SAINT: Sequence-Aware Integration for Spatial Transcriptomics Multi-View Clustering

Zeyu Zhu[1] **Ke Liang**[1*] **Lingyuan Meng**[1] **Meng Liu**[1] **Suyuan Liu**[1]
Renxiang Guan[1] **Miaomiao Li**[2] **Wanwei Liu**[1] **Xinwang Liu**[1*]
[1]National University of Defense Technology, Changsha, China
[2]Changsha College, Changsha, China

Abstract

Spatial transcriptomics (ST) technologies provide gene expression measurements with spatial resolution, enabling the dissection of tissue structure and function. A fundamental challenge in ST analysis is clustering spatial spots into coherent functional regions. While existing models effectively integrate expression and spatial signals, they largely overlook sequence-level biological priors encoded in the DNA sequences of expressed genes. To bridge this gap, we propose SAINT (Sequence-Aware Integration for Nucleotide-informed Transcriptomics), a unified framework that augments spatial representation learning with nucleotide-derived features. We construct sequence-augmented datasets across 14 tissue sections from three widely used ST benchmarks (DLPFC, HBC, and MBA), retrieving reference DNA sequences for each expressed gene and encoding them using a pretrained Nucleotide Transformer. For each spot, gene-level embeddings are aggregated via expression-weighted and attention-based pooling, then fused with spatial-expression representations through a late fusion module. Extensive experiments demonstrate that SAINT consistently improves clustering performance across multiple datasets. Experiments validate the superiority, effectiveness, sensitivity, and transferability of our framework, confirming the complementary value of incorporating sequence-level priors into spatial transcriptomics clustering.

1 Introduction

Spatial transcriptomics (ST) technologies measure gene expression while preserving the spatial layout of tissue sections, enabling the study of molecular organization in space [39, 46, 61, 54]. A key analysis task is clustering spatial spots into biologically meaningful regions, which reveals tissue structures and spatially patterned gene programs. This task can be naturally viewed as a multi-view clustering problem, where each spot is associated with both a gene expression profile and spatial coordinates [52, 48, 44, 11, 29]. Specifically, gene expression captures cell identity and state, while spatial proximity informs local organization and neighborhood relationships. Effectively integrating these heterogeneous views is key to producing biologically meaningful clusters.

Early attempts on spatial transcriptomics clustering primarily relied on unsupervised methods applied to gene expression profiles alone, ignoring the spatial structure inherent in the data [45, 43]. To incorporate spatial context, later approaches introduced regularization terms or handcrafted spatial distances that penalize discontinuities across neighboring spots. While effective to some extent, these methods often struggle to generalize to irregular tissue geometries or heterogeneous microenvironments. More recently, graph-based neural models have become the dominant paradigm due to their flexibility in capturing complex spatial relationships[24, 25, 17, 53, 30, 58]. For example,

*Corresponding authors.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

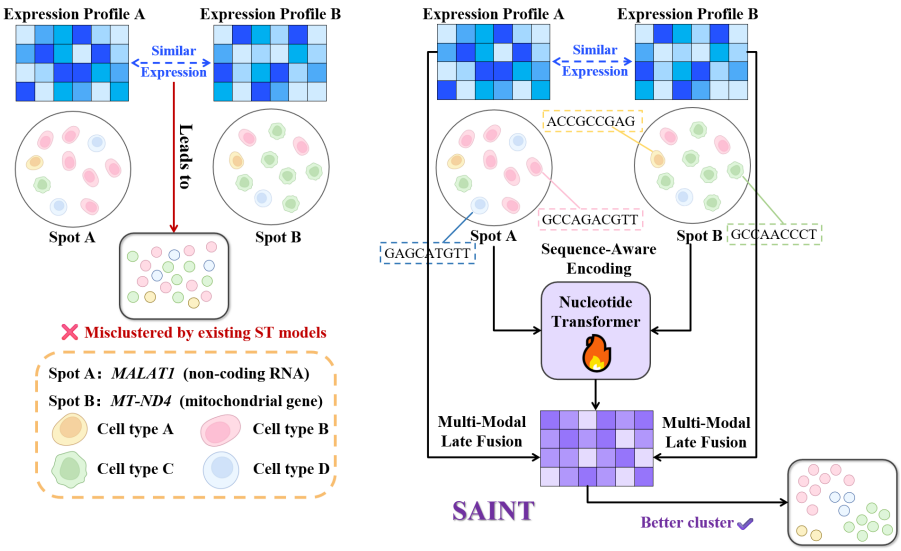

Figure 1: Motivation of SAINT. Existing ST clustering models may miscluster spatial spots with similar expression profiles but divergent gene functions. For example, Spot A and Spot B share highly similar expression profiles, yet express functionally divergent genes such as *MALAT1* and *MT-ND4* , respectively. This leads to incorrect clustering by expression-only models (left). By incorporating gene sequence features via a pretrained nucleotide encoder and multi-modal fusion, SAINT better distinguishes such cases and improves clustering accuracy (right).

SpaGCN [18] constructs a spatial graph based on physical coordinates and applies graph convolutions to jointly encode spatial proximity and gene expression. Building on this, Spatial-MGCN [47] aggregates multiple spatial graphs constructed from different biological priors (e.g., histological similarity, spatial distance), enabling a richer representation of local neighborhoods. Further, MAFN [62] proposes a late-fusion strategy that combines spatial, feature, and contrastive graphs through an adaptive attention module, improving clustering robustness under noise and sparsity. These GNN-based methods have consistently achieved state-of-the-art results across multiple ST benchmarks, highlighting the effectiveness of learning spot embeddings over spatial graphs.

However, existing ST clustering methods rely solely on observed gene expression and spatial proximity, while ignoring the rich biological information encoded in gene sequences. In practice, each spatial spot expresses a set of genes, and each gene is uniquely associated with a DNA sequence composed of nucleotides. These sequences often reflect regulatory elements or biochemical roles that are not evident from expression levels alone. As illustrated in Figure 1, spots A and B exhibit similar expression profiles. However, their dominant genes, *MALAT1* (a non-coding RNA) and *MT-ND4* (a mitochondrial protein-coding gene), indicate distinct biological roles. This discrepancy can lead clustering models to mistakenly group spatial spots into the same region simply because their expression profiles look similar, even though the expressed genes may have fundamentally different biological functions. Despite this, no prior work has systematically incorporated nucleotide-level sequence information into ST clustering, largely due to the following two key challenges in incorporating gene sequence knowledge into spatial transcriptomics clustering.

(1) Lack of Sequence-Annotated Datasets. Existing ST datasets do not contain nucleotide-level annotations, making it difficult to explore how gene sequences influence spatial gene expression or regional identity.

(2) Cross-Modal Representation Integration. Even with sequence information available, it remains unclear how to effectively encode gene sequences and integrate them with spatial and expression features for clustering. Naïvely combining modalities may lead to noise or semantic mismatch.

To fill this gap, we propose **SAINT** (Sequence-Aware Integration for Nucleotide-informed Transcriptomics), a sequence-informed framework that augments spatial transcriptomics with gene-level nucleotide embeddings. **First**, we construct sequence-augmented datasets across 14 tissue

sections spanning three widely used benchmarks: DLPFC (12 slices), HBC, and MBA. For each gene expressed within a spatial spot, we retrieve the corresponding reference DNA sequence from NCBI and organize the data into spotgenesequence mappings, enabling the integration of nucleotide-level information into the modeling pipeline. **Second**, we encode each DNA sequence using the Nucleotide Transformer, a large-scale pretrained language model for genomics. This yields rich, high-dimensional embeddings that capture regulatory signals and sequence-level semantics. These embeddings are projected into a lower-dimensional space and aggregated at the spot level to form sequence-derived spot representations. **After that**, we observe that each spot typically expresses dozens of genes, but not all are equally informative. To mitigate the influence of noisy or redundant sequences, we filter out low-variance genes and apply a lightweight attention mechanism to assign adaptive weights to the remaining gene embeddings, producing compact and spot-specific representations. **Finally**, the learned sequence-aware embeddings are integrated with expression and spatial features using a late fusion module. Experiments are conducted to evaluate the capacity of SAINT from four aspects: superiority, effectiveness, sensitivity, and transferability. The main contributions of our work are summarized as follows.

- **Problem.** Existing spatial transcriptomics (ST) clustering methods rely primarily on expression profiles and spatial coordinates, overlooking the biological priors encoded in gene sequences. This omission limits the semantic expressiveness of current representations and may lead to functionally mismatched clusters.
- **Dataset.** To address this, we construct multiple sequence-augmented ST datasets spanning 14 tissue sections from three benchmarks (DLPFC, HBC, MBA). Each spatial spot is annotated with the reference DNA sequences of its expressed genes, enabling the first systematic exploration of nucleotide-level information in ST clustering.
- **Method.** We propose SAINT, a sequence-informed multi-modal learning framework. SAINT encodes gene sequences using a pretrained genomic transformer model, filters uninformative genes based on expression variability, and applies attention-based aggregation to derive compact spot-level embeddings. These are then fused with expression and spatial representations through a lightweight late-fusion architecture.
- **Experiment.** Extensive experiments demonstrate that SAINT consistently improves clustering performance across multiple datasets. We evaluate SAINT from four perspectives: superiority, effectiveness, sensitivity, and transferability, confirming the complementary value of integrating nucleotide-level features into ST representation learning.

2 Related Work

This section summarizes recent related works from three aspects: (1) multiview clustering methods in spatial transcriptomics data, (2) genomic language modeling for sequence representation, and (3) sequenceaugmented spatial transcriptomics clustering. Due to space limitations, please refer to Appendix A.2 for a more detailed discussion.

3 Method

In this section, we present SAINT, a multi-modal learning framework for spatial transcriptomics clustering that integrates spatial coordinates, gene expression, and sequence-derived biological priors into unified spot-level embeddings. Unlike prior methods that rely solely on spatial and expression features, SAINT incorporates nucleotide-level representations to enhance spatial domain discovery. The framework of SAINT is shown in Fig.2.

The pipeline begins with data preprocessing, where each spatial spot is associated with its expression vector and the DNA sequences of its expressed genes. After normalization and selection of highly variable genes (HVGs), SAINT extracts complementary representations through two parallel modules. The structure-aware graph embedding module constructs spatial, feature, and combined graphs from HVGs, and encodes them using graph convolutional networks (GCNs). These multi-view embeddings are then aggregated via attention to form a graph-derived representation Z^{graph}. In parallel, the sequence-aware encoder tokenizes DNA sequences and feeds them into a pretrained Nucleotide Transformer to produce gene-level embeddings, which are aggregated and projected into

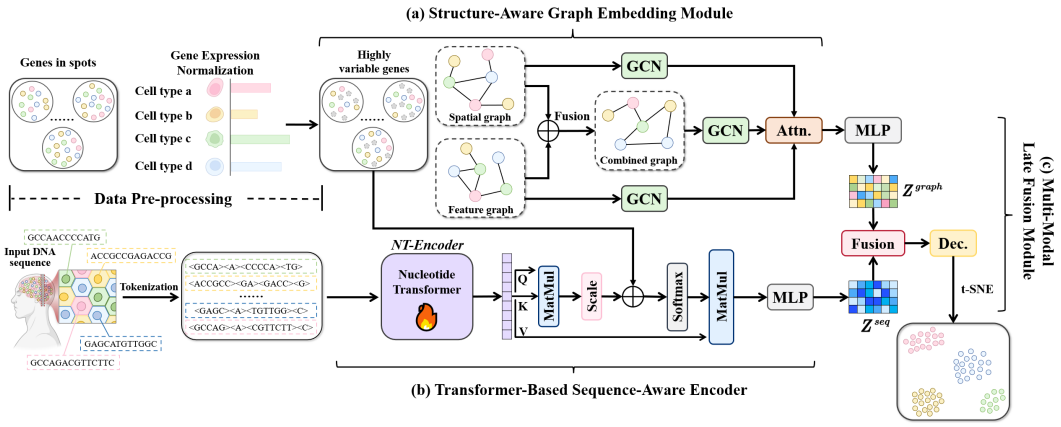

Figure 2: An overview of the SAINT framework. The pipeline starts with preprocessing, where each spatial spot is linked to its gene expression vector and DNA sequences. After normalization and HVG selection, two parallel modules extract structural and sequence-level features. (a) A graph embedding module encodes spatial, feature, and combined graphs via GCNs and fuses them through attention to obtain Z^{graph}. (b) A transformer-based encoder maps tokenized gene sequences to contextual embeddings, which are aggregated to form Z^{seq}. (c) The two representations are fused and passed to a ZINB decoder for expression reconstruction, and the resulting embeddings are used for clustering.

spot-level features Z^{seq}. The two representations are integrated via a late fusion module to obtain unified embeddings, which are then passed to a ZINB decoder for reconstructing gene expression. Clustering is performed in this fused latent space, capturing both structural topology and sequence semantics.

