LEARNING HETEROGENEOUS SINGLE-CELL PERTURBATION RESPONSES WITH MOMENT-CONDITIONED SCHRÖDINGER TRANSPORT

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting single-cell perturbation responses requires learning from unpaired populations with limited coverage of interventions and cellular contexts. A predicted mean response does not determine the corresponding population: translating control cells by a common shift leaves their centered distribution unchanged. We introduce Moment-Conditioned Schrödinger Transport, a divide-and-conquer framework that separates condition-level mean prediction from learning centered population changes. A condition-level predictor estimates the target mean and provides a conditioning signal for a shared conditional Schrödinger bridge trained on centered control and treated populations. The bridge uses a Brownian reference to regularize transport and operates in a response-aware representation incorporating both average-response directions and within-condition variation. At inference, it transforms target-context control cells, followed by expression-space recovery that enforces nonnegativity and consistency between the generated sample mean and the prescribed nonnegative mean. We evaluate the framework on Norman2019 and multi-context perturbation data under unseen gene combinations, unseen perturbations, and few-shot or zero-treated-shot context transfer. In these evaluated settings, the complete framework achieves lower maximum mean discrepancy in expression and PCA spaces than the evaluated distributional baselines, with competitive mean-response accuracy.

1 INTRODUCTION

Predicting cellular responses to genetic interventions is a central goal of functional genomics and an important step toward scalable *in silico* experimentation. Single-cell perturbation screens measure transcriptional responses across many interventions, but experimental coverage remains limited relative to the space of possible perturbations and cellular contexts (Dixit et al., 2016; Norman et al., 2019; Replogle et al., 2022). Moreover, destructive measurements yield separate control and treated populations rather than paired observations of individual cellular responses (Bunne et al., 2023). The challenge is therefore to learn transferable perturbation effects from unpaired populations (Figure 1(a)) and predict outcomes in conditions whose treated populations are unobserved.

Existing methods support unseen-condition prediction through foundation-model representations and biological priors (Cui et al., 2024; Roohani et al., 2024), or through population-level generative and transport models (Bunne et al., 2023; Chi et al., 2026; Yu et al., 2026; Yan et al., 2026). Yet several deep models have failed to outperform linear or additive baselines on unseen perturbations (Ahlmann-Eltze et al., 2025), and improved distributional fit need not improve mean responses (Yan et al., 2026). Aggregate-effect prediction and population modeling therefore remain related but distinct objectives.

Population responses reflect pre-existing cellular states, intervention effects, and their interactions, motivating models that separate cellular background from perturbation-related latent factors (Lotfollahi et al., 2023; Jiang et al., 2026). However, unpaired observations do not directly supervise this attribution: profiles from different cells do not reveal which aspects of an individual cell were preserved or altered. Interpreting such decompositions therefore requires additional generative assumptions (Jiang et al., 2026). This motivates a complementary question: *can population prediction*

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.

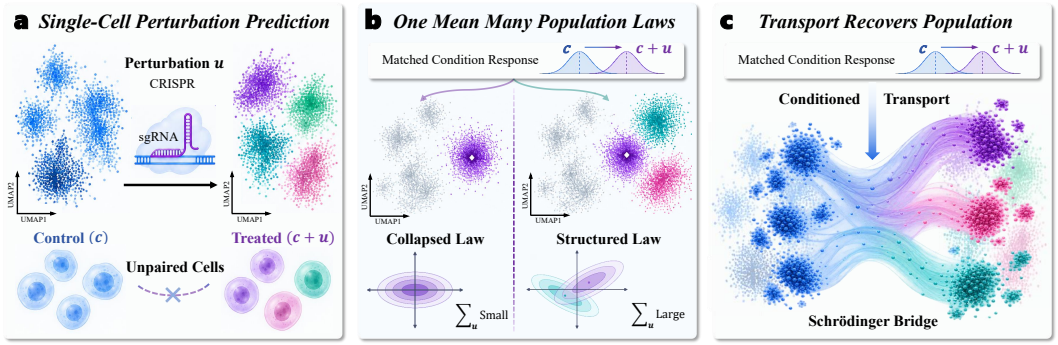


Figure 1: **Connecting mean prediction with population learning.** (a) A condition mean $m(c, u)$ summarizes the aggregate response. (b) Equal means can accompany different dispersion and joint structure. (c) Centered transport and nonnegative reconstruction recover a population whose sample mean matches the predicted target mean.

be organized around statistics estimable from observed populations, rather than first identifying their latent biological sources?

The average response offers a useful starting point: aggregation estimates the mean within observed conditions, while competitive simple predictors suggest that useful average-effect predictions need not require a complete population generator (Ahlmann-Eltze et al., 2025). Unseen-condition mean prediction remains a generalization problem, but can be addressed separately from population generation. Even an accurate mean leaves population structure undetermined (Figure 1(b)): a common predicted shift changes the control population's center but not its centered distribution. Population prediction must therefore determine whether and how the treated population departs from this translated control distribution.

This motivates a *divide-and-conquer* perspective separating condition-level mean prediction from learning centered population changes. It raises two challenges. **First, how can population generation use a separately predicted mean while preserving it in the generated cells? Second, how can it learn transferable changes beyond translation from unpaired populations?** The latter requires capturing changes in dispersion or coordinated expression along directions with little average shift, which mean-only representations may overlook.

We introduce *Moment-Conditioned Schrödinger Transport*, a framework connecting these tasks (Figure 1(c)). A condition-level predictor specifies the target mean, while a shared conditional Schrödinger bridge learns transport between centered control and treated populations from observed training conditions. A Brownian reference regularizes transport ambiguity under unpaired observations. The bridge operates in a response-aware representation retaining both average perturbation shifts and within-condition variation, allowing it to capture distributional changes weak or absent in condition means. At inference, transport acts on target-context controls without treated observations from the test condition. Expression-space recovery enforces nonnegativity and equality between the generated sample mean and the predicted target mean. This is a statistical division of learning responsibilities, not an assumption of independent mechanisms for mean effects and heterogeneity.

We evaluate unseen gene combinations in Norman2019 and multi-context protocols for unseen perturbations and few-shot or zero-treated-shot context transfer. End-to-end comparisons show lower expression- and PCA-space maximum mean discrepancy (MMD) than the evaluated distributional baselines, with competitive or improved mean-response accuracy. Matched-mean comparisons in unseen-combination and unseen-perturbation settings fix the representation and recovery operator, revealing setting-dependent gains beyond control-shape transfer, with larger centered-MMD reductions for combinations. Our contributions are:

- We formulate a divide-and-conquer approach separating transferable condition-level mean prediction from learning the centered population distribution left undetermined by that mean.
- We combine response-aware conditional Schrödinger transport with nonnegative, mean-consistent recovery to connect learned population changes to prescribed aggregate responses.
- We evaluate end-to-end generalization and centered population learning separately through matched-mean comparisons under shared representation and recovery.

