Accurate Spatial Gene Expression Prediction by Integrating Multi-Resolution Features

Youngmin Chung, Ji Hun Ha, Kyeong Chan Im, Joo Sang Lee*
Sungkyunkwan University, South Korea
ymblue@g.skku.edu, joosang.lee@skku.edu

Abstract

Recent advancements in Spatial Transcriptomics (ST) technology have facilitated detailed gene expression analysis within tissue contexts. However, the high costs and methodological limitations of ST necessitate a more robust predictive model. In response, this paper introduces TRIPLEX, a novel deep learning framework designed to predict spatial gene expression from Whole Slide Images (WSIs). TRIPLEX uniquely harnesses multi-resolution features, capturing cellular morphology at individual spots, the local context around these spots, and the global tissue organization. By integrating these features through an effective fusion strategy, TRIPLEX achieves accurate gene expression prediction. Our comprehensive benchmark study, conducted on three public ST datasets and supplemented with Visium data from 10X Genomics, demonstrates that TRIPLEX outperforms current state-of-the-art models in Mean Squared Error (MSE), Mean Absolute Error (MAE), and Pearson Correlation Coefficient (PCC). The model's predictions align closely with ground truth gene expression profiles and tumor annotations, underscoring TRIPLEX's potential in advancing cancer diagnosis and treatment.

1. Introduction

The emergence of large-scale Spatial Transcriptomics (ST) technology has facilitated the quantification of mRNA expression across a multitude of genes within the spatial context of tissue samples [22]. ST technology segments centimeter-scale Whole Slide Images (WSIs) into hundreds of thousands of small spots, each providing its gene expression profile. Considering the substantial cost associated with ST sequencing technology, coupled with the widespread availability of WSIs, a pressing question is how to best predict spatial gene expression based on WSIs using rapidly evolving computer vision techniques.

A number of studies have endeavored to address this challenge [7, 18, 24–26]. Approaches vary, with some predicting gene expression strictly from the tissue image confined within the spot's boundaries [7], while others also take into account spatial dependencies between spot images [18, 26], or consider similarities to reference spots [24, 25]. However, we have noted several limitations inherent to these existing methodologies. Firstly, current methods primarily focus on spot images, neglecting the wealth of biological information available in the wider image context. By integrating both the specific spot and its surrounding environment, along with the holistic view of the entire histology image, we can access richer information, encompassing varied biological contexts. Secondly, models that consider interactions between spots [18, 26] face a limitation in processing the embedding of all patches in a WSI simultaneously. This approach, common in handling hundreds to thousands of patches within a WSI, limits the scalability of the patch embedding model due to resource constraints. Such limitations significantly impede the extraction of fine-grained, rich representations from each spot, thereby affecting the model's ability to perform detailed analysis of WSIs. Thirdly, model performance is frequently overestimated because of inadequate validation, such as using the limited size of dataset [7] sometimes without cross-validation [24] and training/testing with replicates from the same patient [18, 25, 26]. The limited size of ST datasets means that exclusive reliance on a single dataset for model evaluation can hinder an accurate assessment of the model's capabilities, thereby emphasizing the necessity for cross-validation. The issue is compounded when replicate data from the same patient, often featuring nearly identical image-gene expression pairs, are used in both training and testing phases. This can lead to an inflated perception of a model's effectiveness, as it may not accurately reflect the model's ability to generalize to new, unseen data. Lastly, the use of disparate datasets, diverse normalization methods, and varied evaluation techniques in existing research studies compounds the challenge of conducting fair comparisons of the models.

Addressing these limitations, we present TRIPLEX, an innovative deep learning framework designed to leverage

*Corresponding author.

multi-resolution features from histology images for robustly predicting spatial gene expression levels. TRIPLEX extracts three distinct types of features corresponding to different resolutions: the target spot image (representing the specific spot, whose gene expression to be predicted), the neighbor view (encompassing a wider area around the spot), and the global view (comprising the aggregate of all spot images). These features capture varying levels of biological information—ranging from the detailed cell morphology in the target spot image, to the surrounding tissue phenotype, and the overall tissue microenvironment in the WSI. Each is integral to understanding the spatial gene expression levels of the given spot. TRIPLEX employs separate encoders to extract these features from WSIs, each focusing on its assigned resolution to efficiently capture relevant details. For neighbour or global view with larger resolution, pre-extracted features are used to reduce the burden of computational cost, while for target spot images, encoders are fully updated to extract fine-grained information. These features are then integrated via a fusion layer for effective gene expression prediction. This approach allows TRIPLEX to utilize resolution-specific information, thereby enhancing prediction accuracy while avoiding significant increases in computational costs.

Our study sets a new benchmark in spatial gene expression prediction, comparing our model, TRIPLEX, against five prior studies [7, 18, 24–26] under uniform experimental conditions. We conduct internal evaluations using three public Spatial Transcriptomics (ST) datasets [1, 7, 10] and external validations using higher-resolution Visium data from 10X Genomics. Our validation procedure strictly avoids mixing patient sample replicates between training and testing datasets, a significant departure from previous methods [18, 25, 26], and employs rigorous cross-validation. Our results indicate that TRIPLEX surpasses existing models in terms of Mean Squared Error (MSE), Mean Absolute Error (MAE), and Pearson Correlation Coefficient (PCC) in both internal and external evaluations. Furthermore, we provide visualizations of the expression distributions for a specific gene commonly associated with cancer. These visualizations reveal that our model's predictions align more closely with actual gene expression data and tumor annotations, demonstrating its enhanced predictive accuracy.

