MINT-ST: Multi-Dataset Integration and Pretraining for Histology-Based Spatial Transcriptomics Gene Expression Prediction

Keunho Byeon[1] and Jin Tae Kwak[1]

School of Electrical Engineering, Korea University, Seoul 02841, Republic of Korea
{bkh5922,jkwak}@korea.ac.kr

Abstract. Spatial transcriptomics preserves tissue context while profiling gene expression, but its high cost and experimental complexity limit large-scale use. This has motivated methods that predict spatial gene expression directly from H&E histopathology images. However, most existing approaches are trained on individual datasets with fixed gene sets, limiting the joint use of heterogeneous projects. We propose MINT-ST, a multi-dataset pretraining framework for histology-based spatial gene expression prediction. MINT-ST constructs a global output space from the union of project-specific highly variable genes, jointly pretrains a shared model across multiple projects, and fine-tunes it for each target gene panel and expression distribution. A Transformer integrates the target spot, neighboring spots, relative positions, and species and organ metadata. Training combines variance-normalized regression with a local Pearson correlation loss to learn both expression values and spatial patterns. Across ten human and mouse SpaRED datasets, MINT-ST achieved the best average performance across all evaluation metrics. Ablation studies verified the contributions of spatial context, metadata, correlation-based supervision, multi-project pretraining, and project-specific fine-tuning.

Keywords: Spatial transcriptomics · Gene expression prediction · Multi-dataset integration

1 Introduction

Spatial transcriptomics enables gene expression profiling while preserving tissue architecture, supporting the analysis of spatial heterogeneity and region-specific molecular states in tumor, stromal, immune, and normal tissues. However, its specialized equipment, high cost, and complex experimental procedures limit the scale of available cohorts. In contrast, H&E-stained histology images are routinely acquired in clinical practice, motivating efforts to predict spatial gene expression directly from histopathology.

Most existing methods are trained independently on individual datasets and fixed gene panels, limiting the joint use of projects that differ in tissue, species,

platform, expression distribution, and measured genes. To address this limitation, we propose MINT-ST, a multi-dataset integration and pretraining framework for histology-based spatial gene expression prediction. MINT-ST constructs a global output space from the union of project-specific HVGs and jointly pretrains a single model using partially overlapping gene panels. Species and organ information are incorporated as learnable metadata tokens, and the pretrained model is subsequently fine-tuned to each target project's gene panel and expression distribution.

MINT-ST uses a Transformer to jointly process the histological features of a target spot and its spatial neighbors together with their relative positions. Training combines a gene-variance-aware regression loss with a local Pearson correlation loss, enabling the model to learn both spot-level expression values and local spatial expression patterns. We evaluate MINT-ST on diverse human and mouse spatial transcriptomics datasets against existing prediction methods and analyze the effects of spatial context, correlation-based supervision, multi-project pretraining, and project-specific fine-tuning through ablation studies.

2 Related Work

Early studies demonstrated the feasibility of directly predicting spatial gene expression from histopathology images. STNet [4] independently processed local H&E image patches corresponding to individual spots using a convolutional neural network to predict spot-level gene expression. Subsequent studies increasingly incorporated surrounding tissue context and spatial relationships between spots, while more recent approaches have explored richer neighborhood and cross-modal interactions. Hist2ST [14] combined convolutional feature extraction, a Transformer, and a graph neural network to model global and local dependencies between spots. TCGN [12] integrated convolution, Transformer encoding, and graph-node co-embedding into a single-spot vision backbone to enhance local histological representations. EGNv2 [13] introduced exemplar-guided graph learning to retrieve similar tissue regions and transfer their information for improved gene expression prediction. PEKA [7] transferred knowledge from a single-cell transcriptomic foundation model to a pathology foundation model through parameter-efficient adaptation, knowledge distillation, and structural alignment. NH2ST [8] employed target and neighborhood branches together with hypergraph learning, cross-attention, and contrastive learning to jointly model spatial relationships and pathology–transcriptomics interactions.

Despite these advances, most existing methods are trained independently on a predefined dataset and gene panel.

