Reframing Spatial Transcriptomics Prediction: From Regression to Classification Using Swin Transformer

Weiqi Li[1], Fuxin Li[2], Xinzhou Ge[1], and Yuan Jiang[1, *]

[1] *Statistics, Oregon State University, Corvallis, OR, USA, 97331*

[2] *EECS, Oregon State University, Corvallis, OR, USA, 97331*

[*] **Corresponding author:** Yuan Jiang, yuan.jiang@oregonstate.edu.

Abstract. Spatial transcriptomics profiles gene expression while preserving spatial context, offering unprecedented insight into tissue organization. However, high cost and technical complexity limit its routine clinical adoption. A promising alternative is to predict spatially resolved expression directly from widely available histology images using deep learning. Existing approaches formulate this as a regression task, attempting to predict continuous gene expression values from small cropped H&E image patches. We argue this formulation is ill-suited for sparse, discrete, and noisy expression data. In this study, we reframe the task as a multi-class classification problem, predicting discrete expression levels rather than exact values. We demonstrate that this reformulation substantially improves discriminative performance and can be seamlessly incorporated into existing architectures. We further propose a Swin Transformer-based framework that efficiently captures hierarchical spatial context while reducing model complexity and enhancing scalability. The proposed model outperforms the CNN-based benchmark even with limited neighborhood context, offering a viable and scalable solution for near-cellular and subcellular resolution spatial transcriptomics datasets.

Keywords. Deep learning; Multi-class classification; Prediction modeling; Self-attention; Spatial transcriptomics; Transformer.

1. Introduction

Spatial transcriptomics is an emerging RNA sequencing technology that surpasses traditional single-cell and bulk sequencing by preserving spatial information of gene expression. Rather than producing a single aggregated count per gene for an entire tissue sample, spatial transcriptomics measures expression at micrometer-scale spots, or even down to subcellular level (Williams et al., 2022). This spatial dimension, along with the massive increase in data volume, holds great promise for improved diagnostics and deeper insights into disease mechanisms. However, its high cost and technical demands hinder its adoption to routine clinical workflows. In contrast, hematoxylin and eosin (H&E) stained histology images are widely accessible and routinely used in diagnostics (Fischer et al., 2008). To facilitate clinical integration of spatial transcriptomics, recent research turned to deep learning to predict spatially resolved gene expression directly from H&E images.

Over the past three years, a variety of neural architectures have been proposed for this task. Early models treated each sequenced spot independently and focused solely on local features using convolutional neural networks (CNNs) like DenseNet, EfficientNet, ResNet or Vision Transformer (ViT) (He et al., 2020; Rahaman et al., 2023). While these models leveraged the large paired datasets of spot images and gene expression profiles, they largely ignored neighborhood relationships and broader tissue context. To address this limitation, more recent approaches have integrated global context via transformers and local spatial dependencies via graph neural networks (GNNs) (Zeng et al., 2022; Xiao et al., 2024; Pang et al., 2021). Although transformers are capable of capturing long-range dependencies, prior studies often combined them with GNNs to explicitly encode local spatial structure. Despite these architectural advances, predictive performance remains modest. As reviewed by Wang et al. (2025), existing models typically achieve only moderate correlation between predicted and ground truth gene expression, and they struggle to reliably distinguish expressed from non-expressed genes or low from high expression levels.

One main limitation of existing models is that they have all framed this task as a regression problem, seeking to predict hundreds of log-transformed, continuous expression values from each small cropped H&E image patch. We argue that this formulation is ill-suited for noisy, discrete, and highly sparse spatial transcriptomics expression data. Moreover, it is unrealistic to expect a single small image patch to contain sufficient information for predicting continuous expression values at decimal precision across hundreds of genes. We therefore propose reframing the task as a **multi-class classification** problem, predicting discrete expression levels—with zero treated as a separate class and remaining nonzero values divided by quantiles—rather than exact continuous values. The reframing can be applied to any existing regression-based model. We demonstrate its effectiveness by adapting it to the benchmark ST-Net architecture proposed by He et al. (2020), where it yields substantial gains in discriminative performance.

