# PROVE: TRAINING-FREE REFERENCE-GUIDED PARETO-CONSENSUS VELOCITY EDITING FOR CROSS-CONTEXT SINGLE-CELL PERTURBATION PREDICTION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Flow-matching models learn condition-dependent velocity fields from observed context-perturbation pairs. In cross-context covariate transfer, however, the target pair is unobserved, and the training objective provides no direct supervision for its response prediction. The resulting conditional extrapolation may collapse toward the control state, align with another perturbation, or drift beyond the cross-context variation supported by available data. Although responses from related contexts do not uniquely determine the missing target distribution, they provide useful reference constraints for detecting these failures. Motivated by this, we introduce **Pa**ra**R**eto-C**O**nsensus **V**elocity **E**diting (PROVE), a training-free inference-time method for correcting reference-inconsistent predictions from frozen flow-matching models. At each sampling step, PROVE obtains a final-state estimate from the backbone velocity and diagnoses three patterns including effect collapse, perturbation confusion, and context drift. It constructs a local correction for each active pattern, combines the frozen and corrective directions using an MGDA-inspired minimum-norm consensus, and redirects the velocity without changing its instantaneous norm. Experiments across cross-cell-line and cross-state transfer settings show that PROVE improves the original samplers across multiple backbones and strengthens the recovery of perturbation-specific differential-expression and pathway-level effects.

## 1 INTRODUCTION

Single-cell perturbation screens enable systematic studies of gene function Dixit et al. (2016), drug response Srivatsan et al. (2020), and therapeutic targets Yang et al. (2024); Morris et al. (2023). As profiling every perturbation across all biological contexts is infeasible, computational models are used to predict responses in contexts where a perturbation has not been measured. We consider cross-context transfer settings in which post-perturbation cells are unavailable for the target perturbation in the context of interest, while responses to the same perturbation are observed in at least one other context. Controls and responses to other perturbations may also be available in the context of interest.

A broad range of single-cell perturbation prediction methods has been developed, including latent response translation Lotfollahi et al. (2019), compositional autoencoders Lotfollahi et al. (2023), graph-based perturbation models Roohani et al. (2024), and optimal transport methods Bunne et al. (2023). More recently, flow-matching models have used condition-dependent vector fields to generate post-perturbation cell distributions Yu et al. (2026); Klein et al. (2025). As the model does not observe post-perturbation cells for the target context-perturbation combination during training, it must predict the response by extrapolating from the conditions it has observed. The available reference data cannot reveal the exact target response, but they can still constrain extrapolation by identifying predictions that conflict with perturbation and contextual structures observed elsewhere.

We instantiate this reference-constrained view through three diagnosable forms of inconsistency in the predicted response:



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.

- **Effect collapse** occurs when the predicted perturbation effect is substantially weaker in the target context than in the reference contexts, approaching the target-context control.
- **Perturbation confusion** occurs when the predicted response aligns more strongly with a competing perturbation than with the target perturbation.
- **Context drift** occurs when the predicted response deviates beyond the cross-context variation observed for the target perturbation.

Previous works reported similar problems involving attenuated perturbation effects, loss of perturbation specificity, and unsupported variation across biological contexts Ahlmann-Eltze et al. (2025); Viñas Torné et al. (2025). We therefore treat these patterns as soft reference-derived diagnostics rather than universal biological constraints, because a perturbation may have different effects across contexts. PROVE scales each correction by violation severity instead of enforcing a hard constraint. Fig. 1 illustrates the three prediction patterns and their reduction under PROVE on Replogle using scDFM.

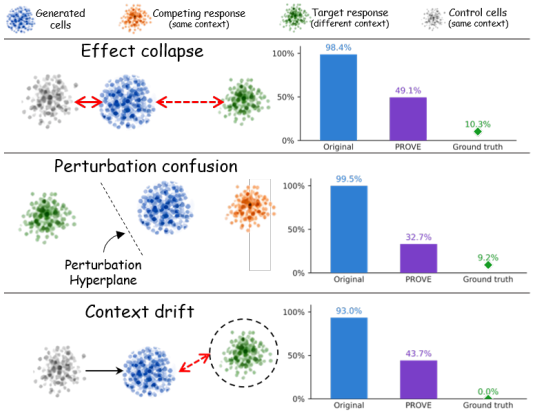


Figure 1: Reference-diagnosable prediction patterns. PROVE reduces their occurrence rates relative to the original.

Motivated by this view, we propose **P**a**R**eto-C**O**nsensus **V**elocity **E**diting (PROVE), a training-free inference-time method for pretrained flow-matching models of perturbation responses. Rather than replacing the backbone prediction with a response transferred directly from the reference contexts, PROVE evaluates at each ODE step whether the response estimate induced by the frozen velocity exhibits effect collapse, perturbation confusion, or context drift. If no pattern is detected, PROVE retains the frozen prediction. Otherwise, each detected pattern produces a local correction direction. Since these directions may conflict with one another or with the frozen transport direction, PROVE reconciles them through minimum-norm Pareto consensus. It then adjusts the response estimate and redirects the frozen velocity while preserving its instantaneous norm. Our contributions are summarized as follows:

- We formulate cross-context perturbation prediction under covariate transfer as reference-constrained conditional extrapolation, where available references do not uniquely determine the missing target response but can diagnose reference-inconsistent predictions.
- We introduce PROVE, a training-free inference-time method that selectively corrects response estimates exhibiting these patterns, reconciles potentially conflicting directions through minimum-norm Pareto consensus, and redirects the frozen velocity while preserving its instantaneous norm.
- We evaluate PROVE across three datasets and three flow-matching backbones, demonstrating consistent improvements in perturbation-specific prediction and biological-signal recovery.

## 2 Related Work

### 2.1 Single-cell Perturbation Prediction

Single-cell perturbation prediction aims to infer cellular response distributions under genetic or chemical interventions. Early approaches modeled perturbation effects through latent response translation or compositional factorization, as exemplified by scGen Lotfollahi et al. (2019) and CPA Lotfollahi et al. (2023). Subsequent methods incorporated biological priors or explicit distributional transport, including the gene-interaction-based GEARS Roohani et al. (2024) and the optimal-transport-based CellOT Bunne et al. (2023). More recently, flow-matching methods such as CellFlow Klein et al. (2025), scDFM Yu et al. (2026), and PRiMeFlow Yan et al. (2026) have learned conditional transport toward perturbation-response distributions. STRAND Fu et al. (2026)

and conditional Monge-gap objectives Driessen et al. (2026) further extend this direction to unseen perturbations and cross-context generalization. These methods primarily improve the predictive model during training. PROVE instead assumes a pretrained flow-matching model and performs correction on its frozen velocity field at inference time.

### 2.2 Challenges in Cross-Context Perturbation Prediction

Generalization to perturbation responses not observed in the target context remains difficult as perturbation effects vary across biological contexts. Recent benchmarks Ahlmann-Eltze et al. (2025) show that deep perturbation-prediction models do not always outperform simple linear baselines on unobserved responses. Systema Viñas Torné et al. (2025) further shows that standard control-referenced metrics can be dominated by systematic variation and proposes evaluation criteria that emphasize perturbation-specific effects and discrimination among perturbations. For conditional generative flows, cross-context prediction requires the learned vector field to extrapolate from observed condition pairs to a target combination without direct post-perturbation supervision Rubbi et al. (2026). PROVE addresses this inference-time extrapolation problem by using available responses from related contexts and perturbations to identify effect collapse, perturbation confusion, and context drift, and to softly redirect predictions away from the corresponding reference-inconsistent regions.

### 2.3 Inference-Time Guidance for Flow Matching

Inference-time guidance modifies the sampling dynamics of pretrained generative models without retraining the backbone. Classical classifier guidance and classifier-free guidance steer generation using an auxiliary classifier or conditional–unconditional predictions Dhariwal & Nichol (2021); Ho & Salimans (2022). Recent work has extended guidance directly to flow matching. Feng et al. (2025) develop a general framework that includes training-free guidance for flow models, while Wang et al. (2025) formulate training-free guided flow matching through optimal control.

Our setting differs primarily in the source and structure of the guidance signal. In cross-context perturbation prediction, post-perturbation cells from the target context–perturbation pair are unavailable, so the missing pair itself provides no direct response-level supervision. PROVE instead constructs guidance from available biological reference populations, using them to quantify reference-relative effect collapse, perturbation confusion, and context drift. Since the resulting correction directions may conflict with one another or with the frozen transport direction, PROVE reconciles them through an MGDA-inspired minimum-norm consensus before redirecting the frozen velocity.

## 3 Method

PROVE retains the frozen sampler as the primary predictor and uses available reference responses only to selectively correct reference-inconsistent extrapolations. It performs four operations at each ODE step: sampler-induced response estimation, reference-based pattern diagnosis, Pareto-consensus direction aggregation, and norm-preserving velocity redirection. Fig. 2 provides an overview, while Algorithm 1 summarizes the inference update.

### 3.1 Problem Setting and Response Estimate

We consider cross-context perturbation prediction under covariate transfer. Let $\mathcal{C}$ denote the set of biological contexts and $\mathcal{G}$ the set of perturbations. For a target pair $(c^\star, g) \in \mathcal{C} \times \mathcal{G}$, where $c^\star$ is the context of interest and $g$ is the target perturbation, all corresponding post-perturbation cells associated with the target combination are excluded from model training and reference construction. The available references may include controls and responses to other perturbations from $c^\star$, together with responses to $g$ from contexts $c' \neq c^\star$.

At each ODE step, PROVE first forms a response estimate from the velocity predicted by the frozen backbone. Let $x_t \in \mathcal{X}$ denote the current state, $x_{\mathrm{con}}$ the control cell supplied to the model, and $v_\theta$ the frozen velocity field. We define:

$$\tilde{x}_1 = x_t + (1-t)v_t, \;\; z_t = \phi(\tilde{x}_1) - \mu_{\mathrm{ctrl}}(c^\star), \tag{1}$$

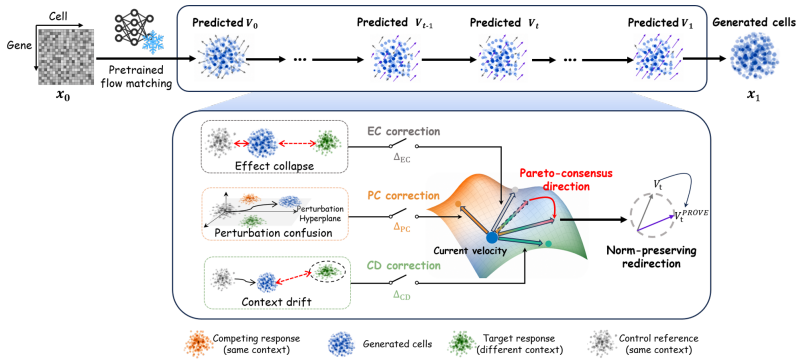


Figure 2: Overview of PROVE. PROVE consists of velocity-induced response estimation, response-pattern correction, Pareto consensus, and norm-preserving velocity redirection.

where $v_t = v_\theta(x_t, t, x_{\text{con}}, g)$ is the frozen-backbone velocity, $\tilde{x}_1$ is the linear final-state estimate induced by $v_t$, and $\phi$ maps backbone states to the response space used for reference construction and diagnosis. $z_t$ is its control-centered response representation. PROVE uses $\tilde{x}_1$ for pattern diagnosis, while the numerical solver advances $x_t$ using the resulting velocity.

While $z_t$ represents an individual generated cell, $p_g$ and $\sigma_g$ summarize context-level mean responses. We therefore treat them as population-informed soft priors rather than calibrated single-cell validity regions. In particular, $\sigma_g$ does not capture within-context cellular variability.

### 3.2 Reference Construction

Although the target response is unobserved, responses from related contexts and perturbations still reveal the typical effect of the target perturbation and its variation across contexts. We summarize these observations into reference statistics that characterize the target-perturbation direction, effect scale, cross-context dispersion, and competing perturbation directions.

For any biological context $c$ and perturbation $q$, let $\bar{\eta}_{c,q}$ denote the mean control-centered response to $q$ in $c$. We use this notation for both the target perturbation $g$ and competing perturbations $h \neq g$.

