RADF: REFERENCE-ANCHORED DYNAMIC FLOW FOR OBSERVED-PERTURBATION SPATIAL COMPLETION

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatial perturbation profiling reveals how a genetic intervention changes transcription across tissue locations, but destructive assays leave many perturbation-by-location response populations unmeasured. We study observed-perturbation spatial completion: for one intervention with a reported response population, use a query coordinate scaffold to estimate the held-out response population. Outputs are unpaired and evaluated as populations; we do not claim unseen-perturbation prediction or cell-to-location matching. RADF is fitted separately for each observed intervention. It first builds a population-valued anchor with equal reference contribution, then applies a bounded coordinate-conditioned residual flow. On matched Perturb-map tasks, RADF reduces macro E-distance, an unordered-population distance, from 138.05 to 40.60 relative to CONCERT (70.6%) and from 62.98 for the strongest adapted single-cell baseline to 40.60. The validation-gated residual improves E-distance over the learned anchor by only 0.24% overall, so reference anchoring supplies the main gain while the flow provides conservative refinement. Code will be made publicly available upon acceptance.

1 INTRODUCTION

Spatial perturbation profiling combines genetic intervention with spatial measurements, revealing how the same intervention can produce different response populations across tissue locations Dhainaut et al. (2022); Binan et al. (2025); Zhang et al. (2026); Shen et al. (2026); Baysoy et al. (2026); Hu et al. (2025). Because the underlying assays use destructive readout and cannot exhaustively cover every perturbation-by-location population, computational completion of missing responses is essential Chi et al. (2026); Hu et al. (2025).

We study a concrete completion problem. For one observed intervention, the input contains a reported response population and a set of query coordinates. The output is the response population requested on that coordinate scaffold. The scaffold is shared context rather than a row-wise truth map: its slots may differ from the native coordinates of held-out observations, and evaluation compares unordered populations. RADF is fitted separately for each observed intervention, so this paper does not claim generalization to an intervention with no same-perturbation reference.

RADF separates support from adaptation: a Sinkhorn-balanced decoder constructs a population-valued anchor, and a bounded coordinate-conditioned residual adapts it to the query geometry. Validation-only selection can return the anchor exactly, and held-out evaluation expression is never used for fitting or selection. The main contributions are:

- **Access-controlled task formulation.** We define observed-perturbation spatial completion with separate fitting, validation, and held-out evaluation expression.
- **Reference-supported population anchoring.** A Sinkhorn-balanced decoder preserves equal reference contribution while constructing a population-valued anchor.
- **Bounded geometry-conditioned refinement.** A dynamic spatial residual flow adapts the anchor within a fixed tube and has an exact anchor fallback.
- **Matched and transfer evaluation.** Matched, cross-slide, and sealed evaluations compare RADF with spatial and task-adapted single-cell predictors, including replicate-shift limitations.

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.

Related work. Many perturbation predictors generate expression from perturbation or cell features; examples include scGen, CPA, GEARS, and CellOT Lotfollahi et al. (2019; 2023); Roohani et al. (2024); Bunne et al. (2023); Chi et al. (2026). In our task-adapted implementations, these models process spots independently and do not use the joint query geometry or a same-perturbation population as explicit support. SpatialProp and CONCERT model tissue-level effects Sun et al. (2025); Lin et al. (2025); empirical-reference and nearest-neighbor baselines retain observed support but do not learn controlled geometry-conditioned refinements. Spatial causal disentanglement and flow-matching methods address complementary causal or distributional questions Megas et al. (2025); Klein et al. (2024). RADF combines reference reuse with bounded spatial adaptation.

2 METHODS

2.1 PROBLEM DEFINITION

Let p denote an observed perturbation and G the number of aligned genes. Its measured target pool contains expression rows with associated spatial coordinates. Let X_p denote the expression matrix of this pool; before model fitting, its rows are partitioned using row identities or coordinates only into a reported reference set, a validation set, and a held-out evaluation set:

$$X_p = R_p \,\dot\cup\, Y_p^{\mathrm{val}} \,\dot\cup\, Y_p^{\mathrm{eval}}. \tag{1}$$

Here $R_p \in \mathbb{R}_+^{m\times G}$, and the three expression subsets in Equation (1) are mutually disjoint; each retains its row identities and coordinates for splitting and evaluation. For n requested outputs, the model input is a query scaffold $S = [s_1, \ldots, s_n]^\top \in \mathbb{R}^{n\times 2}$, which need not equal the native coordinates of held-out rows. RADF predicts a nonnegative expression population of the same output size:

$$\widehat{Y}_p = F_{\theta,p}(S, R_p) \in \mathbb{R}_+^{n\times G}. \tag{2}$$

The perturbation label indexes the task and its reported reference population, and RADF is fitted separately for each observed intervention. The predictor completes an observed response distribution rather than reconstructing paired cells or predicting an intervention for which no same-perturbation reference is available.

No expression-to-location pairing is assumed between rows of R_p and query slots in S. A slot index is an output position, not a biological identity, and the training endpoint is derived from the unordered reference population. Consequently, the method can establish population completion and coordinate dependence, but E-distance alone cannot identify the uniquely correct row-to-location assignment without external spatial supervision.