Another limitation of existing context-aware transformer models is that they process entire tissue sections as single inputs, treating each sequenced spot as a token (e.g., Zeng et al., 2022; Wang et al., 2024; Jia et al., 2024). Limited by the quadratic memory cost of global attention, this strategy becomes computationally infeasible for datasets exceeding a few thousand spots—let alone for the increasingly available near-cellular and subcellular datasets containing tens or hundreds of thousands of cells. In addition, the scarcity of spatial transcriptomics samples limits the ability of global-attention models to learn spatial structure effectively without auxiliary modules such as GNNs (Zeng et al., 2022; Jia et al., 2024). To address these challenges, we introduce a **Swin Transformer**-based framework that considers a defined neighborhood around each target spot, making it viable for datasets of any resolution or scale (Liu et al., 2021). The Swin Transformer captures hierarchical spatial relationships by performing attention within local windows and progressively aggregating features across scales. This design allows for larger token counts and requires less data to learn locality, while eliminating the need for explicit graph-based encoding.

From our experimental results, by reframing the prediction task from regression to clas-

sification, the discriminative power between zero versus nonzero expression and between low versus high expression levels improved markedly. The Swin Transformer further outperforms ST-Net even when contextual information is limited to a relatively small neighborhood (7×7 patch grid). Because Swin Transformer demands substantially less memory than conventional transformers (Liu et al., 2021), larger neighborhoods than the one tested in this study remain computationally feasible and can potentially yield even greater performance gains.

2. Materials and Methods

2.1. Datasets and Preprocessing

This study uses data from a clinical investigation of autoimmune disease (Broad Institute, 2024). To identify potential genetic mutations underlying disease mechanisms, eight samples were collected from five patients diagnosed with inflammatory bowel disease (IBD), one of the most common forms of auto-immune disease. For each sample, there is an H&E image and a matching spatial transcriptomics profile of 460 genes at subcellular resolution sequenced using the 10x Genomics Xenium platform. Each sample contains between 100,000 and 300,000 cells. Every cell is uniquely identified with a cell ID, which allows retrieval of its nucleus location in the H&E image and the corresponding counts of genes detected within its boundary.

Each cell-level observation was preprocessed and paired as one training example. Specifically, a $32 \times 32 \times 3$ RGB image centered on the nucleus was extracted from the H&E image as the model input for each cell; and the 460 gene expression values for the corresponding cell were used as the label. The dataset was split into seven samples for training (approximately 780,000 cells) and one sample for testing, with no overlapping patients between the two sets. Because the testing sample contains over 100,000 cells, evaluating all cells simultaneously is computationally infeasible for memory-intensive metrics such as Pearson's correlation coefficient, which require storing all prediction-label pairs. To maintain computational tractabil-

