This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

# ASIGN: An Anatomy-aware Spatial Imputation Graphic Network for 3D Spatial Transcriptomics

Junchao Zhu[1], Ruining Deng[1], Tianyuan Yao[1], Juming Xiong[1], Chongyu Qu[1], Junlin Guo[1], Siqi Lu[1], Mengmeng Yin[2], Yu Wang[2], Shilin Zhao[2], Haichun Yang[2], Yuankai Huo[1,2]*
1. Vanderbilt University, 2. Vanderbilt Univeristy Medical Center
yuankai.huo@vanderbilt.edu

## Abstract

*Spatial transcriptomics (ST) is an emerging technology that enables medical computer vision scientists to automatically interpret the molecular profiles underlying morphological features. Currently, however, most deep learning-based ST analyses are limited to two-dimensional (2D) sections, which can introduce diagnostic errors due to the heterogeneity of pathological tissues across 3D sections. Expanding ST to three-dimensional (3D) volumes is challenging due to the prohibitive costs; a 2D ST acquisition already costs over 50 times more than whole slide imaging (WSI), and a full 3D volume with 10 sections can be an order of magnitude more expensive. To reduce costs, scientists have attempted to predict ST data directly from WSI without performing actual ST acquisition. However, these methods typically yield unsatisfying results. To address this, we introduce a novel problem setting: 3D ST imputation using 3D WSI histology sections combined with a single 2D ST slide. To do so, we present the Anatomy-aware Spatial Imputation Graph Network (ASIGN) for more precise, yet affordable, 3D ST modeling. The ASIGN architecture extends existing 2D spatial relationships into 3D by leveraging cross-layer overlap and similarity-based expansion. Moreover, a multi-level spatial attention graph network integrates features comprehensively across different data sources. We evaluated ASIGN on three public spatial transcriptomics datasets, with experimental results demonstrating that ASIGN achieves state-of-the-art performance on both 2D and 3D scenarios. The code for this paper is publicly available*[1].

## 1. Introduction

The rapid advancement of spatial transcriptomic (ST) has provided researchers with an unprecedented capability to characterize spatial gene expression patterns at both the tissue and cellular levels [4, 7], transforming our understanding of human tissue development and clinical diagnosis [3, 12, 19]. However, current spatial transcriptomics methods [11, 27, 32] are laborious and costly. This, coupled with the need for specialized equipment and expertise, limits their practical applications in real clinical settings.

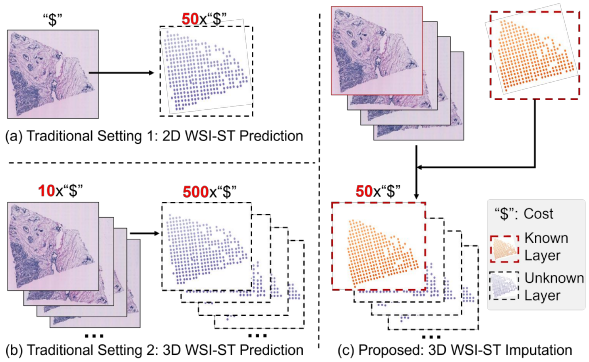


Figure 1. **ASIGN presents a novel problem setting for 3D spatial transcriptomic (ST).** Unlike recent methods that focus on either (a) directly predicting 2D ST from 2D WSI or (b) performing a full 3D WSI to 3D ST prediction, our proposed ASIGN method generates a 3D ST volume by combining 3D WSI histology sections with a single 2D ST slide. This approach provides a balanced solution—more precise than "free but less accurate" 3D predictions, yet far more affordable than the "precise but prohibitively expensive" acquisition of full 3D ST for every slide.

The features of histopathological images are closely correlated with gene expression patterns [2], which offers an opportunity to directly predict spatial transcriptomic expression from images. Driven by the widespread adoption of deep learning models in medical field [12, 16, 25, 41], recently, several studies have employed architectures such as convolutional neural networks (CNNs) [9, 12, 35, 36] and graph neural networks (GNNs) [15, 23, 38] to predict spatial transcriptomic expression from histopathological two-dimensional whole-slide images (WSIs). These methods

[1]Code Availability: https://github.com/hrlblab/ASIGN

incorporate spatial dependencies [35] and image similarities [23], enabling predictions at the WSI level. These approaches provide new pathways for analyzing single-layer tissue sections.

However, there are still many daunting limitations in existing methods and training strategies. First, current approaches only rely on two-dimensional (2D) views of single-layer sections for spatial transcriptomic analysis and training, which neglects the rich structural information of organs and tissues in three-dimensional (3D) space. It might lead to diagnostic errors due to the heterogeneity of pathological tissues across 3D sections. Moreover, existing models exhibit noticeable deficiencies in generalizability, especially at the WSI level. This is mainly due to significant variance in gene expression across patients [28, 33]. Additionally, variations in staining styles [6, 26] and technical discrepancies in sequencing methods [40] across different data sources further contribute to this issue. Consequently, the performance of these models is often degraded when applied to new samples.

In this paper, We introduce an Anatomy-aware Spatial Imputation Graph Network (ASIGN) tailored for 3D ST analysis within a novel problem framework as is described in Figure 1. Unlike recent approaches that either predict 2D ST directly from 2D WSI or perform full 3D WSI to 3D ST prediction, ASIGN creates the 3D ST volume by integrating 3D WSI histology sections with a single 2D ST slide. This method strikes a balance between "free but less accurate" 3D predictions and the "precise yet prohibitively costly" acquisition of full 3D ST on a slide-by-slide basis. ASIGN provides a new cost-effective route towards 3D ST, achieving a significant cost reduction while maintaining robust performance.

To do so, ASIGN employs multiple registration methods to establish inter-layer overlaps at the 3D sample level, effectively constructing spot connections across samples. It also integrates a Multi-Level Spatial Attention Graph Network (MSAGNet) to enhance information fusion across layers, neighboring regions, and resolutions, enabling comprehensive feature integration. Additionally, ASIGN shifts training and prediction from an entirely unknown 2D WSI level to a partially known 3D sample level. To improve the imputation performance, we introduce a Cross-layer Imputation (CLI) block that propagates labels from known to unknown layers, enhancing the model's generalizability on unseen tissue samples.