Fitting and candidate generation may read only (S, R_p). Candidate parameters and predictions are frozen before Y_p^{val} is opened. Validation E-distance selects the residual radius and checkpoint, including an anchor fallback, after which the selection is frozen. Y_p^{eval} is read only once for final scoring and cannot affect fitting or selection. Because the endpoint is an unordered population, no slot-wise truth is assumed or scored.

Figure 1 summarizes this input/output contract and the two-stage predictor.

2.2 E-DISTANCE AND ANCHORED FLOW

Because predicted and observed cells are unpaired and their population sizes may differ, the primary objective is empirical E-distance Székely & Rizzo (2013). Its empirical V-statistic is stated in the supplementary material and uses Euclidean distance in aligned gene space. Evaluation clamps only negative numerical round-off to zero; lower values are better. Metrics are computed per task/split unit and macro-averaged so that larger target pools do not dominate. Because E-distance compares unordered populations, it is invariant to output-row permutations and cannot by itself establish coordinate sensitivity.

RADF begins from an anchor population $A_p \in \mathbb{R}_+^{n\times G}$ produced from R_p. Both the learned anchor weights and the fixed training coupling use balanced matrices Cuturi (2013) in

$$\mathcal{U}_{n,m} = \left\{ C \geq 0 : C\mathbf{1}_m = \mathbf{1}_n,\; C^\top \mathbf{1}_n = \frac{n}{m}\mathbf{1}_m \right\}. \tag{3}$$

Equivalently, C/n is a unit-mass balanced coupling with uniform marginals $1/n$ and $1/m$; meanwhile, non-negativity and $C\mathbf{1}_m = \mathbf{1}_n$ make each row of CR_p a convex combination of reference profiles, while the column constraint assigns every reference equal total contribution n/m. For a fixed entropic optimal-transport coupling $\Gamma_p \in \mathcal{U}_{n,m}$ between A_p and R_p, the coupled endpoint is $B_p = \Gamma_p R_p$. Conditional flow matching uses the straight path and its constant target velocity Lipman et al. (2023); Liu et al. (2023):

$$Z_p(t) = (1-t)A_p + tB_p, \qquad t \in [0,1]. \tag{4}$$

$$u_p(t) = \frac{\mathrm{d}Z_p(t)}{\mathrm{d}t} = B_p - A_p. \tag{5}$$

The learned field $v_\theta(Z_p(t), t; S, R_p)$ approximates this velocity. At inference, projected Euler updates keep every state inside the elementwise anchor tube

$$\mathcal{T}_\rho(A_p) = \{Z \geq 0 : |Z - A_p| \leq \rho(A_p + \mathbf{1})\}. \tag{6}$$

Consequently, zero velocity returns A_p exactly, while every nonzero contextual refinement is explicitly bounded.

2.3 Overview

RADF has a deliberately asymmetric pipeline. The reported set R_p carries response support; the learned flow handles only its spatial adjustment. A Sinkhorn-balanced allocation first produces n anchor profiles A_p, with fixed total contribution from every reference. A conditional velocity network then rebuilds relations from the evolving expression and query coordinates S, followed by projected Euler updates inside an anchor tube. Finally, validation E-distance chooses the radius and checkpoint, or keeps A_p when none of the learned updates helps. The prediction map is $R_p \to W_\phi \to A_p \to \widehat{Y}_p$, with $\widehat{Y}_p$ conditioned on (S, R_p, t). All transformations remain in expression space. Coordinates organize relations among output slots but are not themselves transported. Figure 1 follows the same division through training, validation-only selection, and frozen inference.

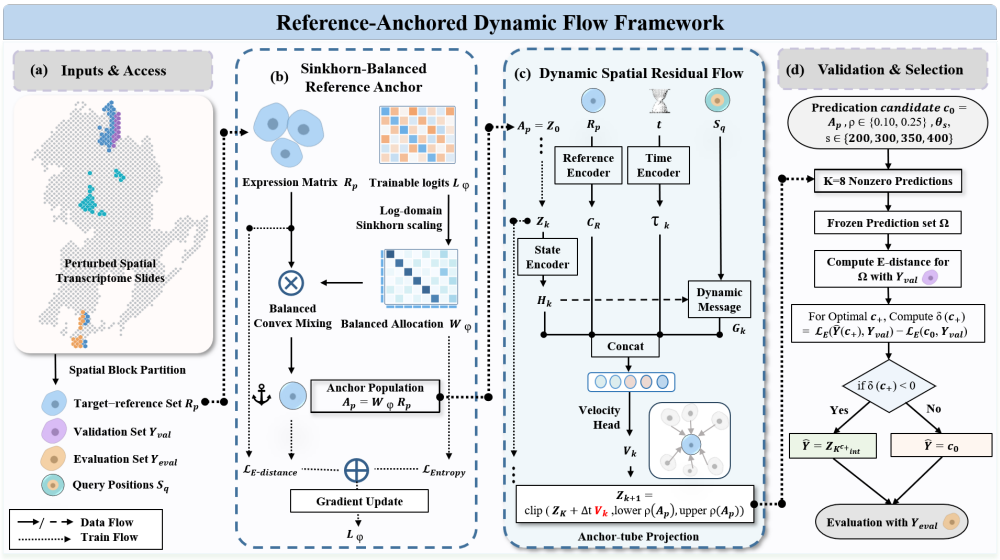


Figure 1: Overview of RADF. The method first reuses the reported response population and then applies only a bounded spatial adjustment. (a) Spatial-block partitioning separates the target-reference, validation, and held-out evaluation sets, with query positions provided as spatial context. (b) Sinkhorn-balanced convex allocation transforms reference profiles into the anchor population A_p. (c) Starting from A_p, dynamic spatial messages, reference context, and time determine bounded velocity updates within the anchor tube. (d) Validation E-distance selects among frozen radius–checkpoint candidates, retaining a nonzero flow only when it improves upon the anchor before final evaluation.

