BSPP PREDICTS PROTEIN EXPRESSION FROM SPATIAL TRANSCRIPTOMICS ACROSS TISSUE SECTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatial transcriptomics (ST) has emerged as a powerful method for linking gene expression to tissue architecture. While recent efforts have focused on ST platforms that measure both RNA and protein levels simultaneously, these approaches suffer from several limitations: they are often restricted by the availability of protein markers, high cost, and the inability to capture large-scale tissue information due to memory and computational constraints. Furthermore, existing models either fail to incorporate spatial dependencies or oversimplify the relationship between RNA and protein expression by treating them as independent variables. In this work, we propose BSPP, a framework combining a frozen RNA-derived proteomic prior with benefit-aware spatial residual routing. BSPP leverages Universal Cell Embedding (UCE) to incorporate cross-species biological knowledge while requiring neither manual annotations nor anatomical registration. We evaluate BSPP on multiple benchmark datasets, where it outperforms baselines by achieving better improvement in protein expression prediction accuracy and recovering spatially organized protein patterns. Our results highlight the potential of BSPP in bridging the gap between RNA expression and protein prediction in spatial transcriptomics.

1 INTRODUCTION

Spatial transcriptomics connects molecular measurements to their locations within tissues, enabling the study of cellular states in relation to tissue architecture and local microenvironments. By retaining positional information alongside RNA profiles, these technologies support analyses of spatial heterogeneity and the organization of cellular neighborhoods (Longo et al., 2021). However, transcript abundance alone does not fully describe the molecular phenotype of a cell or tissue region. Protein abundance also depends on translation and protein turnover, and the correspondence between RNA and protein measurements varies across genes and biological contexts (Vogel & Marcotte, 2012; Liu et al., 2016). Inferring protein expression from spatial RNA measurements therefore requires learning cross-modal relationships beyond the expression of the corresponding genes alone.

Spatial assays that jointly measure RNA and proteins provide experimental resources for learning these relationships. SPOTS combines spatial transcriptome profiling with protein-marker measurements, while spatial CITE-seq enables multiplexed protein and transcriptome co-mapping at cellular resolution (Ben-Chetrit et al., 2023; Liu et al., 2023). These approaches provide complementary molecular views of tissue organization, but their protein measurements remain tied to the selected antibody panels and experimental workflows. For a tissue section measured only at the RNA level, protein information must therefore be acquired through an additional assay or inferred from a suitable reference. This motivates a practical prediction setting: learning from a source section with paired RNA and protein measurements and predicting the assayed protein panel in a target section for which only RNA is available. The objective is to extend access to protein-level information within the reference panel, rather than to reconstruct an unrestricted proteome.

Existing methods approach this problem through complementary strategies. cTP-net learns a nonlinear mapping from single-cell transcriptomes to surface-protein abundance using paired reference measurements (Zhou et al., 2020b). totalVI jointly models RNA and protein observations in a probabilistic latent representation and supports integration of datasets with missing protein measurements (Gayoso et al., 2021a). Spatial approaches additionally exploit information about tissue or-

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.

ganization. SpaMosaic combines graph neural networks and contrastive learning for integration and missing-modality imputation across spatial datasets with partially overlapping measurements (Yan et al., 2026b). NicheTrans incorporates cellular microenvironment information into a Transformer-based framework for spatial cross- omics translation (Wang et al., 2026). Together, these methods establish the feasibility of transferring information across molecular modalities. They also motivate a complementary question: how accurately can an RNA-derived predictor transfer protein expression across tissue sections without requiring neighborhood-based correction, anatomical correspondence or manually annotated cell types?

Answering this question requires separating predictive modeling from access to target protein measurements. In the setting considered here, target proteins are unavailable for model fitting, hyperparameter selection and checkpoint selection, and are revealed only for final evaluation. We focus on a representation that maps RNA-derived features into a learned protein latent space and decodes them into protein expression. This formulation allows the predictor to represent relationships among the measured proteins while retaining an RNA-only input requirement at deployment. Pretrained transcriptomic representations provide an additional source of input features: Universal Cell Embedding (UCE) maps RNA profiles into a shared representation of cellular variation (Rosen et al., 2026). Whether these features improve cross-section protein prediction must be assessed through controlled ablations, rather than assumed from their pretraining alone.

Here, we present BSPP, a framework for cross-section protein expression prediction using an RNA-derived proteomic prior. BSPP combines an RNA encoder with projected UCE features through a learned fusion gate, predicts a protein latent representation and reconstructs protein expression through a protein variational autoencoder decoder. The evaluated model uses the intrinsic prediction path, with no spatial residual correction, and its parameters are fixed before target protein evaluation. We assess BSPP on eight directed prediction tasks across four spatial RNA–protein datasets, comparing it with cTP-net, totalVI, NicheTrans, SpaMosaic and a multilayer perceptron under a common scoring framework. The evaluation measures protein-wise and cell-wise Pearson correlation, protein-wise Spearman rank correlation and expression-profile structural similarity, alongside ablations of the UCE component. BSPP achieves the highest average scores among the compared methods on the two Pearson metrics and profile structural similarity, while the relative performance on Spearman correlation remains method-dependent. These experiments characterize both the predictive value and the limits of the RNA-derived prior for protein inference in the evaluated cross-section setting.

2 Related work

2.1 VAE in Cross-Modal Prediction

Variational Autoencoders (VAEs) (Kingma & Welling, 2014) have been widely used for learning latent representations in various cross-modal prediction tasks, including multi-omics data integration. In particular, VAEs provide a flexible framework for modeling high-dimensional data by learning probabilistic latent variable models. They have been applied to predict missing omics data and to integrate multiple data modalities, such as gene expression and protein abundance.

For example, methods like totalVI (Gayoso et al., 2021b) use VAEs to predict protein expression from single-cell RNA-seq data by learning a joint latent space for both transcriptomic and proteomic features. However, VAEs generally assume relatively simple latent structures between data modalities, which can be limiting when handling complex spatial dependencies and non-linear relationships in multi-omics data. Despite their flexibility, VAEs are less effective at explicitly modeling intricate spatial interactions or continuous transformations across distributions, which are crucial for accurate RNA-to-protein prediction.