3.1 Problem Formulation

ST datasets consist of N spatial spots, each profiled with transcriptome-wide gene expression and associated spatial coordinates. Formally, each spot $i \in 1, \ldots, N$ is described by three components: (1) a gene expression vector $\mathbf{x}i \in \mathbb{R}^G$ over G genes, (2) a 2D spatial location $s_i \in \mathbb{R}^2$, and (3) a set of expressed genes $\mathcal{G}i = g_{i1}, g_{i2}, \ldots, g_{iM_i}$, where each gene g_{ij} is associated with a DNA sequence d_{ij} of variable length. These sequences are encoded into fixed-length embeddings using a pretrained genomic language model. The objective of spatial transcriptomics clustering is to partition the N spots into K spatial domains $\mathcal{C} = C_1, C_2, \ldots, C_K$, such that spots within the same cluster share similar expression programs, spatial context, and sequence-informed regulatory features. Unlike traditional transcriptomics clustering, this task involves integrating heterogeneous modalities, including gene expression, physical location, and sequence-derived embeddings, each with potentially different dimensionality, semantic structure, and noise characteristics.

Formally, our aim is to learn a unified embedding function $f : (\mathbf{x}_i, s_i, d_{ij}) \mapsto \mathbf{z}_i \in \mathbb{R}^d$ for each spot i, such that the resulting embeddings $\mathbf{z}_1, \ldots, \mathbf{z}_N$ can be effectively clustered into biologically meaningful domains. This requires (1) extracting semantically rich features from raw DNA sequences, (2) adapting to the varying number of genes per spot, and (3) designing a robust fusion strategy that preserves complementary signals while mitigating cross-modal redundancy.

3.2 Structure-Aware Graph Embedding Module

In parallel, SAINT models the spatial and expression views using graph-based encoders. We construct a spatial neighbor graph $\mathcal{G}_s = (\mathcal{V}, \mathcal{E}_s)$ where nodes correspond to spatial spots and edges connect neighboring spots based on 2D distance thresholding. We also build a feature graph $\mathcal{G}_f$ based on k-nearest neighbors in the gene expression space. Both graphs are encoded via GCNs.

To capture spatial proximity, expression similarity, and their interaction, we construct three graphs: the spatial graph $\mathcal{G}_s$, the feature graph $\mathcal{G}_f$, and the combined graph $\mathcal{G}_c$. All graphs share the same node set (spots), but differ in adjacency structure.

Each graph is encoded separately using a Graph Convolutional Network (GCN). Given a graph $\mathcal{G}_v$ with adjacency matrix $\mathbf{A}_v$ and node features $\mathbf{X}$, the GCN propagates features via as follows.

$$\mathbf{H}_v = \sigma\left(\tilde{\mathbf{D}}_v^{-1/2}\tilde{\mathbf{A}}_v\tilde{\mathbf{D}}_v^{-1/2}\mathbf{X}\mathbf{W}_v\right), \tag{1}$$

where $\tilde{\mathbf{A}}_v = \mathbf{A}_v + \mathbf{I}$ adds self-loops, $\tilde{\mathbf{D}}_v$ is the degree matrix, $\mathbf{W}_v$ is a trainable weight matrix, and $\sigma(\cdot)$ is a ReLU activation.

To model complementary topology between views, we define the combined graph $\mathcal{G}_c$ by aggregating the adjacency matrices of the spatial and feature graphs.

$$\mathbf{A}_c = \mathbf{A}_s + \mathbf{A}_f, \tag{2}$$

This combined structure captures both spatial continuity and expression similarity in a unified topology. The corresponding node embeddings $\mathbf{H}_s$, $\mathbf{H}_f$, and $\mathbf{H}c$ are then fed into an attention-based fusion module.

$$\mathbf{H}_{\text{g}} = \mathcal{F}_{\text{attn}}([\mathbf{H}_s \,\|\, \mathbf{H}_f \,\|\, \mathbf{H}_c]), \tag{3}$$

where $[\cdot,|,\cdot]$ denotes feature-wise concatenation, and $\mathcal{F}_{\text{attn}}(\cdot)$ is an attention-based MLP that computes normalized weights across views to adaptively control the fusion process.

3.3 Transformer-Based Sequence-Aware Encoder

To capture biological priors at the sequence level, we employ a pretrained genomic language model, Nucleotide Transformer [10], to encode raw DNA sequences into dense vector embeddings. For each expressed gene g_{ij} in spot i, we retrieve its reference DNA sequence d_{ij} from curated databases (e.g., NCBI), and encode it into a dense vector.

$$\mathbf{z}_{ij} = \text{NT-Encoder}(d_{ij}) \in \mathbb{R}^D, \tag{4}$$

where NT-Encoder$(\cdot)$ denotes the pretrained transformer model, and D is the output dimension.

Since each spot may express a variable number of genes, we aggregate these gene-level embeddings into a fixed-dimensional representation using a lightweight attention pooling module. Specifically, for spot i, let $\mathbf{Z}_i = [\mathbf{z}_{i1}, \ldots, \mathbf{z}_{iM_i}] \in \mathbb{R}^{D\times M_i}$ denote the matrix of gene embeddings. We compute attention weights α_{ij} over genes based on a softmax-normalized scoring function.

$$\alpha_{ij} = \frac{\exp(\mathbf{w}^\top \tanh(\mathbf{W}\mathbf{z}_{ij}))}{\sum_{j'=1}^{M_i}\exp(\mathbf{w}^\top \tanh(\mathbf{W}\mathbf{z}_{ij'}))}, \tag{5}$$

where $\mathbf{W} \in \mathbb{R}^{d_a\times D}$ and $\mathbf{w} \in \mathbb{R}^{d_a}$ are learnable parameters.

The sequence-derived embedding for spot i is as follows.

$$\mathbf{z}_i^{\text{seq}} = \sum_{j=1}^{M_i}\alpha_{ij}\cdot\mathbf{z}_{ij}, \tag{6}$$

To ensure compatibility with downstream fusion and to reduce computational overhead, we project the sequence embedding into a lower-dimensional space via a two-layer multilayer perceptron (MLP) with ReLU activation[16].

$$\mathbf{h}_i^{\text{seq}} = \text{MLP}_{\text{seq}}(\mathbf{z}_i^{\text{seq}}) \in \mathbb{R}^{d_s}, \tag{7}$$

where d_s is a hyper-parameter denoting the projected dimension.

3.4 Multi-Modal Late Fusion Module

To obtain a comprehensive spot-level representation, we integrate information from both spatial graphs and gene sequences. Specifically, for each spot i, we concatenate the structural embedding $\mathbf{h}_i^{\text{gcn}}$ with the sequence-derived embedding $\mathbf{h}_i^{\text{seq}}$, capturing complementary features from spatial expression patterns and nucleotide-level signals. The concatenated vector is then projected into a shared latent space using a lightweight multilayer perceptron.

$$\mathbf{h}_i = \mathcal{F}_{\text{fuse}}([\mathbf{h}_i^{\text{gcn}}, |, \mathbf{h}_i^{\text{seq}}]), \tag{8}$$

where $[\cdot, |, \cdot]$ denotes concatenation, and $\mathcal{F}_{\text{fuse}}(\cdot)$ is a two-layer MLP with ReLU activation. This fusion module aligns the structural and sequence information into a unified representation, enabling the model to make better use of both expression and sequence-derived features during clustering.

3.5 Training Objective and Loss Functions

To jointly promote expression reconstruction, structural consistency, and cross-modal complementarity, we incorporate three loss components into the final training objective.

ZINB Reconstruction Loss. To model over-dispersion and dropout noise in ST data, we adopt a Zero-Inflated Negative Binomial (ZINB) decoder. The probability of observing count x_{ig} for spot i and gene g is given as follows.

$$\text{ZINB}(x_{ig} \mid \mu_{ig}, \theta_{ig}, \pi_{ig}) = \begin{cases} \pi_{ig} + (1-\pi_{ig}) \left(\frac{\theta_{ig}}{\theta_{ig}+\mu_{ig}}\right)^{\theta_{ig}}, & \text{if } x_{ig} = 0 \\ (1-\pi_{ig}) \cdot \text{NB}(x_{ig} \mid \mu_{ig}, \theta_{ig}), & \text{if } x_{ig} > 0 \end{cases} \tag{9}$$

Here, μ_{ig} is the predicted mean expression, θ_{ig} is the dispersion parameter, and π_{ig} models the dropout probability. $\text{NB}(\cdot)$ denotes the negative binomial distribution.

The total reconstruction loss over all N spots and G genes is computed as follows.

$$\mathcal{L}_{\text{ZINB}} = \sum_{i=1}^{N} \sum_{g=1}^{G} -\log \text{ZINB}(x_{ig} \mid \mu_{ig}, \theta_{ig}, \pi_{ig}), \tag{10}$$

where x_{ig} denotes the observed count for gene g in spot i.

Graph Contrastive Regularization. To preserve local structural smoothness in the learned embeddings, we encourage proximity between neighboring nodes and separation between non-neighbors via a contrastive loss.

$$\mathcal{L}_{\text{Reg}} = \mathbb{E}_{(i,j)\in\mathcal{N}} \left[-\log \sigma(\cos(\mathbf{h}_i, \mathbf{h}_j))\right] + \mathbb{E}_{(i,j)\notin\mathcal{N}} \left[-\log(1-\sigma(\cos(\mathbf{h}_i, \mathbf{h}_j)))\right]. \tag{11}$$

Here, $\mathcal{N}$ denotes neighbor pairs in the spatial or feature graph, $\cos(\cdot,\cdot)$ denotes cosine similarity, and $\sigma(\cdot)$ is the sigmoid function. $\mathbf{h}_i$ is the final embedding of spot i after fusion.

Cross-Modal Redundancy Reduction (DICR). To encourage complementary information between the structural and sequence branches, we minimize feature redundancy using a decorrelation loss inspired by [33]. Let C be the cross-correlation matrix computed between ℓ_2-normalized embeddings from the two modalities. The loss is defined as follows.

$$\mathcal{L}_{\text{DICR}} = \sum_{i \neq j} C_{ij}^2 + \sum_i (C_{ii} - 1)^2, \tag{12}$$

where the first term penalizes off-diagonal correlations and the second term enforces identity alignment on the diagonal.

Final Objective. We jointly optimize the model by minimizing the following weighted sum of losses.

$$\mathcal{L} = \mathcal{L}_{\text{ZINB}} + \alpha \cdot \mathcal{L}_{\text{Reg}} + \gamma \cdot \mathcal{L}_{\text{DICR}}, \tag{13}$$

where α and γ are hyper-parameters balancing the topology regularization and cross-modal decorrelation objectives. For consistency and fair comparison, we adopt the same hyperparameter configuration as MAFN [62], using its default values across all experiments without further tuning.

4 Experiment

We conduct comprehensive experiments to evaluate the performance and robustness of our SAINT across multiple dimensions, i.e., superiority, effectiveness, transferability, sensitivity and case Study. Specifically, we aim to answer the following five questions.