3 Methods

MINT-ST predicts target-spot gene expression from H&E features, spatial coordinates, and project metadata of the target and its neighboring spots. A frozen pathology foundation model extracts spot-level histological features, which are

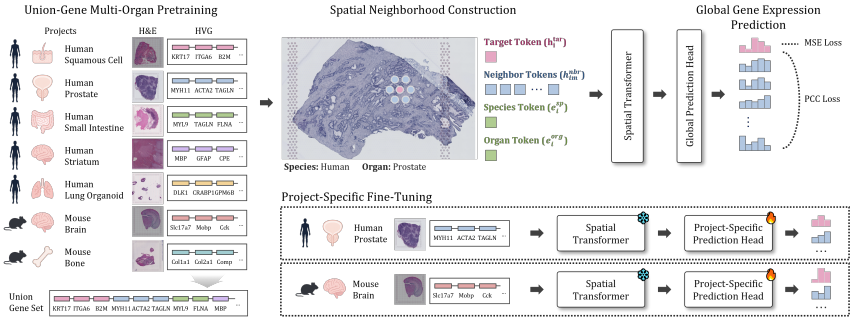

Fig. 1. Overview of the proposed MINT-ST framework.

combined with role, relative-position, species, and organ embeddings and processed by a spatial Transformer. The target-token output predicts the target expression vector, while neighbor-token outputs provide supervision for local expression patterns (Fig. 1).

3.1 Spatial Neighborhood Construction

Each spot is represented by a histological feature, a two-dimensional coordinate, and a gene expression vector. For target spot i, we select the K nearest spots from the same tissue section. When fewer than K valid neighbors are available because of tissue boundaries or empty regions, the remaining slots are padded and masked in both attention and loss computation. The histological encoder remains frozen during multi-project pretraining and project-specific fine-tuning to provide consistent representations across projects.

Let $\mathbf{x}_i \in \mathbb{R}^D$ be the feature of target spot i, j_{im} the spot assigned to its m-th neighbor slot, and $a_{im} \in \{0, 1\}$ the corresponding validity indicator. Target and neighbor features are projected into a shared hidden space by f_{spot}:

$$\mathbf{h}_i^{\text{tar}} = f_{\text{spot}}(\mathbf{x}_i) + \mathbf{e}_{\text{tar}}, \qquad \mathbf{h}_{im}^{\text{nbr}} = f_{\text{spot}}(\mathbf{x}_{j_{im}}) + \mathbf{e}_{\text{nbr}} + \mathbf{p}_{im}, \tag{1}$$

where $\mathbf{e}_{\text{tar}}$ and $\mathbf{e}_{\text{nbr}}$ are learnable role embeddings distinguishing the prediction location from contextual neighbors. The local token sequence is

$$\mathbf{H}_i^{(0)} = \left[\mathbf{h}_i^{\text{tar}}, \mathbf{h}_{i1}^{\text{nbr}}, \ldots, \mathbf{h}_{iK}^{\text{nbr}}\right]. \tag{2}$$

The spatial embedding $\mathbf{p}_{im}$ represents the displacement of neighbor j_{im} from target i. To reduce project-specific differences in coordinate scale and spot spacing, the displacement is normalized by the median target-to-neighbor distance, expanded with sinusoidal Fourier features, and projected by a multilayer perceptron [9]. Padded slots use zero spatial embeddings and are excluded by a_{im}.

3.2 Metadata-Conditioned Spatial Transformer

Species and organ are encoded as learnable metadata tokens and appended to the local sequence:

$$\mathbf{Z}_i^{(0)} = \left[\mathbf{H}_i^{(0)}, \mathbf{e}_i^{\mathrm{sp}}, \mathbf{e}_i^{\mathrm{org}}\right], \qquad \mathbf{Z}_i^{(L)} = \mathrm{Transformer}\Big(\mathbf{Z}_i^{(0)}\Big). \tag{3}$$

The Transformer therefore integrates histological, spatial, and biological information. A shared prediction head is applied to the target and all valid neighbor outputs. The target prediction is used for direct regression, whereas target and neighbor predictions jointly define the local Pearson correlation loss. During pretraining, metadata tokens are randomly replaced with learnable unknown-category tokens, preventing excessive reliance on project identity and encouraging the use of morphology and spatial context.

3.3 Union-Gene Multi-Project Pretraining

MINT-ST forms a global output vocabulary from the union of project-specific highly variable genes (HVGs) across all training projects. A single model is jointly optimized on projects differing in species, organ, expression distribution, and target gene set. The global head predicts every union gene, but each sample contributes loss only for genes observed in its project.