Our key contributions can be summarized as follows:

- We introduce an innovative approach to predict spatial gene expression levels from WSIs by integrating multiple biological contexts.
- Our proposed framework seamlessly integrates multi-resolution features. This integration is facilitated by a feature extraction strategy, the use of various types of transformers, and a fusion loss technique, all while keeping the additional computational costs to a minimum.
- Through comprehensive experiments on three public ST datasets and additional external evaluations using three Visium data, our study establishes a new benchmark in the field of spatial gene expression prediction. The results consistently show that our proposed method outperforms all existing models included in our comparative analysis.

2. Related Work

In this section, we delve into studies pertinent to our research. For clarity, the term 'spot' will be used to denote a predefined unit region within a WSI where gene expression is quantified. Moreover, we will use 'target spot' to specifically refer to the spot within a WSI for which we seek to predict gene expression.

Spatial gene expression prediction from WSIs via deep learning We review the pioneering works in this field, which aim to predict spatial gene expression from WSIs. ST-Net [7] utilizes a standard transfer learning strategy, training a Densenet121 model [8]—pretrained on ImageNet—using histology images as input and gene expression as labels. Following this, HisToGene [18] leverages Vision Transformers (ViT) [6] to account for correlations among patches in a WSI, thereby predicting gene expression from global-context aware features. Further developing this concept, Hist2ST [26] enhances the approach by emphasizing patch embedding using ConvMixer [23] and aggregating neighborhood information through graph convolution network [14]. While these previous works [18, 26] share similarities with our methodology, they predominantly process patch and global embedding sequentially, often overlooking the neighboring information around the target spot. In contrast, our method sets itself apart by concurrently extracting critical features at three distinct resolutions, including the neighbor view, and integrating them for gene expression prediction. In a different vein, EGN [25] adopts exemplar learning for predicting gene expression from histology images. This method dynamically selects the most analogous exemplars from a target spot within a WSI to enhance prediction accuracy. Additionally, BLEEP [24] introduces a bi-modal embedding framework similar to CLIP [19] to co-embed spot images and gene expression. After training, this model imputes the gene expression of a query spot using the retrieved gene expression set from a reference dataset.

Deep learning for WSIs Due to their gigapixel resolution, WSIs present a significant challenge for conventional deep learning frameworks in computer vision. To tackle this, Multiple Instance Learning (MIL) has been employed, enabling the handling of high-resolution data with sparse local annotations. In the context of WSIs, MIL approaches typically predict bag labels, such as distinguishing slides from cancer patients versus healthy individuals, by aggregating information from numerous small patches within the

WSIs [2, 12, 15, 20]. Recently, attention-based networks have been employed in MIL to aggregate all patches in WSIs, achieving state-of-the-art performance [15, 16, 20]. Moreover, Chen et al. [3] introduced a Hierarchical Image Pyramid Transformer (HIPT) to adapt Vision Transformers (ViT) for WSIs. They effectively captured the hierarchical structures of WSIs by sequentially training ViTs across images of various resolutions, employing a self-supervised learning approach. This method has shown superior results in cancer subtyping and survival prediction, surpassing previous models. Drawing inspiration from this, our proposed method is specifically designed to simultaneously handle information from these multi-resolution features for predicting spatial gene expression.

3. Method

3.1. Preliminary

In this section, we present our problem formulation and detail our proposed method, concurrently assessing it in contrast to previous methodologies. Our task is a multi-output regression problem, where we input a set of spot images from a WSI, denoted as $\mathbf{X} \in \mathbb{R}^{n \times H \times W \times 3}$, and aim to predict the gene expression levels of individual spots, represented as $\mathbf{Y} \in \mathbb{R}^{n \times m}$. Here, n denotes the number of spot images in a WSI, m represents the number of genes whose expression levels to be predicted, and H and W signify the height and width of each spot image, respectively.

ST-Net [7] approaches the prediction task by estimating $\hat{Y}_i \in \mathbb{R}^m$ based solely on the information from a single spot image $X_i \in \mathbb{R}^{H \times W \times 3}$. This is represented as:

$$\hat{Y}_i = f(X_i) \in \mathbb{R}^m \tag{1}$$

where *i* indexes a target spot within a WSI. ST-Net is formulated under general supervised learning principles, where each input corresponds to a unique label, and each input is treated independently.

On the other side, methods employing ViT [18, 26] predict the gene expression of all spot images concurrently, expressed as:

$$\hat{Y} = f(X_1, X_2, ..., X_n) \in \mathbb{R}^{n \times m} \tag{2}$$

In this case, the problem is defined within the context of supervised learning, but with a key difference: the inputs are interdependent, affecting the prediction outcome collectively.

EGN [25] approaches the problem from a different perspective and predicts $\hat{Y}_i$ using X_i together with its global view G_i and its exemplar set $\{g_j, y_j\}_{j=1}^k \in K_i$. This can be formulated as:

$$\hat{Y}_i = f(X_i, G_i, K_i) \in \mathbb{R}^m \tag{3}$$

where G_i represents the features extracted from the target spot image X_i using a pretrained model. The set K_i comprises the k-nearest global views to G_i and their associated gene expression levels.