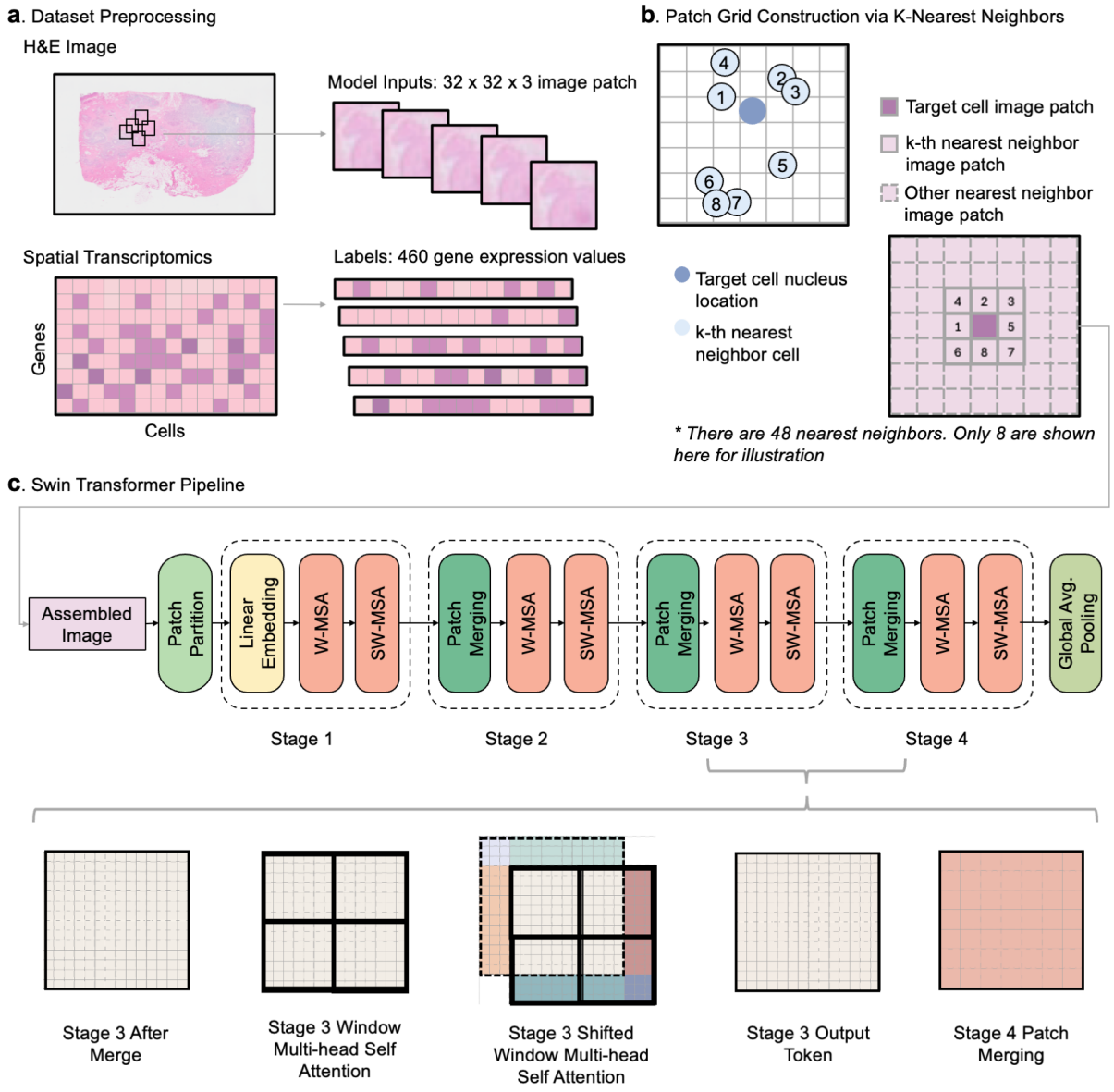

Figure 1: Overview of the Spatial Gene Expression Prediction Framework.

a. Dataset Preprocessing: H&E images (10x Genomics, 2024) were divided into patches as model inputs. Each patch corresponds to a cell with 460-dimensional gene expression labels. The image shown is illustrative only and not from the dataset used in this study.

b. Patch Grid Construction: For each target cell (dark blue), its patch and those of 48 nearest neighbors (light blue) form a spatial grid (8 shown for clarity).

c. Swin Transformer Pipeline: The assembled patch grid passes through four stages of hierarchical encoding with window-based (W-MSA) and shifted window-based (SW-MSA) self-attention, interleaved with patch merging. A global average pooling layer aggregates the final representations for expression prediction.

ity, we therefore evaluated performance on a subset of 10,000 cells. To ensure comparability between architectures, we selected testing cells forming a contiguous neighborhood. While spatial proximity is irrelevant for CNN-based models such as the benchmark ST-Net, it is essential for architectures incorporating spatial context, such as the Swin Transformer.

2.2. Formulation of the Multi-class Classification

To evaluate the impact of different classification formulations and hyperparameter settings, we designed three classification head variants (A-C) and attached them to a pretrained DenseNet-112 backbone. The only modification from the benchmark model ST-Net lies in the prediction head, which performs classification or a combination of classification and regression instead of regression alone.

