ON PROBABILITY MISALLOCATION IN SINGLE-CELL PERTURBATION PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Single-cell perturbation prediction aims to model how interventions reshape cellular populations. Despite substantial progress across diverse modeling paradigms, little is known about the structure of the distributional errors that remain. We uncover a recurring failure across modeling paradigms: models misallocate probability across expression levels, even when mean expression is accurately predicted. We term the shared failure mode **probability misallocation**. In this work, we systematically characterize where misallocation occurs and show that the error can materially alter the biological conclusions that perturbation prediction is intended to support. Guided by this diagnosis, we propose several simple yet effective corrections that substantially reduce probability misallocation and improve perturbation prediction quality across modeling paradigms and datasets. Further analysis reveals why misallocation persists under existing objectives and how data characteristics shape the supervision required. Together, our work provides a new perspective on single-cell perturbation prediction and practical guidance for building more reliable virtual-cell models to accelerate biological discovery.

1 INTRODUCTION

Single-cell perturbation prediction is a central problem for building virtual-cell systems, with the potential to enable large-scale studies of cellular responses and accelerate biological discovery (Bunne et al., 2024). Current methods span a diverse range of modeling paradigms, including deterministic and generative models at the cell level (Ahlmann-Eltze et al., 2025; Lotfollahi et al., 2019), alongside transport and distribution-matching approaches at the population level (Bunne et al., 2023; Adduri et al., 2026). Regardless of formulation, these approaches ultimately seek to predict how a perturbation reshapes the distribution of cellular states across a population. In stimulated immune cells, response heterogeneity involves both the fraction of cells with detectable expression of a given transcript and its expression levels among those cells (Shalek et al., 2014).

Despite this progress, the field has largely focused on developing powerful ways to match predicted perturbation responses to observed data. Much less is understood about the errors that remain: when a predicted population deviates from its target, how is that discrepancy distributed across cells and expression levels? We examine this by comparing the predicted and observed fractions of cells exceeding each expression level. Surprisingly, across modeling paradigms, these fractions systematically deviate from the target across the expression range. Fig. 1 provides a concrete example: despite nearly identical mean expression, the target exhibits strong expression in a small subset of cells, whereas the prediction spreads weaker expression across more cells. We refer to this recurring error as **probability misallocation**.

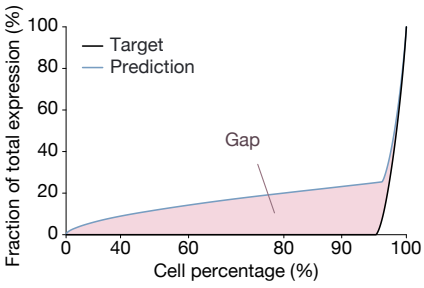


Figure 1: **Similar mean expression can coincide with probability misallocation.** The shaded gap highlights the allocation difference.

We conduct systematic analyses to characterize probability misallocation and determine whether it affects the biological conclusions that virtual-cell models aim to support. First, we rule out mean expression error as an explanation by empirically show-

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

ing that accurate mean prediction does not alleviate probability misallocation. Additionally, we localize misallocation across the expression range. While the largest errors often occur at the zero boundary, positive-expression misallocation remains substantial and becomes more prominent on some datasets. More importantly, misallocation can directly affect differential expression evidence, changing which genes are identified as responsive, with the potential to misdirect experiments and obscure true perturbation mechanisms.

Motivated by these observations, we consider three interventions: (1) hurdle parameterization, testing whether model architecture limits expressiveness; (2) incidence supervision, targeting the dominant zero-boundary error; and (3) multi-threshold supervision, extending probability supervision to positive expression. The results are striking: across modeling paradigms and datasets, parameterization provides limited gains, whereas probability supervision dramatically reduces misallocation and improves perturbation prediction quality. Even a lightweight Gaussian predictor augmented only with incidence supervision achieves the best or tied-best performance on 11 of 14 metrics on each dataset when compared with the original baselines.

To better understand these findings, we further conduct in-depth analyses of the mechanisms underlying them. Specifically, we dissect how different training objectives give rise to misallocation, revealing distinct failure modes behind the shared phenomenon. Finally, we derive a principle for choosing appropriate probability supervision based on data characteristics, guiding the design of future perturbation models.

We highlight the main contributions of this paper below:

(1) We identify and characterize **probability misallocation** as a recurring failure across modeling paradigms, exposing a fundamental blind spot in how single-cell perturbation prediction is understood and evaluated.

(2) We propose several simple yet effective misallocation corrections and validate them across models and datasets, turning the diagnosis into a practical route for improving perturbation prediction.

(3) We dissect why misallocation persists under existing objectives and derive principles for supervision choice, guiding future model design and ultimately accelerating biological discovery.

2 PRELIMINARIES

2.1 TASK FORMULATION

We first present a unified formulation of single-cell perturbation prediction. Given a control cell population, a perturbation, and the associated biological context, single-cell perturbation prediction models the post-perturbation transcriptomic distribution (Bunne et al., 2023; Adduri et al., 2026).

Let $\mathcal{X}_{\text{ctrl}} = \{\mathbf{x}_i\}_{i=1}^{N}$ denote the control cell population, where $\mathbf{x}_i \in \mathbb{R}^G$ denotes the gene-expression vector of a single cell, p denotes the perturbation, and c denotes the biological context. We denote the target post-perturbation distribution by $\mathcal{X}_{\text{pert}}$ and learn a perturbation-response mapping $\mathcal{T}_\theta$ to predict this distribution:

$$\widehat{\mathcal{X}}_{\text{pert}} = \mathcal{T}_\theta(\mathcal{X}_{\text{ctrl}}, p, c), \qquad \widehat{\mathcal{X}}_{\text{pert}} \approx \mathcal{X}_{\text{pert}}. \tag{1}$$

2.2 PERTURBATION RESPONSE MODELING

In this section, we categorize existing single-cell perturbation methods according to their modeling and supervision strategies.

2.2.1 CELL-LEVEL MODELING

Cell-level modeling represents perturbation responses at the resolution of individual cells, using either deterministic mappings or conditional generative models.

