SEPAL: Spatial Gene Expression Prediction from Local Graphs

Anonymous ICCV submission

Paper ID 3

Abstract

Spatial transcriptomics is an emerging technology that aligns histopathology images with spatially resolved gene expression profiling. It holds the potential for understanding many diseases but faces significant bottlenecks such as specialized equipment and domain expertise. In this work, we present SEPAL, a new model for predicting genetic profiles from visual tissue appearance. Our method exploits the biological biases of the problem by directly supervising relative differences with respect to mean expression, and leverages local visual context at every coordinate to make predictions using a graph neural network. This approach closes the gap between complete locality and complete globality in current methods. In addition, we propose a novel benchmark that aims to better define the task by following current best practices in transcriptomics and restricting the prediction variables to only those with clear spatial patterns. Our extensive evaluation in two different human breast cancer datasets indicates that SEPAL outperforms previous state-of-the-art methods and other forms of including spatial context.

1. Introduction

Histopathology is the study of diseases in tissues through microscopic sample examination. Among the different staining methods, Hematoxylin and Eosin (H&E) is the most common one and is currently considered the gold standard for diagnosing a wide range of diseases [25, 23, 19]. More recently, this diagnostic approach has been complemented with molecular biomarkers, such as mRNA expression profiling, offering high specificity and the ability to directly predict prognosis and determine treatments [24, 4]. Interestingly, these two data types prove complementary: while H&E imaging lacks the specificity of transcriptomics, gene profiling lacks the physiological insights derived from morphology. By aligning dense spatial mRNA profiling with H&E histopathological images, Spatial Transcriptomics (ST) provides comprehensive insights into the spatial organization of gene expression within tissues [3].

The advent of direct gene expression assessment on tissue harbors the potential for an unprecedented understanding of the mechanistic causes behind many diseases. However, applying these datasets in real clinical practice encounters major bottlenecks, primarily stemming from the need for specialized equipment, domain expertise, and considerable time requirements [30]. To overcome these burdens and leverage the fact that H&E images are ubiquitous in medical settings, the computer vision community has recently delved into predicting gene expression from tissue images. Although various works demonstrate promising results [18, 16, 2, 9, 28, 29], existing methods are still far from clinical deployment.

Upon closer examination of the problem, it becomes evident that changes in gene expression are typically associated with alterations in tissue appearance. However, it is important to note that this correlation does not universally apply to all genes. For example, constitutive genes that exhibit constant expression within the spatial context [5] are unsuitable for prediction based solely on visual information. Hence, methods should focus on genes with a verifiable dependence on tissue appearance to achieve more accurate predictions. Another challenge lies in the scarcity of data. The current publicly available datasets encompass $2-70$ Whole Slide Images (WSI) with $5,000-15,000$ genes for a set of $300-3500$ coordinates, depending on the technology [3]. Consequently, generating such high-dimensional predictions with such limited samples is intrinsically difficult. Finally, as the technology is still in development, ground-truth data is sparse and may contain occasional missing values [30].

Current approaches present a dichotomy between complete globality, which uses the WSI to jointly predict an expression map for all the coordinates at once (WSI-based methods [18, 16, 2], Fig.1.A), and complete locality, which only uses visual information available at each coordinate to predict gene expression (patch-based methods [9, 28, 29], Fig.1.B). While complete globality leverages spatial information and long-range interactions, it suffers from severe data scarcity, making models prone to overfitting. In contrast, complete locality benefits from abundant data for deep

