PREDICTING SPATIAL GENE EXPRESSION FROM HISTOLOGY WITH A SINGLE-CELL REFERENCE AND ORTHOGONAL RESIDUALS

Anonymous authors
Paper under double-blind review

ABSTRACT

H&E slides are routinely available, whereas spatial transcriptomics (ST) is costly and limited in scale. This has motivated models that predict spatial gene expression from histology. Most existing methods, however, predict spot-level expression directly from image features without modeling the cell types that contribute to each spot. We introduce ScComp, which predicts composition weights over a fixed single-cell reference and combines the reference profiles accordingly. A residual captures expression variation not represented by the reference and is restricted to the orthogonal complement of the reference space. Across three patient-disjoint ST benchmarks, ScComp achieves the best equal-dataset macro performance among the compared methods with only 0.87M trainable parameters. More importantly, the composition weights can be applied beyond the genes used during training. They predict approximately 11,000 reference-covered genes never used as training targets, at both the gene and pathway levels, and agree with cell-type proportions measured by Xenium without composition supervision. Ablation results further show that the orthogonal projection helps preserve this zero-shot behavior by keeping the residual separate from variation represented by the reference.

1 INTRODUCTION

Cell morphology and gene expression provide complementary information about tissue state and disease. Characterizing their relationship can help connect morphological patterns with underlying molecular variation (Haghighi et al., 2022; Chandrasekaran et al., 2021; Lim et al., 2023). Traditionally, tissue morphology has been examined using hematoxylin-and-eosin (H&E) histology, while molecular changes have been studied separately using gene-expression measurements such as RNA sequencing. However, linking these two types of information within the same tissue has been difficult.

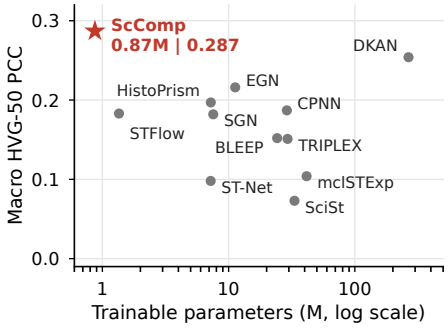

Figure 1: **Accuracy–parameter trade-off.** ScComp attains the highest macro HVG-50 PCC with 0.87M trainable parameters.

Spatial transcriptomics (ST) measures gene expression while preserving its spatial location within tissue (Ståhl et al., 2016), allowing molecular patterns to be examined together with tissue morphology. Histology and ST thus provide complementary views of the same tissue: H&E reveals cell morphology and tissue organization, whereas ST reveals spatial patterns of gene expression. However, ST is expensive and difficult to collect at scale, while H&E slides are routinely acquired and widely available in pathology archives. Because molecular variation is often associated with tissue morphology, this difference in availability raises a natural question: *Can spatial gene expression be predicted directly from histology?* This problem is particularly challenging in spot-based ST, where each measurement contains multiple nearby cells and reflects their combined gene expression.

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.

Most histology-to-ST methods directly predict a predefined set of genes from image features (He et al., 2020; Yang et al., 2023; Chung et al., 2024; Huang et al., 2025; Xie et al., 2023; Min et al., 2024; Zhang et al., 2026; Hu et al., 2026). Although their architectures differ, these methods generally do not explicitly model the cellular composition of each spot, making their predictions difficult to interpret in terms of cell types or states. Moreover, because their prediction targets are defined by a fixed gene panel during training, they cannot directly predict genes outside that panel.

Single-cell transcriptomics has provided gene-expression profiles for many cell types and states (Regev et al., 2017; Chan Zuckerberg Initiative, 2025). These profiles are widely used as references to estimate cellular composition in spot-based ST. Recent histology-to-ST methods have also incorporated single-cell references into gene-expression prediction (Li et al., 2025; Nishimura et al., 2026). However, existing methods either update the reference during training or combine reference-derived information with learned image features. In these settings, the predicted expression is not constrained to remain aligned with the original cell-type reference.

We introduce **ScComp** (**Single-cell Composition**), a histology-to-ST model that predicts composition weights over a fixed single-cell reference matrix. The weighted combination of the reference profiles forms the *reference component* of the predicted expression. Keeping the reference fixed preserves the association between each weight and its corresponding cell type or state. Because the reference cannot explain all expression variation in tissue, ScComp additionally predicts a *reference-orthogonal residual*, constrained to the orthogonal complement of the space spanned by the reference profiles. Together, the two components capture reference-explainable and reference-orthogonal expression variation.

This decomposition gives the composition weights a role beyond predicting the genes used for training. Because the weights are defined over fixed reference profiles rather than individual target genes, they can be combined with reference expression for genes that were never used as prediction targets. Across three patient-disjoint benchmarks, ScComp achieves the best equal-dataset macro performance among 11 baselines with only 0.87M trainable parameters. The learned weights further enable prediction of approximately 11,000 additional reference-covered genes and their pathway activities, and correlate with cell-type proportions measured by Xenium despite receiving no composition supervision (Janesick et al., 2023). Ablation results additionally show that the reference-orthogonal residual helps preserve both the reference component and zero-shot prediction performance.

Our contributions are summarized as follows:

- We propose **ScComp**, which decomposes histology-based gene-expression prediction into a reference component derived from composition weights over a fixed single-cell reference matrix and a reference-orthogonal residual.
- We demonstrate that the learned composition weights can be reused beyond the training gene panel to predict reference-covered genes that were not used during training, as well as their pathway activities.
- We validate the biological meaning of the learned composition weights by comparing them with cell-type proportions measured by Xenium, without using cell-composition labels for training.

2 Related Work

2.1 Histology-to-ST Prediction

Histology-to-ST prediction has been studied using several types of image and spatial representations. ST-Net (He et al., 2020) predicts spatial gene expression from local H&E patches, and several methods incorporate neighboring spots through spatial or graph-based modeling (Zeng et al., 2022; Jia et al., 2024), retrieve morphologically similar tissue regions (Yang et al., 2023), or use multi-scale image features (Chung et al., 2024). Other methods align image and expression representations with contrastive learning (Xie et al., 2023; Min et al., 2024), model gene-specific structure (Zhang et al., 2026; Hu et al., 2026), or use generative objectives for spatial expression prediction (Huang et al., 2025). Despite these differences, most methods directly map image representations to a predefined set of target genes. They therefore do not provide a cell-type-based decomposition of the predicted expression, and their outputs are generally tied to the gene panel used during training.

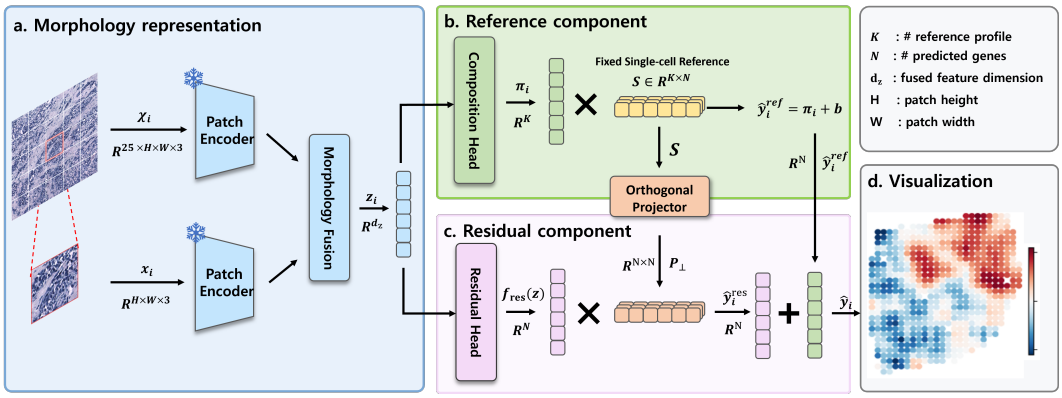

Figure 2: ScComp architecture. A frozen UNI encoder extracts features from the target spot and its local 5×5 neighborhood, which are fused into a morphology representation. Composition weights predicted from this representation are combined with a fixed single-cell reference matrix to form the reference component. A reference-orthogonal residual captures additional expression variation outside the space spanned by the reference profiles. The two components are combined to obtain the final gene-expression prediction.

2.2 Single-Cell Reference-Informed Prediction

Single-cell references have also been used to incorporate cell-type information into histology-to-ST prediction. SciSt (Li et al., 2025) estimates cell-type counts from histology and combines a single-cell-based expression estimate with learned image features. CPNN (Nishimura et al., 2026) predicts mixture weights over cell-type expression profiles derived from scRNA-seq and updates these profiles during training. These approaches use single-cell information in the prediction process, but they do not retain the original reference profiles as a fixed representation throughout training. ScComp instead uses a fixed single-cell reference matrix to form the reference component and models the remaining variation with a reference-orthogonal residual.

2.3 Prediction beyond the Training Gene Panel

Most histology-to-ST methods are trained with a fixed set of target genes and are not designed to generalize to genes outside the training panel. SGN (Yang et al., 2024) represents genes using functional and phenotypic descriptions and conditions expression prediction on these representations. ALIGN-ST (Zhou et al., 2025) models image–gene alignment together with relationships among gene descriptions. These methods enable prediction of unseen genes through learned gene representations. ScComp uses a different formulation based on composition weights over a fixed single-cell reference matrix. The same composition weights can be applied to other genes covered by the reference, without learning gene-specific representations for those genes.

