

TrajLens: Visual Analysis for Constructing Cell Developmental Trajectories in Cross-Sample Exploration

Qipeng Wang , Shaolun Ruan , Rui Sheng , Yong Wang , Min Zhu, and Huamin Qu 

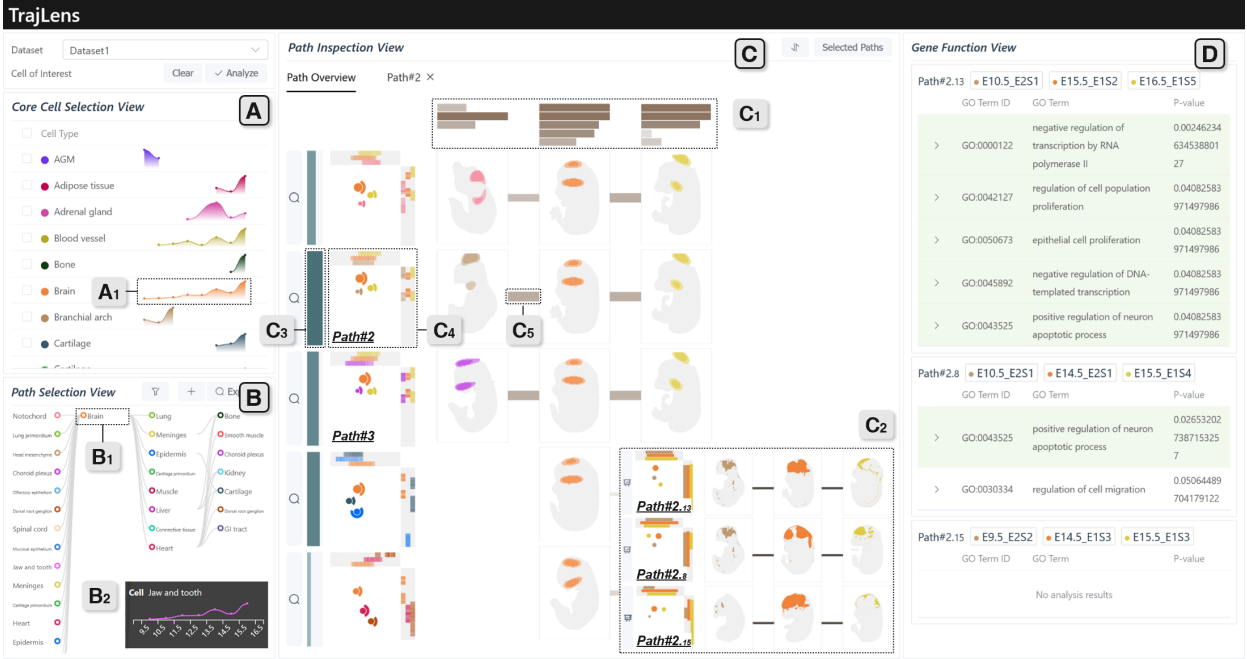


Fig. 1: *TrajLens*: (A) The Core Cell Selection View helps users select the core cell type based on temporal occurrences and quantitative changes during biological development. (B) The Path Selection View assist users in intuitively selecting and analyzing complex cell developmental paths with a focus on high-frequency and diverse trajectories. (C) The Path Inspection View reveals spatial evolutionary patterns through a multi-row synchronized view, facilitating comprehensive validation of biological evolutionary relationships with features for sequence inspection, single-sample inspection, and similarity assessment. (D) The Gene Function View present significant gene functions identified through the analysis of single-sample sequences.