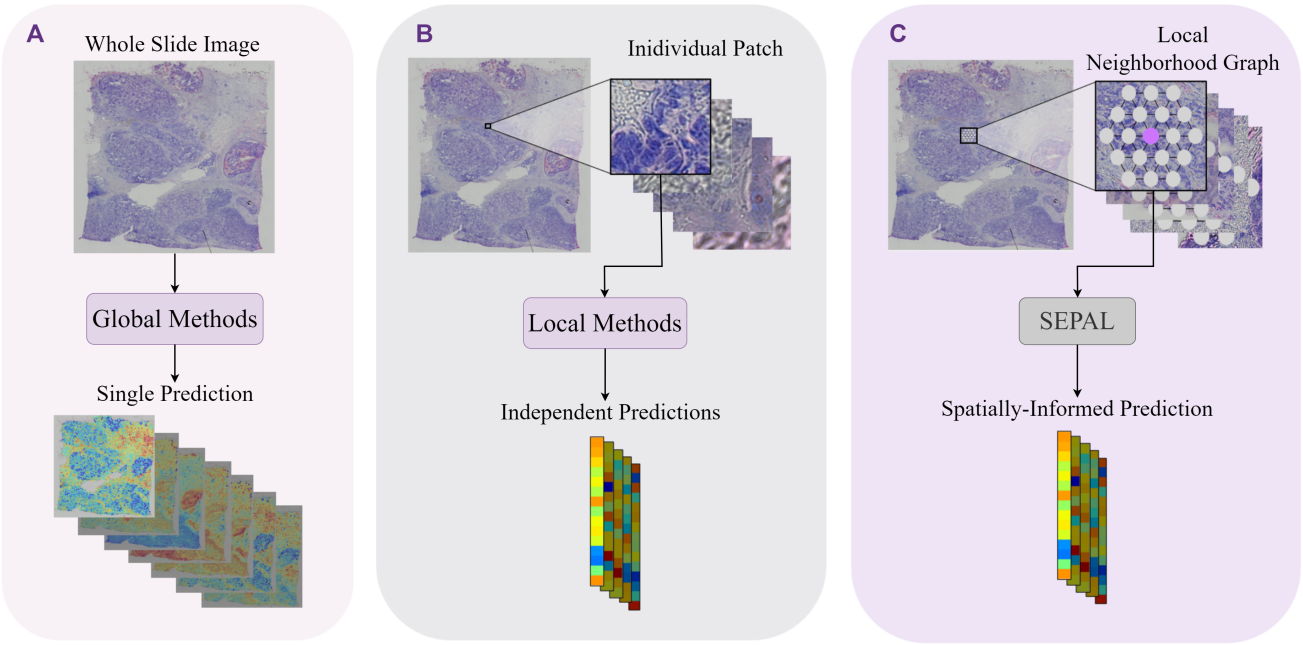

Figure 1. Different approaches for predicting gene expression from tissue images. (A) *Global methods* analyze a whole slide image and make a prediction about the tissue expression in every spot at once. (B) *Local methods* process the image by patches and predict the expression of each individual patch, one at a time. (C) SEPAL uses graphs that contain information from multiple patches to represent spatial information and predict gene expression for the central node of each graph.

learning training, yet it disregards spatial relations, resulting in suboptimal performance.

To overcome these challenges, we propose a new problem formulation, benchmark, and state-of-the-art method for **S**patial **E**xpression **P**rediction by **A**nalysing **L**ocal graphs (**SEPAL**). Our problem formulation strategically exploits the biological nature of the problem. Our benchmark uses a robust bioinformatic pipeline to overcome acquisition issues. Finally, our model bridges the gap between locality and globality by performing local spatial analysis.

In terms of problem formulation, we leverage a domain-specific advantage: the expression of a gene is expected to be within a specific range of values, and the variations inside that range are the ones with physiological significance. Rather than solely focusing on the absolute value of gene expression, we exploit this knowledge by bounding the prediction space within a defined box, using its center as an inductive bias. By estimating this bias from the training data, we can focus on learning relative differences instead of absolute values. Specifically, we supervise expression changes w.r.t. the mean expression of each gene in the training dataset. This novel approach differs from previous works since they directly predict the absolute gene expression.

We build our benchmark by first incorporating standard bioinformatic processing normalizations (TPM [1]), which were previously lacking. To ensure the selection of relevant prediction genes, we filter by Moran's I [15] value, a statistic designed to identify significant spatial patterns over a graph. By leveraging Moran's I, we ensure our focus remains on genes that depend on tissue appearance.

Finally, we introduce a novel approach that harnesses the power of local spatial analysis with the help of a graph neural network. By leveraging this strategy, we integrate information from local neighborhoods surrounding each patch (see Fig.1.C). Our key hypothesis is that gene expression is predominantly influenced by nearby visual characteristics rather than long-range interactions. SEPAL benefits from the advantages of local-based and global-based training (spatial relations and enough data) without succumbing to their respective limitations. We conduct extensive experimentation on two different human breast cancer datasets obtained with different technologies and report favorable results relative to existing techniques.

Our contributions can be summarized as follows:

- We propose a paradigm shift to supervise gene expression changes relative to the mean rather than absolute values.
- We propose a benchmark that follows current best practices in transcriptomics along with selected prediction genes with clear spatial patterns.
- We develop a new state-of-the-art method that applies local spatial analysis via graph neural networks.

To promote further research on ST, our project's benchmark and source code will be publicly released upon acceptance.

2. Related Work

Multiple approaches have been proposed to tackle the gene expression prediction task, with works focusing on different aspects of the visual data. State-of-the-art methods can be divided into two paradigms: local (patch-based) and global (WSI-based) focused.

2.1. Local Methods

Local methods estimate the gene expression one spot at a time by dividing the WSI into individual patches. Some examples of this approach include STNet [9], EGN [28], and EGGN [29]. The focal point of local methods is the visual information in the patch of interest, and they do not take into consideration characteristics such as the vicinity of the patch.

For instance, STNet [9], which is one of the most popular methods, formulates the task as a multivariate regression problem, and its architecture consists of a finetuned CNN (DenseNet-121 [10]) whose final layer is now a linear layer that predicts the expression of 250 genes. A characteristic strategy of STNet is that during inference, it predicts the gene expression for 8 different symmetries of that image (4 rotation angles and their respective reflections) and returns the mean result as the final estimation. This model generalizes well across datasets and has high performance when predicting the spatial variation in the expression of well-known cancer biomarkers [9].

Other examples of this approach include EGN [28] and its upgraded version, EGGN [29]. The core of these methods is exemplar guidance learning [28], a tool that they apply to base their predictions on the expressions of the patches that are most visually similar to the patch of interest. These reference patches are known as the exemplars and correspond to the nearest neighbors of a given patch in the latent space of an image encoder. The difference between these two models lies in the main processing of the input, where EGN uses the exemplars to guide a ViT [6], while EGGN uses the exemplars to build visual similarity graphs that are fed to a GraphSAGE-based backbone [8].

The key hypothesis of EGN and EGGN is that similar images have similar gene expression patterns, no matter their location within a tissue. Nevertheless, depending on the scale of the patches, this assumption could neglect their local context. For instance, if each patch contains a single cell, several similar patches with different physiological contexts might differ in their transcriptomic profile. However, our approach achieves robust and context-aware predictions, considering the physiological significance of relative expression changes within the gene expression range. This not only aligns with biological expectations but also enables us to glean valuable insights into the dynamic behavior of genes within tissues.

2.2. Global Methods

Opposed to local methods, global methods predict the gene expression of all the spots of a WSI at once, meaning that their input corresponds to the complete data from a high-dimension histopathology image. The most notable work of this family of methods is HisToGene [16], which receives a WSI and divides it into patches that are represented through image and positional embeddings fed to a Vision Transformer (ViT) architecture [6].

