GraphFusionNN: an AI-Driven Framework for Context-Aware Cell Type Prediction Using Graphs, Spatial Features, and Embeddings

Faith A. Shim
Department of Computer Science and Biology, Brown University

Primary Advisor: Dr. Ying Ma
Nominal Advisor: Professor Sorin Istrail

A Thesis submitted in partial fulfillment of the requirements for Honors in the Department of Computer Science at Brown University

Department of Computer Science
Brown University
Providence, Rhode Island
April 2025

1 TABLE OF CONTENTS

TABLE OF FIGURES

1 ABSTRACT

We present GraphFusionNN, a multi-modal neural network designed to integrate biological signals from single-cell and spatial transcriptomics data. GraphFusionNN unifies graph-based topological structures, proxy spatial features, and context-aware transcriptomic embeddings into a cohesive predictive framework for robust cell type classification.

This model flexibly incorporates pre-trained embeddings such as those from scGPT or scFoundation, serving as high-dimensional representations of gene expression profiles. These components are seamlessly integrated within a modular architecture that supports efficient experimentation across embedding strategies. Combined with graph connectivity and spatial proximity features, these modalities are fused through a unified integration module that supports cross-attention, MLP, cross-attention, gated, and concatenation mechanisms. This design enables GraphFusionNN to adaptively learn from multiple biological domains—molecular, spatial, and structural—while ensuring extensibility for future enhancements.

Experiments across diverse single-cell and spatial datasets demonstrate that our fusion strategy—integrating graph topology, spatial context, and external embeddings—significantly improves classification accuracy, macro-F1, and robustness compared to single-modality models. Notably, the model excels in settings with imbalanced or overlapping cell populations, where unimodal approaches such as scGPT may struggle.

GraphFusionNN supports both precomputed and jointly fine-tuned embeddings, offering scalable integration under varying computational resources. Its modular and lightweight design provides a resource-efficient alternative to full model fine-tuning, positioning GraphFusionNN as a powerful and extensible foundation for multi-modal representation learning in single-cell omics.

2 INTRODUCTION

Single-cell omics has revolutionized biological research by enabling high-resolution exploration of cellular heterogeneity, molecular dynamics, and gene regulatory networks. However, comprehensively understanding complex biological systems requires the integration of multiple omics modalities, including transcriptomics, and spatial data. A key challenge lies in effectively fusing these diverse data sources to extract meaningful insights while preserving both molecular and spatial contexts.

Recent advancements in transformer-based models, such as scGPT [1], have demonstrated strong performance in capturing transcriptomic relationships and predicting cell types from large-scale single-cell datasets. These models leverage pretraining strategies to learn gene-expression patterns, making them robust to missing data. However, scGPT is inherently focused on transcriptomic features and lacks explicit spatial modeling, limiting its ability to capture cell-cell interactions necessary for understanding tissue-level biological processes.

Spatial transcriptomics has emerged as a powerful tool for studying cellular organization and interactions within tissue microenvironments. Traditional computational approaches, such as KNN-based spatial graphs and PCA-derived embeddings, attempt to approximate spatial structure but often lack biological fidelity. Meanwhile, Graph Neural Networks (GNNs) have demonstrated significant promise in modeling spatial relationships by leveraging neighborhood aggregation techniques. Landmark studies, such as Kipf and Welling's Graph Convolutional Networks (GCNs) [2] and Velickovic et al.'s Graph Attention Networks (GATs) [3], have laid the foundation for integrating spatial data into predictive models. Within single-cell analysis, GNNs have been applied to model cellular neighborhoods [4] and infer cell-cell communication networks [5]. However, the integration of GNNs with large-scale transformer models such as scGPT remains largely unexplored, leaving much room for improvement in effectively capturing both transcriptomic and spatial relationships within a unified framework.

Despite advancements in single-cell multi-omics integration, existing approaches fail to effectively unify GNNs (spatial relationship modeling), scGPT (large-scale transcriptomic learning), and additional multimodal inputs within a cohesive framework. To address this gap, we introduce GraphFusionNN, a novel fusion framework designed to integrate GNNs, scGPT embeddings, proxy spatial representations, real spatial data, and scFoundation embeddings, enabling a comprehensive and adaptive approach to multi-omics data analysis.

Our framework integrates several key components to model complex cellular and spatial relationships. It leverages Graph Neural Networks (GCN and GAT) to capture spatial patterns and cell-cell interactions. For representing high-dimensional gene expression data, it incorporates pre-trained embeddings such as scGPT and scFoundation. In the absence of true spatial coordinates, the framework supports proxy spatial representations using methods such as KNN, Grid, UMAP, and PCA to approximate spatial structures. When available, real spatial data can be incorporated to reflect the true organization of tissue architecture.

This modular design of GraphFusionNN allows easy switching between precomputed and jointly trained embedding strategies. Precomputed embeddings enable fast, resource-efficient training, while joint fine-tuning supports end-to-end learning under more flexible compute settings.

To support a flexible and robust integration of multi-modal single-cell data, GraphFusionNN implements three core fusion strategies—Cross-Attention, MLP, and GCN—each capturing different aspects of biological context.

Cross-Attention computes dynamic attention scores across modalities (e.g., gene, graph, spatial), enabling the model to learn inter-modality dependencies. MLP fusion projects each modality through dense layers with learnable weights and blends them in a shared embedding space, offering a

lightweight and standard baseline. GCN-based integration introduces topological awareness by encoding local neighborhood information, making it especially useful for spatially structured or microenvironment-sensitive datasets. Together, these strategies form a modular fusion framework that can be tailored to the structure and complexity of each dataset.

This approach enables us to identify the optimal combination of inputs for high performance, recognize inputs that do not contribute to improvements, and understand the underlying reasons for their impact.

We evaluate GraphFusionNN across diverse single-cell and spatial transcriptomics datasets. Our results show that integrating spatial topology, transcriptomic embeddings, and graph structure significantly improves classification accuracy and biological interpretability. The framework's plug-and-play architecture enables users to incorporate or replace components with minimal overhead, making it suitable for large-scale single-cell applications.

In this paper, we present the GraphFusionNN model, fusion strategies, and empirical evaluations. We demonstrate that the joint modeling of spatial, graph, and transcriptomic signals leads to robust, interpretable, and scalable cell type classification. GraphFusionNN offers a practical and flexible foundation for multi-modal representation learning in single-cell biology.

3 MULTI-MODAL FUSION FRAMEWORK DESIGN

The GraphFusionNN framework is designed to flexibly integrate heterogeneous biological signals for robust single-cell and spatial transcriptomic analysis. A central component of our framework is its ability to incorporate transformer-based transcriptomic embeddings from scGPT in two distinct operational modes: Option A, precomputed embeddings (frozen), and Option B, end-to-end fine-tuning.

In Option A (Precomputed/Frozen Embeddings), scGPT is used solely as a feature extractor. Embeddings are generated once from raw gene expression profiles using a pre-trained scGPT model and then held fixed throughout the training of GraphFusionNN. This structure simplifies training and avoids costly computations associated with backpropagating through the transformer. As a result, this mode offers a faster, more resource-efficient solution with a strong baseline performance and is particularly well-suited for large-scale experiments or limited-GPU environments. ***Figure 1*** shows the frozen-embedding design (Option A).

Figure 1: GraphFusionNN with Precomputed/Frozen Embeddings (Option A)

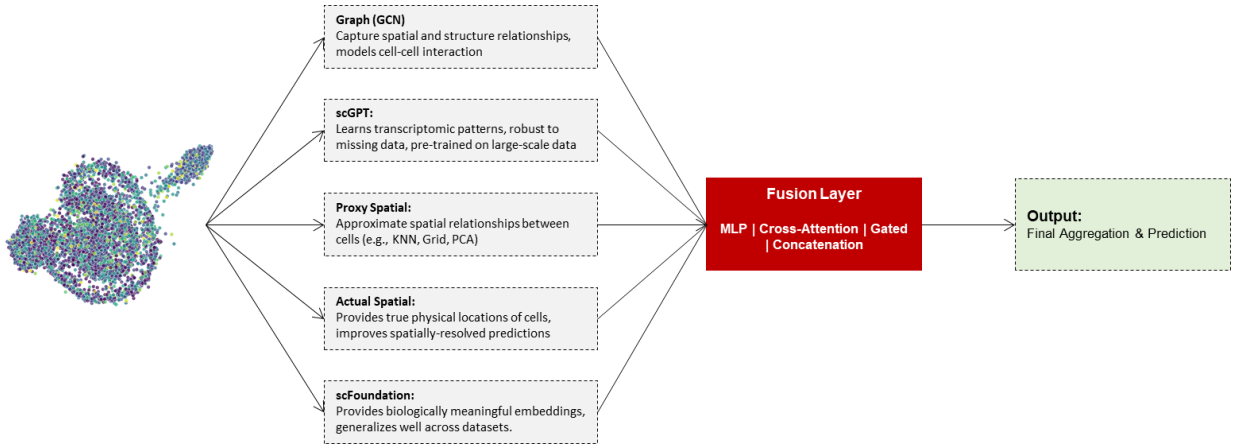
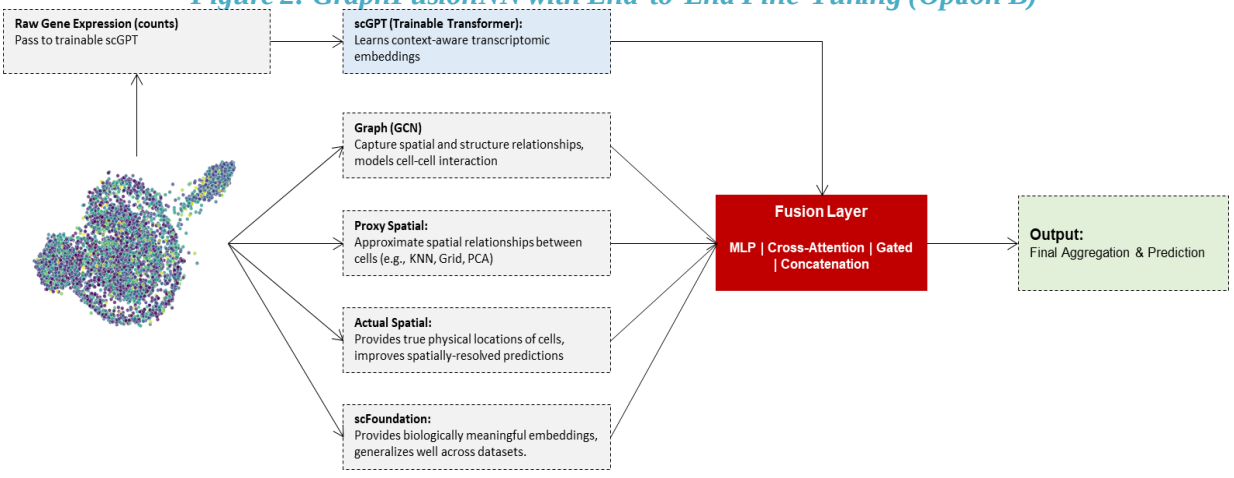


