SILTA: Reference Learning and Target Adaptation for Spatial Transcriptomic Deconvolution

Anonymous authors
Paper under double-blind review

Abstract

Most sequencing-based spatial transcriptomics platforms pool multiple cell-types per capture location, requiring deconvolution to estimate cell-type proportions. Existing methods suffer from spot compositions that are sparse and long-tailed: one type often dominates, yet many methods predict smooth mixtures, and similar gene signatures blur related types. Although neighboring spots share spatial context, spot-wise inference is sensitive to local noise, and these errors accumulate into sample-level composition drift. Platform differences complicate applying composition mappings learned from scRNA-seq references to unlabeled ST targets. To this end, we propose a Two-stage learning model SILTA, **S**patial **I**nference by combining composition **L**earning and **T**arget **A**daptation, which jointly tackles issues above. Stage I learns fine-type boundaries from structured, diffusion-augmented pseudo-spots with the spatial encoder frozen. Stage II refines target proportions from expression and tissue geometry through hierarchical parameter updates, gradual unfreezing, and reference replay. Expression-spatial fusion combines reference composition with target context, while qNB and post-hoc calibration align predictions with measured expression and correct global composition bias. SILTA achieves the best mean for all four primary metrics among evaluated methods on seqFISH, osmFISH and cross-animal Moffitt. We verify both stages' roles through perturbation experiments. Without ground-truth proportions, SILTA demonstrates strong performance in marker-gene concordance across three datasets and in pathology-region classification on HBC, extending its gains to real tissues across assay platforms. The code is at: https://anonymous.4open.science/r/SILTA-SpaDeconv-A38C/.

1 Introduction

Spatial transcriptomics (ST) captures gene expression while retaining tissue coordinates. Imaging assays measure selected transcripts near single-cell resolution, whereas sequencing assays capture broader gene panels within fixed spots (Ståhl et al., 2016). A spot can contain several cell types, so its expression represents a cellular mixture. Cell-type deconvolution estimates the proportion of each type per spot and supports tissue mapping, rare-cell localization, and spatial disease analysis, all of which depend on accurate spot proportions (Gaspard-Boulinc et al., 2025; Li et al., 2023).

Reference-based methods pair an annotated single-cell RNA-seq (scRNA-seq) reference with target ST expression and model each spot as a mixture of reference cell-type profiles. RCTD and cell2location fit this formulation with Poisson or negative-binomial count models (Cable et al., 2022; Kleshchevnikov et al., 2022). These remain established statistical approaches for reference-based deconvolution across current benchmarks.

Despite progress, three categories of difficulty remain. The first concerns within-spot composition: one or two cell types dominate most spots, yet models without explicit composition priors predict diffuse proportions whose per-spot estimates may appear biologically plausible while the tissue-wide type distribution is over-smoothed. Fine-type collinearity compounds this issue, as closely related types share nearly collinear expression signatures that make their individual contributions hard to separate (Li et al., 2022; Gaspard-Boulinc et al., 2025). The second concerns spatial organization: methods that infer each spot independently discard the spatial gradients that distinguish cortical layers, tumor boundaries, and vascular niches, fragmenting composition maps and weakening the re-

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.

gional patterns that downstream analyses depend on (Ma & Zhou, 2022; Li & Luo, 2024). The third concerns reference-target isolation: scRNA-seq references and spatial targets differ in sequencing depth, gene panel, and capture efficiency, yet most pipelines do not explicitly separate the roles of the two data sources. Direct joint training can use platform differences to compensate for errors in inferred composition. The model may then fit target expression while systematically shifting tissue-wide proportions away from biologically plausible patterns. This risk motivates decoupling reference supervision from target adaptation. SILTA preserves the reference-learned proportion mapping through role-separated parameter updates and reference replay while learning from expression and tissue geometry in target tissues without proportion labels.

Our contributions. First, we develop reference-preserving target adaptation through role-specific parameter updates and reference replay. Stage I learns a reference composition mapping from structured, diffusion-augmented pseudo-spots. Stage II combines target-expression alignment and spatial context with continued reference supervision to refine target proportions without requiring proportion labels for the target tissue, while retaining reference-learned cell-type information.

Second, we develop a cooperative dual-expert architecture supporting target adaptation through expression-spatial fusion with an explicit, removable spatial correction. The expression expert supplies reference composition information, while the spatial expert models target spot relationships through distance-aware attention. Their fusion refines target proportions from expression and tissue geometry, and its removable spatial path makes the correction testable.

Third, we combine qNB alignment using fixed reference signatures with post-hoc hierarchical calibration as consistency constraints for unlabeled adaptation. qNB constrains predicted proportions to explain measured target expression, while calibration adjusts confidence, fine-type bias, and mean proportions on reserved reference pseudo-spots after target adaptation.

2 Method

2.1 Overview

We align reference and target expression over shared genes. Reference labels and sampled cell counts define pseudo-spot proportions for supervision. Target expression and coordinates constrain unlabeled proportion estimation through observation alignment and spatial context. The inputs are reference expression $X^r \in \mathbb{R}_+^{N_r \times G}$, cell labels $y^r \in \{1, \dots, K\}^{N_r}$, target expression $X^t \in \mathbb{R}_+^{N_t \times G}$, and coordinates $S^t \in \mathbb{R}^{N_t \times 2}$. Here N_r, N_t, G, K count reference cells, target spots, shared genes, and cell types. Matrices store observations in rows, with column vectors $x_i^t = (X_{i,:}^t)^\top$ and $s_i = (S_{i,:}^t)^\top$. Two encoders, fusion, and a fraction head produce $p_i \in \Delta^{K-1}$, followed by logit calibration to p_i^{final}, where $\Delta^{K-1} = \{p \in \mathbb{R}_+^K : \mathbf{1}^\top p = 1\}$. Figure 1 and Appendix B.1 detail inference and training.

2.2 Composition-aware pseudo-spot construction

Composition sampling. Four supports cover composition boundaries that random aggregation misses. SILTA samples n_i reference cells per pseudo-spot within each dataset's aggregate-size range. Empirical mixtures retain observed abundance, balanced mixtures expose rare types, near-pure anchors isolate individual types, and symmetric hard pairs reverse similar-signature type order. Let $x_{ij} \in \mathbb{R}_+^G$ denote sampled cell j's raw counts and c_{ik} count cells of type k. We construct

$$x_i^{\text{ord}} = \sum_{j=1}^{n_i} x_{ij}, \qquad y_{ik} = \frac{c_{ik}}{n_i}, \qquad n_i = \sum_{k=1}^{K} c_{ik} \tag{1}$$

The composition label $y_i = (y_{i1}, \dots, y_{iK})^\top$ remains fixed during expression augmentation.

Bounded diffusion. Expression varies with cell state, capture, depth, and platform noise. Composition-conditioned diffusion augments a fraction ρ_{diff} of pseudo-spots without changing y_i. Between-type signature suppression, gene-wise clipping, and scaling by α_{diff} bound each residual. Augmented x_i^{hybrid} (Appendix A.4) form N_p rows of $X^p \in \mathbb{R}_+^{N_p \times G}$. Fixed dataset preprocessing $\mathcal{T}$ gives encoder inputs $\bar{X}^p = \mathcal{T}(X^p)$ and $\bar{X}^t = \mathcal{T}(X^t)$, while qNB retains the observation scale.

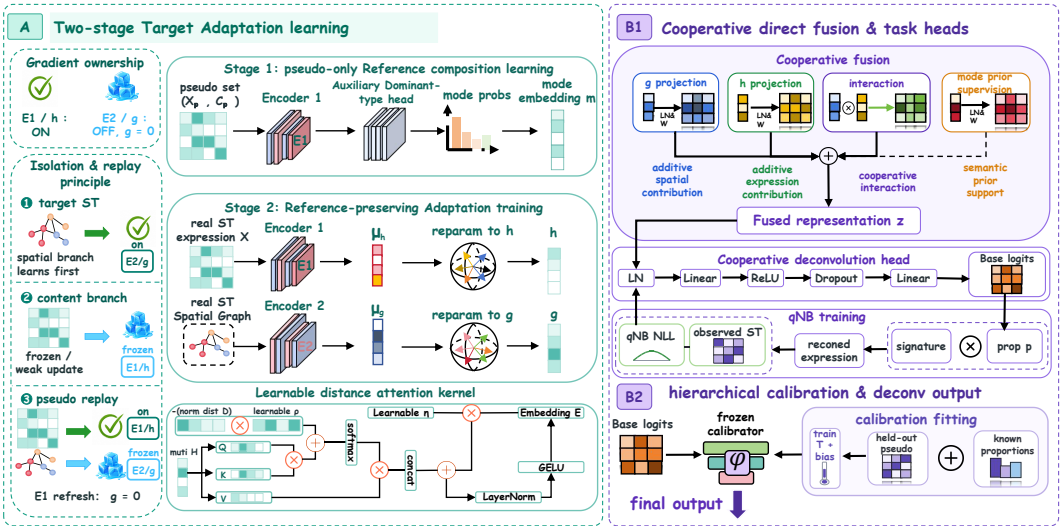


Figure 1: **SILTA workflow.** (A) Two-stage training separates reference composition learning (Stage I) from target adaptation (Stage II). (B1) Cooperative expression-spatial fusion feeds the task heads and qNB observation objective. (B2) Hierarchical calibration uses reserved reference pseudo-spots with the backbone fixed. Implementation details are in Appendix B.

2.3 Expression and Spatial Graphs

Expression graph. A_{expr} uses mutual $k_{\text{expr}} = 6$ neighbors under cosine similarity of preprocessed expression. We fill degrees below six, symmetrize, and add self-loops. The $N_p \times N_p$ pseudo-spot and $N_t \times N_t$ target graphs mask E1, while dense E1 uses all input pairs (Appendix B).

Spatial graph. Median nearest-neighbor distance sets the scale

$$d_{\text{nn}} = \text{median}_i \min_{j \neq i} \|s_i - s_j\|_2 \tag{2}$$

Radius-$1.5d_{\text{nn}}$ edges, nearest-neighbor completion to degree four, symmetrization, and self-loop removal give $A_{\text{sp}} \in \{0,1\}^{N_t \times N_t}$. This graph defines smoothness, boundary-aware total variation (BATV), and edge-removal tests for spatially adjacent target-spot pairs.

Distance-aware attention. Dense E2 head r combines target expression and distance:

$$\begin{aligned} \bar{d}_{ij} &= \|s_i - s_j\|_2 / d_{\text{att}}, \quad a_{ij}^{(r)} = \frac{(q_i^{(r)})^\top k_j^{(r)}}{\sqrt{d_r}} - \frac{\bar{d}_{ij}^2}{2\sigma_r^2}, \\ w_{ij}^{(r)} &= \frac{\exp a_{ij}^{(r)}}{\sum_{v=1}^{N_t} \exp a_{iv}^{(r)}}, \quad o_i^{(r)} = \sum_{j=1}^{N_t} w_{ij}^{(r)} v_j^{(r)} \end{aligned} \tag{3}$$

Head r maps features affinely to d_r-dimensional queries $q_i^{(r)}$, keys $k_i^{(r)}$, and values $v_i^{(r)}$. Here $\bar{d}_{ij}, a_{ij}^{(r)}, w_{ij}^{(r)}, o_i^{(r)}$ denote normalized distance, attention score, weight, and output. Positive scales d_{att}, σ_r control distance normalization and decay. E2 projects concatenated outputs into residual updates over all target pairs. Appendix B gives fixed-bandwidth and incidence variants.

2.4 Stage I Learns the Reference Proportion Mapping

E1 encodes reference pseudo-spots through attention widths 256, 256, and $d_e = 512$, giving $e_i^{(1)} \in \mathbb{R}^{d_e}$. Gaussian heads follow layer normalization:

$$q_{h,i} = \mathcal{N}\big(\mu_{h,i}, \text{diag}(\sigma_{h,i}^2)\big), \qquad h_i = \mu_{h,i} \in \mathbb{R}^{d_h} \tag{4}$$

Both encoders use Gaussian means μ and scales $\sigma > 0$. Expression latent h_i uses the mean, with configured stochastic training samples (Appendix B.1).

Auxiliary classifier C_{mode} maps $e_i^{(1)}$ to K dominant-type logits:

$$q_i^m = \text{softmax}\Big(C_{\text{mode}}(e_i^{(1)})/\tau_m\Big), \qquad m_i = E_m^{\mathsf{T}} q_i^m \in \mathbb{R}^{d_m} \tag{5}$$

Type probabilities q_i^m at temperature $\tau_m > 0$ yield embedding m_i via row-wise type embeddings $E_m \in \mathbb{R}^{K \times d_m}$. Auxiliary supervision combines dominant-type cross-entropy, soft-label KL, and prototype alignment. Fusion uses m_i when $\lambda_m \neq 0$.

For pseudo-spot predictions $\hat{p}_i$ with the spatial route disabled, Stage I minimizes:

$$\mathcal{L}_{\text{frac}} = \frac{1}{N_p} \sum_{i=1}^{N_p} D_{\text{KL}}(y_i \| \hat{p}_i), \quad \mathcal{L}_{\text{content}} = \lambda_{\text{frac}} \mathcal{L}_{\text{frac}} + \mathcal{L}_{\text{mode}} + \mathcal{L}_{\text{reg}}^p \tag{6}$$

D_{KL} denotes KL divergence, λ_{frac} its weight, and $\mathcal{L}_{\text{mode}}$, $\mathcal{L}_{\text{reg}}^p$ weighted auxiliary and Gaussian losses (Appendix B.3). E2 is frozen, with spatial switch $\chi = 0$ (Equation 9).

2.5 Stage II adapts to target expression and spatial context

E2 yields normalized features $e_i^{(2)} \in \mathbb{R}^{d_e}$ and spatial Gaussian parameters:

$$q_{g,i} = \mathcal{N}\big(\mu_{g,i}, \text{diag}(\sigma_{g,i}^2)\big), \qquad g_i = \mu_{g,i} \in \mathbb{R}^{d_g} \tag{7}$$

Parameters depend on $e_i^{(2)}$, and spatial latent g_i enters prediction when $\chi = 1$.

Gradual unfreezing. The spatial-only interval updates E2, its fusion parameters, and active qNB responses while freezing the reference path and fraction head to learn target-specific spatial corrections. **Joint update.** Selected expression and readout parameters then unfreeze at a lower learning rate than E2 to refine the reference-trained prediction pathway. Either interval can have zero length. **Reference replay.** Pseudo-spots supervise the reference mapping with cell-number proportion labels, $\chi = 0$, and E2 frozen to retain reference-learned cell-type information.

For uncalibrated target predictions p_i, the adaptation terms are

$$\mathcal{L}_{\text{target}} = \lambda_{\text{obs}} \mathcal{L}_{\text{qNB}} + \lambda_{\text{smooth}} \mathcal{L}_{\text{smooth}} + \lambda_{\text{BATV}} \mathcal{L}_{\text{BATV}} + \mathcal{L}_{\text{reg}}^t \tag{8}$$

Nonnegative weights select qNB alignment, A_{sp} smoothness, and BATV, which down-weights expression-supported boundary pairs within the BATV term. $\mathcal{L}_{\text{reg}}^t$ collects enabled Gaussian penalties. Each training step updates only its assigned parameters. Replay keeps qNB response parameters frozen. Reference signatures and dispersion remain fixed throughout training.

2.6 Cooperative expression-spatial fusion

Affine maps $P_h : \mathbb{R}^{d_h} \to \mathbb{R}^{D_f}$, $P_g : \mathbb{R}^{d_g} \to \mathbb{R}^{D_f}$, and $P_m : \mathbb{R}^{d_m} \to \mathbb{R}^{D_f}$ align branches for fusion. The interaction map $P_x : \mathbb{R}^{D_f} \to \mathbb{R}^{D_f}$ is also affine. Using layer normalization LN and elementwise product $\odot$,

$$\begin{aligned} h_{0,i} = \text{LN}(P_h(h_i)), \quad g_{0,i} = \text{LN}(P_g(g_i)), \quad m_{0,i} = \text{LN}(P_m(m_i)) \\ f_i = \text{LN}[h_{0,i} + \chi\{\lambda_g g_{0,i} + \lambda_x P_x(h_{0,i} \odot g_{0,i})\} + \lambda_m m_{0,i}], \\ z_i = F_z(f_i), \quad \ell_i = C_{\text{frac}}(z_i), \quad p_i = \text{softmax}(\ell_i/\tau_{\text{frac}}) \end{aligned} \tag{9}$$

Subscript 0 marks aligned branches of width D_f. Weights $\lambda_g, \lambda_x, \lambda_m$ scale spatial, interaction, and auxiliary terms. The fused feature f_i passes through F_z (identity or Gaussian readout) to $z_i \in \mathbb{R}^{D_f}$ (Appendix B.1). Fraction head $C_{\text{frac}} : \mathbb{R}^{D_f} \to \mathbb{R}^K$ gives raw logits ℓ_i, and $\tau_{\text{frac}} > 0$ sets training temperature. Setting $\chi = 0$ bypasses both spatial terms and their affine offsets, retaining the readout.

2.7 Observation alignment and calibration

Observation alignment. The fixed reference dictionary $R \in \mathbb{R}_+^{K \times G}$ stores gene-normalized type signatures, so $R^{\mathsf{T}} p_i \in \mathbb{R}_+^G$ predicts target gene fractions. Bounded gene responses adjust this mixture, and observed library size scales it to the qNB expression mean (Appendix A.5).

Post-hoc calibration. With the adapted backbone fixed, reserved reference pseudo-spots calibrate raw type logits ℓ_{ik} to a_{ik}:

$$a_{ik} = \frac{\ell_{ik}}{T_{c(k)}} + b_{c(k)} + \tilde{b}_k, \qquad p_i^{\text{final}} = \text{softmax}(\mathcal{C}_{\text{end}}(\ell_i)/\tau_{\text{inf}}) \tag{10}$$

Here $c(k)$ groups types, $T_c > 0$ and b_c are group temperatures and biases, and $\tilde{b}_k$ is a centered within-group type bias. The endpoint map $\mathcal{C}_{\text{end}}$ applies this calibration and reference-fitted corrections (Appendix A.5), and $\tau_{\text{inf}} > 0$ sets inference temperature.

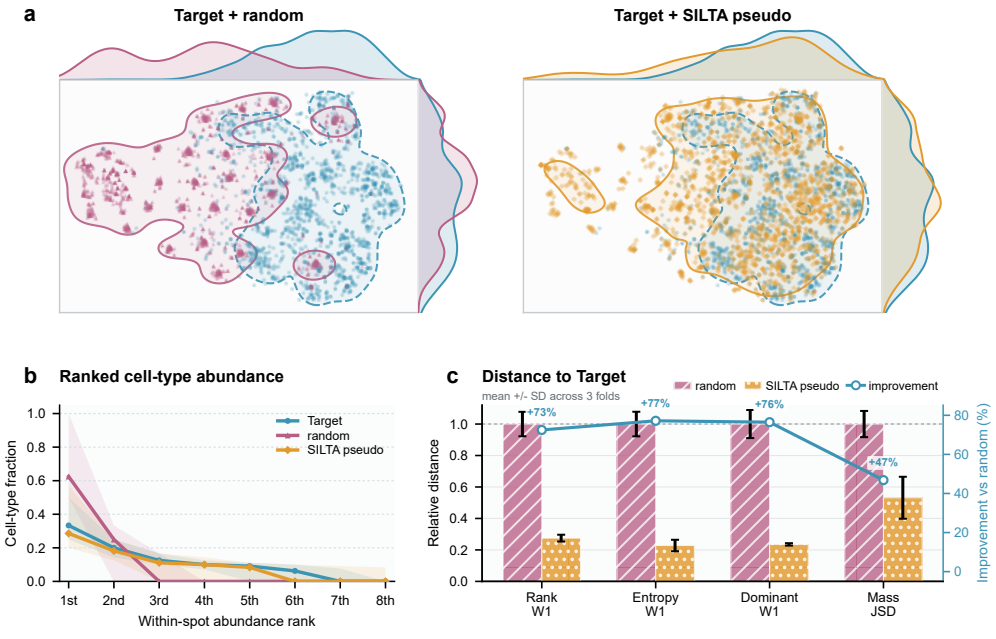


Figure 2: **Pseudo-spot abundance audit on Moffitt.** (a) Target, random, and SILTA pseudo-spots in a fixed embedding. (b) Median rank abundance with interquartile bands. (c) Composition distances across the three cross-animal folds, normalized to random (1.0) on a common scale.