Let $\mathbf{z}_i$ be the target representation and $\mathbf{I}_p$ the global indices of the HVGs for project p. The project-observed prediction is

$$\widehat{\mathbf{y}}_i^{(p)} = \mathbf{W}_{\mathbf{I}_p}\mathbf{z}_i + \mathbf{b}_{\mathbf{I}_p}, \tag{4}$$

where $\mathbf{W}_{\mathbf{I}_p}$ and $\mathbf{b}_{\mathbf{I}_p}$ select the corresponding rows of the global head. Optimization compares $\widehat{\mathbf{y}}_i^{(p)}$ only with the observed project-specific vector $\mathbf{y}_i^{(p)}$. Gene dropout further removes a random subset of observed genes from each sample loss, reducing dependence on a particular panel and improving robustness to different target subsets. This design preserves a unified output space while learning transferable histological and spatial representations across heterogeneous projects.

3.4 Project-Specific Fine-Tuning

After pretraining, the feature projection, relative-position encoder, Transformer, role embeddings, and metadata embeddings are initialized from the shared checkpoint and fine-tuned independently for each target project. A project-specific head is initialized from the matching rows of the global head:

$$\mathbf{W}^{(p)} \leftarrow \mathbf{W}_{\mathbf{I}_p}, \qquad \mathbf{b}^{(p)} \leftarrow \mathbf{b}_{\mathbf{I}_p}. \tag{5}$$

Fine-tuning uses only the target project's training split, adapting the shared representation and output head to its gene panel and expression distribution. Gene dropout and metadata masking are disabled so that all target genes and available metadata are used.

3.5 Training Objectives

Both stages optimize a variance-normalized regression loss and a local Pearson correlation loss. The former matches absolute spot-level expression while balancing genes with different dynamic ranges; the latter preserves relative spatial patterns within each neighborhood.

For project p, the regression loss is

$$\mathcal{L}_{\text{reg}} = \frac{1}{|\mathcal{V}_p|} \sum_{(i,g)\in\mathcal{V}_p} \frac{(\widehat{y}_{ig} - y_{ig})^2}{\sigma_{p,g}^2 + \epsilon}, \tag{6}$$

where $\mathcal{V}_p$ is the set of valid spot–gene pairs and $\sigma_{p,g}^2$ is the variance of gene g estimated only from the training split of project p. Project-specific statistics are used during multi-project pretraining, and ϵ ensures numerical stability.

For each target spot i and gene g, let $\widehat{\mathbf{v}}_{ig}$ and $\mathbf{v}_{ig}$ denote predicted and observed expression over the target and its valid neighbors. The local correlation loss is

$$\mathcal{L}_{\text{corr}} = 1 - \frac{1}{|\mathcal{C}|} \sum_{(i,g)\in\mathcal{C}} \text{corr}(\widehat{\mathbf{v}}_{ig}, \mathbf{v}_{ig}), \tag{7}$$

where $\mathcal{C}$ contains target–gene pairs with at least two valid observations and nonzero expression variation. The final objective is

$$\mathcal{L}_{\text{total}} = \lambda_{\text{reg}}\mathcal{L}_{\text{reg}} + \lambda_{\text{corr}}\mathcal{L}_{\text{corr}}, \tag{8}$$

with $\lambda_{\text{reg}} = \lambda_{\text{corr}} = 1.0$ by default.

4 Experiments

4.1 Implementation Details

Histological features for each spot were extracted using the pretrained pathology foundation model UNI [2], and the feature encoder was kept frozen throughout training. H&E patches of 224×224 pixels were extracted, and the resulting 1024-dimensional spot features were projected into a 512-dimensional hidden space. All data preprocessing procedures, including gene expression preprocessing, target HVG selection, and the predefined training, validation, and test splits, followed the SpaRED framework [5]. The spatial Transformer consisted of four encoder layers with eight attention heads. For each target spot, we used its $K = 36$ nearest neighbors from the same tissue section. Multi-project pretraining was performed for 20 epochs using the AdamW optimizer with a batch size of 32 and a learning rate of 0.001. The model was then fine-tuned independently for each target project for 10 epochs with a batch size of 1 and a learning rate of 0.0002. The weight decay was set to 10^{-4}. During pretraining, gene dropout and metadata masking were applied with probabilities of 0.5 and 0.1, respectively, and both were disabled during fine-tuning. All experiments used a random seed of 42 and were conducted on an NVIDIA GeForce RTX 3090 GPU. The checkpoint achieving the highest validation PCC was selected for test evaluation.

