WHIOCE PUBLISHING PTE. LTD.
PROVIDING FIRST-CLASS SCIENTIFIC INFORMATION FOR TOP SCHOLARS

2026 Volume 7, Issue 2
Cell Biology Research
ISSN:2529-7627

VirtualST: Morphology- and Cell-Composition-Conditioned Diffusion for Predicting Spatial Gene Expression from H&E Images

Yuping Liang, Siwen Xu*
School of Medical Information and Engineering, Guangdong Pharmaceutical University, Guangzhou 510006, Guangdong, China

**Corresponding author:* Siwen Xu, siwxu@gdpu.edu.cn

Abstract: Spatial transcriptomics (ST) measures gene expression while preserving tissue spatial information, but its high cost and limited throughput restrict large-scale application. In contrast, hematoxylin and eosin (H&E)-stained histology images are widely available in routine pathology. We propose VirtualST, a conditional diffusion model for predicting spot-level spatial gene expression from H&E images. The model integrates histological features extracted by a pathology foundation model, local cell-type composition derived from nuclei segmentation, and spatial information from neighboring spots. VirtualST was evaluated on multiple cancer cohorts from HEST-bench and compared with representative histology-to-expression prediction methods. The results showed that VirtualST achieved competitive performance across different cancer types and performed well for representative colorectal cancer marker genes. These findings suggest that VirtualST provides an effective approach for spatial gene-expression prediction from routine histology images.

Keywords: spatial transcriptomics; spatial gene expression prediction; computational pathology; conditional diffusion; cell composition; spatial graph refinement

Online publication: June 26, 2026

1. Introduction

Spatial transcriptomics (ST) enables the measurement of gene expression while preserving tissue spatial information, providing important insights into tissue organization and tumor microenvironments[1]. However, its high cost and limited throughput restrict its routine application[2]. In contrast, hematoxylin and eosin (H&E)-stained histology images are widely available in clinical pathology. Predicting spatial gene expression from H&E images therefore provides a practical way to obtain molecular information from routinely collected tissue samples.

Several methods have been developed for histology-based spatial gene-expression prediction. ST-Net uses convolutional image features[3], while HisToGene[4] and Hist2ST[5] incorporate attention or graph-based structures. BLEEP further introduces contrastive learning to align histological and gene-expression features[6]. Although these methods have achieved promising results, most of them use discriminative prediction frameworks and do not simultaneously consider histological morphology, cellular composition, and spatial neighborhood information.

To address these limitations, we propose VirtualST, a conditional diffusion framework for predicting spot-level spatial gene expression from H&E images. The model performs diffusion directly in normalized gene-expression space and combines morphology features extracted by a frozen pathology foundation model with cell-type composition obtained from nuclei segmentation. A spatial graph-refinement module is further introduced to incorporate information from neighboring spots.

The main contributions of this study are as follows:

(1) A conditional diffusion framework that integrates histological morphology and local cell composition for spatial gene-expression prediction.

(2) A spatial graph-refinement module that uses neighboring spot predictions to improve spatial expression estimation.

(3) A multi-cancer evaluation comparing VirtualST with representative histology-to-expression prediction methods using consistent data splits and evaluation metrics.

2. Method

2.1. Overview

As shown in **Figure 1**, the overall workflow includes spot-centered patch extraction, histology feature encoding, spatially informed expression prediction, and generation of the spatial gene-expression matrix. Given an H&E image patch corresponding to a spatial transcriptomics spot, VirtualST predicts a normalized G-dimensional gene-expression vector, where G is the number of target genes.

VirtualST combines morphology features extracted by a frozen Phikon-v2 encoder[7], local cell-composition information obtained using HoVer-Net[8], a conditional diffusion network, and a spatial graph-refinement module. The diffusion process is performed directly in normalized gene-expression space. During training, the model learns to recover clean expression profiles from noisy inputs, while during inference, predictions are generated from random Gaussian noise. Neighboring spot information is then incorporated through the graph-refinement module, which is jointly trained with the gene-prediction network.