In **Setup A**, the predicted class is obtained directly from the logits produced by the final linear layer, representing a straightforward multi-class prediction for each gene. The predicted expression for each gene is given by the mean expression value of its assigned class, which is defined separately for each gene. Specifically, let $\mathbf{h}_i \in \mathbb{R}^d$ denote the feature embedding extracted for cell i, and let K denote the number of discrete expression classes. For each gene g, $\mathbf{W}_{\text{cls},g} \in \mathbb{R}^{K\times d}$ and $\mathbf{b}_{\text{cls},g} \in \mathbb{R}^K$ represent the classification head weights and bias, and each class k has an associated mean expression $\bar{y}_{g,k}$. The class probabilities are given by

$$\mathbf{p}_{ig} = \text{softmax}(\mathbf{W}_{\text{cls},g}\mathbf{h}_i + \mathbf{b}_{\text{cls},g}), \tag{1}$$

and the predicted class is

$$\hat{c}_{ig} = \underset{k\in\{1,\dots,K\}}{\arg\max}\, p_{igk}. \tag{2}$$

Finally, the predicted gene expression value is $\bar{y}_{g,\hat{c}_{ig}}$.

Compared to Setup A, Setups B and C integrate classification and regression prediction heads in hopes of leveraging the classification head's discriminative power while preserving regression's ability to refine continuous predictions.

In **Setup B**, the predicted class $\hat{c}_{ig}$ determines the class-specific mean and an auxiliary regression head refines deviation within each class. Let $\mathbf{W}_{\text{reg},g} \in \mathbb{R}^d$ and $b_{\text{reg},g} \in \mathbb{R}$ denote regression parameters for gene g. The predicted value is:

$$\hat{y}_{ig} = \bar{y}_{g,\hat{c}_{ig}} + (\mathbf{W}_{\text{reg},g}^T \mathbf{h}_i + b_{\text{reg},g}). \tag{3}$$

In **Setup C**, two separate regression heads estimate deviations for lowly and highly expressed genes respectively, while a classification head distinguishes between non-expressed, low and high expression levels, as follows:

$$\hat{y}_{ig} = \begin{cases} 0, & \hat{c}_{ig} = \text{zero}, \\ \bar{y}_{g,\text{low}} + (\mathbf{W}_{\text{low},g}^T \mathbf{h}_i + b_{\text{low},g}), & \hat{c}_{ig} = \text{low}, \\ \bar{y}_{g,\text{high}} + (\mathbf{W}_{\text{high},g}^T \mathbf{h}_i + b_{\text{high},g}), & \hat{c}_{ig} = \text{high}. \end{cases} \tag{4}$$

Cross-entropy (CE) is used as the loss function for classification head while mean squared error (MSE) is used for regression head. In Setups B and C, the two components are combined in a weighted sum:

$$\mathcal{L} = \alpha \mathcal{L}_{\text{CE}} + \beta \mathcal{L}_{\text{MSE}}, \tag{5}$$

where the weights α and β control the relative contribution of classification and regression to the total objective. Because the two components differ in scale, we select α and β empirically such that their average magnitudes during training are of similar order. Specifically, multiple combinations were tested, and the pair yielding balanced loss magnitudes across training epoch was adopted to prevent one component from dominating optimization.

Gene expression values are highly sparse, with approximately 94% of observations in the dataset equal to zero. We treat zero expression as a distinct class, and discretize nonzero continuous expression values into additional levels based on quantiles. To mitigate the extreme class imbalance, we incorporate class weights into the cross-entropy loss function.

The class weight for each level is adjusted using a tunable weight exponent hyperparameter γ, for example,

$$w_{\text{zero}} = \left(\frac{\sum_i^N \sum_g^G 1\{\text{Count}_{ig} \neq 0\}}{NG} \right)^{\gamma}, \quad w_{\text{nonzero}} = \left(\frac{\sum_i^N \sum_g^G 1\{\text{Count}_{ig} = 0\}}{NG} \right)^{\gamma},$$

where N denotes the number of cells and G denotes the number of genes. The higher γ is, the less weight is allocated to zero expression observations in the loss function.

While this weighting scheme may not represent the optimal approach for class imbalance correction, empirical testing (see section 3 and Figure 2) indicates that it does not substantially affect overall model performance when appropriate thresholds are selected for distinguishing expressed versus non-expressed and low versus high expression levels based on receiver operating characteristic (ROC) analysis of the training data result.

