

Gene Expression Prediction from Histology Images via Hypergraph Neural Networks

Bo Li¹, Yong Zhang^{1,*}, Qing Wang², Chengyang Zhang¹, Mengran Li³, Guangyu Wang^{4,5}, Qianqian Song^{6,7,*}

¹Beijing Key Laboratory of Multimedia and Intelligent Software Technology, Faculty of Information Technology, Beijing Institute of Artificial Intelligence, Beijing University of Technology, Beijing, China. ²Key Laboratory of Intelligent Education Technology and Application of Zhejiang Province, Zhejiang Normal University, Jinhua, China. ³School of Intelligent Systems Engineering, Sun Yat-sen University, Guangdong, China. ⁴Center for Bioinformatics and Computational Biology, Houston Methodist Research Institute, Houston, TX, USA. ⁵Department of Cardiothoracic Surgery, Weill Cornell Medicine, Cornell University, New York, NY, USA. ⁶Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida, Florida, USA. ⁷Department of Cancer Biology, Wake Forest School of Medicine, Winston Salem, NC, USA.

* Corresponding authors: Qianqian Song, PhD: qsong1@ufl.edu; Yong Zhang, PhD: zhangyong2010@bjut.edu.cn.

Biographical Note

Bo Li is a graduate student at the Beijing Key Laboratory of Multimedia and Intelligent Software Technology, Faculty of Information Technology, Beijing Institute of Artificial Intelligence, Beijing University of Technology, Beijing, China. His research focuses on biomedical image analysis and bioinformatics.

Yong Zhang is a Professor in the Faculty of Information Technology, Beijing Institute of Artificial Intelligence, Beijing University of Technology, Beijing, China. His research focuses on big data analysis and visualization, and the applications of artificial intelligence.

Qing Wang is a graduate student at the Key Laboratory of Intelligent Education Technology and Application of Zhejiang Province, Jinhua, China. His research interests include data mining, artificial intelligence, and their applications.

Chengyang Zhang is a graduate student at the Beijing Key Laboratory of Multimedia and Intelligent Software Technology, Faculty of Information Technology, Beijing Institute of Artificial Intelligence, Beijing University of Technology, Beijing, China. His research focuses on computer vision and deep generative model.

Mengran Li is Ph.D. Candidate at the School of Intelligent Systems Engineering, Sun Yat-sen University, Guangzhou, Shenzhen, China. His research interests include big data and artificial intelligence.

Guangyu Wang is an Assistant Professor of Cardiovascular Sciences at Houston Methodist Research Institute, Weill Cornell Medical College. His research focuses on developing computational models for analyzing high-throughput multi-omics datasets and interrogating how cell fate is regulated.

Qianqian Song is an Assistant Professor in the Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida, Florida, USA. She is also an adjunct Assistant Professor in the Department of Cancer Biology, Wake Forest University School of Medicine, North Carolina, USA. Her research focuses on developing innovative computational methods to decipher disease mechanisms and identify therapeutic biomarkers.

Word count: 4,739

Key Points

We develop a novel histology image-based gene prediction model named HGGEP, which demonstrates high accuracy and robust performance.

To reveal the intricate relationship between cell morphology and gene expression in images, we propose a gradient enhancement module, which effectively improves the model's capability in perceiving cell morphology in images.

HGGEP includes a hypergraph module that efficiently models higher-order associations among latent features across multiple latent stages, resulting in significant performance improvement.

ABSTRACT

Spatial transcriptomics reveals the spatial distribution of genes in complex tissues, providing crucial insights into biological processes, disease mechanisms, and drug development. The prediction of gene expression based on cost-effective histology images is a promising yet challenging field of research. Existing methods for gene prediction from histology images exhibit two major limitations. First, they ignore the intricate relationship between cell morphological information and gene expression. Second, these methods do not fully utilize the different latent stages of features extracted from the images. To address these limitations, we propose a novel hypergraph neural network model, HGGEP, to predict gene expressions from histology images. HGGEP includes a gradient enhancement module to enhance the model's perception of cell morphological information. A lightweight backbone network extracts multiple latent stage features from the image, followed by attention mechanisms to refine the representation of features at each latent stage and capture their relations with nearby features. To explore higher-order associations among multiple latent stage features, we stack them and feed into the hypergraph to establish associations among features at different scales. Experimental results on multiple datasets from disease samples including cancers and tumor disease, demonstrate the superior performance of our HGGEP model than existing methods.

Keywords: spatial transcriptomics, gene expression prediction, histology image, gradient enhancement, attention mechanism, hypergraph.

INTRODUCTION

Diverse cell types are intricately arranged both spatially and structurally within tissues to fulfill their specific functions. Revealing the intricate spatial architecture and cell activities within heterogeneous tissues holds considerable importance in comprehending the cellular mechanisms and functions associated with diseases¹⁻⁴. Spatial Transcriptomics (ST) serves as an advanced technology that can be utilized to elucidate the spatial distribution of genes at both tissue and spot levels. This technology has significantly advanced our understanding of gene expressions in biological processes⁵, playing a crucial role in exploring disease mechanisms⁶ and revealing novel drug targets⁷. The rapid progress of ST technology allows for the simultaneous analysis of gene expression, cell or spot locations, and corresponding histology images. Currently, numerous researchers are actively engaged in related studies, covering spatial domain recognition⁸⁻¹⁰, spatial transcriptomics deconvolution¹¹⁻¹³, and inference of spatial cellular interactions^{14, 15}.

However, the considerable cost associated with acquiring spatial transcriptomics data limits the widespread pursuit of research on ST technologies. In contrast, histology images of various disease tissues are more accessible. Recently, researchers have shifted their focus toward predicting gene expression from whole-slide image (WSI) data. Some methods, such as ST-Net¹⁶, HisToGene¹⁷, and Hist2ST¹⁸, have emerged for this purpose. Initially, ST-Net pioneers the use of deep learning techniques to predict spatial gene expression from WSI, yielding promising results. HisToGene and Hist2ST improve the prediction performance by incorporating transformer models to capture global associations of image features across different spots in a WSI. Meanwhile, Hist2ST leverages graph neural networks¹⁹ to enhance local associations of image features between spots. However, these existing methods still face two primary limitations: (1) they ignore the intricate relationship between cell morphological information and gene expression; (2) insufficient utilization of image-based features at multiple latent stages, coupled with the oversight of high-order associations among those features.

Regarding the first limitation, existing methods based on traditional convolution primarily concentrate on semantic information, i.e., pixel values in the image. They do not sufficiently consider the gradient relationship between the current position and its neighboring positions, which leads to the difficulty of the model to perceive the cell morphological information related to gene expression. To address this limitation, our HGGEP model includes the gradient enhancement module to refine the extracted imaging features and generate latent feature maps with prominent cell morphological information. To address the second limitation and enhance the utilization of features at multiple latent stages within WSI, our HGGEP model employs a two-step strategy. Specifically, HGGEP first extracts multiple latent features from WSI through a lightweight backbone network²⁰, and subsequently refine the representation of features at each latent stage using the attention

mechanism^{21, 22}. To explore higher-order associations among multiple latent stage features, we innovatively introduce a hypergraph association module based on multiple metrics. Collectively, we propose a novel HGGE model that overcomes existing challenges and achieves superior performance in gene expression prediction from histology images.

RESULTS

Overview of HGGE model

Our hypergraph neural network-based model, HGGE (HyperGraph Gene Expression Prediction), is depicted in **Figure 1**. The process initiates with the partitioning of a whole slide image into multiple patches centered around spots. Acknowledging the intricate relationship between cell morphology and gene expression, the Gradient Enhancement Module (GEM) is utilized to enhance the model's perception of cell morphology. The latent features preprocessed by GEM undergo processing through a lightweight backbone network²⁰ to extract features at multiple latent stages. These latent features are then fed into the Convolutional Block Attention Module (CBAM²¹) and Vision Transformer (ViT²²), utilizing attention mechanisms to optimize the representation of features at each latent stage. To uncover high-order associations among features across different stages, a Hypergraph Association Module (HAM) based on nearby positions and distance relations is employed for global associations modeling and local representation. Following this, the latent features output from the ViT and HAM are concatenated and fed into the Long Short-Term Memory (LSTM^{23, 24}) module, enhancing information exchange among latent features. Ultimately, the latent features from the last layer are input into a gene regression head to yield the predicted gene expression on this histology image.