Table 1. Comparison with existing histology-based spatial gene expression prediction methods. Results are reported as the mean ± standard deviation across ten datasets. The best and second-best results are shown in bold and underlined, respectively.

Model	PCC ↑	MI ↑	AUC_{0vNZ} ↑	AUC_{Q50} ↑	NRMSE ↓
STNet	0.0246 ± 0.04	0.0226 ± 0.01	0.4860 ± 0.03	0.5139 ± 0.03	0.1987 ± 0.06
Hist2ST	0.0503 ± 0.01	0.0775 ± 0.03	0.4934 ± 0.01	0.4999 ± 0.01	0.1451 ± 0.03
TCGN	0.1638 ± 0.09	0.0585 ± 0.02	0.5459 ± 0.06	0.5896 ± 0.05	0.2450 ± 0.15
EGNv2	0.2116 ± 0.08	0.0604 ± 0.03	0.5929 ± 0.05	0.6208 ± 0.05	$\underline{0.1433 \pm 0.03}$
NH2ST	0.2894 ± 0.09	0.0889 ± 0.03	0.5986 ± 0.06	0.6442 ± 0.05	0.2344 ± 0.27
PEKA	$\underline{0.3458 \pm 0.10}$	$\underline{0.0982 \pm 0.04}$	$\underline{0.6289 \pm 0.05}$	$\underline{0.6670 \pm 0.04}$	0.1633 ± 0.10
MINT-ST	$\mathbf{0.3686 \pm 0.15}$	$\mathbf{0.1344 \pm 0.05}$	$\mathbf{0.6572 \pm 0.06}$	$\mathbf{0.6905 \pm 0.06}$	$\mathbf{0.1250 \pm 0.02}$

4.2 Datasets

We used ten datasets curated in SpaRED [5], originating from Abalo et al. [1], Erickson et al. [3], Mirzazadeh et al. [6], Vicari et al. [10], and Villacampa et al. [11]. The datasets include human and mouse samples from skin, prostate, intestine, bone, brain, and lung organoid tissues. SpaRED provides either 32 or 128 highly variable genes (HVGs) for each dataset, which were used as the project-specific target genes during fine-tuning and evaluation. Across the ten datasets, these project-specific HVG sets were combined to form the union-gene vocabulary used for multi-project pretraining. The 32- or 128-gene project panels form the 695-gene union vocabulary. For multi-project pretraining, we jointly used the predefined training splits of all ten datasets. The pretrained model was subsequently fine-tuned independently for each dataset using its corresponding training split.

4.3 Evaluation Metrics

We evaluated gene expression prediction using Pearson correlation coefficient (PCC), mutual information (MI), two area-under-the-curve (AUC) metrics, and normalized root mean squared error (NRMSE). AUC_{0vNZ} measures the ability to distinguish zero from nonzero expression, and AUC_{Q50} evaluates discrimination between expression values below and above the gene-specific median. All metrics were first computed independently for each target gene across all spots in the test split and then averaged over the target genes.

4.4 Comparison Results

We compared MINT-ST with six baseline models: STNet, Hist2ST, TCGN, EGNv2, NH2ST, and PEKA. Each baseline was trained and evaluated independently on each dataset using its corresponding project-specific split and target HVGs. In contrast, MINT-ST was jointly pretrained across the ten datasets and subsequently fine-tuned and evaluated separately for each project. Table 1 summarizes the mean performance across the ten SpaRED datasets. MINT-ST achieved the best average performance across all five evaluation metrics.

Table 2. Ablation study of the main components of MINT-ST. Results are reported as the mean ± standard deviation across ten datasets. Project adapt. denotes project-specific fine-tuning for the pretrained model and project-specific training from scratch when pretraining is disabled. O and X indicate whether each component is enabled or disabled. The best and second-best results are shown in bold and underlined, respectively.

Meta Trans.	Local PCC	Pretraining	Project Adapt.	PCC ↑	MI ↑	NRMSE ↓
X	X	X	O	0.3474 ± 0.10	0.1114 ± 0.04	0.1498 ± 0.07
O	X	X	O	0.3470 ± 0.14	$\underline{0.1270 \pm 0.05}$	0.1491 ± 0.07
O	O	X	O	$\underline{0.3619 \pm 0.13}$	0.1250 ± 0.04	$\underline{0.1384 \pm 0.04}$
O	O	O	X	0.1979 ± 0.10	0.0547 ± 0.03	0.1479 ± 0.03
O	O	O	O	$\mathbf{0.3686 \pm 0.15}$	$\mathbf{0.1344 \pm 0.05}$	$\mathbf{0.1250 \pm 0.02}$