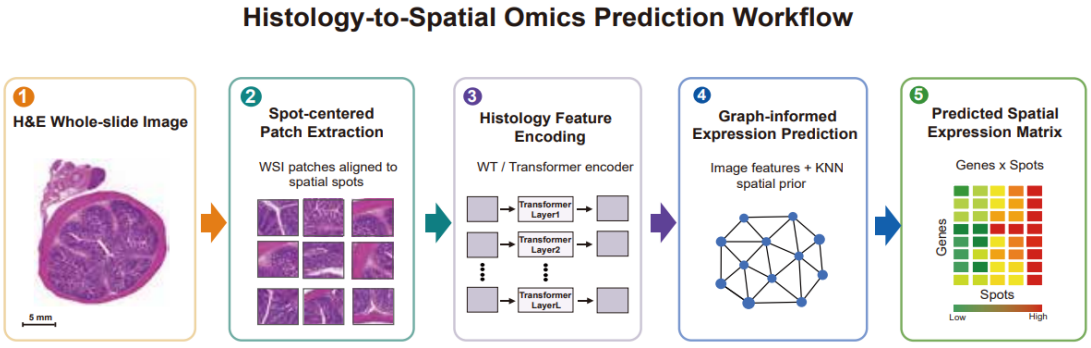

Figure 1. Overall workflow for histology-based spatial gene-expression prediction.

2.2. Morphology- and Cell-Composition Conditioning

For each spatial spot, the corresponding H&E image patch is encoded using a frozen Phikon-v2 pathology foundation model to obtain morphological features. HoVer-Net is used to segment and classify nuclei within the same patch, and the proportions of different cell types are calculated to represent local cellular composition. These two types of features are combined and used as conditional information for the gene-expression diffusion network.

2.3. Conditional Diffusion in Gene-Expression Space

VirtualST performs diffusion directly in normalized gene-expression space[9]. Given a clean expression vector y_0, Gaussian noise is gradually added at diffusion step *t*:

$$y_t = \sqrt{\bar{\alpha}_t}\, y_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon, \qquad \epsilon \sim N(0, I) \tag{1}$$

where $\bar{\alpha}_t$ is determined by the predefined noise schedule.

The reverse process is conditioned on the morphology- and cell-composition features c:

$$p_\theta(y_{t-1} | y_t, c) = N\left(\mu_\theta(y_t, t, c), \sigma_t^2 I\right) \tag{2}$$

where μ_θ is predicted by the conditional denoising network. During training, the model learns to recover the clean expression profile from noisy inputs. During inference, sampling starts from Gaussian noise $y_T \sim N(0, I)$, and the expression vector is gradually generated through the reverse diffusion process.

2.4. Zero-Inflated Output Gating and Bounded Normalization

To account for the large number of zero values in spatial gene-expression data, VirtualST applies a hard gating operation to the predicted expression vector. A threshold τ is calculated from the statistics of the current batch, and values below this threshold are set to zero:

$$\tilde{y}_i = \begin{cases} y_{i,} y_i > \tau \\ 0, y_i \le \tau \end{cases} \tag{3}$$

The gated output is then transformed and restricted to the normalized expression range:

$$\hat{y} = clip(a\tilde{y} + b, 0, 1) \tag{4}$$

where a and b are fixed affine parameters, set to 1 and 0, respectively. This operation encourages sparse predictions while ensuring that the final expression values remain within$[0,1]$.

2.5. Integrated Spatial Graph Refinement

To incorporate spatial context, a graph is constructed according to the coordinates of spatial spots. Each spot is connected to its $k_{nn} = 6$ nearest neighbors. For spot i, the average prediction of its neighboring spots is calculated as

$$m_i = \frac{1}{|N_i|} \sum_{j \in N_i} \hat{\mathrm{y}}_j \tag{5}$$

where N_i denotes the neighbor set of spot i, and $\hat{\mathrm{y}}_j$ is the predicted expression vector of neighboring spot *j*.

The prediction of the target spot and its neighborhood representation are then combined through a learnable refinement function:

$$\mathrm{y}_i^{\mathrm{ref}} = g_\Phi\left(\left[\hat{\mathrm{y}}_i \| \mathrm{m}_i\right]\right) \tag{6}$$

where $[\cdot\|\cdot]$denotes feature concatenation and $g_\phi$is the graph-refinement module. The module is integrated into the forward process and jointly optimized with the gene-expression prediction network.

2.6. Inference and Multi-Sample Prediction

During inference, the generation process starts from independent Gaussian noise without using the observed gene-expression values of the test spots:

$$y_T^{(s)} \sim N(0, I), \qquad s = 1, \ldots, S \tag{7}$$

where s denotes the s-th sampling run. The expression profile is generated through the reverse diffusion process

conditioned on the histological and cell-composition features. To reduce the randomness of diffusion sampling, VirtualST generates $S=10$ predictions for each spot and calculates their average:

$$\hat{y}_i = \frac{1}{s}\sum_{s=1}^{S} y_i^{(S)}, \quad S=10 \tag{8}$$

where $y_i^{(s)}$ is the prediction of spot i from the s-th sampling run. The averaged result is used as the final predicted expression profile.

