Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 18, No. 2, 2026

HiFi-ST: Histology-Conditioned Local Coordinate Modeling for Spatial Gene Expression Prediction

Lei Tang *

School of Mathematics and Computer Science, Yunnan Minzu University, Kunming, Yunnan, 650500, China
*** Corresponding author:** Lei Tang

Abstract: Spatial transcriptomics (ST) enables the characterization of spatial gene expression patterns in tissues, but its discrete measurement locations limit the availability of spatially resolved molecular information. To address this challenge, we propose HiFi-ST, a histology-conditioned framework for spatial gene expression prediction based on local coordinate modeling. HiFi-ST integrates multiscale hematoxylin and eosin (H&E) image representations with spatial coordinate information. Multiple local coordinate queries are performed within each spot neighborhood, and their expression predictions are combined through stochastic aggregation to generate spot-level estimates under observed expression supervision. Experiments on three public datasets, HER2+, cSCC, and Alex_NatGen, achieved mean gene-wise Pearson correlation coefficients (PCC (ACG)) of 0.3547, 0.3314, and 0.3385, respectively. Compared with previously published benchmark reference results, HiFi-ST achieved higher overall correlation on HER2+ and Alex_NatGen, while obtaining a slightly higher PCC (ACG) and lower prediction errors on cSCC. Gene-level spatial expression analyses further demonstrated that the predicted expression preserved selected spatial patterns associated with tissue morphology. In addition, spatial-domain clustering across six HER2+ sections achieved a mean adjusted Rand index of 0.3610 and a mean normalized mutual information of 0.3847, indicating the preservation of spatial information relevant to tissue-region organization. Overall, HiFi-ST provides an effective approach to histology-driven spatial gene expression prediction by integrating multiscale morphological features, local coordinate modeling, and spot-level aggregation.

Keywords: Histology-based Expression Prediction; Local Coordinate Modeling; Multiscale Histology; Spatial Transcriptomics.

1. Introduction

Spatial transcriptomics (ST) enables the investigation of gene expression within the spatial context of tissues, providing valuable information for understanding tumor heterogeneity and tissue microenvironments [1, 2]. However, current ST technologies typically measure gene expression at discrete spatial locations, limiting the availability of spatially resolved molecular information. In contrast, hematoxylin and eosin (H&E) images provide detailed morphological information at higher spatial resolution. Predicting spatial gene expression from histological images has therefore emerged as a promising computational approach for integrating tissue morphology with molecular information.

Recent studies have developed various deep learning methods for histology-based spatial gene expression prediction. HisToGene employs attention-based image modeling[3], whereas Hist2ST combines Transformer architectures with spatial graph learning[4]. THItoGene integrates multiple histological feature modeling mechanisms[5], while BLEEP and mclSTExp utilize cross-modal representation learning to associate histological features with gene expression[6, 7]. Although these methods adopt different architectures, their predictions are generally formulated at the spot level. Consequently, the relationship between local coordinate-level expression representations and observed spot-level measurements remains insufficiently explored.

Coordinate-based representations provide a flexible approach to modeling spatially varying signals[8]. However, introducing spatial coordinates alone does not explicitly connect predictions at nearby locations with the expression measurements obtained at individual spots. To address this issue, we propose HiFi-ST, a histology-conditioned local coordinate modeling framework for spatial gene expression prediction. HiFi-ST integrates multiscale H&E representations, spatial positional encoding, and conditional feature modulation. A shared coordinate decoder generates expression predictions from multiple local queries within each spot neighborhood, which are subsequently aggregated to produce the final spot-level prediction under observed expression supervision.

The main contribution of this study is a modeling strategy that connects local coordinate queries with spot-level supervision through stochastic aggregation. We evaluate HiFi-ST on three public datasets, HER2+, cSCC, and Alex_NatGen, using quantitative expression prediction metrics, gene-level spatial patterns, and spatial-domain clustering concordance. These evaluations examine the ability of the proposed framework to predict spatial gene expression while preserving biologically relevant spatial organization.