Abstract—Constructing cell developmental trajectories is a critical task in single-cell RNA sequencing (scRNA-seq) analysis, enabling the inference of potential cellular progression paths. However, current automated methods are limited to establishing cell developmental trajectories within individual samples, necessitating biologists to manually link cells across samples to construct complete cross-sample evolutionary trajectories that consider cellular spatial dynamics. This process demands substantial human effort due to the complex spatial correspondence between each pair of samples. To address this challenge, we first proposed a GNN-based model to predict cross-sample cell developmental trajectories. We then developed *TrajLens*, a visual analytics system that supports biologists in exploring and refining the cell developmental trajectories based on predicted links. Specifically, we designed the visualization that integrates features on cell distribution and developmental direction across multiple samples, providing an overview of the spatial evolutionary patterns of cell populations along trajectories. Additionally, we included contour maps superimposed on the original cell distribution data, enabling biologists to explore them intuitively. To demonstrate our system's performance, we conducted quantitative evaluations of our model with two case studies and expert interviews to validate its usefulness and effectiveness.

Index Terms—Visual Analytics, Single-cell RNA Sequencing, Cell Developmental Trajectories

1 INTRODUCTION

- Q. Wang, M. Zhu are with Sichuan University. E-mail: wangqipengscu@stu.scu.edu.cn, zhumin@scu.edu.cn
- S. Ruan is with Singapore Management University. E-mail: slruan.2021@phdcs.smu.edu.sg. S. Ruan is also with Monash University.
- R. Sheng, H. Qu are with Hong Kong University of Science and Technology. E-mail: rshengac@connect.ust.hk, huamin@cse.ust.hk.
- Y. Wang is with Nanyang Technological University. E-mail: yong-wang@ntu.edu.sg.

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org.

In the field of single-cell RNA sequencing (scRNA-seq), constructing cross-sample cell developmental trajectories is a crucial analytical task that helps reveal the progression of cells through multiple developmental stages [57]. These trajectories provide insights into cellular differentiation, thereby advancing our understanding of biological development [47], disease progression [28], and targeted therapy [63]. The construction process typically involves two key phases: first, establishing cellular trajectories merely within single biological sample and then linking these trajectories across multiple samples to reveal

Digital Object Identifier: [xx.xxxx/TVCG.201x.xxxxxx](https://doi.org/10.1109/TVCG.201x.xxxxxx)

cross-sample trajectories. While current computational tools like Monocle [7] and PAGA [61] have automated the first phase by effectively modelling intra-sample cellular dynamics, they lack capabilities for cross-sample connection. Consequently, biologists often make every effort to manually connecting these trajectories between samples, which is subjective and prone to errors.

Specifically, constructing and exploring cell developmental trajectories across multiple samples presents significant challenges. Currently, biologists must manually analyze which cell populations share the same expressed genes and connect them, thereby creating a many-to-many mapping problem that complicates the analytic task. Additionally, experts need to assess the spatial distribution correspondence of cell populations in these trajectories, which is complicated by their variable distributions across samples. Consequently, it is difficult to identify developmental continuities and biological evolutionary relationships of these trajectories accurately, which highlights the need for automated tools to improve the reliability of cross-sample trajectory connections.

However, developing such tools requires resolving two critical aspects. First, developing a computational model that integrates multi-dimensional cellular features, such as gene expression, spatial correspondence, and cross-sample dynamics, is essential for predicting developmental trajectories between cell populations across samples. Moreover, it is particularly challenging for a model to capture cross-sample dynamics due to its computational complexity. Second, biologists need to manually understand and refine model predictions to validate their biological evolutionary relationships, requiring the integration of original single-cell data with model outputs. The spatiotemporal complexity of the single-cell data and the large number of connections between different samples often make this task multifaceted.

To tackle the above tasks, we propose *TrajLens*, a visual analytics system with a computational model that assists biologists in predicting and exploring cross-sample cell developmental trajectories. Firstly, we construct a GNN-based model that comprehensively captures cross-sample cellular developmental features, including cellular spatial distributions, gene expression profiles, and cross-sample dynamics, to predict potential developmental trajectories for multi-sample scRNA-seq datasets. Secondly, *TrajLens* enables users to explore predicted trajectories and validate their biological evolutionary relationships through interactive explorations, where users can analyze cellular spatial distribution patterns, developmental directions, and the functions of expressed genes along these trajectories. Finally, we quantitatively evaluated our proposed model on three multi-sample scRNA-seq datasets, demonstrating its effectiveness in predicting cross-sample cellular developmental trajectories. We also conducted two case studies and expert interviews to validate the effectiveness of our system qualitatively.