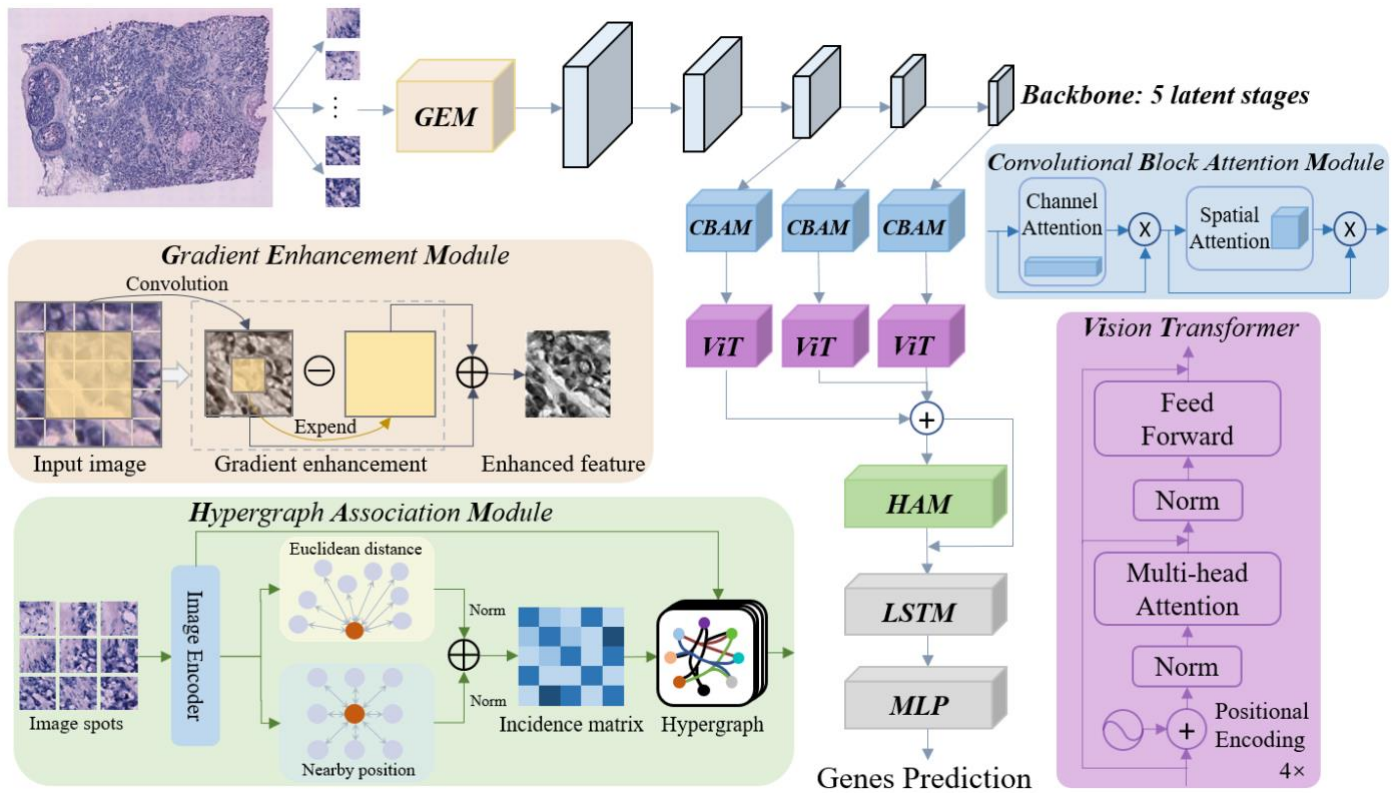


Figure 1. Overview of our HGGE model. This model consists of three pivotal components: GEM, designed to capture the intricate relationship between cell morphology and gene expression; CBAM and Vision Transformer modules, employed for extracting internal features at each latent stage; and a Hypergraph Association Module (HAM), dedicated to revealing higher-order associations among multiple latent stage features.

Performance evaluation using diverse spatial transcriptomics datasets

To comprehensively evaluate the performance of HGGE, we conducted experiments using a leave-one-out validation approach on a total of 44 sections from human HER2-positive breast tumor (HER2⁺²⁵) dataset (section A1-G3) and human cutaneous squamous cell carcinoma (cSCC²⁶) dataset (section P2_ST_rep1-P10_ST_rep3). The performance of HGGE is compared with existing methods including Hist2ST, HisToGene, and STNet. **Figure 2** illustrates the comparison results based on the Pearson Correlation Coefficient (PCC) index for each method. Specifically, on the HER2+ datasets, the HGGE model exhibits

an average PCC index and median PCC index approximately 5% higher than the second-best-performing Hist2ST model. Similarly, on the cSCC datasets, the HGGEp model outperforms the Hist2ST model by around 4%. **Figure 2** reveals that most of the compared models perform less favorably in sections A2-A6 and F1-F3, with PCC indices consistently below 0.07. In contrast, our HGGEp model successfully improves the performance by approximately 5%. Notably, the most substantial improvement relative to the compared models occurs in sections C1-C5, with the PCC increase of about 7%. These results demonstrate that compared with existing methods, our HGGEp model presents superior capabilities of predicting gene expressions from histology images.

