Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

GraphTransformer-CoGNN: A two-stage dynamic graph learning framework for label-scarce cell-type annotation in spatial transcriptomics

Yuanyuan Dang
Changchun University of Technology
Xinyi Han
2202403053@stu.ccut.edu.cn
Changchun University of Technology
Bing Liu
Changchun University of Technology

Research Article

Keywords: Spatial transcriptomics, Semi-supervised learning, Graph neural network, Graph transformer, Dynamic graphs tructure learning

Posted Date: August 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10407073/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

GraphTransformer-CoGNN: A two-stage dynamic graph learning framework for label-scarce cell-type annotation in spatial transcriptomics

Yuanyuan Dang[1], Xinyi Han[1*], Bing Liu[1]

[1]School of Computer Science and Engineering, Changchun University of Technology, Changchun, 130012, Jilin, China.

*Corresponding author(s). E-mail(s): 2202403053@stu.ccut.edu.cn;
Contributing authors: dangyuanyuan@ccut.edu.cn; liubing@ccut.edu.cn;

Abstract

Accurate cell-type identification in single-cell and spatial transcriptomics is crucial for understanding tissue heterogeneity and disease mechanisms. In semi-supervised settings, however, performance is frequently limited by label scarcity, distribution mismatch between labeled and unlabeled cells, and noisy connections in pre-defined cell graphs. To overcome these challenges, we propose GraphTransformer-CoGNN, a two-stage dynamic graph learning framework. First, a Graph Transformer learns an adaptive graph structure, incorporating global context through resistance-distance-based relative positional encoding. Second, Cooperative Graph Neural Network (CoGNN) refines this structure by suppressing spurious edges and enhancing biologically relevant interactions. The model is trained end-to-end with a joint objective combining supervised classification loss and triplet metric loss. Experiments on NanoString-based spatial transcriptomics and the human DLPFC benchmark demonstrate that with only 18%labeled cells, our method achieves state-of-the-art accuracy and Cohen's kappa, while significantly improving the discrimination of rare cell types.

Keywords: Spatial transcriptomics, Semi-supervised learning, Graph neural network, Graph transformer, Dynamic graphs tructure learning

1 Introduction

Single-cell RNA sequencing (scRNA-seq) technology has revolutionized the study of cellular heterogeneity, providing robust support for precise cell type classification, disease mechanism elucidation, and tissue function exploration (Li and Wang 2021). To decipher complex cellular relationships in scRNA-seq data, graph neural networks (GNNs) have emerged as a powerful tool by modeling cells as nodes and cellular similarities or spatial proximities as edges, enabling simultaneous utilization of gene expression profiles and spatial relationship information (Ge et al. 2025).

Despite the success of GNNs, their application in semi-supervised scenarios still faces significant challenges (Feng et al. 2020). First, traditional GNNs rely on static graph structures constructed during the preprocessing phase based on heuristic methods such as Euclidean distance or k-nearest neighbors (k-NN). This fixed topology cannot adapt to the inherent biological complexity and technical noise of scRNA-seq data, potentially leading to misleading information propagation (Buterez et al. 2022). Secondly, in semi-supervised tasks, the impact of structural deficiencies becomes more pronounced: the isolated "hub" nodes where highly variable genes (HVGs) reside (which are crucial for revealing tissue heterogeneity) (Wang and Du 2025) often lack labels, leading to failed message passing during training. Finally, GNNs are prone to over-smoothing and over-squashing issues when handling long-range dependencies.

In recent years, graph structure learning (GSL) and graph Transformer architectures have provided new approaches to address the limitations of static graphs. GSL methods aim to jointly optimize graph topology and node representations (Zhou et al. 2023), while graph Transformers leverage self-attention mechanisms to dynamically compute association strengths between nodes. However, in highly heterogeneous single-cell transcriptomic data, relying solely on a single, fixed information aggregation paradigm may still be insufficient for accurately modeling complex cell microenvironment interactions (Fan et al. 2024). The recently proposed Collaborative Graph Neural Network (CoGNN) (Finkelshtein et al. 2024) introduces new flexibility to graph learning by allowing each node to dynamically select information propagation strategies (such as receiving, broadcasting, or isolating) based on its state.

This paper proposes GraphTransformer-CoGNN, a novel two-stage framework designed to collaboratively address structural learning and information propagation in semi-supervised single-cell classification. In the first stage, a spatial-aware graph Transformer module learns graph structures from gene expression and spatial coordinates, incorporating relative positional encoding biases based on resistance distance (RD) (Zhang et al. 2023). In the second stage, the CoGNN module performs learning on the graph structure, adaptively emphasizing critical connections while suppressing noise, further rectifying structural deficiencies to enhance the generalization performance of semi-supervised cell classification. Furthermore, we conducted systematic evaluations on real-world datasets such as NanoString CosMx SMI (Tang et al. 2023) and DLPFC(spatialLIBD) (Pardo et al. 2022), validating the method's advantages in robustness and generalization.

The remainder of this paper is organized as follows: Section 2 introduces the proposed GraphTransformer-CoGNN framework, including data preprocessing, relative

positional encoding of resistance distance, dynamic graph structure learning in spatial-aware Graph Transformer, the collaborative mechanism of CoGNN, and the joint optimization objective. Section 3 presents the datasets, experimental settings, baseline methods, and evaluation metrics. Section 4 reports the results of comparative experiments, ablation studies, and visualization analyses. Finally, Section 5 concludes the work of this paper.

2 Methods

To enhance the accuracy of single-cell classification and address the limitations in existing graph structure learning, this paper proposes GraphTransformer-CoGNN: a two-stage dynamic graph learning framework for semi-supervised single-cell classification. The specific process is illustrated in Fig. 1.

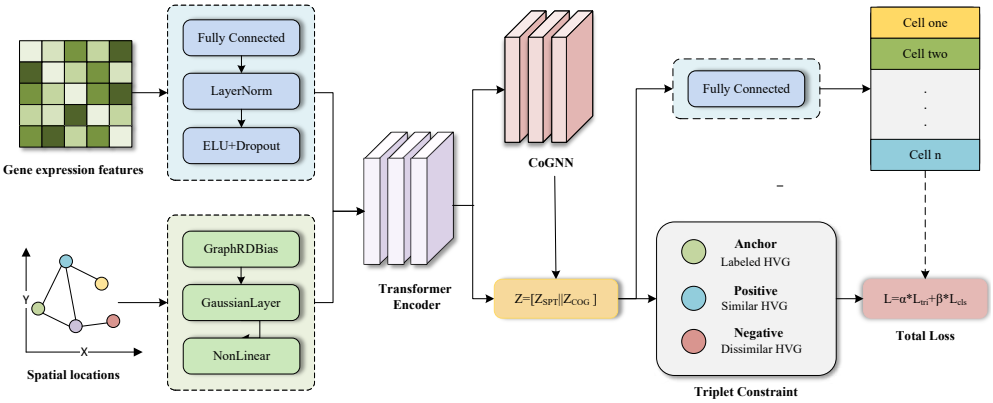


Fig. 1: Schematic overview of the GraphTransformer-CoGNN framework

In this framework, the gene expression features X_g are first linearly projected to map them into a unified hidden dimension. They are then sequentially processed through layer normalization, ELU activation function, and Dropout to obtain the initial node embedding representation. Meanwhile, the 2D spatial coordinates of cells, spatial_locs $\in \mathbb{R}^{N\times 2}$, are initially processed by the GraphRDBias module to compute pairwise resistance distances. These distance measurements are subsequently mapped into Gaussian kernel space through the GaussianLayer, thereby enhancing the model's nonlinear representation capability. After transformation by the NonLinear module, they generate the attention bias attn_bias, which is then injected into the multi-layer Spatial-Aware Transformer Encoder to further produce spatially aware features Z_{SPT}.

After the graph structure learning is completed by Graph-Transformer, the obtained graph structural representations are fed into the multi-layer CoGNN message passing module. Within this module, each node combines its own features with those of its neighbors, generates action distributions through a lightweight policy network, and employs straight-through Gumbel-softmax(Jang et al. 2017; Maddison

et al. 2017) sampling to obtain discrete communication actions, resulting in node-level "retention scores". These scores are used to determine whether a node performs Listen, Broadcast, Standard, or Isolate behavior at each layer, thereby selectively enabling or blocking information reception and broadcasting in a hard-gating manner. This suppresses unreliable connections in the graph while enhancing edges that are more critical for classification tasks, thus optimizing the message-passing process and enabling effective learning on the graph structure.