In conclusion, the contributions of our work are:

- Problem formulation and design requirements for defining and analyzing cross-sample cell developmental trajectories.
- A GNN-based model to predict cross-sample trajectories by capturing cross-sample cellular developmental features.
- *TrajLens*, a visual analytics system for exploring and identifying biological evolutionary relationships of the predicted trajectories.
- Quantitative evaluation of our model’s performance and qualitative validation of our system’s effectiveness through two case studies and interviews with eight experts.

2 RELATED WORK

In this section, we summarize the prior work related to our research on cell developmental trajectory inference tools, single-cell data visualization, and spatiotemporal data visualization.

2.1 Cell Developmental Trajectory Inference Tools

Methods for constructing cell developmental trajectories based on scRNA-seq data, known as “pseudo-time” analysis, temporally order cells within individual samples by comparing gene expression patterns and assign pseudo-time values to individual cells [40]. Tools

like PAGA [61] build graph-based trajectories, with nodes as cell clusters and edges as developmental paths. DTFLOW [57] uses feature decomposition and reverse search, while VeloViz [3] integrates scRNA-seq data and RNA velocity data [35]. However, these tools limit in cross-sample trajectory construction, as they primarily rely on single-sample analysis and fail to capture multi-dimensional features like spatial dynamics and gene expression across samples. Consequently, there is a need for a specialized tool to construct and analyzing complex cross-sample developmental trajectories.

2.2 Single-cell Data Visualization

Single-cell data, primarily scRNA-seq, captures individual cell characteristics [50], often integrated with spatial transcriptomics and protein expression data for comprehensive cellular state analysis. Current scRNA-seq data analytical tools like Seurat [23] and Cell Ranger [18] provide visualization for dimensionality reduction, cell clustering, and differential gene expression analysis. In addition, some tools explain complex cell dynamics through specialized visualizations. For example, Cellchat [27] maps intercellular communication networks via ligand-receptor interaction matrices, and PAGA [61] visualizes developmental trajectories as graph structures. Visual analytics tools enhance single-cell data exploration by integrating additional metadata, such as temporal dynamics [38], spatial contexts [43], and multi-modal datasets [26, 29]. For instance, MulteeSum [38] supports comparisons across multiple samples over time series. Visinity [43] incorporates spatial information to help users understand the positional relationships of cells within tissues. Additionally, some tools leverage user interactions to improve data exploration accuracy. Polyphony [13] uses user-selected anchors to refine dataset alignment. LineageD [25] assists users in constructing cell lineages, enhancing the precision of lineage mapping tasks. However, critical challenges persist in cross-sample trajectory analysis, as current tools struggle to integrate and track cellular state changes across multiple samples and time points, underscoring the need for advanced tools to address these limitations.

2.3 Spatiotemporal Data Visualization

Spatiotemporal data encompasses entities with inherent spatial and temporal dimensions [21], enabling the study of dynamic systems like biological trajectories [16], environmental transitions [42], and urban dynamics [39]. Spatiotemporal visualization tools typically address domain-specific needs. For geospatial-temporal phenomena, GeoChron [68] presents the evolution of geographic entities over time with the combination of maps and narrative timelines. BNVA [59] optimizes bus routes by discovering passenger flow patterns and comparing characteristics of different routes. UcVe [67] assists users in comparing multiple visualized geographic units within spatiotemporal space. Additionally, spatiotemporal visualization is also used to explore relationships between entities and model relational dynamics. MoRe-Vis [17] visualized moving entities to facilitate analysis applications in fields such as climate science. SpreadLine [34] focuses on dynamically illustrating influence within egocentric networks by employing a storyline-based design to represent entities with their evolving relationships. Wallner et al. [56] utilized storyline techniques to integrate the development of entity relationships over time and geographical trajectories in MOBA games to facilitate game replay analysis. Spatiotemporal data visualization can also be applied to analyze and compare cell spatial distributions across multiple biology samples, thereby facilitating the exploration of how cellular patterns dynamically change and develop over both time and space. However, current spatiotemporal visualization methods have limitations in effectively integrating discrete biological events with continuous spatial dynamics, which calls for a visual analytics system that integrates multi-dimensional data and supports interactive temporal-spatial linking to enable biologists to validate the biological evolutionary relationship of cross-sample cell developmental trajectories more accurately.