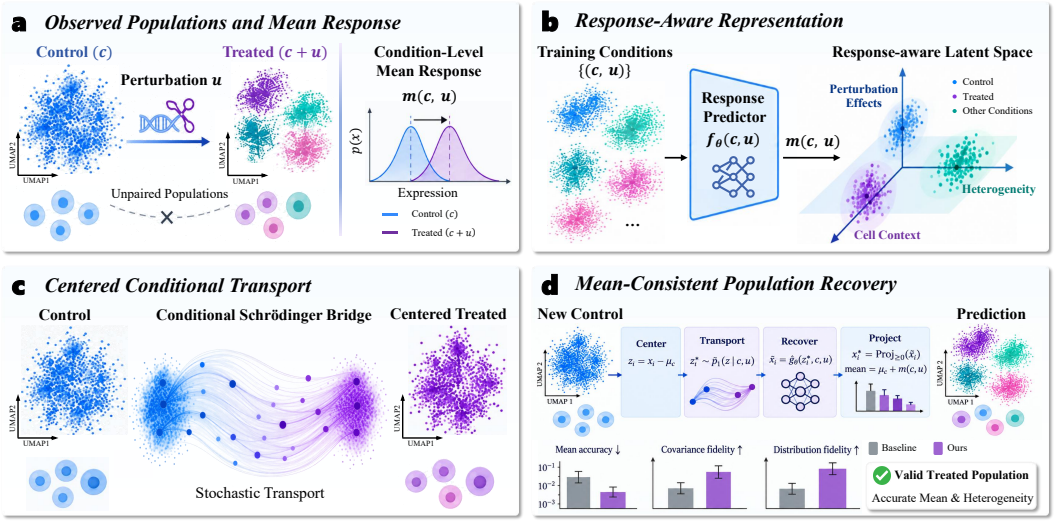


Figure 2: **Overview of Moment-Conditioned Schrödinger Transport.** A protocol-specific predictor estimates the target mean from permitted observations and covariates. Its response coordinates explicitly condition transport between centered control and treated training populations. Transported samples are decoded together with source variation outside the response chart, then projected onto the nonnegative set with the prescribed population mean.

2 Related Work

Foundation Models for Single-Cell Biology. Single-cell foundation models, including scBERT (Yang et al., 2022), Geneformer (Theodoris et al., 2023), scGPT (Cui et al., 2024), scFoundation (Hao et al., 2024), and TranscriptFormer (Pearce et al., 2026), learn reusable representations from large transcriptomic corpora. However, generalization across cellular states does not ensure transfer to unseen interventions: evaluated models such as scGPT and scFoundation have struggled to outperform linear or additive baselines in mean-response prediction (Ahlmann-Eltze et al., 2025).

Perturbation-Specific Models. Perturbation models use latent and compositional representations (Lotfollahi et al., 2019; 2023; Hetzel et al., 2022), biological priors (Roohani et al., 2024), and invariant–responsive factor separation (Jiang et al., 2026), while State scales transition learning across contexts (Adduri et al., 2026). Complementary transport and generative models, including CellOT, PerturbNet, scDFM, and PRiMeFlow, emphasize population fidelity (Bunne et al., 2023; Yu et al., 2025; 2026; Yan et al., 2026). Accurate means nevertheless leave population structure undetermined, and improved distributional fit need not improve mean prediction. Our divide-and-conquer formulation coordinates these objectives by preserving a separately predicted condition mean while learning centered population changes beyond a common shift.

3 Method

We introduce *Moment-Conditioned Schrödinger Transport*, which combines condition-level mean prediction with learning changes in centered single-cell populations. The mean predictor supplies both a conditioning signal for transport and the aggregate response used in expression-space recovery. A conditional bridge learns population transformations in a response-aware representation, and a final projection produces nonnegative cells with the prescribed sample mean (Figure 2).

3.1 Problem Formulation

Let c denote a cellular context, u a perturbation or a set of perturbed genes, and $x \in \mathbb{R}^G_+$ an expression vector in the stored, preprocessed expression space. Control and treated distributions are denoted by ν^0_c and $\nu^1_{c,u}$, with population means μ_c and $m_{c,u}$. Their statistical decomposition is

$$X^0_c = \mu_c + R^0_c, \qquad X^1_{c,u} = m_{c,u} + R^1_{c,u}, \qquad \mathbb{E}[R^0_c] = \mathbb{E}[R^1_{c,u}] = 0. \tag{1}$$

This separates the population center from its centered distribution, without assuming independent biological sources for these components.

Let $\mathcal{T}$ contain the observed training conditions. We distinguish the sample mean $\bar{\mu}_c$ of the available controls, the treated training mean $\bar{m}_{c,u}$ for $(c,u) \in \mathcal{T}$, and a predicted mean for an unobserved condition. The training effect is $\Delta_{c,u} = \bar{m}_{c,u} - \bar{\mu}_c$. Control and treated cells are unpaired. Target-context controls and known intervention identities remain available under the transfer protocols, but held-out treated cells are not used to fit the predictor or transport model. All mean constraints below refer to the stored expression space, not to an unspecified raw-count space.

3.2 Condition-Level Mean Prediction

Condition covariates. Let f_g be a fixed gene feature and a_c a context feature derived from available controls. Features for a multi-gene intervention are summed, and the predictor uses both the marginal features and their interaction:

$$f_u = \sum_{g \in u} f_g, \qquad \phi_{c,u} = \big[f_u;\, a_c;\, \mathrm{vec}(f_u a_c^\top)\big]. \tag{2}$$

The multi-context implementation uses principal-component features of the external ESM2 gene vocabulary; the Norman and Replogle implementations use fixed intervention encodings. These covariates do not require treated observations from the target test condition.