The mechanism in HisToGene enables the model to consider spatial associations between spots [16]. Nonetheless, this method demands a large number of WSIs in the dataset, posing a challenge as WSIs are often scarce in most datasets. Additionally, processing the entire WSI incurs a high computational cost. Therefore, we propose a more efficient spatial analysis at a smaller scale. Instead of using an entire sample as a single data element, we adopt a patch-based strategy, enabling us to execute predictions one patch at a time. This granular approach not only conserves computational resources but also mitigates the overfitting risks associated with using large WSIs.

3. SEPAL

3.1. Problem Formulation

Given an input image patch $X \in \mathbb{R}^{[H,W,3]}$, and k spatial neighbors $Z \in \mathbb{R}^{[k,H,W,3]}$, we want to train an estimator $F_\theta(\cdot)$ that predicts the difference between the gene expression y of patch X and the mean expressions in the training set $\bar{y}_{\text{train}}$. Consequently, we aim to optimize a set of parameters θ^* such that:

$$F_{\theta^*}(X, Z) \approx \Delta y = y - \bar{y}_{\text{train}} \tag{1}$$

Where, $\Delta y \in \mathbb{R}^{[n_g,1]}$ is the disparity between $\bar{y}_{\text{train}} \in \mathbb{R}^{[n_g,1]}$ and the real gene expression $y \in \mathbb{R}^{[n_g,1]}$ of the patch. This paradigm shift of predicting Δy instead of y, has the purpose of allowing our method to focus directly on the nuances in the data since we are standardizing the dynamic range of the prediction space around zero.

3.2. Architecture Overview

SEPAL is comprised of two stages: local embeddings and spatial learning, which are shown in Fig.2. The basis of SEPAL is the representation of the input patch and its neighbors as a graph, where the central node corresponds to the image for which we want to predict the gene expression. With this representation, our model has access to the visual features in the current location and to its surroundings.