2. Methods

2.1. Data and Preprocessing

Three public spatial transcriptomics datasets were used to evaluate HiFi-ST: HER2+[9], cutaneous squamous cell carcinoma (cSCC)[10], and Alex_NatGen[11]. The HER2+ dataset contains 32 tissue sections with 35,328 spots, the cSCC dataset contains 12 sections with 12,382 spots, and Alex_NatGen comprises six breast cancer sections with 29,855 spots. HER2+ and cSCC were generated using the original spatial transcriptomics platform, whereas

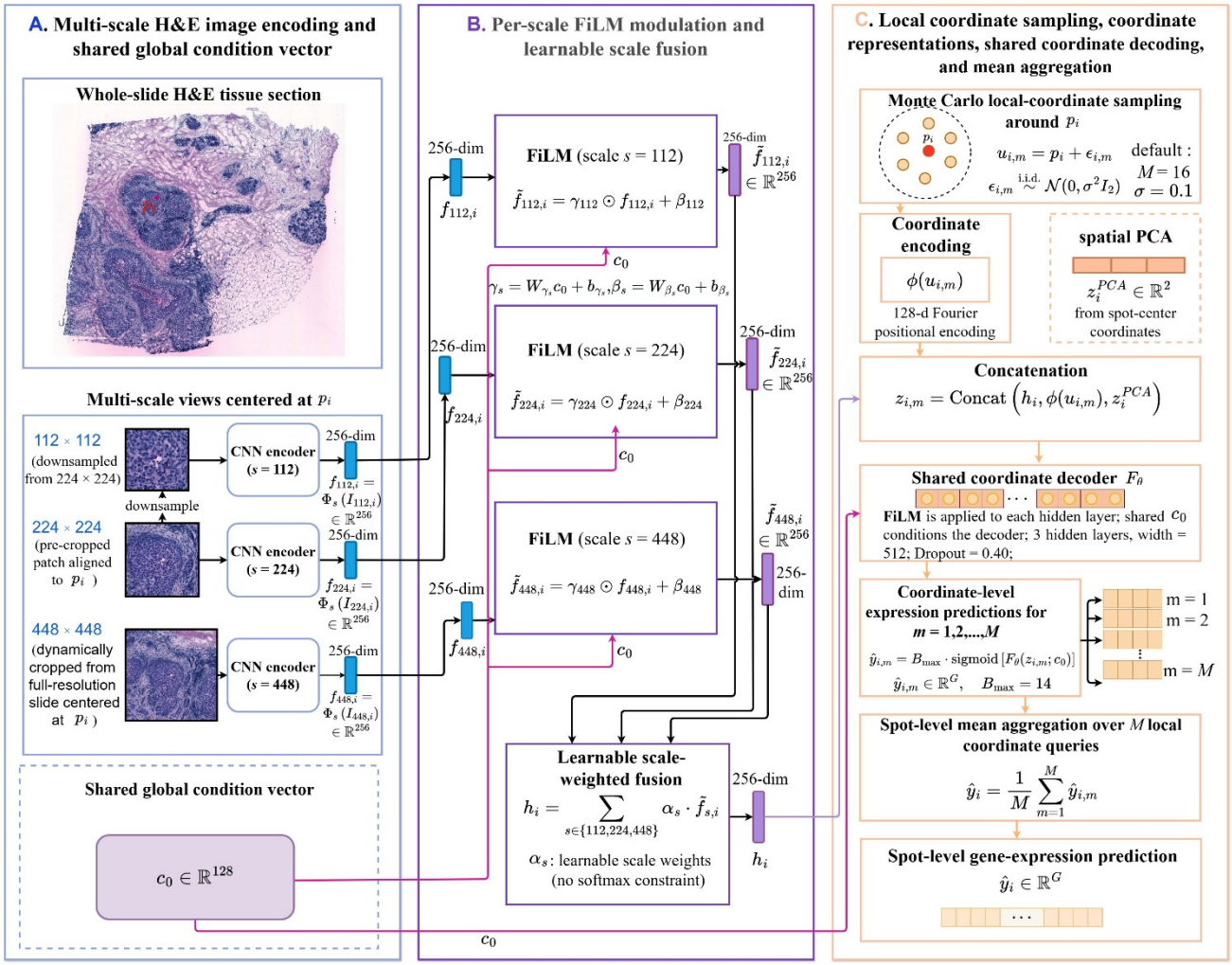


Figure 1. HiFi-ST framework: (A) multiscale H&E encoding; (B) FiLM modulation and feature fusion; and (C) local-coordinate decoding followed by spot-level mean aggregation.