- **Q1: Superiority.** Does SAINT outperform existing state-of-the-art models on spatial transcriptomics clustering benchmarks?
- **Q2: Effectiveness.** How effective are the introduced sequence-aware augmentation strategies in enhancing clustering quality?
- **Q3: Transferability.** Can SAINT be flexibly integrated into different clustering backbones?
- **Q4: Sensitivity.** How sensitive is SAINT to variations in hyper-parameters?
- **Q5: Case Study.** Does SAINT produce biologically meaningful clustering results in real-world spatial transcriptomics datasets?

4.1 Experiment Setting

This section introduces the details of the experiment setting from four aspects, i.e., datasets, implementation details, compared baselines and evaluation metrics. Due to space limitations, details are provided in Appendix A.3.

Table 1: Clustering performance of competing spatial transcriptomics models. Bold entries indicate the best results, and underlined values denote the second-best.

Method	Adjusted Rand Index (ARI)									
	151507	151508	151509	151510	151669	151670	151671	151672	HBC	MBA
SCANPY[49]	0.20	0.15	0.19	0.14	0.10	0.09	0.12	0.12	0.49	0.23
SpaGCN[18]	0.43	0.33	0.41	0.37	0.23	0.35	0.51	0.53	0.56	0.34
DeepST[51]	0.55	0.42	0.43	0.50	0.44	0.33	0.52	0.48	0.53	0.25
SCGDL[31]	0.49	0.34	0.32	0.31	0.24	0.26	0.31	0.34	0.35	0.26
stLearn[36]	0.49	0.31	0.45	0.44	0.32	0.23	0.39	0.34	0.55	0.38
Spatial-MGCN[47]	0.63	0.46	0.54	0.51	0.39	0.35	0.60	0.77	0.64	0.42
GraphST[34]	0.48	0.49	0.52	0.50	0.48	0.46	0.61	0.63	0.54	0.41
stMMR[56]	0.59	0.51	0.58	0.69	0.49	0.48	0.68	0.63	0.62	0.44
MAFN[62]	0.68	0.51	0.71	0.61	0.56	0.48	0.82	0.76	0.60	0.43
SAINT-G	0.74	0.64	0.73	0.71	0.56	0.56	0.83	0.80	0.64	0.45
SAINT-SA	**0.75**	**0.68**	**0.74**	**0.76**	**0.58**	**0.57**	**0.90**	**0.85**	**0.66**	**0.46**
Method	Normalized Mutual Information (NMI)									
	151507	151508	151509	151510	151669	151670	151671	151672	HBC	MBA
SCANPY[49]	0.21	0.21	0.27	0.22	0.16	0.16	0.24	0.23	0.52	0.45
SpaGCN[18]	0.54	0.42	0.55	0.50	0.42	0.45	0.60	0.61	0.56	0.62
DeepST[51]	0.62	0.57	0.62	0.62	0.57	0.51	0.59	0.60	0.68	0.57
SCGDL[31]	0.55	0.44	0.48	0.45	0.38	0.36	0.41	0.46	0.43	0.64
stLearn[36]	0.64	0.53	0.62	0.59	0.49	0.41	0.54	0.47	0.63	0.66
Spatial-MGCN[47]	0.74	0.60	0.68	0.67	0.58	0.56	0.72	0.75	0.69	0.71
GraphST[34]	0.64	0.54	0.64	0.64	0.59	**0.68**	0.70	0.61	0.67	0.71
stMMR[56]	0.72	0.65	0.71	0.71	0.56	0.56	0.72	0.72	0.65	0.68
MAFN[62]	0.74	0.51	0.72	0.68	0.63	0.60	0.78	0.75	0.67	0.73
SAINT-G	0.77	0.69	0.73	0.72	0.62	0.63	0.78	0.79	0.69	0.72
SAINT-SA	**0.78**	**0.71**	**0.74**	**0.73**	**0.64**	0.64	**0.84**	**0.80**	**0.70**	**0.74**

4.2 Main Performance(RQ1)

To assess the effectiveness of SAINT, we compare it against nine state-of-the-art STC(Spatial Transcriptomics Clustering) methods on three benchmark datasets: DLPFC, HBC, and MBA. Table 1 reports ARI and NMI scores; full slice results are provided in Appendix A.4. SAINT includes two variants: **SAINT-G**, which averages gene-level nucleotide embeddings, and **SAINT-SA**, which incorporates sequence-aware attention for dynamic aggregation. As shown in Table 1, SAINT-SA consistently achieves the best or second-best results across nearly all slices, improving ARI from 0.64 to 0.68 and NMI from 0.69 to 0.71 on average. On slice 151507, SAINT-SA achieves an ARI

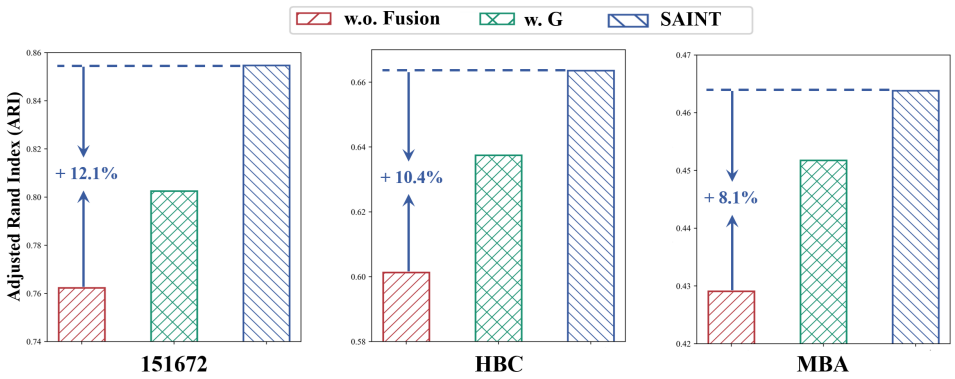

Figure 3: Ablation study of different components in SAINT.

Table 2: Clustering performance of competing spatial transcriptomics models. Bold entries indicate the best results, and underlined values denote the second-best.

Method	151507		151508		151509		151510	
	ARI	NMI	ARI	NMI	ARI	NMI	ARI	NMI
Spatial-MGCN	0.6305	0.7443	0.4622	0.6023	0.5441	0.6812	0.5158	0.6668
Spatial-MGCN+SAINT-G	0.7125	0.7594	0.5888	0.6692	0.6766	0.6840	0.6189	0.6758
Spatial-MGCN+SAINT-SA	**0.7357**	**0.7690**	**0.6476**	**0.6840**	**0.6986**	**0.7065**	**0.7060**	**0.7042**
MAFN	0.6812	0.7402	0.5134	0.5183	0.7128	0.7213	0.6121	0.6822
MAFN+SAINT-G	0.7441	0.7723	0.6414	0.6933	0.7320	0.7279	0.7150	0.7165
MAFN+SAINT-SA	**0.7471**	**0.7790**	**0.6843**	**0.7146**	**0.7426**	**0.7380**	**0.7649**	**0.7318**

of 0.75 (+10.3% vs. MAFN), and an NMI of 0.84 on 151671 (+16.2% vs. GraphST). Notably, SAINT-G also surpasses most baselines. For example, it improves ARI by 8.8% over MAFN and by 10.3% on slice 151671. These results highlight the value of integrating nucleotide-level priors, even with simple aggregation. Additionally, SAINT maintains robust performance across both homogeneous (MBA) and heterogeneous (HBC) tissue contexts.

4.3 Ablation Study (RQ2)

To evaluate the contribution of different components in our framework, we conduct an ablation study comparing three model variants: (1) **w.o. Fusion**, a baseline without sequence integration; (2) **w. SA**, a variant that uses average-pooled gene sequence embeddings; and (3) **SAINT**, our complete model that incorporates sequence-aware attention and late-stage fusion. As shown in Figure 3, SAINT consistently outperforms the reduced variants across all benchmarks. For example, in terms of ARI, it yields relative improvements of +12.1% on slice 151672, +10.4% on HBC, and +8.1% on MBA. Even the average-pooling variant (*w. G*) surpasses the no-fusion baseline in most cases, indicating that sequence representations carry biologically meaningful information that complements expression-based features. Additional NMI results are reported in Appendix A.5.

4.4 Sensitivity Analysis (RQ3)

To evaluate the robustness of SAINT under different sequence embedding dimensions, we conduct a sensitivity analysis by varying d_1 and d_2, which represent the embedding dimensions used in SAINT-G and SAINT-SA, respectively. Each is selected from [16, 32, 64, 128, 256], and experiments are performed on three representative datasets, i.e., DLPFC (151507 as an example), HBC, and MBA. The ARI results are shown in Figure 4. Specifically, on the 151507 data slice, ARI fluctuates between 0.7351 and 0.7471, with a relative deviation of only 1.63%. For HBC, ARI varies from 0.6289 to 0.6433, corresponding to a 2.30% range. On the more challenging MBA dataset, the ARI spans 0.4269 to 0.4623, yielding a slightly wider but still manageable fluctuation of 3.30%. These small variations confirm that SAINT delivers stable clustering performance without requiring extensive hyperparameter tuning. Due to space constraints, the corresponding sensitivity results for NMI are reported in Appendix A.6, which exhibit similar trends.

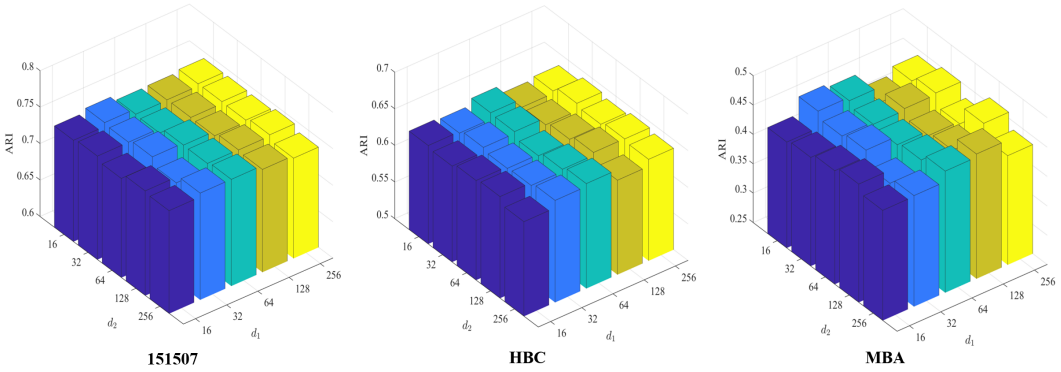

Figure 4: Parameter sensitivity analysis of the proposed SAINT on DLPFC, HBC and MBA datasets.

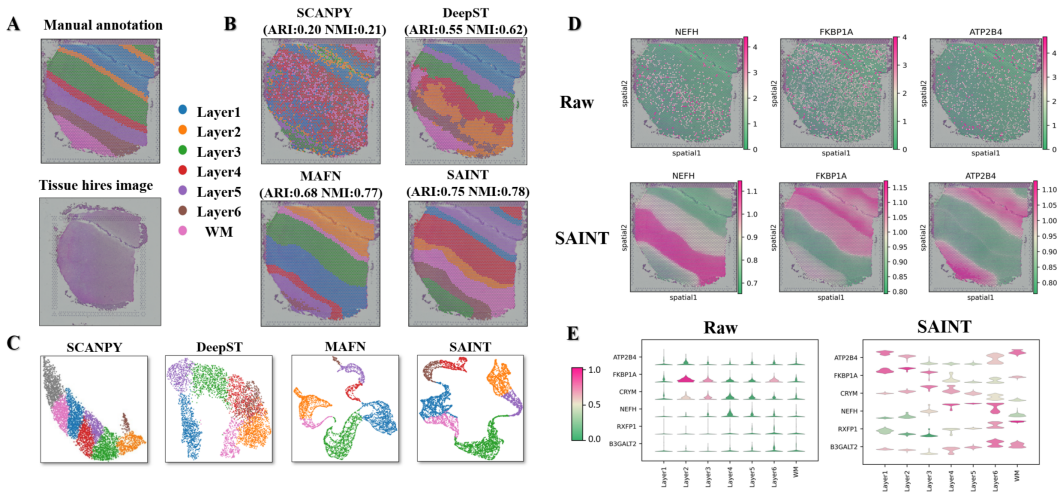

Figure 5: Case study visualizations on DLPFC slice 151507. (A) Manual tissue annotation overlaid on histological image. (B) Comparison of spatial domain identification results produced by competing methods. (C) UMAP projection of learned embeddings. (D) Spatial expression patterns of representative marker genes before and after SAINT enhancement. (E) Violin plots illustrating gene-level expression differences across identified domains.