Prediction branches. The mean predictor uses observed effects when available, then compositional or cross-context transfer, and otherwise a feature-based regression. Let $\mathcal{S}_{c,u} = \{c' \neq c : (c',u) \in \mathcal{T}\}$. The following cases are applied in order:

$$\widehat{\Delta}_{c,u} = \begin{cases} \Delta_{c,u}, & (c,u) \in \mathcal{T}, \\ \displaystyle\sum_{g \in u} \Delta_{c,\{g\}}, & (c,\{g\}) \in \mathcal{T} \text{ for every } g \in u, \\ \displaystyle\frac{1}{|\mathcal{S}_{c,u}|} \sum_{c' \in \mathcal{S}_{c,u}} \Delta_{c',u}, & \mathcal{S}_{c,u} \neq \varnothing, \\ \widehat{\Delta}^{\mathrm{reg}}_{c,u}, & \text{otherwise.} \end{cases} \tag{3}$$

For the regression branch, a ridge regressor r_ψ predicts coefficients in a training-fitted mean-response basis:

$$\widehat{\Delta}^{\mathrm{reg}}_{c,u} = b_\Delta + A_\Delta r_\psi(\phi_{c,u}). \tag{4}$$

Here b_Δ and A_Δ are the stored response center and basis; the regressor is fitted to training response coefficients with regularization 10 and the LSQR solver. For a single target g represented in the expression panel, the fallback additionally replaces its target coordinate by $\widehat{\Delta}^{\mathrm{reg}}_{c,\{g\},j(g)} = \rho_{\mathrm{tar}} \bar{\mu}_{c,j(g)}$ when stored target-effect ratios are available, where ρ_{tar} is their median. This correction is part of the mean predictor, not of transport.

We distinguish the raw prediction from the nonnegative recovery target:

$$\widetilde{m}_{c,u} = \bar{\mu}_c + \widehat{\Delta}_{c,u}, \qquad \widehat{m}_{c,u} = [\widetilde{m}_{c,u}]_+, \tag{5}$$

where $[\cdot]_+$ acts coordinate-wise. Norman's held-out double perturbations use additive single effects; E1 uses the regression branch with the target-coordinate correction; E2 and E3 transfer the mean effect of the same intervention from observed source contexts. The Replogle held-out-cell protocol uses the training mean of the observed condition. All branches supply a condition-level mean to the same downstream transport and recovery interface.

3.3 Response Geometry and Moment Conditioning

A representation based only on condition means can omit changes in within-condition variability. We therefore combine mean-response directions, residual variation, and known target-gene coordinates. Let $D = [\Delta_{c,u}]_{(c,u)\in\mathcal{T}}$, and let $\mathcal{R}$ collect centered training treated cells and available control cells as columns. Writing Basis_k for the implementation's SVD-based basis extraction, we construct

$$\begin{aligned} B_{\mathrm{mean}} &= \mathrm{Basis}_{d_{\mathrm{mean}}}(D), & P^{\perp}_{\mathrm{mean}} &= I - B_{\mathrm{mean}} B_{\mathrm{mean}}^\top, \\ B_{\mathrm{het}} &= \mathrm{Basis}_{d_{\mathrm{het}}}(P^{\perp}_{\mathrm{mean}} \mathcal{R}), & B &= \mathrm{orth}[B_{\mathrm{mean}}, B_{\mathrm{het}}, B_{\mathrm{tar}}]. \end{aligned} \tag{6}$$

Here B_{tar} contains coordinate vectors for known intervention targets present in the expression panel, including held-out target identities but not their outcomes, and $B^\top B = I_d$. Available target-context controls may enter $\mathcal{R}$. For each response direction, the scale is the standard deviation of projected, subsampled training treated residuals, floored at 0.05; these scales form the diagonal matrix S. The sampling limits and retained ranks are specified in Appendix B.

A cell centered at a is encoded as $z(x; a) = S^{-1}B^\top(x - a)$. The empirical training endpoints are

$$\widehat{q}^0_c = \frac{1}{N^0_c}\sum_i \delta_{z(x^0_{c,i};\bar{\mu}_c)}, \qquad \widehat{q}^1_{c,u} = \frac{1}{N^1_{c,u}}\sum_j \delta_{z(x^1_{c,u,j};\bar{m}_{c,u})}. \tag{7}$$

Thus, treated endpoints are centered by their observed training means, not by estimated test-condition means. The network explicitly receives the predicted response in addition to the condition covariates:

$$h_{c,u} = \left[\phi_{c,u};\ S^{-1}_{\mathrm{cond}}B^\top_{\mathrm{cond}}(\widetilde{m}_{c,u} - \bar{\mu}_c)\right]. \tag{8}$$

The conditioning chart is the chart stored with the mean predictor. It equals the transport chart in the default configurations, but is kept fixed when a matched geometry ablation changes the transport representation. Importantly, conditioning uses the raw prediction $\widetilde{m}$, whereas nonnegative recovery uses $\widehat{m}$. Observed training conditions supply their training means to this input; unobserved test conditions use Eq. 3. The implementation does not cross-fit the training mean inputs.

3.4 Conditional Schrödinger Transport

The centered populations do not identify a cell-wise coupling. We learn conditional transport by combining entropy-regularized endpoint matching with Brownian-bridge interpolation. For empirical source and target minibatch laws $\widehat{q}^0_{c,\mathrm{mb}}$ and $\widehat{q}^1_{c,u,\mathrm{mb}}$, the endpoint coupling is

$$\pi^\star_{\mathrm{mb}} = \underset{\pi\in\Pi(\widehat{q}^0_{c,\mathrm{mb}},\widehat{q}^1_{c,u,\mathrm{mb}})}{\arg\min} \ \mathbb{E}_\pi\|z_1 - z_0\|_2^2 + 2\sigma^2\,\mathrm{KL}\left(\pi \,\middle\|\, \widehat{q}^0_{c,\mathrm{mb}} \otimes \widehat{q}^1_{c,u,\mathrm{mb}}\right). \tag{9}$$

The same σ determines the transport regularization and bridge noise. We approximate this coupling using log-domain Sinkhorn iterations, then sample endpoint pairs from the plan with replacement. This is a finite-sample neural approximation motivated by Brownian Schrödinger transport, not a claim of an exact finite-path-KL solution for atomic empirical endpoints.