3 Method

Method overview. ScComp predicts the gene-expression profile of an ST spot from its pathology image patch and surrounding tissue. Before training, we construct a fixed single-cell reference matrix from public scRNA-seq data, with each row corresponding to an expression profile associated with a cell type or state. Given the target patch and its local neighborhood, ScComp predicts composition weights over these reference profiles. Their weighted combination forms the reference component of the predicted expression. ScComp also predicts a reference-orthogonal residual, which is constrained to the orthogonal complement of the space spanned by the reference profiles. The reference component and residual are combined to produce the final gene-expression profile. Figure 2 summarizes the overall framework.

Problem setup. Given paired histology and gene-expression data, for each spatial spot i, let $\mathcal{X}_i = \{x_{i,p}\}_{p=1}^{25}$ denote the 5×5 local histology neighborhood centered at spot i, with $x_i \in \mathcal{X}_i$ denoting the target patch. Let $y_i \in \mathbb{R}^{1\times N}$ denote the corresponding log-normalized expression profile over N genes. We use a fixed single-cell reference matrix $S \in \mathbb{R}^{K\times N}$, where each row is a single-cell reference profile associated with a cell type or state and K is the number of reference profiles. We learn a model F_θ that predicts the spot-level gene-expression profile from the local histology and the single-cell reference:

$$\hat{y}_i = F_\theta(\mathcal{X}_i, S), \tag{1}$$

where $\hat{y}_i \in \mathbb{R}^{1\times N}$ is the predicted expression profile and S remains fixed during training.

Single-cell reference construction. We construct the fixed single-cell reference matrix S from single-cell RNA-seq data in the CZ CELLxGENE Census (CZI Cell Science Program et al., 2025). We use breast tissue data for HER2ST and STNet and skin tissue data for cSCC. For each cell, raw counts are normalized by the total counts across all measured genes and log-transformed as

$$\bar{C}_{i,j} = \log\left(1 + 10^4 \frac{C_{i,j}}{\sum_{j'} C_{i,j'}}\right), \tag{2}$$

where $C_{i,j}$ denotes the raw count of gene j in cell i. Because the number of cells differs across donors, we first average the normalized expression within each donor and then average the donor-level profiles with equal weight. For reference profile k, let $\mathcal{D}_k$ denote the set of donors and $\mathcal{I}_{k,d}$ the cells from donor d. The reference matrix is constructed as

$$S_{k,j} = \frac{1}{|\mathcal{D}_k|} \sum_{d\in\mathcal{D}_k} \frac{1}{|\mathcal{I}_{k,d}|} \sum_{i\in\mathcal{I}_{k,d}} \bar{C}_{i,j}. \tag{3}$$

For the experiments in this study, we retain normal cell types and selected cancer-derived epithelial or malignant cell states and select the N genes used for prediction, resulting in $S \in \mathbb{R}^{K\times N}$. Detailed criteria for cell selection and sampling are provided in Appendix C. Once constructed, S remains fixed throughout model training.

Morphology representation. For each target spot i, we use a 5×5 local neighborhood of 25 H&E patches centered on the target patch x_i. A shared frozen UNI encoder (Chen et al., 2024) extracts an image feature from each patch, $e_{i,p} = \mathrm{UNI}(x_{i,p})$ for $p = 1, \ldots, 25$, with e_i denoting the feature of the central target patch. To summarize the local tissue context while retaining the target feature separately, we aggregate the 25 patch features using attention pooling:

$$\bar{e}_i = \sum_{p=1}^{25} \omega_{i,p} e_{i,p}, \qquad \sum_{p=1}^{25} \omega_{i,p} = 1, \tag{4}$$

where the attention weights $\omega_{i,p}$ are learned from the patch features. The target feature and pooled neighborhood feature are normalized, concatenated, and projected through a two-layer MLP:

$$z_i = \mathrm{MLP}_{\mathrm{fuse}}\left([\mathrm{LN}(e_i); \mathrm{LN}(\bar{e}_i)]\right) \in \mathbb{R}^{d_z}. \tag{5}$$

The representation z_i serves as the morphology representation of spot i, combining the morphology of the target spot with its local tissue context. It is used to predict both the reference component and the reference-orthogonal residual.

Reference–residual based gene expression prediction. ScComp predicts gene expression by separating it into a reference component, represented by a weighted combination of the fixed single-cell reference profiles, and a reference-orthogonal residual component, which captures variation outside the reference space. This design keeps the composition weights tied to the reference profiles while allowing additional expression variation to be modeled separately.

Reference component. The morphology representation z_i is used to predict composition weights over the K reference profiles:

$$\pi_i = \mathrm{softmax}\left(f_{\mathrm{ref}}(z_i)\right) \in \mathbb{R}^K, \tag{6}$$

where the entries of π_i are nonnegative and sum to one. The reference component is given by $\pi_i S$, the weighted combination of the fixed single-cell reference profiles. We then add a trainable gene-wise bias $b \in \mathbb{R}^N$ to obtain the reference-based prediction:

$$\hat{y}_i^{\text{ref}} = \pi_i S + b. \tag{7}$$

Reference-orthogonal residual component. ScComp models expression variation outside the space spanned by the fixed single-cell reference profiles. To separate this variation from the reference component, we constrain the residual to the orthogonal complement of the reference space. We first compute the singular value decomposition of S and retain the right singular vectors corresponding to its numerical rank r:

$$S = U\Sigma V^\top, \qquad Q = V_{:,1:r} \in \mathbb{R}^{N\times r}, \tag{8}$$

where the columns of Q form an orthonormal basis for the row space of S. The projector onto the orthogonal complement of this space is

$$P_\perp = I_N - QQ^\top. \tag{9}$$

The reference-orthogonal residual is predicted as

$$\hat{y}_i^{\text{res}} = f_{\text{res}}(z_i)P_\perp, \tag{10}$$

where f_{res} is a two-layer MLP. By construction, $\hat{y}_i^{\text{res}}Q = 0$, making the residual orthogonal to every reference profile and therefore to the reference component $\pi_i S$. The reference component and residual thus model expression variation within and outside the reference space, respectively. The gene-wise bias b is not constrained by this projection.

The numerical rank r is determined from the singular values of S. The residual predictor uses LayerNorm, GELU, and dropout, with its final layer initialized near zero to keep the residual contribution small at the beginning of training.

Final prediction. The reference-based prediction and the reference-orthogonal residual are combined to produce the final gene-expression prediction:

$$\hat{y}_i = \hat{y}_i^{\text{ref}} + \hat{y}_i^{\text{res}}. \tag{11}$$

Training objective. For a training batch of B spots, let $Y, \hat{Y} \in \mathbb{R}^{B\times N}$ denote the target and predicted gene-expression profiles, respectively. The training objective combines a Huber loss for point-wise prediction error with a correlation loss based on the Pearson correlation across spots for each gene. The Huber loss is defined as

$$\mathcal{L}_{\text{Huber}} = \frac{1}{BN}\sum_{i=1}^{B}\sum_{j=1}^{N} h_\delta(\hat{Y}_{i,j} - Y_{i,j}), \tag{12}$$

where h_δ denotes the Huber loss with threshold δ. The correlation loss is defined as

$$\mathcal{L}_{\text{PCC}} = 1 - \frac{1}{N}\sum_{j=1}^{N} \text{PCC}\left(\hat{Y}_{:,j}, Y_{:,j}\right). \tag{13}$$

The final objective is

$$\mathcal{L} = \mathcal{L}_{\text{Huber}} + \lambda_{\text{PCC}}\mathcal{L}_{\text{PCC}}, \tag{14}$$

where λ_{PCC} controls the relative contribution of the correlation loss.

Patient-balanced optimization. Training is performed with one tissue section per batch. Because patients may contribute different numbers of sections, standard batch-wise optimization can give greater influence to patients represented by more sections. To balance optimization at the patient level, we use a patient-balanced local-update averaging procedure after the initial AdamW epochs. Let $\mathcal{P}$ denote the training patients and θ_t the model parameters at outer iteration t. Each patient starts from the same model parameters and performs L local gradient steps:

$$\theta_{t,p}^{(0)} = \theta_t, \qquad \theta_{t,p}^{(k+1)} = \theta_{t,p}^{(k)} - \eta_{\text{in}}\nabla\mathcal{L}_p\left(\theta_{t,p}^{(k)}\right). \tag{15}$$

The patient-specific displacement $\Delta_{t,p} = \theta_{t,p}^{(L)} - \theta_t$ is then averaged across patients:

$$\theta_{t+1} = \theta_t + \frac{\epsilon_t}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \Delta_{t,p}. \tag{16}$$

All patients therefore contribute equally weighted displacements computed from the same initialization, regardless of the number of sections they provide. Since $L > 1$, this procedure is not equivalent to taking a single gradient step on the patient-averaged loss.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. We evaluate ScComp on three public spatial transcriptomics datasets: HER2ST (Andersson et al., 2021), cSCC (Ji et al., 2020), and STNet (He et al., 2020). HER2ST contains 36 breast cancer tissue sections from 8 patients, cSCC contains 12 cutaneous squamous cell carcinoma sections from 4 patients, and STNet contains 68 breast cancer tissue sections from 23 patients. Following the evaluation setting of DKAN (Zhang et al., 2026), all datasets use smoothed $\log(1 + \mathrm{CP10k})$ expression values over a common panel of 250 genes.