Figure 2: GraphFusionNN with End-to-End Fine-Tuning (Option B)



In contrast, Option B (End-to-End Fine-Tuning) activates the full transformer learning pipeline. Instead of holding the scGPT embeddings constant, the raw gene expression matrix is passed into the trainable scGPT model, whose parameters are jointly optimized with those of GraphFusionNN during training. This means that gradients are backpropagated through both the fusion network and the scGPT

transformer layers, allowing the model to adapt transcriptomic embeddings specifically to the structure of the dataset, task objectives, and fusion strategy. In effect, the scGPT encoder becomes a fully integrated component of the predictive model.

This end-to-end approach enables context-aware, task-adaptive embedding learning, offering enhanced performance and greater flexibility. However, it comes at the cost of increased memory usage and longer training times due to the inclusion of high-dimensional transformer layers in gradient computation. ***Figure 2*** illustrates the end-to-end fine-tuning flow (Option B), where scGPT is actively trained as part of the model pipeline.

By supporting both modes, GraphFusionNN offers users a choice between speed and specialization, making it a versatile platform for both rapid prototyping and high-performance multi-modal representation learning.

3.1 CHARACTERISTICS OF INPUT FEATURES

Note: In this context, gene expression refers to high-dimensional count matrices that capture the number of transcripts (e.g., RNA molecules) per gene per cell. Embeddings are low-dimensional learned representations derived from these counts that summarize underlying biological patterns, enabling more effective integration and prediction in downstream models.

GraphFusionNN integrates a diverse set of input features designed to capture complementary biological signals at the molecular, spatial, and structural levels. These features form the core components of our multi-modal fusion pipeline and are visualized in ***Figure 3***.

Figure 3: Input Feature Characteristics

Graph Feature

- **Definition**: Captures relationships between nodes using gene expression similarity.
- **Derived by**: Gene expression graph using KNN graph or adjacency matrix (e.g., adata.obsp["connectivities"]).
- **Edge Connection**: Edges are based on gene expression similarity (cells with similar gene expression profiles are linked)
- **Learnable**: Uses GCN/GAT to refine graph embeddings
- **Use Case**: Helps identify functionally similar cells (even if they are not physically close).

scGPT Feature

- **Definition**: Captures transcriptomic relationships by leveraging transformer-based pretraining on large-scale single-cell datasets.
- **Derived by**: Gene expression profiles, learned through deep transformer architecture.
- **Learnable**: Uses self-attention mechanisms to extract transcriptomic features, making it robust to missing data.
- **Use Case**: Enables accurate cell-type annotation, gene expression imputation, and feature extraction by capturing global transcriptomic patterns

Spatial Feature

- **Definition**: Captures spatial proximity relationships between cells (physical locations).
- **Derived by**: Physical spatial coordinates (adata.obsm["spatial"]) or custom embeddings (PCA/Grid/KNN) as spatial proxies.
- **Edge Connection**: Edges are based on spatial distances (cells that are physically close are linked)
- **Learnable**: Can be trainable (trainable_spatial=True) or static input
- **Use Case**: Helps understand tissue organization (cells closer together in real space should behave similarly)

scFoundation Feature

- **Definition**: Captures both spatial and non-spatial gene expression patterns, acting as a complementary feature extractor to scGPT.
- **Derived by**: Pre-trained embeddings from large-scale single-cell datasets, designed for transfer learning in transcriptomics.
- **Learnable**: Provides deep embeddings that can generalize across datasets, offering biologically meaningful feature representations.
- **Use Case**: Helps improve model generalization and robustness by integrating prior biological knowledge into feature extraction.

GraphFusionNN leverages four major categories of input features, Graph, Spatial, scGPT, and scFoundation, with each capturing a unique dimension of biological variation. These feature types are modular and configurable, allowing for a flexible model design that can be tailored to different data types and analytical goals.

Graph features capture topological relationships between cells based on gene expression similarities. Derived from adjacency matrices, such as k-nearest neighbor (KNN) graphs or connectivity matrices, these features enable modeling of cell–cell interactions through GCN or GAT mechanisms. They are particularly useful for identifying functionally similar cells that may not be physically adjacent in space.

Spatial features encode the physical proximity between cells using either real spatial coordinates or proxy representations, including PCA, UMAP, or KNN-derived layouts. Depending on the `trainable_spatial` setting, these features can be either static or trainable. By incorporating spatial context, they allow the model to better account for tissue architecture and cellular neighborhoods in downstream predictions.

scGPT features provide high-dimensional embeddings learned from raw gene expression using a transformer-based model. These embeddings are robust to noise and missing data and can be configured as static (precomputed) or trainable (through end-to-end fine-tuning). They support context-aware modeling of transcriptomic patterns across the dataset, offering a powerful representation of cellular states.

Lastly, scFoundation features offer complementary biological embeddings that are pre-trained on large-scale, multi-organ datasets. These features enhance model generalizability and robustness by encoding biological priors learned across diverse conditions. Like scGPT, scFoundation embeddings can be used as either static or trainable inputs, contributing an additional layer of biological insight to the model.

These input modalities are processed and integrated in the Fusion Layer, which adaptively combines them using MLP weighting, cross-attention, gated fusion, or simple concatenation. This modular, AI-powered fusion mechanism enhances predictive performance while maintaining model interpretability and scalability across diverse datasets.

3.2 GRAPHFUSIONNN ARCHITECTURE AND MULTI-MODAL LEARNING PIPELINE

GraphFusionNN is designed as a fully modular and extensible architecture, supporting a comprehensive multi-modal learning pipeline tailored for single-cell and spatial transcriptomics. Unlike prior models with rigid input pathways, GraphFusionNN enables dynamic configuration and flexible experimentation with a wide array of modality combinations, fusion mechanisms, and embedding strategies. This adaptability positions it as a unified framework capable of addressing diverse biological questions under varying computational constraints.

The architecture allows researchers to toggle between real and proxy spatial features, enable or disable graph connectivity layers (GCN), and incorporate context-aware embeddings from pretrained models such as scGPT and scFoundation. Two embedding strategies are supported for scGPT: a frozen mode using precomputed embeddings, and a trainable mode where the embeddings are fine-tuned jointly with the classifier. Users can select among four fusion strategies—cross-attention, MLP, gated or simple concatenation—depending on the complexity and resource budget of their task. Furthermore, GraphFusionNN allows for trainable integration of spatial features, enhancing adaptability to datasets with varying spatial structure.

3.2.1 Data Loading and Modal Construction

The model's pipeline begins with structured loading of `.h5` or `.h5ad` through a unified preprocessing function. This process extracts gene expression counts, spatial coordinates, and graph relationships, such as neighborhood graphs using `adata.obsp` or k-nearest neighbor graphs. If ground-truth cell type annotations are absent, the pipeline performs Leiden clustering to assign surrogate labels. A balanced split of training and validation sets is ensured through stratified sampling, maintaining proportional cell-type representation.

Spatial features are constructed using both real coordinates (when available) and proxy approximations like PCA, UMAP, or KNN graphs. These spatial graphs are created with tunable parameters such as `k_neighbors` and `distance_cutoff`. Additionally, class imbalance is addressed through configurable class weighting schemes, including none, direct, balanced, soft, and logarithmic scaling. These

components establish a robust and reproducible foundation for downstream model training and evaluation.

3.2.2 Modular Architecture and Embedding Strategies

GraphFusionNN's core model is composed of modular components designed to take in and process multiple modalities. Graph-based input is encoded using GCN layers that learn neighborhood-aware topological features. The spatial encoder accepts real or proxy spatial features and, when `trainable_spatial=True`, applies a dedicated GCN to learn latent spatial structures.

Transcriptomic data is handled through a lightweight transformer encoder when `use_scgpt=True`. The scGPT module processes gene sequences to produce contextual embeddings that reflect gene-gene relationships. These embeddings can either be frozen—precomputed externally and treated as static input—or jointly trained with the rest of the network, allowing them to adapt to specific datasets during training. When available, scFoundation embeddings are passed through a projection layer to align with other feature dimensions before fusion.

The fusion layer combines encoded graph, spatial, and transcriptomic features using one of four configurable strategies: cross-attention, MLP, gated fusion, or concatenation. Each strategy offers a distinct mechanism for integrating modality-specific signals. The prediction head applies a fully connected linear layer to the fused representation to produce cell type logits. To support efficient and stable training, the pipeline includes mixed-precision training, learning rate scheduling, early stopping, and model check-pointing.

3.2.3 Fusion Mechanisms and Multi-Modal Integration

GraphFusionNN supports four distinct strategies for integrating multi-modal features. MLP fusion learns adaptive weights for each modality through fully connected layers, enabling parametric blending. Cross-attention fusion captures inter-modality dependencies by computing attention scores through multi-head attention mechanisms. Gated fusion introduces a learnable gating controller that dynamically balances outputs from the MLP and attention pathways. Finally, concatenation fusion, though non-parametric, serves as a baseline by directly merging feature vectors without any learnable transformation.

The model supports user-defined switching between fusion modes, enabling researchers to tradeoff between computational simplicity and integration sophistication. This makes it possible to tailor the fusion approach based on dataset complexity, desired model interpretability, or available GPU resources.