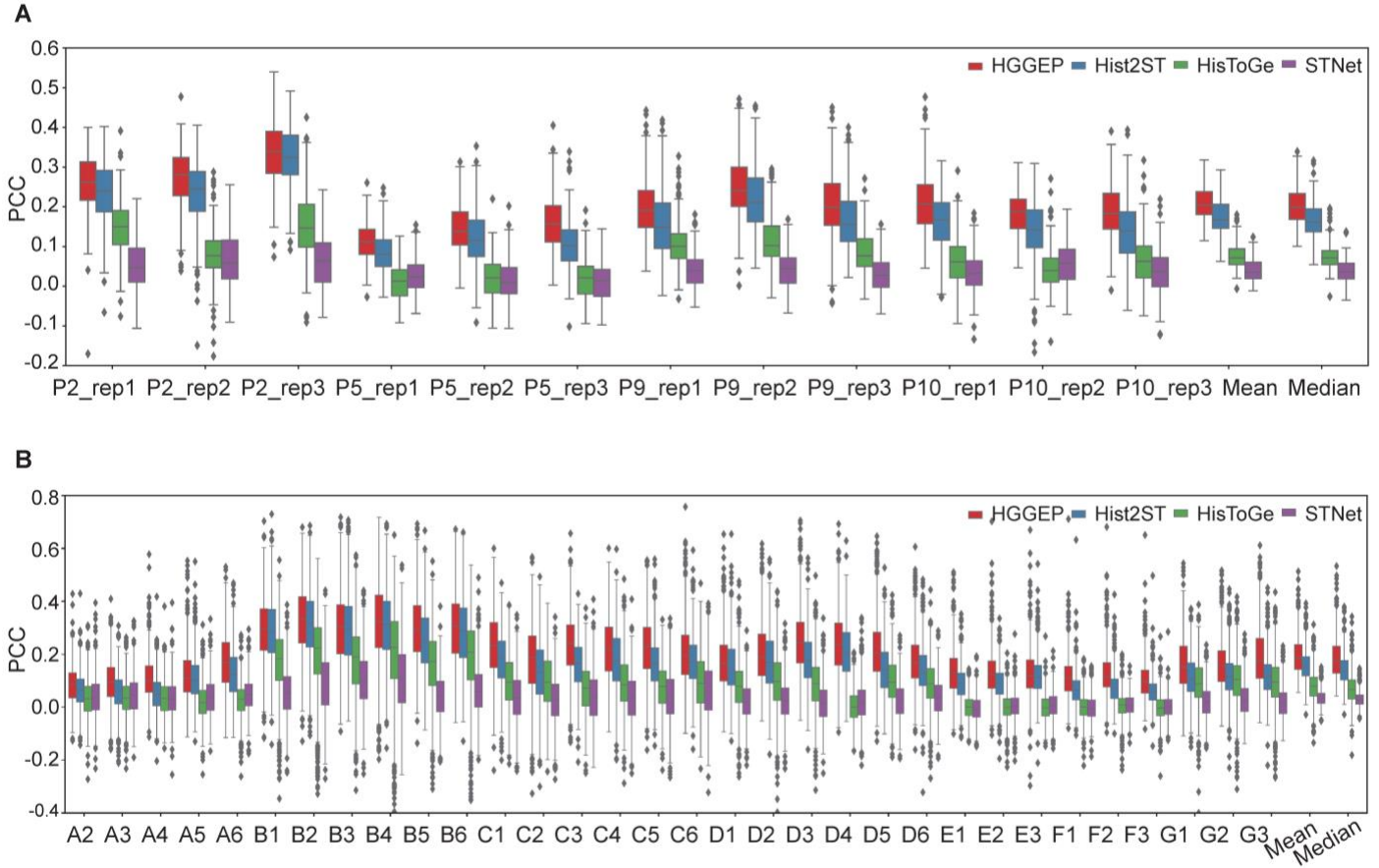


Figure 2. Benchmark of the gene expression prediction performance. Comparison results between our HGGEp model and existing methods on the (a) HER2+ datasets and (b) and cSCC datasets.

To assess the contribution of each module's performance, ablation experiments were conducted on the HER2+ datasets, and the results are presented in **Figure 3**. In this figure, red color represents the performance of ablation experiments conducted on the HGGEp model, while the other three colors correspond to the compared methods (Hist2ST, HisToGene, STNet). The observations in **Figure 3** lead to the following three conclusions: 1) The model's overall performance exhibits an increasing trend of PCC with the gradual inclusion of each new module, indicating that the combination of those modules contributes significantly to the overall performance of HGGEp. 2) Based on the performance of ablation experiments, the inclusion of three modules—GEM, ViT, and HAM—demonstrates more pronounced improvement effects than the other components. Such observation highlights the capability of these three modules in capturing cell morphology within WSI, optimizing the representation of features at each latent stage, and capturing high-order associations among features across multiple latent stages. 3) A clear performance improvement is observed when comparing our ablated model to the other three models (Hist2ST, HisToGene, STNet).

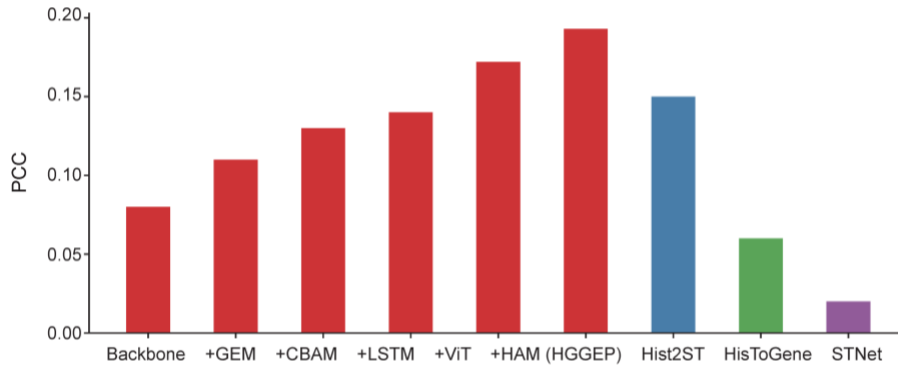


Figure 3. Ablation studies of our HGGEp model. Module ablation experiments for the HGGEp model. The observed performance improvement with the addition of modules underscores their contributions to the performance of HGGEp model.

Evaluation of predicted gene expression

In this section, we delve into the visualization of the top predicted genes by the HGGEp model on the HER2+ and cSCC datasets. To offer a more visually impactful representation of the model's performance, we include comparisons with two additional models, Hist2ST and HisToGene, as depicted in **Figure 4**. Specifically, the upper and lower panels showcase the top gene prediction performance of different models on the HER2+ and cSCC datasets, respectively. Each dot in the figures represents a specific gene, with the vertical axis indicating the PCC index of genes predicted by the HGGEp model, and the horizontal axis corresponding to the HisToGene (left panel) and Hist2ST (right panel) models. Therefore, data points above the diagonal line in the figure indicate that the HGGEp model achieves higher PCC values for those genes compared to the other methods (HisToGene or Hist2ST). Additionally, data points positioned in the upper-left corner signify superior gene prediction performance of HGGEp compared to the comparative models (HisToGene or Hist2ST). These results show that the HGGEp model presented superior performance in gene expression prediction.

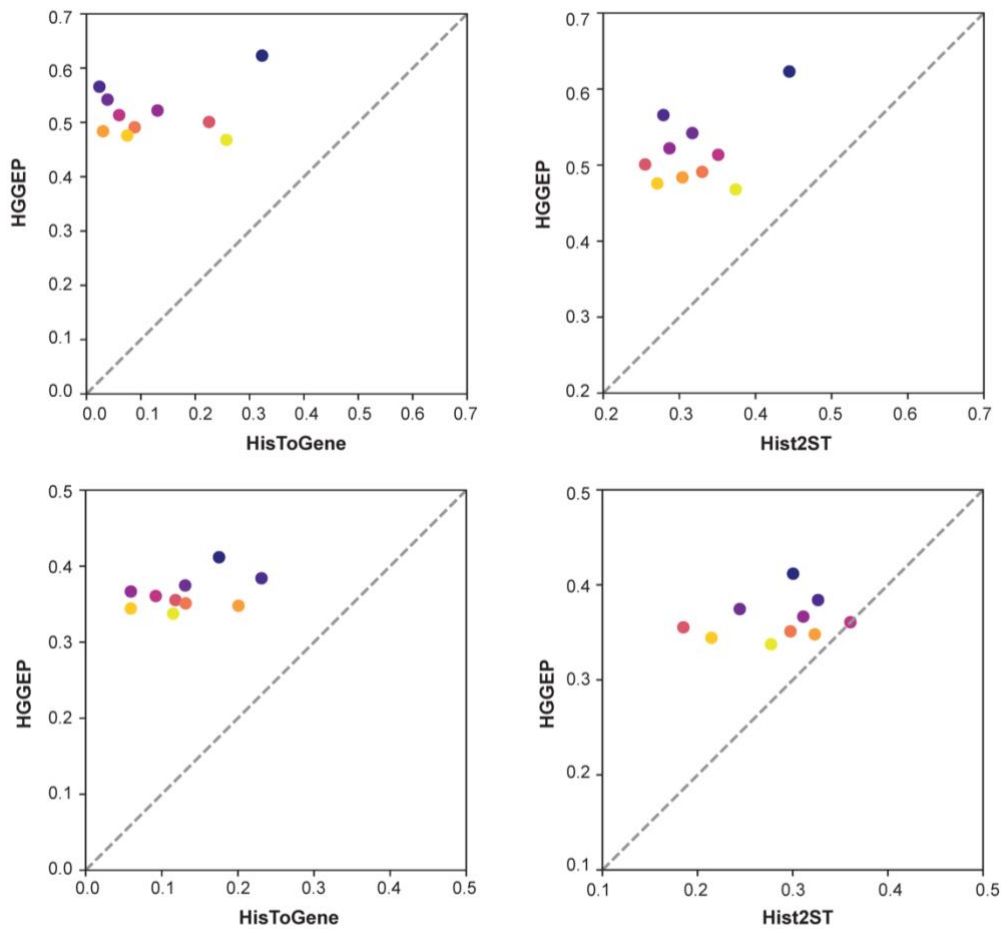


Figure 4: Performance comparison of top 10 gene predictions across two datasets. Data points positioned in the upper-left corner signify superior gene prediction performance of HGGEp compared to the comparative models (HisToGene or Hist2ST).