4.5 Transferability Analysis (RQ4)

We evaluate the transferability of SAINT by integrating it into two representative backbonesSpatial-MGCN and MAFN. As shown in Table 2, both SAINT variants consistently improve clustering performance. For example, Spatial-MGCN+SAINT-G achieves an average ARI gain of +6.2%, while SAINT-SA further boosts it to +8.3%. On slice 151508, SAINT-SA raises the ARI to 0.6476, a relative improvement of +40.2% over the baseline (0.4622). Similar gains are observed with MAFN. On slice 151509, MAFN+SAINT-SA achieves an NMI of 0.7380, outperforming MAFN+SAINT-G by 1.67% and the vanilla MAFN by 2.93%. These results demonstrate that SAINT functions as a transferable and effective plug-in module. Meanwhile, this generality arises from the models modular design, where the sequence-aware encoder and cross-modal fusion can be seamlessly attached to existing spatial frameworks without retraining from scratch. Such adaptability highlights SAINTs potential as a unifying layer for future multi-modal spatial transcriptomics methods. Additional analysis is provided in Appendix A.7.

4.6 Case Study (RQ5)

We conduct a case study on the DLPFC 151507 slice to assess the interpretability and biological relevance of SAINT. As shown in Fig. 5(C), SAINT more accurately recovers cortical layer bound-

aries compared to competing methods. For instance, it captures the transition between Layer 5 and Layer 6 in the lower-right region that is oversmoothed by DeepST and fragmented in SCANPY. To further validate biological plausibility, Fig. 5(D) shows spatial expression maps of representative marker genes (*NEFH*, *FKBP1A*, *ATP2B4*). SAINT enhances spatial coherence and alignment with anatomical structures. Violin plots in Fig. 5(E) also demonstrate sharper inter-layer specificity and lower intra-layer variance. Notably, *ATP2B4* and *CRYM* show clearer separation across domains, while *B3GALT2* displays improved compactness within WM. These results highlight SAINTs ability to generate biologically meaningful and anatomically consistent representations. Additionally, this improvement mainly stems from the incorporation of nucleotide-informed embeddings, which enable SAINT to better distinguish functionally divergent genes with similar expression levels and thus refine boundary delineation. Such sequence-aware representations provide a mechanistic link between gene regulation patterns and observed spatial organization, further supporting the biological interpretability of the model. Additional case studies are provided in Appendix A.8.

5 Conclusion

In this work, we present SAINT, a sequence-aware multi-modal framework for spatial transcriptomics clustering. Unlike previous methods that utilize only gene expression and spatial proximity, SAINT introduces gene-level nucleotide embeddings to capture additional biological priors. To enable this integration, we construct spotgenesequence mappings across three benchmark datasets and encode sequences using a pretrained genomic language model. We design an attention-based aggregation module to summarize sequence features at the spot level, and employ a late fusion strategy to combine them with spatial-expression embeddings. Extensive experiments across multiple datasets demonstrate that SAINT consistently improves clustering accuracy.

Acknowledgments

This work was supported by the National Key R & D Program of China No.2022YFA1005101, the National Natural Science Foundation of China (project No.61872371, 62032024, 62325604, 62441618, 62276271, 62506371).

References

[1] Alma Andersson, Joseph Bergenstråhle, Michaela Asp, Ludvig Bergenstråhle, Aleksandra Jurek, José Fernández Navarro, and Joakim Lundeberg. Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography. *Communications biology*, 3(1):565, 2020.

[2] 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] Ricard Argelaguet, Britta Velten, Damien Arnol, Sascha Dietrich, Thorsten Zenz, John C Marioni, Florian Buettner, Wolfgang Huber, and Oliver Stegle. Multi-omics factor analysisa framework for unsupervised integration of multi-omics data sets. *Molecular systems biology*, 14(6):e8124, 2018.

[4] Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R Ledsam, Agnieszka Grabska-Barwinska, Kyle R Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods*, 18(10):1196–1203, 2021.

[5] Joseph Bergenstråhle, Ludvig Larsson, and Joakim Lundeberg. Seamless integration of image and molecular analysis for spatial transcriptomics workflows. *BMC genomics*, 21:1–7, 2020.

[6] Zhi-Jie Cao and Ge Gao. Multi-omics integration and regulatory inference for unpaired single-cell data with a graph-linked unified embedding framework. *bioRxiv*, pages 2021–08, 2021.

[7] Zhi-Jie Cao and Ge Gao. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nature Biotechnology*, 40(10):1458–1466, 2022.

[8] Zhi-Jie Cao and Ge Gao. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nature Biotechnology*, 40(10):1458–1466, 2022.

[9] Haotian Cui, Chloe Wang, Hassaan Maan, Nan Duan, and Bo Wang. scformer: a universal representation learning approach for single-cell data using transformers. *bioRxiv*, pages 2022–11, 2022.

[10] Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Bernardo P de Almeida, Hassan Sirelkhatim, et al. Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, 22(2):287–297, 2025.

[11] Shijie Deng, Jie Wen, Chengliang Liu, Ke Yan, Gehui Xu, and Yong Xu. Projective incomplete multi-view clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[13] Kangning Dong and Shihua Zhang. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nature communications*, 13(1):1739, 2022.

[14] Ruben Dries, Qian Zhu, Rui Dong, Chee-Huat Linus Eng, Huipeng Li, Kan Liu, Yuntian Fu, Tianxiao Zhao, Arpan Sarkar, Feng Bao, et al. Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome biology*, 22:1–31, 2021.

[15] Jonathan Frazer, Pascal Notin, Mafalda Dias, Aidan Gomez, Joseph K Min, Kelly Brock, Yarin Gal, and Debora S Marks. Disease variant prediction with deep generative models of evolutionary data. *Nature*, 599(7883):91–95, 2021.

[16] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323. JMLR Workshop and Conference Proceedings, 2011.

[17] Kaveh Hassani and Amir H. Khasahmadi. Contrastive multi-view representation learning on graphs. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119, pages 4116–4126, 2020.

[18] Jian Hu, Xiangjie Li, Kyle Coleman, Amelia Schroeder, Nan Ma, David J Irwin, Edward B Lee, Russell T Shinohara, and Mingyao Li. Spagcn: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nature methods*, 18(11):1342–1351, 2021.

[19] Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37(15):2112–2120, 2021.

[20] David R Kelley, Yakir A Reshef, Maxwell Bileschi, David Belanger, Cory Y McLean, and Jasper Snoek. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome research*, 28(5):739–750, 2018.

[21] David R Kelley, Jasper Snoek, and John L Rinn. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome research*, 26(7):990–999, 2016.

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[23] Zeger F Knops, JB Antoine Maintz, Max A Viergever, and Josien PW Pluim. Normalized mutual information based registration using k-means clustering and shading correction. *Medical image analysis*, 10(3):432–439, 2006.

[24] Xuelong Li, Hongyuan Zhang, and Rui Zhang. Adaptive graph auto-encoder for general data clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9725–9732, 2021.

[25] Ke Liang, Yue Liu, Sihang Zhou, Wenxuan Tu, Yi Wen, Xihong Yang, Xiangjun Dong, and Xinwang Liu. Knowledge graph contrastive learning based on relation-symmetrical structure. *IEEE Transactions on Knowledge and Data Engineering*, 36(1):226–238, 2023.

[26] Hanqing Liu, Jingtian Zhou, Wei Tian, Chongyuan Luo, Anna Bartlett, Andrew Aldridge, Jacinta Lucero, Julia K Osteen, Joseph R Nery, Huaming Chen, et al. Dna methylation atlas of the mouse brain at single-cell resolution. *Nature*, 598(7879):120–128, 2021.

[27] Jie Liu, Yuanhao Huang, Ritambhara Singh, Jean-Philippe Vert, and William Stafford Noble. Jointly embedding multiple single-cell omics measurements. In *Algorithms in bioinformatics:... International Workshop, WABI..., proceedings. WABI (Workshop)*, volume 143, page 10, 2019.

[28] Jie Liu, Yuanhao Huang, Ritambhara Singh, Jean-Philippe Vert, and William Stafford Noble. Jointly embedding multiple single-cell omics measurements. In *Algorithms in bioinformatics:... International Workshop, WABI..., proceedings. WABI (Workshop)*, volume 143, page 10, 2019.

[29] Jiyuan Liu, Xinwang Liu, Yuexiang Yang, Qing Liao, and Yuanqing Xia. Contrastive multi-view kernel learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):9552–9566, 2023.

[30] Meng Liu, Yue Liu, Ke Liang, Wenxuan Tu, Siwei Wang, Sihang Zhou, and Xinwang Liu. Deep temporal graph clustering. In *The 12th International Conference on Learning Representations*, 2024.

[31] Teng Liu, Zhao-Yu Fang, Xin Li, Li-Ning Zhang, Dong-Sheng Cao, and Ming-Zhu Yin. Graph deep learning enabled spatial domains identification for spatial transcriptomics. *Briefings in Bioinformatics*, 24(3):bbad146, 2023.

[32] Yang Liu, Mingyu Yang, Yanxiang Deng, Graham Su, Archibald Enninful, Cindy C Guo, Toma Tebaldi, Di Zhang, Dongjoo Kim, Zhiliang Bai, et al. High-spatial-resolution multi-omics sequencing via deterministic barcoding in tissue. *Cell*, 183(6):1665–1681, 2020.

[33] Yue Liu, Wenxuan Tu, Sihang Zhou, Xinwang Liu, Linxuan Song, Xihong Yang, and En Zhu. Deep graph clustering via dual correlation reduction. In *Proc. of AAAI*, volume 36, pages 7603–7611, 2022.

[34] Yahui Long, Kok Siong Ang, Mengwei Li, Kian Long Kelvin Chong, Raman Sethi, Chengwei Zhong, Hang Xu, Zhiwei Ong, Karishma Sachaphibulkij, Ao Chen, et al. Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with graphst. *Nature Communications*, 14(1):1155, 2023.

[35] 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.

[36] Duy Pham, Xiao Tan, Jun Xu, Laura F Grice, Pui Yeng Lam, Arti Raghubar, Jana Vukovic, Marc J Ruitenberg, and Quan Nguyen. stlearn: integrating spatial location, tissue morphology and gene expression to find cell types, cell-cell interactions and spatial trajectories within undissociated tissues. *biorxiv*, pages 2020–05, 2020.

[37] William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

[38] Philipp Rentzsch, Daniela Witten, Gregory M Cooper, Jay Shendure, and Martin Kircher. Cadd: predicting the deleteriousness of variants throughout the human genome. *Nucleic acids research*, 47(D1):D886–D894, 2019.

[39] Samuel G Rodriques, Robert R Stickels, Aleksandrina Goeva, Carly A Martin, Evan Murray, Charles R Vanderburg, Joshua Welch, Linlin M Chen, Fei Chen, and Evan Z Macosko. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, 363(6434):1463–1467, 2019.

[40] Xin Shao, Haihong Yang, Xiang Zhuang, Jie Liao, Penghui Yang, Junyun Cheng, Xiaoyan Lu, Huajun Chen, and Xiaohui Fan. scdeepsort: a pre-trained cell-type annotation method for single-cell transcriptomics using deep learning with a weighted graph neural network. *Nucleic acids research*, 49(21):e122–e122, 2021.

[41] Ritambhara Singh, Jack Lanchantin, Gabriel Robins, and Yanjun Qi. Deepchrome: deep-learning for predicting gene expression from histone modifications. *Bioinformatics*, 32(17):i639–i648, 2016.

