHITECTURE

Given a perturbation, we model the induced *distributional shift* in gene expression relative to an empirically observed control (unperturbed) distribution. This reflects the biological intuition that most perturbations modify, rather than overwrite, a cell's baseline expression profile. Let $h_{\text{ctrl}} \in \mathbb{R}^{G \times B}$

denote the control histogram, where each row (summing to 1) contains the baseline distribution of a gene over the $B$ bins. For a perturbation $p$, we apply shifts $\Delta \in \mathbb{R}^{G\times B}$ in log space as follows:

$$\hat{h}_\theta(p) = \mathrm{softmax}\left(\log h_{\mathrm{ctrl}} + \Delta_\theta(\mathbf{e}_p)\right), \tag{2}$$

where $\Delta_\theta : \mathbb{R}^d \to \mathbb{R}^{G\times B}$ is a multilayer perceptron (MLP) with learnable parameters $\theta$ and output initialized close to zero (He et al., 2015). The softmax is applied over each row. Note that $\Delta = 0$ recovers the control histogram.

### 3.5 Training Objective

We train the model by minimizing a loss that compares the predicted ($\hat{h}$) and target ($h$) distributions using two components: (i) the Wasserstein-1 (W1) distance for distributional alignment and (ii) mean squared error (MSE) to match mean expressions. While cross-entropy is used for histogram prediction in other contexts (Imani et al., 2024), it treats bins independently and ignores the geometry of the support. In contrast, the W1 distance accounts for mass transport across bins and better reflects the continuity of expression data.

**W1 term.** For general 1D distributions, the W1 distance is given by $\mathrm{W1} = \int_{-\infty}^{\infty} |\hat{F}(x) - F(x)|dx$, where $\hat{F}$ and $F$ are the predicted and true cumulative distribution functions (CDFs). For histogram distributions, we can compute this integral in closed form, as described below.

Let $I_{gb} = [x_{gb}^{(l)}, x_{gb}^{(r)}]$ be the $b$-th bin interval for gene $g$ (with width $w_g$), to which the predicted and true histograms assign masses $\hat{h}_{gb}$ and $h_{gb}$ given a perturbation. The CDF gaps at the left and right bin edges are

$$C_{gb}^{-} = \hat{F}_g(x_{gb}^{(l)}) - F_g(x_{gb}^{(l)}), \quad C_{gb}^{+} = \hat{F}_g(x_{gb}^{(r)}) - F_g(x_{gb}^{(r)}). \tag{3}$$

Assuming the densities are uniform inside $I_{gb}$, the gap varies linearly as follows (for $x \in I_{gb}$):

$$\hat{F}_g(x) - F_g(x) = C_{gb}^{-} + \frac{\delta_{gb}}{w_g}\big(x - x_{gb}^{(l)}\big), \tag{4}$$

where $\delta_{gb} = C_{gb}^{+} - C_{gb}^{-} = \hat{h}_{gb} - h_{gb}$ represents the overall change. Integrating Eq. (4) over $I_{gb}$ gives its exact contribution to the W1 distance:

$$\mathrm{W1}_{gb} = \int_{I_{gb}} |\hat{F}_g(x) - F_g(x)|dx = \begin{cases} \dfrac{w_g}{2}\left(|C_{gb}^{-}| + |C_{gb}^{+}|\right) & \text{if} \quad C_{gb}^{-}C_{gb}^{+} \geq 0 \\ \dfrac{w_g}{2}\dfrac{|C_{gb}^{-}|^2 + |C_{gb}^{+}|^2}{|C_{gb}^{-}| + |C_{gb}^{+}|} & \text{if} \quad C_{gb}^{-}C_{gb}^{+} < 0. \end{cases} \tag{5}$$

The W1 loss is thus obtained by summing Eq. 5 over the $B$ bins and averaging over all genes:

$$\mathcal{L}_{\mathrm{wass}} = \frac{1}{G}\sum_{g=1}^{G}\sum_{b=1}^{B} \mathrm{W1}_{gb}. \tag{6}$$

**MSE term.** This term resolves cases where several histograms share the same W1 distance from the control distribution and is defined as

$$\mathcal{L}_{\mathrm{mse}} = \frac{1}{G}\sum_{g=1}^{G}(\hat{\mu}_g - \mu_g)^2 \tag{7}$$

where $\hat{\mu}_g = \sum_{b=1}^{B} \frac{1}{2}(x_{gb}^{(l)} + x_{gb}^{(r)})\, \hat{h}_{gb}$ is the predicted mean and $\mu_g$ is the target mean expression of gene $g$ (for the given perturbation).

**Total loss.** We average the two components over perturbations and combine them with fixed weights:

$$\mathcal{L}_{\mathrm{total}} = \lambda_{\mathrm{wass}}\mathcal{L}_{\mathrm{wass}} + \lambda_{\mathrm{mse}}\mathcal{L}_{\mathrm{mse}}, \tag{8}$$

where $\lambda_{\mathrm{wass}} + \lambda_{\mathrm{mse}} = 1$. The loss $\mathcal{L}_{\mathrm{total}}$ is minimized over model parameters $\theta$.

We provide Python code for our training procedure on GitHub[1].

[1] https://github.com/Kalyan0821/LLMHistPert

# 4 Experiments

## 4.1 Experimental Details

We evaluate our method on three single-cell Perturb-seq datasets across two cell lines. These include 102 single-gene perturbations in the K562 cell line from Norman et al. (2019), as well as 1076 (K562) and 1517 (RPE1) perturbations from Replogle et al. (2022). Gene expression values are log-normalized. We train and evaluate all models on single-gene perturbations. We use 9-fold cross-validation to evaluate generalization to unseen perturbations, given the limited dataset size. Training is done on a single NVIDIA A40 GPU (45 GB) and takes 10-15 minutes per fold, depending on histogram resolution. We use 15 bins by default.

Histogram models are trained with a weighted sum of the W1 and MSE losses. A coarse grid sweep identified the best configuration as $\lambda_{\text{wass}} = 0.75$ and $\lambda_{\text{mse}} = 0.25$ (see Appendix C for details). We vary the MLP size with dataset scale to reduce overfitting: hidden dimensions are $(1024, 512, 1024)$ for the Replogle data and $(128, 64)$ for the smaller Norman dataset. Hidden layers use Layer Normalization (Ba et al., 2016) followed by a Leaky ReLU activation with negative slope $0.01$. The output layer is linear. Models are trained using the AdamW optimizer (Loshchilov & Hutter, 2017) with learning rate $10^{-3}$, weight decay $10^{-3}$, dropout $0.2$, and batch size 32, for $500$ epochs per fold. Resource usage details for training are provided in Appendix A.

## 4.2 Distributional Baselines & Evaluation

**Baselines.** To assess how well our model captures post-perturbation gene expression distributions, we compare against several baselines spanning no-effect, pooled, and perturbation-conditioned parameteric models.

First, *Ctrl Hist* predicts the control distribution $h_{\text{ctrl}}$ for all perturbations. This represents a naive baseline that assumes no response to intervention. Second, *Non-Ctrl Hist* pools all perturbed cells in the training set and fits common per-gene histograms to the resulting data. This captures the aggregate expression distribution across perturbations. Though this may fare well on average likelihood metrics, it cannot model perturbation-specific effects.