It is noteworthy that the top genes identified in the HER2+ dataset (*GNAS*, *UBA52*, *MUCL1*) and the cSCC dataset (*EIF5*, *MSMO1*, *STMNI*) indicate significant biological relevance. In the HER2+ tumor microenvironment, *GNAS* is implicated with critical roles in cell signal transduction²⁷, potentially influencing the biological processes including the proliferation and metastasis of tumor cells²⁸. *UBA52*, closely associated with ubiquitin-protein conjugation²⁹, may significantly impact the treatment response and prognosis of HER2+ tumors due to its aberrant expression³⁰. Additionally, *MUCL1*, involved in cell adhesion and tumor microenvironment regulation, plays a crucial role in HER2+ tumors and offers potential insights for precision therapy³¹. On the other hand, in the cSCC environment, *EIF5* is implicated in protein synthesis regulation and cell proliferation³². *MSMO1* participates in cholesterol biosynthesis³³, and *STMNI* may influence biological processes such as cell division and migration³⁴. Therefore, accurate prediction of those genes contributes to exploring the molecular mechanisms driving cSCC development, thus gaining a better understanding of tumor cell activities in the microenvironment.

For a more intuitive comparison of top gene prediction performance across different models, we conducted the visualization of predicted genes. To ensure a fair comparison, we select four genes mentioned in Hist2ST (*GNAS*, *FASN*, *MY12B*, and *SCD*) and present their predicted gene expressions by each model in **Figure 5**. This figure demonstrates our model achieves the best prediction performance. For example, our model achieves better prediction of the *GNAS* gene expression (PCC= 0.637), compared with the competitors (Hist2ST, PCC = 0.591; HisToGene, PCC = 0.493). Similarly, for the genes *FASN*, *MY12B*, and *SCD*, our model demonstrates consistently better prediction performance. These results support that our model not only excels in overall performance but also demonstrates outstanding performance in predicting specific genes with biological significance.

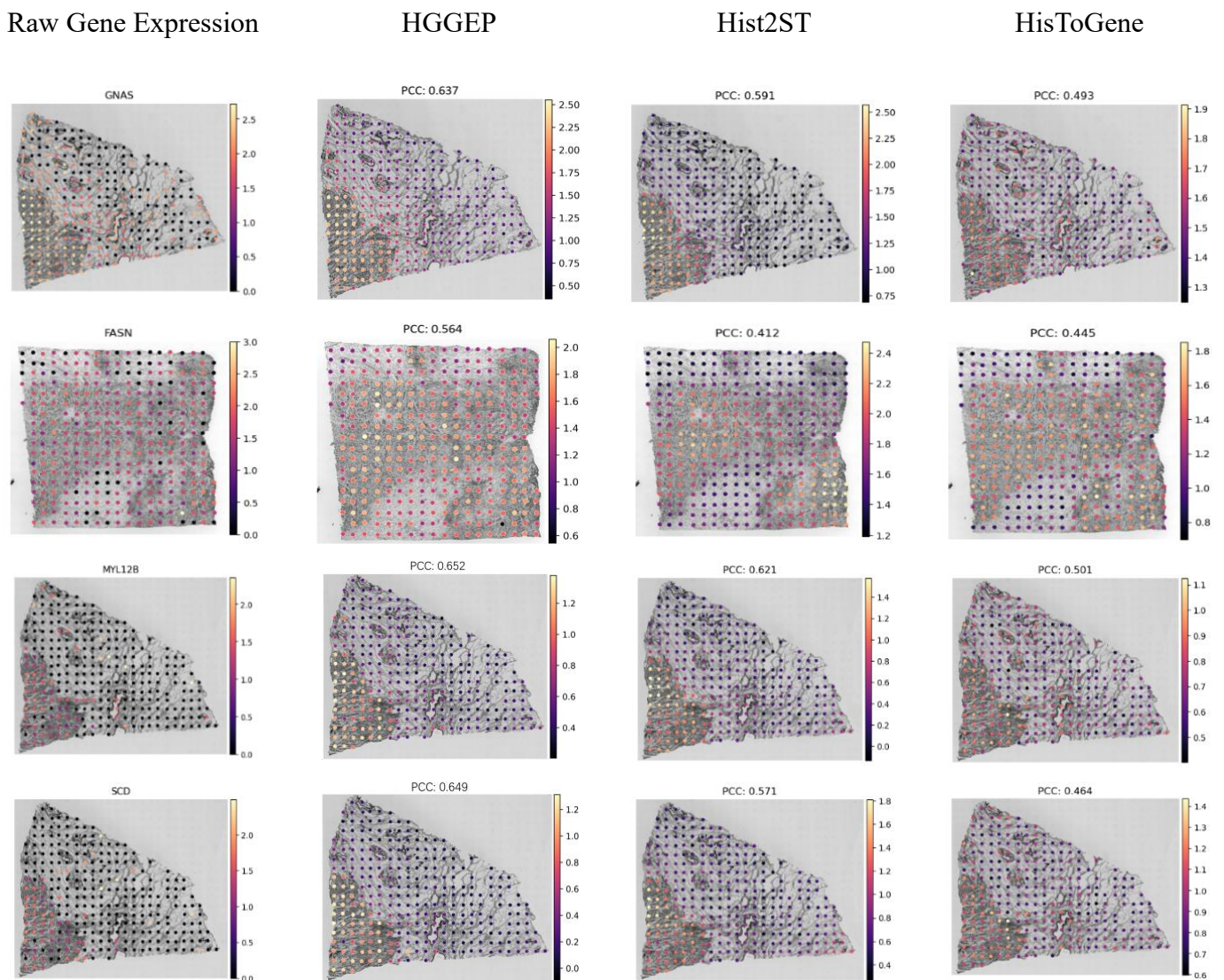
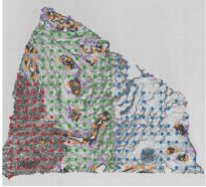
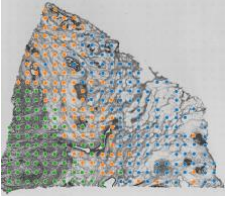
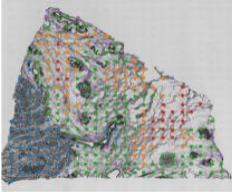
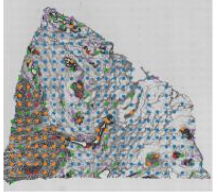
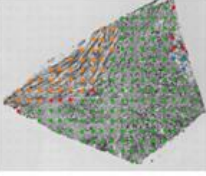
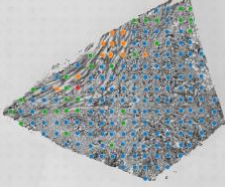
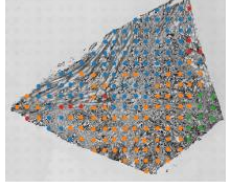
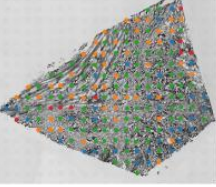
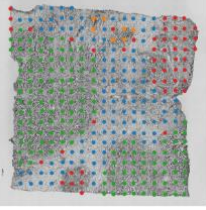
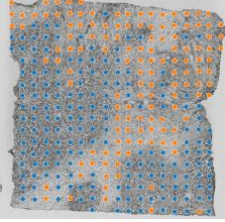
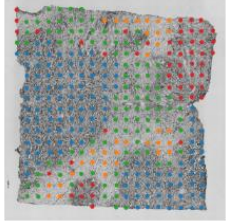
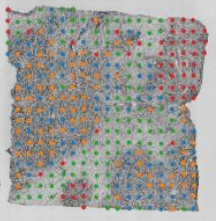
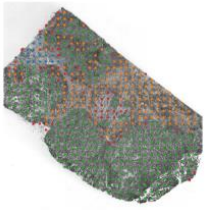
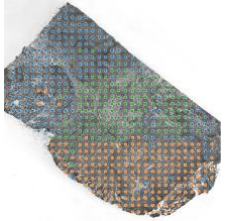