Evaluation protocol and metrics. Following the evaluation protocol of DKAN (Zhang et al., 2026), we use patient-disjoint splits for all datasets: HER2ST and cSCC use leave-one-patient-out cross-validation, while STNet uses eight patient-disjoint folds, giving 20 held-out folds in total. We repeat each experiment with three random seeds (2021, 2022, and 2023) using the same folds. We consider two evaluation settings: in-panel prediction and zero-shot gene prediction. In-panel prediction uses the 250 genes included as prediction targets during training. For zero-shot prediction, held-out genes are excluded from training, checkpoint selection, and fine-tuning, and gene subsets are defined within each fold using training slides only. Under this setting, we compare ScComp with SGN (Yang et al., 2024), BLEEP (Xie et al., 2023), and mclSTExp (Min et al., 2024) using the same held-out genes.

For in-panel prediction, we report mean absolute error (MAE), mean squared error (MSE), and gene-wise Pearson correlation coefficient (PCC). PCC is evaluated over all 250 genes (ALL), the 50 highly expressed genes (HEG-50), and the 50 highly variable genes (HVG-50) used in the DKAN evaluation. Metrics are computed for each tissue section and averaged within each fold and then across folds for each dataset. We also report equal-dataset macro averages, giving equal weight to the three datasets. Statistical comparisons use an exact two-sided Wilcoxon signed-rank test over the 20 seed-averaged folds, with Holm correction across baselines. For zero-shot prediction, we report per-gene PCC on the HVG-1000 subset of the held-out genes. At the pathway level, we compute background-centered activity scores for the 50 Hallmark(Liberzon et al., 2015) gene sets over the full held-out gene set and report spot-wise PCC between predicted and observed activities. Detailed metric definitions and gene-set construction are provided in Appendix D, and zero-shot gene selection and pathway scoring are described in Appendix F.

Implementation details. We use 224×224 image patches and a 256-dimensional morphology representation. Training consists of 60 epochs of AdamW followed by 60 outer iterations of patient-balanced local-update averaging; all trainable modules are trained jointly with the Huber and PCC objective throughout. Full details are in Appendix B.

Baselines. We compare ScComp with representative histology-to-spatial transcriptomics methods from three methodological groups. Morphology- and context-based methods include ST-Net (He et al., 2020), EGN (Yang et al., 2023), TRIPLEX (Chung et al., 2024), STFlow (Huang et al., 2025), and HistoPrism (Hu et al., 2026). Multimodal alignment methods include BLEEP (Xie et al., 2023), mclSTExp (Min et al., 2024), SGN (Yang et al., 2024), and DKAN (Zhang et al., 2026). Reference-informed methods include SciSt (Li et al., 2025) and CPNN (Nishimura et al., 2026). All baselines use the same gene panels, splits, and targets, with their original architectures; DKAN uses the same frozen UNI encoder as ScComp, giving an encoder-matched comparison.

Dataset	Metric	ScComp	DKAN	STFlow	HistoPrism	ST-Net	EGN	TRIPLEX	BLEEP	mclSTExp	SGN	SciSt	CPNN
HER2ST	MAE↓	.379	.371	**.363**	.389	.403	.383	.417	.391	.387	.399	.411	.537
	MSE↓	.242	.242	**.224**	.259	.261	.245	.312	.266	.258	.268	.288	.474
	PCC(ALL)↑	**.303**	.301	.249	.247	.181	.274	.211	.178	.170	.227	.130	.246
	PCC(HEG)↑	.287	**.298**	.214	.214	.120	.248	.186	.150	.147	.200	.097	.206
	PCC(HVG)↑	**.297**	.291	.196	.211	.107	.255	.163	.132	.124	.209	.081	.194
cSCC	MAE↓	**.376**	.402	.402	.419	.446	.441	.465	.427	.418	.438	.468	.650
	MSE↓	**.231**	.266	.255	.278	.300	.312	.356	.293	.277	.307	.336	.666
	PCC(ALL)↑	**.358**	.350	.282	.281	.129	.273	.226	.270	.157	.210	.114	.282
	PCC(HEG)↑	**.329**	.265	.227	.197	.046	.191	.137	.204	.078	.125	.039	.170
	PCC(HVG)↑	**.307**	.268	.201	.180	.075	.194	.150	.215	.093	.144	.067	.184
STNet	MAE↓	**.317**	.335	.345	.326	.322	.328	.376	.357	.335	.337	.336	.492
	MSE↓	**.168**	.194	.203	.182	.175	.180	.264	.221	.193	.188	.192	.401
	PCC(ALL)↑	**.211**	.173	.143	.175	.118	.175	.131	.103	.097	.162	.071	.175
	PCC(HEG)↑	**.190**	.170	.112	.150	.072	.152	.092	.086	.071	.143	.040	.150
	PCC(HVG)↑	**.257**	.204	.153	.199	.113	.200	.140	.109	.096	.191	.070	.182
Macro Avg.	MAE↓	**.357**	.369	.370	.378	.390	.384	.419	.392	.380	.391	.405	.560
	MSE↓	**.214**	.234	.227	.240	.245	.246	.311	.260	.243	.254	.272	.514
	PCC(ALL)↑	**.291**	.275	.225	.234	.143	.241	.189	.184	.141	.200	.105	.234
	PCC(HEG)↑	**.269**	.244	.184	.187	.079	.197	.138	.147	.099	.156	.059	.175
	PCC(HVG)↑	**.287**	.254	.183	.197	.098	.216	.151	.152	.104	.181	.073	.187

Table 1: In-panel gene-expression prediction on HER2ST, cSCC, and STNet. Results are averaged over three seeds. Best results are shown in bold and second-best results are underlined. Macro Avg. denotes the equal-dataset average across the three datasets.

4.2 Comparison with Existing Methods

Table 1 compares ScComp with 11 baseline methods. ScComp achieves the best equal-dataset macro average across all five evaluation metrics. For HVG-50 PCC, ScComp achieves 0.297 on HER2ST, 0.307 on cSCC, and 0.257 on STNet, compared with 0.291, 0.268, and 0.204 for DKAN, respectively. Across the 20 patient-disjoint folds, ScComp achieves higher HVG-50 PCC than all baselines with available paired results, with significant differences after Holm correction ($p_{\mathrm{adjusted}} < 0.05$ for all comparisons).

The difference between ScComp and DKAN in HVG-50 PCC is 0.006 on HER2ST, 0.039 on cSCC, and 0.053 on STNet; the difference on HER2ST alone is not significant. STFlow achieves the lowest MAE and MSE on HER2ST, whereas its HVG-50 PCC is 0.196 compared with 0.297 for ScComp. This result reflects the difference between point-wise expression error and correlation of spatial expression patterns. Qualitative comparisons are provided in Appendix I.

ScComp uses 0.87M trainable parameters on top of the frozen encoder, the fewest among the compared methods (Fig. 1). The effects of local neighborhood context, patient-balanced optimization, and the reference–residual decomposition are examined in Sec. 4.5.

4.3 Zero-Shot Gene Prediction

We evaluate whether the learned composition weights can be applied to genes outside the 250-gene training panel. For each dataset, we collect the reference-covered genes outside the training panel (10,978 for HER2ST, 12,122 for cSCC, and 11,049 for STNet) and construct the corresponding reference matrix $S^{\mathrm{zs}} \in \mathbb{R}^{K \times G_{\mathrm{zs}}}$. Using the same composition weights π_i predicted from histology, expression for these genes is obtained as

$$\hat{y}_i^{\mathrm{zs}} = \pi_i S^{\mathrm{zs}}. \tag{17}$$

The held-out genes are not used for training, checkpoint selection, or fine-tuning, and zero-shot prediction uses no additional parameters or residual component. Table 2 reports zero-shot prediction at the gene and pathway levels. On the HVG-1000 subset, ScComp achieves a mean gene-wise PCC of 0.187, compared with 0.157 for BLEEP and 0.167 for SGN. The difference from BLEEP is significant after correction ($p_{\mathrm{adjusted}} = 0.011$), whereas the difference from SGN is not significant.

Method	HER2ST		cSCC		STNet		Macro	
	Gene	Pathway	Gene	Pathway	Gene	Pathway	Gene	Pathway
ScComp (πS^{zs})	**.209**	**.263**	.220	**.194**	**.132**	**.169**	**.187**	**.209**
ScComp (w/o projection)	.192	.259	.219	.171	.113	.154	.175	.195
ScComp (permuted π)	.002	-.001	.000	.000	.000	.000	.001	.000
SGN	.197	.154	.201	.014	.102	.016	.167	.061
BLEEP	.149	.172	**.237**	.154	.086	.128	.157	.151
mclSTExp	.111	.109	.189	.127	.069	.103	.123	.113

Table 2: Zero-shot gene and pathway prediction. Gene denotes gene-wise PCC on the HVG-1000 subset of held-out genes, and Pathway denotes spot-wise PCC of Hallmark pathway activities computed over the full held-out gene set. Results are averaged over three seeds; the best result in each column is shown in bold.