Our contributions can be summarized as three folds:

- We introduce a new learning paradigm that shifts from 2D WSI-ST predictions to partially known 3D volumetric WSI-ST imputation. This is facilitated by a Cross-layer Imputation (CLI) block, which propagates information across layers to achieve more precise, yet affordable, 3D ST modeling.
- We present ASIGN, a novel framework for 3D ST prediction that transforms isolated 2D spatial relationships into a cohesive 3D structure through inter-layer overlap and similarity, integrating 3D spatial information into 3D ST imputation.
- A new multi-level spatial attention graphic network is proposed to facilitate comprehensive feature integration across different layers, neighboring regions, and multiple resolutions, thus enabling precise sample-level predictions.

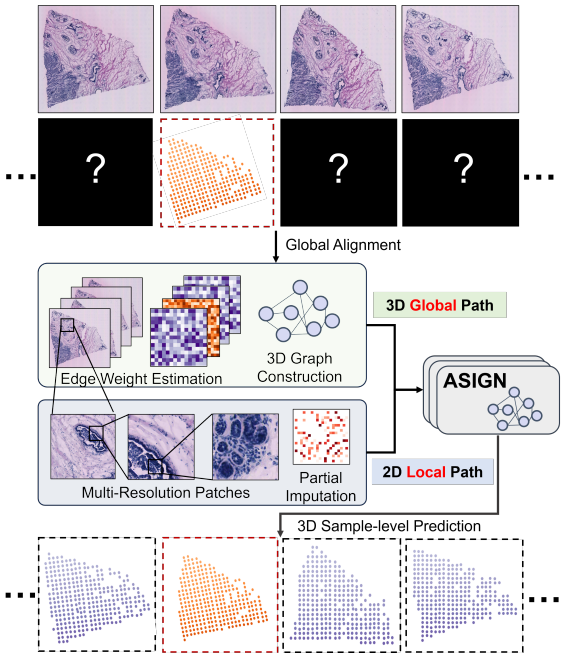


Figure 2. **Overall framework of our proposed ASIGN approach.** ASIGN begins with a global alignment process to estimate overlap and similarity across layers, constructing 3D graph connections. Then, MSAGNet, integrated within ASIGN, enhances information fusion across layers, neighboring regions, and resolutions for comprehensive feature integration. Finally, a CLI block propagates predictions from partially known labels, which are weighted with MSAGNet's model predictions to generate the final output.

## 2. Related Work

### 2.1. Deep Learning Approaches for ST Prediction

The advancement of artificial intelligence in recent years has provided powerful tools for medical image analysis [8, 17, 18, 35]. Deep learning methods have led to the emergence of various approaches for spatial gene expression prediction [9, 12, 23, 35, 36, 38]. ST-Net [12] is the earliest proposed method. It employs a transfer learning strategy and uses DenseNet121 [14] as the backbone to train a patch-to-spot regression model for predicting gene expression.

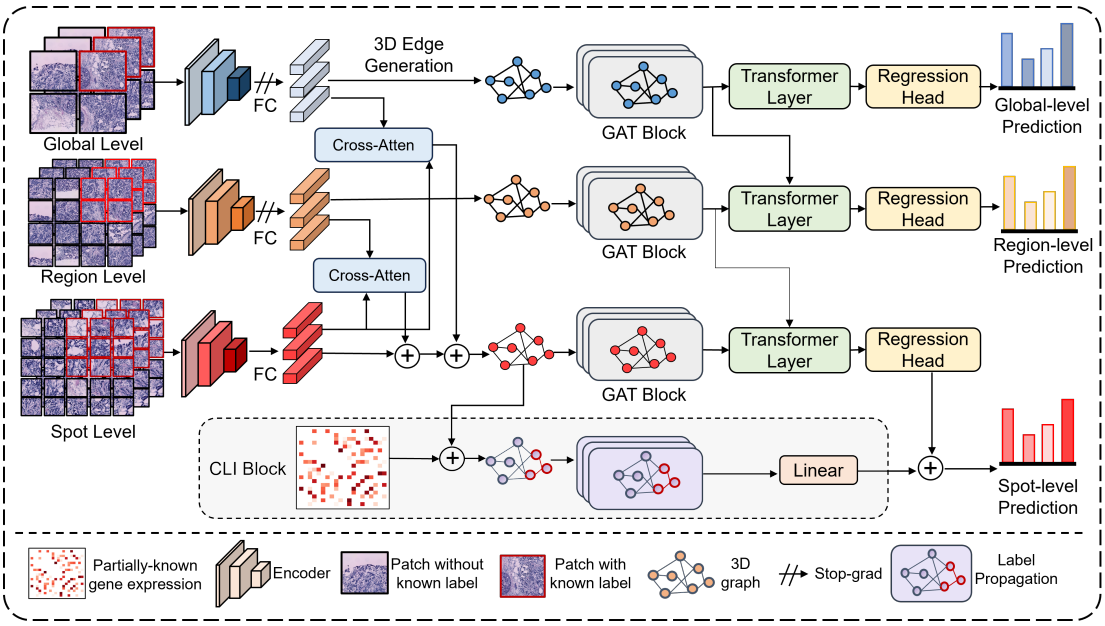


Figure 3. **Architecture of MSAGNet and the CLI block.** MSAGNet comprises cross-attention layers, GAT blocks, and Transformer layers to integrate and aggregate features across multiple resolution levels, 3D sample levels, and patch-self levels, respectively. The CLI block employs a label propagation strategy to diffuse known labels to unknown patches. A self-adaptive weighting mechanism merges the final results from both MSAGNet and the CLI block.

Following this, mainstream methods have primarily focused on leveraging image similarity [35, 36] and spatial relationships [23, 38] to enhance model performance. EGN [36] introduces an exemplar learning mechanism to dynamically select the most similar point within the WSI for a target point. In contrast, BLEEP [35] utilizes a bi-modal embedding architecture to estimate gene expression at queried points by identifying similar expression profiles. However, variations in staining styles and specimen preparation can lead to significant differences at the gene expression level for points identified through image similarity. This limits generalization capability when models encounter entirely new, unseen samples.

On the other hand, methods such as HisToGene [23] and His2ST[38] enhance performance by incorporating spatial relationships between spots within a single-layer WSI, utilizing Transformers and graph neural networks to capture global and neighborhood spatial information. TRIPLEX [9] introduces multi-resolution information centered on spots, integrating spatial information between patches and WSIs for prediction. However, these methods are limited to single-layer 2D spatial relationships and overlook the overall distribution of tissues in 3D space, thereby constraining a comprehensive understanding of spatial interactions.