Third, to create a perturbation-aware baseline, we train an MLP ($\mathbb{R}^d \to \mathbb{R}^{G\times 2}$) to predict the parameters of a truncated Gaussian (*TG*) for each gene conditioned on the perturbation. This can capture coarse changes in the mean and variance, with truncation enforcing non-negative expressions and enabling fair comparison with the bounded supports used by the histogram model (Eq. 1). The TG density has the following general form on its support $[a, b]$:

$$f_{\text{TG}}(x; \mu, \sigma) = \frac{\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)} \tag{9}$$

where $\Phi(x) = \frac{1}{2}\left(1 + \operatorname{erf}(x/\sqrt{2})\right)$, and parameters $(\mu, \sigma)$ describe the underlying Gaussian truncated to $[a, b]$. In our case, given perturbation $p$, the model outputs $G$ pairs $(\mu_{g,\theta}(p),\ \log \sigma_{g,\theta}(p))$ corresponding to supports $r_g = [a_g, b_g]$. We also constrain $\mu_{g,\theta}$ to $[a_g - 1.5\sigma_U, b_g + 1.5\sigma_U]$, where $\sigma_U = (b_g - a_g)/\sqrt{12}$ is the standard deviation of a uniform distribution over $r_g$. This prevents the model from pushing $\mu_{g,\theta}$ far outside the observed data range, which may cause vanishing likelihoods and unstable training.

Finally, to account for the prevalence of zero-valued observations in single-cell data, we define a zero-inflated truncated Gaussian (*ZITG*) density with the following general form:

$$f_{\text{ZITG}}(x; \mu, \sigma, \pi) = \pi\, \mathcal{U}_{[-\epsilon,\epsilon]}(x) + (1 - \pi)\, f_{\text{TG}}(x; \mu, \sigma), \tag{10}$$

where the extra parameter $\pi \in [0, 1]$ represents the zero-inflation probability and $\mathcal{U}_{[-\epsilon,\epsilon]}$ is a narrow uniform density of width $2\epsilon$ centred at the origin. In our case, given perturbation $p$, the MLP ($\mathbb{R}^d \to \mathbb{R}^{G\times 2} \times [0, 1]^G$) produces $G$ triplets $(\mu_{g,\theta}(p),\ \log \sigma_{g,\theta}(p),\ \pi_{g,\theta}(p))$ corresponding to the supports $r_g$ and half-widths $\epsilon_g$ defined in Eq. 1.

| Cell line | Method | NLL (↓) | W1 (↓) | RMAE (↓) | | | | Train time (↓) |
|---|---|---|---|---|---|---|---|---|
| | | | | Mean | Var | Skew | Kurt | (mins/fold) |
| K562 | Ctrl Hist | −0.379 | 0.066 | 1.000 | 0.184 | 0.237 | 0.572 | N/A |
| | Non-Ctrl Hist | −**0.393** | 0.063 | 0.980 | 0.213 | 0.235 | 0.590 | N/A |
| | TG (LLM) | +0.016 | 0.139 | 0.931 | 0.250 | 0.472 | 0.940 | 20.49 |
| | ZITG (LLM) | −0.381 | 0.097 | 0.891 | 0.181 | 0.240 | 0.597 | 24.52 |
| | scLAMBDA* | +2.728 | 0.267 | **0.865** | 0.939 | 0.769 | 1.151 | 13.42 |
| | MLP+Hist (LLM) | −0.389 | **0.056** | 0.876 | **0.174** | **0.224** | **0.537** | **9.89** |
| RPE1 | Ctrl Hist | −0.437 | 0.091 | 1.000 | 0.269 | 0.308 | 0.685 | N/A |
| | Non-Ctrl Hist | −**0.459** | 0.078 | 0.895 | 0.266 | 0.284 | 0.665 | N/A |
| | TG (LLM) | −0.068 | 0.148 | 0.897 | 0.303 | 0.461 | 0.899 | 41.16 |
| | ZITG (LLM) | −0.455 | 0.114 | 0.869 | 0.243 | 0.282 | 0.657 | 49.49 |
| | scLAMBDA* | +2.077 | 0.235 | **0.855** | 0.882 | 0.729 | 1.205 | 17.60 |
| | MLP+Hist (LLM) | −0.441 | **0.072** | 0.863 | **0.239** | **0.276** | **0.633** | **14.93** |

Table 1: Distributional model performance on unseen perturbations from the two Replogle cell lines. For each method, we report the gene-averaged NLL and RMAEs of post-perturbation statistics. *For scLAMBDA, NLL is approximated by binning generated samples with the same bin edges as *MLP+Hist* and using Eq. (12); W1 is computed directly from generated samples via a quantile-grid approximation of the 1D-Wasserstein distance; all moments are computed directly from samples.

**Baseline training.** For the TG and ZITG baselines, we minimize the average negative log-likelihood (NLL) of the training data. Given a perturbation, the NLL of gene $g$'s expressions is

$$\mathrm{NLL}_g = -\frac{1}{N_g}\sum_{n=1}^{N_g} \log f(x_{gn};\theta), \tag{11}$$

where $N_g$ is the number of post-perturbation samples of gene $g$. The average NLL is then obtained by averaging this over all genes and perturbations in the dataset.

**Evaluating distributions.** For each method, we report the average NLL over unseen perturbations to test if the predicted distributions fit the observed expression values. Note that for the histogram model, NLLs can be obtained efficiently without evaluating on the samples $\{x_{gn}\}$ each time as follows:

$$\mathrm{NLL}_g^{\mathrm{hist}} = -\sum_{b=1}^{B} h_{gb} \log\left(\frac{\hat{h}_{gb}}{w_g}\right). \tag{12}$$

We also test the ability to represent perturbation-specific structure by reporting errors in *statistics* of the predicted distributions compared to the observed data. Specifically, we compute the W1 distance and the relative mean absolute errors (RMAE) of the predicted mean, variance, skewness, and excess kurtosis. Statistics are obtained from the parameters *of TG / ZITG* and the bin probabilities of the histogram model (see Appendix B for details). Given an unseen perturbation, the RMAE of statistic $\tau$ is given by

$$\mathrm{RMAE}(\hat{\tau}, \tau) = \frac{\sum_{g=1}^{G} |\hat{\tau}_g - \tau_g|}{\sum_{g=1}^{G} |\tau_g|}, \tag{13}$$

where $\hat{\tau}_g$ and $\tau_g$ denote its predicted and observed value for gene $g$. We average this error over all perturbations in the test set, serving as a "shape-aware" distributional metric.

### 4.3 Quantitative Results

Table 1 compares our histogram model (*MLP+Hist*) to the distributional baselines on unseen perturbations from the two Replogle cell lines. On both, *MLP+Hist* achieves the lowest errors on distributional metrics overall, with particularly strong performance on distributional shape (W1) and higher-order moments. Relative to *ZITG*, the gains increase with moment order on K562, consistent with the histogram representation capturing perturbation-specific changes in distributional shape entirely from the expression data, that are difficult to represent with a fixed parametric form.