Alex_NatGen was obtained using 10x Visium. Each dataset provides gene expression measurements, spatial coordinates, and corresponding hematoxylin and eosin (H&E) images.

For preprocessing, genes shared across tissue sections were identified, followed by highly variable gene (HVG) selection using Scanpy. Expression counts were normalized to a library size of 10,000 and log-transformed using log(1+x) for HVG selection, with a maximum of 1,000 candidate genes. After gene filtering and quality control, 785, 171, and 685 target genes were retained for HER2+, cSCC, and Alex_NatGen, respectively. The prediction targets were then reconstructed from the original counts of the retained genes using CPM-like normalization with a scaling factor of (10^6), followed by the log(1+x) transformation.. Data partitioning was performed at the tissue-section level to prevent spots from the same section from appearing in both training and test sets.

2.2. Histology-conditioned Representation

For each spatial transcriptomics spot, HiFi-ST constructs three H&E image views at resolutions of 112 × 112, 224 × 224, and 448 × 448 pixels to capture tissue morphology at different spatial scales. The 112 × 112 view is downsampled from the aligned 224 × 224 patch, while the 448 × 448 view is cropped from the original histological section. Three independent convolutional neural networks (CNNs) extract 256-dimensional feature representations from these views. A shared learnable 128-dimensional conditioning vector is then used to generate scale-specific feature-wise linear modulation (FiLM) parameters[12], enabling conditional adjustment of the histological features. The modulated representations are combined through learnable linear fusion coefficients initialized to 1/3 without softmax normalization. A regularization term encourages the coefficients to remain close to their initial values. The resulting fused representation provides histological context for subsequent local coordinate prediction, as illustrated in Fig. 1(A–B).

2.3. Coordinate Queries and Aggregation

To model local spatial variations, HiFi-ST generates multiple coordinate queries around each spot center. For the i-th spot with center coordinate p_i, the m-th query location is defined as:

$$u_{i,m} = p_i + \epsilon_{i,m}, \epsilon_{i,m} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2 I_2) \quad (1)$$

where $M = 16$ local queries and $\sigma = 0.1$ are used by default. The Gaussian perturbation defines a local sampling kernel in the model's coordinate system rather than the physical capture geometry of the spatial transcriptomics platform.

Each query coordinate is represented using a 128-dimensional Fourier positional encoding. Additionally, a two-dimensional spatial PCA representation is derived from the spot center coordinates, retaining both principal components without using expression information.

Table 1. Spot-level expression prediction performance of HiFi-ST and benchmark reference methods.

Dataset	Method	PCC (ACG)	PCC (HEG)	MSE	MAE
HER2+	HisToGene	0.0842	0.0711	0.5202	0.6422
	Hist2ST	0.1443	0.1849	0.5135	0.6087
	THItoGene	0.1736	0.2809	0.5012	0.5956
	BLEEP	0.1873	0.2909	0.6015	0.5824
	mclSTExp	0.2304	0.3866	0.5897	0.5813
	HiFi-ST	**0.3547**	**0.5178**	**0.3539**	**0.3451**
cSCC	HisToGene	0.0771	0.0919	0.6805	0.6234
	Hist2ST	0.1838	0.2175	0.6748	0.6107
	THItoGene	0.2373	0.2719	0.6546	0.6012
	BLEEP	0.2449	0.3122	0.5163	0.5399
	mclSTExp	0.3235	0.4261	0.4302	0.5208
	HiFi-ST	**0.3314**	**0.3831**	**0.2282**	**0.2178**
Alex_NatGen	HisToGene	0.0618	0.0984	0.4565	0.4973
	Hist2ST	0.1299	0.1784	0.3788	0.4492
	THItoGene	0.1384	0.2156	0.3672	0.4315
	BLEEP	0.1552	0.2825	0.2593	0.4050
	mclSTExp	0.1949	0.3611	0.2329	0.3897
	HiFi-ST	**0.3385**	**0.5873**	**0.3115**	**0.3032**

Note: Baseline values are taken from the previously published mclSTExp benchmark, whereas HiFi-ST results are obtained from the present study. These results provide a reference-level comparison rather than a strictly paired evaluation under identical protocols.