3 DESIGN

In this section, we introduce the problem formulation of our research, as Fig. 2 shows. We also detail the collaboration process we adopted

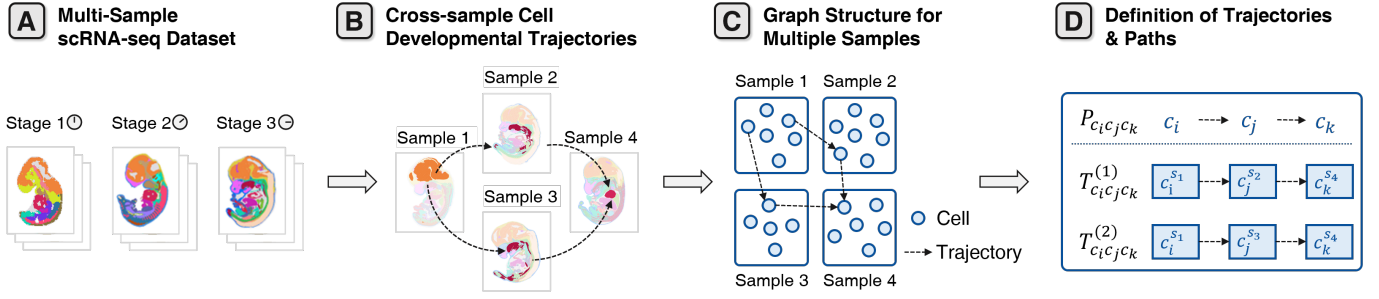


Fig. 2: We define the evolutionary relationships between cells across samples as cross-sample cell developmental trajectories. Next, we abstract the dataset into a graph, with each sample as a subgraph forming the whole graph, and refer to sequences of cross-subgraph links as paths.

with experts during the design process as well as the specific analysis tasks we identified to guide our system design effectively.

3.1 Terminology Definitions

Single-cell RNA sequencing (**scRNA-seq**) is a widely used technique for detecting and quantifies messenger RNA (mRNA) [24, 52], the transcriptional product of DNA [9], within individual cells, thereby helping biologists understand gene expression patterns [1]. **ScRNA-seq datasets** typically consist of multiple biological samples (Fig.3(A)), representing the gene expression profile of a single cell [48]. Common analytical methods include clustering [62], cell-cell communication analysis [33], and developmental trajectory inference [61]. A **cell developmental trajectory** refers to the path that a cell population follows from an undifferentiated state to a differentiated one [49] (Fig.3(B)). Since real-time tracking of all cells across continuous time points is experimentally impractical, current trajectory inference methods simulate a “pseudo-temporal” timeline by sorting cells based on gene expression similarity [57], allowing researchers to infer the developmental sequence and identify key cell types involved in biological processes [58].

3.2 Problem Formulation

Based on the above definitions, we model cell developmental trajectories as graphs from multi-sample scRNA-seq datasets, with nodes as cell types and edges as cell developmental trajectories (Fig.3(C)). Cross-sample developmental trajectories for specific cell populations were defined and unified into cross-sample edges (Fig.3(D)).