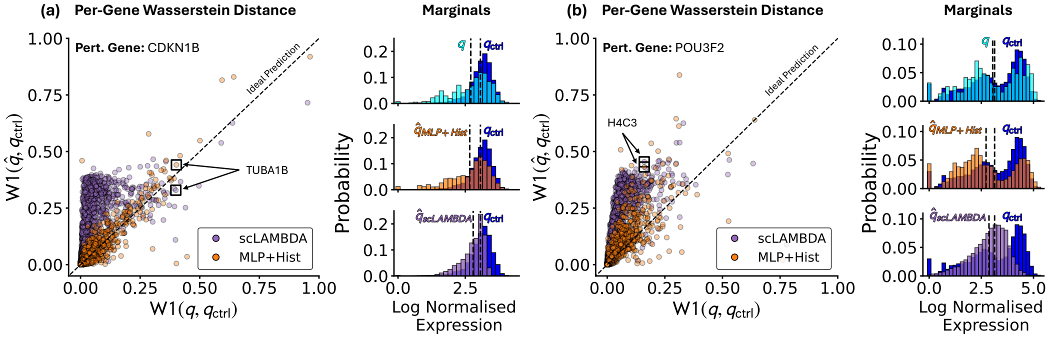

Figure 2: Distributional shift plots (left) and highlighted gene expression distributions (right) (using 35 bins) following two example unseen perturbations, (a) CDKN1B and (b) POU3F2 from the Norman data for MLP+Hist and scLAMBDA. Each circle represents a gene $g$ with coordinates $x = \mathrm{W1}(q_g, q_g^{\mathrm{ctrl}})$ and $y = \mathrm{W1}(\hat{q}_g, q_g^{\mathrm{ctrl}})$), measuring the observed and predicted shifts from the control distribution (dark blue), respectively. We show histograms for two genes that exhibit non-mean responses: (i) TUBA1B, showing increased variance, and (ii) H4C3, showing a multimodal response. Here, post-perturbation cell samples (cyan), predicted histograms by MLP+Hist (orange), and predicted distributions by scLAMBDA (purple) are compared to control cell samples (blue). Vertical dashed lines mark the means.

Performance improves with histogram resolution and saturates quickly; we use 15 bins throughout, and report the full resolution sweeps in Appendix D. As expected, *ZITG* outperforms *TG* across metrics due to the added flexibility of the zero-inflation component. *MLP+Hist* also trains substantially faster than parametric baselines, with additional resource analysis in Appendix A.

As a recent distributional baseline, *scLAMBDA* (Wang et al., 2024a) provides a useful point of comparison. Although it is competitive on mean estimation, achieving the lowest mean RMAE on both K562 and RPE1, it performs poorly on distributional metrics, with substantially worse W1 and higher-moment errors than MLP+Hist and markedly higher (approximated) NLL. This suggests that *scLAMBDA* can recover average effects but does not reliably capture distributional structure beyond the mean, casting doubt on its ability to approximate the joint distribution.

Interestingly, *Non-Ctrl Hist* attains the best NLL on both cell lines. This can be attributed to the pooled (global) histogram assigning a substantial mass to the zero-expression bin (∼61% in K562 and 64% in RPE1), which accounts for the majority of single-cell measurement data. However, since it predicts the same histogram for every perturbation, it cannot represent perturbation-specific shifts and performs significantly worse on the shape-aware metrics (Table 1).

### 4.4 Qualitative Results

Beyond aggregate metrics, we examine whether predicted distributions capture biologically interpretable effects. We plot W1 distances between the control and predicted distributions $\mathrm{W1}(\hat{q}_g, q_g^{\mathrm{ctrl}})$ vs. those between the control and observed distributions $\mathrm{W1}(q_g, q_g^{\mathrm{ctrl}})$ for all genes $g$ given an unseen perturbation for both MLP+Hist (orange) and scLAMBDA (purple). Fig. 2 (left) shows this plot for two example perturbations (CDKN1B[2] and POU3F2[3]), where each circle represents a gene. The plot provides a global view of distributional shift accuracy, helping identify genes where predictions succeed or fail (ideal predictions lie along the diagonal). We examine two genes (TUBA1B[4] and H4C3[5]) exhibiting distinct non-mean effects recovered effectively by our model.

TUBA1B encodes a tubulin alpha chain involved in building the cell's internal structure and regulating cell-cycle progression (Wang et al., 2024b). Its expression in control cells is tightly clustered. CDKN1B (perturbed gene) restrains cell division by putting a checkpoint on the DNA replication phase – when this gene is knocked down, the checkpoint is lost, causing cells to progress through the cycle asynchronously (Sun et al., 2016). As a result, the TUBA1B expression becomes more variable across the population. As seen in Fig. 2 (top right), *MLP+Hist* captures this increase in

[2] /gene/1027 [3] /gene/5454 [4] /gene/10376 [5] /gene/8364

| Cell line | Method | Pearson correlation (↑) | | | RMAE (↓) |
|---|---|---|---|---|---|
| | | Top 20 | Top 50 | Top 100 | |
| K562 | Non-Ctrl Mean | 0.460 ± 0.03 | 0.495 ± 0.02 | 0.513 ± 0.02 | 0.981 ± 0.01 |
| | GEARS | 0.472 ± 0.03 | 0.504 ± 0.03 | 0.513 ± 0.03 | 1.050 ± 0.02 |
| | scLAMBDA | 0.618 ± 0.03 | 0.666 ± 0.01 | 0.675 ± 0.01 | **0.865 ± 0.02** |
| | GP (LLM) | 0.615 ± 0.02 | 0.659 ± 0.02 | 0.666 ± 0.02 | 0.911 ± 0.02 |
| | MLP+Mean (LLM) | **0.658 ± 0.02** | **0.701 ± 0.02** | **0.710 ± 0.02** | 0.871 ± 0.03 |
| | MLP+Hist (LLM) | 0.653 ± 0.03 | 0.697 ± 0.02 | 0.703 ± 0.02 | 0.876 ± 0.02 |
| RPE1 | Non-Ctrl Mean | 0.654 ± 0.03 | 0.702 ± 0.02 | 0.723 ± 0.02 | 0.893 ± 0.03 |
| | GEARS | 0.553 ± 0.02 | 0.621 ± 0.03 | 0.660 ± 0.03 | 0.955 ± 0.03 |
| | scLAMBDA | 0.678 ± 0.02 | 0.725 ± 0.02 | 0.746 ± 0.02 | **0.855 ± 0.03** |
| | GP (LLM) | 0.675 ± 0.03 | 0.726 ± 0.02 | 0.746 ± 0.02 | 0.891 ± 0.03 |
| | MLP+Mean (LLM) | **0.685 ± 0.03** | **0.733 ± 0.02** | **0.756 ± 0.02** | 0.860 ± 0.03 |
| | MLP+Hist (LLM) | 0.672 ± 0.03 | 0.726 ± 0.02 | 0.749 ± 0.02 | 0.859 ± 0.02 |

Table 2: Mean prediction performance (± standard deviation across folds) on unseen perturbations from the two Replogle cell lines. For each method, we report correlations (between predicted and observed mean shifts) over the top DE genes and RMAE over all genes.

variance (and reduction in mean), demonstrating the ability to predict higher-order effects beyond the mean. *scLAMBDA* predicts a similar effect, but does not capture the same magnitude of shift in variance.

H4C3, a histone gene involved in DNA packaging and gene regulation (Seal et al., 2022), shows a bimodal distribution in control conditions, indicating distinct transcriptional states. When POU3F2 is knocked down, more cells enter the low-expression state, and fewer remain high (Eisen et al., 1995). As seen in Fig. 2 (bottom right), *MLP+Hist* captures this change in proportions (although the mean hardly changes and peak positions remain stable), highlighting the ability to predict subtle yet biologically relevant shifts. *scLAMBDA* however neglects this bimodality altogether in its prediction.