2.7. Training Objective and Optimization

At each training iteration, a diffusion step t is randomly sampled and Gaussian noise is added to the normalized expression vector. The denoising network predicts the clean expression profile:

$$\hat{y}_0 = f_\theta\left(y_t, t, c\right) \tag{9}$$

where c represents the morphology- and cell-composition conditions.

The model is trained using the mean squared error between the predicted and observed expression profiles:

$$L_{diff} = \frac{1}{B}\sum_{i=1}^{B}\left\|\hat{y}_{0,i} - y_{0,i}\right\|_2^2 \tag{10}$$

where B is the batch size.

The conditional diffusion network and spatial graph-refinement module are jointly optimized, while the Phikon-v2 pathology encoder remains frozen during training.

3. Experiments and Results

3.1. Datasets and Preprocessing

Experiments were conducted using multiple cancer cohorts from HEST-bench[2]. Each cohort contained paired H&E image patches, spatial coordinates, and spot-level gene-expression profiles. A leave-one-sample-out cross-validation strategy was adopted, in which one sample was used for testing and the remaining samples were used for training.

Gene-expression values were first transformed using $log(1+x)$ normalization and then scaled to the range $[0,1]$ for each sample. Only genes shared across samples within the same cohort were retained. Cohort-specific marker genes were further selected for gene-level evaluation.

For each spatial spot, the corresponding H&E patch was processed using Phikon-v2 to extract morphology features. HoVer-Net was applied to segment and classify nuclei, and the proportions of seven nucleus classes were calculated as cell-composition features. These features were precomputed before model training.

3.2. Evaluation Protocol

Model performance was evaluated using the mean gene-wise Pearson correlation coefficient (PCC). Gene-wise PCC was also examined for representative colorectal cancer marker genes. VirtualST was compared with ST-Net, BLEEP, and CMRCNet[3, 6, 10] under the same data splits and gene sets. For all methods, the model from the final training epoch was used for evaluation. VirtualST generated ten independent predictions and used their average as the final result.

3.3. Main Results

Table 1 compares VirtualST with three representative methods across four selected cancer cohorts. VirtualST achieved the highest mean gene-wise PCC on IDC, LYMPH_IDC, COAD, and PAAD. Overall, these results indicate that VirtualST provides competitive and consistent prediction performance across different cancer types.

Table 1. Mean gene-wise PCC across selected cancer cohorts.

Datasets	ST-Net	BLEEP	CMRCNet	VirtualST
IDC	0.3815	0.3870	0.3668	**0.4238**
LYMPH_IDC	0.3754	0.3580	0.3902	**0.4169**
COAD	0.3470	0.3020	0.2830	**0.3840**
PAAD	0.2242	0.2114	0.1830	0.2340

3.4. Effect of Spatial Graph Refinement

To evaluate the contribution of spatial neighborhood information, we compared the VirtualST backbone with the graph-enabled model. As shown in **Table 2**, spatial graph refinement improved the mean gene-wise PCC across all four selected cohorts. The largest improvement was observed on LYMPH_IDC, followed by IDC and PAAD, while a smaller gain was obtained on COAD. These results suggest that incorporating neighboring spot information can improve spatial gene-expression prediction.

Table 2. Effect of spatial graph refinement on mean gene-wise PCC.

Datasets	Backbone	Graph-enabled E2E	ΔPCC
IDC	0.3857	**0.4238**	+0.0381
LYMPH_IDC	0.3310	**0.4169**	+0.0859
COAD	0.3800	**0.3840**	+0.0040
PAAD	0.2024	**0.2340**	+0.0316

3.5. Performance on Representative Marker Genes

To further evaluate gene-level performance on biologically relevant targets, we compared the gene-wise PCC of VirtualST with ST-Net, BLEEP, and CMRCNet[3, 6, 10] for seven representative colorectal marker genes[11, 12]. As shown in **Figure 2**, VirtualST achieved the highest PCC for LGR5, CEACAM5, CEACAM6, GPX2, and MKI67. Its performance on OLFM4 was comparable to that of ST-Net, whereas ST-Net achieved a higher PCC for REG4. These results indicate that VirtualST provides strong prediction performance for multiple representative colorectal markers, although its relative advantage remains gene dependent.