2.4 SINKHORN-BALANCED REFERENCE ANCHOR

Let $R_p = [r_1, \ldots, r_m]^\top \in \mathbb{R}_+^{m \times G}$, and let n be the requested number of output samples. The decoder maintains unconstrained trainable logits $L_\phi \in \mathbb{R}^{n \times m}$. Log-domain Sinkhorn scaling converts $\exp(L_\phi)$ to a plan with uniform row and column marginals Cuturi (2013). Alternating row and column log-normalizations remain differentiable with respect to L_ϕ; after scaling the unit-mass plan by n, the resulting allocation matrix satisfies $W_\phi \in \mathcal{U}_{n,m}$, with the balance constraints already defined in Equation (3). The exact log-domain recurrences are recorded in the separate supplementary document.

Each row is a probability vector over references, and each reference contributes total mass n/m across the output population. Their nonnegative convex allocation is

$$A_p = W_\phi R_p \in \mathbb{R}_+^{n \times G}. \tag{7}$$

Every anchor row lies in the convex hull of the reported expressions. Row i is an output slot, not a biological cell paired to r_i. Associating that slot with query coordinate s_i lets the relational flow adjust the population by context without inventing its support.

The allocation logits are learned from reported reference expression alone by minimizing

$$\mathcal{L}_{\text{anchor}}(\phi) = D_{\text{E}}(A_p, R_p) + \lambda_H H(W_\phi). \tag{8}$$

We use stabilized negative entropy,

$$\begin{aligned} H(W_\phi) &= -\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} \overline{W}_{\phi,ij} \log \overline{W}_{\phi,ij}, \\ \overline{W}_{\phi,ij} &= \max(W_{\phi,ij}, 10^{-12}). \end{aligned} \tag{9}$$

Here $\lambda_H = 0.05$. Because H enters with a positive coefficient and the objective is minimized, it discourages overly diffuse row allocations. Without this term, convex mixtures can smooth distinct references into similar outputs while still matching their center. The column constraint prevents sharper rows from dropping any reference cell at the population level. This is an allocation regularizer, not a biological similarity prior. The implementation uses 160 Sinkhorn iterations at each decoder update and fits the logits for 400 Adam steps with learning rate 0.05.

In exact arithmetic, balance gives the population-mean identity $(1/n)\mathbf{1}_n^\top A_p = (1/m)\mathbf{1}_m^\top R_p$. The implementation uses a finite number of log-domain Sinkhorn iterations, so empirical marginal statements hold up to the resulting numerical residual. The anchor preserves the reference mean and uniform reference contribution by construction, including when n differs from m. Its n rows retain a population-valued representation rather than collapsing to one centroid. This identity preserves allocation over reference samples, not biological RNA mass, total expression, or gene-wise library size; throughout, balanced therefore refers only to the empirical sample measure represented by R_p. The flow then provides bounded query-context refinement.

2.5 BOUNDED DYNAMIC RELATIONAL FLOW

Let $Z(t) \in \mathbb{R}_+^{n \times G}$ be the evolving population, initialized by $Z(0) = A_p$. RADF parameterizes a conditional velocity field

$$\frac{\mathrm{d}Z(t)}{\mathrm{d}t} = v_\theta(Z(t), t; S, R_p). \tag{10}$$

The field recomputes its relation matrix from the current state at every integration step. For row $z_i(t)$, the state encoder is

$$h_i(t) = \text{Enc}_x(\log(\mathbf{1} + \max(z_i(t), 0))) \in \mathbb{R}^d. \tag{11}$$

Query coordinates are centered and scaled per axis, with the scale lower-bounded by 10^{-3}:

$$\widetilde{s}_i = \frac{s_i - \text{mean}(S)}{\max(\text{std}(S), 10^{-3})}. \tag{12}$$

With learned linear maps Q_h and K_h, let $q_i(t) = Q_h h_i(t)$ and $k_j(t) = K_h h_j(t)$. The off-diagonal relation scores are

$$e_{ij}(t) = \frac{q_i(t)^\top k_j(t)}{\sqrt{d}} - \|\widetilde{s}_i - \widetilde{s}_j\|_2^2, \quad i \neq j. \tag{13}$$

For $n > 1$, self-scores are masked by setting $e_{ii}(t) = -10^9$; the normalized weights are

$$P_{ij}(t) = \frac{\exp(e_{ij}(t))}{\sum_{\ell=1}^{n} \exp(e_{i\ell}(t))}. \tag{14}$$

The spatial penalty has fixed coefficient one after coordinate standardization. Because Q_h and K_h are distinct and normalization is row-wise, $P(t)$ is generally directed, dense, and asymmetric. Consequently, $P(t)\mathbf{1}_n = \mathbf{1}_n$, and the message $g_i(t) = \sum_j P_{ij}(t)h_j(t)$ is a row-stochastic aggregation. For a single output, the unique relation weight is set to one. Since $h_i(t)$ changes with $Z(t)$, $P(t)$ is state-dependent even though the coordinates remain fixed.

