PERTURBBRIDGE: REFERENCE-PROCESS DESIGN FOR HETEROGENEITY-AWARE SINGLE-CELL PERTURBATION PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting single-cell perturbation responses requires transferring a known perturbation to new cellular contexts, where control and perturbed populations are measured separately and remain unpaired. Coupling-based prediction faces two connected challenges. First, standard transport costs can let heterogeneous response amplitudes dominate state correspondence. Second, endpoint optimal transport provides a static coupling but no process over intermediate states, so the coupling alone cannot ensure consistency with the paths required for training and generation. Changing the coupling cost while retaining a fixed bridge further separates the pairing geometry from the reference process used for training and generation. We therefore recast the problem as designing a reference process that jointly defines the coupling and supplies the intermediate states. We propose PerturbBridge, a heterogeneity-aware count reference whose two-parameter negative-binomial-difference family separates displacement scale from tail shape at fixed variance. Each member of the family induces a Schrödinger endpoint coupling and supplies its matching tractable integer bridge, so that coupling and generation remain reference-consistent as the geometry changes. Theoretically, we analyze two baseline states with different response amplitudes and establish conditions under which the reference preserves state correspondence regardless of response magnitude, and we connect this pairing preference to prediction error under composition shift. Empirically, PerturbBridge achieves strong overall rankings across 80 metrics spanning four expression views, ranking first among six learned models in all three PBMC settings. These results establish reference-process design as a principled approach to unpaired perturbation prediction under heterogeneous response amplitudes.

1 INTRODUCTION

Predicting cellular responses to genetic and chemical perturbations is essential for understanding gene regulation and drug action. Perturb-seq links pooled genetic screens to single-cell transcriptomes at increasing scale (Dixit et al., 2016; Replogle et al., 2022). Responses depend strongly on cellular context, and the same perturbation can induce markedly different transcriptional responses across cell types and states, making exhaustive measurement of perturbation–context combinations infeasible. This motivates predicting perturbation responses from observed control populations. Control and perturbed populations are measured separately, and their cell-level correspondence is unobserved. The task is unpaired conditional distribution matching: given control expression x, context e, and perturbation a, learn $Q(y \mid x, e, a)$ whose induced marginal matches the observed perturbed population. Coupling-based methods construct pseudo-pairs between these populations, assigning perturbed cells to control cells. This coupling directly shapes what the model learns.

Optimal transport is commonly used to select couplings, as in CellOT and CINEMA-OT (Bunne et al., 2023; Dong et al., 2023), with costs defined on molecular states or learned representations. A central challenge is response heterogeneity: the same perturbation can induce markedly different response amplitudes across baseline states. Large responses can dominate squared Euclidean costs L_2^2 and favor assignments to a different baseline state; L_1 limits this influence but leaves different allocations of the same total displacement indistinguishable. Designing a better endpoint cost addresses the pairing

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.

decision, but conditional generation also requires a compatible path construction. The endpoint cost specifies the coupling, while the reference bridge supplies intermediate states between paired cells and the sampling updates. Changing the coupling cost while retaining a fixed reference bridge makes the two constructions reference-inconsistent: pairing and generation follow different geometries.

We study this problem through a reference process: a stochastic process whose endpoint law defines the coupling and whose conditional paths supply intermediate states, ensuring consistency by construction. The design challenge is to combine bounded influence for large displacements with sensitivity to smaller state differences within a tractable reference process (Figure 1).

To resolve this tension, we propose PerturbBridge, a conditional count-generation framework whose reference separates the overall displacement scale from the shape of the displacement distribution. The reference is defined by a two-parameter negative-binomial-difference family, where κ controls displacement variance and h adjusts tail shape at fixed variance. This shape control allows the reference to limit the influence of large displacements while keeping smaller displacement patterns distinguishable. As $h \to 0$, the family recovers the finite-κ Poisson reference. For every parameter choice, the endpoint transition law induces a Schrödinger coupling and the conditional paths yield its matching tractable integer bridge. Thus, adjusting the reference geometry changes the coupling and bridge together while preserving their consistency. During training, the coupling supplies pseudo-pairs and the bridge supplies intermediate states for a perturbation- and context-conditioned generator. At inference, generation starts from an observed control cell and alternates endpoint predictions with reference-bridge updates to produce nonnegative expression counts.

We validate the design of PerturbBridge both theoretically and empirically. Theoretically, we analyze a simplified setting with two baseline states that respond with different amplitudes. We show that the cost difference between the state-consistent assignment and the swapped assignment admits a lower bound that depends only on the small state difference and is independent of the large response amplitude. When this bound is positive, state-consistent pairing is preferred regardless of that amplitude. A total-variation bound further links pairing errors to prediction error under changes in baseline-state proportions. Empirically, we evaluate four tasks: in-distribution prediction on NADIG genetic perturbations and PBMC cytokine treatments, and transfer to unseen cell types and donors on PBMC. PerturbBridge achieves the lowest raw-count Wasserstein-2 distance in all four tasks and the best mean rank among six learned models in all three PBMC settings. Controlled sweeps show that shape adjustment at fixed variance improves response recovery beyond the tested scale-only Poisson settings. Ablations show response-recovery gains over the tested alternative pairing costs and noise-source generation; with the selected coupling fixed, its matching bridge also improves recovery over the tested alternative bridge shapes.

Our central contribution is to recast unpaired perturbation prediction as a reference-process design problem. We construct a family of heterogeneity-aware reference processes whose endpoint couplings and integer bridges are matched throughout the family. We introduce PerturbBridge, whose two-parameter negative-binomial-difference reference separates displacement scale from tail shape, induces the Schrödinger coupling through its endpoint law, and yields tractable integer bridges through its conditional paths. We validate this design both theoretically and empirically, showing that it preserves state correspondence under heterogeneous response amplitudes and improves response recovery across four prediction tasks on NADIG and PBMC.

2 Related Work

Conditional perturbation prediction. scGen transfers responses across cell types through latent-space vector arithmetic (Lotfollahi et al., 2019), while CPA learns compositional response representations (Lotfollahi et al., 2023). GEARS uses gene-relation graphs to predict unseen genetic perturbations (Roohani et al., 2023). Here, transfer concerns new contexts for training-observed perturbations. CellFlow uses conditional flow matching and STATE learns transferable responses (Klein et al., 2025; Adduri et al., 2025). DOLORIS aligns populations through shared latent structure, while DEPARTURES combines minibatch OT with neural Schrödinger bridges for expression and activation (Chi et al., 2026b;a). We study the geometry of conditional supervision for count generation.

Cell correspondence and heterogeneous effects. Waddington-OT infers developmental couplings from unpaired temporal populations (Schiebinger et al., 2019). For perturbations, CellOT learns neu-

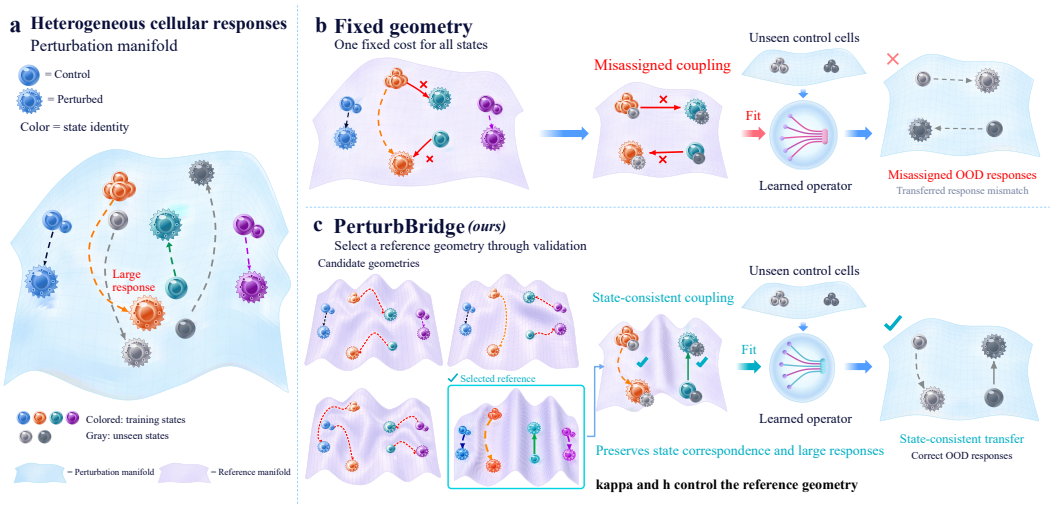

Figure 1: **Reference geometry shapes conditional supervision and OOD prediction.** (a) Baseline states exhibit heterogeneous response amplitudes, including a large response. (b) A fixed geometry can favor state-inconsistent pairs and propagate misassigned responses into the learned conditional operator. (c) PerturbBridge selects a global scale–shape reference through validation, supplying matched coupling and bridge supervision; the illustrated regime preserves state correspondence and the large response. Different couplings of the same training populations supervise different response operators, which receive the same controls from an unseen context but predict different state-dependent responses. Colors denote baseline states; cell morphologies indicate response phenotypes.

ral transport maps with state-dependent responses (Bunne et al., 2023), and CINEMA-OT separates treatment-associated and confounding components before matching cells to estimate causal effects (Dong et al., 2023). GENOT learns stochastic transport plans with general costs, and W_1OT studies efficient nonquadratic transport (Klein et al., 2024; Chen et al., 2025). Flow matching trains continuous flows along prescribed probability paths, while minibatch OT shapes their endpoint couplings (Lipman et al., 2022; Tong et al., 2023). PerturbBridge uses one count-reference geometry for coupling and bridge supervision.

Count generation and robust references. scVI uses a generative count observation model (Lopez et al., 2018); regularized NB regression supports count normalization and variance stabilization (Hafemeister & Satija, 2019). Continuous-time discrete denoising and discrete flow matching model dynamics on discrete state spaces (Campbell et al., 2022; Gat et al., 2024). Lingshu-Cell uses masked discrete transcriptome generation (Zhang et al., 2026). Count Bridges provides tractable integer bridges and Poisson-reference couplings; count-FM uses local count transitions and a Poisson-based coupling cost (Fishman et al., 2026; Wei & Pearson, 2026). We build on integer-process theory (Barndorff-Nielsen et al., 2012) to control reference shape. Adaptive robust losses address large residuals (Barron, 2019); our reference also defines a transition semigroup and bridge sampler. The static–dynamic connection follows Schrödinger theory (Léonard, 2013), developed computationally through diffusion Schrödinger bridges and bridge matching (De Bortoli et al., 2021; Shi et al., 2023).

3 Method

3.1 Prediction as conditional-relation selection

Let $X, Y \in \mathbb{N}_0^G$ denote control and perturbed readouts, a a perturbation, and e an experimental context. Within each training (e, a), observations are independent samples from $\mu_{e,0}$ and $\mu_{e,a}$. The OOD task supplies control readouts from a new context e_* and a training-observed perturbation a. Its treated readouts are reserved for evaluation. We learn a shared kernel $Q_\theta(dy \mid x, e, a)$, $\widehat{\mu}_{e,a}(dy) = \int Q_\theta(dy \mid x, e, a)\mu_{e,0}(dx)$. The corresponding coupling is $\mu_{e,0}(dx)Q_\theta(dy \mid x, e, a)$. The measured baseline x is an informative condition within a context, whose cell-type or donor label need not

resolve individual cellular states. Baseline state and shared readout conditions can be represented conceptually by a latent U, with $X = f_0(U, \varepsilon_0)$ and $Y = f_a(U, \varepsilon_a)$. We learn the observed-readout relation without requiring labels for U.

Coupling specifies conditional supervision. For a selected endpoint coupling γ, let $K_\gamma(dy \mid x, e, a)$ be its treated-given-control conditional law. Direct conditional likelihood training on its pseudo-pairs targets K_γ: on finite support, its loss is conditional entropy plus $\mathbb{E}_X \mathrm{KL}(K_\gamma(\cdot \mid X, e, a) \| Q(\cdot \mid X, e, a))$. Thus changing the coupling can change what the same baseline expression is taught to predict. Independent pairing within (e, a) makes this target independent of x; a systematic reassignment can instead preserve dependence while reversing the state–response association. PerturbBridge realizes the selected supervision through intermediate bridge states and endpoint denoising. Appendix F connects this likelihood view to bridge training and sampling.

Why preserve count scale when selecting supervision? If a coupling cost only uses a representation ϕ, two sources with $\phi(x) = \phi(x')$ receive the same conditional row of the positive entropic plan. Library-size normalization has precisely this invariance for $x' = cx$, $c > 0$. Appendix A.2 proves the statement. Raw-count coupling retains this distinction before the model decides how much it matters.

We orient auxiliary paths as $X_0 = Y$ and $X_1 = X$, so generation starts at the observed control and proceeds toward time zero. Observed endpoints are nonnegative; auxiliary states lie on $\mathbb{Z}^G$. This separates the support of an experimental readout from the support of the computational reference.

3.2 A heterogeneity-aware count reference

The reference must accommodate different response amplitudes across baseline states while retaining the smaller differences that distinguish those states. We encode this requirement in a shared displacement law with separate scale and shape controls. Its endpoint law will define the coupling cost, and its process extension will provide conditional bridge paths. We construct the displacement law using exchangeable reference counts with random positive intensities: $\Lambda^\pm \overset{\text{iid}}{\sim} \mathrm{Gamma}\big(h^{-1}, (hm)^{-1}\big)$, $U^\pm \mid \Lambda^\pm \sim \mathrm{Pois}(\Lambda^\pm)$, $D = U^+ - U^-$. The Gamma distribution uses shape and rate. Its squared coefficient of variation is h. These auxiliary counts define a symmetric discrepancy prior; treatment direction is supplied by the endpoint populations. Set $\kappa = m + hm^2$, $m = \frac{2\kappa}{1+\sqrt{1+4\kappa h}}$, $q = \frac{\sqrt{1+4\kappa h}-1}{\sqrt{1+4\kappa h}+1}$. For $h > 0$, $U^\pm$ are independent negative-binomial variables with shape $r = 1/h$ and probability mass $b_{r,q}(n) = \frac{\Gamma(n+r)}{\Gamma(r)\, n!}(1-q)^r q^n$, $n \in \mathbb{N}_0$. The induced kernel and normalized cost are $K_{\kappa,h}(d) = \sum_{n=0}^{\infty} b_{1/h,q}(n) b_{1/h,q}(n + |d|)$ and

$$\rho_{\kappa,h}(d) = -\log \frac{K_{\kappa,h}(d)}{K_{\kappa,h}(0)}, \qquad c_{\kappa,h}(y, x) = \sum_{g=1}^{G} \rho_{\kappa,h}(x_g - y_g). \tag{1}$$

Validation selects global reference parameters (κ, h) that score candidate displacements. Pair probabilities also depend on observed cell states and endpoint populations.

Proposition 3.1 (Independent scale and shape)**.** *The kernel $K_{\kappa,h}$ has characteristic function $\widehat{K}_{\kappa,h}(\xi) = \big[1 + 2\kappa h(1 - \cos\xi)\big]^{-1/h}$, variance 2κ, and excess kurtosis $(2\kappa)^{-1} + 3h$. As $h \downarrow 0$, $K_{\kappa,h}(d) \to e^{-2\kappa} I_{|d|}(2\kappa)$, recovering the finite-$\kappa$ Poisson reference. At $h = 1$, $\rho_{\kappa,1}$ is a scaled absolute-value cost. For every $h > 0$, the reference law is outside the one-parameter symmetric Skellam family.*

The proof and exact kernel formulas are in Appendix B. Thus κ sets the reference displacement variance, while h changes how probability is distributed across displacement amplitudes at that variance. The next result connects reference shape to correspondence under heterogeneous responses.

3.3 State correspondence under state-dependent response amplitudes

To isolate how conditional supervision affects prediction, we analyze two baseline states with different response amplitudes. The composition-shift setting below holds their state-specific responses

fixed and changes their population weights. Consider the reference slice $h = 1/2$. Writing $\eta = \operatorname{arcosh}(1 + 1/\kappa)$,

$$\rho_{\kappa,1/2}(d) = \eta|d| - \log(1 + |d| \tanh \eta)\,. \tag{2}$$

Its extension to nonnegative real arguments is strictly convex and has derivative below η. Thus, it distinguishes ordinary displacement allocations while limiting the change in cost caused by one additional count, even at a large displacement.

Theorem 3.2 (Correspondence stability and composition transfer)**.** *Consider source states $x_1 = (b, b)$, $x_2 = (b + m, b + \ell)$ and their state-consistent targets $y_1 = (b, b + M)$, $y_2 = (b + m, b)$, with $b \in \mathbb{N}_0$, positive integers m, ℓ, and integer $M \geq \ell$. For a separable cost ρ, let $\Delta_\rho = 2\rho(m) + \rho(M - \ell) - \rho(M) - \rho(\ell)$ be the swapped-minus-consistent assignment cost. Squared Euclidean cost gives $\Delta = 2m^2 - 2M\ell$ and prefers the swapped assignment whenever $M > m^2/\ell$. For Eq. (2), $\Delta_\rho \geq \delta_{\text{state}} := 2\rho(m) - \rho(\ell) - \eta\ell$. If $\delta_{\text{state}} > 0$, the state-consistent assignment is preferred for every $M \geq \ell$. With balanced training marginals and objective $\mathbb{E}c - \tau H$, the conditional probability of consistent pairing is $p = \sigma(\Delta_\rho/(2\tau))$, where $\sigma(u) = (1 + e^{-u})^{-1}$. Transferring this conditional kernel to state weights $(\omega, 1 - \omega)$ gives target-marginal total-variation error $\mathrm{TV}(\widehat{\mu}_{\omega,a}, \mu_{\omega,a}) = |2\omega - 1|(1 - p)$.*

For $\kappa = 4$, $m = 3$, and $\ell = 1$, $\delta_{\text{state}} > 1.18$. This is a concrete regime in which increased response amplitude does not overturn state correspondence. The bound controls *pairing influence*; the large response remains part of the target marginal. Appendix C also gives a translation example where L_1 is indifferent but Eq. (2) prefers the consistent assignment. We next construct the matching integer process and bridge.

3.4 One reference for endpoint coupling and integer paths

For $h > 0$, let N_t^+ and N_t^- be independent negative-binomial Lévy processes whose increments over an operational-time interval of length Δ follow $b_{\Delta/h,q}$. Define $X_t = X_0 + N_t^+ - N_t^-$, independently across genes. The transition characteristic function is

$$\widehat{K}_\Delta^{\kappa,h}(\xi) = \left[1 + 2\kappa h(1 - \cos\xi)\right]^{-\Delta/h}, \qquad K_{\Delta_1+\Delta_2} = K_{\Delta_1} * K_{\Delta_2}. \tag{3}$$

A deterministic schedule $w(0) = 0, w(1) = 1$ replaces Δ by $w(v) - w(u)$; we use $w(t) = (1 - \cos \pi t)/2$. This process uses integer jumps, not discretized continuous trajectories. Its unit-time displacement law is exactly the kernel defined in Section 3.2.

Let $R^{\kappa,h}$ start from $\mu_{e,a}$. The dynamic Schrödinger bridge and its endpoint reduction are $P_{e,a}^* = \arg\min_{P:\, P_0=\mu_{e,a},\, P_1=\mu_{e,0}} \mathrm{KL}(P\|R^{\kappa,h})$ and

$$\gamma_{e,a}^{\kappa,h} = \operatorname*{arg\,min}_{\gamma\in\Gamma(\mu_{e,a},\mu_{e,0})} \left\{\mathbb{E}_\gamma c_{\kappa,h}(Y, X) - H(\gamma)\right\}. \tag{4}$$

The path is $P^*(d\omega) = \gamma^{\kappa,h}(dy, dx)R^{\kappa,h}(d\omega \mid y, x)$; Appendix D gives the KL decomposition. We use the natural entropy coefficient one, so pairing and bridge conditionals refer to the same process.

For finite minibatches, log-domain Sinkhorn solves Eq. (4) with uniform cell weights. The plan has the form $\Pi_{ji} = u_j K_1^{\kappa,h}(x_i - y_j)v_i$, where the multivariate kernel is a product over genes. Thus, $\frac{\Pi_{ji}\Pi_{j'i'}}{\Pi_{ji'}\Pi_{j'i}} = \frac{K_1(x_i-y_j)K_1(x_{i'}-y_{j'})}{K_1(x_{i'}-y_j)K_1(x_i-y_{j'})}$. The two reference parameters restrict pairing odds. Validation-context prediction selects (κ, h) before testing.

3.5 Tractable bridges and conditional generation

The reference bridge over $[u, v]$ satisfies

$$R(X_t = z \mid X_u = a, X_v = b) = \frac{K_{w(t)-w(u)}(z - a)K_{w(v)-w(t)}(b - z)}{K_{w(v)-w(u)}(b - a)}. \tag{5}$$

It can be sampled without enumerating intermediate vectors. For one coordinate, write $d = b - a$, $r = [w(v) - w(u)]/h$, $d_+ = \max(d, 0)$, and $d_- = \max(-d, 0)$. Draw $\Pr(M = m \mid d) \propto b_{r,q}(m + d_+)b_{r,q}(m + d_-)$, set $n_+ = M + d_+$, $n_- = M + d_-$, and sample independent $A_\pm \sim$

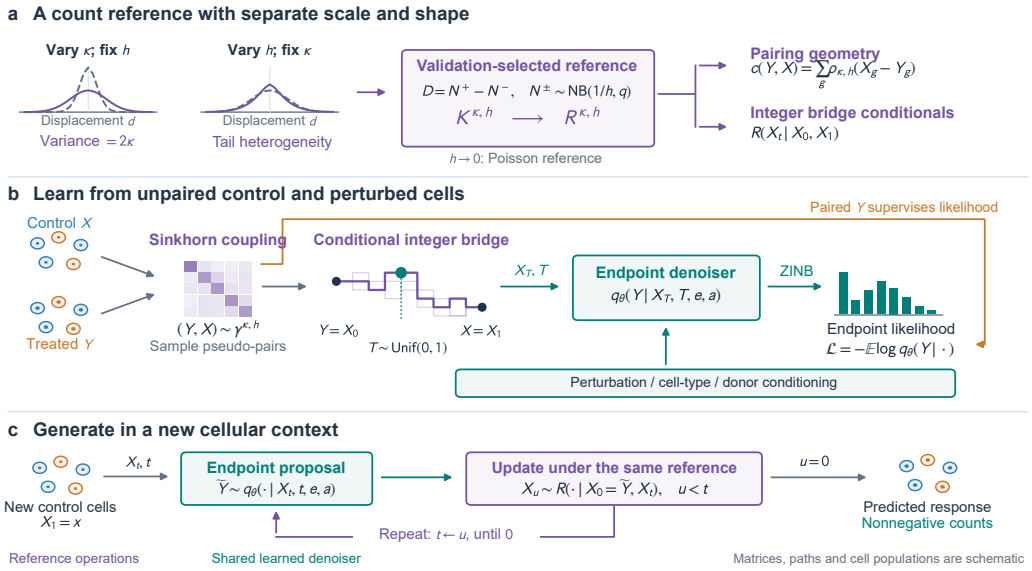

Figure 2: **PerturbBridge pairs and generates with one count reference.** (a) Scale–shape reference family. (b) Reference-induced coupling and integer bridges train the endpoint denoiser. (c) Generation alternates endpoint proposals and matching bridge updates from control cells.

$\text{BetaBinomial}\left(n_{\pm}; \frac{w(t)-w(u)}{h}, \frac{w(v)-w(t)}{h}\right)$. Then $X_t = a + A_+ - A_-$. Appendix E proves this construction and gives stable recurrences. At $h = 0$, negative-binomial slack and beta-binomial allocation reduce to Bessel slack and binomial allocation.

For training, sample $(Y, X) \sim \gamma_{e,a}^{\kappa,h}$, $T \sim \text{Unif}(0, 1)$, and $X_T \sim R(\cdot \mid Y, X)$. A metadata-conditioned denoiser predicts the target endpoint with a gene-factorized ZINB distribution: $q_\theta(y \mid X_T, T, e, a) = \prod_g \text{ZINB}(y_g; \mu_{\theta g}, r_{\theta g}, \pi_{\theta g})$. The training objective is

$$\mathcal{L}(\theta) = \mathbb{E}[-\log q_\theta(Y \mid X_T, T, e, a)]. \tag{6}$$

The network uses a set-based metadata condition encoder and a cross-attention MLP. Control expression conditions the full response kernel through $X_1 = x$ and the subsequent bridge states; the denoiser receives the current state, time, and context tokens. In PBMC OOD tasks, these tokens also include a fixed PCA representation of control pseudobulks (Appendix G.2). The endpoint likelihood enforces nonnegative count outputs. The unrestricted likelihood optimum is the conditional endpoint law under the selected path; the ZINB model learns its likelihood projection.

At inference, initialize $X_1 = x$ and descend a time grid. At each $t > u$, draw $\widetilde{Y} \sim q_\theta(\cdot \mid X_t, t, e, a)$ and then $X_u \sim R(\cdot \mid X_0 = \widetilde{Y}, X_t)$. At time zero this returns the proposed nonnegative endpoint. For an exact reference-consistent coupling and endpoint posterior, these updates are the reverse transitions of the Schrödinger bridge. Appendix F gives the exact result and neural/minibatch approximations.

4 Experiments

4.1 Evaluation setup

We evaluate genetic perturbations on NADIG (Nadig et al., 2025) and cytokine treatments on PBMC (Oesinghaus et al., 2025). Both use cell-level random splits for in-distribution (ID) prediction. PBMC additionally tests transfer to eight unseen cell types and two unseen donors, with controls from each evaluation context supplied as inputs. Cell-type validation uses the ten training types. The four tasks contain 4,141, 15,025, 8,545, and 2,899 evaluated condition groups, respectively. Appendix G details splits and sampling.