These examples illustrate how distributional predictions yield mechanistic insights, such as variance inflation in cell-cycle genes and shifts between chromatin states, that point-estimate methods cannot detect. While *scLAMBDA* is distribution-aware, it does not reliably capture these shifts in our setting. Appendix E provides per-perturbation distributional shift plots comparing the performance of *MLP+Hist* and *scLAMBDA* on the Replogle K562 dataset, serving as a visual counterpart to the performance gap already quantified in Table 1. Thus, our histogram model offers a more informative basis for biological hypothesis generation.

### 4.5 Comparison to Mean-based Methods

As a sanity check, we test whether modeling distributions compromises mean prediction accuracy. We compare against the graph-based *GEARS*(Roohani et al., 2024), a Gaussian process regressor (*GP*) conditioned on concatenated LLM embeddings(Märtens et al., 2024), and *scLAMBDA*Wang et al. (2024a). We also include two mean baselines: *Non-Ctrl Mean* (global perturbed mean) and *MLP+Mean* (an MLP mean regressor, $\mathbb{R}^d \to \mathbb{R}^G$, with the same architecture as *MLP+Hist*). We exclude foundation models (e.g., scGPT(Cui et al., 2024)) as they are easily outperformed by *GEARS* and *Non-Ctrl Mean* on unseen perturbations (Ahlmann-Eltze et al., 2024; Märtens et al., 2025).

Following prior work (Bereket & Karaletsos, 2023; Roohani et al., 2024), we report the perturbation-averaged (i) Pearson correlation between predicted ($\hat{\mu} - \mu_{\text{ctrl}}$) and observed ($\mu - \mu_{\text{ctrl}}$) mean shift, computed over the top 20, 50, and 100 differentially expressed (DE) genes, and (ii) RMAE of the predicted mean shift over all genes. Evaluating shifts relative to control penalizes models that reproduce the control mean, which can inflate correlations (Märtens et al., 2024).

Table 2 compares methods on unseen perturbations from the two Replogle cell lines. On both, *MLP+Hist* matches other LLM-based models, including the dedicated mean regressor *MLP+Mean*, while simultaneously modeling a richer output space. It also outperforms *Non-Ctrl Mean* and *GEARS* on all metrics. For *MLP+Hist*, mean-shift correlations show the same saturation behaviour as distributional metrics with increasing number of bins (Appendix D).

| **Split** | **Method** | **W1 (↓)** | **Pearson correlation (↑)** | | | **RMAE (↓)** | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | **Top 20** | **Top 50** | **Top 100** | **Mean** | **Var** | **Skew** | **Kurt** |
| Keep | Non-Ctrl Hist | 0.0224 | 0.626 | 0.629 | 0.633 | 0.902 | 0.351 | **0.390** | **0.560** |
| | scLAMBDA | 0.0599 | 0.909 | 0.925 | 0.926 | 0.594 | 0.897 | 0.884 | 0.971 |
| | MLP+Mean (LLM)$^{\dagger}$ | N/A | **0.918** | **0.935** | **0.942** | 0.595 | N/A | N/A | N/A |
| | MLP+Hist (LLM) | **0.0150** | 0.908 | 0.915 | 0.921 | **0.586** | **0.229** | 0.555 | 0.705 |
| Drop | Non-Ctrl Hist | **0.0226** | 0.565 | 0.572 | 0.583 | 0.914 | 0.342 | **0.367** | **0.568** |
| | scLAMBDA | 0.0668 | **0.587** | **0.608** | **0.617** | 1.046 | 0.916 | 0.917 | 0.983 |
| | MLP+Mean (LLM)$^{\dagger}$ | N/A | 0.533 | 0.529 | 0.548 | 0.971 | N/A | N/A | N/A |
| | MLP+Hist (LLM) | 0.0228 | 0.478 | 0.485 | 0.504 | **0.969** | **0.339** | 0.542 | 0.691 |

Table 3: Double Perturbation Performance on the Norman et al. (2019) dataset. For each method, we report both distributional and mean metrics; the Wasserstein-1 (W1) distance between predicted and target distributions, correlations (between predicted and observed mean shifts) over the top DE genes, and RMAE in statistical moments over all genes. $^{\dagger}$Distributional metrics are not applicable to mean models.

Despite weaker distributional performance, scLAMBDA is competitive on mean-shift correlation and achieves the best mean-shift RMAE on both K562 and RPE1. This underscores that mean accuracy alone cannot validate distributional models. The simple *Non-Ctrl Mean* baseline outperforms GEARS on RPE1 and is competitive on K562, highlighting the strength of global response trends; nevertheless, the LLM-based predictors consistently exceed these baselines, highlighting the benefit of LLM-derived biological priors for generalizable perturbation prediction.

Finally, our conclusions do not depend on a particular embedding: across alternative text/sequence representations, performance is broadly consistent, with modest gains from newer text models (Appendix F, Table 6).

### 4.6 Extending to Combinatorial Perturbations

Our approach extends naturally from single-gene to combinatorial perturbations that target a set of genes $\mathcal{P}$. We represent a perturbation $p$ acting on $\mathcal{P}$ by composing the corresponding gene embeddings via mean pooling,

$$\mathbf{e}_p = \frac{1}{|\mathcal{P}|} \sum_{p_i \in \mathcal{P}} \mathbf{e}_{p_i}. \tag{14}$$

For a double perturbation, $\mathcal{P} = \{p_1, p_2\}$, and Eq. (14) reduces to $\mathbf{e}_p = \frac{1}{2}(\mathbf{e}_{p_1} + \mathbf{e}_{p_2})$. This composition rule is biologically plausible as many pairwise perturbation effects are approximately additive, and it provides a simple mechanism to generalise beyond single perturbations.

We evaluate this strategy on the Norman dataset under two standard splits: (i) *Keep*, where the training set may include perturbations of each constituent gene individually and the test set contains only the held-out combination $p_1 + p_2$; and (ii) *Drop*, where the training set excludes all perturbations involving either constituent gene, and the test set contains $p_1 + p_2$. Results are reported in Table 3.

In the *Keep* split, the MLP models perform strongly. *MLP+Mean* achieves the best top-DE correlations, while MLP+Hist is best on distributional shape (lowest W1) and strong low-order moment accuracy; *scLAMBDA* is also competitive on mean shifts, but lags on distributional metrics. In *Drop*, performance decreases across all methods due to both unseen composition and fewer training perturbations (109–134 per fold vs. 215–216 in *Keep*). Here, *scLAMBDA* achieves the best top-DE correlations, while *MLP+Hist* remains competitive on W1 and low-order moments, with higher-moment errors expected to improve at higher histogram resolution (Appendix D). Given the modest size of the Norman dataset, results are sensitive to data scarcity and split variance and should be interpreted cautiously; regardless, they support embedding composition as a simple route to combinatorial prediction and motivate evaluation on larger multi-gene benchmarks as they become available. Scaling this evaluation to datasets with greater combinatorial coverage will be important for isolating if gains come from compositional structure or data availability.