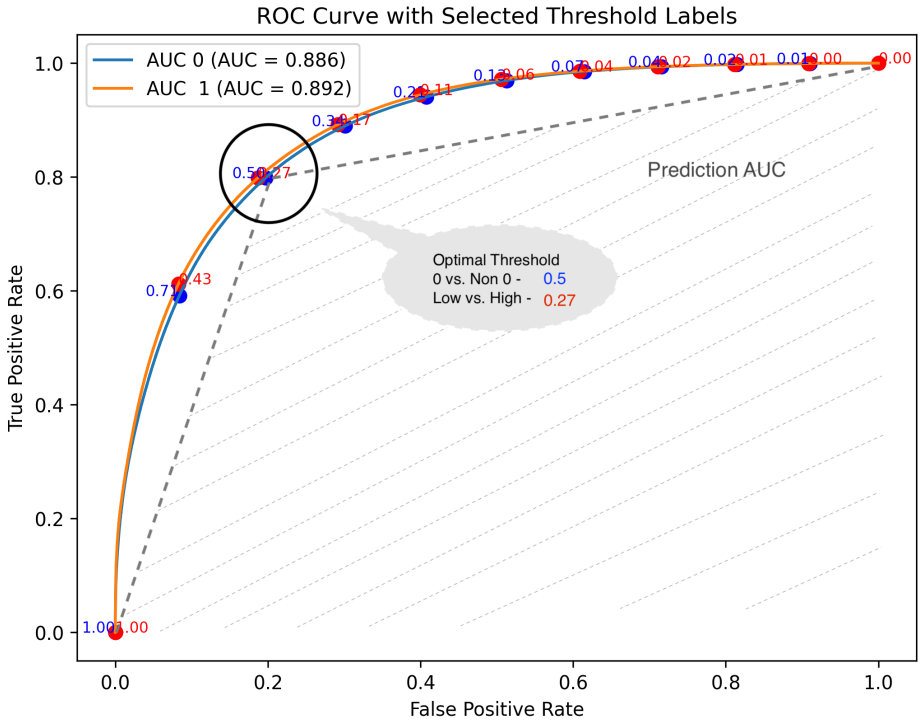

Figure 2: ROC curve analysis identifying thresholds that maximize prediction performance. The highlighted points indicate the optimal single thresholds corresponding to the maximum area under the ROC curve (AUC).

2.3. Swin Transformer

Unlike natural language processing, vision tasks possess an inherent hierarchical structure: contextual information must be integrated across multiple spatial scales, from local neighborhoods to global tissue organization. Vanilla transformers and Vision Transformers (ViTs) employ a global attention mechanism in which all tokens attend to one another, relying entirely on the training process to learn this hierarchy. However, without an inductive bias for local structure, such models often require datasets with millions of images to implicitly capture convolution-like locality (Dosovitskiy et al., 2021). In spatial transcriptomics, training data are comparatively scarce, and transformer-based models trained solely with global attention frequently underperform. Thus, we adopt the Swin Transformer, a hierarchical vision transformer specifically designed to emphasize localized contextual relationships while retaining transformer's ability to capture long-range dependencies (Liu et al., 2021).

Unlike conventional implementations that process patches from an entire tissue image as a single input, our framework constructs a patch grid for each target cell. This design supports scalability across spatial resolutions—from spot-level to near-cellular and subcellular-resolution data. The grid size can be adjusted depending on available GPU memory; in our implementation, each grid contains a 7×7 arrangement of patches, consisting of the target cell and its 48 nearest neighbors. Because cells are distributed irregularly rather than on a fixed lattice, we identify neighbors for each target cell using Euclidean distance in spatial coordinates and sort them into an approximate grid based on their (x, y) positions (See Figure 1b).