Comparators and evaluation. We compare with CellFlow (Klein et al., 2025), STATE (Adduri et al., 2025), DOLORIS (Chi et al., 2026b), DEPARTURES (Chi et al., 2026a), and Lingshu-Cell

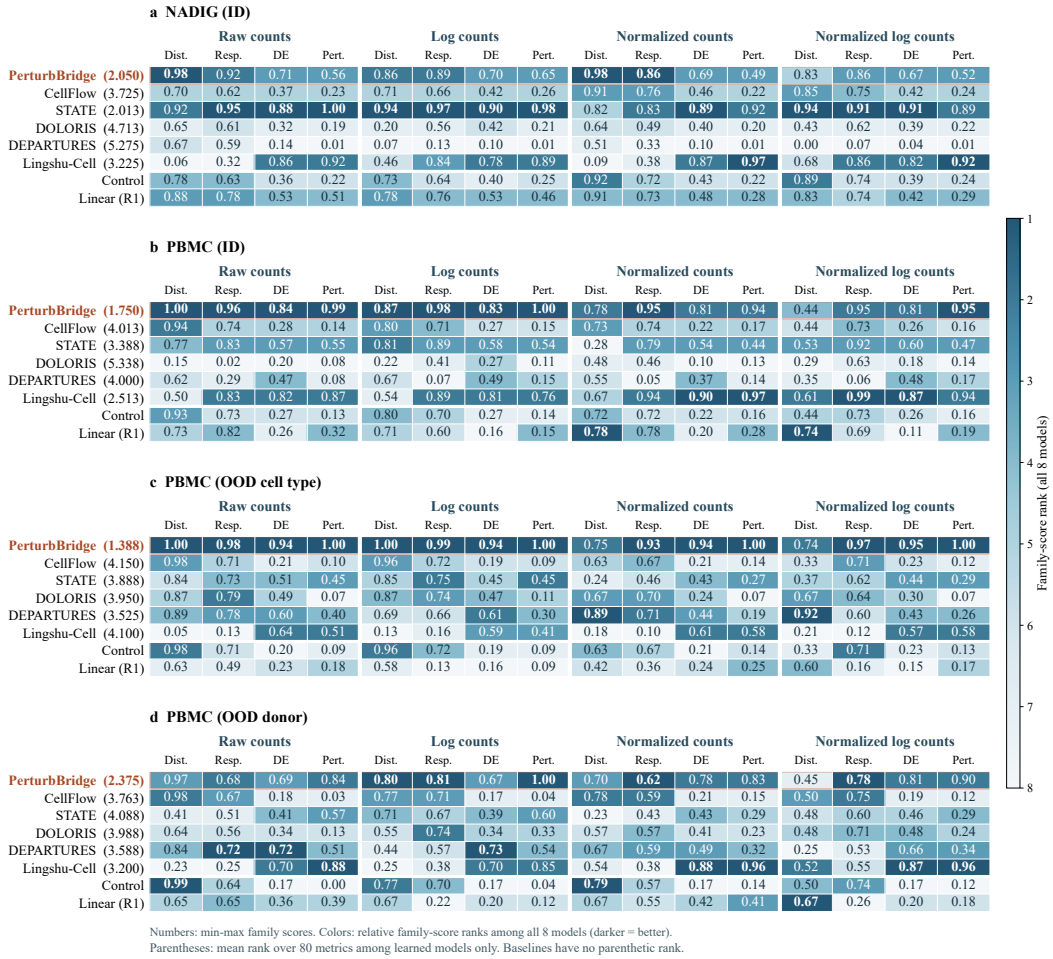

Figure 3: **Multi-view performance across four prediction tasks.** Columns group distribution, response, DE, and perturbation-level metrics in four expression views. Values average direction-aligned min–max scores within each family, normalized across all eight methods; darker blue indicates better family rank and bold marks the best value. Parentheses report mean rank over 80 metrics among the six learned models. Definitions and raw results appear in Appendices L and M.

(Zhang et al., 2026), together with Control and Linear (R1). Validation Sinkhorn scores select checkpoints and (κ, h). All methods use training-observed perturbations without external priors. The six learned models share control expression and perturbation, cell-type, and donor information, including fixed control-derived PCA embeddings for held-out identities (Appendix G.2). A common evaluator scores four expression views: raw counts, log counts, normalized counts, and normalized log counts. W_2 and squared MMD assess distribution fit; response Pearson, perturbation discrimination score (PDS), DE average precision (AP), and DE-breadth ranking score (ES) assess response recovery. Metric definitions and aggregation appear in Appendix H.

4.2 Main results

Response recovery across contexts. PerturbBridge achieves the lowest raw-count W_2 in all four tasks and the best six-model mean rank in all three PBMC settings (Figure 3). Its response gains concentrate on unseen cell types: leading Pearson, cosine PDS, and ES across all views indicates recovery of response patterns, perturbation identity, and relative DE breadth. Persistence after normalization shows that these gains extend to relative expression structure. STATE ranks slightly ahead on NADIG; mixed donor response rankings distinguish distribution fit from response recovery.

Table 1: Stronger relative performance on shape-changing groups. PerturbBridge across 64 shared metrics and eight methods. Subset scores use equal-case means; full-test scores retain benchmark weights. Protocol and metric selection: Appendix I.

Evaluation set	Pairwise wins (%)	First ranks (%)	Mean rank
Full OOD cell type	90.62	64.06	1.656
20 shape-changing cases	97.32	84.38	1.188

Reference-shape, pairing-cost, and source ablations

Raw-count space | 1,080 condition groups | Sinkhorn $\tau = 1$ unless shown

	Distribution			Response errors			Response agreement			Differential expression			Recovery	
	Exact W_2 ↓	Marginal W_1 ↓	MMD^2 ↓	Sq. error sum ↓	MAE / gene ↓	MSE / gene ↓	Pearson ↑	Spearman ↑	Direction agreement ↑	AUROC ↑	DE overlap ↑	DE precision ↑	Mean PDS ↑	ES ↑
Reference $h = 0.5$	89.16	**0.0331**	**0.0094**	**364.58**	**0.0347**	**0.0200**	**0.3949**	**0.5525**	**0.9370**	**0.6616**	**0.2184**	**0.2118**	**0.8038**	**0.6231**
Reference $h = 1$	89.24	0.0366	0.0213	583.22	0.0394	0.0320	0.2199	0.3929	0.8288	0.6568	0.1978	0.1850	0.6247	0.5350
L_1	88.71	0.0372	0.0216	590.73	0.0399	0.0324	0.2036	0.3718	0.8169	0.6560	0.1952	0.1860	0.6044	0.5341
L_2^2	**88.63**	0.0370	0.0214	574.44	0.0396	0.0315	0.2149	0.3891	0.8269	0.6566	0.1974	0.1826	0.6137	0.5511
L_1 ($\tau = 10$)	89.01	0.0370	0.0217	592.35	0.0396	0.0325	0.2093	0.3842	0.8263	0.6551	0.1947	0.1858	0.6097	0.5513
L_2^2 ($\tau = 10$)	91.37	0.0382	0.0209	603.83	0.0411	0.0332	0.2248	0.3653	0.8160	0.6481	0.1913	0.1744	0.6498	0.4903
Huber $\delta = 1$	89.11	0.0368	0.0213	583.89	0.0394	0.0321	0.2199	0.4003	0.8361	0.6553	0.1966	0.1837	0.6212	0.5427
Random pairing	96.04	0.0369	0.0182	700.77	0.0388	0.0385	0.2403	0.3823	0.8488	0.6414	0.1846	0.1798	0.6795	0.4391
Euclidean	89.05	0.0355	0.0179	503.79	0.0379	0.0277	0.2360	0.4028	0.8458	0.6528	0.1957	0.1878	0.6326	0.5726
Noise source	97.01	0.0690	0.0804	1673.70	0.0754	0.0919	0.0843	0.2860	0.7164	0.6942	0.2001	0.1213	0.5153	0.0210

Raw values; column min-max excludes Noise source (white).

Worst — Best

Figure 4: **Reference-shape, pairing-cost, and source ablations in count space.** Reference rows match coupling and bridge at $\kappa = 8$; cost variants use the $(8, 0.5)$ bridge. Sinkhorn uses $\tau = 1$ unless marked otherwise. Values are raw scores. Colors normalize each column across the nine non-noise variants; darker is better and bold marks the best. Noise source stays white and is excluded from normalization. Mean PDS averages cosine-, L_1-, and L_2-based PDS. Four-view results appear in Appendix K.

Responses beyond a population-wide shift. In a shared, training-fitted PCA space, we center treated and PBS populations separately and compare their sliced Wasserstein distance with PBS resampling noise. A shape/noise ratio above two selects 35 OODct and three OOD-donor groups; requiring at least 500 treated cells leaves 20 OODct cases (Appendix I). PerturbBridge's advantage concentrates on these cases: it leads 54 of 64 shared metrics, versus 41 on the full OODct test (Table 1). Both MMD and response Pearson lead in every expression view, linking distribution fit to response recovery beyond total-count scale. More extreme shape changes in OODct, together with stronger relative performance on these cases, help explain its larger advantage on this task. Responses that reshape populations make bounded pairing influence and state discrimination especially relevant.

4.3 Ablation study

We test reference shape, pairing cost, and control-source generation on a PBMC subset with six unseen cell types, two donors, and 1,080 cytokine–context groups (Figure 4). Appendix K gives costs and four-view results.

Observed controls anchor response prediction. Replacing controls with independent gene-wise Poisson draws retains metadata conditioning but reduces cosine PDS toward random-ranking levels and nearly removes DE-breadth ordering. Yet raw-count DE AP remains comparable to the full model. Thus, within-condition DE ranking can survive while treatment-specific patterns and cross-treatment breadth are lost. Observed controls supply structure absent from metadata and training-average Poisson sources.

Pairing geometry determines useful supervision. With the $(8, 0.5)$ bridge fixed, squared-Euclidean pairing gives slightly lower W_2 than the reference coupling, but worse MMD, response

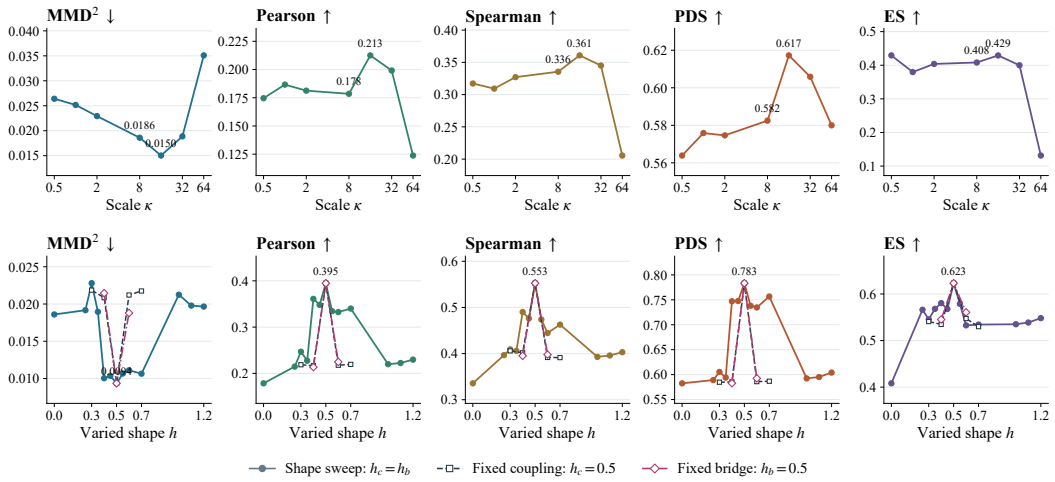

Figure 5: **Scale, coupling, and bridge sensitivity in count space.** Top: vary κ at $h = 0$ on a logarithmic scale. Bottom: at $\kappa = 8$, vary matched shapes $h = h_c = h_b$ (solid), bridge shape $h = h_b$ with $h_c = 0.5$ (dashed), or coupling shape $h = h_c \in \{0.4, 0.5, 0.6\}$ with $h_b = 0.5$ (pink dash-dotted). All three curves share the $(0.5, 0.5)$ run. Columns show MMD^2, response Pearson, response Spearman on the real DE set, cosine PDS, and ES.

correlations, and perturbation recovery. Low Euclidean distribution error therefore need not identify the most useful conditional supervision. Random pairing even exceeds the tested alternative costs in raw-count Pearson and cosine PDS, while remaining below the reference. These results make pairing geometry consequential for response learning. Cost substitutions also break reference matching, examined below.

Shape adjustment contributes beyond scale. Across the 25 settings in Figure 5, the best scale-only Poisson reference reaches cosine PDS 0.617, compared with 0.783 for $(\kappa, h) = (8, 0.5)$. At fixed κ, shape changes preserve displacement variance, so the gain cannot be explained by increasing that variance. Several matched shapes between 0.4 and 0.7 improve response recovery, whereas the linear-cost reference at $h = 1$ loses much of this benefit. This pattern is consistent with balancing bounded influence on large displacements and discrimination among smaller ones, rather than favoring progressively heavier tails (Appendix J).

Coupling and bridge act jointly. Changing either component while fixing the other at $h = 0.5$ reduces response recovery. The alternative bridges remain effective when paired with their own couplings: at $h_b = 0.7$, matching raises cosine PDS from 0.587 to 0.757. All four tested alternative bridge shapes similarly gain Pearson and PDS from matching. Thus, a useful endpoint relation and the intermediate states used to learn it must be considered together. These comparisons support joint reference design beyond one preferred bridge shape (Appendix J.1).

5 Conclusion

We introduced PerturbBridge, a count reference that jointly defines heterogeneity-aware cell coupling and matching integer bridges for unpaired perturbation prediction. Its scale–shape family separates displacement variance from tail shape, with theory connecting correspondence stability to prediction error under composition shift. Across genetic and cytokine tasks, PerturbBridge achieves the lowest raw-count W_2 in all four tasks and the best mean rank among six learned models in all three PBMC settings. Response gains are strongest on unseen cell types and persist across expression views, including normalization. Ablations show that observed controls supply response-relevant structure, shape adjustment improves recovery beyond scale tuning, and matching the coupling to its bridge improves response prediction. Together, these findings support joint reference-process design for transferring perturbation responses across cellular contexts.

AI USE STATEMENT

We used OpenAI Codex to assist with language editing, manuscript review, interpretation of supplied experimental results, preparation of plotting and table-generation scripts, and LaTeX formatting and compilation. The authors remain responsible for all scientific claims, reported results, citations, and the final manuscript.

REFERENCES

Abhinav K. Adduri, Dhruv Gautam, Beatrice Bevilacqua, Alishba Imran, Rohan Shah, Mohsen Naghipourfar, Noam Teyssier, Rajesh Ilango, Sanjay Nagaraj, Mingze Dong, Chiara Ricci-Tam, Christopher Carpenter, Vishvak Subramanyam, Aidan Winters, Sravya Tirukkovular, Jeremy Sullivan, Brian S. Plosky, Basak Eraslan, Nicholas D. Youngblut, Jure Leskovec, Luke A. Gilbert, Silvana Konermann, Patrick D. Hsu, Alexander Dobin, Dave P. Burke, Hani Goodarzi, and Yusuf H. Roohani. Predicting cellular responses to perturbation across diverse contexts with state. *bioRxiv*, 2025. doi: 10.1101/2025.06.26.661135. URL `https://www.biorxiv.org/content/early/2025/06/27/2025.06.26.661135`.

Ole E Barndorff-Nielsen, David G Pollard, and Neil Shephard. Integer-valued lévy processes and low latency financial econometrics. *Quantitative Finance*, 12(4):587–605, 2012.

Jonathan T Barron. A general and adaptive robust loss function. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4326–4334. IEEE, 2019.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

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.

Andrew Campbell, Joe Benton, Valentin De Bortoli, Thomas Rainforth, George Deligiannidis, and Arnaud Doucet. A continuous time framework for discrete denoising models. *Advances in Neural Information Processing Systems*, 35:28266–28279, 2022.

Yanshuo Chen, Zhengmian Hu, Wei Chen, and Heng Huang. Fast and scalable wasserstein-1 neural optimal transport solver for single-cell perturbation prediction. *Bioinformatics*, 41(Supplement_1): i513–i522, 2025.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(25):20544–20552, 2026a.

Changxi Chi, Jun Xia, Yufei Huang, Zhuoli Ouyang, Cheng Tan, Yunfan Liu, Jingbo Zhou, Chang Yu, Liangyu Yuan, Siyuan Li, et al. Doloris: Dual conditional diffusion implicit bridges with sparsity masking strategy for unpaired single-cell perturbation estimation. In *International Conference on Learning Representations*, volume 2026, pp. 52808–52823, 2026b.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. *Advances in neural information processing systems*, 34:17695–17709, 2021.

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

Mingze Dong, Bao Wang, Jessica Wei, Antonio H de O. Fonseca, Curtis J Perry, Alexander Frey, Feriel Ouerghi, Ellen F Foxman, Jeffrey J Ishizuka, Rahul M Dhodapkar, et al. Causal identification of single-cell experimental perturbation effects with cinema-ot. *Nature methods*, 20(11): 1769–1779, 2023.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd international conference on artificial intelligence and statistics*, pp. 2681–2690. PMLR, 2019.

Nic Fishman, Gokul Gowri, Tanush Kumar, Jiaqi Lu, Valentin De Bortoli, Jonathan Gootenberg, and Omar Abudayyeh. Count bridges enable modeling and deconvolving transcriptomic data. In *International Conference on Learning Representations*, volume 2026, pp. 129437–129477, 2026.

Itai Gat, Tal Remez, Neta Shaul, Felix Kreuk, Ricky TQ Chen, Gabriel Synnaeve, Yossi Adi, and Yaron Lipman. Discrete flow matching. *Advances in Neural Information Processing Systems*, 37: 133345–133385, 2024.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The journal of machine learning research*, 13:723–773, 2012.

Christoph Hafemeister and Rahul Satija. Normalization and variance stabilization of single-cell rna-seq data using regularized negative binomial regression. *Genome biology*, 20(1):296, 2019.

Dominik Klein, Théo Uscidda, Fabian Theis, and Marco Cuturi. Genot: Entropic (gromov) wasserstein flow matching with applications to single-cell genomics. *Advances in Neural Information Processing Systems*, 37:103897–103944, 2024.

Dominik Klein, Jonas Simon Fleck, Daniil Bobrovskiy, Lea Zimmermann, Sören Becker, Alessandro Palma, Leander Dony, Alejandro Tejada-Lapuerta, Guillaume Huguet, Hsiu-Chuan Lin, Nadezhda Azbukina, Fátima Sanchís-Calleja, Theo Uscidda, Artur Szalata, Manuel Gander, Aviv Regev, Barbara Treutlein, J. Gray Camp, and Fabian J. Theis. Cellflow enables generative single-cell phenotype modeling with flow matching. *bioRxiv*, 2025. doi: 10.1101/2025.04.11.648220. URL `https://www.biorxiv.org/content/early/2025/04/17/2025.04.11.648220`.

Christian Léonard. A survey of the Schrödinger problem and some of its connections with optimal transport. *arXiv preprint arXiv:1308.0215*, 2013.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058, 2018.

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*, pp. 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.

Lukas Oesinghaus, Sören Becker, Larsen Vornholz, Efthymia Papalexi, Joey Pangallo, Amir Ali Moinfar, Jenni Liu, Alyssa La Fleur, Maiia Shulman, Simone Marrujo, et al. A single-cell cytokine dictionary of human peripheral blood. *bioRxiv*, 2025.

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 multi-gene perturbations with gears. *Nature Biotechnology*, 2023.

Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berube, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

Yuyang Shi, Valentin De Bortoli, Andrew Campbell, and Arnaud Doucet. Diffusion schrödinger bridge matching. *Advances in neural information processing systems*, 36:62183–62223, 2023.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *arXiv preprint arXiv:2302.00482*, 2023.

Ganchao Wei and John Pearson. Flow matching for count data. *arXiv preprint arXiv:2605.07746*, 2026.

B. L. Welch. The generalization of 'student's' problem when several different population variances are involved. *Biometrika*, 34(1-2):28–35, 1947.

Han Zhang, Guo-Hua Yuan, Chaohao Yuan, Tingyang Xu, Tian Bian, Hong Cheng, Wenbing Huang, Deli Zhao, and Yu Rong. Lingshu-cell: A generative cellular world model for transcriptome modeling toward virtual cells. *arXiv preprint arXiv:2603.25240*, 2026.

A Predictive target and information retained by coupling

A.1 Notation and the role of context

An experimental context e groups observed metadata, whereas a baseline state U may vary within that group. The readout kernel $Q_\theta(dy \mid x, e, a)$ conditions on the measured baseline, not an observed label for every biological or technical factor. A state-composition illustration uses

$$\mu_{e,0} = \sum_{z=1}^{J} \alpha_e(z) P_z^0, \qquad \mu_{e,a} = \sum_{z=1}^{J} \alpha_e(z) P_z^a. \tag{7}$$

Here z indexes baseline origin; P_z^a need not retain a post-treatment annotation equal to z. Equation (7) describes composition shift with fixed within-origin response laws. The main predictor does not require this finite-mixture model; it is the setting used to isolate correspondence transfer in Theorem 3.2.

On finite latent supports, a shared response matrix Q maps a row of source probabilities to target probabilities. Collecting training source rows in A gives $V = AQ$. Candidate kernels are indistinguishable on these contexts when $A(Q_1 - Q_2) = 0$. New source compositions test additional directions. This explains the use of held-out-context prediction as an operational test of conditional transfer without making the training marginals the sole selection criterion.

Symbol	Meaning
X, Y	Nonnegative control and perturbed readouts
X_t	Auxiliary integer bridge state; $X_0 = Y$, $X_1 = X$
e, a	Experimental context and perturbation
κ, h	Reference scale and amplitude-heterogeneity parameter
q, r	NB probability parameter and shape; $r = \Delta/h$
$w(t)$	Cumulative operational time
$K_\Delta^{\kappa,h}$	Increment probability mass over operational duration Δ
$\gamma_{e,a}^{\kappa,h}$	Reference-selected endpoint coupling
q_θ, Q_θ	Learned endpoint posterior and induced final response kernel

Table 2: Notation. The endpoint dispersion $r_{\theta g}$ is distinct from the reference NB shape r.

A.2 Representation collisions induce identical conditional supervision

Proposition A.1 (Information preservation at the pairing stage)**.** *Let Π^* minimize $\langle \Pi, C\rangle - \tau H(\Pi)$ over finite supports with strictly positive marginals a, b, where $\tau > 0$ and $C_{ij} = c(\phi(x_i), \phi(y_j))$ is finite. If $\phi(x_i) = \phi(x_{i'})$, then $\Pi^*_{i,:}/a_i = \Pi^*_{i',:}/a_{i'}$.*

Proof. Positive-kernel Sinkhorn scaling (Cuturi, 2013) gives $\Pi^*_{ij} = u_i \exp(-C_{ij}/\tau) v_j$. Enforcing the row constraint yields

$$\frac{\Pi^*_{ij}}{a_i} = \frac{\exp(-C_{ij}/\tau) v_j}{\sum_k \exp(-C_{ik}/\tau) v_k}. \tag{8}$$

Equal represented inputs give equal cost rows, so their normalized conditional rows coincide. □

For $\phi(x) = \log(1 + cx/\|x\|_1)$ and positive library size, $\phi(sx) = \phi(x)$ for $s > 0$. For example, sources $(1, 1), (10, 10)$ collapse to the same representation, as do targets $(1, 3), (10, 30)$. Every pair has the same represented cost and the entropic solution is independent. Raw counts retain the distinction.

B Derivation and properties of the scale–shape family

B.1 Random-effects construction and characteristic function

Let $h > 0$. The Gamma–Poisson mixture in Section 3.2 has probability generating function

$$\mathbb{E} z^{U^\pm} = [1 + hm(1 - z)]^{-1/h}. \tag{9}$$

Since U^+ and U^- are independent, the characteristic function of their difference is

$$\mathbb{E}e^{i\xi D} = \left[(1 + hm - hme^{i\xi})(1 + hm - hme^{-i\xi})\right]^{-1/h} \tag{10}$$

$$= [1 + 2hm(1 + hm)(1 - \cos\xi)]^{-1/h}. \tag{11}$$

Using $m + hm^2 = \kappa$ proves the characteristic-function formula in Proposition 3.1. The NB parameter is $q = hm/(1 + hm)$, which equals the expression given in Section 3.2.

An equivalent unit-time representation is useful for intuition. Let $V \sim \text{Gamma}(1/h, 1/(\kappa h))$ and, conditional on V, let $D \sim \text{Skellam}(V, V)$. It has the same characteristic function as in Proposition 3.1. The independent-intensity and common-activity representations have the same *net-difference* law, although they do not give the same joint distribution of the auxiliary positive and negative counts.

B.2 Moments and strict extension of the Poisson reference

For $|t|$ sufficiently small,

$$\log \mathbb{E}e^{tD} = -\frac{1}{h}\log[1 - 2\kappa h(\cosh t - 1)] = \kappa t^2 + \left(\frac{\kappa}{12} + \frac{\kappa^2 h}{2}\right)t^4 + O(t^6). \tag{12}$$

Hence the second and fourth cumulants are

$$\varkappa_2(D) = 2\kappa, \qquad \varkappa_4(D) = 2\kappa + 12\kappa^2 h. \tag{13}$$

Dividing $\varkappa_4$ by $\varkappa_2^2$ gives the stated excess kurtosis. If a member with $h > 0$ were equal to a symmetric Skellam law with parameter κ', equality of variances would force $\kappa' = \kappa$, while equality of fourth cumulants would force $h = 0$. Thus the positive-h reference laws are genuinely outside the original family.

As $h \downarrow 0$, $[1 + ha]^{-1/h} \to e^{-a}$, so

$$\widehat{K}_{\kappa,h}(\xi) \to \exp[2\kappa(\cos\xi - 1)]. \tag{14}$$

Fourier inversion on $[-\pi, \pi]$, bounded convergence, and the generating function of the modified Bessel functions yield $K_{\kappa,h}(d) \to e^{-2\kappa} I_{|d|}(2\kappa)$. On any fixed finite pair support, normalized costs converge uniformly. Strict convexity of the entropic objective and compactness of the transport polytope then imply convergence of the unique coupling to the finite-κ coupling.

B.3 Explicit kernel, special slices, and tails

For $r = 1/h$, $n = |d|$, and $0 < q < 1$, the kernel defined in Section 3.2 can be expressed as

$$K_{\kappa,h}(d) = (1 - q)^{2r} q^n \frac{(r)_n}{n!}\, {}_2F_1(r + n, r; n + 1; q^2), \tag{15}$$

where $(r)_n = \Gamma(r + n)/\Gamma(r)$. This follows by taking the terms independent of the summation index out of the convolution of the two NB masses.

For $h = 1$, $r = 1$ and $b_{1,q}(n) = (1 - q)q^n$. Summing the geometric series gives

$$K_{\kappa,1}(d) = \frac{1 - q}{1 + q} q^{|d|}, \qquad \rho_{\kappa,1}(d) = \eta|d|, \quad \eta = -\log q = \operatorname{arcosh}\left(1 + \frac{1}{2\kappa}\right). \tag{16}$$

For $h = 1/2$, $r = 2$ and $b_{2,q}(n) = (n + 1)(1 - q)^2 q^n$. Therefore

$$K_{\kappa,1/2}(d) = (1 - q)^4 q^n \frac{(n + 1) + (1 - n)q^2}{(1 - q^2)^3}. \tag{17}$$

Dividing by the mass at zero yields

$$\frac{K_{\kappa,1/2}(d)}{K_{\kappa,1/2}(0)} = q^n \left(1 + n\frac{1 - q^2}{1 + q^2}\right), \tag{18}$$

which proves Eq. (2), since $(1 - q^2)/(1 + q^2) = \tanh\eta$.