Deterministic Modeling. Deterministic formulations represent the perturbation response with a point estimate,

$$\hat{\mathbf{y}} = f_\theta(\mathbf{x}, p, c), \tag{2}$$

where $\mathbf{x}$ denotes the gene-expression vector of a control cell and $\hat{\mathbf{y}}$ denotes the predicted gene-expression vector of a perturbed cell. Training minimizes regression objectives between predicted and target expression, such as MAE or MSE (Ahlmann-Eltze et al., 2025).

Generative Modeling. Generative formulations model a conditional distribution over perturbed cellular states,

$$\mathbf{y} \sim q_\theta(\mathbf{y} \mid \mathbf{x}, p, c), \tag{3}$$

where q_θ denotes the conditional distribution of the perturbed transcriptomic state. Generative modeling learns the conditional response distribution using generative objectives such as the ELBO (Kingma & Welling, 2014) or denoising loss (He et al., 2026).

2.2.2 Population-level Modeling

Population-level modeling represents perturbation responses directly through transformations or distributions over cellular populations.

Transport-based Modeling. Transport-based formulations learn a transformation from the control distribution to the perturbed distribution,

$$\widehat{P}_{c,p} = \mathcal{T}_\theta(P_{\text{ctrl}}, p, c), \tag{4}$$

where $\mathcal{T}_\theta$ defines the population-level transport under perturbation p and context c. Transport-based formulations learn the population transformation using objectives defined along the transport path, such as optimal-transport objectives (Bunne et al., 2023) or conditional flow-matching losses (Klein et al., 2025).

Distribution-based Modeling. Distribution-based formulations directly constrain the predicted perturbed distribution against the target distribution,

$$D\Big(\widehat{P}_{c,p}, P_{c,p}\Big)\,, \tag{5}$$

where D measures the discrepancy between the predicted and target populations. Training minimizes this discrepancy using objectives such as MMD (Gretton et al., 2012) or energy distance (Székely & Rizzo, 2013).

2.3 Quantifying distribution Discrepancies Across Expression Levels

Survival Function. To characterize how gene expression is distributed across a cell population, we use the survival function. For gene g under context c and perturbation p, we define

$$S_{c,p,g}(\tau) = \Pr_{\mathbf{x}\sim P_{c,p}}(x_g > \tau)\,, \tag{6}$$

where τ denotes an expression level. $S_{c,p,g}(\tau)$ gives the fraction of cells whose expression of gene g exceeds τ. The corresponding survival function of the predicted population is

$$\widehat{S}_{c,p,g}(\tau) = \Pr_{\mathbf{x}\sim \widehat{P}_{c,p}}(x_g > \tau)\,. \tag{7}$$

We define the discrepancy between the predicted and target populations at expression level τ as

$$d_{c,p,g}(\tau) = \widehat{S}_{c,p,g}(\tau) - S_{c,p,g}(\tau). \tag{8}$$

A positive $d_{c,p,g}(\tau)$ indicates that the predicted population contains a larger fraction of cells above expression level τ, while a negative value indicates the opposite. For clarity, we omit the subscripts (c, p, g) below when no ambiguity arises.

Kolmogorov–Smirnov Diagnostics. We use the Kolmogorov–Smirnov (KS) statistic (Gretton et al., 2012) to quantify the largest probability discrepancy between the predicted and target expression distributions:

$$K = \sup_{\tau\geq 0} |\widehat{S}(\tau) - S(\tau)|. \tag{9}$$

To separately measure misallocation at zero and at positive expression levels, we define two additional statistics. Let $S_+(\tau) = \Pr(X > \tau \mid X > 0)$ and $\widehat{S}_+(\tau) = \Pr(\widehat{X} > \tau \mid \widehat{X} > 0)$ denote the target and predicted survival functions conditional on positive expression. We define

$$K_0 = |\widehat{S}(0) - S(0)|,$$
$$K_+ = \sup_{\tau>0} |\widehat{S}_+(\tau) - S_+(\tau)|. \tag{10}$$

We refer to K_0, K_+, and K as **Zero KS**, **Positive KS**, and **Overall KS**, respectively. Zero KS measures misallocation in the nonzero rate, Positive KS measures misallocation in the distribution of positive expression values, and Overall KS measures misallocation over the full expression distribution.

Mean Expression Error. To isolate distribution discrepancies from expression-magnitude errors, we additionally report the normalized mean-expression error

$$e_\mu = \frac{|\widehat{\mu} - \mu|}{s_g}, \tag{11}$$

where s_g is a gene-wise normalization scale estimated from training data and fixed across models.

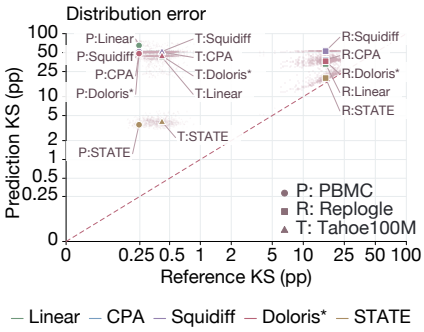


Figure 2: **Probability misallocation persists across modeling paradigms and datasets.** Above the diagonal, prediction KS exceeds reference KS, with larger gaps indicating more severe probability misallocation. Doloris* denotes the modified Doloris variant that removes the sparsity mask and is trained without BCE supervision.

3 Observation

3.1 Probability Misallocation Across Modeling Paradigms

We applied the metrics in Sec. 2.3 to representative perturbation models on PBMC, Replogle, and Tahoe100M. For each model–dataset pair, we compared the mean KS between the predicted and target populations (prediction KS) with the KS between two disjoint target populations (reference KS).

Fig. 2 shows substantial probability misallocation in predictions across modeling paradigms and datasets. Cell-level models show the largest deviations, with Linear, CPA, and Squidiff lying far above reference KS on PBMC and Tahoe100M. Population-level models maintain the same tendency, though it remains less extreme, while the KS for STATE still remains approximately $10\times$ the reference on PBMC and Tahoe100M. On Replogle, the pattern is less pronounced, yet model predictions consistently display marked discrepancies against the target reference.

3.2 Characterizing the Misallocation

3.2.1 Misallocation Across Mean-Error Levels

To further isolate the effect of mean expression error on probability misallocation, we conduct a controlled experiment in which we group each perturbation–gene pair according to the gene-wise normalized mean error. We then split the target population into two subsets and compute the KS and mean expression error between them as reference.