In BLEEP [24], a bi-modal pretraining phase is leveraged, employing contrastive loss [19] to generate embeddings for both images and gene expression levels. During the inference phase, a given query image X_i is processed through an image encoder, Enc^{img}, to produce its d-dimensional embedding vector v_i:

$$v_i = Enc^{img}(X_i) \in \mathbb{R}^d \tag{4}$$

Simultaneously, an expression encoder, Enc^{exp}, is applied to the gene expression levels of the reference dataset, yielding an d-dimensional embedding vectors e^{ref} represented as:

$$e^{ref} = Enc^{exp}(Y^{ref}) \in \mathbb{R}^{l \times d} \tag{5}$$

Here, l signifies the number of gene expression levels in the reference dataset. The process continues by identifying the top-k closest embeddings to v_i, denoted as $e^{top} \in \mathbb{R}^{k \times d}$. The corresponding gene expression values for these top-k embeddings, $Y^{top} \in \mathbb{R}^{k \times m}$, are then retrieved. The gene expression level for the query spot is estimated by computing the average of Y^{top}:

$$\hat{Y}_i = Average(Y^{top}) \in \mathbb{R}^m \tag{6}$$

Our proposed method diverges from the previously mentioned approaches by employing a unique combination of three input types to predict $\hat{Y}_i$. These are the target spot image X_i^{Ta}, the local neighbor views X_i^{Ne}, and a collection of all the global views G: $g_1, g_2, ..., gn$.

$$z_i^{Ta} = Enc^{Ta}(X_i^{Ta}) \in \mathbb{R}^{n^{Ta} \times d} \tag{7}$$

$$z_i^{Ne} = Enc^{Ne}(X_i^{Ne}) \in \mathbb{R}^{n^{Ne} \times d} \tag{8}$$

$$z^{Gl} = Enc^{Gl}(G) \in \mathbb{R}^{n \times d} \tag{9}$$

$$\hat{Y}i = f(z_i^{Ta}, z_i^{Ne}, z_i^{Gl}) \in \mathbb{R}^m \tag{10}$$

Here, Enc^{Ta}, Enc^{Ne}, and Enc^{Gl} represent models that independently embed each type of input. The dimension of each embedded token is denoted by d. The local neighbor views X_i^{Ne} consist of the n^{Ne} adjacent patches around the target spot image X_i^{Ta}, and n^{Ta} signifies the number of tokens derived from X_i^{Ta}. In all cases, f refers to a neural network function outputting gene expression levels.

3.2. TRIPLEX

The overall workflow of our method is illustrated in Figure 1. Initially, we process the global view, derived from all spot images of a WSI, through a global encoder to produce global tokens. Although these global tokens capture

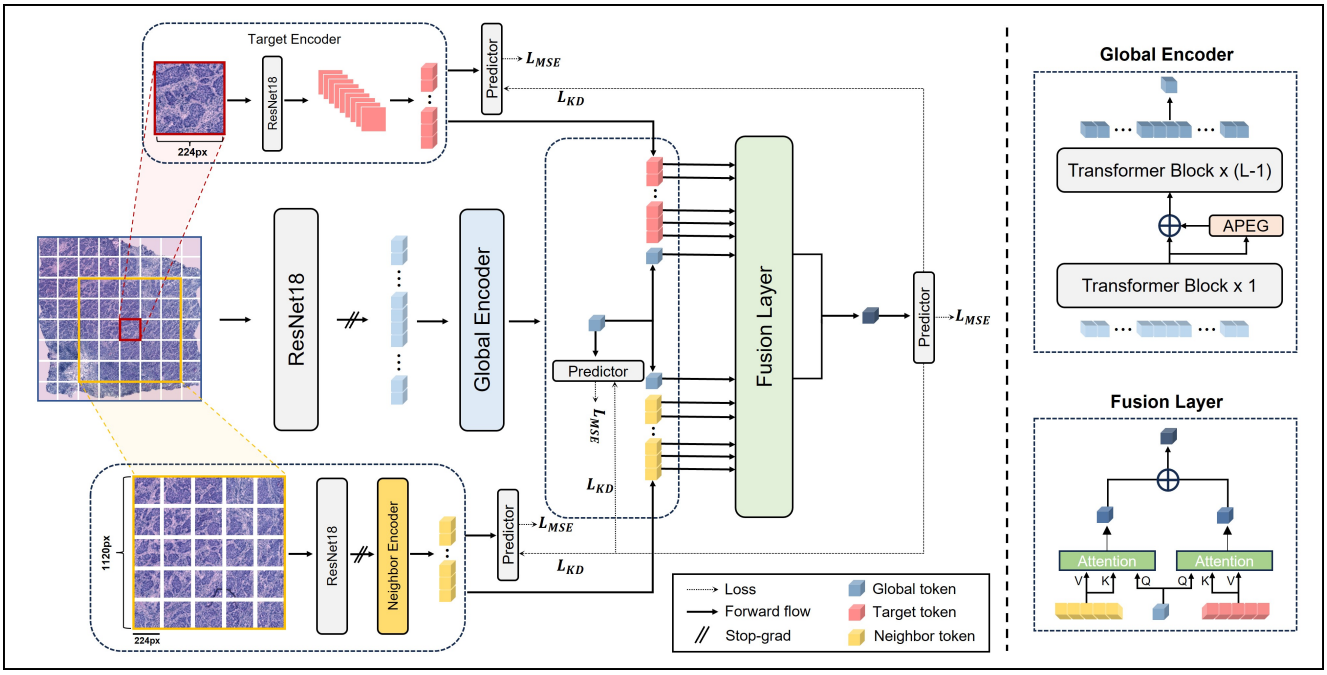

Figure 1. Schematic representation of the TRIPLEX. The global encoder processes the global view, while separate encoders handle the target spot image and neighbor view. A fusion layer, incorporated with fusion loss, facilitates the effective integration of these tokens to predict gene expression levels.

the macroscopic spatial distribution and inter-spot correlations, they might not adequately represent the detailed information specific to each target spot. To address this, we independently encode the target spot image and its neighboring views, generating tokens that encapsulate finer details. These are then integrated with the global tokens via fusion layers, enriching the global representation with specific target-related information. Additionally, considering the diverse contextual information provided by different input sources, we have implemented a fusion loss mechanism, inspired by knowledge distillation, to enhance the efficacy of the fusion process. Detailed descriptions of the individual components involved in our proposed model will be provided in the following sections.