For fixed $h > 0$ and $n \to \infty$, the convolution gives

$$K_{\kappa,h}(n) \sim \frac{1}{\Gamma(r)} \left(\frac{1-q}{1+q}\right)^r q^n n^{r-1}. \tag{19}$$

To see this, write the summand as a constant times $q^n q^{2m}\Gamma(m+r)/[\Gamma(r)m!]\times\Gamma(n+m+r)/[\Gamma(r)(n+m)!]$. For fixed m, the last Gamma ratio divided by n^{r-1} tends to $1/\Gamma(r)$. The remaining factors are dominated by a polynomial in m times q^{2m}, so dominated convergence applies; the sum is $(1-q^2)^{-r}$. Thus

$$\rho_{\kappa,h}(n) = \eta n - (r-1)\log n + O(1), \qquad \eta = -\log q. \tag{20}$$

For $h = 0$, in contrast, the fixed-κ Bessel series and Stirling's formula give

$$\rho_{\kappa,0}(n) = n\log(n/\kappa) - n + O(\log n). \tag{21}$$

The two families therefore differ in asymptotic marginal cost, even when their displacement variances agree. The limits $h \downarrow 0$ and $n \to \infty$ need not commute.

C Proofs for state-correspondence geometry

C.1 Bounded marginal penalty with strict convexity

Let $a = \tanh\eta \in (0,1)$. For $u \geq 0$, the right-sided extension of Eq. (2) is $\rho(u) = \eta u - \log(1+au)$. Differentiating gives

$$\rho'(u) = \eta - \frac{a}{1+au}, \qquad \rho''(u) = \frac{a^2}{(1+au)^2} > 0. \tag{22}$$

Since $\tanh\eta < \eta$, $0 < \rho'(u) < \eta$. In particular, for integers $n \geq \ell \geq 0$,

$$0 \leq \rho(n) - \rho(n-\ell) \leq \eta\ell. \tag{23}$$

This proves bounded influence along count displacements while preserving strict convexity on the nonnegative half-line.

C.2 Proof of Theorem 3.2

The consistent assignment has cost $\rho(M)+\rho(\ell)$; the swapped assignment has cost $2\rho(m)+\rho(M-\ell)$, proving the assignment-cost gap formula in Theorem 3.2. For $\rho(n) = n^2$, subtraction yields $2m^2 - 2M\ell$. For Eq. (2), apply Eq. (23) to obtain the margin bound in Theorem 3.2.

A balanced 2×2 coupling can be parameterized as

$$\Pi(p) = \frac{1}{2}\begin{pmatrix} p & 1-p \\ 1-p & p \end{pmatrix}, \qquad 0 \leq p \leq 1. \tag{24}$$

Let C_{same} and C_{swap} denote the two assignment totals. The cost is $[pC_{\text{same}} + (1-p)C_{\text{swap}}]/2$, and $H(\Pi) = -p\log p - (1-p)\log(1-p) + \log 2$. Differentiating gives

$$-\frac{\Delta_\rho}{2} + \tau\log\frac{p}{1-p} = 0, \tag{25}$$

so $p = \sigma(\Delta_\rho/(2\tau))$. A test source mixture $(\omega, 1-\omega)$ is mapped to target weight

$$\widehat{\omega} = \omega p + (1-\omega)(1-p). \tag{26}$$

The true weight of y_1 is ω, and total variation between two distributions on the two distinct targets is the absolute difference of their first weights. Thus

$$|\widehat{\omega} - \omega| = |2\omega - 1|(1-p). \tag{27}$$

Consequently, if $\delta_{\text{state}} > 0$,

$$\text{TV}(\widehat{\mu}_{\omega,a}, \mu_{\omega,a}) \leq \frac{|2\omega-1|}{1+\exp(\delta_{\text{state}}/(2\tau))}. \tag{28}$$

This concerns exact realization of the chosen conditional kernel. If a learned kernel differs from it by at most ϵ in source-averaged conditional TV on the test population, the triangle inequality adds at most ϵ to Eq. (28).

C.3 Ordinary differences, strong responses, and comparison costs

For sources $(1, 2)$ and targets $(3, 4)$ in one dimension, the consistent assignment has displacements $(2, 2)$ and the swapped assignment $(3, 1)$. The L_1 totals are equal. Equation (22) instead gives $2\rho(2) < \rho(1) + \rho(3)$. In this example the linear cost has no conditional discrimination, whereas the nonlinear slice has a positive matching margin.

For the heterogeneous construction, fixed-κ Bessel cost has $\rho_{\kappa,0}(M) - \rho_{\kappa,0}(M - \ell) = \ell \log(M/\kappa) + o(1)$. Thus its swapped-minus-consistent margin eventually becomes negative for any fixed m, ℓ, κ. A positive-h choice changes this tail behavior rather than only its finite scale. At $(\kappa, h, m, \ell) = (4, 1/2, 3, 1)$,

$$\delta_{\text{state}} = 2\rho(3) - \rho(1) - \log 2 \approx 1.18335. \tag{29}$$

At $M = 100$, the exact margins are -182 for squared Euclidean, approximately -2.0995 for Bessel with $\kappa = 4$, and approximately 1.1932 for Eq. (2).

D Integer semigroup and Schrödinger construction

D.1 Existence and jump generator

For $h > 0$, independent NB increments with common q add their shape parameters. Therefore two independent processes $N^{\pm}$ with interval law $b_{\Delta/h,q}$ define the semigroup in Eq. (3). Its log characteristic function can be expanded as

$$\log \widehat{K}_{\Delta}^{\kappa,h}(\xi) = \frac{\Delta}{h} \sum_{k=1}^{\infty} \frac{q^k}{k} \left(e^{ik\xi} + e^{-ik\xi} - 2 \right). \tag{30}$$

This gives the one-coordinate generator

$$\mathcal{L}f(x) = \sum_{k \geq 1} \frac{q^k}{hk} \left[f(x + k) + f(x - k) - 2f(x) \right]. \tag{31}$$

The total jump rate is $-2\log(1 - q)/h < \infty$. The multigene generator is the sum of these coordinate generators. This is an integer-valued compound-Poisson construction within the established Lévy-process theory (Barndorff-Nielsen et al., 2012).

At fixed κ, $q = \kappa h + O(h^2)$ as $h \downarrow 0$. Hence the ± 1 rates tend to κ, while all $\pm k$, $k \geq 2$, rates vanish. Equivalently, Eq. (3) converges to the symmetric Poisson difference semigroup. The limiting relation therefore holds for finite-dimensional path distributions, not just the unit-time cost. A deterministic schedule multiplies the generator by $w'(t)$ where the derivative exists.

D.2 Static–dynamic equivalence

We state the construction on finite empirical endpoint supports with positive cell weights; this is the setting used by the solver. Since every integer increment has positive reference probability, every endpoint pair is accessible and all finite-support coupling KL divergences are finite. Let R have initial marginal $\mu_{e,a}$ and transition K_1. For a candidate path law P with endpoint law γ, disintegration gives

$$\mathrm{KL}(P\|R) = \mathrm{KL}(\gamma\|R_{01}) + \mathbb{E}_{\gamma}\, \mathrm{KL}\left(P(\cdot \mid X_0, X_1) \| R(\cdot \mid X_0, X_1) \right). \tag{32}$$

The second term is minimized at the reference bridges. Since $R_{01}(y, x) = \mu_{e,a}(y) K_1(x - y)$ and the source marginal is fixed,

$$\mathrm{KL}(\gamma\|R_{01}) = \mathbb{E}_{\gamma}[-\log K_1(X - Y)] - H(\gamma) + \text{constant}. \tag{33}$$

This proves Eq. (4) and the reference-bridge mixture representation using the standard Schrödinger decomposition (Léonard, 2013).

The optimal finite plan has positive scaling factors f, g with $\gamma^*(y, x) = f(y)\mu_{e,a}(y)K_1(x - y)g(x)$. Define

$$\varphi_t(z) = \sum_x K_{1-w(t)}(x - z) g(x). \tag{34}$$

At interior times, the forward transition is the Doob transform

$$P^*(X_t = z' \mid X_s = z) = K_{w(t)-w(s)}(z' - z)\frac{\varphi_t(z')}{\varphi_s(z)}. \tag{35}$$

It is normalized by the semigroup property. This supplies the Markov path law associated with the chosen endpoint coupling.

D.3 Entropy, units, and the restricted search space

Adding any source-only or target-only term to a finite cost does not change the plan. This justifies removing the parameter-only normalization term when using Eq. (1). Scaling the entire cost is different: using c/τ changes the Gibbs kernel to $K_1^{1/\tau}$. Dividing a summed gene cost by G while keeping the entropy coefficient fixed therefore changes the reference rather than merely stabilizing computation.

For fixed positive marginals, strict convexity of $\sum_{ij} \Pi_{ij} \log \Pi_{ij}$ gives uniqueness, and Lagrange multipliers give the Sinkhorn form. Taking the four-point ratio cancels all marginal scaling factors, yielding the pairing-odds identity in Section 3.4. The family changes relative pairing odds through (κ, h) while remaining in a restricted subset of the transport polytope.

The theoretical reference is supported on $\mathbb{Z}^G$ and the trained endpoint model on $\mathbb{N}_0^G$. Integer-valued intermediate paths require neither token quantization nor output rounding. Their time is computational; the likelihood of a reference path is not identified with a measured biological trajectory.

E Exact conditional bridges and numerical evaluation

E.1 Conditional slack distribution

On an operational-time interval of length Δ, let $r = \Delta/h$ and let N^+, N^- be independent with mass $b_{r,q}$. Conditioning on $d = N^+ - N^-$ restricts the counts to $N^+ = m + d_+$ and $N^- = m + d_-$. Their joint mass immediately yields the conditional slack distribution in Section 3.5, with normalizer $K_\Delta^{\kappa,h}(d)$. Notice that $r = \Delta/h$ changes with the interval length, whereas q is fixed by (κ, h).

For $n = |d|$ and unnormalized weights $v_m = b_{r,q}(m)b_{r,q}(m+n)$,

$$\frac{v_{m+1}}{v_m} = q^2\frac{(m+r)(m+n+r)}{(m+1)(m+n+1)}. \tag{36}$$

This provides a log-domain recursion. Let $a = (r-1)_+$ and

$$R_m = q^2\left(1 + \frac{a}{m+1}\right)\left(1 + \frac{a}{m+n+1}\right). \tag{37}$$

Every subsequent ratio is bounded by R_m. Once $R_m < 1$, the omitted tail after index m obeys

$$\sum_{j=m+1}^{\infty} v_j \le \frac{v_m R_m}{1 - R_m}. \tag{38}$$

Comparing this bound with the accumulated mass gives a certified tolerance for kernel evaluation or truncated-CDF sampling. Displacements themselves are not clipped to a maximum count; tables are extended when needed.

E.2 Conditional allocation is beta-binomial

Suppose $A \sim b_{\alpha,q}$ and $B \sim b_{\beta,q}$ independently. Since their sum follows $b_{\alpha+\beta,q}$,

$$\Pr(A = k \mid A + B = n) = \binom{n}{k}\frac{(\alpha)_k(\beta)_{n-k}}{(\alpha+\beta)_n} = \mathrm{BetaBinomial}(k; n, \alpha, \beta). \tag{39}$$

Given N^+ and N^-, apply this relation independently to the early and late increments of each NB process. Choosing $\alpha = [w(t) - w(u)]/h$ and $\beta = [w(v) - w(t)]/h$ proves the beta-binomial allocation rule in Section 3.5. The positive and negative allocations use independent beta-binomial draws; sharing an auxiliary Beta variable would change the bridge law.

Algorithm 1 One-coordinate reference bridge, $h > 0$

Require: Endpoints $a, b \in \mathbb{Z}$; $u \le t \le v$; $\kappa, h > 0$; schedule w
if $t = u$ **then**
  return a
end if
if $t = v$ **then**
  return b
end if
$d \leftarrow b - a$; $r \leftarrow [w(v) - w(u)]/h$; compute q from (κ, h) as in Section 3.2
Sample M with mass proportional to $b_{r,q}(m + d_+)b_{r,q}(m + d_-)$
$n_+ \leftarrow M + d_+$; $n_- \leftarrow M + d_-$
$\alpha \leftarrow [w(t) - w(u)]/h$; $\beta \leftarrow [w(v) - w(t)]/h$
Independently draw $A_+ \sim \text{BetaBinomial}(n_+; \alpha, \beta)$ and $A_- \sim \text{BetaBinomial}(n_-; \alpha, \beta)$
return $a + A_+ - A_-$

E.3 A rejection-free special case

When $r = \Delta/h = 2$, the slack posterior has weights proportional to $(m + 1)(m + n + 1)a^m$, where $a = q^2$ and $n = |d|$. Since

$$(m + 1)(m + n + 1) = m(m + 1) + (n + 1)(m + 1), \tag{40}$$

its distribution is the exact mixture

$$M \sim \begin{cases} \text{NB}(2, a), & \text{with probability } 1 - \zeta, \\ 1 + \text{NB}(3, a), & \text{with probability } \zeta, \end{cases} \qquad \zeta = \frac{2a}{(n + 1)(1 - a) + 2a}. \tag{41}$$

Here $b_{r,q}$ uses the NB parameterization defined in Section 3.2. This shortcut applies whenever the *interval shape* equals two, including the full interval at $h = 1/2$. It is not valid for every intermediate interval merely because the global shape parameter is $h = 1/2$.

E.4 Poisson limit and stable parameterization

For $h = 0$, on duration Δ, use independent Poisson totals with mean $\kappa\Delta$. The slack weights are proportional to

$$v_m^{(0)} = \frac{(\kappa\Delta)^{2m+|d|}}{m!(m + |d|)!}, \qquad \frac{v_{m+1}^{(0)}}{v_m^{(0)}} = \frac{(\kappa\Delta)^2}{(m + 1)(m + |d| + 1)}. \tag{42}$$

Each positive and negative total is independently thinned with probability $p = [w(t)-w(u)]/[w(v)-w(u)]$. This is the Count Bridge limit (Fishman et al., 2026). For $h \downarrow 0$, the NB totals converge to the Poisson totals and beta-binomial allocations converge to these binomial allocations.

For small positive h, compute

$$q = \frac{4\kappa h}{(\sqrt{1 + 4\kappa h} + 1)^2} \tag{43}$$

rather than subtracting nearly equal quantities. Use log-Gamma functions for NB masses and log-sum-exp for convolution sums. Exact endpoint cases are handled before evaluating beta parameters. The analytic $h = 0$ branch is distinct from numerical approximations at small positive h.

E.5 Computational structure

The one-dimensional cost table is reusable across all genes for a fixed global reference. Pairwise mini-batch costs require $O(B_0B_1G)$ lookup-and-sum operations; L Sinkhorn iterations cost $O(LB_0B_1)$. These orders match a coordinatewise raw-count pairing implementation. Bridge sampling factorizes across coordinates under the reference, and the denoiser predicts a fixed number of parameters per gene. No categorical output dimension scales with the largest observed count.

F Endpoint learning, sampling, and reference consistency

F.1 Conditional supervision selected by a coupling

Fix (e, a) and suppress these indices. Let γ have control marginal μ and treated marginal ν, with treated-given-control law K_γ. For direct endpoint prediction on finite support,

$$\mathcal{L}_\gamma(Q) = \mathbb{E}_{(Y,X)\sim\gamma}[-\log Q(Y \mid X)] \tag{44}$$

$$= H_\gamma(Y \mid X) + \mathbb{E}_{X\sim\mu}\,\mathrm{KL}\big(K_\gamma(\cdot \mid X)\|Q(\cdot \mid X)\big). \tag{45}$$

This follows by adding and subtracting $\log K_\gamma(Y \mid X)$ inside the expectation. The unrestricted optimum equals K_γ on the training control support. Under independent endpoint pairing, $K_\gamma(dy \mid x) = \nu(dy)$, so the target does not distinguish baseline states. By contrast, swapping strong and weak responses between two states can retain input dependence while reversing its meaning.

The bridge denoiser learns the endpoint posterior conditioned on the intermediate state. Proposition F.1 establishes how the matched-reference sampler realizes the selected endpoint relation.

F.2 Likelihood target

Write $H_t = (X_t, t, e, a)$ and let P denote the training bridge mixture. If the relevant conditional entropies are finite, the endpoint NLL decomposes as

$$\mathbb{E}[-\log q(Y \mid H_T)] = H_P(Y \mid H_T) + \mathbb{E}_{H_T}\,\mathrm{KL}\big(P(Y \mid H_T)\|q(Y \mid H_T)\big). \tag{46}$$

Thus an unrestricted denoiser learns the endpoint posterior. A factorized ZINB family yields its conditional KL projection. Endpoint independence in this head does not imply independent final gene marginals: the shared network and random intermediate states induce a mixture of count distributions.

The count likelihood uses mean $\mu > 0$, inverse dispersion $r > 0$, and zero-inflation probability $\pi \in [0, 1]$:

$$\mathrm{NB}(y; \mu, r) = \frac{\Gamma(y + r)}{\Gamma(r)y!}\left(\frac{r}{r+\mu}\right)^r\left(\frac{\mu}{r+\mu}\right)^y, \tag{47}$$

$$\mathrm{ZINB}(y; \mu, r, \pi) = \pi\,\mathbf{1}_{\{y=0\}} + (1-\pi)\mathrm{NB}(y; \mu, r). \tag{48}$$

Softplus and sigmoid links enforce the parameter domains. A sample is generated by a Bernoulli zero gate followed, when active, by a Gamma–Poisson NB draw. This head specifies the endpoint readout law rather than the auxiliary increment law.

F.3 Exact reverse transition with the matched reference

Proposition F.1 (Reference-consistent endpoint resampling)**.** *Let P^* be the exact Schrödinger bridge for a positive reference kernel and fixed endpoint marginals. For $0 \le u < t \le 1$,*

$$P^*(X_u \in A \mid X_t = z) = \sum_y P^*(Y = y \mid X_t = z)R(X_u \in A \mid X_0 = y, X_t = z). \tag{49}$$

Initializing at P^*_1 and applying these transitions on any finite decreasing grid reproduces the corresponding finite-dimensional marginals of P^*, including its endpoint coupling.

Proof. The optimal path density relative to R factors into a function of X_0 times a function of X_1. Conditional on $X_0 = y$ and $X_t = z$, future endpoint factors integrate to a function of z, so the law of the earlier bridge segment is $R(\cdot \mid X_0 = y, X_t = z)$. Disintegrating over Y proves Eq. (49). The optimal path is Markov, so the reverse transition chain reconstructs its grid law. □

In practice, minibatch OT, finite network capacity, and optimization approximate this construction.

Algorithm 2 PerturbBridge training and generation

Require: Reference (κ, h), condition groups, network q_θ, schedule w
for each training update **do**
  Sample control and perturbed batches within each observed (e, a)
  Form $C_{ji} = \sum_g \rho_{\kappa,h}(x_{ig} - y_{jg})$; solve log-Sinkhorn with natural entropy coefficient one
  Sample endpoint indices $(j, i) \sim \Pi$ and set $(X_0, X_1) = (y_j, x_i)$
  Sample $t \sim \mathrm{Unif}(0, 1)$; sample X_t coordinatewise with Algorithm 1 or its $h = 0$ limit
  Update θ using $-\log q_\theta(X_0 \mid X_t, t, e, a)$
end for
Generation: Given control x, initialize $z \leftarrow x$ at $t_K = 1$
for $k = K, K-1, \ldots, 1$ **do**
  Draw $\widetilde{y} \sim q_\theta(\cdot \mid z, t_k, e, a)$
  Sample the state at t_{k-1} in the reference bridge from $\widetilde{y}$ at 0 to z at t_k; replace z by this sample
end for
return $z \in \mathbb{N}_0^G$

F.4 What a mismatched path can change

Suppose a different coupling γ is mixed with the old reference bridges. Write S for the original control, Y for the target, and Z for an intermediate state. The three-time training law is

$$P_\gamma(s, z, y) = \gamma(y, s)R(z \mid s, y). \tag{50}$$

A two-step sampler that keeps only its current state instead realizes

$$M_\gamma(s, z, y) = P_\gamma(s)P_\gamma(z \mid s)P_\gamma(y \mid z). \tag{51}$$

Direct expansion gives

$$\mathrm{KL}(P_\gamma \| M_\gamma) = I_{P_\gamma}(S; Y \mid Z). \tag{52}$$

Thus, matching intermediate marginals does not automatically preserve the endpoint relation. For example, a balanced swap coupling $0 \to 2, 2 \to 0$ under a symmetric reference has identical midpoint distributions for the two sources. A midpoint-only endpoint posterior must then return the two endpoints with equal probability, losing the swap relation.

G Datasets, training, and evaluation protocol

NADIG contains 408,429 cells, 7,632 genes, two cell lines, and 2,394 targeting perturbations. PBMC contains 9,697,974 cells, 18,213 genes, 18 cell types, 12 donors, and 90 cytokines.

G.1 Datasets and split assignments

Table 3 gives the cell counts before excluding control groups and small test groups from scoring. Random splits use seed 42 and cell-level 60/20/20 allocation. A condition group is a cell-line–perturbation pair for NADIG and a cell-type–donor–cytokine triplet for PBMC. Within each task, all methods use the same group eligibility rules and evaluator.

Table 3: Split sizes and numbers of evaluated perturbed condition groups. Validation for PBMC OOD cell type uses source types, whereas its test set uses unseen target types.

Task	Train cells	Validation cells	Test cells	Groups
NADIG ID	245,057	81,685	81,687	4,141
PBMC ID	5,818,784	1,939,594	1,939,596	15,025
PBMC OOD cell type	4,237,908	224,829	5,235,237	8,545
PBMC OOD donor	5,758,816	2,375,248	1,563,910	2,899

NADIG. After unified preprocessing, the Perturb-seq matrix contains 408,429 cells and 7,632 genes from HepG2 and Jurkat, with 2,394 targeting perturbations and non-targeting controls. All 4,747 cell-line–perturbation combinations present in the test partition also occur in training; this is an ID task, not a held-out-perturbation task. Evaluation removes non-targeting groups and perturbed groups with fewer than five test cells, leaving 4,141 groups and 76,623 perturbed cells. For each eligible group, we sample controls with replacement from the same cell line's test-partition non-targeting pool and generate the same number of predicted cells as real cells. The largest group contains 483 test cells, so the 1,000-cell distribution-evaluation cap is never active.

PBMC. The cytokine matrix contains 9,697,974 cells and 18,213 genes, spanning 18 cell types, 12 donors, 90 cytokines, and PBS controls. Evaluation excludes PBS as a target condition and requires at least five test cells per condition group. Inputs are PBS cells matched to the target cell-type–donor context. Target and predicted populations have matched sizes; distribution metrics use at most 1,000 cells per group. Response summaries and DE tests use the full group. The three main tasks share this evaluator with the reference-sensitivity study.

PBMC donor split. Training donors are 1, 4, 6, 7, 8, 10, and 12; validation donors are 3, 5, and 11; test donors are 2 and 9. These identity sets are disjoint.

PBMC cell-type split. The ten source types are B Intermediate/Memory, CD4 Naive, CD14 Mono, CD16 Mono, ILC, MAIT, NK CD56bright, Plasmablast, cDC, and pDC. The eight target types are B Naive, CD4 Memory, CD8 Memory, CD8 Naive, HSPC, NK, NKT, and Treg. Validation holds out 5% of source-type cells, stratified by perturbation, cell type, and donor. Training and validation share source-type identities but have disjoint cells; test-type identities are absent from both. The source pool contains 4,462,737 cells before removing the 224,829 validation cells, leaving 4,237,908 training cells. The sensitivity runs use this same exclusion.

G.2 Implementation and comparator scope

The condition encoder embeds perturbation and available context metadata, applies a shared pre-pooling MLP, and pools through a learned attention query. A cross-attention MLP combines this condition with the signed bridge state and time. Table 4 records the backbone and optimization settings retained from the model specification.

Control-derived context embeddings. All six learned models receive the same fixed control-derived PCA embeddings for held-out identities in the PBMC OOD tasks. The construction below is shared across models; lookup and denoising details describe their use in PerturbBridge. In PBMC OOD-cell-type runs, the cell-type lookup is filled with offline PBS-control summaries and remains fixed during training; donor tokens use a separate learned lookup. For each cell-type–donor group (c, d), at most 512 PBS cells are sampled with seed 42. Counts are summed across the sampled cells on all $G = 18{,}213$ genes, normalized to library size 10^4, and log-transformed:

$$p_{c,d,g} = \log\left(1 + 10^4 \frac{\sum_{i \in \mathcal{S}_{c,d}} x_{i,g}}{\sum_{g'=1}^{G} \sum_{i \in \mathcal{S}_{c,d}} x_{i,g'}}\right). \tag{53}$$

Let $\mathcal{T}$ be the training control groups, and $C_d = \{c : (c, d) \in \mathcal{T}\}$. Training groups determine the donor mean offsets

$$\delta_d = \frac{1}{|C_d|} \sum_{c \in C_d} p_{c,d} - \frac{1}{|\mathcal{T}|} \sum_{(c,d) \in \mathcal{T}} p_{c,d}, \qquad \widetilde{p}_{c,d} = p_{c,d} - \delta_d. \tag{54}$$

This donor mean-centering step reduces donor variation in the cell-type representation. An independent learned donor token remains in the condition encoder. Let μ_{train} be the mean adjusted training pseudobulk. SVD of the matrix with training rows $\widetilde{p}_{c,d} - \mu_{\text{train}}$ supplies 64 principal directions V_k. The scores are standardized using training-group standard deviations and averaged across donors for each cell type:

$$z_{c,d,k} = \frac{(\widetilde{p}_{c,d} - \mu_{\text{train}})^\top V_k}{\text{Std}_{(c,d) \in \mathcal{T}}[(\widetilde{p}_{c,d} - \mu_{\text{train}})^\top V_k]}, \qquad e_c = \frac{1}{|D_c|} \sum_{d \in D_c} z_{c,d}. \tag{55}$$

Here D_c is the set of donors with available PBS-control pseudobulks used to construct the embedding for cell type c; it is the averaging set, while $\mathcal{T}$ specifies the groups used to fit the correction and PCA transform. The resulting 18×64 matrix supplies the fixed cell-type lookup. The training-fitted PCA transform is reused for evaluation-context controls. Pseudobulk construction and PCA are performed offline; denoising uses the fixed lookup tokens alongside the current bridge state and the independently learned metadata tokens. For OOD-donor prediction, the same construction interchanges cell type and donor. Define $D_c^{\text{train}} = \{d : (c, d) \in \mathcal{T}\}$ and remove training-estimated cell-type offsets:

$$\alpha_c = \frac{1}{|D_c^{\text{train}}|} \sum_{d \in D_c^{\text{train}}} p_{c,d} - \frac{1}{|\mathcal{T}|} \sum_{(c,d) \in \mathcal{T}} p_{c,d}, \qquad \widetilde{p}_{c,d}^{\text{donor}} = p_{c,d} - \alpha_c. \tag{56}$$

A separate 64-dimensional PCA is fitted and standardized on these adjusted training control groups. Its scores are averaged across cell types for each donor, yielding a fixed 12×64 donor lookup; the cell-type lookup remains learned. In both OOD tasks, the transform, centering, and standardization are fitted on training controls and then applied to PBS controls from evaluation contexts. The derived embeddings remain frozen throughout model training and inference.