[42] Ritambhara Singh, Jack Lanchantin, Arshdeep Sekhon, and Yanjun Qi. Attend and predict: Understanding gene regulation by selective attention on chromatin. *Advances in neural information processing systems*, 30, 2017.

[43] 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.

[44] Yuan Sun, Yang Qin, Yongxiang Li, Dezhong Peng, Xi Peng, and Peng Hu. Robust multi-view clustering with noisy correspondence. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[45] Valentine Svensson, Sarah A Teichmann, and Oliver Stegle. Spatialde: identification of spatially variable genes. *Nature methods*, 15(5):343–346, 2018.

[46] Sanja Vickovic, Gökcen Eraslan, Fredrik Salmén, Johanna Klughammer, Linnea Stenbeck, Denis Schapiro, Tarmo Äijö, Richard Bonneau, Ludvig Bergenstråhle, José Fernandéz Navarro, et al. High-definition spatial transcriptomics for in situ tissue profiling. *Nature methods*, 16(10):987–990, 2019.

[47] Bo Wang, Jiawei Luo, Ying Liu, Wanwan Shi, Zehao Xiong, Cong Shen, and Yahui Long. Spatial-mgcn: a novel multi-view graph convolutional network for identifying spatial domains with attention mechanism. *Briefings in Bioinformatics*, 24(5):bbad262, 2023.

[48] Siwei Wang, Xinwang Liu, Suyuan Liu, Jiaqi Jin, Wenxuan Tu, Xinzhong Zhu, and En Zhu. Align then fusion: Generalized large-scale multi-view clustering with anchor matching correspondences. *Advances in Neural Information Processing Systems*, 35:5882–5895, 2022.

[49] F Alexander Wolf, Philipp Angerer, and Fabian J Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19:1–5, 2018.

[50] 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.

[51] Chang Xu, Xiyun Jin, Songren Wei, Pingping Wang, Meng Luo, Zhaochun Xu, Wenyi Yang, Yideng Cai, Lixing Xiao, Xiaoyu Lin, et al. Deepst: identifying spatial domains in spatial transcriptomics by deep learning. *Nucleic Acids Research*, 50(22):e131–e131, 2022.

[52] Mouxing Yang, Yunfan Li, Peng Hu, Jinfeng Bai, Jiancheng Lv, and Xi Peng. Robust multi-view clustering with incomplete information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):1055–1069, 2022.

[53] Yunsheng You, Tianlong Chen, Ximing Liu, Yang Ji, Yan Gao, and Jure Leskovec. Graph contrastive learning with augmentations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 5811–5823, 2020.

[54] Xin Yuan, Yanran Ma, Ruitian Gao, Shuya Cui, Yifan Wang, Botao Fa, Shiyang Ma, Ting Wei, Shuangge Ma, and Zhangsheng Yu. Heartsvg: a fast and accurate method for identifying spatially variable genes in large-scale spatial transcriptomics. *Nature Communications*, 15(1):5700, 2024.

[55] Zhiyuan Yuan, Yisi Li, Minglei Shi, Fan Yang, Juntao Gao, Jianhua Yao, and Michael Q Zhang. Sotip is a versatile method for microenvironment modeling with spatial omics data. *Nature Communications*, 13(1):7330, 2022.

[56] Daoliang Zhang, Na Yu, Zhiyuan Yuan, Wenrui Li, Xue Sun, Qi Zou, Xiangyu Li, Zhiping Liu, Wei Zhang, and Rui Gao. stmmr: accurate and robust spatial domain identification from spatially resolved transcriptomics with multimodal feature representation. *GigaScience*, 13:giae089, 2024.

[57] Edward Zhao, Matthew R Stone, Xing Ren, Jamie Guenthoer, Kimberly S Smythe, Thomas Pulliam, Stephen R Williams, Cedric R Uytingco, Sarah EB Taylor, Paul Nghiem, et al. Spatial transcriptomics at subspot resolution with bayesspace. *Nature biotechnology*, 39(11):1375–1384, 2021.

[58] Zhongying Zhao, Zhan Yang, Chao Li, Qingtian Zeng, Weili Guan, and Mengchu Zhou. Dual feature interaction-based graph convolutional network. *IEEE Transactions on Knowledge and Data Engineering*, 35(9):9019–9030, 2023.

[59] Jian Zhou, Chandra L Theesfeld, Kevin Yao, Kathleen M Chen, Aaron K Wong, and Olga G Troyanskaya. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nature genetics*, 50(8):1171–1179, 2018.

[60] Jian Zhou and Olga G Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature methods*, 12(10):931–934, 2015.

[61] Jiaqiang Zhu, Shiquan Sun, and Xiang Zhou. Spark-x: non-parametric modeling enables scalable and robust detection of spatial expression patterns for large spatial transcriptomic studies. *Genome biology*, 22(1):184, 2021.

[62] Yanran Zhu, Xiao He, Chang Tang, Xinwang Liu, Yuanyuan Liu, and Kunlun He. Multi-view adaptive fusion network for spatially resolved transcriptomics data clustering. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

A Appendix

A.1 Limitations

While SAINT demonstrates strong performance and broad applicability, several limitations remain. First, the reliance on pretrained genomic models, such as the Nucleotide Transformer, may introduce biases from the training corpus, which predominantly includes well-annotated genes and reference sequences. This could limit generalizability to under-characterized genes or non-model organisms. Second, our current framework treats all expressed genes equally during sequence aggregation after HVG filtering. Although attention mechanisms mitigate some noise, rare but biologically significant genes may still be down-weighted or omitted. Future work may explore more context-aware gene selection strategies.

A.2 Related Work

In this section, we summarize related work along three aspects. First, we review spatial transcriptomics clustering methods through the lens of multi-view learning, highlighting how expression and spatial features have been combined. Second, we introduce genomic language models that extract meaningful representations from DNA sequences. Finally, we discuss recent efforts toward integrating sequence information into spatial clustering, which remains an underexplored yet promising direction.

MVC in Spatial Transcriptomics Data. Clustering is a central task in spatial transcriptomics (ST), aiming to delineate spatially coherent tissue domains that reflect underlying biological structure and organization. Existing methods for ST clustering can be broadly grouped into three paradigms. (1) Expression-based approaches [5, 14] perform unsupervised clustering purely based on transcriptional profiles, typically using K-means or community detection on PCA-reduced features. (2) Spatially regularized models [57, 1] augment these strategies with spatial smoothing, either through distance-based penalties or spatial Laplacians to encourage neighboring spots to share cluster assignments. (3) In contrast, graph neural network (GNN)-based methods [13, 18, 55, 62] explicitly model spatial structure via graphs and propagate information using neural message passing. Among these, GNNs have emerged as particularly effective due to their flexibility in modeling complex tissue architectures and incorporating multimodal inputs. STAGATE [13] introduces a graph attention autoencoder that jointly learns from spatial adjacency and gene expression similarity, achieving state-of-the-art clustering performance across multiple platforms. SpaGCN [18], in a similar spirit, integrates histological information into the spatial graph, enabling more anatomically consistent clustering via graph convolutions. Meanwhile, SOTIP [55] formulates a spatial multi-task framework that captures micro-environmental structure and intercellular context through local neighborhood graphs. Moreover, MAFN [62] leverages a multi-view fusion strategy, adaptively combining representations from both spatial graphs and gene-gene similarity graphs to enhance robustness and tissue-domain separation. While these methods have shown strong empirical performance, they are all fundamentally built upon expression-derived features and spatial graphs. They often overlook additional layers of biological prior knowledge, i.e., such as the regulatory or structural information embedded in gene sequences, which may influence spatial gene expression patterns but remain underexplored in current models.

Genomic Language Modeling for Sequence Representations. While spatial transcriptomics clustering has traditionally focused on expression-level and spatial features, another promising source of biological prior lies in the gene sequences themselves. DNA sequences encode regulatory, structural, and evolutionary signals that can influence gene activity and co-expression. Recent years have seen a surge in the development of genomic language models (GLMs), inspired by advances in self-supervised learning from natural language processing. These models treat nucleotide sequences as a form of structured text and learn contextualized embeddings using masked language modeling or next-token prediction objectives. For example, DNABERT [19] adapts the BERT architecture [12] to k-mer tokenized DNA sequences, showing strong performance on transcription factor binding prediction and enhancer classification. Building on this foundation, Nucleotide Transformer [10] scales the model size and diversity of training data to cover multiple species and larger sequence contexts, achieving robust generalization across downstream genomics tasks. Unlike traditional handcrafted motif features or one-hot encodings, GLMs capture long-range dependencies, compositional signals, and shared patterns across different genomic loci. These properties make them attractive for repre-

sentation learning in biological applications, especially when labeled data are limited. Moreover, the pretrained embeddings can be transferred and fine-tuned to serve various predictive tasks, such as variant effect prediction [4, 60, 38, 21], epigenomic state modeling [41, 59, 20, 42], and multi-omics integration [15, 27, 7, 6]. Despite these advances, the use of sequence-based embeddings in spatial transcriptomics remains largely unexplored. Existing ST models rarely utilize the nucleotide sequences of expressed genes, thereby missing the opportunity to incorporate regulatory priors that may underlie the observed expression patterns. A few recent efforts have begun exploring protein-level embeddings for single-cell data [32], but nucleotide-level integration in spatial contexts has not been systematically studied.

Sequence-Augmented Spatial Transcriptomics Clustering. Recent advances in genomic language modeling have enabled the extraction of rich sequence-level representations from raw DNA. While these sequence-level embeddings have shown utility in a range of genomics tasks, their incorporation into ST clustering remains largely unexplored. This subsection reviews recent advances in augmenting ST models with sequence-derived features. A straightforward approach is late fusion, where sequence features are concatenated with spatial representations prior to downstream prediction. This method is modular and simple to implement, and has been widely applied in multi-modal omics integration [3, 28]. However, naive concatenation may fail to capture interactions between modalities and cannot dynamically adjust to varying gene importance across spatial contexts. To mitigate this, attention-based mechanisms have been introduced to assign adaptive weights to gene embeddings based on their contextual relevance. In spatial omics, attention modules have been used to integrate expression profiles with histological context or neighborhood structure [18, 8]. Soft attention mechanisms have also been employed to aggregate gene-level embeddings within each spot, yielding compact and informative representations that emphasize contextually relevant sequences. Transformer-based architectures offer a more flexible alternative by modeling interactions across gene sequences through self-attention. Though effective in natural language and genomics [9, 19], such models demand substantial training resources and are not yet widely used in spatial transcriptomics. Lightweight alternatives include expression-weighted averaging, which gives more influence to highly expressed genes in embedding aggregation [40]. Filtering by HVGs provides another practical benefit, reducing redundancy and focusing on the most informative sequence signals [2]. Despite these developments, existing ST clustering methods rarely incorporate sequence-level priors, leaving untapped the potential of regulatory DNA features in spatial organization. To bridge this gap, we introduce a novel framework that integrates gene sequence embeddings into ST clustering.

A.3 Experiment Setting

Experiment settings are introduced from four aspects, i.e., datasets, implementation details, compared baselines and evaluation metrics.

Datasets. We conduct experiments on three benchmark datasets commonly used in spatial transcriptomics clustering:

We evaluate our method on three publicly available and widely used spatial transcriptomics datasets, spanning both human and mouse tissues. These datasets provide diverse anatomical and pathological contexts for robust model evaluation.

LIBD Human Dorsolateral Prefrontal Cortex (DLPFC) Dataset. The DLPFC dataset, curated by the LIBD research group [35], provides high-resolution spatial transcriptomic profiles from post-mortem human brain tissue, generated using the 10x Genomics Visium platform. It comprises 12 sagittal tissue sections, each covering approximately 3,6004,000 barcoded spots with over 33,000 protein-coding genes profiled per section. These slices encompass the full laminar structure of the neocortex (L1L6) as well as the white matter beneath. The dataset includes expert-annotated spatial domains based on histological examination, facilitating benchmarking of computational clustering models.