2.2 Universal Cell Embedding (UCE)

The Universal Cell Embedding (UCE) (Rosen et al., 2023) is a single-cell foundation model that maps individual cells into a unified biological embedding space. Unlike conventional expression encoders that treat genes as independent input features, UCE introduces biologically informed gene tokens. Specifically, expressed genes are sampled according to their expression levels, represented by embeddings of their corresponding protein products derived from a large protein language model,

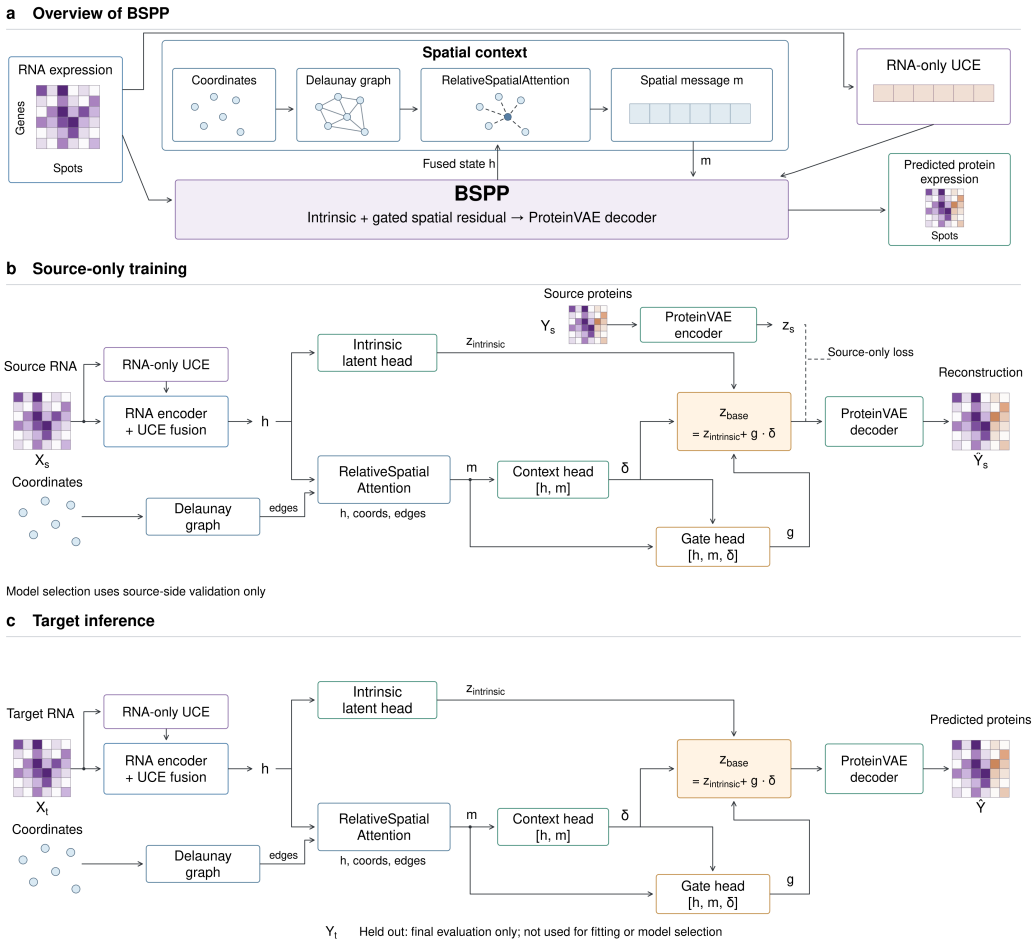

Figure 1: **Overview of BSPP. (a)** BSPP predicts protein expression by integrating RNA expression, RNA-derived UCE representations, and spatial context. A Delaunay graph is constructed from spot coordinates, and RelativeSpatialAttention aggregates relative node states and positions to capture spatial context. **(b)** During source-only training, RNA encoding and UCE fusion support intrinsic protein prediction. A context head produces a spatial residual, and a learned gate controls its contribution to the intrinsic prediction. A protein VAE encoder provides target representations for supervision, while the decoder maps predicted latents to protein expression. Model selection uses source-side validation only. **(c)** At target inference, the trained model uses target RNA, UCE representations, and spatial coordinates to predict protein abundance. Target proteins are excluded from fitting and model selection and are used only for final evaluation.

and organized according to genomic location and chromosome structure before being processed by a Transformer encoder. The model is pretrained in a self-supervised manner to distinguish truly expressed genes from unexpressed genes, allowing it to learn cell-state representations without requiring manual cell-type annotations.

This design provides several forms of biological prior knowledge. First, using protein-product embeddings allows UCE to encode sequence-level similarity between genes and supports transfer across homologous genes from different species. Second, organizing gene tokens by genomic location introduces structural information about the genome rather than treating the transcriptome as an unordered feature vector. Third, pretraining on large-scale cell atlas data across multiple species and tissues enables UCE to capture broad patterns of cellular identity, lineage, and tissue context. Therefore, in our model, UCE embeddings are used as biologically informed cell-state priors rather than direct predictors of protein abundance. By incorporating UCE-derived representations, STPFlow can condition RNA-to-protein prediction on a broader cellular context learned from large-scale single-cell atlases, which helps regularize the latent representation and improves the modeling of context-dependent protein expression in spatial tissues.

3 METHOD

3.1 PROBLEM FORMULATION

We consider cross-section protein expression prediction from spatial transcriptomic measurements. Let $X^s \in \mathbb{R}^{N_s \times G}$ and $Y^s \in \mathbb{R}^{N_s \times P}$ denote the preprocessed RNA and protein expression matrices of a source tissue section, respectively. Here, N_s is the number of spatial spots or cells, G is the number of input genes, and P is the number of measured proteins. For a target section, only the RNA expression matrix $X^t \in \mathbb{R}^{N_t \times G}$ is provided as predictive input. The objective is to estimate $\widehat{Y}^t \in \mathbb{R}^{N_t \times P}$ for the protein panel represented in the source data.

The source and target sections share a consistent input gene order and output protein order. Each prediction remains associated with the spatial location of its corresponding spot or cell. However, the evaluated model does not require a spatial neighborhood graph, anatomical correspondence between sections, or manually annotated cell types to compute its intrinsic predictions.

We additionally derive a pretrained Universal Cell Embedding (UCE) representation from each RNA profile (Rosen et al., 2026). Denoting this representation by u_i, the prediction task is

$$\widehat{y}_i = f_\theta(x_i, u_i), \qquad u_i = E_{\mathrm{UCE}}(x_i). \tag{1}$$

Both inputs are derived from RNA measurements; target protein measurements are unavailable for model fitting and checkpoint selection.

3.2 OVERVIEW OF BSPP

BSPP learns an RNA-derived proteomic prior through a mapping from transcriptomic features to a learned protein latent representation. The prediction pathway consists of an RNA encoder, a projection of pretrained UCE features, a learned fusion gate, an intrinsic latent prediction head, and a frozen protein variational autoencoder (VAE) decoder.