# 5 Conclusion

We introduced a simple, efficient method for predicting gene expression distributions following genetic perturbations. Combining LLM-informed gene embeddings with a non-parametric histogram output, we showed on standard Perturb-seq datasets that this approach (i) outperforms baselines in predicting distributional structure at a fraction of the training cost, (ii) reveals mechanistic insights such as variance inflation and multimodality that enable improved hypothesis generation, and (iii) matches state-of-the-art mean prediction, demonstrating that distributional accuracy need not trade off against point-estimate performance. Our results also confirm the benefit of LLM-derived biological priors for generalization to unseen perturbations. More broadly, we highlight a gap between the distributional objectives of existing generative models and their empirical marginal accuracy, suggesting that simpler, explicitly validated approaches may be preferable until joint models demonstrate stronger marginal fidelity.

Several limitations remain. First, the histogram representation constrains predictions to the expression range observed during training, limiting extrapolation; continuous alternatives (e.g., KDE) could address this at higher complexity. Second, while we focus on marginals, extending to lightweight joint models (e.g., copulas) that preserve marginals is a natural next step. Finally, while our qualitative results suggest distributional outputs yield biological insights, systematic evaluation on downstream tasks remains important. One promising application is optimal experimental design, where distributional predictions provide estimates of aleatoric uncertainty, supporting more informed perturbation prioritization under fixed experimental budgets (Huang et al., 2023).

### Acknowledgements

KR is supported by the EPSRC Centre for Doctoral Training in Autonomous Intelligent Machines and Systems EP/S024050/1. JGH is supported by a Novo Nordisk Postdoctoral Fellowship in partnership with the University of Oxford, and the David Cockayne Junior Research Fellowship from Linacre College, Oxford. SQ is supported by Wellcome Trust in partnership with the University of Oxford.

## References

Abhinav K. Adduri, Dhruv Gautam, Beatrice Bevilacqua, Alishba Imran, Rohan Shah, Mohsen Naghipourfar, Noam Teyssier, Rajesh Ilango, Sanjay Nagaraj, Mingze Dong, Chiara Ricci-Tam, Christopher Carpenter, Vishvak Subramanyam, Aidan Winters, Sravya Tirukkovular, Jeremy Sullivan, Brian S. Plosky, Basak Eraslan, Nicholas D. Youngblut, Jure Leskovec, Luke A. Gilbert, Silvana Konermann, Patrick D. Hsu, Alexander Dobin, Dave P. Burke, Hani Goodarzi, and Yusuf H. Roohani. Predicting cellular responses to perturbation across diverse contexts with state. *bioRxiv*, 2025. doi: 10.1101/2025.06.26.661135. URL `https://www.biorxiv.org/content/early/2025/07/10/2025.06.26.661135`.

C. Ahlmann-Eltze, W. Huber, and S. Anders. Deep learning-based predictions of gene perturbations effects do not yet outperform simple linear baselines. *bioRxiv*, 2024.09.16.613342, 2024.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer Normalization. *arXiv preprint arXiv:1607.06450*, 2016.

M. Bereket and T. Karaletsos. Modelling cellular perturbations with the sparse additive mechanism shift variational autoencoder. *NeurIPS*, 2023.

C. Bunne, S. G. Start, and G. Gut et al. Learning single-cell perturbation responses using neural optimal transport. *Nature Methods*, 20:1759–1768, 2023.

Y. Chen and J. Zou. GenePT: A simple but effective foundation model for genes and cells built from ChatGPT. *bioRxiv*, 2023.10.16.562533, 2024.

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

L. de Torrenté, S. Zimmerman, M. Suzuki, M. Christopeit, J. M. Greally, and J. C. Mar. The shape of gene expression distributions matter: how incorporating distribution shape improves the interpretation of cancer transcriptomic data. *BMC Bioinformatics*, 21:562, 2020.

A. Dixit, O. Parnas, B. Li, J. Chen, C. P. Fulco, L. Jerby-Arnon, N. D. Marjanovic, D. Dionne, T. Burks, R. Raychowdhury, B. Adamson, T. M. Norman, E. S. Lander, J. S. Weissman, N. Friedman, and A. Regev. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell*, 167:1853–1866, 2016.

T. Eisen, D. J. Easty, D. C. Bennett, and C. R. Goding. The POU domain transcription factor Brn-2: elevated expression in malignant melanoma and regulation of melanocyte-specific gene expression. *Oncogene*, 11(10):2157–2164, 1995.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:7112–7127, 2022.

M. Hao, J. Gong, X. Zeng, C. Liu, Y. Guo, X. Cheng, T. Wang, J. Ma, X. Zhang, and L. Song. Large-scale foundation model on single-cell transcriptomics. *Nature Methods*, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

Kexin Huang, Romain Lopez, Jan-Christian Hütter, Takamasa Kudo, Antonio Rios, and Aviv Regev. Sequential optimal experimental design of perturbation screens guided by multi-modal priors. *bioRxiv*, 2023. doi: 10.1101/2023.12.12.571389. URL https://www.biorxiv.org/content/early/2023/12/13/2023.12.12.571389.

Ehsan Imani, Kai Luedemann, Sam Scholnick-Hughes, Esraa Elelimy, and Martha White. Investigating the Histogram Loss in Regression. *arXiv*, 2402.13425, 2024.

Diederik P Kingma, Max Welling, et al. Auto-Encoding Variational Bayes, 2013.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574.

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

R. Lopez, N. Tagasovska, S. Ra, K. Cho, J. Pritchard, and A. Regev. Learning Causal Representations of Single Cells via Sparse Mechanism Shift Modeling. In *Proceedings of the Second Conference on Causal Learning and Reasoning*, volume 213 of *Proceedings of Machine Learning Research*, pp. 662–691, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. *arXiv preprint arXiv:1711.05101*, 2017.

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

M. Lotfollahi, A. K. Susmelj, and C. De Donno et al. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology*, 19:e11517, 2023.

K. Märtens, R. Donovan-Maiye, and J. Ferkinghoff-Borg. Enhancing generative perturbation models with LLM-informed gene embeddings. In *ICLR 2024 Workshop on Machine Learning for Genomics Explorations*, 2024.

Kaspar Märtens, Marc Boubnovski Martell, Cesar A. Prada-Medina, and Rory Donovan-Maiye. LangPert: LLM-Driven Contextual Synthesis for Unseen Perturbation Prediction. In *ICLR 2025 Workshop on Machine Learning for Genomics Explorations*, 2025.

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

J. Paulsson. Models of stochastic gene expression. *Physics of Life Reviews*, 2:157–175, 2005.

A. Raj and A. van Oudenaarden. Nature, Nurture, or Chance: Stochastic Gene Expression and its Consequences. *Cell*, 135:216–226, 2008.

J. M. Replogle, R. A. Saunders, and A. N. Pogson et al. Mapping information-rich genotype-phenotype landscapes with genome-scale perturb-seq. *Cell*, 185:2559–2575, 2022.

A. Rives, J. Meier, T. Sercu, and R. Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. U.S.A.*, 118: e2016239118, 2021.

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

R. L. Seal, P. Denny, E. A. Bruford, A. K. Gribkova, D. Landsman, W. F. Marzluff, M. McAndrews, A. R. Panchenko, A. K. Shaytan, and P. B. Talbert. A standardized nomenclature for mammalian histone genes. *Epigenetics & Chromatin*, 15:34, 2022.

C. Sun, G. Wang, K. H. Wrighton, H. Lin, Z. Songyang, X.-H. Feng, and X. Lin. Regulation of p27Kip1 phosphorylation and G1 cell cycle progression by protein phosphatase PPM1G. *American Journal of Cancer Research*, 6(10):2207–2220, 2016.