10x Visium Human Breast Cancer (HBC) Dataset. The HBC dataset [50] contains spatially resolved gene expression measurements from human breast tumor sections. Each spot captures the expression of approximately 36,000 genes across diverse histological structures. The dataset is annotated with 20 spatial domains, encompassing pre-invasive ductal carcinoma in situ (DCIS), lobular carcinoma in situ (LCIS), invasive ductal carcinoma (IDC), tumor stroma, immune infiltration zones,

Table 3: Statistics of the constructed sequence-augmented datasets. Each dataset corresponds to a tissue section or benchmark. "#B. class" denotes the number of unique barcodes (spots), "#G. class" indicates the number of distinct genes with available nucleotide sequences, and "Num" refers to the total number of `(barcode, gene, sequence)` triplets after filtering.

Dataset	#B. class	#G. class	Num	Dataset	#B. class	#G. class	Num
151507	721	237	24,968	151672	578	319	15,551
151508	849	190	11,940	151673	465	423	16,855
151509	697	249	12,489	151674	373	561	20,542
151510	737	233	12,092	151675	572	386	14,486
151669	563	341	16,034	151676	517	382	14,606
151670	596	316	15,089	HBC	182	250	32,734
151671	531	330	15,999	MBA	182	300	40,729

and adjacent normal tissue. This rich spatial annotation supports fine-grained exploration of tumor heterogeneity and microenvironmental interactions.

Mouse Brain Anterior Tissue (MBA) Dataset. The MBA dataset [26] includes gene expression profiles from a sagittal-anterior section of the adult mouse brain, obtained using the Illumina NovaSeq 6000 platform. It contains 2,695 spatial spots and over 32,000 genes, capturing molecular patterns across anatomical regions such as the cortex, hippocampus, and basal forebrain. This dataset enables detailed investigation of spatial gene regulation and inter-regional signaling in a mammalian neural context. It is publicly available through the 10x Genomics data repository.

Sequence-Augmented Triplet Construction.

To integrate nucleotide-level information into spatial transcriptomics clustering, we construct a structured sequence-augmented dataset for each benchmark. For every expressed gene within a spatial spot (barcode), we retrieve its corresponding reference DNA sequence from the NCBI nucleotide database using standardized gene identifiers. Each data entry is formatted as a `(barcode, gene, sequence)` triplet. To ensure biological relevance and avoid noisy or sparse entries, we remove barcodes expressing fewer than 10 genes with valid sequences. This filtering step ensures that each spatial spot contributes sufficient nucleotide-level context for representation learning. Table 3 summarizes the resulting datasets. For each tissue slice or benchmark, we report the number of unique barcodes (#B. class), the number of distinct genes with matched sequences (#G. class), and the total number of triplets (Num). These triplets serve as the input for sequence encoder modules and enable the downstream modeling of spatial domains with nucleotide-informed priors.

Implementation Settings. All models are implemented in PyTorch 2.0.1 and trained using the Adam optimizer [22] on a workstation with an Intel Core i9-9900K CPU, 64GB RAM, and an NVIDIA RTX 3090 Ti GPU. Following MAFN [62], we adopt consistent training settings and learning rate schedules. For sequence embedding, we evaluate two aggregation variants:

- **SAINT-G**: gene sequence embeddings are averaged without expression weighting.
- **SAINT-SA**: expression-aware attention pooling is applied to gene embeddings per spot.

The projection dimensions d_1 (for SAINT-G) and d_2 (for SAINT-SA) are selected from $\{16, 32, 64, 128, 256\}$.

Compared Baselines. To comprehensively evaluate the effectiveness of SAINT, we compare it against a wide range of state-of-the-art spatial transcriptomics clustering methods, including both traditional approaches and recent GNN-based frameworks.

- **SCANPY** [49] is a widely used single-cell analysis toolkit that performs PCA-based dimensionality reduction and graphbased clustering on highly variable genes without incorporating spatial context.
- **SpaGCN** [18] constructs a spatial graph from tissue coordinates and employs graph convolution to jointly model gene expression and spatial dependencies for anatomically coherent clustering.

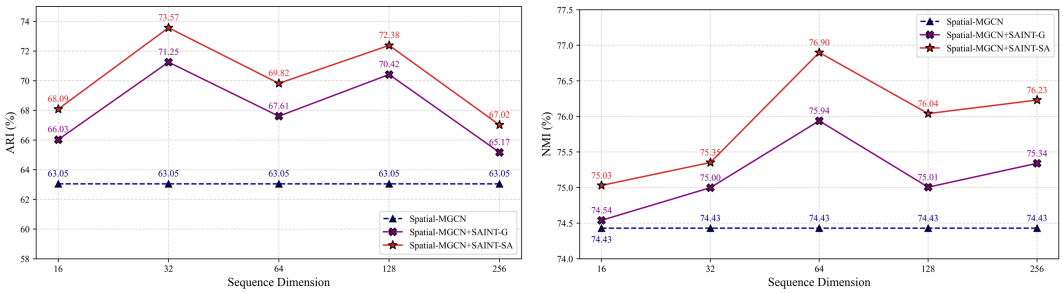

Figure 6: Transfer study of SAINT variants on the DLPFC 151507 slice under the Spatial-MGCN backbone.

- **DeepST** [51] learns spatially informed spot embeddings by jointly modeling transcriptional profiles and physical locations, enabling accurate tissue domain segmentation.
- **SCGDL** [31] utilizes a residual gated graph neural network augmented with a deep graph information maximization module to capture hierarchical and long-range dependencies in spatial transcriptomic graphs.
- **stLearn** [36] integrates histological morphology with gene expression and spatial coordinates through a self-supervised learning pipeline that regularizes clustering via spatially constrained random fields.
- **stMMR** [56] applies Markov random field regularization to smooth clustering labels and enforce spatial coherence in noisy or irregular tissue regions.
- **GraphST** [34] proposes a spatially guided contrastive learning framework that enhances intra-domain cohesion and inter-domain separation across spot embeddings.
- **Spatial-MGCN** [47] fuses multiple spatial and expression graphs using multi-view graph convolutions and an attention mechanism to capture heterogeneous tissue patterns.
- **MAFN** [62] employs a multi-branch graph convolutional architecture with adaptive late-fusion, allowing flexible integration of spatial and gene-gene similarity representations for robust clustering.

Evaluation Metrics. We adopt two widely used metrics for evaluating clustering performance:

Adjusted Rand Index (ARI) [37]: Measures similarity between predicted and ground-truth cluster assignments, adjusted for chance. Given a contingency table between true labels and predicted clusters, ARI is computed as:

$$\text{ARI} = \frac{RI_{\text{obs}} - RI_{\text{rand}}}{\max(RI) - RI_{\text{rand}}} \tag{14}$$

where RI_{obs} is the observed Rand index measuring the similarity between predicted and ground-truth clusterings, RI_{rand} denotes the expected index under random labeling, and $\max(RI)$ is the maximum attainable value.

Normalized Mutual Information (NMI) [23]: Quantifies mutual dependence between true and predicted clusters. Defined as follows.

$$\text{NMI} = \frac{2 \cdot I(Y; \hat{Y})}{H(Y) + H(\hat{Y})}, \tag{15}$$

where $I(Y; \hat{Y})$ is the mutual information between true labels Y and predicted labels $\hat{Y}$, and $H(\cdot)$ denotes entropy. Higher ARI and NMI values indicate better clustering alignment with biological ground truth.

A.4 Extended Analysis for Main Performance (RQ1)

To supplement the main performance discussion in Section 4.2, we provide a detailed analysis of SAINT across all individual slices from the DLPFC, HBC, and MBA datasets, as reported in Table 4.

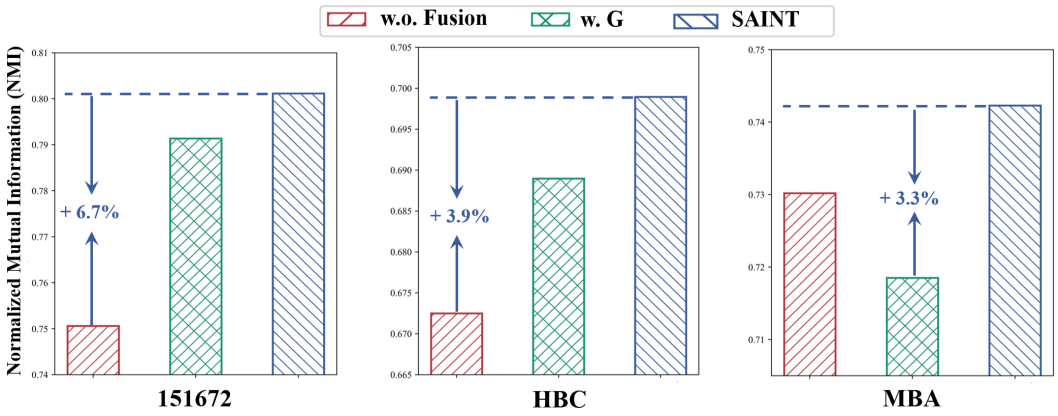

Figure 7: Ablation study of different components in SAINT.

Table 4: Full clustering performance of competing spatial transcriptomics models. Bold entries indicate the best results, and underlined values denote the second-best.

Method	Adjusted Rand Index (ARI)													
	151507	151508	151509	151510	151669	151670	151671	151672	151673	151674	151675	151676	HBC	MBA
SCANPY	0.20	0.15	0.19	0.14	0.10	0.09	0.12	0.12	0.20	0.22	0.23	0.22	0.49	0.23
SpaGCN	0.43	0.33	0.41	0.37	0.23	0.35	0.51	0.53	0.40	0.31	0.33	0.28	0.56	0.34
DeepST	0.55	0.42	0.43	0.50	0.44	0.33	0.52	0.48	0.54	0.55	0.53	0.56	0.53	0.25
SCGDL	0.49	0.34	0.32	0.31	0.24	0.26	0.31	0.34	0.33	0.27	0.30	0.29	0.35	0.26
stLearn	0.49	0.31	0.45	0.44	0.32	0.23	0.39	0.34	0.30	0.38	0.38	0.40	0.55	0.38
Spatial-MGCN	0.63	0.46	0.54	0.51	0.39	0.35	0.60	0.77	0.61	**0.60**	0.54	0.57	0.64	0.42
GraphST	0.48	0.49	0.52	0.50	0.48	0.46	0.61	0.63	**0.63**	0.43	0.55	0.55	0.54	0.41
stMMR	0.59	0.51	0.58	0.69	0.49	0.48	0.68	0.63	0.60	0.51	0.57	0.55	0.62	0.44
MAFN	0.68	0.51	0.71	0.61	0.56	0.48	0.82	0.76	0.57	0.50	0.46	0.53	0.60	0.43
SAINT-G	0.74	0.64	0.73	0.71	0.56	0.56	0.83	0.80	0.61	0.56	0.57	0.58	0.64	0.45
SAINT-SA	**0.75**	**0.68**	**0.74**	**0.76**	**0.58**	**0.57**	**0.90**	**0.85**	0.62	0.57	**0.58**	**0.60**	**0.66**	**0.46**
Method	Normalized Mutual Information (NMI)													
	151507	151508	151509	151510	151669	151670	151671	151672	151673	151674	151675	151676	HBC	MBA
SCANPY	0.21	0.21	0.27	0.22	0.16	0.16	0.24	0.23	0.29	0.31	0.32	0.31	0.52	0.45
SpaGCN	0.54	0.42	0.55	0.50	0.42	0.45	0.60	0.61	0.55	0.46	0.46	0.46	0.56	0.62
DeepST	0.62	0.57	0.62	0.62	0.57	0.51	0.59	0.60	0.69	0.69	0.66	0.68	0.68	0.57
SCGDL	0.55	0.44	0.48	0.45	0.38	0.36	0.41	0.46	0.42	0.38	0.41	0.42	0.43	0.64
stLearn	0.64	0.53	0.62	0.59	0.49	0.41	0.54	0.47	0.49	0.54	0.56	0.56	0.63	0.66
Spatial-MGCN	0.74	0.60	0.68	0.67	0.58	0.56	0.72	0.75	0.68	**0.69**	0.67	0.67	0.69	0.71
GraphST	0.64	0.54	0.64	0.64	0.59	**0.68**	0.70	0.61	**0.74**	0.61	0.62	0.66	0.67	0.71
stMMR	0.72	0.65	0.71	0.71	0.56	0.56	0.72	0.72	0.68	0.62	0.66	0.66	0.65	0.68
MAFN	0.74	0.51	0.72	0.68	0.63	0.60	0.78	0.75	0.67	0.62	0.60	0.67	0.67	0.73
SAINT-G	0.77	0.69	0.73	0.72	0.62	0.63	0.78	0.79	0.68	0.65	0.66	0.68	0.69	0.72
SAINT-SA	**0.78**	**0.71**	**0.74**	**0.73**	**0.64**	0.64	**0.84**	**0.80**	0.69	0.66	**0.67**	**0.69**	**0.70**	**0.74**

Overall Trends. SAINT-SA consistently achieves either the best or second-best performance across the vast majority of the 12 DLPFC slices. On average, it improves ARI from 0.64 (Spatial-MGCN) to 0.68 and NMI from 0.69 to 0.71. Notably, SAINT-Gdespite its simpler averaging-based sequence integrationalready outperforms all baselines in several slices, demonstrating the standalone benefits of incorporating gene-level nucleotide priors.