During predictor training, the protein VAE encoder provides latent supervision from source protein measurements. The RNA encoder and the UCE branch produce feature representations that are combined through a learned scalar gate. The intrinsic prediction head maps the fused representation into the standardized protein latent space. The predicted latent vector is subsequently transformed back to the decoder's original latent scale and decoded into protein expression.

The evaluated configuration uses only this intrinsic pathway. Spatial residual corrections and anatomical anchor predictions do not contribute to its output. After source-validation checkpoint selection, the complete predictor is fixed and applied to the target section.

3.3 PROTEIN LATENT REPRESENTATION

We use a pretrained protein VAE with encoder q_ϕ and decoder D_ψ to define the protein latent representation. For a protein expression vector $y_i \in \mathbb{R}^P$, the encoder parameterizes a diagonal

Gaussian distribution:

$$q_\phi(z \mid y_i) = \mathcal{N}\left(\mu_\phi(y_i), \mathrm{diag}\left(\sigma_\phi^2(y_i)\right)\right), \tag{2}$$

where $z \in \mathbb{R}^d$ and d is the protein latent dimension.

For training the RNA-to-protein predictor, we use the posterior mean

$$z_i^{\mathrm{raw}} = \mu_\phi(y_i) \tag{3}$$

as a deterministic supervision target, rather than sampling from the posterior.

Let $\mathcal{I}_{\mathrm{tr}}$ denote the source training indices. We compute the coordinate-wise mean m_s and standard deviation s_s of the latent targets using these indices:

$$m_s = \frac{1}{|\mathcal{I}_{\mathrm{tr}}|} \sum_{i \in \mathcal{I}_{\mathrm{tr}}} z_i^{\mathrm{raw}}, \qquad s_s = \max\left(\mathrm{Std}_{i \in \mathcal{I}_{\mathrm{tr}}}\left[z_i^{\mathrm{raw}}\right], 10^{-6}\right), \tag{4}$$

where the maximum is applied element-wise. The standardized supervision target is

$$z_i^* = (z_i^{\mathrm{raw}} - m_s) \oslash s_s, \tag{5}$$

where $\oslash$ denotes element-wise division.

The protein VAE is held fixed during BSPP predictor training. The same source-derived latent statistics are used for source validation and target inference. The protein encoder is required to construct training and validation targets, but it is not needed to predict proteins from target RNA.

3.4 RNA Encoding and UCE Fusion

RNA encoder. For each preprocessed RNA vector x_i, a multilayer perceptron produces an RNA representation

$$h_i^R = E_{\theta_R}(x_i) \in \mathbb{R}^H. \tag{6}$$

The encoder contains two linear layers, each followed by layer normalization and a GELU activation, with dropout after the first activation. The evaluated configuration uses a hidden dimension of $H = 128$ and a dropout rate of 0.1.

UCE representation. Precomputed UCE embeddings provide an additional representation of the RNA input. These embeddings are aligned to the expression matrices by spot or cell identifiers. A projection maps each embedding to the same dimension as the RNA representation:

$$h_i^U = P_\eta(u_i) = \mathrm{GELU}\left(\mathrm{LN}(W_U u_i + b_U)\right). \tag{7}$$

The UCE embeddings and projection parameters are frozen during predictor training.

The implementation also supplies the fusion gate with a scalar input statistic. For a finite UCE vector of dimension d_U, this statistic is

$$c_i = \tanh\left(\sqrt{\frac{1}{d_U}\sum_{k=1}^{d_U} u_{ik}^2}\right). \tag{8}$$

It is set to zero when the original embedding contains non-finite values, and non-finite entries are sanitized before projection. This statistic summarizes input magnitude and validity; it is not interpreted as a calibrated measure of biological confidence.

Gated feature fusion. A trainable gate combines the RNA and UCE representations:

$$a_i = \mathrm{sigmoid}\left(g_{\theta_g}\left([h_i^R; h_i^U; c_i]\right)\right), \tag{9}$$

where $[\,;\,]$ denotes concatenation and g_{θ_g} consists of a linear layer, a GELU activation, and a scalar output layer. The fused representation is

$$h_i = a_i h_i^R + (1 - a_i) h_i^U. \tag{10}$$

The scalar $a_i \in (0, 1)$ is shared across feature dimensions for each spot or cell. A larger value assigns greater weight to the RNA branch. The gate is learned using source protein supervision.

3.5 Intrinsic Protein Prediction and Decoding

The fused representation is mapped to the standardized protein latent space by an intrinsic prediction head:

$$\widehat{z}_i = H_{\theta_z}(h_i). \tag{11}$$

The head consists of two linear layers with an intermediate GELU activation, mapping from H hidden dimensions to d protein latent dimensions.

Before decoding, the predicted vector is restored to the original protein latent scale:

$$\widehat{z}_i^{\mathrm{raw}} = s_s \odot \widehat{z}_i + m_s, \tag{12}$$

where $\odot$ denotes element-wise multiplication. Protein expression is then predicted as

$$\widehat{y}_i = D_\psi\left(\widehat{z}_i^{\mathrm{raw}}\right). \tag{13}$$

The decoder outputs a vector in the preprocessed protein expression space. Although the protein representation originates from a VAE, the evaluated RNA-to-protein pathway produces a deterministic point prediction. It does not sample a target protein posterior or provide calibrated predictive uncertainty.

3.6 Training Objective and Target-blind Inference

Predictor objective. We supervise the predictor in both the protein latent space and the decoded expression space. For a source training set or batch $\mathcal{B}$ containing n observations, the latent mean squared error is

$$\mathcal{L}_{\mathrm{latent}} = \frac{1}{nd}\sum_{i\in\mathcal{B}} \|\widehat{z}_i - z_i^*\|_2^2, \tag{14}$$

and the protein expression mean squared error is

$$\mathcal{L}_{\mathrm{protein}} = \frac{1}{nP}\sum_{i\in\mathcal{B}} \|\widehat{y}_i - y_i\|_2^2. \tag{15}$$

We additionally use a protein-wise Pearson correlation loss. Let $\widetilde{\widehat{y}}_{ip} = \widehat{y}_{ip} - \overline{\widehat{y}}_p$ and $\widetilde{y}_{ip} = y_{ip} - \overline{y}_p$, where the means are taken over $\mathcal{B}$. The stabilized correlation for protein p is

$$\rho_p = \frac{\sum_{i\in\mathcal{B}} \widetilde{\widehat{y}}_{ip}\widetilde{y}_{ip}}{\sqrt{\max\left[\left(\sum_{i\in\mathcal{B}} \widetilde{\widehat{y}}_{ip}^2\right)\left(\sum_{i\in\mathcal{B}} \widetilde{y}_{ip}^2\right), \epsilon\right]}}, \qquad \epsilon = 10^{-8}. \tag{16}$$