In contrast, our proposed ASIGN framework leverages the integration of 3D-level information associations and sample-level training strategies to better capture the spatial distribution of information, thus significantly enhancing the generalization capability for cross-sample predictions.

## 2.2. 3D ST Analysis and Label Imputation

As spatial transcriptomics technology has gradually developed in recent years, some studies have successfully constructed 3D spatial transcriptomic data for the mouse brain [22] and cancer tissues [1]. Researchers conducted an in-depth analysis of 3D spatial transcriptomic data to construct a dynamic map of tissue changes across temporal and spatial dimensions, revealing the pathways of disease progression [34] and the complex mechanisms of gene expression and spatial structural changes during organismal development [20, 31].

However, building up a 3D ST dataset is extremely costly. To narrow this gap, other studies have attempted to use this 3D spatial transcriptomic information to reconstruct and predict gene expression patterns. Existing 3D spatial transcriptomic methods primarily align different slice layers using gene-level expression similarities between spots [30, 37]. For example, PASTE [37] employs Gromov-Wasserstein optimal transport to probabilistically align neighboring slices based on transcriptional and spatial similarities. Stitch3D [30] uses a graph neural network to map gene expression and spatial data from multiple slices into a shared latent space to reconstruct 3D spatial transcriptomic. However, due to deformation and displacement during slicing, spot-level 3D reconstruction often fails to capture true spot distributions. BLEEP [35] proposes query-

reference imputation to directly weight and predict labels for unknown spots, but information shifts between spots and questionable query reliability have limited its effectiveness.

On the contrary, our approach performs registration at the WSI level to avoid spot-level deformation and displacement. We also optimize label information by combining deep learning with a cross-layer imputation block to improve label imputation accuracy.

# 3. Method

## 3.1. Overall Framework

Our proposed ASIGN architecture consists of three main components, including 3D graph construction, MSAGNet, and the CLI block. Figure 2 shows the whole process of our ASIGN framework. First, we align the original 2D single-layer WSIs using a global alignment method to estimate cross-layer overlap and similarity, thereby facilitating the construction of spot-level 3D graph connections. Then, we leverage MSAGNet to integrate multi-resolution image information and incorporate 3D spatial relationships, enhancing the model's representation capabilities. Finally, the CLI module propagates known labels to unknown spots, combining adaptive weights with model predictions to generate the final results.

## 3.2. 3D Graph Construction

WSI-wise alignment is a critical prerequisite for obtaining 3D information distribution. First, we utilize XFeat [24] to extract features from each WSI sample and align them using Advanced Normalization Tools (ANTs) [29]. We take the middle layer of each sample as a reference for image registration.

After alignment, for the $j$-th spot in the $i$-th layer, denoted as $s_{ij}$, we draw a circle centered at its midpoint, denoted as $c_{ij}$. Subsequently, we calculate the spatial connection weight between this point and any spot $s_{pq}$ on other layers using Intersection over Union and cosine similarity, as follows:

$$e_{ij}^{pq} = \frac{|c_{ij} \cap c_{pq}|}{|c_{ij} \cup c_{pq}|} + \frac{\vec{c}_{ij} \cdot \vec{c}_{pq}}{\|\vec{c}_{ij}\| \times \|\vec{c}_{pq}\|} \tag{1}$$

where $\vec{c}_{ij}$ is the extracted feature from $c_{ij}$. For spots $s_{ik}$ on the same layer$i$, we employ the Euclidean Distance to compute their connection weight, defined as $e_{ij}^{ik}$. Finally, for each $s_{ij}$, we select the Top-k values of $e_{ij}^{pq}$ and $e_{ij}^{ik}$ to build up connections with spots other layers and current layer respectively. Then the constructed 3D graph is utilized in the training of MSAGNet.

## 3.3. MASGNet

Figure 3 illustrates the network structure of MSAGNet and the CLI block. To effectively integrate multi-level information, MSAGNet leverages cross-attention layers to fuse features from multi-resolution patches and employs a GAT-Transformer block to aggregate sample-level 3D information. Subsequently, a regression head generates predictions for each resolution level independently.

### 3.3.1. Hierarchical Fusion of Multi-Level Features

For each spot $s$, the paired images consist of spot-level, region-level, and global-level patches, denoted as $i_s$, $i_r$, and $i_g$ respectively. A pretrained ResNet50 [13] is employed as the encoder for the feature extraction process to obtain features at different resolution levels, denoted as $f_s$, $f_r$, and $f_g$. To reduce computational costs, we follow the approach proposed by TRIPLEX [9] and freeze the parameters of the encoders for $i_r$ and $i_g$, only updating the parameters of the spot-level encoder.

We employ a cross-attention layer to incorporate information from $f_r$ and $f_g$ into $f_s$, optimizing the representation of $f_s$ by aligning multi-source information. Taking the fusion with region-level information as an example, we treat $f_s$ as the query matrix and $f_r$ as the key and value matrices. The fused feature $f_s'$ is given by the following formula:

$$f_s' = \text{softmax}\left(\frac{f_s f_r^T}{\sqrt{d}}\right) f_r \tag{2}$$

where $\sqrt{d}$ is a scaling factor based on the feature dimension. Similarly, we can obtain the fused feature $f_s''$ by aligning $f_s$ with global-level information. Finally, we concatenate the features from the three levels to obtain the fused multi-level feature $F$, represented as:

$$F = \text{concat}(f_s, f_s', f_s'') \tag{3}$$

### 3.3.2. Spatial-aware Graph Attention Block

To integrate feature information in 3D space, we employ a Spatial-aware Graph Attention Block that combines multi-layer Graph Attention Networks and Transformer layers to achieve feature fusion.

Based on the 3D graph construction strategy proposed in Section 3.1, we establish edge connections for each resolution level, forming a graph structure in the following GAT block for processing. Taking the feature $F_i$ corresponding to spot $s_i$ as an example, after rounds of graph convolution, the processed feature $F_i'$ is expressed as follows:

$$\mathbf{F_i'} = \Big\|_{k=1}^{K} \sigma\left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij}^k \mathbf{W}^k \mathbf{F}_j\right) \tag{4}$$

where $\mathcal{N}(i)$ denotes the set of neighboring nodes, $\Big\|$ represents the concatenation operation of multi-head attention, $\sigma$ is the activation function, $\alpha_{ij}^k$ is the attention weight of the $k$-th attention head, and $\mathbf{W}^k$ is a linear transformation matrix determined by the connections between nodes.