Graph Construction. Our scRNA-seq datasets consist of multiple samples representing different developmental stages of a biological subject, with each sample tied to a specific stage, though stages may include multiple samples. Collaborative discussions with biologists, revealed that developmental trajectories within samples are often abstracted as graph structures [57]. Inspired by this, we formally built a global graph $G = (V, E)$ to depict a complete cell developmental trajectory. The vertex set V can be represented as:

$$V = \bigcup_{s=1}^k V^{(s)}, \text{ where } V^{(s)} = \{v_i^{(s)}\}.$$

Here a vertex $v_i^{(s)}$ represents cell population i (identified by annotated cell names) in sample s . And the edge set E can be defined as:

$$E = \{e \mid e = (v_i^{(p)}, v_j^{(q)}), \text{ stage}(p) \neq \text{stage}(q)\},$$

where $\text{stage}(p)$ denotes sample p 's developmental stage, ensuring edges only connect cell populations from temporally distinct stages.

Path Definition. To comprehensively characterize cellular differentiation dynamics across samples and developmental stages [15], we formally defined a developmental trajectory as a temporally ordered sequence of cell type progressions, represented as:

$$T_{c_1 c_2 \dots c_{m+1}} = [c_1^{s_a} \xrightarrow{e_1} c_2^{s_b} \xrightarrow{e_2} \dots \xrightarrow{e_m} c_{m+1}^{s_c}],$$

where c_i^s represents a cell type i in sample s , and e_i represents a developmental trajectory between cell types c_i and c_{i+1} . As several trajectories may share the same cell-type sequences but originate from different samples, we defined a cross-sample cell developmental path as a collection of trajectories with identical same cell-type sequences across different samples (e.g., $c_1^{s_1} \rightarrow c_2^{s_2} \rightarrow c_3^{s_3}$ and $c_1^{s_1} \rightarrow c_2^{s_4} \rightarrow c_3^{s_5}$):

$$P_{c_1 c_2 c_3} = \{T_{c_1 c_2 c_3}^{(1)}, T_{c_1 c_2 c_3}^{(2)}, \dots, T_{c_1 c_2 c_3}^{(n)}\},$$

where n represents the frequency of the paths, i.e., the number of distinct sample combinations supporting the same cell type sequence.

3.3 Design Process

We closely collaborated with six biologists (E1-E6) with varying years of experience across specialized fields. Specifically, E1 and E2 specialize in bioinformatics, with 7 and 4 years of experience, respectively. E3 has 5 years of experience in microbiology, and E4 brings 5 years of expertise in cell biology. E5 and E6 are medical researchers with experience in using bioinformatics tools, possessing 4 and 5 years of research experience. All participants have experience using tools to analyze scRNA-seq data by conducting cell developmental trajectories. We conducted at least two rounds of systematic interviews with each expert, obtaining insights into their workflows and the challenges they encountered in studying cell developmental trajectories. This collaboration ensures that our systems are aligned with the needs of the biological field, while the diversity of expert backgrounds helps mitigate bias arising from individual perspectives.

3.4 Analysis Tasks

Based on interviews with six experienced biologists, we have summarized five key analysis tasks (**T1** - **T5**) that users need to explore when investigating cross-sample cell developmental trajectories.

T1 Select a specific cell type for analysis based on its temporal occurrence. Domain experts emphasized that selecting a core cell type for cross-sample developmental trajectory analysis should prioritize cell types persisting consistently across multiple developmental stages. Their consistent presence across stages suggests they play critical roles in biological development, thereby determining them as key candidates for further investigation.

T2 Identify high-frequency predicted cell developmental paths related to the selected cell type. After selecting a core cell type, biologists aim to identify high-frequency developmental paths connecting it to other cells undergoing critical dynamics, such as those that emerge, disappear, or experience quantitative changes during key developmental stages. As E2 noted, such cells are central to driving biological processes, as their temporal dynamics often reflect pivotal mechanisms: “Cells that emerge or disappear during key developmental processes typically play important roles in shaping the underlying mechanisms.”