The corresponding loss is

$$\mathcal{L}_{\mathrm{Pearson}} = 1 - \frac{1}{P}\sum_{p=1}^{P} \rho_p. \tag{17}$$

For the intrinsic configuration, the effective optimization objective is

$$\mathcal{L}_{\mathrm{BSPP}} = 1.5\,\mathcal{L}_{\mathrm{latent}} + 0.25\,\mathcal{L}_{\mathrm{protein}} + 0.025\,\mathcal{L}_{\mathrm{Pearson}}. \tag{18}$$

The coefficient of the latent term combines the final-output and intrinsic-output supervision terms, which coincide in this configuration. Any routing bookkeeping term in the shared implementation contributes no gradient to the intrinsic predictor when the spatial gate is fixed to zero.

Optimization and checkpoint selection. The RNA encoder, fusion gate, and intrinsic prediction head are optimized using source protein supervision. The UCE representation, UCE projection, and protein VAE remain fixed. Although the decoder parameters are frozen, gradients propagate through the decoder to train the RNA-to-latent predictor.

The evaluated runs use AdamW with an initial learning rate of 3×10^{-4} and weight decay of 5×10^{-4}. Learning-rate scheduling and early stopping are driven by source-validation protein-wise Pearson correlation. Source training and validation indices are fixed and recorded for each run. Checkpoints are selected using source-validation performance rather than target protein measurements.

4 EXPERIMENTS

We evaluate frozen intrinsic BSPP through three complementary experiments: (1) strict target-blind benchmark comparisons with external RNA-to-protein prediction methods; (2) target-blind ablation experiments examining the contribution of the UCE-enhanced intrinsic predictor; and (3) one-factor hyperparameter sensitivity experiments.

All experiments use the same RNA representation, protein ordering, coordinate conventions, and evaluation metrics. Target protein measurements are not used for model fitting, checkpoint selection, or hyperparameter selection.

4.1 BENCHMARK EXPERIMENTS

To systematically evaluate the performance of BSPP, we conduct benchmark experiments on four spatial transcriptomics datasets with paired RNA and protein measurements, including SPOTS Mouse Spleen, Mouse Thymus, Human Lymphnode, and Human Renal. These datasets represent different prediction settings. SPOTS Mouse Spleen corresponds to a replicate-sample setting, where two biological replicates are used for bidirectional prediction. Mouse Thymus and Human Lymphnode correspond to cross-slice prediction, where the model is trained on one tissue slice and evaluated on another slice. Human Renal corresponds to a same-slice setting, where a single tissue section is split into two spatial regions and evaluated bidirectionally. For each dataset, we perform prediction in both directions, resulting in a total of eight benchmark tasks. The complete benchmark results are reported in Table 1.

Baselines. There are three categories of baseline methods:

- **MLP baseline.** We first include a simple multi-layer perceptron (MLP) as a non-spatial neural baseline. The MLP directly learns a mapping from RNA expression to protein expression without using spatial coordinates, neighborhood information, or tissue-level context.
- **Single-cell RNA-to-protein prediction methods.** We compare BSPP with two single-cell multi-omics methods, totalVI (Gayoso et al., 2021b) and cTP-net (Zhou et al., 2020a). totalVI is a probabilistic deep generative model originally designed for joint modeling of paired single-cell RNA and protein measurements, such as CITE-seq data. cTP-net predicts cell-surface protein abundance from single-cell transcriptomes using a deep transfer learning framework. Since these methods are developed for single-cell RNA-to-protein translation, they do not explicitly model spatial tissue organization.
- **Spatial transcriptomics methods.** We further compare BSPP with spatially aware methods, including SpaMosaic (Yan et al., 2026a) and NicheTrans (Wang, Zhikang and others, 2024). SpaMosaic performs spatial multi-omics integration by learning a modality-agnostic latent representation and can be applied to missing-modality imputation. NicheTrans is a spatial-aware cross-omics translation framework that incorporates neighborhood context for modality translation. Compared with these methods, BSPP explicitly models RNA-to-protein prediction through a benefit-aware spatial routing mechanism, enabling protein expression generation while preserving cell-intrinsic proteomic priors and spatial organization.

For each prediction task, we evaluate performance using four complementary metrics: PCC (protein), PCC (cell), SSIM, and SPCC. Let $\mathbf{Y} \in \mathbb{R}^{N \times P}$ denote the ground-truth protein expression matrix and $\hat{\mathbf{Y}} \in \mathbb{R}^{N \times P}$ denote the predicted protein expression matrix, where N is the number of spatial spots or cells and P is the number of proteins.

The benchmark is strictly target-blind. BSPP is selected using source-side RNA and source-side validation data only. Target ADT is opened only after checkpoint and protocol freezing for final scoring. External predictions are included in the main table only when their prediction artifacts satisfy the same audited target-blind criterion.

BSPP integrates RNA-derived representations and UCE features with spatial context through benefit-aware spatial routing. As shown in Table 1, BSPP ranks first in 21 of the 32 metric columns

Mouse Spleen

Method	rep1 → rep2				rep2 → rep1			
	PCC (protein)	PCC (cell)	SSIM	SPCC	PCC (protein)	PCC (cell)	SSIM	SPCC
BSPP	**0.542**	**0.942**	**0.928**	0.535	0.541	**0.948**	**0.939**	0.520
NicheTrans	0.495	0.114	0.063	**0.574**	0.505	0.491	0.221	**0.609**
SpaMosaic	0.411	0.914	0.000	0.390	0.411	0.911	0.001	0.389
cTP-net	0.298	0.145	0.046	0.316	0.338	0.167	0.051	0.355
totalVI	0.421	0.925	0.298	0.412	**0.531**	0.940	0.279	0.485
MLP	-0.203	0.887	0.397	-0.178	-0.130	0.913	0.419	-0.113

Mouse Thymus

Method	2 → 3				3 → 2			
	PCC (protein)	PCC (cell)	SSIM	SPCC	PCC (protein)	PCC (cell)	SSIM	SPCC
BSPP	0.712	**0.978**	**0.968**	0.454	0.710	**0.978**	**0.967**	0.485
NicheTrans	0.561	0.158	0.064	**0.503**	0.597	0.130	0.058	0.500
SpaMosaic	0.431	0.925	0.001	0.409	0.406	0.927	0.001	0.377
cTP-net	0.226	0.149	0.056	0.346	0.227	0.136	0.051	0.352
totalVI	**0.747**	0.969	0.208	0.463	**0.780**	0.972	0.228	**0.532**
MLP	-0.012	0.271	0.246	0.130	-0.178	0.894	0.477	-0.045