Subsequently, the spatially-aware Transformer representation Z_{SPT} and the graph structural representation Z_{COG} output by CoGNN are concatenated along the feature dimension to obtain the final node representation $Z = [Z_{\mathrm{SPT}} \| Z_{\mathrm{CoG}}]$. This is then mapped to logits for each cell type through a fully connected classification head(Kipf and Welling 2017) for final cell type prediction. The node classification prediction is optimized using the cross-entropy loss function to improve classification accuracy. Meanwhile, a triplet loss $\mathcal{L}_{\mathrm{tri}}$ is introduced to constrain the distance relationships among Anchor, Positive, and Negative samples, ensuring that cells of the same type are closer in the representation space while those of different types are farther apart. The weighted combination form of the total model loss: $\mathcal{L} = \alpha \cdot \mathcal{L}_{\mathrm{tri}} + \beta \cdot \mathcal{L}_{\mathrm{cel}}$, which combines triplet loss with cross-entropy loss to jointly optimize node features, improve the accuracy of cell classification, and facilitate the discovery of spatial heterogeneity in cells.

2.1 Problem Definition and Data Representation

This study aims to accurately classify cell types from scRNA-seq data by leveraging gene expression profiles and spatial structural information. Specifically, we seek to enhance the graph model's ability to represent the relational structures among cells in a semi-supervised setting with only limited cell annotations, thereby enabling reliable cell type discrimination in complex tissues. To achieve this goal, we utilize two core types of input data:.

The gene expression matrix $\mathbf{X}_g \in \mathbb{R}^{n \times d}$, where n is the number of cells and d is the gene-feature dimension. The i-th row $\mathbf{X}_g(i,:)$ corresponds to the expression profile of cell i. The data are typically obtained from sequencing-based spatial barcode capture (e.g., 10x Genomics Visium)(Maynard et al. 2021; Oliveira et al. 2025) or high-throughput multiplexed in situ imaging (e.g., MERFISH(Kumar et al. 2024), CosMx SMI(He et al. 2022)), which provide spatially resolved single-cell(-like) transcriptomic measurements.

The spatial coordinate matrix $\mathbf{X}_s \in \mathbb{R}^{n \times 2}$ records the 2D physical coordinates of each cell in the tissue section (e.g., the centroid position of the cell).

Under the graph-learning framework, we model the data as an expression graph $G_g = (V, E, \mathbf{X}_g)$, where the node set V contains n cells and the edge set $E \subseteq V \times V$ encodes cell–cell associations (typically initialized from spatial proximity). Each node $v_i \in V$ carries both expression features $\mathbf{X}_g(v_i)$ and spatial coordinates $\mathbf{X}_s(v_i)$.

In traditional graph neural networks, such as the Graph Convolutional Network (GCN) (Kipf and Welling 2017) and GraphSAGE (Hamilton et al. 2017), models rely on a statically constructed adjacency matrix $\mathbf{A}_g \in \mathbb{R}^{n \times n}$ that remains fixed during training. Such fixed structures can capture local neighborhood relations but often

struggle with context-dependent cell–cell interactions under strong spatial heterogeneity, overlapping cell-type distributions, and rare subpopulations (Feng et al. 2020; Buterez et al. 2022; Zhou et al. 2023).

To address this, our study proposes the GraphTransformer-CoGNN model, which instead learns an input-dependent dynamic adjacency relationship. Specifically, we integrate the collaborative learning mechanism of CoGNN based on a spatially-aware Graph Transformer. This design enables the model to adaptively adjust the strength of inter-node connections according to cellular feature contexts, thereby achieving more accurate identification of cell subpopulations in scenarios with sparse annotations, complex spatial structures, or highly intertwined cell types. Compared to methods such as STAGATE (relying on fixed spatial adjacency)(Dong and Zhang 2022) or SpaGCN (based on predefined spatial graphs)(Hu et al. 2021), our framework focuses on task-oriented dynamic optimization of the initial graph structure, thereby providing more adaptive and robust topological modeling.

Formally, in semi-supervised classification, we have a labeled subset of nodes $V_l \subset V$ with $|V_l| = \ell \ll n$ and labels $Y_l = \{y_i \mid i \in V_l\}$. The objective is to learn task-sensitive representations by jointly leveraging expression and spatial information, and to propagate supervision from V_l to unlabeled nodes through learned graph structures.

Overall, GraphTransformer-CoGNN improves generalization and classification accuracy in complex tissue environments by dynamically optimizing edge weights and learning structured representations that correct biases in the initial graph and reinforce informative intercellular interactions.

2.2 Graph Transformer

To overcome the dependency of traditional Graph Neural Networks (GNNs) on predefined static graph structures, this study employs a Graph Transformer architecture for global relationship modeling of single-cell RNA sequencing data. The core of this module lies in utilizing multi-head self-attention mechanisms to adaptively learn the association strength between nodes, thereby constructing a dynamic graph structure (Shehzad et al. 2024) dependent on input features. This approach can effectively filter out noisy connections or irrelevant edges introduced when constructing initial graphs based on simple spatial heuristic rules (e.g., k-nearest neighbors), while providing more expressive initial node representations for the subsequent CoGNN.

Although existing studies have attempted to mitigate the limitations of static graphs through joint optimization of graph topology and node representations (e.g., GLCN (Jiang et al. 2019)and the adaptive graph structure learning method proposed by Li (Zhang et al. 2025)et al.), most of these approaches are designed for general graph data. They have not fully accounted for the noise and bias inherent in imperfect initial graph structures, making it particularly challenging to provide stable and reliable graph structural support for cell type identification under label-sparse conditions.

As shown in Fig. 2, we construct a spatially aware Transformer encoder. For the node feature input $X^{(l)} \in \mathbb{R}^{n \times d}$ at the l-th layer, the encoder first maps the node features into key, query, and value representations through three sets of learnable linear transformations W_k, W_q, and W_v(Vaswani et al. 2017), respectively. Building upon this, the multi-head self-attention mechanism is employed to compute similarity

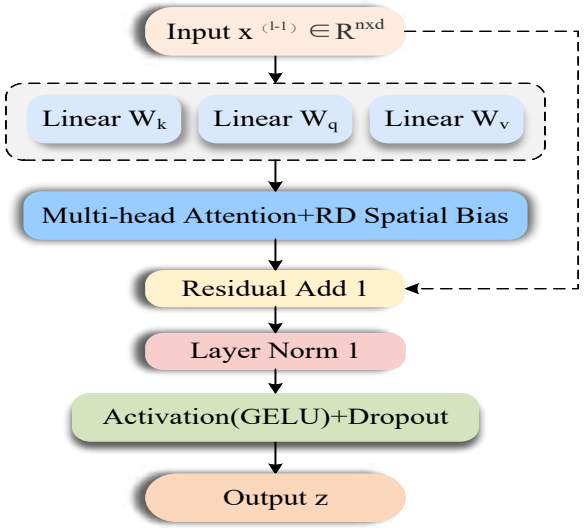


Fig. 2: Schematic illustration of the spatially aware Transformer encoder

scores between node pairs, while incorporating spatial bias derived from resistance distance. This constructs an adjacency relationship that simultaneously integrates gene expression similarity and global spatial structure.

Specifically, the forward propagation process of multi-head attention in a single encoder layer can be formally expressed as:.

$$\mathbf{X}_g^{(l+1)} = \mathbf{X}_g^{(l)} + \mathrm{GELU}\left(\sum_{h=1}^{H} \mathrm{softmax}\left(\frac{\mathbf{Q}_h^{(l)}\mathbf{K}_h^{(l)\top}}{\sqrt{d_h}}\right)\mathbf{V}_h^{(l)}\mathbf{W}^{(l)}\right) \tag{1}$$

Here, H denotes the number of attention heads, and $Q^{(h,l)}$, $K^{(h,l)}$, $V^{(h,l)}$ represent the query, key, and value representations of the h-th head at the l-th layer, respectively, with d_h being the hidden dimension of a single head. The attention weights, normalized via softmax, are used to perform weighted aggregation of the value vectors. The resulting output is then added to the layer input through residual connection, followed by sequential processing through layer normalization (LayerNorm)(Xu et al. 2019), exponential linear unit (GELU) activation function (Hendrycks and Gimpel 2016), and Dropout, ultimately yielding the output features $X^{(l+1)}$ of the current layer. The incorporation of residual connections and GELU activation function helps stabilize the training of deep networks and enhances the model's capability to fit complex nonlinear expression patterns.

Compared to traditional GNNs that rely on fixed adjacency lists and only aggregate information within local neighborhoods, graph Transformers offer two key advantages: First, their global receptive field enables each node to adaptively aggregate information from all other nodes based on attention weights (Liu et al. 2025), which is crucial for capturing spatially non-adjacent but functionally relevant cellular interactions. Second, by combining gene expression similarity with spatially biased resistance distance,

the model can simultaneously achieve long-range dependency mining and initial graph noise suppression when constructing dynamic adjacency. Compared to methods relying on fixed spatial graphs, this design can learn more robust and discriminative node representations under conditions of sparse annotations and structural noise, thereby providing a more reliable graph structure foundation for the downstream CoGNN module.