The Fourier encoding, spatial PCA representation, and fused histological features are concatenated and passed to a shared coordinate decoder. The decoder contains three hidden layers with FiLM-based modulation and generates bounded gene expression estimates using a sigmoid output transformation.

The final spot-level prediction is obtained by averaging the local coordinate-level predictions:

$$\hat{y}_i = \frac{1}{M}\sum_{m=1}^{M} \hat{y}_{i,m} \quad (2)$$

The aggregated prediction is supervised using the observed spot-level gene expression. This formulation connects local coordinate modeling with spot-level measurements without requiring sub-spot expression annotations.

2.4. Training and Evaluation

HiFi-ST is trained using a composite objective combining regression and correlation losses. Before loss computation, predicted and observed expression values are clipped to [0, 14] and further transformed using log(1+x). The overall training objective is defined as:

$$\mathcal{L} = \mathcal{L}_{MSE} + 0.5\mathcal{L}_{PCC} + 0.01\mathcal{L}_{scale} + \lambda_{HEG}\mathcal{L}_{HEG} \quad (3)$$

where $\mathcal{L}_{MSE}$ denotes the mean squared error, $\mathcal{L}_{PCC}$ is the negative Pearson correlation computed over flattened batch-level predictions, and $\mathcal{L}_{scale}$ regularizes the learnable scale-fusion coefficients toward 1/3. For cSCC, an auxiliary highly expressed gene (HEG) prediction head introduces an additional regression–correlation loss for the 50 most highly expressed genes selected within each training batch, with $\lambda_{HEG} = 1$. This auxiliary loss is disabled for HER2+ and Alex_NatGen ($\lambda_{HEG} = 0$), and only the main prediction head is used during evaluation.

The model is optimized using Adam with a batch size of 256, weight decay of 1×10^{-5}, and a maximum of 100 epochs. The initial learning rates are 5×10^{-5} for HER2+ and Alex_NatGen and 2×10^{-5} for cSCC. The best model is selected according to validation PCC (ACG), with early stopping after 30 epochs without improvement.

Prediction performance is evaluated using PCC (ACG), PCC (HEG), MSE, and MAE. All metrics are calculated from aggregated spot-level predictions, and evaluation errors are computed without an additional logarithmic transformation. PCC (HEG) is calculated as the mean gene-wise PCC over the evaluable genes among the 50 target genes with the highest observed mean expression.

3. Experimental Results

3.1. Overall Spot-level Expression Prediction

We evaluated HiFi-ST on three public datasets, HER2+, cSCC, and Alex_NatGen, and compared its performance with five existing methods: HisToGene, Hist2ST, THItoGene, BLEEP, and mclSTExp. Four evaluation metrics were employed: PCC (ACG), PCC (HEG), mean squared error (MSE), and mean absolute error (MAE). The results are

summarized in Table 1.

On HER2+, HiFi-ST achieved the highest PCC (ACG) and PCC (HEG), together with the lowest MSE and MAE among the reported results. On cSCC, HiFi-ST obtained a slightly higher PCC (ACG) and substantially lower prediction errors, although its PCC (HEG) was lower than that of mclSTExp. On Alex_NatGen, HiFi-ST achieved higher correlation values and lower MAE, whereas mclSTExp obtained a lower MSE. These results indicate that HiFi-ST provides competitive spot-level expression prediction performance across different tissue datasets, with performance advantages varying across evaluation metrics.

3.2. Gene-wise Correlation Distributions

To further evaluate gene-level prediction performance, we analyzed the distributions of gene-wise Pearson correlation coefficients (PCC) between predicted and observed expression across tissue sections. Figures 2–4 present the results for HER2+, cSCC, and Alex_NatGen, respectively. Compared with the displayed baseline methods, HiFi-ST exhibited higher median gene-wise PCC values in multiple tissue sections, suggesting improved recovery of spatial expression patterns for many evaluated genes. Nevertheless, prediction performance varied across datasets and tissue sections, indicating that the improvements were not uniform across all genes. These findings complement the overall evaluation in Table 1 and demonstrate the ability of HiFi-ST to capture gene-level spatial expression variations.

