



This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

# Enhancing Gene Expression Prediction from Histology Images with Spatial Transcriptomics Completion

Gabriel Mejia\*<sup>[0000-0003-4382-6390]</sup>, Daniela Ruiz\*<sup>[0000-0001-6636-1173]</sup>, Paula Cárdenas<sup>[0009-0005-1185-548X]</sup>, Leonardo Manrique<sup>[0009-0008-9428-6009]</sup>, Daniela Vega<sup>[0009-0002-9731-7591]</sup>, and Pablo Arbeláez<sup>[0000-0001-5244-2407]</sup>

Center for Research and Formation in Artificial Intelligence  
Universidad de los Andes, Colombia

{gm.mejia,da.ruiz11,p.cardenasg,dl.manrique,d.vegaa}@uniandes.edu.co

**Abstract.** Spatial Transcriptomics is a novel technology that aligns histology images with spatially resolved gene expression profiles. Although groundbreaking, it struggles with gene capture yielding high corruption in acquired data. Given potential applications, recent efforts have focused on predicting transcriptomic profiles solely from histology images. However, differences in databases, preprocessing techniques, and training hyperparameters hinder a fair comparison between methods. To address these challenges, we present a systematically curated and processed database collected from 26 public sources, representing an 8.6-fold increase compared to previous works. Additionally, we propose a state-of-the-art transformer-based completion technique for inferring missing gene expression, which significantly boosts the performance of transcriptomic profile predictions across all datasets. Altogether, our contributions constitute the most comprehensive benchmark of gene expression prediction from histology images to date and a stepping stone for future research on spatial transcriptomics.

**Keywords:** Spatial transcriptomics · completion · transformers · histology

## 1 Introduction

Spatial Transcriptomics (ST) is an emerging technology that precisely localizes gene expression profiles within histological images [10]. While histology analysis is the gold standard for diagnosis of many diseases [27], transcriptomics unlocks molecular insights that unveil causal pathways behind pathologies [30,10]. Together, they open a new spectrum of possibilities to understand how physiological alterations start, evolve and respond to treatment directly in spatial coordinates [25].

Although potentially groundbreaking, the deployment of ST presents significant practical challenges. On the technical side, it inherits the problem of not

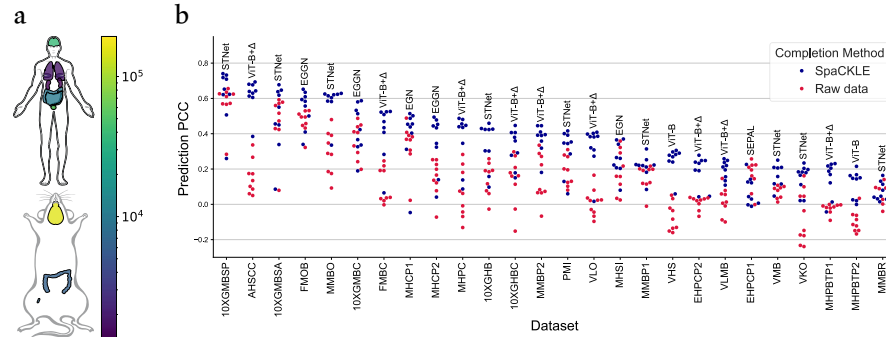


Fig. 1: (a) Organisms and tissues available in SpaRED, along with the number of spots available from each tissue. (b) Prediction Pearson Correlation Coefficient for each model across all the datasets in SpaRED. For each dataset, the state-of-the-art model that obtains the highest Pearson Correlation Coefficient is included.

detecting transcripts even though they were present in the source tissue, a failure mode known as dropout in bulk and single-cell transcriptomics [18,2]. In practice, this phenomenon is observed as pepper noise in the acquired gene expression maps and is so severe that single-cell reference datasets are commonly required to handle missing data [2].

Among the social challenges, expensive equipment, the need for domain expertise, and a slow adoption in clinical settings impede diagnostic benefits from reaching most patients [16]. Acknowledging these problems, the deep learning community has delved into democratizing ST by studying gene expression prediction from histology images [10]. Solving this problem would allow patients to obtain all the molecular insights from the image of an ordinary biopsy.