The reported reference set enters through a global distributional summary. Let μ_R and σ_R be its gene-wise mean and standard deviation. We encode them as $c_R = \mathrm{Enc}_R(\log(\mathbf{1} + \max(\mathrm{concat}(\mu_R, \sigma_R), 0)))$, while $\tau_i(t) = \mathrm{Enc}_t(t)$ encodes time. The same c_R is broadcast to all output slots, providing a shared summary of the observed response distribution. Slot-specific adaptation is then determined by the evolving state, its spatial message, and continuous time.

The velocity head combines the local state, its dynamic relational message, the shared reference summary, and time:

$$v_{\theta,i}(t) = \mathrm{MLP}_v(\mathrm{concat}(h_i(t), g_i(t), c_R, \tau_i(t))) \in \mathbb{R}^G. \tag{15}$$

The state encoder maps G genes to d features, the reference encoder maps $2G$ moments to d, and the time encoder maps one scalar to d. The velocity head maps the concatenated $4d$ features through a $2d$ GELU hidden layer to G outputs. All three encoders use GELU activations; the state and reference encoders additionally use layer normalization. The benchmark uses $d = 32$. The final linear layer of MLP_v is initialized to zero, so the untrained residual path initially returns the anchor.

Rather than solving an unrestricted neural ODE Chen et al. (2018), RADF uses K_{int} projected Euler steps. The fixed elementwise limits in Equation (6) are $\mathrm{lower}_\rho = \max(0, A_p - \rho(A_p + \mathbf{1}))$ and $\mathrm{upper}_\rho = A_p + \rho(A_p + \mathbf{1})$. The additive one permits a bounded update even for genes whose anchor value is near zero, while the multiplicative part scales the allowed deviation with expression magnitude. These bounds are fixed from A_p and are not recomputed from later states.

With $Z_0 = A_p$, $\Delta t = 1/K_{\mathrm{int}}$, and midpoint time $t_k = (k + \frac{1}{2})\Delta t$, inference applies

$$\begin{aligned} Z_{k+1} = \mathrm{clip}\Big(& Z_k + \Delta t\, v_\theta(Z_k, t_k; S, R_p), \\ & \mathrm{lower}_\rho, \mathrm{upper}_\rho \Big), \\ & k = 0, \ldots, K_{\mathrm{int}} - 1. \end{aligned} \tag{16}$$

The final prediction is $\widehat{Y}_p = Z_{K_{\mathrm{int}}}$. Clipping is elementwise and is performed after every step; $K_{\mathrm{int}} = 8$ in the benchmark. Hence every finite prediction is nonnegative and remains inside the chosen anchor tube. The tube bounds coordinate-conditioned deviations from the anchor; it does not imply that a refined prediction remains inside the convex hull or empirical support of the reference population. Zero velocity leaves $Z_k = A_p$, giving an exact structural fallback; nonzero updates are coordinate-conditioned relational refinements in expression space.

2.6 Training and Conservative Selection

RADF trains the residual field against a fixed balanced endpoint derived from A_p and R_p. We standardize their pooled expression gene-wise using $Z_p = \mathrm{concat}(A_p, R_p)$, $\mu = \mathrm{mean}(Z_p)$, $\sigma = \max\big(\mathrm{std}(Z_p), 1\big)$, $\overline{A} = (A_p - \mu)/\sigma$, and $\overline{R} = (R_p - \mu)/\sigma$. The coupling cost is $C_{ij} = \|\overline{A}_i - \overline{R}_j\|_2/\sqrt{G}$. Log-domain Sinkhorn scaling of $-C/\epsilon$, with $\epsilon = 0.5$ and 120 scaling iterations, produces a fixed approximately balanced Γ_p, and $B_p = \Gamma_p R_p$. This endpoint supervises reference-supported distributional refinement. Because both A_p and B_p are derived from R_p, the CFM target is not location-supervised; coordinates influence the field only through Equations (13)–(14).

For each output row, sample $t_i \sim \mathrm{Uniform}(0, 1)$, define $z_i(t_i) = (1 - t_i)a_i + t_i b_i$ and $u_i = b_i - a_i$, and stack these rows as $Z_{\mathbf{t}}$ and U. This row-wise time sampling is a training approximation to

the population-coupled field; inference uses shared midpoint times for the whole population. We therefore do not interpret the flow as an exact discretization of the independently sampled training paths. The detached normalization scales are $s_v = \max\Big(\sqrt{\mathrm{mean}((B_p - A_p)^2)}, 1\Big)$ and $s_E = \max(D_{\mathrm{E}}(A_p, R_p), 1)$. Training minimizes the normalized conditional flow-matching loss and the E-distance of the projected prediction to the reported references Lipman et al. (2023); Liu et al. (2023):

$$\mathcal{L}_{\mathrm{CFM}}(\theta) = \mathrm{mean}\left[\left(\frac{v_\theta(Z_{\mathbf{t}}, \mathbf{t}; S, R_p) - U}{s_v}\right)^2\right]. \tag{17}$$

$$\mathcal{L}_{\mathrm{ref}}(\theta, \rho) = \frac{D_{\mathrm{E}}(\widehat{Y}_p(\theta, \rho), R_p)}{s_E}. \tag{18}$$

$$\mathcal{L}_{\mathrm{RADF}} = \mathcal{L}_{\mathrm{CFM}} + 0.5\,\mathcal{L}_{\mathrm{ref}}. \tag{19}$$

The fixed coupling, anchor, and scales are detached from gradient updates. Models are optimized for 400 Adam steps with learning rate 10^{-3}, with gradient norm clipped to five. The differentiable algebraic E-distance is used during optimization without the evaluation-only nonnegative clamp. Separate models with the same deterministic initialization are trained for $\rho \in \{0.10, 0.25\}$ because projection changes $\mathcal{L}_{\mathrm{ref}}$; predictions and states are saved at checkpoints $\{200, 300, 350, 400\}$. Together with the anchor candidate c_0, this yields a frozen candidate set before validation expression is read.