#### 3.2.1 Target-Perturbation References

Responses to $g$ in non-target contexts provide information about its typical effect and its variation across contexts. We collect one context-level response vector from each available non-target context as $\mathcal{R}_g = \{\bar{\eta}_{c',g} : c' \in \mathcal{C} \setminus \{c^\star\},\ g \text{ is observed in } c'\}$, where each element of $\mathcal{R}_g$ summarizes the response to the target perturbation in one non-target context. We summarize these responses by their mean, normalized direction, and cross-context dispersion:

$$p_g = \frac{1}{|\mathcal{R}_g|} \sum_{r \in \mathcal{R}_g} r, \quad u_g = \frac{p_g}{\|p_g\|_2}, \qquad \sigma_g = \left( \frac{1}{|\mathcal{R}_g|} \sum_{r \in \mathcal{R}_g} \|r - p_g\|_2^2 \right)^{1/2}, \tag{2}$$

where $p_g$ is the equally weighted mean of the available context-level responses, $u_g$ is its normalized direction, and $\sigma_g$ is their root-mean-square distance from $p_g$. The radius $\sigma_g$ is an average dispersion statistic, not a support boundary, confidence region, or calibrated prediction radius. It need not contain every observed reference mean.

#### 3.2.2 Competing-Perturbation References

Responses to non-target perturbations in the target context provide competing directions for detecting perturbation confusion. For each perturbation $h \in \mathcal{G} \setminus \{g\}$ observed in $c^\star$, the normalized

direction is defined as:

$$u_h = \frac{\bar{\eta}_{c^\star,h}}{\|\bar{\eta}_{c^\star,h}\|_2}, \tag{3}$$

where $\bar{\eta}_{c^\star,h}$ is the mean control-centered response to $h$ in the target context. $\mathcal{H}_g$ contains the perturbations $h \neq g$ for which $\|\bar{\eta}_{c^\star,h}\|_2 > \epsilon$ and $\|u_h - u_g\|_2 > \epsilon$.

#### 3.2.3 Reference Roles and Availability

Target-context controls define the origin of the control-centered response representation. The cross-context target responses determine the reference direction $u_g$, effect scale $\|p_g\|_2$, and dispersion $\sigma_g$, while the target-context non-target responses determine the competing directions $\{u_h\}_{h\in\mathcal{H}_g}$. Effect collapse is evaluated when $u_g$ is available, perturbation confusion additionally requires $\mathcal{H}_g \neq \emptyset$, and context drift requires at least two target-perturbation reference contexts with $\sigma_g > \epsilon$. Detailed cell-level definitions are provided in the Appendix.

### 3.3 Reference-Diagnosable Response Patterns

Using these reference statistics, PROVE computes three nonnegative, reference-relative discrepancy scores. Effect collapse measures a shortfall in signed projection onto $u_g$; perturbation confusion measures greater alignment with an available competing direction; and context drift measures distance from $p_g$ beyond the reference RMS dispersion. These scores define soft preferences for inference, not tests that distinguish biologically valid from invalid responses. These scores quantify disagreement with the reference priors rather than necessarily indicating prediction errors:

$$\begin{aligned}
\ell_{\mathrm{EC}}(z_t) &= \left[\beta\|p_g\|_2 - u_g^\top z_t\right]_+, \\
\ell_{\mathrm{PC}}(z_t) &= \left[u_{h_t^\star}^\top z_t - u_g^\top z_t\right]_+, \\
\ell_{\mathrm{CD}}(z_t) &= \left[\|z_t - p_g\|_2 - \sigma_g\right]_+,
\end{aligned} \tag{4}$$

where $\ell_{\mathrm{EC}}(z_t)$, $\ell_{\mathrm{PC}}(z_t)$, and $\ell_{\mathrm{CD}}(z_t)$ denote the respective violation scores at $z_t$, which determine pattern activation; correction magnitudes are defined below. Moreover, $h_t^\star = \arg\max_{h\in\mathcal{H}_g} u_h^\top z_t$ is the competing perturbation most strongly aligned with $z_t$, $[a]_+ = \max(a, 0)$, and $\beta \in (0, 1)$ specifies the minimum retained fraction of the reference effect magnitude. Since each $u_h$ is unit-normalized, selecting $h_t^\star$ is equivalent to maximizing cosine similarity whenever $\|z_t\|_2 > 0$. Geometrically, effect collapse moves the estimate forward along $u_g$, perturbation confusion moves it toward $u_g$ and away from $u_{h_t^\star}$, and context drift moves an estimate outside the reference ball back toward its boundary around $p_g$.

For each active pattern, we construct the minimum-distance correction associated with its active constraint. Let $A_t = \{k \in \{\mathrm{EC}, \mathrm{PC}, \mathrm{CD}\} : \ell_k(z_t) > \epsilon\}$ denote the set of active patterns. For each $k \in A_t$, the corresponding minimum-distance correction $\Delta_k$ is defined as:

$$\Delta_k = \Pi_k(z_t) - z_t, \;\; m_k = \|\Delta_k\|_2, \;\; d_k = \frac{\Delta_k}{m_k}, \tag{5}$$

where $m_k$ and $d_k$ are the magnitude and unit direction of the correction. For EC and CD, $\Pi_k$ is the Euclidean projection onto the corresponding feasible region. For PC, it projects onto the half-space induced by the active competitor $h_t^\star$. Closed-form corrections are given in Eqs. equation 25–equation 27 in the Appendix.

### 3.4 Pareto-Consensus Direction Aggregation

PROVE retains the remaining displacement proposed by the frozen sampler as one of the directions entering the consensus problem. We define:

$$\delta_t = \phi(\tilde{x}_1) - \phi(x_t), \;\; d_{\mathrm{frz}} = \frac{\delta_t}{\|\delta_t\|_2}, \tag{6}$$

where $\delta_t$ is the remaining response-space displacement induced by the frozen velocity and $d_{\mathrm{frz}}$ is its normalized direction. If $A_t = \emptyset$ or $\|\delta_t\|_2 \leq \epsilon$, PROVE retains the frozen velocity.

**Algorithm 1** PROVE update at one ODE step

**Require:** Current state $x_t$, time $t$, target perturbation $g$, frozen field $v_\theta$, and precomputed references
**Ensure:** Updated velocity $v_t^{\mathrm{PROVE}}$
1: Estimate the frozen response $(\tilde{x}_1, z_t)$ using Eq. equation 1
2: Diagnose the three response patterns using Eq. equation 4 and form the active set $A_t$
3: **if** $A_t = \emptyset$ **then return** $v_t$
4: Construct the correction directions $\{d_k\}_{k \in A_t}$ using Eq. equation 5
5: Obtain the frozen direction $d_{\mathrm{frz}}$ using Eq. equation 6
6: Reconcile $d_{\mathrm{frz}}$ and corrective directions using Eq. equation 7
7: **if** $a_t \leq \epsilon$ **then return** $v_t$
8: Scale the consensus edit using Eq. equation 8
9: Map and clip the edit using Eq. equation 9
10: Redirect the velocity using Eq. equation 10
11: **return** $v_t^{\mathrm{PROVE}}$

The frozen transport direction and the active correction directions may disagree. PROVE therefore computes the minimum-norm point in their convex hull, following the multi-objective formulation of Sener & Koltun (2018):

$$\bar{d}_t = \arg \min_{d \in \mathrm{conv}(\{d_{\mathrm{frz}}\} \cup \{d_k : k \in A_t\})} \|d\|_2^2, \tag{7}$$

where $\bar{d}_t$ is the minimum-norm convex combination of the frozen transport direction and the active correction directions, $a_t = \|\bar{d}_t\|_2$ measures their directional agreement, and $d_{\mathrm{mix}} = \bar{d}_t / a_t$ is the unit consensus direction. When the input directions agree, their convex hull remains far from the origin and $a_t$ is large. When they conflict, the convex hull approaches the origin and $a_t$ becomes small, which automatically weakens the subsequent edit. Proposition 3 further shows that every nonzero consensus direction is positively aligned with all input directions.

### 3.5 Norm-Preserving Velocity Redirection

The Pareto-consensus direction specifies where the remaining transport should be redirected, while the edit strength is determined by the guidance schedule, pattern severity, and directional agreement:

$$\gamma_t = \rho_t a_t \min \left\{ 1, \frac{\max_{k \in A_t} m_k}{\max(\|p_g\|_2, \epsilon)} \right\}, \tag{8}$$

where $\rho_t$ denotes the time schedule used to control when guidance is applied, $\max_{k \in A_t} m_k$ measures the strongest active violation relative to the target-perturbation scale, and $a_t$ weakens the edit when the proposed directions conflict.

To redirect the frozen response-space displacement $\delta_t$, we retain its length but replace its direction with the consensus direction. The resulting change is mapped back to the state space and clipped:

$$\tilde{c}_t = M_\phi \left( \|\delta_t\|_2 d_{\mathrm{mix}} - \delta_t \right), \;\; c_t = \tilde{c}_t \min \left\{ 1, \frac{\|\tilde{x}_1 - x_t\|_2}{\|\tilde{c}_t\|_2 + \epsilon} \right\}. \tag{9}$$

where $M_\phi$ is the associated displacement map from the response space back to the backbone state space. For scDFM, $\phi$ maps model states to standardized gene expression, and $M_\phi$ applies the corresponding inverse feature scaling. For CellFlow and CFM-GP, $\phi$ maps model states to their preprocessed gene-expression spaces, and $M_\phi$ is the identity map. The term $\|\delta_t\|_2 d_{\mathrm{mix}} - \delta_t$ is the change required to replace the frozen displacement by an equal-length displacement in the consensus direction. The clipping bounds the state-space correction by the remaining displacement proposed by the frozen sampler.

Finally, we define the corrected endpoint $\hat{x}_1 = \tilde{x}_1 + \gamma_t c_t$ and redirect the velocity toward it:

$$v_t^{\mathrm{PROVE}} = \|v_t\|_2 \frac{\hat{x}_1 - x_t}{\max\left(\|\hat{x}_1 - x_t\|_2, \epsilon\right)}, \tag{10}$$

Whenever the corrected displacement is nondegenerate, $\|v_t^{\mathrm{PROVE}}\|_2 = \|v_t\|_2$. Thus, PROVE changes the local transport direction while preserving the velocity scale of the frozen sampler.

Table 1: Comparison between vanilla models and their PROVE-enhanced variants on three datasets. ↓ indicates lower is better, while ↑ indicates higher is better. Boldface indicates the better result.

| | Replogle | | | | | | Frangieh21 | | | | | | Srivatsan20 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | scDFM | | CellFlow | | CFM-GP | | scDFM | | CellFlow | | CFM-GP | | scDFM | | CellFlow | | CFM-GP | |
| **Metric** | *Vanilla* | *+PROVE* | *Vanilla* | *+PROVE* | *Vanilla* | *+PROVE* | *Vanilla* | *+PROVE* | *Vanilla* | *+PROVE* | *Vanilla* | *+PROVE* | *Vanilla* | *+PROVE* | *Vanilla* | *+PROVE* | *Vanilla* | *+PROVE* |
| MAE ↓ | 0.1936 | **0.1274** | 0.0705 | **0.0686** | 0.0690 | **0.0649** | 0.1285 | **0.1014** | **0.0184** | 0.0224 | 0.0251 | **0.0225** | 0.1092 | **0.0598** | 0.0137 | **0.0128** | 0.0164 | **0.0147** |
| Pearson-Δ ↑ | 0.0643 | **0.2994** | 0.0320 | **0.2918** | 0.2675 | **0.3183** | -0.5337 | **0.4637** | 0.8320 | **0.8583** | 0.7983 | **0.8037** | -0.6440 | **0.8059** | 0.8637 | **0.8772** | 0.8228 | **0.8420** |
| DE-Spearman ↑ | 0.0316 | **0.3759** | 0.0185 | **0.5707** | 0.3007 | **0.4119** | -0.5618 | **0.5967** | 0.9232 | **0.9281** | 0.9071 | **0.9109** | -0.7094 | **0.8322** | 0.8475 | **0.8754** | 0.8156 | **0.8495** |
| DS ↑ | 0.5013 | **0.6199** | 0.5082 | **0.6253** | 0.5950 | **0.6411** | 0.5269 | **0.6272** | 0.5173 | **0.5878** | 0.5372 | **0.6008** | 0.2052 | **0.8519** | 0.8432 | **0.8597** | 0.8422 | **0.8610** |
| MMD-GEX ↓ | 0.1019 | **0.0620** | 0.0126 | **0.0076** | **0.0094** | 0.0112 | 0.1050 | **0.0805** | 0.0036 | **0.0015** | 0.0057 | **0.0046** | 0.1276 | **0.1219** | 0.0119 | **0.0100** | 0.1510 | **0.1240** |
| MMD-PCA ↓ | 0.2206 | **0.1681** | 0.0458 | **0.0271** | 0.0343 | **0.0330** | 0.2712 | **0.1643** | 0.0111 | **0.0091** | 0.0227 | **0.0187** | 0.2975 | **0.2693** | **0.0992** | 0.1251 | 0.0722 | **0.0642** |
| AUPRC-DE ↑ | 0.1985 | **0.3005** | 0.1511 | **0.2975** | 0.2294 | **0.3152** | 0.5022 | **0.5198** | 0.6761 | **0.7260** | 0.6175 | **0.6287** | 0.5314 | **0.7242** | 0.6507 | **0.6690** | 0.5936 | **0.6206** |
| DEG Recall ↑ | 0.6598 | **0.7130** | 0.0030 | **0.2953** | 0.1732 | **0.3430** | 0.2362 | **0.7007** | 0.6752 | **0.7021** | 0.6368 | **0.6915** | 0.2799 | **0.5762** | 0.6974 | **0.7523** | 0.6666 | **0.7412** |

# 4 EXPERIMENTS

We evaluate PROVE by answering the following questions: (**Q1**) *Does PROVE improve performance across backbones and datasets?* (**Q2**) *Does PROVE reduce the detected response patterns in the final generated responses?* (**Q3**) *Which components contribute to PROVE's performance?* (**Q4**) *Could PROVE be extended to diffusion-based models?* (**Q5**) *Does PROVE recover perturbation-specific biological signals?* We also validate the parameter sensitivity in the Appendix.