Human Lymphnode

Method	1 → 2				2 → 1			
	PCC (protein)	PCC (cell)	SSIM	SPCC	PCC (protein)	PCC (cell)	SSIM	SPCC
BSPP	0.354	**0.915**	**0.887**	0.400	0.350	**0.899**	**0.867**	0.439
NicheTrans	0.128	0.059	0.016	0.331	0.213	0.240	0.070	**0.474**
SpaMosaic	**0.372**	0.893	0.000	0.378	**0.355**	0.883	0.000	0.368
cTP-net	0.305	-0.054	0.042	0.343	0.339	-0.035	0.038	0.422
totalVI	0.202	0.839	0.135	**0.415**	0.216	0.834	0.136	0.468
MLP	-0.089	-0.070	0.086	-0.097	0.076	0.274	0.182	0.207

Human Renal

Method	A → B				B → A			
	PCC (protein)	PCC (cell)	SSIM	SPCC	PCC (protein)	PCC (cell)	SSIM	SPCC
BSPP	**0.511**	**0.744**	**0.766**	0.338	**0.484**	**0.746**	**0.766**	**0.315**
NicheTrans	0.420	0.610	0.467	0.294	0.428	0.641	0.536	0.295
SpaMosaic	0.368	0.618	0.290	0.289	0.330	0.625	0.283	0.259
cTP-net	0.401	0.529	0.274	0.270	0.228	0.315	0.364	0.173
totalVI	0.209	0.516	0.136	0.143	0.167	0.483	0.102	0.111
MLP	0.507	0.633	0.444	**0.358**	0.431	0.563	0.400	0.293

Table 1: Benchmark results for RNA-to-protein prediction. Each dataset is evaluated in two directions. Each direction reports PCC (protein), PCC (cell), SSIM, and SPCC. BSPP achieves the best or second-best performance across most metrics and datasets, demonstrating strong robustness across different prediction settings.

and second in eight, achieving the highest PCC (cell) and SSIM across all eight prediction directions. Its most pronounced advantage lies in SSIM: BSPP achieves scores of 0.928–0.939 on Mouse Spleen, 0.967–0.968 on Mouse Thymus, 0.867–0.887 on Human Lymphnode, and 0.766 in both directions on Human Renal. These results exceed the strongest competing method in each direction by absolute margins of 0.230–0.752, demonstrating consistently higher structural similarity to the ground truth across replicate-sample, cross-slice, and same-slice settings.

This improvement is particularly relevant to spatial omics, where recovering the tissue organization of protein expression is a central objective beyond predicting abundance at individual spots. The consistently higher SSIM suggests that BSPP better preserves spatial expression structure and tissue-level heterogeneity, complementing its strong expression correlations. Although BSPP does not lead every protein-wise PCC or SPCC comparison, its consistent advantage in structural similarity highlights its ability to recover spatially organized protein patterns—a key requirement for studying biological variation within tissues.

4.2 Ablation Studies

We evaluate the contributions of UCE and spatial graph components using four datasets and eight prediction directions. As shown in Table 2, removing UCE consistently degrades performance by 20%–36% across all metrics, demonstrating that the UCE-enhanced RNA-derived proteomic prior is essential for BSPP's robust performance.

In contrast, ablating the spatial graph context by permuting node labels has minimal impact on most directions. Performance decreases by only 0.04%–5.5% in PCC, with an average reduction of approximately 0.4%. The no_spatial variant performs similarly to the full model on SPOTS, SpatialCOC, and SpatialGluex datasets, indicating that the frozen UCE representation provides the core predictive power while the spatial graph serves as a complementary refinement mechanism.

Notably, the Renal B→A direction shows a relatively larger performance drop (5.5% PCC decrease) when the graph is permuted, suggesting that spatial context plays a more important role in certain data configurations.

Overall, these results demonstrate that the UCE-enhanced RNA-derived proteomic prior is essential for BSPP's robust performance, while spatial context provides additional but less critical benefits.

Mouse Spleen

Method	rep1 → rep2				rep2 → rep1			
	PCC (protein)	PCC (cell)	SSIM	SPCC	PCC (protein)	PCC (cell)	SSIM	SPCC
BSPP (full)	**0.5422**	**0.9416**	**0.9285**	**0.5349**	**0.5391**	**0.9469**	**0.9385**	**0.5201**
w/o spatial	0.5420	0.9415	0.9282	0.5338	0.5384	0.9468	0.9384	0.5197
w/o UCE	0.4203	0.9122	0.8952	0.3676	0.3635	0.9206	0.9108	0.2655

Mouse Thymus

Method	2 → 3				3 → 2			
	PCC (protein)	PCC (cell)	SSIM	SPCC	PCC (protein)	PCC (cell)	SSIM	SPCC
BSPP (full)	**0.7076**	**0.9783**	**0.9681**	**0.4535**	**0.7098**	**0.9783**	**0.9668**	**0.4847**
w/o spatial	0.7039	0.9774	0.9673	0.4487	0.7038	0.9780	0.9663	0.4804
w/o UCE	0.3471	0.9770	0.9624	0.3842	0.4265	0.9794	0.9644	0.3899

Human Lymphnode

Method	1 → 2				2 → 1			
	PCC (protein)	PCC (cell)	SSIM	SPCC	PCC (protein)	PCC (cell)	SSIM	SPCC
BSPP (full)	**0.5542**	**0.9156**	**0.8994**	**0.4942**	**0.5457**	**0.8971**	**0.8773**	**0.4829**
w/o spatial	0.5526	0.9136	0.8968	0.4905	0.5443	0.8960	0.8740	0.4810
w/o UCE	0.3541	**0.9156**	0.8870	0.3904	0.3408	**0.8971**	0.8651	0.4279

Human Renal

Method	A → B				B → A			
	PCC (protein)	PCC (cell)	SSIM	SPCC	PCC (protein)	PCC (cell)	SSIM	SPCC
BSPP (full)	**0.5113**	**0.7438**	**0.7662**	**0.3381**	**0.5109**	**0.7555**	**0.7945**	**0.3677**
w/o spatial	0.5100	0.7419	0.7651	0.3375	0.4830	0.6316	0.6011	0.2321
w/o UCE	0.3219	0.6137	0.4469	0.2321	0.3273	0.6316	0.5718	0.2321

Table 2: Ablation study of BSPP. We compare the full model with two variants: w/o spatial removes the spatial graph context (permuted node labels), and w/o UCE removes the universal cell embedding branch. Each dataset is evaluated in two transfer directions, and each direction reports PCC (protein), PCC (cell), SSIM, and SPCC. Higher values indicate better performance. The best result in each column is shown in bold, and the second-best result is underlined.