We report the ratio of prediction KS to reference KS. As shown in Fig. 3, the KS ratio shows no consistent monotonic trend with mean expression error across models and datasets. Crucially, substantial probability misallocation persists even when the model's mean expression error falls within the 95th percentile of reference mean error. Thus, probability misallocation remains even when mean expression is accurately predicted.

3.2.2 Locating the Misallocation

To investigate the internal structure of probability misallocation, we ask where the discrepancy occurs and how severe it is across the expression range. For localization, we identify the expression

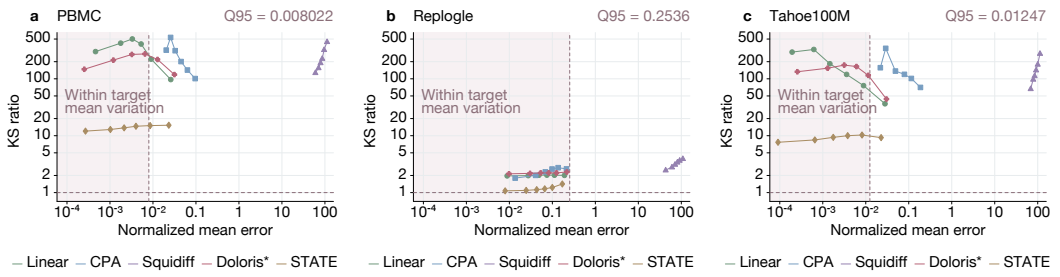


Figure 3: **Accurate mean prediction does not eliminate probability misallocation.** Ratio of prediction KS to reference KS across normalized mean error bins on (a) PBMC, (b) Replogle, and (c) Tahoe100M. Higher KS ratios suggest relatively more severe misallocation. Shading indicates relatively accurate mean prediction: errors below the 95th percentile of reference mean error.

thresholds with the largest probability discrepancy (KS) and report the frequency of these peak locations. For severity, we assess the relative contribution of each peak location to the overall KS. We further compute KS over progressively higher expression thresholds to measure how much discrepancy remains beyond the low-expression range.

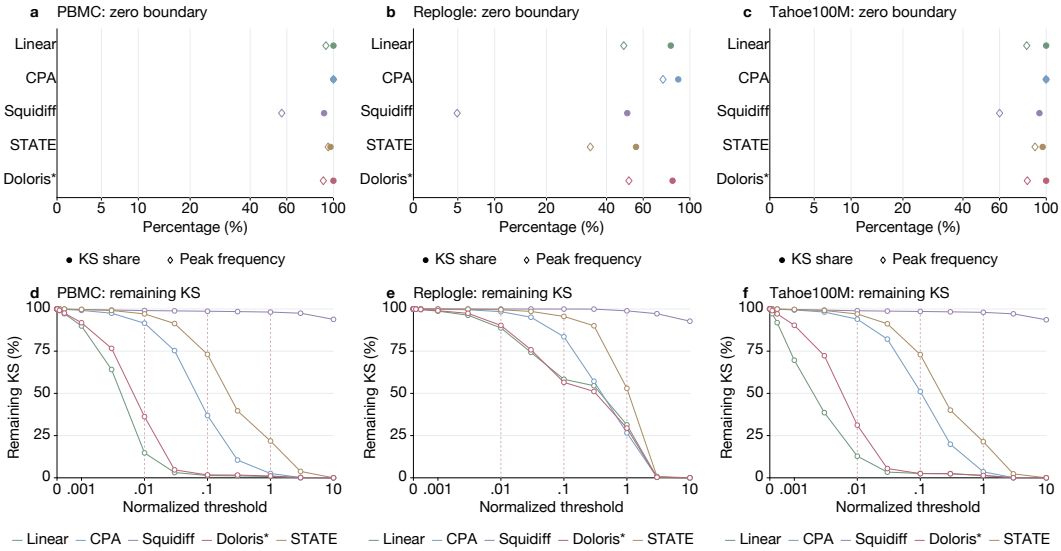


Figure 4: **Misallocation at zero and positive expression.** Columns show PBMC, Replogle, and Tahoe100M. (a–c) Zero-boundary contribution to mean KS (filled circles) and frequency of KS maxima at zero (open diamonds). (d–f) Fraction of overall KS remaining above increasing expression thresholds. Zero boundary errors dominate misallocation in many settings, but substantial mismatch persists at positive expression levels.

Prominent Misallocation at the Expression Boundary. As shown in Fig. 4a–c, we observe pronounced probability misallocation at the zero-expression boundary across models and datasets. On PBMC and Tahoe100M, the boundary discrepancy accounts for nearly the entire mean KS for Linear, CPA, STATE, and Doloris*, and zero-peak frequencies are consistently high. Although this pattern is less pronounced on Replogle, the Zero KS accounts for over 50% of the Overall KS for all models, indicating that the zero-expression boundary is still an important source of misallocation.

Substantial Misallocation Remains Within Positive Expression. Despite the strong discrepancy at zero, substantial probability misallocation exists at positive expression levels. As the minimum expression threshold increases, a considerable fraction of the KS remains across models (Fig. 4d–f), showing that the discrepancy is not restricted to the zero-expression boundary. This pattern is particularly evident for CPA, Squidiff, and STATE on PBMC and Tahoe100M, while Linear and

Doloris* decay more rapidly. Additionally, Replogle exhibits a more salient discrepancy within the positive expression range, with substantial KS remaining across increasing expression thresholds.

3.3 Why Does Misallocation Matter?

Single-cell perturbation screens have revealed regulatory relationships and mechanisms of genetic interaction (Norman et al., 2019). Single-cell perturbation prediction is ultimately intended to support biological discovery by identifying responsive genes and prioritizing interventions for experimental follow-up. Prior work has shown that false discoveries in single-cell differential expression can mislead biological interpretation and experimental follow-up (Squair et al., 2021). We therefore ask whether the probability misallocation above can affect these biological conclusions. Thus, we establish analytically that misallocation can alter differential expression evidence through errors in expression incidence and expression ordering, changing gene rankings and responsive gene selection (Appendix D.1). Inaccurate gene-level discoveries can consequently distort which interventions are prioritized for experimental follow-up. More broadly, such distortions can mislead biological interpretation, obscure true perturbation mechanisms, and divert experimental effort toward spurious hypotheses. Such distortions risk directing experimental resources toward unsupported hypotheses, making reliable response inference a central requirement for virtual-cell models.