3.3. Embedding Global Information

Processing all 224x224-sized images of spots starting from patch embedding is computationally intensive since the number of spots typically ranges from hundreds to thousands. To address this, we employ a feature extraction strategy commonly used in the MIL approach for WSIs. Specifically, we utilize a ResNet18 model pretrained on large-scale histology images for feature extraction [5]. The features thus obtained serve as input to the global encoder, which generates global tokens. This encoder comprises a series of transformer blocks and a Position Encoding Generator (PEG) [4], adept at encoding positional information for a variable number of spots. Given that PEG was originally designed for natural images with regular, square shapes, we modify it for our specific use-case where the image shape is irregular. Our adaptation, termed the Atypical Position Encoding Generator (APEG), imbues tokens with absolute positional information pertinent to all spots in a WSI. This modification is crucial for effectively capturing the spatial distribution within the WSI.

Atypical Position Encoding Generator (APEG) In our APEG framework, after the initial transformation of tokens through one transformer block, we employ a technique to re-establish their relative positional context. This involves reshaping the global tokens $z^{Gl} \in \mathbb{R}^{n \times d}$ into a spatial format $\hat{z}^{Gl} \in \mathbb{R}^{h \times w \times d}$. Here, h and w represent the maximum values of the x- and y-coordinates, respectively. During this process, any voids in the spatial arrangement are temporarily filled with zeros. Subsequently, we apply convolutional layers to these reshaped tokens. This step includes refilling the areas that were previously vacant with zeros. After this convolutional processing, we revert the tokens back to their original format. This method enables us to effectively incorporate the relative spatial information of the tokens, enhancing the positional encoding within the WSI framework.

3.4. Embedding Target/Neighbor Information

We independently encode the image of the target spot and its neighboring view, generating tokens for both target

and neighbor that are rich in contextual information.

Embedding target information To encode the target spot image, we utilize a ResNet18 architecture, excluding the global average pooling and the fully-connected layer. This encoder processes each 224x224 target spot image by embedding it into 49 distinct features, with each feature having a dimension of 512. Notably, while this instance of ResNet18 is initialized with the same weights as before, it undergoes unique updates during the training process, ensuring tailored feature extraction for each target spot image.

Embedding neighbor information For the neighbor view, we use the surrounding 1120x1120 image of each target spot. This choice is based on images directly adjacent to the target spot, rather than a group of neighboring spot images. This approach addresses the issue of non-uniform alignment and spacing between spots in ST data (refer to supplementary Figure 1 for details). In cases where a target spot is at the edge of the slide, zero padding is applied to maintain the required image size. This 1120x1120 neighbor view is then embedded into twenty-five 512-dimensional feature vectors using the ResNet18. These vectors then serve as input for the neighbor encoder. The embedding process within the neighbor encoder involves a sequence of self-attention blocks, each integrated with relative position encoding [21]. Although the ResNet18 weights used for feature extraction remain fixed, the weights of the neighbor encoder are dynamically updated during training. Further details about the neighbor encoder are provided in the supplementary material.

3.5. Integrating global, neighbor, and target information

To achieve an effective exchange of information and enhanced contextual understanding between different token types, we implement cross-attention layers. In this setup, the global token acts as the query (Q), with the target and neighboring tokens within the WSI forming the key (K) and value (V) pairs. These Q, K, and V components are utilized in a dot-product attention mechanism, which is crucial for assessing the relative importance of each target and neighbor token in comparison to the global token. This process is instrumental in developing a comprehensive understanding of the local and global contexts at each spot, thereby augmenting the model's performance in complex gene expression prediction tasks. The integration of these insights is realized by summing the resultant tokens from the cross-attention layers:

$$z_i^{GT} = CrossAttn(z_i^{Gl}, z_i^{Ta}) \in \mathbb{R}^d \tag{11}$$

$$z_i^{GN} = CrossAttn(z_i^{Gl}, z_i^{Ne}) \in \mathbb{R}^d \tag{12}$$

$$z_i^{GTN} = Sum(z_i^{GT}, z_i^{GN}) \in \mathbb{R}^d \tag{13}$$

In these equations, z_i^{GT} and z_i^{GN} represent the cross-attention results for the target and neighbor tokens, respectively, relative to the global token. z_i^{GTN} is the summation of these two tokens, which is used to estimate the gene expression levels. During the process, information exchange between the neighbor token and the target token occurs only through the global tokens. In this manner, we can efficiently integrate the three pieces of information at a minimal additional cost.

Fusion Loss and Objective Function To optimize the integration of information from multiple tokens, we have introduced a fusion loss mechanism. This approach capitalizes on the rich, gene expression-relevant information inherent in the fusion token, which synthesizes data from the target, neighbor, and global tokens. By transferring knowledge from this fusion token to other individual tokens, we significantly enhance the model's predictive accuracy for gene expression levels. In practice, fully-connected layers, tasked with predicting gene expression levels, are attached to each of the target, neighbor, and global tokens, with average pooling applied beforehand to the target and neighbor tokens. The optimization process involves two key components: 1) minimizing Mean Squared Error (MSE) loss between individual predictors' outputs and ground-truth values, 2) reducing the MSE loss between each predictor's output and the 'soft target,' which are the predictions derived from the fusion token.