For validation set Y_p^{val}, define each candidate's improvement as

$$\delta_c = D_{\mathrm{E}}(\widehat{Y}_p^{\,c}, Y_p^{\mathrm{val}}) - D_{\mathrm{E}}(A_p, Y_p^{\mathrm{val}}). \tag{20}$$

The best nonzero candidate is

$$c^+ = \arg\min_{c \in \Omega_{\mathrm{nonzero}}} D_{\mathrm{E}}(\widehat{Y}_p^{\,c}, Y_p^{\mathrm{val}}). \tag{21}$$

Conservative selection chooses

$$c^\star = \begin{cases} c^+, & \delta_{c^+} < 0, \\ c_0, & \text{otherwise.} \end{cases} \tag{22}$$

Ties among nonzero candidates prefer the smaller radius and then the earlier checkpoint. This strict comparison implements conservative selection: a residual is retained only when validation supports improvement over zero residual. The selected prediction is frozen before held-out evaluation expression is opened. PCC and MAE are reported descriptively but do not enter training, selection, or fallback.

3 Experiments

Evaluation covers matched Perturb-map tasks, adapted single-cell predictors, cross-slide transfer, and sealed SPAC-seq. E-distance is the primary endpoint. Mean PCC correlates predicted and observed gene-wise population means, and mean MAE measures their absolute difference. For spatial diagnostics, pooled query and truth coordinates are divided at the median projection onto their first spatial principal axis. SS-ED averages the two within-stratum E-distances with truth-count weights; spatial contrast error compares the predicted and observed separation between strata. Each stratum must contain at least four samples.

3.1 Experimental Setup

The matched Perturb-map benchmark Dhainaut et al. (2022) contained five slide-perturbation tasks: Tgfbr2 on GSM5808054, GSM5808055, and GSM5808057, and Ifngr2 on GSM5808055 and GSM5808056. Each task used the 31 nearest KP/tumor query spots and 2,000 ordered genes in frozen raw-count space. Target rows were split into 60% references, 20% validation, and the remainder evaluation. Random splits used seed 20260731; spatial-block splits followed coordinate order, producing ten units. Predictions use 31 fixed scaffold slots, whereas evaluation populations contain 9–18 unpaired rows; we therefore use population endpoints and make no unique row-to-location claim. Official CONCERT used its fixed multi-kernel implementation after held-out rows

Table 1: Matched Perturb-map E-distance, reported as mean ± population standard deviation across task/split units (overall $n = 10$; each partition $n = 5$). Direct-anchor and simpler-flow control means are reported in the text and supplementary material. Lower is better. PCC and MAE are reported in the text and supplementary material.

Method	Overall	Random	Spatial block
CONCERT	138.05 ± 55.74	117.15 ± 28.02	158.95 ± 67.50
scGen	$\underline{62.98} \pm 34.29$	$\underline{43.79} \pm 14.79$	$\underline{82.17} \pm 37.36$
biolord	102.02 ± 41.26	86.65 ± 21.72	117.40 ± 49.59
CellOT	461.26 ± 250.84	447.28 ± 307.12	475.25 ± 176.43
RADF (Ours)	$\mathbf{40.60} \pm 23.92$	$\mathbf{22.64} \pm 11.02$	$\mathbf{58.55} \pm 19.46$

were removed from its fit file Lin et al. (2025). CellOT, scGen, and biolord used fixed configurations with the same source, reference, genes, and held-out-expression access Bunne et al. (2023); Lotfollahi et al. (2019); Piran et al. (2024). The matched ablation set additionally compared the direct balanced anchor, a self-only residual flow, and a fixed-spatial flow under the same split units.

All candidate predictions were generated from the reported reference expression and query coordinates before target expression was opened. The candidate grid (two tube radii, four checkpoints, and the exact anchor) was frozen in advance. Validation E-distance selected a nonzero candidate only when it strictly improved over the anchor; otherwise the prediction was set to the exact anchor. After this selection was frozen, evaluation expression was opened once for scoring. Thus the protocol permits transductive use of query coordinates, but never uses evaluation expression for fitting, candidate generation, or selection.

External experiments added five leave-one-slide-out Perturb-map units on 1,053 shared genes, each with 32 references from other slides. SPAC-seq Zhang et al. (2026) used the released expression matrix on 1,718 genes intersecting the frozen panel for sgZc3h12a, sgZhx2, and sgGata3 across days 4, 7, and 10. The sealed same-slide track contained 18 units with 32 references and 64 disjoint evaluation spots; the cross-replicate track contained nine units with 64 replicate-1 references and 128 replicate-2 evaluation spots. Radius 0.25 and checkpoint 300 were frozen from Perturb-map validation. Absolute E-distances are interpreted only within a dataset and preprocessing space.