## 4.1 EXPERIMENTAL SETUP

### 4.1.1 BASELINES AND DATASETS

As PROVE is an inference-time velocity-editing method, we evaluate it as a plug-in module for several flow-matching-based methods. Specifically, we compare the original samplers of CellFlow Klein et al. (2025), scDFM Yu et al. (2026), and CFM-GP Abir et al. (2025) with their PROVE-augmented counterparts.

We evaluate PROVE in three cross-context transfer settings, each of which excludes all post-perturbation cells for a target context-perturbation pair while retaining responses to the same perturbation in other contexts. For Replogle Replogle et al. (2022), we evaluate cross-cell-line transfer following the STATE split Adduri et al. (2025). Specifically, we exclude the HepG2 responses to 638 perturbations that are observed in other cell lines, while using the remaining cells for training. We additionally use two covariate-transfer benchmarks from PerturBench Wu et al. (2024): Srivatsan20 Srivatsan et al. (2020), where the context variable is the cell line and the task involves drug-perturbation transfer, and Frangieh21 Frangieh et al. (2021), where the context variable is the biological state and the task involves genetic-perturbation transfer. For both PerturBench datasets, we follow the official train/test splits.

### 4.1.2 IMPLEMENTATION DETAILS

We follow the official implementations and training protocols of scDFM, CellFlow, and CFM-GP and keep all backbones frozen during PROVE inference. All post-perturbation cells from the held-out target pairs are excluded from backbone training, reference construction, and inference-time guidance. Hyperparameter selection and the use of evaluation labels are documented in Appendix.

We evaluate predictions from three complementary perspectives, following Yan et al. (2026). Global expression and distributional fidelity are measured by *MAE, MMD-GEX*, and *MMD-PCA*. Perturbation-specific response recovery is assessed using *Pearson-*$\Delta$, *DE-Spearman*, and *discriminability score (DS)*, while differential-expression recovery is evaluated by *AUPRC-DE* and *DEG Recall*.

## 4.2 MAIN RESULTS

### 4.2.1 COMPARISON WITH BASELINES (Q1)

Table 1 shows consistent gains in perturbation-specific recovery. Pearson-$\Delta$, DE-Spearman, DS, AUPRC-DE, and DEG Recall improve for every backbone–dataset pair across all nine evaluation

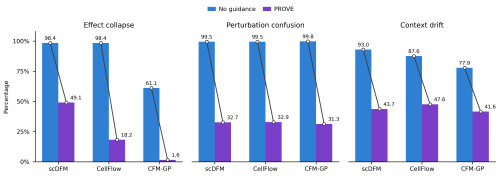


Figure 3: Final response-pattern rates on Replogle. Each bar reports the percentage of predictions exhibiting inconsistency after completion of the sampling trajectory.

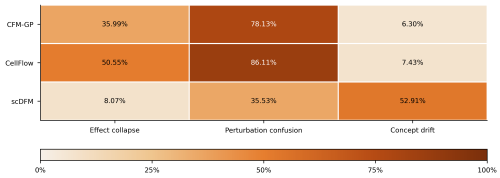


Figure 4: Pattern activation frequencies during sampling on Replogle. Each cell reports the percentage of sample ODE-step pairs for which the corresponding pattern is detected.

Table 2: Ablation study of PROVE. EC, PC, and CD denote the corrections associated with effect collapse, perturbation confusion, and context drift, respectively.

| **Metric** | *w/o PC* | *w/o EC* | *w/o Norm* | *w/o CD* | *w/o Pareto* | ***PROVE*** |
|---|---|---|---|---|---|---|
| MAE ↓ | 0.0688 | 0.0693 | 0.0703 | 0.0710 | 0.0712 | **0.0686** |
| Pearson-Δ ↑ | 0.2196 | 0.1971 | 0.2131 | 0.1936 | 0.2053 | **0.2918** |
| DE-Spearman ↑ | 0.2395 | 0.3356 | 0.2990 | 0.2943 | 0.3202 | **0.5707** |
| DS ↑ | 0.5725 | 0.5931 | 0.5957 | 0.5679 | 0.5899 | **0.6253** |
| MMD-GEX ↓ | 0.0172 | 0.0125 | 0.0128 | 0.0136 | 0.0122 | **0.0076** |
| MMD-PCA ↓ | 0.0360 | 0.0458 | 0.0446 | 0.0443 | 0.0450 | **0.0271** |
| AUPRC-DE ↑ | 0.2960 | 0.2855 | 0.2917 | 0.2915 | 0.2944 | **0.2975** |
| DEG Recall ↑ | 0.2479 | 0.2616 | 0.2904 | 0.2143 | 0.2781 | **0.2953** |

Table 3: Extension of PROVE to PerturbDiff and Squidiff on Replogle.

| | **PerturbDiff** | | **Squidiff** | |
|---|---|---|---|---|
| **Metric** | *Vanilla* | *+PROVE* | *Vanilla* | *+PROVE* |
| MAE ↓ | 0.0598 | **0.0497** | 0.5451 | **0.1417** |
| Pearson-Δ ↑ | 0.3265 | **0.3499** | -0.0012 | **0.3291** |
| DE-Spearman ↑ | 0.6067 | **0.7366** | 0.0824 | **0.5744** |
| DS ↑ | 0.6099 | **0.6919** | 0.5073 | **0.6574** |
| MMD-GEX ↓ | 0.0562 | **0.0465** | 0.2096 | **0.0718** |
| MMD-PCA ↓ | 0.2045 | **0.1716** | 0.2685 | **0.1283** |
| AUPRC-DE ↑ | 0.1939 | **0.2650** | 0.2351 | **0.3011** |
| DEG Recall ↑ | 0.6976 | **0.7256** | 0.1494 | **0.2360** |

settings. The largest gains occur when the frozen sampler performs poorly. For scDFM on Frangieh21 and Srivatsan20, PROVE changes negative Pearson-Δ and DE-Spearman correlations into positive ones. MAE, MMD-GEX, and MMD-PCA each improve in eight of the nine settings. Overall, PROVE consistently strengthens perturbation-specific and differential-expression recovery while generally preserving or improving global expression and distributional fidelity.

### 4.2.2 Response-Pattern Reduction Analysis (Q2)

To verify whether step-wise edits persist across completed trajectories, we activation of the reference-relative scores after sampling and during integration. As shown in Fig. 3, PROVE consistently reduces all three pattern rates across backbones: relative rates of the overall mean occurrence lowering from $90.6\%$ to $33.2\%$ ($57.4$ percentage points reduction). These analyses describe agreement with the same priors used for guidance and they are not independent tests of error detection. Lower activation rates therefore show stronger conformity to the reference preferences, not necessarily fewer biological prediction errors.

We next evaluate pattern activation frequencies across all sample–step observations during ODE integration (Fig. 4). The perturbation confusion is the most frequently detected pattern, occurring in an average of $66.6\%$ of the evaluated steps. The dominant activation pattern nevertheless varies across models, with context drift occurring more frequently for scDFM and perturbation confusion dominating for CellFlow and CFM-GP. This variation suggests that no single diagnostic dominates across backbones and provides empirical motivation for combining complementary, potentially conflicting corrections during sampling. This analysis evaluates reference-constraint satisfaction rather than predictive accuracy, which is assessed by the ground-truth metrics in Table 1.

### 4.2.3 Ablation Study (Q3)

The ablations are conducted on CellFlow under the Replogle cross-cell-line setting. We evaluate the contribution of the three pattern-specific corrections, Pareto consensus, and instantaneous-norm preservation. *w/o EC*, *w/o PC*, and *w/o CD* remove the corrections associated with effect collapse, perturbation confusion, and context drift, respectively. *w/o Pareto* replaces the minimum-norm con-

Table 4: Average pathway recovery on HepG2.

| | CellFlow | | CFM-GP | | scDFM | |
|---|---|---|---|---|---|---|
| **Metric** | *Van.* | *+PR.* | *Van.* | *+PR.* | *Van.* | *+PR.* |
| NES Spear. ↑ | 0.032 | **0.328** | 0.357 | **0.404** | -0.035 | **0.428** |
| Sign Agree. ↑ | 0.554 | **0.761** | 0.764 | **0.813** | 0.450 | **0.764** |
| MRR ↑ | 0.010 | **0.053** | 0.022 | **0.049** | 0.013 | **0.017** |

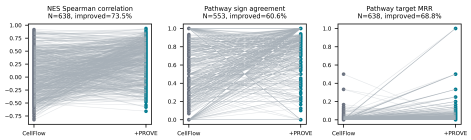


Figure 5: Pathway recovery for CellFlow.

vex consensus with the normalized arithmetic mean of the frozen transport direction and all active correction directions. *w/o Norm Pres.* removes the final velocity normalization.

As shown in Table 2, the complete method has the best reported mean on all eight metrics performing CellFlow with Replogle. Relative to normalized arithmetic averaging, minimum-norm aggregation increases Pearson-$\Delta$ from $0.2053$ to $0.2918$ and DE-Spearman from $0.3202$ to $0.5707$, whereas the AUPRC-DE difference is smaller ($0.2944$ versus $0.2975$). These results support the tested design choices in this configuration. They do not establish that each component is necessary across all backbones and datasets, or that stepwise editing outperforms endpoint-only correction, direct interpolation, sequential projection, or violation-weighted aggregation, which are not evaluated here.

#### 4.2.4 Extension to Diffusion Models (Q4)

To examine whether PROVE extends beyond flow matching, we apply it to the diffusion-based PerturbDiff Yuan et al. (2026) and Squidiff He et al. (2026). Both expose step-wise denoised-data estimates, allowing the same reference diagnostics, corrections, and Pareto-consensus procedure to be applied in the predicted data space. We use the released Replogle checkpoint for PerturbDiff and retrain Squidiff on Replogle following its official implementation, evaluating both with the Q1 metrics. As shown in Table 3, PROVE improves all eight metrics for both models. DE-Spearman increases from $0.6067$ to $0.7366$ for PerturbDiff and from $0.0824$ to $0.5744$ for Squidiff, while MAE decreases from $0.0598$ to $0.0497$ and from $0.5451$ to $0.1417$, respectively. These results suggest that reference-guided correction extends to diffusion samplers with step-wise denoised-data estimates, although its implementation remains sampler dependent.

#### 4.2.5 Biological Pathway Recovery (Q5)

Gene-level accuracy does not necessarily imply recovery of perturbation-induced biological programs. We evaluate pathway-level recovery on Replogle for all three backbones, with CellFlow used for perturbation-level visualization. Using the MSigDB Hallmark collection Liberzon et al. (2015), we rank genes by control-centered mean expression changes and apply preranked GSEA Subramanian et al. (2005) to obtain normalized enrichment scores (NES).

We compare predicted and ground-truth NES profiles using NES Spearman for pathway-ranking agreement, Pathway Sign Agreement for enrichment-direction consistency, and Perturbation Retrieval MRR based on cosine-similarity ranking. Table 4 shows that PROVE improves all three metrics across all three backbones, while Fig. 5 shows improved recovery for most test perturbations with CellFlow. For CellFlow, NES Spearman increases from 0.032 to 0.328 and MRR from 0.010 to 0.053; for scDFM, NES Spearman changes from negative to positive. Together with improved pathway-sign agreement, these results provide complementary evidence that PROVE better recovers coherent and perturbation-specific pathway responses, although absolute MRR remains modest.

