Proceedings of Anticancer Research, 2026, Volume 10, Issue 4
http://ojs.bbwpublisher.com/index.php/PAR
Online ISSN: 2208-3553
Print ISSN: 2208-3545

Inferring Tumor Spatial Gene Expression from Routine H&E Images: Pan-Cancer Validation and Spatial Profiling of Primary Breast Tumors and Lymph Node Metastases

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) enables gene expression profiling while preserving tissue architecture, but its high cost and limited throughput constrain large-scale applications. This study evaluated the biological utility of VirtualST, a conditional diffusion model for inferring spatial gene expression from H&E images, in colon adenocarcinoma (COAD), primary invasive ductal carcinoma of the breast (IDC), and breast cancer lymph node metastasis (LYMPH_IDC). Twelve samples with paired H&E images and ground-truth ST data were evaluated using leave-one-sample-out cross-validation. Inference started from pure noise, with ground-truth expression used only for post hoc evaluation, and the final prediction for each sample was obtained by averaging ten sampling runs. The mean gene-wise Pearson correlation coefficients (PCCs) were 0.467, 0.439, and 0.496 for COAD, IDC, and LYMPH_IDC, respectively, with better reconstruction of tumor epithelial, secretory, and proliferative programs. Using ground-truth ST epithelial scores from the held-out samples as an independent molecular reference, the predicted and ground-truth scores achieved within-sample Pearson correlations of 0.79 and 0.66 in IDC and LYMPH_IDC, respectively, with corresponding hotspot Dice coefficients of 0.64 and 0.44. These results indicate that VirtualST can recover within-sample spatial epithelial signals but is not suitable for absolute quantification across tissues or direct pathological region detection.

Keywords: Spatial gene expression prediction; H&E histology; Conditional diffusion; Breast cancer; Lymph node metastasis

Online publication: August 11, 2026

1. Introduction

Tumor tissues consist of malignant cells, immune cells, fibroblasts, and other components of the tumor

microenvironment. The non-random spatial distribution of these cell populations creates substantial spatial heterogeneity and is closely associated with tumor growth, local invasion, and treatment response. Spatial transcriptomics (ST) captures gene expression information while preserving tissue coordinates, providing an important tool for characterizing tumor cell compartments and their microenvironments [1]. However, the high experimental cost, complex sample-processing procedures, and limited throughput of ST hinder its direct application to large clinical cohorts. In contrast, routine hematoxylin and eosin (H&E)-stained sections are inexpensive, readily accessible, and extensively accumulated in clinical pathology workflows. Inferring spatial gene expression from H&E images may therefore offer a cost-effective approach to scaling up spatial molecular analyses.

In recent years, several computational methods have been developed to infer spatial gene expression from H&E images. ST-Net uses convolutional neural networks to extract histological image features and predict expression levels at different spatial locations [2]. HisToGene further introduces an attention mechanism to enhance the modeling of relationships among tissue regions [3], whereas Hist2ST combines a Transformer with a graph neural network to jointly incorporate morphological features and spatial neighborhood information [4]. These studies demonstrate that tissue morphology contains molecular cues that can be used to infer spatial expression. However, existing evaluations generally focus on overall correlation metrics, with limited investigation into the predictability of specific gene programs or whether model-inferred expression retains biological consistency across different tissue contexts, such as primary tumors and metastatic lesions.

In this study, the study employed VirtualST, a conditional diffusion model that uses morphological features extracted by a pathology foundation model, cell composition derived from nuclei segmentation, and spatial neighborhood information as conditioning inputs to generate predictions in gene-expression space progressively. Unlike deterministic regression models, diffusion-based inference is inherently stochastic. If noisy ground-truth expression is used as the initial input during testing, test information may leak into the inference process and lead to overestimated performance. Therefore, all predictions in this study were generated through a leakage-free sampling procedure starting from pure Gaussian noise. Ground-truth ST expression was used only for evaluation after inference, thereby maintaining independence between model predictions and the molecular reference.

Accordingly, the study evaluated the biological utility of VirtualST using leave-one-sample-out cross-validation across 12 samples with paired H&E images and ground-truth ST data from three cohorts: COAD, IDC, and LYMPH_IDC. The study addressed three main questions. First, the study assessed the overall fidelity of spatial expression reconstruction across different tissue contexts. Second, the study compared the predictability of tumor epithelial, secretory, proliferative, hormone-receptor-related, and immune-related gene programs. Third, the study used ground-truth ST epithelial scores from held-out samples, which were not involved in inference, as an independent molecular reference to evaluate the spatial consistency of the predicted epithelial programs. Importantly, the IDC and LYMPH_IDC samples were not patient-matched; therefore, this study compared spatial recovery across two tissue contexts without concluding metastatic evolution or molecular mechanisms.