Table 4: PerturbBridge backbone and optimization settings.

Component	Setting
Denoiser	6 layers; width 1,024; 8 cross-attention heads; 8 condition tokens
Condition encoder	Internal width 64; hidden width 128; output width 144; 4 pooling heads
OOD context tokens	Fixed control-derived cell-type vectors for OOD cell types; fixed control-derived donor vectors for OOD donors
State pathway	Projected bridge state concatenated at each denoiser layer
Output	$3G$ ZINB parameters; positive mean and inverse dispersion
Optimizer	AdamW; learning rate 3×10^{-4}; weight decay 0.01
Schedule	Cosine decay to 10^{-6}; gradient clipping at 1.0
EMA	Decay 0.999; EMA weights evaluated
NADIG	Microbatch 256; accumulation 4; 200 epochs
PBMC	Microbatch 128; accumulation 4; 30 epochs
Reference	Global (κ, h) within a run; natural OT entropy coefficient one

We compare with five learned methods, Control, and Linear (R1). Linear (R1) preserves the identity of the exported `Tiny Linear R1` baseline. DOLORIS and DEPARTURES scores describe the implementations evaluated in this benchmark, rather than numbers quoted from their original papers.

H Metric definitions and aggregation

H.1 Expression views and response vectors

For a condition group q, let X_q, $\widehat{X}_q$, and $X_{0,c}$ denote real perturbed cells, generated cells, and controls in context c. A context is a cell line for NADIG and a cell-type–donor pair for PBMC. The four views are

$$v_{\text{count}}(x) = x, \quad v_{\log}(x) = \log(1+x), \quad v_{\text{norm},q}(x) = \frac{L_q x}{\sum_g x_g}, \quad v_{\text{normlog},q}(x) = \log\big(1 + v_{\text{norm},q}(x)\big), \tag{57}$$

where L_q is the median library size of the real perturbed test group, and logarithms act coordinate-wise. The same group-specific transformation is used for real, predicted, and control cells when evaluating that group. Thus, normalized scores assess expression at a target-informed scale and do not measure a model's ability to predict total counts. This evaluation transformation is separate from model fitting.

In any one view, let μ_q, $\widehat{\mu}_q$, and $\mu_{0,c}$ denote the corresponding gene-wise means and define

$$\delta_q = \mu_q - \mu_{0,c}, \qquad \widehat{\delta}_q = \widehat{\mu}_q - \mu_{0,c}. \tag{58}$$

For normalized views the control mean uses the same L_q as the target group; its dependence on q is suppressed in the notation.

H.2 Distribution metrics

Let $P = \sum_i a_i \delta_{x_i}$ and $Q = \sum_j b_j \delta_{y_j}$ be uniformly weighted real and generated empirical measures in the chosen view. The exact joint transport scores are

$$W_p(P, Q) = \left[\min_{\pi \in \Pi(a,b)} \sum_{ij} \pi_{ij} \|x_i - y_j\|_2^p \right]^{1/p}, \qquad p \in \{1, 2\}. \tag{59}$$

These correspond to `strict_w1` and `strict_w2`. The marginal score averages one-dimensional exact transport over genes:

$$W_{1,\text{marginal}} = \frac{1}{G} \sum_{g=1}^{G} W_1(P_g, Q_g). \tag{60}$$

It measures gene-wise marginals without measuring joint cross-gene dependence.

The Sinkhorn fields retain the evaluator's GeomLoss `SamplesLoss` outputs (Feydy et al., 2019) with `blur=0.05` and `debias=False`; the $p = 2$ field takes the square root of the regularized output. In the library's native convention,

$$\text{OT}_{p,\epsilon} = \min_{\pi \in \Pi(a,b)} \sum_{ij} \pi_{ij} \frac{\|x_i - y_j\|_2^p}{p} + \epsilon \, \text{KL}(\pi \| a \otimes b), \qquad \epsilon = 0.05^p. \tag{61}$$

In particular, the native $p = 2$ ground cost contains a factor of $1/2$, so its square-root output should not be numerically equated to the exact W_2 field. We retain the exported values without rescaling. The convention follows the GeomLoss implementation documentation.

The MMD field is the nonnegative part of the unbiased squared RBF-kernel estimator (Gretton et al., 2012):

$$\widehat{\text{MMD}}_u^2 = \frac{\sum_{i \neq i'} k(x_i, x_{i'})}{m(m-1)} + \frac{\sum_{j \neq j'} k(y_j, y_{j'})}{n(n-1)} - \frac{2 \sum_{ij} k(x_i, y_j)}{mn}, \tag{62}$$

$$k(x, y) = \exp\left(-\frac{\|x - y\|_2^2}{2\sigma^2} \right), \qquad \texttt{mmd} = \max(0, \widehat{\text{MMD}}_u^2). \tag{63}$$

The bandwidth σ is the median pairwise Euclidean distance in the combined real and predicted sample. Every distribution metric is lower-is-better. Distribution estimators use at most 1,000 real and 1,000 generated cells per group.

H.3 Pseudobulk response metrics

Writing $e_q = \widehat{\delta}_q - \delta_q$ and $\bar{\delta}_q = G^{-1} \sum_g \delta_{qg}$,

$$R_q^2 = 1 - \frac{\|e_q\|_2^2}{\|\delta_q - \bar{\delta}_q \mathbf{1}\|_2^2}, \qquad \texttt{pdcorr}_q = \text{Corr}_g(\widehat{\delta}_{qg}, \delta_{qg}). \tag{64}$$

Both are higher-is-better. The fields named `mae` and `mse` are sums, $\|e_q\|_1$ and $\|e_q\|_2^2$, respectively. Their `_per_gene` counterparts divide these sums by G and correspond to the usual mean absolute and mean squared errors. All four error fields are lower-is-better and use the full condition group.

H.4 Differential-expression metrics

For each gene, we compare real perturbed cells with matched controls and, separately, generated cells with those controls using two-sided Welch tests (Welch, 1947). For each comparison,

$$t_g = \frac{\mu_{1g} - \mu_{0g}}{\sqrt{s_{1g}^2/n_1 + s_{0g}^2/n_0}}, \qquad \nu_g = \frac{(s_{1g}^2/n_1 + s_{0g}^2/n_0)^2}{(s_{1g}^2/n_1)^2/(n_1 - 1) + (s_{0g}^2/n_0)^2/(n_0 - 1)}. \tag{65}$$

Benjamini–Hochberg correction (Benjamini & Hochberg, 1995) is applied across all genes within a comparison. Let T_q and $\widehat{T}_q$ be real and predicted DE sets at adjusted q-value below 0.05. Their sizes give `true_de_n` and `pred_de_n`.

For DE overlap, set $K = |T_q|$ and rank genes by adjusted q-value ascending, breaking ties by absolute effect descending. Then

$$\text{DEOverlap}_q = \frac{|\,\text{TopK}_{\text{real}} \cap \text{TopK}_{\text{pred}}\,|}{K}, \tag{66}$$

$$\text{DEPrecision}_q = \frac{|T_q \cap \widehat{T}_q|}{|\widehat{T}_q|}, \tag{67}$$

$$\text{DirectionAgreement}_q = \frac{\sum_{g \in T_q \cap \widehat{T}_q} \mathbf{1}[\text{sign}(\delta_{qg}) = \text{sign}(\widehat{\delta}_{qg})]}{|T_q \cap \widehat{T}_q|}. \tag{68}$$

The `lfcspear` field computes Spearman correlation between δ_q and $\widehat{\delta}_q$ on the real DE set. Despite its legacy name, it is a correlation of response values in the evaluated expression view, not an additional log-fold-change transformation. AUROC and AP use labels $\mathbf{1}[g \in T_q]$ and prediction scores $-\log_{10}(\widehat{q}_g)$; `auprc` stores average precision. All these scores are higher-is-better, whereas predicted DE breadth is descriptive and must be compared with real breadth. DE tests use all cells of the eligible group.

H.5 Perturbation discrimination and effect breadth

Within a context containing M eligible perturbations, the evaluator computes

$$\text{PDS} = 1 - \frac{1}{M} \sum_{i=1}^{M} \frac{\#\{j \neq i : d(\widehat{\delta}_i, \delta_j) < d(\widehat{\delta}_i, \delta_i)\}}{M}. \tag{69}$$

The three distances are squared L_2, L_1, and cosine distance. The denominator is M, not $M - 1$, and only strictly closer mismatches lower the score. Higher PDS means that predicted responses more often retrieve the correct real perturbation response. In the absence of ties, random ranks have expectation $(M + 1)/(2M)$, approximately 0.5; identical distances receive favorable ties under this convention. PDS therefore needs to be read alongside response magnitude and DE metrics.

For count and log-count views, HVG-PDS restricts the response vectors to the 2,000 genes of greatest variance across real perturbation responses within the context. This feature selection is part of test scoring, not model fitting; `hvg_n_genes` records the number used. The effect-breadth score is

$$\text{ES} = \rho_S\left(\{|\widehat{T}_i|\}_{i=1}^{M}, \{|T_i|\}_{i=1}^{M}\right). \tag{70}$$

It measures the ranking of perturbations by DE breadth and is higher-is-better; it is not an energy distance.

H.6 Aggregation and field accounting

Let n_q^{dist} be the number of cells used for distribution scoring in group q. Distribution metrics aggregate as

$$\overline{m}_{\text{dist}} = \frac{\sum_q n_q^{\text{dist}} m_q}{\sum_q n_q^{\text{dist}}}. \tag{71}$$

Pseudobulk and DE fields use equal group weights. PDS, HVG-PDS, and ES are computed within context, then averaged equally over contexts: cell lines for NADIG and cell-type–donor blocks for PBMC. Consequently, a task-level distribution score and a task-level response score weight the underlying data differently.

The 108 exported fields comprise 98 distinct performance fields, seven metadata fields, and three compatibility aliases. Each expression view contributes 23 performance fields and the true DE count; count and log-count additionally contribute three HVG-PDS scores and one HVG gene count each. The remaining metadata field is the number of evaluated groups. The 98-field schema includes predicted DE counts and scaled copies of error metrics. The aliases are

$$\begin{aligned} \texttt{test_mmd} &= \texttt{test_count_mmd}, \\ \texttt{test_sinkhorn} &= \texttt{test_count_sinkhorn_w2}, \\ \texttt{test_w1} &= \texttt{test_count_sinkhorn_w1}. \end{aligned}$$

They are not counted again in comparative summaries. Missing exported values are shown as dashes and are not imputed.

I Population-shape changes and response prediction

This analysis asks whether the unseen-cell-type advantage coincides with perturbations that change population structure beyond a common mean shift. The shape diagnostic uses observed treated and PBS populations to select cases, independently of model predictions. It is a post-hoc test-set analysis and is separate from the control-derived context embeddings used as model inputs.

I.1 A common PCA coordinate system

We use all 18,213 protein-coding genes and transform each cell as

$$z_{ig} = \log\left(1 + 10^4 x_{ig} / \sum_{g'} x_{ig'}\right). \tag{72}$$

The diagnostic uses a fixed normalization target of 10^4, distinct from the target-group median scale in the main evaluator. PCA is fitted on the intersection of the OOD-cell-type and OOD-donor training splits, containing 2,748,574 cells. Sampling is stratified by cell type, donor, and perturbation: each PBS group contributes at most 64 cells and each non-PBS group at most two, followed by a fixed sample of 16,000 cells with seed 42. This limits the dominance of abundant groups in the fitted coordinate system. PCA subtracts the fitting-sample gene means, without gene-wise variance standardization or whitening. The first two and first 30 components explain approximately 7.96% and 19.79% of fitting-sample variance, respectively. All populations are projected into the same 30-dimensional space, so the diagnostic characterizes structure within this retained subspace.

I.2 Centered distance relative to PBS sampling variation

Each non-PBS group is compared with PBS cells of the same cell type and donor. We sample at most 128 treated cells and a PBS pool of at most 1,024 cells. Both sides use

$$n = \min\left(128, n_{\text{treated}}, \lfloor n_{\text{PBS pool}}/2 \rfloor\right), \tag{73}$$

and groups with $n < 20$ are excluded. Let $A, B \in \mathbb{R}^{n\times 30}$ be the treated and PBS coordinates. Centering each population separately gives $\widetilde{A} = A - \bar{A}$ and $\widetilde{B} = B - \bar{B}$, removing their mean displacement. For 32 fixed random unit directions u_j, shared by all groups, sort the projected coordinates $a_{ij} = \widetilde{A}_i^\top u_j$ and $b_{ij} = \widetilde{B}_i^\top u_j$. Define

$$d_{\text{shape}}(A, B) = \left[\frac{1}{32n}\sum_{j=1}^{32}\sum_{i=1}^{n}\left(a_{(i)j} - b_{(i)j}\right)^2\right]^{1/2}. \tag{74}$$

This is a random-projection approximation to centered sliced Wasserstein-2, rather than exact joint W_2. It detects changes in dispersion, covariance structure, and non-Gaussian shape after mean removal. Because the PCs are not whitened, high-variance directions retain greater influence.

For each cell-type–donor–n combination, we draw $2n$ PBS cells without replacement and split them into two disjoint groups, repeating this procedure 16 times. Different repetitions may reuse cells. Both halves are separately centered and compared with the same directions. The normalized diagnostic is

$$s_{\text{shape}} = \frac{d_{\text{shape}}(A, B)}{\text{median}_{r=1}^{16}\, d_{\text{shape}}(B_{r,1}, B_{r,2})}. \tag{75}$$

Perturbations with the same context and sample size share this PBS baseline. Values near one indicate a distance comparable to typical PBS sampling variation. We use the descriptive cutoff $s_{\text{shape}} > 2$ to select extreme shape changes. The 5th, 50th, and 95th percentiles over the 16 PBS splits describe the empirical resampling distribution.

I.3 Screening and selected cases

The screen selects 35 OOD-cell-type and three OOD-donor groups with $s_{\text{shape}} > 2$, providing a collection of extreme cases with pronounced changes relative to PBS sampling variation. Filtering

the 35 cell-type cases to those with at least 500 treated cells before diagnostic subsampling leaves 20 groups (Table 6). Selection uses observed population shape and group size independently of predictive performance.

For example, B Naive–Donor4–SCF has centered distance 0.87547 and median PBS distance 0.27910, giving $s_{\text{shape}} = 3.13678$ at $n = 128$. Its treated-to-PBS RMS-radius ratio is 1.208, illustrating that broadening contributes to the detected shape change. The three donor cases are CD4 Naive–Donor9–IL-15, CD14 Mono–Donor2–IFN-omega, and CD14 Mono–Donor9–IL-32-beta.

I.4 Prediction performance and interpretation

We evaluate the eight methods on the 20 selected groups using the supplied per-case metrics in all four expression views. Table 5 reports equal-case means. PerturbBridge has the best mean MMD, response Pearson, DE-set Spearman, and DE AP in each view. It also leads W_2 in raw and log counts, while Linear (R1) leads W_2 in the normalized views. In raw counts PerturbBridge leads all seven comparators separately in 16 of 20 cases for Pearson and 16 for MMD, compared with nine cases for W_2 and nine for DE AP. This distinction matters: the aggregate advantage combines broad response-pattern recovery with more variable case-level gains in transport distance and DE ranking. Table 6 includes every selected case, including the four where another method has higher Pearson.

For a wider descriptive summary, Table 1 uses 16 directional metrics per view: six distribution scores (W_1, W_2, their Sinkhorn counterparts, marginal W_1, and MMD), four response scores (MAE, MSE, Pearson, and R^2), and six DE scores (AP, AUROC, overlap, precision, direction agreement, and DE-set Spearman). We exclude predicted DE counts, which have no intrinsic better direction, and per-gene error duplicates. The per-case export has no PDS or ES, so these are absent from this summary. Means omit undefined precision or direction-agreement entries without imputation; all other included fields are finite. Finite-case counts accompany the machine-readable summaries. Each metric is ranked among all eight methods using average ranks for exact ties. First-rank frequency counts rank one, and pairwise win frequency counts strict improvements over the seven comparators. This 64-metric, eight-method summary is distinct from the main figure's 80-metric, six-model ranking.

PerturbBridge leads 54 of these 64 aggregate scores, compared with 41 on the full OOD-cell-type evaluation under the same metric and comparator selection. The full-test distribution scores retain their cell-weighted aggregation, whereas subset means weight cases equally. These summaries compare relative performance across evaluation sets with different group sizes, context compositions, and distribution-metric weights.

Centering makes the diagnostic invariant to a common additive shift in PCA space, exposing changes in population structure that a shared translation cannot capture. Cell-type transfer supplies more extreme examples of these changes, and PerturbBridge's relative advantage becomes more pronounced on the selected subset. Together, these observations suggest that nonuniform population responses help explain its stronger benchmark advantage on unseen cell types than on unseen donors. In this regime, learning how baseline states relate to different responses becomes especially relevant. A reference that limits the pairing influence of large displacements while retaining state discrimination provides a mechanism consistent with the observed gains in response correlation and distribution fit.

Table 5: Equal-case means on the 20 selected groups. Bold marks the best unrounded value among all eight methods within each view. r_δ is response Pearson; ρ_{DE} is response Spearman on the real DE set.

Model	$W_2 \downarrow$	$\mathrm{MMD}^2 \downarrow$	$r_\delta \uparrow$	$\rho_{\mathrm{DE}} \uparrow$	AP↑
Raw counts					
PerturbBridge	**104.16**	**0.0239**	**0.7786**	**0.7991**	**0.6420**
CellFlow	109.13	0.0570	-0.0800	-0.0507	0.3187
STATE	123.76	0.0461	0.4402	0.5321	0.5259
DOLORIS	167.39	0.0835	0.2476	0.3421	0.5996
DEPARTURES	2122.22	0.7261	0.3138	0.1211	0.5708
Lingshu-Cell	447.81	0.2293	0.4489	0.5589	0.6223
Control	108.93	0.0568	-0.0519	-0.0477	0.3190
Linear (R1)	114.77	0.4380	0.3107	0.4579	0.4711
Log counts					
PerturbBridge	**40.82**	**0.0131**	**0.7851**	**0.7713**	**0.6222**
CellFlow	41.73	0.0321	-0.0574	-0.0411	0.3204
STATE	43.04	0.0248	0.5433	0.5554	0.5177
DOLORIS	50.62	0.0478	0.1691	0.2835	0.5723
DEPARTURES	156.56	0.8569	0.1836	0.0305	0.5925
Lingshu-Cell	60.72	0.1664	0.4980	0.5318	0.6057
Control	41.75	0.0320	-0.0408	-0.0414	0.3207
Linear (R1)	41.40	0.3791	0.3998	0.4690	0.4928
Normalized counts					
PerturbBridge	86.30	**0.0217**	**0.5976**	**0.7403**	**0.3964**
CellFlow	96.89	0.0242	-0.0064	0.0418	0.1364
STATE	111.39	0.0405	0.3476	0.4616	0.2882
DOLORIS	103.11	0.0511	0.0238	-0.0510	0.2319
DEPARTURES	189.41	0.4210	0.0572	0.1814	0.2567
Lingshu-Cell	120.08	0.1297	0.3415	0.5413	0.3433
Control	96.78	0.0242	0.0119	0.0149	0.1367
Linear (R1)	**80.67**	0.2681	0.2628	0.4256	0.2333
Normalized log counts					
PerturbBridge	36.83	**0.0125**	**0.6527**	**0.7091**	**0.4144**
CellFlow	40.67	0.0137	0.0011	0.0235	0.1589
STATE	39.24	0.0161	0.4476	0.4875	0.3218
DOLORIS	41.65	0.0272	-0.0072	-0.0365	0.2629
DEPARTURES	38.19	0.2779	0.0823	0.0360	0.2941
Lingshu-Cell	39.36	0.0874	0.3814	0.4911	0.3643
Control	40.65	0.0138	-0.0020	0.0053	0.1591
Linear (R1)	**32.42**	0.2420	0.3011	0.4184	0.2470

Table 6: All 20 selected cases and raw-count response Pearson. Best other denotes the strongest of the seven other methods separately for each case. Bold marks the larger value.

Cell type	Cytokine	Donor	PerturbBridge	Best other
B Naive	IL-7	6	**0.6855**	0.5611
B Naive	SCF	4	**0.9293**	0.7888
CD4 Memory	4-1BBL	4	**0.6286**	0.3532
CD4 Memory	IL-15	11	**0.7739**	0.7412
CD4 Memory	IL-1Ra	4	0.9301	**0.9479**
CD4 Memory	IL-36-alpha	5	**0.9254**	0.8873
CD4 Memory	IL-4	6	**0.7260**	0.4366
CD4 Memory	IL-7	11	**0.8423**	0.7664
CD4 Memory	IL-7	5	0.9364	**0.9424**
CD8 Memory	IL-36-alpha	7	**0.8609**	0.7666
CD8 Naive	IL-20	5	**0.9281**	0.5580
CD8 Naive	IL-33	5	**0.9394**	0.5225
CD8 Naive	IL-3	5	**0.9155**	0.5674
CD8 Naive	TSLP	5	**0.9175**	0.6391
NK	IL-10	6	0.3461	**0.6860**
NK	IL-13	6	**0.6976**	0.4165
NK	IL-15	5	**0.8574**	0.7599
NK	IL-31	5	**0.7581**	0.5311
NK	LIF	5	**0.7138**	0.6331
NK	RANKL	6	0.2591	**0.5420**

J Reference sensitivity on the PBMC subset

The sensitivity test partition contains 671,234 cells from B Naive, CD4 Memory, CD8 Memory, CD8 Naive, NK, and Treg, restricted to donors 9 and 11. Its 90 non-control cytokines form $6 \times 2 \times 90 = 1,080$ eligible condition groups, each with at least five cells. PBS cells provide matched context inputs but are excluded as target groups. This restricted test partition is distinct from the 8,545-group main OOD-cell-type evaluation.

For PerturbBridge, a pairing block contains a single perturbation, cell type, and donor. Each mini-batch contains four distinct perturbation blocks of 32 cells each. Source cells are drawn from the training PBS pool of the same cell type and donor. The independent gene-wise random-effects specification is

$$A_g \sim \text{Gamma}(1/h, \text{scale} = \kappa h), \qquad D_g \mid A_g \sim \text{Skellam}(A_g, A_g). \tag{76}$$

Here the exported scale parameter s equals κ. Thus $\mathbb{E}[A_g] = \kappa$, $\text{Var}(A_g) = \kappa^2 h$, and $\text{Var}(D_g) = 2\kappa$. This endpoint-displacement law is the common-activity representation of the reference family derived in Appendix B. Network architecture and optimization settings are held fixed across the scale–shape grid.

The validation pack contains 60 groups: the 20 largest, the 20 smallest with at least five cells, and 20 intermediate-sized groups sampled with seed 42. It contains 29,654 target cells, with distribution evaluation capped at 1,000 per group. Runs share the validation target and PBS packs, and each uses the checkpoint with lowest validation count-space Sinkhorn-W_2. The same checkpoint is then evaluated for every test metric, without metric-wise checkpoint selection. Training excludes all 224,829 held-out validation cells.

Table 7 reports all 25 configurations in a single (κ, h) grid. The seven scale-sweep settings are $h = 0$ with $\kappa \in \{0.5, 1, 2, 8, 16, 32, 64\}$. At $\kappa = 8$, the positive shape settings are $h \in \{0.25, 0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6, 0.7, 1, 1.1, 1.2\}$. Additional settings are $(4, 0.25)$, $(4, 0.5)$, $(12, 0.45)$, $(12, 0.5)$, $(12, 0.55)$, and $(16, 0.5)$. All 12 positive-shape runs at $\kappa = 8$ use the same parameters for coupling and bridge.

The scale sweep at $h = 0$ reaches its best W_2 at $\kappa = 8$, and its best cosine PDS at $\kappa = 16$. Shape adjustment further improves response recovery: $(8, 0.5)$ achieves PDS 0.783, above the maximum of 0.617 in the entire zero-shape sweep. At $\kappa = 12$, all three tested positive shape settings achieve PDS above 0.742. The grid also delineates the scale dependence of these gains: $(4, 0.25)$ and $(4, 0.5)$ have W_2 of 337.73 and 216.79, whereas the evaluated positive-shape settings at $\kappa = 8$ and 12 remain below 89.54. The observed useful region therefore depends on both parameters.

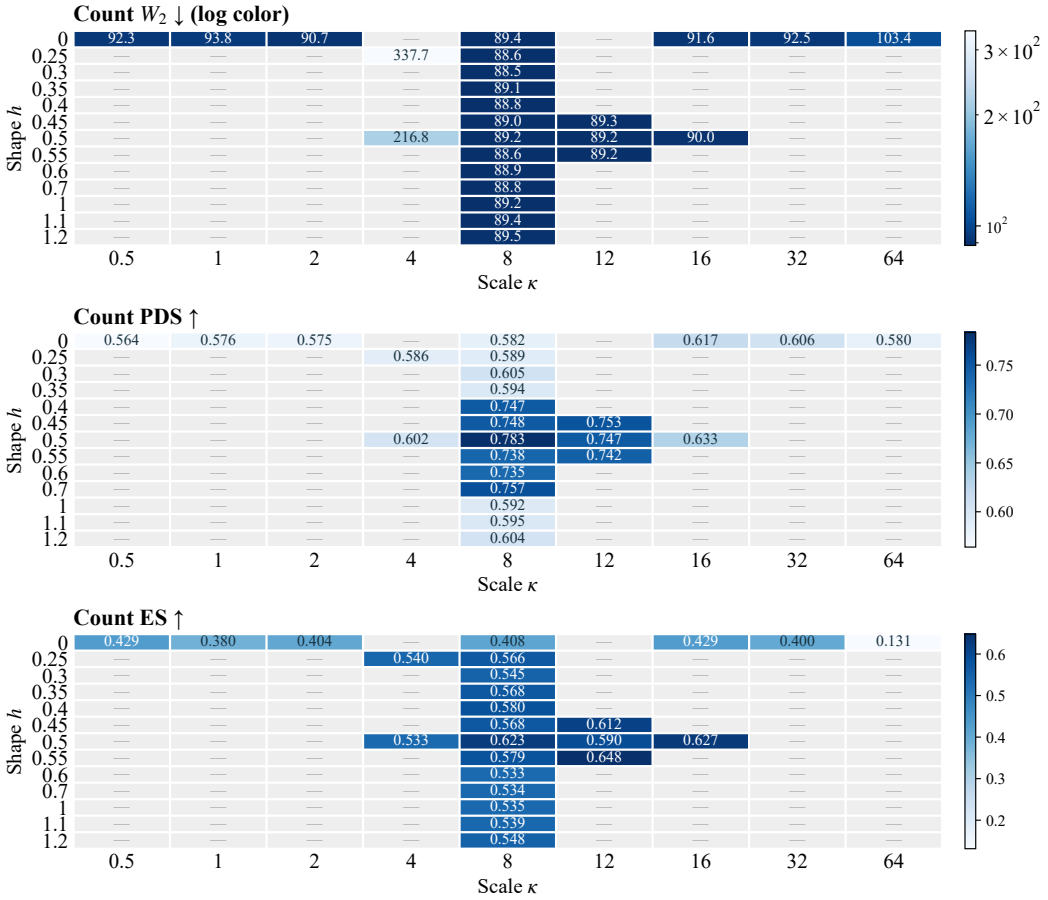

Figure 6: The complete PerturbBridge scale–shape grid, including the $h = 0$ scale sweep. Gray cells are unmeasured configurations; values are not interpolated. Each panel colors its own raw metric, with logarithmic color scaling for W_2 and linear scaling for PDS and ES. Darker blue denotes better scores. All 25 numerical results appear in Table 7.