As with any novel technology, multiple variations of ST are currently in use [19,4,21]. However, as demonstrated by the number of entries in the comprehensive repository of spatial transcriptomics [25], 10X Visium [21] is by far the most popular. Leveraging the abundance of public Visium data, several deep learning methods have been developed for this task [8,16,29,28,27,30,13], varying greatly in architecture and always reporting favorable results against state-of-the-art. Nonetheless, differences in databases, data preprocessing, and training hyperparameters impede fair comparison between approaches and compromise the validity of new results.

To address these challenges, we present two contributions. First, we systematically compile, curate and standardize 26 public ST databases into the **S**patially **R**esolved **E**xpression **D**atabase (SpaRED), an extensive Visium resource with human and mouse samples from 9 tissue types as shown in Fig. 1.a. Secondly, we evaluate 7 state-of-the-art prediction methods in SpaRED, establishing a new

benchmark with an 8.6-fold increase in analyzed datasets compared to previous works [10].

As main technical contribution, inspired by the unrivaled power of self-attention mechanisms for next token prediction in natural language processing [5], we design a transformer-based completion model for corrupted gene expression vectors, which we call **S**patial transcriptomics **C**ompletion with **K**nowledge from the **L**ocal **E**nvironment (SpaCKLE). SpaCKLE not only outperforms previous gene completion strategies but significantly enhances the gene prediction performance of all state-of-the-art methods across every dataset of our benchmark, as evidenced by the improvement from red to blue points in Fig.1b. Our project’s benchmark and source code is publicly available at <https://github.com/BCV-Uniandes/SpaRED/tree/main>.

## 2 Related Work

### 2.1 Integrated Databases

Recent advancements in ST have led to the development of multiple databases. For instance, CROST [25] is a comprehensive repository with 1033 spatial transcriptomics samples from 8 species, 35 tissues, and 56 diseases. Other databases include SpatialDB, Aquila, SPASCEr, SODB, and STomicsDB [25], each offering unique datasets and analytical tools. Although these databases facilitate advanced spatial analyses, they are not designed for the expression profile prediction task. SpaRED tackles this limitation by implementing best practices in bioinformatics analysis, including the selection of Moran genes, standardization of reference genomes, TPM normalization, and batch correction, making SpaRED particularly valuable for clinical applications.

### 2.2 Completion strategies

To address the missing value problem, [13] employs a modified adaptive median filter as a completion strategy, replacing dropout values with circular region medians or, if unsuccessful, the whole slide image (WSI)’s median. Alternatively, stLearn [18] uses genetic and morphological similarity to adjust existing spots or predict gene expression for missing values. Moreover, although there are alternative methods that integrate knowledge from single cell RNAseq (Seurat [20], Harmony [12], LIGER [26] and Tangram [3]), these strategies require a paired single-cell dataset that hampers their usability and practicality. In contrast, SpaCKLE is a reference-free completion method, which stands out from alternatives by leveraging the complete genetic profile of adjacent spots and taking advantage of the transformer capacity to predict missing values.

### 2.3 Gene Expression Prediction Benchmarks

A recent study by [10] reviews 6 deep learning methods for gene expression profile prediction, testing their performance on three distinct breast cancer datasets.

Although the study presents a solid model performance analysis, it only focuses on human breast cancer tissue. Hence, our benchmark represents a substantial advancement, featuring an 8.6-fold increase in the number of datasets and including 9 different tissue types from human and mouse subjects.

### 3 Spatially Resolved Expression Database

#### 3.1 Original Datasets and Curation

To build SpaRED, we collect raw data from 7 independent publications [1], [17], [24], [23], [14], [6], [7] and complement them using 5 demonstration datasets from 10X Genomics (available through the SquidPy python package [15]). We only include datasets with more than one WSI and split the publications’ data by tissue type, resulting in 26 distinct datasets: 14 from human and 12 from mouse, showcasing a variety of tissue samples, as illustrated in Fig. 1.a. According to the number of patients in a dataset, we define two types of tasks: intra-patient generalization (where WSIs are generally consecutive cuts of a single tissue) and inter-patient generalization (where WSIs correspond to the same tissue in different subjects). We manually assign WSIs into train, validation, or test sets seeking similar visual distributions in all splits. In 11 out of 26 cases, a test set is defined. Otherwise, due to the limited number of patients/slides, we split the data into train and validation sets.