The loss for a given j_{th} token (target, neighbor, or global) is computed as follows:

$$L^j = (1-\alpha)\frac{1}{m}\sum_{k=1}^{m}\left\|q_k^j - y_k\right\|_2^2 + \alpha\frac{1}{m}\sum_{k=1}^{m}\left\|q_k^j - q_k^F\right\|_2^2, \tag{14}$$

where q_k^j is the prediction for the k_{th} gene by the j_{th} token, $q^F k$ is the fusion token's prediction for the k_{th} gene, and α is a hyperparameter balancing the two aspects of the loss. For the fusion token, we calculate the MSE loss in relation to the actual labels:

$$L^F = \frac{1}{m}\sum_{k=1}^{m}\left\|q_k^F - y_k\right\|_2^2 \tag{15}$$

Ultimately, we optimize the following object function:

$$L = \sum_{j}^{3} L^j + L^F. \tag{16}$$

4. Experiments

In this section, we outline the specifics of the ST data employed in our model's training, detail our experimental setup and evaluation metrics, and provide implementation details. For more details on experiment settings, please refer to Section 2 of the supplementary material.

Source	Model	BC1			BC2			SCC		
		MSE	PCC(M)	PCC(H)	MSE	PCC(M)	PCC(H)	MSE	PCC(M)	PCC(H)
Local	ST-Net [7]	0.260 ± 0.04	0.194 ± 0.11	0.345 ± 0.16	0.209 ± 0.02	0.116 ± 0.06	0.223 ± 0.10	0.294 ± 0.07	0.274 ± 0.08	0.382 ± 0.08
	EGN [25]	0.241 ± 0.06	0.197 ± 0.11	0.328 ± 0.17	0.192 ± 0.02	0.111 ± 0.05	0.203 ± 0.09	0.281 ± 0.08	0.281 ± 0.06	0.388 ± 0.06
	BLEEP [24]	0.277 ± 0.05	0.151 ± 0.11	0.277 ± 0.16	0.235 ± 0.02	0.095 ± 0.05	0.193 ± 0.10	0.297 ± 0.05	0.269 ± 0.07	0.396 ± 0.08
	TEM	0.252 ± 0.04	0.208 ± 0.11	0.365 ± 0.15	0.190 ± 0.02	0.119 ± 0.06	0.227 ± 0.10	0.290 ± 0.06	0.296 ± 0.07	0.402 ± 0.08
	NEM	0.278 ± 0.08	0.255 ± 0.13	0.424 ± 0.18	0.193 ± 0.03	0.152 ± 0.05	0.277 ± 0.09	0.373 ± 0.14	0.308 ± 0.05	0.444 ± 0.06
Global	HistoGene [18]	0.314 ± 0.09	0.168 ± 0.12	0.302 ± 0.19	0.194 ± 0.05	0.100 ± 0.05	0.219 ± 0.12	0.270 ± 0.09	0.133 ± 0.06	0.261 ± 0.13
	GEM	0.253 ± 0.06	0.295 ± 0.14	0.491 ± 0.17	0.221 ± 0.03	0.193 ± 0.06	0.341 ± 0.08	0.317 ± 0.16	0.276 ± 0.09	0.392 ± 0.08
Local+Global	Hist2ST [26]	0.285 ± 0.08	0.118 ± 0.10	0.248 ± 0.17	$\mathbf{0.181} \pm 0.02$	0.044 ± 0.02	0.099 ± 0.03	1.291 ± 0.65	0.004 ± 0.01	0.053 ± 0.01
Multiple	TRIPLEX	$\mathbf{0.228} \pm 0.07$	$\mathbf{0.314} \pm 0.14$	$\mathbf{0.497} \pm 0.17$	0.202 ± 0.02	$\mathbf{0.206} \pm 0.07$	$\mathbf{0.352} \pm 0.10$	$\mathbf{0.268} \pm 0.09$	$\mathbf{0.374} \pm 0.07$	$\mathbf{0.490} \pm 0.07$

Table 1. Cross validation result on each ST dataset. PCC(M) and PCC(H) denote the mean PCC for all genes and the mean PCC for highly predictive genes, respectively. The mean and standard deviation of cross-validation results are displayed. MAEs are excluded due to space limitations and can be found in the supplementary.

ST dataset Spatial Transcriptomics (ST) data is characterized by its compilation of numerous spot images within a single slide, each accompanied by corresponding gene expression values. A typical ST dataset contains several hundred spatially resolved spots, with each spot representing the expression values of around 20,000 genes. Visium, the next iteration of ST data, expands this scope to include thousands of spots, each still characterized by the expressions of a similar number of genes. For our internal validation, we utilize two breast cancer ST datasets [1, 7] and one skin cancer ST dataset [10]. External validation is conducted using Visium data from three independent breast cancer patients. We refer to the breast cancer ST dataset from [1] as the BC1 dataset, the one from [7] as the BC2 dataset, and the skin cancer dataset from [10] as the SCC dataset.