Table 7: All 25 PerturbBridge scale–shape configurations on the 1,080-group PBMC subset. The scale parameter s equals κ; the seven scale-only runs fix $h = 0$. Each row uses one best checkpoint.

κ	h	$W_2 \downarrow$	Pearson↑	PDS↑	ES↑	AP↑
0.5	0	92.3272	0.1746	0.5639	0.4293	0.1613
1	0	93.7945	0.1866	0.5759	0.3797	0.1567
2	0	90.7430	0.1813	0.5746	0.4039	0.1593
4	0.25	337.7280	0.2244	0.5864	0.5405	0.1706
4	0.5	216.7933	0.2313	0.6023	0.5328	0.1678
8	0	89.4459	0.1785	0.5825	0.4082	0.1460
8	0.25	88.5981	0.2147	0.5890	0.5662	0.1640
8	0.3	88.5247	0.2468	0.6055	0.5455	0.1770
8	0.35	89.0646	0.2274	0.5941	0.5681	0.1642
8	0.4	88.7904	0.3610	0.7471	0.5805	0.1659
8	0.45	89.0429	0.3481	0.7477	0.5679	0.1672
8	0.5	89.1649	0.3949	0.7834	0.6231	0.1803
8	0.55	88.5527	0.3344	0.7381	0.5788	0.1686
8	0.6	88.9223	0.3325	0.7347	0.5326	0.1650
8	0.7	88.8160	0.3401	0.7567	0.5343	0.1674
8	1	89.2356	0.2199	0.5924	0.5350	0.1677
8	1.1	89.4165	0.2224	0.5952	0.5387	0.1668
8	1.2	89.5309	0.2298	0.6041	0.5481	0.1674
12	0.45	89.2501	0.3639	0.7533	0.6122	0.1732
12	0.5	89.1713	0.3633	0.7470	0.5895	0.1692
12	0.55	89.1849	0.3464	0.7425	0.6479	0.1685
16	0	91.5837	0.2125	0.6173	0.4291	0.1438
16	0.5	90.0041	0.2768	0.6331	0.6269	0.1650
32	0	92.5229	0.1991	0.6059	0.3999	0.1441
64	0	103.4335	0.1238	0.5800	0.1314	0.1264

J.1 Coupling and bridge shape ablations

Four runs fix $h_c = 0.5$ for the coupling and set $h_b \in \{0.3, 0.4, 0.6, 0.7\}$ for the bridge. Two further runs fix $h_b = 0.5$ and set $h_c \in \{0.4, 0.6\}$. Both reference scales are $\kappa = 8$, and the Sinkhorn temperature is $\tau = 1$. Each run uses one validation-selected checkpoint for all test metrics. Tables 8–11 report all 18 configurations: 12 matched shapes and six asymmetric settings on the same 1,080 groups. The shared $(0.5, 0.5)$ result is listed once. Asymmetric settings are kept separate from the 25-setting scale–shape grid. Each bridge-shape and component-ablation configuration uses one training run.

The fixed-coupling comparison varies the bridge shape at $h_c = 0.5$. The matching $h_b = 0.5$ bridge achieves lower MMD and higher Pearson, Spearman, cosine PDS, ES, and DE AP than all four alternative bridges across all four expression views. At fixed $h_b = 0.5$, changing h_c to 0.4 or 0.6 also reduces response recovery. Crucially, each alternative bridge gains raw-count Pearson and cosine PDS when its coupling is matched: the same $h_b = 0.7$ bridge reaches PDS 0.7567 with $h_c = 0.7$, versus 0.5867 with $h_c = 0.5$. This comparison separates an ineffective bridge shape from an ineffective coupling–bridge combination. Intermediate-state supervision depends on the bridge as well as its endpoints, making joint reference selection consequential. The benefit is metric dependent: for $h_b = 0.6$, matching improves Pearson and PDS but slightly lowers ES (0.5326 versus 0.5473). All component and bridge comparisons use single training runs, so the reported differences are descriptive rather than estimates of variation across seeds.

Table 8: Coupling–bridge shape comparison in raw counts on 1,080 PBMC groups. Both references use $\kappa = 8$; Sinkhorn uses $\tau = 1$. Blocks: 12 matched shapes, four altered bridges at $h_c = 0.5$, and two altered couplings at $h_b = 0.5$. The shared $(0.5, 0.5)$ run is listed once. PDS is cosine-based; bold marks column-best values.

h_c	h_b	$W_2 \downarrow$	$\text{MMD}^2 \downarrow$	Pearson↑	PDS↑	ES↑	AP↑
0.25	0.25	88.5981	0.0192	0.2147	0.5890	0.5662	0.1640
0.3	0.3	**88.5247**	0.0228	0.2468	0.6055	0.5455	0.1770
0.35	0.35	89.0646	0.0190	0.2274	0.5941	0.5681	0.1642
0.4	0.4	88.7904	0.0100	0.3610	0.7471	0.5805	0.1659
0.45	0.45	89.0429	0.0104	0.3481	0.7477	0.5679	0.1672
0.5	0.5	89.1649	**0.0094**	**0.3949**	**0.7834**	**0.6231**	**0.1803**
0.55	0.55	88.5527	0.0107	0.3344	0.7381	0.5788	0.1686
0.6	0.6	88.9223	0.0111	0.3325	0.7347	0.5326	0.1650
0.7	0.7	88.8160	0.0106	0.3401	0.7567	0.5343	0.1674
1	1	89.2356	0.0213	0.2199	0.5924	0.5350	0.1677
1.1	1.1	89.4165	0.0198	0.2224	0.5952	0.5387	0.1668
1.2	1.2	89.5309	0.0197	0.2298	0.6041	0.5481	0.1674
0.5	0.3	88.7441	0.0219	0.2187	0.5847	0.5408	0.1683
0.5	0.4	88.6676	0.0208	0.2166	0.5854	0.5347	0.1642
0.5	0.6	88.6930	0.0212	0.2176	0.5859	0.5473	0.1665
0.5	0.7	88.7289	0.0217	0.2195	0.5867	0.5300	0.1683
0.4	0.5	188.2113	0.0215	0.2135	0.5828	0.5452	0.1667
0.6	0.5	89.1262	0.0188	0.2248	0.5925	0.5606	0.1639

Table 9: Coupling–bridge shape comparison in log counts on 1,080 PBMC groups. Both references use $\kappa = 8$; Sinkhorn uses $\tau = 1$. Blocks: 12 matched shapes, four altered bridges at $h_c = 0.5$, and two altered couplings at $h_b = 0.5$. The shared $(0.5, 0.5)$ run is listed once. PDS is cosine-based; bold marks column-best values.

h_c	h_b	$W_2 \downarrow$	$\text{MMD}^2 \downarrow$	Pearson↑	PDS↑	ES↑	AP↑
0.25	0.25	38.3066	0.0080	0.2580	0.6447	0.5670	0.1566
0.3	0.3	38.2045	0.0092	0.2748	0.6413	0.5522	0.1693
0.35	0.35	38.4051	0.0078	0.2735	0.6625	0.5642	0.1572
0.4	0.4	38.3112	0.0058	0.3519	0.7556	0.6055	0.1681
0.45	0.45	38.3891	0.0059	0.3487	0.7585	0.5812	0.1692
0.5	0.5	38.3833	**0.0052**	**0.3954**	**0.8004**	**0.6559**	**0.1821**
0.55	0.55	**38.1943**	0.0059	0.3377	0.7502	0.5952	0.1635
0.6	0.6	38.3374	0.0061	0.3403	0.7546	0.5371	0.1627
0.7	0.7	38.2854	0.0058	0.3482	0.7693	0.5643	0.1644
1	1	38.3052	0.0085	0.2635	0.6451	0.5433	0.1601
1.1	1.1	38.3346	0.0080	0.2663	0.6550	0.5438	0.1596
1.2	1.2	38.3957	0.0078	0.2719	0.6623	0.5428	0.1597
0.5	0.3	38.2690	0.0087	0.2641	0.6370	0.5467	0.1607
0.5	0.4	38.2903	0.0081	0.2648	0.6497	0.5446	0.1562
0.5	0.6	38.2785	0.0083	0.2624	0.6424	0.5527	0.1590
0.5	0.7	38.2728	0.0084	0.2630	0.6397	0.5398	0.1604
0.4	0.5	38.3061	0.0085	0.2608	0.6406	0.5503	0.1586
0.6	0.5	38.3694	0.0077	0.2691	0.6531	0.5571	0.1568

Table 10: Coupling–bridge shape comparison in normalized counts on 1,080 PBMC groups. Both references use $\kappa = 8$; Sinkhorn uses $\tau = 1$. Blocks: 12 matched shapes, four altered bridges at $h_c = 0.5$, and two altered couplings at $h_b = 0.5$. The shared $(0.5, 0.5)$ run is listed once. PDS is cosine-based; bold marks column-best values.

h_c	h_b	$W_2 \downarrow$	$\text{MMD}^2 \downarrow$	Pearson↑	PDS↑	ES↑	AP↑
0.25	0.25	77.5497	0.0170	0.1741	0.6082	0.5025	0.0770
0.3	0.3	77.7052	0.0196	0.1642	0.5971	0.4833	0.0801
0.35	0.35	77.4147	0.0178	0.1766	0.6006	0.4747	0.0793
0.4	0.4	77.4014	0.0106	0.2571	0.7275	0.5947	0.0853
0.45	0.45	77.4871	0.0111	0.2530	0.7200	0.5829	0.0872
0.5	0.5	77.4622	**0.0104**	**0.2894**	**0.7712**	**0.6690**	**0.0967**
0.55	0.55	77.7158	0.0109	0.2418	0.7176	0.5929	0.0836
0.6	0.6	77.3827	0.0113	0.2414	0.6874	0.5647	0.0855
0.7	0.7	77.4223	0.0111	0.2487	0.7248	0.6110	0.0852
1	1	77.5908	0.0192	0.1650	0.5869	0.4456	0.0774
1.1	1.1	77.6169	0.0183	0.1628	0.5989	0.4453	0.0767
1.2	1.2	77.3014	0.0185	0.1645	0.5915	0.4509	0.0778
0.5	0.3	77.5918	0.0195	0.1739	0.5841	0.4740	0.0786
0.5	0.4	**77.0477**	0.0188	0.1727	0.5917	0.4519	0.0768
0.5	0.6	77.4224	0.0191	0.1664	0.5850	0.4544	0.0773
0.5	0.7	77.3213	0.0190	0.1682	0.5789	0.4513	0.0783
0.4	0.5	77.3884	0.0189	0.1704	0.5928	0.4651	0.0776
0.6	0.5	77.5308	0.0176	0.1727	0.5931	0.4881	0.0776

Table 11: Coupling–bridge shape comparison in normalized log counts on 1,080 PBMC groups. Both references use $\kappa = 8$; Sinkhorn uses $\tau = 1$. Blocks: 12 matched shapes, four altered bridges at $h_c = 0.5$, and two altered couplings at $h_b = 0.5$. The shared $(0.5, 0.5)$ run is listed once. PDS is cosine-based; bold marks column-best values.

h_c	h_b	$W_2 \downarrow$	$\text{MMD}^2 \downarrow$	Pearson↑	PDS↑	ES↑	AP↑
0.25	0.25	36.6069	0.0070	0.2104	0.6586	0.5189	0.0831
0.3	0.3	36.7284	0.0079	0.2056	0.6447	0.5163	0.0862
0.35	0.35	36.5545	0.0071	0.2177	0.6582	0.4951	0.0855
0.4	0.4	36.6308	0.0053	0.2715	0.7365	0.6240	0.0925
0.45	0.45	36.5583	0.0055	0.2703	0.7343	0.6125	0.0944
0.5	0.5	**36.2855**	**0.0049**	**0.3062**	**0.7814**	**0.6927**	**0.1040**
0.55	0.55	36.6809	0.0053	0.2643	0.7343	0.6199	0.0908
0.6	0.6	36.6034	0.0055	0.2635	0.7154	0.5864	0.0924
0.7	0.7	36.6223	0.0053	0.2693	0.7406	0.6264	0.0920
1	1	36.6476	0.0075	0.2057	0.6358	0.4758	0.0833
1.1	1.1	36.6158	0.0072	0.2050	0.6495	0.4856	0.0826
1.2	1.2	36.5520	0.0071	0.2071	0.6466	0.4850	0.0836
0.5	0.3	36.6780	0.0077	0.2124	0.6314	0.4996	0.0850
0.5	0.4	36.5101	0.0073	0.2120	0.6488	0.4725	0.0827
0.5	0.6	36.6201	0.0074	0.2080	0.6365	0.4895	0.0836
0.5	0.7	36.5945	0.0073	0.2083	0.6269	0.4833	0.0845
0.4	0.5	36.6470	0.0075	0.2088	0.6407	0.4825	0.0836
0.6	0.5	36.5484	0.0070	0.2114	0.6451	0.5058	0.0838

K Control-source and pairing-cost ablations

The ten component variants use the 1,080-group PBMC subset described in Appendix J. The two reference rows use $\kappa = 8$ and set both coupling and bridge shape to $h = 0.5$ or $h = 1$. They are reported alongside six cost variants, random pairing, and the noise-source variant. These component runs are separate from the six-model benchmark ranking. Tables 12, 13, 14, and 15 report six metrics across four expression views, followed by all 108 exported fields.

Pairing-cost definitions. For control cell i and perturbed cell j, let $r_{jig} = y_{j,g} - x_{i,g}$ denote the raw-count residual on gene g. Costs use all $G = 18{,}213$ genes:

$$C_{ji}^{L_1} = \sum_g |r_{jig}|, \qquad C_{ji}^{L_2^2} = \sum_g r_{jig}^2, \qquad C_{ji}^{\text{Euclidean}} = \sqrt{\sum_g r_{jig}^2}, \tag{77}$$

$$C_{ji}^{\text{Huber}} = \sum_g H_1(r_{jig}), \qquad H_\delta(r) = \begin{cases} \frac{1}{2}r^2, & |r| \le \delta, \\ \delta(|r| - \frac{1}{2}\delta), & |r| > \delta. \end{cases} \tag{78}$$

The Huber experiment uses one global threshold $\delta = 1$. It sums gene-wise penalties rather than applying Huber to a cell-level norm. All cost variants solve the coupling with Sinkhorn and generate intermediate states with the $(\kappa, h) = (8, 0.5)$ reference bridge. The default temperature is $\tau = 1$; the two additional temperature runs use L_1 or L_2^2 with $\tau = 10$. Random pairing neither computes a cost nor solves OT. At $h = 1$, the matched reference also has an absolute-value endpoint cost, but its scale and compatible $h = 1$ bridge distinguish it from the L_1 cost substitution.

Poisson-source construction. A subset $\mathcal{S}$ of at most 200,000 cells is sampled from the strict training split with seed 73. For each of the 18,213 genes, define

$$\lambda_g = \max\left(0.01, \frac{1}{|\mathcal{S}|}\sum_{i \in \mathcal{S}} x_{i,g}\right), \qquad X_{j,g}^{\text{source}} \sim \text{Pois}(\lambda_g). \tag{79}$$

Draws are independent across genes and generated cells. The largest rate in this run is 16.92. Training sources are generated dynamically by data-loader workers; validation uses a fixed Poisson-source pack generated with seed 42. Inference also uses Poisson sources with the training-derived rates. Perturbation and context metadata conditioning remain available. The source readouts encode training-average expression, and no separate per-cell control-expression input is supplied.

Interpreting the metric differences. The reference improves response recovery even when squared-Euclidean pairing gives slightly lower W_2. Distribution and response scores also use different aggregation weights (Appendix H), so their task-level disagreement does not directly measure cell-pairing accuracy. The cost substitutions change both the coupling and its relationship to the fixed bridge. The shape comparisons in Appendix J.1 complement this test by examining matched and asymmetric combinations within one reference family. For the tested Huber threshold, integer residuals of magnitude at least one already receive the linear penalty $|r| - 1/2$; this setting does not retain curvature across larger integer displacements.

The noise-source result further distinguishes gene ranking from response structure: raw-count AP is 0.1806, close to the reference's 0.1803, while Pearson, cosine PDS, and ES decline. AP ranks DE genes within each condition, whereas ES ranks treatments by DE breadth. Preserving one ordering therefore need not preserve the other. The Poisson substitution jointly removes observed source-state information and cross-gene dependence, so its effect concerns the observed control source as a whole.

Table 12: Reference-shape, coupling-cost, and source ablations on 1,080 PBMC unseen-cell-type groups, in raw-count space. Reference rows use $\kappa = 8$ and the same h for coupling and bridge. Cost variants use Sinkhorn and retain the $(8, 0.5)$ bridge. Sinkhorn temperature is $\tau = 1$ unless $\tau = 10$ is shown. PDS uses cosine similarity. Bold marks the best value in each column.

Variant	$W_2 \downarrow$	$\text{MMD}^2 \downarrow$	Pearson↑	PDS↑	ES↑	AP↑
Reference $h = 0.5$	89.1649	**0.0094**	**0.3949**	**0.7834**	**0.6231**	0.1803
Reference $h = 1$	89.2356	0.0213	0.2199	0.5924	0.5350	0.1677
L_1	88.7150	0.0216	0.2036	0.5774	0.5341	0.1658
L_2^2	**88.6338**	0.0214	0.2149	0.5815	0.5511	0.1683
L_1 ($\tau = 10$)	89.0107	0.0217	0.2093	0.5812	0.5513	0.1647
L_2^2 ($\tau = 10$)	91.3743	0.0209	0.2248	0.6037	0.4903	0.1612
Huber $\delta = 1$	89.1134	0.0213	0.2199	0.5865	0.5427	0.1658
Random pairing	96.0406	0.0182	0.2403	0.6462	0.4391	0.1502
Euclidean	89.0523	0.0179	0.2360	0.5941	0.5726	0.1648
Noise source	97.0066	0.0804	0.0843	0.5147	0.0210	**0.1806**

Table 13: Component ablations in log-count space. PDS uses cosine similarity.

Variant	$W_2 \downarrow$	$\text{MMD}^2 \downarrow$	Pearson↑	PDS↑	ES↑	AP↑
Reference $h = 0.5$	38.3833	**0.0052**	**0.3954**	**0.8004**	**0.6559**	**0.1821**
Reference $h = 1$	38.3052	0.0085	0.2635	0.6451	0.5433	0.1601
L_1	38.2720	0.0086	0.2484	0.6292	0.5300	0.1573
L_2^2	38.2911	0.0085	0.2644	0.6418	0.5559	0.1601
L_1 ($\tau = 10$)	38.3394	0.0086	0.2553	0.6377	0.5582	0.1567
L_2^2 ($\tau = 10$)	38.8419	0.0088	0.2635	0.6755	0.4761	0.1550
Huber $\delta = 1$	38.3670	0.0084	0.2665	0.6531	0.5468	0.1578
Random pairing	39.1359	0.0082	0.2678	0.7225	0.4128	0.1460
Euclidean	38.3834	0.0076	0.2777	0.6606	0.5779	0.1586
Noise source	**37.4628**	0.0482	0.1062	0.5155	0.0209	0.1710

Table 14: Component ablations in normalized-count space. PDS uses cosine similarity.

Variant	$W_2 \downarrow$	$\text{MMD}^2 \downarrow$	Pearson↑	PDS↑	ES↑	AP↑
Reference $h = 0.5$	77.4622	**0.0104**	**0.2894**	**0.7712**	**0.6690**	**0.0967**
Reference $h = 1$	77.5908	0.0192	0.1650	0.5869	0.4456	0.0774
L_1	77.3747	0.0186	0.1700	0.5909	0.4526	0.0767
L_2^2	77.2590	0.0186	0.1668	0.5962	0.4508	0.0772
L_1 ($\tau = 10$)	77.2509	0.0195	0.1704	0.5903	0.5009	0.0786
L_2^2 ($\tau = 10$)	76.9067	0.0216	0.1649	0.6033	0.4537	0.0815
Huber $\delta = 1$	77.2273	0.0196	0.1692	0.5898	0.4551	0.0773
Random pairing	**76.5289**	0.0195	0.1773	0.6080	0.5448	0.0792
Euclidean	77.4908	0.0166	0.1825	0.5880	0.5105	0.0789
Noise source	89.9744	0.0443	0.0961	0.5196	0.2066	0.0695

Table 15: Component ablations in normalized-log space. PDS uses cosine similarity.

Variant	$W_2 \downarrow$	$\text{MMD}^2 \downarrow$	Pearson↑	PDS↑	ES↑	AP↑
Reference $h = 0.5$	36.2855	**0.0049**	**0.3062**	**0.7814**	**0.6927**	**0.1040**
Reference $h = 1$	36.6476	0.0075	0.2057	0.6358	0.4758	0.0833
L_1	36.6399	0.0074	0.2059	0.6357	0.4787	0.0824
L_2^2	36.6267	0.0075	0.2064	0.6448	0.4738	0.0830
L_1 ($\tau = 10$)	36.5903	0.0076	0.2079	0.6369	0.5206	0.0842
L_2^2 ($\tau = 10$)	36.5469	0.0082	0.2068	0.6497	0.4406	0.0863
Huber $\delta = 1$	36.5594	0.0076	0.2100	0.6472	0.4673	0.0832
Random pairing	**36.1526**	0.0074	0.2202	0.6822	0.5241	0.0838
Euclidean	36.5594	0.0069	0.2182	0.6389	0.5219	0.0853
Noise source	38.7725	0.0231	0.1106	0.5176	0.1184	0.0734

Table 16: Full component-ablation export, variants 1–4. Columns: 1: Reference $h = 0.5$; 2: Reference $h = 1$; 3: L_1; 4: L_2^2.

Metric	1	2	3	4
test_count_auprc	0.180282	0.167659	0.165762	0.1683
test_count_auroc	0.66161	0.656817	0.656019	0.656589
test_count_deover	0.218429	0.197842	0.195225	0.197403
test_count_deprec	0.211806	0.184986	0.186017	0.182608
test_count_diragr	0.936998	0.828793	0.816853	0.826904
test_count_es	0.623142	0.535004	0.534125	0.551065
test_count_hvg_n_genes	2000	2000	2000	2000
test_count_hvg_pds_cosine	0.780628	0.587994	0.574465	0.577716
test_count_hvg_pds_l1	0.811646	0.637366	0.609496	0.623426
test_count_hvg_pds_l2	0.799311	0.610802	0.590545	0.599084
test_count_lfcspear	0.552526	0.392921	0.371811	0.389066
test_count_mae	631.784	717.101	726.498	720.427
test_count_mae_per_gene	0.0346886	0.039373	0.039889	0.0395556
test_count_marginal_w1	0.0331148	0.0366038	0.0371877	0.0370318
test_count_mmd	0.00935446	0.0212545	0.0216254	0.0213541
test_count_mse	364.577	583.215	590.733	574.436
test_count_mse_per_gene	0.0200174	0.0320219	0.0324347	0.0315399
test_count_pdcorr	0.394878	0.219861	0.2036	0.214859
test_count_pds_cosine	0.783364	0.592397	0.577449	0.581481
test_count_pds_l1	0.82607	0.667644	0.641749	0.656821
test_count_pds_l2	0.801852	0.614198	0.593909	0.602819
test_count_pred_de_n	1407.12	1977.13	1950.7	2038.03
test_count_r2	-2.86062	-5.71237	-5.67359	-5.62957
test_count_sinkhorn_w1	82.4801	82.3041	82.0254	81.9841
test_count_sinkhorn_w2	62.8047	62.8454	62.4901	62.4284
test_count_strict_w1	82.3377	82.1893	81.9243	81.8772
test_count_strict_w2	89.1649	89.2356	88.715	88.6338
test_count_true_de_n	1213.29	1213.29	1213.29	1213.29
test_log1p_auprc	0.182122	0.160081	0.157347	0.160117
test_log1p_auroc	0.664077	0.651359	0.65	0.650497
test_log1p_deover	0.219842	0.19299	0.189867	0.191822
test_log1p_deprec	0.206224	0.180744	0.181515	0.178535
test_log1p_diragr	0.92172	0.824048	0.813118	0.822825
test_log1p_es	0.655916	0.543287	0.530002	0.555901
test_log1p_hvg_n_genes	2000	2000	2000	2000
test_log1p_hvg_pds_cosine	0.775792	0.631677	0.616101	0.627541
test_log1p_hvg_pds_l1	0.817068	0.676574	0.657284	0.673374
test_log1p_hvg_pds_l2	0.807562	0.670905	0.654362	0.666409
test_log1p_lfcspear	0.517649	0.375605	0.356355	0.374804
test_log1p_mae	284.853	316.183	317.704	315.62
test_log1p_mae_per_gene	0.0156401	0.0173603	0.0174438	0.0173294
test_log1p_marginal_w1	0.0142798	0.0156037	0.0157241	0.0156902
test_log1p_mmd	0.00516301	0.0084621	0.00855899	0.0084573
test_log1p_mse	26.8579	37.8197	38.2729	37.4994
test_log1p_mse_per_gene	0.00147466	0.00207652	0.00210141	0.00205894
test_log1p_pdcorr	0.395424	0.263548	0.248417	0.264386
test_log1p_pds_cosine	0.80037	0.645123	0.629156	0.64177
test_log1p_pds_l1	0.853066	0.712634	0.692243	0.711019
test_log1p_pds_l2	0.829918	0.685658	0.667901	0.681039
test_log1p_pred_de_n	1565.43	1921.26	1866.88	1958.39
test_log1p_r2	-1.66722	-3.01031	-2.99447	-2.99092
test_log1p_sinkhorn_w1	37.7567	37.6797	37.6714	37.6871
test_log1p_sinkhorn_w2	27.1134	27.0562	27.0331	27.0461
test_log1p_strict_w1	37.4851	37.4101	37.4024	37.4182
test_log1p_strict_w2	38.3833	38.3052	38.272	38.2911
test_log1p_true_de_n	1188.1	1188.1	1188.1	1188.1
test_mmd	0.00935446	0.0212545	0.0216254	0.0213541
test_n_groups	1080	1080	1080	1080
test_norm_count_auprc	0.0966678	0.0774459	0.0766741	0.0772205
test_norm_count_auroc	0.650955	0.631334	0.629579	0.629268
test_norm_count_deover	0.133897	0.106993	0.106022	0.106199
test_norm_count_deprec	0.103945	0.0853138	0.0850077	0.0841631
test_norm_count_diragr	0.940381	0.799821	0.80369	0.800098
test_norm_count_es	0.669029	0.445553	0.452601	0.450849
test_norm_count_lfcspear	0.542664	0.323531	0.324437	0.323154
test_norm_count_mae	498.522	574.233	576.485	573.858
test_norm_count_mae_per_gene	0.0273718	0.0315287	0.0316524	0.0315082
test_norm_count_marginal_w1	0.0271647	0.0296506	0.0298055	0.0297723
test_norm_count_mmd	0.0104295	0.0191823	0.01855	0.0186308
test_norm_count_mse	201.399	344.867	338.945	329.931
test_norm_count_mse_per_gene	0.011058	0.0189352	0.0186101	0.0181151
test_norm_count_pdcorr	0.289412	0.164998	0.169991	0.166806
test_norm_count_pds_cosine	0.771173	0.586924	0.590854	0.596183
test_norm_count_pds_l1	0.844198	0.707675	0.705226	0.709311
test_norm_count_pds_l2	0.814352	0.657243	0.663086	0.660823
test_norm_count_pred_de_n	1016.92	1304.2	1289.74	1303.97

Table 16: Full component-ablation export, variants 1–4. Columns: 1: Reference $h = 0.5$; 2: Reference $h = 1$; 3: L_1; 4: L_2^2.