3.4 Correcting the Misallocation

Motivated by the preceding observations, we consider three progressively stronger interventions in the model parameterization and training objective. They address probability misallocation from complementary perspectives: parameterization, expression incidence, and distributional structure.

Hurdle Parameterization. We first test whether the output parameterization limits expressiveness and contributes to misallocation. We separate expression occurrence from positive expression using a hurdle representation (Finak et al., 2015):

$$Z \sim \mathrm{Bernoulli}(q), \qquad X = ZX_+, \qquad X_+ > 0, \tag{12}$$

where q controls expression occurrence and X_+ models the conditional positive expression distribution. We retain the original training objective to isolate limitations imposed by the model architecture.

Incidence Supervision. Motivated by the prominent zero-boundary discrepancy, we next directly supervise expression incidence. Let $\pi = \Pr_{X\sim P}(X > 0)$ denote the target incidence, and let q denote its prediction. We augment the training loss with the expected binary logarithmic loss (Gneiting & Raftery, 2007):

$$\mathcal{L}_{\mathrm{inc}} = -\pi \log q - (1-\pi)\log(1-q), \tag{13}$$

yielding

$$\mathcal{L} = \mathcal{L}_{\mathrm{base}} + \lambda_{\mathrm{inc}}\mathcal{L}_{\mathrm{inc}}. \tag{14}$$

Multi-Threshold Supervision. Finally, as apparent misallocation remains beyond the zero boundary, we broaden the supervision across the expression range. For M thresholds $\{\tau_k\}_{k=1}^{M}$, we constrain the predicted and target exceedance probabilities, $\widehat{S}(\tau_k)$ and $S(\tau_k)$, using a distributional discrepancy over multiple thresholds:

$$\mathcal{L}_{\mathrm{MT}} = \sum_{k=1}^{M} w_k \, \mathcal{D}\left(\widehat{S}(\tau_k), S(\tau_k)\right). \tag{15}$$

Here, $\mathcal{D}$ can be instantiated with objectives such as Brier-style losses (Gneiting & Raftery, 2007) or distributional distances such as MMD over threshold indicator features. The final objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{base}} + \lambda_{\mathrm{MT}}\mathcal{L}_{\mathrm{MT}}. \tag{16}$$

4 Experiments

We conduct two complementary sets of experiments. First, we assess how far a simple model can go by correcting only zero-boundary misallocation. Specifically, we use a lightweight Gaussian predictor with an auxiliary BCE loss (architecture: Appendix B.1.2; training: Appendix B.3). Second, we apply different correction strategies to existing models across modeling paradigms.

4.1 Experimental Setup

Datasets. We evaluate on three well-established single-cell perturbation datasets: PBMC for cytokine perturbations, Tahoe100M (Zhang et al., 2026) for drug perturbations, and Replogle (Replogle et al., 2022; Nadig et al., 2025) for genetic perturbations. To ensure a fair and consistent comparison, we adopt the same preprocessing, 2,000-HVG gene space, and dataset splits for all methods, following the evaluation framework of STATE and PerturbDiff.

Baselines. We use representative perturbation models as baselines. These baselines cover the four modeling paradigms introduced in Sec. 2.2. (1) Cell-level deterministic models: Mean (Yuan et al., 2026), Linear (Ahlmann-Eltze et al., 2025), and scGPT (Cui et al., 2024). (2) Cell-level generative models: CPA (Lotfollahi et al., 2023), Squidiff (He et al., 2026), and Doloris* (Chi et al., 2026). (3) Transport-based population models: CellFlow (Klein et al., 2025). (4) Distribution-based population models: STATE (Adduri et al., 2026) and PerturbDiff (Yuan et al., 2026).

Training. We follow the same training protocol as PerturbDiff (Yuan et al., 2026) and optimize all configurations for 20,000 steps, evaluating on the validation set every 2,000 steps. We report the results from the checkpoint that achieves the lowest validation loss. We repeat each benchmark experiment with three random seeds and report the mean for every metric. For the correction experiments, we keep all training settings unchanged and modify only the correction component.

Metrics. To maintain comparability with STATE and PerturbDiff (Adduri et al., 2026; Yuan et al., 2026), we retain the fourteen metrics and group them by evaluation granularity: (1) Expression Accuracy, measured by R^2, PDCorr, MSE, and MAE; (2) Differential Expression, evaluated by DEOver, DEPrec, DirAgr, LFCSpear, AUROC, and AUPRC; and (3) Perturbation Effects, captured by ES, PDS-L_1, PDS-L_2, and PDS-cos. For the correction experiments, we additionally evaluate the predicted distributions using Zero KS, Positive KS, and Overall KS, defined in Sec. 2.3.

Further experimental details are provided in Appendices A and B.1.

4.2 Main Results

4.2.1 How Far Can Incidence Supervision Alone Go?

As shown in Fig. 5, a lightweight Gaussian predictor with an auxiliary BCE loss surprisingly outperforms nearly all baselines on most metrics across all three datasets, achieving the best or tied-best performance on 11/14 metrics in each dataset. Notably, we observe substantial gains on differential expression metrics, including DEOver, DEPrec, AUPRC, and AUROC. For example, AUPRC improves from 0.607 to 0.782 on PBMC and from 0.605 to 0.817 on Tahoe100M compared with the strongest baseline. Complete results are reported in Appendix C.1. Together, these results echo our observation of pronounced zero-boundary misallocation, showing that correcting this single failure mode alone can already yield substantial gains in perturbation prediction.

4.2.2 Correcting Existing Models

According to Table 1, our correction strategies can yield substantial gains across both perturbation prediction and probability misallocation metrics. Beyond the overall trend, the results reveal three recurring patterns.

Parameterization alone provides limited overall gains. We observe mixed results with very modest overall gains from introducing a hurdle parameterization without auxiliary supervision, suggesting that model parameterization alone is insufficient to explain misallocation.

Incidence supervision leads to marked benefits. On PBMC and Tahoe100M, incidence supervision substantially alleviates misallocation and improves the reliability of perturbation prediction. Across both STATE and Doloris*, we observe gains in differential expression and perturbation effect metrics, together with a sharp reduction in Zero KS. For STATE, Zero KS decreases from 0.035 to 0.007 on PBMC and from 0.038 to 0.008 on Tahoe100M, while AUPRC increases from 0.533 to 0.691 and from 0.528 to 0.784, respectively. Doloris* follows a similar overall trend.