2. Materials and methods

2.1. Data sources and preprocessing

This study used three tumor cohorts from the HEST-bench dataset [5]: colon adenocarcinoma (COAD), primary invasive ductal carcinoma of the breast (IDC), and breast cancer lymph node metastasis (LYMPH_IDC). Each cohort contained four samples with paired H&E images and ground-truth ST data, yielding a total of 12 samples. H&E image regions corresponding to individual spatial locations were used as morphological inputs, whereas ground-truth ST expression served as the training target and as a post hoc evaluation reference for held-out samples. Gene-expression values were first transformed using log(1+x) and then independently min–max normalized within each sample. Because normalization parameters were determined separately for each sample, this study evaluated only relative within-sample spatial patterns and did not perform absolute expression comparisons across samples or tissue types.

2.2. VirtualST model

VirtualST uses H&E images as its primary information source and performs conditional diffusion directly in gene-expression space. First, a frozen Phikon-v2 pathology foundation model extracts morphological features from the image region corresponding to each spatial location [6]. HoVer-Net is then used to segment and classify nuclei and to calculate the proportions of different cell types at each location [7]. The morphological features and cell-composition profiles jointly condition the diffusion process, guiding the model to progressively generate spatial gene-expression predictions from random noise [8]. The model output is further refined through a residual spatial graph module, in which each location is connected to its six nearest spatial neighbors to integrate local tissue-structure information. The trained graph-refinement module was applied to IDC and LYMPH_IDC, whereas the corresponding module in COAD remained an identity mapping because it was not separately trained.

2.3. Leakage-free cross-validation and model inference

Leave-one-sample-out cross-validation was conducted separately within each of the three cohorts. Because each cohort contained four samples, four independent folds were generated: three samples were used for training and the remaining sample was held out for testing in each fold, until every sample had served once as the test sample. The last checkpoint obtained at the end of training was consistently used for inference in all folds. During testing, the model followed a leakage-free sampling procedure, starting from pure Gaussian noise with the same dimensionality as the target expression and generating predicted expression through reverse diffusion. Ground-truth expression from the held-out sample was not used for noise initialization, reverse diffusion, or spatial graph refinement and was used only for performance evaluation after inference. To reduce stochastic variation in diffusion sampling, ten predictions were independently generated for each test sample using ten fixed random seeds, and their average was used as the final model-inferred output.

2.4. Evaluation metrics and statistical analysis

Overall reconstruction performance was evaluated using the mean gene-wise Pearson correlation coefficient (PCC) and Spearman rank correlation coefficient. For each gene (g), the correlation between predicted expression $\hat{x}_{ig}$ and ground-truth expression x_{ig} was calculated across all spatial locations within a sample. The correlations were then averaged across all genes in the evaluation set to obtain the mean gene-wise

correlation for that sample. In addition to the complete evaluation gene set, performance was separately assessed for highly expressed genes (HEGs), highly variable genes (HVGs), and cancer-specific marker genes. To prevent gene-selection leakage, HVGs in each cross-validation fold were identified using only the expression variances of the three training samples and were subsequently evaluated in the held-out sample. All metrics are reported as the mean ± standard deviation across four held-out samples, with the sample treated as the statistical unit (n = 4 per cohort).

The breast cancer epithelial program score was constructed using EPCAM, KRT8, and KRT7, which were shared across both cohorts. For each spatial location i, the epithelial score was defined as the mean normalized expression of the three genes:

$$S_i = \frac{x_{i,EPCAM} + x_{i,KRT8} + x_{i,KRT7}}{3} \quad (1)$$

The predicted epithelial score $\hat{S}_i$ and ground-truth ST epithelial score S_i were calculated separately, and their Pearson and Spearman correlations were evaluated within each sample. Spatial locations falling within the top 20% of predicted and ground-truth scores were further defined as the high-value hotspot sets H_{pred} and H_{ST}, respectively. Hotspot overlap was quantified using the Dice coefficient:

$$Dice = \frac{2 \mid H_{pred} \cap H_{ST} \mid}{\mid H_{pred} \mid + \mid H_{ST} \mid} \quad (2)$$

Because the ground-truth ST epithelial scores were not involved in model inference for the held-out samples, they served as an independent molecular reference. However, these scores should not be considered equivalent to tumor regions annotated by pathologists.