For data preprocessing, we follow the protocol proposed in [13], which handles batch correction, normalization, and gene selection based on Moran’s I calculations. Depending on data quality, we select either 32 (6/26 datasets) or 128 (20/26 datasets) prediction genes in each refined dataset. In total, SpaRED contains 105 slides (308,843 spots) collected from 35 patients. See supplementary Table 1 for detailed datasets statistics.

#### 3.2 Benchmark of Existing Gene Prediction Methods

We use SpaRED to evaluate 7 state-of-the-art expression profile prediction methods. Among these, HisToGene [16] splits a WSI into patches that are processed by a Visual Transformer (ViT) model. The output is the genetic profile of the WSI. STNet [8] inputs individual patches into a fine-tuned DenseNet-121 with a linear layer for prediction. Additionally, STNet averages predictions across 8 symmetries of each patch to determine the final output. EGN [29] and its improved version EGGN [28] apply exemplar-guided learning, a prediction strategy that bases its estimations on patches that are visually similar to the target patch within a latent space. BLEEP [27], employs contrastive learning to map image patches and expression profiles in a shared latent space. Hist2ST [30] applies a Convolutional Neural Network (CNN) to extract local patch features, followed by a transformer and a Graph Neural Network (GNN) to handle long and near-range dependencies, respectively. Finally, SEPAL[13] finetunes a ViT backbone, and subsequently refines its predictions applying a GNN that processes a neighborhood graph for each patch. Additionally, SEPAL supervises

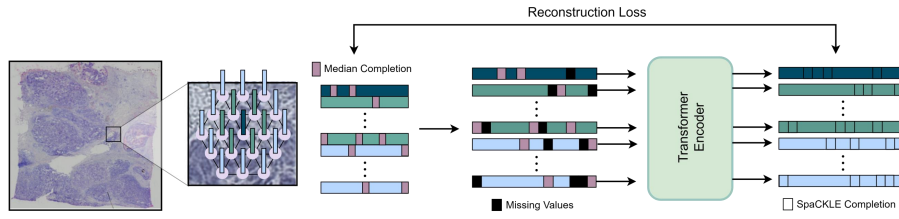


Fig. 2: Overview of our data completion framework using a transformer-based model.

expression changes relative to the mean expression in the training data instead of the absolute expression value, a strategy denoted ( $\Delta$ ) prediction.

Alongside these models, our comprehensive benchmark also includes the performance of three baseline methods: a ShuffleNet [31] architecture that finetunes an image encoder with low computational cost, a ViT-B encoder [5] that reflects the impact of fine-tuning a state-of-the-art backbone for this task, and a ViT-B+ $\Delta$  approach as suggested by [13].

We search for the optimal learning rate in every dataset, and then, with this value fixed, we explore two training scenarios: using raw data directly and SpaCKLE-completed data.

## 4 Gene Completion with Transformers

Inspired by the disruptive success of the transformer architecture for completion tasks such as language next token prediction [22] and visual reconstruction [9], we adapt these ideas to the ST domain. Fig. 2 illustrates SpaCKLE’s training, a process that takes as a starting point data that we pre-completed using the median method proposed in [13]. This process ensures faster training convergence and guarantees non-zero predictions. Given the median-completed expression vector  $x \in \mathbb{R}^{[g,1]}$  of a particular spot with  $g$  prediction genes, we start by extracting the expression matrix  $E_x = [x; V_x] \in \mathbb{R}^{[g,n+1]}$ , which concatenates  $x$  with the expression matrix  $V_x \in \mathbb{R}^{[g,n]}$  of its  $n$  2-hop neighbors in the Visium hexagonal geometry. We then randomly mask  $E_x$  and process it with a transformer encoder  $T(\cdot)$  that leverages the self-attention mechanism:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (1)$$

to get a reconstructed version  $\hat{E}_x$  as follows:

$$E_m = E_x \odot M(\rho) \quad (2)$$

$$\hat{E}_x = L_{out}(T(L_{in}(E_m))), \quad (3)$$

where  $\odot$  represents the Hadamard product between  $E_x$  and a binary masking matrix  $M \in \mathbb{R}^{[g,n+1]}$  that does not overlap with median-completed values and