3 Experiments

We evaluate six datasets across five ST platforms. Moffitt, osmFISH (OSM), and seqFISH provide known proportions, while melanoma, the dorsolateral prefrontal cortex (DLPFC), and human breast cancer (HBC) support marker-gene or downstream analyses. Appendix Tables 4 and 5 summarize datasets, roles, and replication, with full protocols in Appendix A.2. Each SILTA fold or split averages three model seeds before the three top-level means. The ten baselines are AddaGCN (Ding et al., 2026), cell2location (Kleshchevnikov et al., 2022), HarmoDecon (Wang et al., 2025b), Tangram (Biancalani et al., 2021), STRIDE (Sun et al., 2022), STdGCN (Li & Luo, 2024), SPOTlight (Elosua-Bayes et al., 2021), CARD (Ma & Zhou, 2022), RCTD (Cable et al., 2022), and SpatialDecon (Danaher et al., 2022). All methods share target spots, genes, and type order. Reference-based methods also receive the same fold-specific reference cells, with native defaults retained. We report RMSE, JSD, SSIM, and PCC, with directions and definitions in Appendix A.3. Balanced accuracy, macro-F1, and Rare-F1 over prespecified rare types provide complementary structural checks of dominant-type recovery and fine-type boundary preservation.

3.1 Deconvolution accuracy and pseudo-spot supervision

Composition coverage. Structured pseudo-spots should cover sparse, long-tailed target compositions, whereas random aggregation can concentrate around average mixtures. We compare both strategies with known target fractions using a fixed embedding, rank-abundance profiles, and fold-level composition distances. Figure 2 shows that SILTA's structured pseudo-spots more closely

Table 1: **Moffitt pseudo-spot curriculum ablation.** We average three model seeds within each of three held-out-animal folds. Both strategies use the same model settings, training budget and other conditions. Relative improvement uses unrounded fold means for consistent comparison.

Strategy	RMSE↓	JSD↓	SSIM↑	PCC↑	Source macro-type RMSE↓	Target Rare-F1↑	Rare-rich RMSE↓
Random aggregation	.04092	.45401	.81498	.80015	.04622	.30001	.04899
SILTA curriculum	**.04073**	**.45285**	**.81636**	**.80161**	**.04463**	**.31071**	**.04849**
Improvement	**0.46%**	**0.26%**	**0.17%**	**0.18%**	**3.44%**	**3.57%**	**1.03%**

Table 2: **Deconvolution benchmark.** Bold marks the best mean per dataset and metric. SILTA and AddaGCN average three seeds within each of three folds or splits before the top-level mean. C2L, Harmo, SPOT, and SpDecon denote cell2location, HarmoDecon, SPOTlight, and SpatialDecon.

Method	**OSM**				**seqFISH**				**Moffitt**			
	RMSE↓	JSD↓	SSIM↑	PCC↑	RMSE↓	JSD↓	SSIM↑	PCC↑	RMSE↓	JSD↓	SSIM↑	PCC↑
SILTA	**.06429**	**.45866**	**.80526**	**.79928**	**.06787**	**.36552**	**.90185**	**.90136**	**.03892**	**.44687**	**.83377**	**.82060**
AddaGCN	.09131	.67145	.24438	.46260	.13034	.64367	.29864	.60805	.05434	.63754	.49478	.59791
Harmo	.11759	.71334	.23132	.20933	.10529	.44201	.79893	.80285	.08599	—	.08715	.06530
Tangram	.08108	.47295	.68291	.67361	.09302	.41351	.83025	.82840	.06146	.48046	.65909	.63249
STRIDE	.07787	.58583	.63463	.65857	.11073	.58733	.55297	.73014	.04901	.59612	.59795	.70766
STdGCN	.07565	.55456	.70857	.70646	.07254	.40367	.88814	.88947	.04534	.55817	.72183	.71955
SPOT	.08064	.50485	.61087	.62437	.11129	.49671	.66027	.69354	.05513	.55243	.58883	.56120
CARD	.07792	.51082	.67347	.67045	.07944	.40464	.83662	.85692	.04596	.49868	.73843	.72255
RCTD	.07213	.48353	.75143	.74241	.06925	.38605	.90145	.90125	.04334	.45472	.78636	.76980
SpDecon	.07537	.50496	.72293	.71372	.07798	.40160	.86947	.87154	.04490	.48498	.77124	.75335
C2L	.09306	.68926	.15211	.53963	.06875	.41373	.88547	.90082	.05217	.64024	.54636	.62019

match the Moffitt target-composition distribution than random aggregation across these views, including the dominant-type abundance pattern. Appendix D.9 reports OSM and seqFISH analyses.

Curriculum comparison. The matched curriculum experiment compares empirical random aggregation with the complete SILTA curriculum under matched model settings and training budgets. This matched design isolates pseudo-spot construction from model capacity and optimization effects. Across the three Moffitt animal folds, the structured curriculum improves all seven reported means, with its largest relative gains in source macro-type RMSE and target Rare-F1 (Table 1). These metrics measure coarse source composition and rare-type recovery at broad and fine cell-type levels across the matched source and target evaluations in all three folds.

Deconvolution benchmarks. We compare SILTA with ten baselines on Moffitt, OSM, and seqFISH under matched inputs. AddaGCN tests graph/domain adaptation, while cell2location provides a hierarchical Bayesian count model (Ding et al., 2026; Kleshchevnikov et al., 2022). OSM has 33 genes and 31 types, while Moffitt evaluates 45 types with cross-animal folds. SILTA attains the lowest mean RMSE/JSD across datasets and highest mean SSIM/PCC on OSM and Moffitt (Table 2). Figure 3 and Appendix D.15 show OSM and seqFISH maps. All completed outputs enter Table 2, although some AddaGCN maps are structurally flat.

3.2 Perturbations isolate stage-specific recovery

We assess SILTA through target accuracy, reference retention, and perturbation recovery, with marker concordance and pathology-region classification providing complementary biological evaluation in Section 3.3. Paired controls compare dominant predictions and full proportion vectors. Update trajectories test whether target gains preserve the source-learned mapping. Appendices C and D provide the full controls and cross-dataset results.

Stage I boundary recovery. A fixed logit margin switches 30% of reserved reference pseudo-spot predictions toward similar types. With E2 frozen, source supervision improves all five measured criteria without target data (Appendix Table 11). RMSE and JSD decline, while hard-pair accuracy and rare-type recall improve for confusable and infrequent types. These gains connect lower proportion error with recovery of reference-defined boundaries before target adaptation.

Stage II prediction recovery. This experiment switches target spots' dominant predictions before Stage II. Adaptation restores agreement with paired clean Stage II predictions in 89.62% of switched spots. On these spots, ground-truth dominant accuracy increases from 12.20% to 54.57% and balanced accuracy from 12.73% to 53.75% (Appendix Table 12). These ground-truth gains show improved dominant-type identification on switched spots in addition to prediction agreement. Across all target spots, the four fraction metrics approach clean-adapted values.

Matched adaptation and replay. During Moffitt Stage II, matched schedules reach similar target endpoints within fold-level variation, but direct joint adaptation weakens the source-learned mapping (Figure 4). Appendix Tables 15 and 13 report the matched Moffitt and OSM update endpoints.

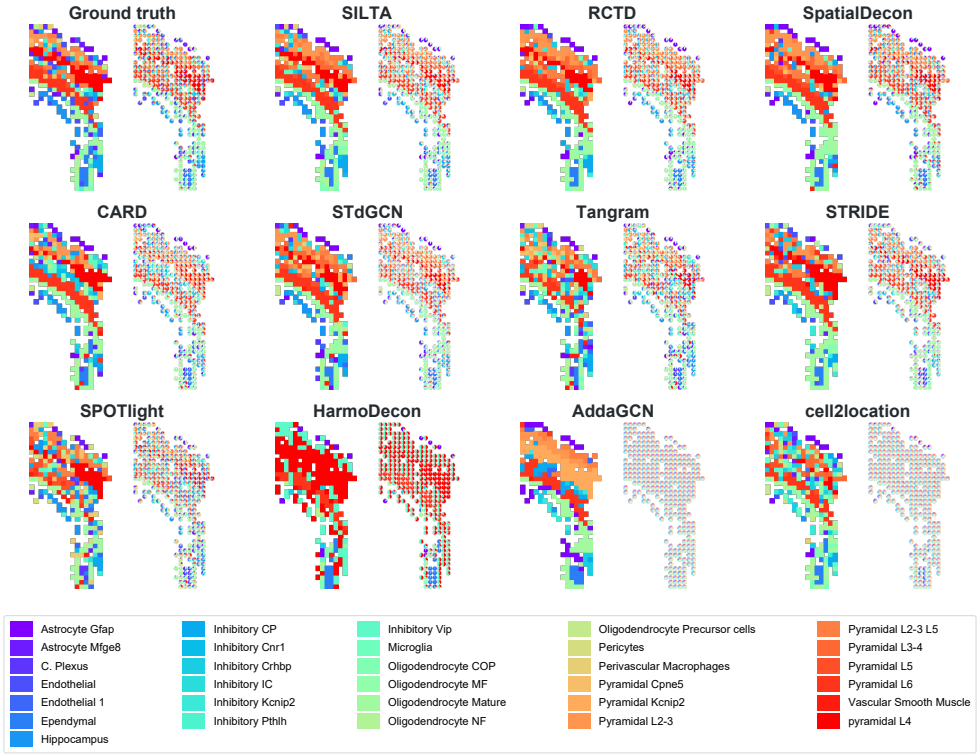


Figure 3: **OSM spatial deconvolution.** Ground truth, SILTA, and ten baselines appear as dominant-type maps and spot-level proportion pies. The bottom gives the type legend.

Experiments on OSM, Moffitt, and seqFISH show replay's distinct roles. It gives the largest additional target gain among OSM updates (Appendix Table 13) and lowers reference RMSE and global composition drift versus direct joint adaptation (Appendix Table 15). It reduces source prediction drift under joint and protected seqFISH updates (Appendix Table 23).

Spot-level allocation recovery. We permute complete proportion rows across distant Moffitt spots, preserving global composition but disrupting spot-level allocation. Stage II restores dominant-prediction agreement with paired clean adaptation in 83.93% of selected, row-permuted spots. All four fraction metrics approach the clean adapted values (Table 3), indicating recovery of complete proportion vectors. On the same selected spots, ground-truth dominant accuracy rises from 5.51% to 58.00% with gains in every fold, close to 58.05% for clean adaptation. Appendix C defines recovery and aggregation. Reference supervision supports target adaptation and helps retain reference-derived composition information, while schedule benefits remain dataset- and metric-dependent.

Table 3: **Moffitt allocation recovery.** Long-distance permutation moves complete proportion/logit rows across 30% of spots, preserving global composition. On/off denote Stage II, with matched adaptation settings.

Condition	RMSE↓	JSD↓	SSIM↑	PCC↑
Clean / off	.05181	.60033	.57113	.63879
Clean / on	**.03798**	**.44957**	**.82333**	**.81177**
Permuted / off	.05814	.64844	.45213	.47329
Permuted / on	.03833	.45583	.82016	.80840

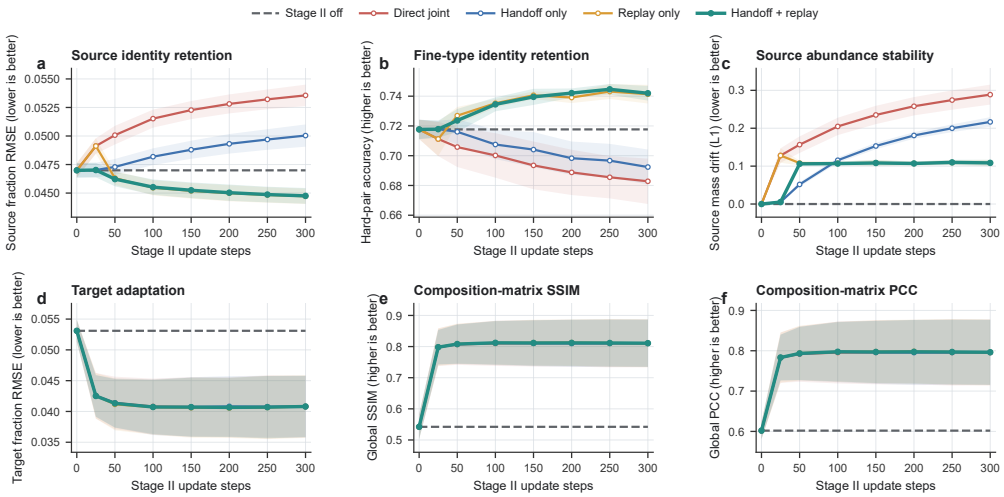


Figure 4: **Reference retention and target improvement during Stage II on Moffitt.** The top row shows held-out pseudo-spot RMSE, hard-pair accuracy, and global composition drift, whereas the bottom row shows target RMSE and composition-matrix SSIM and PCC. Shading marks fold SD.

3.3 Marker-gene concordance and pathology-region classification

Marker concordance. Besides the known-composition benchmarks, we assess melanoma tissues without ground-truth proportions. With fixed marker panels and a shared protocol, we compare predicted macrophage and T-cell distributions with measured expression as an indirect consistency check. SILTA achieves the highest PCC and Spearman correlations among evaluated methods for macrophages on Patient 3 replicate 1 (Fig. 5) and T cells on Patient 4 replicate 1 (Appendix Table 8). SILTA's higher Pearson and Spearman correlations indicate stronger linear correspondence with marker-associated expression variation and closer agreement in the relative ordering of spots. These results support predicted immune-cell distributions and local enrichment in melanoma.

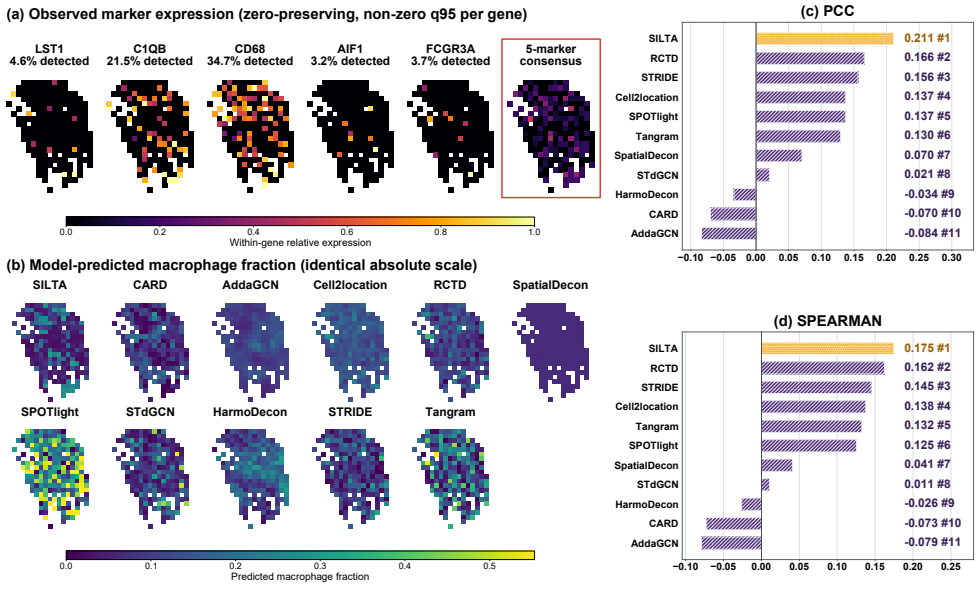


Figure 5: **Macrophage marker concordance on Patient 3 replicate 1.** (a) Five marker-expression maps, whose pointwise mean is the consensus. (b) SILTA and ten baseline fractions on one absolute fraction scale. (c,d) Pearson (PCC) and Spearman correlation rankings.

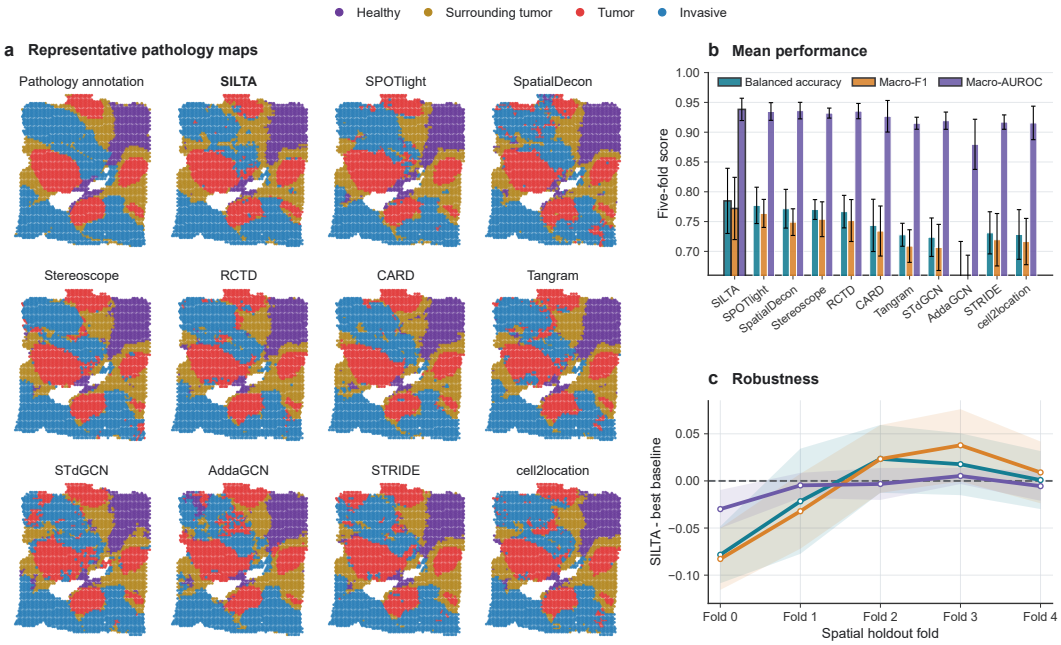


Figure 6: **HBC pathology-region classification.** (a) Public labels and shared classifier maps for SILTA and ten baselines. (b) Five-fold means and fold SD for balanced accuracy, macro-F1, and macro-AUROC. (c) Sealed-fold SILTA-minus-best-baseline deltas by metric and fold, with ribbons giving ± 1 SD from 1,000 paired within-fold spatial-block bootstrap replicates.

Pathology-region classification. Using a shared nested classifier-selection protocol, we test whether features derived from inferred HBC compositions support pathology-region classification. SILTA achieves the highest five-fold mean balanced accuracy, macro-F1, and macro-AUROC. The Fold 3 prediction map retains the annotated surrounding-tumor zone (Fig. 6), and Appendix Table 9 reports five-fold means and SDs for all three classification metrics. The class-balanced results indicate that composition-derived features remain informative for distinguishing pathology regions in spatial blocks withheld from classifier training. The retained surrounding-tumor zone links classification performance to annotated tissue structure.

Appendix C presents matched tests of reference recovery, target correction, replay, and spatial assignment. Appendix D includes cross-dataset ablations, hyperparameter analyses, qNB and calibration controls, pseudo-spot coverage, spatial interventions, and run-level results.

4 Conclusion

Reference-based spatial deconvolution requires retaining cell-type information from annotated references while accounting for expression differences and spatial variation in target tissues. In this paper, we introduce SILTA to address these demands by separating composition learning from target adaptation, combining diffusion-augmented reference learning with expression-spatial expert cooperation to refine proportions in unlabeled target tissues. Across three known-composition benchmarks, including cross-animal validation, SILTA improves deconvolution accuracy over evaluated baselines. Evaluations in melanoma, breast cancer, and dorsolateral prefrontal cortex tissues further support the biological relevance of its predictions. Marker concordance links inferred cell distributions to measured molecular patterns, while pathology-region classification demonstrates the value of composition-derived features for distinguishing tissue regions. Together, these findings connect improved proportion estimation with biological information retention and downstream utility, highlighting the value of coordinated reference learning and target adaptation. Our future work will extend SILTA to larger tissue cohorts and improve adaptation across different measurement platforms, supporting broader investigations of tissue heterogeneity.

AI USE STATEMENT

We used generative AI for language editing, bilingual consistency checks, LaTeX formatting, plotting, and analysis-code assistance. The authors verified all AI-assisted text, code, figures, and calculations and retain responsibility for the final content and all reported claims.