Multi-threshold supervision also works, particularly for positive misallocation. Multi-threshold supervision also delivers considerable improvements, with many metrics matching or surpassing

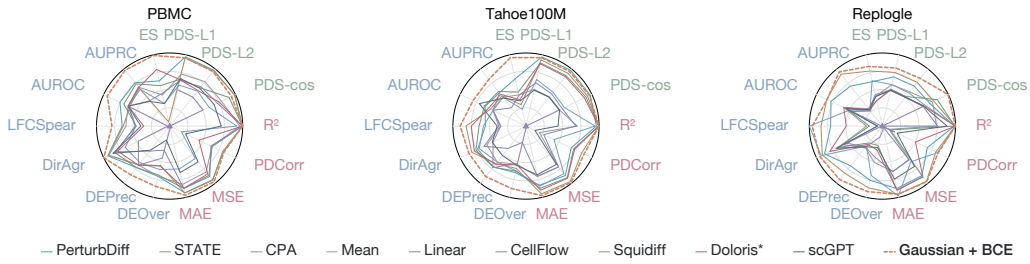


Figure 5: **Comparison of perturbation prediction methods across 14 metrics.** From left to right: PBMC, Tahoe100M, and Replogle. Larger radii indicate better performance.

those achieved by incidence supervision. This benefit is especially evident on Replogle, where misallocation extends more strongly into the positive region. While STATE + BCE reduces Zero KS, Positive KS increases from 0.312 to 0.373, with Overall KS worsening from 0.197 to 0.220. Multi-threshold supervision ($M = 9$) instead reduces Positive KS to 0.272 and Overall KS to 0.184, outperforming BCE on both measures. These distributional improvements are also reflected in superior perturbation prediction performance.

Table 1: **Comparison of correction strategies for Doloris* and STATE.** Darker blue indicates better performance. Shades mark the top three ranks including ties among variants.

Model	Expression Accuracy				Differential Expression						Perturbation Effects				Probability Misallocation		
	R^2 ↑	PDCorr ↑	MSE ↓	MAE ↓	DEOver ↑	DEPrec ↑	DirAgr ↑	LFCSpear ↑	AUROC ↑	AUPRC ↑	ES ↑	PDS_{L1} ↑	PDS_{L2} ↑	PDS_{cos} ↑	Zero KS ↓	Positive KS ↓	Overall KS ↓
PBMC																	
Doloris*	0.997	0.445	5.774e-04	0.009	0.449	0.547	0.631	0.283	0.501	0.542	0.512	0.651	0.621	0.756	0.482	0.705	0.482
Doloris* + BCE	0.991	0.535	4.303e-04	0.009	0.427	0.557	0.692	0.242	0.580	0.606	0.497	0.787	0.767	0.766	0.008	0.315	0.017
Doloris* + multi-threshold ($M = 4$)	0.999	0.533	5.407e-04	0.007	0.621	0.617	0.730	0.436	0.683	0.677	0.601	0.683	0.646	0.703	0.024	0.170	0.025
Doloris* + multi-threshold ($M = 9$)	0.999	0.528	5.436e-04	0.007	0.601	0.598	0.731	0.435	0.666	0.661	0.600	0.683	0.645	0.696	0.021	0.141	0.022
Tahoe100M																	
Doloris*	0.996	0.737	3.804e-04	0.008	0.436	0.534	0.784	0.599	0.568	0.555	0.296	0.875	0.865	0.888	0.442	0.656	0.443
Doloris* + BCE	0.928	0.505	9.574e-04	0.014	0.493	0.580	0.615	0.250	0.731	0.715	0.456	0.641	0.719	0.706	0.017	0.462	0.026
Doloris* + multi-threshold ($M = 4$)	0.997	0.742	3.648e-04	0.007	0.632	0.630	0.799	0.598	0.707	0.668	0.452	0.878	0.866	0.887	0.157	0.402	0.159
Doloris* + multi-threshold ($M = 9$)	0.997	0.742	3.665e-04	0.007	0.555	0.579	0.803	0.615	0.637	0.611	0.328	0.877	0.866	0.887	0.253	0.472	0.254
Replogle																	
Doloris*	0.998	0.250	1.257e-02	0.075	0.083	0.080	0.639	0.227	0.469	0.086	0.357	0.506	0.508	0.507	0.302	0.368	0.365
Doloris* + BCE	0.997	0.277	1.271e-02	0.075	0.106	0.147	0.652	0.336	0.597	0.189	0.410	0.513	0.509	0.500	0.100	0.389	0.217
Doloris* + multi-threshold ($M = 4$)	0.996	0.242	1.230e-02	0.073	0.091	0.078	0.634	0.222	0.590	0.159	0.357	0.506	0.505	0.503	0.119	0.307	0.209
Doloris* + multi-threshold ($M = 9$)	0.996	0.310	1.148e-02	0.071	0.110	0.110	0.652	0.281	0.620	0.194	0.366	0.506	0.504	0.501	0.109	0.295	0.199