Metric	1	2	3	4
test_norm_count_r2	-5.23303	-10.0509	-9.44343	-9.57237
test_norm_count_sinkhorn_w1	76.2702	76.4696	76.2836	76.1686
test_norm_count_sinkhorn_w2	54.7167	54.8113	54.6561	54.5771
test_norm_count_strict_w1	76.0116	76.2098	76.0246	75.9068
test_norm_count_strict_w2	77.4622	77.5908	77.3747	77.259
test_norm_count_true_de_n	417.773	417.773	417.773	417.773
test_norm_log1p_auprc	0.104029	0.0832554	0.0824132	0.0830061
test_norm_log1p_auroc	0.654097	0.629588	0.631177	0.629678
test_norm_log1p_deover	0.139963	0.113951	0.113468	0.113485
test_norm_log1p_deprec	0.109976	0.0934578	0.0934931	0.0924123
test_norm_log1p_diragr	0.926372	0.809697	0.809619	0.808713
test_norm_log1p_es	0.692726	0.475777	0.478677	0.473799
test_norm_log1p_lfcspear	0.534049	0.338219	0.331861	0.333261
test_norm_log1p_mae	248.904	275.87	276.61	276.002
test_norm_log1p_mae_per_gene	0.0136663	0.0151468	0.0151875	0.0151541
test_norm_log1p_marginal_w1	0.0126276	0.0139851	0.0140652	0.0140685
test_norm_log1p_mmd	0.00492728	0.00749409	0.00737193	0.00745856
test_norm_log1p_mse	22.4902	30.7429	30.82	30.4778
test_norm_log1p_mse_per_gene	0.00123484	0.00168797	0.0016922	0.00167341
test_norm_log1p_pdcorr	0.306228	0.205664	0.205921	0.206444
test_norm_log1p_pds_cosine	0.781399	0.635802	0.63573	0.644763
test_norm_log1p_pds_l1	0.836132	0.72713	0.725669	0.730741
test_norm_log1p_pds_l2	0.822078	0.706996	0.705545	0.711409
test_norm_log1p_pred_de_n	1213.4	1424.45	1403.99	1425.31
test_norm_log1p_r2	-2.93748	-4.55953	-4.40919	-4.47585
test_norm_log1p_sinkhorn_w1	36.2608	36.6495	36.6523	36.6383
test_norm_log1p_sinkhorn_w2	25.6497	25.9059	25.9005	25.8912
test_norm_log1p_strict_w1	35.9803	36.3697	36.3727	36.3587
test_norm_log1p_strict_w2	36.2855	36.6476	36.6399	36.6267
test_norm_log1p_true_de_n	479.547	479.547	479.547	479.547
test_sinkhorn	62.8047	62.8454	62.4901	62.4284
test_w1	82.4801	82.3041	82.0254	81.9841

Table 17: Full component-ablation export, variants 5–8. Columns: 1: L_1 ($\tau = 10$); 2: L_2^2 ($\tau = 10$); 3: Huber $\delta = 1$; 4: Random pairing.

Metric	1	2	3	4
test_count_auprc	0.164726	0.161214	0.165836	0.15022
test_count_auroc	0.655123	0.64814	0.655345	0.641438
test_count_deover	0.194733	0.191286	0.196621	0.184584
test_count_deprec	0.185774	0.174448	0.183725	0.179848
test_count_diragr	0.826325	0.81601	0.836075	0.848778
test_count_es	0.551255	0.49033	0.542689	0.439077
test_count_hvg_n_genes	2000	2000	2000	2000
test_count_hvg_pds_cosine	0.577912	0.59894	0.583282	0.639208
test_count_hvg_pds_l1	0.617562	0.68	0.633405	0.71251
test_count_hvg_pds_l2	0.595967	0.623899	0.604681	0.669753
test_count_lfcspear	0.384182	0.365315	0.40028	0.382309
test_count_mae	720.336	748.81	716.804	707.257
test_count_mae_per_gene	0.0395506	0.0411141	0.0393567	0.0388326
test_count_marginal_w1	0.0369592	0.0381661	0.0368389	0.0368912
test_count_mmd	0.0217484	0.0209033	0.0212965	0.0182295
test_count_mse	592.35	603.833	583.886	700.766
test_count_mse_per_gene	0.0325235	0.0331539	0.0320587	0.0384761
test_count_pdcorr	0.209334	0.224777	0.2199	0.240346
test_count_pds_cosine	0.581193	0.603663	0.586502	0.646173
test_count_pds_l1	0.648755	0.716512	0.668632	0.717675
test_count_pds_l2	0.599002	0.629146	0.608385	0.674712
test_count_pred_de_n	1924.33	2050.52	1975.33	1677.72
test_count_r2	-5.71096	-5.74555	-5.653	-4.95437
test_count_sinkhorn_w1	82.2928	83.8411	82.412	85.7374
test_count_sinkhorn_w2	62.6876	64.2967	62.7825	67.5785
test_count_strict_w1	82.1859	83.7612	82.2915	85.6008
test_count_strict_w2	89.0107	91.3743	89.1134	96.0406
test_count_true_de_n	1213.29	1213.29	1213.29	1213.29
test_log1p_auprc	0.156719	0.154974	0.157833	0.145984
test_log1p_auroc	0.649455	0.643219	0.649136	0.6394
test_log1p_deover	0.189933	0.18661	0.191426	0.180936
test_log1p_deprec	0.18112	0.168164	0.178871	0.172861
test_log1p_diragr	0.822745	0.810605	0.831669	0.839221
test_log1p_es	0.558238	0.476142	0.546828	0.412842
test_log1p_hvg_n_genes	2000	2000	2000	2000
test_log1p_hvg_pds_cosine	0.62463	0.656327	0.636852	0.713158

Table 17: Full component-ablation export, variants 5–8. Columns: 1: L_1 ($\tau = 10$); 2: L_2^2 ($\tau = 10$); 3: Huber $\delta = 1$; 4: Random pairing.

Metric	1	2	3	4
test_log1p_hvg_pds_l1	0.66536	0.720494	0.686564	0.733827
test_log1p_hvg_pds_l2	0.66284	0.702706	0.677963	0.740041
test_log1p_lfcspear	0.367476	0.346941	0.383	0.367354
test_log1p_mae	315.405	334.266	314.744	321.408
test_log1p_mae_per_gene	0.0173176	0.0183532	0.0172813	0.0176472
test_log1p_marginal_w1	0.0156686	0.0164646	0.0156403	0.0161102
test_log1p_mmd	0.00856236	0.00875486	0.00839556	0.00820344
test_log1p_mse	37.9757	40.3808	37.6425	37.3655
test_log1p_mse_per_gene	0.00208509	0.00221714	0.00206679	0.00205159
test_log1p_pdcorr	0.25529	0.263507	0.266512	0.267819
test_log1p_pds_cosine	0.637685	0.675504	0.653117	0.722469
test_log1p_pds_l1	0.703097	0.747171	0.722078	0.726296
test_log1p_pds_l2	0.676523	0.722243	0.694095	0.741255
test_log1p_pred_de_n	1850.45	2077.45	1912.03	1759.89
test_log1p_r2	-2.99127	-3.11929	-2.9766	-2.72923
test_log1p_sinkhorn_w1	37.7416	38.1792	37.7714	38.5121
test_log1p_sinkhorn_w2	27.0797	27.4224	27.1019	27.6389
test_log1p_strict_w1	37.4722	37.9174	37.5004	38.241
test_log1p_strict_w2	38.3394	38.8419	38.367	39.1359
test_log1p_true_de_n	1188.1	1188.1	1188.1	1188.1
test_mmd	0.0217484	0.0209033	0.0212965	0.0182295
test_n_groups	1080	1080	1080	1080
test_norm_count_auprc	0.0785624	0.0814781	0.0772935	0.0792012
test_norm_count_auroc	0.630565	0.631706	0.629107	0.636441
test_norm_count_deover	0.108246	0.11162	0.106992	0.109612
test_norm_count_deprec	0.0858162	0.0799262	0.0836629	0.0883974
test_norm_count_diragr	0.808625	0.790567	0.812641	0.848158
test_norm_count_es	0.500945	0.453655	0.455082	0.544756
test_norm_count_lfcspear	0.328001	0.308431	0.347285	0.397311
test_norm_count_mae	573.697	593.209	577.076	552.037
test_norm_count_mae_per_gene	0.0314993	0.0325706	0.0316848	0.0303101
test_norm_count_marginal_w1	0.0298149	0.0308356	0.0298363	0.0292043
test_norm_count_mmd	0.0194874	0.0216213	0.0195589	0.0194748
test_norm_count_mse	346.224	368.623	350.331	333.53
test_norm_count_mse_per_gene	0.0190097	0.0202395	0.0192352	0.0183127
test_norm_count_pdcorr	0.170369	0.164879	0.169208	0.17727
test_norm_count_pds_cosine	0.59035	0.603313	0.589805	0.608045
test_norm_count_pds_l1	0.713786	0.715525	0.720525	0.733981
test_norm_count_pds_l2	0.660628	0.663302	0.666656	0.675226
test_norm_count_pred_de_n	1298.56	1463.07	1330.42	1263.62
test_norm_count_r2	-9.94545	-10.5856	-10.0051	-9.71076
test_norm_count_sinkhorn_w1	76.1835	75.8975	76.1409	75.4038
test_norm_count_sinkhorn_w2	54.5717	54.3296	54.5542	54.0537
test_norm_count_strict_w1	75.9221	75.6361	75.8804	75.1477
test_norm_count_strict_w2	77.2509	76.9067	77.2273	76.5289
test_norm_count_true_de_n	417.773	417.773	417.773	417.773
test_norm_log1p_auprc	0.0842277	0.0863405	0.0831908	0.0837814
test_norm_log1p_auroc	0.62887	0.629318	0.627931	0.632384
test_norm_log1p_deover	0.115111	0.117266	0.114645	0.114435
test_norm_log1p_deprec	0.0941735	0.0865329	0.0917967	0.0945023
test_norm_log1p_diragr	0.815141	0.789799	0.821194	0.840812
test_norm_log1p_es	0.520597	0.440637	0.467276	0.524139
test_norm_log1p_lfcspear	0.333516	0.310528	0.354188	0.380833
test_norm_log1p_mae	275.593	287.232	276.909	269.103
test_norm_log1p_mae_per_gene	0.0151317	0.0157707	0.0152039	0.0147753
test_norm_log1p_marginal_w1	0.0140707	0.01467	0.0140781	0.0139444
test_norm_log1p_mmd	0.00758844	0.00823019	0.00762312	0.00742434
test_norm_log1p_mse	31.0366	32.9967	31.2822	29.8694
test_norm_log1p_mse_per_gene	0.00170409	0.00181171	0.00171757	0.00164
test_norm_log1p_pdcorr	0.207877	0.206773	0.209974	0.220228
test_norm_log1p_pds_cosine	0.636883	0.649681	0.647191	0.682191
test_norm_log1p_pds_l1	0.734835	0.74536	0.743097	0.775761
test_norm_log1p_pds_l2	0.71393	0.71537	0.72213	0.749743
test_norm_log1p_pred_de_n	1413.39	1613.54	1452.32	1399.91
test_norm_log1p_r2	-4.54655	-4.83252	-4.59646	-4.32422
test_norm_log1p_sinkhorn_w1	36.6057	36.5575	36.5733	36.1546
test_norm_log1p_sinkhorn_w2	25.8655	25.8348	25.8436	25.556
test_norm_log1p_strict_w1	36.3262	36.2784	36.2937	35.875
test_norm_log1p_strict_w2	36.5903	36.5469	36.5594	36.1526
test_norm_log1p_true_de_n	479.547	479.547	479.547	479.547
test_sinkhorn	62.6876	64.2967	62.7825	67.5785
test_w1	82.2928	83.8411	82.412	85.7374

Table 18: Full component-ablation export, variants 9–10. Columns: 1: Euclidean; 2: Noise source.

Metric	1	2
test_count_auprc	0.164769	0.180584
test_count_auroc	0.652821	0.694175
test_count_deover	0.195739	0.200134
test_count_deprec	0.187803	0.121277
test_count_diragr	0.845783	0.716379
test_count_es	0.57262	0.0209571
test_count_hvg_n_genes	2000	2000
test_count_hvg_pds_cosine	0.590031	0.514907
test_count_hvg_pds_l1	0.648745	0.513745
test_count_hvg_pds_l2	0.613817	0.519177
test_count_lfcspear	0.402839	0.286017
test_count_mae	689.773	1374.08
test_count_mae_per_gene	0.0378725	0.0754448
test_count_marginal_w1	0.0355343	0.0689901
test_count_mmd	0.0178678	0.0804033
test_count_mse	503.789	1673.7
test_count_mse_per_gene	0.027661	0.0918961
test_count_pdcorr	0.235985	0.0843242
test_count_pds_cosine	0.594115	0.514691
test_count_pds_l1	0.685617	0.512418
test_count_pds_l2	0.618045	0.518786
test_count_pred_de_n	1812.8	6790.33
test_count_r2	-4.70593	-19.6018
test_count_sinkhorn_w1	82.3411	83.4633
test_count_sinkhorn_w2	62.716	68.314
test_count_strict_w1	82.211	83.4802
test_count_strict_w2	89.0523	97.0066
test_count_true_de_n	1213.29	1213.29
test_log1p_auprc	0.15856	0.171049
test_log1p_auroc	0.647708	0.688432
test_log1p_deover	0.192501	0.193151
test_log1p_deprec	0.18383	0.118377
test_log1p_diragr	0.838633	0.700715
test_log1p_es	0.577862	0.0209264
test_log1p_hvg_n_genes	2000	2000
test_log1p_hvg_pds_cosine	0.64179	0.516245
test_log1p_hvg_pds_l1	0.700247	0.515278
test_log1p_hvg_pds_l2	0.689486	0.520381
test_log1p_lfcspear	0.386254	0.252573
test_log1p_mae	306.458	616.12
test_log1p_mae_per_gene	0.0168264	0.0338286
test_log1p_marginal_w1	0.0152322	0.0307849
test_log1p_mmd	0.00755776	0.0482442
test_log1p_mse	34.4928	142.418
test_log1p_mse_per_gene	0.00189386	0.00781956
test_log1p_pdcorr	0.27772	0.106169
test_log1p_pds_cosine	0.660586	0.515504
test_log1p_pds_l1	0.734506	0.512593
test_log1p_pds_l2	0.705381	0.519198
test_log1p_pred_de_n	1779.5	6662.01
test_log1p_r2	-2.59716	-16.6586
test_log1p_sinkhorn_w1	37.7481	36.749
test_log1p_sinkhorn_w2	27.1116	26.4509
test_log1p_strict_w1	37.477	36.5023
test_log1p_strict_w2	38.3834	37.4628
test_log1p_true_de_n	1188.1	1188.1
test_mmd	0.0178678	0.0804033
test_n_groups	1080	1080
test_norm_count_auprc	0.0788647	0.0694969
test_norm_count_auroc	0.63035	0.64215
test_norm_count_deover	0.109345	0.0980461
test_norm_count_deprec	0.0872139	0.0643473
test_norm_count_diragr	0.818396	0.711671
test_norm_count_es	0.510492	0.206567
test_norm_count_lfcspear	0.346831	0.202773
test_norm_count_mae	558.255	870.335
test_norm_count_mae_per_gene	0.0306515	0.0477865
test_norm_count_marginal_w1	0.0289438	0.054241
test_norm_count_mmd	0.0165554	0.0443119
test_norm_count_mse	299.138	859.92
test_norm_count_mse_per_gene	0.0164244	0.0472146
test_norm_count_pdcorr	0.182539	0.0960979
test_norm_count_pds_cosine	0.587973	0.519578
test_norm_count_pds_l1	0.714496	0.562706
test_norm_count_pds_l2	0.669156	0.550782
test_norm_count_pred_de_n	1246.67	2420.01
test_norm_count_r2	-8.5138	-28.3671

Table 18: Full component-ablation export, variants 9–10. Columns: 1: Euclidean; 2: Noise source.

Metric	1	2
test_norm_count_sinkhorn_w1	76.2977	88.5824
test_norm_count_sinkhorn_w2	54.7377	63.5568
test_norm_count_strict_w1	76.0388	88.3452
test_norm_count_strict_w2	77.4908	89.9744
test_norm_count_true_de_n	417.773	417.773
test_norm_log1p_auprc	0.0852775	0.0733941
test_norm_log1p_auroc	0.629563	0.643183
test_norm_log1p_deover	0.116924	0.0995706
test_norm_log1p_deprec	0.0959818	0.0654451
test_norm_log1p_diragr	0.82426	0.734443
test_norm_log1p_es	0.521902	0.11845
test_norm_log1p_lfcspear	0.355037	0.248684
test_norm_log1p_mae	269.18	413.978
test_norm_log1p_mae_per_gene	0.0147795	0.0227298
test_norm_log1p_marginal_w1	0.0136511	0.0251664
test_norm_log1p_mmd	0.00687516	0.0230676
test_norm_log1p_mse	28.5274	80.8651
test_norm_log1p_mse_per_gene	0.00156632	0.00443997
test_norm_log1p_pdcorr	0.218233	0.110627
test_norm_log1p_pds_cosine	0.638879	0.517644
test_norm_log1p_pds_l1	0.739619	0.5418
test_norm_log1p_pds_l2	0.716183	0.546728
test_norm_log1p_pred_de_n	1364.03	3150.83
test_norm_log1p_r2	-4.09476	-15.6038
test_norm_log1p_sinkhorn_w1	36.5552	38.8689
test_norm_log1p_sinkhorn_w2	25.8436	27.4055
test_norm_log1p_strict_w1	36.2753	38.5916
test_norm_log1p_strict_w2	36.5594	38.7725
test_norm_log1p_true_de_n	479.547	479.547
test_sinkhorn	62.716	68.314
test_w1	82.3411	83.4633

L Grouped scores and six-model ranking

The heatmap uses 20 directed metrics per expression view, giving 80 metrics across four views. The distribution family contains exact W_1, exact W_2, Sinkhorn $p = 1$ and $p = 2$, marginal W_1, and squared MMD. The pseudobulk family contains R^2, response Pearson, MAE per gene, and MSE per gene. The DE family contains AUROC, AP, overlap, precision, direction agreement, and response Spearman. The perturbation-level family contains three PDS distances and ES. Family sizes are therefore 6, 4, 6, and 4. All eight methods have finite values for this common set on every task.

Compatibility aliases and summed MAE/MSE are excluded to avoid counting duplicate scalings. Descriptive counts have no universal better direction and are excluded. The main summaries use 20 metrics shared across the four expression views, yielding 80 metrics in total. HVG-PDS scores, computed only in the count and log-count views, are reported separately as supplementary results. The same grouping and eligibility rules apply to all tasks and views.

Heatmap cell numbers. Let a_{tvjm} denote a metric value after reversing the sign of lower-is-better metrics. For the set $\mathcal{M}$ of all eight displayed methods, define

$$z_{tvjm} = \frac{a_{tvjm} - \min_{m' \in \mathcal{M}} a_{tvjm'}}{\max_{m' \in \mathcal{M}} a_{tvjm'} - \min_{m' \in \mathcal{M}} a_{tvjm'}}, \qquad S_{tvfm} = \frac{1}{|\mathcal{J}_f|} \sum_{j \in \mathcal{J}_f} z_{tvjm}. \tag{80}$$

If all methods tie on a metric, each receives 0.5. All 80 eligible metrics enter the family means. The cell numbers show S_{tvfm} rounded to two decimal places. The contribution table retains raw values and per-metric normalized scores; full raw results appear in Appendix M.

Rank-based colors. Within each task, view, and family, rank all eight methods by unrounded S_{tvfm}, with larger scores better and average ranks for exact ties. Control and Linear (R1) participate in this ranking. The shared color scale maps rank one to the darkest blue and rank eight to the lightest blue. Bold numbers identify the highest family score among all eight methods. Colors represent relative positions, whereas cell numbers summarize min–max scores; color differences do not represent effect sizes. The color table records every method's family rank.

Mean ranks beside model names. Rank each metric among PerturbBridge, CellFlow, STATE, DOLORIS, DEPARTURES, and Lingshu-Cell only, assigning one to the best and average ranks to exact ties. If $r^{(6)}_{tvjm}$ denotes this rank, the annotation is

$$R^{(6)}_{tm} = \frac{1}{80} \sum_{v=1}^{4} \sum_{j=1}^{20} r^{(6)}_{tvjm}. \tag{81}$$

Control and Linear (R1) are excluded from this computation and receive no rank annotation. The mean ranks beside names are computed from the original metric values before normalization and family averaging. Cell colors instead rank family-averaged min–max scores among all eight methods. Thus the color ranks and the parenthetic mean ranks use different comparison sets and aggregation rules. PerturbBridge ranks first in all three PBMC settings and second on NADIG under this six-model rule. Mean ranks provide a descriptive summary across metrics. Machine-readable family scores, per-metric contributions, and six-model ranks accompany the figure.

M Full multi-view comparisons and numerical results

Figure 3 combines all four tasks and all four expression spaces in a single family-level heatmap. The following tables retain every non-alias field, including metadata, for all eight methods. Values are rounded to four significant digits for readability; accompanying machine-readable tables retain source precision. Best-value formatting uses unrounded values. Individual scores complement the grouped scores and six-model rank annotations defined in Appendix L.

Table 19: NADIG (ID): raw counts. Raw evaluator values; best directed scores in bold. PB: PerturbBridge; CF: CellFlow; DOL: DOLORIS; DEP: DEPARTURES; LC: Lingshu-Cell; Ctrl: Control; Lin: Linear (R1). Counts are descriptive.

Metric	PB	CF	STATE	DOL	DEP	LC	Ctrl	Lin
Exact $W_1 \downarrow$	**301**	344.1	312.2	338.6	327.7	468	330.9	317.4
Exact $W_2 \downarrow$	**358**	435.8	369.1	404.5	393.9	596.6	401	384.1
Sinkhorn $p = 1 \downarrow$	**296**	336.5	307.2	332.5	321.4	465.9	324	311.1
Sinkhorn $p = 2$ (sqrt) $\downarrow$	**247.7**	299.4	254.5	278.9	271.5	419.8	275.7	264
Marginal $W_1 \downarrow$	**0.5481**	0.6584	0.6155	0.773	0.7797	0.7024	0.6302	0.5985
$\text{MMD}^2 \downarrow$	0.02676	0.03065	**0.02165**	0.03839	0.04204	0.07184	0.03455	0.02873
Response $R^2 \uparrow$	**-1.059**	-2.198	-1.363	-1.827	-1.973	-25.89	-1.448	-1.677
Response Pearson $\uparrow$	0.282	0.09916	0.3999	0.1463	0.1953	**0.4**	0.01823	0.1926
L_1 error (sum) $\downarrow$	**4196**	5007	4424	5350	5606	5226	4777	4573
L_1 error / gene $\downarrow$	**0.5498**	0.656	0.5797	0.7011	0.7346	0.6847	0.626	0.5992
Squared error (sum) $\downarrow$	5.50e4	8.22e4	6.28e4	6.86e4	6.70e4	2.80e5	6.86e4	6.69e4
Squared error / gene $\downarrow$	**7.213**	10.78	8.232	8.987	8.775	36.75	8.987	8.768
DE AUROC $\uparrow$	0.532	0.5127	**0.5674**	0.5261	0.4956	0.5439	0.5133	0.5203
DE average precision $\uparrow$	0.1195	0.1027	0.1321	0.1047	0.1032	**0.1338**	0.102	0.1102
DE overlap $\uparrow$	0.1549	0.1375	**0.173**	0.1333	0.1286	0.1693	0.1371	0.1428
DE precision $\uparrow$	0.1555	0.1528	0.1322	0.1264	0.1054	0.1427	0.1513	**0.1685**
Direction agreement $\uparrow$	**0.9983**	0.9969	0.9982	0.9907	0.8921	0.9964	0.997	0.9972
DE response Spearman $\uparrow$	0.4105	0.06451	0.4463	0.06987	0.4154	**0.4962**	0.05684	0.1661
Predicted DE genes	1051	419.5	1831	763.9	1897	1695	411.6	507.8
PDS cosine $\uparrow$	0.5481	0.5025	0.6734	0.5052	0.5035	**0.6751**	0.4968	0.5434
PDS $L_1 \uparrow$	0.6011	0.5148	**0.6834**	0.5139	0.5043	0.6777	0.514	0.582
PDS squared $L_2 \uparrow$	0.5914	0.5073	**0.7057**	0.5069	0.5059	0.6998	0.5064	0.5759
ES $\uparrow$	0.4873	0.3712	**0.4957**	0.283	-0.1561	0.3328	0.3778	0.4853
HVG PDS cosine $\uparrow$	0.548	0.5026	0.6738	0.5053	0.5034	**0.6758**	0.4969	0.5431
HVG PDS $L_1 \uparrow$	0.6007	0.5105	**0.6856**	0.5112	0.5004	0.6827	0.5096	0.5831
HVG PDS squared $L_2 \uparrow$	0.5914	0.5072	**0.7055**	0.5068	0.5057	0.6998	0.5062	0.5759
True DE genes	699.8	699.8	699.8	699.8	699.8	699.8	699.8	699.8
HVG genes	2000	2000	2000	2000	2000	2000	2000	2000

Table 20: NADIG (ID): log counts. Raw evaluator values; best directed scores in bold. PB: PerturbBridge; CF: CellFlow; DOL: DOLORIS; DEP: DEPARTURES; LC: Lingshu-Cell; Ctrl: Control; Lin: Linear (R1). Counts are descriptive.