REFERENCES

Tommaso Biancalani, Gabriele Scalia, Lorenzo Buffoni, Raghav Avasthi, Ziqing Lu, Aman Sanger, Neriman Tokcan, Charles R. Vanderburg, et al. Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. *Nature Methods*, 18(11):1352–1362, 2021. doi: 10.1038/s41592-021-01264-7.

Dylan M. Cable, Evan Murray, Luli S. Zou, Aleksandrina Goeva, Evan Z. Macosko, Fei Chen, and Rafael A. Irizarry. Robust decomposition of cell type mixtures in spatial transcriptomics. *Nature Biotechnology*, 40(4):517–526, 2022. doi: 10.1038/s41587-021-00830-w.

Simone Codeluppi, Lars E. Borm, Amit Zeisel, Gioele La Manno, Josina A. van Lunteren, Camilla I. Svensson, and Sten Linnarsson. Spatial organization of the somatosensory cortex revealed by osmFISH. *Nature Methods*, 15(11):932–935, 2018. doi: 10.1038/s41592-018-0175-z.

Patrick Danaher, Youngmi Kim, Brenn Nelson, Maddy Griswold, Zhi Yang, Erin Piazza, and Joseph M. Beechem. Advances in mixed cell deconvolution enable quantification of cell types in spatial transcriptomic data. *Nature Communications*, 13:385, 2022. doi: 10.1038/s41467-022-28020-5.

Shuzhen Ding, Zhou Yu, and Jingsi Ming. AddaGCN: Spatial transcriptomics deconvolution using graph convolutional networks with adversarial discriminative domain adaptation. *PLOS Computational Biology*, 22(8):e1014609, 2026. doi: 10.1371/journal.pcbi.1014609.

Rui Dong and Guo-Cheng Yuan. SpatialDWLS: accurate deconvolution of spatial transcriptomic data. *Genome Biology*, 22(1):145, 2021. doi: 10.1186/s13059-021-02362-7.

Marc Elosua-Bayes, Paula Nieto, Elisabetta Mereu, Ivo Gut, and Holger Heyn. SPOTlight: seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. *Nucleic Acids Research*, 49(9):e50, 2021. doi: 10.1093/nar/gkab043.

Lucie C. Gaspard-Boulinc, Luca Gortana, Thomas Walter, Emmanuel Barillot, and Florence M. G. Cavalli. Cell-type deconvolution methods for spatial transcriptomics. *Nature Reviews Genetics*, 26:828–846, 2025. doi: 10.1038/s41576-025-00845-y.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1321–1330, 2017.

Jonathan Ho, Ajay N. Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Vitalii Kleshchevnikov, Artem Shmatko, Emma Dann, Alexander Aivazidis, Hamish W. King, Tong Li, Rasa Elmentaite, Artem Lomakin, Veronika Kedlian, et al. Cell2location maps fine-grained cell types in spatial transcriptomics. *Nature Biotechnology*, 40(5):661–671, 2022. doi: 10.1038/s41587-021-01139-4.

Bin Li, Wen Zhang, Chuang Guo, Hao Xu, Longfei Li, Minghao Fang, Yinlei Hu, Xinye Zhang, Xinfeng Yao, Meifang Tang, et al. Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution. *Nature Methods*, 19(6):662–670, 2022. doi: 10.1038/s41592-022-01480-9.

Haoyang Li, Juexiao Zhou, Zhongxiao Li, Siyuan Chen, Xingyu Liao, Bin Zhang, Ruochi Zhang, Yu Wang, Shiwei Sun, and Xin Gao. A comprehensive benchmarking with practical guidelines for cellular deconvolution of spatial transcriptomics. *Nature Communications*, 14(1):1548, 2023. doi: 10.1038/s41467-023-37168-7.

Yawei Li and Yuan Luo. STdGCN: spatial transcriptomic cell-type deconvolution using graph convolutional networks. *Genome Biology*, 25(1):206, 2024. doi: 10.1186/s13059-024-03353-0.

Tim Lohoff, Shila Ghazanfar, Alexandra Missarova, Noushin Koulena, Nicole Pierson, Jonathan A. Griffiths, Evan S. Bardot, Chee-Huat Linus Eng, Richard C. V. Tyser, Ricard Argelaguet, et al. Integration of spatial and single-cell transcriptomic data elucidates mouse organogenesis. *Nature Biotechnology*, 40:74–85, 2022. doi: 10.1038/s41587-021-01006-2.

Romain Lopez, Baoguo Li, Hadas Keren-Shaul, Pierre Boyeau, Merav Kedmi, David Pilzer, Adam Jelinski, Ido Yofe, Eyal David, et al. DestVI identifies continuums of cell types in spatial transcriptomics data. *Nature Biotechnology*, 40(9):1360–1369, 2022. doi: 10.1038/s41587-022-01272-8.

Yaxiong Ma, Zengfa Dou, Yuhong Zha, and Xiaoke Ma. SA2E: Spatial-aware auto-encoder for cell type deconvolution of spatial transcriptomics data. *Bioinformatics*, 42(4):btag133, 2026. doi: 10.1093/bioinformatics/btag133.

Ying Ma and Xiang Zhou. Spatially informed cell-type deconvolution for spatial transcriptomics. *Nature Biotechnology*, 40(9):1349–1359, 2022. doi: 10.1038/s41587-022-01273-7.

Kristen R. Maynard, Leonardo Collado-Torres, Lukas M. Weber, Cedric Uytingco, Brianna K. Barry, Stephen R. Williams, Joseph L. Catallini, Matthew N. Tran, Zachary Besich, et al. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nature Neuroscience*, 24(3):425–436, 2021. doi: 10.1038/s41593-020-00787-0.

Yuanyuan Mo, Juan Liu, and Lihua Zhang. Deconvolution of spatial transcriptomics data via graph contrastive learning and partial least square regression. *Briefings in Bioinformatics*, 26 (1):bbaf052, 2025. doi: 10.1093/bib/bbaf052.

Jeffrey R. Moffitt, Dhananjay Bambah-Mukku, Stephen W. Eichhorn, Eric Vaughn, Karthik Shekhar, Julio D. Perez, Nimrod D. Rubinstein, Junjie Hao, Aviv Regev, Catherine Dulac, and Xiaowei Zhuang. Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science*, 362(6416):eaau5324, 2018. doi: 10.1126/science.aau5324.

Jingyang Qian, Hudong Bao, Xin Shao, Yin Fang, Jie Liao, Zhuo Chen, Chengyu Li, Wenbo Guo, Yining Hu, Anyao Li, Yue Yao, Xiaohui Fan, and Yiyu Cheng. Simulating multiple variability in spatially resolved transcriptomics with scCube. *Nature Communications*, 15(1):5021, 2024. doi: 10.1038/s41467-024-49445-0.

Qianqian Song and Jing Su. DSTG: deconvoluting spatial transcriptomics data through graph-based artificial intelligence. *Briefings in Bioinformatics*, 22(5):bbaa414, 2021. doi: 10.1093/bib/bbaa414.

Patrik L. Ståhl, Fredrik Salmén, Sanja Vickovic, Anna Lundmark, José Fernández Navarro, Jens Magnusson, Stefania Giacomello, Michaela Asp, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82, 2016. doi: 10.1126/science.aaf2403.

Dongqing Sun, Zhaoyang Liu, Taiwen Li, Qiu Wu, and Chenfei Wang. STRIDE: Accurately decomposing and integrating spatial transcriptomics using single-cell RNA sequencing. *Nucleic Acids Research*, 50(7):e42, 2022. doi: 10.1093/nar/gkac150.

Kim Thrane, Hanna Eriksson, Jonas Maaskola, Johan Hansson, and Joakim Lundeberg. Spatially resolved transcriptomics enables dissection of genetic heterogeneity in stage III cutaneous malignant melanoma. *Cancer Research*, 78(20):5970–5979, 2018. doi: 10.1158/0008-5472.CAN-18-0747.

Bo Wang, Yahui Long, Yuting Bai, Jiawei Luo, and Chee Keong Kwoh. STCGAN: a novel cycle-consistent generative adversarial network for spatial transcriptomics cellular deconvolution. *Briefings in Bioinformatics*, 26(1):bbae670, 2025a. doi: 10.1093/bib/bbae670.

Zirui Wang, Ke Xu, Yang Liu, Yu Xu, and Lu Zhang. Mitigation of multi-scale biases in cell-type deconvolution for spatially resolved transcriptomics using HarmoDecon. *Bioinformatics*, 41(9): btaf451, 2025b. doi: 10.1093/bioinformatics/btaf451.

Sunny Z. Wu, Ghamdan Al-Eryani, Daniel Lee Roden, et al. A single-cell and spatially resolved atlas of human breast cancers. *Nature Genetics*, 53:1334–1347, 2021. doi: 10.1038/s41588-021-00911-1.

Yichen Xu, Ryumei Nakada, Linjun Zhang, and Lexin Li. Residual feature integration is sufficient to prevent negative transfer. In Carl Vondrick, Bharath Hariharan, Colin Raffel, Lerrel Pinto, Diyi Yang, and Aleksandra Faust (eds.), *International Conference on Learning Representations*, pp. 77738–77763, 2026. URL `https://proceedings.iclr.cc/paper_files/paper/2026/file/7d998ff3e6ecce821986c9dd876be9bc-Paper-Conference.pdf`.

A Implementation and Evaluation Details

A.1 Notation and data handling

For a pseudo-spot, n_i is the number of aggregated cells and $y_i \in \Delta^{K-1}$ is its known composition. The graph matrices are A_{expr} and A_{sp}, and d_{nn} is the median nearest spatial-neighbor distance. In the observation model, x_{ij}^t is the nonnegative qNB input for gene j in spot i, R_{kj} is the normalized reference expression of gene j in type k, $L_i = \sum_j x_{ij}^t$ is the row sum, and ϑ_j is the fixed reference-estimated negative-binomial size.

We intersect genes once and keep the same order for the reference, pseudo-spots, and target assay. We aggregate pseudo-spots in nonnegative raw-count space, then apply the dataset's fixed preprocessing map $\mathcal{T}$. Appendix B.1 distinguishes standardization from count-normalization and `log1p` routes. The encoders use these transformed matrices, whereas qNB uses nonnegative counts or unrounded quasi-counts and computes each library scale from the row sum of that same input matrix. Reference signatures, dispersions, and calibration statistics use reference data only.

A.2 Evaluation protocol details

Table 4: **Dataset characteristics.** Fixed manifests specify the species, tissue, platform, gene panel, cell-type vocabulary, and target evaluation scale for every dataset used in this study.

Dataset	Species and tissue	Platform	Evaluation scale
Moffitt (Moffitt et al., 2018)	Mouse hypothalamic preoptic region	MERFISH	135 genes, 45 types, 533-568 target spots
OSM (Codeluppi et al., 2018)	Mouse somatosensory cortex	osmFISH	33 genes, 31 types, 296-305 target spots
seqFISH (Lohoff et al., 2022)	Mouse embryo	seqFISH	351 genes, 21 types, 668-676 target spots
DLPFC (Maynard et al., 2021)	Human dorsolateral prefrontal cortex	10x Visium	Fine cell types and cortical-layer markers
Melanoma (Thrane et al., 2018)	Human cutaneous melanoma	Legacy ST	Patient 3 replicate 1 and Patient 4 replicate 1
HBC (Wu et al., 2021)	Human breast carcinoma	10x Visium	3,798 spots and 13 reference types

Spot construction. For the primary OSM and seqFISH evaluations, we aggregate imaging cells into fixed spatial bins instead of fixing the number of cells per spot. Each bin represents a tissue area and retains natural variation in integer cell occupancy. The separate OSM C800 protocol models partial-cell capture. A cell crossing a bin boundary contributes cell-equivalent mass to each intersected bin in proportion to overlap area, while a seeded multinomial draw allocates its integer molecules. C800 isolates capture geometry and does not replace the primary whole-cell comparison.

Table 5: **Dataset use and evaluation design.** Known-composition benchmarks use fraction metrics, real-tissue analyses use observed markers or downstream pathology labels.

Dataset	Primary use	Replication or holdout design	Reported evidence
Moffitt	Deconvolution benchmark, pseudo-spot and Stage II analyses	Three leave-one-animal-out folds, each with three model seeds	RMSE, JSD, SSIM, PCC, and matched perturbations
OSM	Low-panel deconvolution benchmark, Stage II analysis	Three 60/40 cell-disjoint splits, each with three model seeds	RMSE, JSD, SSIM, PCC, and matched perturbations
seqFISH	Deconvolution benchmark, Stage II analysis	Three 60/40 cell-disjoint splits, each with three model seeds	RMSE, JSD, SSIM, PCC, and matched perturbations
DLPFC	Real-tissue deconvolution-derived analysis	Target-slice evaluation	Marker-gene concordance and cortical-layer organization
Melanoma	Real-tissue marker comparison	Macrophage and T-cell endpoints	PCC and Spearman marker-gene concordance
HBC	Downstream tissue analysis	Five contiguous held-out spatial folds	Pathology-region balanced accuracy, macro-F1, and macro-AUROC

Partitions and aggregation. Moffitt contains three biological sources. In each of three folds, two animals form the training reference and the remaining animal forms the held-out validation target, yielding three cross-animal training runs. OSM and seqFISH use predeclared 60/40 stratified cell-disjoint splits with seeds 17, 42, and 89. We split cells before spatial aggregation, generate pseudo-spots and validation samples from reference cells, and enforce zero cell-ID overlap within each split. The three split seeds define separate partitions of one source pool, so a cell can change roles across partitions. Methods that use transductive fitting receive target expression and coordinates.

Every primary known-composition SILTA benchmark contains three top-level folds or splits with three model seeds nested within each unit. We first average the model seeds within a unit and then average the three unit means, while benchmark standard deviations use those three fold or split means. Deterministic methods contribute one prediction per unit, while stochastic baselines follow their fixed native seed protocols. This rule does not govern the HBC downstream classifier, which reports five outer spatial folds under the nested selection protocol described with Table 9.

Low-panel baseline note. SpatialDecon's public implementation rejects panels with fewer than 100 genes. The OSM compatibility run removes this input-size guard while retaining its published optimizer. We lock every AddaGCN prediction before the private evaluator reads target proportions. Aggregate means include every completed run, including flat outputs.

A.3 Metric definitions

For S spots and K types, let $\mathbf{c}_s$ and $\hat{\mathbf{c}}_s$ denote the true and predicted composition vectors at spot s, and define $\mathbf{m}_s = (\mathbf{c}_s + \hat{\mathbf{c}}_s)/2$. We use $0 \log(0/q) = 0$ for zero-probability KL terms:

$$\text{RMSE} = \sqrt{\frac{1}{SK} \sum_{s,k} (\hat{c}_{sk} - c_{sk})^2} \tag{11}$$

$$\text{JSD} = \frac{1}{S} \sum_{s} \left[\frac{1}{2} D_{\text{KL}}(\mathbf{c}_s \| \mathbf{m}_s) + \frac{1}{2} D_{\text{KL}}(\hat{\mathbf{c}}_s \| \mathbf{m}_s) \right] \tag{12}$$

For SSIM and PCC, we row-normalize the spot-by-type matrices and vectorize corresponding entries in fixed spot-major, canonical-type order. With paired vectors y and $\hat{y}$, range $R = \max(y, \hat{y}) - \min(y, \hat{y})$, $C_1 = (0.01R)^2$, and $C_2 = (0.03R)^2$, we compute the following global structural similarity score over all paired entries and shared range constants:

$$\text{SSIM}_{\text{global}} = \frac{(2\mu_y \mu_{\hat{y}} + C_1)(2\sigma_{y\hat{y}} + C_2)}{(\mu_y^2 + \mu_{\hat{y}}^2 + C_1)(\sigma_y^2 + \sigma_{\hat{y}}^2 + C_2)} \tag{13}$$

PCC is the Pearson correlation of the same paired entries. These two scores measure agreement of the aligned composition matrix; they do not use spatial coordinates, graph adjacency, or local two-

dimensional windows. Because each spot sums to one over K types, both vectorized matrices have mean $1/K$, so the luminance factor in Eq. 13 equals one. We evaluate spatial structure separately through coordinate interventions, graph-edge interventions, and Moran's I. Dominant-label agreement is argmax accuracy, $S^{-1}\sum_s \mathbf{1}[\arg\max_k \hat{c}_{sk} = \arg\max_k c_{sk}]$, and macro-F1 assigns equal weight to each resulting type label when aggregating classwise F1 scores.

A.4 Partial-diffusion construction details

A Bernoulli indicator $\delta_i^{\text{diff}} \in \{0,1\}$ selects pseudo-spots for composition-conditioned diffusion. With generator $\mathcal{D}_{\theta_{\text{diff}}}$, generator parameters θ_{diff}, and noise ϵ_i^{diff},

$$\delta_i^{\text{diff}} \sim \text{Bernoulli}(\rho_{\text{diff}}), \qquad r_i^{\text{diff}} = \mathcal{D}_{\theta_{\text{diff}}}(x_i^{\text{ord}}, y_i, \epsilon_i^{\text{diff}}) - x_i^{\text{ord}} \tag{14}$$

Here $r_i^{\text{diff}} \in \mathbb{R}^G$ is the expression residual. Let $B \in \mathbb{R}^{G\times d_{\text{sig}}}$ contain an orthonormal basis of the centered reference-signature span, whose rank is d_{sig}. With the $G \times G$ identity I_G, SILTA forms

$$\begin{aligned} \bar{r}_i^{\text{diff}} &= \text{clip}\big((I_G - \gamma BB^{\mathsf{T}})r_i^{\text{diff}}, -c_{\text{clip}}, c_{\text{clip}}\big)\,, \\ x_i^{\text{hybrid}} &= \big[x_i^{\text{ord}} + \delta_i^{\text{diff}}\alpha_{\text{diff}}\bar{r}_i^{\text{diff}}\big]_+ \end{aligned} \tag{15}$$

The coefficient $\gamma \in [0,1]$ controls signature-direction suppression, $c_{\text{clip}} > 0$ bounds each gene correction, and $\alpha_{\text{diff}} \geq 0$ scales the residual. Both clip and the nonnegative part $[\cdot]_+$ act elementwise. Because $x_i^{\text{ord}} \geq 0$, an inactive indicator returns the ordinary pseudo-spot exactly. Augmentation changes expression while preserving its label y_i and the number of pseudo-spots.

A.5 qNB and calibration details

The frozen signature matrix $R \in \mathbb{R}_+^{K\times G}$ stores one reference profile per type, normalized so that $\sum_j R_{kj} = 1$. We use j as the gene index here, reserving g_i for the spatial latent. For uncalibrated target proportions p_i, the observation path is

$$\rho_{ij} = \sum_{k=1}^{K} p_{ik}R_{kj}, \qquad \tilde{\rho}_{ij} = \frac{\beta_j\rho_{ij}}{\sum_{v=1}^{G}\beta_v\rho_{iv}}, \qquad \mu_{ij}^{\text{obs}} = L_i\tilde{\rho}_{ij} \tag{16}$$

With row-normalized signatures, the base mixture treats cell-number fractions as expression-mixture weights, approximating equal per-cell expression totals across types on the assayed genes. Reference supervision and replay use cell-number proportion labels from reference pseudo-spots.

The input x_{ij}^t is the raw target count when available and an unrounded nonnegative quasi-count otherwise. Its library size is $L_i = \sum_j x_{ij}^t$. The factors $\beta_j > 0$ adjust relative gene response without changing this library total:

$$\log\beta_j = b_{\max}\left(\tanh u_j - \frac{1}{G}\sum_{v=1}^{G}\tanh u_v\right) \tag{17}$$

Here $u \in \mathbb{R}^G$ contains response parameters and $b_{\max} \geq 0$ bounds their range. Setting $b_{\max} = 0$ gives $\beta_j = 1$. For fixed negative-binomial size $\vartheta_j > 0$, the count model has mean μ_{ij}^{obs} and variance $\mu_{ij}^{\text{obs}} + (\mu_{ij}^{\text{obs}})^2/\vartheta_j$. Its average negative log-likelihood is

$$\begin{aligned} \mathcal{L}_{\text{qNB}} = -\frac{1}{N_tG}\sum_{i=1}^{N_t}\sum_{j=1}^{G}\Bigg[&\log\Gamma(x_{ij}^t + \vartheta_j) - \log\Gamma(\vartheta_j) - \log\Gamma(x_{ij}^t + 1) \\ &+ \vartheta_j\log\frac{\vartheta_j}{\vartheta_j + \mu_{ij}^{\text{obs}}} + x_{ij}^t\log\frac{\mu_{ij}^{\text{obs}}}{\vartheta_j + \mu_{ij}^{\text{obs}}}\Bigg] \end{aligned} \tag{18}$$

Γ denotes the gamma function. For continuous nonnegative inputs, SILTA uses the same expression as a quasi-likelihood. The qNB implementation clamps the mean and dispersion to at least 10^{-8}. No zero-inflation term enters this qNB objective.