2.3 Cooperative Graph Neural Networks(CoGNN)

To further strengthen the structural support for cell type discrimination, we incorporate the cooperative graph structure learning mechanism of CoGNN on top of the spatially aware node representations Z_{SPT} produced by the Graph Transformer. In CoGNN, each node is viewed as a player that selects a communication strategy at each layer. Specifically, the action network decides how a node interacts with its neighbors, while the environment network updates node states accordingly, resulting in a flexible, layer-wise message passing process.

Taking Z_{SPT} as input, CoGNN refines the pre-constructed cell adjacency graph by learning layer-dependent edge strengths, which modulate neighborhood aggregation during training (Gutteridge et al. 2023). This adaptive refinement suppresses noisy or uninformative interactions and emphasizes task-relevant cell–cell dependencies, yielding more robust and discriminative structural representations and improving the accuracy and stability of semi-supervised cell type prediction.

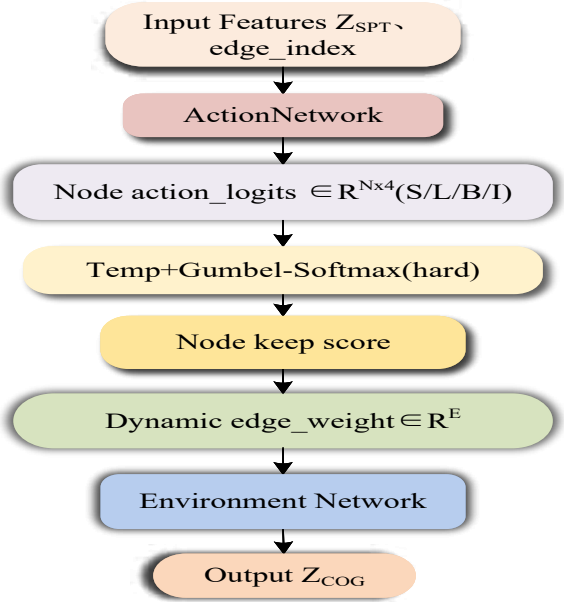


Fig. 3: CoGNN Flowchart

As shown in Fig. 3, the CoGNN adopts an iterative mechanism of Action Decision - Environment Update (Faber and Wattenhofer 2022). The specific workflow is as follows: First, a lightweight Action Network (Lai et al. 2020) generates 4-dimensional action logits $\in \mathbb{R}^{N\times 4}$ for each node based on the current layer's node representations, corresponding to four actions: Listen, Broadcast, Standard, and Isolate respectively. The discrete actions are obtained through the Gumbel-Softmax (hard) with temperature τ; subsequently, the action results are mapped to node keep scores, which quantify the nodes' willingness to participate in information propagation.

Based on this, dynamic edge weights are generated on the pre-constructed edge set E. The keep score is derived from a specific dimension of the action distribution, and the edge weight edge_weight $\in \mathbb{R}^E$ is constructed by multiplying the keep scores of the endpoints, thereby performing task-driven re-weighting of edge importance without altering the edge set. The weighted adjacency relations and node features are then fed into the environment network, a neighborhood aggregation module based on sparse multi-head attention (Shirzad et al. 2023). Here, dynamic edge weights serve as structural priors to adaptively modulate the intensity of message passing: signals from unreliable connections are weakened, while neighbor interactions highly relevant to the current task are strengthened. Ultimately, CoGNN outputs structurally refined node representations Z_{CoG}.

Formally, at the l-th layer, the policy network π outputs an action probability distribution based on the current node representations and their neighbor features:

$$\mathbf{p}_v^{(l)} = \pi\Big(\mathbf{h}_v^{(l)}, \{\mathbf{h}_u^{(l)} \mid u \in \mathcal{N}(v)\}\Big), \quad \mathbf{p}_v^{(l)} \in \mathbb{R}^4 \tag{2}$$

Use the straight-through Gumbel-Softmax estimator to discretely sample the actions:

$$\mathbf{a}_v^{(l)} = \text{ST-GS}\Big(\mathbf{p}_v^{(l)}; \tau\Big) \tag{3}$$

where τ denotes the temperature parameter.

In graph modeling of spatially resolved single-cell transcriptomics, the initial adjacency graph often contains redundant or erroneous connections due to tissue boundaries, cell density variations, and technical noise. Under semi-supervised scenarios, sparse annotations further amplify the impact of such structural defects, hindering effective propagation of supervisory signals (Dai et al. 2022). The fixed edge set with learnable weights refinement mechanism adopted by CoGNN provides an elegant solution: while preserving the initial spatial neighborhood priors, it allows the model to adaptively adjust the utility of each edge based on dynamic node representations and global training objectives. This flexibility enables the model to learn more precise and robust representations of intercellular relationships despite challenges posed by scarce annotations and noise presence, thereby providing more reliable structural foundations for final classification decisions.

2.4 Resistance Distance (RD) Positional Encoding

The self-attention mechanism in Graph Transformers is equivariant to node permutations, meaning that permuting the input node order results in the same permutation

of the outputs. Without additional position information injection, the model cannot distinguish nodes located at different structural positions in the graph, thereby limiting its expressive power for topological structures. Therefore, incorporating position encoding into Graph Transformers is crucial. Existing positional encodings in Graph Transformers are mainly divided into two categories (Black et al. 2024):

1. Absolute Position Encoding (APE): Assigns a position vector to each node as a node-level feature input to the Transformer (Dwivedi and Bresson 2020).
2. Relative Position Encoding (RPE): Defines the relative relationship between node pairs (i,j) and uses it to modulate attention (Shaw et al. 2018).

Studies indicate that for graph data with rich node features, RPE typically outperforms APE as it directly models relationships between nodes without relying on fixed node identifiers , thereby better capturing structural patterns and enhancing cross-graph generalization capability.

To further theoretically understand the structural expressive power of different positional encodings, this paper employs the Weisfeiler-Lehman (WL) (Morris et al. 2019) graph isomorphism test hierarchy as a theoretical reference for measuring the structural expressive capabilities of various positional encodings (PEs). Intuitively, 1-WL (WL1) corresponds to the discriminative ability based on iterative aggregation of local neighborhoods, hence many relative positional encodings relying on local topology (e.g., hop distance (Zhang et al. 2020) or shortest path distance SPD (Ying et al. 2021))typically belong to this level. In contrast, 2-WL (WL2) can capture more fine-grained edge- and path-level structural differences, exhibiting stronger expressive power in certain graph structure discrimination tasks.

Based on this, we propose adopting relative positional encoding based on effective resistance (RD). RD is a global graph metric derived from circuit theory, which quantifies the connectivity convenience between two nodes by considering all possible paths in the graph.

It is defined as:.

$$\mathrm{RD}(i,j) = (\mathbf{e}_i - \mathbf{e}_j)^\top \mathbf{L}^\dagger (\mathbf{e}_i - \mathbf{e}_j) \tag{4}$$

Here, $\mathbf{L}^\dagger$ denotes the Moore–Penrose pseudoinverse of the Laplacian matrix $\mathbf{L}$, and $\mathbf{e}_i$ represents the one-hot vector of node i. To reduce computational overhead, RD can be approximated through spectral decomposition:

$$\mathrm{RD}(i,j) \approx \sum_{k=2}^{K} \frac{1}{\lambda_k} \left(u_k(i) - u_k(j)\right)^2 \tag{5}$$

Here λ_k and u_k represent the non-zero eigenvalues and corresponding eigenvectors of the Laplacian matrix respectively. In practical implementation, the first K non-zero spectral components are selected to balance accuracy and efficiency.

Compared with other positional encoding methods, RD can capture the global connectivity of multiple paths between nodes (Black et al. 2024), thus surpassing the expressive power of the 1-WL hierarchy and approaching the discriminative capability of 2-WL(Ellens et al. 2011). Table 1 systematically compares the characteristics of several positional encoding approaches.

Table 1: Positional encoding method comparison.

Encoding	Category	Where Injected	WL View
Absolute PE	APE	Node Feature	$<$ 1-WL
Hop Distance	RPE	Attention bias	$\sim$ 1-WL
SPD	RPE	Attention bias	$\sim$ 1-WL
RD	RPE	Attention bias	$>$ 1-WL

In spatially resolved single-cell analysis, cells within local spatial neighborhoods may exhibit similar microenvironments, yet their topological positions and functional associations within the overall tissue architecture can differ significantly. The sensitivity of RD to global connectivity structures enables the model to more effectively identify such long-range functional coordination relationships. Consequently, under conditions of noisy initial graph construction and sparse annotations, RD-RPE can provide the model with richer and more robust structural priors, thereby facilitating the learning of more discriminative cellular representations.