Metric	PB	CF	STATE	DOL	DEP	LC	Ctrl	Lin
Exact $W_1 \downarrow$	50.52	51.75	**48.83**	56.68	55.71	54.12	51.62	51.23
Exact $W_2 \downarrow$	50.88	52.22	**49.14**	57.26	56.44	54.6	52.01	51.62
Sinkhorn $p = 1 \downarrow$	50.6	51.81	**48.91**	56.76	55.79	54.25	51.68	51.3
Sinkhorn $p = 2$ (sqrt) $\downarrow$	35.87	36.78	**34.65**	40.38	39.83	38.59	36.65	36.37
Marginal $W_1 \downarrow$	**0.1378**	0.1507	0.1644	0.1933	0.2301	0.1528	0.1494	0.1463
$\text{MMD}^2 \downarrow$	**0.006611**	0.01064	0.009312	0.01534	0.05361	0.0227	0.01149	0.008177
Response $R^2 \uparrow$	-0.9939	-1.248	**-0.9069**	-1.699	-3.997	-1.524	-1.215	-1.124
Response Pearson $\uparrow$	0.1565	0.02658	0.2249	0.04518	0.1295	**0.2515**	0.001218	0.07725
L_1 error (sum) $\downarrow$	1107	1194	**1104**	1279	1618	1210	1187	1158
L_1 error / gene $\downarrow$	0.145	0.1564	**0.1446**	0.1676	0.212	0.1586	0.1555	0.1518
Squared error (sum) $\downarrow$	310.8	362.6	**307.7**	412.1	618.6	379.8	358.6	337
Squared error / gene $\downarrow$	0.04073	0.04751	**0.04032**	0.05399	0.08105	0.04976	0.04699	0.04415
DE AUROC $\uparrow$	0.5339	0.5193	**0.5761**	0.5251	0.4852	0.5406	0.5199	0.5237
DE average precision $\uparrow$	0.09521	0.08143	**0.1106**	0.08325	0.0783	0.1058	0.08101	0.08685
DE overlap $\uparrow$	0.1378	0.1185	**0.1612**	0.1207	0.111	0.1447	0.1173	0.1247
DE precision $\uparrow$	0.1541	0.1527	0.1322	0.1267	0.09155	0.1324	0.1517	**0.1632**
Direction agreement $\uparrow$	**0.9973**	0.9951	0.9952	0.9874	0.7585	0.9933	0.9943	0.9955
DE response Spearman $\uparrow$	0.2565	0.04183	0.3008	0.1116	0.2294	**0.356**	0.01837	0.09181
Predicted DE genes	695.5	334.8	1040	550.6	1584	1238	333.5	372.8
PDS cosine $\uparrow$	0.574	0.5013	**0.6905**	0.5005	0.5026	0.6556	0.4974	0.5517
PDS $L_1 \uparrow$	0.6088	0.5163	**0.6764**	0.5106	0.5007	0.6723	0.5155	0.5581
PDS squared $L_2 \uparrow$	0.6121	0.5109	0.6921	0.5088	0.4979	**0.6962**	0.5098	0.5609
ES $\uparrow$	**0.7028**	0.5553	0.6292	0.4271	-0.2945	0.4833	0.558	0.634
HVG PDS cosine $\uparrow$	0.5767	0.5028	**0.7008**	0.5001	0.5018	0.6744	0.4969	0.5568
HVG PDS $L_1 \uparrow$	0.631	0.5127	**0.718**	0.5122	0.4931	0.7154	0.5114	0.5792
HVG PDS squared $L_2 \uparrow$	0.6347	0.5119	0.7312	0.5109	0.4965	**0.7339**	0.5105	0.5788
True DE genes	520.5	520.5	520.5	520.5	520.5	520.5	520.5	520.5
HVG genes	2000	2000	2000	2000	2000	2000	2000	2000

Table 21: NADIG (ID): normalized counts. Raw evaluator values; best directed scores in bold. PB: PerturbBridge; CF: CellFlow; DOL: DOLORIS; DEP: DEPARTURES; LC: Lingshu-Cell; Ctrl: Control; Lin: Linear (R1). Counts are descriptive.

Metric	PB	CF	STATE	DOL	DEP	LC	Ctrl	Lin
Exact W_1 ↓	226.2	227.9	242.8	244.6	254.6	331.2	**225**	225.6
Exact W_2 ↓	232.8	236.7	261.2	250.9	260.6	372.7	**232.5**	233
Sinkhorn $p = 1$ ↓	225.4	227	241.9	243.8	253.8	330.3	**224.1**	224.8
Sinkhorn $p = 2$ (sqrt) ↓	163.5	166.1	182.9	176.4	183.1	262.5	**163.2**	163.6
Marginal W_1 ↓	**0.3893**	0.4057	0.4591	0.5035	0.5844	0.4812	0.3993	0.4072
MMD^2 ↓	0.02933	0.04462	**0.02517**	0.07449	0.07793	0.07685	0.04717	0.04927
Response R^2 ↑	**-0.5493**	-0.8481	-1.383	-1.888	-2.474	-3.893	-0.5653	-0.7198
Response Pearson ↑	0.1769	0.1057	0.3969	0.04427	0.09947	**0.4047**	-0.002191	0.06699
L_1 error (sum) ↓	**2935**	3031	3280	3626	4295	3598	2997	3076
L_1 error / gene ↓	**0.3845**	0.3971	0.4297	0.4751	0.5628	0.4715	0.3927	0.4031
Squared error (sum) ↓	1.44e4	1.59e4	1.81e4	2.06e4	2.27e4	3.89e4	1.60e4	1.67e4
Squared error / gene ↓	**1.888**	2.086	2.365	2.7	2.97	5.096	2.1	2.183
DE AUROC ↑	0.5394	0.5261	**0.568**	0.5264	0.5132	0.5438	0.5234	0.5247
DE average precision ↑	0.09838	0.08999	**0.1135**	0.09036	0.08175	0.1105	0.08877	0.09042
DE overlap ↑	0.146	0.1285	0.1639	0.129	0.09994	**0.1667**	0.1257	0.1297
DE precision ↑	**0.1597**	0.1458	0.1272	0.1271	0.08919	0.1437	0.1434	0.1521
Direction agreement ↑	**0.9977**	0.9941	0.997	0.9769	0.7788	0.9968	0.9958	0.9913
DE response Spearman ↑	0.1668	0.05316	0.2793	0.04231	0.2025	**0.3138**	0.03905	0.0555
Predicted DE genes	555.9	445.2	1195	665.4	1174	988	438.1	440
PDS cosine ↑	0.6006	0.5042	0.7698	0.5081	0.5103	**0.7704**	0.5014	0.5295
PDS L_1 ↑	0.591	0.5043	0.7613	0.5056	0.5067	**0.7908**	0.5068	0.5258
PDS squared L_2 ↑	0.5969	0.5056	0.809	0.5065	0.5071	**0.8302**	0.5058	0.5238
ES ↑	**0.8258**	0.694	0.6506	0.555	-0.2633	0.6814	0.6913	0.6794
True DE genes	555.2	555.2	555.2	555.2	555.2	555.2	555.2	555.2

Table 22: NADIG (ID): normalized log counts. Raw evaluator values; best directed scores in bold. PB: PerturbBridge; CF: CellFlow; DOL: DOLORIS; DEP: DEPARTURES; LC: Lingshu-Cell; Ctrl: Control; Lin: Linear (R1). Counts are descriptive.

Metric	PB	CF	STATE	DOL	DEP	LC	Ctrl	Lin
Exact W_1 ↓	50.05	49.53	**47.96**	53.47	55.62	50.97	49.06	49.71
Exact W_2 ↓	50.55	50.16	**48.76**	53.8	55.95	51.15	49.65	50.3
Sinkhorn $p = 1$ ↓	50.17	49.65	**48.08**	53.59	55.75	51.1	49.18	49.83
Sinkhorn $p = 2$ (sqrt) ↓	35.72	35.44	**34.45**	38.02	39.55	36.16	35.08	35.54
Marginal W_1 ↓	**0.1263**	0.1301	0.1453	0.1534	0.2015	0.1415	0.1289	0.131
MMD^2 ↓	**0.003225**	0.006829	0.008128	0.01287	0.04755	0.01488	0.007266	0.007081
Response R^2 ↑	**-0.7759**	-0.836	-0.8448	-1.358	-3.43	-1.211	-0.8038	-0.8708
Response Pearson ↑	0.09901	0.01664	0.1817	0.03498	0.06061	**0.225**	−3.87e−4	0.02831
L_1 error (sum) ↓	**1001**	1016	1027	1119	1495	1084	1009	1031
L_1 error / gene ↓	**0.1312**	0.1331	0.1346	0.1466	0.1959	0.1421	0.1322	0.1351
Squared error (sum) ↓	**252**	259.1	273.2	313.6	513.7	307	257.2	266
Squared error / gene ↓	**0.03302**	0.03395	0.03579	0.04109	0.06731	0.04022	0.0337	0.03485
DE AUROC ↑	0.54	0.5256	**0.5702**	0.5242	0.5158	0.5434	0.5237	0.524
DE average precision ↑	0.08705	0.07831	**0.1034**	0.07905	0.07357	0.09801	0.07737	0.07852
DE overlap ↑	0.1357	0.1186	**0.1572**	0.1207	0.1012	0.1505	0.1168	0.1188
DE precision ↑	**0.1552**	0.1437	0.1331	0.1294	0.09497	0.1366	0.1433	0.148
Direction agreement ↑	**0.9973**	0.9929	0.9939	0.969	0.6664	0.9949	0.9939	0.9895
DE response Spearman ↑	0.153	0.03006	0.2452	0.05999	0.08616	**0.2813**	0.01754	0.02763
Predicted DE genes	502.6	376.6	813.4	517	1112	914.1	371.8	373.9
PDS cosine ↑	0.6234	0.5025	**0.8352**	0.5045	0.5059	0.7791	0.4989	0.5363
PDS L_1 ↑	0.5682	0.5043	0.6635	0.5033	0.4963	**0.6944**	0.5046	0.5133
PDS squared L_2 ↑	0.5742	0.5012	0.6836	0.5003	0.4937	**0.7259**	0.5012	0.5147
ES ↑	**0.8391**	0.7039	0.726	0.588	-0.3032	0.681	0.7011	0.709
True DE genes	476	476	476	476	476	476	476	476

Table 23: PBMC (ID): raw counts. Raw evaluator values; best directed scores in bold. PB: PerturbBridge; CF: CellFlow; DOL: DOLORIS; DEP: DEPARTURES; LC: Lingshu-Cell; Ctrl: Control; Lin: Linear (R1). Counts are descriptive.

Metric	PB	CF	STATE	DOL	DEP	LC	Ctrl	Lin
Exact W_1 ↓	**108.3**	113.7	127.2	179.2	113.1	148.7	114.1	110.3
Exact W_2 ↓	**118.4**	125.2	146.1	209.8	125.1	185.8	125.4	121.4
Sinkhorn $p = 1$ ↓	**108**	113.4	126.9	177.8	113.1	148.5	113.8	110.2
Sinkhorn $p = 2$ (sqrt) ↓	**83.28**	87.92	102.3	147.7	88.23	130.5	88.11	85.56
Marginal W_1 ↓	**0.06173**	0.06586	0.08742	0.1604	0.1788	0.08211	0.06591	0.1574
MMD^2 ↓	**0.004813**	0.01571	0.01274	0.06804	0.2268	0.04827	0.01585	0.1598
Response R^2 ↑	-0.8059	-0.5249	-1.613	-13.24	-8.642	-2.963	**-0.5221**	-1.155
Response Pearson ↑	0.417	0.01946	0.3086	0.03921	0.04607	**0.4755**	-0.002842	0.187
L_1 error (sum) ↓	**1554**	1690	1798	3393	3054	1850	1688	1656
L_1 error / gene ↓	**0.08534**	0.09279	0.0987	0.1863	0.1677	0.1016	0.09268	0.09093
Squared error (sum) ↓	2661	2664	3794	1.32e4	7818	5955	2643	**2608**
Squared error / gene ↓	0.1461	0.1463	0.2083	0.7241	0.4292	0.3269	0.1451	**0.1432**
DE AUROC ↑	0.6242	0.5112	0.5992	0.5976	**0.723**	0.6292	0.5107	0.5844
DE average precision ↑	0.128	0.04924	0.09213	0.0749	0.07764	**0.1543**	0.04904	0.064
DE overlap ↑	0.1807	0.09344	0.1217	0.1103	0.07572	**0.1964**	0.09349	0.09088
DE precision ↑	**0.2062**	0.09857	0.1669	0.1013	0.06204	0.1473	0.09594	0.08603
Direction agreement ↑	**0.9992**	0.9769	0.9609	0.7544	0.8925	0.9904	0.9807	0.811
DE response Spearman ↑	0.4545	0.04146	0.275	-0.2506	**0.5607**	0.3963	0.02627	0.2052
Predicted DE genes	587.7	191.7	569.3	1396	8432	1144	192	1879
PDS cosine ↑	**0.7542**	0.5155	0.6411	0.515	0.5398	0.7228	0.5099	0.6061
PDS L_1 ↑	**0.7522**	0.5231	0.6533	0.5195	0.5426	0.7426	0.5196	0.5897
PDS squared L_2 ↑	0.7338	0.517	0.6426	0.5138	0.5368	**0.746**	0.513	0.5866
ES ↑	**0.7492**	0.3789	0.4175	0.2303	0.01558	0.4895	0.385	0.2267
HVG PDS cosine ↑	**0.7532**	0.5158	0.6384	0.5149	0.5388	0.7232	0.5092	0.6052
HVG PDS L_1 ↑	**0.747**	0.5181	0.6469	0.5165	0.548	0.7368	0.5128	0.5881
HVG PDS squared L_2 ↑	0.7315	0.5169	0.6415	0.5137	0.5366	**0.7455**	0.5131	0.5862
True DE genes	643.7	643.7	643.7	643.7	643.7	643.7	643.7	643.7
HVG genes	2000	2000	2000	2000	2000	2000	2000	2000

Table 24: PBMC (ID): log counts. Raw evaluator values; best directed scores in bold. PB: PerturbBridge; CF: CellFlow; DOL: DOLORIS; DEP: DEPARTURES; LC: Lingshu-Cell; Ctrl: Control; Lin: Linear (R1). Counts are descriptive.

Metric	PB	CF	STATE	DOL	DEP	LC	Ctrl	Lin
Exact W_1 ↓	41.5	42.3	41.81	51.18	39.3	46.91	42.32	**39.16**
Exact W_2 ↓	42.35	43.22	42.69	52.26	39.9	47.15	43.24	**39.78**
Sinkhorn $p = 1$ ↓	41.71	42.5	42.02	51.36	39.51	47.13	42.52	**39.27**
Sinkhorn $p = 2$ (sqrt) ↓	29.9	30.52	30.13	36.91	28.19	33.32	30.53	**28.11**
Marginal W_1 ↓	**0.02098**	0.02499	0.03335	0.04969	0.07507	0.03003	0.025	0.07859
MMD^2 ↓	**0.002527**	0.008705	0.005914	0.03294	0.1729	0.0198	0.008738	0.1304
Response R^2 ↑	**-0.4139**	-0.6209	-0.7366	-4.741	-11.36	-1.71	-0.6251	-3.438
Response Pearson ↑	0.3799	0.002954	0.3288	0.02856	0.1099	**0.4136**	-0.001274	0.1251
L_1 error (sum) ↓	**606.7**	671.9	682.4	1009	1297	731.5	672.1	851.3
L_1 error / gene ↓	**0.03331**	0.03689	0.03747	0.0554	0.07124	0.04016	0.0369	0.04674
Squared error (sum) ↓	**107.4**	134.3	143.4	279.3	474	157.7	134.4	207.7
Squared error / gene ↓	**0.005899**	0.007372	0.007873	0.01534	0.02603	0.008658	0.007379	0.0114
DE AUROC ↑	0.6187	0.5102	0.5971	0.5691	**0.7134**	0.6307	0.5101	0.5682
DE average precision ↑	0.1288	0.04948	0.09615	0.07061	0.07638	**0.1544**	0.04947	0.06319
DE overlap ↑	0.1864	0.09606	0.125	0.1122	0.07208	**0.1987**	0.09629	0.09188
DE precision ↑	**0.2103**	0.1034	0.172	0.1157	0.06369	0.1432	0.103	0.08406
Direction agreement ↑	**0.9987**	0.9799	0.9453	0.8254	0.8923	0.9866	0.982	0.6825
DE response Spearman ↑	0.4122	0.01867	0.2928	-0.1206	**0.564**	0.3883	0.01222	0.03827
Predicted DE genes	615.5	204	526.4	912	8370	1279	204.5	3851
PDS cosine ↑	**0.7773**	0.5149	0.6549	0.5169	0.5723	0.7074	0.5153	0.5554
PDS L_1 ↑	**0.7588**	0.5265	0.6536	0.5174	0.5551	0.7205	0.5227	0.5462
PDS squared L_2 ↑	**0.7601**	0.518	0.6603	0.5149	0.571	0.7251	0.5138	0.5497
ES ↑	**0.7461**	0.4014	0.3633	0.3352	0.01553	0.4704	0.4089	0.1572
HVG PDS cosine ↑	**0.7777**	0.5153	0.654	0.517	0.5693	0.7088	0.513	0.5567
HVG PDS L_1 ↑	**0.7591**	0.5179	0.6579	0.5154	0.5672	0.7053	0.514	0.541
HVG PDS squared L_2 ↑	**0.7621**	0.5181	0.6617	0.5139	0.5699	0.7245	0.514	0.5497
True DE genes	653	653	653	653	653	653	653	653
HVG genes	2000	2000	2000	2000	2000	2000	2000	2000

Table 25: PBMC (ID): normalized counts. Raw evaluator values; best directed scores in bold. PB: PerturbBridge; CF: CellFlow; DOL: DOLORIS; DEP: DEPARTURES; LC: Lingshu-Cell; Ctrl: Control; Lin: Linear (R1). Counts are descriptive.

Metric	PB	CF	STATE	DOL	DEP	LC	Ctrl	Lin
Exact W_1 ↓	104.2	105.9	122.2	114.1	99.96	104.9	106.1	**94.62**
Exact W_2 ↓	106.2	107.8	127.6	116.7	101.2	113.5	108	**96.4**
Sinkhorn $p = 1$ ↓	104.4	106	122.2	114.1	100.1	105.1	106.2	**94.66**
Sinkhorn $p = 2$ (sqrt) ↓	74.99	76.07	89.81	82.32	71.49	79.81	76.2	**68.08**
Marginal W_1 ↓	**0.04673**	0.05168	0.07075	0.07182	0.1528	0.05584	0.05167	0.1263
MMD^2 ↓	**0.004745**	0.01054	0.02095	0.03726	0.1785	0.0116	0.01066	0.1046
Response R^2 ↑	**-0.5262**	-0.5362	-1.456	-4.285	-13.75	-0.7526	-0.5369	-1.537
Response Pearson ↑	0.3605	0.01766	0.2322	-0.01805	0.07958	**0.4362**	−8.26e−4	0.1227
L_1 error (sum) ↓	1267	1350	1368	1713	2706	**1247**	1350	1313
L_1 error / gene ↓	0.06956	0.07413	0.07511	0.09408	0.1486	**0.06849**	0.07414	0.07207
Squared error (sum) ↓	**1312**	1527	2064	3093	4533	1969	1546	1569
Squared error / gene ↓	**0.07205**	0.08385	0.1133	0.1699	0.2489	0.1081	0.08486	0.08617
DE AUROC ↑	0.6277	0.5132	0.6181	0.5571	**0.7265**	0.6629	0.5123	0.5987
DE average precision ↑	0.1015	0.02846	0.06145	0.0337	0.04277	**0.1374**	0.02825	0.0374
DE overlap ↑	0.1522	0.05817	0.09573	0.0549	0.03862	**0.1814**	0.05745	0.05444
DE precision ↑	**0.1606**	0.07065	0.1017	0.05832	0.03779	0.1265	0.06699	0.05764
Direction agreement ↑	**0.9993**	0.9834	0.9721	0.8531	0.8948	0.9977	0.9895	0.8483
DE response Spearman ↑	0.4785	0.0462	0.3843	0.02361	0.4168	**0.5449**	0.03358	0.2592
Predicted DE genes	331.6	137.3	421.1	609.4	7863	613.5	137.6	1447
PDS cosine ↑	0.79	0.5177	0.6396	0.5142	0.5334	**0.7952**	0.5101	0.578
PDS L_1 ↑	0.7339	0.5228	0.6164	0.5236	0.5965	**0.7582**	0.5228	0.5797
PDS squared L_2 ↑	0.7649	0.5173	0.63	0.5221	0.5594	**0.8015**	0.5184	0.5787
ES ↑	**0.8445**	0.5461	0.4135	0.4158	-0.007342	0.7267	0.5326	0.3662
True DE genes	359.5	359.5	359.5	359.5	359.5	359.5	359.5	359.5

Table 26: PBMC (ID): normalized log counts. Raw evaluator values; best directed scores in bold. PB: PerturbBridge; CF: CellFlow; DOL: DOLORIS; DEP: DEPARTURES; LC: Lingshu-Cell; Ctrl: Control; Lin: Linear (R1). Counts are descriptive.

Metric	PB	CF	STATE	DOL	DEP	LC	Ctrl	Lin
Exact W_1 ↓	41.99	41.93	40.92	42.98	39.83	40.33	41.9	**36.89**
Exact W_2 ↓	42.32	42.25	41.27	43.29	40.02	40.46	42.22	**37.2**
Sinkhorn $p = 1$ ↓	42.22	42.15	41.14	43.2	40.04	40.56	42.12	**37**
Sinkhorn $p = 2$ (sqrt) ↓	29.91	29.86	29.16	30.59	28.29	28.6	29.84	**26.29**
Marginal W_1 ↓	**0.01937**	0.02179	0.02697	0.02865	0.07	0.02519	0.02178	0.06788
MMD^2 ↓	**0.001865**	0.005178	0.004539	0.01662	0.1705	0.006373	0.005215	0.102
Response R^2 ↑	-0.4947	-0.612	**-0.4068**	-2.653	-18.37	-0.5607	-0.6131	-3.111
Response Pearson ↑	0.3195	0.002929	0.271	-0.008146	0.08052	**0.3887**	−2.31e−4	0.09601
L_1 error (sum) ↓	551.4	582.1	**539.2**	688.1	1286	550.7	582.3	708.1
L_1 error / gene ↓	0.03027	0.03196	**0.0296**	0.03778	0.0706	0.03024	0.03197	0.03888
Squared error (sum) ↓	**89.9**	104.8	92.6	148.6	494.8	91.01	105	144.8
Squared error / gene ↓	**0.004936**	0.005755	0.005084	0.008161	0.02717	0.004997	0.005762	0.007948
DE AUROC ↑	0.6171	0.513	0.5986	0.5428	**0.715**	0.633	0.5129	0.556
DE average precision ↑	0.1033	0.03195	0.06968	0.03627	0.04561	**0.1346**	0.03188	0.03643
DE overlap ↑	0.1597	0.06645	0.1029	0.0641	0.03993	**0.1829**	0.06633	0.05823
DE precision ↑	**0.1665**	0.08111	0.1431	0.07241	0.0422	0.1341	0.07983	0.05895
Direction agreement ↑	**0.999**	0.9866	0.9581	0.8261	0.9078	0.9978	0.9884	0.6669
DE response Spearman ↑	0.427	0.02398	0.356	-0.02578	**0.5238**	0.4502	0.01608	0.04563
Predicted DE genes	390	168.8	332.2	499.2	8134	645.4	169.4	3445
PDS cosine ↑	**0.8281**	0.5156	0.6771	0.5193	0.5504	0.7934	0.5099	0.5657
PDS L_1 ↑	0.7327	0.5261	0.6257	0.5254	0.6031	**0.7594**	0.5261	0.5781
PDS squared L_2 ↑	0.7553	0.5171	0.6367	0.5166	0.5787	**0.7834**	0.5157	0.5784
ES ↑	**0.8117**	0.5076	0.3878	0.4409	-0.002453	0.7039	0.5052	0.1142
True DE genes	400.4	400.4	400.4	400.4	400.4	400.4	400.4	400.4

Table 27: PBMC (OOD cell type): raw counts. Raw evaluator values; best directed scores in bold. PB: PerturbBridge; CF: CellFlow; DOL: DOLORIS; DEP: DEPARTURES; LC: Lingshu-Cell; Ctrl: Control; Lin: Linear (R1). Counts are descriptive.