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

Gefei Wang, Tianyu Liu, Jia Zhao, Youshu Cheng, and Hongyu Zhao. Modeling and predicting single-cell multi-gene perturbation responses with scLAMBDA. *bioRxiv*, 2024a. doi: 10.1101/2024.12.04.626878. URL `https://www.biorxiv.org/content/early/2024/12/08/2024.12.04.626878`.

Y. Wang, Y. Li, Y. Jing, Y. Yang, H. Wang, D. Ismtula, and C. Guo. TUBA1B promotes proliferation and migration in hepatocellular carcinoma. *Scientific Reports*, 14:8201, 2024b.

# APPENDIX

## A RESOURCE USE FOR TRAINING

Table 4 compares the training time, peak GPU memory usage, and learnable parameters of several methods described in this work for the two Replogle cell lines. Among the learnable distributional models, *MLP+Hist* is both the quickest to train and the lightest in terms of memory usage despite carrying more parameters. For example, on the RPE1 data, it trains ∼2.7 - 3.3x faster than the *TG / ZITG* baselines and requires ∼8.2x less memory to train.

The extra cost for the distributional baselines is due to the need to calculate and optimize log-likelihoods for every post-perturbation sample per gene. Moreover, since the number of cells varies across perturbations, the data is padded up to the largest sample size in its mini-batch, wasting computation and increasing memory usage. In contrast, *MLP+Hist* learns from a fixed-size compression of the data, enabling simple and scalable modeling of genetic perturbation effects.

| **Cell line** | **Method** | **Train time**<br>(mins per fold) | **GPU Memory Usage**<br>(MB) | **Learnable Params** |
|---|---|---|---|---|
| K562 | Non-Ctrl Mean | N/A | 1,013 | N/A |
| | Non-Ctrl Hist | N/A | 1,617 | N/A |
| | MLP+Mean (LLM) | 3.14 | 1,575 | 8,802,696 |
| | MLP+Hist (LLM) | 9.89 | 3,219 | 80,552,696 |
| | TG (LLM) | 20.49 | 10,735 | 13,927,696 |
| | ZITG (LLM) | 24.52 | 12,023 | 19,052,696 |
| RPE1 | Non-Ctrl Mean | N/A | 1,091 | N/A |
| | Non-Ctrl Hist | N/A | 1,885 | N/A |
| | MLP+Mean (LLM) | 4.47 | 1,655 | 8,802,696 |
| | MLP+Hist (LLM) | 14.93 | 3,503 | 80,552,696 |
| | TG (LLM) | 41.16 | 28,970 | 13,927,696 |
| | ZITG (LLM) | 49.49 | 28,710 | 19,052,696 |

Table 4: Training costs of methods when trained on a single NVIDIA A40 GPU.

## B Higher-order Statistics

We calculate the variance, skewness, and excess kurtosis of a distribution from its raw moments $m_k = \mathbb{E}[X^k]$, where $k \in \{1, 2, 3, 4\}$, using the following formulas:

$$\begin{aligned}
\mathrm{Var}[X] &= m_2 - m_1^2 \\
\mathrm{Skew}[X] &= \frac{m_3 - 3\, m_1 m_2 + 2\, m_1^3}{\mathrm{Var}[X]^{3/2}} \\
\mathrm{Kurt}[X] &= \frac{m_4 - 4\, m_1 m_3 + 6\, m_1^2 m_2 - 3 m_1^4}{\mathrm{Var}[X]^2} - 3.
\end{aligned}$$

We specify the moment expressions for each distribution below.

**Histogram.** Given a histogram with $B$ bin intervals $[x_b^{(l)}, x_b^{(r)}]$ of widths $w_b$ having probabilities $h_b$,

$$m_k^{\text{hist}} = \sum_{b=1}^{B} \frac{(x_b^{(r)})^{k+1} - (x_b^{(l)})^{k+1}}{k+1} \frac{h_b}{w_b}.$$

**TG.** Given parameters $(\mu, \sigma)$ and support $[a, b]$,

$$m_k^{\text{TG}} = \sum_{r=0}^{k} \binom{k}{r} \mu^{k-r} \sigma^r L_r,$$

with

$$\begin{aligned}
L_0 &= 1 \\
L_1 &= -\frac{\varphi(\beta) - \varphi(\alpha)}{\Phi(\beta) - \Phi(\alpha)} \\
L_r &= -\frac{\beta^{r-1}\varphi(\beta) - \alpha^{r-1}\varphi(\alpha)}{\Phi(\beta) - \Phi(\alpha)} + (r-1)\, L_{r-2}, \quad r \geq 2,
\end{aligned}$$

where $\alpha = (a - \mu)/\sigma$, $\beta = (b - \mu)/\sigma$, and $\varphi(\cdot)$ and $\Phi(\cdot)$ are the density and CDF of the standard normal distribution[6].

**ZITG.** Given zero-inflation probability $\pi$, TG parameters $(\mu, \sigma)$, and support $[a, b]$,

$$m_k^{\text{ZITG}} = \pi\, m_k^{\mathcal{U}} + (1 - \pi)\, m_k^{\text{TG}},$$

where the moments of the uniform spike $\mathcal{U}_{[-\epsilon, \epsilon]}$ are given by

$$m_k^{\mathcal{U}} = \frac{1 - (-1)^{k+1}}{k+1} \frac{\epsilon^k}{2}.$$

[6] `https://people.sc.fsu.edu/~jburkardt/presentations/truncated_normal.pdf`

# C Hyperparameters

We experiment with using a weighted sum of three loss terms,

$$\mathcal{L}_{\text{total}} = \lambda_{\text{ce}}\mathcal{L}_{\text{ce}} + \lambda_{\text{wass}}\mathcal{L}_{\text{wass}} + \lambda_{\text{mse}}\mathcal{L}_{\text{mse}},$$

where $\lambda_{\text{ce}} + \lambda_{\text{wass}} + \lambda_{\text{mse}} = 1$ and $\mathcal{L}_{\text{ce}}$ is a simple cross-entropy loss (which we average over perturbations in the training set):

$$\mathcal{L}_{\text{ce}} = -\frac{1}{G}\sum_{g=1}^{G}\sum_{b=1}^{B} h_{gb} \log \hat{h}_{gb}.$$

We perform a coarse grid search (with step size $0.25$) over valid triples $(\lambda_{\text{ce}}, \lambda_{\text{wass}}, \lambda_{\text{mse}})$ and report mean-shift prediction performance for the Replogle K562 cell line in Table 5. The best configuration ($\lambda_{\text{wass}} = 0.75$ and $\lambda_{\text{mse}} = 0.25$) assigns a majority weight to the Wasserstein term and retains the MSE contribution. Adding even a small weight to the cross-entropy loss hurts performance, with the worst configuration assigning all of the weight to it.

| $\lambda_{\text{ce}}$ | $\lambda_{\text{wass}}$ | $\lambda_{\text{mse}}$ | **Corr. Top 20** (↑) |
|---|---|---|---|
| 1 | 0 | 0 | 0.628 |
| 0.5 | 0 | 0.5 | 0.635 |
| 0.25 | 0 | 0.75 | 0.636 |
| 0.75 | 0 | 0.25 | 0.638 |
| 0.75 | 0.25 | 0 | 0.639 |
| 0.25 | 0.5 | 0.25 | 0.639 |
| 0.25 | 0.75 | 0 | 0.644 |
| 0.25 | 0.25 | 0.5 | 0.644 |
| 0.5 | 0.5 | 0 | 0.647 |
| 0 | 1 | 0 | 0.647 |
| 0 | 0.25 | 0.75 | 0.647 |
| 0 | 0.5 | 0.5 | 0.648 |
| **0** | **0.75** | **0.25** | **0.651** |

Table 5: Effect of the loss weights used to train the histogram model.