2.5 Triplet Loss

In the semi-supervised setting, the number of available labeled nodes is limited. Relying solely on classification loss often provides supervisory signals only at a few annotated points, making it difficult to directly constrain the geometric structure of the latent representation space. To enhance the discriminative power of the representations, we introduce triplet metric learning constraints (Schroff et al. 2015): nodes belonging to the same cell type are brought closer together in the embedding space, while nodes from different categories are pushed apart, thereby providing additional supervisory signals for classification tasks under low annotation rates.

Specifically, we construct triplets from the labeled node set V_l. Each triplet consists of three nodes:

1. Anchor node: $V_a = \{v \in V_l \mid \text{HVGs} \subseteq X(v)\}$
2. Positive node (shares the same label as the anchor): $V_p = \{v \in V \setminus \{v_a\} \mid \text{HVGs} \subseteq X(v), \text{dis}(v_a, v) \geq d\}$
3. Negative node (has a different label from the anchor): $V_n = \{v \in V \setminus \{v_a\} \mid \text{HVGs} \nsubseteq X(v), \text{dis}(v_a, v) \geq d\}$

Here, X(v) denotes the gene expression profile of node v, and HVGs refer to biologically informative highly variable genes selected based on expression variability across cells. During the construction process, we introduce a spatial distance-based constraint: the Euclidean spatial distance between positive/negative sample nodes and the anchor node must exceed a predefined threshold d. This constraint ensures that the selected samples are not immediate spatial neighbors of the anchor, thereby encouraging the model to learn biologically relevant associations that may be spatially separated but functionally or typologically similar .

The triplet loss function is:.

$$\mathcal{L}_{\text{tri}} = \sum_{(a,p,n)\in T} \left[\|h_{v_i^a} - h_{v_i^p}\|_2^2 - \|h_{v_i^a} - v_i^n\|_2^2 + \alpha \right]_+ \quad (6)$$

This loss function enforces the constraint that the distance between the anchor and the positive node in the embedding space should be smaller than the distance between the anchor and the negative node by at least α. More importantly, by incorporating spatial distance constraints during the triplet construction process, the model learns to identify and construct biologically coherent cell relationships that may not be spatially adjacent. This stands in sharp contrast to traditional graph neural networks (e.g., GCN, GAT), which typically rely heavily on predefined static graph structures based on spatial proximity, with their message-passing mechanisms confined to local neighborhoods(Kipf and Welling 2017).

In spatially resolved single-cell data, cells of the same type may be spatially distant due to factors such as cell differentiation, tissue functional zoning, or functional coordination. The spatial constraints introduced in the triplet loss effectively mitigate the model's overreliance on the heuristic prior that spatial proximity implies association, thereby reducing noise caused by initial graph construction bias. This design significantly enhances the model's generalization capability and stability in cell type identification under scenarios with high tissue heterogeneity and complex spatial structures (Yang et al. 2025).

2.6 Overall Objective

To simultaneously enhance the discriminative power of spatial perception representations and the accuracy of cell type classification, this paper adopts a joint objective function composed of triplet loss and cross-entropy classification loss for end-to-end training (Hermans et al. 2017). This design constrains the geometric structure of the embedding space through metric learning, making samples of the same class more compact and samples of different classes more separated. On the other hand, it directly improves the accuracy of node category prediction through supervised classification objectives, thereby achieving more stable representation learning effects under semi-supervised settings.

Triplet loss: For a triplet consisting of an anchor point v_a, a positive sample v_p, and a negative sample v_n, let $z(\cdot)$ denote the node embedding extracted by the model. The triplet loss is defined as follows:

$$\mathcal{L}_{\text{tri}} = \sum_{(v_a,v_p,v_n)\in\mathcal{T}} \left[\|\mathbf{z}_a - \mathbf{z}_p\|_2^2 - \|\mathbf{z}_a - \mathbf{z}_n\|_2^2 + \lambda \right]_+ \quad (7)$$

Here, $\|\cdot\|_2$ denotes the ℓ_2 norm, $\lambda > 0$ is the margin, $[x]_+ = \max(0, x)$ is the hinge function, and $\mathcal{T}$ is the set of triplets constructed from labeled cells. The loss pulls anchor/positive embeddings together while pushing negatives at least λ away, improving inter-class separability.

Classification loss: For labeled nodes $v_i \in V_l$, the final representations are fed into a classifier (e.g., an MLP) to produce logits and a probability distribution $\hat{\mathbf{y}}_i \in \mathbb{R}^C$ via softmax, where C is the number of cell types. We use cross-entropy:

$$\mathcal{L}_{\text{cls}} = -\frac{1}{|V_l|} \sum_{v_i \in V_l} \mathbf{y}_i^\top \log(\hat{\mathbf{y}}_i) \tag{8}$$

Joint Optimization: The ultimate objective of the model is a weighted combination of triplet loss and classification loss:

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{\text{tri}} + (1 - \alpha) \cdot \mathcal{L}_{\text{cls}}, \tag{9}$$

where $\alpha \in [0, 1]$ serves as a tunable hyperparameter to balance the contributions of metric learning constraints and supervised classification objectives. Through joint optimization, the model can not only learn more discriminative embedding space structures under low-labeling conditions but also enhance the overall accuracy and generalization capability of cell type prediction.

3 Results

3.1 Datasets

To comprehensively evaluate the performance and generalization capability of the proposed model, we assessed our GraphTransformer model on two real single-cell transcriptomics datasets: NanoString CosMx SMI (Tang et al. 2023)and DLPFC (SpatialLIBD) (Pardo et al. 2022). The CosMx SMI dataset contains 83,621 cells from 20 fields of view (FOVs) in human non-small cell lung cancer (NSCLC) tissues. Each FOV provides spatial coordinates and detailed RNA expression data. In addition to tumor cells, this dataset encompasses various stromal and immune cell types that constitute the complex tumor microenvironment, such as fibroblasts, macrophages, lymphocytes, and vascular endothelial cells, which collectively form the intricate tumor microenvironment. The DLPFC dataset is provided by spatialLIBD and originates from the human dorsolateral prefrontal cortex, comprising 12 tissue sections with 47,681 spatial spots. Its cell type annotations cover major cortical cell classes, including excitatory neurons, inhibitory neurons, astrocytes, and oligodendrocytes. By selecting datasets with diverse tissue types and biological backgrounds, we aim to test our model's generalization capability across different domains. All experiments were conducted in a semi-supervised node classification setting with 18% labeled data, and we evaluated the performance under both random and disjoint label splitting patterns.

During the data preprocessing phase, we performed normalization on the gene expression matrices of both datasets: the expression values of each gene were subjected to min-max scaling, linearly mapped to the (Li and Wang 2021) interval. For the NanoString data, to further suppress technical noise and enhance biological signals, we additionally employed a self-supervised reconstruction model to reconstruct the expression profiles, with negative values in the reconstruction results truncated to zero to maintain biological plausibility (non-negativity) of expression values.

All experiments were conducted under the semi-supervised node classification setting, employing two labeling partition strategies: random and disjoint. To visually

demonstrate the enhancement effect of self-supervised reconstruction on expression profiles, Fig. 4 presents stacked violin plots of representative marker genes (e.g., MZT2A, HLA-A). Compared to the original data, the reconstructed expression profiles exhibit clearer and more distinct distribution patterns across different cell types, validating the effectiveness of the preprocessing step in noise removal and biological signal enhancement.

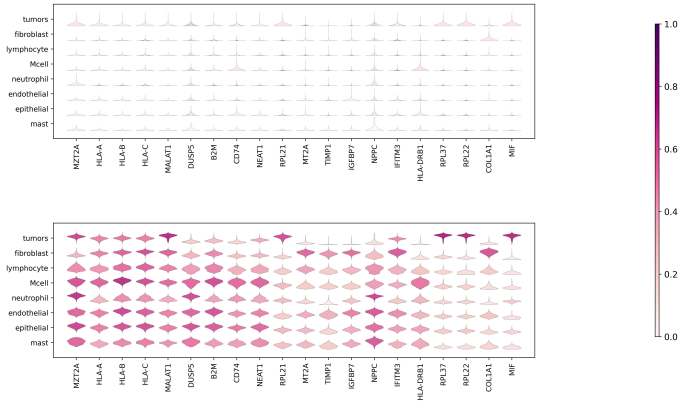


Fig. 4: Comparison of stacked violin plots between raw and enhanced expression in NanoString data

3.2 Implementation Settings

In our experiments, the encoder and decoder of PEL are implemented as 256-dimensional multilayer perceptrons (MLPs) (Berahmand et al. 2024). We optimize the overall objective using Adam. The learning rate is selected by grid search over $[5 \times 10^{-4}, 10^{-2}]$ with progressively coarser steps at larger values. During inference, the model outputs a pseudo-edge matrix $\mathbf{P} \in \mathbb{R}^{n \times n}$ that represents learned potential intercellular associations.