Metrics were computed independently for each unit and then macro-averaged. Paired intervals used 10,000 cluster-bootstrap draws over tasks, held-out slides, or days. These intervals describe cross-unit consistency; no spot-level significance tests were used.

3.2 RADF Outperforms CONCERT on Matched Tasks

RADF has lower E-distance than CONCERT in every matched unit (Table 1). Overall macro error falls from 138.05 to 40.60, a reduction of 70.60%. The reduction is 80.70% on random splits and 63.20% on spatial blocks; the task-cluster interval for CONCERT minus RADF is $[64.39, 132.39]$. Population means improve as well: PCC rises from 0.92 to 0.97 and MAE drops from 0.68 to 0.38. The scale and consistency of this gap favor a reference-supported predictor over relearning the full response distribution with a general conditional generator.

The same separation is visible in Table 1: overall E-distance is 1.55 times RADF's error for the strongest adapted single-cell baseline, 3.40 times for CONCERT, and 11.36 times for CellOT.

3.3 Why Single-Cell Predictors Fall Short

This spatial completion setting lies outside the native design of single-cell perturbation predictors. scGen applies a shared latent displacement, biolord models condition attributes without tissue relations, and CellOT maps dissociated populations without query coordinates. Their Visium adaptation treats spots as independent pseudo-cells. Two query sets with the same spot profiles but different tissue arrangements therefore receive the same prediction. RADF instead retains the observed response population and rebuilds relations over the supplied query geometry.

That difference in task design is reflected in the matched results. RADF records 40.60 overall E-distance, compared with 62.98 for scGen, 102.02 for biolord, and 461.26 for CellOT, and wins all

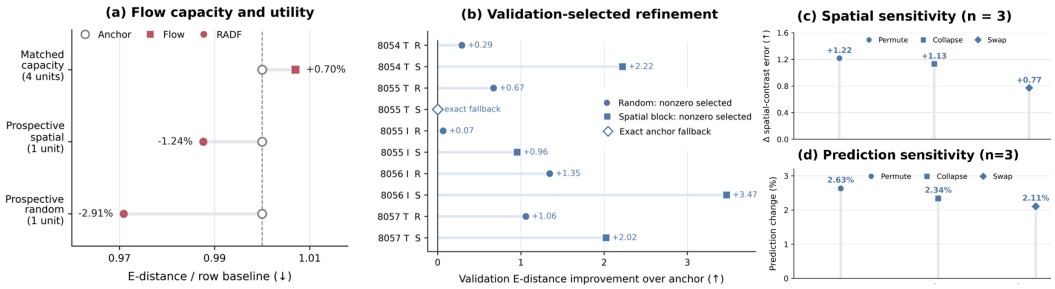


Figure 2: Flow evidence separated by protocol. (a) Candidate E-distance normalized to the row baseline for a matched-capacity audit and two untouched prospective units; the protocols are not pooled. (b) Validation-selected E-distance improvement over the learned anchor for all ten matched units; the exact-anchor fallback is shown explicitly. Panel (b) plots the positive quantity $D_E(A_p, Y^{\mathrm{val}}) - D_E(\widehat{Y}^c, Y^{\mathrm{val}}) = -\delta_c$. (c) Coordinate interventions on the three units satisfying the common-stratum support rule; these are diagnostic units, not a new main benchmark. The interventions change both regional contrast error and predictions, establishing coordinate sensitivity without implying a unique expression-to-location assignment.

ten units against each method. Equal expression and reference access is therefore not enough when the predictor has no mechanism to condition a population completion on joint tissue geometry.

The advantage transfers beyond the matched split. RADF cuts leave-one-slide-out CONCERT error from 155.30 to 71.20. On sealed same-slide SPAC-seq, it is the strongest learned method at 0.47 E-distance, against 3.69 for CONCERT and 1.15 for scGen. Under cross-replicate variation, RADF remains ahead of CONCERT and has the lowest spatial contrast error.

3.4 What the Flow Contributes

Figure 2 summarizes the capacity audit, validation-selected marginal gains, and coordinate interventions. The anchor-versus-RADF average measures the realized marginal value of validation-gated refinement above an already strong anchor; it does not measure the representational capacity of the flow. In a matched-architecture development audit over four earlier units, a relational flow reached 35.92 E-distance versus 35.67 for the reference decoder, a gap of 0.70%. On two untouched prospective units, an anchored relational flow reduced E-distance by 1.24% on a spatial split and 2.91% on a random split. These protocols are reported separately from the paper benchmark in the supplementary material.

On the ten matched units, validation selected nonzero flow in nine and returned the exact anchor in one. The direct anchor, self-only flow, fixed-spatial flow, and full RADF have overall E-distances of 40.693, 40.624, 40.804, and 40.595, respectively; on spatial blocks they score 58.209, 58.572, 58.770, and 58.550. Full RADF therefore has the lowest overall error, but no flow variant improves the anchor uniformly on spatial blocks. The selected RADF residuals changed 2.2–7.9% of the anchor magnitude and improved over the learned anchor by 0.24% overall, decomposing into +2.31% on random splits and -0.58% on spatial blocks. Evaluation favored the selected residual in five units, tied it in one, and favored the anchor in four. These results support a conservative, task-dependent refinement interpretation rather than a claim that dynamic flow supplies the headline gain.