The assembled image grid is then processed by a pretrained Swin Transformer (Liu et al., 2021; Microsoft Research, 2021), which first divides the image into $4 \times 4 \times 3$ tokens to capture fine details and then sequentially passes through multiple hierarchical stages of attention and downsampling. As illustrated in Figure 1c, in each stage, the tokens are first downsampled by merging into 2×2 local groups to increase receptive field, after which self-attention is computed within non-overlapping windows. To propagate global context, each stage in-

cludes two consecutive Swin Transformer blocks: one with window multi-head self-attention (W-MSA) and another with shifted-window multi-head self-attention (SW-MSA), allowing informative exchange beyond local regions while maintaining computational efficiency. The outputs are again merged at the start of the subsequent stage, progressively aggregating contextual information across larger spatial scales. After four hierarchical stages, a global average pooling layer aggregates the learned features to obtain a compact representation for the target cell.

2.4. Evaluation Criteria

The downstream analyses that use the model predictions—such as differential expression analysis, pathway enrichment, or treatment outcome prediction—typically focus on relative scale or discriminative ability rather than exact agreement with fixed expression thresholds. Thus, it is common practice to evaluate the prediction performance based on how well model outputs, $\mathbf{Y}_{\text{predicted}}$, correlate with observed expression values, $\mathbf{Y}_{\text{observed}}$ (Wang et al., 2025). One standard metric is Pearson's correlation coefficient (PCC):

$$\text{PCC} = \frac{\text{Cov}(\mathbf{Y}_{\text{observed}}, \mathbf{Y}_{\text{predicted}})}{\sqrt{\text{Var}(\mathbf{Y}_{\text{observed}})\text{Var}(\mathbf{Y}_{\text{predicted}})}}. \tag{6}$$

In addition, to assess the model's ability to distinguish expressed versus non-expressed genes (AUC_0) and low versus high expression levels (AUC_1), we compute the area under the receiver operating characteristic curve. For each binary comparison, true positive rate (TPR) and false positive rate (FPR) are evaluated across a range of thresholds, and their relationship defines the AUC score.

Although AUC is traditionally designed for classification, prior studies have also relied on it to measure the discriminative power of regression models (Wang et al., 2025). In the regression setting, predicted values are first converted to binary outcomes (e.g., expressed versus non-expressed), yielding a single TPR and FPR. In contrast, classification models

output class probabilities, allowing TPR and FPR to vary with the decision threshold. To enable a fair comparison between regression and classification, we report two variants for classification models: probability AUC computed with class probabilities and prediction AUC computed from hard class assignments.

2.5. Model Specification

We compare four model variants: the original **ST-Net** as the benchmark model (He et al., 2020), its classification reformulation (**ST-Classy**), and two Swin Transformer adaptations for regression (**Swin-Reg**) and classification (**Swin-Classy**). In particular, with ST-Net as a well-established model in this application, we use its backbone and test regression setups A-C, various number of classes and weight exponents to compare performance gain with the benchmark. The best performing combination is then used in Swin-Classy. The implementation details of the four models are listed as follows.

1. **ST-Net:** We re-implemented ST-Net following its original design, which employs a DenseNet backbone for feature extraction and a regression head predicting continuous log-transformed expression. All hyperparameters were kept consistent with the original paper, including a learning rate of 10^{-6}, momentum 0.9, 50 training epochs and batch size 32. We also implement the image transformation specified in the paper for training data augmentation, $0, 90, 180, 270°$ rotation and random mirroring with 50% probability (He et al., 2020).

2. **ST-Classy:** To evaluate our classification reformulation directly with the benchmark model, we replaced ST-Net's regression head with the multi-class prediction heads (A-C) described in section 2.2 while retaining the pretrained backbone. We examined class discretizations (3 or 7 levels) and class-weight exponents (0, 0.5, 1).

3. **Swin-Reg:** For regression using Swin Transformer, we adopted the pretrained Swin-Tiny variant. The model was fine-tuned with different learning rates for the backbone

and the regression layer. Specifically, the learning rate for the backbone was fixed at 10^{-5} to ensure stability, and regression layer used a cosine-decay schedule with a peak 0.001, including linear warm-up during the first 10 epochs. The same image transformations defined in the pretrained Swin-Tiny implementation were applied during training and testing, including random resizing, horizontal flipping, color jittering and normalization (Microsoft Research, 2021).

4. **Swin-Classy:** Following findings from ST-Classy, Swin-Classy employed the 3-class setup and classification head A. All other hyperparameters mirrored Swin-Reg for consistency.