The outputs $F_i'$ from each round are concatenated and, together with features processed at the Global-level and Region-level, are fed into a Transformer layer to achieve flexible information aggregation. Finally, a regression head generates predictions for the global-level, region-level, and spot-level, denoted as $p_g$, $p_l$, and $p_s$ respectively.

### 3.4. Cross-layer Imputation

To further enhance the generalization capability of the 3D framework, we introduce a Cross-layer Imputation block designed to propagate known gene expression label information from one layer to unknown spots in other layers. For nodes with known labels, their labels remain fixed during the propagation process. For an unknown node $n_i$, the prediction result $p_i'$ after $t+1$ iterations can be given as:

$$p_i'^{(t+1)} = \frac{\sum_{j \in \mathcal{N}(i)} \alpha_j \cdot e_{ij} \cdot p_j'^{(t)}}{\sum_{j \in \mathcal{N}(i)} e_{ij}} \tag{5}$$

where $\mathcal{N}(i)$ denotes the set of neighbors of node $i$, $\alpha_j$ is a learnable weight for node j, $e_{ij}$ represents the edge weight between node $i$ and its neighboring node $j$ as mentioned in Section 3.1, and $p_j'^{(t)}$ is the label distribution of neighboring node $j$ at the $t$-th iteration. The denominator $\sum_{j \in \mathcal{N}(i)} e_{ij}$ serves as a normalization factor.

Subsequently, we combine the model prediction results and the imputation results using an adaptive weight $\alpha$, yielding the final spot-level prediction result $p_s'$ given by:

$$p_s' = \alpha \cdot p_s + (1 - \alpha) \cdot p_i' \tag{6}$$

### 3.5. Loss Function

We adopt a hybrid loss function to optimize the model's integration and learning of multi-level information. The overall loss function consists of multi-level prediction loss ($L_p$) and multi-level consistency loss ($L_c$). The multi-level prediction loss evaluates the discrepancy between the model's predictions and the true labels. At the spot level, we use Mean Squared Error (MSE) and Pearson Correlation Coefficient loss (PCC loss) to assess the model's prediction performance. For the region-level and global-level, we only use PCC loss ($P$) to estimate the model's prediction performance, thereby avoiding the introduction of extra noise. Therefore, $L_p$ can be described as:

$$L_p = MSE(p_s, y_s) + \sum_{i=s,r,g} \lambda_i \cdot P(p_i, y_i) \tag{7}$$

where $s$, $r$, and $g$ represent spot-level, region-level, and global-level, respectively. $p_i$ and $y_i$ denote the model's predictions and labels, and $\lambda_i$ is a hyperparameter used to balance the PCC loss across resolutions.

We further constrain the predictions across different resolutions by a cross-resolution consistency loss. Since the gene expressions of patches at different resolutions share similar trends, we employ PCC loss to assess the differences between spot-level prediction and region-level as well as global-level predictions. $L_c$ is formulated as:

$$L_c = \lambda_1 \cdot P(p_s, p_r) + \lambda_2 \cdot P(p_s, p_g) \tag{8}$$

Thus, the overall loss of the model $L$ is described as:

$$L = \gamma_1 \cdot L_p + \gamma_2 \cdot L_c \tag{9}$$

where $\gamma_1$ and $\gamma_2$ are hyperparameters used to balance the two losses and are set as 1 in the following experiments.

## 4. Data and Experiments

**Dataset and pre-processing procedure.** We utilized three public datasets for fair cross-validation: two breast cancer datasets [1, 12] and one human dorsolateral prefrontal cortex (DLPFC) dataset [20]. The HER2-positive breast tumor dataset is referred to as HER2, while the other breast dataset proposed by ST-Net is denoted as ST-Data. The HER2 dataset consists of 4 samples with 6-layer sections and 4 samples with 3-layer sections, and a total of 13,620 spots. The DLPFC dataset contains 3 samples with 4-layer sections and 42,474 spots. For optimal global alignment, we selected 16 three-layer samples from the 23 available samples in ST-Data, totaling 41,544 spots.

With over 20,000 original gene dimensions, predicting all gene expressions from histological image patches is impractical. Therefore, we referred to the method mentioned in ST-Net [12] selecting the top 250 genes with the highest average expression for prediction. The normalization of gene expression values was performed following the strategy proposed by TRIPLEX [9], with a proportional normalization followed by a log transformation.

We cropped patches at different levels into sizes: spot-level, region-level, and global-level, with dimensions $224 \times 224$, $512 \times 512$, and $1024 \times 1024$ pixels, respectively. Spot-level patches were centered on the spot, while other patches used overlaps to find corresponding spot patches.

**Evaluation metrics.** The Pearson correlation coefficient (PCC), mean squared error (MSE), and mean absolute error (MAE) is employed as the evaluation metrics to provide a comprehensive assessment of the models' performance in gene expression prediction tasks.

**Implementation.** The experiments were conducted using two NVIDIA RTX A6000 GPU cards. We utilized an SGD optimizer configured with a momentum of 0.9 and a weight decay parameter set to $10^{-4}$. The initial learning rate, denoted as $lr_0$, was set to $10^{-4}$. A cosine decay strategy was applied to reduce the learning rate to 0.01 of its initial value during training. We employed a batch size of 256 for training and fine-tuned the hyperparameters $\lambda_1$, $\lambda_2$, $\lambda_s$, $\lambda_r$, and $\lambda_g$ in our hybrid loss function to values of 0.25, 0.25, 0.75,

Table 1. Quantitative comparisons in both 2D WSI-wise and 3D Sample-wise methods across different public datasets.