For an endpoint pair (z_0, z_1), we sample $t \sim \mathcal{U}(\tau, 1-\tau)$ with $\tau = 0.01$ and $\epsilon \sim \mathcal{N}(0, I_d)$. The interpolation and conditional flow target are

$$\begin{aligned} z_t &= (1-t)z_0 + tz_1 + \sigma\sqrt{t(1-t)}\,\epsilon, \\ v^\star_t &= z_1 - z_0 + \frac{\sigma(1-2t)}{2\sqrt{t(1-t)}}\,\epsilon. \end{aligned} \tag{10}$$

A shared network outputs a flow field $v_\theta(z, t \mid h)$ and a scaled score $e_\theta(z, t \mid h)$, with

$$e^\star_t = -\epsilon, \qquad s_\theta(z, t \mid h) = \frac{e_\theta(z, t \mid h)}{\sigma\sqrt{t(1-t)}}. \tag{11}$$

We optimize the implemented score–flow loss

$$\mathcal{L}_{\mathrm{SB}}(\theta) = \frac{1}{d}\,\mathbb{E}\left[\|v_\theta(z_t, t \mid h_{c,u}) - v^\star_t\|_2^2 + \lambda_s\|e_\theta(z_t, t \mid h_{c,u}) + \epsilon\|_2^2\right], \qquad \lambda_s = 0.25. \tag{12}$$

The expectation is over training conditions, sampled endpoint pairs, interior times, and interpolation noise. Predicting the scaled score avoids directly regressing a target that diverges at the endpoints; truncating the time interval also bounds the conditional flow target's time-dependent coefficient. The mean predictor and response charts are held fixed during bridge optimization.

At inference, the selected control batch is centered by its own sample mean $\bar{x}_{\mathrm{src}}$, and $Z^{(i)}_0 = z(x^0_i; \bar{x}_{\mathrm{src}})$. We approximate the conditional SDE

$$dZ_t = \left[v_\theta(Z_t, t \mid h_{c,u}) + \frac{\sigma^2}{2}s_\theta(Z_t, t \mid h_{c,u})\right]dt + \sigma\,dW_t \tag{13}$$

with $K = 50$ Euler–Maruyama updates and coefficients evaluated at $t_k = (k + \frac{1}{2})/K$. The probability-flow ablation uses only v_θ, omitting both the score correction and Brownian increments. No treated samples from the target test condition enter this procedure.

3.5 Moment-Consistent Population Recovery

Let $\widehat{z}_i^1$ be the terminal response coordinates, $\bar{z}^1 = n^{-1}\sum_i \widehat{z}_i^1$, and $r_i^0 = x_i^0 - \bar{x}_{\mathrm{src}}$. The implemented decoder centers the transported coordinates and retains source variation outside the response chart:

$$\begin{aligned}\widetilde{x}_i &= \widetilde{m}_{c,u} + (I - BB^\top)r_i^0 + BS(\widehat{z}_i^1 - \bar{z}^1) \\ &= \widetilde{m}_{c,u} + r_i^0 + B\left[S(\widehat{z}_i^1 - \bar{z}^1) - B^\top r_i^0\right].\end{aligned} \tag{14}$$

Before the support projection, this construction has sample mean $\widetilde{m}_{c,u}$ up to floating-point error.

For the matrix $\widetilde{X}$ with rows $\widetilde{x}_i^\top$, the final output is

$$\widehat{X}_{c,u} = \underset{Y\geq 0:\ n^{-1}\mathbf{1}^\top Y = \widehat{m}_{c,u}^\top}{\arg\min} \|Y - \widetilde{X}\|_F^2, \qquad \frac{1}{n}\mathbf{1}^\top \widehat{X}_{c,u} = \widehat{m}_{c,u}^\top. \tag{15}$$

Since $\widehat{m}_{c,u} \geq 0$, the feasible set is nonempty, closed, and convex, and the Euclidean projection is unique. It separates into gene-wise simplex projections: $\widehat{X}_{ij} = \max\{\widetilde{X}_{ij} - \theta_j, 0\}$, where θ_j enforces column sum $n\widehat{m}_{c,u,j}$; zero-mean genes yield zero columns. The equality in Eq. 15 holds in exact arithmetic and is checked numerically in the implementation.

Consequently, mean-response accuracy is determined by the nonnegative mean prediction, not by the transport sampler. Transport changes the population structure supplied to recovery. The projection can alter variance, covariance, and the source component outside the response chart, so those quantities are not claimed to be preserved in the final output. Matched-mean comparisons therefore apply the same recovery operator to control-shape transfer and learned transport.

4 Experiments

4.1 Experimental Setup

Datasets and Protocols. We construct three generalization protocols from the public support resources of the 2025 Virtual Cell Challenge (VCC) (Roohani et al., 2025), rather than its official leaderboard split. The data combine K562 and RPE1 screens (Replogle et al., 2022) with HepG2 and Jurkat screens (Nadig et al., 2025), retaining 53 shared perturbations covered by the common representation. E1 uses 10 perturbation genes for training, five for validation, and the remainder for testing across contexts. E2 uses all 53 perturbations in three source contexts and 10 treated training conditions in HepG2, with separate target validation and test conditions. E3 excludes HepG2 treated outcomes from both training and model selection. Target-context controls remain available. After cell-count filtering, E1 contains 150 test context–perturbation pairs, while E2 and E3 share 37 HepG2 test perturbations. Norman2019 (Norman et al., 2019) evaluates generalization from observed single perturbations to 125 held-out double perturbations. A separate Replogle2022 held-out-cell evaluation covers 256 observed conditions.

Baselines and Metrics. We compare against the representation-based, generative, and transport baselines listed in Tables 1 and 2. All baselines in Tables 1 and 2 are reproduced under the same experimental protocol within each setting. Native CellOT is excluded from unseen-gene and unseen-combination experiments. E1–E3 use 5,000 genes selected from control variability, retaining intervention targets. MAE measures condition-mean accuracy; DES and PDS provide additional response-evaluation scores. MMD-GEX and MMD-PCA measure population discrepancy in expression and control-fitted PCA spaces. Centered MMD-PCA (cMMD-PCA) removes each population's empirical mean before applying the same evaluation map and kernel. We additionally report off-diagonal covariance error $\mathrm{Cov}_{\mathrm{off}}$ and response-strength Wasserstein distance $W_{1,\mathrm{resp}}$. These evaluate unpaired populations, not individual-cell counterfactual accuracy. Substitution intervals use condition-cluster bootstrap resampling.

4.2 Perturbation and Context Generalization

Table 1 evaluates the complete prediction pipeline. Ours achieves the lowest MAE and both MMD scores, and the highest DES and PDS, among the displayed methods in E1–E3. Relative to the

Table 1: Results of perturbation and context generalization on protocols constructed from VCC (Roohani et al., 2025).