At the pathway level, ScComp achieves a PCC of 0.209, compared with 0.151 for BLEEP, 0.113 for mclSTExp, and 0.061 for SGN, with significant differences in all paired comparisons ($p_{\text{adjusted}} \leq 0.001$). Permuting the correspondence between the entries of π_i and the rows of the reference matrix reduces the gene-wise PCC from 0.187 to 0.001, indicating that zero-shot prediction depends on the correspondence between the composition weights and the reference profiles.

We next examine the contribution of the reference-orthogonal constraint. Removing the orthogonal projection reduces gene-wise zero-shot PCC from 0.187 to 0.175 and pathway-level PCC from 0.209 to 0.195 ($p_{\text{adjusted}} = 0.0001$ and 0.0005, respectively; Table 2). In-panel HVG-50 PCC also decreases from 0.287 to 0.278 ($p_{\text{adjusted}} = 0.009$; Table 3). A random rank-matched projection does not produce the same improvement (Appendix 9). These results indicate that constraining the residual outside the reference space helps maintain the composition weights used for prediction beyond the training gene panel.

4.4 What the Composition Coordinate Means

We evaluate whether the predicted weights π reflect the cell types represented in the fixed reference. Using 12 Xenium Prime breast samples, we derive the proportions of 13 reference cell types within $100\,\mu\text{m}$ tiles and assess their association with π predicted from the registered H&E images. As a supervised reference, we train a ridge regressor on the same image features using the measured cell-type proportions as targets. We also include a random-reference baseline. We evaluate all models using four-fold cross-validation, with donors held out from training in each fold.

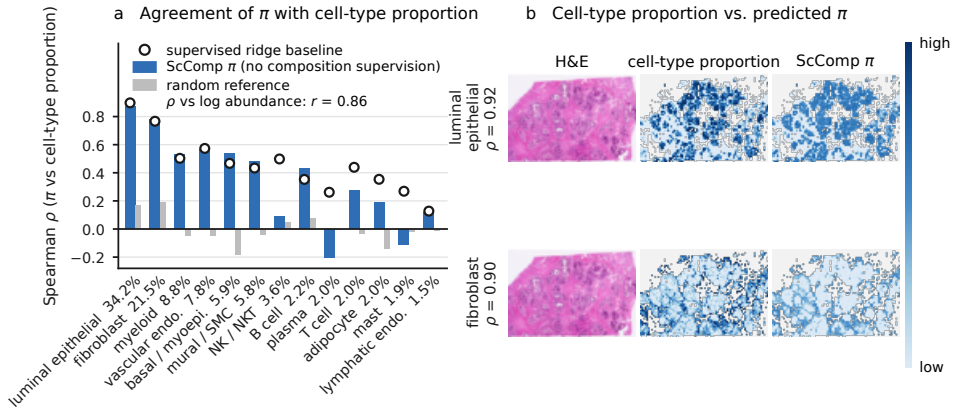

Figure 3: **Composition weights correspond to cell-type proportions without composition supervision.** (a) Spearman correlation between π and measured cell-type proportions for 13 cell types on held-out Xenium tiles, compared with a supervised ridge baseline and a random-reference control. Cell types are ordered by abundance, shown after each name. (b) H&E, measured cell-type proportion, and predicted π for luminal epithelial cells and fibroblasts on one held-out patient.

As shown in Fig. 3a, the predicted weights are correlated with the measured cell-type proportions, with a Spearman correlation of 0.66. The supervised ridge regressor reaches 0.71, while the random-reference baseline reaches 0.08. Figure 3b shows that this association is also observed spatially. For luminal epithelial cells and fibroblasts, regions with high predicted weights correspond to regions with high measured cell-type proportions. These results show that π retains the cell-type correspondence of the fixed reference without using cell-composition labels during training. Additional details are provided in Appendix G.

4.5 Ablation Studies

We evaluate the contribution of each model component to in-panel prediction. All variants use the same evaluation protocol and are compared using the equal-dataset macro HVG-50 PCC over three seeds.

As shown in Table 3, using only the target patch reduces HVG-50 PCC from 0.287 to 0.221, indicating that the local neighborhood provides substantial information for gene-expression prediction. Replacing the reference–residual prediction with a parameter-matched direct expression head gives a PCC of 0.282. Using the reference component only, without the residual, gives 0.279, while retaining the residual without the orthogonal projection gives 0.278. Removing the patient-balanced update reduces the PCC to 0.277.

Setting	HVG-50
ScComp	**.287**
target patch only	.221
direct expression head	.282
reference component only	.279
w/o orthogonal projection	.278
w/o patient-balanced update	.277

Table 3: Ablation of the main model components. HVG-50 is equal-dataset macro PCC on the training gene panel

These results indicate that the local neighborhood and patient-balanced update contribute most to in-panel prediction performance. The relatively small difference between ScComp and the direct expression head suggests that the main role of the reference–residual formulation is not limited to improving prediction on the training gene panel. In particular, the orthogonal projection supports prediction based on the learned composition weights beyond the training gene panel, as examined in Sec. 4.3.

5 Conclusion

We present ScComp, which predicts spatial gene expression from H&E using composition weights over a fixed single-cell reference matrix together with a reference-orthogonal residual. Across three public benchmarks, ScComp achieves the best equal-dataset macro performance among the compared methods while using only 0.87M trainable parameters. The ablation shows that a direct expression head achieves similar in-panel accuracy, indicating that the main value of the reference–residual formulation is not limited to improving prediction on the training gene panel. The composition weights can instead be applied to reference-covered genes that were never used during training and agree with measured cell-type proportions on Xenium data without composition supervision. The orthogonal projection helps preserve this behavior by preventing the residual from explaining variation already represented by the reference component.

The method also has two limitations. First, the composition weights depend on the coverage and quality of the single-cell reference. Cell types or states missing from the reference cannot be represented explicitly by the composition weights, and variation not captured by the available reference profiles must instead be modeled by the residual. Second, the Xenium cell-type labels are derived from gene expression using the same reference atlas, so they provide a silver standard rather than fully independent validation of the composition weights. Extending the reference to additional tissues and cell states and evaluating the predicted weights against independent cell-type measurements are natural directions for future work.

AI USE STATEMENT

We used large language models to assist with manuscript editing, code refactoring, and feedback on the presentation and analysis of experiments. The authors made all research decisions, ran the experiments, verified the code and numerical results, and reviewed and revised all AI-assisted text. No experimental results or figures were generated by an AI system. The authors take full responsibility for the final content of this work.

ETHICS STATEMENT

This work uses only previously released, publicly available datasets and does not involve the collection of new human-subject data. We use de-identified spatial transcriptomics, histology, and single-cell datasets from HER2ST, cSCC, STNet, 10x Genomics Xenium, and CZ CELLxGENE in accordance with their respective data-access and usage terms. No identifiable participant information was accessed. The predicted gene-expression profiles and composition weights are research outputs and have not been validated for clinical diagnosis or decision making.

REPRODUCIBILITY STATEMENT

All datasets used in this work are publicly available. Dataset splits, evaluation metrics, and gene sets are described in Sec. 4.1 and Appendix D. Single-cell reference construction is described in Appendix C, and architecture and optimization details are provided in Appendix B. The zero-shot gene pool and pathway evaluation are described in Appendix F, and the Xenium composition evaluation is detailed in Appendix G. All reported experiments use fixed patient-disjoint folds and three fixed random seeds unless otherwise stated. An anonymized implementation, configuration files, and the reference matrices are provided with the supplementary material.

REFERENCES

A. Andersson et al. Spatial deconvolution of HER2-positive breast cancer delineates tumor-associated cell type interactions. *Nature Communications*, 12:6012, 2021.

Chan Zuckerberg Initiative. CZ CELLxGENE Discover Census: Lts release 2025-11-08. `https://chanzuckerberg.github.io/cellxgene-census/cellxgene_census_docsite_data_release_info.html`, 2025. LTS data release.

S. N. Chandrasekaran, H. Ceulemans, J. D. Boyd, and A. E. Carpenter. Image-based profiling for drug discovery: due for a machine-learning upgrade? *Nature Reviews Drug Discovery*, 20(2): 145–159, 2021.

R. J. Chen et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30:850–862, 2024.

Y. Chung, J. H. Ha, K. C. Im, and J. S. Lee. Accurate spatial gene expression prediction by integrating multi-resolution features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11591–11600, 2024.

CZI Cell Science Program, Shibla Abdulla, Brian Aevermann, Pedro Assis, Seve Badajoz, Sidney M. Bell, Emanuele Bezzi, Batuhan Cakir, Jim Chaffer, Signe Chambers, J. Michael Cherry, et al. CZ CELLxGENE Discover: a single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *Nucleic Acids Research*, 53(D1):D886–D900, 2025. doi: 10.1093/nar/gkae1142.

M. Haghighi, J. C. Caicedo, B. A. Cimini, A. E. Carpenter, and S. Singh. High-dimensional gene expression and morphology profiles of cells across 28,000 genetic and chemical perturbations. *Nature Methods*, 19(12):1550–1557, 2022.

B. He et al. Integrating spatial gene expression and breast tumour morphology via deep learning. *Nature Biomedical Engineering*, 4:827–834, 2020.