Experiment Setup and Evaluation Metrics To mitigate potential overfitting due to the limited size of our datasets, we employ cross-validation for model performance evaluation across the three ST datasets. Consistent with our earlier mention, we ensure that samples from the same patients are exclusively allocated to either the training or the test dataset, avoiding any overlap. Specifically, we adopt a leave-one-patient-out cross-validation approach for the BC1 dataset (n sample=36, n patient=8) and the SCC dataset (n sample=12, n patient=4), using samples from a single patient for sequential validation. For the BC2 dataset, which has a larger sample size (n sample=68, n patient=23), we conduct 8-fold cross-validation, with careful consideration to keep samples from the same patient within the same data partition. To extend our model's evaluation to independent datasets, we use three breast cancer Visium datasets from 10x Genomics, training on the BC1 dataset and testing on each Visium dataset. Our evaluation metrics include the Pearson correlation coefficient (PCC), mean squared error (MSE), and mean absolute error (MAE). PCC is computed for each gene across all spots in a sample, and we report both the mean PCC for all genes (PCC(M)) and the mean PCC for highly predictive genes (PCC(H)). The highly predictive genes are identified through a cross-validation ranking process, where the top 50 genes are determined based on their average rank across all folds.

Implementation Details For preprocessing, we crop each spot image to 224x224 pixels using the center coordinates of each spot. Neighbor views are obtained by capturing a 1120x1120 image centered on the spot, which is then subdivided into 25 equal-sized sub-images. We select 250 genes for each dataset following the criteria in [7]. The gene expression values are normalized by dividing by the sum of expressions in each spot, followed by a log transformation. To mitigate experimental noise, we adopt the smoothing approach from [7], averaging the gene expression values of each spot with those of its adjacent neighbors. Our model is optimized using the Adam optimizer [13] with an initial learning rate of 0.0001. The learning rate is adjusted dynamically using a Step LR scheduler, with a step size of 50 and a decay rate of 0.9. During training, we use a batch size of 128. In testing, the model's performance is evaluated on all spots in each WSI per batch, and we report the mean of all validation performances.

4.1. Cross-validation Performance of TRIPLEX

We conduct cross-validation using the three ST datasets to assess TRIPLEX's performance in comparison to baseline models. The total counts of spot images in the BC1, BC2, and SCC datasets are 13,620, 68,050, and 23,205, respectively.

Baselines Our model's performance was benchmarked against existing models, including 1) local-based models (ST-Net[7], EGN[25], BLEEP[24]) and 2) global-based models (HisToGene[18], Hist2ST[26]). For a consistent evaluation, the same ResNet18 used in TRIPLEX was applied as the feature extractor in EGN and the image encoder in BLEEP. We also compared TRIPLEX with simpler models focusing on single information types: Target Encoding Model (TEM), Neighbor Encoding Model (NEM), and Global Encoding Model (GEM). Implementation details for all baseline models are available in Section 2 of the supplementary material.

Result Comparison As shown in Table 1, TRIPLEX outperforms all previous models and demonstrates superior

Source	Model	10X Visium-1				10X Visium-2				10X Visium-3			
		MSE	MAE	PCC(M)	PCC(H)	MSE	MAE	PCC(M)	PCC(H)	MSE	MAE	PCC(M)	PCC(H)
Local	ST-Net [7]	0.423	0.505	-0.026	-0.000	0.395	0.492	0.091	0.193	0.424	0.508	-0.032	0.008
	EGN [25]	0.421	0.512	0.003	0.024	0.328	0.443	0.102	0.157	0.303	0.425	0.106	0.220
	BLEEP [24]	0.367	0.470	0.106	0.221	0.289	0.406	0.104	0.260	0.298	0.415	0.114	0.229
	TEM	**0.339**	**0.453**	0.024	0.093	**0.278**	**0.402**	0.106	0.218	0.290	0.412	0.078	0.193
	NEM	0.444	0.515	0.089	0.259	0.391	0.482	0.105	0.290	0.393	0.483	0.036	0.175
Global	GEM	0.392	0.494	0.132	**0.269**	0.397	0.482	0.056	0.166	0.394	0.488	0.082	0.191
Multiple	TRIPLEX	0.351	0.464	**0.136**	0.241	0.282	0.407	**0.155**	**0.356**	**0.285**	**0.410**	**0.118**	**0.282**

Table 2. Generalization performance comparison between other models and ours by PCC(M), PCC(H), MSE, and MAE.

performance over the individual modules within TRIPLEX across most evaluation metrics. Notably, GEM demonstrates substantial effectiveness, underscoring the crucial role of global interactions in accurately predicting gene expression. However, TRIPLEX, which integrates the target and neighbor view information with the global view, yields the most notable improvement. Compared to EGN, one of the best-performing existing models, TRIPLEX achieves substantial increases in PCC(M) and PCC(H) across all datasets. Specifically, in the BC1 dataset, there is an improvement of 0.117 in PCC(M) and 0.169 in PCC(H); in the BC2 dataset, an increase of 0.095 in PCC(M) and 0.149 in PCC(H); and in the SCC dataset, a rise of 0.093 in PCC(M) and 0.102 in PCC(H). Despite integrating three types of information, TRIPLEX maintains a parameter count comparable to other top-performing models (see supplementary table 1). The variances between our results and those reported in the original works of the baseline models can be attributed to differences in 1) cross-validation strategies, 2) normalization methods, and 3) metric calculations. For an in-depth explanation of these differences, refer to Section 3 of the supplementary material.

4.2. Generalization performance of TRIPLEX

For external validation, we preprocess Visium data similarly to the ST data and evaluate model performance on three individual Visium samples.

Result comparison In this set of experiments, TRIPLEX is benchmarked against the three best-performing models identified in prior tests. According to Table 2, TRIPLEX shows robust performance on unseen Visium data, which is formatted differently from the training data. It outperforms existing models (ST-Net, EGN, BLEEP) across MSE, MAE, PCC (M), and PCC (H). Notably, TRIPLEX consistently achieves impressive PCC (H) scores (average 0.293), indicating its potential applicability in clinical settings.