Method	MAE ↓	DES ↑	PDS ↑	MMD-GEX ↓	MMD-PCA ↓
E1: unseen perturbations across four cellular contexts					
PerturbNet-dec. (Yu et al., 2025)	0.2173	0.00501	0.5230	13.4218	11.3383
PerturbNet-pred. (Yu et al., 2025)	0.3102	0.00264	0.5148	8.2507	12.8078
Conditional Monge (Driessen et al., 2026)	0.1117	0.00472	0.5143	9.0416	1.7135
Departures (Chi et al., 2026)	0.2251	0.00402	0.5152	6.1868	5.8179
PRiMeFlow (Yan et al., 2026)	0.1379	0.00236	0.5179	3.1775	2.0231
scDFM (Yu et al., 2026)	0.3162	0.00249	0.5062	8.5809	12.8727
PerturbedVAE (Jiang et al., 2026)	0.2148	0.00483	0.5162	12.8497	10.5365
Ours	**0.1065** (−4.7%)	**0.00595** (+18.8%)	**0.5271** (+0.8%)	**2.4805** (−21.9%)	**1.1137** (−35.0%)
E2: few-shot transfer to HepG2					
CellOT (Bunne et al., 2023)	0.3090	0.00032	0.5200	15.9347	12.5642
PerturbNet-dec. (Yu et al., 2025)	0.2796	0.00016	0.5140	18.7240	19.3980
PerturbNet-pred. (Yu et al., 2025)	0.3643	0.00016	0.5096	12.5879	18.2878
Conditional Monge (Driessen et al., 2026)	0.1313	0.00003	0.5164	8.8090	2.1398
Departures (Chi et al., 2026)	0.2313	0.00019	0.5220	6.9805	3.7087
PRiMeFlow (Yan et al., 2026)	0.1718	0.00003	0.5174	4.8224	2.6856
scDFM (Yu et al., 2026)	0.2799	0.00016	0.5091	8.4644	9.4443
PerturbedVAE (Jiang et al., 2026)	0.2975	0.00016	0.5147	20.1288	21.2480
Ours	**0.1307** (−0.5%)	**0.00091** (+184.4%)	**0.5464** (+4.7%)	**3.7212** (−22.8%)	**1.5306** (−28.5%)
E3: transfer to HepG2 without treated target-context training cells					
CellOT (Bunne et al., 2023)	0.3248	0.00031	0.5190	16.7884	10.8943
PerturbNet-dec. (Yu et al., 2025)	0.3102	0.00006	0.5103	20.9381	22.5284
PerturbNet-pred. (Yu et al., 2025)	0.3879	0.00019	0.5084	13.6018	20.6729
Conditional Monge (Driessen et al., 2026)	0.1538	0.00006	0.5123	9.2567	2.9076
Departures (Chi et al., 2026)	0.2640	0.00003	0.5140	8.0471	3.8721
PRiMeFlow (Yan et al., 2026)	0.3394	0.00017	0.5181	13.3473	20.7370
scDFM (Yu et al., 2026)	0.2861	0.00016	0.5259	8.8327	10.3959
PerturbedVAE (Jiang et al., 2026)	0.3214	0.00013	0.5138	21.5450	23.5519
Ours	**0.1316** (−14.4%)	**0.00062** (+100.0%)	**0.5444** (+3.5%)	**3.7331** (−53.6%)	**1.5326** (−47.3%)

Table 2: Response and population prediction on Norman2019 (Norman et al., 2019) double perturbations. C-RMSE denotes condition-mean RMSE, ΔP. denotes ΔPearson. All methods are reproduced under the same experimental protocol.

Method	L2 ↓	ΔP. ↑	MAE ↓	C-RMSE ↓	Cell RMSE ↓	MMD-GEX ↓	MMD-PCA ↓
GEARS (Roohani et al., 2024)	4.6797	0.4631	—	0.1514	0.5861	—	—
PerturbNet-dec. (Yu et al., 2025)	5.0544	0.4946	0.0331	0.0715	0.2830	3.9258	0.9105
PerturbNet-pred. (Yu et al., 2025)	25.3648	0.2043	0.3395	0.3587	0.4537	9.9837	3.5182
Conditional Monge (Driessen et al., 2026)	5.0477	0.5338	0.0326	0.0714	0.2829	4.0255	1.2322
Departures (Chi et al., 2026)	5.9356	0.2930	0.0429	0.0839	0.2863	0.8669	0.7366
PRiMeFlow (Yan et al., 2026)	5.6327	0.2844	0.0388	0.0797	0.2849	0.8794	1.1106
scDFM (Yu et al., 2026)	8.9727	0.1595	0.0903	0.1269	0.3024	1.8335	1.4085
PerturbedVAE (Jiang et al., 2026)	5.0452	0.5173	0.0332	0.0714	0.2829	5.8188	2.3094
Ours	**3.4171** (−27.0%)	**0.8346** (+56.4%)	**0.0257** (−21.2%)	**0.0483** (−32.4%)	**0.2788** (−1.4%)	**0.5528** (−36.2%)	**0.2563** (−65.2%)

strongest baseline for each distribution metric, MMD-GEX decreases by 21.9%, 22.8%, and 53.6%, while MMD-PCA decreases by 35.0%, 28.5%, and 47.3%, respectively. The mean-accuracy gain is small in E2: MAE is 0.1307 versus 0.1313 for Conditional Monge, despite substantially lower MMD. Our MMD-PCA is also similar in E2 and E3 (1.5306 versus 1.5326), showing limited degradation under the tested zero-treated-shot protocol. These are full-pipeline comparisons; the additional contribution of transport is examined through the controlled analyses below.

4.3 Combinatorial Perturbation Generalization

Table 2 evaluates generalization to 125 held-out double perturbations on Norman2019. Under the common experimental protocol, Ours reduces condition L2 from 4.6797 for GEARS, the strongest baseline on this metric, to 3.4171, a reduction of 27.0%. Compared with Conditional Monge, ΔPearson increases from 0.5338 to 0.8346 and MAE decreases from 0.0326 to 0.0257. At the population level, MMD-GEX and MMD-PCA decrease by 36.2% and 65.2%, respectively, relative to Departures, the strongest baseline for both metrics. The smaller improvement in Cell RMSE, from 0.2829 to 0.2788, motivates assessing population structure separately from mean-response and cell-level errors.