3.2.4 Evaluation and Prediction Performance

GraphFusionNN includes a comprehensive evaluation framework to measure classification performance across diverse input configurations. Standard classification metrics—such as accuracy, precision, recall, and F1 scores (macro and micro)—are computed on the validation or test sets. Confusion matrices are used to inspect misclassification patterns, and classification reports summarize label-wise performance.

To ensure consistent evaluation across datasets, the model aligns predicted indices with ground-truth label names and supports remapping in cases where label spaces do not match. The evaluation module is fully modular, capable of assessing both GraphFusionNN and hybrid models that include scGPT or scFoundation embeddings. Stored model checkpoints can be reloaded for reproducibility and benchmarking.

In addition to numerical metrics, the evaluation stage outputs key visualizations such as confusion matrices, label distribution plots, and formatted reports suitable for publication. This holistic evaluation

process allows users to interpret and compare different modality combinations, fusion strategies, and embedding configurations.

3.2.5 Interpretability and Performance Analysis

To provide insight into how GraphFusionNN makes predictions and integrates different data modalities, we implemented a suite of interpretability tools. Dimensionality reduction techniques like UMAP and t-SNE are applied to visualize cell embeddings at various stages—raw inputs, scGPT outputs, and final fused representations—highlighting the clustering structure and separability of cell types.

Cluster composition analysis quantifies how well the model preserves biological groupings during prediction, while confusion matrices and classification reports offer a class-wise breakdown of model accuracy. Additional plots showing label distributions and gene expression statistics provide context about class imbalance and expression sparsity.

These interpretability tools support deeper analysis of model performance, guiding the selection of optimal fusion strategies for specific datasets and helping users identify potential limitations or failure modes in their prediction pipeline.

3.3 IMPLEMENTATION OF MULTI-MODAL FUSION IN GRAPHFUSIONNN

GraphFusionNN is implemented as a modular and extensible PyTorch-based framework, purpose-built to support seamless integration of heterogeneous single-cell data modalities. Its architecture enables plug-and-play configuration of graph-based structures, spatial features, and transcriptomic embeddings, allowing researchers to experiment with various fusion strategies, embedding modes, and spatial encodings efficiently. This section details how key components—input encoders, adaptive fusion modules, and output heads—interact within the system to support scalable, high-performance cell type prediction.

3.3.1 Encoding Multi-Modal Inputs: Graph, Spatial and Embedding Features

GraphFusionNN accepts five types of inputs that collectively represent structural, molecular, and spatial dimensions of single-cell data. First, graph-based representations are constructed from adjacency matrices derived using either k-nearest neighbor (KNN) graphs or spatial distance thresholds. These graphs are then encoded through Graph Convolutional Networks (GCNs) to capture neighborhood-level dependencies. Second, scGPT-based embeddings are used to encode transcriptomic profiles. These embeddings may either be precomputed and frozen (in Option A) or trained end-to-end as part of the network (in Option B). Third, proxy spatial approximations—such as PCA, UMAP, or regular grid layouts—serve to provide latent spatial structure when direct coordinates are unavailable. Fourth, real spatial coordinates, when provided by spatial transcriptomics platforms, offer biologically accurate locational context. Finally, scFoundation embeddings can be optionally added as biologically pretrained features, which are passed through a projection layer before being fused with other inputs.

3.3.2 Modular Fusion Layer with Adaptive Integration Strategies

At the core of GraphFusionNN is a configurable fusion module designed to integrate diverse input modalities. The Cross-Attention strategy applies multi-head attention to explicitly model dependencies and information flow across modalities, enabling richer and context-aware integration. The Multi-Layer Perceptron (MLP) strategy uses fully connected layers to project each modality into a shared space, learning modality-specific transformations and adaptive weights. The Gated Fusion strategy introduces a trainable gating mechanism that dynamically balances contributions from MLP and attention outputs based on input relevance. Finally, the Concatenation strategy provides a non-parametric baseline by

directly merging input features without any learned transformation, serving as a control for evaluating more sophisticated fusion techniques. This modular design allows GraphFusionNN to adapt its integration strategy to match the structure and complexity of each dataset.

3.3.3 Output Projection with Cell Type Prediction

Once the fusion module has integrated the multi-modal representations, the resulting features are passed through the output head for cell type classification. This head includes a transformation layer that adjusts the fused feature dimensions to match the classifier input, followed by dropout for regularization and ReLU activations to introduce non-linearity and prevent overfitting. The final classification layer is a fully connected linear projection that maps the processed features to the target number of cell types, producing class logits for downstream loss calculation. This architecture supports both standard and weighted cross-entropy loss functions, enabling the model to handle imbalanced datasets effectively.

4 EXPERIMENTS AND RESULTS

We evaluate the performance of GraphFusionNN across multiple single-cell and spatial transcriptomics datasets to assess its predictive accuracy, generalizability, and adaptability across diverse biological contexts. Our experiments are designed to systematically analyze the effect of multi-modal fusion, spatial encodings, graph-based topologies, and embedding strategies (precomputed vs. fine-tuned) on cell type classification. We also compare different fusion mechanisms and conduct ablation studies to understand the contribution of each modality.

To ensure reproducibility and fair comparison, we use consistent preprocessing pipelines, stratified train-validation splits, and multiple evaluation metrics including accuracy, macro F1-score, precision, and recall. Visualizations such as UMAP projections and confusion matrices further support the interpretability of the model's learned representations.

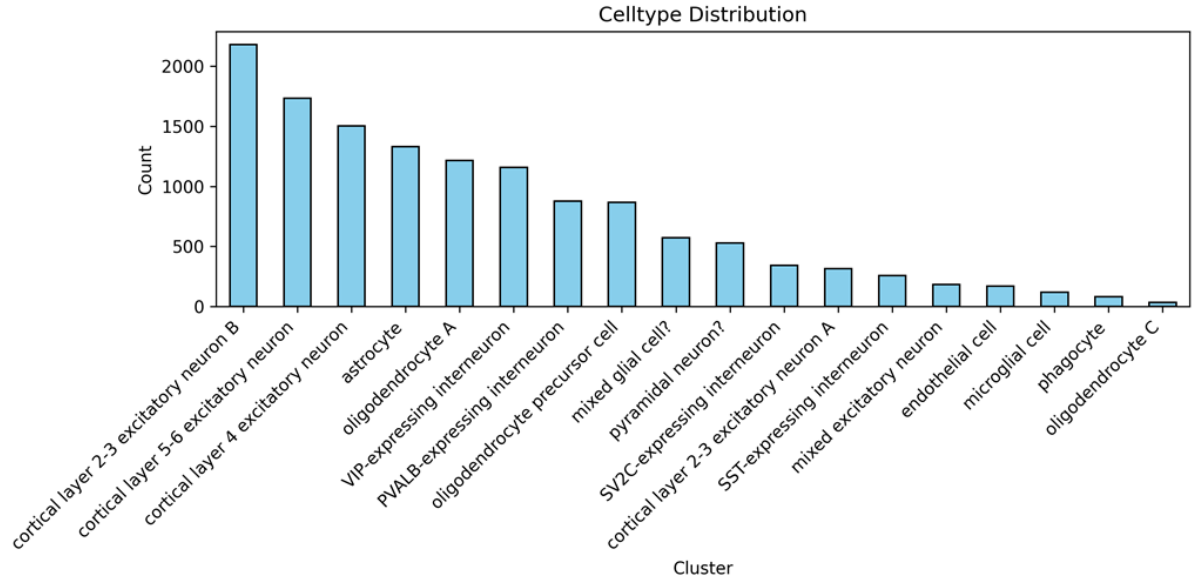
4.1 DATASET CHARACTERISTICS

We conduct experiments on a collection of publicly available and biologically diverse single-cell and spatial transcriptomics datasets. These datasets include varying degrees of complexity in terms of tissue types, cell heterogeneity, spatial resolution, and labeling quality. Each dataset is preprocessed using a consistent pipeline that includes gene filtering, normalization, optional dimensionality reduction, and label assignment. The datasets used in this study are available in the References section, which includes the dataset of PBMC (Peripheral Blood Mononuclear Cells) [7], Human Lymph Node [8], Breast Cancer (FFPE Human Breast Tumor) [9], FFPE Human Lung Tissue [10], and Multiple Sclerosis (M.S.) [11].

4.1.1 Cell Class Distribution Analysis

To better understand the cellular diversity within the Multiple Sclerosis (M.S.) dataset [11] as an example, we analyzed the distribution of annotated cell types based on available biological labels (`celltype` in `adata.obs`). As shown in ***Figure 4***, the dataset exhibits a heterogeneous mix of neuronal and glial cell types.

Figure 4: Cell Type Distribution Analysis



Notably, three dominant excitatory neuron cell types—cortical layer 2-3, layer 4, and layer 5-6 excitatory neurons—comprise the majority of cells. These cell types play critical roles in cortical signaling and are highly relevant in neuroinflammatory conditions, such as multiple sclerosis. Among the glial subtypes, astrocytes, oligodendrocyte type A, and oligodendrocyte precursor cells are prominently represented, consistent with the demyelination pathology characteristic of the disease.

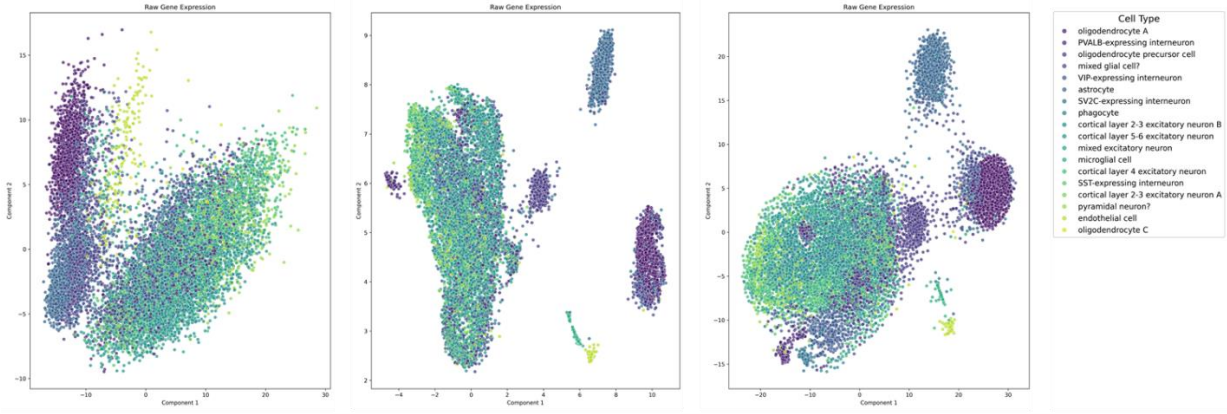
Less frequent, but biologically important, cell types, such as microglial cells, phagocytes, and endothelial cells, are present in smaller numbers. These provide valuable insight into the immune and vascular activity within the context of the disease.