To suppress the propagation of noisy edges while maintaining the basic connectivity of the graph, we perform a two-stage post-processing on P:

1. Row-wise normalization(Pinheiro et al. 2025): For each node, retain its strongest connection such that $\max_j P_{ij} = 1$, ensuring every node has at least one valid information pathway.
2. Confidence threshold filtering(Li et al. 2023): For isolated nodes, only pseudo-edges satisfying $P_{ij} > \tau$ are retained ($\tau = 0.8$ in experiments).

The final augmented adjacency matrix is defined as:

$$\tilde{A} = A + \alpha P^* \tag{10}$$

Here, P^* denotes the processed pseudo-edges, and α controls the adjacency enhancement strength. This enhanced graph representation serves as input for downstream modules.

3.3 Hyperparameter Tuning and Sensitivity Analysis

The model performance is influenced by several key hyperparameters, including: the labeling ratio r_l, which controls the number of labeled nodes participating in supervised training for each category; the anchor ratio r_a, which determines the proportion of labeled nodes selected as anchors for triplets and consequently affects the composition of triplet samples and the strength of training signals; and the loss weight α, used to balance the relative contributions of the triplet loss $\mathcal{L}_{\text{tri}}$ and the classification loss $\mathcal{L}_{\text{cls}}$ in the overall objective function. To systematically investigate the impact of these parameters and identify the optimal configuration, we conducted a series of controlled experiments.

All hyperparameter tuning experiments were conducted on the first field of view of the NanoString dataset. First, we evaluated the model's sensitivity to the labeling ratio within the range of $r_l \in \{0.10, 0.12, 0.14, 0.16, 0.18, 0.20\}$. As shown in Table 2, it can be observed that when labeled data is relatively limited ($r_l < 0.16$), the model performance is relatively low. Within the range of $r_l = 0.16 \sim 0.20$, the performance reaches a higher level, achieving the highest accuracy (77.57%) at $r_l = 0.18$. Therefore, $r_l = 0.18$ was fixed for subsequent experiments.

Based on this, we searched for the optimal anchor ratio within the range of $r_a \in \{0.05, 0.10, 0.15, 0.20, 0.25\}$. As shown in Table 3, the model achieved optimal performance when $r_a = 0.15$, while either too small or too large r_a values resulted in varying degrees of performance degradation.

Finally, we fix r_l and r_a to investigate the impact of loss weight α, with a candidate set of $\{0.1, 0.2, 0.3, 0.4, \text{w/o}\}$, where w/o denotes the control setting without using the loss weight balancing strategy. The results in Table 4 demonstrate that model performance degrades significantly when α is too small, while reaching its peak when α ranges between 0.3–0.4. Specifically, $\alpha = 0.4$ achieves the highest accuracy (78.11%), outperforming the w/o setting. Therefore, we determine $\alpha = 0.4$ as the optimal weight.

Table 2: Impact of marker ratio r_l on classification performance

r_l	**0.10**	**0.12**	**0.14**	**0.16**	**0.18**	**0.20**
Accuracy (%)	74.04	72.51	72.90	74.79	77.57	77.43

In conclusion, through systematic grid search, we determined the optimal hyperparameter combination for the model as: $r_l = 0.18$, $r_a = 0.15$, $\alpha = 0.4$. All subsequent comparative experiments and ablation studies adopted this configuration.

Table 3: Effect of anchor ratio r_a on classification performance

r_a	**0.05**	**0.10**	**0.15**	**0.20**	**0.25**
Accuracy (%)	75.23	77.12	77.39	76.78	75.57

Table 4: Sensitivity of weight loss α to classification performance

r_l	**0.10**	**0.20**	**0.30**	**0.40**	**w/o**
Accuracy (%)	21.28	19.59	78.03	78.11	76.36

3.4 Classification Performance

To highlight the challenges faced by semi-supervised learning in single-cell transcriptomics scenarios, we adopted a more stringent label sampling strategy (Disjoint mode). The core objective of this strategy is to maximize the spatial distance between labeled and unlabeled nodes within each category by concentrating labeled nodes in mutually non-overlapping spatial regions, thereby isolating a large number of unlabeled nodes from the labeled areas spatially (Cao et al. 2023). This setup simulates the common yet highly challenging local annotation scenario encountered in practical research, which significantly hinders the message-passing mechanism based on local neighborhoods. As a result, numerous unlabeled nodes become information islands, substantially increasing classification difficulty.

To visually demonstrate the specific impact of the disjoint pattern, we take FOV1 and FOV2 from the NanoString dataset as examples. The initial graph structure is constructed based on the traditional 7NN approach, as detailed in Fig 5 and 6. In the figures, nodes with borders represent annotated cells, while gray edges indicate the initially constructed, unweighted adjacency relationships. It can be clearly observed that a large number of unannotated nodes (without borders) are spatially isolated from the annotated regions.

Building upon this, we further visualize the learned structural information to analyze the structural correction effects brought by dynamic graph learning. To enhance visualization readability during graph reconstruction, we retain only the top 1,000 gray edges with highest weights, using edge thickness to represent weight differences. Meanwhile, red edges connect anchor points with their corresponding positive samples, reflecting key cell pairs guided by triplet constraints. Comparing Fig. 5 Fig. 6 with Fig. 7 Fig. 8 demonstrates that our model effectively reduces isolated nodes and corrects numerous erroneous or redundant connections present in the initial 7NN graph, thereby learning a clearer and more robust graph structure that better reflects underlying biological interactions.

To ensure the comprehensiveness and rigor of experimental evaluation, we also conducted comparative experiments under the more conventional random mode. In

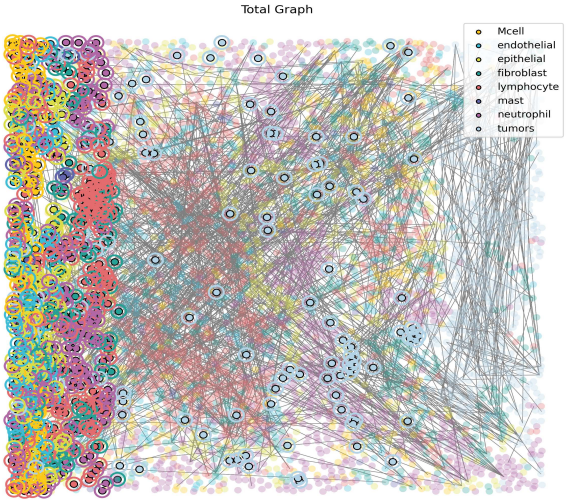


Fig. 5: 7NN graph of FOV1 under disjoint mode in the NanoString dataset

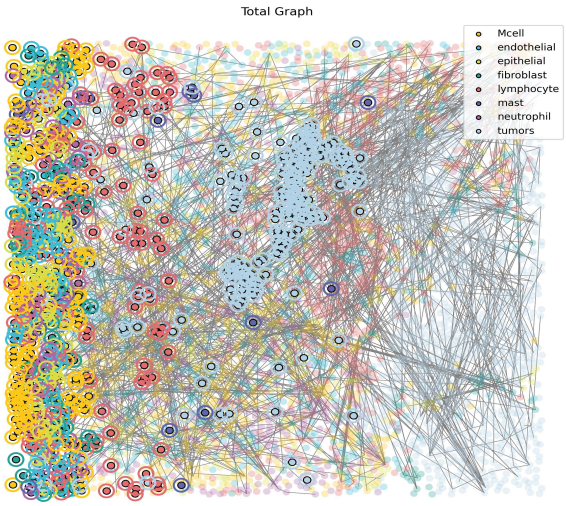


Fig. 6: 7NN graph of FOV2 under disjoint mode in the NanoString dataset

the random mode, labeled nodes of various categories are randomly distributed in the spatial domain, allowing supervision signals to more easily cover the entire area (Feng et al. 2023). Notably, our method maintains stable and excellent performance under both disjoint and random settings, further validating the model's generalization capability and adaptability across different label distributions.