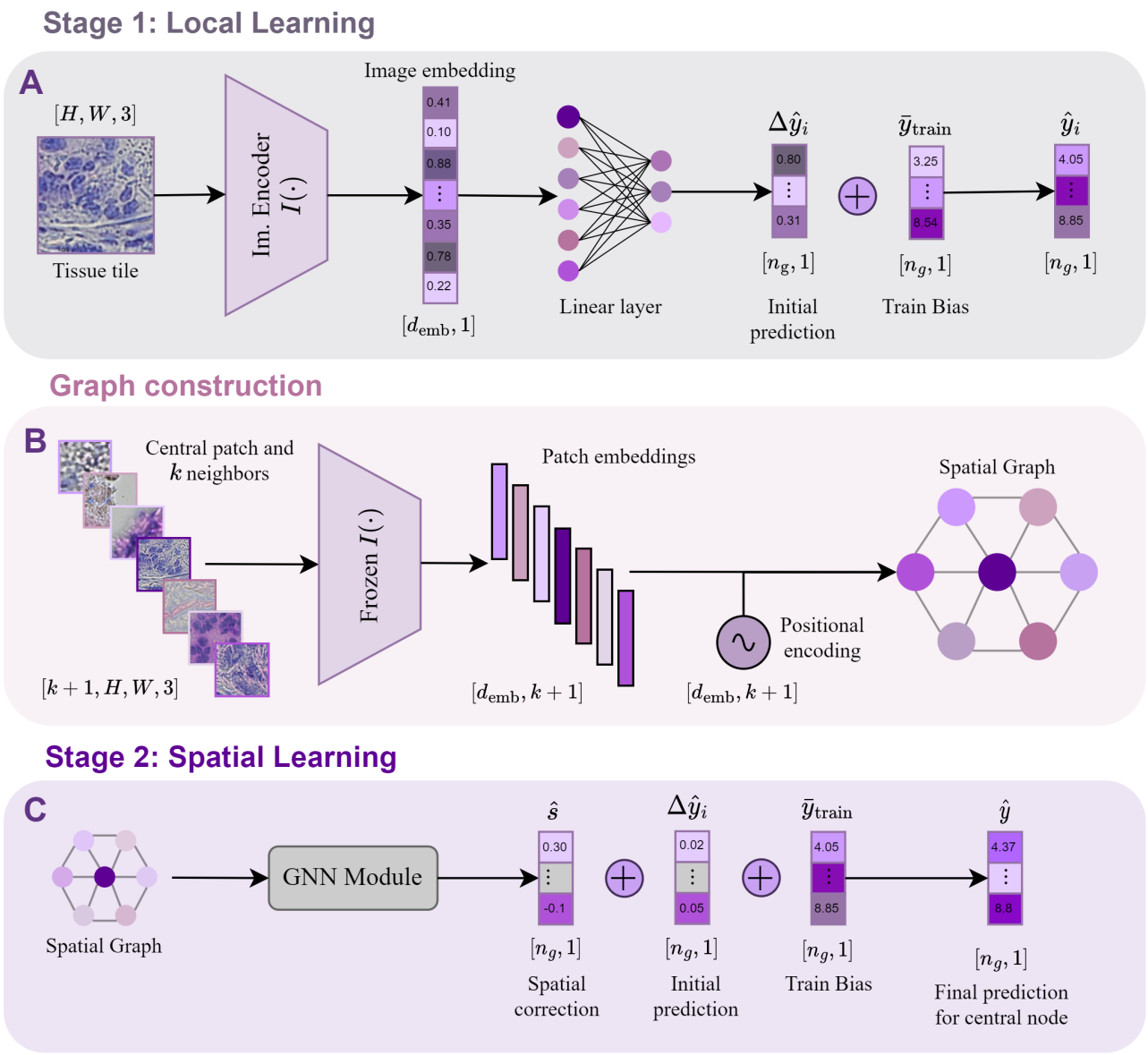

Figure 2. (A)First stage of our proposal. Pretraining of the Image Encoder $I(\cdot)$ and a linear layer $L(\cdot)$ to output the Image Embedding (I_{emb}) of a patch X, along with a preliminar prediction $\Delta\hat{y}_i$ of the difference between the expression in the patch and the mean expression in the train dataset. (B) The Graph Construction process begins with an image patch of interest and its spatial neighbors to build the graph representation based on the patch embeddings returned by the frozen $I(\cdot)$ and the positional encoding of each neighbor. (C) Architecture of SEPAL, which receives as input a Spatial Graph of the patch neighborhood and applies a GNN Module to predict the spatial correction $\hat{s}$ that further improves the $\Delta\hat{y}_i$ to get the $\Delta\hat{y}$ associated to the center patch of the graph and obtain the final gene expression prediction $\hat{y}$.

Prior to the construction of the graphs, in the first stage of our proposal (Fig.2.A), we train a feature extractor $I(\cdot)$ to process an input image patch X and return a low-dimensional representation $I_{\text{emb}} \in \mathbb{R}^{[d_{\text{emb}},1]}$. Besides, this module also outputs a local prediction $\Delta\hat{y}_i \in \mathbb{R}^{[n_g,1]}$ obtained by applying a linear layer $L(\cdot)$ to I_{emb} as follows:

$$I(X) = I_{\text{emb}} \tag{2}$$

$$L(I_{\text{emb}}) = \Delta\hat{y}_i \approx y - \bar{y}_{\text{train}} \tag{3}$$

Consequently, the preliminary prediction $\Delta\hat{y}_i$ is completely based on X and is later refined in the Spatial Learning stage. After training $I(\cdot)$, we fix it and use it to obtain the visual embeddings of all the patches in the dataset. We integrate these embeddings, together with a transformer-like positional encoding, to construct a local neighborhood graph $\mathcal{G}(X)$ for each patch (Fig.2.B).

Lastly, in the Spatial Learning stage (Fig.2.C), input graphs are processed by a GNN Module to obtain a spatial

correction vector $\hat{s} \in \mathbb{R}^{[n_g,1]}$ which is then added to $\Delta\hat{y}_i$ to obtain $\Delta\hat{y}$. This spatially aware prediction is summed with the bias $\bar{y}_{\text{train}}$ to present the final gene expression estimation $\hat{y}$ for the input patch:

$$\Delta\hat{y} = \hat{s} + \Delta\hat{y}_i \tag{4}$$

$$\hat{y} = \Delta\hat{y} + \bar{y}_{\text{train}} \tag{5}$$

3.3. Graph construction

The process of building the graphs is shown in Fig.2.B and aims to follow the spatial connectivity of the WSI while conserving the visual richness of local regions. Therefore, for a patch of interest X, we first select the k neighbors within an $m-$hop vicinity of X. For example, in Fig.2B $m = 1$ and $k = 6$ because of the hexagonal coordinate geometry. We join the patch and its neighbors in a single set $P = \{X, Z\} \in \mathbb{R}^{[k+1,H,W,3]}$ and compute the visual embedding matrix $M_i \in \mathbb{R}^{[d_{\text{emb}},k+1]}$ using our frozen image encoder $I(\cdot)$. Additionally, to enrich the spatial information beyond the topology of our graphs, we calculate a positional embedding $E_{\text{pos}} \in \mathbb{R}^{[d_{\text{emb}},1]}$ for each patch in P. Moreover, we use the 2D transformer-like positional encoder from [27] to process the relative coordinates of each neighbor w.r.t. the center patch. This computation gives us a positional matrix $M_p \in \mathbb{R}^{[d_{\text{emb}},k+1]}$ that is added with M_i to give the final graph features. Summarizing, we define graphs as:

$$G(X) = \mathcal{G}(P, E, M) \tag{6}$$

$$M = M_i + M_p \tag{7}$$

Where E is a binary and undirected set of edges defined by dataset geometry.

3.4. Spatial Learning Module

Once a graph $\mathcal{G}(X)$ is fed to the spatial learning module, it is passed through a series of h Graph Convolutional Operators ($\text{GNN}_i(\cdot)$) with a sequence $C = \{d_{\text{emb}}, c_1, c_2, \ldots, c_{h-1}, n_g\}$ of hidden channels following the recursive expression:

$$g_0 = \mathcal{G}(X) \tag{8}$$

$$g_{i+1} = \sigma\left(\text{GNN}_i(g_i)\right) \tag{9}$$

$$\hat{s} = \text{Pooling}(g_h) \tag{10}$$

Where g_i is the representation of $\mathcal{G}(X)$ at layer $i \in \{0, 1, 2, \ldots, h\}$, $\sigma(\cdot)$ is an activation function, and the $\text{Pooling}(\cdot)$ operator represents a global graph pooling operator. The correction vector $\hat{s}$ represents the contribution of local spatial information to the final prediction.

4. Experiments

4.1. Datasets

We evaluate our performance in two breast cancer datasets produced with different technologies: (1) the 10x Genomics breast cancer spatial transcriptomic [Section 1, Section 2] (referred to as *Visium* because of the experimental protocol), and (2) the human breast cancer in situ capturing transcriptomics dataset [22, 21] (referred to as *STNet dataset* because of the first deep learning method that used this data). The Visium dataset contains two slide images from a breast tissue sample with invasive ductal carcinoma from one patient, each with 3987 spots of $\approx 55\mu$m detected under the tissue. On the other hand, STNet dataset consists of 68 slide images of H&E-stained tissue from 23 patients with breast cancer and their corresponding spatial transcriptomics data. Specifically, the number of spots of size $\approx 150\mu$m varies between 256 and 712 in each replication, so the complete dataset contains $30,612$ gene expression data points spatially correlated with image patches. For both datasets, we take reshaped patches of dimension $[224, 224, 3]$ as input for SEPAL.