Hierarchical correction. Let $\mathcal{G}_c = \{k : c(k) = c\}$ be broad group $c \in \{1, \dots, C\}$. The calibrator learns raw type biases b_k^{raw}, optional group biases b_c, and optional temperature parameters v_c. It sets $T_c = \exp(0.5 \tanh v_c)$ and $\tilde{b}_k = b_k^{\text{raw}} - |\mathcal{G}_{c(k)}|^{-1} \sum_{v \in \mathcal{G}_{c(k)}} b_v^{\text{raw}}$ for the hierarchical mode. The ungrouped-bias mode instead uses b_k^{raw} directly and omits b_c. A disabled temperature correction sets $T_c = 1$. These choices define the basic logit map $\mathcal{C}_0(\ell_i) = a_i$ in Equation 10.

For reserved pseudo-spots, let $p_i^{\text{cal}} = \text{softmax}(\mathcal{C}_0(\ell_i)/\tau_{\text{fit}})$, where $\tau_{\text{fit}} > 0$ is the fixed temperature used during fitting. With N_v reserved observations and nonnegative weights η, the fitting loss is

$$\begin{aligned}\mathcal{L}_{\text{cal}} = \frac{1}{N_v} \sum_i D_{\text{KL}}(y_i \| p_i^{\text{cal}}) + \frac{\eta_{\text{macro}}}{K} \sum_k w_k \left[\frac{1}{N_v} \sum_i H_1(p_{ik}^{\text{cal}} - y_{ik}) \right] \\ + \frac{\eta_{\text{mass}}}{K} \sum_k |\bar{p}_k^{\text{cal}} - \bar{y}_k| + \eta_{\text{id}} \Omega \end{aligned} \tag{19}$$

Bars denote means over the reserved observations. The smooth-L_1 penalty is $H_1(e) = e^2/2$ for $|e| \le 1$ and $H_1(e) = |e| - 1/2$ otherwise. For $f_k = \max(\bar{y}_k, 10^{-3})$, define $r_k^{\text{freq}} = f_k^{-1/2}$, $u_k^{\text{freq}} = \text{clip}(r_k^{\text{freq}}/\bar{r}^{\text{freq}}, 0.25, 4)$, and $w_k = u_k^{\text{freq}}/\bar{u}^{\text{freq}}$, where these bars average over types. The identity penalty sums mean squares of the raw parameters: $\Omega = K^{-1} \sum_k (b_k^{\text{raw}})^2 + C^{-1} \sum_c b_c^2 + C^{-1} \sum_c v_c^2$, omitting terms for inactive parameter groups.

Endpoint variants. A scaled correction uses $\mathcal{C}_{\text{end}}(\ell) = \ell + s_{\text{cal}}[\mathcal{C}_0(\ell) - \ell]$ with fixed strength s_{cal}, including identity at zero and full correction at one. The released OSM endpoint instead composes its first group-temperature/type-bias correction with a within-group shape correction. Define the centering operator $[\text{ctr}(v)]_k = v_k - |\mathcal{G}_{c(k)}|^{-1} \sum_{j \in \mathcal{G}_{c(k)}} v_j$. For learned vectors $u^b, u^T \in \mathbb{R}^K$ and fixed nonnegative caps $B_{\text{cap}}, T_{\text{cap}}$,

$$\begin{aligned} d^b = B_{\text{cap}} \, \text{ctr}(\tanh u^b), \qquad d^T = T_{\text{cap}} \, \text{ctr}(\tanh u^T), \\ [\mathcal{C}_{\text{shape}}(a_i)]_k = \frac{a_{ik} + d_k^b}{\text{clip}(\exp d_k^T, 0.5, 2)}, \quad \mathcal{C}_{\text{end}} = \mathcal{C}_{\text{shape}} \circ \mathcal{C}_0 \end{aligned} \tag{20}$$

This refinement fits reference-label KL plus an identity penalty proportional to $K^{-1}(\|d^b\|_2^2 + \|d^T\|_2^2)$, with the first correction fixed. The endpoint temperature τ_{inf} in Equation 10 is separate from the training temperature τ_{frac}. All fitted maps and endpoint settings remain fixed during target inference.

A.6 Computational cost

Let d_e, d_h, d_g, d_m, and D_f denote the encoder-output, expression-latent, spatial-latent, auxiliary, and fused widths, and let H denote the number of attention heads. We use K types, G genes, K auxiliary dominant-type classes, and C calibration groups. Table 6 reports incremental costs relative to an expression-only proportion prediction head. The dense E2 variants materialize all-pairs attention, giving the quadratic target-size term shown in the table. Hypergraph-incidence operators have a different message structure, so this dense bound does not describe every E2 configuration. For the fixed three-block encoder, let $D_E = \max(G, 256, 512)$ bound all block widths. The table uses D_E for attention computation and reserves d_g for the projected spatial latent. Diffusion adds source-training cost but no target inference cost. The auxiliary dominant-type path adds target inference cost only when $\lambda_m > 0$. qNB adds a Stage II training objective, while post-hoc calibration with the backbone fixed adds only a type-level affine transform at inference.

A.7 Evaluation integrity checks

The evaluator checks every configured prediction for shape mismatch, non-finite proportions, negative values, and structural collapse before metric aggregation.

A.8 Native baselines

Table 2 reports the native means for all evaluated methods. OSM SpatialDecon uses the published optimizer with only its 100-gene input guard removed for the 33-gene panel. Moffitt STRIDE

Table 6: **Analytical incremental complexity.** N_p is the number of pseudo-spots, T_{diff} the diffusion steps, and $C_{\mathcal{D}}$ the cost of one diffusion step. These bounds isolate where complexity enters, although they do not replace a matched hardware and tuning-budget study.

Component	Incremental parameters	Incremental time	Activation memory	Prediction-time status
Dense distance-decay E2	$O(D_E^2 + H)$	$O(N_t D_E^2 + N_t^2 D_E)$	$O(HN_t^2 + N_t D_E)$	Dense target spatial path
Additive g projection	$O(D_f d_g)$	$O(N_t D_f d_g)$	$O(N_t D_f)$	Core additive prediction head
Bounded $h \odot g$ interaction	$O(D_f^2)$	$O(N_t D_f^2)$	$O(N_t D_f)$	Controlled, removed by $\lambda_x = 0$
qNB reconstruction regularizer	$O(G)$ trainable response terms	$O(N_t K G)$	$O(N_t G)$	Stage II training only
Hierarchical calibration	$O(K + C)$ for C groups	$O(N_t K)$	$O(N_t K)$	Post-hoc affine logit transform
Auxiliary dominant-type path	$O(Kd_e + Kd_m + D_f d_m)$	$O(N_t(Kd_e + Kd_m + D_f d_m))$ when active	$O(N_t(K + d_m))$	Direct when $\lambda_m > 0$, inactive when $\lambda_m = 0$
Partial diffusion	Generator stored separately	$O(\rho_{\text{diff}} N_p T_{\text{diff}} C_{\mathcal{D}})$	Generator-dependent	Source-side augmentation, absent at target inference

Table 7: **Recorded computational cost across the three evaluation datasets.** Values report mean ± SD across three folds or splits at fixed algorithm seed on one NVIDIA GPU. The records use different hardware models and timing scopes: SILTA seqFISH reports Stage I + Stage II training, whereas the baselines report native end-to-end time. The values document implementation scale and do not define a controlled speed ranking across the evaluated methods.

Dataset	Method	Time (s)	Peak GPU (MiB)	Params
Moffitt	SILTA	119 ± 15	721 ± 15	9.3 M
	cell2location	3,142 ± 139	38 ± 0	127 K
	STdGCN	21 ± 0	431 ± 3	45 K
seqFISH	SILTA	∼63	N/A	9.3 M
	cell2location	2,991 ± 337	71 ± 0	120 K
	STdGCN	18 ± 1	486 ± 17	91 K
OSM	SILTA	1,490 ± 1,169	N/A	9.3 M
	cell2location	5,222 ± 959	22 ± 0	56 K
	STdGCN	22 ± 1	262 ± 1	3 K

returns all three fold predictions through an SC-reference validation fallback. AddaGCN (Ding et al., 2026) follows its native graph construction and alternating adversarial updates. Cell2location (Kleshchevnikov et al., 2022) follows its hierarchical Bayesian count model and native variational-inference protocol under published defaults across all three benchmarks.

A.9 Melanoma observed-marker concordance protocol

The melanoma analysis reports Patient 3 replicate 1 and Patient 4 replicate 1. The macrophage panel contains LST1, C1QB, CD68, AIF1, and FCGR3A, whereas the T-cell panel contains CD3E, CD2, CD8B, CD4, and CD5. The evaluator applies the same log1p(CPM) normalization and correlation implementation to SILTA and ten fixed baselines. Figure 5 reports the main macrophage comparison on 219 aligned spots. Figure 14 reports one T-cell prediction per method on 169 aligned spots.

A.10 Supplementary melanoma and HBC results

Nested selection standardizes HBC evaluation. The nested protocol derives candidate features from 13 predicted cell-type proportions, coarse cell groups, and neighborhood compositions. Five contiguous outer folds evaluate unseen tissue blocks. Within each outer fold, inner spatial cross-validation selects one of three neighborhood sets: $k = 8$, $k = 4, 12$, or $k = 4, 8, 20$. It also selects one of three shallow classifiers: elastic-net classifiers at $C = 0.1$ and 1, or depth-2 histogram gradient boosting. Public pathology labels train and select only these downstream classifiers and never enter deconvolution training or deconvolution-model selection. All methods share the labels, outer folds, candidate grid, and selection protocol. Candidate classifiers use neither absolute coordinates

Table 8: **Observed-marker concordance on two melanoma sections.** We identify the best baseline separately for every endpoint and metric among ten baselines. Margin is SILTA minus the corresponding best baseline under the shared observed-marker protocol.

Section	Cell type	Metric	SILTA	Best baseline (method)	Margin
P3-r1	Macrophages	PCC	**.2111**	.1658 (RCTD)	+.0452
		Spearman	**.1746**	.1624 (RCTD)	+.0123
P4-r1	T cells	PCC	**.1151**	.0645 (HarmoDecon)	+.0506
		Spearman	**.1105**	.0402 (SPOTlight)	+.0703

Table 9: **HBC pathology-region classification under five-fold spatial generalization.** Entries are five-fold mean ± sample SD over the same held-out contiguous spatial folds shown in Fig. 6(b). All methods use identical public labels and folds and the same candidate-grid and inner-selection protocol. Bold marks the leading entry within this matched comparison.

Method	Balanced Acc	Macro-F1	Macro-AUROC
SILTA	**0.7847 ± 0.0547**	**0.7720 ± 0.0522**	**0.9383 ± 0.0187**
SPOTlight	0.7772 ± 0.0305	0.7637 ± 0.0235	0.9347 ± 0.0148
SpatialDecon	0.7716 ± 0.0325	0.7491 ± 0.0224	0.9362 ± 0.0139
Stereoscope	0.7702 ± 0.0167	0.7539 ± 0.0292	0.9321 ± 0.0084
RCTD	0.7667 ± 0.0273	0.7517 ± 0.0352	0.9354 ± 0.0128
CARD	0.7436 ± 0.0439	0.7341 ± 0.0420	0.9267 ± 0.0265
STRIDE	0.7311 ± 0.0354	0.7195 ± 0.0440	0.9170 ± 0.0120
cell2location	0.7284 ± 0.0417	0.7165 ± 0.0388	0.9156 ± 0.0281
Tangram	0.7279 ± 0.0193	0.7089 ± 0.0274	0.9150 ± 0.0101
STdGCN	0.7238 ± 0.0323	0.7066 ± 0.0387	0.9193 ± 0.0145
AddaGCN	0.6512 ± 0.0655	0.6267 ± 0.0668	0.8796 ± 0.0421

nor probability post-processing. Within each fold, 1,000 paired spatial-block bootstrap resamples estimate uncertainty for SILTA-minus-best-baseline differences, and the source data retain the full 95% intervals for every outer fold and metric.

A.11 SILTA Leads Selected Spearman Endpoints

Observed HBC marker expression provides an indirect spatial check because spot-level cell-type fraction ground truth is unavailable. We align 3,798 spots and evaluate SILTA with ten baselines for two selected-compartment endpoints. For each endpoint, five marker maps and a five-marker consensus share the same tissue coordinates. The evaluator correlates each predicted fraction map with each observed marker map. For each selected compartment, we report the mean of five marker-wise PCC values and the mean of five marker-wise Spearman values under the same endpoint-specific scoring rule with fixed spot alignment and identical aggregation.

At the Fibroblast endpoint, SILTA ranks first among eleven methods by both mean PCC and mean Spearman (0.1223, Fig. 7). At the Macrophage/DC/monocyte endpoint, SILTA ranks first only by mean Spearman (0.1028, Fig. 8). These scores summarize spatial concordance between predicted fractions and observed expression. Because HBC supplies no spot-level fraction labels, this comparison remains an indirect spatial consistency check.

Each marker display divides positive values by its nonzero 95th percentile, clips the result to $[0, 1]$, and keeps zeros at zero. The displayed consensus averages the five normalized maps and remains outside the ranking metric. Within each cell-type label, all eleven prediction maps share a color scale from zero to their pooled 99th percentile, with larger values clipped at the upper color limit. The spot set and coordinate orientation remain fixed across maps. The rank panels report independent PCC and Spearman orderings for the same eleven methods and aligned spots.

DLPFC provides a selected Mix_5 endpoint across 3,639 aligned spots. The evaluator correlates each predicted fraction map with observed expression for GRIK1, ELFN1, ROBO2, PDIA2, and UPP1, then averages the five PCC values and five Spearman values separately. SILTA ranks first among eleven methods by mean PCC (0.08952) and mean Spearman (0.13388) (Fig. 9). These scores

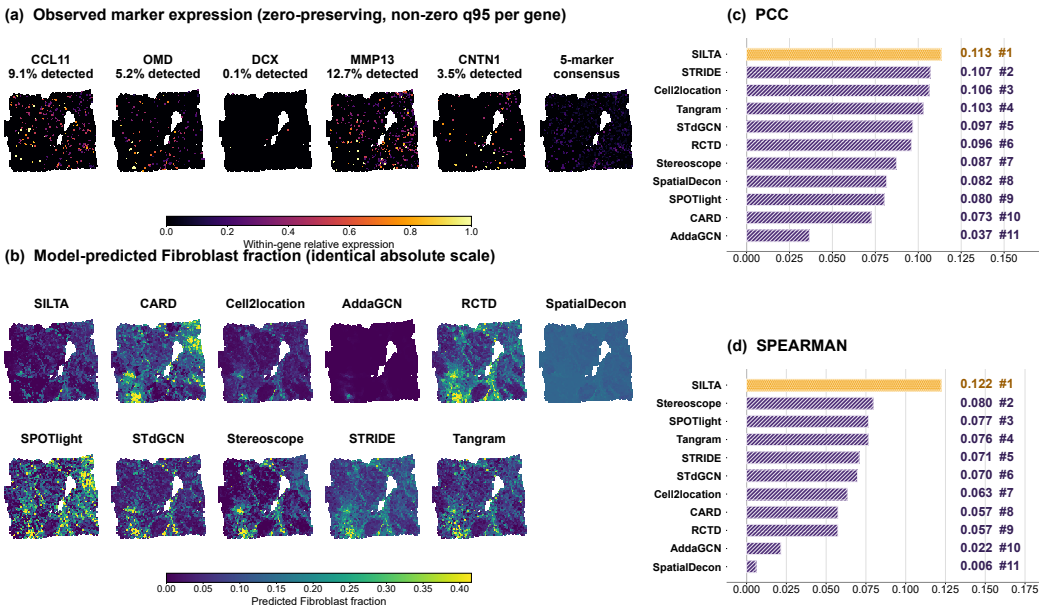


Figure 7: **HBC Fibroblast marker spatial comparison.** (a) Five observed markers use separate zero-preserving nonzero-q95 display scales, and their mean forms the displayed consensus. (b) SILTA and ten baseline maps share a color scale from zero to the pooled prediction q99 for Fibroblast, with larger values clipped at the upper limit. (c,d) Independent rankings report mean per-marker PCC and Spearman. At this selected-compartment endpoint, SILTA ranks first by both PCC and Spearman. HBC lacks spot-level fraction ground truth, so the figure provides an indirect spatial-consistency check rather than fraction-accuracy evidence.

measure per-marker spatial concordance for the selected Mix_5 label. DLPFC provides no spot-level fraction ground truth for this comparison on the aligned target spots.

A.12 Observation-model specification

Reference signatures have nonnegative rows that sum to one and share the target gene order. Integer raw inputs use the NB likelihood. For continuous nonnegative inputs, we preserve the values and extend the NB kernel with the Gamma function. qNB retains the reference signature, library-size, and gene-dispersion constraints while accepting finite nonnegative continuous observations. Primary benchmark SILTA outputs use qNB, while qNB-off configurations appear only as ablation controls. We center and bound gene-response corrections. The evaluator records correction saturation, predicted sharpness, branch gradients, and zero/shuffle responses.

B Additional Method Definitions

This appendix connects input preprocessing, latent representations, prediction, observation alignment, and calibration. The shared notation below separates the forward computation from the parameter masks that govern each training event.

B.1 Shared network and graph construction

Notation and dimensions. Uppercase X, S, R, E_m, B denote matrices. Lowercase per-observation quantities are column vectors, with coordinates indexed by a second subscript. Batched implementations store transposed vectors in rows. $\mathrm{diag}(v)$ creates a diagonal matrix, I_d is the d-dimensional identity, $\odot$ is the elementwise product, and concat stacks head outputs along their feature dimension. Every affine map includes a bias unless its configuration omits it. For $P : \mathbb{R}^a \to \mathbb{R}^b$, $P(v) = W_P v + b_P$, with $W_P \in \mathbb{R}^{b\times a}$ and $b_P \in \mathbb{R}^b$. Layer normalization acts along features.

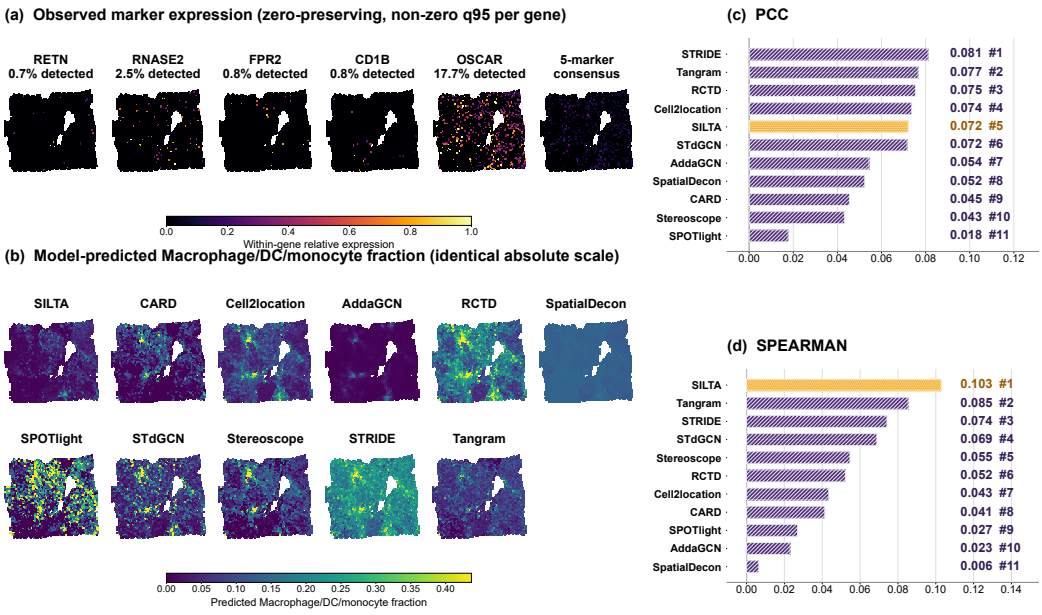


Figure 8: **HBC Macrophage/DC/monocyte marker spatial comparison.** (a) Five observed markers use separate zero-preserving nonzero-q95 display scales, and their mean forms the displayed consensus. (b) SILTA and ten baseline maps share a color scale from zero to the pooled prediction q99 for Macrophage/DC/monocyte, with larger values clipped at the upper limit. (c,d) Independent rankings report mean per-marker PCC and Spearman. At this selected-compartment endpoint, SILTA ranks first by Spearman. HBC lacks spot-level fraction ground truth, so the figure provides an indirect spatial-consistency check rather than fraction-accuracy evidence.