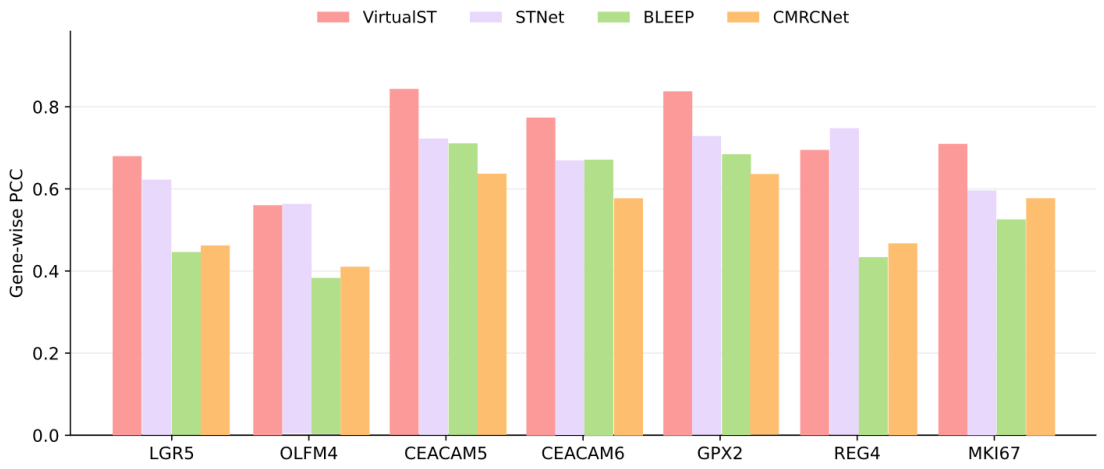

Figure 2. Gene-wise PCC comparison for representative colorectal marker genes.

4. Conclusion

We presented VirtualST, a conditional diffusion framework for predicting spot-level spatial gene expression from H&E images. The model combines histological morphology, local cell composition, and spatial neighborhood information. Experiments on multiple cancer cohorts showed that VirtualST achieved competitive performance, with particularly strong results on several cancer types. The spatial graph-refinement module further improved prediction performance in the evaluated cohorts. VirtualST also showed good gene-level performance for representative colorectal marker genes. These results suggest that combining conditional diffusion with morphological, cellular, and spatial information is effective for histology-based spatial gene-expression prediction. Further evaluation on larger and independent datasets is needed to assess the broader applicability of the proposed method.

Disclosure statement

The author declares no conflict of interest.

References

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

[2] Jaume G, Doucet P, Song AH, et al., 2024, HEST-1k: A dataset for spatial transcriptomics and histology image analysis. Advances in Neural Information Processing Systems, 37. DOI:10.52202/079017-1704.

[3] He B, Bergenstråhle L, Stenbeck L, et al., 2020, Integrating spatial gene expression and breast tumour morphology via deep learning. Nature Biomedical Engineering, 4: 827-834. DOI:10.1038/s41551-020-0578-x.

[4] Pang M, Su K, Li M, 2021, Leveraging information in spatial transcriptomics to predict super-resolution gene expression from histology images in tumors. bioRxiv. DOI:10.1101/2021.11.28.470212.

[5] 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. DOI:10.1093/bib/bbac297.

[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.

[7] Filiot A, Jacob P, Mac Kain A, et al., 2024, Phikon-v2, a large and public feature extractor for biomarker prediction. arXiv preprint, arXiv:2409.09173. DOI:10.48550/arXiv.2409.09173.

[8] Graham S, Vu QD, Raza SEA, et al., 2019, HoVer-Net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. Medical Image Analysis, 58: 101563. DOI:10.1016/j.media.2019.101563.

[9] Ho J, Jain AN, Abbeel P, 2020, Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems, 33: 6840-6851.

[10] Liu J, Eckstein M, Wang Z, et al., 2026, Spatial transcriptomics expression prediction from histopathology based on cross-modal mask reconstruction and contrastive learning. Medical Image Analysis, 108: 103889. DOI:10.1016/j.media.2025.103889.

[11] van der Flier LG, Haegebarth A, Stange DE, et al., 2009, OLFM4 is a robust marker for stem cells in human intestine and marks a subset of colorectal cancer cells. Gastroenterology, 137(1): 15-17. DOI:10.1053/j.gastro.2009.05.035.

[12] Becker WR, Nevins SA, Chen DC, et al., 2022, Single-cell analyses define a continuum of cell state and composition changes in the malignant transformation of polyps to colorectal cancer. Nature Genetics, 54: 985-995. DOI:10.1038/s41588-022-01088-x.