| Dimension | Model | HER2[1] | | | DLPFC[20] | | | ST-Data[12] | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | MSE | MAE | PCC | MSE | MAE | PCC | MSE | MAE | PCC |
| 2D WSI-wise | ST-Net[12] | 0.663 | 0.633 | 0.430 | 0.496 | 0.552 | 0.738 | 0.804 | 0.705 | 0.517 |
| | EGN [36] | 0.580 | 0.585 | 0.442 | 0.401 | 0.490 | 0.757 | 0.674 | 0.644 | 0.554 |
| | HisToGene[23] | 0.669 | 0.644 | 0.432 | 0.348 | 0.472 | 0.753 | 0.575 | 0.598 | 0.552 |
| | BLEEP [35] | 0.743 | 0.665 | 0.340 | 0.304 | 0.429 | 0.737 | 0.581 | 0.590 | 0.496 |
| | His2ST [38] | 0.591 | 0.591 | 0.441 | 0.409 | 0.494 | 0.764 | 0.534 | 0.568 | 0.539 |
| | **ASIGN 2D (Ours)** | 0.579 | 0.590 | 0.476 | **0.275** | **0.411** | 0.766 | 0.528 | 0.566 | 0.552 |
| 3D Sample-wise | Overlap | 0.461 | 0.532 | 0.625 | 0.292 | 0.422 | 0.749 | 0.447 | 0.525 | 0.663 |
| | Similarity | 0.449 | 0.517 | 0.649 | 0.278 | 0.412 | 0.762 | 0.579 | 0.548 | 0.696 |
| | **ASIGN 3D (Ours)** | **0.363** | **0.473** | **0.693** | 0.353 | 0.458 | **0.773** | **0.332** | **0.459** | **0.739** |

0.5, and 0.5, respectively. We set the ratio of known labels as 0.2 in the CLI block to achieve the ideal performance. The top k selection for 3D graph construction is set as 8 for inter-layer spots and 12 for cross-layer spots.

### 4.1. Experimental Setting

**Sample-level cross-validation strategy** We performed cross-validation at the sample level across the three datasets, utilizing four-fold cross-validation for HER2 and ST-Data, and three-fold cross-validation for DLPFC to account for dataset size differences. During validation, one layer was selected as the known label layer and included in the training set, while other unknown layers were reserved for the test set. This approach ensures that all models have equitable access to known information, thereby maintaining a consistent and fair evaluation framework and allowing for a comprehensive assessment of model performance across varying data distributions.

**Baseline methods** We benchmarked our model's performance against current SOTA methods, including 1) 2D-level CNN-based models such as ST-Net [12], EGN [36], and BLEEP [35], and 2) 2D-level Transformer and GNN-based networks like HisToGene[23] and His2ST[38]. Additionally, we designed and introduced two 3D-based imputation methods inspired by PASTE [37], which utilize spot overlap information and patch similarity to enhance predictions. All baseline models were trained and evaluated under consistent conditions for fair comparison.

## 5. Results

### 5.1. Empirical Validation

The cross-validation performance of different baseline models across three datasets is summarized in Table 1. Our proposed ASIGN framework consistently outperformed existing approaches in terms of MSE, MAE, and PCC metrics across both the 2D WSI level and the 3D sample level experiments, with particularly strong results observed for the ASIGN 3D model. For instance, in the HER2 dataset, known for its high inter-sample heterogeneity, the ASIGN 3D model achieved MSE, MAE, and PCC values of 0.363±0.054, 0.473±0.053, and 0.693±0.091, respectively. This marks a substantial improvement over other methods, such as the 3D similarity-based approach (PCC = 0.649±0.130) and the 2D EGN model (PCC = 0.442±0.152). These results indicate that the ASIGN framework is particularly adept at capturing complex data relationships that are critical for accurate modeling in heterogeneous datasets.

Moreover, the boxplot presented in Figure 4 highlights the generalization capability and prediction stability of the models under cross-dataset validation. Our ASIGN architecture exhibited the smallest prediction deviation across different datasets, underscoring its robust cross-sample prediction generalization capabilities. As it ensures consistent performance across diverse patient data, further enhancing the framework's practical value.

This significant improvement stems from the ASIGN framework's exceptional ability to accurately capture and integrate multi-dimensional information through the construction of cross-layer connections, rather than focusing solely on intra-layer information as is typical with existing methods. The ASIGN framework also facilitates the effective propagation of known layer information into unexplored regions, thereby enhancing the model's generalization ability and providing a more applicable approach to clinical uses.

### 5.2. Cancer Marker Gene Expression Prediction

We identified two key cancer markers, Human Epidermal Growth Factor Receptor 2 (ERBB2) [10, 21] and Midkine (MDK)[5, 39], which are highly associated with HER2 and have significant clinical relevance, to evaluate the model's applicability in real clinical scenarios.

Figure 5 visualizes the spatial distribution of ERBB2 and MDK in WSIs, along with the MSE metrics of different

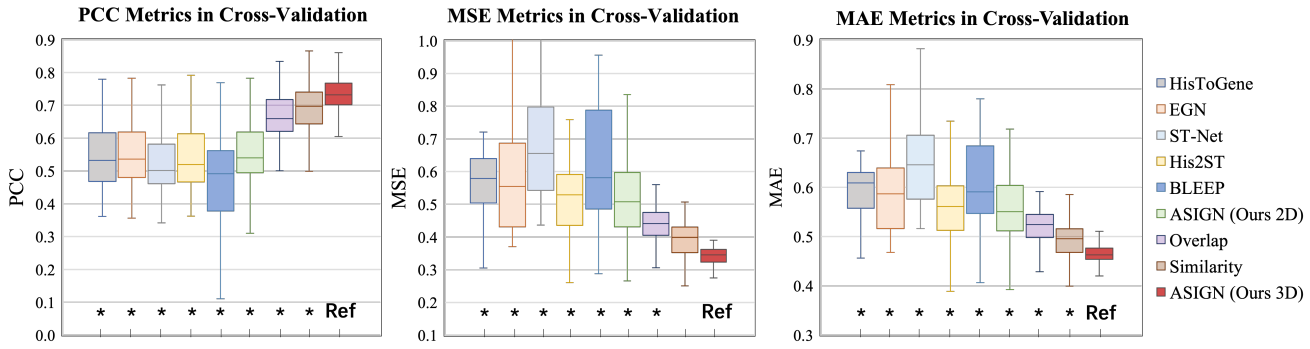


Figure 4. **Quantitative comparison between different methods.** This figure presents the box plot of different methods with Pearson correlation coefficient (PCC), mean squared error (MSE), and mean absolute error (MAE). Welch's t-test is used to assess the statistical significance of differences between other method and ASIGN. P-values below 0.05 are considered significant and are annotated with $*$.

models for these critical genes. Our model achieved the lowest MSE for ERBB2 and MDK, at 0.1421 and 0.1791, respectively. In contrast, some 2D-based methods, such as HisToGene (0.8200) and His2ST (0.7410), failed to effectively predict these key cancer markers due to significant variance in cross-sample prediction, underscoring their limitations in capturing cross-sample spatial heterogeneity. By comparison, our ASIGN framework was able to nearly replicate the gene expression spatial distribution of these key markers, demonstrating its robustness and superior capacity to capture complex patterns. This makes it more applicable for use by pathologists in clinical practice and further enhances its practical utility in real-world scenarios.