Forward-path notation. N denotes the number of observations in the current source or target forward pass, with each observation occupying one expression-matrix row.

Quantity	Shape	Meaning
X^r, y^r	$N_r \times G,\ N_r$	Reference expression and type-label vector
X^p, Y^p	$N_p \times G,\ N_p \times K$	Augmented pseudo expression and sampled proportions
X^t, S^t	$N_t \times G,\ N_t \times 2$	Nonnegative target observations and coordinates
$\bar{X} = \mathcal{T}(X)$	$N \times G$	Encoder input under the dataset's fixed preprocessing
$A_{\text{expr}}, A_{\text{sp}}$	$N \times N,\ N_t \times N_t$	Expression adjacency and target spatial adjacency
$e_i^{(1)}, e_i^{(2)}$	d_e	Normalized encoder outputs
$\mu_{b,i}, \sigma^2_{b,i}$	d_b	Mean and diagonal variance of latent $b \in \{h, g, z\}$
h_i, g_i, m_i	d_h, d_g, d_m	Expression, spatial, and auxiliary representations
E_m, q_i^m	$K \times d_m,\ K$	Type embeddings and dominant-type probabilities
$h_{0,i}, g_{0,i}, m_{0,i}, f_i, z_i$	D_f	Aligned branches, fused features, and readout input
$\ell_i, p_i, p_i^{\text{final}}$	K	Raw logits, training proportions, and final proportions
R, u, ϑ	$K \times G,\ G,\ G$	Frozen signatures, response parameters, and fixed NB sizes
μ_i^{obs}	G	qNB mean on the nonnegative observation scale

Preprocessing and encoder outputs. The reference and target use the same gene order. The map $\mathcal{T}$ denotes the fixed preprocessing recipe for that dataset and does not act on composition labels. It distinguishes encoder features from the observation-scale X^t used by qNB. The Moffitt release uses per-gene standardization without `log1p`, whereas count-normalized routes apply `log1p` once after aggregation. No equation assumes that standardized encoder features remain nonnegative.

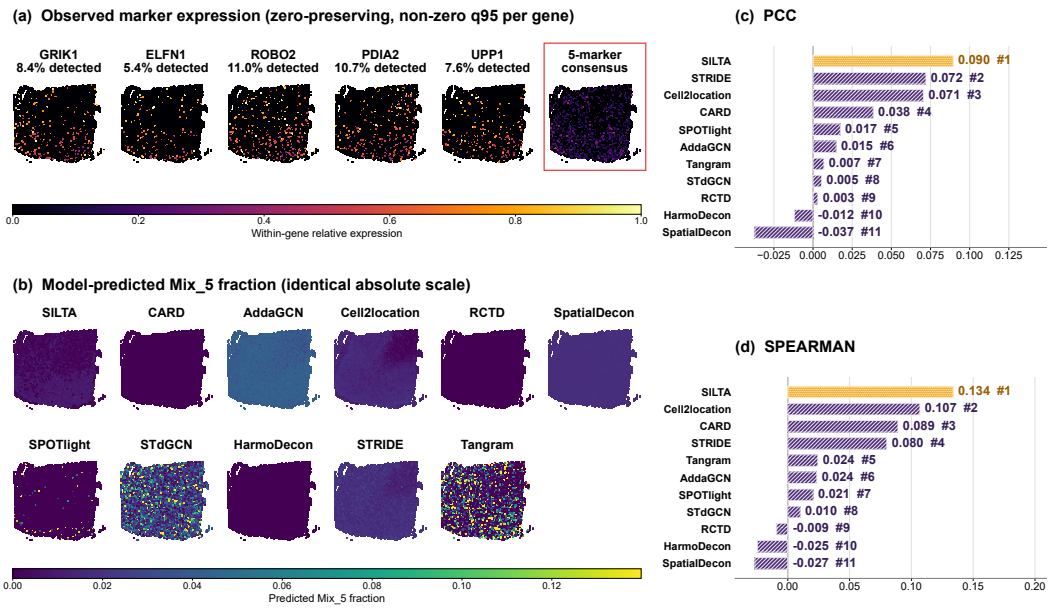


Figure 9: **DLPFC Mix_5 marker spatial comparison.** (a) Five observed marker maps cover 3,639 aligned spots and use separate zero-preserving nonzero-q95 display scales, and their mean forms the displayed consensus. (b) SILTA and ten baseline maps share a color scale from zero to the pooled prediction q99 (0.13989), with larger values clipped at the upper limit. (c,d) Independent rankings report mean per-marker PCC and Spearman. SILTA ranks first for both metrics at the selected Mix_5 endpoint. DLPFC provides no spot-level fraction ground truth, so the figure supplies an indirect per-marker check rather than spot-level fraction-accuracy evidence.

Table 10: **Shared architecture.** Configuration choices select graph operators, fused readout, and optional routes without changing the type order.

Component	Shape	Operation
E1/E2 output	$N \times 512$	Three attention blocks and encoder normalization
Gaussian parameter heads	$512 \to 192$ for h, g	Separate means and log variances, log variance clipped to $[-6, 2]$
Auxiliary head/embedding	$512 \to K,\ K \times 64$	Dominant-type probabilities and $E_m^{\mathsf{T}} q_i^m$
Branch projections	$P_h, P_g : 192 \to D_f, P_m : 64 \to D_f$	Affine maps and branch-specific LayerNorm
Interaction	$P_x : D_f \to D_f$	Project $h_{0,i} \odot g_{0,i}$, then weight by λ_x
Fused readout	$F_z : D_f \to D_f$	Identity or normalized Gaussian latent
Fraction head	$D_f \to 256 \to K$	LayerNorm, linear, ReLU, dropout 0.1, linear logits
Residual diffusion	$G \to 512 \to 128 \to 64$ encoder	Symmetric decoder and conditional denoising MLP

For the graph-masked expression route, $E^{(1)} = \mathrm{E1}(\bar{X}, A_{\mathrm{expr}})$. For target data, $E^{(2)} = \mathrm{E2}(\bar{X}^t, S^t, A_{\mathrm{sp}})$. Their rows are $(e_i^{(1)})^{\mathsf{T}}$ and $(e_i^{(2)})^{\mathsf{T}}$. Dense E1 variants omit the adjacency mask. The three encoder block widths are $G \to 256 \to 256 \to 512$, with normalization before the latent projectors. E1 graph blocks use ReLU. Distance-aware E2 blocks use their configured residual and activation operators. The shared latent dimensions are $d_h = d_g = 192$ and $d_m = 64$, while D_f is the configured fusion width. The dominant-type temperature is $\tau_m = 1.5$.

Attention output. The graph-construction scale d_{nn} and attention scale d_{att} have separate roles. The Moffitt release sets d_{att} to the median positive edge length in A_{sp}, while the seqFISH dense operator uses $d_{\mathrm{att}} = d_{\mathrm{nn}}$. Both normalize physical distances before attention. For the dense operator

in Equation 3, let $t_i \in \mathbb{R}^{d_{\text{in}}}$ be a block's input and let H be its number of heads. Each head projects $\text{LN}_{\text{in}}(t_i)$ into $q_i^{(r)}, k_i^{(r)}, v_i^{(r)}$. With affine maps $P_{\text{res}} : \mathbb{R}^{d_{\text{in}}} \to \mathbb{R}^{d_{\text{out}}}$ and $P_{\text{out}} : \mathbb{R}^{Hd_r} \to \mathbb{R}^{d_{\text{out}}}$,

$$t_i^{\text{next}} = \text{GELU}\Big(\text{LN}_{\text{out}}\Big[P_{\text{res}}(t_i) + \text{sigmoid}(\omega) P_{\text{out}}\Big(\text{concat}_{r=1}^{H} o_i^{(r)}\Big)\Big]\Big) \tag{21}$$

The scalar ω controls the learned message gate. Each successive block consumes t_i^{next} from the preceding attention block. The fixed-bandwidth form holds σ_r constant. Hypergraph-incidence forms use coordinate-derived incidence messages instead of this dense sum. Graph regularization uses the non-diagonal edges of A_{sp} regardless of the selected encoder operator.

Gaussian and fused readout paths. For $b \in \{h, g\}$, the projector produces mean $\mu_{b,i}$ and log-variance $\nu_{b,i}$, with $\sigma_{b,i} = \exp(\nu_{b,i}/2)$. An enabled stochastic route uses $b_i = \mu_{b,i} + s_b \sigma_{b,i} \odot \epsilon_{b,i}$, where $\epsilon_{b,i} \sim \mathcal{N}(0, I_{d_b})$ and $s_b \geq 0$ is the configured sampling scale. Mean routes use $b_i = \mu_{b,i}$. The aligned vectors $h_{0,i}, g_{0,i}, m_{0,i}$ have width D_f. Subscript 0 identifies normalized projections.

For the direct readout, $F_z(f_i) = f_i$. For the Gaussian readout, affine heads $P_{\mu_z}, P_{\nu_z} : \mathbb{R}^{D_f} \to \mathbb{R}^{D_f}$ produce $\mu_{z,i} = P_{\mu_z}(f_i)$ and $\nu_{z,i} = \text{clip}(P_{\nu_z}(f_i), -10, 10)$, followed by

$$z_i = \text{LN}_z\left(\mu_{z,i} + \exp(\nu_{z,i}/2) \odot \epsilon_{z,i}\right), \qquad \epsilon_{z,i} \sim \mathcal{N}(0, I_{D_f}) \tag{22}$$

Training samples this route when enabled, whereas evaluation uses $z_i = \text{LN}_z(\mu_{z,i})$. Thus $d_z = D_f$ for its Gaussian penalty. Source and target share the fraction head, but source reference passes bypass the spatial contribution with $\chi = 0$. This explicit bypass removes projection and interaction offsets as well as spatial features in both training and evaluation. The shorthand $g = 0$ in the experimental descriptions denotes this whole-path bypass $\chi = 0$.

The resulting chain is $X \to \mathcal{T}(X) \to (E^{(1)}, E^{(2)}) \to (h, g, m) \to f \to z \to \ell$. Training uses $\text{softmax}(\ell/\tau_{\text{frac}})$. The qNB path maps these proportions through R to μ^{obs}. Final inference instead applies the fixed endpoint calibrator to raw logits and then uses τ_{inf}, as in Equation 10. An identity endpoint map covers configurations with temperature-only inference.

B.2 Pseudo-spot and diffusion contracts

Generators aggregate reference-cell counts before applying $\mathcal{T}$. Empirical spots sample cells by reference frequency. Balanced spots sample cell types uniformly. Near-pure anchors place most cells in one balanced type. Hard pairs select similar-signature types within a broad group and include both composition orders. Each generator records y_i from sampled cell counts.

The diffusion indicator δ_i^{diff}, residual r_i^{diff}, and clipping operator in Appendix A.4 change only expression while preserving sampled proportion labels. They are distinct from the fused representation z_i, spatial coordinates s_i, and numerical stabilizers. Reserved reference pseudo-spots use the same type vocabulary and fixed broad groups as the corresponding training fold.

B.3 Complete objective definitions

Labelled fraction and auxiliary losses. For probability vectors $y, p \in \Delta^{K-1}$, $D_{\text{KL}}(y\|p) = \sum_{k:y_k>0} y_k \log(y_k/p_k)$. We use the same average KL definition of $\mathcal{L}_{\text{frac}}$ in Equation 6 and this appendix. Cross-entropy differs from this value by the label entropy, even though fixed labels give the same prediction gradients. For a batch of N labelled pseudo-spots, let $k_i^* = \arg\max_k y_{ik}$ and $v_i^m = C_{\text{mode}}(e_i^{(1)})$. With prototype $e_k^m = (E_m[k,:])^\intercal$ and auxiliary classifier $C_m : \mathbb{R}^{d_m} \to \mathbb{R}^K$,

$$\mathcal{L}_{\text{CE}}^m = -\frac{1}{N}\sum_i \log[\text{softmax}(v_i^m)]_{k_i^*}, \quad \mathcal{L}_{\text{soft}}^m = \frac{1}{N}\sum_i D_{\text{KL}}(y_i \| q_i^m),$$

$$\mathcal{L}_{\text{proto}} = \frac{1}{N}\sum_i [1 - \cos(m_i, e_{k_i^*}^m)], \qquad \mathcal{L}_{\text{CE}}^{\text{emb}} = -\frac{1}{N}\sum_i \log[\text{softmax}(C_m(m_i))]_{k_i^*} \tag{23}$$

Thus $\mathcal{L}_{\text{mode}} = \lambda_{\text{CE}}^m \mathcal{L}_{\text{CE}}^m + \lambda_{\text{soft}}^m \mathcal{L}_{\text{soft}}^m + \lambda_{\text{proto}} \mathcal{L}_{\text{proto}} + \lambda_{\text{CE}}^{\text{emb}} \mathcal{L}_{\text{CE}}^{\text{emb}}$. These nonnegative coefficients select active auxiliary terms. A zero coefficient disables that term. Configured class weights and balanced sampling affect dominant-label supervision, not the composition labels. Cosine similarity uses normalized feature vectors, with a positive numerical floor on vector norms.

Gaussian regularization. For latent b of width d_b, its mean per-element Gaussian penalty is

$$K_b = \frac{1}{2Nd_b} \sum_{i=1}^{N} \sum_{j=1}^{d_b} \left[\mu_{b,ij}^2 + \exp(\nu_{b,ij}) - 1 - \nu_{b,ij}\right]$$
$$S_c(K) = \begin{cases} K, & K \le c \\ c + \log(1 + K - c), & K > c \end{cases} \tag{24}$$

Here $\nu_b = \log \sigma_b^2$, and $c > 0$ is a configured soft cap. A disabled cap gives $S_\infty(K) = K$. For source or target domain $d \in \{p, t\}$, $\mathcal{L}_{\text{reg}}^d = \sum_{b \in \{h,g,z\}} \kappa_b^d S_{c_b}(K_b)$. The weights $\kappa_b^d \ge 0$ and active latent routes come from the run configuration. An inactive spatial path contributes no g penalty. This definition includes the fused-z regularizer and makes the $1/(Nd_b)$ reduction explicit. It does not assume identical KL weights across datasets.

Spatial regularization. Let $\mathcal{E} = \{(i,j) : A_{\text{sp},ij} = 1,\ i \ne j\}$ and $\bar{x}_i^t = (\bar{X}_{i,:}^t)^\mathsf{T}$. The source of boundary weights is expression, whereas the physical graph determines the eligible pairs:

$$\mathcal{L}_{\text{smooth}} = \frac{1}{|\mathcal{E}|} \sum_{(i,j)\in\mathcal{E}} \|p_i - p_j\|_2^2,$$
$$d_{ij}^{\text{expr}} = \max\{0, 1 - \cos(\bar{x}_i^t, \bar{x}_j^t)\}, \qquad \sigma_{\text{expr}} = \text{median}_{(i,j)\in\mathcal{E}}\, d_{ij}^{\text{expr}},$$
$$w_{ij}^{\text{B}} = \max\{w_{\min}, \exp[-(d_{ij}^{\text{expr}})^2/(\sigma_{\text{expr}}^2 + \epsilon_{\text{num}})]\}, \quad \mathcal{L}_{\text{BATV}} = \frac{\sum_{(i,j)\in\mathcal{E}} w_{ij}^{\text{B}} \|p_i - p_j\|_1}{\sum_{(i,j)\in\mathcal{E}} w_{ij}^{\text{B}}} \tag{25}$$

Here $w_{\min} > 0$ is the boundary-weight floor and $\epsilon_{\text{num}} > 0$ stabilizes division. Boundary weights apply only to $\mathcal{L}_{\text{BATV}}$, leaving $\mathcal{L}_{\text{smooth}}$ unchanged. Empty edge sets contribute zero spatial loss. Equations 16-18 define the observation term. Reference pseudo-spots determine fixed dispersion ϑ_j using clipped method-of-moments estimation, with $\vartheta_j = 10^4$ for genes without positive excess variance.

Objective scope. Equations 6 and 8 connect the fraction, auxiliary, Gaussian, observation, and spatial terms to the forward path. Their coefficients and activation masks vary across the benchmark and matched experiments. The diffusion generator has its own reference-side training: noise prediction compares predicted and sampled diffusion noise, and autoencoder reconstruction compares decoded and input expression. These generator losses do not constitute additional target-proportion labels. Optional reconstruction and domain-adversarial modules in the general codebase are separate configuration choices, not mandatory terms in the two equations. The matched experiments in Appendix D isolate their reported settings.

Replay and calibration losses. For a replay batch $\mathcal{B}_r$ of labelled pseudo-spots, define

$$\mathcal{L}_{\text{replay}} = \frac{1}{|\mathcal{B}_r|} \sum_{i\in\mathcal{B}_r} D_{\text{KL}}\left(y_i \| \text{softmax}(\ell_i^{\chi=0})\right) \tag{26}$$

Replay uses raw logits with unit temperature in the matched replay implementation. Its nonnegative weight λ_{replay} controls the source supervision added during adaptation. A separate replay step or an accumulated target-plus-replay optimizer step can implement this loss, but both must preserve the replay gradient mask. It does not carry target qNB gradients into reference events.

After adaptation, reserved reference logits and labels fit the calibration objectives in Appendix A.5. A 70/30 split of reference-validation pseudo-spots selects calibrator settings, followed by refitting on that full set. Those pseudo-spots exclude the 20% subset used for Stage II reference-retention evaluation. Both derive from the fold-specific reference cell pool without target fraction labels.

B.4 Optimization, stopping, and inference

Let θ_{ref} contain E1, its projectors, the auxiliary heads and embeddings, and reference-side fusion maps. Let θ_{sp} contain E2, its projectors, and spatial/interaction maps. Let θ_{out} contain the shared fused readout and fraction head. Only trainable parameters in these sets enter an optimizer. The

qNB response vector is u, while calibration parameters are ϕ. Normalization parameters follow the block that owns them for optimizer membership and gradient masking.

For event e at update t, let M_e be a diagonal zero-one mask over the concatenated parameter vector θ and let η_e contain per-parameter learning rates. The optimizer receives $M_e \nabla_\theta \mathcal{L}_e$, with zero learning rate for frozen entries. In gradient-descent notation,

$$\theta^{(t+1)} = \theta^{(t)} - \eta_e \odot M_e \nabla_\theta \mathcal{L}_e \tag{27}$$

This equation specifies gradient ownership, not a replacement for the configured optimizer and its moment estimates. The allowed parameter set is the intersection of the event mask, the optimizer's parameter groups, and the unfrozen parameters.

Algorithm 1 SILTA forward computation and role-specific training

Input: X^r, y^r, X^t, S^t, fixed gene/type order, preprocessing $\mathcal{T}$, and the run's operator, loss, and update configuration.
Output: $P^{\text{final}} \in [0,1]^{N_t \times K}$ with unit row sums.

1. Construct labelled pseudo-spots (X^p, Y^p) with Equations 1 and 15. Reserve reference pseudo-spots for validation and calibration. Build expression and spatial graphs on the corresponding inputs.
2. **Stage I:** Compute the source forward path with $\chi = 0$. Optimize $\mathcal{L}_{\text{content}}$ over eligible $\theta_{\text{ref}}, \theta_{\text{out}}$. Freeze θ_{sp} and u.
3. **Stage II target events:** Compute target logits with $\chi = 1$, then p_i and the qNB mean μ_i^{obs}. During a configured spatial-only interval, optimize $\mathcal{L}_{\text{target}}$ over eligible θ_{sp} and u. During joint refinement, also allow selected $\theta_{\text{ref}}, \theta_{\text{out}}$ at their configured lower rate. Skip an interval when its configured length is zero.
4. **Reference replay:** At scheduled replay events, recompute source logits with $\chi = 0$ and use $\lambda_{\text{replay}} \mathcal{L}_{\text{replay}}$. Allow eligible reference/readout parameters, with zero replay gradient to θ_{sp} and u. Accumulated target and replay losses retain their separate gradient masks.
5. **Calibration:** Fix the trained backbone. Obtain reserved reference logits, fit the configured endpoint map $\mathcal{C}_{\text{end}}$ using reference labels, and fix its parameters ϕ.
6. **Inference:** Use posterior means on target data, keep $\chi = 1$, and return $p_i^{\text{final}} = \text{softmax}[\mathcal{C}_{\text{end}}(\ell_i)/\tau_{\text{inf}}]$ for every target spot.

