BEYOND RECONSTRUCTION: LIKELIHOOD-FREE REINFORCEMENT LEARNING FOR VIRTUAL CELLS

Anonymous authors
Paper under double-blind review

ABSTRACT

Virtual cell models predict how cells respond to genetic perturbations. A central use of these predictions is deciding which perturbations can shift disease-relevant pathways in a desired direction and accelerate therapeutic target discovery. Because these models are trained to reconstruct gene-level expression, they can reduce reconstruction error without ranking the perturbations with the strongest pathway effects near the top. Pathway-level effects are defined over gene sets, relative to a background set, rather than individual genes so they are difficult to express as supervised losses but can be directly scored as rewards, which suggests benefits from reinforcement learning (RL). Standard RL, however, requires policy likelihoods, whereas virtual cell models generate high-dimensional cell populations without tractable likelihoods. We introduce Maximum Mean Discrepancy Policy Optimization (MMD-PO), a likelihood-free reinforcement learning method that weights rollouts sampled from a frozen policy by biological rewards, and moves the current policy toward this reward-weighted empirical target by minimizing maximum mean discrepancy between rollouts. We pair MMD-PO with Gene-Effect and Pathway-Enrichment rewards. In a controlled synthetic benchmark with known latent effects, MMD-PO improves held-out ranking over continued supervised training, GRPO, and online DPO, and its gains depend on feedback reliability, candidate diversity, and the headroom of the starting model. On held-out perturbations in HepG2 cells and zero-shot transfer from resting to stimulated primary T-Cells, MMD-PO improves top-1% hit enrichment over GRPO by 53.4% and 15.6%, respectively, while retaining prediction fidelity. These results establish likelihood-free policy optimization as a practical interface between virtual cell models and experimental prioritization.

1 INTRODUCTION

High-throughput single-cell perturbation assays reveal how genetic and chemical perturbations reshape cellular states (Dixit et al., 2016) and enable the discovery of disease-associated expression signatures and pathways. Yet the number of possible perturbations and cellular contexts is far too large for exhaustive *in vitro* experiments. Virtual cell models aim to narrow this search space by predicting responses to unmeasured perturbations and transferring predictions across contexts (Lotfollahi et al., 2023; Roohani et al., 2024; Bunne et al., 2023; Adduri et al., 2026). A central use of these predictions is a decision: which perturbations should be tested next? We call this *perturbation prioritization*. Given predicted responses for many candidate perturbations, the model should rank first the true hits that shift disease-relevant pathways in the desired direction.

This decision problem differs from the gene-level reconstruction losses used to train most virtual cell models. A model can improve its average prediction accuracy by fitting highly variable genes that are unrelated to disease-relevant pathways without improving the ranking of the strongest pathway hits from prior data, challenging its generalization for virtual screening and planning. Conversely, correctly promoting a few strong hits may barely change genome-wide reconstruction error. The gap matters most at the top of the ranking because, in practice, only a small fraction of candidates can be followed up on experimentally. We call this mismatch the *prediction-to-decision gap*.

A controlled synthetic study, in which the latent effect of every perturbation is known, makes this gap concrete (Section 5.4). Because ranking in the simulator is scored against latent effects rather than

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

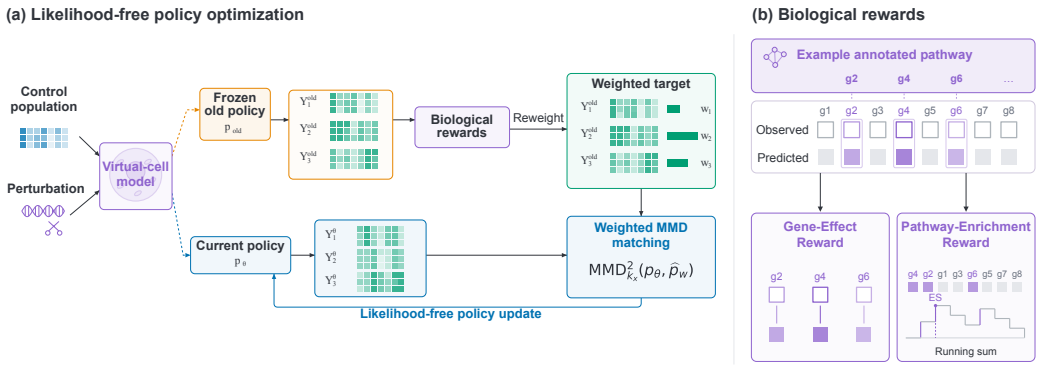


Figure 1: **(a)** MMD-PO. **(b)** Gene-Effect and Pathway-Enrichment rewards.

any reward statistic, it isolates whether an update improves the decision itself. Three observations motivate our approach. First, ranking quality and predictive fidelity separate. The reward-guided update that most improves held-out ranking also slightly lowers Pearson-Δ relative to its starting checkpoint. This suggests better reconstruction is neither necessary nor sufficient for better prioritization. Second, how the feedback is used matters. Given the same calibration data and update budget, reward-guided distribution matching improves held-out ranking over continued supervised training, likelihood-surrogate GRPO, and online DPO. Reward-guided post-training is therefore a tool for the prediction-to-decision gap.

Reinforcement learning (RL) is a natural way to optimize such task-specific biological feedback. However, virtual cell models generate high-dimensional and unordered cell populations without tractable output likelihoods. Standard RL methods such as PPO and GRPO therefore cannot be applied directly, since they require policy likelihood ratios (Schulman et al., 2017; Shao et al., 2024). Existing likelihood-free approaches remain tied to specific generators. DiffusionNFT and PerturbCellRL reward-weight diffusion training objectives, but optimize sample-wise denoising targets rather than directly matching output distributions (Zheng et al., 2026; Wu et al., 2026). DMPO formulates post-training as fitting a reward-tilted target, but implements this projection through denoising cross-entropy for diffusion language models (Zhu et al., 2025b). These limitations motivate a general question: *Can we align an implicit generator with a reward-tilted empirical distribution without policy likelihoods or generator-specific objectives?*

To address this question, we introduce **Maximum Mean Discrepancy Policy Optimization (MMD-PO)**, a likelihood-free method for reward-guided post-training. MMD-PO samples predictions from a frozen behavior policy and weights them according to their rewards, forming an empirical reward-tilted target distribution. MMD-PO then projects the current policy onto this target by minimizing maximum mean discrepancy in output space (Gretton et al., 2012; Li et al., 2015). The update differentiates through model samples rather than output likelihoods, so it supports non-differentiable rewards and implicit generators. In contrast to standard MMD training, which fits an observed data distribution, MMD-PO shifts the pretrained model toward the reward-tilted target while anchoring it to the supervised reference.

For single-cell perturbation prediction, we combine MMD-PO with two biological rewards. The Gene-Effect Reward measures agreement in per-gene pseudobulk perturbation effects over the union of annotated pathway genes, providing credit assignment to individual genes. The Pathway-Enrichment Reward measures agreement in pathway regulation relative to the background transcriptome (Subramanian et al., 2005). We evaluate decision quality separately, by measuring how many significant pathway hits are recovered within a fixed screening budget on held-out perturbations and contexts. Our contributions include:

- **Likelihood-free policy optimization for virtual cells.** We introduce MMD-PO, a sample-based projection onto a reward- and reference-weighted rollout target that requires neither tractable output likelihoods nor reward gradients, allowing implicit virtual cell models to move toward reward-tilted target distributions through distribution matching.

- **Biologically aligned reward design.** We develop complementary Gene-Effect and Pathway-Enrichment rewards that capture fine-grained transcriptional responses and aggregate pathway activity for perturbation prioritization.
- **Controlled evidence for when reward-guided post-training helps.** In synthetic worlds with known latent effects, MMD-PO outperforms continued SFT, GRPO, and online DPO. Gains depend on feedback reliability, candidate diversity, and checkpoint headroom.
- **Improved budgeted perturbation prioritization.** At a top-1% selection budget, MMD-PO improves hit enrichment relative to GRPO by 53.4% in few-shot HepG2 evaluations and 15.6% in zero-shot primary T-cell transfer, while retaining competitive cell-state prediction fidelity.

2 Related Work

Virtual-cell modeling and biological alignment Single-cell perturbation models span latent-shift and compositional approaches (Lotfollahi et al., 2019; 2023), graph- and optimal-transport-based methods (Roohani et al., 2024; Bunne et al., 2023), and large-scale systems such as State, X-Cell, and scDFM (Adduri et al., 2026; Wang et al., 2026; Yu et al., 2026). Most are trained to reconstruct expression or match cell distributions, which promotes predictive fidelity without directly optimizing biologically relevant pathway effects.

Reward-guided post-training and distribution matching PPO and GRPO use policy likelihood ratios to optimize reward (Schulman et al., 2017; Shao et al., 2024); DDPO and Flow-GRPO extend this approach to diffusion and flow models using sampling-transition likelihoods (Black et al., 2024; Liu et al., 2025). These updates are difficult to apply when a virtual-cell model's population-level likelihood is unavailable. Likelihood-free alternatives remain tied to specific generators: DiffusionNFT uses reward-weighted forward-process regression, LFPO modifies masked-diffusion denoising logits, and PerturbCellRL applies DiffusionNFT to a flow-matching virtual-cell model (Zheng et al., 2026; Wei et al., 2026; Wu et al., 2026).

MMD-PO builds on reward-weighted distribution matching, which constructs a target from high-reward samples and projects the policy toward it. REPS uses a KL-constrained projection, while DMPO uses weighted denoising cross-entropy for diffusion language models (Peters et al., 2010; Zhu et al., 2025b). kCGM uses MMD with a fixed external target and tractable sample or trajectory log-probabilities (Diamant & Trippe, 2026). In contrast, MMD-PO constructs its target online from reward-weighted rollouts and matches complete population samples through pathwise MMD, without output or transition likelihoods or reward gradients.

3 Preliminaries

Let $X^{\mathrm{ctrl}} \in \mathbb{R}^{S\times G}$ denote S control cells measured over G genes, and let t_{pert} denote a perturbation. Together they define a conditioning context $x = (X^{\mathrm{ctrl}}, t_{\mathrm{pert}})$ drawn from the training distribution $\mathcal{D}$. A virtual-cell model with parameters θ induces a conditional distribution $p_\theta(\cdot \mid x)$ over post-perturbation cell sets $Y = \{y_s\}_{s=1}^{S}$. We call each sampled cell set a *rollout*. For a given context, the model can generate multiple rollouts from $p_\theta(\cdot \mid x)$. Let $p_{\mathrm{ref}}(\cdot \mid x)$ denote the frozen supervised reference policy, and let $R(x, Y)$ be the biological reward assigned to a rollout.

3.1 Reward-tilted distribution matching

KL-regularized reward maximization is a standard formulation for improving a policy while controlling its deviation from a pretrained reference distribution in the field of reinforcement learning (Peters et al., 2010; Rafailov et al., 2023; Shao et al., 2024; Schulman et al., 2017). For each context x, it considers

$$\arg\max_{\theta}\ \mathbb{E}_{x\sim\mathcal{D}}\left[\mathbb{E}_{Y\sim p_\theta(\cdot|x)}\big[R(x,Y)\big] - \alpha D_{\mathrm{KL}}(p_\theta(\cdot \mid x)\,\|\,p_{\mathrm{ref}}(\cdot \mid x))\right]. \tag{1}$$

where $\alpha > 0$ controls the strength of the reference regularization. Prior work shows that the optimizer has the closed form (Rafailov et al., 2023; Zhu et al., 2025b).

$$p^\star(Y \mid x) = \frac{1}{Z(x)} p_{\mathrm{ref}}(Y \mid x) \exp\left(\frac{R(x,Y)}{\alpha}\right), \qquad Z(x) = \mathbb{E}_{Y \sim p_{\mathrm{ref}}(\cdot|x)}\left[\exp\left(\frac{R(x,Y)}{\alpha}\right)\right] \tag{2}$$

Thus, $p^\star$ preserves the support of the reference policy while assigning greater probability to high-reward populations. We provide the derivation in Appendix A.1.

Given this target, policy optimization can be viewed as distribution matching:

$$\min_\theta \ \mathbb{E}_{x \sim \mathcal{D}}\left[\mathcal{F}(p_\theta(\cdot \mid x), p^\star(\cdot \mid x))\right], \tag{3}$$

where $\mathcal{F}$ measures the discrepancy between the learned policy and the reward-tilted target. A standard choice for $\mathcal{F}$ is the reverse KL divergence $D_{\mathrm{KL}}(p_\theta \| p^\star)$ (Peters et al., 2010; Rafailov et al., 2023; Zhu et al., 2025b). Minimizing this reverse KL is equivalent to maximizing Eq. 1. The derivation details are in Appendix A.1. For an implicit virtual cell model, however, this projection requires computing likelihoods that cannot generally be evaluated. Section 4.3 instead uses a sample-based maximum mean discrepancy (MMD).

4 Maximum Mean Discrepancy Policy Optimization

Overview. MMD-PO instantiates Eq. 3 using MMD because the likelihoods needed for a KL projection are generally unavailable for implicit virtual-cell models. Each rollout is a population of cells, while the policy defines a distribution over such populations. As shown in Figure 1(a), a frozen old policy generates candidate rollouts. Biological rewards and a reference penalty weight these candidates to form an empirical target. A kernel compares pairs of cell populations, and MMD aggregates these comparisons to match current-policy rollouts to the weighted target. The update differentiates through current-policy samples without requiring likelihoods or reward gradients. We first derive the ideal objective, then present its practical weights and update.

