xMINT: A Multimodal Integration Transformer for Xenium Gene Imputation

Anonymous Authors[1]

Abstract

Xenium provides multimodal data with pathology images and corresponding spatial gene expressions to enhance biomedical studies. However, its limited ability to sequence only around 500 genes introduces complexity in panel design and restricts its capacity for exploration analysis. To address this challenge, some methods are developed to impute genes based on external single-cell RNA sequencing (scRNA-seq) data; however, they have neglected the rich cellular morphology and location information available in the Xenium pathology images. We introduce xMINT (**M**ultimodal **In**tegration **T**ransformer for **X**enium), a novel gene imputation method utilizing both gene expression profiles and corresponding pathology images to enhance imputation accuracy for Xenium data. xMINT is small and efficient; yet it has a superior imputation accuracy compared to competing methods.

1. Introduction

Spatial transcriptomics technologies have revolutionized biological studies by providing spatially resolved gene expression data (Marx, 2021). Xenium, a commercial high-resolution spatial transcriptomics technology by 10x Genomics, is among the most popular ones. In Xenium, imaging-based sequencing precisely localize each RNA molecule, while paired high-resolution Hematoxylin and Eosin (H&E) Whole Slide Imaging (WSI) provides further morphology information. However, Xenium can only sequence around 500 genes, imposing challenges on panel design and limit its capacity for downstream exploratory analysis.

To address this issue, researchers have developed several methods for imputing the missing genes in the Xenium panel. Current methods utilize external single-cell RNA sequencing (scRNA-seq) data (Lopez et al., 2019; Stuart et al., 2019; Abdelaal et al., 2020), establishing a joint embedding space between Xenium and scRNA-seq data to facilitate imputation. For instance, Seurat employs Canonical Correlation Analysis (CCA), gimVI leverages a deep generative model, and SpaGE adopts a domain adaptation approach called PRECISE to identify the joint space. However, scRNA-seq and Xenium use different technologies, which may introduce discrepancies in gene transcript levels. Efforts to reduce the discrepancies typically involve batch correction, which may weaken signal-to-noise ratios and reduce imputation accuracy. Besides, these methods have not utilized the cell morphology and location information from pathology images available for Xenium data, which could provide additional clues about cell types and states.

Pathological imaging has long been fundamental to biological studies, primarily through the examination of tissue sections based on their visual characteristics. Recent studies have identified relationships between pathology image features and spatial gene expressions, suggesting that image data can enhance gene expression imputation. For example, ImageCCA explores the relationship between image features and high-dimensional genomic markers (Ash et al., 2021), and iSTAR leverages image information to improve gene expression resolution in low-resolution spatial transcriptomics (ST) data, achieving a resolution close to the single-cell level (Zhang et al., 2024).

Recognizing the potential contribution of pathology images in enhancing gene imputation, we propose a new multimodal gene imputation method, xMINT. Our new method utilizes one Xenium dataset (the origin dataset) with coupled high-resolution gene expression data and pathology images, to impute some missing genes in another Xenium dataset (the target dataset). These two Xenium panels contain shared genes and unique genes. xMINT adopts Transformer-based models, which have shown great success in integrating multiple data modalities (Xie et al., 2023; Xu et al., 2023). xMINT first constructs sequences using the gene expression data and pathology images, then uses a multimodal Transformer to integrate the two data modalities for gene imputation. Our experiments show that xMINT outperforms the existing scRNA-seq-based methods in gene imputation accuracy.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the ICML 2024 Workshop on Accessible and Efficient Foundation Models for Biological Discovery. Do not distribute.

2. Methods

2.1. Data sequentizer and notations

Xenium data contain rich multimodal information. Here, we introduce the data modalities used as inputs for xMINT. Specifically, we construct sequences of cells in local regions, which contain both gene expression data and pathology images.

First, we divide the WSI into small regions with 1024×1024 pixels, each containing approximately 1000 to 10,000 cells. The ith regional image is denoted by R_i. The number of cells in local region i is denoted by N_i. We order N_i cells into a long sequence according to their cell ID in the data. The distances between cell IDs partially reflect their spational locations but the relationship is not completely monotone. Then we repeat this long sequence $B = 5$ times to further expand the long sequence to contain N_iB cells. Then, we sequentially segment this long sequence into K-cell short sequence ($K = 1024$). Thus, the total number of short sequences for each local region is $S_i = \lfloor \frac{N_i \times B}{K} \rfloor$. Each cell ID sequence is denoted by $\text{Ind}_{i,j}$, where $j \in [S_i]$.