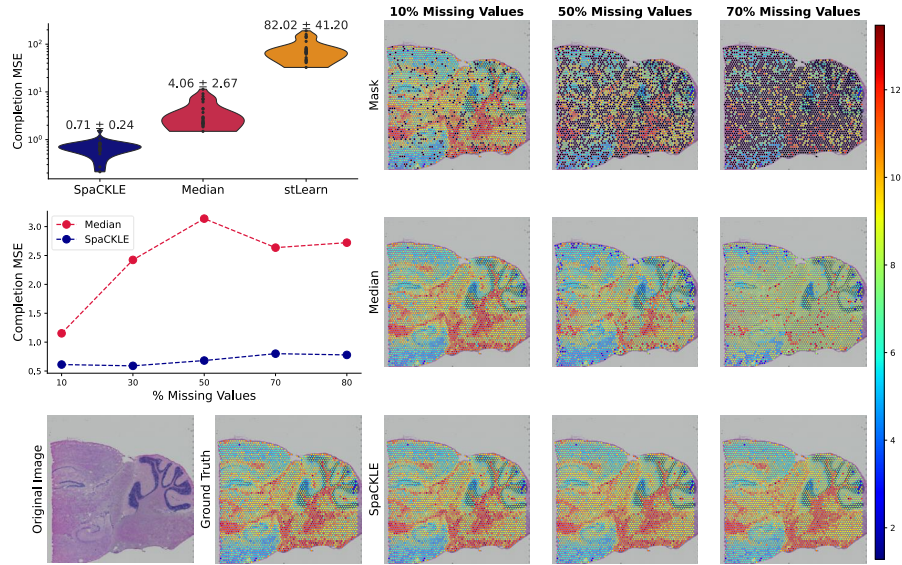


Fig. 3: Completion results: Violin plot displaying completion MSE scores for each method (SpaCKLE, Median and stLearn) across all datasets in SpaRED (upper left). Line plot displaying completion MSE for the median and SpaCKLE methods across different percentages of synthetically masked data (middle left). Qualitative results showing gene completion for increasing synthetic masking percentages (row 1) with the median method (row 2) and SpaCKLE (row 3).

exhibits a zero value probability  $\rho = 30\%$ . To accommodate different gene dimensionalities to a fixed transformer dimension  $d_k = 128$ , we use the  $L_{in}(\cdot)$  and  $L_{out}(\cdot)$  linear adapters. We optimize a Mean Square Error (MSE) loss between the two complete matrices  $\mathcal{L} = \left\| E_x - \hat{E}_x \right\|_2$  but we only compute metrics and complete missing values using the masked elements from the first vector of the output. Hence, each component of the completed version  $\hat{x}$  can be expressed as:

$$\hat{x}_i = \begin{cases} x_i, & M[1, i] = 1 \\ \hat{E}_x[1, i], & M[1, i] = 0 \end{cases} \quad (4)$$

At inference, we replace  $M$  with a binary matrix indicating original missing values to get a refined version of the gene expression profiles.

**Implementation Details:** We train all our models on a NVIDIA Quadro RTX 8000 with a batch size of 256 and use an Adam [11] optimizer with default PyTorch library parameters. We handle both regression and completion problems as multivariate regression tasks and evaluate them using MSE and Pearson Correlation Coefficient (PCC). To select the best model, we save the one with the

lowest validation MSE after 1,000 and 10,000 iterations for prediction and completion, respectively. All metrics are computed exclusively on real data for both the completion and the prediction task.