This uneven distribution highlights the importance of robust handling of class imbalance during model training. Accordingly, GraphFusionNN incorporates dynamic class weighting to ensure that underrepresented cell types receive appropriate attention during learning and are not neglected in prediction.

4.1.2 Embedding-Based Feature Space Visualization

To explore the intrinsic structure of cell populations in the Multiple Sclerosis (M.S.) dataset [11], we also visualize the raw gene expression features using three widely-used dimensionality reduction techniques: PCA, t-SNE, and UMAP, as shown in ***Figure 5***.

Figure 5: Visualizing the Raw Gene Expression Features with PCA, UMAP, t-SNE



(a) PCA Raw Gene Expression (b) UMAP Raw Gene Expression (c) t_SNE Raw Gene Expression

Each cell is colored according to its ground-truth cell type annotation (`adata.obs["celltype"]`), enabling biologically meaningful interpretation of cell type distinction.

PCA provides a linear projection of the expression space, capturing overall variance. While PCA reveals some general separation, cell type clusters appear partially overlapping, suggesting limited representation of fine-grained transcriptional differences.

t-SNE preserves local neighborhoods, highlighting clearer groupings of similar cell types. Some cell types form well-defined clusters, reflecting the local structure of the gene expression space.

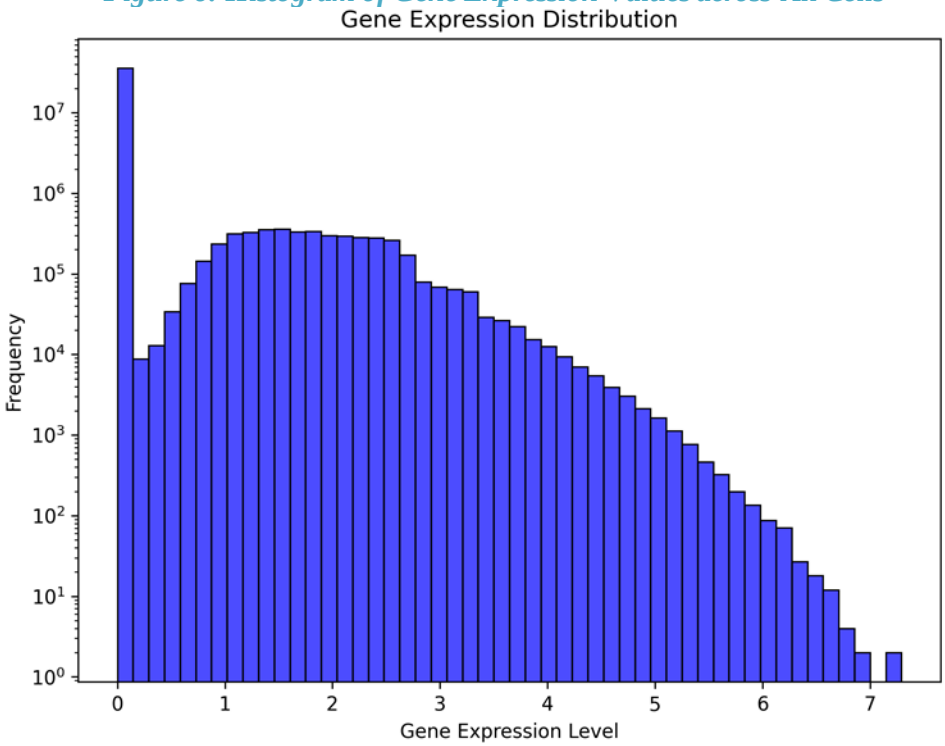
UMAP provides the most distinct separation among cell types, preserving both local and global relationships. The UMAP plot reveals compact, distinctly separated clusters corresponding to unique biological cell types, supporting its strength in capturing non-linear structure in high-dimensional space.

These visualizations indicate that raw gene expression captures meaningful transcriptional variation across cell types. However, the separation remains suboptimal, highlighting the need for multi-modal fusion approaches such as GraphFusionNN, which leverage graph-based topology, spatial proximity, and context-aware embeddings to enhance biological classification accuracy.

4.1.3 Gene Expression Distribution

To further analyze the underlying characteristics of the dataset, we examine the overall distribution of gene expression values across all cells. This is visualized through a histogram that plots expression frequency across gene expression levels.

Figure 6: Histogram of Gene Expression Values across All Cells



The gene expression distribution is computed using the `adata.X` matrix, which contains the expression levels of each gene in every cell. Depending on the preprocessing pipeline, these values may have been transformed through methods such as log1p normalization.

The X-Axis, Gene Expression Level, represents the expression level of individual genes across the dataset. If log-transformed, a value of 0 indicates the gene is not expressed in the corresponding cell. Higher values denote increasing levels of gene activity. These values are binned to form a continuous histogram. This axis effectively bins expression values into ranges (e.g., 0–0.2, 0.2–0.4 ...), so we can count the frequency of these ranges.

The Y-Axis, Frequency, displays the frequency distribution of gene expression values across the entire dataset, including all genes and cells. Given that expression matrices typically contain millions of data points, the y-axis is shown on a logarithmic scale to reveal patterns spanning several orders of magnitude. For example, a bar with a height of 10^7 indicates that approximately 10 million values fall within that specific expression range (e.g., near zero), while smaller bars—such as those with heights around 10 or 100—represent rare high-expression events.

From the resulting histogram (***Figure 6***), we observe a very high peak near 0, indicating that a large number of genes are not expressed in most cells, highlighting the sparsity of the dataset. The concentration of values between log-expression levels of 1 to 3 indicates that most genes exhibit moderate expression. The extended right tail highlights a smaller subset of highly expressed genes.

This is consistent with typical single-cell RNA-seq datasets, where dropout effects and biological sparsity are common. Understanding this distribution helps shape preprocessing, normalization, and modeling strategies for downstream analysis.

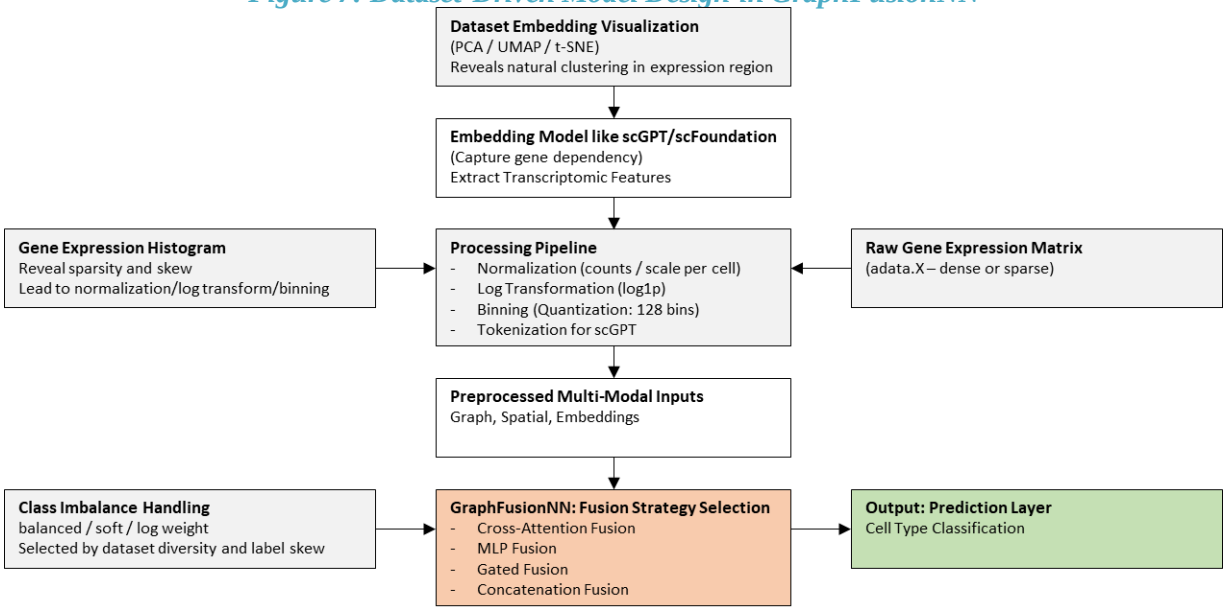
4.1.4 Implications for Model Design

The following dataset analysis procedures directly influenced the modular architecture and learning strategies adopted in GraphFusionNN: Cell Type Distribution, Embedding Structure, and Gene Expression Landscape.

Cell Type Distribution (Section 4.1.1)

Leveraging the diversity of cell types and transcriptomic structure, GraphFusionNN incorporates a flexible fusion strategy layer. As illustrated in ***Figure 7***, this layer enables dynamic selection among cross-attention, MLP, gated, and concatenation. Each fusion method models inter-modality dependencies in a distinct way, improving feature alignment across heterogeneous inputs such as scGPT embeddings, graph-based topologies, and spatial features. To further address class imbalance, GraphFusionNN integrates configurable loss-level weighting strategies—such as `"balanced"`, `"log"`, and `"soft"` reweighting—which ensure that underrepresented cell types receive proper gradient updates during training. Together, the adaptive fusion layer and class reweighting contribute to improved classification robustness and generalization, especially in imbalanced datasets.