S. Hu, Q. Zeng, N. Bhasker, J. N. Kather, and S. Speidel. HistoPrism: Unlocking functional pathway analysis from pan-cancer histology via gene expression prediction. In *International Conference on Learning Representations*, 2026.

T. Huang, T. Liu, M. Babadi, W. Jin, and Z. Ying. Scalable generation of spatial transcriptomics from histology images via whole-slide flow matching. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 25550–25565, 2025.

A. Janesick, R. Shelansky, A. D. Gottscho, F. Wagner, S. R. Williams, M. Rouault, G. Beliakoff, C. A. Morrison, M. F. Oliveira, J. T. Sicherman, et al. High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis. *Nature Communications*, 14:8353, 2023.

A. L. Ji et al. Multimodal analysis of composition and spatial architecture in human squamous cell carcinoma. *Cell*, 182(2):497–514.e22, 2020.

Y. Jia, J. Liu, L. Chen, T. Zhao, and Y. Wang. THItoGene: a deep learning method for predicting spatial transcriptomics from histological images. *Briefings in Bioinformatics*, 25(1):bbad464, 2024.

Y. Li, F. Zhong, and L. Liu. SciSt: single-cell reference-informed spatial gene expression prediction from pathological images. *Briefings in Bioinformatics*, 26(6):bbaf613, 2025.

A. Liberzon, C. Birger, H. Thorvaldsdóttir, M. Ghandi, J. P. Mesirov, and P. Tamayo. The molecular signatures database (MSigDB) hallmark gene set collection. *Cell Systems*, 1(6):417–425, 2015.

J. Lim, V. Chin, K. Fairfax, C. Moutinho, D. Suan, H. Ji, and J. E. Powell. Transitioning single-cell genomics into the clinic. *Nature Reviews Genetics*, 24(8):573–584, 2023.

W. Min, Z. Shi, J. Zhang, J. Wan, and C. Wang. Multimodal contrastive learning for spatial gene expression prediction using histology images. *Briefings in Bioinformatics*, 25(6):bbae551, 2024.

K. Nishimura, R. Bise, S. Matsuo, H. Hirose, and Y. Kojima. Cell-type prototype-informed neural network for gene expression estimation from pathology images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19801–19811, 2026.

A. Regev, S. A. Teichmann, E. S. Lander, I. Amit, C. Benoist, E. Birney, B. Bodenmiller, P. Campbell, P. Carninci, M. Clatworthy, et al. The Human Cell Atlas. *eLife*, 6:e27041, 2017.

P. L. Ståhl et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82, 2016.

R. Xie, K. Pang, S. W. Chung, C. T. Perciani, S. A. MacParland, B. Wang, and G. D. Bader. Spatially resolved gene expression prediction from histology images via bi-modal contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 70626–70637, 2023.

Y. Yang, M. Z. Hossain, E. A. Stone, and S. Rahman. Exemplar guided deep neural network for spatial transcriptomics analysis of gene expression prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5039–5048, 2023.

Y. Yang, M. Z. Hossain, X. Li, S. Rahman, and E. Stone. Spatial transcriptomics analysis of zero-shot gene expression prediction. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 492–502, 2024.

Yuansong Zeng, Zhuoyi Wei, Weijiang Yu, Rui Yin, Yuchen Yuan, Bingling Li, Zhonghui Tang, Yutong Lu, and Yuedong Yang. Spatial transcriptomics prediction from histology jointly through Transformer and graph neural networks. *Briefings in Bioinformatics*, 23(5):bbac297, 2022. doi: 10.1093/bib/bbac297.

W. Zhang, J. Chu, X. Liu, C. Tong, and X. Li. Dual-path knowledge-augmented contrastive alignment network for spatially resolved transcriptomics. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 12807–12815, 2026. doi: 10.1609/aaai.v40i15.38278.

Yijing Zhou, Yadong Lu, Qingli Li, Xinxing Li, and Yan Wang. Deep association multimodal learning for zero-shot spatial transcriptomics prediction. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, pp. 131–140. Springer Nature Switzerland, 2025. doi: 10.1007/978-3-032-04978-0_13.

A Overview

This appendix provides additional details on the implementation and computational cost, single-cell reference construction, evaluation protocol, statistical analysis, zero-shot prediction, Xenium composition evaluation, residual projection controls, and qualitative results. Unless otherwise stated, experiments use three random seeds, $\{2021, 2022, 2023\}$, under the patient-disjoint folds used in the main experiments.

B Implementation Details

Implementation settings. We use frozen 1,024-dimensional UNI features from the target patch and its 5×5 local neighborhood. A learned query pools the neighborhood features using 128-dimensional keys and queries. The target and pooled neighborhood features are fused by a two-layer LayerNorm–GELU–dropout trunk with hidden width 256 and dropout rate 0.5. The composition head is a linear layer followed by softmax. The residual head is a two-layer MLP with hidden width 256. Its output layer is initialized near zero, with weights drawn from $\mathcal{N}(0, (10^{-3})^2)$ and zero bias, so that the initial prediction is dominated by the composition term and the residual starts near zero.

Training begins with 60 epochs of AdamW over all training spots, using a learning rate of 10^{-3}, weight decay of 0.01, and gradient-norm clipping at 1.0. Each batch contains one complete tissue section (176–1,182 spots; median 320, 629, and 447 for HER2ST, cSCC, and STNet), and section order is reshuffled every epoch. We use complete sections because the objective includes a within-section Pearson correlation term.

This stage is followed by 60 iterations of patient-balanced local-update averaging. In each iteration, every training patient starts from the shared parameters and takes $L = 2$ gradient steps with $\eta_{\text{in}} = 10^{-2}$ on all of its spots. Patient updates are averaged with equal weight, and the resulting displacement is applied with $\epsilon_t = 0.5 \min(1, (t+1)/15)$. All trainable modules, including the residual head, are optimized throughout training. We use a Huber threshold of 1 and set $\lambda_{\text{PCC}} = 1$. In the training PCC term, we add $\varepsilon_{\text{corr}} = 10^{-8}$ to the denominator, so genes with zero measured or predicted variance within a section contribute zero correlation. All experiments are run on a single NVIDIA RTX PRO 6000 GPU.

Numerical rank. We retain singular directions satisfying

$$\frac{\sigma_j}{\sigma_1} > \tau, \qquad \tau = \max(K, N)\epsilon_{\text{float32}}. \tag{18}$$

The reference matrices $S \in \mathbb{R}^{K \times N}$ with $N = 250$ are numerically full row rank in all three datasets: $r = K = 68$ for HER2ST and STNet and $r = K = 76$ for cSCC, leaving orthogonal-complement dimensions of 182 and 174, respectively.

Computational efficiency. Table 4 compares the number of trainable predictor parameters in ScComp with those of the evaluated baselines.

C Single-Cell Reference Construction

Source and filtering. We construct tissue-specific references from the public CZ CELLxGENE Discover Census LTS snapshot 2025-11-08 (Chan Zuckerberg Initiative, 2025). We query *Homo sapiens* cells with `is_primary_data` set to true and retain cells with a finite positive full-transcriptome library size and a non-null cell-type annotation. Breast tissue (`tissue_general` = *breast*) is used for HER2ST and STNet, whereas *skin of body* is used for cSCC. For each tissue, we include healthy donors and donors with dataset-relevant malignancies. The breast reference

Method	Trainable (M)	Method	Trainable (M)
ScComp	0.87	DKAN	265.92
ST-Net	7.21	SGN	7.55
EGN	11.28	BLEEP	24.23
CPNN	28.93	TRIPLEX	29.29
SciSt	33.25	mclSTExp	41.46
STFlow	1.35	HistoPrism	7.23

Table 4: Predictor-only efficiency on the 250-gene HER2ST configuration.

includes breast cancer, invasive ductal breast carcinoma, triple-negative breast carcinoma, estrogen-receptor positive breast cancer, invasive lobular breast carcinoma, HER2-positive breast carcinoma, and breast carcinoma. The skin reference includes squamous cell carcinoma and basal cell carcinoma. Each cell is labeled as *normal-source* or *cancer-source* according to the donor diagnosis. This label indicates the source donor and does not imply that every cancer-source cell is malignant.

Reference rows. Each reference row corresponds to a (source, cell type) pair, where the cell type is the Cell Ontology annotation provided by Census. We retain all normal-source cell types with at least 150 eligible cells. From cancer sources, we retain epithelial and malignant states and omit immune, stromal, and endothelial profiles to avoid duplicating broad cellular populations already represented in the normal-source reference. For breast, the retained cancer-source states are *basal-myoepithelial cell of mammary gland*, *epithelial cell*, *luminal epithelial cell of mammary gland*, *malignant cell*, and *mammary gland epithelial cell*. For skin, they are *basal cell*, *basal cell of epidermis*, *malignant cell*, and *suprabasal keratinocyte*. This yields $K = 68$ rows for the breast datasets (63 normal-source and 5 cancer-source) and $K = 76$ rows for cSCC (72 normal-source and 4 cancer-source). HER2ST and STNet share the same breast reference rows, while their target-gene columns are dataset-specific. The complete row list is given in Table 6.