## 5 Results and Discussion

### 5.1 Gene Completion Evaluation

The violin plot in Fig. 3 presents a comparison of the logarithmic MSE for data completion using SpaCKLE, the median completion method, and stLearn across SpaRED. The results indicate that SpaCKLE outperforms alternative completion methods, with a relative 82.5% MSE reduction compared to the median method and by two orders of magnitude concerning stLearn. Notably, stLearn presents the highest MSE in the entirety of SpaRED, which conveys its inability to restore masked data. These results are consistent with those reported in [2], where stLearn’s completion predictions included a high proportion of zero values. It is noteworthy that the median method is based solely on the adjacent expression of a single gene, an approach that, although straightforward, does not consider the broader genetic context. In contrast, SpaCKLE has access to the complete genetic profile of the neighboring spots. Thus, we hypothesize that our transformer architecture is leveraging the full expression profile of the empty spot’s vicinity to enhance completion predictions.

To thoroughly assess the robustness of our approach, we characterize the completion performance when synthetically corrupting increasing percentages of data in the 10XGMBSP dataset. The MSE results in Fig. 3 show how the completion’s accuracy changes with various masking percentages. For visualization purposes, we only show SpaCKLE and the median method since stLearn has a significantly higher MSE. We observe that, as the task gets more challenging with a greater percentage of missing data, SpaCKLE outperforms the median completion method by a larger margin. The predicted expression maps support these observations, showing that SpaCKLE strongly approximates the ground truth patterns even at a missing value percentage of 70%. Conversely, the uniformity in the color pattern of the predictions from the median method demonstrates that this strategy repeatedly imposes the global median when it cannot find a local value due to the high fraction of missing data. This behavior impairs the expression profiles by homogenizing the gene’s activity in the tissue and removing spatial information.

### 5.2 Gene Prediction Benchmark

Figure 1.b shows the performance of all methods for every dataset when trained under our two scenarios (SpaCKLE-completed and raw). It is clear that the prediction performance significantly improves when applying SpaCKLE to every dataset and, in some cases, the best PCC increases to 0.36 points (AHSCC). This result pinpoints the importance of acknowledging missing data for the prediction task and proves the significance of including gene completion in ST pipelines.

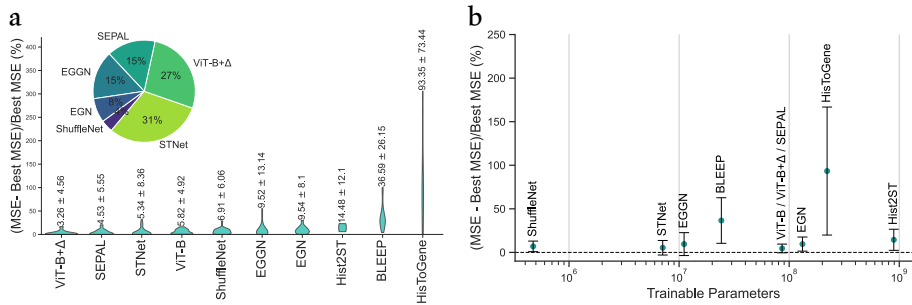


Fig. 4: (a) Violin plot: normalized prediction MSE of each model across all datasets within SpaRED, with normalization done against the best MSE obtained on each dataset. The mean and standard deviation of the methods are included at the top of each violin. Pie chart: percentage of datasets within SpaRED for which each model achieves the best prediction MSE. (b) Mean normalized prediction MSE against the number of trainable parameters for each model.

Comparing datasets’ difficulty, we find that the most challenging dataset to predict was MMBR (PCC=0.16), while 10XGMBSP emerged as the least difficult (PCC=0.74). When inspecting each dataset’s characteristics, we observe that, although the organism doesn’t seem to have an impact on the difficulty of the task, a larger amount of available genes (due to better quality) and generalizing in an intra-patient fashion typically makes the prediction easier (Supplementary Fig. 1.a and 1.b for tissue type analysis).