Model	Expression Accuracy				Differential Expression						Perturbation Effects				Probability Misallocation		
	R^2 ↑	PDCorr ↑	MSE ↓	MAE ↓	DEOver ↑	DEPrec ↑	DirAgr ↑	LFCSpear ↑	AUROC ↑	AUPRC ↑	ES ↑	PDS_{L1} ↑	PDS_{L2} ↑	PDS_{cos} ↑	Zero KS ↓	Positive KS ↓	Overall KS ↓
PBMC																	
STATE	0.997	0.804	1.736e-04	0.006	0.443	0.539	0.739	0.508	0.484	0.533	-0.519	0.955	0.965	0.986	0.035	0.443	0.036
STATE + hurdle	0.997	0.786	1.739e-04	0.006	0.486	0.539	0.766	0.556	0.491	0.537	0.550	0.944	0.938	0.927	0.022	0.473	0.026
STATE + BCE	0.998	0.793	1.620e-04	0.006	0.575	0.604	0.748	0.479	0.677	0.691	0.833	0.945	0.947	0.945	0.007	0.385	0.016
STATE + multi-threshold ($M = 4$)	0.996	0.732	2.006e-04	0.006	0.489	0.550	0.730	0.441	0.521	0.554	-0.043	0.941	0.937	0.980	0.021	0.405	0.022
STATE + multi-threshold ($M = 9$)	0.998	0.811	1.418e-04	0.005	0.519	0.563	0.768	0.512	0.556	0.575	0.088	0.974	0.965	0.987	0.020	0.386	0.021
Tahoe100M																	
STATE	0.997	0.846	2.043e-04	0.005	0.476	0.526	0.824	0.657	0.537	0.528	0.240	0.929	0.923	0.937	0.038	0.562	0.040
STATE + hurdle	0.997	0.837	2.139e-04	0.005	0.459	0.491	0.845	0.723	0.429	0.462	0.530	0.918	0.911	0.928	0.021	0.506	0.025
STATE + BCE	0.997	0.830	2.133e-04	0.006	0.670	0.666	0.837	0.679	0.767	0.784	0.640	0.920	0.912	0.924	0.008	0.383	0.015
STATE + multi-threshold ($M = 4$)	0.997	0.844	1.992e-04	0.005	0.549	0.561	0.811	0.589	0.603	0.565	0.387	0.930	0.924	0.933	0.021	0.477	0.024
STATE + multi-threshold ($M = 9$)	0.996	0.842	2.123e-04	0.005	0.547	0.559	0.815	0.597	0.599	0.562	0.368	0.928	0.921	0.933	0.022	0.477	0.024
Replogle																	
STATE	0.998	0.405	7.467e-03	0.059	0.216	0.206	0.731	0.413	0.675	0.272	0.773	0.753	0.766	0.626	0.109	0.312	0.197
STATE + hurdle	0.998	0.368	9.089e-03	0.065	0.150	0.114	0.728	0.425	0.545	0.139	0.664	0.714	0.724	0.599	0.236	0.611	0.332
STATE + BCE	0.999	0.335	8.825e-03	0.065	0.197	0.196	0.710	0.376	0.645	0.254	0.698	0.715	0.722	0.616	0.095	0.373	0.220
STATE + multi-threshold ($M = 4$)	0.998	0.405	7.277e-03	0.058	0.221	0.207	0.739	0.426	0.685	0.291	0.774	0.766	0.778	0.654	0.093	0.281	0.186
STATE + multi-threshold ($M = 9$)	0.998	0.406	7.042e-03	0.057	0.213	0.199	0.738	0.424	0.690	0.293	0.770	0.769	0.781	0.661	0.092	0.272	0.184

5 Analysis

5.1 Why Do Existing Objectives Leave Misallocation?

Our observations demonstrate that probability misallocation persists across different training objectives. We combine analytical derivations with targeted diagnostics to examine why these objectives can still leave probability misallocation.

Mean Supervision. Single-cell expression differences can involve changes in the number of expression modes or in the proportions of cells occupying different expression states (Korthauer et al., 2016). Matching population means does not determine how expression is distributed across cells. We provide a simple analysis and counterexample in Appendix D.2.

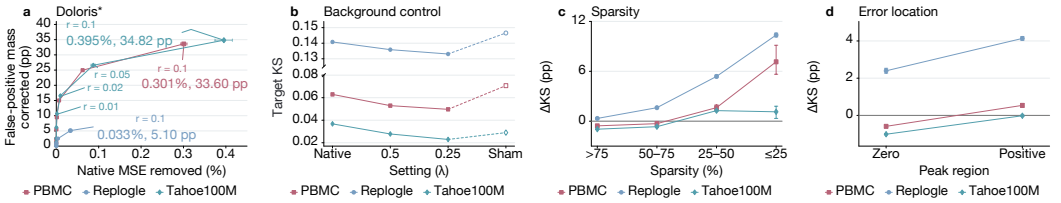


Figure 6: **Understanding and mitigating probability misallocation.** (a–b) Failure analysis of existing objectives. (a) In Doloris*, small positive predictions at zero entries have little MSE cost but create large false positive mass. (b) In STATE, downweighting background genes reduces KS on target genes. (c–d) $\Delta KS = 100(K_{\mathrm{BCE}} - K_{\mathrm{MT9}})$, with positive values favoring multi-threshold supervision. (c) Preference shifts to multi-threshold supervision as expression becomes denser. (d) Preference shifts to multi-threshold supervision when KS occurs at positive expression levels.

Generative Model Objectives. We find that the surrogate optimized by generative objectives can be poorly aligned with the generated probability distribution. As a result, small errors under the surrogate can produce large misallocation. We analyze a family of generative models spanning reconstruction, denoising, and likelihood objectives (Appendix D.3) and take Doloris* as a representative example. For Doloris*, this mismatch manifests at the zero boundary, where small positive errors incur small x_0-reconstruction MSE while creating substantial false-positive mass. We gradually expand the range of small positive predictions reset to zero. Across all three datasets, false-positive probability decreases much faster than MSE (Fig. 6a).

Distribution Matching. Distribution matching materially reduces misallocation, but small expression differences can carry little loss, and their contribution can be further diluted when the loss is computed jointly across many dimensions, allowing misallocation to remain. We formalize how these factors weaken distributional supervision for STATE's MMD (energy distance) objective (Adduri et al., 2026) in Appendix D.4. We test this by downweighting 1,800 genes in the distance computation while leaving the remaining 200 target genes unchanged. As illustrated in Fig. 6b, progressively reducing the weights of the background genes consistently lowers the KS discrepancy of the target genes across all three datasets.

5.2 Choosing the Right Probability Supervision

Despite the overall improvements from probability supervision, we observe a striking cross-dataset discrepancy. Incidence supervision substantially improves performance on PBMC and Tahoe100M, while multi-threshold supervision is markedly more effective on Replogle. We next ask what drives this preference and whether it reveals a broader principle for choosing supervision.

Sparsity. We note that Replogle is substantially denser, with a nonzero rate of 44.67%, compared with only 6.85% on PBMC and 5.26% on Tahoe100M. We therefore investigate analytically how sparsity affects the choice of supervision. Our analysis shows that, as expression becomes less sparse, positive-expression misallocation contributes more to overall misallocation (Appendix D.5), suggesting a stronger preference for multi-threshold supervision. We further validate this empirically by stratifying perturbation–gene pairs by sparsity within each dataset and comparing BCE with multi-threshold supervision ($M = 9$). As Fig. 6c shows, supervision preference shifts from BCE toward multi-threshold supervision as expression becomes denser. This is especially significant in the densest regime ($s \leq 25\%$), where multi-threshold supervision improves over BCE by 10.36 percentage points on Replogle and 7.14 percentage points on PBMC.