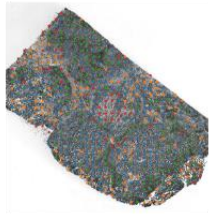
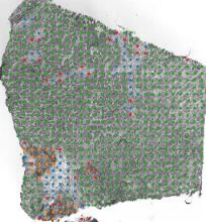
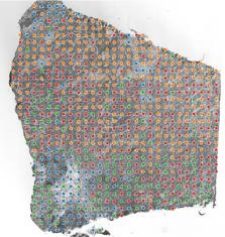
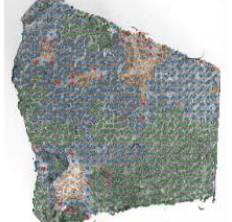
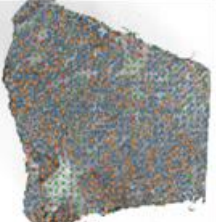
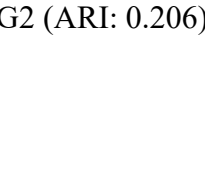
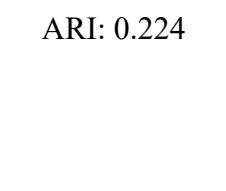
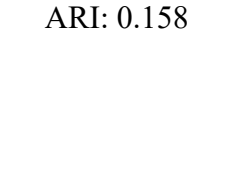
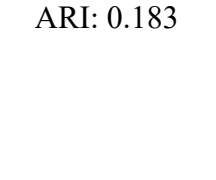


Figure 5. Visualization of predicted genes. The top predicted genes by HisToGene in the HER2+ dataset, where the p-value for each tissue section was obtained in the association test between the predicted and observed gene expression.

Evaluation of spatial region detection

To further demonstrate the accurate gene expressions predicted by our model, we evaluate the spatial region detection capability across the entire histology image accordingly. We performed clustering analysis on the predicted gene expressions of each method, as illustrated in **Figure 6**. Here we utilized six sections (B1, C1, D1, E1, F1, and G2) of the HER2+ dataset, with available annotations from expert professionals. As depicted in **Figure 6**, our model consistently outperforms the latest Hist2ST and HisToGene models, achieving nearly optimal detection performance. Across the six sections, our model surpasses the runner-up model Hist2ST by approximately 11% in average performance. Particularly noteworthy is the outstanding performance in section C1, where our model significantly improves the ARI index by 30% compared to Hist2ST, demonstrating the superior precision of our HGGEF model in region detection.

GT	HGGEF	Hist2ST	HisToGene
B1 (ARI: 0.217)	ARI: 0.399	ARI: 0.286	ARI: 0.311
			
C1 (ARI: 0.12)	ARI: 0.317	ARI: 0.017	ARI: 0.014
			
D1 (ARI: 0.228)	ARI: 0.509	ARI: 0.500	ARI: 0.297
			
E1 (ARI: 0.037)	ARI: 0.214	ARI: 0.089	ARI: 0.040
			
F1 (ARI: 0.079)	ARI: 0.148	ARI: 0.118	ARI: 0.114
			
G2 (ARI: 0.206)	ARI: 0.224	ARI: 0.158	ARI: 0.183
			

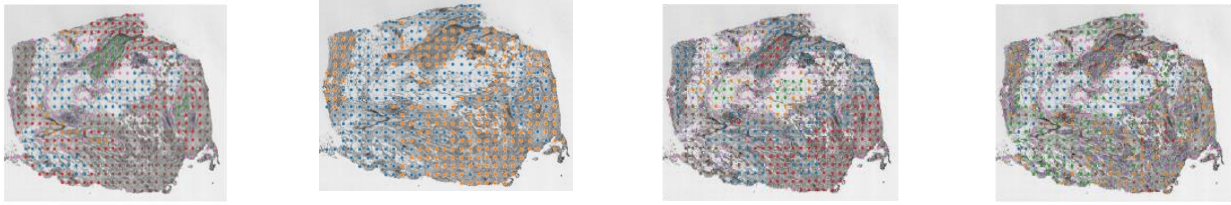


Figure 6. Spatial domain detection based on predicted gene expressions. The accuracy of spatial domain detection on the HER2+ dataset using the gene expressions predicted by each model. GT represents the ground truth labels from the pathology annotations.

DISCUSSION

Gene expression prediction is a pivotal focus in current research, with its extensive applications making it a central theme in scientific exploration. However, many existing methods face limitations due to their reliance on costly data, which constrains further research progress. In response to this challenge, some researchers have turned to a more cost-effective strategy, utilizing histology images to unveil gene expression. In this work, we have introduced HGGEF, a cutting-edge hypergraph neural network model tailored for the precise prediction of gene expression from histology images. HGGEF has been applied to a total of 44 sections, showcasing robust performance in predicting gene expression and underscoring the pivotal role of histology images in this process.

Our HGGEF model outperforms existing methods in three aspects. First, we enhance the model's ability to perceive morphological information through the gradient enhancement module, thus enabling the model to capture the relationship between cell morphology and gene expression. Second, features at different stages in the image have different information, usually high-level features have more semantic information, while low-level features carry more detailed features. Therefore, we use a lightweight backbone to extract different levels of information and then use this attention mechanism to optimize the features at each stage. Lastly, we fuse features from various stages, establishing global associations among spots to achieve comprehensive information integration.

Though HGGEF presents superior performance, we anticipate key areas for future exploration and improvement. First, most existing methods rely on large amounts of labeled data for learning, thus limiting the ability of the model to generalize to small sample domains. To overcome this limitation, future work could consider fine-tuning based on large vision models in the biomedical field^{35, 36}. By leveraging the wealth of knowledge that already exists in large vision models, we can enhance the generalizability of the models to a wider range of application scenarios. Second, existing methods typically use histology images as input for predicting gene expression. Future research will consider introducing more modal information as input to help establish mapping relationships between images and gene expression, such as molecular biology data or pathology data. By synthesizing multimodal information, it is expected to improve the accuracy and comprehensiveness of gene expression prediction. In conclusion, the experiments in this paper demonstrate the power of HGGEF in predicting gene expression based on histology images. Moreover, future work could focus on enhancing model generalization capabilities, including the use of large vision models and multimodal information, to better accommodate the diversity of biomedical data and scenarios.

MATERIALS AND METHODS

Data processing

To validate the efficacy of our proposed method, we utilized the spatial transcriptomics datasets comprising histology images and gene expression data at spot locations. We primarily leveraged the HER2+ and cSCC datasets profiled from 32 and 12 tissue slides, respectively, with a total of 9,612 spots and 6,630 spots respectively. For each histology image in the datasets, we segment it into multiple sub-images based on the positions of spots. Each sub-image is cropped by 112×112 pixels around the spot's center. The input sub-image features are annotated as $\mathbf{x}_{in} \in \mathbb{R}^{N \times 3 \times 112 \times 112}$, where N is the number of spots within the histology image.

During the validation phase, we adopt a consistent leave-one-out cross-validation strategy. Taking the HER2+ dataset as an example, for each section, we train the model on the remaining 31 sections and validate it on that leave-out section.

HGGEF model

In contrast to prior methods, HGGEp significantly improves the prediction accuracy of gene expression from histology images. The detailed structure of HGGEp model is illustrated below.

Gradient Enhancement Module

To enrich the cell morphological information closely associated with gene expression, this paper proposes the GEM, which enhances the model’s perception of cell morphology through difference convolution^{37, 38}. This module comprises two key components: 1) the convolution process that convolves the sub-images; 2) the gradient enhancement process that enhances the cell morphology information by difference convolution.