Using $\text{Ind}_{i,j}$ as a baseline cell sequence, we define this sequence's corresponding gene expression sequences: the shared gene sequence $\text{Sha}_{i,j}$ and the imputed gene sequence $\text{Imp}_{i,j}$. The shared gene sequence contains the expressions of the genes shared by both Xenium datasets; the values are ordered based on the cells in $\text{Ind}_{i,j}$, forming a $K \times |\mathcal{G}_{\text{Sha}}|$ feature matrix, where $\mathcal{G}_{\text{Sha}}$ represents the common gene set. Similarly, the imputed gene sequence, which is a $K \times |\mathcal{G}_{\text{Imp}}|$ feature matrix, contains the expressions of the genes present in the origin dataset but missing in the target dataset, and thus need to be imputed.

One technical variation source in Xenium data is the cell library size. To address this issue, we normalize the gene expressions based on the library size. Because two Xenium panels do not have the identify genes, to unify both datasets, we use the following restricted-gene-set library sizes to normalize gene expressions. For cell c, and $\hat{g} \in \mathcal{G}_{\text{Sha}}$, the normalized gene expression is calculated as $X_{c\hat{g}} = Y_{c\hat{g}} / \sum_{g \in \mathcal{G}_{\text{Sha}}} Y_{cg}$, where Y_{cg} represents the UMI counts of gene g in cell c. Similarly, for $\hat{g} \in \mathcal{G}_{\text{Imp}}$, the normalized imputed gene expression is calculated as $X_{c\hat{g}} = Y_{c\hat{g}} / \sum_{g \in \mathcal{G}_{\text{Imp}}} Y_{cg}$. These values are used to construct the gene expression sequences $\text{Sha}_{i,j}$ and $\text{Imp}_{i,j}$.

2.2. xMINT Framework

As shown in Figure 1, the xMINT framework employs an Image Tokenizer and a Gene Transformer to create sequences from pathology images and gene expression data, respectively. These sequences are then integrated using a Transformer model, then the outputs are further processed to impute missing gene expressions. The architecture comprises the following key steps.

2.2.1. Constructing pathology sequences

The single-cell-level image tokenization contains large redundancy and is computationally expensive. Thus, we divide each local region R_i into smaller patches, and use the patch's morphology features to represent all their containing cells' morphology features. This approach substantially reduces computational cost; yet still maintains high performance in downstream tasks.

Specifically, the **Image Tokenizer** module uses a custom ResNet architecture to process local region images R_i of size $1024 \times 1024 \times 3$. The ResNet architecture contains three ResNet block layers. It extracts f feature maps without altering the original width and height dimensions, resulting in $1024 \times 1024 \times f$ feature maps. Then, the local region feature maps are divided into 16×16 patches, which is approxiamately single cell level. For any given cell in $\text{Ind}_{i,j}$, we find its residing patch and use the patch features after max pooling to represent the cell's morphology features ($1 \times f$). These cells' morphology features form the pathology sequence, a $K \times f$ feature matrix.

2.2.2. Contructing gene expression sequences

The shared gene sequence $\text{Sha}_{i,j}$ contains the normalized gene expressions of the shared genes in both datasets. To integrate local information in model training, we use a **Gene Transformer** module to process the normalized shared genes and output a $K \times f$ feature matrix.

2.2.3. Integrating multimodal sequences and constructing final output

Next, in the **Integrated Multimodal Transformer** module, we first integrate the outputs from the Image Tokenizer and the Gene Transformer, both having dimensions of $K \times f$. By concatenating these two matrices along the feature dimension, we form a new multimodal sequence with dimensions $K \times 2f$. This sequence is input into the Integrated Multimodal Transformer to capture interactions between pathological and gene expression features. The output multimodal sequence is a $K \times 2f$ matrix. Next, each token is projected into a $|\mathcal{G}_{\text{Imp}}|$-dimensional space for the imputation task using a feed-forward neuro network.

Finally, for each cell, we record the gene imputation results in all sequences containing this cell; the final imputated gene expression is the average of the corresponding cell in all sequences.