Figure 7: Dataset-Driven Model Design in GraphFusionNN



Embedding Structure (Section 4.1.2)

The PCA, t-SNE, and UMAP plots reveal that cells naturally cluster based on their gene expression profiles. These clusters often align with known cell types, even in the absence of spatial or graph information. This indicates that transcriptomic data alone already contain strong biological signals. Embedding models like scGPT, which are trained to capture gene-gene relationships and contextual dependencies, amplify these signals and provide more informative features. By capturing expression context, scGPT embeddings improve input quality for downstream fusion in GraphFusionNN, leading to more accurate and biologically meaningful predictions.

Gene Expression Landscape (Section 4.1.3)
The log-scaled distribution of gene expression values is skewed towards low-expression genes, underscoring the importance of robust preprocessing. In our implementation, we apply normalization and log1p transformation to stabilize variance, followed by binning (e.g., 128 discrete levels) to handle sparsity and reduce noise. These preprocessing steps, integrated into the scGPT tokenization pipeline, help preserve biologically relevant signals while minimizing the influence of low-abundance or zero-expression noise. This preprocessing is essential to support downstream attention-based and non-linear fusion strategies, which selectively emphasize informative signals during multi-modal integration.

4.2 GRAPHFUSIONNN CONFIGURATION AND DATASETS

To efficiently evaluate the capability of multi-modal fusion, we configured the GraphFusionNN model with a flexible set of options that control modality inclusion, spatial processing, and fusion strategies. These configurations enabled fine-grained ablation studies and allowed the model to adapt to datasets with varying biological complexity.

Multi-Modal Input Settings

The GraphFusionNN model was instantiated with specific modality flags to configure its behavior across data types. Transcriptomics input was enabled using scGPT embeddings (with `USE_SCGPT = True`), which served as the core representation of gene expression in all hybrid experiments. Graph structure modeling was incorporated through k-nearest neighbor (kNN) graph features (`USE_GRAPH = True`) to capture relationships among cells. The optional scFoundation second-layer model was disabled (`USE_SCFOUNDATION = False`) to isolate and analyze the contributions of scGPT and spatial features.

Spatial Feature Settings

To integrate spatial context, proxy spatial features were activated using kNN-based graphs (`USE_PROXY_SPATIAL = "knn"`), allowing the model to infer spatial relationships even without true coordinates. In contrast, actual spatial features were disabled for this study (`USE_ACTUAL_SPATIAL = False`), although the model supports them. Additionally, trainable spatial layers were enabled (`TRAINABLE_SPATIAL = True`), leveraging graph convolutional network (GCN) layers to refine spatial feature learning.

Fusion Strategy

We used attention-based fusion (`USE_FUSION_WEIGHTS = "attention"`), which dynamically learns modality importance for each sample. This strategy promotes robustness and interpretability by suppressing noisy modalities and emphasizing informative ones.

Class Imbalance Handling
We applied balanced class weights (`CLASS_WEIGHT = "balanced"`) to address imbalance in single-cell labels, which is especially beneficial for underrepresented cell types.

Modality Ablation Table

The following table summarizes the model variants and configurations used in our experiments:

Model Variant	scGPT	Graph	Proxy Spatial	Fusion	Purpose
scGPT Only	Used	Not Used	Not Used	–	Baseline model that uses only gene expression data (transcriptomics) to

					establish reference performance.
Graph Only	Not Used	Used	Not Used	None	Ablation study to isolate and assess the predictive power of spatial graph structure alone, without transcriptomic or spatial proxy input.
Proxy Spatial Only	Not Used	Not Used	Used	None	Ablation study to evaluate the effect of spatial proximity features independently of graph or gene expression data.
GraphFusionNN	Used	Used	Used	Attention	Full multi-modal hybrid model combining transcriptomics, spatial proximity, and graph structure with attention-based fusion.
GraphFusionNN (No scGPT)	Not Used	Used	Used	Attention	Variant that excludes gene expression data to measure the contribution of graph and spatial modalities only.
GraphFusionNN (No Graph)	Used	Not Used	Used	Attention	Variant that excludes graph structure to assess the role of transcriptomics and spatial proximity in model performance.
GraphFusionNN (No Spatial)	Used	Used	Not Used	Attention	Variant that excludes spatial proximity information, retaining only gene expression and graph structure to test the spatial modality's impact.

Datasets

We benchmarked GraphFusionNN across five publicly available single-cell RNA-seq datasets from the cellxgene portal, encompassing both tumor microenvironment and immune-related biological contexts.

RESULT	DATASET ID	TISSUE	REFERENCE
Figure 8	HTAPP-982-SMP-7629	Breast cancer	[11]
Figure 9	HTAPP-944-SMP-7479	Breast cancer	[12]
Figure 10	HTAPP-364-SMP-1321	Breast cancer	[13]
Figure 11	Immune Cells	Blood/Immune	[14]
Figure 12	Bladder Stromal	Urinary Bladder	[15]

These variants were used to evaluate the relative contributions of each modality across diverse datasets and to demonstrate the effectiveness of the proposed fusion strategies.

4.3 SCGPT PERFORMANCE

In this section, we systematically evaluate the fine-tuned scGPT model across a diverse set of single-cell RNA-seq datasets, examining its ability to accurately classify cell types under varying biological conditions. First, we highlight datasets where the model performs well, exhibiting clear cluster boundaries and high generalization accuracy. Next, we explore edge cases where scGPT struggles to resolve subtle transcriptomic differences or overlapping cell populations. Through quantitative metrics and qualitative embeddings, we aim to identify both the strengths and limitations of scGPT, and to motivate the integration of complementary modalities in subsequent sections.

4.3.1 Performance of scGPT

The fine-tuned scGPT model demonstrated strong convergence behavior across training, validation, and evaluation sets (*Figure 8*, right). Training accuracy consistently increased over 20 epochs, surpassing 99%, while validation accuracy stabilized at approximately 94–95% after epoch 10. This suggests effective generalization with minimal overfitting. The evaluation accuracy, computed on a stratified, 10% test subset withheld during training and validation, reached 94.1%, indicating robust model performance on unseen data.

In addition, we visualized the learned latent representations using UMAP (*Figure 8*, left), based on the final hidden states from the scGPT encoder. The resulting embedding space reveals distinct and well-separated clusters corresponding to annotated cell types, confirming that the model preserves biologically relevant structures in the high-dimensional gene expression space.

To further characterize performance, a classification report was computed on the evaluation set, highlighting per-class accuracy, precision, recall, and F1-scores. However, the macro-averaged F1-score was notably lower than the accuracy, indicating that while scGPT performs well overall, its predictions may be biased toward dominant classes. Specifically, major cell populations exhibited strong F1-scores, while rarer or transcriptionally similar cell types showed reduced recall and precision—suggesting challenges in resolving minority or ambiguous classes. These findings highlight a key limitation of accuracy alone and underscore the importance of macro-F1 as a more balanced metric for performance evaluation in imbalanced biological datasets.

Figure 8: Breast Cancer Predication I (clearly isolated cells)

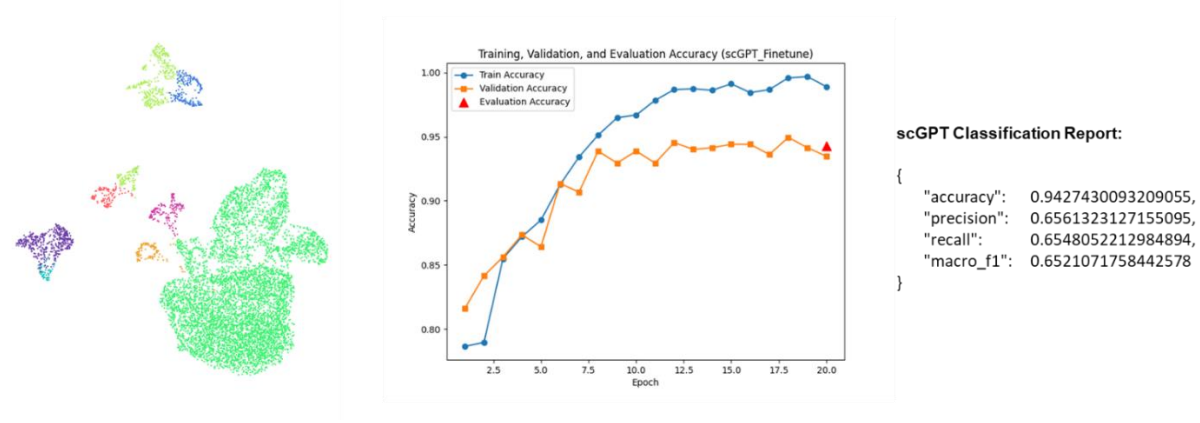
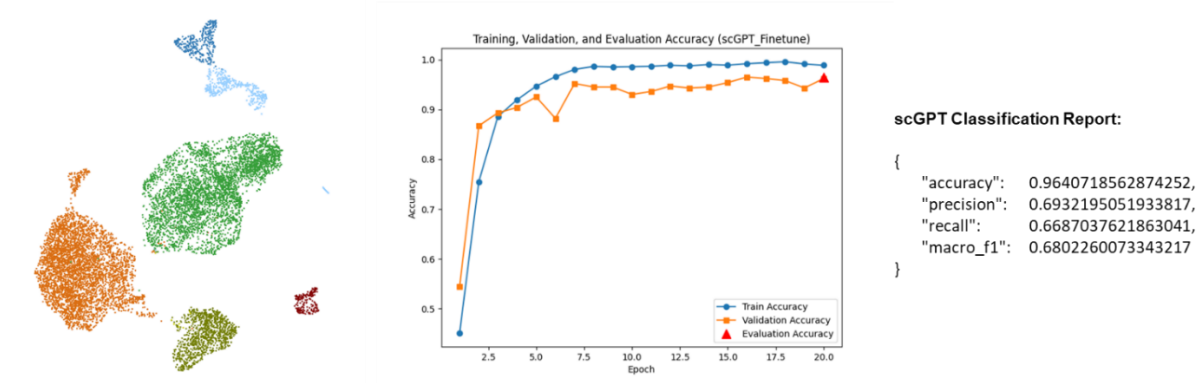


Figure 9: Breast Cancer Predication II (clearly isolated cells)