4.2. Benchmark

To design a robust benchmark, we focus on two main characteristics: (1) a bioinformatic pipeline in pair with current best practices in transcriptomic analysis, and (2) a selection strategy to ensure that all genes have spatial patterns.

In terms of the processing pipeline, we first filter out both genes and samples with total counts outside a defined range (See Supplementary Table 1 for detailed values in each dataset). Then, we discard genes based on the number of samples where their sparsity. Here, we ensure that the remaining variables are expressed in at least ε_T percent of the total dataset and ε_{WSI} percent of all WSIs. Following the filtering, we perform TPM [26] gene normalization and a $\log_2(x + 1)$ transformation. Finally, if batch effects are observed in UMAP [14] embeddings (Supplementary Figures 1-3) of the data (only seen in the STNet dataset), they are corrected with ComBat [11].

Once the bioinformatic pipeline is complete, we select the final prediction variables with the help of Moran's I. This statistic is a spatial autocorrelation measure and can detect if a given gene has a pattern over spatial graphs. The closer its value to one, the more autocorrelated the variable is. For our benchmark, we compute Moran's I for every gene and WSI and average across the slide dimension. We select the top $n_g = 256$ genes with the highest general Moran's I value as our final prediction variables (See supplementary Figures 4-7).

Summarizing, the processed Visium and STNet datasets have a total of 7,777 and 29,820 samples, respectively, along with a set of 256 prediction genes. As the Visium dataset only contains two WSIs, we use one for training (3795 samples) and the other one as the validation/test set (3982 samples). For the STNet dataset, from the 23 patients, we randomly choose 15 for training (20,734 samples), 4 for validation (3,397 samples), and 4 for testing

(5,689 samples).

4.3. Evaluation Metrics

We use three standard metrics in multivariate regression problems: global standard errors (MSE, MAE), Pearson Correlation Coefficients (PCC-Gene, PCC-Patch), and linear regression determination coefficients (R2-Gene, R2-Patch). Both PCC and R2 have the gene and patch variants of the metric since they address two aspects of the problem. The gene type metrics aim to quantify how good expression maps are in general, while the patch type metrics evaluate how good multiple gene predictions are for a specific patch. For instance, to compute PCC-Gene, we obtain PCC values for each one of the n_g gene maps and then average over the gene dimension. Conversely, computing PCC-Patch involves calculating PCC values for each patch and the average over that dimension. The errors are expected to take values close to zero, whereas PCC and R2 should be close to 1.0.

4.4. State-of-the-art Methods

We compare SEPAL to four of the most popular methods in this task, including three local options (STNet, EGN, EGGN), as well as one global method (HisToGene). For a fair comparison, we choose the best performance between 50 different training protocols. If the method allows batch size as a hyperparameter (STNet, EGN), we test combinations with an empirical Bayes approach by selecting learning rates in the logarithmic range $[10^{-2}, 10^{-6}]$ and batch sizes from the list $[32, 64, 128, 256, 320]$. If the method only accepts the learning rate, we perform a logarithmic grid search within the range $[10^{-2}, 10^{-6}]$. Both the best epoch during training and the best model of the sweep are selected based on the validation MSE. The only exception to this protocol (due to computational cost) is the STNet method in the STNet dataset, for which we report the best between the original hyperparameters and the best Visium hyperparameters.

4.5. Architecture Optimization

We extensively experiment with our spatial module, aiming to select the most effective architecture to integrate local information. For this purpose, we: (1) optionally introduce pre-processing and post-processing stages via multilayer perceptrons of varying sizes, (2) allow the positional encoding to be added or concatenated during the graph construction, (3) change the number of hops m from one to three, (4) try six different convolutional operators, and (5) vary the hidden dimensions h of our graph convolutional network going from one to four layers. Furthermore, we train all architecture variations with 12 different settings of learning rate and batch size. For a detailed explanation of every tunned hyperparameter, we refer the reader to the Supplementary Material (Sec. 2).

With this systematic procedure, we generate 3888 hyperparameter combinations from 324 module variations.

4.6. Implementation Details

We choose ViT [6] as our image encoder and select the best training protocol with a grid search of learning rate $[10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}]$ and batch size $[320, 256, 128, 64]$. We use ELU as our activation function and SAGPooling [13] as our pooling function. We replace occasional missing measurements in gene expression [30] with the median value in a neighborhood to take full advantage of the training data. We implement SEPAL using Pytorch [17] and Pytorch geometric [7] for graph operators. All experiments run on a single NVIDIA Quadro RTX 8000 GPU.

5. Results

Table 1 presents the final hyperparameter configurations of SEPAL for the Visium and the STNet datasets.

Hyperparameters	STNet Dataset	Visium
Number of hops	1	3
Embeddings aggregation	sum	sum
Graph operator	GraphConv[12]	TransformerConv[20]
Preprocessing Stage	-	-
Graph hidden channels	0,512,256,128,64	0,512,256
Postprocessing Stage	64,128,256	-
Learning rate	10^{-4}	10^{-5}
Batch size	512	64

Table 1. Hyperparameters with the best performance for both datasets.