E-distance alone cannot reveal where an output belongs because row permutation leaves it unchanged. Across the three supported units, permuting, collapsing, or swapping coordinates changed predictions by 2.11–2.63% and increased spatial-contrast error by 0.77–1.22. The fitted flow therefore uses spatial inputs rather than acting as a coordinate-invariant decoder. Complementary SS-ED outcomes and all unit-level selection results are reported in the separate supplementary document.

3.5 Transfer Across Spatial Contexts

The advantage over CONCERT persists as contexts separate: RADF reduces E-distance by 80.70% on random partitions, 63.20% on held-out spatial blocks, and 54.2% when an entire slide is left out.

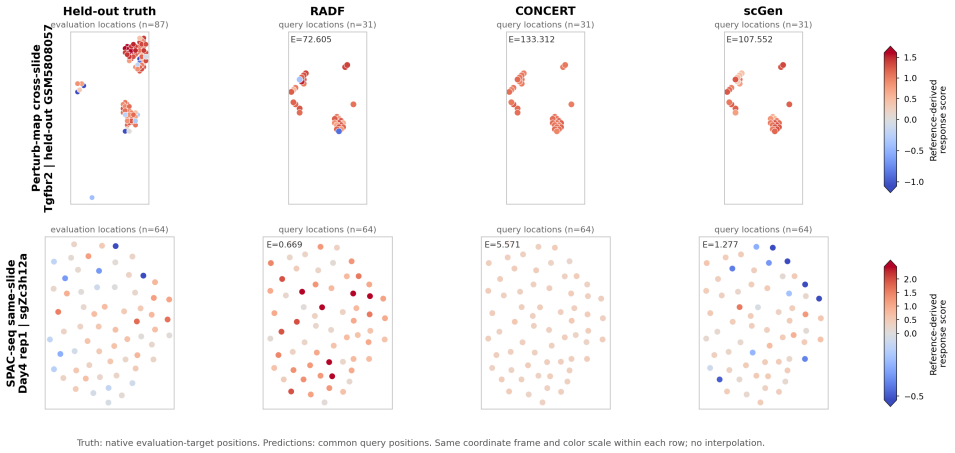


Figure 3: Qualitative spatial completion on deterministic first-manifest units. Colors are fit-side reference-derived response scores, defined without held-out expression. Truth uses native evaluation positions; predictions use common query positions and are unpaired. The plot shows population-level coordinate-conditioned agreement, not a unique row-to-location correspondence.

On supported units, spatially stratified SS-ED falls from 138.40 to 65.91 on matched tasks and from 3.91 to 0.84 on sealed same-slide SPAC-seq; full regional results are reported in the supplementary material. Figure 3 gives a qualitative view with native truth coordinates separated from common query coordinates. Cross-slide evaluation takes 38.9 seconds for RADF versus 63.6 for CONCERT, and SPAC-seq takes 39.6 versus 84.2 seconds. Cross-replicate variation remains the hardest setting and a target for slide-aware normalization.

3.6 Downstream Utility: Held-Out Response Recovery

The practical use of RADF is to complete an unmeasured response population and then compare gene or pathway summaries across spatial regions. In a retrospective audit that was frozen after prediction, RADF reached response Spearman 0.690, Top-50 response recovery 0.638, and pathway-gene Spearman 0.841, compared with 0.423, 0.330, and 0.569 for CONCERT. It matched the held-out pathway-response direction in 9/10 units. These measurements evaluate recovery of observed response structure and support region or pathway prioritization; they do not establish causal activation, a unique cell-to-location assignment, or pathway discovery. The learned anchor was essentially tied with RADF on these biological outcomes, reinforcing that the downstream utility primarily comes from reference-supported anchoring.

4 Conclusion

RADF completes observed-perturbation responses from same-perturbation references. Its balanced anchor supplies the main gain, while a bounded flow adds conservative coordinate-conditioned refinement. This is population-level completion, not unique row-to-location recovery.

Reproducibility Statement

All data sources, gene panels, split rules, candidate grids, optimization settings, and leakage controls are specified in Methods and the supplementary material. Anonymous source release is planned upon acceptance.

AI USE STATEMENT

For this revision, OpenAI Codex was used to assist with language editing for readability, manuscript organization, LaTeX format conversion, and checking the placement and wording of visual materials. It was not used for data generation, data cleaning or reformatting, theory or concept development, hypothesis formation, mathematical claims or proofs, experiment design, implementation or evaluation, qualitative or thematic analysis, result interpretation, or translation. The authors reviewed all AI-assisted text and checks and take responsibility for the final content, claims, and artifacts.

REFERENCES

Alev Baysoy, Xiaolong Tian, Paul Renauer, Feifei Zhang, Zhiliang Bai, Hao Shi, Mingyu Yang, Dingyao Zhang, Miao Liu, Haikuo Li, Bo Tao, Archibald Enninful, Yao Lu, Fu Gao, Guangchuan Wang, Wanqiu Zhang, Thao Tran, Nathan Heath Patterson, Jie Sheng, Shuozhen Bao, Chuanpeng Dong, Shan Xin, Binfan Chen, Mei Zhong, Sherri Rankin, Cliff Guy, Yan Wang, Jon P. Connelly, Shondra M. Pruett-Miller, Daifeng Wang, Mina Xu, Mark B. Gerstein, Hongbo Chi, Sidi Chen, and Rong Fan. Large-scale, spatially resolved panoramic CRISPR screening in native tissue environments using Perturb-DBiT. Nature Biotechnology, 2026. Advance online publication. doi: 10.1038/s41587-026-03127-y.