Table 3: Hyperparameter sensitivity analysis of BSPP. Each value reports the range of mean performance across different settings, where the mean is computed over all transfer settings.

Hyperparameter	Settings	PCC (protein)	PCC (cell)	SPCC	SSIM
Latent dimension	$z_{\text{dim}} \in \{64, 256, 512\}$	0.563–0.614	0.942–0.947	0.562–0.594	0.936–0.941
Dropout rate	$p \in \{0.0, 0.2, 0.3\}$	0.583–0.586	0.945–0.946	0.569–0.578	0.939–0.940
Protein weight	$\lambda_{\text{protein}} \in \{0.1, 0.5, 1.0\}$	0.575–0.587	0.946–0.948	0.568–0.575	0.940–0.941
Protein Pearson weight	$\lambda_{\text{Pearson}} \in \{0.0, 0.05, 0.2\}$	0.580–0.587	0.946–0.946	0.569–0.574	0.940–0.940

4.3 Hyperparameter Experiments

We next examine the impact of various hyperparameters on the performance of BSPP. Table 3 summarizes the performance range under different method-specific hyperparameter settings. Across latent dimension, dropout rate, protein weight, and protein Pearson weight, BSPP shows only limited performance variation, suggesting that the model is generally robust to these hyperparameters.

In these experiments, we vary each hyperparameter and report the corresponding performance metrics on the source validation set. The results show that larger latent dimension z_{dim} significantly improves prediction accuracy. Notably, the intrinsic RNA encoder architecture plays a crucial role in determining the model's predictive capacity.

For instance, increasing the latent space dimension allows the model to capture finer variations in protein expression, while adjusting the routing-related hyperparameters has minimal impact on overall performance. These findings help optimize BSPP for better performance in practical applications.

5 Conclusion

In this paper, we study spatial RNA-to-protein prediction, aiming to infer spatially resolved protein abundance from spatial transcriptomic profiles. We propose BSPP, a frozen intrinsic predictor that combines RNA expression with UCE-derived biological priors to estimate protein abundance under a strict target-blind protocol. Across four paired spatial multi-omics resources and five external baselines, BSPP achieves the strongest and most consistent performance for cell-wise protein recovery and spatial similarity, obtaining macro-average Cell PCC of 0.8937 and SSIM of 0.8860 across eight dataset-direction tasks. Ablation experiments further show that removing the UCE-enhanced prior substantially reduces prediction quality. These results support BSPP as a simple, reproducible, and target-blind baseline for spatial RNA-to-protein prediction.

AI USE STATEMENT

In this work, we used generative AI tools for helping generate codes, polishing the paper and figures. We have not used generative AI tools for generating research ideas, designing the experiments, explaining experiments results and generating/reorganizing data are not applicable to this work. Additionally, we used generative AI tools for polishing the paper and figures. We have reviewed all AI-assisted work. we checked LLM-generated sentences, LLM-generated code was verified and tested for correctness by authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The Method section and the Appendix provide descriptions of the proposed approach and implementation details to support reproduction of our results. We will publicly release the source code upon acceptance of the paper.

REFERENCES

Nir Ben-Chetrit, Xiang Niu, Ariel D. Swett, Jesus Sotelo, Maria S. Jiao, Caitlin M. Stewart, Catherine Potenski, Paulius Mielinis, Patrick Roelli, Marlon Stoeckius, and Dan A. Landau. Integration of whole transcriptome spatial profiling with protein markers. *Nature Biotechnology*, 41:788–793, 2023. doi: 10.1038/s41587-022-01536-3.

Adam Gayoso, Zoë Steier, Romain Lopez, Jeffrey Regier, Kristopher L. Nazor, Aaron Streets, and Nir Yosef. Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nature Methods*, 18:272–282, 2021a. doi: 10.1038/s41592-020-01050-x.

Adam Gayoso, Zoë Steier, Romain Lopez, Jeffrey Regier, Kristopher L Nazor, Aaron Streets, and Nir Yosef. Joint probabilistic modeling of single-cell multi-omic data with totalvi. *Nature Methods*, 18:272–282, 2021b.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014. doi: 10.48550/arXiv.1312.6114. URL https://arxiv.org/abs/1312.6114.

Yang Liu, Marcello DiStasio, Graham Su, Hiromitsu Asashima, Archibald Enninful, Xiaoyu Qin, Yanxiang Deng, Jungmin Nam, Fu Gao, Pino Bordignon, Marco Cassano, Mary Tomayko, Mina Xu, Stephanie Halene, Joseph E. Craft, David Hafler, and Rong Fan. High-plex protein and whole transcriptome co-mapping at cellular resolution with spatial CITE-seq. *Nature Biotechnology*, 41: 1405–1409, 2023. doi: 10.1038/s41587-023-01676-0.

Yansheng Liu, Andreas Beyer, and Ruedi Aebersold. On the dependency of cellular protein levels on mRNA abundance. *Cell*, 165:535–550, 2016. doi: 10.1016/j.cell.2016.03.014.

Sophia K. Longo, Margaret G. Guo, Andrew L. Ji, and Paul A. Khavari. Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics. *Nature Reviews Genetics*, 22: 627–644, 2021. doi: 10.1038/s41576-021-00370-8.

Yanay Rosen, Yusuf Roohani, Ayush Agrawal, Leon Samotorcan, Tabula Sapiens Consortium, Stephen R. Quake, and Jure Leskovec. Universal cell embeddings: A foundation model for cell biology. *bioRxiv*, pp. 2023.11.28.568918, 2023. doi: 10.1101/2023.11.28.568918. URL https://www.biorxiv.org/content/10.1101/2023.11.28.568918.

Yanay Rosen, Yusuf Roohani, Ayush Agrawal, Leon Samotorčan, Tabula Sapiens Consortium, Stephen R. Quake, and Jure Leskovec. Universal cell embedding provides a foundation model for cell biology. *Nature*, 656:183–191, 2026. doi: 10.1038/s41586-026-10689-z.

Christine Vogel and Edward M. Marcotte. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature Reviews Genetics*, 13:227–232, 2012. doi: 10.1038/nrg3185.