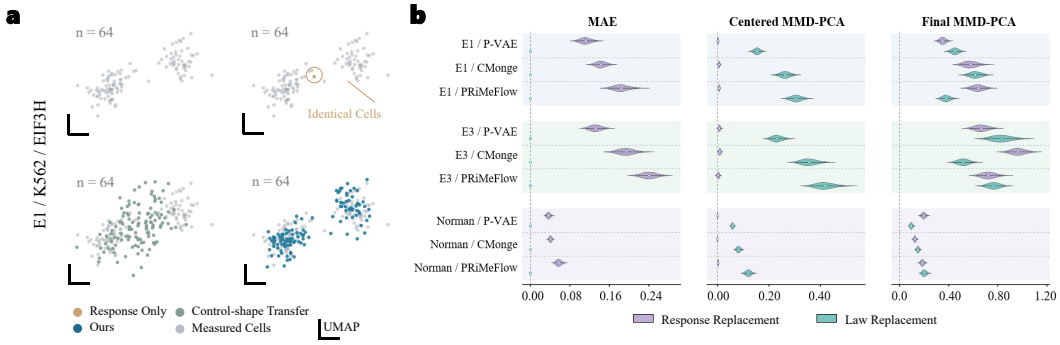


Figure 3: **Attribution of aggregate-response and population-law contributions.** (a) Predicted populations share the same expression-space mean but differ in structure ($n = 64$; E1, K562/EIF3H). (b) Mean (purple) and centered-law (teal) replacements under a common recovery operator, with condition-cluster bootstrap intervals.

Table 3: Matched component ablation on Norman2019 double perturbations at $d = 35$ and training seed 42. The comparison separates representation, endpoint coupling, and inference.

Variant	MMD-GEX ↓	MMD-PCA ↓	Var. error ↓	Cov_{off} ↓	$W_{1,\text{resp}}$ ↓
Mean-only response geometry	0.54780	0.25368	0.35847	1.1843	1.2339
Independent endpoint pairing	0.55486	0.25525	0.36293	1.1228	1.2331
Probability-flow inference	0.54177	0.25340	0.35291	1.1793	1.2191
Full model	0.54936	0.25214	0.36059	1.1403	1.2479

4.4 Understanding Response-Constrained Transport

Separating Mean and Population-Law Contributions. Figure 3(a) compares response-only prediction, control-shape transfer, and learned transport at the same expression-space mean. Response-only prediction assigns identical profiles to all cells, illustrating why mean accuracy does not determine population structure. Panel (b) crosses predicted means and centered samples from our model and each comparator under a common recovery operator. For example, with our mean fixed in E1, replacing Conditional Monge's centered samples with ours reduces cMMD-PCA from 0.70217 to 0.41992 and final MMD-PCA from 1.5042 to 1.1137, while MAE remains 0.10652. Thus, the improvement is not explained by changing the predicted mean. Because nonnegative recovery can alter centered structure, these substitutions measure controlled effects rather than an additive error decomposition.

Response Geometry, Coupling, and Inference. Table 3 compares matched variants on 125 Norman double perturbations at $d = 35$ and training seed 42. Mixed response geometry lowers Cov_{off} from 1.1843 to 1.1403 relative to mean-only directions. Entropic coupling yields lower MMD-PCA than independent pairing (0.25214 versus 0.25525), although independent pairing has lower covariance error. Stochastic inference reduces Cov_{off} from 1.1793 to 1.1403 relative to probability-flow inference, whereas the latter has lower MMD-GEX, variance error, and $W_{1,\text{resp}}$. The full configuration therefore gives the lowest MMD-PCA in this ablation but does not dominate all structural metrics. These results show metric-specific trade-offs, not that stochastic sampling is necessary to produce heterogeneous populations.

Mean-Consistent Recovery. Table 4 uses a separate 139-condition Norman diagnostic set and a common checkpoint. Without the response constraint, the mean discrepancy is 1.08×10^{-2}. Recentering reduces it to 2.20×10^{-8} but leaves 32.47% of expression entries negative. Full recovery gives a mean error of 2.51×10^{-8} with no negative entries. This validates agreement with the prescribed mean and nonnegative support, not preservation of population shape or agreement with the unknown true mean.

Table 4: Response-consistent recovery on the recorded Norman diagnostic set.

Recovery	Mean error ↓	Negative fraction ↓
Without response constraint	1.08×10^{-2}	0.0000
Recenter only	2.20×10^{-8}	0.3247
Full recovery	2.51×10^{-8}	0.0000

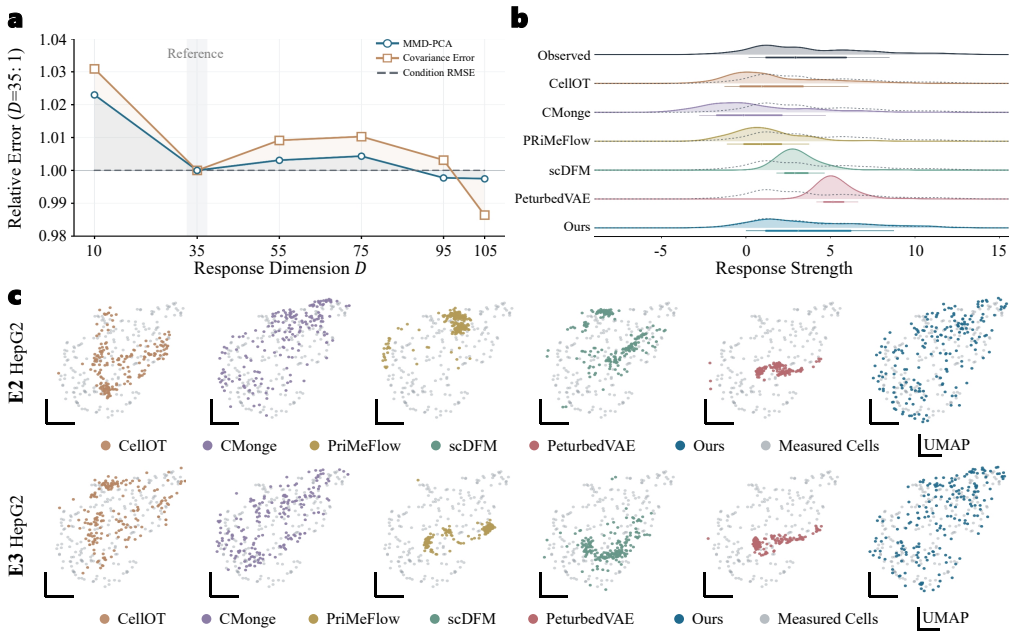


Figure 4: Population structure beyond the aggregate response. (a) Response-space capacity changes distributional diagnostics at a fixed predicted mean. (b) Empirical response-strength distributions expose differences in spread and quantiles. (c) Common-reference projections visualize population coverage in a fixed cross-context case. All panels use recorded outputs; coordinates and cases are fixed independently of method ranking, and two-dimensional visual proximity is not used as a substitute for high-dimensional distribution metrics.