5.1. Main Results

		Local			Global	
	Method	STNet[9]	EGN[28]	EGGN[29]	HisToGene[16]	SEPAL
Visium	MAE (↓)	0.660	0.659	0.660	0.678	**0.639**
	MSE (↓)	0.788	0.782	0.772	0.822	**0.743**
	PCC-Gene (↑)	0.343	0.326	0.316	0.186	**0.381**
	R2-Gene (↑)	0.054	0.062	0.061	0.013	**0.105**
	PCC-Patch (↑)	0.924	0.924	0.926	0.921	**0.928**
	R2-Patch (↑)	0.845	0.846	0.849	0.840	**0.854**
STNet dataset	MAE (↓)	0.508	0.508	0.530	**0.505**	0.507
	MSE (↓)	0.617	0.619	0.667	0.619	**0.604**
	PCC-Gene (↑)	0.080	0.076	0.028	0.055	**0.156**
	R2-Gene (↑)	0.005	0.004	-0.103	0.002	**0.024**
	PCC-Patch (↑)	**0.894**	0.893	0.890	**0.894**	**0.894**
	R2-Patch (↑)	0.794	0.793	0.776	0.794	**0.798**

Table 2. Quantitative comparison with state-of-the-art methods on Visium and STNet datasets. The best performance is written in **bold**, and the second best result is underlined for each metric.

Table 2 depicts the performance of local and global state-of-the-art methods against SEPAL on the Visium and

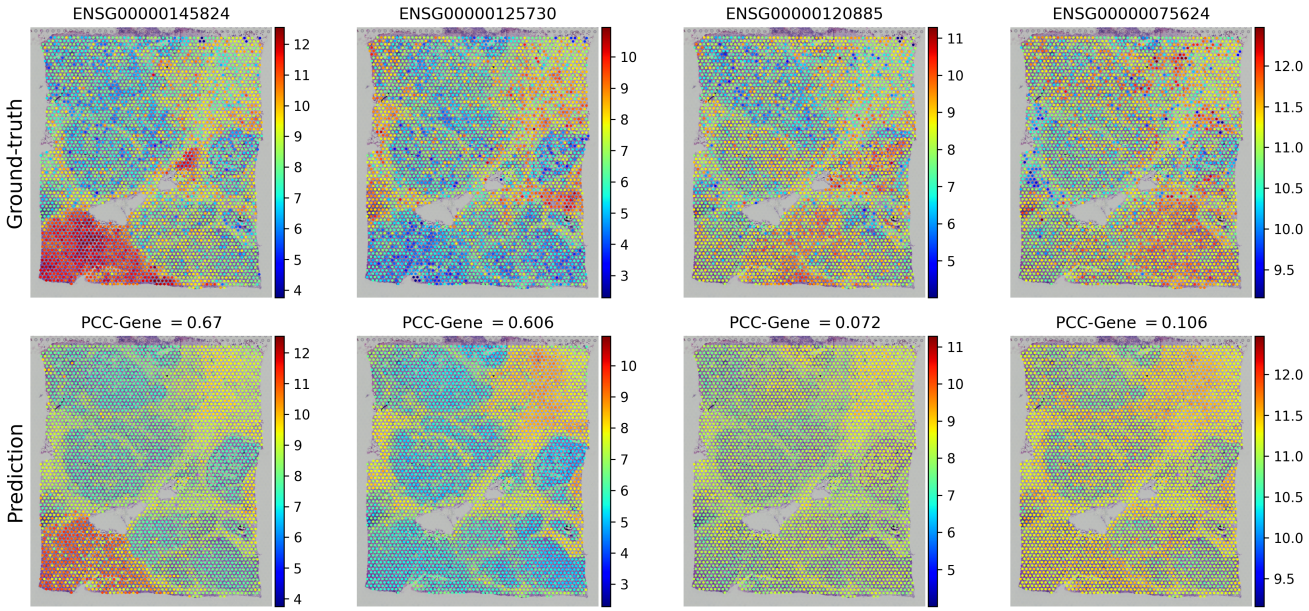

Figure 3. Visualization of the two genes with the highest (left) and lowest (right) Pearson Correlation Coefficient. At the top is the Ground-Truth of the expression and at the bottom is the qualitative prediction of our method with its respective PCC.

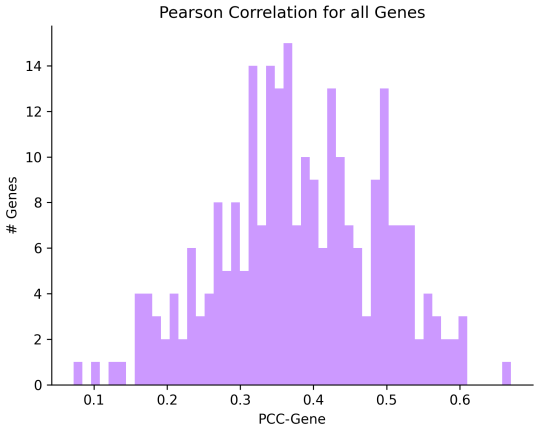

Figure 4. Histogram of the Pearson correlation between the ground-truth and the prediction of each gene. The X-axis displays the values of the Pearson correlation coefficient, while the Y-axis shows the number of genes that have that particular correlation.

STNet datasets. Our method consistently outperforms these methods on all but one evaluation metric. In particular, we attend primarily to the gene-wise metrics and find that SEPAL presents significant improvements in the gene correlation on the two datasets. Especially for the STNet dataset, where the gene-wise performance across all the models is low, SEPAL achieves a correlation 1.95 times higher than the second best model. Likewise, the R2 metric calculated on the genes increases when using SEPAL.