Algorithm 1 distinguishes four learning events and the final calibration fit. A spatial-only interval, joint interval, or optional loss can be inactive in a particular experiment. seqFISH and OSM use three cell-disjoint reference/target splits, while Moffitt uses three leave-one-animal-out folds. We average three model seeds within each fold or split before summarizing the three top-level means.

C Stage-Wise Matched Comparisons

These matched comparisons keep the checkpoint, split, perturbation rule, and update budget fixed within each experiment. Tables 11 and 12 isolate reference-side and target-side type recovery, Table 13 separates the OSM update strategies, and Table 3 tests recovery after complete spot-level compositions are displaced while global composition remains fixed.

Table 11: **Stage I correction of perturbed composition boundaries on reserved reference pseudo-spots.** A fixed margin switches 30% of predictions to a reference-similar type. Correction updates E1, the auxiliary objective, and the fraction readout while E2 stays frozen, and it never uses target data. This isolates source-side boundary recovery under reference supervision alone.

Composition state	Correction	RMSE↓	JSD↓	Dom. BAcc↑	Hard-pair↑	Rare recall↑
Clean	Off	.04698	.19931	.61187	.71917	.67338
Clean	On	**.04520**	**.18320**	**.61270**	**.73672**	**.68563**
Perturbed	Off	.05684	.23245	.45891	.57310	.48940
Perturbed	On	**.05171**	**.19753**	**.55553**	**.65276**	**.64230**

Table 12: **Dominant-prediction perturbation and Stage II adaptation.** Within each matched run, the four experiments share the Stage I checkpoint, target split, and perturbation rule within the evaluation fold and an identical Stage II update budget. (a) Fraction metrics over all target spots. (b) Ground-truth (GT) classification on the same switched spots for dominant-prediction recovery.

(a) Target composition estimates

Composition state	Stage II	RMSE↓	JSD↓	SSIM↑	PCC↑
Clean	Off	.051812	.600331	.571133	.638790
Perturbed	Off	.054449	.610619	.513647	.570801
Clean	On	**.037983**	**.449568**	**.823328**	**.811772**
Perturbed	On	.038217	.451008	.820616	.808951

(b) GT classification on switched spots (%)

Composition state	Stage II	Dominant Acc.↑	Balanced Acc.↑	Macro-F1↑
Clean	Off	42.12	33.89	30.30
Perturbed	Off	12.20	12.73	11.25
Clean	On	53.59	52.01	45.28
Perturbed	On	**54.57**	**53.75**	**46.79**

For each run, dominant recovery is the fraction of spots whose perturbation changed the dominant prediction and whose Stage II output matches the dominant prediction from the paired clean Stage II run. The switched-spot denominators are 169, 170, and 160 in B-M1, B-M2, and B-M3, respectively, for each model seed. We first average the three seed-level rates within each fold and then report the mean ± SD across three folds: $89.62\% \pm 1.43$ percentage points. Panel (b) uses the same fold-first aggregation and compares locked predictions with evaluator-private GT labels on those switched spots. Balanced accuracy and macro-F1 average over GT-present types in the switched subset.

Table 13: **Matched Stage II factorial on OSM.** The five experiments share one Stage I checkpoint, target split, update budget, and calibration procedure. This matched configuration differs from the benchmark configuration in Table 2 and is interpreted only through within-factorial contrasts.

Configuration	RMSE↓	JSD↓	SSIM↑	PCC↑
No Stage II update	.07079	.50089	.73734	.73694
Direct joint	.06863	.47754	.78040	.77276
Spatial-only	.06761	.47360	.78721	.77979
Replay only	**.06636**	**.46462**	**.79590**	**.78871**
Spatial-only + replay	.06650	.46501	.79498	.78777

Allocation-recovery protocol. Table 3 reports a deterministic long-distance derangement of full proportion and logit rows across 30% of target spots while preserving global composition. For each run, allocation recovery is the fraction of selected, row-permuted spots whose post-Stage II dominant prediction matches the paired clean Stage II prediction. The selected-spot denominators are 169, 170, and 160 in B-M1, B-M2, and B-M3, respectively, for each model seed. We first average the three seed-level rates within each fold and then report the mean ± SD across three folds: $83.93\% \pm 1.12$ percentage points. The denominator includes selected spots whose dominant label remains unchanged after permutation. GT classification compares predictions with evaluator-private labels on the same selected spots, using the same fold-first aggregation.

D Configured Component Ablations

All appendix tables use the cross-fold evaluation protocol. Moffitt uses three leave-one-animal-out folds, whereas OSM and seqFISH use three predeclared cell-disjoint splits. Each top-level unit contains three nested model seeds. Values first average nested seeds within each top-level unit and then summarize the three unit means except where noted. Each table reports the configuration and aggregation rule used by its source experiment for direct auditability.

Table 14: **Consolidated ablation summary.** Each row reports a matched contrast between the named configuration and its control. Positive Δ values favor the configuration named on the left. Results aggregate three folds or splits, with three model seeds averaged within each fold or split before cross-unit aggregation for every reported comparison.

Component	Dataset	ΔRMSE	ΔJSD	ΔSSIM	ΔPCC
Stage II parameter-update rules					
Full Stage II vs no Stage II update	Moffitt	+.0123	+.1638	+.2686	+.1943
Direct joint vs no Stage II update	OSM	+.0022	+.0234	+.0431	+.0358
Spatial-only vs direct joint	OSM	+.0010	+.0039	+.0068	+.0070
Replay only vs direct joint	OSM	+.0023	+.0129	+.0155	+.0160
Expression-spatial branch cooperation					
$h + g$ vs h-only	Moffitt	+.002	+.037	+.021	+.021
$h + g$ vs h-only	OSM	+.002	+.006	+.012	+.012
Target-expression reconstruction and post-hoc calibration					
qNB on vs off ($C = 0$)	Moffitt	+.00868	+.04386	+.07838	+.08329
Calibration on vs off ($Q = 1$)	Moffitt	+.00233	−.00410	+.01870	+.02012
qNB on vs off	OSM	+.004	+.047	+.097	+.056
Calibration on vs off	OSM	+.006	.000	+.035	+.037

Table 15: **Stage II endpoints on Moffitt at step 300.** Reference-side metrics use reserved reference pseudo-spots, whereas target metrics use ground-truth proportions. All experimental conditions share the Stage I checkpoint in the first row. Global composition drift is the L_1 change in mean predicted composition from that checkpoint before Stage II begins.

	Reference retention				Target accuracy	
Arm	RMSE↓	Hard-pair↑	Rare recall↑	Global drift↓	RMSE↓	SSIM↑
Stage I output	.0470	.718	.668	N/A	.0531	.5422
Direct joint	.0536	.683	.685	.289	**.040789**	.8104
Spatial-only	.0500	.692	.682	.217	.040805	.8106
Replay only	**.0447**	.741	.689	**.109**	.040803	.8107
Spatial-only + replay	.0448	**.742**	**.690**	**.109**	.040804	**.8108**

D.1 Matched ablations isolate update roles, branch fusion, and expression reconstruction

We verify each design choice through matched contrasts where only one component changes. Table 14 consolidates the key results, and matched controls follow in Appendix D.

Stage II operations. Section 3.2 reports matched Stage II controls. On OSM's 33-gene, 31-type panel, reference replay provides additional supervision during target adaptation (Table 13). On Moffitt, the matched comparison highlights reference retention: adding replay to direct joint adaptation reduces source RMSE from .0536 to .0447 and global drift from .289 to .109 (Table 15).

Spatial branch and interventions. Adding g improves RMSE by about 0.002 on both Moffitt and OSM and improves all four primary metrics (Table 14). Coordinate shuffle degrades all four Moffitt metrics (Appendix D.12), showing that the spatial branch produces location-dependent corrections, although this intervention does not isolate one particular graph topology.

Expression reconstruction and calibration. The matched Moffitt 2×2 factorial separates qNB from post-hoc calibration. qNB improves all four primary metrics with calibration either off or on, and all three held-out animals agree in direction. With qNB active, calibration further improves RMSE, SSIM, and PCC but increases JSD by 0.00410. We therefore use qNB with calibration as the integrated benchmark setting and reserve qNB without calibration for the JSD-specific ablation. On OSM, qNB gives the largest single-component change in JSD (+0.047) and SSIM (+0.097), while calibration mainly improves RMSE, SSIM, and PCC. On seqFISH, calibration has the larger readout effect and qNB adds a smaller gain. Appendix D.3 reports the full Moffitt factorial under the matched design, with separate contrasts for both components.

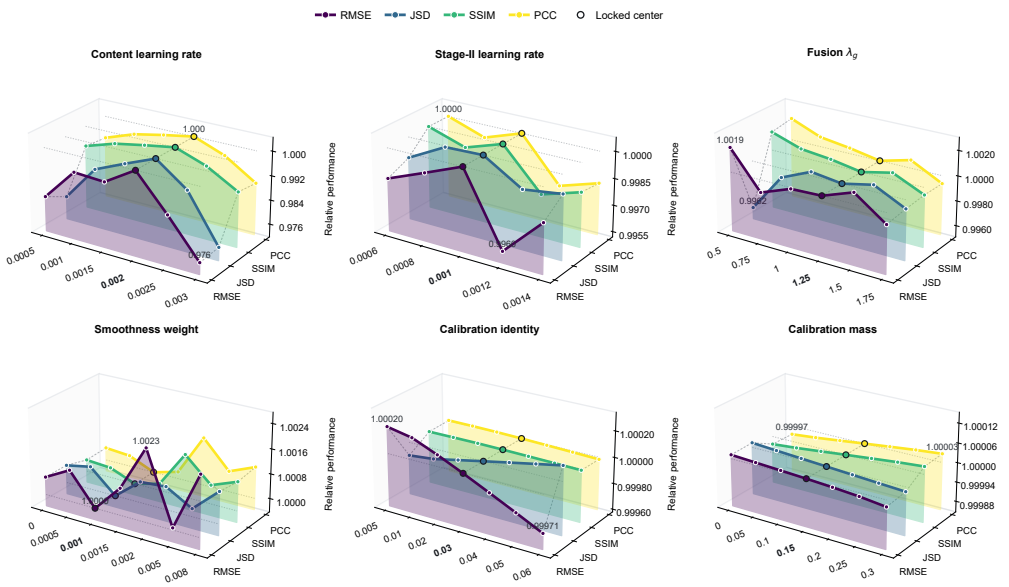


Figure 10: **Local hyperparameter plateaus on Moffitt.** Each bar gives the mean of three Moffitt benchmark folds. The zero plane marks the full model's selected setting in each sweep.

D.2 Hyperparameter sensitivity

We sweep one parameter at a time around the selected Moffitt configuration. Each bar summarizes the three leave-one-animal-out folds under the matched benchmark configuration.

Four training parameters form local plateaus across the primary metrics (Figure 10): expression-branch learning rate 0.0015-0.0025, Stage II learning rate 0.0008-0.0012, $\lambda_g = 1.0$-1.5, and smoothness weight 0.0005-0.0015. Calibration shrinkage weight 0.02-0.04 and mass weight 0.10-0.20 form two further plateaus. The selected result remains close to the center across each interval.

Appendix D.17 reports the remaining sweeps. Stage I steps, Stage II steps, spatial-only steps, BATV weight, diffusion fraction, and diffusion residual scale keep the primary metrics close to the selected setting over their stated intervals, while Rare-F1 varies more. qNB weight produces the largest monotonic improvement in the primary metrics over 0.30-0.60. Calibration strength over 0.75-1.25 trades RMSE, SSIM, and PCC against JSD and Rare-F1. The effective-temperature scan below shows a related metric trade-off: higher temperatures improve RMSE and PCC but worsen JSD.

Table 16: **Effective-temperature sensitivity on Moffitt.** Each row reports the benchmark mean at the specified effective inference temperature. Bold marks the best value in each metric column.

T	RMSE↓	JSD↓	SSIM↑	PCC↑
0.95	0.040596	**0.439904**	0.826773	0.813748
1.00	0.039656	0.442921	0.830189	0.816649
1.05	0.038917	0.446641	0.832263	0.819006
1.10	0.038366	0.450935	**0.833035**	0.820867
1.15	**0.037986**	0.455687	0.832546	**0.822280**

The scan shows a metric trade-off, with higher temperatures improving RMSE/PCC but worsening JSD, and the prespecified benchmark configuration remains the headline setting.

For display, we invert the RMSE and JSD ratios before normalization so that a positive relative change always denotes an improvement for every displayed metric.

Table 17: **Moffitt qNB × post-hoc-calibration factorial.** The qNB-on/calibration-on row is the integrated benchmark configuration. Bold marks the column optimum within this factorial, and the integrated row retains one configuration across all reported metric columns.

qNB	Calibration	RMSE↓	JSD↓	SSIM↑	PCC↑	Rare-F1↑
Off	Off	.049925	.486622	.736683	.717186	.282021
Off	On	.046879	.488466	.767775	.749394	.286871
On	Off	.041245	**.442765**	.815061	.800477	.304041
On	On	**.038916**	.446869	**.833765**	**.820599**	**.321290**

D.3 Moffitt separates qNB regularization from post-hoc calibration

We use a matched 2×2 factorial to separate the qNB observation term from post-hoc hierarchical calibration fitted with the backbone fixed. All four experiments share the expression checkpoint, and within each qNB stratum, the calibration contrast shares the Stage II backbone. We first average three model seeds within each held-out animal and then average the three animal folds.

qNB improves RMSE, JSD, SSIM, and PCC with calibration either off or on. The paired effects have the same direction in all three folds, and the fold-level confidence intervals do not cross zero. Relative to the both-off condition, the both-on condition lowers RMSE by 22.0% and JSD by 8.2%, and raises SSIM and PCC by 0.0971 and 0.1034. With the same qNB-trained backbone, calibration gives the lowest RMSE and highest SSIM and PCC among the matched conditions. The uncalibrated endpoint gives the lowest full-simplex JSD (.442765), providing an alternative when compositional divergence is the primary evaluation criterion.

D.4 Supplementary stage-isolation and fraction-replay controls

This matched factorial compares stage isolation and labelled reference replay under a shared Moffitt checkpoint and update budget for controlled comparison of updates.

Table 18: **Moffitt Stage isolation × labelled reference replay.** Means cover three held-out animals with three model seeds per animal. This factorial table compares the update rules and does not define a single winning configuration for the full benchmark evaluation.

Isolation / replay	RMSE↓	JSD↓	SSIM↑	PCC↑	Rare-F1↑
On / off	.040894	.448838	.814872	.800024	.295975
On / on	.040868	.449182	.815098	.800272	.290932
Off / off	.040951	.450682	.815080	.800227	.302634
Off / on	.040888	.451195	.815229	.800385	.308401

D.5 OSM supervised fraction-readout training

The Moffitt audit in Section 3.2 perturbs an established reference-defined cell-type mapping. On OSM, we instead test whether source supervision can establish that mapping under a persistent logit bias, because the 33-gene panel makes supervised fraction-readout training itself informative.

Before supervised fraction-readout training, the auxiliary-pretrained encoder has no trained fraction readout, so dominant balanced accuracy sits near the 31-type chance level. Source supervision raises it to 0.697 and dominant-type accuracy on hard-pair pseudo-spots to 0.714. The biased condition reaches nearly the same endpoint. The injected bias switches the predicted dominant label for a subset of spots, and after training, 88.3% return to the unbiased-model prediction. Source supervision establishes reference-defined cell-type semantics under a 33-gene panel. This audit measures supervised fraction-readout training, and target adaptation begins afterward.

Table 19: **OSM supervised fraction-readout training on reserved reference pseudo-spots.** The protocol applies 180 source-supervised updates to E1, the auxiliary classifier, and the fraction readout while E2 remains frozen at $g = 0$. The biased condition applies a persistent logit bias to 30% of samples selected before source training begins in every split. Hard-pair dom. acc. measures 31-class dominant-type accuracy on pseudo-spots with two fractions above 10^{-7} and a largest fraction in $[0.55, 0.65]$. Pairwise ordering accuracy instead compares the two supported types.

Condition	Fraction-readout training	RMSE↓	JSD↓	Dom. BAcc↑	Hard-pair dom. acc.↑	Rare recall↑
Unbiased	Before	.12884	.56219	.0288	.0311	.0335
Unbiased	After	**.06333**	**.16437**	**.6971**	**.7139**	**.6836**
Biased	Before	.12844	.56146	.0300	.0316	.0340
Biased	After	**.06433**	**.16590**	**.6938**	**.7099**	**.6853**

Table 20: **OSM source retention without calibration.** Direct joint adaptation improves every source measure on this assay under the matched protocol setting.

Arm	RMSE↓	JSD↓	SSIM↑	PCC↑	Hard-pair↑
No Stage II update	.07962	.48097	.57077	.64452	.59295
Direct joint	.07664	.43982	.58802	.71387	.67128
Replay only	**.07346**	**.42213**	**.62067**	**.73563**	**.69250**
Spatial-only	.07538	.43269	.59052	.72429	.67751
Spatial-only + replay	.07358	.42252	.61993	.73485	**.69250**

D.6 Source retention before and after calibration

Each experimental condition fits its own calibrator with the same reference-only procedure and hyperparameters, so calibrated source metrics include a refitting effect. We therefore use the uncalibrated values as the primary source-retention evidence reported below.

The uncalibrated results preserve the target-metric ordering in Table 13, so calibration does not determine that ordering across the five Stage II conditions tested here.

D.7 OSM spot-composition permutation

A parallel permutation study on OSM reached 86.3% dominant-type recovery among permuted spots after Stage II target adaptation. This study used a different Stage I state from the factorial above.

D.8 seqFISH separates stable from conditional components

Rare AUPRC is mean precision-recall AUC over predeclared rare types in each split.

Table 21: **seqFISH latent, fusion, and readout ablations.** Matched comparisons across three splits and three seeds each contrast deterministic and sampled latent paths, h-only/additive/bilinear fusion, and a qNB-by-calibration readout factorial.

Block	Arm	RMSE↓	JSD↓	SSIM↑	PCC↑	Rare AUPRC↑
Latent	deterministic	.07507	.40035	.87905	.87782	.65134
Latent	sampled-h, KL=0	.07458	.39536	.88702	.88524	.64467
Latent	sampled-g, KL=0	.07470	.39872	.87899	.87788	**.65899**
Latent	sampled-$h + g$, KL=0	**.07444**	**.39306**	.88631	.88455	.64968
Fusion	h only	**.07430**	.39987	.88326	.88176	-
Fusion	$h + g$ additive	.07432	**.39462**	**.88750**	**.88575**	-
Fusion	$h + g + h \times g$	.07458	.39536	.88702	.88524	-
Readout	qNB off, cal. off	.07554	.39913	.87946	.87804	-
Readout	qNB on, cal. off	.07528	.39893	.88066	.87925	-
Readout	qNB off, cal. on	.07476	.39550	.88631	.88453	-
Readout	qNB on, cal. on	.07458	**.39536**	**.88702**	**.88524**	-

Table 22: **seqFISH composition-sampling and semantic-path ablations.** The composition block compares random aggregation with the structured 70/10/10/10 mixture under a matched model and training budget. The normal, zero, and shuffled m conditions yield small metric differences.

Block	Arm	RMSE↓	JSD↓	SSIM↑	PCC↑	Rare AUPRC↑	Active types↑
Pseudo	Random aggregation	**.07181**	.40797	**.88754**	**.88741**	.62245	16.67
Pseudo	Structured composition sampling	.07458	**.39536**	.88702	.88524	**.64467**	**17.56**
m audit	normal	.07458	.39515	.88711	.88534	N/A	N/A
m audit	shuffle	.07453	.39510	.88725	.88548	N/A	N/A
m audit	zero	.07454	.39521	.88725	.88548	N/A	N/A

Table 23: **seqFISH Stage II protection and replay.** Means ± fold SD are computed across three cell-disjoint splits after averaging three model seeds within each split. S0 omits Stage II; F freezes the reference-prediction path; J updates jointly; H uses 150 protected steps followed by 150 joint steps; and the +R suffix adds labelled reference replay during the joint phase. Source ΔMAE is the absolute prediction change from the shared source-supervised checkpoint. Endpoints use posterior means at $T = 1$ before calibration; target labels are used only for evaluation.