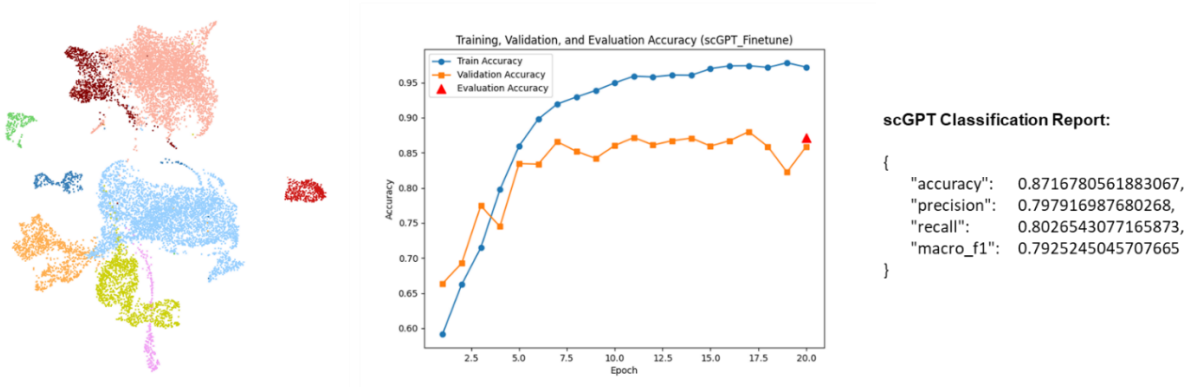


Following fine-tuning on the new dataset, the scGPT model achieved rapid convergence within the first 5 epochs, as shown by steep improvements in both training and validation accuracy (***Figure 9***, right). Training accuracy plateaued near 99%, while validation accuracy stabilized around 95%, indicating effective learning with minimal overfitting. The final evaluation accuracy, measured on an independent test split, reached 96.4%, confirming strong generalization. The UMAP visualization of the learned scGPT embeddings (***Figure 9***, left) revealed distinct, non-overlapping clusters corresponding to known cell types, suggesting that the model successfully differentiated between transcriptomic variations and preserved biological separability.

The classification report for this experiment provides a detailed breakdown of the model's performance across all annotated cell types for this dataset. Dominant clusters achieved relatively high F1-scores, while minor or transcriptionally similar populations showed lower precision and recall—likely reflecting challenges due to class imbalance and overlapping gene expression profiles. The macro-averaged F1-score of 0.68, as reported, indicates moderate overall performance, highlighting the model's partial success in generalizing across a diverse cellular landscape.

As shown in the right panel (***Figure 10***), training accuracy steadily increased to 97% over 20 epochs, while validation accuracy plateaued around 87%, suggesting moderate overfitting toward later epochs. Despite this, the final evaluation accuracy on the 10% held-out test set reached 86.8%, indicating strong generalization to unseen data. Notably, the performance gap between training and validation trajectories underscores the increased difficulty of the classification task compared to previous datasets.

Figure 10: Breast Cancer Predication III (slightly overlapped cells)



The UMAP visualization of the learned gene embeddings (***Figure 10***, left) revealed discrete clusters corresponding to major annotated cell types, with a few overlapping or closely positioned subpopulations. This suggests that while the model successfully captured dominant transcriptional signatures, subtle differences among similar or rare cell types remain challenging. The classification report supports this interpretation, showing a macro-averaged F1-score of ~0.79 and an overall accuracy of ~87.2%. While the model performs well on average, the precision (0.80) and recall (0.80) indicate variability across classes, with likely reductions in performance for underrepresented or transcriptionally ambiguous cell types. This is consistent with the UMAP visualization, which reveals partially overlapping regions. Still, the results suggest that the model maintains a reasonable level of consistency across diverse cellular phenotypes.

In summary, the fine-tuned scGPT model exhibited strong performance when the underlying dataset displayed well-separated clusters, effectively capturing dominant transcriptomic patterns across major cell types. Although classification performance declined for rare or transcriptionally similar cell populations, the overall macro-averaged F1-score reflected moderate generalization performance,

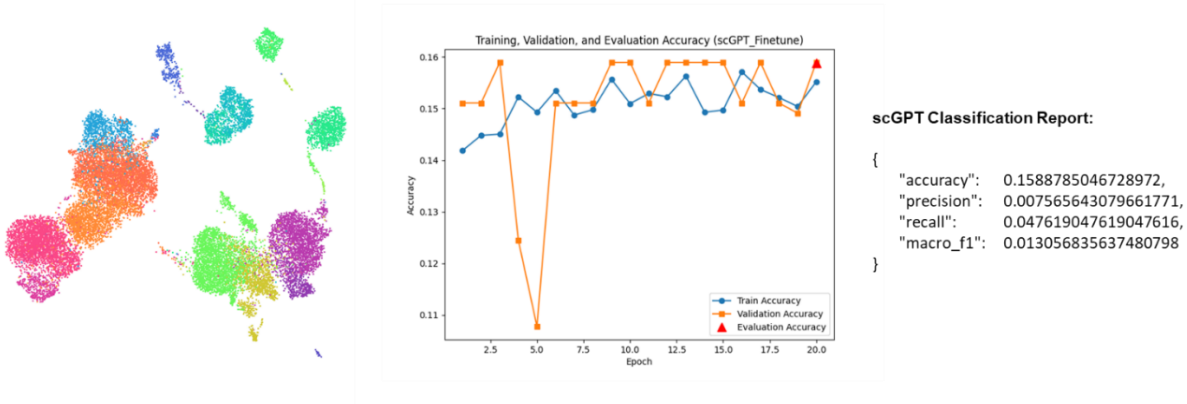
suggesting that scGPT captures key biological structure while leaving room for improvement in handling less distinct cell types.

4.3.2 Limitations of scGPT

In contrast to previous datasets, the fine-tuned scGPT model exhibited limited predictive performance on this immune cell dataset (***Figure 11***). Training and validation accuracies remained low across epochs, stabilizing around ~15%, indicating that the model struggled to extract meaningful discriminative features. The evaluation accuracy also reflected this challenge, peaking at only ~15.8%. This plateau suggests that the model was unable to learn a robust latent representation for this dataset. The UMAP visualization of the learned embeddings (***Figure 11***, left) revealed overlapping clusters with diffuse boundaries, suggesting poor transcriptomic separability between annotated cell types.

The lack of a distinct structure in the learned latent space likely contributed to the model's difficulty in generalizing. These findings highlight the limitations of scGPT when applied to datasets with minimal transcriptional separation or subtle inter-cell type variations.

Figure 11: Immune Cells (overlapped cells)



As illustrated in ***Figure 12***, the fine-tuned scGPT model exhibited stagnant learning behavior on this dataset. Training and validation accuracies plateaued around ~52.8% after only a few epochs, with negligible improvement throughout the rest of the training cycle. The evaluation accuracy matched this trend, indicating a failure to generalize beyond random or majority-class predictions. Despite a UMAP projection (***Figure 12***, left) showing some degree of cluster separability, the learned embeddings did not result in meaningful classification boundaries. This discrepancy suggests that while low-dimensional structure is present, the model failed to capture discriminative transcriptomic features required for accurate cell type classification. These findings further highlight the model's sensitivity to dataset characteristics and suggest that superficial clustering in the latent space may not necessarily translate into effective downstream prediction.

Figure 12: Bladder Stromal (overlapped cells)

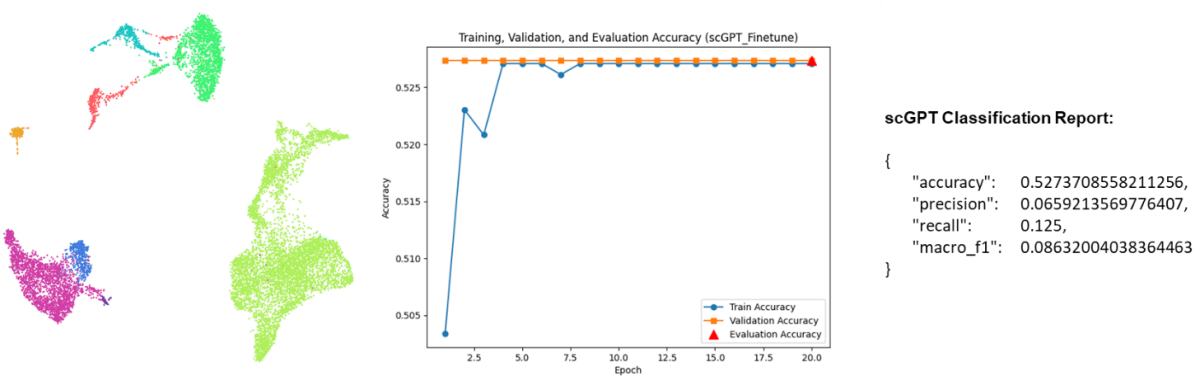
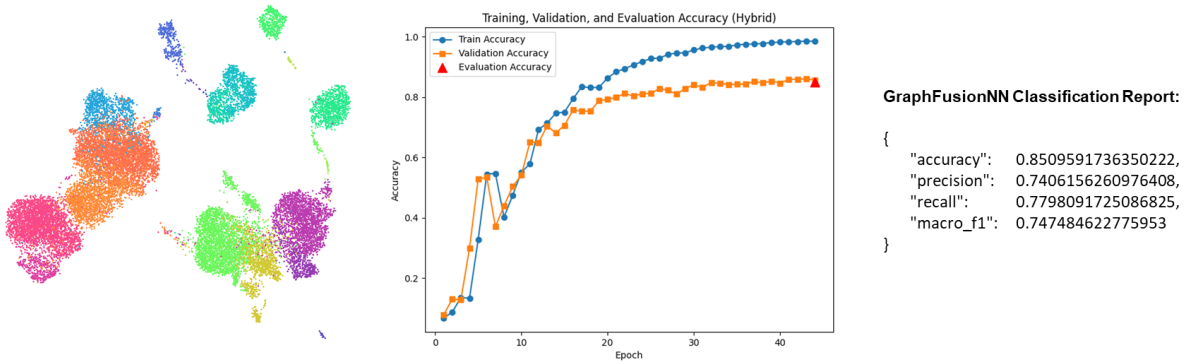


Figure 11 and *Figure 12* highlight cases where the fine-tuned scGPT model failed to achieve meaningful learning, with training, validation, and evaluation accuracies remaining low and stagnant throughout. Despite some degree of visual clustering in the UMAP embeddings, the model struggled to extract discriminative transcriptomic features, particularly in datasets with overlapping or poorly separable cell type profiles. These results highlight the model's limitations when working with datasets characterized by weak signals, ambiguous class boundaries, or imbalanced data distributions.