4.3. Visualization of Cancer Marker Genes

Among the highly predictive genes identified by our model are known breast cancer markers such as CLDN4 and GNAS [11, 17], which could aid pathological diagnosis. In the cross-validation for the GNAS gene, TRIPLEX significantly outperforms all existing models in both BC1 and BC2 datasets. For instance, in the BC1 dataset, the PCCs are 0.359 (HisToGene), 0.286 (Hist2ST), 0.411 (ST-Net), 0.374 (EGN), 0.338 (BLEEP), and 0.583 (TRIPLEX). Similarly, in the BC2 dataset, the PCCs are 0.282 (HisToGene), 0.138 (Hist2ST), 0.371 (ST-Net), 0.341 (EGN), 0.0669 (BLEEP), and 0.554 (TRIPLEX). We further provide visualizations of the predicted values for GNAS alongside its ground truth values for each dataset. As depicted in Figure 2, we notice that TRIPLEX not only exhibits a higher PCC with the actual ground truth values but also demonstrates a greater visual congruence with the actual gene expression patterns. Consequently, it appears to be more effective in assisting pathologists in clinical diagnostics. More visualizations are available in the supplementary material.

4.4. Ablation study

In this section, we demonstrate the contributions of each method to gene expression prediction through an ablation study of our proposed model. Specifically, we aim to observe the contributions of the three modules in our model and investigate the extent to which the position encoding generator and our proposed fusion approach contribute to gene expression prediction. Here we only present experimental results on the SCC dataset, but we have observed similar outcomes on other datasets as well. Additional experiment results can be found in supplementary material.

Individual Modules We assess how the fusion of information from TEM, NEM, and GEM enhances gene expression prediction accuracy. According to Table 3, it is clear that incorporating features from all three modules achieves the best performance. Notably, the exclusion of the NEM module results in a substantial decrease in PCC(M), highlighting the critical influence of neighboring interactions in gene expression prediction in TRIPLEX. While the absence of the GEM module does not significantly impact performance metrics such as MSE and MAE in this specific dataset, ablation studies conducted on alternative datasets reveal the significant contribution of global interactions to the accuracy of gene expression level predictions.

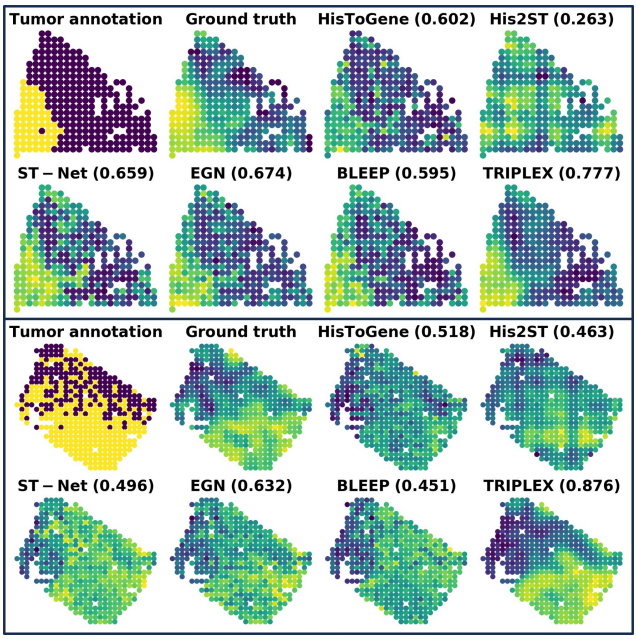

Figure 2. The visualization includes tumor region annotations by pathologists, ground truth for GNAS expression levels, and predicted GNAS expression levels from HisToGene, Hist2ST, ST-Net, EGN, BLEEP, and TRIPLEX, in samples from datasets BC1 and BC2. The PCC between the ground truth and predicted values is displayed for each model.

	MSE	MAE	PCC(M)	PCC(T)
w/o TEM	0.289	0.419	0.352	0.471
w/o NEM	0.271	0.408	0.330	0.439
w/o GEM	**0.263**	**0.402**	0.358	0.481
TRIPLEX	0.268	0.404	**0.374**	**0.490**

Table 3. Ablation study for the individual modules

Position Encoding Generator (PEG) We further compare our APEG to different methods of positional encoding, including (1) without PEG and (2) with PEG [4], as shown in Table 4. In the conventional PEG implementation, we add zero padding to the global token to make the number of tokens a perfect square, followed by the standard squaring procedure. The results indicate that APEG not only outperforms the model without any PEG but also surpasses the conventional PEG. This implies that APEG more effectively encodes spatial distribution within the global tokens, enhancing the overall model performance.

Fusion Method and Fusion Loss We also investigate the impact of different fusion methods for multi-resolution features, comparing our model's fusion layer with traditional feature fusion techniques like summation, concatenation, and attentional pooling [9]. Additionally, we assess the contribution of fusion loss to the model's performance. As shown in Table 5, integrating multi-resolution features using our fusion layer yields superior results compared to these conventional techniques. Moreover, the marked improvement in performance when incorporating fusion loss indicates its effectiveness in integrating the three types of features. This result highlights the significance of our fusion approach in achieving high accuracy in gene expression prediction.