Furthermore, we find that calculating the PCC and the R2 metrics in a gene-wise fashion results in a remarkably poorer performance compared to the patch-wise evaluation. This means that predicting the distribution of the expression of a single gene in a WSI is a significantly more difficult task than aiming to obtain the expression of all the genes in one single spot. Nevertheless, despite this different trend for gene or patch-focused evaluations, our method consistently achieves the best results.

HisToGene has poorer performance on Visium than on STNet, and overall it shows the worst results on the Visium dataset. These differences within the results of HisToGene support the observation that data scarcity of small datasets like Visium leads to deficient results in global methods. Conversely, we demonstrate that our method is able to retrieve important information from the input despite the difference in the data acquisition technologies and number of samples since it achieves high performance on both datasets.

Finally, Fig.4 shows a histogram of the PCC between the ground-truth and the predictions of each gene on Visium. None of the genes has a negative correlation, and the lowest PCC is 0.1 and goes as far as 0.67. Overall, our model has a satisfactory performance for the evaluation of the genes selected. We observe that the PCC has an approximately normal distribution, with no evident outliers.

5.2. Control Experiments

Table 3 shows the results for the ablation experiments. Comparing the results between predicting the absolute expression (ViT) and predicting the expression variations of

Method	ViT	ViT+Δ	ViT+Δ+S7	SEPAL
MAE (↓)	0.661	0.644	0.653	**0.639**
MSE (↓)	0.788	0.756	0.769	**0.743**
PCC-Gene (↑)	0.302	0.372	**0.396**	0.381
R2-Gene (↑)	0.055	0.090	0.075	**0.105**
PCC-Patch (↑)	0.924	0.927	0.925	**0.928**
R2-Patch (↑)	0.846	0.851	0.848	**0.854**

Table 3. Control experiments on the Visium validation/test set. Δ: predicting differences with respect to the mean expression $\hat{y}_{\text{train}}$. S7: input patch is 7 times bigger than the original one.

the genes (ViT+Δ), we notice that the latter option has a better performance in every metric. For instance, when predicting delta variations, the MSE is 0.032 points below that of the absolute expression prediction. The PCC-Gene also increased 0.07 points with our problem formulation. These results reflect the suitability of the paradigm shift that we propose by learning the difference between y and $\bar{y}_{\text{train}}$ instead of directly predicting y.

We evaluate the benefit of using a larger neighborhood to determine how raw spatial information affects gene prediction. Table 3 compares the behavior of the exact same image encoder while solely altering the scale of the patches. Except for PCC-Gene, keeping a scale of 1.0 remains the best option among the ViT architectures tested. Our findings suggest that increasing the visual coverage of an image encoder does not yield significant improvements in gene prediction, given the marginal differences observed in the metrics.

In addition, the results from SEPAL show an improvement in overall metrics but PCC-Gene, with respect to ViT+Δ+S7. Notably, both SEPAL and ViT+Δ+S7 have access to the same visual context in the WSI and are differentiated only by how spatial information is represented. This compelling outcome underscores the importance of incorporating spatial features in the description of each patch and constructing graphs to glean highly relevant information for accurate expression prediction. The performance of SEPAL shows that the predictions from the spatial module do further improve the preliminary predictions obtained during the local learning stage. Our results validate the efficacy of our novel approach, emphasizing the value of spatial interactions in gene expression prediction.

As a sidenote observation, when predicting on the STNet dataset, ViT+Δ converges in earlier stages of the training process than ViT, and the training of the model is notably more stable than when predicting absolute gene expression.

5.3. Qualitative Results

Figure.3 shows the heatmaps for the real and the predicted expression distribution of the genes with the best and worst performances. Focusing on the genes with the highest PCC, we see that for the second best gene, the expressions both on the ground-truth and on the prediction are highly associated with the tissue color. Note that the regions with darker tissue obtain a similar gene expression prediction, and the same happens for the regions with lighter tissue. These results suggest that our model might be basing the predictions solely on the color of the patches rather than looking for specific morphology patterns. Nevertheless, for the best gene, the predicted expressions are not uniformly the same for all dark or light tissue sections, conveying that our model does not rely only on the appearance of the images and is actually learning from the spatial context of the patches and tissue morphology.

The predicted expressions show a lower intensity than the ground-truth for both genes, indicating that the dynamic range of SEPAL predictions may not match that of the real expression levels. Notably, for the two genes with the highest PCC, the output of our method appears over-smoothed compared to the ground-truth. An evident distinction arises when comparing the real expression, which exhibits adjacent spots with drastically different expression levels, to the predictions, where no regions display sudden changes in expression tendencies. While our model's consistent predictions showcase its strength, this attribute may also be considered a drawback when seeking to detect gene expression deviations with high spatial resolution.

Regarding the worst genes, Fig.3 shows in their ground-truth that these are cases where the expression does not have a particularly clear spatial pattern, even if these genes surpassed our Moran's I Statistic threshold. Consequently, our model's poor performance on these genes can be attributed to the absence of a major spatial pattern to predict. Additionally, we see that for bad-performing genes, the predictions tend to correspond to the mean expression value of each gene and are practically constant throughout the entire WSI.

6. Conclusions

In this work, we develop a novel framework to approach the spatial gene expression prediction task by integrating local context and exploiting inductive biases inherent to the biological nature of the problem. Our proposed SEPAL consistently outperforms state-of-the-art models and closes the gap between completely global and completely local analysis. Furthermore, aligning with biological expectations, it is capable of recognizing patterns in histological data that go beyond simple color intensities. Consequently, our approach represents a significant step forward in spatial expression prediction, enhancing the applicability of deep learning methods in the context of disease analysis and precision medicine.