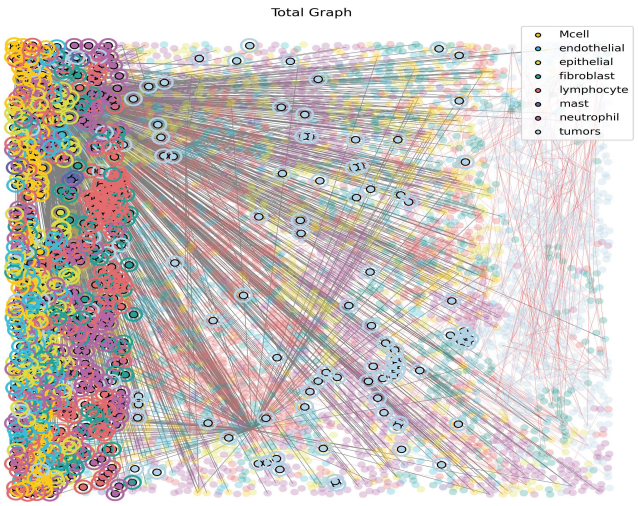


Fig. 7: Reconstructed graph of FOV1 under disjoint mode in the NanoString dataset

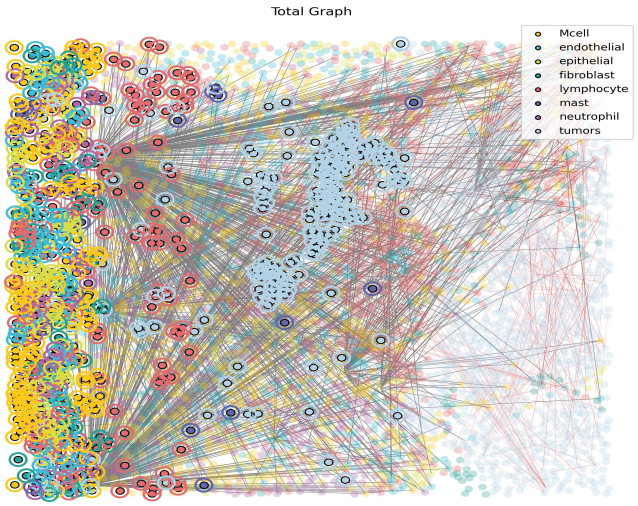


Fig. 8: Reconstructed graph of FOV2 under disjoint mode in the NanoString dataset

In terms of classification performance comparison, we comprehensively compared the proposed GraphTransformer-CoGNN with various state-of-the-art methods on two datasets: NanoString CosMx SMI and DLPFC. Including TransformerConv (Shi et al. 2021) and GAT (Veličković et al. 2018) based on standard graph message-passing mechanisms, SiGra (Tang et al. 2023) that integrates spatial morphological information, graph Transformer architectures scGraphformer (Fan et al. 2024) and

GraphTrans (Wu et al. 2021), as well as specialized semi-supervised models GraphCS (Zeng et al. 2022) and scSemiGCN (Yang et al. 2024). All experiments were conducted under unified data preprocessing and evaluation protocols, with performance measured by Accuracy and Cohen's Kappa under both random and disjoint annotation partitions.

Based on the data from Table 5 and Table 6, the following conclusions can be drawn:

Under the relatively friendly Random partitioning, the performance of all models improved, but our method still maintained its leading position. This indicates that even with relatively uniform labeling, our approach can consistently benefit from dynamic structure learning and refined information propagation.

Under either partitioning mode, our method demonstrates robust performance on both the NanoString and DLPFC datasets, which exhibit significant differences, thereby validating its strong cross-dataset generalization capability.

In summary, the experimental results indicate that GraphTransformer-CoGNN consistently delivers outstanding and stable cell type classification performance across varying levels of supervision scarcity and diverse spatial annotation distributions. This provides an effective solution to address key challenges in semi-supervised single-cell transcriptomic analysis.

Table 5: Comparison of classification accuracy among different models on the Nanostring dataset

Mode	Model	Accuracy (%)	Kappa
Disjoint	TransformerConv	74.62	67.31
Disjoint	GAT	77.20	70.10
Disjoint	SiGra	79.24	72.06
Disjoint	scGraphformer	81.45	74.23
Disjoint	GraphTrans	82.91	75.47
Disjoint	GraphCS	80.59	73.16
Disjoint	scSemiGCN	83.26	76.75
Disjoint	Ours	**84.70**	**78.42**
Random	TransformerConv	77.51	70.25
Random	GAT	80.09	73.51
Random	SiGra	81.46	74.12
Random	scGraphformer	83.62	76.65
Random	GraphTrans	84.91	77.61
Random	GraphCS	82.59	75.23
Random	scSemiGCN	85.61	77.97
Random	Ours	**87.01**	**80.04**

Under the highly challenging Disjoint partitioning, our method achieved optimal performance on both datasets, with higher accuracy and Kappa values than other models. This fully demonstrates that GraphTransformer-CoGNN can maintain strong

discriminative capability through two-stage dynamic graph learning, even under conditions of severely limited supervision signals and extremely uneven spatial distribution.

Table 6: Comparison of classification accuracy among different models on the DLPFC dataset

Mode	Model	Accuracy (%)	Kappa
Disjoint	TransformerConv	75.36	68.91
Disjoint	GAT	77.94	71.42
Disjoint	SiGra	78.53	72.43
Disjoint	scGraphformer	80.66	74.63
Disjoint	GraphTrans	81.88	76.15
Disjoint	GraphCS	79.66	73.73
Disjoint	scSemiGCN	82.08	76.29
Disjoint	Ours	**83.52**	**77.75**
Random	TransformerConv	77.18	70.96
Random	GAT	79.52	73.43
Random	SiGra	80.02	73.63
Random	scGraphformer	82.10	75.64
Random	GraphTrans	83.41	76.95
Random	GraphCS	81.19	74.83
Random	scSemiGCN	83.84	77.16
Random	Ours	**85.28**	**78.67**

3.5 Ablation Study

To assess the contribution of each key component in GraphTransformer-CoGNN, we conducted ablation experiments on the DLPFC dataset. The experiments employed disjoint annotation splits and were evaluated with 18% labeled node ratio, with results shown in Table 7.

The complete model consists of the following key designs: (1) Dynamic graph structure learning (DynGraph), (2) Triplet metric supervision (Triplet), (3) Resistance distance-based relative position encoding (RD-RPE), (4) Collaborative graph neural network module (CoGNN), (5) A unified Graph Transformer backbone for representation learning and classification.

We constructed a baseline version as a control, which removes all the aforementioned enhancement modules, retaining only a basic Transformer encoder with a classification head and is trained solely using cross-entropy loss. In this context, w/o indicates the removal of a specific component. It is important to emphasize that when the Triplet constraint is removed, the model's training supervision will rely solely on the node classification loss. This ablation setup ensures controlled and fair comparisons, thereby enabling a clear evaluation of each module's independent contribution to the final classification performance and validating the necessity and effectiveness of each component in the semi-supervised single-cell transcriptome cell type identification task.

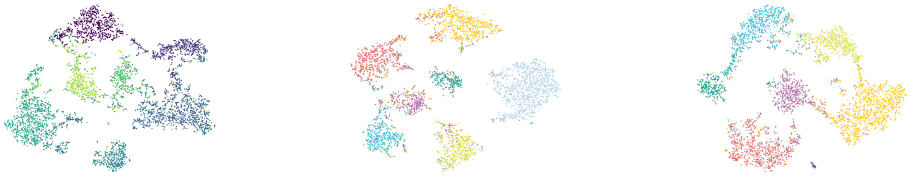
Overall, the ablation experiments validate the complementary nature of our framework: DynGraph provides a more reliable structural foundation, RD-RPE injects global structural priors, CoGNN further refines information flow, while Triplet reinforces discriminative structures in the representation space. These modules work synergistically to collectively ensure the stable and superior classification performance of GraphTransformer-CoGNN in extreme semi-supervised scenarios.

Table 7: Ablation study of the model on the DLPFC dataset with 18% labeled nodes under disjoint conditions

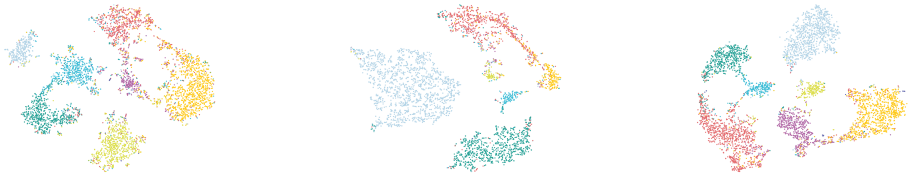
Variants	Accuracy (%)	Kappa
GraphTransformer w/o DynGraph	70.25±0.75	64.14±0.80
GraphTransformer w/o Triplet	72.46±1.01	69.05±0.91
GraphTransformer w/o RD-RPE	74.34±0.92	70.39±0.93
GraphTransformer w/o CoGNN	75.67±0.79	71.85±1.01
Our GraphTransformer	77.70±0.95	73.37±1.12

3.6 Feature Visualization

To intuitively demonstrate the effectiveness of the method from a graph representation learning perspective, we performed t-SNE visualization analysis using the node embeddings output by the final layer of our model on multiple different FOVs from the NanoString dataset, with results shown in Fig. 9(a)-(i). Different colors in the figures correspond to distinct cell types. Overall, in the vast majority of FOVs, major cell categories formed compact and well-separated cluster structures. Cells of the same type were tightly aggregated in the low-dimensional space, while clear boundaries were maintained between different cell categories, indicating that the node embeddings learned by our model possess high discriminative power and consistency. Even in tissues with more complex cell type compositions and higher spatial structural heterogeneity, the learned representations maintain clear inter-class separation without significant category confusion. Furthermore, for certain rare cell types with limited sample sizes or located at tissue margins, they still form relatively independent local clusters in the t-SNE plot and are neither obscured nor assimilated by the major cell populations. This demonstrates that our method can not only effectively distinguish major cell categories but also preserve the unique phenotypic information of rare and marginal cells, thereby enhancing the overall completeness of cell atlas construction. In summary, the consistent visualization results across multiple perspectives demonstrate that GraphTransformer-CoGNN can learn robust and highly discriminative cell representations, further validating its generalization capability for structural learning and feature modeling in complex single-cell transcriptomic data.