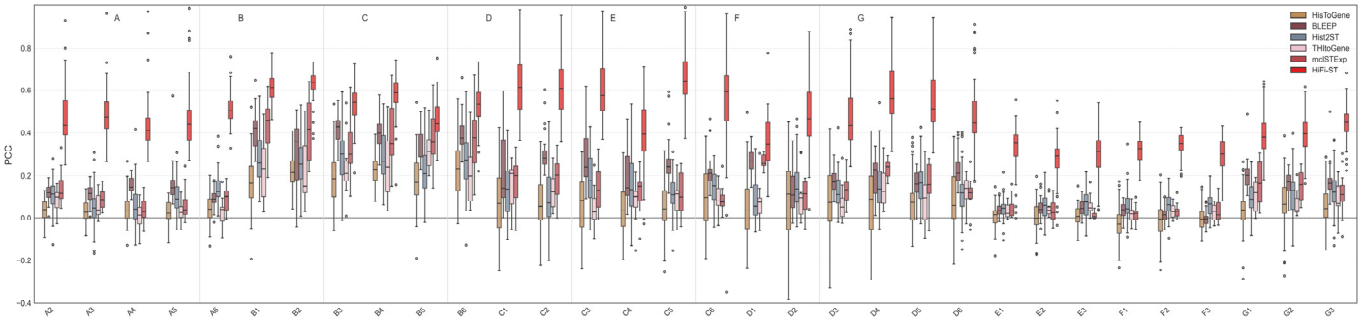


Figure 2. Distribution of gene-wise PCC values across different methods in the HER2+ dataset.

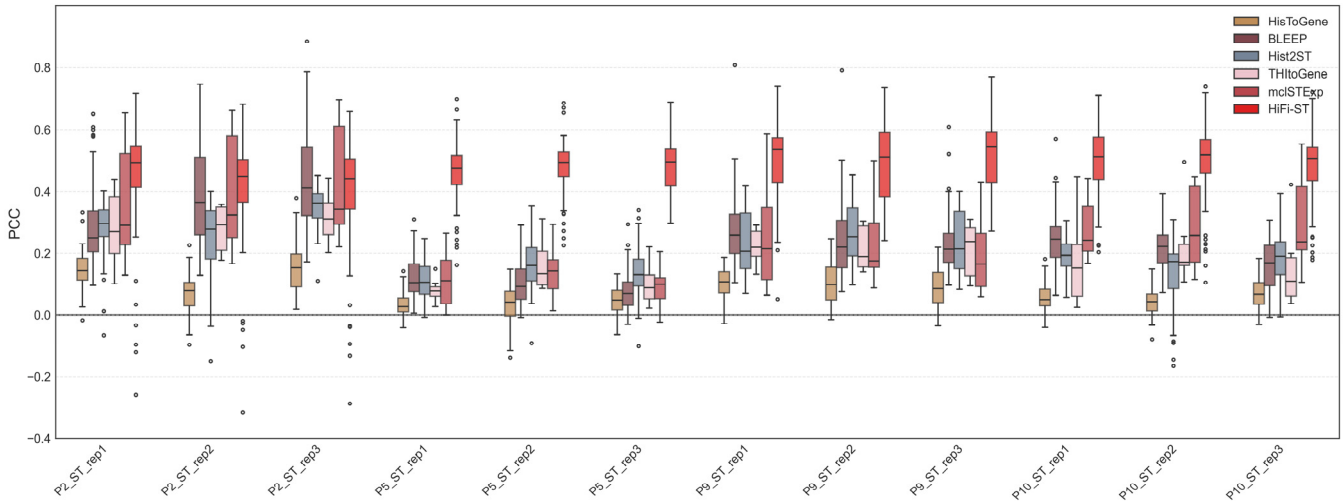


Figure 3. Distribution of gene-wise PCC values across different methods in the cSCC dataset.