## Conclusion

We introduced PROVE, a training-free method for correcting reference-inconsistent extrapolations in cross-context perturbation prediction. PROVE uses three reference-derived diagnostics to construct local edits and reconciles them with the frozen sampler through a minimum-norm consensus, while preserving the backbone's instantaneous velocity scale. Across three datasets and three flow-matching backbones, PROVE consistently improves perturbation-specific and differential-expression recovery. Pathway-level analyses and experiments on two diffusion backbones further support the broader applicability of reference-guided correction.

## AI USE STATEMENT

Large language models were utilized to assist in drafting and rephrasing portions of this manuscript, as well as refining grammar, readability, and sentence flow. The underlying research questions, theoretical derivations, experimental designs, and data analyses were conceived and conducted entirely by the human authors. All AI-assisted text was thoroughly reviewed, verified, and edited by the authors, who take full responsibility for the accuracy, novelty, and integrity of the entire paper.

## REFERENCES

Abrar Rahman Abir, Sajib Acharjee Dip, and Liqing Zhang. CFM-GP: Unified conditional flow matching to learn gene perturbation across cell types. *arXiv preprint arXiv:2508.08312*, 2025.

Abhinav K Adduri, Dhruv Gautam, Beatrice Bevilacqua, Alishba Imran, Rohan Shah, Mohsen Naghipourfar, Noam Teyssier, Rajesh Ilango, Sanjay Nagaraj, Mingze Dong, et al. Predicting cellular responses to perturbation across diverse contexts with state. *BioRxiv*, pp. 2025–06, 2025.

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.

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*, 20:1759–1768, 2023. doi: 10.1038/s41592-023-01969-x.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794, 2021.

Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P Fulco, Livnat Jerby-Arnon, Nemanja D Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, et al. Perturb-seq: dissecting molecular circuits with scalable single-cell rna profiling of pooled genetic screens. *cell*, 167(7): 1853–1866, 2016.

Alice Driessen, Dhruva Abhijit Rajwade, Benedek Harsanyi, Marianna Rapsomaniki, and Jannis Born. Conditional monge gap enables generalizable single-cell perturbation modelling. *Nature Machine Intelligence*, pp. 1–13, 2026.

Ruiqi Feng, Chenglei Yu, Wenhao Deng, Peiyan Hu, and Tailin Wu. On the guidance of flow matching. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 16993–17029. PMLR, 2025. URL https://proceedings.mlr.press/v267/feng25s.html.

Chris J. Frangieh, Johannes C. Melms, Pratiksha I. Thakore, Kathryn R. Geiger-Schuller, Patricia Ho, Adrienne M. Luoma, Brian Cleary, Livnat Jerby-Arnon, Shruti Malu, Michael S. Cuoco, Maryann Zhao, Casey R. Ager, Meri Rogava, Lila Hovey, Asaf Rotem, Chantale Bernatchez, Kai W. Wucherpfennig, Bruce E. Johnson, Orit Rozenblatt-Rosen, Dirk Schadendorf, Aviv Regev, and Benjamin Izar. Multimodal pooled Perturb-CITE-seq screens in patient models define mechanisms of cancer immune evasion. *Nature Genetics*, 53(3):332–341, 2021.

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

Siyu He, Yuefei Zhu, Daniel Naveed Tavakol, Haotian Ye, Yeh-Hsing Lao, Zixian Zhu, Cong Xu, Shradha Chauhan, Guy Garty, Raju Tomer, et al. Squidiff: predicting cellular development and responses to perturbations using a diffusion model. *Nature methods*, 23(1):65–77, 2026.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. URL https://arxiv.org/abs/2207.12598.

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

Arthur Liberzon, Chet Birger, Helga Thorvaldsdottir, Mahmoud Ghandi, Jill P. Mesirov, and Pablo Tamayo. The molecular signatures database hallmark gene set collection. *Cell Systems*, 1(6): 417–425, 2015.

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.

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):e11517, 2023. doi: 10.15252/msb.202211517.

John A Morris, Christina Caragine, Zharko Daniloski, Júlia Domingo, Timothy Barry, Lu Lu, Kyrie Davis, Marcello Ziosi, Dafni A Glinos, Stephanie Hao, et al. Discovery of target genes and pathways at gwas loci by pooled single-cell crispr screens. *Science*, 380(6646):eadh7699, 2023.

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.

Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nature Biotechnology*, 42:927–935, 2024. doi: 10.1038/s41587-023-01905-6.

Andrea Rubbi, Amir Akbarnejad, Mohammad Vali Sanian, Aryan Yazdan Parast, Hesam Asadollahzadeh, Arian Amani, Naveed Akhtar, Sarah Cooper, Andrew Bassett, Pietro Liò, et al. Mixflow: Mixture-conditioned flow matching for out-of-distribution generalization. *arXiv preprint arXiv:2601.11827*, 2026.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Sanjay R Srivatsan, José L McFaline-Figueroa, Vijay Ramani, Lauren Saunders, Junyue Cao, Jonathan Packer, Hannah A Pliner, Dana L Jackson, Riza M Daza, Lena Christiansen, et al. Massively multiplex chemical transcriptomics at single-cell resolution. *Science*, 367(6473):45–51, 2020.

Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43): 15545–15550, 2005.

Ramon Viñas Torné, Maciej Wiatrak, Zoe Piran, Shuyang Fan, Liangze Jiang, Sarah A Teichmann, Mor Nitzan, and Maria Brbić. Systema: a framework for evaluating genetic perturbation response prediction beyond systematic variation. *Nature Biotechnology*, pp. 1–10, 2025.

Luran Wang, Chaoran Cheng, Yizhen Liao, Yanru Qu, and Ge Liu. Training free guided flow-matching with optimal control. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/19a94fdf9e1c5b387830e4c4fef6972a-Abstract-Conference.html.

Yan Wu, Esther Wershof, Sebastian M. Schmon, Marcel Nassar, Błażej Osiński, Ridvan Eksi, Zichao Yan, Rory Stark, Kun Zhang, and Thore Graepel. PerturBench: Benchmarking machine learning models for cellular perturbation analysis. *arXiv preprint arXiv:2408.10609*, 2024.

Zichao Yan, Yan Wu, Mica Xu Ji, Chaitra Agrahar, Esther Wershof, Marcel Nassar, Mehrshad Sadria, Ridvan Eksi, Vladimir Trifonov, Ignacio Ibarra, et al. Primeflow: Capturing complex expression heterogeneity in perturbation response modelling. *arXiv preprint arXiv:2604.13986*, 2026.

Xiaodong Yang, Guole Liu, Guihai Feng, Dechao Bu, Pengfei Wang, Jie Jiang, Shubai Chen, Qinmeng Yang, Hefan Miao, Yiyang Zhang, et al. Genecompass: deciphering universal gene regulatory mechanisms with a knowledge-informed cross-species foundation model. *Cell Research*, 34 (12):830–845, 2024.

Chenglei Yu, Chuanrui Wang, Bangyan Liao, and Tailin Wu. scDFM: Distributional flow matching model for robust single-cell perturbation prediction. *arXiv preprint arXiv:2602.07103*, 2026. doi: 10.48550/arXiv.2602.07103. URL https://arxiv.org/abs/2602.07103.

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. doi: 10.48550/arXiv.2602.19685. URL https://arxiv.org/abs/2602.19685.

# A Additional Method and Theoretical Details

## A.1 Scope of the analysis.

The above results characterize key geometric and numerical properties of PROVE: the reference-defined regions are mutually compatible, the linear endpoint estimate becomes increasingly accurate near the terminal time under the stated regularity assumptions, the exact minimum-norm solution is jointly aligned with its input directions, and the final redirection is bounded while preserving the instantaneous velocity norm. Together, these properties provide geometric consistency and numerical control for the inference-time update. They should be interpreted as guarantees on the update itself rather than as an end-to-end guarantee of predictive accuracy.

## A.2 Theoretical Properties

For theoretical analysis, we associate each response pattern with the set of estimates that do not trigger it. We define:

$$\begin{aligned} C_{\mathrm{EC}} &= \{z : u_g^\top z \geq \beta \|p_g\|_2\}, \\ C_{\mathrm{PC}} &= \left\{ z : u_g^\top z \geq \max_{h \in \mathcal{H}_g} u_h^\top z \right\}, \\ C_{\mathrm{CD}} &= \{z : \|z - p_g\|_2 \leq \sigma_g\}. \end{aligned} \tag{11}$$

where $C_{\mathrm{EC}}$, $C_{\mathrm{PC}}$, and $C_{\mathrm{CD}}$ are the zero-score regions associated with effect collapse, perturbation confusion, and context drift, respectively.

**Proposition 1** (Compatibility of the pattern-specific regions)**.** *Assume $p_g \neq 0$, $\beta \in (0, 1)$, and $u_g = p_g / \|p_g\|_2$. Whenever the corresponding patterns are available,*

$$p_g \in C_{\mathrm{EC}} \cap C_{\mathrm{PC}} \cap C_{\mathrm{CD}}. \tag{12}$$

*Consequently, the intersection of the available pattern-specific regions is nonempty.*

*Proof.* Since $u_g^\top p_g = \|p_g\|_2 \geq \beta \|p_g\|_2$, we have $p_g \in C_{\mathrm{EC}}$. For every $h \in \mathcal{H}_g$, $u_h^\top p_g = \|p_g\|_2 u_h^\top u_g \leq \|p_g\|_2 = u_g^\top p_g$, so $p_g \in C_{\mathrm{PC}}$. Finally, $\|p_g - p_g\|_2 = 0 \leq \sigma_g$, and therefore $p_g \in C_{\mathrm{CD}}$. □

The pattern scores are evaluated on the linear final-state estimate in Eq. equation 1. The following result characterizes its approximation error along the remaining ODE trajectory.

**Proposition 2** (Accuracy of the linear final-state estimate)**.** *Let $x_s$ satisfy $\dot{x}_s = v(x_s, s)$ for $s \in [t, 1]$, and define $\tilde{x}_1 = x_t + (1 - t)v(x_t, t)$. Assume:*

$$\begin{aligned} \|v(x, s) - v(y, s)\|_2 &\leq L_x \|x - y\|_2, \\ \|v(x, s) - v(x, r)\|_2 &\leq L_t |s - r|, \\ \|v(x, s)\|_2 &\leq B. \end{aligned} \tag{13}$$

*where $L_x$ and $L_t$ are the state and time Lipschitz constants and $B$ is a uniform velocity bound. Then:*

$$\|x_1 - \tilde{x}_1\|_2 \leq \frac{L_x B + L_t}{2} (1 - t)^2. \tag{14}$$

*If $\phi$ is $L_\phi$-Lipschitz, then :*

$$\|\phi(x_1) - \phi(\tilde{x}_1)\|_2 \leq \frac{L_\phi (L_x B + L_t)}{2} (1 - t)^2. \tag{15}$$

*Proof.* Using the integral form of the ODE, we have $x_1 - \tilde{x}_1 = \int_t^1 (v(x_s, s) - v(x_t, t))\, ds$. The velocity bound gives $\|x_s - x_t\|_2 \leq B(s - t)$, and the Lipschitz assumptions therefore imply $\|v(x_s, s) - v(x_t, t)\|_2 \leq (L_x B + L_t)(s - t)$. Integrating over $s \in [t, 1]$ yields Eq. equation 14. Equation equation 15 follows from the Lipschitz continuity of $\phi$. □

The minimum-norm consensus is intended to retain agreement with both the frozen transport and all active corrections. We define :

$$\mathcal{G}_t = \mathrm{conv}\left(\{d_{\mathrm{frz}}\} \cup \{d_k : k \in A_t\}\right), \tag{16}$$

where $\mathcal{G}_t$ is the convex hull of the directions entering the consensus problem.

When applied to the estimate used by PROVE, this bound concerns the frozen ODE continuation initialized at the current state $x_t$. It does not bound the difference from the final guided sample or from an observed biological target cell. The constants are not estimated in our experiments, and the bound may be loose early in the trajectory. Full-trajectory guidance is therefore an empirical design choice, not a consequence of this approximation guarantee; the current evaluation does not measure endpoint-estimation error as a function of time.