Donor-balanced sampling. We define a donor unit as a (study, donor) pair, namely (`dataset_id`, `donor_id`), so that donor identifiers reused across studies are not merged. Each reference row is capped at $1{,}500$ cells, with sampling distributed across donor units. For a row with U donor units, we set

$$q = \left\lceil \frac{1500}{U} \right\rceil . \tag{19}$$

We shuffle cells within each donor unit and sample up to q cells from each unit. If the resulting total exceeds $1{,}500$ cells, we retain cells in round-robin order across donor units visited in random order using seed 0. Because many donor units contain fewer than q cells, most rows do not reach the cap. The median row contains $1{,}252$ sampled cells for breast and 762 for skin; 16 of 68 breast rows and 5 of 76 skin rows reach the $1{,}500$-cell cap. Table 5 summarizes the resulting reference sizes.

Table 5: Reference construction statistics. Available cells are eligible Census cells before sampling, and donor units are defined as (study, donor) pairs.

	Breast (HER2ST, STNet)	Skin (cSCC)
Reference rows K (normal / cancer)	68 (63 / 5)	76 (72 / 4)
Available cells per row (min / median / max)	184 / 7,384 / 399,392	160 / 1,661 / 115,159
Donor units per row (min / median / max)	2 / 114 / 294	2 / 13 / 189
Sampled cells per row (min / median / max)	184 / 1,252 / 1,500	160 / 762 / 1,500
Rows reaching the $1{,}500$-cell cap	16 / 68	5 / 76
Total sampled cells	76,130	59,628

Normalization and aggregation. For cell c and gene g, raw counts are normalized as

$$\bar{X}_{cg} = \log\left(1 + 10^4 \frac{X_{cg}}{\sum_m X_{cm}}\right), \tag{20}$$

where the denominator is the full-transcriptome library size, i.e., the sum over all genes measured in the assay rather than over the target gene panel. The logarithm is natural. For each reference row,

profiles are first averaged within donor units and then averaged equally across donor units:

$$S_{k,g} = \frac{1}{|\mathcal{D}_k|} \sum_{d \in \mathcal{D}_k} \frac{1}{|\mathcal{I}_{k,d}|} \sum_{c \in \mathcal{I}_{k,d}} \bar{X}_{cg}. \tag{21}$$

Here, $\mathcal{D}_k$ denotes the donor units contributing to row k, and $\mathcal{I}_{k,d}$ denotes the sampled cells from donor unit d. This prevents donors with many cells from dominating a reference row. We apply no batch correction, gene-wise rescaling, or cross-row normalization, and the resulting reference matrix S remains fixed throughout model training.

Gene columns. Target-panel gene symbols are matched to Census `feature_name`. When multiple Census features share the same symbol, we retain the feature with the largest total raw count. Thirteen legacy panel symbols (AES, ATP5B, ATP5E, ATP5G2, C4orf48, H1FX, H2AFJ, H3F3B, KIAA0100, PHB, SEPW1, SLC9A3R1, and TCEB2) are mapped to their current Census symbols. Two clone-based panel names without a Census counterpart remain all-zero columns. For zero-shot prediction, we reuse the same reference rows and sampled cells and extract held-out gene columns from the same underlying reference. Thus, S and S^{zs} differ only in their gene columns rather than being estimated from independently sampled cells.

D Evaluation Protocol and Gene Sets

Evaluation metrics. We report mean absolute error (MAE), mean squared error (MSE), and Pearson correlation coefficient (PCC) between measured and predicted log-normalized expression. For each held-out section s, MAE and MSE are averaged over all spot–gene pairs. PCC is computed gene-wise across the n_s spots of the section,

$$r_{sg} = \frac{(\widehat{\mathbf{y}}_{sg} - \widehat{\mu}_{sg}\mathbf{1}_{n_s})^\top (\mathbf{y}_{sg} - \mu_{sg}\mathbf{1}_{n_s})}{\|\widehat{\mathbf{y}}_{sg} - \widehat{\mu}_{sg}\mathbf{1}_{n_s}\|_2 \, \|\mathbf{y}_{sg} - \mu_{sg}\mathbf{1}_{n_s}\|_2}, \tag{22}$$

where $\mathbf{y}_{sg}, \widehat{\mathbf{y}}_{sg} \in \mathbb{R}^{n_s}$ are the measured and predicted expression of gene g and $\mu_{sg}, \widehat{\mu}_{sg}$ their means across spots. The section-level PCC averages r_{sg} over genes with nonzero measured and predicted variance in that section; genes with zero variance have undefined PCC and are excluded.

For each random seed, section-level scores are averaged within each held-out fold, and fold-level scores are averaged within each dataset. The three dataset-level means are then weighted equally to obtain the equal-dataset macro score. We report the mean and population standard deviation across three seeds.

Gene sets. ALL contains all 250 supervised genes. Following the DKAN-compatible benchmark (Zhang et al., 2026), HEG-50 contains the 50 genes with the highest mean expression, whereas HVG-50 contains the 50 genes with the highest variance after per-spot total normalization. The same fixed dataset-level gene lists are used for all methods and folds.

E Additional Statistical Analysis

Each statistical unit is one held-out fold after averaging the three seeds. The three datasets contain 8, 4, and 8 folds, respectively. We use two-sided exact Wilcoxon tests for fold-paired comparisons and apply Holm correction over the family of 11 ScComp–baseline comparisons. Table 7 summarizes the macro HVG-50 results, variability across seeds, and fold-paired significance.

F Zero-Shot Prediction Details

Held-out gene pool and strata. The zero-shot pool is defined separately for each dataset as every gene of the single-cell reference that is present in the raw ST data and lies outside the 250-gene supervised panel (10,978, 12,122, and 11,049 genes for HER2ST, cSCC, and STNet). The same spatial preprocessing as for the supervised targets is applied. Gene subsets are defined per fold from the training slides only: the HVG-1000 subset contains the 1,000 pool genes with the highest