References

[1] Zachary B Abrams, Travis S Johnson, Kun Huang, Philip RO Payne, and Kevin Coombes. A protocol to evaluate rna sequencing normalization methods. *BMC bioinformatics*, 20(24):1–7, 2019.

[2] Areej Alsaafin, Amir Safarpoor, Milad Sikaroudi, Jason D. Hipp, and H. R. Tizhoosh. Learning to predict rna sequence expressions from whole slide images with applications for search and classification. *Communications Biology 2023 6:1*, 6:1–9, 3 2023.

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

[4] Anton Buzdin, Maxim Sorokin, Andrew Garazha, Alexander Glusker, Alex Aleshin, Elena Poddubskaya, Marina Sekacheva, Ella Kim, Nurshat Gaifullin, Alf Giese, Alexander Seryakov, Pavel Rumiantsev, Sergey Moshkovskii, and Alexey Moiseev. Rna sequencing for research and diagnostics in clinical oncology. *Seminars in Cancer Biology*, 60:311–323, 2 2020.

[5] Joanne R Chapman and Jonas Waldenström. With reference to reference genes: A systematic review of endogenous controls in gene expression studies. *PloS one*, 10:e0141853, 11 2015.

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

[7] Matthias Fey and Jan Eric Lenssen. Fast graph representation learning with pytorch geometric. *arXiv preprint arXiv:1903.02428*, 2019.

[8] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

[9] 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*, 4:827–834, 6 2020.

[10] 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.

[11] W Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127, 2007.

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

[13] Junhyun Lee, Inyeop Lee, and Jaewoo Kang. Self-attention graph pooling. In *International conference on machine learning*, pages 3734–3743. PMLR, 2019.

[14] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[15] Patrick AP Moran. Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23, 1950.

[16] Minxing Pang, Kenong Su, and Mingyao Li. Leveraging information in spatial transcriptomics to predict super-resolution gene expression from histology images in tumors. *bioRxiv*, page 2021.11.28.470212, 11 2021.

[17] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[18] Benoît Schmauch, Alberto Romagnoni, Elodie Pronier, Charlie Saillard, Pascale Maillé, Julien Calderaro, Aurélie Kamoun, Meriem Sefta, Sylvain Toldo, Mikhail Zaslavskiy, Thomas Clozel, Matahi Moarii, Pierre Courtiol, and Gilles Wainrib. A deep learning model to predict rna-seq expression of tumours from whole slide images. *Nature Communications 2020 11:1*, 11:1–15, 8 2020.

[19] Neil J. Sebire. Oncology: histopathology and imaging in the future. *Pediatric Radiology*, 41:170–171, 5 2011.

[20] Yunsheng Shi, Zhengjie Huang, Shikun Feng, Hui Zhong, Wenjin Wang, and Yu Sun. Masked label prediction: Unified message passing model for semi-supervised classification. *arXiv preprint arXiv:2009.03509*, 2020.

[21] Linnea Stenbeck, Ludvig Bergenstråhle, Joakim Lundeberg, and Åke Borg. Human breast cancer in situ capturing transcriptomics. 5, 2021.

[22] Patrik L. Ståhl, Fredrik Salmén, Sanja Vickovic, Anna Lundmark, José Fernández Navarro, Jens Magnusson, Stefania Giacomello, Michaela Asp, Jakub O. Westholm, Mikael Huss, Annelie Mollbrink, Sten Linnarsson, Simone Codeluppi, Åke Borg, Fredrik Pontén, Paul Igor Costea, Pelin Sahlén, Jan Mulder, Olaf Bergmann, Joakim Lundeberg, and Jonas Frisén. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353:78–82, 7 2016.

[23] Yoshihisa Takahashi. Histopathology of nonalcoholic fatty liver disease/nonalcoholic steatohepatitis. *World Journal of Gastroenterology*, 20:15539, 11 2014.

[24] Laura J van 't Veer, Hongyue Dai, Marc J van de Vijver, Yudong D He, Augustinus A M Hart, Mao Mao, Hans L Peterse, Karin van der Kooy, Matthew J Marton, Anke T Witteveen, George J Schreiber, Ron M Kerkhoven, Chris Roberts, Peter S Linsley, René Bernards, and Stephen H Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530–6, 1 2002.

[25] Vincenzo Villanacci, Alessandro Vanoli, Giuseppe Leoncini, Giovanni Arpa, Tiziana Salviato, Luca Reggiani Bonetti, Carla Baronchelli, Luca Saragoni, and Paola Parente. Celiac disease: histology-differential diagnosis-complications. a practical approach. *Pathologica*, 112:186–196, 9 2020.

[26] Günter P Wagner, Koryu Kin, and Vincent J Lynch. Measurement of mrna abundance using rna-seq data: Rpkm measure is inconsistent among samples. *Theory in biosciences*, 131:281–285, 2012.

[27] Zelun Wang and Jyh-Charn Liu. Translating math formula images to latex sequences using deep neural networks with sequence-level training, 2019.

[28] Yan Yang, Md Zakir Hossain, Eric A Stone, and Shafin Rahman. Exemplar guided deep neural network for spatial transcriptomics analysis of gene expression prediction. *arXiv*, 10 2022.

[29] Yan Yang, Zakir Hossain, Eric Stone, and Shafin Rahman Projector. Spatial transcriptomics analysis of gene expression prediction using exemplar guided graph neural network. *bioRxiv*, page 2023.03.30.534914, 3 2023.

[30] Qichao Yu, Miaomiao Jiang, and Liang Wu. Spatial transcriptomics technology in cancer research. *Frontiers in Oncology*, 12:1019111, 10 2022.