(a) FOV 1 (b) FOV 2 (c) FOV 3



(d) FOV 4 (e) FOV 6 (f) FOV 7



(g) FOV 8 (h) FOV 9 (i) FOV 10

Fig. 9: t-SNE feature visualization of multiple FOVs by NanoString

4 Conclusion

This paper presents GraphTransformer-CoGNN, a two-stage dynamic learning framework for semi-supervised single-cell transcriptome cell type classification. The framework first employs a spatially-aware graph transformer to learn input-dependent dynamic graph structures under the joint constraints of expressive features and spatial structure bias (RD-RPE), while introducing triplet metric supervision to enhance intra-class compactness and inter-class separability in the embedding space. Subsequently, the study introduces the CoGNN module, which performs node-level collaborative decision-making and dynamic optimization of edge weights on the learned graph structure, effectively suppressing noisy connections and reinforcing task-critical intercellular interactions. This method enhances the modeling capability of long-range dependencies and mitigates the accumulation of structural noise caused by static graph construction.

We conducted systematic evaluations on two real-world datasets, testing scenarios including random partitions and more challenging non-overlapping label partitions. Experimental results demonstrate that even with only 18% of labeled nodes, our method consistently outperforms various representative baseline methods in metrics such as accuracy and Kappa coefficient. This approach not only reliably identifies common cell types but also improves the detection sensitivity of rare cell subpopulations, providing an efficient and scalable tool for single-cell research.

The method holds significant promise for advancing the construction of tissue maps at single-cell resolution and elucidating intercellular.

Author contributions. **Yuanyuan Dang**: Conceptualization, Writing – review & editing. **Xinyi Han**: Conceptualization, Writing – review & editing. **Bing Liu**: Data curation.

Statements and Declarations

Competing interests. The authors declare that they have no competing interests.

References

Buterez, D., Bica, I., Tariq, I., Andrés-Terré, H., Liò, P.: Cellvgae: an unsupervised scRNA-seq analysis workflow with graph attention networks. Bioinformatics **38**(5), 1277–1286 (2022) https://doi.org/10.1093/bioinformatics/btab804

Berahmand, K., Daneshfar, F., Salehi, E., Li, Y., Xu, Y.: Autoencoders and their applications in machine learning: a survey. Artificial Intelligence Review **57**, 28 (2024) https://doi.org/10.1007/s10462-023-10662-6

Black, M., Wan, Z., Mishne, G., Nayyeri, A., Wang, Y.: Comparing graph transformers via positional encodings. In: Proceedings of the 41st International Conference on Machine Learning (ICML 2024). Proceedings of Machine Learning Research, vol. 235, pp. 4103–4139 (2024)

Cao, X., Li, C., Feng, J., Jiao, L.: Semi-supervised feature learning for disjoint hyperspectral imagery classification. Neurocomputing **526**, 9–18 (2023) https://doi.org/10.1016/j.neucom.2023.01.054

Dwivedi, V., Bresson, X.: A generalization of transformer networks to graphs. arXiv preprint arXiv:2012.09699 (2020). https://doi.org/10.48550/arXiv.2012.09699 . https://doi.org/10.48550/arXiv.2012.09699

Dai, E., Jin, W., Liu, H., Wang, S.: Towards robust graph neural networks for noisy graphs with sparse labels. In: Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining (WSDM 2022), pp. 181–191 (2022). https://doi.org/10.1145/3488560.3498408 . https://doi.org/10.1145/3488560.3498408

Dong, K., Zhang, S.: Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. Nature Communications **13**, 1739 (2022) https://doi.org/10.1038/s41467-022-29439-6

Ellens, W., Spieksma, F.M., Mieghem, P., Jamakovic, A., Kooij, R.E.: Effective graph resistance. Linear Algebra and its Applications **435**(10), 2491–2506 (2011) https://doi.org/10.1016/j.laa.2011.02.024

Finkelshtein, B., Huang, X., Bronstein, M., Ceylan, I.: Cooperative graph neural networks. In: Proceedings of the 41st International Conference on Machine Learning (ICML 2024). Proceedings of Machine Learning Research, vol. 235, pp. 13633–13659 (2024)

Fan, X., Liu, J., Yang, Y., Gu, C., Han, Y., Wu, B., Jiang, Y., Chen, G., Heng, P.-A.: scgraphformer: unveiling cellular heterogeneity and interactions in scRNA-seq data using a scalable graph transformer network. Commun Biol **7**, 1463 (2024) https://doi.org/10.1038/s42003-024-07154-w

Faber, L., Wattenhofer, R.: Asynchronous neural networks for learning in graphs. arXiv preprint arXiv:2205.12245 (2022). https://doi.org/10.48550/arXiv.2205.12245 . https://doi.org/10.48550/arXiv.2205.12245

Feng, H., Wang, Y., Li, Z., Zhang, N., Zhang, Y., Gao, Y.: Information leakage in deep learning-based hyperspectral image classification: a survey. Remote Sensing **15**(15), 3793 (2023) https://doi.org/10.3390/rs15153793

Feng, W., Zhang, J., Dong, Y., Han, Y., Luan, H., Xu, Q., Yang, Q., Kharlamov, E., Tang, J.: Graph random neural networks for semi-supervised learning on graphs. In: Advances in Neural Information Processing Systems, vol. 33 (2020). https://doi.org/10.5555/3495724.3497577 . arXiv:2005.11079. https://doi.org/10.5555/3495724.3497577

Gutteridge, B., Dong, X., Bronstein, M., Di, G.F.: Drew: dynamically rewired message passing with delay. In: Proceedings of the 40th International Conference on Machine Learning (ICML 2023). Proceedings of Machine Learning Research, vol. 202, pp. 12252–12267 (2023)

Ge, S., Sun, S., Xu, H., Cheng, Q., Ren, Z.: Deep learning in single-cell and spatial transcriptomics data analysis: advances and challenges from a data science perspective. Briefings in Bioinformatics **26**(2), 136 (2025) https://doi.org/10.1093/bib/bbaf136

He, S., Bhatt, R., Brown, C., *et al.*: High-plex imaging of RNA and proteins at subcellular resolution in fixed tissue by spatial molecular imaging. Nature Biotechnology **40**(12), 1794–1806 (2022) https://doi.org/10.1038/s41587-022-01483-z

Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017). https://doi.org/10.48550/

arXiv.1703.07737 . https://doi.org/10.48550/arXiv.1703.07737

Hendrycks, D., Gimpel, K.: Gaussian error linear units (GELUs). arXiv preprint arXiv:1606.08415 (2016). https://doi.org/10.48550/arXiv.1606.08415 . https://doi.org/10.48550/arXiv.1606.08415

Hu, J., Li, X., Coleman, K., *et al.*: SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. Nature Methods **18**, 1342–1351 (2021) https://doi.org/10.1038/s41592-021-01255-8

Hamilton, W., Ying, Z., Leskovec, J.: Inductive representation learning on large graphs. In: Advances in Neural Information Processing Systems, vol. 30, pp. 1024–1034 (2017). https://doi.org/10.48550/arXiv.1706.02216 . https://doi.org/10.48550/arXiv.1706.02216

Jang, E., Gu, S., Poole, B.: Categorical reparameterization with gumbel-softmax. In: Proceedings of the International Conference on Learning Representations (ICLR 2017) (2017). https://doi.org/10.48550/arXiv.1611.01144 . https://doi.org/10.48550/arXiv.1611.01144

Jiang, B., Zhang, Z., Lin, D., Tang, J., Luo, B.: Semi-supervised learning with graph learning-convolutional networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2019), pp. 11313–11320 (2019). https://doi.org/10.1109/CVPR.2019.01157 . https://doi.org/10.1109/CVPR.2019.01157