We display the results of evaluating the 7 state-of-the-art models on SpaRED, as well as the baseline experiments on Fig. 4.a sorted by best average performance. The normalized MSE metric indicates how close every model’s results are to the best performance achieved on each dataset. Results show that ViT-B+ $\Delta$  attains the best gene expression predictions on average, despite being one of the most straightforward approaches for the prediction task. Moreover, the pie chart showcases that STNet and ViT-B+ $\Delta$  emerge most frequently as the best methods. Interestingly, we notice that SEPAL, which is built on top of ViT-B+ $\Delta$ , falls behind the latter. This contrast reveals that incorporating local vicinity information does not necessarily improve the outputs and that focusing on predicting the  $\Delta$  from the mean expression is already a powerful strategy. Supplementary Table 2 illustrates the statistical differences in MSE performance across all datasets.

In general, most prediction methods exhibit roughly comparable performance. Moreover, our results also indicate that more complex architectures do not necessarily provide superior predictions on our benchmark. This behavior is also supported by Fig. 4.b, where Hist2ST ranks as the method with the most trainable parameters but performs worse than methods with orders of magnitude fewer parameters. In contrast, ShuffleNet is the method with the fewest parameters and offers a competitive performance. We hypothesize that this counterintuitive



trend is caused by the limited scale of publicly available datasets (the biggest SpaRED dataset contains 43,804 spots), probably leading to overfitting in bigger models.

## 6 Conclusions

In this paper, we present SpaRED, an extensive Visium database composed of 26 curated and standardized datasets, which emerges as a novel standard point of comparison for gene expression prediction from histology images methods. We also introduce SpaCKLE, a transformer-based model that successfully overcomes the dropout limitations, completing gene expression values even when the missing data fraction is up to 70%. Finally, we benchmark 7 state-of-the-art methods in SpaRED, demonstrating that completing training data beforehand using SpaCKLE significantly improves the performance of prediction models. Consequently, our work represents a significant advancement in the automation of Spatial Transcriptomics and is intended to promote further research in this field.

**Acknowledgments.** Gabriel Mejia and Daniela Vega acknowledge the support of UniAndes-GoogleDeepMind Scholarships 2022 and 2024 respectively.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. Abalo, X., Thrane, K., Ji, A.L., et al.: Human squamous cell carcinoma, visium **1** (2021). <https://doi.org/10.17632/2bh5fchcv6.1>
2. Avşar, G., Pir, P.: A comparative performance evaluation of imputation methods in spatially resolved transcriptomics data. *Molecular Omics* **19**(2), 162–173 (2023). <https://doi.org/10.1039/d2mo00266c>
3. Biancalani, T., Scalia, G., Buffoni, L., et al.: Deep learning and alignment of spatially resolved single-cell transcriptomes with tangram. *Nature methods* **18**(11), 1352–1362 (2021)
4. Chen, K.H., Boettiger, A.N., Moffitt, J.R., et al.: Spatially resolved, highly multiplexed rna profiling in single cells. *Science* **348** (4 2015). <https://doi.org/10.1126/science.aaa6090>, <https://www.science.org/doi/10.1126/science.aaa6090>
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
6. Erickson, A., He, M., Berglund, E., et al.: Spatially resolved clonal copy number alterations in benign and malignant tissue. *Nature* **608**(7922), 360–367 (2022)
7. Fan, Y., Andrusivová, Ž., Wu, Y., et al.: Expansion spatial transcriptomics. *Nature Methods* pp. 1–4 (2023)