4.1 Importance-Weighted MMD Projection

Consider an arbitrary positive-semidefinite kernel k_x on pairs of rollouts. For a fixed context x, let Y, Y' be independent draws from the current policy p_θ, and let W, W' be independent draws from the ideal reward-tilted target $p^\star$. Choosing $\mathcal{F} = \mathrm{MMD}^2_{k_x}$ in Eq. 3 gives

$$\mathrm{MMD}^2_{k_x}(p_\theta, p^\star) = \mathbb{E}_{Y,Y' \sim p_\theta}[k_x(Y,Y')] - 2\,\mathbb{E}_{\substack{Y \sim p_\theta \\ W \sim p^\star}}[k_x(Y,W)] + \mathbb{E}_{W,W' \sim p^\star}[k_x(W,W')]. \tag{4}$$

The variables W, W' are target rollouts, but direct draws from $p^\star$ are generally unavailable, so we estimate the target-side expectations by importance sampling. We freeze a copy of the current policy, $p_{\theta_{\mathrm{old}}}$, and draw N candidate rollouts $Y_j^{\mathrm{old}} \sim p_{\theta_{\mathrm{old}}}(\cdot \mid x)$, $j = 1, \ldots, N$. For any integrable f,

$$\mathbb{E}_{W \sim p^\star}[f(W)] = \mathbb{E}_{Y \sim p_{\theta_{\mathrm{old}}}}\left[\frac{p^\star(Y|x)}{p_{\theta_{\mathrm{old}}}(Y|x)} f(Y)\right].$$

Following Zhu et al. (2025b), we substitute Eq. 2 into the importance ratio $p^\star / p_{\theta_{\mathrm{old}}}$, so that the log ratio of candidate j equals $s_j^\star - \log Z(x)$, where

$$s_j^\star = \frac{R(x, Y_j^{\mathrm{old}})}{\alpha} + \log \frac{p_{\mathrm{ref}}(Y_j^{\mathrm{old}} \mid x)}{p_{\theta_{\mathrm{old}}}(Y_j^{\mathrm{old}} \mid x)}, \qquad w_j^\star = \frac{\exp(s_j^\star)}{\sum_{r=1}^N \exp(s_r^\star)}. \tag{5}$$

Because $Z(x)$ is shared by all candidates, it cancels when the weights are normalized, and self-normalized importance sampling gives $\mathbb{E}_{W \sim p^\star}[f(W)] \approx \sum_{j=1}^N w_j^\star f(Y_j^{\mathrm{old}})$. Equivalently, $p^\star$ is approximated by the discrete distribution $\hat{p}_{w^\star}$ that places probability $w_j^\star$ on each candidate Y_j^{old}. Replacing $p^\star$ with $\hat{p}_{w^\star}$ in Eq. 4 gives the importance-weighted MMD

$$\mathrm{MMD}^2_{k_x}(p_\theta, \hat{p}_{w^\star}) = \mathbb{E}[k_x(Y,Y')] - 2\sum_j w_j^\star\, \mathbb{E}[k_x(Y, Y_j^{\mathrm{old}})] + \sum_{j,j'} w_j^\star w_{j'}^\star\, k_x(Y_j^{\mathrm{old}}, Y_{j'}^{\mathrm{old}}), \tag{6}$$

where the expectations are over $Y, Y' \sim p_\theta$ as in Eq. 4. For a fixed current rollout Y, taking $f(W) = k_x(Y, W)$ gives the cross term, and applying the approximation to both W and W' gives the target–target term.

4.2 LIKELIHOOD-FREE MMD-PO UPDATE

The importance-weighted MMD in Eq. 6 still depends on the ideal weights in Eq. 5, which require a reference-to-old-policy likelihood ratio that is generally unavailable for implicit virtual-cell models. We instead construct practical weights from biological rewards and a reference penalty. For each candidate Y_j^{old}, we obtain a reference rollout Y_j^{ref} for the same context and compute

$$s_j = \frac{R_j}{\alpha} - \beta_{\mathrm{ref}}\, d_E(Y_j^{\mathrm{old}}, Y_j^{\mathrm{ref}}), \qquad w_j = \frac{\exp(s_j)}{\sum_{r=1}^{N} \exp(s_r)}, \tag{7}$$

where $R_j = R(x, Y_j^{\mathrm{old}})$, $\beta_{\mathrm{ref}} \geq 0$ controls the reference penalty, and d_E is the energy distance between the empirical cell distributions of two rollouts. The resulting weights favor high-reward candidates close to their reference predictions. They provide a practical surrogate for the ideal weights, rather than an exact estimate of the likelihood ratio.

Replacing $w_j^\star$ with w_j in Eq. 6 gives the practical weighted MMD, $\mathrm{MMD}^2_{k_x}(p_\theta, \hat{p}_w)$, where $\hat{p}_w$ places probability w_j on each candidate Y_j^{old}. Estimating its current-policy expectations with N rollouts $Y_i^\theta \sim p_\theta(\cdot \mid x)$ gives its empirical estimate

$$\widehat{\mathrm{MMD}}^2_{k_x} = \frac{1}{N^2}\sum_{i,i'} k_x(Y_i^\theta, Y_{i'}^\theta) - \frac{2}{N}\sum_{i,j} w_j\, k_x(Y_i^\theta, Y_j^{\mathrm{old}}) + \sum_{j,j'} w_j w_{j'}\, k_x(Y_j^{\mathrm{old}}, Y_{j'}^{\mathrm{old}}). \tag{8}$$

The three terms compare current rollouts with one another, current rollouts with the weighted candidates, and the weighted candidates with one another; Appendix A.2 gives the full expansion. The rewards enter only through the second term, which pulls current rollouts toward high-weight candidates; the first term penalizes current rollouts that are too similar to one another, discouraging collapse onto a single candidate. The third term depends only on the candidates, their weights, and the kernel bandwidth, all of which are fixed during the update, so dropping it leaves the gradient unchanged and gives the MMD-PO loss:

$$\mathcal{L}_{\text{MMD-PO}}(\theta; x) = \frac{1}{N^2}\sum_{i,i'} k_x(Y_i^\theta, Y_{i'}^\theta) - \frac{2}{N}\sum_{i,j} w_j\, k_x(Y_i^\theta, Y_j^{\mathrm{old}}). \tag{9}$$

We instantiate k_x as a Gaussian kernel over energy distance, $k_x(A, B) = \exp[-d_E(A, B)/(2h_x^2)]$ for any two rollouts A and B, where $h_x > 0$ is a context-specific bandwidth. The kernel is invariant to the ordering of cells within either rollout; its construction and properties are given in Appendix A.3. During the update, we differentiate through current-policy rollouts while holding the old-policy rollouts, their weights, and the bandwidth fixed. This requires neither policy likelihoods nor gradients through the reward. Algorithm 1 summarizes the procedure details.

4.3 BIOLOGICAL REWARD FUNCTIONS

For a context $x = (X^{\mathrm{ctrl}}, t_{\mathrm{pert}})$, let Y^{obs} denote the measured post-perturbation population. We write $\bar{X}_g^{\mathrm{ctrl}}$, $\bar{y}_g$, and $\bar{y}_g^{\mathrm{obs}}$ for the mean expression of gene g in the control, predicted, and observed populations, respectively. On the log-expression scale used by the model, we define the predicted and observed pseudobulk perturbation effects as $\Delta_g^{\mathrm{pred}} = \bar{y}_g - \bar{X}_g^{\mathrm{ctrl}}$ and $\Delta_g^{\mathrm{obs}} = \bar{y}_g^{\mathrm{obs}} - \bar{X}_g^{\mathrm{ctrl}}$. The reward combines fine-grained agreement over pathway-associated genes with pathway-level agreement in coordinated transcriptional activity.

Gene-Effect Reward. Let $\mathcal{G}_\ell$ be the gene set of pathway ℓ and $\mathcal{U} = \bigcup_{\ell=1}^{P} \mathcal{G}_\ell$. We measure agreement over the union of pathway-associated genes with a Huber loss H_δ of threshold δ,

$$R_{\mathrm{gene}}(x, Y) = -\frac{1}{|\mathcal{U}|}\sum_{g\in\mathcal{U}} H_\delta\big(\Delta_g^{\mathrm{pred}} - \Delta_g^{\mathrm{obs}}\big), \qquad H_\delta(e) = \begin{cases} \frac{1}{2}e^2, & |e| \leq \delta, \\ \delta\big(|e| - \frac{1}{2}\delta\big), & |e| > \delta. \end{cases} \tag{10}$$

This aligns individual pathway-gene responses while capping the influence of unusually large errors.

Pathway-Enrichment reward. Gene-wise agreement on pathway genes may miss their response relative to the measured background. GSEA (Subramanian et al., 2005) ranks all measured genes by perturbation effect, so its enrichment score reflects where pathway genes fall relative to non-pathway genes. Applying GSEA to the predicted and observed effect vectors gives $\widehat{\mathrm{ES}}_\ell$ and ES_ℓ for each selected pathway ℓ. We penalize differences outside a relative tolerance band τ:

$$R_{\mathrm{path}}(x, Y) = -\sum_{\ell=1}^{P} a_\ell \max\Big(\Big|\widehat{\mathrm{ES}}_\ell - \mathrm{ES}_\ell\Big| - \tau\,|\mathrm{ES}_\ell|\,,\ 0\Big)^2, \tag{11}$$

Here $a_\ell = \omega_\ell / \sum_{r=1}^{P} \omega_r$, with $\omega_\ell = \sum_{g \in \mathcal{G}_\ell} 1/m_g$, and m_g is the number of selected pathways containing gene g. These weights account for overlap between pathway gene sets.

Combined multiple rewards. For each perturbation context, we independently standardize the two rewards across the N rollouts from that context. The combined reward for each rollout is the sum of its two standardized scores, which places the components on comparable scales without requiring a manually tuned mixing coefficient.

5 Experiments

5.1 Experimental Setup

Datasets and pathway panels. We evaluate MMD-PO on Replogle–Nadig (Replogle et al., 2022; Nadig et al., 2025) and Marson (Zhu et al., 2025a) dataset on protein-coding highly variable genes (HVGs). Following the X-Cell protocol (Wang et al., 2026), we train on multiple cell lines in Replogle–Nadig and evaluate held-out perturbation–context pairs in HepG2. Training includes HepG2 cells and may include the same perturbations in other contexts, but excludes the held-out responses in HepG2, defining *few-shot perturbation generalization*. For Marson, we train on resting donor D3 CD4$^+$ T cells and evaluate the same perturbations after CD3/CD28 stimulation. Stimulated perturbed profiles are excluded from training and post-training, defining *zero-shot context transfer*. The pathway panels contain 37 MSigDB Hallmark pathways for Replogle–Nadig and 17 MSigDB/KEGG pathways for Marson (Liberzon et al., 2015; Kanehisa et al., 2016). Dataset and pathway-selection details appear in Appendix B.

Baselines. We evaluate X-Cell (Wang et al., 2026), scGPT (Cui et al., 2024), and scDFM (Yu et al., 2026). For each backbone, we compare its fine-tuned initialization with continued SFT, a policy-gradient baseline, and MMD-PO. Post-training methods start from the same checkpoint and use matched numbers of optimizer steps. We use GRPO (Shao et al., 2024) for X-Cell and scGPT, and Flow-GRPO (Liu et al., 2025) for scDFM. Implementation details appear in Appendix C.

Evaluation metrics. In order to evaluate the gains from MMD-PO, we assess virtual cell models on two complementary axes: perturbation prioritization and predictive fidelity. Perturbation prioritization asks whether predicted responses place the held-out candidates with the strongest measured pathway effects near the top of the ranking. Predictive fidelity asks whether these gains preserve the accuracy of predicted transcriptional responses. Together, these metrics evaluate whether MMD-PO achieves the goal of prioritizing promising perturbations while preserving the accurate perturbation predictions of virtual cell models.

Perturbation prioritization. We measure pathway effects using gene set enrichment analysis (GSEA), which scores whether a pathway's genes lie toward the up- or down-regulated end of a gene-response ranking relative to background genes (Subramanian et al., 2005). Pathway effects are relevant to follow-up decisions made at the level of biological programs. For each pathway, we rank candidates by the magnitude of their predicted enrichment score (ES), capturing both activation and suppression. We evaluate the top 1% and 5%, representing limited follow-up screening budgets.

We report three metrics. Enrichment factor (EF) measures how many significant, high-effect hits the shortlist recovers relative to random selection. Adjusted normalized discounted cumulative gain (AnDCG) (Jeunen et al., 2024) also assesses their order, rewarding relevant candidates near the top. Rank AnDCG uses discrete tiers of observed effect strength, while Geometric AnDCG uses a continuous score combining effect strength and significance. Formal definitions appear in Appendix E.2.

Table 1: Perturbation prioritization and predictive fidelity across virtual-cell backbones and post-training methods. Bold and underlined values indicate the best and second-best distinct displayed values within each backbone and dataset, respectively.