	Reference retention		Target accuracy			
Condition	RMSE↓	ΔMAE↓	RMSE↓	JSD↓	SSIM↑	PCC↑
S0	**.06200**±.00041	**.00000**±.00000	.11528±.00384	.58367±.01010	.58886±.03142	.64142±.02508
F	**.06200**±.00041	**.00000**±.00000	.06979±.00134	**.36552**±.00394	.90513±.00467	.90401±.00483
J	.08327±.00165	.02975±.00052	.06928±.00107	.37036±.00206	.90266±.00229	.90145±.00230
J+R	.06400±.00028	.01362±.00042	.06863±.00172	.37029±.00299	.90496±.00493	.90354±.00502
H	.06752±.00129	.01852±.00087	.06935±.00075	.37042±.00184	.90326±.00199	.90192±.00204
H+R	.06259±.00053	.01183±.00018	**.06828**±.00144	.36817±.00336	**.90678**±.00446	**.90532**±.00453

Takeaway. The matched ablations assign distinct roles to the retained components. Fraction replay provides the largest matched OSM effect. The Moffitt factorial shows that qNB improves all four primary metrics, whereas post-hoc calibration trades a small JSD increase for lower RMSE and higher SSIM and PCC. On seqFISH, sampled-h, additive g, and calibration improve their matched endpoints. Structured composition sampling improves JSD, Rare AUPRC, and active-type coverage at the cost of higher RMSE and lower SSIM and PCC than random aggregation. In the seqFISH Stage II mechanism protocol, spatial-only adaptation leaves source predictions unchanged and gives the lowest target JSD. Protected handoff with replay gives the best mean target RMSE, SSIM, and PCC and reduces source drift relative to handoff without replay. Joint adaptation shows the same retention benefit from replay, while the best target endpoint remains metric-dependent.

D.9 Cross-dataset pseudo-spot abundance supplements

The main text summarizes the three-dataset audit and presents the Moffitt visualization. Table 24 reports the strict three-fold distances, followed by the matched seqFISH partial-diffusion experiment.

Table 24: **Cross-dataset pseudo-spot abundance comparison.** We compare ordinary aggregation with random sampling. Lower W_1 and JSD indicate closer agreement with the target composition distribution under the same dataset-specific evaluation protocol.

Dataset	Rank W_1	Entropy W_1	Dominant W_1	Mass JSD
SILTA ordinary				
Moffitt	**0.0045**	**0.054**	**0.070**	**0.161**
OSM	**0.0076**	**0.073**	**0.086**	0.235
seqFISH	0.0374	0.283	0.372	**0.071**
Random sampling baseline				
Moffitt	0.0164	0.238	0.296	0.303
OSM	0.0143	0.142	0.185	**0.233**
seqFISH	**0.0085**	**0.029**	**0.058**	0.283

Table 25: **Complete Moffitt curriculum endpoints.** The matched comparison uses the same model settings and training budget. Values summarize three held-out-animal folds after averaging three model seeds within each fold before fold-level aggregation.

Pseudo strategy	RMSE↓	JSD↓	SSIM↑	PCC↑	Rare-F1↑
Ordinary	.04092	.45401	.81498	.80015	.30001
Structured curriculum	**.04073**	**.45285**	**.81636**	**.80161**	**.31071**

Table 26: **Diffusion-augmented configurations versus ordinary controls on seqFISH.** Values report mean ± sample SD across three split means, each averaging three seeds. Each metric pairs diffusion and ordinary runs at the same epoch and temperature, with positive differences indicating improvement. This matched 10% diffusion ablation uses settings distinct from the benchmark.

Metric	Partial-diffusion endpoint	Paired improvement
RMSE↓	0.067936 ± 0.001090	+0.001669
JSD↓	0.383946 ± 0.006037	+0.005300
SSIM↑	0.901852 ± 0.003461	+0.007202
PCC↑	0.901358 ± 0.003315	+0.006931

On OSM, ordinary aggregation lowers rank W_1 from 0.0143 to 0.0076, entropy W_1 from 0.142 to 0.073, and dominant-fraction W_1 from 0.185 to 0.086. Sample-level mass JSD remains comparable between the two strategies (0.235 versus 0.233 overall).

On seqFISH, ordinary aggregation spreads mass across more within-spot types than the target: its local rank, entropy, and dominant-fraction distances exceed the random baseline. The same pseudo-spots improve global type coverage, lowering mass JSD from 0.283 to 0.071. This result records the intended trade-off between local sharpness and coverage, while the matched curriculum and diffusion experiments test whether that support improves learning.

Matched partial diffusion on seqFISH. This experiment isolates expression diversification from composition sampling. Stage I starts with ordinary pseudo-spots. In a later Stage I segment, reference-trained diffusion replaces the expression realization of 10% of the pseudo-spots while preserving their composition labels. The ordinary control and partial-diffusion experiment use the same pseudo count, labels, model, optimizer, and training budget. Table 26 reports fold-first descriptive endpoints and paired effects against the matched ordinary path. We average three model seeds within each split before split-level aggregation shown below.

Partial diffusion improves the reported paired mean for all four metrics. Because the intervention changes only a bounded subset of expression realizations, the comparison isolates expression diversification from changes in labels, pseudo count, or optimization budget.

Figures 11 and 12 report the complete distributions for both datasets.

D.10 Auxiliary dominant-type representation audit

We assess whether auxiliary dominant-type supervision produces a representation that encodes cell-type composition, and whether the cooperative prediction head uses that representation directly. The audit consists of two parts: a linear probe that measures the information content of m, and counterfactual interventions that measure its causal contribution to the final prediction.

Linear probe. We freeze the trained SILTA model and fit a single linear layer $m \to$ softmax $\to$ fraction on reference validation pseudo-spots (6,400 per fold on Moffitt and 3 splits $\times$ 3 seeds on OSM). Table 27 reports the results under the shared linear-probe protocol.

The probe links dominant-type recognition to mixture estimation within the same auxiliary representation. Dominant balanced accuracy exceeds the corresponding chance level on both datasets, and the linear readout also estimates continuous cell-type proportions. Together with hard-pair ordering, these results show that m contains information about both cell-type identity and relative abundance in the evaluated reference pseudo-spots.



Figure 11: **seqFISH pseudo-spot abundance audit.** The v10 panels show the fixed composition embedding with high-density contours, marginal densities, and corner bridges, followed by rank-abundance profiles and four distances normalized to random. The distance bars summarize three splits after averaging three pseudo seeds within each split.

Table 27: **Auxiliary dominant-type representation probe.** We train a linear probe on m alone while SILTA remains frozen. Hard-pair ordering accuracy measures whether m correctly ranks the relative fractions of the two most confusable types in each spot. Random baseline for dominant balanced accuracy is $1/K$ for the corresponding cell-type vocabulary.

Metric	Moffitt (K=45)	OSM (K=31)
Hard-pair ordering accuracy	0.689 ± 0.053	0.696 ± 0.041
Dominant balanced accuracy	0.559 ± 0.043	0.428 ± 0.045
Fraction RMSE	0.056 ± 0.004	0.078 ± 0.005

Counterfactual interventions. We replace m with shuffled or zeroed vectors in the trained model's forward pass and measure the resulting change in predicted fractions (Table 28).

Table 28: **Auxiliary-representation counterfactual interventions.** Mean absolute proportion change and dominant-type switch rate under shuffle(m) and zero(m). OSM values are exactly zero because $\lambda_m = 0$ in the selected configuration used for target evaluation.

Intervention	Metric	Moffitt	OSM
shuffle(m)	Mean fraction change	6.4×10^{-5}	0
shuffle(m)	Dominant switch rate	0.38%	0%
zero(m)	Mean fraction change	1.5×10^{-4}	0
zero(m)	Dominant switch rate	0.83%	0%

On Moffitt, where $\lambda_m = 0.025$, m is a direct but small prediction path: the interventions change fewer than 1% of dominant-type assignments. The primary target contribution comes from h and g, while auxiliary dominant-type supervision also shapes E1 and h during Stage I. On OSM, $\lambda_m = 0$, so m makes no direct contribution to the selected target readout.

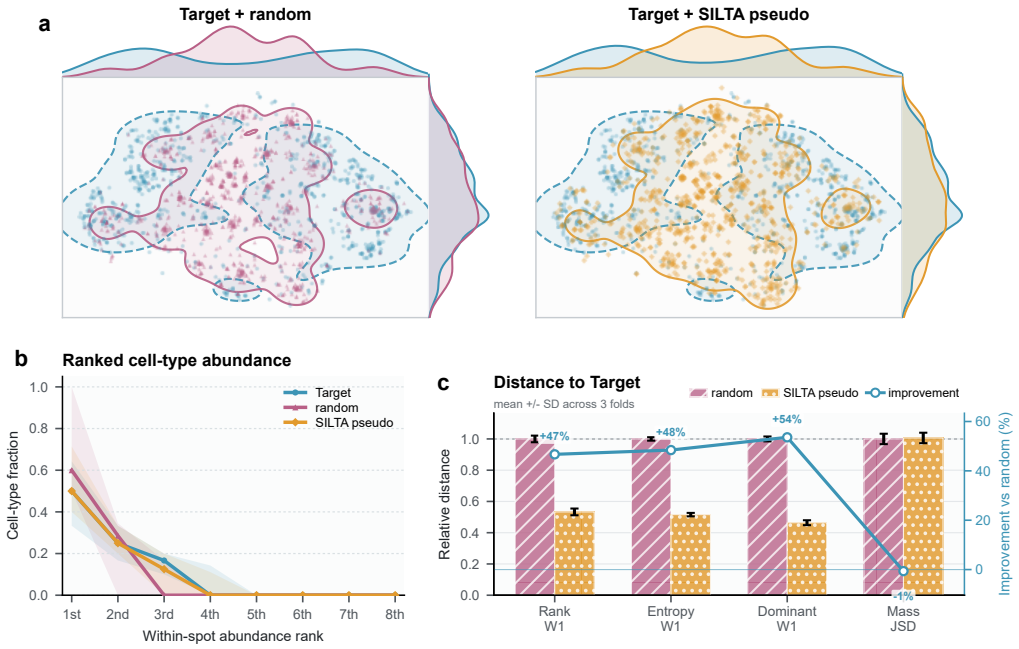


Figure 12: **OSM pseudo-spot abundance audit.** The same v10 protocol compares the joint and marginal composition views, rank-abundance profiles, and normalized composition distances across three cell-disjoint splits, each averaged over three pseudo seeds.

D.11 Stage II spatial correction: Moran's I process audit

For each fold and seed, Moran's I measures clustering of per-type signed errors before the first Stage II update and after adaptation; spot MAE measures error magnitude.

OSM results. Table 29 reports the OSM audit across three cell-disjoint splits at fixed model seed.

Table 29: **OSM Stage II Moran's I audit.** Signed-error Moran's I decreases across all three splits.

Split	Entry	Final	Change
17	0.068	0.058	−15%
42	0.069	0.056	−19%
89	0.059	0.048	−18%
Mean	0.065	0.054	−17%

All three splits show consistent reduction in signed-error spatial autocorrelation, supporting the conclusion that Stage II performs spatial correction beyond global error reduction.

Moffitt results. Table 30 reports the Moffitt audit across three cross-animal folds at the fixed algorithm seed used throughout this process audit and every reported checkpoint.

On Moffitt, Stage II reduces spot-level MAE by about 33% on average, with improvements in all three folds. However, the Moran's I response is fold-dependent: B-M3 shows strong reduction while B-M1 and B-M2 show increases. This matches the Moffitt schedule: 300 spatial-only steps and no low-rate joint refinement. Among the prediction-backbone branches, only the spatial branch trains while the expression branch stays frozen, whereas active qNB response parameters update separately. This schedule redistributes spatial error structure unevenly. MAE shows that Stage II improves fraction accuracy across all folds, regardless of the Moran's I direction.

Table 30: **Moffitt Stage II process audit.** Spot MAE improves across all folds, while the Moran's I response varies across the three folds under the shared schedule.

Fold	MAE (entry)	MAE (final)	Moran's I (entry)	Moran's I (final)
B-M1	0.0235	0.0159	0.223	0.288
B-M2	0.0247	0.0170	0.133	0.182
B-M3	0.0242	0.0156	0.419	0.304
Mean	0.0241	0.0162	0.258	0.258

D.12 Spatial intervention details

Table 31 reports the SILTA geometry interventions. Both interventions alter individual proportion vectors and worsen every mean metric on Moffitt and OSM.

Across the nine fold-seed pairs, coordinate shuffle degrades all four Moffitt metrics in 9/9 pairs. On OSM, the corresponding counts are 9/9 for RMSE, SSIM, and PCC and 8/9 for JSD. Edge removal degrades Moffitt RMSE/JSD in 9/9 pairs and SSIM/PCC in 8/9, while the OSM counts are 8/9 for RMSE/SSIM/PCC and 6/9 for JSD. The latent-path audit gives the same interpretation. On Moffitt, shuffle(g) changes the predicted proportions by .01592 and decreases PCC by .11126, whereas zero(g) changes them by .01399 and decreases PCC by .05028. Reassigning the spatial representation therefore produces a larger location-specific effect than removing it.

Training-time geometry controls. Matched Moffitt controls separate the contribution of the g path from its distance bias (Table 32). Each fold-seed pair shares a Stage I checkpoint and a 300-step spatial-only adaptation budget. The reference path and fraction head remain frozen. We compare correct geometry, shuffled training distances, E2 attention without distance bias, and a disabled g path. The shuffle permutes distance-matrix rows and columns together during training and restores correct distances for evaluation. The no-distance control retains E2 attention and its parameters. All four rows report the post-calibration output under the same evaluation protocol.

The g path with correct geometry lowers RMSE by 16.0% relative to the g-off control. Correct geometry also improves all four primary metrics over shuffled training distances in all three folds and all nine fold-seed runs. Removing distance bias leaves performance close to the correct-geometry condition: the mean gains from distance bias are .000016 in RMSE, .000671 in JSD, .000584 in SSIM, and .000497 in PCC. These controls support the g path's role in target adaptation, with a small incremental benefit from distance bias in this Moffitt setting.

Table 31: **SILTA spatial-geometry interventions.** Differences are intervention minus intact-model values. Positive RMSE/JSD and negative SSIM/PCC indicate degradation. Prediction L_1 and dominant-type flips measure the change from the intact prediction. Values average three folds or splits after averaging three model seeds within each unit before cross-unit aggregation.

Dataset	Intervention	Prediction L_1	Dominant flips	ΔRMSE	ΔJSD	ΔSSIM	ΔPCC
Moffitt	Coordinate shuffle	.1705	.1149	+.0013	+.0064	-.0104	-.0091
Moffitt	Edge removal	.1571	.1069	+.0009	+.0046	-.0062	-.0049
OSM	Coordinate shuffle	.1477	.0861	+.0009	+.0031	-.0085	-.0092
OSM	Edge removal	.1296	.0700	+.0003	+.0010	-.0033	-.0036

Table 32: **Moffitt training-time geometry controls.** Values are means ± fold SD, after averaging three model seeds within each of three cross-animal folds. The g-off control disables both spatial and interaction contributions to fusion.

Condition	RMSE↓	JSD↓	SSIM↑	PCC↑
Correct geometry	.03798 ± .00430	.44957 ± .02079	.82333 ± .06982	.81177 ± .07210
Shuffled training distances	.03886 ± .00437	.45261 ± .02169	.81844 ± .07128	.80570 ± .07391
No distance bias	.03800 ± .00432	.45024 ± .02087	.82274 ± .07044	.81127 ± .07261
g path off	.04523 ± .00286	.53835 ± .01807	.75276 ± .03491	.74737 ± .03100

D.13 Supplementary melanoma marker evidence

The appendix separates the per-marker SILTA audit for Patient 3 replicate 1 from the full Patient 4 replicate 1 T-cell comparison. Both use the same observed-marker normalization and fixed evaluation protocol and scoring implementation used in the main text.

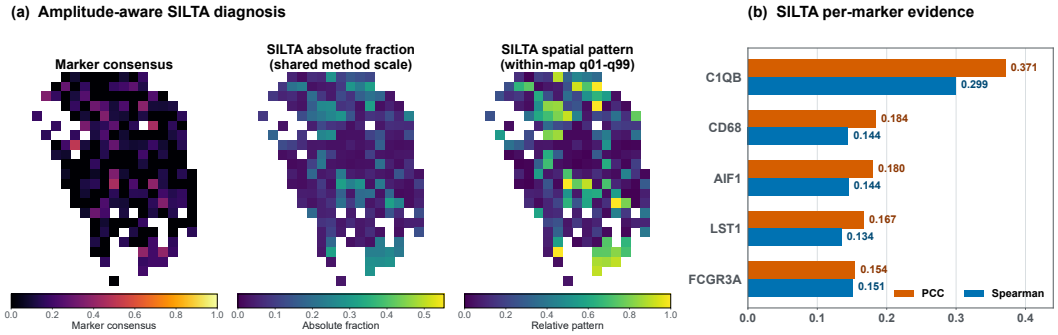


Figure 13: **SILTA macrophage marker audit on Patient 3 replicate 1.** (a) Marker consensus, the absolute predicted proportion map, and the within-map spatial pattern. (b) Per-marker PCC and Spearman values across all five markers. The marker-specific scores show whether the aggregate macrophage concordance is supported across the complete panel.

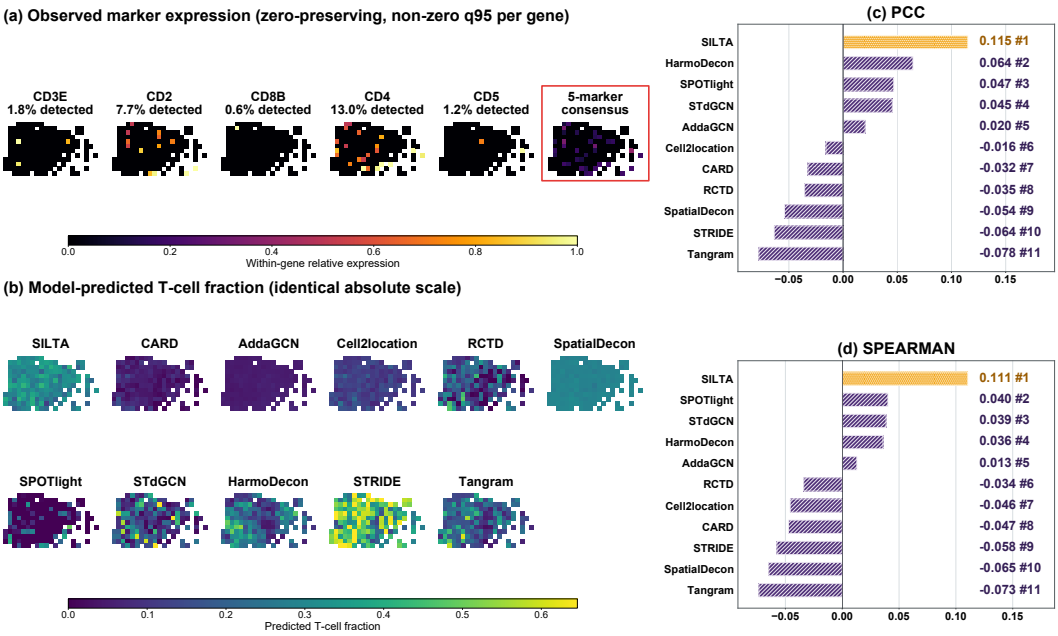


Figure 14: **T-cell spatial concordance on Patient 4 replicate 1.** The figure shows five observed T-cell marker maps, T-cell proportion maps from SILTA and ten baselines, and the corresponding PCC and Spearman rankings under the shared protocol. Independent rankings make the two concordance criteria directly comparable across methods in both displayed panels.