Loïc Binan, Aiping Jiang, Serwah A. Danquah, Vera Valakh, Brooke Simonton, Jon Bezney, Robert T. Manguso, Kathleen B. Yates, Ralda Nehme, Brian Cleary, and Samouil L. Farhi. Simultaneous CRISPR screening and spatial transcriptomics reveal intracellular, intercellular, and functional transcriptional circuits. *Cell*, 188(8):2141–2158.e18, 2025. doi: 10.1016/j.cell.2025.02.012.

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

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural Ordinary Differential Equations. In *Advances in Neural Information Processing Systems*, 2018.

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, 2026. doi: 10.1609/aaai.v40i25.39190.

Marco Cuturi. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In *Advances in Neural Information Processing Systems*, 2013.

Maxime Dhainaut, Samuel A. Rose, Guray Akturk, Aleksandra Wroblewska, Sebastian R. Nielsen, Eun Sook Park, Mark Buckup, Vladimir Roudko, Luisanna Pia, Robert Sweeney, Jessica Le Berichel, C. Matthias Wilk, Anela Bektesevic, Brian H. Lee, Nina Bhardwaj, Adeeb H. Rahman, Alessia Baccarini, Sacha Gnjatic, Dana Pe'er, Miriam Merad, and Brian D. Brown. Spatial CRISPR genomics identifies regulators of the tumor microenvironment. *Cell*, 185(7):1223–1239.e20, 2022. doi: 10.1016/j.cell.2022.02.015.

Mengwei Hu, Yi Cui, Qianhui Huang, Khoi Chu, Sierra McKinzie, Michael Patrick, Sharanya Iyengar, Maerjianghan Abuduli, Marianne Spatz, Nandita Joshi, Brendan Miller, Shams Vellarikkal, Timothy Riordan, Danny Bitton, Jan Lubojacky, Iya Khalil, Federica Piccioni, Michael Rhodes, Alex Tamburino, Shanshan He, Joseph Beechem, and Vanessa Peterson. SPACE: multimodal spatial CRISPR screening with whole-transcriptome readout at subcellular resolution in 3D models, 2025.

D. Klein, U. Düssica, T. Theis, and F. Cuturi. GENOT: Entropic (Gromov) Wasserstein flow matching with applications to single-cell genomics. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Xiang Lin, Zhenglun Kong, Soumya Ghosh, Manolis Kellis, and Marinka Zitnik. CONCERT predicts niche-aware perturbation responses in spatial transcriptomics, 2025.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow Matching for Generative Modeling. In *International Conference on Learning Representations*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. In *International Conference on Learning Representations*, 2023.

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

Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Leon Hetzel, Yuge Ji, Ignacio L Ibarra, Sanjay R Srivatsan, Mohsen Naghipourfar, Riza M Daza, Beth Martin, Jay Shendure, Jose L McFaline-Figueroa, Pierre Boyeau, F Alexander Wolf, Nafissa Yakubova, Stephan Günnemann, Cole Trapnell, David Lopez-Paz, and Fabian J Theis. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology*, 19(6):MSB202211517, 2023. doi: 10.15252/msb.202211517.

S. Megas, D. Chen, K. Polanski, M. Eliasof, C.-B. Schönlieb, and S. Teichmann. Estimation of single-cell and tissue perturbation effect in spatial transcriptomics via spatial causal disentanglement. In *International Conference on Learning Representations*, 2025.

Zoe Piran, Niv Cohen, Yedid Hoshen, and Mor Nitzan. Disentanglement of single-cell data with biolord. *Nature Biotechnology*, 42(11):1678–1683, 2024. doi: 10.1038/s41587-023-02079-x.

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

Kimberle Shen, Wan Yi Seow, Choong Tat Keng, Michelle Gek Liang Lim, Daryl Shern Lim, Ke Guo, Amine Meliani, Muhammad Irfan Bin Hajis, Bolun Wang, Shyam Prabhakar, Kok Hao Chen, and Wei Leong Chew. Spatial perturb-seq: single-cell functional genomics within intact tissue architecture. Nature Communications, 2026. Article in press. doi: 10.1038/s41467-026-69677-6.

Eric D. Sun, Alejandro Buendia, Anne Brunet, and James Zou. SpatialProp: tissue perturbation modeling with spatially resolved single-cell transcriptomics, 2025.

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

Haorui Zhang, Zongxu Zhang, Peiyu Wang, Tian Xu, Xiaoyu Chen, Yanping Zhao, Siyu Lin, Wenjie Cai, Pengfei Ren, Ce Luo, Peng Zhang, Yunfeng Wang, Sen Hou, Yahui Zhao, Hu Zeng, Zhihua Liu, Cunyu Wang, Zhidong Gao, Yu Feng, Deng Pan, and Zexian Zeng. Uncovering spatially resolved functional genomics with CRISPR screen sequencing. *Cell*, 189(15):4594–4618.e48, 2026. doi: 10.1016/j.cell.2026.04.049.