Figure 1 illustrates the implementation of GEM with the steps of convolution and difference operations. Specifically, a 3×3 convolutional kernel is applied to a 5×5 input split histology image, resulting in a downsampled feature map of size 3×3 . Subsequently, a gradient enhancement operation is performed on the resulting feature map, aiming to utilize the differences between neighboring pixels in the feature map to enhance its gradient information. Traditional convolution process can be expressed as:

$$TC(p_0) = \sum_{p_n \in R(p_0)} \mathbf{w}_{p_n} \cdot \mathbf{x}_{p_n} \quad (1)$$

, where p_0 represents the central position of the local receptive field $R(p_0)$, p_n denotes the relative position of other pixels within the receptive field $R(p_0) = \{(-1, -1), (-1, 0), \dots, (0, 1), (1, 1)\}$, \mathbf{x}_{p_n} denotes the pixel value at position p_n in the input feature map, \mathbf{w}_{p_n} is a learnable parameter. To achieve gradient enhancement, we deform the traditional convolution $TC(p_0)$ to difference convolution $DC(p_0)$:

$$DC(p_0) = \sum_{p_n \in R(p_0)} (\mathbf{w}_{p_n} \cdot \mathbf{x}_{p_n}) + \theta \left(-\mathbf{x}_{p_0} \cdot \sum_{p_n \in R(p_0)} \mathbf{w}_{p_n} \right) \quad (2)$$

, where θ is a hyperparameter that regulates the balance between semantic and gradient information, and we set it to 0.7 default. When $\theta = 0$, the difference convolution is the same as the traditional convolution operation. Given the input sub-images $\mathbf{x}_{in} \in \mathbb{R}^{N \times 3 \times 112 \times 112}$, each of the convolution (F_{TC} and F_{DC}) exports the $F_{TC} \in \mathbb{R}^{N \times 6 \times 56 \times 56}$ and $F_{DC} \in \mathbb{R}^{N \times 6 \times 56 \times 56}$. The latent features from GEM module are $z^{GEM} \in \mathbb{R}^{N \times 3 \times 56 \times 56}$. The entire GEM operates as follows:

$$\mathbf{F}_{TC} = \text{BN}(\sigma(\text{TC}(\mathbf{x}_{in}))), \quad (3)$$

$$\mathbf{F}_{DC} = \text{BN}(\sigma(\text{DC}(\mathbf{x}_{in}))), \quad (4)$$

$$z^{GEM} = \text{BN}(\sigma(\text{MLP}([\mathbf{F}_{TC}, \mathbf{F}_{DC}]))), \quad (5)$$

where z^{GEM} is the enhanced latent feature map, σ is ReLU activation function, BN is Batch Norm. Here the Multi-Layer Perceptron (MLP) layer is used to adjust the number of feature channels, which then serve as input into the subsequent backbone network.

Backbone and Multiple Latent Stage Feature

With the GEM-enhanced feature map z^{GEM} , the backbone network, CBAM²¹, and ViT²² are employed to extract and optimize multiple latent stage feature information from the z^{GEM} , enhance the model’s prowess in latent feature modeling, and capture global relationships among spots at each latent stage.

For the backbone network, we chose a lightweight shufflenet V2²⁰ to optimize computing efficiency. This network produces output across five stages, with diminishing latent feature map sizes as the network deepens. For subsequent processing, we leverage features from the last three stages from the backbone, feeding them into the CBAM. As shown in **Figure 1**, CBAM utilizes both Channel Attention (CA) and Spatial Attention (SA) for adjusting the weight of each channel/spot by considering global information. With these attention mechanisms, CBAM enhances the network’s comprehension of vital features in the image and improves the model’s performance and generalization capabilities.

Specifically, the GEM output features $z^{GEM} \in \mathbb{R}^{N \times 3 \times 56 \times 56}$ are first fed into the shufflenet V2 backbone network to get the embeddings at 5 latent stages [$\mathbf{S}_1 \in \mathbb{R}^{N \times 24 \times 29 \times 29}$, $\mathbf{S}_2 \in \mathbb{R}^{N \times 48 \times 15 \times 15}$, $\mathbf{S}_3 \in \mathbb{R}^{N \times 96 \times 8 \times 8}$, $\mathbf{S}_4 \in \mathbb{R}^{N \times 192 \times 4 \times 4}$, $\mathbf{S}_5 \in \mathbb{R}^{N \times 1024 \times 4 \times 4}$]. The embeddings of the last 3 latent stages are selected for subsequent analysis. Subsequently, we input these embedding features into the CBAM to optimize the features of each latent stage

separately. Specifically, for \mathcal{S}_i , $i \in \{3,4,5\}$, the CBAM module performs as below:

$$\text{CA}(\mathcal{S}_i) = \sigma \left(\text{MLP}_{\text{channel}} \left(\text{AvgPool}_{\text{spatial}}(\mathcal{S}_i) \oplus \text{MaxPool}_{\text{spatial}}(\mathcal{S}_i) \right) \right), \quad (6)$$

$$\mathbf{h}_i^{\text{tmp}} = \mathcal{S}_i \odot \text{CA}(\mathcal{S}_i), \quad (7)$$

$$\text{SA}(\mathbf{h}_i^{\text{tmp}}) = \sigma \left(\text{Conv}_{7 \times 7}([\text{AvgPool}_{\text{channel}}(\mathbf{h}_i^{\text{tmp}}), \text{MaxPool}_{\text{channel}}(\mathbf{h}_i^{\text{tmp}})]) \right), \quad (8)$$

$$\mathbf{h}_i = \mathbf{h}_{\text{tmp}4} \odot \text{SA}(\mathbf{h}_i^{\text{tmp}}), \quad (9)$$

where the symbol \oplus denotes element-wise addition, \odot represents element-wise multiplication, and σ is the ReLU activation function. $\text{AvgPool}_{\text{spatial}}(\mathcal{S}_i)$ and $\text{MaxPool}_{\text{spatial}}(\mathcal{S}_i)$ denote the average and maximum pooling operations on the embedding feature \mathcal{S}_i , respectively, and the output of both is $\mathbb{R}^{N \times 192 \times 1 \times 1}$. As shown in equation (6), the outputs of the two pooling operations are element-wise added. They subsequently enter the $\text{MLP}_{\text{channel}}$ for channel scaling, with the scaling factor set to 16 by default (equation 6). For spatial attention $\text{SA}(\mathbf{h}_i^{\text{tmp}})$, $\text{AvgPool}_{\text{channel}}(\mathbf{h}_i^{\text{tmp}})$ and $\text{MaxPool}_{\text{channel}}(\mathbf{h}_i^{\text{tmp}})$ perform average and maximum pooling for the channels, respectively, followed by the 7×7 convolution. It is worth noting that CBAM-optimized hidden features maintain their original input dimension, i.e., the hidden features \mathbf{h}_i after CBAM optimization are $\mathbf{h}_3 \in \mathbb{R}^{N \times 96 \times 8 \times 8}$, $\mathbf{h}_4 \in \mathbb{R}^{N \times 192 \times 4 \times 4}$, $\mathbf{h}_5 \in \mathbb{R}^{N \times 1024 \times 4 \times 4}$.