**Proposition 3** (Positive alignment of the consensus direction)**.** *Let $\bar{d}_t = \Pi_{\mathcal{G}_t}(0)$ be the minimum-norm point in $\mathcal{G}_t$. If $a_t = \|\bar{d}_t\|_2 > 0$ and $d_{\mathrm{mix}} = \bar{d}_t / a_t$, then every input direction $d_i \in \{d_{\mathrm{frz}}\} \cup \{d_k : k \in A_t\}$ satisfies :*

$$d_i^\top d_{\mathrm{mix}} \geq a_t > 0. \tag{17}$$

*Proof.* The variational inequality for the Euclidean projection of the origin onto $\mathcal{G}_t$ gives $\langle \bar{d}_t, d_i - \bar{d}_t \rangle \geq 0$. Therefore, $d_i^\top \bar{d}_t \geq \|\bar{d}_t\|_2^2$. Dividing by $a_t = \|\bar{d}_t\|_2 > 0$ yields Eq. equation 17. □

The clipping and normalization steps control the size of each correction and preserve the velocity scale of the frozen sampler.

**Proposition 4** (Bounded editing and exact norm preservation)**.** *Assume $A_t \neq \emptyset$, $\rho_t \in [0, 1]$, and $a_t \in (0, 1]$. Then :*

$$0 \leq \gamma_t \leq 1, \;\; \|\gamma_t c_t\|_2 \leq \|\tilde{x}_1 - x_t\|_2. \tag{18}$$

*Under the numerical fallback in Eq. equation 40, either $v_t^{\mathrm{PROVE}} = v_t$ or :*

$$\|v_t^{\mathrm{PROVE}}\|_2 = \|v_t\|_2. \tag{19}$$

*Proof.* Equation equation 8 gives $0 \leq \gamma_t \leq 1$, while Eq. equation 9 gives $\|c_t\|_2 \leq \|\tilde{x}_1 - x_t\|_2$. Therefore :

$$\|\gamma_t c_t\|_2 \leq \gamma_t \|\tilde{x}_1 - x_t\|_2 \leq \|\tilde{x}_1 - x_t\|_2. \tag{20}$$

If $\|\hat{x}_1 - x_t\|_2 > \epsilon$, the normalized direction in Eq. equation 40 has unit norm, which gives $\|v_t^{\mathrm{PROVE}}\|_2 = \|v_t\|_2$. Otherwise, the fallback directly returns $v_t$. □

### A.3 Reference Construction and Pattern Availability

The main text uses context-level mean responses as references. For any reference cell $x$, let $c(x)$ and $\pi(x)$ denote its context and perturbation labels, and define its control-centered response as $\eta(x) = \phi(x) - \mu_{\mathrm{ctrl}}(c(x))$. For each non-target context $c' \neq c^\star$ in which perturbation $g$ is observed, we define :

$$\begin{aligned} S_{c',g} &= \{x \in \mathcal{D}_{\mathrm{ref}} : c(x) = c', \pi(x) = g\}, \\ \bar{\eta}_{c',g} &= \frac{1}{|S_{c',g}|} \sum_{x \in S_{c',g}} \eta(x), \end{aligned} \tag{21}$$

where $S_{c',g}$ contains the reference cells receiving perturbation $g$ in context $c'$, and $\bar{\eta}_{c',g}$ is their mean control-centered response. We collect the available context-level responses in $\mathcal{R}_g = \{\bar{\eta}_{c',g} : c' \neq c^\star, |S_{c',g}| > 0\}$ and define :

$$\begin{aligned} p_g &= \frac{1}{|\mathcal{R}_g|} \sum_{r \in \mathcal{R}_g} r, \\ \sigma_g &= \left( \frac{1}{|\mathcal{R}_g|} \sum_{r \in \mathcal{R}_g} \|r - p_g\|_2^2 \right)^{1/2}, \\ u_g &= \frac{p_g}{\|p_g\|_2} \end{aligned} \tag{22}$$

where $p_g$ is the mean response to the target perturbation across reference contexts, $\sigma_g$ is its cross-context dispersion, and $u_g$ is its normalized direction when $\|p_g\|_2 > \epsilon$. For each non-target perturbation $h \neq g$ observed in $c^\star$, we define :

$$\begin{aligned} S_{c^\star,h} &= \{x \in \mathcal{D}_{\text{ref}} : c(x) = c^\star, \pi(x) = h\}, \\ \bar{\eta}_{c^\star,h} &= \frac{1}{|S_{c^\star,h}|} \sum_{x \in S_{c^\star,h}} \eta(x), \\ u_h &= \frac{\bar{\eta}_{c^\star,h}}{\|\bar{\eta}_{c^\star,h}\|_2} \end{aligned} \tag{23}$$

where $u_h$ is defined when $\|\bar{\eta}_{c^\star,h}\|_2 > \epsilon$. We let $\mathcal{H}_g$ contain the perturbations whose directions are defined and satisfy $\|u_h - u_g\|_2 > \epsilon$. At ODE step $t$, we define :

$$h_t^\star = \arg\max_{h \in \mathcal{H}_g} u_h^\top z_t. \tag{24}$$

where $h_t^\star$ is the competing perturbation most strongly aligned with the current response estimate. Effect collapse is available when $\mathcal{R}_g \neq \emptyset$ and $\|p_g\|_2 > \epsilon$; perturbation confusion additionally requires $\mathcal{H}_g \neq \emptyset$; and context drift requires $|\mathcal{R}_g| \geq 2$ and $\sigma_g > \epsilon$. A pattern is omitted from $A_t$ whenever its required reference statistics are unavailable.

### A.4 Closed-Form Pattern Corrections

The generic projection in Eq. equation 5 has a closed-form solution for each response pattern. For effect collapse, we define :

$$\Delta_{\text{EC}} = \left[\beta\|p_g\|_2 - u_g^\top z_t\right]_+ u_g. \tag{25}$$

For perturbation confusion, we define :

$$\Delta_{\text{PC}} = \begin{cases} \dfrac{u_{h_t^\star}^\top z_t - u_g^\top z_t}{\|u_g - u_{h_t^\star}\|_2^2}(u_g - u_{h_t^\star}), & u_{h_t^\star}^\top z_t > u_g^\top z_t, \\ 0, & \text{otherwise}, \end{cases} \tag{26}$$

For context drift, we define :

$$\Delta_{\text{CD}} = \begin{cases} p_g + \sigma_g \dfrac{z_t - p_g}{\|z_t - p_g\|_2} - z_t, & \|z_t - p_g\|_2 > \sigma_g, \\ 0, & \text{otherwise}. \end{cases} \tag{27}$$

where $\Delta_{\text{EC}}$ is the Euclidean projection displacement onto the minimum-effect half-space, $\Delta_{\text{PC}}$ is the projection displacement onto the half-space defined by the active competitor $h_t^\star$, and $\Delta_{\text{CD}}$ is the projection displacement onto the cross-context reference ball. Thus, the projection operator in Eq. equation 5 satisfies $\Pi_k(z_t) = z_t + \Delta_k$.

### A.5 Pareto-Consensus Optimization

The convex-hull problem in Eq. equation 7 can be written as a simplex-constrained quadratic program. Let $J = |A_t| + 1$, and let $D_t \in \mathbb{R}^{d \times J}$ contain $d_{\text{frz}}$ and the active correction directions $\{d_k : k \in A_t\}$ as columns. We define :

$$\begin{aligned} \alpha^\star &= \arg\min_{\alpha \in \mathbb{R}^J} \|D_t\alpha\|_2^2, \\ &\text{s.t.} \quad \mathbf{1}^\top\alpha = 1, \;\; \alpha \geq 0, \end{aligned} \tag{28}$$

where $\bar{d}_t = D_t\alpha^\star$. Equivalently,

$$\bar{d}_t = \alpha_{\text{frz}}^\star d_{\text{frz}} + \sum_{k \in A_t} \alpha_k^\star d_k, \;\; \alpha_{\text{frz}}^\star + \sum_{k \in A_t} \alpha_k^\star = 1, \;\; \alpha_{\text{frz}}^\star, \alpha_k^\star \geq 0. \tag{29}$$

When $J = 2$, for input directions $q_1$ and $q_2$, we define :

$$\lambda^\star = \Pi_{[0,1]}\left(\frac{\|q_2\|_2^2 - q_1^\top q_2}{\|q_1 - q_2\|_2^2}\right), \;\; \bar{d}_t = \lambda^\star q_1 + (1 - \lambda^\star)q_2. \tag{30}$$

where $\Pi_{[0,1]}$ denotes scalar projection onto $[0, 1]$. When $J \geq 3$, we use Frank–Wolfe optimization. Let $f(\alpha) = \|D_t\alpha\|_2^2$. At iteration $j$, we define :

$$\begin{aligned} i_j &= \underset{i\in\{1,\ldots,J\}}{\arg\min} [\nabla f(\alpha^{(j)})]_i, \\ \alpha^{(j+1)} &= (1-\eta_j)\alpha^{(j)} + \eta_j e_{i_j}, \\ \eta_j &= \frac{2}{j+2} \end{aligned} \tag{31}$$

where $e_{i_j}$ is the selected simplex vertex. We use $N_{\mathrm{FW}} = 10$ iterations in all experiments.

### A.6 Finite-iteration qualification.

For any feasible approximate solution $q = D_t\alpha$ with $a = \|q\|_2 > 0$, define the Frank–Wolfe gap:

$$G(\alpha) = 2\left(\|q\|_2^2 - \min_i d_i^\top q\right), \tag{32}$$

where $d_i$ denotes the $i$th column of $D_t$. Then

$$\min_i d_i^\top \frac{q}{a} = a - \frac{G(\alpha)}{2a}. \tag{33}$$

Thus positive alignment is certified when $G(\alpha) < 2a^2$, but not merely by using a fixed iteration count. We do not claim that ten iterations certify this condition for every update.

### A.7 Geometric Interpretation of the Velocity Edit

Recall that $\delta_t = \phi(\tilde{x}_1) - \phi(x_t)$ is the remaining response-space displacement proposed by the frozen sampler. Since $d_{\mathrm{mix}}$ is unit-normalized, the displacement :

$$\delta_t^{\mathrm{con}} = \|\delta_t\|_2 d_{\mathrm{mix}} \tag{34}$$

has the same length as $\delta_t$ but points in the Pareto-consensus direction. Therefore,

$$e_t^{\mathrm{resp}} = \delta_t^{\mathrm{con}} - \delta_t \tag{35}$$

is the response-space change required to replace the frozen displacement with the equal-length consensus displacement. Equation equation 9 maps this change to the model state space and clips its magnitude before modifying the endpoint estimate.

Let :

$$\kappa_t = \min\left\{1, \frac{\|\tilde{x}_1 - x_t\|_2}{\|\tilde{c}_t\|_2 + \epsilon}\right\}, \quad \lambda_t = \gamma_t\kappa_t, \tag{36}$$

so that $c_t = \kappa_t\tilde{c}_t$ and $\lambda_t \in [0, 1]$. For the identity and affine response maps used in our experiments, $M_\phi$ is an exact inverse displacement map and satisfies :

$$\phi(x + M_\phi(e)) - \phi(x) = e. \tag{37}$$

The corrected endpoint $\hat{x}_1 = \tilde{x}_1 + \gamma_t c_t$ therefore satisfies :

$$\phi(\hat{x}_1) - \phi(x_t) = (1-\lambda_t)\delta_t + \lambda_t\|\delta_t\|_2 d_{\mathrm{mix}}. \tag{38}$$

Thus, the edit interpolates between the frozen response-space displacement and an equal-length displacement in the consensus direction. If $\gamma_t = 0$, the endpoint remains unchanged. If $d_{\mathrm{mix}} = d_{\mathrm{frz}}$, then $\|\delta_t\|_2 d_{\mathrm{mix}} = \delta_t$ and the correction is zero. In either case, PROVE recovers the frozen update.

### A.8 Response-to-State Mapping and Numerical Safeguards

The consensus correction is computed in the response space and mapped to the model state space through $M_\phi$. For gene-wise standardized expression, we use :

$$\phi(x) = D_s^{-1}(x - b), \quad M_\phi(e) = D_s e, \tag{39}$$

where $D_s$ is the diagonal matrix of gene-wise scales and $b$ is the centering vector. The map $M_\phi$ acts on displacements rather than absolute states. When the ODE state is already represented in the response space, both $\phi$ and $M_\phi$ are the identity map. In our experiments, scDFM uses the affine map in Eq. equation 39, whereas CellFlow and CFM-GP use the identity map in their respective preprocessed state spaces.

We use a shared tolerance $\epsilon = 10^{-6}$ for reference availability, pattern activation, direction normalization, and numerical fallback. A direction is normalized only when its norm exceeds $\epsilon$. The set $H_g$ excludes competing directions satisfying $\|u_h - u_g\|_2 \leq \epsilon$, preventing a degenerate denominator in the perturbation-confusion correction. The denominator in the two-direction Pareto solution is also lower-bounded by $\epsilon$ in implementation.