Response Capacity and Population Structure. In Figure 4(a), increasing the response dimension from 10 to 105 lowers covariance error from 0.7251 to 0.6938 while mean accuracy remains unchanged. Panels (b) and (c) illustrate response-strength distributions and population coverage in fixed cellular examples. These views complement the quantitative diagnostics; two-dimensional proximity does not substitute for high-dimensional distributional fit.

5 Conclusion

We introduced Moment-Conditioned Schrödinger Transport, which separates condition-level mean prediction from learning centered changes in single-cell populations. A shared conditional bridge operates in a response-aware representation, while expression-space recovery produces nonnegative cell populations consistent with the predicted mean. Our evaluation distinguishes end-to-end prediction accuracy from the additional contribution of population learning under matched means. The current comparisons indicate that the benefit of learned transport depends on the prediction setting, with larger gains over control-shape transfer in combinatorial generalization than in unseen-perturbation prediction. The framework does not identify individual-cell counterfactual trajectories, and its predictions remain limited by mean-estimation error, the chosen response space, and the effects of nonnegative recovery on population structure.

Limitation. Our framework depends on the accuracy of the condition-level mean predictor. Mean-consistent recovery enforces the prescribed nonnegative mean but cannot correct prediction errors, which may also affect population structure through conditioning and recovery. Using observed means during training and predicted means at inference may introduce a conditioning mismatch. Future work could investigate cross-fitted conditioning and the propagation of mean-prediction uncertainty.

REFERENCES

Abhinav K Adduri, Dhruv Gautam, Beatrice Bevilacqua, Mohsen Naghipourfar, Alishba Imran, Rohan Shah, Noam Teyssier, Rishi Verma, Christopher Carpenter, Basak Eraslan, et al. Predicting cellular responses to perturbation across diverse contexts with state. *Cell*, 2026.

Constantin Ahlmann-Eltze, Wolfgang Huber, and Simon Anders. Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines. *Nature Methods*, 22(8): 1657–1661, 2025.

Charlotte Bunne, Stefan G Stark, Gabriele Gut, Jacobo Sarabia Del Castillo, Mitch Levesque, Kjong-Van Lehmann, Lucas Pelkmans, Andreas Krause, and Gunnar Rätsch. Learning single-cell perturbation responses using neural optimal transport. *Nature methods*, 20(11):1759–1768, 2023.

Changxi Chi, Yufei Huang, Jun Xia, Jiangbin Zheng, Yunfan Liu, Zelin Zang, and Stan Z Li. Departures: Distributional transport for single-cell perturbation prediction with neural schrödinger bridges. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 20544–20552, 2026.

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.

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

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

Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, Xuegong Zhang, and Le Song. Large-scale foundation model on single-cell transcriptomics. *Nature methods*, 21(8):1481–1491, 2024.

Leon Hetzel, Simon Boehm, Niki Kilbertus, Stephan Günnemann, Fabian Theis, et al. Predicting cellular responses to novel drug perturbations at a single-cell resolution. *Advances in Neural Information Processing Systems*, 35:26711–26722, 2022.

Wenkang Jiang, Yuhang Liu, Yichao Cai, Erdun Gao, Jiayi Dong, Ehsan Abbasnejad, Lina Yao, and Javen Qinfeng Shi. What makes a representation good for single-cell perturbation prediction? *arXiv preprint arXiv:2605.19343*, 2026.

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

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, et al. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular systems biology*, 19(6):e11517, 2023.

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

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.

James D Pearce, Sara E Simmonds, Gita Mahmoudabadi, Lakshmi Krishnan, Giovanni Palla, Ana-Maria Istrate, Alexander Tarashansky, Benjamin Nelson, Omar Valenzuela, Donghui Li, et al. Transcriptformer: A generative cell atlas across 1.5 billion years of evolution. *Science*, 393 (6806):aec8514, 2026.

Joseph M Replogle, Reuben A Saunders, Angela N Pogson, Jeffrey A Hussmann, Alexander Lenail, Alina Guna, Lauren Mascibroda, Eric J Wagner, Karen Adelman, Gila Lithwick-Yanai, 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(6):927–935, 2024.

Yusuf H Roohani, Tony J Hua, Po-Yuan Tung, Lexi R Bounds, Feiqiao B Yu, Alexander Dobin, Noam Teyssier, Abhinav Adduri, Alden Woodrow, Brian S Plosky, et al. Virtual Cell Challenge: Toward a turing test for the virtual cell. *Cell*, 188(13):3370–3374, 2025.

Christina V Theodoris, Ling Xiao, Anant Chopra, Mark D Chaffin, Zeina R Al Sayed, Matthew C Hill, Helene Mantineo, Elizabeth M Brydon, Zexian Zeng, X Shirley Liu, et al. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, 2023.

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

Fan Yang, Wenchuan Wang, Fang Wang, Yuan Fang, Duyu Tang, Junzhou Huang, Hui Lu, and Jianhua Yao. scbert as a large-scale pretrained deep language model for cell type annotation of single-cell rna-seq data. *Nature Machine Intelligence*, 4(10):852–866, 2022.

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.

Hengshi Yu, Weizhou Qian, Yuxuan Song, and Joshua D Welch. Perturbnet predicts single-cell responses to unseen chemical and genetic perturbations. *Molecular systems biology*, 21(8):960–982, 2025.

A Usage of Large Language Models

In this work, Large Language Models (LLMs) are only used for polishing the paper writing. They are not involved in research ideation, experimental design, data analysis, or the formulation of conclusions. All substantive intellectual contributions are made by the authors. The authors have checked the content throughout the paper after polishing, and take full responsibility for the current form.

Reproducibility Statement

The model formulation, training objective, and inference procedure are described in Section 3.

B Implementation Details

B.1 Information Sources and Predictor Branches

The mean predictor and transport representation are prepared before bridge optimization. For each context, the response-chart fitting routine requests up to 2,048 available control cells and stores their mean. The mean of each training treated condition is computed from the training cells returned for that condition, whereas the residual basis uses at most 64 treated cells per condition. Consequently, the mean-estimation sample and the residual-basis sample need not have the same size. No validation or test treated populations enter this chart-fitting routine. Available controls from the target context do enter both the chart and the control-derived condition information.