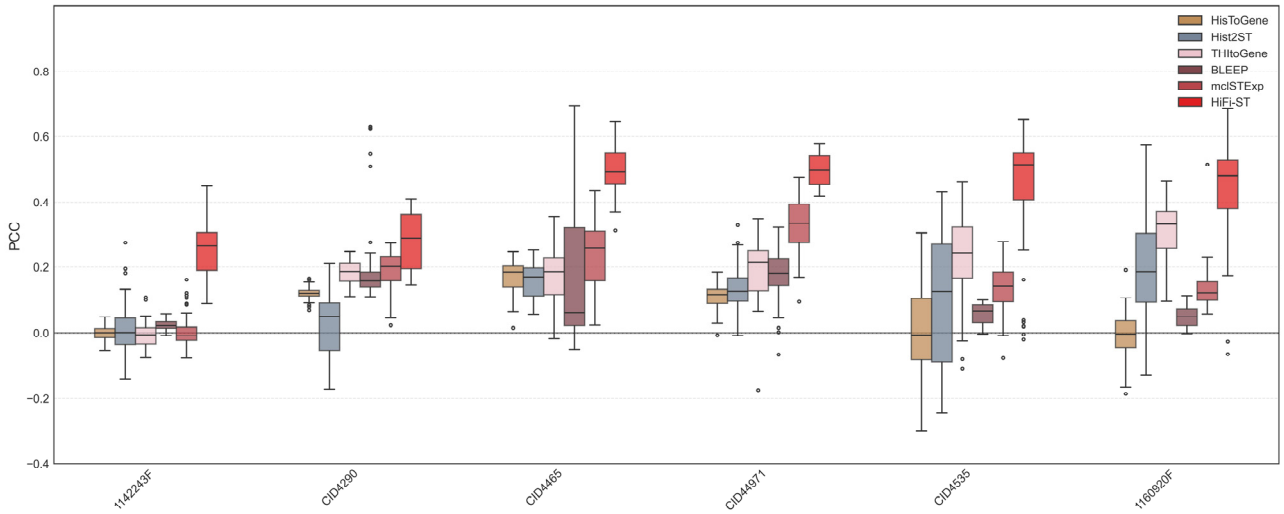


Figure 4. Distribution of gene-wise PCC values across different methods in the Alex_NatGen dataset.

3.3. Representative Gene-level Spatial Patterns

To assess the spatial consistency of gene expression predictions, we visualized representative genes in the HER2+ dataset, as shown in Figure 5. The predicted expression patterns of genes such as PSMB3, FN1, and IGKC were broadly consistent with the corresponding observed spatial distributions. Among these genes, FN1 is associated with extracellular matrix organization and tumor stromal remodeling, while IGKC is related to humoral immune responses and tumor-infiltrating plasma cells. These representative results suggest that HiFi-ST can preserve selected spatial expression patterns associated with biologically relevant tissue characteristics. However, the agreement varies across genes, and the visual comparisons should be interpreted together with the corresponding gene-wise PCC values rather than as evidence of uniform prediction accuracy.



Figure 5. Spatial expression prediction results for representative genes in the HER2+ dataset. The reported PCC values are calculated for individual gene–section combinations.

3.4. Spatial-domain Clustering Concordance

To evaluate the preservation of tissue-level spatial organization, spatial-domain clustering was performed on six HER2+ sections (B1, C1, D1, E1, F1, and G2) with expert pathological annotations. A unified PCA–KMeans pipeline was applied to observed and predicted expression using identical annotated spots, and clustering agreement was assessed using the adjusted Rand index (ARI) and normalized mutual information (NMI). As shown in Table 2, HiFi-ST achieved a mean ARI of 0.3610 and a mean NMI of 0.3847, exceeding the results of the compared prediction methods. At the individual-section level, HiFi-ST achieved the highest ARI in five of six sections and the highest NMI in all six sections among the prediction methods. The spatial visualizations in Figure 6 further illustrate the correspondence between predicted spatial domains and pathological annotations. These results suggest that HiFi-ST preserves spatial expression information relevant to tissue-region organization.