For numerical stability, the velocity redirection is implemented as :

$$v_t^{\mathrm{PROVE}} = \begin{cases} \|v_t\|_2 \dfrac{\hat{x}_1 - x_t}{\|\hat{x}_1 - x_t\|_2}, & \|\hat{x}_1 - x_t\|_2 > \epsilon, \\ v_t, & \text{otherwise.} \end{cases} \tag{40}$$

PROVE also returns $v_t$ when $A_t = \emptyset$, $\|\delta_t\|_2 \leq \epsilon$, or $a_t \leq \epsilon$. Hence, every applied redirection preserves the instantaneous velocity norm exactly, while degenerate cases leave the frozen velocity unchanged. All pretrained backbones remain frozen during inference, and all reference statistics are computed exclusively from split-valid training cells, with post-perturbation cells from the target context–perturbation pair excluded.

# B Implementation Details

The ODE solvers, integration settings, preprocessing procedures, and other inference configurations follow the corresponding official implementations. Each backbone is trained with seeds $\{0, 42, 2026\}$ following the official protocol. For each trained model, the vanilla and PROVE samplers use matched inference randomness. All experiments are conducted on NVIDIA RTX PRO 6000 GPUs. Main tables report means, with standard deviations provided in the Appendix. We fix $\beta = 0.5$ and use $\epsilon = 10^{-6}$ for numerical stability. We apply guidance $\rho_t = 1$ throughout the full sampling trajectory $[0, 1]$. For Pareto consensus, we use the closed-form solution when $J = 2$ and Frank-Wolfe optimization otherwise, with $N_{\mathrm{FW}} = 10$ iterations.

## B.1 Inference cost

Let $d$ be response dimension, $H = |\mathcal{H}_g|$ the number of competing perturbations, and $J \leq 4$ the number of consensus directions. Per generated cell and guided velocity evaluation, scanning competitors costs $O(dH)$. With a precomputed consensus Gram matrix, the remaining optimization costs $O(dJ^2 + N_{\mathrm{FW}}J^2)$. PROVE reuses the current frozen prediction and does not require backpropagation through the backbone. These operation counts are not wall-clock measurements, and training-free does not mean cost-free. For adaptive solvers, changing the field may also change the total number of function evaluations.

# C Additional Experimental Results

## C.1 Dataset Descriptions

### C.1.1 Dataset 1 (Replogle)

We use the STATE-processed Replogle benchmark for large-scale cross-cell-line genetic perturbation prediction Replogle et al. (2022); Adduri et al. (2025). Following the STATE covariate-transfer split, HepG2 is treated as the target context, and all post-perturbation cells corresponding to 638 perturbations in HepG2 are excluded from model training and reference construction. Each target perturbation remains observed in at least one non-target cell line, providing cross-context target-perturbation references, while HepG2 controls and responses to non-target perturbations are retained when permitted by the split. This setting evaluates whether a model can transfer genetic perturbation effects across cell lines without observing the target cell-line–perturbation combinations.

Table 5: Percentage of ground-truth target context–perturbation pairs whose responses do not trigger each reference diagnostic. Bold values are at least $90\%$.

| Dataset | No effect collapse | No perturbation confusion | No context drift |
|---|---|---|---|
| Replogle | 89.7 | **90.8** | **100.0** |
| Frangieh21 | 82.8 | **90.8** | **92.0** |
| Srivatsan20 | 74.4 | **91.9** | **90.7** |

#### C.1.2 DATASET 2 (FRANGIEH21)

Frangieh21 is derived from pooled Perturb-CITE-seq screens of patient-derived melanoma models designed to characterize genetic mechanisms of tumor immune evasion Frangieh et al. (2021). The dataset contains single-cell responses to CRISPR-based genetic perturbations measured under distinct biological states, including baseline, interferon-$\gamma$ stimulation, and immune-cell co-culture conditions. We use the PerturBench-processed version and follow its official covariate-transfer split Wu et al. (2024). The biological state is treated as the context variable: cells from each held-out state–perturbation combination are excluded from training and reference construction, while responses to the same perturbation in other states remain available. This benchmark evaluates whether perturbation effects can be transferred across immune-related states with substantially different transcriptional programs.

#### C.1.3 DATASET 3 (SRIVATSAN20)

Srivatsan20 is based on the sci-Plex single-cell chemical perturbation screen, which profiled transcriptional responses to a large collection of compounds across the A549, K562, and MCF7 cancer cell lines Srivatsan et al. (2020). We use the PerturBench-processed version and follow its official cross-cell-line covariate-transfer split Wu et al. (2024). The cell line is treated as the context variable, and selected drug–cell-line combinations are held out for evaluation while each target drug remains observed in at least one other cell line. Target-cell-line controls and responses to non-target drugs are retained according to the split. This setting evaluates whether chemically induced transcriptional effects can be transferred across cell lines with distinct basal expression profiles.

### C.2 GROUND-TRUTH PATTERN COVERAGE

We evaluate whether the reference-based checks retain the ground-truth target responses. These responses are used only for post-hoc evaluation and are excluded from reference construction and PROVE inference. For each pattern, coverage is defined as the percentage of test context–perturbation pairs whose ground-truth control-centered response does not trigger the corresponding check.

As shown in Table 5, the non-activation rates vary across diagnostics and datasets. In particular, the effect-collapse score activates on $25.6\%$ of the Srivatsan20 ground-truth target responses. These rates quantify agreement with the chosen reference priors rather than validating the scores as error detectors. Thus, legitimate target responses may disagree with the reference priors. Since we do not evaluate predictive performance separately on such target pairs, the aggregate improvements do not rule out degradation on this subset.

### C.3 PARAMETER SENSITIVITY

Figures 6 and 7 evaluate the guidance schedule and effect-collapse threshold on Replogle with CellFlow. Full-trajectory guidance performs best on seven of eight metrics, while $\beta = 0.5$ provides the strongest overall trade-off between perturbation-specific recovery and global fidelity.

The defaults $\beta = 0.5$ and $\rho_t = 1$ were fixed before inspecting performance on the held-out target pairs. Figures 6 and 7 are post-hoc sensitivity analyses using the Replogle evaluation responses; they were not used to select the reported configuration. The same settings were retained across datasets. Their purpose is to describe sensitivity, not to provide an independent estimate of a configuration selected from these curves.

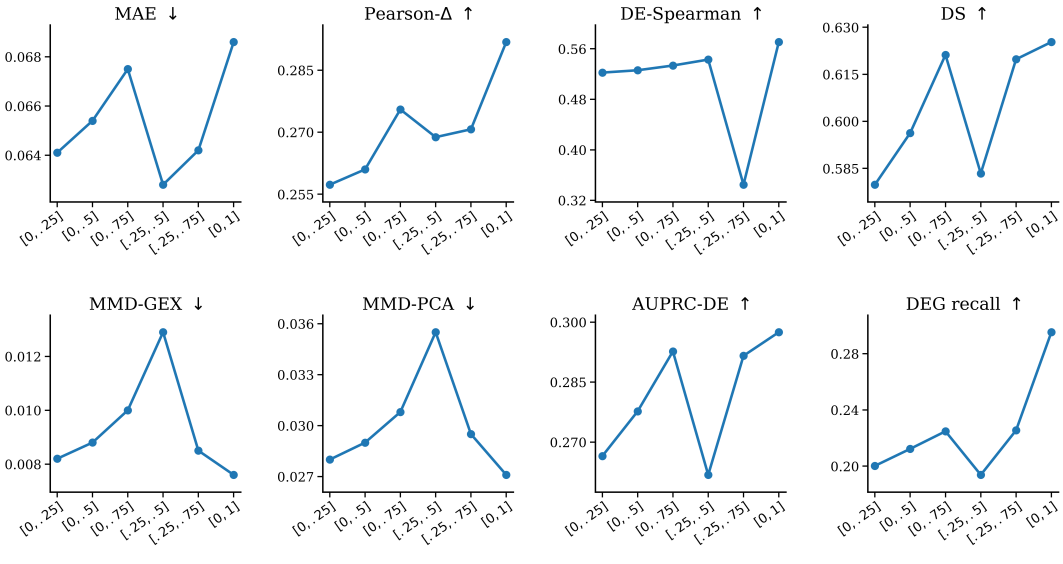


Figure 6: Sensitivity to the guidance schedule $\rho_t$.

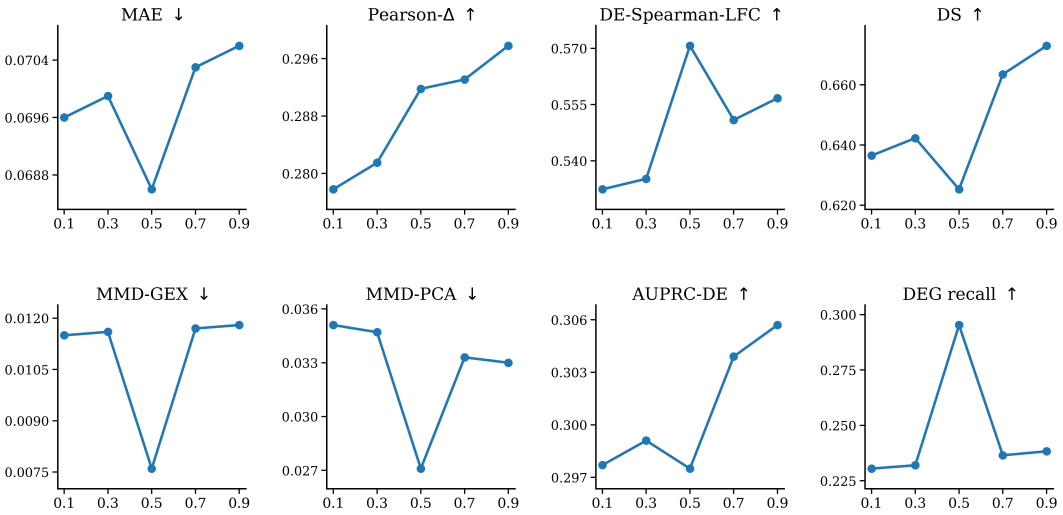


Figure 7: Sensitivity to the effect-collapse threshold $\beta$.

### C.4 Comparison with Direct Reference Transfer

An important question is whether the improvements obtained by PROVE can be reproduced by directly transferring the available source-context responses to the unobserved target context. Reference-only predictors treat source-context responses as direct estimates of the target-context response, effectively replacing the conditional transport model. In contrast, PROVE retains the transport proposed by a pretrained backbone and uses the available references only to identify and correct reference-inconsistent predictions.