Model	Rank AnDCG ↑		Geometric AnDCG ↑		Enrichment Factor ↑		Predictive Fidelity		
	1%	5%	1%	5%	1%	5%	Pearson-Δ ↑	Centroid Accuracy ↑	Pseudobulk MAE ↓
Replogle–Nadig HepG2									
X-Cell	0.294	0.280	0.369	0.368	5.942	4.580	0.531	0.828	0.060
+ Continued SFT	0.340	0.302	0.397	0.396	5.371	4.903	0.528	**0.837**	0.060
+ GRPO	0.376	0.323	0.447	0.422	6.714	5.026	0.536	0.826	0.057
+ MMD-PO	**0.425**	**0.339**	**0.481**	**0.445**	**10.300**	**5.610**	**0.538**	0.815	**0.054**
scGPT	0.094	0.075	0.179	0.139	2.711	1.715	0.276	0.554	0.138
+ Continued SFT	0.089	0.073	0.186	0.141	2.940	1.649	0.277	0.553	0.139
+ GRPO	0.153	0.117	0.266	0.204	3.453	2.310	0.275	0.573	0.124
+ MMD-PO	**0.284**	**0.227**	**0.364**	**0.326**	**4.755**	**3.655**	**0.307**	**0.697**	**0.073**
scDFM	0.064	0.063	0.108	0.110	1.959	1.847	**0.348**	**0.577**	0.093
+ Continued SFT	0.060	0.042	0.132	0.089	1.565	1.203	0.279	0.507	0.081
+ Flow-GRPO	**0.168**	0.091	**0.234**	0.160	3.265	2.428	0.269	0.552	0.075
+ MMD-PO	0.165	**0.120**	0.178	**0.169**	**6.565**	**3.015**	0.342	0.564	**0.074**
Marson primary CD4$^+$ T cells									
X-Cell	0.083	0.052	0.106	0.076	1.903	1.533	0.058	0.504	0.592
+ Continued SFT	0.101	0.065	0.148	0.098	2.114	1.668	0.060	0.506	0.594
+ GRPO	0.104	0.077	0.150	0.117	2.214	1.839	0.058	0.502	0.479
+ MMD-PO	**0.128**	**0.081**	**0.175**	**0.119**	**2.560**	**1.854**	**0.060**	**0.509**	**0.414**
scGPT	0.032	0.022	**0.063**	**0.038**	1.270	1.211	0.036	0.501	0.354
+ Continued SFT	0.029	0.016	0.052	0.029	1.308	1.193	0.034	0.501	0.364
+ GRPO	0.033	0.025	0.045	0.033	**1.359**	**1.263**	0.039	0.501	0.339
+ MMD-PO	**0.036**	**0.025**	0.055	0.035	1.331	1.253	**0.057**	**0.517**	**0.122**
scDFM	0.077	0.044	0.110	0.067	1.775	1.429	**0.058**	**0.501**	0.466
+ Continued SFT	0.074	0.043	0.101	0.064	1.836	1.412	0.057	0.501	0.463
+ Flow-GRPO	0.067	0.042	0.098	0.065	1.650	1.425	0.049	0.501	**0.450**
+ MMD-PO	**0.080**	**0.048**	**0.120**	**0.073**	**1.993**	**1.503**	0.051	0.500	0.480

Predictive fidelity. We separately compare predicted and measured responses across the evaluated gene panel using Cell-Eval metrics, including Pearson-Δ, Centroid Accuracy, and pseudobulk MAE (Adduri et al., 2026). These assess whether improved prioritization comes with reduced transcriptional accuracy. Detailed definitions appear in Appendix E.1.

5.2 Main Results

Table 1 summarizes performance in the few-shot HepG2 and zero-shot primary T-cell settings. For X-Cell, MMD-PO achieves the best prioritization on all six metrics in both settings. At the 1% budget, it raises EF over GRPO from 6.714 to 10.300 on HepG2 and from 2.214 to 2.560 on primary T cells, gains of 53.4% and 15.6%. These gains shrink at the 5% budget, indicating that reward weighting mainly reorders the strongest candidates, where a small screening budget is spent. Continued SFT does not reproduce these gains: on HepG2, it raises X-Cell's centroid accuracy from 0.828 to 0.837 but lowers EF@1% from 5.942 to 5.371, illustrating the prediction-to-decision gap.

MMD-PO improves perturbation prioritization across all three backbones. On HepG2, it roughly triples scGPT's rank AnDCG@1% from 0.094 to 0.284, compared with 0.153 for GRPO, and doubles scDFM's EF@1% relative to Flow-GRPO from 3.265 to 6.565. For scDFM, the margin over Flow-GRPO is smaller. One likely reason is how scDFM generates populations: each cell is produced independently by integrating a flow ODE from noise, conditioned on a single control cell (Yu et al., 2026). Candidate populations for the same context therefore differ mainly through per-cell noise, which largely averages out in the pseudobulk profiles scored by our rewards, so the population kernel separates candidates only weakly on the features that determine their rewards. Flow-GRPO instead assigns credit to individual sampling transitions through likelihood ratios. Even so, MMD-PO outperforms Flow-GRPO on 10 of 12 prioritization metrics, leads all three 5%-budget metrics in

Table 2: Ablation of supervised objectives and MMD-PO rewards on Replogle–Nadig HepG2.

Method	AnDCG Rank ↑		AnDCG Geo ↑		EF ↑	
	1%	5%	1%	5%	1%	5%
Supervised-objective ablation						
Continued SFT	0.340	0.302	0.397	0.396	5.371	4.903
Pathway-gene SFT	0.211	0.157	0.320	0.258	3.991	2.764
Reward-weighted SFT	0.255	0.185	0.330	0.277	4.463	3.343
MMD-PO reward ablation						
Gene-Effect reward	0.418	0.336	0.465	0.435	9.210	5.550
Pathway-Enrichment reward	0.394	0.334	0.447	0.445	6.200	5.160
Combined reward	**0.425**	**0.339**	**0.481**	**0.445**	**10.300**	**5.610**

both settings, and achieves higher Pearson-Δ, showing that a single backbone-agnostic update can outperform an RL method designed specifically for flow models. The full Cell-Eval metrics are in Appendix F.1.

MMD-PO improves data efficiency. MMD-PO consistently outperforms SFT when both methods use the same amount of training data and achieves comparable performance with substantially fewer examples. As shown in Figure 2, MMD-PO trained on 60% of the data nearly matches the rank AnDCG of SFT trained on 80%, while MMD-PO at 40% similarly approaches the geometric AnDCG of SFT at 60%. These results indicate that MMD-PO can attain similar perturbation-ranking quality with less training data, demonstrating improved data efficiency over SFT.

5.3 Ablation Studies

We first examine whether biological rewards improve perturbation prioritization when incorporated into supervised training. All variants use X-Cell on the same Replogle–Nadig HepG2 split. Continued SFT extends training with the original supervised objective. Pathway-gene SFT uses a supervised loss focused on pathway-associated genes, while reward-weighted SFT applies reward-dependent weights to the supervised objective. The implementation details are in Appendix F.2. These comparisons test whether incorporating biological rewards into supervised training can match the gains from using them to guide MMD-PO's rollout-based policy updates. As shown in Table 2, both pathway-gene SFT and reward-weighted SFT underperform continued SFT on all six prioritization metrics, whereas every MMD-PO variant outperforms the supervised controls. The corresponding results in Table 7 show that both reward-informed SFT variants also reduce predictive fidelity. Thus, incorporating these signals into the tested supervised objectives does not bring the gains from MMD-PO.

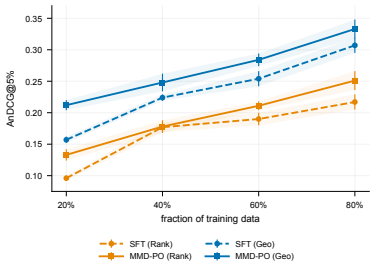


Figure 2: Data efficiency vs. SFT.

We next assess the contribution of each reward component to MMD-PO. The Gene-Effect and Pathway-Enrichment rewards each improve all six prioritization metrics over the supervised controls. Their combination achieves the best or tied-best result on every metric, suggesting that gene-level effect agreement and pathway-level enrichment provide complementary signals for perturbation prioritization.

5.4 Controlled Synthetic Study

Empirical benchmarks leave two questions open. Measured pathway effects are noisy and the benchmarks cannot vary the quality of the feedback or the strength of the starting model. We therefore

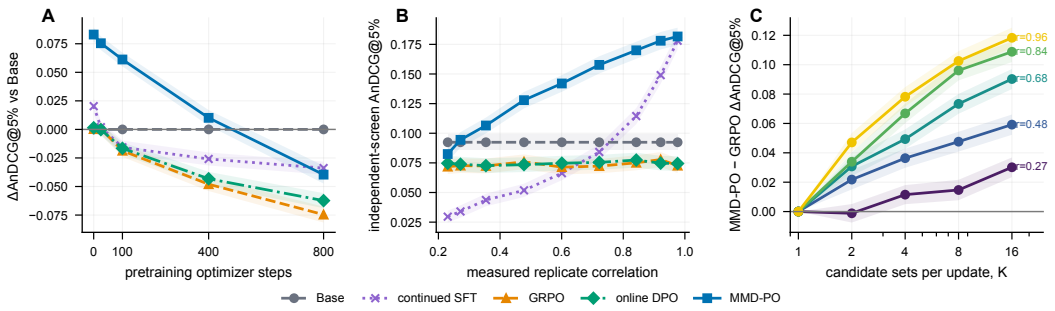


Figure 3: Synthetic AnDCG@5%. (**A**) Gain over starting checkpoint vs. pretraining. (**B**) Score vs. calibration reliability. (**C**) MMD-PO−GRPO vs. candidate count K.

construct a synthetic benchmark with known latent effects to examine when reward-guided post-training improves ranking. In each simulated world, a structural teacher generates noisy measurements for 64 calibration perturbations and effects for 256 unseen evaluation perturbations. Methods use the calibration measurements for post-training, then rank the unseen perturbations. Within each world and starting-checkpoint condition, all methods share the same checkpoint and receive 40 update steps. Noise-free calibration targets and evaluation effects are withheld during post-training. We repeat the comparison across 2,048 independently generated worlds.

We calculate AnDCG@5% for each method within each world and average the within-world differences across worlds. MMD-PO gains 0.0799 over Gaussian-surrogate GRPO, 0.0778 over online DPO, and 0.0769 over continued SFT. Its advantage over GRPO remains positive when evaluated in held-out synthetic cellular contexts (0.0422).

Figure 3 shows when these gains arise. In panel A, MMD-PO's change from its starting checkpoint falls from $+0.083$ at the weakest checkpoint to -0.040 at the strongest, although the strongest retains the best absolute ranking. Panel B shows higher screen performance with more reproducible calibration measurements. In panel C, MMD-PO's advantage over GRPO grows with candidate count K, especially when measurements are reliable, although larger K also increases computation. The results point to the importance of reliable feedback, candidate choice, and room for improvement in the starting model. These controlled simulations do not provide additional biological validation.

6 Conclusion and Limitations

We propose MMD-PO, a likelihood-free RL framework for post-training virtual-cell models that constructs a reward-weighted target from generated cell populations and optimizes it through output-space MMD. This enables population-level policy updates across different generative architectures without tractable output likelihoods. On X-Cell, MMD-PO improves top-1% hit enrichment over adapted GRPO by 53.4% on HepG2 and 15.6% in stimulated T-cell transfer while retaining competitive predictive fidelity, providing evidence that the ranking gains do not come from sacrificing expression-level predictive fidelity to hack the rewards. Our main limitation is that perturbation prioritization was evaluated only on selected pathway responses; these retrospective results may therefore not generalize to other pathways or biological objectives. Future work includes applying MMD-PO to larger pretrained virtual cell models, extending pathway rewards to combinatorial perturbations, developing verifiers informed by functional assays or expert feedback, and testing MMD-PO in prospective screens. By aligning virtual cell models with the pathway-level effects that prioritize disease-modifying perturbations, MMD-PO moves these models from reconstructing expression toward prioritizing candidates for therapeutic target discovery.

AI USE STATEMENT

In this work, we used generative AI tools to provide feedback on research methodology and experiments, implement methods, and interpret results. We have not used generative AI tools to generate synthetic data, develop theoretical models or conceptual frameworks, clean or reformat datasets, or support qualitative. Translation is not applicable to this work. Additionally, we have used generative AI tools to assist with experimental-method feedback, software implementation and review, result analysis, figure preparation, proof verification and drafting and editing the manuscript. The authors executed the experiments, checked the generated code and analyses against the experiment logs, reviewed all AI-assisted text and claims, and take responsibility for the the complete submission and final content of this work.

REPRODUCIBILITY STATEMENT

We have made several efforts to ensure the reproducibility of our results. Appendix B specifies the dataset splits and pathway selection rules. Appendix D gives hyperparameter and optimization settings. Appendix E defines the evaluation metrics. All tables and figures are generated by a single script from logged training runs. Upon publication, we will publicly release code and configuration files needed to reproduce experiments.

ETHICS STATEMENT

This work analyzes previously published single-cell perturbation data, including profiles from primary human CD4+ T cells. The evaluations are retrospective: predicted perturbation rankings have not been validated in prospective screens and should not be interpreted as clinical or therapeutic recommendations. The original dataset publications provide the provenance of the underlying human-derived data.

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*, 2026. doi: 10.1016/j.cell.2026.07.052. URL https://pubmed.ncbi.nlm.nih.gov/42673963/. Published online 31 August 2026.

Joshua D. Backman, Alexander H. Li, Anthony Marcketta, et al. Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature*, 599(7886):628–634, 2021. doi: 10.1038/s41586-021-04103-z.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training Diffusion Models with Reinforcement Learning. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2305.13301.

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. URL https://www.nature.com/articles/s41592-023-01969-x.

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.

Edward De Brouwer, Carl Edwards, Alexander Wu, Jenna Collier, Graham Heimberg, Xiner Li, Meena Subramaniam, Ehsan Hajiramezanali, David Richmond, Jan-Christian Hütter, et al. Assaybench: An assay-level virtual cell benchmark for llms and agents. *arXiv preprint arXiv:2605.10876*, 2026.

Nathaniel L. Diamant and Brian L. Trippe. Calibrating Generative Models to Feature Distributions with MMD Finetuning. *arXiv preprint arXiv:2606.19496*, 2026. URL https://arxiv.org/html/2606.19496v1.

Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P. Fulco, Livnat Jerby-Arnon, Nemanja D. Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, Britt Adamson, Thomas M. Norman, Eric S. Lander, Jonathan S. Weissman, Nir Friedman, and Aviv Regev. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell*, 167:1853–1866.e17, 2016. doi: 10.1016/j.cell.2016.11.038. URL https://pubmed.ncbi.nlm.nih.gov/27984732/.

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.