4.4 GRAPHFUSIONNN PERFORMANCE

Figure 13 demonstrates a substantial performance improvement achieved by integrating GraphFusionNN in place of the standalone scGPT model shown in *Figure 11*. While scGPT alone struggled to extract meaningful representations—yielding near-random classification performance—the hybrid GraphFusionNN architecture effectively fused transcriptomic, spatial, and graph-derived features, resulting in robust convergence across training, validation, and evaluation subsets. The model achieved a final evaluation accuracy of 85%, a drastic increase compared to the ~15% baseline in *Figure 11*. The UMAP visualization of the learned hybrid embeddings reveals significantly enhanced cluster separation, indicating that GraphFusionNN successfully differentiated between overlapping cell types that scGPT could not resolve in isolation. These results highlight the importance of integrating multi-modal context for challenging datasets with subtle or ambiguous transcriptional structures.

Figure 13: Performance Improvement by GraphFusionNN with scGPT (Immune Cells)

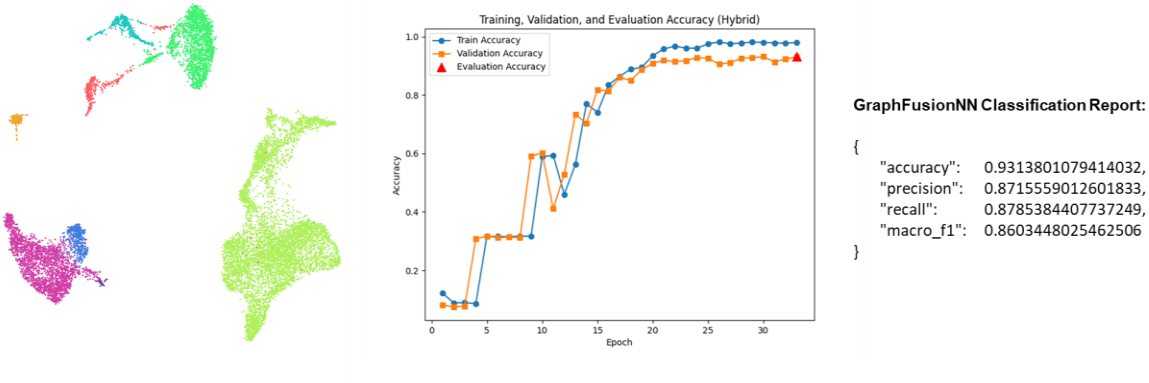


This substantial performance improvement observed with GraphFusionNN stems from its ability to integrate complementary modalities—namely, scGPT embeddings, local neighborhood structure (graph topology), and proxy spatial context—to construct a richer and more discriminative feature space. In datasets like the one shown in ***Figure 11***, where transcriptomic signals alone are ambiguous or insufficient to delineate cell types, the standalone scGPT model struggles to resolve overlapping or transcriptionally similar populations. By contrast, GraphFusionNN leverages graph structures, proxy spatial information, and scGPT embeddings in order to capture connectivity patterns and local relationships among cells (e.g., k-nearest neighbors in gene expression space), encode approximate tissue organization or spatial proximity, and provide powerful sequence-level representations of gene activity.

The combination of these modalities allows the model to differentiate the subtle differences between cell types that may be transcriptionally similar but spatially or structurally distinct. The model's cross-attention fusion mechanism dynamically weighs each modality's contribution, enabling adaptive feature integration based on context. This multi-view representation learning is particularly beneficial in complex biological settings where no single modality is sufficient to fully characterize cellular identity.

In contrast to the poor performance observed when using scGPT alone on the same dataset (***Figure 12***), the hybrid GraphFusionNN model yielded a substantial improvement in accuracy, with evaluation performance rising from approximately 52% to 93% (***Figure 14***, right). This performance gain can be attributed to GraphFusionNN's ability to integrate multiple complementary modalities—including transcriptomic embeddings from scGPT, proxy spatial features, and graph-based neighborhood information. While gene expression alone may fail to sufficiently differentiate cell types with overlapping or subtle transcriptional differences, the additional structural and spatial context helps clarify these relationships. This underscores the critical role of multi-modal fusion in enhancing classification performance, particularly in challenging scenarios where transcriptomic signals alone are insufficient for robust cell-type discrimination. The UMAP visualization (***Figure 14***, left) further supports this, showing improved cluster separation and alignment with known cell type annotations, reflecting the model's enhanced capacity to capture biologically meaningful patterns.

Figure 14: Performance Improvement by GraphFusionNN with scGPT (Bladder Stromal)

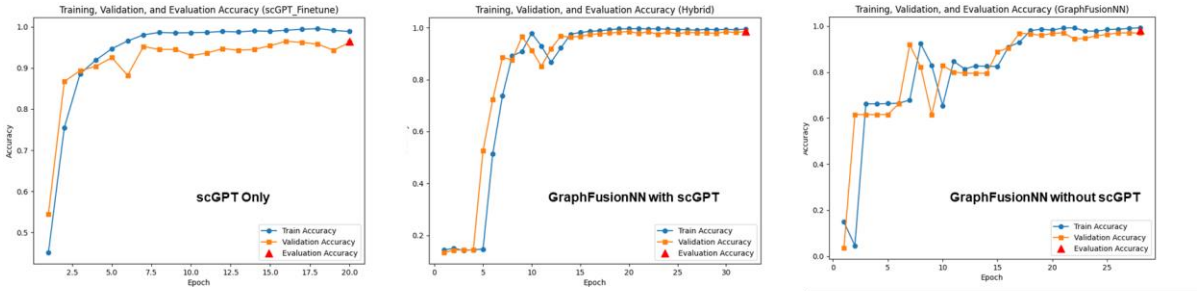


4.5 GRAPHFUSIONNN AND SCGPT COMPARISONS

Figure 15 presents a focused comparison of classification performance across three modeling strategies: standalone scGPT (***Figure 9***), GraphFusionNN with multi-modal inputs including scGPT (hybrid model), and GraphFusionNN without scGPT. While the scGPT-only model converged rapidly and achieved strong evaluation accuracy, its lower precision, recall, and macro-F1 scores reveal an

imbalanced performance biased toward dominant cell types. This suggests that scGPT performs well on abundant classes but struggles with minority populations—errors that have minimal impact on overall accuracy due to class imbalance. As a result, accuracy can be misleading in such settings, as it disproportionately reflects performance on majority classes while obscuring poor predictions on rarer ones.

Figure 15: Comparison: Breast Cancer II



Classification Report:

```
{
    "accuracy":    0.9640718562874252,
    "precision":   0.6932195051933817,
    "recall":      0.6687037621863041,
    "macro_f1":    0.6802260073343217
}
```

Classification Report:

```
{
    "accuracy":    0.9870259481037924,
    "precision":   0.7377025057935175,
    "recall":      0.7337052403945746,
    "macro_f1":    0.7353600678692609
}
```

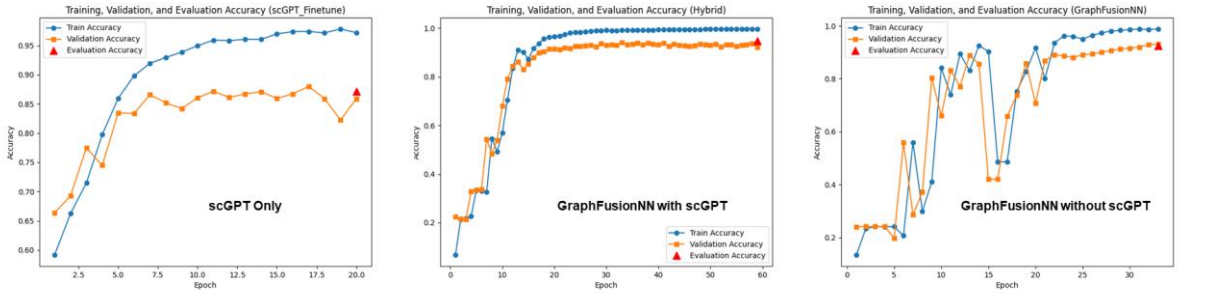
Classification Report:

```
{
    "accuracy":    0.9820359281437125,
    "precision":   0.7126889690362948,
    "recall":      0.7276303068317439,
    "macro_f1":    0.7192796126076665
}
```

In contrast, both GraphFusionNN variants demonstrated not only high accuracy but also more stable training curves and stronger convergence in validation performance. Notably, the hybrid GraphFusionNN model that incorporates scGPT, graph topology, and proxy spatial features yielded the best overall generalization, with high and balanced evaluation performance. Meanwhile, the GraphFusionNN variant without scGPT still outperformed scGPT-only—indicating that even without pretrained expression embeddings, multi-modal fusion of graph and spatial information substantially improves classification robustness.

These results underscore that scGPT captures dominant transcriptional patterns well but struggles with rarer or ambiguous cell types. GraphFusionNN, through modality fusion, provides greater resilience to class imbalance, and ultimately delivers fairer and more generalizable predictions in biologically diverse datasets.