We compare three PROVE-guided backbones with three reference-only predictors, including CC-Mean (predicting the target response by averaging the observed responses to the same perturbation across all available non-target contexts), TwoWay-Ridge (fitting an additive ridge-regression model with context and perturbation effects, and predicting the target condition by summing the estimated main effects), and Analog Residual (retrieving reference perturbations with similar context-centered

Table 6: Comparison of PROVE-guided models and reference-only predictors on Replogle. Direct reference transfer is highly competitive in this setting, while PROVE-guided models attain the best MAE, MMD-GEX, AUPRC-DE, and DEG Recall. Boldface denotes the best result among all six methods.

| Group | Method | MAE ↓ | Pearson-Δ ↑ | DE-Spearman ↑ | DS ↑ | MMD-GEX ↓ | MMD-PCA ↓ | AUPRC-DE ↑ | DEG Recall ↑ |
|---|---|---|---|---|---|---|---|---|---|
| PROVE-guided | scDFM + PROVE | 0.1274 | 0.2994 | 0.3759 | 0.6199 | 0.0620 | 0.1681 | 0.3005 | **0.7130** |
| | CellFlow + PROVE | 0.0686 | 0.2918 | 0.5707 | 0.6253 | **0.0076** | 0.0271 | 0.2975 | 0.2953 |
| | CFM-GP + PROVE | **0.0649** | 0.3183 | 0.4119 | 0.6411 | 0.0112 | 0.0330 | **0.3152** | 0.3430 |
| Reference-only | CC-Mean | 0.0691 | 0.3206 | 0.5945 | 0.7253 | 0.0095 | **0.0264** | 0.3121 | 0.3243 |
| | TwoWay-Ridge | 0.0699 | **0.3656** | **0.6393** | 0.6999 | 0.0080 | 0.0272 | 0.3005 | 0.3203 |
| | Analog Residual | 0.0965 | 0.2667 | 0.5408 | **0.7678** | 0.0276 | 0.0523 | 0.2635 | 0.5287 |

Table 7: Comparison of PROVE-guided models and reference-only predictors on Frangieh21. Reference-only predictors attain the lowest global expression and distributional errors, whereas PROVE-guided models lead all five perturbation-specific metrics, including response direction, differential-expression ranking, and DE-gene recovery. Boldface denotes the best result among all six methods.

| Group | Method | MAE ↓ | Pearson-Δ ↑ | DE-Spearman ↑ | DS ↑ | MMD-GEX ↓ | MMD-PCA ↓ | AUPRC-DE ↑ | DEG Recall ↑ |
|---|---|---|---|---|---|---|---|---|---|
| PROVE-guided | scDFM + PROVE | 0.1014 | 0.4637 | 0.5967 | **0.6272** | 0.0805 | 0.1643 | 0.5198 | 0.7007 |
| | CellFlow + PROVE | 0.0224 | **0.8583** | **0.9281** | 0.5878 | 0.0015 | 0.0091 | **0.7260** | **0.7021** |
| | CFM-GP + PROVE | 0.0225 | 0.8037 | 0.9109 | 0.6008 | 0.0046 | 0.0187 | 0.6287 | 0.6915 |
| Reference-only | CC-Mean | 0.0204 | 0.1865 | 0.7111 | 0.5791 | 0.0027 | 0.0058 | 0.5281 | 0.5405 |
| | TwoWay-Ridge | **0.0177** | 0.0994 | 0.4862 | 0.5178 | **0.0011** | **0.0046** | 0.1217 | 0.0014 |
| | Analog Residual | 0.0233 | 0.1403 | 0.7900 | 0.5992 | 0.0041 | 0.0058 | 0.4886 | 0.6148 |

residual responses and transferring their residual expression change onto the target-context control baseline) on all three datasets. This comparison tests whether direct reference transfer can replace reference-guided transport.

For interpretation, we separate the metrics into two groups. MAE, MMD-GEX, and MMD-PCA evaluate global expression or distributional fidelity, whereas Pearson-Δ, DE-Spearman, DS, AUPRC-DE, and DEG Recall emphasize perturbation-specific response direction, differential-expression structure, and perturbation identity.

### C.4.1 Aggregate Comparison

Tables 6–8 show a metric-dependent distinction. PROVE-guided methods lead 12 of the 15 perturbation-specific dataset–metric comparisons, including all five such metrics on Frangieh21 and Srivatsan20. Reference-only methods achieve the best global-fidelity scores on Frangieh21 and Srivatsan20, whereas PROVE-guided methods lead all five perturbation-specific metrics on both datasets. PROVE-guided models nevertheless attain the best Replogle MAE, MMD-GEX, AUPRC-DE, and DEG Recall. These results indicate that direct transfer is effective when source effects extrapolate well, whereas PROVE is more consistent at preserving perturbation-specific response structure.

### C.4.2 Perturbation-Level Analysis on Srivatsan20

For a perturbation-level illustration, we compare the representative CellFlow+PROVE model with the simple CC-Mean reference-transfer baseline on all Srivatsan20 test perturbations. As shown in Fig. 8, PROVE improves DE-Spearman for 61 of the 64 valid perturbations, with a median paired gain of $0.569$ and a paired Wilcoxon $p$-value of $5.15 \times 10^{-12}$. The improvement is therefore broadly distributed rather than driven by a few outliers.

### C.4.3 Source-Only Predictability

Fig. 9 relates the leave-one-context-out source prediction error to the paired PROVE gain. Larger source-only error is associated with a larger DE-Spearman advantage for PROVE (Spearman $\rho = 0.311$, 95% bootstrap CI $[0.069, 0.542]$). This association is consistent with the intended distinction. Direct transfer is competitive when source effects are reliable, whereas retaining the backbone transport becomes more valuable as source-only prediction error increases.

Table 8: Comparison of PROVE-guided models and reference-only predictors on Srivatsan20. Reference-only methods remain competitive on global expression and distributional fidelity, whereas PROVE-guided models lead all five perturbation-specific metrics by substantial margins. Boldface denotes the best result among all six methods.

| Group | Method | MAE ↓ | Pearson-Δ ↑ | DE-Spearman ↑ | DS ↑ | MMD-GEX ↓ | MMD-PCA ↓ | AUPRC-DE ↑ | DEG Recall ↑ |
|---|---|---|---|---|---|---|---|---|---|
| PROVE-guided | scDFM + PROVE | 0.0598 | 0.8059 | 0.8322 | 0.8519 | 0.1219 | 0.2693 | **0.7242** | 0.5762 |
| | CellFlow + PROVE | 0.0128 | **0.8772** | **0.8754** | 0.8597 | 0.0100 | 0.1251 | 0.6690 | **0.7523** |
| | CFM-GP + PROVE | 0.0147 | 0.8420 | 0.8495 | **0.8610** | 0.1240 | 0.0642 | 0.6206 | 0.7412 |
| Reference-only | CC-Mean | 0.0138 | 0.0985 | 0.2130 | 0.5155 | 0.0084 | 0.0413 | 0.1524 | 0.1470 |
| | TwoWay-Ridge | **0.0110** | 0.1564 | 0.2847 | 0.5627 | **0.0056** | **0.0310** | 0.1996 | 0.0363 |
| | Analog Residual | 0.0169 | 0.2649 | 0.3897 | 0.8018 | 0.0111 | 0.0395 | 0.2728 | 0.4453 |

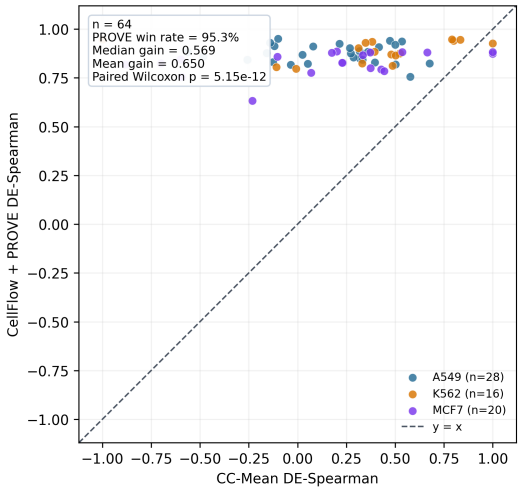


Figure 8: Perturbation-level comparison of CellFlow + PROVE and CC-Mean on Srivatsan20. Each point represents one perturbation with finite DE-Spearman under both methods, and colors denote the target cell line. Points above the dashed diagonal favor PROVE. CellFlow + PROVE outperforms CC-Mean on 95.3% of the 64 valid pairs, with a median paired gain of 0.569 and a mean gain of 0.650. The paired Wilcoxon test gives $p = 5.15 \times 10^{-12}$.

### C.5 Full Results with Standard Deviations

Each model-dataset configuration is evaluated with seeds 0, 42, and 2026. The main tables report mean values, and the complete results with standard deviations are provided below.

Table 9: Full results on the Replogle dataset. Values are reported as mean ± standard deviation over three seeds. Boldface indicates the better mean between each vanilla model and its PROVE-enhanced counterpart.

| | scDFM | | CellFlow | | CFM-GP | |
|---|---|---|---|---|---|---|
| Metric | Vanilla | +PROVE | Vanilla | +PROVE | Vanilla | +PROVE |
| MAE ↓ | $0.1936 \pm 0.00023$ | $\mathbf{0.1274 \pm 0.00017}$ | $0.0705 \pm 0.00054$ | $\mathbf{0.0686 \pm 0.00021}$ | $0.0690 \pm 0.00055$ | $\mathbf{0.0649 \pm 0.0001}$ |
| Pearson-Δ ↑ | $0.0643 \pm 0.0216$ | $\mathbf{0.2994 \pm 0.01249}$ | $0.0320 \pm 0.01207$ | $\mathbf{0.2918 \pm 0.03057}$ | $0.2675 \pm 0.05737$ | $\mathbf{0.3183 \pm 0.02030}$ |
| DE-Spearman ↑ | $0.0316 \pm 0.03958$ | $\mathbf{0.3759 \pm 0.03651}$ | $0.0185 \pm 0.01672$ | $\mathbf{0.5707 \pm 0.01369}$ | $0.3007 \pm 0.05877$ | $\mathbf{0.4119 \pm 0.02205}$ |
| DS ↑ | $0.5013 \pm 0.00397$ | $\mathbf{0.6199 \pm 0.0229}$ | $0.5082 \pm 0.00101$ | $\mathbf{0.6253 \pm 0.01031}$ | $0.5950 \pm 0.04142$ | $\mathbf{0.6411 \pm 0.03639}$ |
| MMD-GEX ↓ | $0.1019 \pm 0.00019$ | $\mathbf{0.0620 \pm 0.00021}$ | $0.0126 \pm 0.00023$ | $\mathbf{0.0076 \pm 0.00039}$ | $\mathbf{0.0094 \pm 0.00010}$ | $0.0112 \pm 0.00050$ |
| MMD-PCA ↓ | $0.2206 \pm 0.00018$ | $\mathbf{0.1681 \pm 0.00006}$ | $0.0458 \pm 0.00039$ | $\mathbf{0.0271 \pm 0.00023}$ | $0.0343 \pm 0.00010$ | $\mathbf{0.0330 \pm 0.00050}$ |
| AUPRC-DE ↑ | $0.1985 \pm 0.00103$ | $\mathbf{0.3005 \pm 0.00152}$ | $0.1511 \pm 0.00125$ | $\mathbf{0.2975 \pm 0.00666}$ | $0.2294 \pm 0.01734$ | $\mathbf{0.3152 \pm 0.00661}$ |
| DEG Recall ↑ | $0.6598 \pm 0.0290$ | $\mathbf{0.7130 \pm 0.0263}$ | $0.0030 \pm 0.00109$ | $\mathbf{0.2953 \pm 0.01124}$ | $0.1732 \pm 0.07204$ | $\mathbf{0.3430 \pm 0.03655}$ |

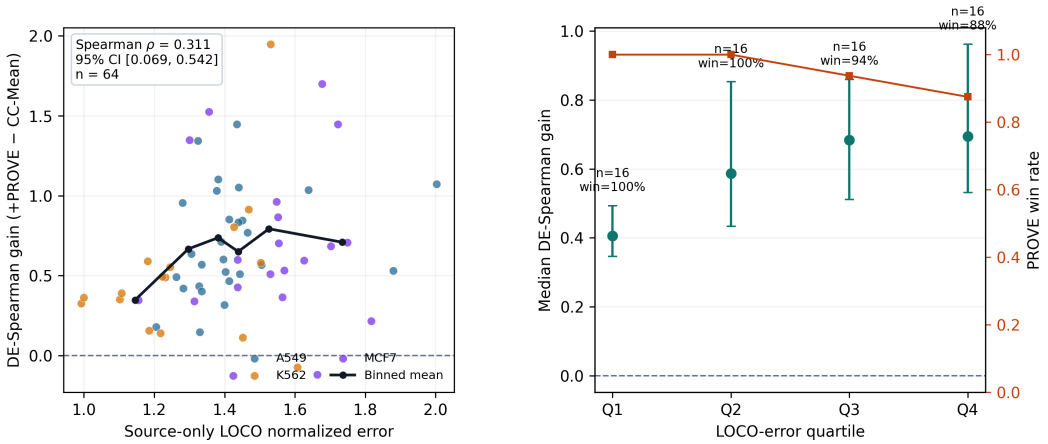


Figure 9: Source-only reference predictability and the relative benefit of PROVE on Srivatsan20. **Left:** Relationship between leave-one-context-out (LOCO) normalized error and the paired DE-Spearman gain of CellFlow + PROVE over CC-Mean. Each point denotes one perturbation, colors indicate target cell lines, and the dark curve shows the binned mean. Larger LOCO error is associated with larger PROVE gain (Spearman $\rho = 0.311$, 95% bootstrap CI $[0.069, 0.542]$, $n = 64$). **Right:** Median paired gain and perturbation-level bootstrap 95% confidence interval within LOCO-error quartiles. The secondary axis reports the PROVE win rate. The median gain increases from approximately 0.41 in the lowest-error quartile to 0.69 in the highest-error quartile, while PROVE maintains a high win rate in every quartile.