Olivier Jeunen, Ivan Potapov, and Aleksei Ustimenko. On (normalised) discounted cumulative gain as an off-policy evaluation metric for top-n recommendation. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 1222–1233, 2024.

Minoru Kanehisa, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. Kegg as a reference resource for gene and protein annotation. *Nucleic acids research*, 44(D1):D457–D462, 2016.

Yujia Li, Kevin Swersky, and Rich Zemel. Generative moment matching networks. In *International conference on machine learning*, pp. 1718–1727. PMLR, 2015.

Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P. Mesirov, and Pablo Tamayo. The molecular signatures database hallmark gene set collection. *Cell Systems*, 1(6): 417–425, 2015. doi: 10.1016/j.cels.2015.12.004.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-GRPO: Training Flow Matching Models via Online RL. *arXiv preprint arXiv:2505.05470*, 2025. URL https://arxiv.org/abs/2505.05470.

Mohammad Lotfollahi, F. Alexander Wolf, and Fabian J. Theis. scGen predicts single-cell perturbation responses. *Nature Methods*, 16:715–721, 2019. doi: 10.1038/s41592-019-0494-8. URL https://www.nature.com/articles/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. URL https://link.springer.com/article/10.15252/msb.202211517.

Ajay Nadig, Joseph M. Replogle, Andrew N. Pogson, Meena 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.

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.

Jan Peters, Katharina Mülling, and Yasemin Altun. Relative entropy policy search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 24, pp. 1607–1612, 2010.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Joseph M. Replogle, Reuben A. Saunders, Andrew N. Pogson, Jeffrey A. Hussmann, Alexander Lenail, Alina Guna, Luisa Mascibroda, Eric J. Wagner, Karen Adelman, Gila Lithwick-Yanai, et al. Mapping information-rich genotype-phenotype landscapes with genome-scale perturb-seq. *Cell*, 185(14):2559–2575, 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. URL https://www.nature.com/articles/s41587-023-01905-6.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The annals of statistics*, pp. 2263–2291, 2013.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

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, et al. 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.

Chloe Wang, Mehran Karimzadeh, Neal G Ravindra, Lexi R Bounds, Nader Alerasool, Ann C Huang, Shihao Ma, Daniel R Gulbranson, Haotian Cui, Yongju Lee, et al. X-cell: Scaling causal perturbation prediction across diverse cellular contexts via diffusion language models. *bioRxiv*, pp. 2026–03, 2026.

Chenxing Wei, Jiazhen Kang, Hong Wang, Jianqing Zhang, Hao Jiang, Xiaolong Xu, Ningyuan Sun, Ying He, F. Richard Yu, Yao Shu, and Bo Jiang. LFPO: Likelihood-Free Policy Optimization for Masked Diffusion Models. *arXiv preprint arXiv:2603.01563*, 2026. URL https://arxiv.org/abs/2603.01563.

Dongxia Wu, Mingyu Li, Yuhui Zhang, Anurendra Kumar, Emma Lundberg, Serena Yeung-Levy, and Emily B. Fox. PerturbCellRL: Verifier-Guided Reinforcement Learning for Single-Cell Perturbation Prediction. *arXiv preprint arXiv:2606.27752*, 2026. URL https://arxiv.org/abs/2606.27752.

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.

Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. DiffusionNFT: Online Diffusion Reinforcement with Forward Process. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2509.16117.

Rui Zhu, Emma Dann, Jing Yan, Jonathan R. Retana, Ryunosuke Goto, Romain C. Guitche, Laura K. Petersen, Motohiro Ota, Jonathan K. Pritchard, and Alexander Marson. Genome-scale Perturb-seq in primary human CD4+ T cells maps context-specific regulators of T cell programs and human immune traits. *bioRxiv*, 2025a. Preprint.

Yuchen Zhu, Wei Guo, Jaemoo Choi, Petr Molodyk, Bo Yuan, Molei Tao, and Yongxin Chen. Enhancing reasoning for diffusion llms via distribution matching policy optimization. *arXiv preprint arXiv:2510.08233*, 2025b.

A Additional Derivations and Implementation Details

A.1 Derivation of the reward-tilted optimum

We adapt the derivation of the KL-regularized policy optimum from Rafailov et al. (2023, Appendix A.1) to conditional distributions over cell populations. Starting from Eq. 1 and expanding the KL divergence, we obtain

$$\begin{aligned}\theta^\star &\in \arg\max_\theta \ \mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{Y\sim p_\theta(\cdot|x)}\left[R(x,Y) - \alpha \log \frac{p_\theta(Y \mid x)}{p_{\mathrm{ref}}(Y \mid x)}\right] \\ &= \arg\min_\theta \ \mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{Y\sim p_\theta(\cdot|x)}\left[\log \frac{p_\theta(Y \mid x)}{p_{\mathrm{ref}}(Y \mid x)} - \frac{R(x,Y)}{\alpha}\right] \\ &= \arg\min_\theta \ \mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{Y\sim p_\theta(\cdot|x)}\left[\log \frac{p_\theta(Y \mid x)}{p_{\mathrm{ref}}(Y \mid x)\exp(R(x,Y)/\alpha)}\right]. \end{aligned} \tag{12}$$

The denominator in the last line is a nonnegative, unnormalized density over population predictions. Its context-dependent normalizing constant is

$$Z(x) = \int_{\mathcal{Y}} p_{\mathrm{ref}}(Y \mid x)\exp\left(\frac{R(x,Y)}{\alpha}\right)\mathrm{d}Y, \tag{13}$$

where $\mathcal{Y}$ is the population-output space. Assuming $0 < Z(x) < \infty$, normalizing the denominator gives the reward-tilted distribution

$$p^\star(Y \mid x) = \frac{1}{Z(x)} p_{\mathrm{ref}}(Y \mid x)\exp\left(\frac{R(x,Y)}{\alpha}\right). \tag{14}$$

We can now substitute

$$p_{\mathrm{ref}}(Y \mid x)\exp\left(\frac{R(x,Y)}{\alpha}\right) = Z(x)p^\star(Y \mid x) \tag{15}$$

into Eq. 12. This yields

$$\begin{aligned}\theta^\star &\in \arg\min_\theta \ \mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{Y\sim p_\theta(\cdot|x)}\left[\log \frac{p_\theta(Y \mid x)}{Z(x)p^\star(Y \mid x)}\right] \\ &= \arg\min_\theta \ \mathbb{E}_{x\sim\mathcal{D}}\left[D_{\mathrm{KL}}(p_\theta(\cdot \mid x) \,\|\, p^\star(\cdot \mid x)) - \log Z(x)\right]. \end{aligned} \tag{16}$$

Because $Z(x)$ does not depend on θ, the normalizing term does not affect the optimizer. Therefore,

$$\theta^\star \in \arg\min_\theta \ \mathbb{E}_{x\sim\mathcal{D}}\left[D_{\mathrm{KL}}(p_\theta(\cdot \mid x) \,\|\, p^\star(\cdot \mid x))\right]. \tag{17}$$

If $p^\star$ is representable by the policy family, nonnegativity of KL divergence implies $p_{\theta^\star}(\cdot \mid x) = p^\star(\cdot \mid x)$ for $\mathcal{D}$-almost every x. Otherwise, the solution is the reverse-KL projection of $p^\star$ onto the distributions representable by p_θ.

A.2 Derivation of the Weighted MMD Objective

For a fixed context x, let $\widehat{p}_w$ assign probability w_j to each old-policy population Y_j^{old}, where $\sum_j w_j = 1$. Expectations under this distribution are weighted sums. In particular,

$$\begin{aligned}\mathbb{E}_{W\sim\widehat{p}_w}[k_x(Y,W)] &= \sum_{j=1}^N w_j k_x(Y, Y_j^{\mathrm{old}}), \\ \mathbb{E}_{W,W'\sim\widehat{p}_w}[k_x(W,W')] &= \sum_{j,j'=1}^N w_j w_{j'} k_x(Y_j^{\mathrm{old}}, Y_{j'}^{\mathrm{old}}),\end{aligned} \tag{18}$$

where W and W' are drawn independently. Substituting these expressions into Eq. 4 gives

$$\begin{aligned}\mathrm{MMD}^2_{k_x}(p_\theta, \widehat{p}_w) = \mathbb{E}_{Y,Y'\sim p_\theta}[k_x(Y,Y')] \\ - 2\sum_{j=1}^N w_j \mathbb{E}_{Y\sim p_\theta}[k_x(Y, Y_j^{\mathrm{old}})] + C_x,\end{aligned} \tag{19}$$

where $C_x = \sum_{j,j'} w_j w_{j'} k_x(Y_j^{\text{old}}, Y_{j'}^{\text{old}})$.

Using the empirical distribution of the current-policy samples $\{Y_i^\theta\}_{i=1}^N$ to evaluate the remaining expectations yields Eq. 8. Consequently,

$$\widehat{\text{MMD}}^2_{k_x}(p_\theta, \widehat{p}_w) = \mathcal{L}_{\text{MMD-PO}}(\theta; x) + C_x. \tag{20}$$

The old-policy populations, their weights, and the kernel bandwidth are held fixed during each update, so C_x is independent of θ. The loss in Eq. 9 therefore has the same gradients and minimizers as the full empirical MMD objective.

A.3 Population Kernel

Kernel construction. For cell populations $Y = \{y_s\}_{s=1}^S$ and $W = \{z_t\}_{t=1}^T$, with expression vectors in the same gene space, we define

$$\begin{aligned} d_E(Y, W) = \frac{2}{ST} \sum_{s=1}^{S} \sum_{t=1}^{T} \|y_s - z_t\|_2 - \frac{1}{S^2} \sum_{s,s'=1}^{S} \|y_s - y_{s'}\|_2 \\ - \frac{1}{T^2} \sum_{t,t'=1}^{T} \|z_t - z_{t'}\|_2. \end{aligned} \tag{21}$$

This is the energy distance between the empirical cell distributions, using Euclidean distances and including diagonal pairs in the within-population sums (Sejdinovic et al., 2013). The population kernel is

$$k_x(Y, W) = \exp\left[-\frac{d_E(Y, W)}{2h_x^2}\right], \qquad h_x > 0, \tag{22}$$

where h_x is held fixed during each projection update.

Permutation invariance. Each term in Eq. 21 averages over all pairs of cells. Reordering either population therefore leaves $d_E(Y, W)$, and consequently $k_x(Y, W)$, unchanged. The kernel compares populations without requiring a cell-to-cell correspondence.

Positive semidefiniteness. Euclidean distance is of negative type, so there exist a Hilbert space $\mathcal{H}$ and a map ϕ such that $\|y - z\|_2 = \|\phi(y) - \phi(z)\|_{\mathcal{H}}^2$ (Sejdinovic et al., 2013). Define the population embeddings $\mu_Y = S^{-1} \sum_{s=1}^S \phi(y_s)$ and $\mu_W = T^{-1} \sum_{t=1}^T \phi(z_t)$. Substituting this identity into Eq. 21 and expanding the squared norms gives

$$d_E(Y, W) = 2\|\mu_Y - \mu_W\|_{\mathcal{H}}^2, \qquad k_x(Y, W) = \exp\left(-\frac{\|\mu_Y - \mu_W\|_{\mathcal{H}}^2}{h_x^2}\right). \tag{23}$$

Thus, k_x is a Gaussian kernel on the population embeddings and is positive semidefinite.To see this directly, factor the exponential as

$$k_x(Y, W) = e^{-\|\mu_Y\|_{\mathcal{H}}^2/h_x^2} e^{-\|\mu_W\|_{\mathcal{H}}^2/h_x^2} \sum_{m=0}^{\infty} \frac{(2/h_x^2)^m}{m!} \langle \mu_Y, \mu_W \rangle_{\mathcal{H}}^m.$$

Each term in the series is a positive-semidefinite kernel, and multiplication by the two population-specific factors preserves this property.

The embedding identity also establishes $d_E(Y, W) \geq 0$, $0 < k_x(Y, W) \leq 1$, and $k_x(Y, Y) = 1$. Consequently, the kernel defines a valid, permutation-invariant MMD objective over cell populations.

A.4 Algorithm Details

We summarize the complete MMD-PO procedure below, bringing together the reward-guided target and MMD policy update introduced above.