Figure 16: Comparison: Breast Cancer III



Classification Report:

```
{
    "accuracy":    0.8716780561883067,
    "precision":   0.7979169876802680,
    "recall":      0.8026543077165873,
    "macro_f1":    0.7925245045707665
}
```

Classification Report:

```
{
    "accuracy":    0.9476082004555809,
    "precision":   0.9640340180440431,
    "recall":      0.9160211182646203,
    "macro_f1":    0.9376986542099175
}
```

Classification Report:

```
{
    "accuracy":    0.9255884586180714,
    "precision":   0.8810785551305329,
    "recall":      0.9192493137910375,
    "macro_f1":    0.8964869737158639
}
```

In *Figure 16*, we compare the performance of scGPT alone (*Figure 10*) with GraphFusionNN models with and without scGPT inputs. While the standalone scGPT model achieved a strong evaluation accuracy (~87.1%), its improvement in accuracy gain surpassed by that of GraphFusionNN with scGPT, which reached near-perfect convergence and validation stability over a longer training schedule. The model without scGPT also achieved a high evaluation accuracy, though with more variability across

epochs—reflecting greater sensitivity to modality limitations. These results reinforce the advantage of GraphFusionNN's multi-modal design: while scGPT captures dominant transcriptomic signals efficiently, the addition of graph structure and spatial context helps the model generalize more robustly. Even when scGPT is excluded, the model still benefits from contextual information that expression alone does not provide. The hybrid model with scGPT consistently outperforms in both stability and final accuracy, highlighting the value of embedding fusion when navigating complex cellular heterogeneity.

Figure 17: Comparison: Immune Cells

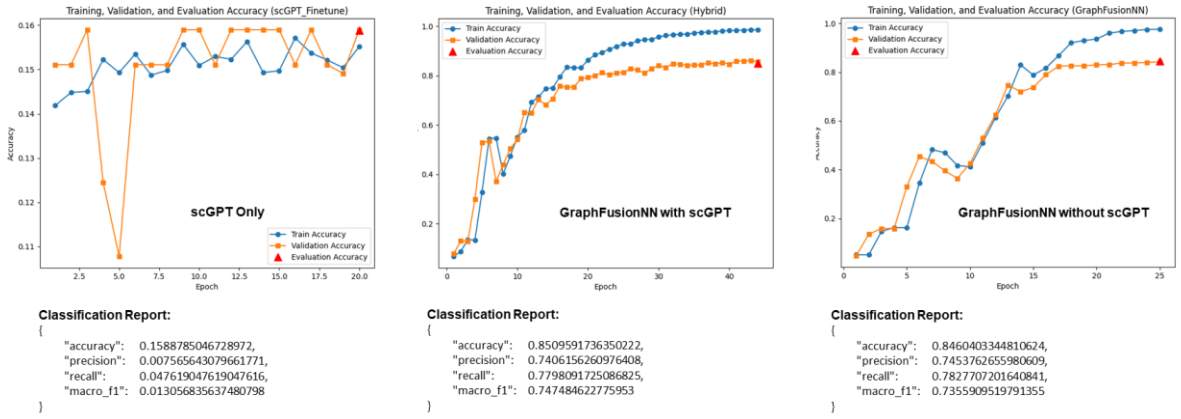


Figure 17 highlights a dramatic performance improvement when transitioning from the scGPT-only model (***Figure 13***) to the multi-modal GraphFusionNN frameworks (center and right). The scGPT-only model plateaued early, with both training and validation accuracies hovering near 15–16%, indicating a failure to extract useful features from the transcriptomic data alone. This suggests that in this dataset, gene expression lacks the discriminatory power needed for accurate classification—possibly due to noise, dropout, or transcriptional similarity between classes.

In contrast, both GraphFusionNN variants demonstrated strong learning dynamics and achieved substantial gains, with an evaluation accuracy of approximately 85%. The hybrid GraphFusionNN model that includes scGPT embeddings, graph-based topology, and proxy spatial features consistently outperformed the version without scGPT, indicating that multi-modal fusion provides a cumulative benefit when integrating expression with structure and location. These findings reinforce the importance of context: when transcriptomic signals are weak or overlapping, spatial and graph-derived cues can reintroduce learnable structure, enabling robust classification where single-modality approaches fail.

Across all three comparison scenarios, the multi-modal GraphFusionNN models consistently outperformed the scGPT-only baseline in both accuracy and balanced classification metrics, especially on challenging datasets with overlapping, noisy, or transcriptionally ambiguous cell types. While scGPT alone performed well on simpler datasets with well-separated classes, it consistently struggled on imbalanced or complex datasets, often overfitting to dominant populations and underperforming on rare or more subtle cell types. In contrast, GraphFusionNN—with or without scGPT—achieved substantially better generalization, driven by its ability to integrate complementary modalities like graph-based cell neighborhoods and spatial proximity. These modalities provide structural context that helps resolve ambiguities in expression profiles. The results reinforce the importance of multi-modal fusion for building robust, fair, and scalable models in single-cell transcriptomic analysis—especially as datasets become increasingly heterogeneous and biologically complex.

5 CONCLUSION

In this study, we systematically evaluated the performance of the fine-tuned scGPT model across multiple single-cell RNA-seq datasets and demonstrated its strong ability to generalize when transcriptomic features are well-separated in the latent space. Evaluation accuracy and macro-F1 scores were higher in datasets with clearer UMAP-based cluster separability, suggesting that while scGPT can capture dominant gene expression patterns, its ability to achieve robust and balanced classification declines in more complex or transcriptionally ambiguous landscapes.

However, we also identified key limitations: in datasets with overlapping or poorly resolved transcriptional boundaries, scGPT struggled to learn discriminative features, resulting in stagnant training dynamics and reduced classification performance. These cases highlight scGPT's reliance on the intrinsic separability of transcriptomic signals, which is often absent in biologically complex or imbalanced datasets. Furthermore, although scGPT-only models achieved high overall accuracy in some scenarios, deeper analysis revealed that this metric often masked poor performance on rare or underrepresented cell types—underscoring the importance of using macro-averaged metrics to evaluate classification fairness and robustness.

To address these challenges, we introduce GraphFusionNN—a multi-modal hybrid model that integrates transcriptomic embeddings from scGPT with graph-based structural context and proxy spatial information. Our experiments show that GraphFusionNN substantially improved classification performance in more complex datasets where scGPT alone failed. Notably, GraphFusionNN achieved the highest precision, recall, and macro-F1 scores across all tested models, indicating more equitable performance across both abundant and rare populations. The fusion of complementary modalities enables the model to distinguish ambiguous transcriptomic profiles and enhance cell-type separability, confirming the advantage of multi-modal learning in single-cell analysis.

Future Work

Building on the promising results of GraphFusionNN, future work will focus on advancing multi-modal learning in single-cell analysis across several key directions:

First, we aim to extend GraphFusionNN to incorporate additional biological modalities such as surface protein abundance (e.g., CITE-seq), chromatin accessibility (e.g., ATAC-seq), and histological image-derived features. Incorporating these complementary data sources will provide richer biological context and enable more accurate classification and discovery in highly heterogeneous datasets.

Second, we plan to develop adaptive modality weighting strategies that dynamically determine which modality is most informative for each individual cell or tissue context. Unlike current global fusion mechanisms (e.g., attention, MLP, gating), this approach would enable cell-specific or region-aware fusion, enhancing robustness in cases where some modalities are noisy, incomplete, or locally irrelevant.

Third, we will explore unsupervised and semi-supervised extensions of GraphFusionNN to better support datasets with limited or noisy cell-type annotations. Incorporating self-supervised or contrastive learning objectives may improve generalization to unseen cell types and tissue conditions. Furthermore, integrating external biological knowledge, such as cell type hierarchies (e.g., from Cell Ontology) or pathway-level priors. could guide representation learning in a biologically meaningful manner.

Finally, we envision adding interpretability modules that can attribute model predictions to specific modality-level features, such as key genes, neighborhood structures, or spatial context. This level of explainability can enhance real-world biomedical applications—such as biomarker discovery, developmental lineage tracing, and precision medicine—by connecting model predictions to biologically significant insights.

REFERENCES

1. Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., & Wang, B. (2023). scGPT: toward building a foundation model for single-cell multi-omics using generative AI. Nature Methods.
2. Kipf, T. N., & Welling, M. (2017). Semi-Supervised Classification with Graph Convolutional Networks. arXiv preprint arXiv:1609.02907.
3. Velickovic, P., et al. (2018). Graph Attention Networks. arXiv preprint arXiv:1710.10903.
4. Hu, H., et al. (2021). Cell-Cell Communication in Cancer Using Single-Cell RNA Sequencing. Nature Reviews Cancer.
5. Svensson, V., et al. (2018). SpatialDE: Identification of Spatially Variable Genes. Nature Methods.
6. 10x Genomics, Human PBMCs, 3' v3.1, Chromium Controller. Retrieved from https://cf.10xgenomics.com/samples/cell-exp/6.1.0/10k_PBMC_3p_nextgem_Chromium_Controller/10k_PBMC_3p_nextgem_Chromium_Controller_filtered_feature_bc_matrix.h5
7. 10x Genomics. (n.d.). Human Lymph Node Dataset. Retrieved from https://www.10xgenomics.com/datasets/human-lymph-node-1-standard-1-1-0.
8. Human Breast Cancer (Block A Section 2) Dataset. Retrieved from https://www.10xgenomics.com/datasets/human-breast-cancer-block-a-section-2-1-standard-1-1-0
9. Human Lung Cancer (FFPE) Dataset. Retrieved from https://www.10xgenomics.com/datasets/human-lung-cancer-ffpe-2-standard
10. Multiple Sclerosis (M.S.) Dataset. Retrieved from https://github.com/bowang-lab/scGPT/blob/main/data/README.md
11. HTAPP-982-SMP-7629, Breast Cancer Dataset. Retried from https://cellxgene.cziscience.com/collections/a96133de-e951-4e2d-ace6-59db8b3bfb1d
12. HTAPP-944-SMP-7479, Breast Cancer Dataset. Retried from https://cellxgene.cziscience.com/collections/a96133de-e951-4e2d-ace6-59db8b3bfb1d
13. HTAPP-364-SMP-1321, Breast Cancer Dataset. Retried from https://cellxgene.cziscience.com/collections/a96133de-e951-4e2d-ace6-59db8b3bfb1d
14. Immune Cells, Blood/Immune Tissue Dataset. Retrieved from https://cellxgene.cziscience.com/collections/2d40e6a7-f2fd-49ba-9db9-6b97e4c6dad5
15. Bladder Stromal, Urinary Bladder Tissue Dataset. Retrieved from https://cellxgene.cziscience.com/collections/0e54d4de-44f0-4d50-8649-b5c2bbe8f5d1