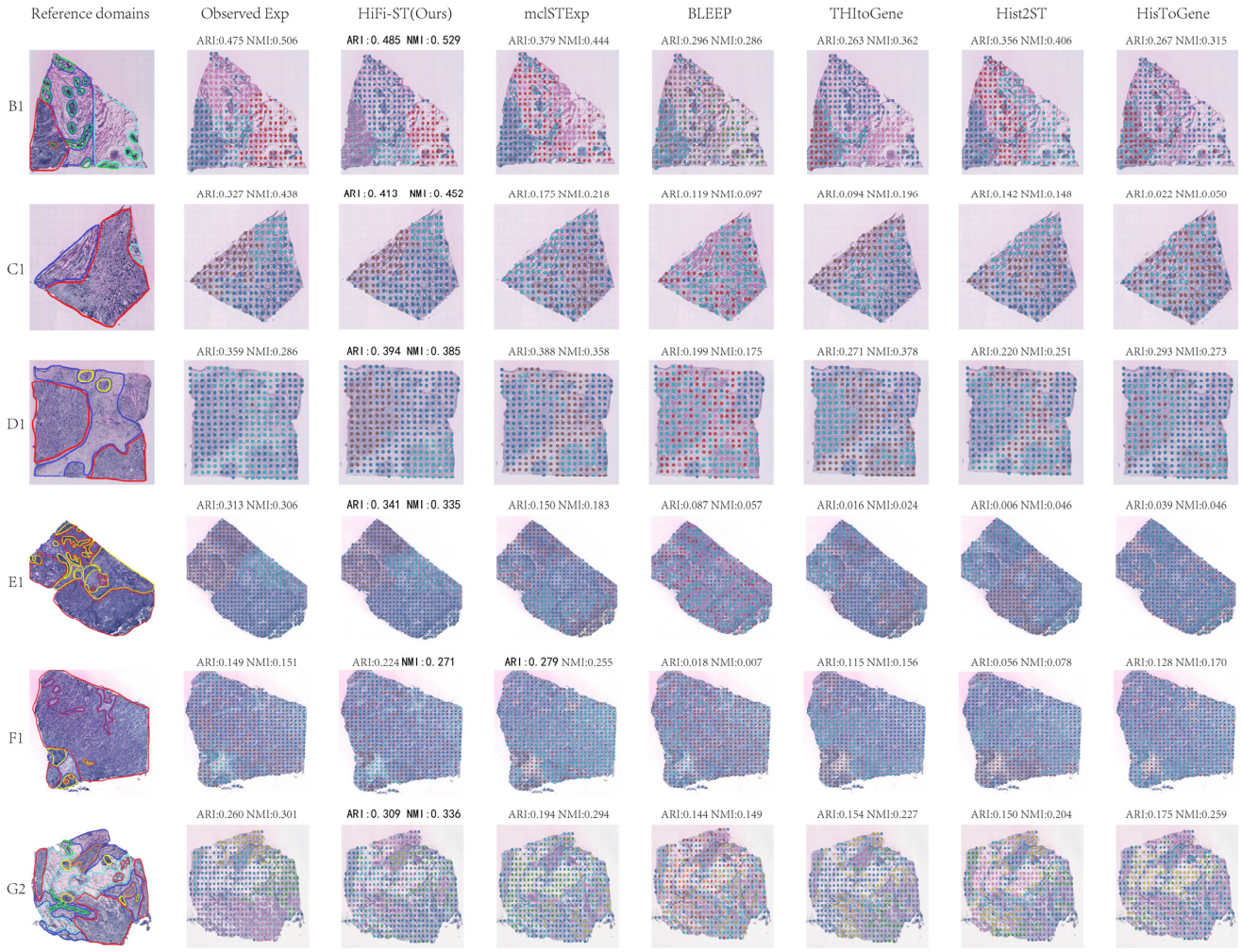


Figure 6. Spatial-domain clustering results across six HER2+ sections. Expert pathological annotations are compared with clustering results derived from observed expression and different prediction methods using a unified PCA–KMeans pipeline.

Table 2. Spatial-domain clustering concordance across six HER2+ sections.

Expression Representation/Method	Mean ARI	Median ARI	Mean NMI	Median NMI
Observed Exp	0.3138	0.3200	0.3313	0.3035
HiFi-ST(Ours)	**0.3610**	**0.3675**	**0.3847**	**0.3605**
mclSTExp	0.2608	0.2365	0.2920	0.2745
BLEEP	0.1438	0.1315	0.1285	0.1230
THItoGene	0.1522	0.1345	0.2238	0.2115
Hist2ST	0.1550	0.1460	0.1888	0.1760
HisToGene	0.1540	0.1515	0.1855	0.2145

Note: Observed Exp represents clustering based on measured expression and serves as an empirical reference rather than a theoretical performance upper bound.