Algorithm 1 MMD-PO: Reward-Weighted MMD Policy Optimization

Require: Reference policy p_{ref}, reward function R, training dataset $\mathcal{D}$, population kernel k_x, rollout count N, inner update count $U \geq 1$, learning rate η, temperature α, and reference coefficient β_{ref}

Initialize: $\theta \leftarrow \theta_{\mathrm{ref}}$; freeze p_{ref}

for each training iteration t **do**
  $\theta_{\mathrm{old}} \leftarrow \mathrm{copy}(\theta)$; freeze $p_{\theta_{\mathrm{old}}}$
  $\mathcal{T} \leftarrow \emptyset$
  for each sampled context $x \in \mathcal{D}$ **do** ▷ *Target construction*
    Sample $Y_j^{\mathrm{old}} \sim p_{\theta_{\mathrm{old}}}(\cdot \mid x)$, $j = 1, \ldots, N$
    Obtain paired $Y_j^{\mathrm{ref}} \sim p_{\mathrm{ref}}(\cdot \mid x)$, $j = 1, \ldots, N$
    Compute rewards $R_j \leftarrow R(x, Y_j^{\mathrm{old}})$, $j = 1, \ldots, N$
    Compute scores $s_j \leftarrow \dfrac{R_j}{\alpha} - \beta_{\mathrm{ref}}\, d_E(Y_j^{\mathrm{old}}, Y_j^{\mathrm{ref}})$
    Normalize weights $w_j \leftarrow \dfrac{\exp(s_j)}{\sum_{r=1}^{N} \exp(s_r)}$
    $\mathcal{T} \leftarrow \mathcal{T} \cup \left\{ \left(x, (Y_j^{\mathrm{old}}, w_j)_{j=1}^{N} \right) \right\}$
  end for
  Keep stored populations and weights fixed
  for $u = 1, \ldots, U$ **do** ▷ *MMD projection*
    for each $(x, (Y_j^{\mathrm{old}}, w_j)_{j=1}^{N}) \in \mathcal{T}$ **do**
      Generate $Y_i^{\theta} \sim p_\theta(\cdot \mid x)$, $i = 1, \ldots, N$
      Compute $\mathcal{L}_{\text{MMD-PO}}(\theta; x)$ using Eq. 9
    end for
    $\theta \leftarrow \theta - \eta \nabla_\theta \left[\dfrac{1}{|\mathcal{T}|} \displaystyle\sum_{(x,\cdot) \in \mathcal{T}} \mathcal{L}_{\text{MMD-PO}}(\theta; x) \right]$
  end for
end for

Output: Updated policy p_θ

A.5 Biological reward details

Pseudobulk perturbation effects. For a context $x = (X^{\mathrm{ctrl}}, t_{\mathrm{pert}})$, let Y and Y^{obs} denote the predicted and observed post-perturbation populations. We average expression over cells and compute the perturbation effect of gene g relative to the control population:

$$\Delta_g^{\mathrm{pred}} = \bar{y}_g - \bar{X}_g^{\mathrm{ctrl}}, \qquad \Delta_g^{\mathrm{obs}} = \bar{y}_g^{\mathrm{obs}} - \bar{X}_g^{\mathrm{ctrl}}. \tag{24}$$

Because the model operates on log-normalized expression, these differences represent pseudobulk effects on the log-expression scale. Averaging over cells also makes both reward components invariant to the ordering of cells within a population.

Gene-Effect reward. Let $\mathcal{G}_\ell$ denote the gene set associated with pathway ℓ, and let $\mathcal{U} = \bigcup_{\ell=1}^{P} \mathcal{G}_\ell$ be the union of genes in the selected pathway collection. We compare the predicted and observed effects over this union using the Huber loss:

$$R_{\mathrm{gene}}(x, Y) = -\frac{1}{|\mathcal{U}|} \sum_{g \in \mathcal{U}} H_\delta\left(\Delta_g^{\mathrm{pred}} - \Delta_g^{\mathrm{obs}}\right), \tag{25}$$

where

$$H_\delta(e) = \begin{cases} \frac{1}{2} e^2, & |e| \leq \delta, \\ \delta\left(|e| - \frac{1}{2}\delta\right), & |e| > \delta. \end{cases} \tag{26}$$

We use $\delta = 0.5$. The quadratic region encourages accurate recovery of moderate effects, while the linear region limits the influence of unusually large errors.

Signed GSEA enrichment scores. We compute pathway-level activity using the weighted GSEA procedure of Subramanian et al. (2005), replacing the original phenotype-correlation statistic with the pseudobulk perturbation effect. For each predicted or observed effect vector, we first rank all measured genes from the most positive to the most negative effect. We then walk down this ranked list, increasing a running statistic whenever a pathway gene is encountered and decreasing it whenever a non-pathway gene is encountered. Increases are weighted by the absolute effect magnitude of the pathway gene, using the standard weighting exponent $p = 1$, whereas decreases are uniform.

The enrichment score is the largest absolute deviation of this running statistic from zero while retaining the sign of the deviation. A positive score therefore indicates that pathway genes are concentrated near the top of the ranked list, corresponding to pathway activation, while a negative score indicates concentration near the bottom, corresponding to pathway suppression. Applying the same procedure to the predicted and observed effect vectors gives

$$\widehat{\mathrm{ES}}_\ell = \mathrm{GSEA}\left(\Delta^{\mathrm{pred}}, \mathcal{G}_\ell\right), \qquad \mathrm{ES}_\ell^{\mathrm{obs}} = \mathrm{GSEA}\left(\Delta^{\mathrm{obs}}, \mathcal{G}_\ell\right). \tag{27}$$

The predicted and observed scores are computed using the same gene universe, pathway definitions, ranking rule, and weighting scheme. For the observed data, GSEA significance is estimated using the standard permutation procedure.

Overlap-aware pathway weights. Because pathways may share genes, we use overlap-aware weights to reduce the influence of highly overlapping gene sets. Let

$$m_g = |\{\ell : g \in \mathcal{G}_\ell\}| \tag{28}$$

denote the number of selected pathways containing gene g. The weight of pathway ℓ is

$$\omega_\ell = \sum_{g \in \mathcal{G}_\ell} \frac{1}{m_g}, \qquad a_\ell = \frac{\omega_\ell}{\sum_{r=1}^{P} \omega_r}, \qquad \sum_{\ell=1}^{P} a_\ell = 1. \tag{29}$$

Thus, a gene shared by multiple pathways divides its contribution among those pathways rather than contributing its full weight to each one.

Pathway-Enrichment reward. For each pathway, we penalize only the enrichment-score error outside a relative tolerance band:

$$\xi_\ell(x, Y) = \max\left(\left|\widehat{\mathrm{ES}}_\ell - \mathrm{ES}_\ell^{\mathrm{obs}}\right| - \tau \left|\mathrm{ES}_\ell^{\mathrm{obs}}\right|, 0\right). \tag{30}$$

The Pathway-Enrichment reward is

$$R_{\mathrm{path}}(x, Y) = -\sum_{\ell=1}^{P} a_\ell \, \xi_\ell(x, Y)^2. \tag{31}$$

We use $\tau = 0.05$, so a prediction incurs no penalty when its enrichment score lies within 5% of the magnitude of the observed score. Errors outside this band are penalized quadratically, encouraging agreement in both the magnitude and direction of pathway regulation.

B Dataset and Pathway-Panel Details

For both benchmarks, models are trained and evaluated exclusively on the dataset-specific protein-coding highly variable gene (HVG) panel. Model inputs, predicted populations, supervised objectives, biological rewards, and evaluation metrics are defined over this feature space. Genes outside the corresponding HVG panel are not provided to or predicted by the model. To ensure that each evaluated pathway is sufficiently represented, we apply the same minimum threshold of 15 measurable genes to both datasets.

B.1 Replogle–Nadig

Dataset and evaluation split. The Replogle–Nadig benchmark combines genome-scale CRISPR Perturb-seq measurements from four cell lines: K562, RPE1, Jurkat, and HepG2 (Replogle et al.,

Table 3: Hallmark pathways retained for the Replogle–Nadig benchmark after applying the minimum measurable-gene threshold.

#	Pathway	#	Pathway
1	G2M checkpoint	20	Estrogen response early
2	E2F targets	21	Unfolded-protein response
3	Mitotic spindle	22	Cholesterol homeostasis
4	mTORC1 signaling	23	PI3K–AKT–mTOR signaling
5	p53 pathway	24	IL2–STAT5 signaling
6	Glycolysis	25	Spermatogenesis
7	Heme metabolism	26	DNA repair
8	TNFα signaling via NF-κB	27	Oxidative phosphorylation
9	Apoptosis	28	UV response down
10	Hypoxia	29	Complement
11	UV response up	30	Peroxisome
12	Adipogenesis	31	Protein secretion
13	Estrogen response late	32	Epithelial–mesenchymal transition
14	Fatty-acid metabolism	33	Myogenesis
15	Interferon-γ response	34	Interferon-α response
16	Androgen response	35	Allograft rejection
17	Xenobiotic metabolism	36	Inflammatory response
18	Apical junction	37	TGF-β signaling
19	MYC targets V1		

2022; Nadig et al., 2025). We use the processed release and HVG representation adopted by X-Cell (Wang et al., 2026). The complete processed dataset contains 643,413 cells. Following the published X-Cell split, 380 perturbations in HepG2 are held out for evaluation, while the remaining 616,535 cells are used for training.

This benchmark tests generalization to unseen perturbation–context combinations. HepG2 cells are represented during training, and a held-out perturbation may be observed in another cell line, but its response in HepG2 is excluded from training and checkpoint selection. All 380 held-out HepG2 perturbations constitute the candidate set for pathway-level prioritization. We therefore refer to this setting as *few-shot perturbation generalization*.

Pathway selection. We begin with all 50 Hallmark gene sets in MSigDB release `2026.1.Hs`. Each pathway is intersected with the protein-coding HVG panel used by the model, and pathways containing fewer than 15 measurable genes are excluded. This procedure retains 37 of the 50 Hallmark pathways. Because the filtering depends only on gene-set membership and feature availability, it does not use the held-out HepG2 perturbation responses.

For evaluation, the observed enrichment significance of each perturbation–pathway pair is determined at FDR $q < 0.05$. This significance criterion is applied after defining the fixed 37-pathway panel and is used to construct the relevance labels for the perturbation-prioritization metrics.

B.2 Marson Primary CD4$^+$ T Cells

Dataset and evaluation split. The Marson dataset contains genome-scale perturbational profiles from primary human CD4$^+$ T cells collected from multiple donors under resting and CD3/CD28-stimulated conditions (Zhu et al., 2025a). We use only the protein-coding HVG panel selected for the Marson experiment.

Perturbational profiles from resting cells of donor D3 are used for training and post-training, whereas the resulting model is evaluated on stimulated cells from the same donor. Thus, the perturbation identities may be observed during training, but their transcriptional effects in the stimulated cellular context are held out. At inference time, non-targeting control cells from the stimulated condition provide the input cellular context. Stimulated perturbed profiles are not used for training, post-training, or checkpoint selection. We therefore refer to this setting as *zero-shot context transfer*.

Table 4: Pathways retained for the Marson benchmark. The panel is the union of significant terms identified across seven T-cell-related seed perturbations after applying the minimum measurable-gene threshold.

Hallmark pathways	**KEGG pathways**
Complement	Ribosome biogenesis in eukaryotes
E2F targets	Proteasome
Interferon-α response	T-cell receptor signaling pathway
G2M checkpoint	Fc gamma R-mediated phagocytosis
Allograft rejection	Cell adhesion molecules
IL-6/JAK/STAT3 signaling	TNF signaling pathway
MYC targets V1	NF-κB signaling pathway
Interferon-γ response	
UV response down	
Inflammatory response	

Pathway selection. Because the reward in a virtual-cell RL loop is defined by the model itself, a policy can succeed by exploiting model error rather than biology. We therefore seed perturbations with genes supported by two complementary lines of human genetic evidence. Six are implicated in T cell activation–related immunodeficiencies catalogued in OMIM (`https://omim.org/`, accessed August 2026): `LAT`, `CD247`, `CD3G`, `PTPRC`, `MALT1`, and `IKBKB`. The seventh, `KLF2`, carries a significant rare-variant loss-of-function burden association with lymphocyte percentage in UK Biobank whole-exome data (Backman et al., 2021). This anchors in causal signals from human biology the model was never trained to reproduce. Together they anchor the search in causal signal the model was never trained to reproduce. For each seed perturbation, we separately perform upstream pathway enrichment against the `MSigDB_Hallmark_2020` and `KEGG_2021_Human` collections.

As in Replogle–Nadig, each candidate pathway is intersected with the protein-coding HVG panel used by the model, and pathways containing fewer than 15 measurable genes are excluded. Among the remaining pathways, we retain terms with FDR $q < 0.05$ and take their union across the seven seed perturbations.

C Backbone and Baseline Implementations

Virtual-cell backbones. We evaluate post-training across three virtual-cell backbones with distinct model architectures. X-Cell is a diffusion language model that generates post-perturbation expression through iterative masked reconstruction (Wang et al., 2026). scGPT is a generatively pretrained Transformer for single-cell expression modeling that can be adapted to perturbation prediction through task-specific fine-tuning (Cui et al., 2024). scDFM models the conditional distribution of perturbed cells using flow matching and incorporates an MMD regularizer during supervised training (Yu et al., 2026). Each backbone is first fine-tuned on the corresponding training split, and all post-training methods applied to a given backbone are initialized from the same supervised checkpoint.

Supervised initialization and continued SFT. The supervised-initialization baseline is the checkpoint obtained after fine-tuning each backbone with its native supervised objective. Continued SFT then trains the same checkpoint for additional optimizer steps using the unchanged supervised objective. We match its number of optimizer steps to the post-training budget used by MMD-PO. This baseline controls for the possibility that improvements arise simply from additional optimization rather than from reward-guided post-training.

GRPO. GRPO is a critic-free policy-gradient method that estimates advantages by standardizing rewards across a group of outputs generated for the same condition (Shao et al., 2024). It optimizes a clipped likelihood-ratio objective while regularizing the updated policy toward a reference policy. To accommodate the output parameterizations of X-Cell and scGPT, which do not expose

tractable likelihoods over complete unordered cell populations, we adapt the likelihood computation by defining a Gaussian policy centered at the model prediction. This provides the likelihood ratios required by GRPO without changing its group-relative advantage estimation or clipped optimization objective. We refer to this adapted implementation as GRPO throughout the paper for simplicity.

Flow-GRPO. For the flow-matching scDFM backbone, we use Flow-GRPO as the corresponding policy-gradient baseline (Liu et al., 2025). Flow-GRPO formulates the iterative flow-generation process as a Markov decision process and converts the deterministic sampling ODE into a stochastic process, enabling exploration and the evaluation of likelihood ratios between successive sampling transitions. Terminal rewards are standardized within each group and propagated through a clipped GRPO objective over the sampled trajectory. We adapt this construction to conditional single-cell generation while retaining the control-cell state and perturbation identity as conditioning variables. Flow-GRPO and MMD-PO are initialized from the same supervised scDFM checkpoint and receive the same biological reward, number of candidate populations, and optimizer-step budget. Their comparison therefore primarily isolates the likelihood-based trajectory update from the likelihood-free output-distribution update used by MMD-PO.