## D Effect of Histogram Resolution

**Distributional Metrics.** Fig. 3 shows how model performance varies with histogram resolution. On the Replogle data, as the number of bins increases, the RMAEs of the mean, variance, and skewness steadily improve before plateauing at 15–20 bins, suggesting that a moderate resolution is sufficient to recover statistical features while maintaining computational efficiency. The improvement is less consistent across statistics on the Norman dataset, likely due to its relatively small training set size of around 90 perturbations per fold, compared to 1000 in Replogle.

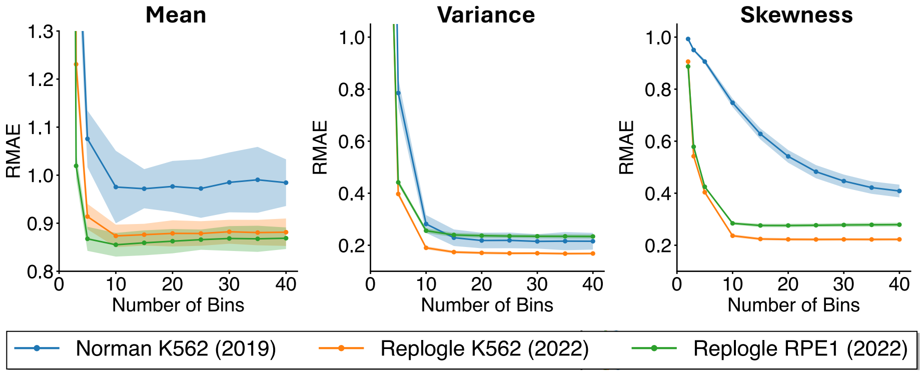

Figure 3: Effect of histogram resolution on the prediction error of post-perturbation statistics for three benchmark datasets. Performance improves with resolution but plateaus at 15-20 bins on the larger Replogle datasets. Shading indicates the standard deviation across folds.

**Mean Metrics.** Fig. 4 shows how *MLP+Hist* performance varies with histogram resolution. Correlations improve with increasing bin count and plateau at 15-20 bins, confirming that a moderate resolution is sufficient to recover shifts in mean expression as well.

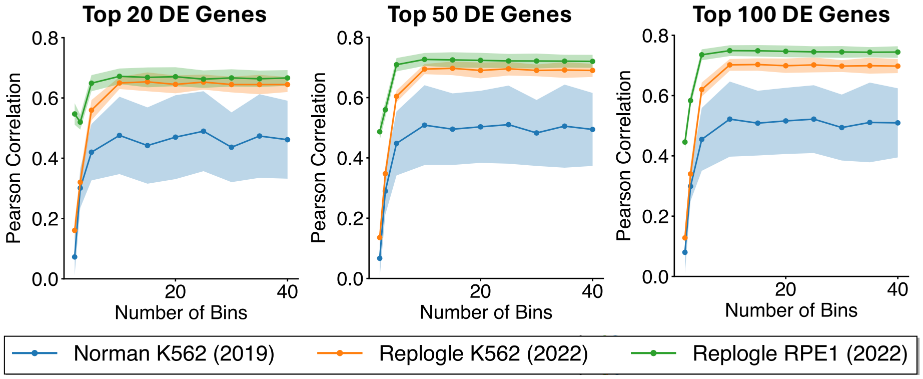

Figure 4: Effect of histogram resolution on the correlation between predicted and true mean shifts over the top DE genes for three benchmark datasets. Performance improves with resolution but plateaus at 15-20 bins. Shading indicates the standard deviation across folds.

## E FURTHER COMPARISON BETWEEN MLP+HIST & SCLAMBDA

Figure 5 visualises the distributional performance differences between *MLP+Hist* and *scLAMBDA* using per-perturbation W1 shift plots on K562. Each panel compares the observed gene-wise shift from control to the shift predicted by each model; tighter alignment with the diagonal indicates better distributional shift prediction. Consistent with Table 1, *scLAMBDA* frequently over-estimates gene-wise distributional shifts, leading to lower per-perturbation correlations, whereas *MLP+Hist* produces more faithful shift magnitudes across genes. This provides an intuitive view of why *scLAMBDA* can remain competitive on mean metrics while underperforming on distributional metrics (NLL, W1 and higher moments).