Table 10: Full results on the Frangieh21 dataset. Values are reported as mean ± standard deviation over three seeds. Boldface indicates the better mean between each vanilla model and its PROVE-enhanced counterpart.

| | scDFM | | CellFlow | | CFM-GP | |
|---|---|---|---|---|---|---|
| Metric | Vanilla | +PROVE | Vanilla | +PROVE | Vanilla | +PROVE |
| MAE ↓ | $0.1285 \pm 0.00039$ | $\mathbf{0.1014 \pm 0.00006}$ | $\mathbf{0.0184 \pm 0.00006}$ | $0.0224 \pm 0.00008$ | $0.0251 \pm 0.00007$ | $\mathbf{0.0225 \pm 0.00014}$ |
| Pearson-Δ ↑ | $-0.5337 \pm 0.00391$ | $\mathbf{0.4637 \pm 0.00161}$ | $0.8320 \pm 0.00142$ | $\mathbf{0.8583 \pm 0.00130}$ | $0.7983 \pm 0.00123$ | $\mathbf{0.8037 \pm 0.00298}$ |
| DE-Spearman ↑ | $-0.5618 \pm 0.00811$ | $\mathbf{0.5967 \pm 0.00191}$ | $0.9232 \pm 0.01336$ | $\mathbf{0.9281 \pm 0.00006}$ | $0.9071 \pm 0.00331$ | $\mathbf{0.9109 \pm 0.00241}$ |
| DS ↑ | $0.5269 \pm 0.00744$ | $\mathbf{0.6272 \pm 0.00140}$ | $0.5173 \pm 0.00713$ | $\mathbf{0.5878 \pm 0.00541}$ | $0.5372 \pm 0.00524$ | $\mathbf{0.6008 \pm 0.00529}$ |
| MMD-GEX ↓ | $0.1050 \pm 0.00074$ | $\mathbf{0.0805 \pm 0.00027}$ | $0.0036 \pm 0.00013$ | $\mathbf{0.0015 \pm 0.00009}$ | $0.0057 \pm 0.00011$ | $\mathbf{0.0046 \pm 0.00021}$ |
| MMD-PCA ↓ | $0.2712 \pm 0.00172$ | $\mathbf{0.1643 \pm 0.00071}$ | $0.0111 \pm 0.00081$ | $\mathbf{0.0091 \pm 0.00066}$ | $0.0227 \pm 0.00047$ | $\mathbf{0.0187 \pm 0.00043}$ |
| AUPRC-DE ↑ | $0.5022 \pm 0.00131$ | $\mathbf{0.5198 \pm 0.00056}$ | $0.6761 \pm 0.00110$ | $\mathbf{0.7260 \pm 0.00085}$ | $0.6175 \pm 0.00039$ | $\mathbf{0.6287 \pm 0.00314}$ |
| DEG Recall ↑ | $0.2362 \pm 0.00371$ | $\mathbf{0.7007 \pm 0.00097}$ | $0.6752 \pm 0.00210$ | $\mathbf{0.7021 \pm 0.00403}$ | $0.6368 \pm 0.00436$ | $\mathbf{0.6915 \pm 0.00140}$ |

Table 11: Full results on the Srivatsan20 dataset. Values are reported as mean ± standard deviation over three seeds. Boldface indicates the better mean between each vanilla model and its PROVE-enhanced counterpart.

| | scDFM | | CellFlow | | CFM-GP | |
|---|---|---|---|---|---|---|
| Metric | Vanilla | +PROVE | Vanilla | +PROVE | Vanilla | +PROVE |
| MAE ↓ | $0.1092 \pm 0.00026$ | $\mathbf{0.0598 \pm 0.00011}$ | $0.0137 \pm 0.00000$ | $\mathbf{0.0128 \pm 0.00001}$ | $0.0164 \pm 0.00001$ | $\mathbf{0.0147 \pm 0.00003}$ |
| Pearson-Δ ↑ | $-0.6440 \pm 0.00364$ | $\mathbf{0.8059 \pm 0.00100}$ | $0.8637 \pm 0.00065$ | $\mathbf{0.8772 \pm 0.00018}$ | $0.8228 \pm 0.00084$ | $\mathbf{0.8420 \pm 0.00032}$ |
| DE-Spearman ↑ | $-0.7094 \pm 0.00267$ | $\mathbf{0.8322 \pm 0.00041}$ | $0.8475 \pm 0.00121$ | $\mathbf{0.8754 \pm 0.00062}$ | $0.8156 \pm 0.00024$ | $\mathbf{0.8495 \pm 0.00032}$ |
| DS ↑ | $0.2052 \pm 0.00087$ | $\mathbf{0.8519 \pm 0.00191}$ | $0.8432 \pm 0.00069$ | $\mathbf{0.8597 \pm 0.00047}$ | $0.8422 \pm 0.00110$ | $\mathbf{0.8610 \pm 0.00068}$ |
| MMD-GEX ↓ | $0.1276 \pm 0.00032$ | $\mathbf{0.1219 \pm 0.00040}$ | $0.0119 \pm 0.00013$ | $\mathbf{0.0100 \pm 0.00007}$ | $0.1510 \pm 0.00019$ | $\mathbf{0.1240 \pm 0.00021}$ |
| MMD-PCA ↓ | $0.2975 \pm 0.00156$ | $\mathbf{0.2693 \pm 0.00021}$ | $\mathbf{0.0992 \pm 0.00065}$ | $0.1251 \pm 0.00088$ | $0.0722 \pm 0.00035$ | $\mathbf{0.0642 \pm 0.00066}$ |
| AUPRC-DE ↑ | $0.5314 \pm 0.00394$ | $\mathbf{0.7242 \pm 0.00025}$ | $0.6507 \pm 0.00119$ | $\mathbf{0.6690 \pm 0.00064}$ | $0.5936 \pm 0.00271$ | $\mathbf{0.6206 \pm 0.00125}$ |
| DEG Recall ↑ | $0.2799 \pm 0.00349$ | $\mathbf{0.5762 \pm 0.00171}$ | $0.6974 \pm 0.00090$ | $\mathbf{0.7523 \pm 0.00077}$ | $0.6666 \pm 0.00118$ | $\mathbf{0.7412 \pm 0.00284}$ |

Table 12: Ablation study of PROVE. EC, PC, and CD denote the corrections associated with effect collapse, perturbation confusion, and context drift, respectively. The remaining variants remove instantaneous-norm preservation or Pareto consensus. Values are reported as mean ± standard deviation over three seeds.

| Metric | w/o PC | w/o EC | w/o Norm Pres. | w/o CD | w/o Pareto | PROVE |
|---|---|---|---|---|---|---|
| MAE ↓ | $0.0688 \pm 0.00007$ | $0.0693 \pm 0.00013$ | $0.0703 \pm 0.00012$ | $0.0710 \pm 0.00013$ | $0.0712 \pm 0.00010$ | $\mathbf{0.0686 \pm 0.00012}$ |
| Pearson-Δ ↑ | $0.2196 \pm 0.00256$ | $0.1971 \pm 0.00484$ | $0.2131 \pm 0.00442$ | $0.1936 \pm 0.00445$ | $0.2053 \pm 0.00419$ | $\mathbf{0.2918 \pm 0.03057}$ |
| DE-Spearman ↑ | $0.2395 \pm 0.00926$ | $0.3356 \pm 0.01799$ | $0.2990 \pm 0.01353$ | $0.2943 \pm 0.01260$ | $0.3202 \pm 0.01309$ | $\mathbf{0.5707 \pm 0.01369}$ |
| DS ↑ | $0.5725 \pm 0.00334$ | $0.5931 \pm 0.00073$ | $0.5957 \pm 0.00137$ | $0.5679 \pm 0.00144$ | $0.5899 \pm 0.00182$ | $\mathbf{0.6253 \pm 0.00137}$ |
| MMD-GEX ↓ | $0.0172 \pm 0.00004$ | $0.0125 \pm 0.00010$ | $0.0128 \pm 0.00009$ | $0.0136 \pm 0.00009$ | $0.0122 \pm 0.00008$ | $\mathbf{0.0076 \pm 0.00009}$ |
| MMD-PCA ↓ | $0.0360 \pm 0.00038$ | $0.0458 \pm 0.00075$ | $0.0446 \pm 0.00071$ | $0.0443 \pm 0.00070$ | $0.0450 \pm 0.00069$ | $\mathbf{0.0271 \pm 0.00072}$ |
| AUPRC-DE ↑ | $0.2960 \pm 0.00161$ | $0.2855 \pm 0.00112$ | $0.2917 \pm 0.00116$ | $0.2915 \pm 0.00093$ | $0.2944 \pm 0.00105$ | $\mathbf{0.2975 \pm 0.00053}$ |
| DEG Recall ↑ | $0.2479 \pm 0.00043$ | $0.2616 \pm 0.00365$ | $0.2904 \pm 0.00794$ | $0.2143 \pm 0.00955$ | $0.2781 \pm 0.00693$ | $\mathbf{0.2953 \pm 0.00832}$ |

Table 13: Generalization of PROVE to PerturbDiff and Squidiff on Replogle. Values are reported as mean ± standard deviation over three seeds.

| Method | MAE↓ | Pearson-Δ↑ | DE-Spearman↑ | DS↑ | MMD-GEX↓ | MMD-PCA↓ | AUPRC-DE↑ | DEG Recall↑ |
|---|---|---|---|---|---|---|---|---|
| *PerturbDiff* | $0.0598 \pm 0.0002$ | $0.3265 \pm 0.0010$ | $0.6067 \pm 0.0067$ | $0.6099 \pm 0.00003$ | $0.0562 \pm 0.0005$ | $0.2045 \pm 0.0028$ | $0.1939 \pm 0.0006$ | $0.6976 \pm 0.0031$ |
| *+PROVE* | $\mathbf{0.0497 \pm 0.0001}$ | $\mathbf{0.3499 \pm 0.0005}$ | $\mathbf{0.7366 \pm 0.0036}$ | $\mathbf{0.6919 \pm 0.0025}$ | $\mathbf{0.0465 \pm 0.0004}$ | $\mathbf{0.1716 \pm 0.0020}$ | $\mathbf{0.2650 \pm 0.0021}$ | $\mathbf{0.7256 \pm 0.0021}$ |
| *Squidiff* | $0.5451 \pm 0.0008$ | $-0.0012 \pm 0.0009$ | $0.0824 \pm 0.0142$ | $0.5073 \pm 0.0004$ | $0.2096 \pm 0.0006$ | $0.2685 \pm 0.0007$ | $0.2351 \pm 0.0012$ | $0.1494 \pm 0.0001$ |
| *+PROVE* | $\mathbf{0.1417 \pm 0.0002}$ | $\mathbf{0.3291 \pm 0.0004}$ | $\mathbf{0.5744 \pm 0.0063}$ | $\mathbf{0.6574 \pm 0.0022}$ | $\mathbf{0.0718 \pm 0.0001}$ | $\mathbf{0.1283 \pm 0.0006}$ | $\mathbf{0.3011 \pm 0.0008}$ | $\mathbf{0.2360 \pm 0.0016}$ |

Table 14: Average pathway recovery results on HepG2 perturbations. Values are reported as mean ± standard deviation over three seeds.

| | scDFM | | CellFlow | | CFM-GP | |
|---|---|---|---|---|---|---|
| Metric | Vanilla | +PROVE | Vanilla | +PROVE | Vanilla | +PROVE |
| NES Spearman ↑ | $-0.0352 \pm 0.0037$ | $\mathbf{0.4279 \pm 0.0014}$ | $0.0322 \pm 0.0055$ | $\mathbf{0.3281 \pm 0.0031}$ | $0.3574 \pm 0.0016$ | $\mathbf{0.4044 \pm 0.0053}$ |
| Pathway Sign Agreement ↑ | $0.4503 \pm 0.0015$ | $\mathbf{0.7640 \pm 0.0008}$ | $0.5537 \pm 0.0053$ | $\mathbf{0.7606 \pm 0.0016}$ | $0.7635 \pm 0.0066$ | $\mathbf{0.8131 \pm 0.0067}$ |
| Perturbation MRR ↑ | $0.0125 \pm 0.0001$ | $\mathbf{0.0170 \pm 0.0005}$ | $0.0101 \pm 0.0004$ | $\mathbf{0.0533 \pm 0.0040}$ | $0.0216 \pm 0.0039$ | $\mathbf{0.0494 \pm 0.0005}$ |

## D LIMITATIONS AND FUTURE WORK.

PROVE requires the target perturbation to be observed in at least one non-target context. Additionally, its fixed effect-collapse threshold and isotropic context-drift region may lack flexibility under strong context-specific attenuation or heterogeneous variation. Future work includes designing uncertainty-aware, non-spherical reference constraints, incorporating population-level objectives, and generalizing PROVE to unseen perturbations via biological priors or embeddings.