The ordered branches in Eq. 3 are important for interpreting the experiments. On Norman's 125 held-out double perturbations, the current mean predictor is already the sum of observed single-perturbation effects added to the control mean. It is therefore identical to the additive mean reference, before and after applying a common nonnegative recovery target. On E2 and E3, the test-condition mean uses the average shift for the same intervention over observed source contexts. The

target-context treated observations available in E2 can affect population learning, but the mean prediction for a held-out target condition follows this source-effect branch rather than a separate target-adapted ridge branch. On E1, the same intervention has no observed training source, so the predictor uses the regression branch and its target-coordinate adjustment. Replogle's held-out-cell predictions reuse training-condition means; this is distinct from unseen-condition generalization.

The raw prediction $\widetilde{m}_{c,u}$ enters the mean-related network condition in Eq. 8; only the recovery target $\widehat{m}_{c,u} = [\widetilde{m}_{c,u}]_+$ is clipped to the nonnegative orthant. During training, observed conditions provide their training means to the conditioning input, and treated endpoints are centered by those training means. At test time, the mean input follows the applicable prediction branch. Training mean inputs are not cross-fitted. The conditioning chart remains fixed when the transport representation is changed in a matched geometry ablation.

B.2 Response-Chart Construction

For the default charts, the requested mean and heterogeneity ranks are 16 and 48. The heterogeneity matrix combines at most 64 centered training cells per treated condition with the first 256 available controls per context, centered by the stored control mean. The component along the mean basis is removed before the heterogeneity-basis routine is called. Coordinate vectors for all known target genes present in the expression panel are then appended. An SVD of the concatenated candidate basis retains singular vectors whose singular values exceed 10^{-5}. The resulting dimension can exceed $16 + 48$ because of the target coordinates.

For each final basis direction, the implementation computes the population-standard-deviation convention (divisor N) over the pooled, subsampled training treated residuals and floors the result at 0.05. Controls contribute to the heterogeneity basis, but not to this final scale estimate. The recorded default checkpoint dimensions are 140 for Norman, 117 for each of E1–E3, and 269 for Replogle. These are distinct from the dimension-matched Norman component experiments at $d = 35$. The independent evaluation PCA is not the learned response chart.

B.3 Endpoint Coupling, Learning, and Sampling

The empirical endpoint measures have uniform weights within each minibatch. The Sinkhorn solver uses squared Euclidean costs and final entropy coefficient $2\sigma^2$. It operates in the log domain, first warming the dual variables with coefficient multipliers $32, 16, 8, 4, 2, 1$, using 50 iterations per multiplier. At the final coefficient, it performs up to $N_{\text{iter}} = \max\{N_{\text{req}}, 4000\}$ additional iterations, where N_{req} is the requested iteration count. The maximum absolute marginal error is checked every 50 iterations, with an early-stopping tolerance of 10^{-5}. The solve raises an error if the final marginal error exceeds 10^{-4}; reaching the iteration limit is therefore not by itself a certificate of the stricter tolerance. The returned plan is normalized and endpoint pairs are sampled from it with replacement.

Training times lie in $[0.01, 0.99]$. The network's second output regresses negative interpolation noise, not the unscaled score. Using $s_t^\star = -\epsilon/(\sigma\sqrt{t(1-t)})$, its loss is algebraically equivalent to the weighted score loss:

$$\lambda_s \|e_\theta + \epsilon\|_2^2 = \lambda_s \sigma^2 t(1-t) \|s_\theta - s_t^\star\|_2^2. \tag{16}$$

The code averages over both sampled cells and response dimensions. It does not differentiate through the sampled endpoint coupling.

The audited default checkpoint arguments specify $\sigma = 1$, hidden width 256, minibatch size 128, a learning rate of 2×10^{-4}, and a budget of 8,000 optimization steps. Validation is requested every 2,000 steps, with a configured cap of 32 validation conditions and 128 evaluation cells. These are checkpoint training/validation settings, not a statement that every saved test prediction contains 128 cells. The final checkpoint is selected through the protocol's validation procedure rather than the test outcomes.

The inference update is explicitly

$$\begin{gathered} Z_{k+1} = Z_k + \Delta t \left[v_\theta(Z_k, t_k \mid h) + \frac{\sigma}{2\sqrt{t_k(1-t_k)}} e_\theta(Z_k, t_k \mid h) \right] + \sigma\sqrt{\Delta t}\, \xi_k, \\ \Delta t = 1/K, \quad t_k = (k + \tfrac{1}{2})\Delta t, \quad \xi_k \sim \mathcal{N}(0, I_d), \quad K = 50. \end{gathered} \tag{17}$$

The coefficients are evaluated at midpoint times, but the state update is an explicit Euler update, not a higher-order midpoint-state integrator. With $K = 50$, the queried times are $0.01, 0.03, \ldots, 0.99$. The probability-flow comparison removes both the second term inside the brackets and the random increment.

B.4 Recovery and Numerical Precision

For a single gene, write $v \in \mathbb{R}^n$ for its decoded values and $a \geq 0$ for the prescribed mean. Projection solves

$$\min_{y \geq 0,\ \mathbf{1}^\top y = na} \|y - v\|_2^2, \qquad y_i = \max\{v_i - \theta, 0\}. \tag{18}$$

The threshold is obtained by sorting the values, computing cumulative sums, and identifying the active coordinates. When $a = 0$, the only feasible output is the zero vector. The full matrix problem separates over genes. This is an observation-space support constraint; it is not an additional Schrödinger-bridge optimality statement.

A common gene-wise translation before projection does not change the result when the prescribed mean is fixed. Indeed, for any vector b and any feasible Y,

$$\|Y - (V + \mathbf{1}b^\top)\|_F^2 - \|Y - V\|_F^2 = -2nb^\top(\widehat{m} - \bar{v}) + n\|b\|_2^2, \tag{19}$$

which does not depend on Y. Thus, explicitly recentering the decoded population to the same prescribed mean before projection is mathematically equivalent, although the supplied decoder centers source and response samples before its projection call.

Exact mean consistency is a property of the mathematical projection. The supplied implementation returns single-precision arrays. The recorded generated-count sensitivity audit reports a maximum coordinate-wise mean discrepancy of approximately 6.34×10^{-6}; six control-shape rows with 192 generated cells exceed the audit's 5×10^{-6} threshold. These numerical exceptions are retained rather than reported as passing that tolerance. They do not imply a negative output or an infeasible mathematical projection, but preclude claiming machine-exact equality for every stored output. The projection can change the covariance and the component outside the response chart, so structural fidelity is evaluated after the same recovery operation for all matched-mean learners.