4. Discussion and Conclusion

In this study, we proposed HiFi-ST, a histology-conditioned framework that integrates multiscale H&E representations, local coordinate queries, and stochastic aggregation for spatial gene expression prediction. Unlike conventional spot-centered regression approaches, HiFi-ST explicitly connects local coordinate-level modeling with observed spot-level expression through aggregation. Experiments on three public datasets demonstrated competitive prediction performance compared with previously published benchmark reference results. HiFi-ST achieved favorable results across all four metrics on HER2+, while its advantages on cSCC and Alex_NatGen varied across correlation- and error-based metrics. Furthermore, gene-level spatial expression analyses and spatial-domain clustering indicated that the predicted expression preserved selected spatial patterns associated with tissue morphology and pathological organization.

Nevertheless, several limitations remain. The model is supervised only by observed spot-level expression, and its coordinate-level predictions cannot be interpreted as independently validated sub-spot expression fields. In addition, the benchmark comparisons are not strictly paired under identical experimental protocols, and further validation across datasets and spatial transcriptomics platforms is needed. Future work will focus on improving the reliability of local coordinate modeling and conducting standardized evaluations. Overall, HiFi-ST provides a promising computational approach for histology-driven spatial gene

expression prediction by combining multiscale morphological information with local spatial modeling and spot-level supervision.

References

[1] Ståhl, P. L., Salmén, F., Vickovic, S., et al. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science, 353(6294), 78–82. https://doi.org/10.1126/science.aaf2403.

[2] Burgess, D. J. (2019). Spatial transcriptomics coming of age. Nature Reviews Genetics, 20(6), 317. https://doi.org/10.1038/s41576-019-0133-8.

[3] Pang, M., Su, K., & Li, M. (2021). Leveraging information in spatial transcriptomics to predict super-resolution gene expression from histology images in tumors. bioRxiv. https://doi.org/10.1101/2021.11.28.470212.

[4] Zeng, Y., Wei, Z., Yu, W., et al. (2022). Spatial transcriptomics prediction from histology jointly through Transformer and graph neural networks. Briefings in Bioinformatics, 23(5), bbac297. https://doi.org/10.1093/bib/bbac297.

[5] Jia, Y., Liu, J., Chen, L., et al. (2023). THItoGene: a deep learning method for predicting spatial transcriptomics from histological images. Briefings in Bioinformatics, 25. https://doi.org/10.1093/bib/bbad446.

[6] Xie, R., Pang, K., Chung, S., et al. (2023). Spatially resolved gene expression prediction from histology images via bi-modal contrastive learning. Advances in Neural Information Processing Systems, 36, 70626–70637.

[7] Min, W., Shi, Z., Zhang, J., et al. (2024). Multimodal contrastive learning for spatial gene expression prediction using histology images. Briefings in Bioinformatics, 25(6). https://doi.org/10.1093/bib/bbae469.

[8] Tancik, M., Srinivasan, P., Mildenhall, B., et al. (2020). Fourier features let networks learn high frequency functions in low dimensional domains. Advances in Neural Information Processing Systems, 33, 7537–7547.

[9] Andersson, A., Bergenstråhle, J., Asp, M., et al. (2020). Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography. Communications Biology, 3(1), 565. https://doi.org/10.1038/s42003-020-01247-y.

[10] Ji, A. L., Rubin, A. J., Thrane, K., et al. (2020). Multimodal Analysis of Composition and Spatial Architecture in Human Squamous Cell Carcinoma. Cell, 182(2), 497–514.e22. https://doi.org/10.1016/j.cell.2020.05.032.

[11] Wu, S. Z., Al-Eryani, G., Roden, D. L., et al. (2021). A single-cell and spatially resolved atlas of human breast cancers. Nature Genetics, 53(9), 1334–1347. https://doi.org/10.1038/s41588-021-00911-1.

[12] Perez, E., Strub, F., De Vries, H., et al. (2018). FILM: Visual reasoning with a general conditioning layer. Proceedings of the AAAI Conference on Artificial Intelligence.