To facilitate subsequent processing, we uniformly adjust the above three hidden features $[\mathbf{h}_3, \mathbf{h}_4, \mathbf{h}_5]$ to $\mathbb{R}^{N \times 1024}$ by linear and dimension transformation. These hidden features are fed into the ViT module separately and its self-attention mechanism is utilized to capture the associations among the spots in WSI. For \mathbf{h}_i , $i \in \{3,4,5\}$, the ViT module is implemented as follows:

$$\mathbf{z}_i^{\text{tmp}} = \text{Multihead}(\mathbf{h}_i) = \text{Concat}(\text{head}_1(\mathbf{h}_i), \dots, \text{head}_8(\mathbf{h}_i)) \mathbf{W}^0, \quad (10)$$

$$\text{For each head } t \in \{1, \dots, 8\}, \text{head}_t(\mathbf{h}_i) = \text{Attention}(\mathbf{Q}_{\mathbf{h}_i}, \mathbf{K}_{\mathbf{h}_i}, \mathbf{V}_{\mathbf{h}_i}) = \text{Softmax} \left(\frac{\mathbf{Q}_{\mathbf{h}_i} \mathbf{K}_{\mathbf{h}_i}^T}{\sqrt{d_k^{h_i}}} \right) \mathbf{V}_{\mathbf{h}_i}, \quad (11)$$

where $\mathbf{Q}_{\mathbf{h}_i} = \mathbf{h}_i \mathbf{W}_q^h$, $\mathbf{K}_{\mathbf{h}_i} = \mathbf{h}_i \mathbf{W}_k^h$, $\mathbf{V}_{\mathbf{h}_i} = \mathbf{h}_i \mathbf{W}_v^h$, \mathbf{h}_i represents one of the input features $[\mathbf{h}_3, \mathbf{h}_4, \mathbf{h}_5]$, $\mathbf{W}^0, \mathbf{W}_q^h, \mathbf{W}_k^h, \mathbf{W}_v^h$ denote the learnable weight matrices, $\mathbf{Q}_{\mathbf{h}_i}, \mathbf{K}_{\mathbf{h}_i}, \mathbf{V}_{\mathbf{h}_i}$ are the matrices of queries, keys, and values obtained by linear transformation of \mathbf{h}_i , respectively. The parameter $d_k^{h_i}$ is used to scale the denominator in the dot-product attention, controlling the scaling of attention weights. These operations enable the model to dynamically model the associations among spots, leading to an optimal allocation of attention. After the optimization of the multi-head attention mechanism, the dimension of the ViT output $[\mathbf{z}_3^{\text{tmp}}, \mathbf{z}_4^{\text{tmp}}, \mathbf{z}_5^{\text{tmp}}]$ remain constant at $\mathbb{R}^{N \times 1024}$.

Finally, a Feed-Forward Network (FFN) is applied for additional non-linear transformations and representation learning of the hidden features from each stage:

$$\mathbf{z}_i^{\text{ViT}} = \text{FFN}(\mathbf{z}_i^{\text{tmp}}) = \text{ReLU}(\mathbf{z}_i^{\text{tmp}} \mathbf{W}^1 + b^1) \mathbf{W}^2 + b^2 \quad (13)$$

, where \mathbf{W}^i and b^i are the weight matrix and bias vector, respectively. After ViT optimization, it is obtained $[\mathbf{z}_3^{\text{ViT}}, \mathbf{z}_4^{\text{ViT}}, \mathbf{z}_5^{\text{ViT}}] \in \mathbb{R}^{N \times 1024}$.

Hypergraph Association Module

The feature representations of spots at each latent stage are derived by encoding histology images through the model's front-end image encoder, which encompasses the GEM, Backbone, CBAM, and ViT modules. However, the above modules are limited to treating the features of each stage independently and do not explore the association among the multiple latent stage features $[\mathbf{z}_3^{\text{ViT}}, \mathbf{z}_4^{\text{ViT}}, \mathbf{z}_5^{\text{ViT}}]$. Herein, we first fuse the multiple latent feature maps described above, and then introduce a Hypergraph Association Module (HAM) aimed at capturing high-order association among spots. To comprehensively model feature associations at varying distances, we adopt a global modeling approach based on Euclidean distance and a local modeling approach based on nearby positions.

Specifically, with the summed element by element of $[z_3^{ViT}, z_4^{ViT}, z_5^{ViT}]$, $H_{in} \in \mathbb{R}^{N \times 1024}$ is obtained and fed into the HAM. We use v_i as the node and the attributes of the node v_i is $m_i \in \mathbb{R}^{1 \times 1024}$. To establish hyperedges among nodes, we propose a method that combines Euclidean distance and nearby positions metrics to generate an incidence matrix. Initially, we measure the Euclidean distance between the current node v_i and other nodes. This effectively models relationships between distant nodes v_i and v_j :

$$\text{Dis}(v_i, v_j) = \sqrt{\sum_{k=1}^{1024} (m_{ik} - m_{jk})^2}. \quad (14)$$

Simultaneously, considering that neighboring nodes on the coordinates usually have similar genetic phenotypes, we also build hyperedges based on the positional relationship between nodes. Assume that the coordinates of node v_i are (x_i, y_i) and the coordinates of node v_j are (x_j, y_j) , the positional relationship is encoded as:

$$\text{Pos}(v_i, v_j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}. \quad (15)$$

By combining the two aforementioned metrics, we establish the relationship between the current node and other nodes. To ensure a balanced contribution of the two metrics to the final incidence matrix, normalization method is applied:

$$\text{Inc_Mat}(v_i, v_j) = \text{Norm}(\text{Dis}(v_i, v_j)) + \text{Norm}(\text{Pos}(v_i, v_j)). \quad (16)$$

Next, the incidence matrix and features $H_{in} \in \mathbb{R}^{N \times 1024}$ are input into a hypergraph convolutional network. Throughout the convolution process, the model enhances its understanding of inter-node dependencies, encompassing relationships among nodes as well as between nodes and hyperedges. For the concrete implementation of hypergraph convolution^{39, 40}, this study leverages the HypergraphConv module from the torch_geometric library:

$$H_{tmp} = \text{HypergraphConv}(H_{in}, v_i, m_i), \quad (17)$$

$$H_{out} = \text{Dropout}\left(\text{Norm}\left(\text{ReLU}(H_{tmp})\right)\right). \quad (18)$$

The output feature obtained after the HAM module is $H_{out} \in \mathbb{R}^{N \times 1024}$.

Finally, to obtain the final gene prediction, a Long Short-Term Memory (LSTM) module and MLP module are introduced in this paper. LSTM models the association among features by treating the input multiple latent stage features as time series. The features of the four stages $[z_3^{ViT}, z_4^{ViT}, z_5^{ViT}, H_{out}]$ are concatenated together to obtain $L_{in} \in \mathbb{R}^{4 \times N \times 1024}$, which is used as an input to the LSTM. Subsequently, the first dimension of the LSTM output features is average pooled to obtain the average vector $L_{out} \in \mathbb{R}^{N \times 1024}$. Finally, MLP is used to map the latent feature $L_{out} \in \mathbb{R}^{N \times 1024}$ to the number of final predicted genes, which is $\mathbb{R}^{N \times 785}$ for HER2+ and $\mathbb{R}^{N \times 171}$ for cSCC dataset.

Model hyperparameter configuration. The hyperparameters used in the HGGE model are listed specifically. (1) GEM: the parameter θ in Equation 2 is set to 0.7 and the feature channel variations progress from $3 \rightarrow 6 \rightarrow 3$. (2) The backbone network: we utilize the shufflenet_v2_x0_5 model, pre-trained on the ImageNet dataset, to extract features from the final three stages for subsequent modules. (3) CBAM: the scaling factor in the channel attention in Equation 6 is configured to shrink by $16 \times$. Subsequently, for ease of subsequent processing, we reduce these features from the last three stages to $\mathbb{R}^{n \times 1024}$, where n represents the number of spots. (4) ViT: the whole ViT module is iterated four times. The number of attention heads is set to 8, and the FeedForward component includes 2 MLP layers. (5) HAM: when constructing the incidence matrix, a fixed number of adjacent nodes is set to 3, and nodes retain their own indices. The convolution process involves 2 layers of hypergraph convolution, with the hidden layer being half the size of the input. (6) LSTM: This encompasses input and hidden state dimensions of 1024 each, along with a layer count of 4. (7) MLP: After output normalization, a single MLP layer is directly employed to predict the final gene output.

Loss function. Mean Squared Error (MSE) loss is utilized to measure the average squared difference between the predicted gene expression values and the ground truth for each gene. Meanwhile, given the prevalent zeros

in the spatial transcriptomics data, we also include the Zero-Inflated Negative Binomial (ZINB⁴¹) loss. This loss is rooted in the zero-inflated negative binomial distribution, amalgamating the negative binomial distribution with a zero-inflation component. The total loss of the HGGE model is a fusion of MSE and ZINB loss, for the task of gene prediction from histology images.