Dominant Misallocation Region. Beyond sparsity, we examine whether supervision preference also depends on where the model's misallocation is most severe. Building on Sec. 3.2.2, we group perturbation–gene pairs according to whether the largest probability discrepancy (KS) occurs at the zero boundary or at positive expression levels, and compare BCE with multi-threshold supervision ($M = 9$) separately. As shown in Fig. 6d, when KS occurs at the zero boundary, incidence supervision is relatively more effective; by contrast, KS at positive expression levels favors multi-threshold supervision ($M = 9$). On PBMC, the relative gain of multi-threshold supervision over BCE increases by 1.13 percentage points, while on Replogle it increases by 1.74 percentage points.

AI USE STATEMENT

We used generative AI tools to refine the clarity, grammar, and presentation of the manuscript. We also used these tools for retrieval and discovery of relevant literature. The authors carefully reviewed and verified all AI-assisted content. The scientific ideas, methodological designs, experimental results, and conclusions were developed and validated by the authors. We assume full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

Appendix A specifies the datasets, preprocessing, splits, and evaluation metrics. Appendix B details the model architectures, objectives, training, checkpoint selection, and sampling procedures. Appendix C reports the complete benchmark results and diagnostic protocols, and Appendix D provides the supporting derivations.

REFERENCES

Abhinav K. Adduri, Dhruv Gautam, Beatrice Bevilacqua, Mohsen Naghipourfar, Alishba Imran, Rohan Shah, Noam Teyssier, Rishi Verma, Christopher Carpenter, Basak Eraslan, Francis Chalissery, Rajesh Ilango, Vishvak Subramanyam, Chiara Ricci-Tam, Sanjay Nagaraj, Aidan Winters, Mingze Dong, Stefanie Fellinger, Adam Krejci, Tilmann Burckstummer, Sravya Tirukkovular, Jeremy Sullivan, Brian S. Plosky, 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. *Cell*, 189(19):5914–5931.e20, 2026. doi: 10.1016/j.cell.2026.07.052. URL https://doi.org/10.1016/j.cell.2026.07.052.

Constantin Ahlmann-Eltze, Wolfgang Huber, and Simon Anders. Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines. *Nature Methods*, 22(8): 1657–1661, 2025. doi: 10.1038/s41592-025-02772-6. URL https://doi.org/10.1038/s41592-025-02772-6.

Arc Institute. Tahoe-100M: Arc Virtual Cell Atlas, n.d. URL https://github.com/ArcInstitute/arc-virtual-cell-atlas/blob/main/tahoe-100M/README.md. Official dataset release. Accessed September 26, 2026.

Charlotte Bunne, Stefan G. Stark, Gabriele Gut, Jacobo Sarabia del Castillo, Mitch Levesque, Kjong-Van Lehmann, Lucas Pelkmans, Andreas Krause, and Gunnar Rätsch. Learning single-cell perturbation responses using neural optimal transport. *Nature Methods*, 20(11):1759–1768, 2023. doi: 10.1038/s41592-023-01969-x. URL https://doi.org/10.1038/s41592-023-01969-x.

Charlotte Bunne, Yusuf Roohani, Yanay Rosen, Ankit Gupta, Xikun Zhang, Marcel Roed, Theo Alexandrov, Mohammed AlQuraishi, Patricia Brennan, Daniel B. Burkhardt, Andrea Califano, Jonah Cool, Abby F. Dernburg, Kirsty Ewing, Emily B. Fox, Matthias Haury, Amy E. Herr, Eric Horvitz, Patrick D. Hsu, Viren Jain, Gregory R. Johnson, Thomas Kalil, David R. Kelley, Shana O. Kelley, Anna Kreshuk, Tim Mitchison, Stephani Otte, Jay Shendure, Nicholas J. Sofroniew, Fabian Theis, Christina V. Theodoris, Srigokul Upadhyayula, Marc Valer, Bo Wang, Eric Xing, Serena Yeung-Levy, Marinka Zitnik, Theofanis Karaletsos, Aviv Regev, Emma Lundberg, Jure Leskovec, and Stephen R. Quake. How to build the virtual cell with artificial intelligence: Priorities and opportunities. *Cell*, 187(25):7045–7063, 2024. doi: 10.1016/j.cell.2024.11.015. URL https://doi.org/10.1016/j.cell.2024.11.015.

Changxi Chi, Jun Xia, Yufei Huang, Zhuoli Ouyang, Cheng Tan, Yunfan Liu, Jingbo Zhou, Chang Yu, Liangyu Yuan, Siyuan Li, Zelin Zang, and Stan Z Li. Doloris: Dual conditional diffusion implicit bridges with sparsity masking strategy for unpaired single-cell perturbation estimation. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 52808–52823, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/568b6cc71889ea0b2aa74152ef9c28db-Paper-Conference.pdf.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21(8):1470–1480, 2024. doi: 10.1038/s41592-024-02201-0. URL https://doi.org/10.1038/s41592-024-02201-0.

Greg Finak, Andrew McDavid, Masanao Yajima, Jingyuan Deng, Vivian Gersuk, Alex K. Shalek, Chloe K. Slichter, Hannah W. Miller, M. Juliana McElrath, Martin Prlic, Peter S. Linsley, and Raphael Gottardo. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology*, 16 (1):278, 2015. doi: 10.1186/s13059-015-0844-5. URL https://doi.org/10.1186/s13059-015-0844-5.

Tilmann Gneiting and Adrian E Raftery. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL https://doi.org/10.1198/016214506000001437.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL https://jmlr.org/papers/v13/gretton12a.html.

Siyu He, Yuefei Zhu, Daniel Naveed Tavakol, Haotian Ye, Yeh-Hsing Lao, Zixian Zhu, Cong Xu, Shradha Chauhan, Guy Garty, Raju Tomer, Gordana Vunjak-Novakovic, James Zou, Elham Azizi, and Kam W. Leong. Squidiff: predicting cellular development and responses to perturbations using a diffusion model. *Nature Methods*, 23(1):65–77, 2026. doi: 10.1038/s41592-025-02877-y. URL https://doi.org/10.1038/s41592-025-02877-y.

Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *International Conference on Learning Representations*, 2014. URL https://arxiv.org/abs/1312.6114.