Breast: 63 N + 5 T	Skin: 72 N + 4 T
N B cell (2,826/591/11)	N B-1 B cell (174/174/10)
N CD4-positive helper T cell (26,379/1,500/210)	N B-2 B cell (959/656/11)
N CD4-positive, alpha-beta T cell (5,199/1,194/119)	N CD4-positive, alpha-beta T cell (359/359/4)
N CD8-positive, alpha-beta memory T cell (23,220/1,490/206)	N CD4-positive, alpha-beta memory T cell (166/166/5)
N IgA plasma cell (5,233/1,136/116)	N CD8-positive, alpha-beta T cell (1,481/1,064/4)
N IgG plasma cell (1,076/687/108)	N CD8-positive, alpha-beta memory T cell, CD45RO-positive (6,659/1,072/20)
N T cell (27,894/732/119)	N Langerhans cell (12,670/625/15)
N Tc1 cell (1,408/838/84)	N Merkel cell (1,140/966/9)
N activated CD4-positive, alpha-beta T cell (2,123/967/94)	N Schwann cell (571/551/24)
N activated CD8-positive, alpha-beta T cell (1,032/539/30)	N T cell (5,597/1,034/24)
N adipocyte (2,712/1,336/16)	N basal cell (519/441/16)
N adipocyte of breast (9,780/1,290/20)	N basal cell of epidermis (29,185/857/32)
N alternatively activated macrophage (9,059/1,428/123)	N cell of skeletal muscle (160/160/13)
N basal cell (125,667/1,500/160)	N central memory CD4-positive, alpha-beta T cell (228/228/5)
N basal-myoepithelial cell of mammary gland (162,976/1,500/170)	N chondrocyte (2,139/514/21)
N blood vessel endothelial cell (133,608/1,500/55)	N contractile cell (190/190/2)
N capillary endothelial cell (33,148/1,500/211)	N conventional dendritic cell (6,000/1,241/31)
N class switched memory B cell (4,493/1,198/205)	N cytotoxic T cell (11,286/1,156/23)
N classical monocyte (4,166/1,278/124)	N dendritic cell (7,063/992/39)
N conventional dendritic cell (3,985/1,337/118)	N endothelial cell (4,899/1,283/23)
N dendritic cell (220/220/33)	N endothelial cell of lymphatic vessel (6,148/968/30)
N effector memory CD4-positive, alpha-beta T cell (4,883/1,192/110)	N endothelial cell of vascular tree (30,322/1,321/33)
N effector memory CD8-positive, alpha-beta T cell (7,900/1,259/200)	N epidermal cell (446/446/3)
N endothelial cell (5,412/1,429/17)	N epithelial cell (339/178/14)
N endothelial cell of artery (18,763/1,500/214)	N epithelial cell of sweat gland (301/301/3)
N endothelial cell of lymphatic vessel (30,119/1,500/294)	N erythrocyte (1,279/523/18)
N endothelial cell of vascular tree (9,241/1,280/23)	N erythroid lineage cell (240/240/2)
N endothelial tip cell (53,780/1,474/89)	N eurydendroid cell (1,244/797/11)
N fibroblast (240,865/1,500/180)	N fibroblast (115,159/1,455/39)
N fibroblast of breast (762/755/2)	N glial cell (4,380/1,371/13)
N fibroblast of mammary gland (270,739/1,500/144)	N granular cell of epidermis (212/212/8)
N gamma-delta T cell (529/386/82)	N granulocyte (215/215/13)
N inflammatory macrophage (8,853/1,477/123)	N group 2 innate lymphoid cell (324/317/11)
N lactocyte (15,973/1,207/16)	N group 3 innate lymphoid cell (1,724/861/16)
N leukocyte (30,259/1,500/71)	N hair follicular keratinocyte (3,396/958/9)
N luminal adaptive secretory precursor cell of mammary gland (321,201/1,500/167)	N helper T cell (34,929/1,282/25)
N luminal epithelial cell of mammary gland (157,543/1,500/203)	N hematopoietic cell (759/713/9)
N luminal hormone-sensing cell of mammary gland (207,411/1,499/167)	N immature B cell (385/385/8)
N lymphocyte (1,667/965/115)	N inflammatory macrophage (1,773/1,423/5)
N macrophage (8,099/1,155/203)	N innate lymphoid cell (7,572/832/36)
N mammary gland epithelial cell (38,668/1,365/256)	N keratinocyte (72,566/1,335/37)
N mast cell (1,618/803/104)	N leukocyte (612/612/4)
N mature NK T cell (3,032/885/159)	N macrophage (43,928/1,421/55)
N mononuclear phagocyte (765/687/5)	N mast cell (4,380/927/48)
N mural cell (2,303/1,245/53)	N megakaryocyte (368/360/12)
N myeloid cell (9,073/655/115)	N melanocyte (9,553/1,228/40)
N myeloid dendritic cell (455/411/88)	N melanocyte of skin (847/675/22)
N naive B cell (2,760/846/170)	N mesenchymal stem cell (1,588/1,025/2)
N naive thymus-derived CD4-positive, alpha-beta T cell (6,869/1,337/88)	N monocyte (8,350/1,271/31)
N natural killer cell (4,418/1,085/159)	N monocyte-derived dendritic cell (12,490/1,500/5)
N neutrophil (508/447/55)	N myofibroblast cell (5,228/1,274/13)
N non-classical monocyte (996/709/106)	N naive thymus-derived CD4-positive, alpha-beta T cell (3,378/780/11)
N pericyte (26,404/1,500/146)	N naive thymus-derived CD8-positive, alpha-beta T cell (1,405/546/13)
N perivascular cell (131,160/1,500/144)	N natural killer cell (11,548/1,218/41)
N plasma cell (712/466/61)	N neuron (670/536/6)
N plasmacytoid dendritic cell (184/184/59)	N neutrophil (344/267/11)
N progenitor cell of mammary luminal epithelium (13,702/1,341/16)	N pericyte (13,284/1,500/34)
N regulatory T cell (2,365/957/114)	N plasma cell (517/465/25)
N subcutaneous adipocyte (1,756/788/2)	N plasmacytoid dendritic cell (1,693/814/32)
N unswitched memory B cell (740/587/138)	N pro-B cell (175/175/10)
N vascular associated smooth muscle cell (27,928/1,500/124)	N promonocyte (582/580/13)
N vascular lymphangioblast (8,796/1,457/16)	N regulatory T cell (12,028/1,161/37)
N vein endothelial cell (62,276/1,477/215)	N sebaceous gland cell (233/233/3)
T basal-myoepithelial cell of mammary gland (1,098/578/11)	N skeletal muscle satellite cell (5,768/745/10)
T epithelial cell (3,325/1,064/19)	N skin fibroblast (111,944/1,500/189)
T luminal epithelial cell of mammary gland (14,367/1,144/21)	N small pre-B-II cell (222/222/7)
T malignant cell (399,392/1,500/163)	N smooth muscle cell (733/480/18)
T mammary gland epithelial cell (13,379/1,273/20)	N spinous cell of epidermis (26,321/1,500/9)
	N stem cell of epidermis (1,187/1,050/5)
	N stromal cell (4,133/318/7)
	N suprabasal keratinocyte (1,629/735/20)
	N vascular associated smooth muscle cell (9,414/1,500/13)
	T basal cell (3,422/1,158/8)
	T basal cell of epidermis (2,745/657/8)
	T malignant cell (3,452/1,061/6)
	T suprabasal keratinocyte (272/272/8)

Table 6: Exact selective-reference rows. Each parenthesis gives available cells/sampled cells/dataset–donor units. N and T denote normal- and tumor-source profiles. The breast row set is shared by HER2ST and STNet; their gene columns differ.

Method	Macro HVG-50 (mean ± SD)	Δ [95% CI]	p_{Holm}
ScComp	.287 ± .002	—	—
DKAN	.254 ± .002	+.032 [.012, .055]	.014
EGN	.216 ± .002	+.071 [.033, .114]	7.9×10^{-4}
HistoPrism	.197 ± .004	+.090 [.056, .134]	2.7×10^{-5}
CPNN	.187 ± .003	+.100 [.067, .136]	2.5×10^{-4}
STFlow	.183 ± .008	+.103 [.079, .134]	2.7×10^{-5}
SGN	.182 ± .016	+.105 [.067, .144]	2.5×10^{-4}
BLEEP	.152 ± .006	+.135 [.117, .154]	2.7×10^{-5}
TRIPLEX	.151 ± .001	+.136 [.100, .173]	2.7×10^{-5}
mclSTExp	.104 ± .004	+.182 [.134, .233]	2.7×10^{-5}
ST-Net	.098 ± .001	+.188 [.152, .230]	2.7×10^{-5}
SciSt	.073 ± .004	+.214 [.178, .252]	2.7×10^{-5}

Table 7: Equal-dataset macro HVG-50 PCC and fold-paired comparisons. Mean and SD are over three seeds; Δ is the macro difference to ScComp with a 95% bootstrap confidence interval over the 20 folds; p_{Holm} is from exact two-sided Wilcoxon tests on the 20 seed-averaged folds with Holm correction over the 11 comparisons. ScComp wins 16–20 of 20 folds against every baseline.

variance across training-slide spots. Pathway-level activity is scored over the full pool. The same pool and strata are used for every method. For SGN, BLEEP, and mclSTExp, as for ScComp, the expression of held-out genes is not used for training, checkpoint selection, hyperparameter tuning, or fine-tuning. SGN predicts held-out genes through gene-name embeddings; BLEEP and mclSTExp predict them by retrieval, averaging the measured expression of the 50 nearest training spots in their image–expression embedding. All baselines use their published hyperparameters, and checkpoints are selected on panel genes only.

Pathway-level evaluation. We use the 50 MSigDB Hallmark gene sets (`h.all.v2026.1.Hs`) and map them to the full held-out gene pool of each dataset. All 50 sets retain at least 12 genes and are included in the evaluation. Unlike the gene-level metric, which is evaluated on the HVG-1000 subset, pathway activity is computed using the full held-out pool.

For each held-out section, observed and predicted expression are standardized separately for each gene across spots. Let $\mathcal{G}_s$ denote the held-out genes with nonzero variance and finite predictions in section s. For Hallmark set S, the background-centered activity at spot i is

$$a_i(S) = \frac{1}{|S \cap \mathcal{G}_s|} \sum_{g \in S \cap \mathcal{G}_s} z_{ig} - \frac{1}{|\mathcal{G}_s|} \sum_{g \in \mathcal{G}_s} z_{ig}, \tag{23}$$

where z_{ig} denotes the section-wise standardized expression of gene g. The score is computed independently for observed and predicted expression. For each Hallmark set, we compute the Pearson correlation between observed and predicted activity across spots, average over the 50 sets, and aggregate section-level scores using the same fold, dataset, and equal-dataset macro averaging as for the gene-level metric.

We use the background-centered score because the uncentered mean is strongly influenced by global expression variation across spots. With the uncentered score, size-matched random gene sets perform as well as or better than the Hallmark sets. After background centering, random-set PCC decreases to 0.01–0.04, whereas the Hallmark sets retain PCCs of 0.17–0.26.

G Xenium Composition Evaluation

Data and folds. We use the 12 sections of the 10x Xenium Prime Human Breast Biomarkers dataset (v4.0.0; S1–S4 × Top/Mid/Bot). Each section is an independent run from a different donor; the S1–S4 grouping is a naming convention rather than a physical slide or tissue block. We form four folds by this grouping, each holding out three sections, and train on the other nine (4 folds × 3 seeds). Because every donor contributes exactly one section, training and held-out sections never share a donor. Tiles are 100 μm bins of cell centroids with at least 10 cells.

Silver-standard cell types. Each cell is assigned one of 13 broad types by nearest-centroid (cosine) label transfer from the same CELLxGENE breast atlas used to build S, using the genes shared between the Xenium panel and the atlas. To keep the labelling genes separate from the evaluation, the shared genes are split into 5 folds and each cell is labelled with prototypes that exclude the genes of its own fold.

The 13-type reference. The 13 types collapse the atlas cell-type labels with the fixed rule in Table 8. The Xenium model uses a 13-row reference built from the atlas with this labelling, so π is 13-dimensional and is compared directly with the measured cell-type proportions. Training targets are the tile-level counts of the 263 panel genes present in the reference; the measured cell-type proportions are never used in training.