### 5.3. Ablation Study

We designed three sets of ablation experiments to verify the effectiveness and necessity of each module within the ASIGN framework. These experiments include 1) the comparison between 2D-based and 3D-based graph construction 2) the network functional block, and 3) the CLI block and the proportion of known labels

We adhered to the cross-validation procedure employed in the comparison experiments to ensure a fair evaluation of the ablation study. We trained the models on the HER2 dataset using identical hyperparameter settings, which allowed us to objectively assess the impact of each module and isolate their individual contributions to the overall performance of ASIGN.

**Function blocks in ASIGN** The experimental results for the ablation study of the functional blocks incorporated in ASIGN are shown in Table 2. The cross-attention layer for fusing multi-resolution features and the GAT blocks for capturing 3D spatial information improved the PCC metric from 0.659 to 0.677 and 0.686, respectively. Moreover, when both blocks were utilized together, the model's performance further increased to 0.693. Notably, the benefits from the two blocks are orthogonal, enabling ASIGN to achieve optimal performance when both are employed together.

This indicates that the cross-attention layer plays a crucial role in effectively integrating 2D multi-resolution information, enabling the model to capture and align details at different scales. Meanwhile, the GAT blocks are responsible for processing and handling 3D spatial information, providing an additional layer of contextual understanding through 3D spatial relationships. By strategically combining these multi-level and multi-source features, the model can develop a more holistic data representation.

Table 2. Ablation study for functional blocks in ASIGN

| Functional Blocks | MSE | MAE | PCC |
|---|---|---|---|
| w.o GAT & Cross-atten | 0.674 | 0.531 | 0.659 |
| w.o Cross-atten | 0.578 | 0.692 | 0.686 |
| w.o GAT | 0.466 | 0.627 | 0.677 |
| w. Both blocks | **0.363** | **0.473** | **0.693** |

**CLI block and proportion for known labels.** Table 3 illustrates the impact of different proportions of known labels on the model's predictive performance. When the CLI block is introduced to propagate known labels to unknown labels, the model's performance sees a significant improvement, with the PCC metric increasing from 0.635 to 0.693. This underscores the essential role of the CLI block and label imputation in enhancing the effectiveness of our proposed ASIGN framework.

Additionally, as the proportion of known labels continues to increase, the model's performance progressively improves. Once the proportion of known labels reaches the range of 20-30%, the model performance stabilizes, indicating convergence. This suggests that when the proportion of known labels reaches a certain threshold, increasing similarity difference and spatial distance will introduce additional noise and limit gains, thus affirming that a relatively

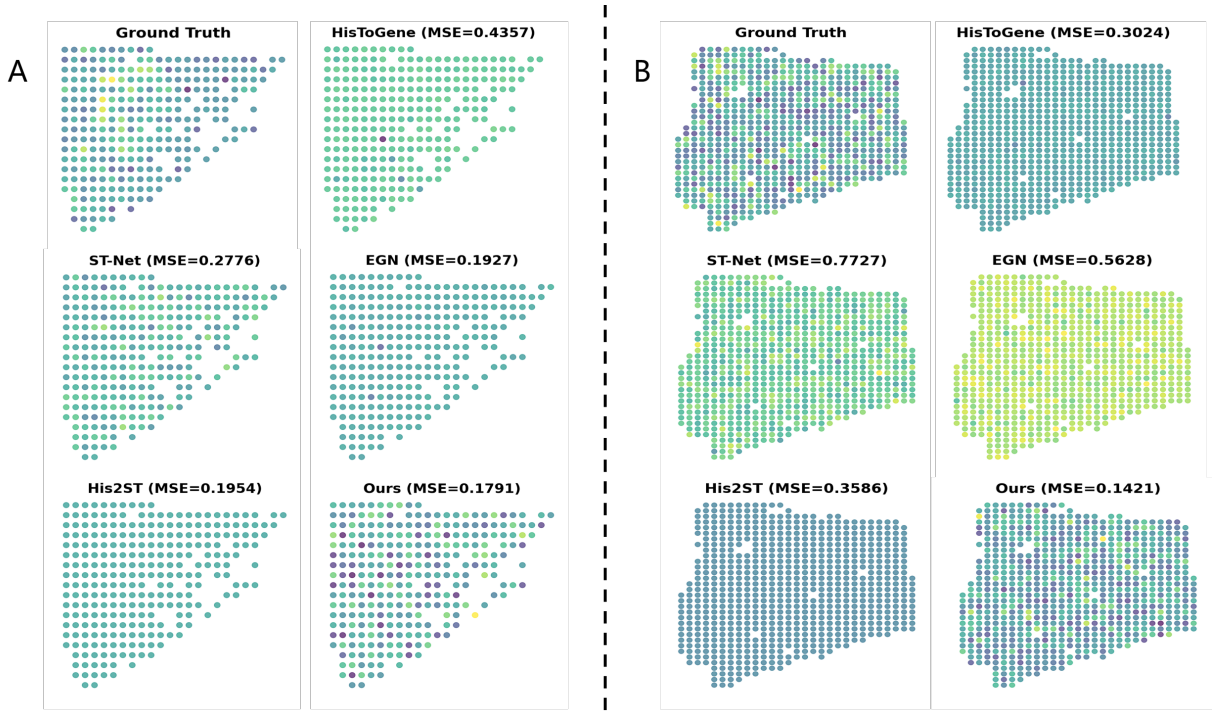


Figure 5. **Qualitative comparison between different methods.** This figure presents the ST imputation performance for predicting the distribution of cancer markers, specifically (A) ERBB2 and (B) MDK, within the WSIs. The MSE between the ground truth and the predicted values is presented for each model.

small proportion of known labels can yield optimal predictive outcomes within our proposed ASIGN framework.