3. Experiment Results

In this section, we summarize the results from our experiment in the following aspects.

First, reframing as classification improves discriminative power across architectures. As shown in Table 1, reframing the prediction task from regression to classification markedly increases discriminative power for both DenseNet-based (ST-Net) and Swin Transformer-based models. Under comparable hyperparameter settings, classification improved predictive AUC from near-chance levels (0.5–0.6) to strong discrimination (around 0.8). These results suggest that the classification reformulation is broadly applicable and can enhance diverse backbones with minimal architectural changes.

Second, Swin Transformer outperforms benchmark even with limited neighborhood context. For the purpose of demonstrating the scalability and efficiency of Swin Transformer framework,we only used a moderate neighborhood size when assembling grid patches around target cells rather than the largest computationally feasible region. Despite this restricted spatial context, Swin Transformer based model still substantially outperformed ST-Net, underscoring the value of hierarchical context integration even at small neighborhood scales.

Third, classification performance remains robust under varying class weight-

ing schemes. Across all settings, models trained with different weight exponents achieved comparable Probability AUCs (≈ 0.9) as shown in Table 2. When predictions were derived simply from the class with the highest probability, larger class counts required stronger weighting of minority classes to maintain AUC. However, applying a data-driven threshold derived from training ROC analysis yielded similar results across weight exponents presented as adjusted prediction AUC in Table 2. This robustness reduces the need for extensive tuning, an important advantage given that gene-expression sparsity can vary substantially across tissues and gene panels.

Finally, improved class granularity and hybrid prediction heads modestly improve correlation. While classification primarily enhances discriminative ability, it may slightly reduce linear concordance measured by Pearson correlation due to discretization by quantiles and balanced tolerance for false positive and false negatives. Nevertheless, Pearson correlation remained comparable to regression baselines, and hybrid models combining regression and classification heads (Setups B and C) partially recovered correlation without compromising AUC. As shown in Table 2, models with more expression levels achieved higher PCC while maintaining similar AUCs, highlighting the flexibility of the proposed formulation.

4. Discussion

Our findings highlight that regression formulations—while intuitive—may be suboptimal for sparse and discrete-like spatial transcriptomics data. Classification-based objectives, by contrast, focus model capacity on discriminating biologically meaningful states such as expressed vs. non-expressed, rather than approximating continuous log-values prone to noise.

Additionally, the Swin Transformer's hierarchical attention allows efficient context modeling within defined neighborhoods, avoiding the quadratic cost of global attention and eliminating the need for graph modules. This design not only improves computational scalability but also aligns more closely with the spatial organization of biological tissues, where local

microenvironments drive expression heterogeneity.

Previous deep-learning models for image-based spatial expression prediction—such as ST-Net and subsequent transformer-GNN hybrids—achieved only modest correlation with observed expression. By demonstrating consistent gains in AUC across both CNN and transformer backbones, our study suggests that objective reformulation (from regression to classification) may be as consequential as architectural complexity. Furthermore, adapting Swin Transformer illustrates that hierarchical attention mechanisms can substitute for explicit spatial graphs, simplifying model design without compromising contextual learning.

It is worth noting several limitations in this work. Our experiments used a moderate-sized gene panel (460 genes) and focused on one disease context (IBD). Broader benchmarking across tissue types, sequencing platforms, and larger gene sets is needed to confirm generalizability. Additionally, while classification improves discrimination, the optimal discretization strategy (e.g., number and definition of expression levels) may vary across datasets and warrants systematic exploration.

Model	**AUC_0**	**AUC_1**	**PCC**
ST-Net	0.601	0.561	0.525
ST-Classy-3A	0.804	0.816	0.352
ST-Classy-7A	0.804	0.813	0.374
ST-Classy-3B	0.783	0.792	0.361
ST-Classy-7B	0.780	0.794	0.384
ST-Classy-3C	0.780	0.792	0.383
Swin-Reg	0.649	0.602	0.553
Swin-Classy-3A	0.760	0.781	0.392

Table 1: Comparative performance of regression and classification variants across architectures. AUC_0 quantifies discrimination between expressed and non-expressed genes, and AUC_1 between low and high expression levels. PCC denotes Pearson correlation between predicted and observed expression. Reframing as classification markedly improves discriminative AUCs over the regression benchmark (ST-Net).