Compared with PEKA, the strongest competing method, MINT-ST improved PCC, MI, $\mathrm{AUC}_{0\mathrm{vNZ}}$ and AUC_{Q50} by 0.0228, 0.0362, 0.0283, and 0.0235, respectively. MINT-ST also reduced NRMSE to 0.1250, compared with the second-best result of 0.1433 obtained by EGNv2. These consistent improvements indicate that multi-dataset pretraining provides transferable information beyond models trained independently on individual datasets. The union-gene model showed limited performance when applied directly without project-specific adaptation. This is expected because the projects differ in their target gene panels, expression distributions, tissues, and species. Although pretraining learns shared histological and spatial representations, its predictions are not directly calibrated to the expression scale and gene composition of an individual project. Project-specific fine-tuning is therefore required to adapt both the representation and prediction head to the target distribution.

4.5 Ablation Studies

Table 2 demonstrates ablation experiments to evaluate the contributions of the metadata-conditioned spatial Transformer, local PCC loss, multi-project pretraining, and project-specific fine-tuning. All configurations used the same frozen histological features. The metadata-conditioned spatial Transformer improved MI from 0.1114 to 0.1270. Adding the local PCC loss further increased PCC from 0.3470 to 0.3619 and reduced NRMSE from 0.1491 to 0.1384. The pretrained model showed limited performance when applied without fine-tuning. In contrast, the complete MINT-ST model improved PCC from 0.3619 to 0.3686 over the model trained without pretraining and achieved the best performance across all metrics. This indicates that multi-project pretraining provides a useful initialization, while project-specific adaptation remains necessary to account for differences in gene sets and expression distributions.

4.6 Qualitative Analysis

To qualitatively evaluate the predicted spatial expression patterns, we visualized spot-level gene expression values overlaid on the corresponding H&E tissue

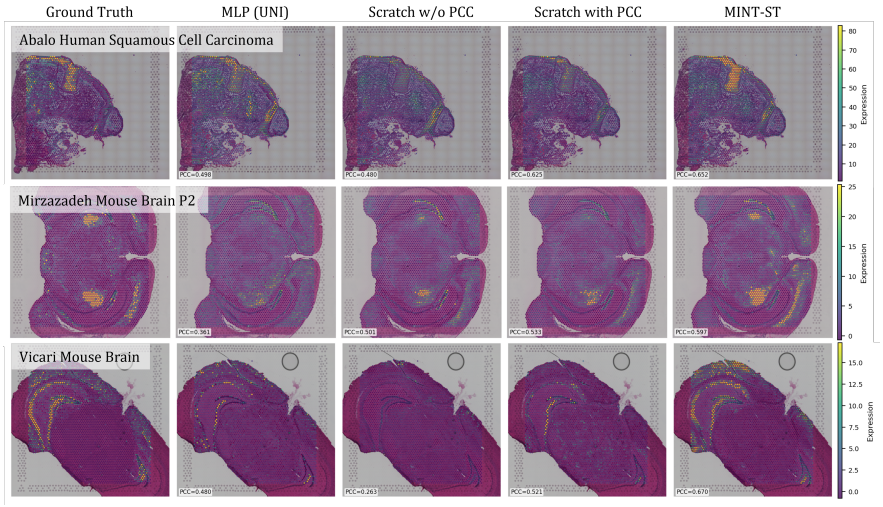

Fig. 2. Qualitative comparison of spot-level spatial gene expression maps overlaid on the corresponding H&E tissue sections.

sections. We compared the ground truth with MLP (UNI), Scratch w/o PCC, Scratch w/ PCC, and the complete MINT-ST model for three representative dataset–gene pairs: *MALAT1* in Abalo human squamous cell carcinoma, *Pcp4* in Mirzazadeh mouse brain P2, and *Slc17a7* in Vicari mouse brain.

As shown in Fig. 2, MLP (UNI) and the scratch-trained models partially recovered the overall expression patterns but exhibited less accurate regional localization or expression intensities. In comparison, MINT-ST most closely reproduced both the spatial distribution of high- and low-expression regions and the corresponding expression magnitudes observed in the ground truth. These qualitative results support the quantitative findings and indicate that multi-project pretraining, spatial context modeling, and local correlation supervision jointly improve the reconstruction of spatial gene expression patterns.