DLPFC Dataset. On DLPFC slices such as 151507, 151508, and 151509, SAINT-SA achieves ARI scores of 0.75, 0.68, and 0.74 respectivelyeach being the highest among all compared methods. Particularly, SAINT-G and SAINT-SA both outperform MAFN and GraphST by large margins. For instance, on slice 151671, SAINT-SA yields an NMI of 0.84, a +16.2% improvement over GraphST (0.72) and even surpasses MAFN by +7.7%. We also observe stable improvements on more challenging slices such as 151669 and 151670, where most baselines underperform (e.g., SCANPY < 0.1 ARI). SAINT-SA boosts ARI to 0.58 and 0.57, respectively, offering clear gains in low-signal scenarios.

HBC Dataset. The Human Breast Cancer (HBC) dataset features heterogeneous tumor microenvironments. Here, SAINT-G and SAINT-SA again dominate, reaching ARI scores of 0.64 and 0.66, and NMI scores of 0.70 and 0.70. These surpass MAFN by +3.1% ARI and +3.7% NMI. Compared to SpaGCN and DeepST, SAINT provides more spatially coherent and functionally aligned clustering, as further confirmed in qualitative case studies (see Appendix A.8).

MBA Dataset. On the more homogeneous mouse brain anterior (MBA) tissue, SAINT maintains strong performance, with ARI and NMI reaching 0.46 and 0.74. These results demonstrate the

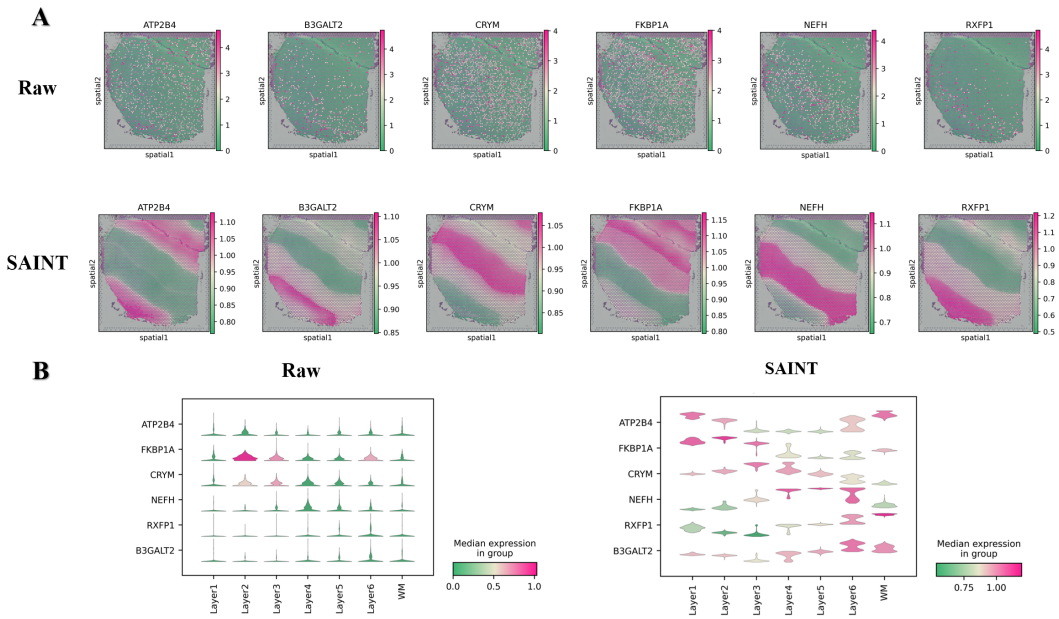

Figure 8: (A) Spatial expression of six marker genes before (Raw) and after SAINT enhancement. (B) Violin plots comparing layer-specific expression distributions of these genes in Raw and SAINT outputs.

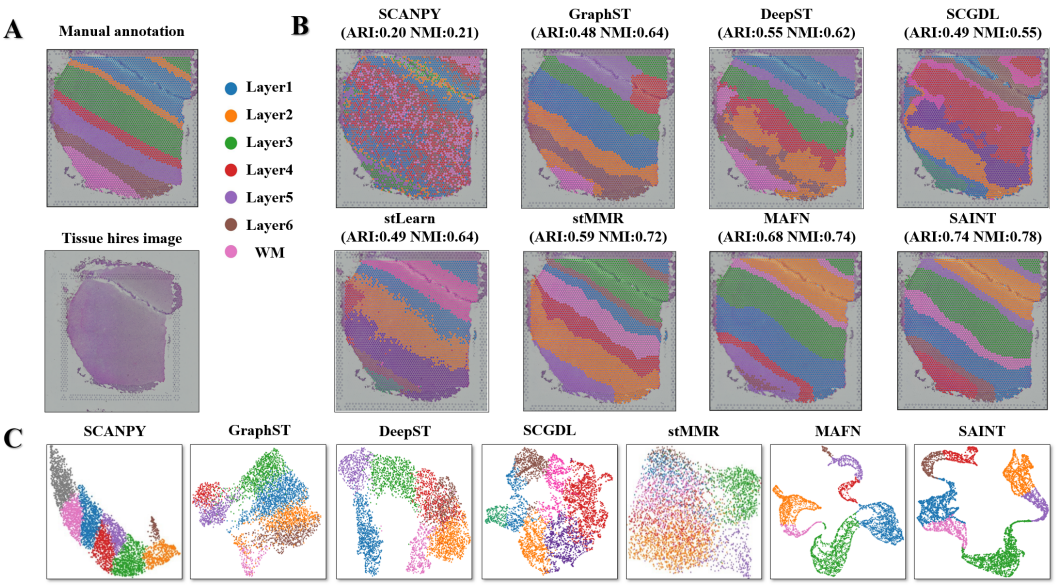

Figure 9: Case study of the proposed SAINT on DLPFC(151507) dataset.

generalizability of SAINT across diverse spatial contexts. The minimal drop in performance from HBC to MBA suggests that the framework is robust to differences in tissue type, gene expression scale, and biological variability.

Overall, these improvements reflect the effectiveness of nucleotide-informed representations in enhancing spatial transcriptomics clustering.

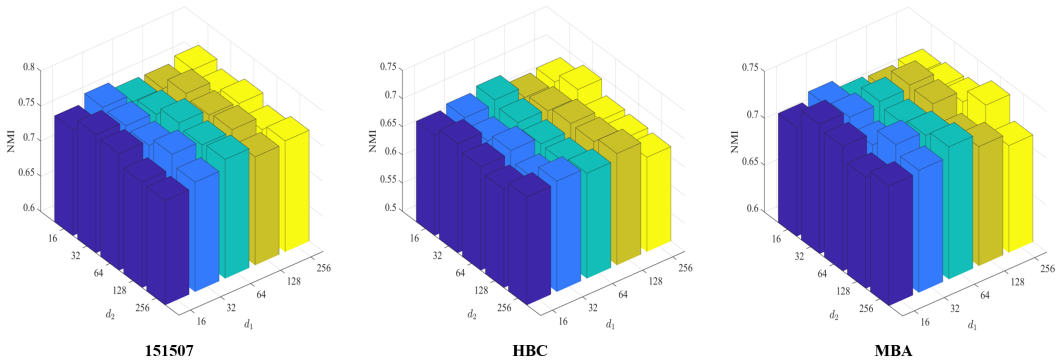

Figure 10: Sensitivity analysis of SAINT on NMI metrics.

A.5 Detailed Ablation Study on NMI

To complement the ARI-based evaluation in the main text, we further analyze the NMI performance of different SAINT variants across three representative datasets: DLPFC 151672, HBC, and MBA. Figure 7 presents the corresponding NMI results.

Across all datasets, the full SAINT model achieves the highest NMI, validating the effectiveness of incorporating nucleotide-level priors through an attention-based fusion mechanism. On DLPFC slice 151672, SAINT improves over the baseline by +6.7% in NMI. Similarly, on HBC and MBA, relative gains of +3.9% and +3.3% are observed. Notably, even the average-pooling variant (denoted as w. SA) provides substantial improvements over the no-fusion baseline. For instance, on slice 151672, w. SA achieves a NMI of 0.79, compared to 0.75 for the baseline. These findings suggest that gene sequence embeddings capture biological context that complements spatial transcriptomic signals, even in the absence of attention. The consistent improvements across both tumor (HBC) and healthy (MBA, DLPFC) tissues further demonstrate that sequence priors enhance clustering generalizability in both homogeneous and heterogeneous spatial domains.

A.6 Extended Sensitivity Analysis (NMI)

To complement the ARI-based sensitivity results, we further report the corresponding Normalized Mutual Information (NMI) scores under varying sequence embedding dimensions. As shown in Figure 10, we evaluate performance across a grid of values for d_1 and $d_2$the embedding dimensions used in the SAINT-G and SAINT-SA modules, respectivelychosen from $\{16, 32, 64, 128, 256\}$.

Across all three datasets (151507, HBC, and MBA), we observe that SAINT consistently maintains stable NMI scores under different dimension combinations. On slice 151507, NMI ranges from 0.7224 to 0.7386, with a relative variation of only 2.25%. For the HBC dataset, NMI varies from 0.6276 to 0.6550, corresponding to a 4.37% fluctuation. On the more structurally complex MBA dataset, scores range between 0.6212 and 0.6677, reflecting a 6.99% difference. These trends align with the ARI results and confirm that SAINT exhibits low sensitivity to the choice of sequence embedding dimensions. The model is capable of achieving strong performance across a broad range of d_1 and d_2 values without requiring delicate tuning.

A.7 Impact of Sequence Embedding Dimension

To further investigate the effect of sequence embedding dimensionality on model performance, we conduct an ablation study by varying the dimensionality d of nucleotide sequence embeddings across five values: $\{16, 32, 64, 128, 256\}$. The results, presented in Figure 6, report ARI and NMI scores on the DLPFC 151507 slice for three models: the original Spatial-MGCN, Spatial-MGCN+SAINT-G, and Spatial-MGCN+SAINT-SA.