3. Results

3.1. Overall reconstruction fidelity across tissue contexts

Across the 12 held-out samples from the three cohorts, the final VirtualST outputs showed interpretable within-sample spatial concordance with ground-truth ST data (**Table 1**). The mean gene-wise PCCswere 0.467±0.054, 0.439±0.076, and 0.496±0.127 for COAD, IDC, and LYMPH_IDC, respectively, corresponding Spearman correlations of 0.475±0.057, 0.456±0.080, and 0.471±0.142. Although the mean overall with PCC in LYMPH_IDC was not lower than that in IDC, its larger standard deviation indicated greater variability in model performance across lymph node metastasis samples. Because each cohort contained only four samples, these findings were interpreted descriptively and were not used to infer statistically significant differences between cohorts.

Across the evaluated gene subsets, the HEG, HVG, and marker-gene correlations in COAD were 0.587±0.069, 0.669±0.061, and 0.557±0.067, respectively, with the training-fold-defined HVGs showing the highest correlation. The corresponding correlations in IDC were 0.596±0.075, 0.564±0.078, and 0.537±0.093, indicating moderate concordance across all three subsets. In LYMPH_IDC, the HEG, HVG, and marker-gene correlations were 0.468±0.130, 0.412±0.102, and 0.473 ±0.138, respectively. These values were generally lower than the corresponding results in COAD and IDC and showed greater between-sample variability. Collectively, these findings indicate that VirtualST recovered part of the spatial expression structure across all three tissue contexts, although performance varied according to

both the gene subset and tissue background. **Table 1** reports the final model outputs after sparse gating a nd spatial graph refinement; the graph-refinement module in COAD was an identity mapping.

Table 1. Spatial expression reconstruction fidelity of VirtualST across the three cohorts

Cohort	PCC	Spearman	HEG	HVG	Marker genes
COAD	0.467 ± 0.054	0.475 ± 0.057	0.587 ± 0.069	0.669 ± 0.061	0.557 ± 0.067
IDC	0.439 ± 0.076	0.456 ± 0.080	0.596 ± 0.075	0.564 ± 0.078	0.537 ± 0.093
LYMPH_IDC	0.496 ± 0.127	0.471 ± 0.142	0.468 ± 0.130	0.412 ± 0.102	0.473 ± 0.138

3.2. Predictability of specific gene programs

Gene-wise analysis revealed substantial differences in the spatial predictability of distinct biological programs. In COAD, genes associated with tumor epithelium, secretion, and proliferation showed relatively strong performance. The mean PCCs for CEACAM5, GPX2, CEACAM6, and MKI67 were 0.84, 0.83, 0.76, and 0.73, respectively, whereas the intestinal stem-cell and tumor-associated markers LGR5 and OLFM4 achieved PCCs of 0.63 and 0.52. In comparison, the PCCs for ACTA2, MS4A1, and VWF were 0.26, 0.30, and 0.40, respectively. A similar pattern was observed in IDC, where the epithelial markers EPCAM and KRT8 achieved PCCs of 0.79 and 0.77, while ERBB2 and the proliferation marker TOP2A achieved PCCs of 0.67 and 0.65. These results indicate that VirtualST more reliably reconstructed tumor epithelial and proliferative programs with relatively clear morphological correlates.

Compared with epithelial programs, hormone-receptor- and immune-related genes were less predictable in IDC. The mean PCCs for ESR1, PGR, and CD8A were 0.42, 0.38, and 0.33, respectively. This weaker performance may be related to their more diffuse spatial distributions, narrower dynamic ranges, or expression variation that is less directly reflected in routine tissue morphology, although these interpretations require validation in larger cohorts. In LYMPH_IDC, KRT18, KRT8, and EPCAM each achieved a PCC of approximately 0.60, indicating that VirtualST recovered part of the breast cancer epithelial program within the lymph node tissue background. The immunoglobulin-related genes IGKC, IGHG1, and IGLC2 achieved PCCs of 0.42, 0.41, and 0.35, respectively, whereas ERBB2 achieved a PCC of 0.33, lower than that observed in IDC. Overall, VirtualST showed more consistent reconstruction of tumor epithelial, secretory, and proliferative programs, whereas its inference of hormone-receptor- and some immune-related programs was more limited.