5 Conclusion

We proposed MINT-ST, a histology-based gene expression prediction framework designed to integrate multiple spatial transcriptomics datasets with different tissues, species, and target gene sets. MINT-ST enables joint learning from heterogeneous ST projects by combining a union-gene output space with spatial neighborhood and metadata conditioning. Across ten human and mouse datasets, it achieved the best mean performance for all evaluated metrics. Ablations show that spatial context and local PCC supervision improve project-specific learning, while union-gene pretraining is most effective when followed by project adaptation. Future work will investigate leave-one-project-out pretraining and transfer to unseen projects, organs, species, platforms, and gene panels.

Acknowledgments. This work was supported by a grant of the National Research Foundation of Korea (NRF) (No. RS-2025-00558322 and RS-2024-00397293).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Abalo, X., Thrane, K., Ji, A., Mirzazadeh, R., Khavari, P.: Human squamous cell carcinoma, visium. Mendeley Data **10** (2021)
2. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., et al.: Towards a general-purpose foundation model for computational pathology. Nature medicine **30**(3), 850–862 (2024)
3. Erickson, A., He, M., Berglund, E., Marklund, M., Mirzazadeh, R., Schultz, N., Kvastad, L., Andersson, A., Bergenstråhle, L., Bergenstråhle, J., et al.: Spatially resolved clonal copy number alterations in benign and malignant tissue. Nature **608**(7922), 360–367 (2022)
4. He, B., Bergenstråhle, L., Stenbeck, L., Abid, A., Andersson, A., Borg, Å., Maaskola, J., Lundeberg, J., Zou, J.: Integrating spatial gene expression and breast tumour morphology via deep learning. Nature biomedical engineering **4**(8), 827–834 (2020)
5. Mejia, G., Ruiz, D., Cárdenas, P., Manrique, L., Vega, D., Arbeláez, P.: Enhancing gene expression prediction from histology images with spatial transcriptomics completion. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 91–101. Springer (2024)
6. Mirzazadeh, R., Andrusivova, Z., Larsson, L., Newton, P.T., Galicia, L.A., Abalo, X.M., Avijgan, M., Kvastad, L., Denadai-Souza, A., Stakenborg, N., et al.: Spatially resolved transcriptomic profiling of degraded and challenging fresh frozen samples. Nature Communications **14**(1), 509 (2023)
7. Pan, S., Chen, J., Secrier, M.: Teaching pathology foundation models to accurately predict gene expression with parameter efficient knowledge transfer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 605–613. Springer (2025)
8. Qu, M., Wu, Y., Di, D., Gao, Y., Su, T., Song, Y., Fan, L.: Spatially gene expression prediction using dual-scale contrastive learning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 574–584. Springer (2025)
9. Tancik, M., Srinivasan, P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J., Ng, R.: Fourier features let networks learn high frequency functions in low dimensional domains. Advances in neural information processing systems **33**, 7537–7547 (2020)
10. Vicari, M., Mirzazadeh, R., Nilsson, A., Shariatgorji, R., Bjärterot, P., Larsson, L., Lee, H., Nilsson, M., Foyer, J., Ekvall, M., et al.: Spatial multimodal analysis of transcriptomes and metabolomes in tissues. Nature Biotechnology **42**(7), 1046–1050 (2024)
11. Villacampa, E.G., Larsson, L., Mirzazadeh, R., Kvastad, L., Andersson, A., Mollbrink, A., Kokaraki, G., Monteil, V., Schultz, N., Appelberg, K.S., et al.: Genome-wide spatial expression profiling in formalin-fixed tissues. Cell Genomics **1**(3) (2021)

12. Xiao, X., Kong, Y., Li, R., Wang, Z., Lu, H.: Transformer with convolution and graph-node co-embedding: An accurate and interpretable vision backbone for predicting gene expressions from local histopathological image. Medical Image Analysis **91**, 103040 (2024)
13. Yang, Y., Hossain, M.Z., Stone, E., Rahman, S.: Spatial transcriptomics analysis of gene expression prediction using exemplar guided graph neural network. Pattern Recognition **145**, 109966 (2024)
14. Zeng, Y., Wei, Z., Yu, W., Yin, R., Yuan, Y., Li, B., Tang, Z., Lu, Y., Yang, Y.: Spatial transcriptomics prediction from histology jointly through transformer and graph neural networks. Briefings in Bioinformatics **23**(5), bbac297 (2022)