Broad type	Atlas labels
luminal epithelial	lactocyte; luminal adaptive secretory precursor; luminal epithelial cell of mammary gland; luminal hormone-sensing cell; mammary gland epithelial cell; progenitor cell of mammary luminal epithelium; tumor luminal epithelial cell; tumor mammary gland epithelial cell
basal / myoepithelial	basal cell; basal-myoepithelial cell of mammary gland; tumor basal-myoepithelial cell
fibroblast	fibroblast; fibroblast of breast; fibroblast of mammary gland
vascular endothelial	blood vessel endothelial; capillary endothelial; endothelial cell; endothelial cell of artery; endothelial cell of vascular tree; endothelial tip cell; vein endothelial cell
lymphatic endothelial	endothelial cell of lymphatic vessel; vascular lymphangioblast
mural / smooth muscle	mural cell; pericyte; perivascular cell; vascular associated smooth muscle cell
adipocyte	adipocyte; adipocyte of breast; subcutaneous adipocyte
myeloid	alternatively activated macrophage; classical monocyte; conventional dendritic cell; dendritic cell; inflammatory macrophage; macrophage; mononuclear phagocyte; myeloid cell; myeloid dendritic cell; neutrophil; non-classical monocyte; plasmacytoid dendritic cell
mast	mast cell
T	$CD4^+$ helper T; $CD4^+$ $\alpha\beta$ T; $CD8^+$ $\alpha\beta$ memory T; T cell; Tc1; activated $CD4^+$ and $CD8^+$ $\alpha\beta$ T; effector memory $CD4^+$ and $CD8^+$ $\alpha\beta$ T; $\gamma\delta$ T; naive thymus-derived $CD4^+$ $\alpha\beta$ T; regulatory T
NK / NKT	mature NK T cell; natural killer cell
B	B cell; class-switched memory B; naive B; unswitched memory B
plasma	IgA plasma cell; IgG plasma cell; plasma cell
excluded	leukocyte; lymphocyte (ambiguous immune). The tumor-source epithelial cell and malignant cell rows are assigned to the luminal epithelial category.

Table 8: Rule collapsing atlas cell-type labels to the 13 broad types used for the Xenium reference and the measured cell-type proportions.

H Residual Projection Analysis

We compare the orthogonal projection used in ScComp with an unconstrained residual and a random rank-matched projection. Only the fixed projector applied to the residual head changes; parameters, training schedule, and random seeds are identical. Table 9 reports in-panel HVG-50 PCC, the HVG-50 PCC of the reference component πS alone, zero-shot gene PCC, and the fraction of residual energy that lies in the row space of S.

Residual projector	HVG-50	Comp. branch alone	Zero-shot	Leakage
identity (unconstrained)	.278	.235	.175	.53
random subspace, rank-matched	.281	.245	.174	.46
complement of rowspace(S) (ScComp)	**.287**	**.256**	**.187**	**0**

Table 9: Residual projection controls (equal-dataset macro, 3 seeds). Reference component alone is the HVG-50 PCC of πS without the residual. Zero-shot is the macro gene-level PCC on held-out genes. Leakage is the fraction of residual energy within the row space of S.

Neither control differs significantly from the unconstrained residual on any column, whereas ScComp differs from each control with $p \leq 10^{-4}$ on the reference-branch, zero-shot, and leakage columns. The orthogonal projection thus removes overlap between the residual and the reference space while giving the highest reference-branch and zero-shot performance. A random rank-matched projection does not provide the same effect, indicating that the gain comes from removing the component aligned with the reference rather than from reducing the residual dimension.

I Additional Qualitative Results

Figure 4 compares the reference-based prediction πS with the final prediction after adding the orthogonal residual.

Figure 5 shows predicted spatial expression maps for representative genes on STNet across the compared methods.

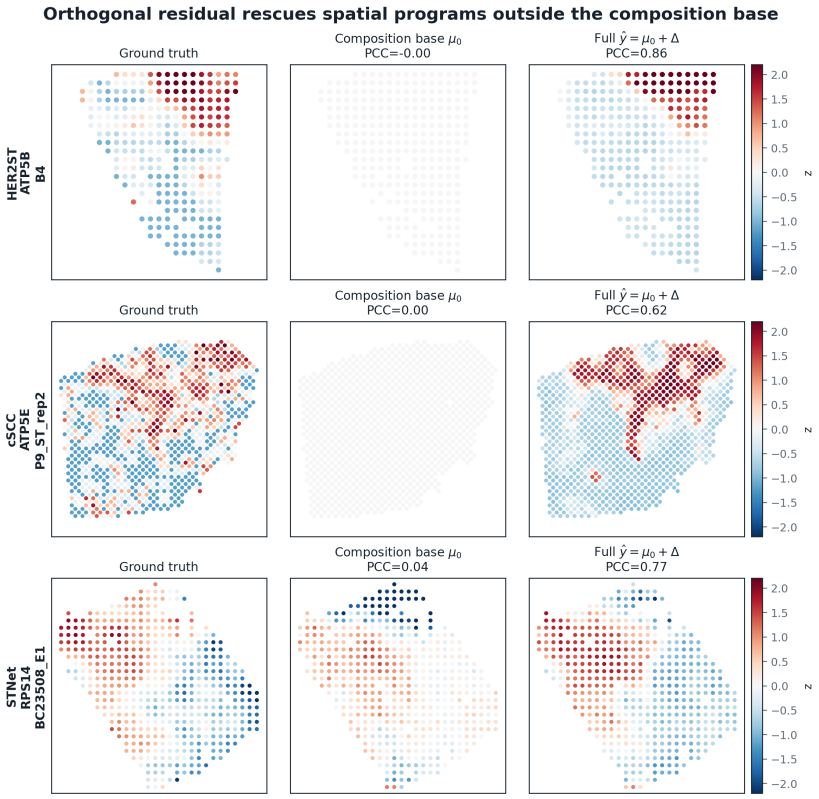

Figure 4: Reference component and final predictions for ATP5B (HER2ST), ATP5E (cSCC), and RPS14 (STNet). The final prediction adds the orthogonal residual to the reference-based prediction πS.

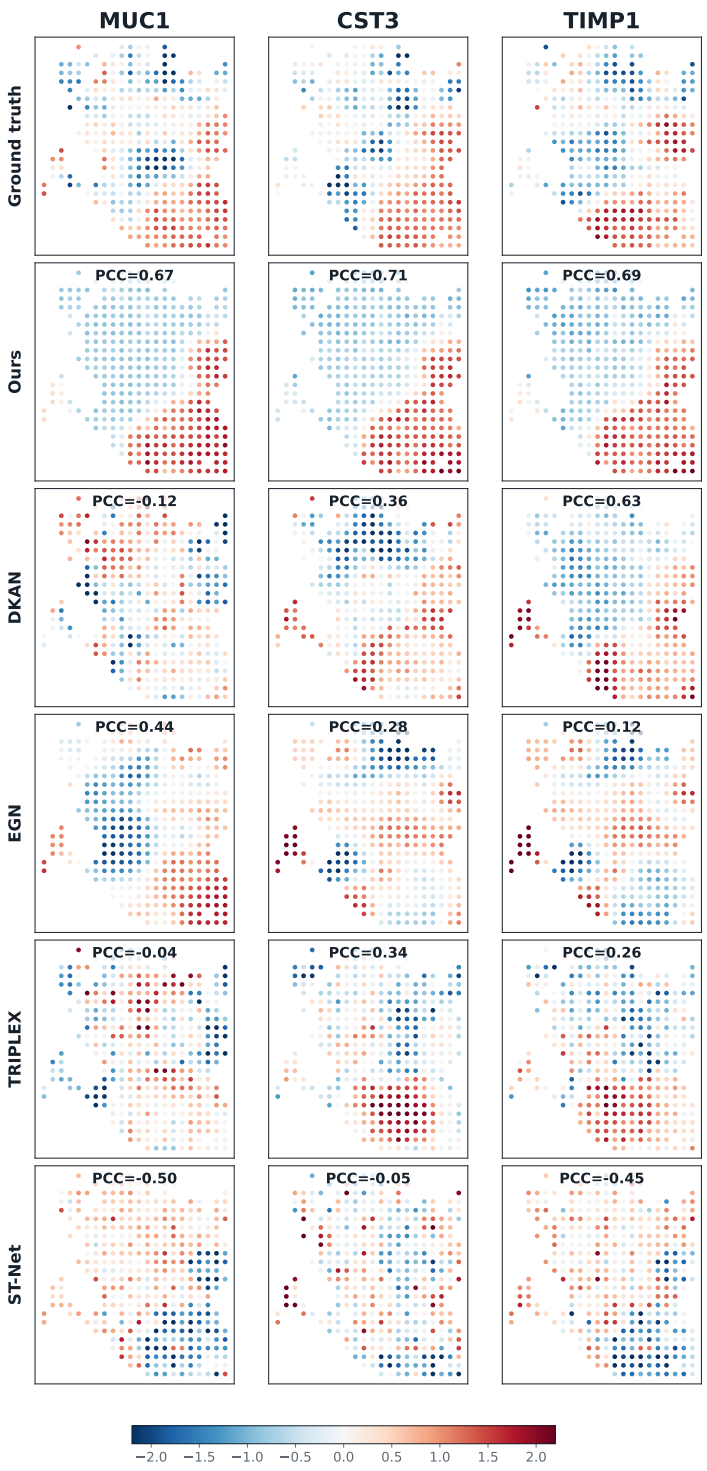

Figure 5: Predicted spatial expression maps for representative genes on STNet, compared with the observed expression and the competing methods.