DiffusionNFT and the PerturbCellRL-inspired comparison. In the supplementary scDFM compatibility analysis, we additionally consider a DiffusionNFT-style baseline. DiffusionNFT avoids reverse-process likelihoods by contrasting high- and low-reward generations through a forward-process flow-matching objective (Zheng et al., 2026). PerturbCellRL applies this strategy to a pretrained single-cell flow-matching generator using Pearson similarity, RMSE proximity, differential-expression Spearman correlation, and pathway-activity verifiers (Wu et al., 2026). Our comparison reconstructs the three target-dependent components available in our evaluation setting: Pearson similarity, normalized inverse RMSE, and differential-expression Spearman correlation. The PROGENy-based pathway verifier and its associated assets are unavailable in our experimental pipeline and are therefore omitted. Consequently, this arm is a PerturbCellRL-inspired DiffusionNFT baseline rather than an exact reproduction.

Comparison protocol. Unless otherwise stated, post-training methods use identical data splits, HVG panels, pathway panels, control-cell populations, biological rewards, rollout counts, and checkpoint-selection procedures. Comparisons are matched by the number of optimizer steps. Method-specific likelihood parameterizations and regularization coefficients are reported together with the remaining hyperparameters in Appendix D.

D Hyperparameters and Implementation Details

We initialize MMD-PO from the X-Cell checkpoint obtained after 20 epochs of supervised fine-tuning and post-train it for 5 epochs. For each input context, the frozen policy generates eight candidate populations using the four-step masked-reveal process. Training is distributed across four compute nodes, each equipped with eight NVIDIA H200 GPUs, for a total of 32 GPUs. Table 5 summarizes the main optimization, rollout, and reward-weighting hyperparameters. The MMD projection uses a Gaussian kernel over the energy distance between unordered cell populations, with the bandwidth selected independently for each context using the median heuristic. We use the reference correction defined in Section 4 with $\beta_{\text{ref}} = 0.1$, without an additional KL penalty or auxiliary regularization term.

E Evaluation Metrics

E.1 Predictive-Fidelity Metrics

We evaluate predictive fidelity using metrics implemented in Cell-Eval (Adduri et al., 2026). These metrics compare predicted and observed perturbation responses at the expression, differential-expression, and perturbation-identity levels. Each metric is computed separately for every perturbation and then macro-averaged across scoreable perturbations.

Table 5: Key hyperparameters for the main MMD-PO experiment.

Hyperparameter	Value
Initialization	SFT checkpoint at epoch 20
Post-training epochs	5
Population size S	64 cells
Optimizer	AdamW
Peak learning rate	1×10^{-5}
Optimizer step size	64
Rollouts per context N	8
Reveal fractions	$\{0.25, 0.50, 0.75, 1.00\}$
Reward temperature α	0.3
Pathway-enrichment tolerance τ	0.05
Gene-effect Huber threshold δ	0.5
Reference coefficient β_{ref}	0.1

Notation. For perturbation i and gene g, let $\bar{x}_{ig}^{\mathrm{pred}}$ and $\bar{x}_{ig}^{\mathrm{gt}}$ denote the predicted and observed pseudobulk expression values. Let $\bar{x}_{0g}^{\mathrm{pred}}$ and $\bar{x}_{0g}^{\mathrm{gt}}$ denote the corresponding control means. The predicted and observed perturbation effects are $\Delta_{ig}^{\mathrm{pred}} = \bar{x}_{ig}^{\mathrm{pred}} - \bar{x}_{0g}^{\mathrm{pred}}$ and $\Delta_{ig}^{\mathrm{gt}} = \bar{x}_{ig}^{\mathrm{gt}} - \bar{x}_{0g}^{\mathrm{gt}}$. We denote their differential-expression log-fold changes by $\ell_{ig}^{\mathrm{pred}}$ and ℓ_{ig}^{gt}.

Overlap@N. For perturbation i, let $\mathcal{D}_i^{\mathrm{gt}}$ be the set of observed differentially expressed genes passing the Cell-Eval significance threshold, and let $n_i = |\mathcal{D}_i^{\mathrm{gt}}|$. Let $\mathcal{T}_{i,n_i}^{\mathrm{pred}}$ contain the top n_i predicted genes ranked by absolute predicted log-fold change. Overlap@N is $|\mathcal{D}_i^{\mathrm{gt}} \cap \mathcal{T}_{i,n_i}^{\mathrm{pred}}|/n_i$. It therefore measures the fraction of observed DE genes recovered among an equally sized set of predicted genes.

DE direction match. Among genes that are significantly differentially expressed in the observed data, DE direction match is the fraction for which the predicted and observed log-fold changes have the same sign: $\mathrm{sign}(\ell_{ig}^{\mathrm{pred}}) = \mathrm{sign}(\ell_{ig}^{\mathrm{gt}})$. This metric evaluates whether the model correctly predicts upregulation versus downregulation.

DE Spearman LFC. DE Spearman LFC is the Spearman rank correlation between $\ell_{ig}^{\mathrm{pred}}$ and ℓ_{ig}^{gt}, evaluated over genes that are significantly differentially expressed in the observed data. It measures whether the model preserves the relative ordering of gene-level perturbation effects.

Pearson-Δ. Pearson-Δ is the Pearson correlation across genes between the predicted and observed perturbation-effect vectors, Δ_i^{pred} and Δ_i^{gt}. Unlike the DE-based metrics, it evaluates the complete pseudobulk response rather than only genes passing a significance threshold.

Mean absolute error. Mean absolute error (MAE) measures expression-level deviation between the predicted and observed pseudobulk profiles: $\mathrm{MAE}_i = G^{-1} \sum_{g=1}^{G} |\bar{x}_{ig}^{\mathrm{pred}} - \bar{x}_{ig}^{\mathrm{gt}}|$, where G is the number of evaluated genes. Lower values indicate better expression-level fidelity.

Perturbation discrimination score. Perturbation discrimination evaluates whether a predicted response retains the identity of its intended perturbation. For each predicted effect vector Δ_i^{pred}, Cell-Eval computes its distance to every observed perturbation-effect vector and determines the rank r_i of the matching observed perturbation. The score is $1 - r_i/M$, where M is the number of candidate perturbations and $r_i = 0$ indicates the closest match. Higher values therefore indicate that the predicted response is more readily distinguishable as the correct perturbation.

E.2 Perturbation-Prioritization Metrics

We evaluate whether perturbations with strong ground-truth pathway effects are ranked highly by the model. Metrics are computed independently for each pathway and then macro-averaged across scoreable pathways.

Notation. For a given pathway, let M be the number of candidate perturbations. For perturbation i, let ES_i^{pred}, ES_i^{gt}, and q_i^{gt} denote its predicted enrichment score, observed enrichment score, and observed GSEA FDR q-value, respectively. We evaluate effect magnitude using

$$a_i^{\text{pred}} = \left|ES_i^{\text{pred}}\right|, \qquad a_i^{\text{gt}} = \left|ES_i^{\text{gt}}\right|. \tag{32}$$

Let π_j denote the perturbation at position j after sorting a_i^{pred} in descending order. Predicted q-values are not used in any prioritization metric.

For a fractional screening budget b, we use $B = \lceil bM \rceil$. We report $b \in \{0.01, 0.05\}$.

E.3 Enrichment factor

Let $\mathcal{S}_B^{\text{pred}}$ contain the top-B perturbations ranked by a_i^{pred}. We define ground-truth hits as perturbations in the top 10% by absolute ground-truth pathway enrichment score, $a_i^{\text{gt}} = |ES_i^{\text{gt}}|$, whose ground-truth GSEA FDR satisfies $q_i^{\text{gt}} < 0.05$:

$$\mathcal{S}^{\text{gt}} = \left\{ i : i \in \text{Top}_{\lceil 0.1M \rceil}(a^{\text{gt}}) \ \wedge \ q_i^{\text{gt}} < 0.05 \right\}. \tag{33}$$

This hit set is fixed across screening budgets. The significance filter is applied after ranking the complete candidate pool; nonsignificant perturbations are removed without promoting lower-ranked candidates into the hit set.

The number of recovered hits at budget B is $F_B = |\mathcal{S}_B^{\text{pred}} \cap \mathcal{S}^{\text{gt}}|$. Enrichment factor compares the hit rate within the predicted shortlist with the prevalence of ground-truth hits in the complete candidate pool:

$$\text{EF@}B = \frac{F_B/B}{|\mathcal{S}^{\text{gt}}|/M}. \tag{34}$$

A uniformly random ranking has expected $\text{EF@}B = 1$. Pathways with no ground-truth hits are excluded from the aggregate.

E.4 Adjusted normalized discounted cumulative gain

Normalized discounted cumulative gain (nDCG) evaluates whether items with high ground-truth relevance appear near the top of a predicted ranking. DCG applies a logarithmic discount so that placing a relevant item near the top contributes more than placing it near the bottom, while nDCG normalizes by the score of the ideal ranking (Jeunen et al., 2024).

Given nonnegative relevance scores $\{r_i\}_{i=1}^M$ and a gain function g, the discounted cumulative gain of the predicted ranking at cutoff B is

$$\text{DCG@}B = \sum_{j=1}^{B} \frac{g(r_{\pi_j})}{\log_2(j+1)}, \tag{35}$$

where π_j is the perturbation at position j in the ranking induced by predicted $|ES|$.

Let $r_{(j)}$ denote the relevance values sorted in descending order. The ideal DCG and normalized DCG are

$$\text{IDCG@}B = \sum_{j=1}^{B} \frac{g(r_{(j)})}{\log_2(j+1)}, \qquad \text{nDCG@}B = \frac{\text{DCG@}B}{\text{IDCG@}B}. \tag{36}$$

Thus, nDCG equals 1 when the predicted ordering matches the ideal ordering induced by the ground-truth relevance scores.

Following AssayBench (De Brouwer et al., 2026), we further adjust nDCG by its expected value under a uniformly random ranking. Let

$$\overline{g} = \frac{1}{M} \sum_{i=1}^{M} g(r_i) \tag{37}$$

denote the mean gain across the candidate perturbations. The expected DCG of a random ranking and its normalized value are

$$\mathbb{E}_{\text{rand}}[\text{DCG@}B] = \overline{g} \sum_{j=1}^{B} \frac{1}{\log_2(j+1)}, \qquad \text{nDCG}^{\text{rand}}\text{@}B = \frac{\mathbb{E}_{\text{rand}}[\text{DCG@}B]}{\text{IDCG@}B}. \tag{38}$$

Adjusted normalized discounted cumulative gain is

$$\text{AnDCG@}B = \frac{\text{nDCG@}B - \text{nDCG}^{\text{rand}}\text{@}B}{1 - \text{nDCG}^{\text{rand}}\text{@}B}. \tag{39}$$

This adjustment maps the ideal ranking to 1 and gives a uniformly random ranking an expected score of 0. Negative values indicate performance below the random baseline. Pathways with $\text{IDCG@}B = 0$ are excluded from the aggregate.

We instantiate this framework with two relevance definitions: a discrete rank-based score and a continuous geometric score. In both cases, statistical significance is incorporated directly into the relevance score by assigning zero relevance to perturbations with $q_i^{\text{gt}} \geq 0.05$.

E.5 Rank-based relevance

AnDCG-Rank assigns discrete relevance according to a perturbation's position in the ground-truth effect-magnitude ranking. Let $\rho_i^{ES} \in \{1, \ldots, M\}$ denote the descending rank of perturbation i according to a_i^{gt}. We define

$$r_i^{\text{rank}} = \begin{cases} 3, & \rho_i^{ES} \leq \lceil 0.05M \rceil \ \wedge \ q_i^{\text{gt}} < 0.05, \\ 2, & \lceil 0.05M \rceil < \rho_i^{ES} \leq \lceil 0.10M \rceil \ \wedge \ q_i^{\text{gt}} < 0.05, \\ 1, & \lceil 0.10M \rceil < \rho_i^{ES} \leq \lceil 0.20M \rceil \ \wedge \ q_i^{\text{gt}} < 0.05, \\ 0, & \text{otherwise.} \end{cases} \tag{40}$$

The percentile bands are determined from the complete ground-truth ranking before applying the significance condition. Therefore, nonsignificant perturbations are assigned zero relevance without changing the percentile bands of other perturbations.

For AnDCG-Rank, we use

$$g_{\text{rank}}(r) = 2^r - 1, \tag{41}$$

giving gains of 7, 3, 1, and 0 for relevance levels 3, 2, 1, and 0, respectively.

E.6 Geometric relevance

AnDCG-Geo uses a continuous relevance score that combines ground-truth effect magnitude and statistical significance. Define the ground-truth effect percentile as

$$p_{ES,i} = \frac{M - \rho_i^{ES} + 1}{M}. \tag{42}$$

Let ρ_i^q denote the rank of perturbation i after sorting q_i^{gt} in ascending order, so that the smallest q-value has rank one. Its significance percentile is

$$p_{q,i} = \frac{M - \rho_i^q + 1}{M}. \tag{43}$$

Tied values receive their average rank. We define geometric relevance as

$$r_i^{\text{geom}} = \sqrt{p_{ES,i} p_{q,i}}\, \mathbb{I}\left[q_i^{\text{gt}} < 0.05\right]. \tag{44}$$

High relevance therefore requires both a strong observed pathway effect and strong statistical significance. For AnDCG-Geo, we use the linear gain

$$g_{\text{geom}}(r) = r. \tag{45}$$

Table 6: Cell-Eval metrics across virtual-cell backbones and post-training methods. Values are rounded to three decimals; dashes indicate unreported metrics. Bold and underlined values denote the best and second-best distinct displayed values within each backbone and dataset, respectively. For ties, only one entry is marked.