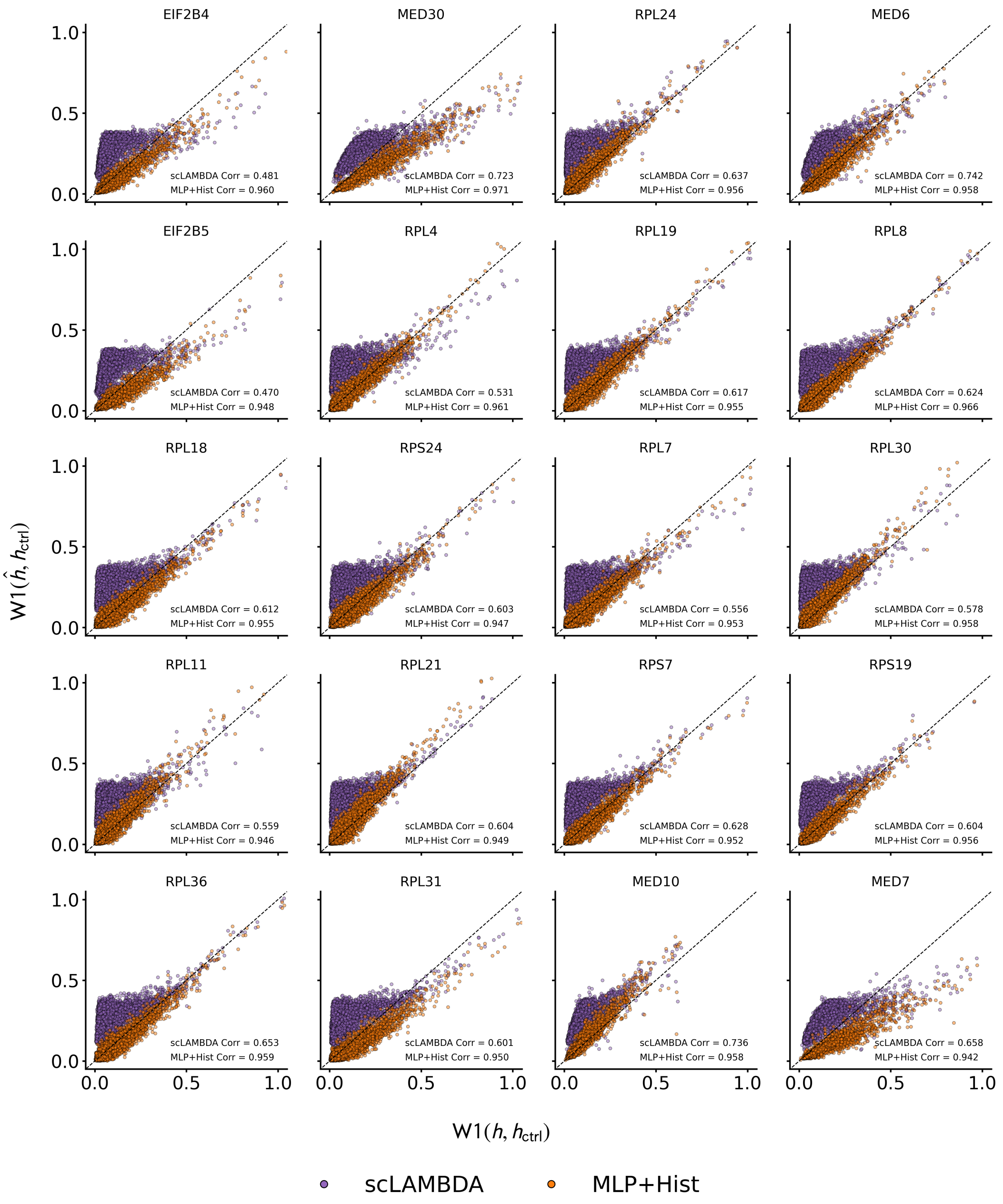

Figure 5: Distributional shift plots on perturbations on K562 (Replogle et al., 2022). For each of the 20 shown perturbations, we compare the observed per-gene distributional shift from control to the shift predicted by both scLAMBDA and MLP+Hist. Each circle represents a gene $g$ with coordinates $x = \mathrm{W1}(h_g, h_g^{\mathrm{ctrl}})$ (target vs. control) and $y = \mathrm{W1}(\hat{h}_g, h_g^{\mathrm{ctrl}})$ (predicted vs. control). For scLAMBDA (blue), W1 is computed via a quantile-based approximation directly from samples; for MLP+Hist (red), W1 is computed between 15-bin histograms. The dashed line denotes perfect agreement ($y = x$), and per-panel Pearson correlations between $x$ and $y$ are reported.

## F Embedding Model Experiments

Table 6 reports an ablation over gene embedding backbones for MLP+Mean (LLM) on unseen perturbations in K562 and RPE1. We compare sequence embeddings alone (ProtT5, ESM2 (Lin et al., 2023)), text embeddings alone (text-embedding-ada-002, llama-embed-nemotron-8b, Qwen3-Embedding-8B), and their concatenations. Across both cell lines, performance is robust to the embedding choice: several text-only models already perform strongly, and combining sequence + text typically yields the best or near-best results. In K562, the strongest overall configuration uses ProtT5 + llama-embed-nemotron-8b (best Top-50/Top-100 correlations and lowest RMAE), while in RPE1 the best results are achieved by ProtT5 + Qwen3-Embedding-8B. Given these relatively small differences, we use ProtT5 + GPT (text-embedding-ada-002) as a consistent default throughout the paper.

| **Cell line** | **Embedding Model** | | **Pearson correlation (↑)** | | | **RMAE (↓)** |
|---|---|---|---|---|---|---|
| | **Sequence** | **Text** | **Top 20** | **Top 50** | **Top 100** | |
| K562 | ProtT5 | – | 0.579 ± 0.020 | 0.625 ± 0.014 | 0.635 ± 0.023 | 0.917 ± 0.018 |
| | ESM2 | – | 0.570 ± 0.041 | 0.610 ± 0.033 | 0.619 ± 0.028 | 0.925 ± 0.020 |
| | – | text-embedding-ada-002 | 0.644 ± 0.048 | 0.685 ± 0.041 | 0.691 ± 0.038 | 0.886 ± 0.029 |
| | – | llama-embed-nemotron-8b | 0.670 ± 0.035 | 0.710 ± 0.028 | 0.716 ± 0.026 | 0.872 ± 0.029 |
| | – | Qwen3-Embedding-8B | 0.660 ± 0.041 | 0.705 ± 0.035 | 0.711 ± 0.033 | 0.875 ± 0.028 |
| | ProtT5 | text-embedding-ada-002 | 0.658 ± 0.029 | 0.703 ± 0.023 | 0.711 ± 0.026 | 0.872 ± 0.029 |
| | ProtT5 | llama-embed-nemotron-8b | 0.670 ± 0.028 | **0.719 ± 0.021** | **0.726 ± 0.020** | **0.862 ± 0.024** |
| | ProtT5 | Qwen3-Embedding-8B | **0.671 ± 0.037** | 0.717 ± 0.029 | 0.724 ± 0.029 | 0.868 ± 0.023 |
| | ESM2 | text-embedding-ada-002 | 0.636 ± 0.032 | 0.677 ± 0.031 | 0.683 ± 0.029 | 0.883 ± 0.024 |
| | ESM2 | llama-embed-nemotron-8b | 0.670 ± 0.031 | 0.712 ± 0.027 | 0.718 ± 0.027 | 0.870 ± 0.024 |
| | ESM2 | Qwen3-Embedding-8B | 0.662 ± 0.029 | 0.707 ± 0.029 | 0.714 ± 0.027 | 0.870 ± 0.024 |
| RPE1 | ProtT5 | – | 0.665 ± 0.030 | 0.713 ± 0.027 | 0.733 ± 0.029 | 0.893 ± 0.047 |
| | ESM2 | – | 0.662 ± 0.026 | 0.709 ± 0.026 | 0.728 ± 0.028 | 0.897 ± 0.039 |
| | – | text-embedding-ada-002 | 0.682 ± 0.031 | 0.736 ± 0.033 | 0.758 ± 0.031 | 0.873 ± 0.040 |
| | – | llama-embed-nemotron-8b | 0.688 ± 0.030 | 0.740 ± 0.030 | 0.762 ± 0.028 | 0.860 ± 0.034 |
| | – | Qwen3-Embedding-8B | 0.687 ± 0.028 | 0.738 ± 0.030 | 0.761 ± 0.027 | 0.865 ± 0.044 |
| | ProtT5 | text-embedding-ada-002 | 0.687 ± 0.029 | 0.735 ± 0.023 | 0.756 ± 0.019 | 0.861 ± 0.027 |
| | ProtT5 | llama-embed-nemotron-8b | 0.689 ± 0.030 | 0.740 ± 0.025 | 0.762 ± 0.021 | 0.859 ± 0.031 |
| | ProtT5 | Qwen3-Embedding-8B | **0.692 ± 0.025** | **0.741 ± 0.022** | **0.763 ± 0.022** | **0.853 ± 0.029** |
| | ESM2 | text-embedding-ada-002 | 0.673 ± 0.034 | 0.724 ± 0.033 | 0.743 ± 0.031 | 0.866 ± 0.037 |
| | ESM2 | llama-embed-nemotron-8b | 0.682 ± 0.030 | 0.735 ± 0.031 | 0.756 ± 0.029 | 0.859 ± 0.039 |
| | ESM2 | Qwen3-Embedding-8B | 0.687 ± 0.034 | 0.737 ± 0.032 | 0.758 ± 0.030 | 0.857 ± 0.039 |

Table 6: Mean prediction performance (± standard deviation across folds) on unseen perturbations from the two Replogle cell lines, using MLP+Mean (LLM). For each embedding model, we report correlations (between predicted and observed mean shifts) over the top DE genes and RMAE over all genes.