Metric	PB	CF	STATE	DOL	DEP	LC	Ctrl	Lin
Exact W_1 ↓	**89.79**	92.28	107.4	102.6	96.52	184.9	92.2	100.7
Exact W_2 ↓	**96.78**	100.3	116.4	112.8	103.1	248	100.2	106.6
Sinkhorn $p = 1$ ↓	**89.79**	92.3	107.5	102.7	96.65	184.9	92.22	100.7
Sinkhorn $p = 2$ (sqrt) ↓	**68.17**	70.59	81.75	79.43	72.7	175.1	70.5	75.25
Marginal W_1 ↓	**0.03654**	0.03923	0.0749	0.06462	0.08102	0.1814	0.03923	0.1607
MMD^2 ↓	**0.007029**	0.01287	0.02773	0.03079	0.03294	0.1591	0.01292	0.2302
Response R^2 ↑	**-0.2337**	-0.3506	-5.039	-7.938	-6.902	-51.56	-0.3463	-22.27
Response Pearson ↑	**0.4405**	0.009579	0.1961	0.1384	0.1897	0.2218	0.001354	0.0577
L_1 error (sum) ↓	1008	1148	1569	**895.2**	1321	3503	1148	2047
L_1 error / gene ↓	0.05535	0.06304	0.08616	**0.04915**	0.07253	0.1924	0.06302	0.1124
Squared error (sum) ↓	1352	1474	3093	705.7	**597.1**	1.41e4	1461	4374
Squared error / gene ↓	0.07426	0.08091	0.1698	0.03875	**0.03279**	0.7762	0.08023	0.2402
DE AUROC ↑	0.6521	0.4997	0.6364	0.6504	**0.7169**	0.6672	0.4992	0.5804
DE average precision ↑	0.1922	0.09282	0.1556	0.1812	0.1778	**0.1933**	0.09271	0.1119
DE overlap ↑	**0.22**	0.1427	0.1744	0.2106	0.1838	0.214	0.1422	0.1468
DE precision ↑	**0.2143**	0.09774	0.1412	0.1548	0.1124	0.1315	0.09705	0.1019
Direction agreement ↑	0.945	**0.9549**	0.7544	0.6191	0.7353	0.6871	0.9528	0.729
DE response Spearman ↑	**0.5011**	0.01934	0.2642	-0.1249	0.3244	0.2714	0.01544	0.1031
Predicted DE genes	1475	150.8	4325	1997	1.19e4	6211	149.4	4970
PDS cosine ↑	**0.8012**	0.5095	0.6471	0.5431	0.6507	0.6913	0.5081	0.5558
PDS L_1 ↑	**0.8344**	0.5106	0.6607	0.5353	0.5507	0.6303	0.5109	0.5516
PDS squared L_2 ↑	**0.8074**	0.5075	0.6462	0.5382	0.6065	0.7065	0.5076	0.5599
ES ↑	**0.4445**	0.149	0.1655	-0.03367	0.281	0.1517	0.1422	0.09032
HVG PDS cosine ↑	0.669	0.5096	0.6431	0.5427	0.6727	**0.6926**	0.5085	0.5544
HVG PDS L_1 ↑	0.6471	0.508	**0.6712**	0.5317	0.5698	0.6508	0.5081	0.5574
HVG PDS squared L_2 ↑	0.6554	0.5076	0.6419	0.5379	0.607	**0.7087**	0.5076	0.5584
True DE genes	1480	1480	1480	1480	1480	1480	1480	1480
HVG genes	2000	2000	2000	2000	2000	2000	2000	2000

Table 28: PBMC (OOD cell type): log counts. Raw evaluator values; best directed scores in bold. PB: PerturbBridge; CF: CellFlow; DOL: DOLORIS; DEP: DEPARTURES; LC: Lingshu-Cell; Ctrl: Control; Lin: Linear (R1). Counts are descriptive.

Metric	PB	CF	STATE	DOL	DEP	LC	Ctrl	Lin
Exact W_1 ↓	**38.66**	39.23	40.72	40.26	43.37	52.69	39.23	40.54
Exact W_2 ↓	**39.48**	40.18	41.62	41.23	43.99	52.86	40.18	41.05
Sinkhorn $p = 1$ ↓	**38.92**	39.5	40.97	40.56	43.67	52.96	39.49	40.67
Sinkhorn $p = 2$ (sqrt) ↓	**27.89**	28.38	29.37	29.13	31.09	37.35	28.38	29.01
Marginal W_1 ↓	**0.01528**	0.01616	0.03281	0.02831	0.03977	0.06766	0.01616	0.08934
MMD^2 ↓	**0.004292**	0.007189	0.01421	0.02634	0.03545	0.1047	0.00719	0.2003
Response R^2 ↑	**-0.3541**	-0.4631	-3.706	-10.78	-15.25	-34.48	-0.4623	-37.01
Response Pearson ↑	**0.4332**	3.66e−4	0.2424	0.1201	0.1754	0.241	9.00e−4	0.07585
L_1 error (sum) ↓	436.6	469.1	638.3	**399.7**	664.6	1298	468.9	1129
L_1 error / gene ↓	0.02397	0.02575	0.03504	**0.02195**	0.03649	0.07125	0.02575	0.062
Squared error (sum) ↓	**60.7**	83.21	154.1	78.58	94.69	534.7	82.95	455.6
Squared error / gene ↓	**0.003333**	0.004569	0.00846	0.004314	0.005199	0.02936	0.004554	0.02502
DE AUROC ↑	0.6492	0.4995	0.636	0.6267	**0.7222**	0.6663	0.499	0.5748
DE average precision ↑	**0.1894**	0.09249	0.1523	0.1708	0.1838	0.1885	0.09236	0.1117
DE overlap ↑	**0.2181**	0.1419	0.171	0.2045	0.2023	0.2102	0.1417	0.1535
DE precision ↑	**0.2162**	0.1004	0.139	0.1743	0.1126	0.1307	0.1003	0.1007
Direction agreement ↑	0.9518	**0.9632**	0.7176	0.6822	0.7292	0.6898	0.9619	0.6911
DE response Spearman ↑	**0.4605**	0.01076	0.2557	-0.06003	0.288	0.2589	0.008675	0.05139
Predicted DE genes	1482	160	4540	1949	1.17e4	6395	158.9	5840
PDS cosine ↑	**0.8146**	0.5087	0.6591	0.5524	0.5766	0.6538	0.506	0.5242
PDS L_1 ↑	**0.7952**	0.5111	0.6418	0.5381	0.5323	0.6055	0.5109	0.5177
PDS squared L_2 ↑	**0.8227**	0.5074	0.6761	0.5665	0.5958	0.6462	0.5075	0.5225
ES ↑	**0.4649**	0.1553	0.1319	-0.02244	0.2686	0.1687	0.155	0.08105
HVG PDS cosine ↑	**0.676**	0.5091	0.6658	0.5544	0.6112	0.6693	0.5069	0.5261
HVG PDS L_1 ↑	0.6419	0.5073	**0.6701**	0.552	0.5644	0.6253	0.5078	0.5182
HVG PDS squared L_2 ↑	0.6634	0.5073	**0.6796**	0.5663	0.6125	0.6636	0.5077	0.5243
True DE genes	1483	1483	1483	1483	1483	1483	1483	1483
HVG genes	2000	2000	2000	2000	2000	2000	2000	2000

Table 29: PBMC (OOD cell type): normalized counts. Raw evaluator values; best directed scores in bold. PB: PerturbBridge; CF: CellFlow; DOL: DOLORIS; DEP: DEPARTURES; LC: Lingshu-Cell; Ctrl: Control; Lin: Linear (R1). Counts are descriptive.

Metric	PB	CF	STATE	DOL	DEP	LC	Ctrl	Lin
Exact W_1 ↓	82.08	87.58	100.8	83.22	**69.82**	95.29	87.57	81.93
Exact W_2 ↓	83.25	89.24	105.7	84.45	**70.63**	102.2	89.23	83.22
Sinkhorn $p = 1$ ↓	82.32	87.81	100.9	83.47	**70.08**	95.51	87.79	82.05
Sinkhorn $p = 2$ (sqrt) ↓	58.79	63.02	74.37	59.61	**49.88**	72.1	63.02	58.81
Marginal W_1 ↓	**0.02825**	0.02872	0.05799	0.04144	0.05294	0.08765	0.02872	0.1059
MMD^2 ↓	**0.005732**	0.006403	0.03376	0.02487	0.04979	0.1062	0.006398	0.1444
Response R^2 ↑	**-0.3549**	-0.3802	-9.578	-10.73	-12.17	-30.66	-0.3807	-18.15
Response Pearson ↑	**0.3238**	0.003163	0.1182	0.04227	0.1165	0.1328	9.84e−4	0.03469
L_1 error (sum) ↓	823.1	840.4	1304	**605.1**	747.7	1778	839.9	1311
L_1 error / gene ↓	0.04519	0.04614	0.07159	**0.03322**	0.04105	0.0976	0.04612	0.07196
Squared error (sum) ↓	754.1	753.2	2651	**347.2**	391.7	4135	756.3	2195
Squared error / gene ↓	0.0414	0.04135	0.1456	**0.01906**	0.0215	0.2271	0.04152	0.1205
DE AUROC ↑	0.6554	0.502	0.625	0.5811	**0.7011**	0.6729	0.5015	0.588
DE average precision ↑	**0.1075**	0.03784	0.07454	0.0586	0.06269	0.09485	0.03784	0.04735
DE overlap ↑	**0.1439**	0.06584	0.09796	0.09129	0.08082	0.1264	0.06572	0.06665
DE precision ↑	**0.1377**	0.06485	0.06876	0.08439	0.046	0.05861	0.06552	0.05602
Direction agreement ↑	0.9491	**0.9837**	0.7985	0.7226	0.7927	0.7885	0.9827	0.8163
DE response Spearman ↑	**0.5664**	0.02167	0.2867	-0.00656	0.4473	0.4506	0.01903	0.2027
Predicted DE genes	158.3	112.3	2685	4926	9803	4779	111.5	3888
PDS cosine ↑	**0.8285**	0.5109	0.5976	0.5315	0.5632	0.6618	0.5112	0.5443
PDS L_1 ↑	**0.836**	0.5103	0.642	0.5604	0.5859	0.7747	0.5106	0.5967
PDS squared L_2 ↑	**0.8145**	0.5075	0.6165	0.5246	0.5731	0.7081	0.5077	0.5594
ES ↑	**0.6047**	0.3531	0.06869	0.03365	0.1159	0.2525	0.354	0.3053
True DE genes	579.2	579.2	579.2	579.2	579.2	579.2	579.2	579.2

Table 30: PBMC (OOD cell type): normalized log counts. Raw evaluator values; best directed scores in bold. PB: PerturbBridge; CF: CellFlow; DOL: DOLORIS; DEP: DEPARTURES; LC: Lingshu-Cell; Ctrl: Control; Lin: Linear (R1). Counts are descriptive.

Metric	PB	CF	STATE	DOL	DEP	LC	Ctrl	Lin
Exact W_1 ↓	36.41	38.74	38.19	36.53	**34.96**	38.48	38.74	35.34
Exact W_2 ↓	36.64	39.02	38.49	36.77	**35.14**	38.59	39.03	35.57
Sinkhorn $p = 1$ ↓	36.69	39.01	38.45	36.83	**35.25**	38.75	39.01	35.48
Sinkhorn $p = 2$ (sqrt) ↓	25.9	27.58	27.2	25.99	**24.84**	27.28	27.58	25.15
Marginal W_1 ↓	**0.01291**	0.01316	0.02523	0.02076	0.02853	0.04282	0.01316	0.06246
MMD^2 ↓	**0.003048**	0.003416	0.01265	0.02302	0.02649	0.06542	0.003413	0.1276
Response R^2 ↑	**-0.4183**	-0.4541	-4.481	-13.82	-14.14	-21.2	-0.453	-19.38
Response Pearson ↑	**0.3456**	−5.75e−5	0.1633	0.0763	0.119	0.1725	5.74e−4	0.05698
L_1 error (sum) ↓	373.6	375.1	516.8	**316.6**	430.6	774.4	374.9	676.5
L_1 error / gene ↓	0.02052	0.02059	0.02837	**0.01738**	0.02364	0.04252	0.02059	0.03714
Squared error (sum) ↓	**49.65**	59.02	124.3	55.06	55.7	249.5	58.92	212.6
Squared error / gene ↓	**0.002726**	0.00324	0.006825	0.003023	0.003059	0.0137	0.003235	0.01167
DE AUROC ↑	0.6447	0.5026	0.6225	0.5741	**0.679**	0.6529	0.5025	0.5647
DE average precision ↑	**0.1106**	0.0443	0.08197	0.06656	0.06954	0.09585	0.04425	0.05095
DE overlap ↑	**0.1472**	0.07691	0.1046	0.1034	0.08769	0.1241	0.07667	0.07435
DE precision ↑	**0.1401**	0.07533	0.07591	0.1077	0.05243	0.06656	0.07495	0.06134
Direction agreement ↑	0.9483	**0.9831**	0.7272	0.6841	0.7645	0.7343	0.983	0.7149
DE response Spearman ↑	**0.4866**	0.01233	0.26	-0.03482	0.3527	0.3569	0.01253	0.0983
Predicted DE genes	1143	133.3	3121	3684	1.06e4	4905	132.2	4700
PDS cosine ↑	**0.9146**	0.5094	0.6419	0.5503	0.5979	0.7318	0.5099	0.5753
PDS L_1 ↑	**0.8639**	0.5114	0.6736	0.5156	0.5891	0.7685	0.5116	0.5944
PDS squared L_2 ↑	**0.8849**	0.5061	0.6403	0.5749	0.6075	0.7692	0.5066	0.5919
ES ↑	**0.5702**	0.3068	0.05306	0.04604	0.2227	0.224	0.3089	0.08141
True DE genes	673.4	673.4	673.4	673.4	673.4	673.4	673.4	673.4

Table 31: PBMC (OOD donor): raw counts. Raw evaluator values; best directed scores in bold. PB: PerturbBridge; CF: CellFlow; DOL: DOLORIS; DEP: DEPARTURES; LC: Lingshu-Cell; Ctrl: Control; Lin: Linear (R1). Counts are descriptive.

Metric	PB	CF	STATE	DOL	DEP	LC	Ctrl	Lin
Exact $W_1 \downarrow$	**108.3**	110	132	133.4	117	152.9	109.5	109.3
Exact $W_2 \downarrow$	**121.1**	123.2	202.3	150.5	128.2	189	122.1	121.8
Sinkhorn $p = 1 \downarrow$	**107.9**	109.9	131.5	133	116.9	152.7	109.4	109.2
Sinkhorn $p = 2$ (sqrt) $\downarrow$	**85.22**	86.47	139.9	105.8	90.14	132.5	85.75	85.84
Marginal $W_1 \downarrow$	0.05495	0.0446	0.08426	0.07053	0.07534	0.09487	**0.04444**	0.1478
$\text{MMD}^2 \downarrow$	0.01726	**0.008658**	0.02096	0.02026	0.02244	0.07982	0.008792	0.1687
Response $R^2 \uparrow$	-3.365	-0.4764	-6.695	-4.665	-7.203	-12.65	**-0.4324**	-3.545
Response Pearson $\uparrow$	0.2192	0.0467	0.1744	0.1247	0.1459	**0.3318**	−5.79e−4	0.1795
L_1 error (sum) $\downarrow$	1571	1364	1732	1633	**1096**	2151	1361	1519
L_1 error / gene $\downarrow$	0.08626	0.07489	0.09509	0.08965	**0.06016**	0.1181	0.07475	0.08341
Squared error (sum) $\downarrow$	3406	2745	4512	3713	**1115**	9813	2714	3496
Squared error / gene $\downarrow$	0.187	0.1507	0.2477	0.2039	**0.06123**	0.5388	0.149	0.1919
DE AUROC $\uparrow$	0.6653	0.5029	0.6214	0.5702	**0.6758**	0.6695	0.5016	0.5944
DE average precision $\uparrow$	0.1826	0.07716	0.1266	0.1208	**0.2791**	0.1955	0.07684	0.1136
DE overlap $\uparrow$	0.2068	0.1243	0.1465	0.1579	**0.3037**	0.224	0.1241	0.1476
DE precision $\uparrow$	0.1841	0.09067	0.1249	0.1876	**0.2506**	0.1467	0.08564	0.1253
Direction agreement $\uparrow$	0.8874	0.9613	0.823	0.771	0.6482	0.8783	**0.9635**	0.817
DE response Spearman $\uparrow$	0.377	0.03611	0.2642	0.0962	0.1415	**0.4333**	0.009812	0.2312
Predicted DE genes	1543	120.8	2865	753.5	5764	3384	120.7	3152
PDS cosine $\uparrow$	0.6425	0.526	0.6154	0.5207	0.6664	**0.6934**	0.5074	0.6085
PDS $L_1 \uparrow$	0.6886	0.5149	0.6301	0.5376	0.6152	**0.7023**	0.5136	0.6021
PDS squared $L_2 \uparrow$	0.6425	0.5119	0.615	0.5273	0.615	**0.7025**	0.51	0.587
ES $\uparrow$	**0.4185**	0.1131	0.2728	0.1789	0.1479	0.2762	0.112	0.1568
HVG PDS cosine $\uparrow$	0.6379	0.5256	0.6126	0.5203	0.6644	**0.6917**	0.5071	0.6064
HVG PDS $L_1 \uparrow$	0.6744	0.512	0.6251	0.5366	0.62	**0.6956**	0.5106	0.5968
HVG PDS squared $L_2 \uparrow$	0.6404	0.5121	0.6137	0.5276	0.6134	**0.701**	0.51	0.5862
True DE genes	1151	1151	1151	1151	1151	1151	1151	1151
HVG genes	2000	2000	2000	2000	2000	2000	2000	2000

Table 32: PBMC (OOD donor): log counts. Raw evaluator values; best directed scores in bold. PB: PerturbBridge; CF: CellFlow; DOL: DOLORIS; DEP: DEPARTURES; LC: Lingshu-Cell; Ctrl: Control; Lin: Linear (R1). Counts are descriptive.

Metric	PB	CF	STATE	DOL	DEP	LC	Ctrl	Lin
Exact $W_1 \downarrow$	41.32	41.79	41.93	44.22	45.2	47.67	41.78	**39.09**
Exact $W_2 \downarrow$	42.33	42.84	42.85	45.31	46.07	47.97	42.82	**39.78**
Sinkhorn $p = 1 \downarrow$	41.58	42.05	42.18	44.49	45.49	47.94	42.04	**39.22**
Sinkhorn $p = 2$ (sqrt) $\downarrow$	29.9	30.26	30.24	32.01	32.55	33.89	30.24	**28.11**
Marginal $W_1 \downarrow$	0.01938	0.01656	0.03195	0.02353	0.0315	0.03319	**0.01655**	0.0757
$\text{MMD}^2 \downarrow$	0.008303	**0.005166**	0.01077	0.006662	0.02484	0.0364	0.005192	0.1334
Response $R^2 \uparrow$	-2.092	**-0.5178**	-2.316	-1.308	-14.22	-8.352	-0.52	-8.783
Response Pearson $\uparrow$	0.2685	0.00802	0.2075	0.1422	0.09698	**0.3333**	-0.001209	0.1421
L_1 error (sum) $\downarrow$	561.2	516.6	616.9	549.9	**458.3**	766.7	516.8	796.2
L_1 error / gene $\downarrow$	0.03081	0.02836	0.03387	0.03019	**0.02516**	0.0421	0.02837	0.04372
Squared error (sum) $\downarrow$	116.5	101.6	139.5	117.9	**100.4**	219.6	101.5	213.3
Squared error / gene $\downarrow$	0.006397	0.005577	0.007657	0.006473	**0.005513**	0.01206	0.005572	0.01171
DE AUROC $\uparrow$	0.6607	0.5016	0.6179	0.5578	0.6681	**0.6682**	0.5012	0.5717
DE average precision $\uparrow$	0.1789	0.0754	0.1279	0.1092	**0.2756**	0.1961	0.07534	0.1084
DE overlap $\uparrow$	0.2034	0.1199	0.1496	0.1447	**0.3007**	0.2227	0.1199	0.1441
DE precision $\uparrow$	0.1807	0.09202	0.1252	0.1825	**0.258**	0.1422	0.09053	0.1131
Direction agreement $\uparrow$	0.8808	0.9684	0.7929	0.819	0.6642	0.8818	**0.9696**	0.7103
DE response Spearman $\uparrow$	0.3468	0.01579	0.2619	0.1497	0.17	**0.4274**	0.002342	0.0958
Predicted DE genes	1524	127.4	2861	619.6	5144	3487	127.4	4984
PDS cosine $\uparrow$	**0.7027**	0.5076	0.6333	0.5688	0.6481	0.6936	0.5035	0.5526
PDS $L_1 \uparrow$	**0.7066**	0.5152	0.6396	0.5497	0.6323	0.6903	0.5149	0.5316
PDS squared $L_2 \uparrow$	**0.7**	0.5103	0.6407	0.5552	0.6143	0.6966	0.5092	0.5368
ES $\uparrow$	**0.4432**	0.1294	0.2218	0.2879	0.1741	0.2718	0.1293	0.07695
HVG PDS cosine $\uparrow$	**0.6915**	0.5088	0.6259	0.5687	0.6383	0.6901	0.5039	0.554
HVG PDS $L_1 \uparrow$	**0.6985**	0.5104	0.636	0.562	0.6295	0.6823	0.51	0.5289
HVG PDS squared $L_2 \uparrow$	0.6936	0.5096	0.639	0.5623	0.6096	**0.6951**	0.5091	0.5371
True DE genes	1157	1157	1157	1157	1157	1157	1157	1157
HVG genes	2000	2000	2000	2000	2000	2000	2000	2000

Table 33: PBMC (OOD donor): normalized counts. Raw evaluator values; best directed scores in bold. PB: PerturbBridge; CF: CellFlow; DOL: DOLORIS; DEP: DEPARTURES; LC: Lingshu-Cell; Ctrl: Control; Lin: Linear (R1). Counts are descriptive.

Metric	PB	CF	STATE	DOL	DEP	LC	Ctrl	Lin
Exact W_1 ↓	101.9	100	117.6	107.7	101.4	103.2	100	**91.11**
Exact W_2 ↓	104.1	102.1	122.8	109.3	102.3	109.8	102.1	**93.02**
Sinkhorn $p = 1$ ↓	102	100.2	117.7	107.8	101.5	103.3	100.2	**91.18**
Sinkhorn $p = 2$ (sqrt) ↓	73.42	72.05	86.34	77.08	72.17	77.1	72.04	**65.69**
Marginal W_1 ↓	0.04162	**0.03236**	0.06452	0.04516	0.06044	0.06218	0.03236	0.1132
MMD^2 ↓	0.01728	**0.0072**	0.03061	0.01414	0.02835	0.04254	0.007255	0.1081
Response R^2 ↑	-4.67	**-0.4151**	-6.21	-3.058	-17.73	-8.598	-0.418	-4.922
Response Pearson ↑	0.2013	0.02668	0.1517	0.1127	0.1035	**0.2842**	0.003906	0.113
L_1 error (sum) ↓	1208	1036	1380	1183	**939.9**	1387	1037	1146
L_1 error / gene ↓	0.06632	0.0569	0.07575	0.06495	**0.05161**	0.07617	0.05693	0.06294
Squared error (sum) ↓	2107	2601	2639	2264	**762.1**	4455	2603	2536
Squared error / gene ↓	0.1157	0.1428	0.1449	0.1243	**0.04184**	0.2446	0.1429	0.1392
DE AUROC ↑	0.646	0.5022	0.6248	0.5877	0.6105	**0.6898**	0.5008	0.6087
DE average precision ↑	0.09964	0.03378	0.0705	0.05901	0.1034	**0.1285**	0.03355	0.05857
DE overlap ↑	0.1404	0.05992	0.09066	0.0861	0.1396	**0.1713**	0.05954	0.0834
DE precision ↑	**0.1029**	0.06545	0.06396	0.0769	0.08722	0.07667	0.05254	0.07242
Direction agreement ↑	0.8821	0.9671	0.825	0.8052	0.5924	0.908	**0.9832**	0.8151
DE response Spearman ↑	0.4101	0.03617	0.2568	0.2733	0.1235	**0.5478**	0.01812	0.2747
Predicted DE genes	1163	85.23	2341	1095	5194	2349	83.06	2658
PDS cosine ↑	0.615	0.5181	0.554	0.5404	0.5878	**0.6681**	0.5098	0.5606
PDS L_1 ↑	0.6968	0.513	0.5973	0.5385	0.5894	**0.7295**	0.5122	0.5967
PDS squared L_2 ↑	0.6757	0.5134	0.6133	0.5311	0.598	**0.7186**	0.5118	0.5727
ES ↑	**0.4975**	0.3059	0.072	0.2974	0.0838	0.4337	0.3039	0.3369
True DE genes	499.7	499.7	499.7	499.7	499.7	499.7	499.7	499.7

Table 34: PBMC (OOD donor): normalized log counts. Raw evaluator values; best directed scores in bold. PB: PerturbBridge; CF: CellFlow; DOL: DOLORIS; DEP: DEPARTURES; LC: Lingshu-Cell; Ctrl: Control; Lin: Linear (R1). Counts are descriptive.

Metric	PB	CF	STATE	DOL	DEP	LC	Ctrl	Lin
Exact W_1 ↓	41.24	40.94	40.66	41	42.55	40.09	40.94	**36.21**
Exact W_2 ↓	41.59	41.3	41	41.26	42.75	40.22	41.3	**36.55**
Sinkhorn $p = 1$ ↓	41.52	41.21	40.93	41.28	42.84	40.37	41.22	**36.35**
Sinkhorn $p = 2$ (sqrt) ↓	29.39	29.19	28.97	29.16	30.22	28.43	29.19	**25.84**
Marginal W_1 ↓	0.0169	0.0137	0.02511	0.01778	0.02732	0.0271	**0.0137**	0.06302
MMD^2 ↓	0.007077	**0.003457**	0.01101	0.005277	0.02594	0.02137	0.003468	0.101
Response R^2 ↑	-2.715	**-0.483**	-3.014	-1.561	-22.64	-5.614	-0.4859	-7.933
Response Pearson ↑	0.2293	0.005358	0.1622	0.1098	0.08202	**0.2922**	7.06e−4	0.1137
L_1 error (sum) ↓	482.6	429.4	518.6	469.7	**424**	549.9	429.5	608.4
L_1 error / gene ↓	0.0265	0.02358	0.02847	0.02579	**0.02328**	0.03019	0.02358	0.0334
Squared error (sum) ↓	94.25	80.57	109.1	94.3	88.94	129.9	**80.52**	136
Squared error / gene ↓	0.005175	0.004424	0.00599	0.005177	0.004884	0.007133	**0.004421**	0.00747
DE AUROC ↑	0.6384	0.5037	0.6223	0.5775	0.6259	**0.6647**	0.503	0.5572
DE average precision ↑	0.104	0.03758	0.07967	0.06655	0.1241	**0.1267**	0.03749	0.05372
DE overlap ↑	0.1429	0.06602	0.1034	0.09757	0.1624	**0.1656**	0.06578	0.08336
DE precision ↑	**0.108**	0.06907	0.07224	0.09978	0.106	0.08419	0.06394	0.07026
Direction agreement ↑	0.8825	0.9743	0.791	0.801	0.6346	0.8997	**0.9779**	0.6925
DE response Spearman ↑	0.3833	0.01959	0.2632	0.2362	0.1603	**0.4842**	0.00571	0.1035
Predicted DE genes	1284	103	2356	778.2	5147	2379	101.5	4504
PDS cosine ↑	0.6785	0.5045	0.575	0.5421	0.6086	**0.7083**	0.5044	0.5566
PDS L_1 ↑	0.7144	0.5135	0.601	0.5296	0.5896	**0.7383**	0.514	0.559
PDS squared L_2 ↑	0.6996	0.5098	0.6039	0.5421	0.6026	**0.7312**	0.51	0.561
ES ↑	**0.4949**	0.2715	0.07196	0.3129	0.1146	0.4322	0.2683	0.08619
True DE genes	559.1	559.1	559.1	559.1	559.1	559.1	559.1	559.1