Evaluation Metrics. Pearson Correlation Coefficient (PCC) is used to assess the performance of benchmarking models. This coefficient, a statistical measure for assessing the linear relationship between two continuous variables, is widely employed to quantify the strength and direction of the linear correlation between these variables. The PCC ranges from -1 to 1, with the progression from -1 to 1 indicating a shift from negative to positive correlation. The specific calculation method is as follows:

$$PCC = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \quad (19)$$

, where \bar{x} and \bar{y} represent the means of variables x and y , respectively, and x_i and y_i denote the i_{th} observations.

Training parameters. During training, the model utilizes a learning rate of 0.00001, runs for a maximum of 400 epochs, and undergoes testing every 5 epochs. The experiments are conducted on an Ubuntu 20.04 system equipped with 128GB of RAM and an A6000 GPU featuring 48GB of memory.

CODE AVAILABILITY

All source codes and trained models in our experiments have been deposited at <https://github.com/QSong-github/HGGE>.

DATA AVAILABILITY

The spatial transcriptomics datasets used in this study include the (1) HER2-positive breast tumor ST datasets, which are available at <https://github.com/almaan/her2st/>; (2) 10x Visium data of human cutaneous squamous cell carcinoma are publicly available in the Gene Expression Omnibus (GEO) (GSE144240).

COMPETING INTERESTS

The authors declare no competing interests.

FUNDING

Q.S. is supported by the National Institute of General Medical Sciences of the National Institutes of Health (R35GM151089). G.W. is supported by the National Institute of General Medical Sciences of the National Institutes of Health (1R35GM150460).

MATERIALS & CORRESPONDENCE

Correspondence and requests for materials should be addressed to QS or YZ.

REFERENCES

1. Asp, M. et al. A Spatiotemporal Organ-Wide Gene Expression and Cell Atlas of the Developing Human Heart. *Cell* **179**, 1647-1660.e1619 (2019).
2. Maynard, K.R. et al. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *bioRxiv*, 2020.2002.2028.969931 (2020).
3. Moncada, R. et al. Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. *Nat Biotechnol* **38**, 333-342 (2020).
4. Maniatis, S. et al. Spatiotemporal dynamics of molecular pathology in amyotrophic lateral sclerosis. *Science* **364**, 89-93 (2019).
5. Ståhl, P.L. et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353**, 78-82 (2016).
6. Schmauch, B. et al. A deep learning model to predict RNA-Seq expression of tumours from whole slide images. *Nat*

- Commun* **11**, 3877 (2020).
7. Tolios, A. et al. Computational approaches in cancer multidrug resistance research: Identification of potential biomarkers, drug targets and drug-target interactions. *Drug Resistance Updates* **48**, 100662 (2020).
 8. Zhao, E. et al. Spatial transcriptomics at subspot resolution with BayesSpace. *Nature biotechnology* **39**, 1375-1384 (2021).
 9. Hu, J. et al. SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nature methods* **18**, 1342-1351 (2021).
 10. Xu, C. et al. DeepST: identifying spatial domains in spatial transcriptomics by deep learning. *Nucleic Acids Research* **50**, e131-e131 (2022).
 11. Song, Q. & Su, J. DSTG: deconvoluting spatial transcriptomics data through graph-based artificial intelligence. *Briefings in bioinformatics* **22**, bbaa414 (2021).
 12. Kleshchevnikov, V. et al. Cell2location maps fine-grained cell types in spatial transcriptomics. *Nature biotechnology* **40**, 661-671 (2022).
 13. Biancalani, T. et al. Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. *Nat Methods* **18**, 1352-1362 (2021).
 14. Cang, Z. et al. Screening cell-cell communication in spatial transcriptomics via collective optimal transport. *Nature Methods* **20**, 218-228 (2023).
 15. Tang, Z., Zhang, T., Yang, B., Su, J. & Song, Q. spaCI: deciphering spatial cellular communications through adaptive graph model. *Briefings in Bioinformatics* **24** (2022).
 16. He, B. et al. Integrating spatial gene expression and breast tumour morphology via deep learning. *Nat Biomed Eng* **4**, 827-834 (2020).
 17. Pang, M., Su, K. & Li, M. Leveraging information in spatial transcriptomics to predict super-resolution gene expression from histology images in tumors. *bioRxiv*, 2021.2011.2028.470212 (2021).
 18. Zeng, Y. et al. Spatial transcriptomics prediction from histology jointly through Transformer and graph neural networks. *Briefings in Bioinformatics* **23** (2022).
 19. Wu, Z. et al. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks learning systems* **32**, 4-24 (2020).
 20. Ma, N., Zhang, X., Zheng, H.-T. & Sun, J. in Proceedings of the European conference on computer vision (ECCV) 116-131 (2018).
 21. Woo, S., Park, J., Lee, J.-Y. & Kweon, I.S. in Proceedings of the European conference on computer vision (ECCV) 3-19 (2018).
 22. Dosovitskiy, A. et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:11929* (2020).
 23. Shi, X. et al. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems* **28** (2015).
 24. Memory, L.S.-T. Long short-term memory. *Neural computation* **9**, 1735-1780 (2010).
 25. Andersson, A. et al. Spatial deconvolution of HER2-positive breast cancer delineates tumor-associated cell type interactions. *Nature communications* **12**, 6012 (2021).
 26. Ji, A.L. et al. Multimodal analysis of composition and spatial architecture in human squamous cell carcinoma. *Cell* **182**, 497-514. e422 (2020).
 27. Jin, X. et al. 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).
 28. Ding, H., Zhang, X., Su, Y., Jia, C. & Dai, C. GNAS promotes inflammation-related hepatocellular carcinoma progression by promoting STAT3 activation. *Cellular & Molecular Biology Letters* **25**, 1-17 (2020).
 29. Yang, S., Wang-Su, S.-T., Cai, H. & Wagner, B. Changes in three types of ubiquitin mRNA and ubiquitin-protein conjugate levels during lens development. *Experimental eye research* **74**, 595-604 (2002).
 30. Zhang, L. et al. Identification and characterization of biomarkers and their functions for Lapatinib-resistant breast cancer. *Medical oncology* **34**, 1-8 (2017).
 31. Kim, J. & Choi, H. The mucin protein MUCL1 regulates melanogenesis and melanoma genes in a manner dependent on threonine content. *British Journal of Dermatology* **186**, 532-543 (2022).

32. Mémin, E. et al. Blocking eIF5A modification in cervical cancer cells alters the expression of cancer-related genes and suppresses cell proliferation. *Cancer research* **74**, 552-562 (2014).
33. Simigdala, N. et al. Cholesterol biosynthesis pathway as a novel mechanism of resistance to estrogen deprivation in estrogen receptor-positive breast cancer. *Breast Cancer Research* **18**, 1-14 (2016).
34. Ni, P.-Z. et al. Overexpression of Stathmin 1 correlates with poor prognosis and promotes cell migration and proliferation in oesophageal squamous cell carcinoma. *Oncology reports* **38**, 3608-3618 (2017).
35. Qiu, J. et al. Large ai models in health informatics: Applications, challenges, and the future. *IEEE Journal of Biomedical Health Informatics* (2023).
36. Yang, Z. et al. The dawn of Imms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:17421* **9**, 1 (2023).
37. Yu, Z. et al. in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 5295-5305 (2020).
38. Yu, Z., Qin, Y., Zhao, H., Li, X. & Zhao, G. Dual-cross central difference network for face anti-spoofing. *arXiv preprint arXiv:01290* (2021).
39. Bai, S., Zhang, F. & Torr, P.H. Hypergraph convolution and hypergraph attention. *Pattern Recognition* **110**, 107637 (2021).
40. Li, B. et al. Multi-scale hypergraph-based feature alignment network for cell localization. *Pattern Recognition*, 110260 (2024).
41. Eraslan, G., Simon, L.M., Mircea, M., Mueller, N.S. & Theis, F.J. Single-cell RNA-seq denoising using a deep count autoencoder. *Nature communications* **10**, 390 (2019).