Zhikang Wang, Qi Zou, Senlin Lin, Sijie Li, Yan Cui, Daoliang Zhang, Chuangyi Han, Yida Li, Jianmin Li, Yi Zhao, Rui Gao, Jiangning Song, and Zhiyuan Yuan. NicheTrans: spatial-aware cross-omics translation. *Nature Methods*, 23:1528–1539, 2026. doi: 10.1038/s41592-026-03153-3.

Wang, Zhikang and others. Nichetrans: Spatial-aware cross-omics translation. *bioRxiv*, 2024.

X. Yan et al. Mosaic integration of spatial multi-omics with spamosaic. *Nature Genetics*, 2026a.

Xuhua Yan, Zhaoyu Fang, Kok Siong Ang, Lynn van Olst, Alex Edwards, Thomas Watson, Ruiqing Zheng, Di Zhang, Rong Fan, Min Li, David Gate, and Jinmiao Chen. Mosaic integration of spatial multi-omics with SpaMosaic. *Nature Genetics*, 58:1126–1137, 2026b. doi: 10.1038/s41588-026-02573-3.

Zilu Zhou, Chao Ye, Jun Wang, and Nancy R Zhang. Surface protein imputation from single cell transcriptomes by deep neural networks. *Nature Communications*, 11:651, 2020a.

Zilu Zhou, Chengzhong Ye, Jingshu Wang, and Nancy R. Zhang. Surface protein imputation from single cell transcriptomes by deep neural networks. *Nature Communications*, 11:651, 2020b. doi: 10.1038/s41467-020-14391-0.

A Implemention Details

Running environment. All experiments were performed on an NVIDIA A800 80GB PCIe GPU. The software environment used NVIDIA driver 550.107.02 and CUDA 12.4 for GPU acceleration.

Data preprocessing. For each dataset, we use paired spatial RNA and protein profiles when available. RNA and protein matrices are first aligned by their shared spot or cell barcodes. For cross-section prediction, we further restrict the input features to the intersection of genes shared by all training and testing RNA samples, and restrict the output space to the shared protein panel across the corresponding protein measurements.For both RNA and protein modalities, we apply a $\log(1+x)$ transformation to the raw count matrices. This transformation reduces the dynamic range of highly expressed features while preserving zero entries. The spatial coordinates of spots are extracted from the metadata when available; otherwise, they are parsed from spot identifiers if coordinates are encoded in the barcode names. A spatial k-nearest-neighbor graph is constructed independently for each tissue section based on the two-dimensional spatial coordinates.

Data splits. For each benchmark dataset, we use the paired spatial RNA and protein measurements from the training sample to train the model, and evaluate on a held-out test sample or held-out tissue region, as summarized in Table 4. Specifically, the test protein measurements are used only for evaluation and are never used during model training, feature selection, hyperparameter tuning, or normalization-statistic estimation. For model selection, each original training sample is first divided into 16 spatial batches, among which 25% are held out as the validation set and the remaining 75% are used for training. The checkpoint with the best validation performance is selected for final testing. All reported test results are computed using the selected checkpoint on the held-out test sample.

Preprocessing and feature alignment. For each dataset, RNA and protein matrices are aligned by shared spot or cell barcodes. Genes are restricted to the common gene set shared between the training and test RNA samples, and proteins are restricted to the shared measured protein panel. Unless otherwise specified, the top 2,000 highly variable genes are selected from the training RNA data and used as model inputs. Both RNA and protein count matrices are transformed using $\log(1 + x)$, where x denotes the raw count value. The same preprocessing procedure is applied to all methods to ensure a fair comparison. Spatial coordinates are extracted from the dataset metadata when available and are used to construct a spatial k-nearest-neighbor graph for spatial methods.

Baseline implementation. We compare STPFlow with six baseline methods: MLP, cTP-net, totalVI, SpaMosaic, and NicheTrans. All baselines use the same train-test split and the same aligned input genes and output proteins as STPFlow whenever possible.

MLP. The MLP baseline directly maps RNA expression to protein expression using fully connected layers. The model is trained with mean squared error loss using the Adam optimizer. The hidden dimension is 128, and learning rate is 1e-3.

cTP-net. For cTP-net, we adapt the original single-cell RNA-to-protein prediction framework to the spatial RNA-to-protein setting. Following the cTP-net preprocessing strategy, RNA expression is normalized using a Seurat-like transformation.

totalVI. For totalVI, we use the implementation from `scvi-tools`. The model is trained only on the training samples with paired RNA and protein measurements. During testing, the test RNA samples are treated as query data with protein measurements masked or unavailable. All parameters follow the default settings of `scvi-tools`.

SpaMosaic. SpaMosaic is evaluated as a spatial multi-omics missing-modality imputation method. We provide RNA and protein modalities for the training samples and mask the protein modality in the test sample. SpaMosaic then imputes the missing protein profiles based on the learned cross-modal and spatial representations. The predicted protein matrix is aligned with the ground-truth protein matrix by spot and protein names before evaluation.

NicheTrans. NicheTrans is evaluated using its spatial-aware cross-omics translation framework. Since our task focuses on RNA-to-protein prediction without assuming paired histology image inputs, we use the omics-based setting of NicheTrans. The model is trained on paired RNA and protein data from the training samples and then applied to the test RNA sample to predict protein abundance. All parameters follow the official training protocol.

B DATASETS

We evaluate STPFlow on four spatial multi-omics datasets that jointly provide spatial RNA and protein measurements. These datasets cover different tissues, species, and evaluation settings, including replicate prediction, cross-section prediction, and within-section regional prediction. For each dataset, RNA profiles are used as input and spatial protein profiles are treated as prediction targets. The detailed dataset information is summarized in Table 4.

Table 4: Summary of spatial RNA-to-protein prediction datasets used in this study. Detailed accession numbers, original references, and preprocessing information will be provided in the final version.

Dataset	Source / Platform	Training sample	Test sample	Evaluation setting	Resolution
Dataset 1	SPOTS	mouse spleen rep1	mouse spleen rep2	Replicate prediction	Spot-level
Dataset 2	spatialCOC	mouse thymus 1	mouse thymus 2	Cross-section prediction	Spot-level
Dataset 3	spatialglue	human lymph node 1	human lymph node 2	Cross-section prediction	Spot-level
Dataset 4	10x Xenium	human renal A	human renal B	Within-section regional prediction	Cell-level

Dataset 1: mouse spleen. The SPOTS mouse spleen data were obtained from the Gene Expression Omnibus (GEO) repository. This setting evaluates whether a model trained on one biological replicate can generalize to another replicate from the same tissue. **Accession**: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE198353.