2.3. Model Parameters and Computation Time

xMINT includes the Gene Transformer, Integrated Multimodal Transformer, and an Image Tokenizer. Adjustable

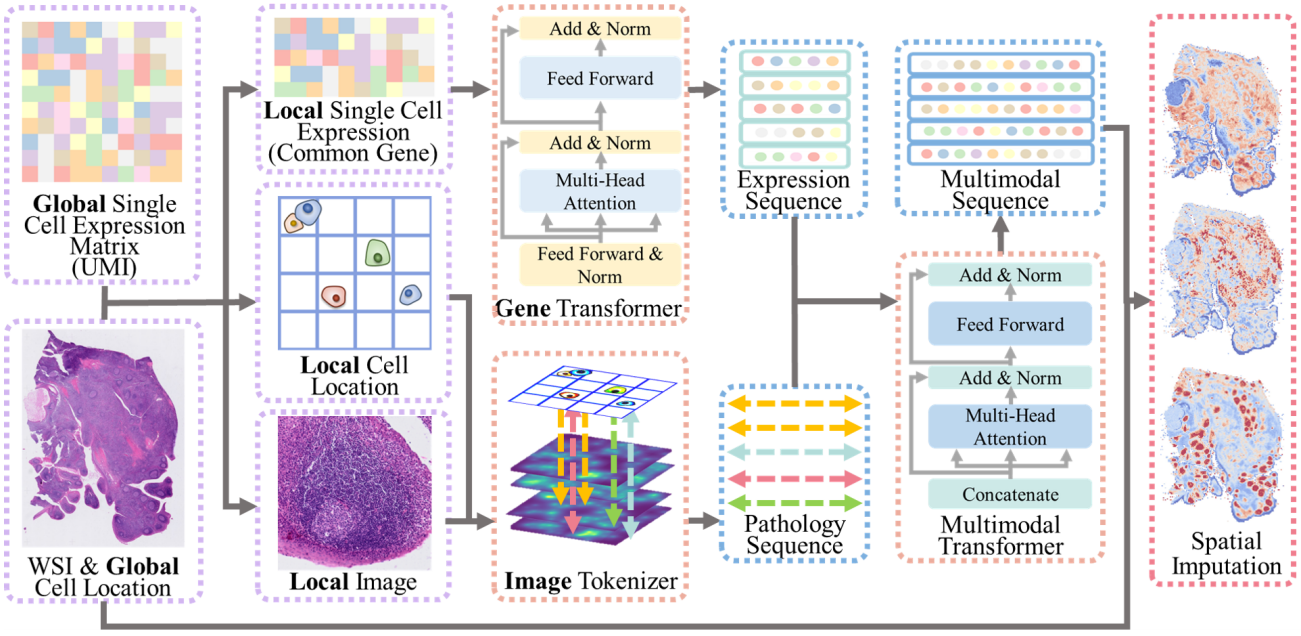

Figure 1. Flowchart of xMINT: Initially, WSI is divided into local regions. Sequences are generated within each local region using morphology features and RNA sequencing values. The image tokenizer uses local image data and cell coordinates to create a pathology sequence, while shared gene expression data are used to generate an expression sequence. These two sequences are concatenated and input into a multimodal transformer, which then outputs imputed gene sequences.

Table 1. xMINT Model Parameters and Computation Time

f	EXP.	HEAD	LAYER	PARAMETERS	TIME (S)
256	4	4	4	16,799,760	1.03
256	8	8	8	53,564,432	1.09
128	4	4	4	4,338,064	0.56
128	8	8	8	13,545,360	0.60

parameters are embedding size (Emb.), forward expansion factor (Exp.), number of attention heads (Head), and number of transformer layers (Layer) for the transformers, and number of feature maps (Maps) for the Image Tokenizer. f is a shared parameter across these three models, indicating the complexity of the features. In the Gene Transformer, Emb. $= f$, and in the Integrated Multimodal Transformer, Emb. $= 2f$; other settings in these two transformers are the same. In the Image Tokenizer, Maps $= f$.

Table 1 shows four different configurations, and the corresponding parameters and computation time per batch with 4 sequences on a single RTX A6000. To generate analysis results in this paper, we used $f = 256$, Head $= 8$, Exp. $= 8$, Layer $= 4$, with a computation time of 1.05 seconds per batch.

3. Results

The study employs two Xenium human tonsil datasets from 10x Genomics website. Follicular lymphoid hyperplasia is used for training, and reactive follicular hyperplasia is used for testing.

To quantify the accuracy of gene imputation in terms of spatial patterns, we define a metric to measure the similarity between the spatial patterns of imputed and true gene values for each gene. Specifically, we first partition the WSI into 128x128 pixel regions and calculate the mean true gene expression values and the mean imputed gene values for each region. Then, we calculate the Spearman correlation between these two sets of mean values.