Table 3. Ablation study for CLI block and known label proportion

| Proportion of Known Label | MSE | MAE | PCC |
|---|---|---|---|
| w. 0% (w.o CLI) | 0.622 | 0.606 | 0.635 |
| w. 10% | 0.542 | 0.563 | 0.677 |
| w. 20% (ASIGN) | **0.363** | **0.473** | 0.693 |
| w. 30% | 0.397 | 0.492 | **0.708** |
| w. 40% | 0.439 | 0.512 | 0.703 |

**Comparison of 2D and 3D graph construction.** Table 4 compares the performance of 2D-based and 3D-based graph construction on the HER2 dataset. After incorporating 3D spatial information, the model's PCC score significantly increased from 0.476 to 0.635. This demonstrates that adding an extra spatial dimension allows the model to better capture gene correlations using 3D spatial information and reconstruct the true three-dimensional structure of tissues. Consequently, this wealth of comprehensive information allows for a more detailed and precise analysis of gene expression patterns, significantly enhancing the model's predictive capability and overall accuracy, thus making reliable predictions across diverse datasets.

Table 4. Ablation study for graph construction

| Graph Construction | MSE | MAE | PCC |
|---|---|---|---|
| 2D WSI-wise | **0.579** | **0.590** | 0.476 |
| 3D Sample-wise | 0.622 | 0.606 | **0.635** |

## 6. Conclusion

We propose a new learning paradigm for 3D ST modeling that imputes full 3D volumetric ST data using 3D WSI sections combined with a single 2D ST acquisition. This approach provides a balanced solution—more precise than "free but less accurate" 3D predictions, yet far more affordable than the "precise but prohibitively expensive" acquisition of full 3D ST for every slide. To this end, our ASIGN framework integrates multi-level features, combining local multi-resolution patches with global 3D sample-wise graph connections to enhance prediction and imputation outcomes. Experimental results demonstrate ASIGN's superior performance and robust generalization in cross-sample validation across multiple public datasets, significantly improving the PCC metric for gene expression tasks from $\approx$0.5 with existing methods to $\approx$0.7. ASIGN offers a promising pathway to achieve accurate and cost-effective 3D ST data for real-world clinical applications.

## 7. Acknowledgement

This research was supported by NIH R01DK135597 (Huo), DoD HT9425-23-1-0003 (HCY), and KPMP Glue Grant. This work was also supported by Vanderbilt Seed Success Grant, Vanderbilt Discovery Grant, and VISE Seed Grant. This project was supported by The Leona M. and Harry B. Helmsley Charitable Trust grant G-1903-03793 and G-2103-05128. This research was also supported by NIH grants R01EB033385, R01DK132338, REB017230, R01MH125931, and NSF 2040462. We extend gratitude to NVIDIA for their support by means of the NVIDIA hardware grant. This work was also supported by NSF NAIRR Pilot Award NAIRR240055.

## References

[1] Alma Andersson, Ludvig Larsson, Linnea Stenbeck, Fredrik Salmén, Anna Ehinger, Sunny Z Wu, Ghamdan Al-Eryani, Daniel Roden, Alex Swarbrick, Åke Borg, et al. Spatial deconvolution of her2-positive breast cancer delineates tumor-associated cell type interactions. *Nature communications*, 12 (1):6012, 2021. 3, 5, 6

[2] Jordan T Ash, Gregory Darnell, Daniel Munro, and Barbara E Engelhardt. Joint analysis of expression levels and histological images identifies genes associated with tissue morphology. *Nature communications*, 12(1):1609, 2021. 1

[3] Michaela Asp, Stefania Giacomello, Ludvig Larsson, Chenglin Wu, Daniel Fürth, Xiaoyan Qian, Eva Wärdell, Joaquin Custodio, Johan Reimegård, Fredrik Salmén, et al. A spatiotemporal organ-wide gene expression and cell atlas of the developing human heart. *Cell*, 179(7):1647–1660, 2019. 1

[4] Michaela Asp, Joseph Bergenstråhle, and Joakim Lundeberg. Spatially resolved transcriptomes—next generation tools for tissue exploration. *Bioessays*, 42(10):1900221, 2020. 1

[5] Ahmet Şükrü Aynacıoğlu, Ayhan Bilir, and Mehmet Yakup Tuna. Involvement of midkine in autoimmune and autoinflammatory diseases. *Modern rheumatology*, 29(4):567–571, 2019. 6

[6] Aïcha BenTaieb and Ghassan Hamarneh. Adversarial stain transfer for histopathology image analysis. *IEEE transactions on medical imaging*, 37(3):792–802, 2017. 2

[7] Darren J Burgess. Spatial transcriptomics coming of age. *Nature Reviews Genetics*, 20(6):317–317, 2019. 1

[8] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021. 2

[9] Youngmin Chung, Ji Hun Ha, Kyeong Chan Im, and Joo Sang Lee. Accurate spatial gene expression prediction by integrating multi-resolution features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11591–11600, 2024. 1, 2, 3, 4, 5

[10] Susan Dent, Basak Oyan, Arnd Honig, Max Mano, and Sacha Howell. Her2-targeted therapy in breast cancer: a systematic review of neoadjuvant trials. *Cancer treatment reviews*, 39(6):622–631, 2013. 6

[11] Chee-Huat Linus Eng, Michael Lawson, Qian Zhu, Ruben Dries, Noushin Koulena, Yodai Takei, Jina Yun, Christopher Cronin, Christoph Karp, Guo-Cheng Yuan, et al. Transcriptome-scale super-resolved imaging in tissues by rna seqfish+. *Nature*, 568(7751):235–239, 2019. 1

[12] 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, 5, 6

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[14] 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

[15] Yuran Jia, Junliang Liu, Li Chen, Tianyi Zhao, and Yadong Wang. Thitogene: a deep learning method for predicting spatial transcriptomics from histological images. *Briefings in Bioinformatics*, 25(1):bbad464, 2024. 1

[16] Jing Ke, Yizhou Lu, Yiqing Shen, Junchao Zhu, Yijin Zhou, Jinghan Huang, Jieteng Yao, Xiaoyao Liang, Yi Guo, Zhonghua Wei, et al. Clusterseg: A crowd cluster pinpointed nucleus segmentation framework with cross-modality datasets. *Medical Image Analysis*, 85:102758, 2023. 1

[17] Jing Ke, Junchao Zhu, Xin Yang, Haolin Zhang, Yuxiang Sun, Jiayi Wang, Yizhou Lu, Yiqing Shen, Sheng Liu, Fusong Jiang, et al. Tshfna-examiner: A nuclei segmentation and cancer assessment framework for thyroid cytology image. *Journal of Shanghai Jiaotong University (Science)*, 29 (6):945–957, 2024. 2