Dominik Klein, Jonas Simon Fleck, Daniil Bobrovskiy, Lea Zimmermann, Sören Becker, Alessandro Palma, Leander Dony, Alejandro Tejada-Lapuerta, Guillaume Huguet, Hsiu-Chuan Lin, Nadezhda Azbukina, Fátima Sanchís-Calleja, Theo Uscidda, Artur Szalata, Manuel Gander, Aviv Regev, Barbara Treutlein, J. Gray Camp, and Fabian J. Theis. CellFlow enables generative single-cell phenotype modeling with flow matching. *bioRxiv*, 2025. doi: 10.1101/2025.04.11.648220. URL https://doi.org/10.1101/2025.04.11.648220. Preprint.

Keegan D. Korthauer, Li-Fang Chu, Michael A. Newton, Yuan Li, James Thomson, Ron Stewart, and Christina Kendziorski. A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome Biology*, 17(1):222, 2016. doi: 10.1186/s13059-016-1077-y. URL https://doi.org/10.1186/s13059-016-1077-y.

Mohammad Lotfollahi, F. Alexander Wolf, and Fabian J. Theis. scGen predicts single-cell perturbation responses. *Nature Methods*, 16(8):715–721, 2019. doi: 10.1038/s41592-019-0494-8. URL https://doi.org/10.1038/s41592-019-0494-8.

Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Leon Hetzel, Yuge Ji, Ignacio L Ibarra, Sanjay R Srivatsan, Mohsen Naghipourfar, Riza M Daza, Beth Martin, Jay Shendure, Jose L McFaline-Figueroa, Pierre Boyeau, F Alexander Wolf, Nafissa Yakubova, Stephan Günnemann, Cole Trapnell, David Lopez-Paz, and Fabian J Theis. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology*, 19(6), 2023. doi: 10.15252/msb.202211517. URL https://doi.org/10.15252/msb.202211517.

Ajay Nadig, Joseph M. Replogle, Angela N. Pogson, Mukundh Murthy, Steven A. McCarroll, Jonathan S. Weissman, Elise B. Robinson, and Luke J. O'Connor. Transcriptome-wide analysis of differential expression in perturbation atlases. *Nature Genetics*, 57(5):1228–1237, 2025. doi: 10.1038/s41588-025-02169-3. URL https://doi.org/10.1038/s41588-025-02169-3.

Thomas M. Norman, Max A. Horlbeck, Joseph M. Replogle, Alex Y. Ge, Albert Xu, Marco Jost, Luke A. Gilbert, and Jonathan S. Weissman. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science*, 365(6455):786–793, 2019. doi: 10.1126/science.aax4438. URL https://doi.org/10.1126/science.aax4438.

Parse Biosciences. 10 Million Human PBMCs in a Single Experiment, n.d. URL https://www.parsebiosciences.com/datasets/10-million-human-pbmcs-in-a-single-experiment/. Official dataset. Accessed September 26, 2026.

Joseph M. Replogle, Reuben A. Saunders, Angela N. Pogson, Jeffrey A. Hussmann, Alexander Lenail, Alina Guna, Lauren Mascibroda, Eric J. Wagner, Karen Adelman, Gila Lithwick-Yanai, Nika Iremadze, Florian Oberstrass, Doron Lipson, Jessica L. Bonnar, Marco Jost, Thomas M. Norman, and Jonathan S. Weissman. Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. *Cell*, 185(14):2559–2575.e28, 2022. doi: 10.1016/j.cell.2022.05.013. URL https://doi.org/10.1016/j.cell.2022.05.013.

Alex K. Shalek, Rahul Satija, Joe Shuga, John J. Trombetta, Dave Gennert, Diana Lu, Peilin Chen, Rona S. Gertner, Jellert T. Gaublomme, Nir Yosef, Schraga Schwartz, Brian Fowler, Suzanne Weaver, Jing Wang, Xiaohui Wang, Ruihua Ding, Raktima Raychowdhury, Nir Friedman, Nir Hacohen, Hongkun Park, Andrew P. May, and Aviv Regev. Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature*, 510(7505):363–369, 2014. doi: 10.1038/nature13437. URL https://doi.org/10.1038/nature13437.

Jordan W. Squair, Matthieu Gautier, Claudia Kathe, Mark A. Anderson, Nicholas D. James, Thomas H. Hutson, Rémi Hudelle, Taha Qaiser, Kaya J. E. Matson, Quentin Barraud, Ariel J. Levine, Gioele La Manno, Michael A. Skinnider, and Grégoire Courtine. Confronting false discoveries in single-cell differential expression. *Nature Communications*, 12 (1):5692, 2021. doi: 10.1038/s41467-021-25960-2. URL https://doi.org/10.1038/s41467-021-25960-2.

Gábor J. Székely and Maria L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013. doi: 10.1016/j.jspi.2013.03.018. URL https://doi.org/10.1016/j.jspi.2013.03.018.

Xinyu Yuan, Xixian Liu, Ya Shi Zhang, Zuobai Zhang, Hongyu Guo, and Jian Tang. PerturbDiff: Functional Diffusion for Single-Cell Perturbation Modeling. *arXiv preprint arXiv:2602.19685*, 2026. URL https://arxiv.org/abs/2602.19685.

Jesse Zhang, Airol A. Ubas, Valentine Svensson, Nicole Thomas, Vishvak Subramanyam, Christopher Carpenter, Aidan Winters, Richard de Borja, Noam Teyssier, Neha Thakar, Umair Khan, Pooja V. Akella, Harper J. Green, John D. Thompson, Vuong Tran, Joseph Pangallo, Efthymia Papalexi, Ajay Sapre, Hoai Nguyen, Oliver Sanderson, Maria Nigos, Olivia Kaplan, Sarah Schroeder, Bryan Hariadi, Simone Marrujo, Crina Curca, Alec Salvino, Guillermo Gallareta Olivares, Ryan Koehler, Alexander Rosenberg, Charles Roco, Chiara Ricci-Tam, Matthew G. Jones, Alexander Dobin, Brian S. Plosky, Danielle E. Lyons, Daniele Merico, Nima Alidoust, Hani Goodarzi, and Johnny Yu. Tahoe-100M: Mapping drug-induced molecular phenotypes at single-cell resolution. *Cell*, 189(19):5945–5961.e8, 2026. doi: 10.1016/j.cell.2026.08.035. URL https://doi.org/10.1016/j.cell.2026.08.035.