We first evaluate the robustness of xMINT with different numbers of shared genes, ranging from 50 to 200 (Figure 2a). We use the first 50 to 200 genes in the Xenium tonsil panel to ensure that when we add more genes, the previous genes are kept in the training set, thus avoiding performance changes due to randomness. We found that, in general, xMINT's performance is stable across different numbers of shared genes. Increasing the number of shared genes from 50 to 100 slightly improves the imputation accuracy in visualization, but little improvement is observed when the number of shared genes is further increased to 150 or 200. In fact, when the number of shared genes increases from 100 to 150, we see a slight drop in median

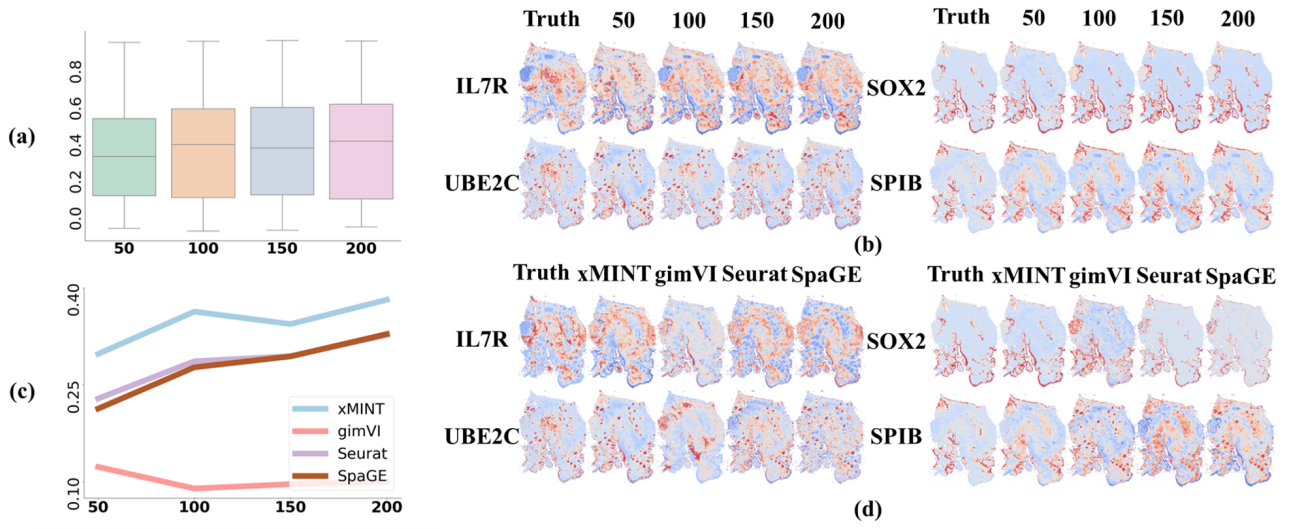

Figure 2. Performance of xMINT. (a) Boxplots of Spearman correlations between xMINT predicted and true gene expressions using 50, 100, 150, and 200 shared genes. (b) Provides examples of xMINT results. (c) Shows the comparison with scRNA-based methods using the same metrics. (d) Shows examples of comparison using 50 shared genes.

performance. This is probably due to some of the shared genes may not be informative for imputation, or even introduce noise. Figure 2b displays four example genes, *IL7R*, *UBE2C*, *SOX2*, and *SPIB*: the spatial patterns of the imputed gene expressions are very similar to the true gene expressions, indicating that xMINT has successfully captured the spatial patterns in gene expressions.

We compared the performance of xMINT with scRNA-based methods, including gimVI, Seurat, and SpaGE. All three methods used the same shared gene set as xMINT, and an online tonsil scRNA-seq dataset for imputation (Massoni-Badosa et al., 2024). The Spearman correlation of gimVI does not increase with the number of shared genes, while Seurat and SpaGE show a consistent increase in Spearman correlation. However, xMINT outperforms all three methods across all numbers of shared genes (Figure 2c). Specifically, with the number of shared gene equal to 50, the spatial patterns of xMINT predicted genes are more similar to the true gene expressions than those of the other methods (Figure 2d). The results indicate that incorporating pathology images can enhance gene imputation accuracy.

4. Discussion

4.1. Contribution

We propose xMINT, a new computational method to impute the missing genes in Xenium data. xMINT utilizes both pathology images and spatial gene expressions to impute genes, thus outperforming methods that only use external scRNA-seq data for imputation.