Dataset 2: mouse thymus. This dataset contains spatial multi-omics measurements from mouse thymus tissue. This setting evaluates the model performance on across tissue sections from the same organ. **Accession:** https://zenodo.org/records/17655345.

Dataset 3: human lymph node. The data were obtained from the Gene Expression Omnibus (GEO) repository. This setting also evaluates cross-section RNA-to-protein prediction, **Accession**: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE26361.

Dataset 4: human renal. This dataset is from 10x Xenium official website. We split the full slice in half: range A and range B, testing the interpolation capability of the model on the same slice. Since the original dataset is large, we downsampled the original slice to 3000 cells in each half. **Accession**: https://www.10xgenomics.com/datasets/xenium-protein-ffpe-human-renal-carcinoma.

C EVALUATION METRICS

PCC (protein) evaluates how well each protein is predicted across spatial locations. It is computed by first calculating the Pearson correlation coefficient for each protein across all spots or cells, and then averaging over all proteins:

$$\mathrm{PCC}_{\mathrm{protein}} = \frac{1}{P}\sum_{p=1}^{P} \frac{\sum_{i=1}^{N}\left(\hat{Y}_{ip} - \overline{\hat{Y}}_{\cdot p}\right)\left(Y_{ip} - \overline{Y}_{\cdot p}\right)}{\sqrt{\sum_{i=1}^{N}\left(\hat{Y}_{ip} - \overline{\hat{Y}}_{\cdot p}\right)^2}\sqrt{\sum_{i=1}^{N}\left(Y_{ip} - \overline{Y}_{\cdot p}\right)^2}}. \tag{19}$$

PCC (cell) evaluates whether the predicted protein profile of each spot or cell matches the corresponding ground-truth protein profile. It is computed by calculating the Pearson correlation coefficient across proteins for each spot or cell, followed by averaging over all spots or cells:

$$\mathrm{PCC}_{\mathrm{cell}} = \frac{1}{N}\sum_{i=1}^{N} \frac{\sum_{p=1}^{P}\left(\hat{Y}_{ip} - \overline{\hat{Y}}_{i\cdot}\right)\left(Y_{ip} - \overline{Y}_{i\cdot}\right)}{\sqrt{\sum_{p=1}^{P}\left(\hat{Y}_{ip} - \overline{\hat{Y}}_{i\cdot}\right)^2}\sqrt{\sum_{p=1}^{P}\left(Y_{ip} - \overline{Y}_{i\cdot}\right)^2}}. \tag{20}$$

To evaluate whether the predicted protein expression preserves spatial morphology, we further compute the structural similarity index, SSIM, on the spatial expression map of each protein. For the predicted spatial map $\hat{\mathbf{I}}_p$ and the corresponding ground-truth spatial map $\mathbf{I}_p$, SSIM is defined as

$$\mathrm{SSIM}(\hat{\mathbf{I}}_p, \mathbf{I}_p) = \frac{\left(2\mu_{\hat{I}}\mu_I + C_1\right)\left(2\sigma_{\hat{I}I} + C_2\right)}{\left(\mu_{\hat{I}}^2 + \mu_I^2 + C_1\right)\left(\sigma_{\hat{I}}^2 + \sigma_I^2 + C_2\right)}, \tag{21}$$

where $\mu_{\hat{I}}$ and μ_I are the mean intensities of the predicted and ground-truth spatial maps, $\sigma_{\hat{I}}^2$ and σ_I^2 are their variances, $\sigma_{\hat{I}I}$ is their covariance, and C_1 and C_2 are small constants for numerical stability. The final SSIM score is averaged over all proteins:

$$\mathrm{SSIM} = \frac{1}{P}\sum_{p=1}^{P}\mathrm{SSIM}(\hat{\mathbf{I}}_p, \mathbf{I}_p). \tag{22}$$

Finally, SPCC measures rank-based spatial consistency between predicted and ground-truth protein expression. For each protein, we compute the Spearman rank correlation across all spots or cells, and then average over proteins:

$$\mathrm{SPCC} = \frac{1}{P}\sum_{p=1}^{P}\rho\left(\mathrm{rank}(\hat{\mathbf{Y}}_{\cdot p}), \mathrm{rank}(\mathbf{Y}_{\cdot p})\right), \tag{23}$$

where $\rho(\cdot,\cdot)$ denotes the Pearson correlation coefficient applied to ranked variables.

The results of all algorithms that have undergone data normalization have been restored to the raw counts data distribution during the evaluation process.

D ADDITIONAL EXPERIMENTAL RESULTS

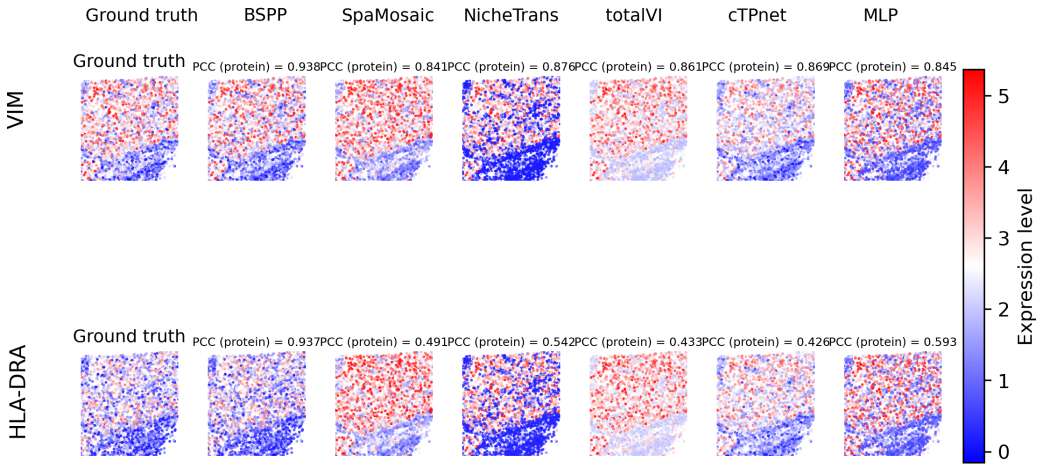

Figure 2: Spatial visualization of protein expression prediction on the Human Renal slice A $\rightarrow$ slice B benchmark. Rows correspond to two representative proteins, VIM and HLA-DRA. The first column shows the raw ADT ground truth, while the remaining columns show predictions from different methods. Panel titles report protein-wise PCC values computed across spatial spots. Color intensity denotes protein expression level. BSPP closely recovers the spatial expression patterns of proteins.