8. He, B., Bergenstr ahle, L., Stenbeck, L., et al.: Integrating spatial gene expression and breast tumour morphology via deep learning. *Nature biomedical engineering* **4**(8), 827–834 (2020)
9. He, K., Chen, X., Xie, S., Li, Y., Dollar, P., Girshick, R.: Masked autoencoders are scalable vision learners. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* **2022-June**, 15979–15988 (11 2021). <https://doi.org/10.1109/CVPR52688.2022.01553>, <https://arxiv.org/abs/2111.06377v3>
10. Jiang, Y., Xie, J., Tan, X., Ye, N., Nguyen, Q.: Generalization of deep learning models for predicting spatial gene expression profiles using histology images: A breast cancer case study. *bioRxiv* (2023). <https://doi.org/10.1101/2023.09.20.558624>, <https://www.biorxiv.org/content/early/2023/09/22/2023.09.20.558624>
11. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization (2017)
12. Korsunsky, I., Millard, N., Fan, J., et al.: Fast, sensitive and accurate integration of single-cell data with harmony. *Nature methods* **16**(12), 1289–1296 (2019)
13. Mejia, G., C ardenas, P., Ruiz, D., Castillo, A., Arbel ez, P.: Sepal: Spatial gene expression prediction from local graphs. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*. pp. 2294–2303 (October 2023)
14. Mirzazadeh, R., Andrusivova, Z., Larsson, L., et al.: Spatially resolved transcriptomic profiling of degraded and challenging fresh frozen samples. *Nature Communications* **14**(1), 509 (2023)
15. Palla, G., Spitzer, H., Klein, M., et al.: Squidpy: A scalable framework for spatial omics analysis. *Nature Methods* **19**(2), 171–178 (Jan 2022). <https://doi.org/10.1038/s41592-021-01358-2>
16. Pang, M., Su, K., Li, M.: Leveraging information in spatial transcriptomics to predict super-resolution gene expression from histology images in tumors. *bioRxiv* pp. 2021–11 (2021)
17. Parigi, S.M., Larsson, L., Das, S., et al.: The spatial transcriptomic landscape of the healing mouse intestine following damage **13**(1), 828 (2022)
18. Pham, D., Tan, X., Balderson, B., et al.: Robust mapping of spatiotemporal trajectories and cell–cell interactions in healthy and diseased tissues. *Nature Communications* **14**(1) (Nov 2023). <https://doi.org/10.1038/s41467-023-43120-6>
19. Stickels, R.R., Murray, E., Kumar, P., et al.: Highly sensitive spatial transcriptomics at near-cellular resolution with slide-seq2. *Nature Biotechnology* **39**, 313–319 (3 2021). <https://doi.org/10.1038/s41587-020-0739-1>, <https://www.nature.com/articles/s41587-020-0739-1>
20. Stuart, T., Butler, A., Hoffman, P., et al.: Comprehensive integration of single-cell data. *cell* **177**(7), 1888–1902 (2019)
21. St ahl, P.L., Salm en, F., Vickovic, S., et al.: Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353**, 78–82 (7 2016). <https://doi.org/10.1126/science.aaf2403>, <https://www.science.org/doi/10.1126/science.aaf2403>
22. Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need (2023)
23. Vicari, M., Mirzazadeh, R., Nilsson, A., et al.: Spatial multimodal analysis of transcriptomes and metabolomes in tissues. *Nature Biotechnology* pp. 1–5 (2023)
24. Villacampa, E.G., Larsson, L., Mirzazadeh, R., et al.: Genome-wide spatial expression profiling in formalin-fixed tissues. *Cell Genomics* **1**(3) (2021)

25. Wang, G., Wu, S., Xiong, Z., et al.: CROST: a comprehensive repository of spatial transcriptomics. *Nucleic Acids Research* **52**(D1), D882–D890 (10 2023). <https://doi.org/10.1093/nar/gkad782>, <https://doi.org/10.1093/nar/gkad782>
26. Welch, J.D., Kozareva, V., Ferreira, A., Vanderburg, C., Martin, C., Macosko, E.Z.: Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell* **177**(7), 1873–1887 (2019)
27. Xie, R., Pang, K., Bader, G.D., Wang, B.: Spatially resolved gene expression prediction from h&e histology images via bi-modal contrastive learning. *arXiv preprint arXiv:2306.01859* (2023)
28. Yang, Y., Hossain, M.Z., Stone, E., Rahman, S.: Spatial transcriptomics analysis of gene expression prediction using exemplar guided graph neural network. *Pattern Recognition* **145**, 109966 (2024)
29. Yang, Y., Hossain, M.Z., Stone, E.A., Rahman, S.: 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*. pp. 5039–5048 (2023)
30. Zeng, Y., Wei, Z., Yu, W., et al.: Spatial transcriptomics prediction from histology jointly through transformer and graph neural networks. *Briefings in Bioinformatics* **23**(5), bbac297 (2022)
31. Zhang, X., Zhou, X., Lin, M., Sun, J.: Shufflenet: An extremely efficient convolutional neural network for mobile devices. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 6848–6856 (2018)