Although xMINT employs multiple Transformer modules, the parameters in these Transformer models are quite affordable. See Section 2.3 for details. It can be trained on a single RTX A6000 GPU within 36 hours (100 epochs), making it cost-effective and time-efficient. Its architecture can be extended to larger Transformer models with more parameters, which may further improve imputation accuracy. Yet, the current slim model has already shown superior performance compared to existing methods. This sheds light on using small and affordable deep learning models to link pathology images with genomic data.

4.2. Limitation

To impute genes in a Xenium sample, our model requires another Xenium sample with the same tissue type. Since Xenium is a new technology, not many samples are publicly available, currently limiting the application scenarios. However, as Xenium is one of the most popular spatial transcriptomics technologies, we expect more samples to be publicly available in the next two or three years. Many large consortia funded by the National Institutes of Health (NIH), such as TOPMed and HTAN (National Cancer Institute, 2024; National Heart, Lung, and Blood Institute, 2024), have plans to collect and release massive Xenium samples in the next few years. Thus, we expect xMINT to be widely applicable in the near future.

References

Abdelaal, T., Mourragui, S., Mahfouz, A., and Reinders, M. J. T. SpaGE: Spatial Gene Enhancement using scRNA-seq. *Nucleic Acids Research*, 48(18):e107–e107, 09 2020. ISSN 0305-1048. doi: 10.1093/nar/gkaa740. URL https://doi.org/10.1093/nar/gkaa740.

Ash, J., Darnell, G., Munro, D., et al. Joint analysis of expression levels and histological images identifies genes associated with tissue morphology. *Nature Communications*, 12:1609, 2021. doi: 10.1038/s41467-021-21727-x.

Lopez, R. et al. A joint model of unpaired data from scrna-seq and spatial transcriptomics for imputing missing gene expression measurements. In *ICML Workshop on Computational Biology*, 2019.

Marx, V. Method of the year: spatially resolved transcriptomics. *Nat Methods*, 18(1):9–14, Jan 2021. doi: 10.1038/s41592-020-01033-y.

Massoni-Badosa, R., Aguilar-Fernández, S., Nieto, J. C., Soler-Vila, P., Elosua-Bayes, M., Marchese, D., Kulis, M., Vilas-Zornoza, A., Bühler, M. M., Rashmi, S., Alsinet, C., Caratù, G., Moutinho, C., Ruiz, S., Lorden, P., Lunazzi, G., Colomer, D., Frigola, G., Blevins, W., Romero-Rivero, L., Jiménez-Martínez, V., Vidal, A., Mateos-Jaimez, J., Maiques-Diaz, A., Ovejero, S., Moreaux, J., Palomino, S., Gomez-Cabrero, D., Agirre, X., Weniger, M. A., King, H. W., Garner, L. C., Marini, F., Cervera-Paz, F. J., Baptista, P. M., Vilaseca, I., Rosales, C., Ruiz-Gaspà, S., Talks, B., Sidhpura, K., Pascual-Reguant, A., Hauser, A. E., Haniffa, M., Prosper, F., Küppers, R., Gut, I. G., Campo, E., Martin-Subero, J. I., and Heyn, H. An atlas of cells in the human tonsil. *Immunity*, 57:379–399.e18, 2024. ISSN 1074-7613. doi: 10.1016/j.immuni.2024.01.006. URL https://www.sciencedirect.com/science/article/pii/S1074761324000311.

National Cancer Institute. Human tumor atlas network (htan). https://humantumoratlas.org/, 2024. Accessed: 2024-05-24.

National Heart, Lung, and Blood Institute. Trans-omics for precision medicine (topmed) program. https://www.nhlbi.nih.gov/science/trans-omics-precision-medicine-topmed-program, 2024. Accessed: 2024-05-24.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. Comprehensive integration of single-cell data. *cell*, 177(7):1888–1902, 2019.

Xie, W., Fang, Y., Yang, G., Yu, K., and Li, W. Transformer-based multi-modal data fusion method for copd classification and physiological and biochemical indicators identification. *Biomolecules*, 13(9), Sep 2023. doi: 10.3390/biom13091391.

Xu, P., Zhu, X., and Clifton, D. A. Multimodal learning with transformers: A survey. *IEEE Trans Pattern Anal Mach Intell*, 45(10):12113–12132, Oct 2023. doi: 10.1109/TPAMI.2023.3275156.

Zhang, D., Schroeder, A., Yan, H., et al. Inferring super-resolution tissue architecture by integrating spatial transcriptomics with histology. *Nature Biotechnology*, 2024. doi: 10.1038/s41587-023-02019-9.