3.3. Spatial recovery in primary breast cancer and lymph node metastasis

In representative samples, VirtualST predictions exhibited epithelial-marker spatial patterns similar to those observed in ground-truth ST data (**Figure 1**). In the IDC sample TENX95, the prediction-to-ground-truth ST correlations for EPCAM and ERBB2 were 0.86 and 0.85, respectively. In the LYMPH_IDC sample NCBI684, the corresponding correlations were 0.75 and 0.65.

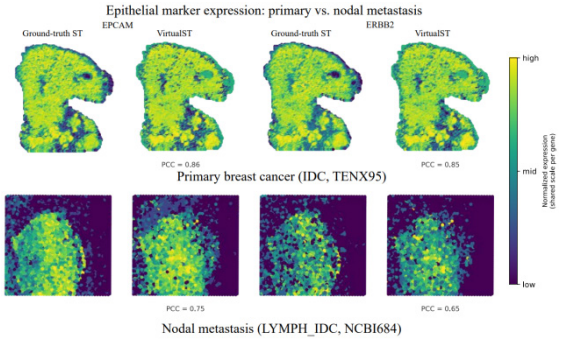

Figure 1. Spatial expression of EPCAM and ERBB2 in primary breast cancer and lymph node metastasis.

The study further used ground-truth ST data from the held-out samples, which were not involved in the inference process, as an independent molecular reference. The Pearson correlations between predicted and ground-truth epithelial scores were 0.79±0.10 in IDC and 0.66±0.11 in LYMPH_IDC. The corresponding Dice coefficients for the top 20% high-value hotspots were 0.64±0.15 and 0.44±0.08, respectively (**Table 2**) . These results indicate that VirtualST recovered the relative spatial patterns of epithelial programs in both tissue contexts.

Table 2. Spatial concordance between predicted and ground-truth ST epithelial-marker scores

Metric	Primary IDC	LYMPH_IDC lymph node metastasis
Pearson correlation	0.79 ± 0.10	0.66 ± 0.11
Spearman correlation	0.74 ± 0.14	0.61 ± 0.13
Top-20% hotspot Dice	0.64 ± 0.15	0.44 ± 0.08

The correlations between predicted EPCAM expression and the proportion of neoplastic cells in representative IDC and LYMPH_IDC samples were 0.59 and 0.52, respectively.

4. Discussion and conclusion

This study evaluated the biological utility of VirtualST using leakage-free sampling across 12 held-out samples from the COAD, IDC, and LYMPH_IDC cohorts. The model preserved moderate and interpretable within-sample spatial expression signals, with relatively strong prediction performance for genes associated with tumor epithelial, secretory, and proliferative programs. Using ground-truth ST data that were not involved in inference as an independent molecular reference, the predicted and ground-truth epithelial scores achieved Pearson correlations of 0.79 in IDC and 0.66 in LYMPH_IDC. These findings indicate that VirtualST recovered epithelial spatial patterns to a certain extent in both breast cancer tissue contexts. Therefore, VirtualST may serve as an exploratory spatial molecular prior when ST data are scarce, but it should not be considered a replacement for ground-truth ST measurements.

This study has several limitations, including the small sample size, the absence of patient-matched IDC and LYMPH_IDC samples, independent normalization of each sample, and saturation in model predictions. Because HoVer-Net-derived cell composition was used as a model input, the corresponding correlation

analysis reflects only directional consistency between the model output and its conditioning information and does not constitute independent validation. Moreover, ground-truth ST epithelial scores are not equivalent to tumor regions annotated by pathologists. Therefore, the current results are not suitable for absolute expression comparisons across tissues or direct pathological tumor-region detection. Future studies should evaluate VirtualST in larger independent cohorts, incorporate pathological region annotations and AUROC, AUPRC, and Dice metrics, and perform retraining-based ablation without cell-composition conditioning.

Disclosure statement

The authors declare 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.

[2] He B, Bergenstråhle L, Stenbeck L, et al., 2020, Integrating Spatial Gene Expression and Breast Tumour Morphology via Deep Learning. Nature Biomedical Engineering, 4(8): 827–834.

[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.

[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.

[5] 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.

[6] Filiot A, Jacob P, Mac KA, Saillard C, 2024, Phikon-v2, A Large and Public Feature Extractor for Biomarker Prediction. arXiv.

[7] 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.

[8] Ho J, Jain A, Abbeel P, 2020, Denoising Diffusion Probabilistic Models. Advances in Neural Information Processing Systems, 33: 6840–6851.