Model	DE direction match ↑	DE Spearman LFC ↑	Overlap@N ↑	Pearson-Δ ↑	Centroid accuracy ↑	Pseudobulk MAE ↓
Replogle–Nadig HepG2						
X-Cell	0.826	0.625	0.229	0.531	<u>0.828</u>	0.060
+ Continued SFT	0.830	0.635	0.230	0.528	**0.837**	0.060
+ GRPO	<u>0.838</u>	<u>0.639</u>	<u>0.234</u>	<u>0.536</u>	0.826	<u>0.057</u>
+ **MMD-PO**	**0.842**	**0.662**	**0.246**	**0.538**	0.815	**0.054**
scGPT	0.646	0.286	**0.215**	0.276	0.554	0.138
+ Continued SFT	0.643	0.288	<u>0.214</u>	<u>0.277</u>	0.553	0.139
+ GRPO	<u>0.652</u>	<u>0.292</u>	0.205	0.275	<u>0.573</u>	<u>0.124</u>
+ **MMD-PO**	**0.706**	**0.416**	0.166	**0.307**	**0.697**	**0.073**
scDFM	<u>0.729</u>	<u>0.453</u>	**0.189**	**0.348**	**0.577**	0.093
+ Continued SFT	0.683	0.383	<u>0.149</u>	0.279	0.507	0.081
+ Flow-GRPO	0.682	0.389	0.133	0.269	0.552	<u>0.075</u>
+ **MMD-PO**	**0.772**	**0.460**	0.146	<u>0.342</u>	<u>0.564</u>	**0.074**
Marson primary CD4$^+$ T cells						
X-Cell	–	–	–	0.058	0.504	0.592
+ Continued SFT	–	–	–	0.060	<u>0.506</u>	0.594
+ GRPO	–	–	–	<u>0.058</u>	0.502	<u>0.479</u>
+ **MMD-PO**	–	–	–	**0.060**	**0.509**	**0.414**
scGPT	–	–	–	0.036	0.501	0.354
+ Continued SFT	–	–	–	0.034	0.501	0.364
+ GRPO	–	–	–	<u>0.039</u>	<u>0.501</u>	<u>0.339</u>
+ **MMD-PO**	–	–	–	**0.057**	**0.517**	**0.122**
scDFM	–	–	–	**0.058**	**0.501**	0.466
+ Continued SFT	–	–	–	<u>0.057</u>	0.501	<u>0.463</u>
+ Flow-GRPO	–	–	–	0.049	0.501	**0.450**
+ **MMD-PO**	–	–	–	0.051	<u>0.500</u>	0.480

F Additional Experiment Results

F.1 Additional Cell-Eval Metrics

Table 6 reports the Cell-Eval metrics for all backbones and post-training methods in Table 1. MMD-PO has lower or comparable pseudobulk MAE than the corresponding RL baseline for most backbone and dataset.

The Marson evaluation contains 6,997 perturbations. Computing DE Direction Match, DE Spearman LFC, and Overlap@N for all predictions would require more than 40 CPU-hours in our evaluation pipeline. We therefore omit these three Marson metrics; dashes indicate values that were not computed or imputed.

F.2 Additional Ablation Studies

Supervised ablation baselines. Let $\mathcal{L}_{\mathrm{XCell}}(\theta; x, Y^{\mathrm{obs}})$ denote the original X-Cell supervised loss for context x and observed perturbed-cell population Y^{obs}. All three SFT variants start from the same checkpoint and use the same training split and post-training optimizer-step budget.

Continued SFT. We continue training with $\mathcal{L}_{\mathrm{XCell}}$ without modifying the objective. This baseline controls for the effect of additional supervised optimization.

Table 7: Ablation of supervised objectives and MMD-PO rewards for predictive fidelity on Replogle–Nadig HepG2. All experiments use X-Cell. The combined reward uses both normalized components. Bold and underlined values indicate the best and second-best results across all rows, respectively.

Method	DE direction match ↑	DE Spearman LFC ↑	Overlap@N ↑	Pearson-Δ ↑	Centroid accuracy ↑	Pseudobulk MAE ↓
Supervised-objective ablation						
Continued SFT	0.830	0.635	0.230	0.528	**0.837**	0.060
Pathway-gene SFT	0.623	0.208	0.140	0.180	0.647	0.121
Reward-weighted SFT	0.711	0.501	0.157	0.400	0.712	0.080
MMD-PO reward ablation						
Gene-Effect reward	**0.851**	**0.673**	**0.248**	**0.546**	0.819	**0.053**
Pathway-Enrichment reward	0.808	0.566	0.176	0.519	<u>0.822</u>	0.064
Combined reward	<u>0.842</u>	<u>0.662</u>	<u>0.246</u>	<u>0.538</u>	0.815	<u>0.054</u>

Pathway-gene SFT. Let U denote the union of the selected pathway gene sets. We apply the original X-Cell loss only to predictions for genes in U:

$$\mathcal{L}_{\text{PG-SFT}} = \mathcal{L}_{\text{XCell}}(\theta; x, Y^{\text{obs}} \mid U). \tag{46}$$

The training targets remain the observed cell profiles, but the supervised objective focuses on pathway-associated genes.

Reward-weighted SFT. We use the combined Gene-Effect and Pathway-Enrichment reward to weight the original X-Cell loss. For a minibatch $\mathcal{B}$, the objective is

$$\mathcal{L}_{\text{RW-SFT}} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} w_i \, \mathcal{L}_{\text{XCell}}(\theta; x_i, Y_i^{\text{obs}}), \tag{47}$$

where w_i is the weight obtained from the combined reward for example i. Thus, the reward changes the contribution of each example to supervised training; it does not define a target distribution for the MMD-PO update.

Table 7 reports predictive fidelity for the supervised-objective and reward ablations. Both pathway-gene and reward-weighted SFT underperform standard SFT across all six metrics. In contrast, MMD-PO with the combined reward improves five of the six metrics over standard SFT, including DE Spearman LFC (0.662 vs. 0.635) and pseudobulk MAE (0.054 vs. 0.060), while centroid accuracy decreases from 0.837 to 0.815.

Among the reward variants, the Gene-Effect reward alone achieves the best fidelity on five metrics, while the Pathway-Enrichment reward yields the highest centroid accuracy among MMD-PO variants. Although combining the rewards slightly reduces predictive fidelity relative to the gene-level reward alone, performance remains comparable while perturbation prioritization improves (Table 2).

G Complete Standardized Empirical Results and Compatibility Checks

Tables 8 and 9 report every paper-facing result from the 16 provenance-pinned scGPT and adapted-scDFM evaluations. All rows are single-training-seed point estimates. The 1% ranking endpoint was obtained by uniformly rescoring the frozen prediction panels; training, checkpoint inference, relevance definitions, gains, and the $q < 0.05$ gate were unchanged. At 1%, $K = \lceil 0.01P \rceil$ is 4 of 380 Replogle candidates and 70 of 6,997 Marson candidates. "Marson D3 Rest" identifies the training condition; evaluation uses D3 Stim 8 h.

Table 8: Complete standardized perturbation-prioritization results. Each cell reports 1% / 5% / 10% / 20%, in that order; higher is better. Rank (band) AnDCG uses exponential gain, and geometric AnDCG uses linear gain. Values are single-training-seed point estimates; dashes indicate unreported values.

Dataset	Backbone	Post-training	Rank AnDCG	Geometric AnDCG	EF
Replogle–Nadig HepG2	scGPT	Supervised initialization	.094 / .075 / .092 / .110	.179 / .139 / .130 / .117	2.711 / 1.715 / 1.636 / 1.424
		Continued SFT	.089 / .073 / .091 / .105	.186 / .141 / .131 / .114	2.940 / 1.649 / 1.635 / 1.379
		Gaussian GRPO	.153 / .117 / .116 / .127	.266 / .204 / .164 / .139	3.453 / 2.310 / 1.773 / 1.440
		MMD-PO	.284 / .227 / .246 / .285	.364 / .326 / .317 / .295	4.755 / 3.655 / 2.939 / 2.221
	adapted scDFM	Supervised initialization	.064 / .063 / .068 / .091	.108 / .110 / .108 / .116	1.959 / 1.847 / 1.506 / 1.534
		Continued SFT	.060 / .042 / – / –	.132 / .089 / – / –	1.565 / 1.203 / – / –
		DiffusionNFT (reconstructed 3R)	.143 / .099 / .106 / .126	.184 / .162 / .139 / .137	3.900 / 2.799 / 2.264 / 1.777
		Flow-GRPO	.168 / .091 / .096 / .106	.234 / .160 / .153 / .139	3.265 / 2.428 / 2.014 / 1.605
		MMD-PO	.165 / .120 / .129 / .145	.178 / .169 / .163 / .158	6.565 / 3.015 / 2.296 / 1.740
Marson D3 Rest	scGPT	Supervised initialization	.032 / .022 / .023 / .030	.063 / .038 / .030 / .026	1.270 / 1.211 / 1.164 / 1.119
		Continued SFT	.029 / .016 / .017 / .022	.052 / .029 / .022 / .019	1.308 / 1.193 / 1.133 / 1.083
		Gaussian GRPO	.033 / .025 / .027 / .032	.045 / .033 / .029 / .027	1.359 / 1.263 / 1.195 / 1.115
		MMD-PO	.036 / .025 / .031 / .041	.055 / .035 / .035 / .034	1.331 / 1.253 / 1.226 / 1.173
	adapted scDFM	Supervised initialization	.077 / .044 / .044 / .053	.110 / .067 / .053 / .044	1.775 / 1.429 / 1.295 / 1.200
		Continued SFT	.074 / .043 / – / –	.101 / .064 / – / –	1.836 / 1.412 / – / –
		DiffusionNFT (reconstructed 3R)	.076 / .046 / .049 / .057	.129 / .074 / .062 / .051	1.822 / 1.454 / 1.345 / 1.233
		Flow-GRPO	.067 / .042 / .044 / .054	.098 / .065 / .055 / .048	1.650 / 1.425 / 1.317 / 1.220
		MMD-PO	.080 / .048 / .057 / .065	.120 / .073 / .064 / .056	1.993 / 1.503 / 1.439 / 1.263

Table 9: Complete standardized Cell-Eval results. Values are single-training-seed point estimates rounded to three decimals. All endpoints are higher-is-better except pseudobulk MAE. The three DE-based Marson endpoints were not computed or imputed.

Dataset	Backbone	Post-training	DE direction match	DE Spearman LFC	Overlap@N	Pearson-Δ	Centroid accuracy	Pseudobulk MAE
Replogle–Nadig HepG2	scGPT	Supervised initialization	.646	.286	.215	.276	.554	.138
		Continued SFT	.643	.288	.214	.277	.553	.139
		Gaussian GRPO	.652	.292	.205	.275	.573	.124
		MMD-PO	.706	.416	.166	.307	.697	.073
	adapted scDFM	Supervised initialization	.729	.453	.189	.348	.577	.093
		Continued SFT	.683	.383	.149	.279	.507	.081
		DiffusionNFT (reconstructed 3R)	.682	.372	.113	.315	.562	.099
		Flow-GRPO	.682	.389	.133	.269	.552	.075
		MMD-PO	.772	.460	.146	.342	.564	.074
Marson D3 Rest	scGPT	Supervised initialization	–	–	–	.036	.501	.354
		Continued SFT	–	–	–	.034	.501	.364
		Gaussian GRPO	–	–	–	.039	.501	.339
		MMD-PO	–	–	–	.057	.517	.122
	adapted scDFM	Supervised initialization	–	–	–	.058	.501	.466
		Continued SFT	–	–	–	.057	.501	.463
		DiffusionNFT (reconstructed 3R)	–	–	–	.054	.501	.430
		Flow-GRPO	–	–	–	.049	.501	.450
		MMD-PO	–	–	–	.051	.500	.480

G.1 ADAPTED-SCDFM INTERPRETATION AND PERTURBCELLRL COMPATIBILITY

Within adapted scDFM, DiffusionNFT with the reconstructed 3R reward and MMD-PO start from the same supervised checkpoint and receive the same number of condition sets, but change both reward and update rule. The reconstructed 3R reward averages target-cell Pearson similarity, normalized inverse RMSE, and DE-fold-change Spearman terms rebuilt from training cells; unavailable PROGENy assets are omitted. It is therefore PerturbCellRL inspired, not an exact reproduction (Wu et al., 2026). FlowGRPO and MMD-PO provide the closer update-rule comparison because they share the base checkpoint, two-term reward, and rollout exposure while retaining method-specific optimization interfaces.

On Replogle, FlowGRPO leads rank and geometric AnDCG at 1%, whereas MMD-PO leads EF at 1% and all three 5%-budget metrics. MMD-PO has the lowest pseudobulk MAE and improves all six displayed fidelity metrics relative to reconstructed 3R. On Marson, MMD-PO leads rank AnDCG and EF at 1% and all three metrics at 5%; reconstructed 3R leads geometric AnDCG at 1% and has lower pseudobulk MAE. These mixed outcomes support portability to an adapted diffusion backbone but do not isolate the effect of the reward or update rule.

Table 10: **Four-fold Norman combination-holdout validation of adapted scDFM.** Values are unweighted mean ± sample SD over four fixed data folds, evaluated separately on held-out single- and double-gene conditions using Cell-Eval v0.6.8. Fold variation is not training-seed uncertainty. MAE is lower-is-better; the remaining metrics are higher-is-better. DS denotes the L_1 discrimination score.

Method	Held out	MAE ↓	Pearson-Δ ↑	DE LFC Spearman ↑	DS ↑
Our scDFM Norman adaptation	single	$.0580 \pm .0054$	$.6008 \pm .0220$	$.7780 \pm .0160$	$.8256 \pm .0200$
Our scDFM Norman adaptation	double	$.0550 \pm .0033$	$.7071 \pm .0119$	$.8673 \pm .0279$	$.9289 \pm .0174$

Table 11: **Same-fold Norman compatibility study.** All values are single-seed point estimates on fold 1 under Cell-Eval v0.6.8. "Compat." denotes evaluation through our adapted scDFM implementation. The 3R row is a no-PROGENy reconstruction using local, train-only DE statistics, not an exact PerturbCellRL reproduction. The post-training arms are neither compute- nor objective-matched. MAE is lower-is-better; other metrics are higher-is-better.