	MSE	MAE	PCC(M)	PCC(T)
w/o PEG	0.280	0.413	0.360	0.480
PEG [4]	0.276	0.411	0.364	0.479
APEG	**0.268**	**0.404**	**0.374**	**0.490**

Table 4. Ablation study PEG

Fusion method	MSE	MAE	PCC(M)	PCC(T)
Summation	0.297	0.425	0.341	0.464
Concatenation	0.293	0.422	0.348	0.474
Attentional pooling	0.293	0.423	0.353	0.473
fusion layer	**0.268**	**0.404**	**0.374**	**0.490**
Fusion loss	MSE	MAE	PCC(M)	PCC(T)
w/o fusion loss	0.292	0.423	0.358	0.469
w/ fusion loss	**0.268**	**0.404**	**0.374**	**0.490**

Table 5. Ablation studies for fusion method and fusion loss

5. Conclusion

We demonstrate a novel approach for predicting spatial gene expression patterns from WSIs. By incorporating multiple sources of information utilizing various types of transformer and our proposed fusion method, TRIPLEX achieves superior performance compared to all existing approaches in both internal and external evaluations. TRIPLEX has the potential to improve the accuracy and robustness of the predictions for spatial gene expression distribution, paving the way for new discoveries at the interface of WSIs and sequencing.

Acknowledgments This work was partly supported by Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No.2019-0-00421, AI Graduate School Support Program(Sungkyunkwan University)) and the Samsung Research Funding & Incubation Center of Samsung Electronics under Project SRFC-MA2102-05. This work is a study partly supported by domestic scholarships funded by the Kwanjeong Educational Foundation (KEF1464).

References

[1] Alma Andersson, Ludvig Larsson, Linnea Stenbeck, Fredrik Salmén, Anna Ehinger, Sunny Wu, Ghamdan Al-Eryani, Daniel Roden, Alex Swarbrick, Åke Borg, et al. Spatial deconvolution of her2-positive breast tumors reveals novel intercellular relationships. *bioRxiv*, 2020. 2, 6

[2] Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8):1301–1309, 2019. 3

[3] Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16144–16155, 2022. 3

[4] Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Conditional positional encodings for vision transformers. *arXiv preprint arXiv:2102.10882*, 2021. 4, 8

[5] Ozan Ciga, Tony Xu, and Anne Louise Martel. Self supervised contrastive learning for digital histopathology. *Machine Learning with Applications*, 7:100198, 2022. 4

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[7] Bryan He, Ludvig Bergenstråhle, Linnea Stenbeck, Abubakar Abid, Alma Andersson, Åke Borg, Jonas Maaskola, Joakim Lundeberg, and James Zou. Integrating spatial gene expression and breast tumour morphology via deep learning. *Nature biomedical engineering*, 4(8): 827–834, 2020. 1, 2, 3, 6, 7

[8] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 2

[9] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018. 8

[10] Andrew L Ji, Adam J Rubin, Kim Thrane, Sizun Jiang, David L Reynolds, Robin M Meyers, Margaret G Guo, Benson M George, Annelie Mollbrink, Joseph Bergenstråhle, et al. Multimodal analysis of composition and spatial architecture in human squamous cell carcinoma. *Cell*, 182(2): 497–514, 2020. 2, 6

[11] X Jin, L Zhu, Z Cui, J Tang, M Xie, and G Ren. Elevated expression of gnas promotes breast cancer cell proliferation and migration via the pi3k/akt/snail1/e-cadherin axis. *Clinical and Translational Oncology*, 21:1207–1219, 2019. 7

[12] Fahdi Kanavati, Gouji Toyokawa, Seiya Momosaki, Michael Rambeau, Yuka Kozuma, Fumihiro Shoji, Koji Yamazaki, Sadanori Takeo, Osamu Iizuka, and Masayuki Tsuneki. Weakly-supervised learning for lung carcinoma classification using deep learning. *Scientific reports*, 10(1):1–11, 2020. 3

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[14] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 2

[15] Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14318–14328, 2021. 3

[16] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021. 3

[17] Patrice J Morin. Claudin proteins in human cancer: promising new targets for diagnosis and therapy. *Cancer research*, 65(21):9603–9606, 2005. 7

[18] Minxing Pang, Kenong Su, and Mingyao Li. Leveraging information in spatial transcriptomics to predict super-resolution gene expression from histology images in tumors. *bioRxiv*, 2021. 1, 2, 3, 6

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3

[20] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in Neural Information Processing Systems*, 34:2136–2147, 2021. 3

[21] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*, 2018. 5

[22] Patrik L Ståhl, Fredrik Salmén, Sanja Vickovic, Anna Lundmark, José Fernández Navarro, Jens Magnusson, Stefania Giacomello, Michaela Asp, Jakub O Westholm, Mikael Huss, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353 (6294):78–82, 2016. 1

[23] Asher Trockman and J Zico Kolter. Patches are all you need? *arXiv preprint arXiv:2201.09792*, 2022. 2

[24] Ronald Xie, Kuan Pang, Gary D. Bader, and Bo Wang. Spatially resolved gene expression prediction from h&e histology images via bi-modal contrastive learning, 2023. 1, 2, 3, 6, 7

[25] Yan Yang, Md Zakir Hossain, Eric A Stone, and Shafin Rahman. Exemplar guided deep neural network for spatial transcriptomics analysis of gene expression prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applica-*

tions of Computer Vision, pages 5039–5048, 2023. 1, 2, 3, 6, 7

[26] Yuansong Zeng, Zhuoyi Wei, Weijiang Yu, Rui Yin, Yuchen Yuan, Bingling Li, Zhonghui Tang, Yutong Lu, and Yuedong Yang. Spatial transcriptomics prediction from histology jointly through transformer and graph neural networks. *Briefings in Bioinformatics*, 23(5):bbac297, 2022. 1, 2, 3, 6