Kumar, A., Schrader, A., Aggarwal, B., *et al.*: Intracellular spatial transcriptomic analysis toolkit (instant). Nature Communications **15**, 7794 (2024) https://doi.org/10.1038/s41467-024-49457-w

Kipf, T., Welling, M.: Semi-supervised classification with graph convolutional networks. In: Proceedings of the International Conference on Learning Representations (ICLR 2017) (2017). https://doi.org/10.48550/arXiv.1609.02907 . https://doi.org/10.48550/arXiv.1609.02907

Liu, R., Hou, X., Liu, S., Zhou, Y., Zhou, J., Qiao, K., *et al.*: Predicting antidepressant response via local-global graph neural network and neuroimaging biomarkers. NPJ Digit Med **8**(1), 515 (2025) https://doi.org/10.1038/s41746-025-01912-8

Li, X., Wang, C.-Y.: From bulk, single-cell to spatial RNA sequencing. Int J Oral Sci **13**(1), 36 (2021) https://doi.org/10.1038/s41368-021-00146-0

Li, Y., Yin, J., Chen, L.: Informative pseudo-labeling for graph neural networks with few labels. Data Mining and Knowledge Discovery **37**, 228–254 (2023) https://doi.org/10.1007/s10618-022-00879-4

Lai, K.-H., Zha, D., Zhou, K., Hu, X.: Policy-GNN: aggregation optimization for graph neural networks. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD 2020), pp. 461–471 (2020). https://doi.org/10.1145/3394486.3403088 . https://doi.org/10.1145/3394486.3403088

Maynard, K., Collado-Torres, L., Weber, L., *et al.*: Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. Nat Neurosci **24**(3), 425–436 (2021) https://doi.org/10.1038/s41593-020-00787-0

Maddison, C., Mnih, A., Teh, Y.: The concrete distribution: a continuous relaxation of discrete random variables. In: Proceedings of the International Conference on Learning Representations (ICLR 2017) (2017). https://doi.org/10.48550/arXiv.1611.00712 . https://doi.org/10.48550/arXiv.1611.00712

Morris, C., Ritzert, M., Fey, M., Hamilton, W., Lenssen, J., Rattan, G., Grohe, M.: Weisfeiler and leman go neural: higher-order graph neural networks. Proc AAAI Conf Artif Intell **33**(1), 4602–4609 (2019) https://doi.org/10.1609/AAAI.V33I01.33014602

Oliveira, M., Romero, J., Chung, M., *et al.*: High-definition spatial transcriptomic profiling of immune cell populations in colorectal cancer. Nat Genet **57**(6), 1512–1523 (2025) https://doi.org/10.1038/s41588-025-02193-3

Pinheiro, J., Oliveira, S.d., Silva, T., Saraiva, P., Souza, E.d., Godoy, R., Ambrosio, L., Becker, M.: The impact of feature scaling in machine learning: effects on regression and classification tasks. IEEE Access **13**, 199903–199931 (2025) https://doi.org/10.1109/ACCESS.2025.3635541

Pardo, B., Spangler, A., Weber, L., *et al.*: spatiallibd: an R/Bioconductor package to visualize spatially-resolved transcriptomics data. BMC Genomics **23**(1), 434 (2022) https://doi.org/10.1186/s12864-022-08601-w

Shi, Y., Huang, Z., Feng, S., Zhong, H., Wang, W., Sun, Y.: Masked label prediction: unified message passing model for semi-supervised classification. In: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21), pp. 1548–1554 (2021). https://doi.org/10.24963/ijcai.2021/214 . https://doi.org/10.24963/ijcai.2021/214

Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: a unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015), pp. 815–823 (2015). https://doi.org/10.1109/CVPR.2015.7298682 . https://doi.org/10.1109/CVPR.2015.7298682

Shaw, P., Uszkoreit, J., Vaswani, A.: Self-attention with relative position representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume

2 (Short Papers). Association for Computational Linguistics,, pp. 464–468 (2018). https://doi.org/10.18653/v1/N18-2074 . https://doi.org/10.18653/v1/N18-2074

Shirzad, H., Velingker, A., Venkatachalam, B., Sutherland, D., Sinop, A.: Exphormer: sparse transformers for graphs. In: Proceedings of the 40th International Conference on Machine Learning (ICML 2023). Proceedings of Machine Learning Research, vol. 202, pp. 31613–31632 (2023)

Shehzad, A., Xia, F., Abid, S., Peng, C., Yu, S., Zhang, D., Verspoor, K.: Graph Transformers: A Survey. arXiv preprint arXiv:2407.09777 (2024). https://doi.org/10.48550/arXiv.2407.09777 . https://doi.org/10.48550/arXiv.2407.09777

Tang, Z., Li, Z., Hou, T., Zhang, T., Yang, B., Su, J., Song, Q.: SiGra: single-cell spatial elucidation through an image-augmented graph transformer. Nature Communications **14**, 5618 (2023) https://doi.org/10.1038/s41467-023-41437-w

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y.: Graph attention networks. In: International Conference on Learning Representations (2018). https://doi.org/10.48550/arXiv.1710.10903 . https://doi.org/10.48550/arXiv.1710.10903

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30, pp. 5998–6008 (2017). https://doi.org/10.48550/arXiv.1706.03762 . https://doi.org/10.48550/arXiv.1706.03762

Wang, Y.-R., Du, P.-F.: Wcsgnet: a graph neural network approach using weighted cell-specific networks for cell-type annotation in scRNA-seq. Frontiers in Genetics **16**, 1553352 (2025) https://doi.org/10.3389/fgene.2025.1553352

Wu, Z., Jain, P., Wright, M., Mirhoseini, A., Gonzalez, J., Stoica, I.: Representing long-range context for graph neural networks with global attention. In: Advances in Neural Information Processing Systems, vol. 34, pp. 13266–13279 (2021). https://doi.org/10.48550/arXiv.2201.08821 . https://doi.org/10.48550/arXiv.2201.08821

Xu, J., Sun, X., Zhang, Z., Zhao, G., Lin, J.: Understanding and improving layer normalization. In: Advances in Neural Information Processing Systems, vol. 32, pp. 4383–4393 (2019). https://doi.org/10.48550/arXiv.1911.07013 . https://doi.org/10.48550/arXiv.1911.07013

Ying, C., Cai, T., Luo, S., Zheng, S., Ke, G., He, D., Shen, Y., Liu, T.: Do transformers really perform bad for graph representation? In: Advances in Neural Information Processing Systems, vol. 34, pp. 28877–28888 (2021). https://doi.org/10.5555/3540261.3542473 . https://doi.org/10.5555/3540261.3542473

Yang, Y., Cui, Y., Zeng, X., Zhang, Y., Loza, M., Park, S., Nakai, K.: Staig: spatial transcriptomics analysis via image-aided graph contrastive learning for domain

exploration and alignment-free integration. Nature Communications **16**, 1067 (2025) https://doi.org/10.1038/s41467-025-56276-0

Yang, J., Wang, W., Zhang, X.: scsemiGCN: boosting cell-type annotation from noise-resistant graph neural networks with extremely limited supervision. Bioinformatics **40**(2), 091 (2024) https://doi.org/10.1093/bioinformatics/btae091

Zhang, Z., Lu, S., Huang, Z., Zhao, Z.: Graph neural networks with adaptive structures. IEEE J Sel Top Signal Process **19**(1), 181–194 (2025) https://doi.org/10.1109/JSTSP.2024.3485892

Zhang, B., Luo, S., Wang, L., He, D.: Rethinking the expressive power of GNNs via graph biconnectivity. In: Proceedings of the Eleventh International Conference on Learning Representations (ICLR 2023) (2023). https://doi.org/10.48550/arXiv.2301.09505 . https://doi.org/10.48550/arXiv.2301.09505

Zeng, Y., Wei, Z., Pan, Z., Lu, Y., Yang, Y.: A robust and scalable graph neural network for accurate single-cell classification. Brief Bioinform **23**(2), 570 (2022) https://doi.org/10.1093/bib/bbab570

Zhou, Z., Zhou, S., Mao, B., Zhou, X., Chen, J., Tan, Q., Zha, D., Feng, Y., Chen, C., Wang, C.: Opengsl: A comprehensive benchmark for graph structure learning. In: Advances in Neural Information Processing Systems, vol. 36 (2023). https://doi.org/10.48550/arXiv.2306.10280 . NeurIPS 2023 Datasets and Benchmarks Track. https://doi.org/10.48550/arXiv.2306.10280

Zhang, J., Zhang, H., Xia, C., Sun, L.: Graph-BERT: only attention is needed for learning graph representations. arXiv preprint arXiv:2001.05140 (2020). https://doi.org/10.48550/arXiv.2001.05140 . https://doi.org/10.48550/arXiv.2001.05140