Metric	**3 Classes**			**7 Classes**	
Weight Exponent	0	0.5	1	0.5	1
Prob AUC_0	0.889	0.888	0.886	0.887	0.886
Prob AUC_1	0.901	0.898	0.892	0.899	0.889
Pred AUC_0	0.573	0.670	0.791	0.607	0.729
Pred AUC_1	0.598	0.680	0.773	0.633	0.716
PCC	0.401	0.406	0.360	0.407	0.403
Adj. Pred AUC_0	0.804	**0.804**	0.782	**0.804**	0.804
Adj. Pred AUC_1	0.813	**0.816**	0.789	**0.813**	0.805
Adj. PCC	0.299	**0.352**	0.385	**0.374**	0.358

Table 2: Performance metrics for 3-class and 7-class models under Setup A. *Probability AUC* uses output probabilities across thresholds; subscripts 0 and 1 denote discrimination between expressed vs. non-expressed and low vs. high expression, respectively. *Pred AUC* uses the most probable class, while *Adj. Pred AUC* applies an optimized threshold. Finer class granularity (7 vs. 3) slightly improves PCC without compromising AUC, and class-weight exponent shows minimal effect after threshold adjustment.

References

10x Genomics (2024). Xenium human breast dataset explorer. https://www.10xgenomics.com/products/xenium-in-situ/human-breast-dataset-explorer. Accessed: 2025-10-01.

Broad Institute (2024). Dataset provided via crunchdao challenge in collaboration with broad institute. Available at https://hub.crunchdao.com/competitions/broad-1.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint*.

Fischer, A. H., Jacobson, K. A., Rose, J., and Zeller, R. (2008). Hematoxylin and eosin staining of tissue and cell sections. *Cold Spring Harbor Protocols*, 2008(5):pdb.prot4986.

He, B., Bergenstråhle, L., Ståhl, P. L., and Lundeberg, J. (2020). Integrating spatial gene expression and breast tumour morphology via deep learning. *Nature Biomedical Engineering*, 4(8):827–834.

Jia, Y., Liu, J., Chen, L., Zhao, T., and Wang, Y. (2024). Thitogene: a deep learning method for predicting spatial transcriptomics from histological images. *Briefings in Bioinformatics*, 25(1):bbad464.

Liu, Z., Lin, Y., Cao, Y., Hu, H., et al. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002.

Microsoft Research (2021). Swin transformer: Official implementation. https://github.com/microsoft/Swin-Transformer. Accessed: 2025-10-01.

Pang, M., Lin, X., and Zhang, J. (2021). Leveraging information in spatial transcriptomics to predict super-resolution gene expression from histology images in tumors. *bioRxiv*.

Rahaman, M. M. et al. (2023). Breast cancer histopathology image-based gene expression prediction using spatial transcriptomics data and deep learning. *Scientific Reports*, 13:40219.

Wang, C., Zhang, Q., and Liu, J. (2025). Benchmarking the translational potential of spatial gene expression prediction from histology. *Nature Communications*, 16(1).

Wang, H., Du, X., Liu, J., Ouyang, S., Chen, Y.-W., and Lin, L. (2024). M2ort: Many-to-one regression transformer for spatial transcriptomics prediction from histopathology images. *arXiv preprint arXiv:2401.10608*. arXiv:2401.10608.

Williams, C. G., Lee, H. J., Asatsuma, T., Vento-Tormo, R., Haque, A., et al. (2022). An introduction to spatial transcriptomics for biomedical research. *Genome Medicine*, 14(1):68.

Xiao, X. et al. (2024). Transformer with convolution and graph-node co-embedding: An accurate and interpretable vision backbone for predicting gene expressions from local histopathological images. *Medical Image Analysis*, 91:103040.

Zeng, Y., Yang, Z., and Chen, L. (2022). Spatial transcriptomics prediction from histology jointly through transformer and graph neural networks. *Briefings in Bioinformatics*, 23(5):bbac297.