Method	Held out	MAE ↓	Pearson-Δ ↑	DE LFC Spearman ↑	DS ↑
Our scDFM Norman adaptation	single	.0526	.5737	.7688	.8328
	double	.0515	.6979	.9023	.9378
Replogle-initialized, then Norman-adapted scDFM	single	.0461	.5628	.7813	.8225
	double	.0419	.6803	.9049	.9022
Released scDFM checkpoint (compat.)	single	.0312	.5385	.7923	.8905
	double	.0348	.7203	.9132	.9422
Released checkpoint + reconstructed 3R (compat.)	single	.0273	.6052	.8078	.8950
	double	.0308	.7655	.9177	.9378
Released checkpoint + MMD-PO (compat.)	single	.0256	.6422	.8233	.8402
	double	.0291	.7851	.9201	.9378

G.2 NORMAN COMBINATION-HOLDOUT VALIDATION

We next evaluated the adapted scDFM implementation on four fixed Norman combination-holdout folds (Norman et al., 2019). The immutable dataset contains 84,986 cells across 227 conditions. Predictions use a fixed 1,000-gene metric panel and are scored separately on held-out single- and double-gene perturbations with Cell-Eval v0.6.8 (Adduri et al., 2026). Table 10 reports the unweighted mean and sample standard deviation of fold-level metrics; this variation reflects held-out data splits rather than repeated training seeds. Norman provides a Cell-Eval validation only and has no q-AnDCG ranking evaluation.

We then compared five systems on fold 1, the only combination-holdout fold completed for every arm (Table 11). Under this common evaluator, both short post-training checks descriptively improve MAE, Pearson-Δ, and DE-LFC Spearman relative to the released-checkpoint compatibility baseline, while discrimination score is mixed. "Compat." means that the public weights were evaluated through our locally adapted implementation (`adapted_scdfm_not_release_equivalent`), not the official release runner. The Replogle-initialized row subsequently receives 200,000 Norman condition sets and is therefore not a zero-shot transfer result.

The fold-1 3R and MMD-PO arms start from the same released checkpoint and each uses 384 condition sets over 12 distributed optimizer steps, but the remaining contracts differ materially: reconstructed 3R uses 32 rollouts per condition, learning rate 2×10^{-6}, and EMA weights for evaluation, whereas MMD-PO uses eight rollouts, learning rate 10^{-5}, raw policy weights, and a distinct combination of the Gene-Effect and Pathway-Enrichment rewards Accordingly, the fold-1 ordering is an exploratory portability result, not evidence that MMD-PO causally outperforms PerturbCellRL. In particular, MMD-PO's discrimination score is lower than the released baseline for held-out singles and tied with reconstructed 3R for doubles.

H SUPPORTING SYNTHETIC AND PATHWAY ANALYSES

H.1 CONTROLLED SYNTHETIC DESIGN AND PATHWAY ELIGIBILITY

Panels B–C use a 2,048-world mechanics ensemble with 2,000 genes, five 200-gene latent programs, and eight cellular contexts. Panels D–E use a separate 2,048-world factorial ensemble with eight disjoint, fully annotated 160-gene pathways, one for each offset $\times$ variability $\times$ reliability combination; context-response shift is disabled in that ensemble. In both, a structural teacher generates responses for a distinct DeepSets student. Both designs use 64 calibration perturbations measured in three 64-cell target sets, 40 post-training steps, $K = 8$ candidates per update by default, and 256 unseen evaluation perturbations. Within each ensemble, all arms branch from the same SFT1 checkpoint and share the update budget. Panels B–C use the source-context independent screen; held-out-context outcomes are serialized separately. Clean latent targets and screen outcomes are evaluation-only.

Continued SFT is a supervised control. GRPO, online DPO, and MMD-PO receive the same detached scalar reward; GRPO and DPO use the same Gaussian conversion of the deterministic student output, whereas MMD-PO operates directly on sampled cell populations. For gene g, the imposed standard deviation is $\sigma_g = s\,\mathrm{SD}_{\mathrm{control},g}$; the prespecified scale is $s = 0.3$, and online DPO uses $\beta = 10$.

Synthetic AnDCG@5% assigns relevance 3/2/1/0 to the top 5%, next 5%, next 10%, and remaining perturbations by exact latent effect, uses linear gain, and subtracts the exact random-ranking expectation. Exact prediction ties receive expected credit. Unlike empirical q-AnDCG, it has no significance gate because the simulator supplies the latent effects directly; the empirical and synthetic values are therefore not numerically interchangeable.

Across the 100-fold scale sweep, the paired MMD-PO advantage remains 0.0735–0.0803 over GRPO and 0.0754–0.0783 over online DPO, with every 95% interval excluding zero. Within the tested Gaussian-surrogate family, the ordering is therefore not explained by the prespecified scale alone. Changing s alters both rollout dispersion and score-gradient magnitude, so this is not a general noise-robustness claim. In the serialized secondary 8×7 DPO β-by-scale grid, the across-world mean at each setting ranges from 0.0634 to 0.0824, below SFT1 (0.0924) and MMD-PO (0.1536); because this is a final-world sensitivity, it is not used to choose β or s.

Reward-temperature sensitivity. The MMD-PO reward temperature α is distinct from the Gaussian conversion scale s used by GRPO and online DPO. In a separate paired confirmation across 512 fresh synthetic worlds, increasing α from 0.3 to 3.0 improved source-context AnDCG@5% by 0.0212 (95% CI, 0.0091–0.0333) and held-out-context AnDCG@5% by 0.0123 (0.0008–0.0237). Normalized effective sample size increased from 0.4150 to 0.8467, whereas source- and held-out-context Pearson-Δ changes were indistinguishable from zero. Thus flatter reward weighting improved ranking in this two-level synthetic comparison, but the tested values do not establish $\alpha = 3.0$ as universally optimal.

Annotation completeness is a distinct boundary condition: MMD-PO minus GRPO is 0.0043 (-0.0016–0.0102) when 2% of causal-program genes are retained and 0.0754 (0.0686–0.0821) at full coverage; MMD-PO is below SFT1 at 2% and above it from 25% coverage onward in the tested grid.

On the calibration-selected pathway, MMD-PO improves over SFT1 by 0.0595 (95% CI, 0.0526–0.0664), over GRPO by 0.0692 (0.0618–0.0766), and over online DPO by 0.0663 (0.0589–0.0737) AnDCG@5%. The corresponding averages over the other seven pathways are slightly negative. In the full factorial, the MMD-PO minus SFT1 contrast changes from $+0.0606$ in the eligible cell to -0.0148 with high common offset, -0.0066 with low variability, and $+0.0068$ with low reliability. Within this constructed factorial, this supports calibration-based eligibility rather than uniform pathway improvement.

Two negative controls delimit that claim. Under high objective conflict, the prespecified 5%-budget MMD-PO minus GRPO oracle-recovery contrast is 0.0039 (-0.0072–0.0149). Moreover, the canonical ranking gain over GRPO ($+0.0799$) accompanies a Pearson-Δ change of -0.0244

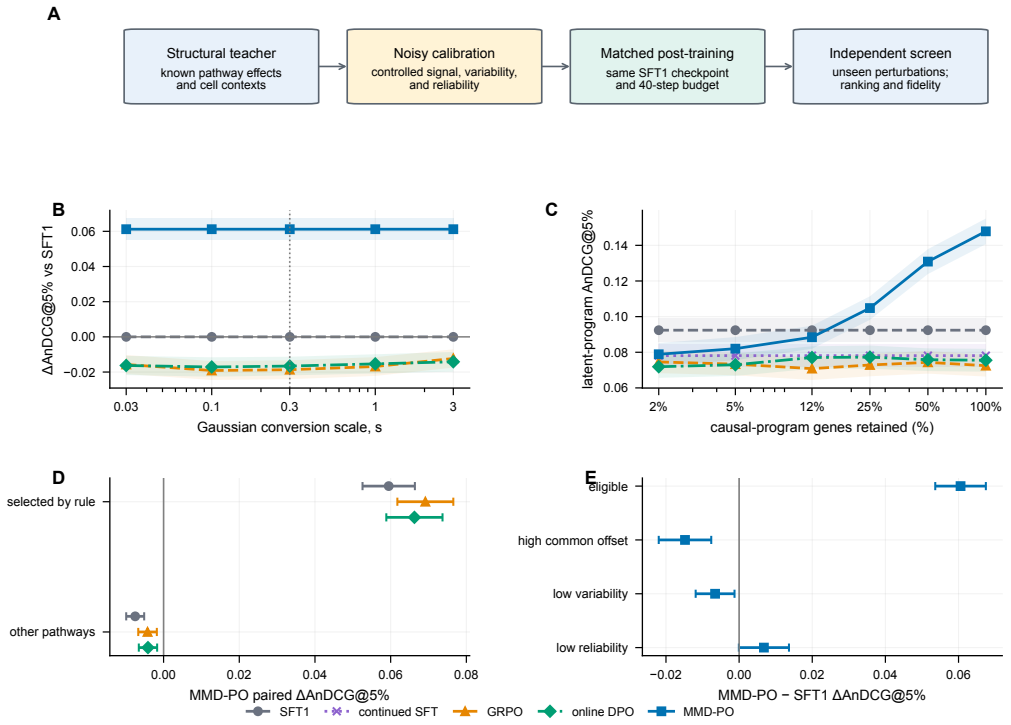


Figure 4: **Synthetic design and boundary conditions. A,** Minimal benchmark workflow. **B,** Sensitivity to the Gaussian policy-conversion scale used by GRPO and online DPO ($\sigma_g = s\,\mathrm{SD}_{\mathrm{control},g}$; DPO $\beta = 10$); SFT1 and MMD-PO are density-free, scale-invariant references, and the dotted line marks $s = 0.3$. **C,** Ranking performance as the retained fraction of causal-program genes varies. **D,** Paired MMD-PO contrasts against SFT1, GRPO, and online DPO on the pathway chosen by a calibration-only rule and on the other seven pathways. **E,** MMD-PO minus SFT1 in the eligible low-offset/high-variability/high-reliability cell and its three one-factor departures. The rule identifies the planted eligible cell in 98.1% of worlds. Panels B–C and D–E come from separate ensembles of 2,048 independently generated worlds; methods and conditions are paired within each world. Lines, bands, and error bars are means and pointwise 95% Student-t intervals across worlds.

$(-0.0256{-}{-}0.0231)$ relative to SFT1. Ranking and predictive fidelity must therefore be reported separately.

H.2 Frozen Hallmark pathway audit

This post hoc audit rescores predictions from three already-frozen HepG2 checkpoints; the audit itself performs no retraining or checkpoint selection. Its pathway AnDCG@5% uses ungated percentile bands with exponential gain and is distinct from the linear-gain synthetic metric.

At 5%, the bidirectional macro-average is 0.348 for SFT-20, 0.368 for GRPO, and 0.392 for MMD-PO. Relative to GRPO, the paired MMD-PO contrasts are $+0.0258$ for activation, $+0.0233$ for inhibition, and $+0.0246$ for their bidirectional mean. All three descriptive pathway-bootstrap intervals lie above zero; they summarize heterogeneity across the fixed, overlapping pathway set and are not biological- or seed-level confidence intervals. Online DPO was not evaluated in this frozen empirical audit and is therefore shown only in the controlled synthetic study.

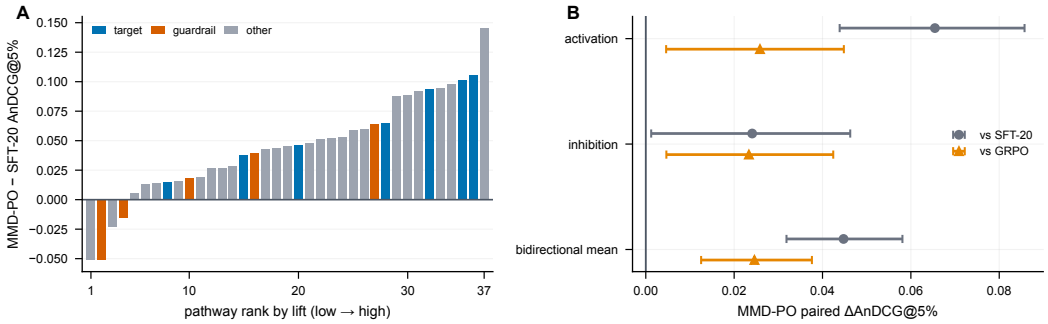


Figure 5: **Frozen Hallmark pathway audit on HepG2.** The audit contains 380 held-out perturbations and 1,898 genes; 37 of 50 MSigDB Hallmark 2026.1 terms pass the > 15 mapped-gene threshold (Liberzon et al., 2015). **A,** Ordered differences in per-pathway bidirectional mean AnDCG@5% (activation and inhibition, each averaged over two matched reveal orders); colors denote seven configured example target programs, five configured guardrails, and the remainder. MMD-PO is higher on 33 of 37 pathway means, increasing the macro-average from 0.348 to 0.392. **B,** Paired pathway-level MMD-PO contrasts against SFT-20 and GRPO for activation, inhibition, and their bidirectional mean; error bars are descriptive 95% pathway-bootstrap intervals over this fixed, overlapping pathway collection, not biological- or training-seed uncertainty. The two matched reveal orders measure inference sensitivity rather than biological replication. All pathways contributed genes to the training reward union, so this is not unseen-pathway generalization.