D.14 SILTA run-level results on known-composition benchmarks

Table 33 reports the four primary metrics for every SILTA fold-seed run on OSM, seqFISH, and Moffitt. Each dataset contains three folds or splits and three model seeds per fold under the shared evaluation protocol used for the benchmark means reported in Table 2.

Across runs, OSM varies narrowly across seeds within each split. Moffitt shows larger between-fold than within-fold variation, with B-M2 harder than B-M1 and B-M3. Within each fold, seed-to-seed differences remain smaller than this fold-level gap. These patterns support fold-first aggregation and avoid treating nested model seeds as independent biological replicates.

Table 33: **SILTA-only fold-seed results on OSM, seqFISH, and Moffitt.** Every fold–seed run has its own trained weights. Means average model seeds within each fold or split before averaging the three folds or splits under the fold-first rule used for every benchmark mean in Table 2.

Fold / model seed	RMSE↓	JSD↓	SSIM↑	PCC↑
OSM				
split 17 / seed 17	.065408	.467246	.801799	.795372
split 17 / seed 42	.063277	.466267	.813441	.808098
split 17 / seed 89	.064464	.467410	.806275	.800246
split 42 / seed 17	.062154	.456917	.821271	.815516
split 42 / seed 42	.063514	.458262	.807553	.802513
split 42 / seed 89	.064003	.454221	.807620	.801638
split 89 / seed 17	.064364	.454031	.798816	.793496
split 89 / seed 42	.066859	.452457	.789357	.781947
split 89 / seed 89	.064581	.451081	.801172	.794678
3 × 3 mean	**.064292**	**.458655**	**.805256**	**.799278**
seqFISH				
split 17 / seed 17	.067187	.371744	.904119	.903534
split 17 / seed 42	.066506	.369899	.906895	.906725
split 17 / seed 89	.067677	.367623	.904695	.903971
split 42 / seed 17	.067776	.360066	.905143	.904348
split 42 / seed 42	.067218	.367603	.902512	.901425
split 42 / seed 89	.068997	.366878	.898345	.897863
split 89 / seed 17	.069589	.362359	.892471	.892668
split 89 / seed 42	.068197	.364606	.897266	.897014
split 89 / seed 89	.067713	.358925	.905220	.904673
3 × 3 mean	**.067873**	**.365523**	**.901852**	**.901358**
Moffitt				
B-M1 / seed 17	.037012	.442021	.843824	.830372
B-M1 / seed 42	.036826	.443369	.843960	.830509
B-M1 / seed 89	.036561	.443254	.846527	.833277
B-M2 / seed 17	.041680	.444857	.806117	.790269
B-M2 / seed 42	.042035	.445842	.804764	.788746
B-M2 / seed 89	.042070	.446946	.802658	.786512
B-M3 / seed 17	.038099	.452149	.851112	.841074
B-M3 / seed 42	.037860	.452752	.852530	.842518
B-M3 / seed 89	.038101	.450628	.852390	.842117
3 × 3 mean	**.038916**	**.446869**	**.833765**	**.820599**

The run-level ranges make this distinction concrete. Across OSM, RMSE spans .062154-.066859, PCC .781947-.815516, and SSIM .789357-.821271. These intervals show that the fold-first means summarize all nine completed runs rather than a favorable algorithm seed. For Moffitt, B-M2 yields RMSE .041680-.042070 and PCC .786512-.790269, whereas B-M1 and B-M3 remain below .038101 RMSE and above .830372 PCC. The narrower within-fold ranges relative to the between-animal separation reinforce animal folds, rather than nested seeds, as the independent evaluation units while preserving fold difficulty in the benchmark aggregate.

D.15 seqFISH benchmark with known cell-type proportions

seqFISH (21 cell types, 351 genes, 668-676 target spots) provides a composition test with recognizable embryonic tissue domains. Across three cell-disjoint splits with three SILTA model seeds per split, the size of SILTA's mean advantage varies across the reported endpoints (Table 2). Figure 15 shows dominant-type and spot-level mixture patterns under a common type order.

The panels also separate two failure modes hidden by a single aggregate score: diffuse within-spot mixtures and spatially misplaced dominant types. Because the maps and rankings use the same predictions and type order, the qualitative and quantitative views support the same comparison across all reported methods and evaluation metrics.

Table 2 also shows that the strongest baseline depends on the metric. RCTD approaches SILTA in mean SSIM and PCC, whereas cell2location gives the next-lowest RMSE and RCTD the next-lowest JSD. This pattern supports a metric-specific reading of the benchmark: SILTA leads the reported means, but the size of its advantage varies across endpoints. Figure 15 complements these aggregate comparisons by showing the corresponding type assignments and within-spot mixtures.

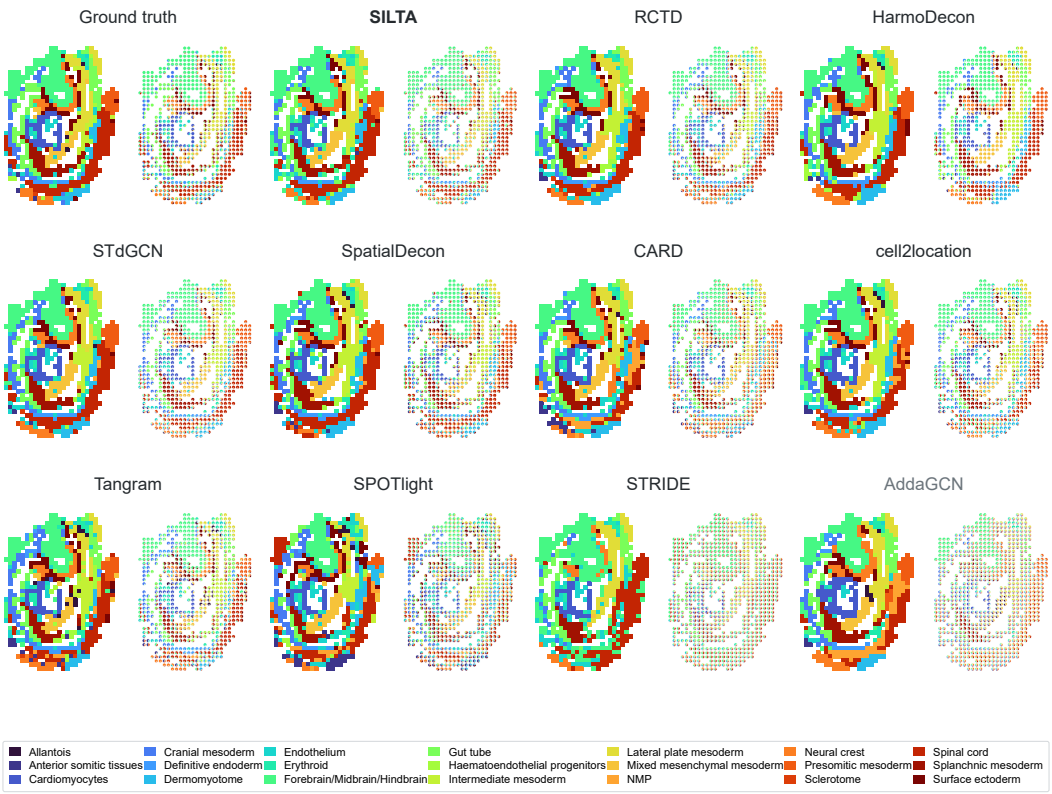


Figure 15: **seqFISH deconvolution.** Ground truth, SILTA, and ten baselines appear as dominant-type maps and spot-level proportion pies.

D.16 Moffitt spatial deconvolution

Figure 16 presents the complete cross-animal spatial comparison used in the Moffitt evaluation across held-out animals. The panel uses the same fold predictions and canonical type order as Table 2, allowing the cross-animal metrics to be read alongside the recovered spatial distribution of fine types. It also exposes diffuse or structurally flat baseline outputs without excluding them from the quantitative mean used in the primary ranking across all folds.

The Moffitt comparison combines two demands: transferring a reference across animals and allocating 45 cell types within the target tissue. The coordinate-shuffling and edge-removal controls in Table 31 complement this cross-animal evaluation. Under these matched interventions, coordinate shuffling causes a larger mean change in predicted proportions than edge removal, and both worsen the four mean metrics. These results connect the maps in Figure 16 to a test of how target geometry affects proportion estimates within the held-out target tissue.

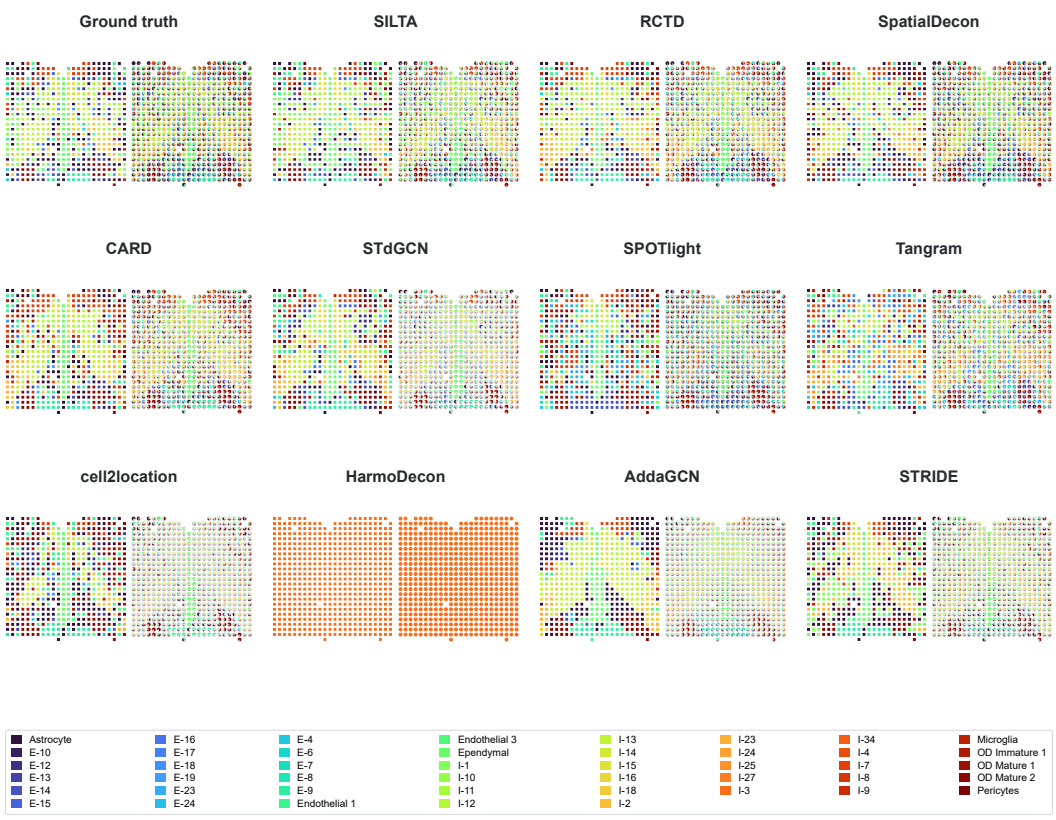


Figure 16: **Moffitt spatial deconvolution.** Ground truth, SILTA, and the evaluated baselines appear as dominant-type maps and spot-level proportion pies under cross-animal evaluation.

D.17 Moffitt sensitivity beyond the local plateaus

All sweeps in this section summarize three leave-one-animal-out Moffitt folds.

The training-schedule intervals are content steps 600-750, Stage II steps 300-420, and spatial-only steps 0-30. The regularization intervals are BATV weight 0.00025-0.001, diffusion fraction 0-0.10, and diffusion residual scale 0-0.10. The four primary metrics remain close to their selected values within these ranges, but Rare-F1 varies enough to prevent a general fine-type stability claim. The stability of the primary metrics supports the selected schedule as a representative operating point, whereas the Rare-F1 variation cautions against extending that stability claim uniformly to rare types.

The qNB weight sweep over 0.30-0.60 gives a monotonic dose response and tests observation-regularization strength within the evaluated Moffitt configuration. We treat this sweep as response evidence. Calibration strength over 0.75-1.25 exposes a separate trade-off: RMSE, SSIM, and PCC favor a different region from JSD and Rare-F1. Table 16 reports the effective-temperature scan separately. These axes diagnose distinct controls: qNB changes target-expression alignment strength, whereas calibration redistributes the final composition readout.

The matched factorial tests the effects of enabling qNB and calibration, while the strength sweeps measure local metric changes around the selected Moffitt configuration (Table 17).

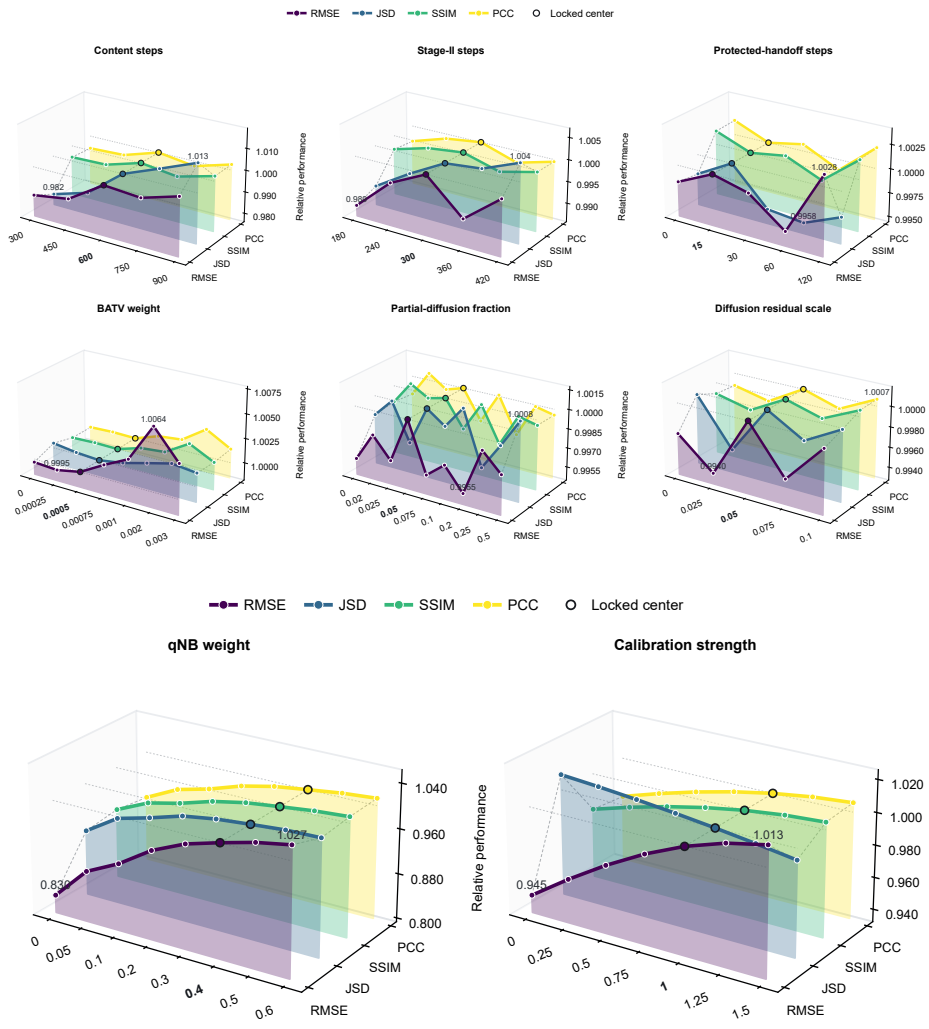


Figure 17: **Conditional and response-axis sensitivity on Moffitt.** (a) Conditional axes whose four primary metrics remain near the selected setting while Rare-F1 varies. (b) Effect of qNB weight and the calibration-strength trade-off. Both panels average three Moffitt benchmark folds.

E Related Work

E.1 Reference-Based Statistical Deconvolution

The dominant approach pairs an ST dataset with an annotated scRNA-seq reference and infers per-spot cell-type proportions through a statistical model. RCTD fits a Poisson model of spot counts as a mixture of reference cell-type profiles while correcting for platform effects (Cable et al., 2022), and cell2location places a hierarchical Bayesian prior over cell abundances and estimates them via variational inference (Kleshchevnikov et al., 2022). Regression formulations include SPOTlight, which uses seeded non-negative matrix factorization over cell-type signatures (Elosua-Bayes et al., 2021), and SpatialDWLS, which extends dampened weighted least squares to the spatial setting (Dong & Yuan, 2021). Tangram instead aligns single cells to spatial locations by learning a mapping that matches expression (Biancalani et al., 2021), while DestVI introduces a variational latent-variable model that resolves continuous cell-type variation (Lopez et al., 2022). Most relevant to the role of spatial context, CARD augments deconvolution with a conditional autoregressive prior that encourages neighboring spots to share similar compositions (Ma & Zhou, 2022). SILTA learns its reference

mapping before adapting target proportions. Stage I trains an expression-to-proportion mapping on structured pseudo-spots that cover sparse, near-pure, and confusable mixtures. Bounded diffusion varies their expression without changing cell-number proportion labels, exposing the reference mapping to expression variation before target adaptation. Stage II refines target proportions using target expression and tissue geometry through designated parameter updates, while reference replay supervises the reference mapping (Section 2.5).

E.2 Deep and Graph-Based Deconvolution

A second line of work learns the expression-to-composition mapping with neural networks, most often trained on pseudo-spots synthesized from the reference. DSTG constructs a link graph between real and pseudo spots and propagates composition labels with a graph convolutional network (Song & Su, 2021). STdGCN combines expression and spatial graphs within a GCN (Li & Luo, 2024), while Mo et al. couple graph contrastive learning with partial least-squares regression (Mo et al., 2025). STCGAN uses a cycle-consistent adversarial mapping between expression and composition (Wang et al., 2025a). HarmoDecon combines supervised pseudo-spots with expression and spatial graphs, using entropy and sample-level composition losses on pseudo-spots and adversarial domain alignment to address composition biases and platform differences (Wang et al., 2025b).

AddaGCN uses adversarial alignment of pseudo-spot and target representations in a joint graph (Ding et al., 2026). Its domain classifier promotes domain-invariant features while a shared GCN and source classifier predict proportions. SA2E learns spot representations and cell-type signatures through spatially regularized autoencoding without preset marker panels (Ma et al., 2026).

SILTA coordinates reference supervision and target adaptation through role-specific updates. Its expression path supplies reference-derived composition information, while its spatial path uses target tissue relationships to refine proportions. Target expression and geometry update assigned parameters; labelled replay supervises the reference mapping with $\chi = 0$ and E2 frozen. Spatial-only updates precede lower-rate joint refinement when both intervals run (Section 2.5).

E.3 Transfer, Augmentation, and Calibration Primitives

General transfer methods clarify one part of this separation. REFINE augments a frozen source representation with a trainable target residual encoder and fits a predictor on their combined representation (Xu et al., 2026). Its theory addresses supervised target prediction and includes a lymph-node anatomical domain classification example. SILTA operates with unlabeled target compositions, whereas REFINE's negative-transfer guarantee assumes a different target-supervision setting. All model selection in SILTA uses reference data only, and target fractions remain hidden.

Generative models provide another adjacent tool. DDPM pairs a diffusion probabilistic process with a denoising parameterization and a variational training objective (Ho et al., 2020), while scCube uses a variational autoencoder and spatial placement rules to simulate complete spatial transcriptomic datasets (Qian et al., 2024). SILTA uses a bounded diffusion residual for controlled expression augmentation of Stage I reference pseudo-spots. Cell state, capture, depth, and platform noise motivate this augmentation before target adaptation (Appendix A.4).

Cell2location uses negative-binomial count models for expression-based inference (Kleshchevnikov et al., 2022). During Stage II, SILTA uses qNB to compare the expression implied by predicted proportions with measured target expression. Fixed reference signatures and dispersions define this observation constraint across all target spots, while the quasi-likelihood accepts nonnegative continuous inputs. This supplies an expression-space constraint alongside role-specific updates and replay to guide proportion refinement in unlabeled target tissues (Appendix A.5).

Post-hoc calibration adjusts predictions after training. Guo et al. fit a single temperature to classification logits (Guo et al., 2017). After target adaptation, SILTA freezes its backbone and fits a calibrator on held-out reference pseudo-spots. Configured corrections can include group temperatures, group biases, and centered fine-type biases, with an optional mass penalty on mean predicted proportions. Thus, qNB constrains target predictions during adaptation, while reference-fitted calibration adjusts the final readout with the backbone fixed (Appendix A.5).