We observe that both SAINT variants consistently outperform the base Spatial-MGCN across all dimensions. Notably, SAINT-SA achieves the best performance at $d = 32$ and $d = 64$, reaching an ARI of 73.57% and 72.38%, respectively. This corresponds to a relative ARI gain of +10.5% and +9.3% compared to SAINT-G at the same dimensions, and +16.7% and +15.0% over the baseline

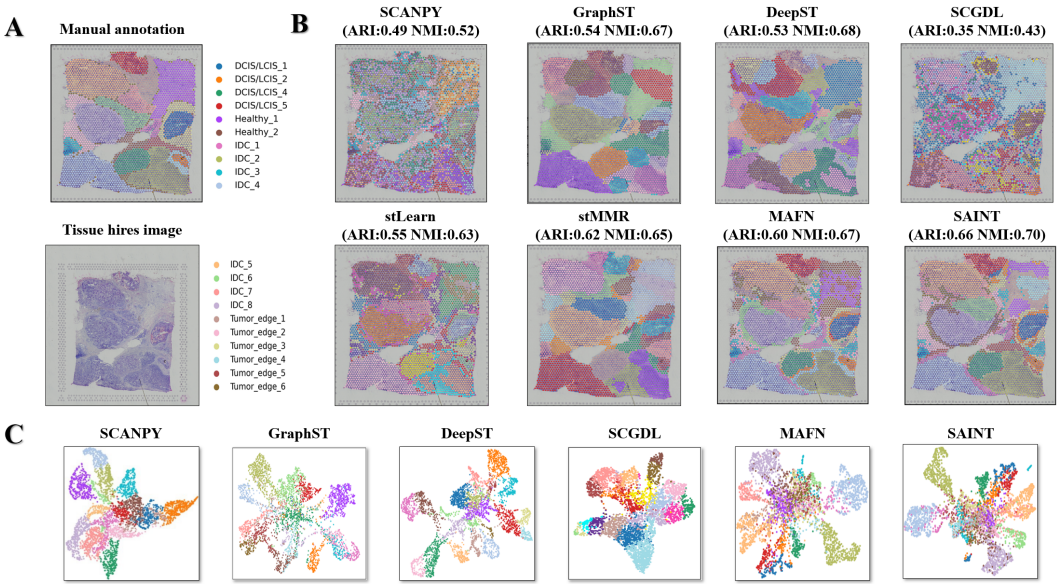

Figure 11: Case study of the proposed SAINT on HBC dataset.

Spatial-MGCN. Similarly, for NMI, SAINT-SA peaks at $d = 64$ with a score of 76.90%, surpassing SAINT-G and Spatial-MGCN by +0.96% and +2.47%, respectively. SAINT-G also consistently improves upon the base model, particularly at lower embedding sizes. At $d = 32$, SAINT-G achieves 71.25% ARI and 75.00% NMIgains of +8.2% (ARI) and +0.57% (NMI) over the base model. However, its performance tends to plateau or slightly decline at higher dimensions ($d = 128$ or 256), suggesting potential overfitting or redundancy without the attention-guided fusion mechanism.

These findings indicate that the sequence-aware attention module in SAINT-SA enables more effective utilization of high-dimensional embeddings, maintaining robust and discriminative representations even as dimensionality increases. In contrast, the simple averaging strategy used in SAINT-G is more sensitive to dimensionality, showing diminishing returns beyond $d = 64$. Overall, the dimensional analysis validates the robustness and scalability of SAINT, especially in its full attention-based variant. It further confirms that integrating gene sequence representations can enhance spatial clustering performance in a dimension-aware manner.

A.8 Case Study Analysis

To further investigate the biological interpretability of SAINT, we perform case studies on three representative spatial transcriptomics datasets (DLPFC, HBC, and MBA) to visually examine the clustering quality, gene expression continuity, and alignment with anatomical ground truth.

DLPFC: Layer-specific Marker Recovery and Cortical Architecture. Figure 8A shows the spatial expression patterns of six canonical marker genes*ATP2B4*, *B3GALT2*, *CRYM*, *FKBP1A*, *NEFH*, and *RXFP1*on slice 151507 before (Raw) and after (SAINT) model reconstruction. The raw data presents noisy, fragmented spatial signals. In contrast, SAINT smooths out spurious variations and reveals clear laminar boundaries that correspond well to cortical layers. In Figure 8B, violin plots further demonstrate improved expression stratification across layers. Genes like *ATP2B4* and *CRYM* exhibit sharper peaks in layer-specific distributions, indicating enhanced intra-cluster consistency. This suggests that SAINT captures biologically meaningful spatial organization that is obscured in raw measurements.

To further validate the spatial fidelity of SAINT, Figure 9 compares the clustering outputs of SAINT and six baseline methods on slice 151507 of the DLPFC dataset. Panel A displays the manually annotated ground truth and histological image for reference, while Panel B shows clustering results from competing models. As observed, classical methods such as SCANPY fail to capture spatial structure, resulting in noisy and biologically implausible regions. More advanced models like GraphST and DeepST partially recover cortical lamination but exhibit irregular boundaries or fragmented do-

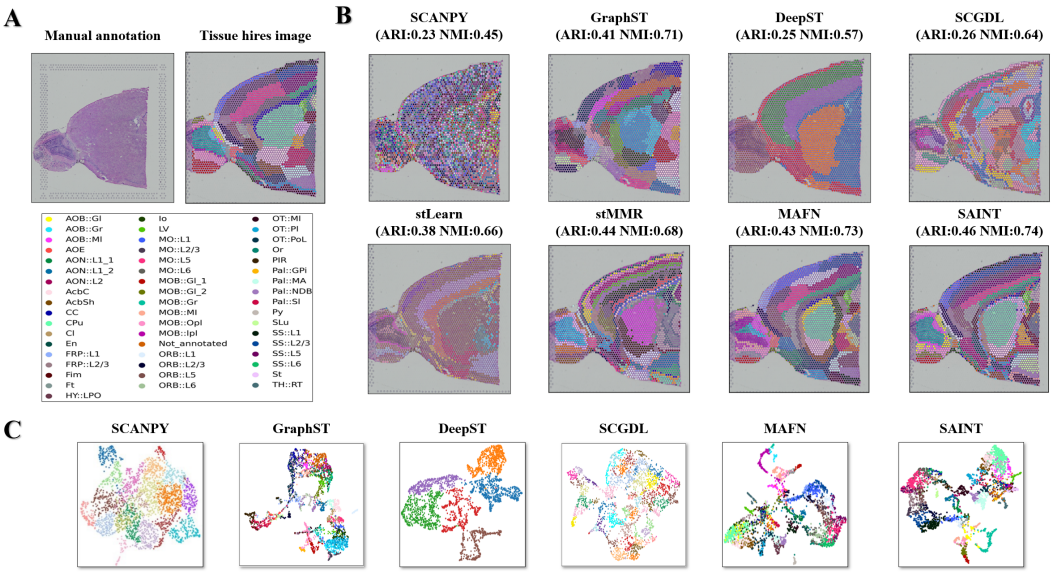

Figure 12: Case study of the proposed SAINT on MBA dataset.

mains. In contrast, SAINT produces well-aligned spatial clusters that closely match the manually defined cortical layers. Each layer is sharply delineated with minimal mixing, particularly in deeper layers such as Layer 5 and WM (white matter), where most other models show noticeable confusion. This suggests that the incorporation of nucleotide-level priors improves spatial coherence and biological interpretability. Panel C presents the 2D UMAP projections of spot embeddings. Compared to scattered or overlapping clusters generated by prior methods, SAINT yields more compact and clearly separable groups, further confirming its ability to preserve anatomical structure in the latent space.

HBC: Tumor-edge Delineation and Microenvironment Disentanglement.

As illustrated in Figure 11, SAINT shows improved resolution of complex tumor microenvironments in the HBC dataset. Compared to other methods that often over-fragment or blur tumor boundaries, SAINT delineates ductal carcinoma in situ (DCIS), invasive ductal carcinoma (IDC), and tumor-edge zones with higher coherence and spatial continuity. UMAP embeddings (Figure 11C) confirm this observation. Clusters identified by SAINT are compact, well-separated, and show clearer boundaries between IDC subtypes (e.g., IDC_1 to IDC_8). Notably, transitional tumor-edge areas are correctly positioned at cluster boundaries, reflecting subtle expression gradients across tissue regions. These results demonstrate SAINTs ability to model fine-grained spatial heterogeneity in complex pathological samples.

MBA: Resolving Anatomical Hierarchies in the Mouse Brain. In the MBA dataset (Figure 12), SAINT excels in reconstructing intricate anatomical subregions such as olfactory cortex (MOB), thalamus (TH), and hypothalamus (HY). While prior methods (e.g., SCANPY, SCGDL) tend to over-fragment or mix transitional areas, SAINT preserves spatial coherence and respects anatomical continuity. For instance, in regions like FRP::L2/3 and TH::RT, SAINT accurately recovers localized clusters that align with the brains hierarchical organization. The corresponding UMAP projection reveals compact and non-overlapping clusters, indicating that the learned embedding reflects both macro-structure and local transcriptional variation.

A.9 Theoretical Complexity Analysis

The computational complexity of SAINT is mainly determined by the multi-view GCN layers, where each branch processes a feature matrix $X \in \mathbb{R}^{N\times F}$ with adjacency $A \in \mathbb{R}^{N\times N}$, resulting in $O(3|E|F)$, where N is the number of spots, F the feature dimension, and $|E|$ the number of edges. The sequence embedding branch only adds a lightweight projection from precomputed embeddings of dimension d_s to d_p with complexity $O(Nd_sd_p)$, and the ZINB decoder adds a negligible $O(Nd_p)$

term. Thus, the total complexity remains $O(|E|F)$, comparable to GCN-based methods like MAFN, with only a small linear overhead from the sequence branch.

A.10 Comparison with Spatial-MGCN and MAFN

This study follows the experimental setup, datasets, and benchmarking protocols used in MAFN[62] to ensure fair comparison. The DLPFC, HBC, and MBA datasets, together with the same evaluation metrics, are adopted for consistency. Nevertheless, the proposed SAINT framework introduces several essential innovations beyond Spatial-MGCN[47] and MAFN[62] .

First, previous methods mainly integrate spatial and gene expression information through graph convolutional networks and adaptive fusion strategies, without considering the biological knowledge contained in gene nucleotide sequences. SAINT explicitly incorporates nucleotide-level representations into the clustering process, allowing biologically distinct genes with similar expression profiles to be differentiated.

Second, SAINT employs a sequence-aware encoder based on the pretrained Nucleotide Transformer, which converts gene sequences into high-dimensional embeddings that capture functional and regulatory genomic information not accessible from expression data alone.

Third, while Spatial-MGCN and MAFN both use expression and spatial graphs, SAINT integrates sequence-derived representations through a cross-modal decorrelation loss (DICR). This loss promotes complementary information across modalities rather than simple consistency.

Fourth, SAINT further contributes by constructing sequence-augmented datasets for widely used benchmarks, enabling reproducible evaluation of sequence-informed spatial models.

While the overall architectural design inherits certain effective components from previous framework, such as the graph embedding backbone and the zero-inflated negative binomial (ZINB) reconstruction loss, these elements serve as well-established foundations in spatial transcriptomics modeling. The ZINB loss is particularly suited to the sparse and overdispersed characteristics of gene expression data, offering better statistical fidelity than alternatives such as mean squared error (MSE). On top of this foundation, SAINT introduces biologically motivated enhancements, including attention-based aggregation that dynamically weights nucleotide features according to their relevance, thereby emphasizing informative signals and suppressing noise.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (12 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly summarize the main contributions, including the integration of gene sequence embeddings and the proposed fusion strategy. These claims are consistent with the methods and results presented.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: The limitations of the proposed method, including its reliance on pretrained sequence embeddings and sensitivity to gene selection, are discussed in Appendix A.1.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: [NA]

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

Justification: The experimental settings, datasets, and implementation details are clearly described in the main text and appendix. The authors also state that the source code and pretrained models will be released upon publication, ensuring reproducibility of the results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The authors mention that the source code will be released after the double-blind review.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so No is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.

- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The details of experimental settings are carefully described in the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Five aspects of experiments are conducted to evaluate the proposed model.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The details of experimental settings are carefully described in the Appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The authors preserve anonymity and obey the code of ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: This work focuses on methodological advancements for spatial transcriptomics data clustering, which is a foundational task in computational biology. While the approach may indirectly support biomedical discoveries or disease understanding, it is not tied to direct applications or deployments that raise immediate societal concerns. As such, broader societal impacts were not discussed in the main text.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All of the backbone models and datasets are cited with their references.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.