[18] Wenxuan Li, Chongyu Qu, Xiaoxi Chen, Pedro RAS Bassi, Yijia Shi, Yuxiang Lai, Qian Yu, Huimin Xue, Yixiong Chen, Xiaorui Lin, et al. Abdomenatlas: A large-scale, detailed-annotated, & multi-center dataset for efficient transfer learning and open algorithmic benchmarking. *Medical Image Analysis*, 97:103285, 2024. 2

[19] Silas Maniatis, Tarmo Äijö, Sanja Vickovic, Catherine Braine, Kristy Kang, Annelie Mollbrink, Delphine Fagegaltier, Žaneta Andrusivová, Sami Saarenpää, Gonzalo Saiz-Castro, et al. Spatiotemporal dynamics of molecular pathology in amyotrophic lateral sclerosis. *Science*, 364(6435): 89–93, 2019. 1

[20] Kristen R Maynard, Leonardo Collado-Torres, Lukas M Weber, Cedric Uytingco, Brianna K Barry, Stephen R Williams, Joseph L Catallini, Matthew N Tran, Zachary Besich, Madhavi Tippani, et al. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nature neuroscience*, 24(3):425–436, 2021. 3, 5, 6

[21] Arjun Mehta and Debu Tripathy. Co-targeting estrogen receptor and her2 pathways in breast cancer. *The breast*, 23(1): 2–9, 2014. 6

[22] Cantin Ortiz, Jose Fernandez Navarro, Aleksandra Jurek, Antje Märtin, Joakim Lundeberg, and Konstantinos Meletis. Molecular atlas of the adult mouse brain. *Science advances*, 6(26):eabb3446, 2020. 3

[23] Minxing Pang, Kenong Su, and Mingyao Li. Leveraging information in spatial transcriptomics to predict super-resolution gene expression from histology images in tumors. *BioRxiv*, pages 2021–11, 2021. 1, 2, 3, 6

[24] Guilherme Potje, Felipe Cadar, André Araujo, Renato Martins, and Erickson R Nascimento. Xfeat: Accelerated features for lightweight image matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2682–2691, 2024. 4

[25] Chongyu Qu, Tiezheng Zhang, Hualin Qiao, Yucheng Tang, Alan L Yuille, Zongwei Zhou, et al. Abdomenatlas-8k: Annotating 8,000 ct volumes for multi-organ segmentation in three weeks. *Advances in Neural Information Processing Systems*, 36:36620–36636, 2023. 1

[26] Marlen Runz, Daniel Rusche, Stefan Schmidt, Martin R Weihrauch, Jürgen Hesser, and Cleo-Aron Weis. Normalization of he-stained histological images using cycle consistent generative adversarial networks. *Diagnostic Pathology*, 16: 1–10, 2021. 2

[27] 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

[28] Dylan J Taylor, Surya B Chhetri, Michael G Tassia, Arjun Biddanda, Stephanie M Yan, Genevieve L Wojcik, Alexis Battle, and Rajiv C McCoy. Sources of gene expression variation in a globally diverse human cohort. *Nature*, 632(8023): 122–130, 2024. 2

[29] Nicholas J Tustison, Philip A Cook, Andrew J Holbrook, Hans J Johnson, John Muschelli, Gabriel A Devenyi, Jeffrey T Duda, Sandhitsu R Das, Nicholas C Cullen, Daniel L Gillen, et al. The antsx ecosystem for quantitative biological and medical imaging. *Scientific reports*, 11(1):9068, 2021. 4

[30] Gefei Wang, Jia Zhao, Yan Yan, Yang Wang, Angela Ruohao Wu, and Can Yang. Construction of a 3d whole organism spatial atlas by joint modelling of multiple slices with deep neural networks. *Nature Machine Intelligence*, 5(11):1200–1213, 2023. 3

[31] Mingyue Wang, Qinan Hu, Tianhang Lv, Yuhang Wang, Qing Lan, Rong Xiang, Zhencheng Tu, Yanrong Wei, Kai Han, Chang Shi, et al. High-resolution 3d spatiotemporal transcriptomic maps of developing drosophila embryos and larvae. *Developmental Cell*, 57(10):1271–1283, 2022. 3

[32] Xiao Wang, William E Allen, Matthew A Wright, Emily L Sylwestrak, Nikolay Samusik, Sam Vesuna, Kathryn Evans, Cindy Liu, Charu Ramakrishnan, Jia Liu, et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science*, 361(6400):eaat5691, 2018. 1

[33] Scott Wolf, Diogo Melo, Kristina M Garske, Luisa F Pallares, Amanda J Lea, and Julien F Ayroles. Characterizing the landscape of gene expression variance in humans. *PLoS genetics*, 19(7):e1010833, 2023. 2

[34] Sunny Z Wu, Ghamdan Al-Eryani, Daniel Lee Roden, Simon Junankar, Kate Harvey, Alma Andersson, Aatish Thennavan, Chenfei Wang, James R Torpy, Nenad Bartonicek, et al. A single-cell and spatially resolved atlas of human breast cancers. *Nature genetics*, 53(9):1334–1347, 2021. 3

[35] Ronald Xie, Kuan Pang, Sai Chung, Catia Perciani, Sonya MacParland, Bo Wang, and Gary Bader. Spatially resolved gene expression prediction from histology images via bimodal contrastive learning. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2, 3, 6

[36] 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 Applications of Computer Vision*, pages 5039–5048, 2023. 1, 2, 3, 6

[37] Ron Zeira, Max Land, Alexander Strzalkowski, and Benjamin J Raphael. Alignment and integration of spatial transcriptomics data. *Nature Methods*, 19(5):567–575, 2022. 3, 6

[38] 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

[39] Zi-Zhen Zhang, Gang Wang, Shan-Hui Yin, and Xiao-Hua Yu. Midkine: A multifaceted driver of atherosclerosis. *Clinica chimica acta*, 521:251–257, 2021. 6

[40] Chongyue Zhao, Zhongli Xu, Xinjun Wang, Shiyue Tao, William A MacDonald, Kun He, Amanda C Poholek, Kong Chen, Heng Huang, and Wei Chen. Innovative super-resolution in spatial transcriptomics: a transformer model exploiting histology images and spatial gene expression. *Briefings in Bioinformatics*, 25(2):bbae052, 2024. 2

[41] Junchao Zhu, Yiqing Shen, Haolin Zhang, and Jing Ke. An anti-biased tbsrtc-category aware nuclei segmentation framework with a multi-label thyroid cytology benchmark. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 580–590. Springer, 2023. 1