T3 Assess cell developmental paths based on cellular spatial continuity. Biologists assess the spatial continuity and directional

consistency of cellular populations along developmental trajectories to validate their evolutionary relationships and dynamic behaviors. As noted by E4, “*Trajectories with biological evolutionary relationships typically maintain spatial continuity between early and subsequent cell populations.*” By integrating these spatial and directional features, biologists can enhance the understanding of the biological evolutionary relationships of these paths.

T4 Examine original samples as context information. For validated developmental paths, each of them typically encompasses multiple trajectories. Due to sample-level differences, such as developmental stages or cell scale differences, developmental trajectories may exhibit differences in gene expression. Therefore, experts need to select multiple trajectories within each path to determine if they exhibit consistent or divergent patterns.

T5 Identifying differentially expressed genes in specific trajectories and their biological functions. Experts select multiple trajectories within each path to examine gene functions separately, as different trajectories may exhibit distinct gene expression patterns. Specifically, in bioinformatics practices, identifying differentially expressed genes (DEGs) is a pivotal approach to delineating the genetic foundations of cell developmental trajectories [65]. To achieve this goal, biologists typically utilize statistical methods [6] and gene annotation databases [53] to uncover the biological processes in which these DEGs are involved.

4 DATA PROCESSING

In this section, we introduce our data processing workflow, which analyzes multi-sample scRNA-seq datasets and predicts cross-sample cell developmental paths (Fig.3). Firstly, we collected three multi-sample scRNA-seq datasets, preprocessed them following biological practices, and constructed them into graphs ((Fig.3(A))). Secondly, we implemented a GNN-based model and conducted training and prediction phases ((Fig.3(B, C))). Subsequently, we filtered predicted edges according to specific biological criteria and identified high-frequency paths ((Fig.3(D))). Finally, we provided a comprehensive summary of paths, including path overviews and gene function analysis ((Fig.3(E, F))). Through these steps, *TrajLens* provides a complete workflow for processing and analyzing multi-sample scRNA-seq data, enabling biologists to better understand the dynamic changes in cell development.

4.1 Dataset

Datasets Description. We collected a multi-sample scRNA-seq dataset for the following analysis process, referred to as Mouse-Embryo Dataset, whose detailed descriptions are provided in Supplement Materials. Each sample in the dataset contains a gene expression matrix and single-cell-level metadata. In the gene expression matrix, each row corresponds to a unique single cell, with each column representing a gene type and the values indicating the expression levels of a specific gene within the cell. Additionally, the single-cell-level metadata includes spatial coordinates of every single cell (i.e., x and y coordinates in a two-dimensional plane) for describing its extract position within the sample. Furthermore, cells are annotated by their cell types, such as neuronal progenitor cells, endothelial cells, and many others.

Dataset Preprocessing. For each dataset, we normalized the expression values of each sample [70] and employed the widely-used bioinformatics tool Harmony [32] to mitigate technical variations across multiple samples. For each sample, we constructed a graph-based representation $G_{sub} = (V, E_{none})$ to model cellular relationships, where nodes $v \in V$ correspond to annotated cell types. Notably, edges E_{none} in each graph were not predefined, as we will construct a model to predict them. We also stored metadata for each node, including a set of differentially expressed genes and spatial coordinates of all cells within the node, which provides critical biological context for downstream analysis. Finally, we construct a global graph $G = \{G_{sub}^{(1)}, G_{sub}^{(2)}, \dots, G_{sub}^{(s)}\}$ composed of multiple sample-specific subgraphs G_{sub} .

4.2 Model Setup

Adoption Rationale. Existing methods, such as PAGA and linear models, fail to adequately capture implicit semantic relationships [10], such as node appearance/disappearance events and cell gene expression profiles. Consequently, they face challenges in effectively performing cross-sample cell developmental trajectory prediction. Consequently, they limit in predicting cross-sample cell developmental trajectories effectively. In contrast, the Graph neural network (GNN) architectures [44] excels at preserving both node-level features and complex topological relationships [22], making them particularly suitable for representing dynamic cellular development. Consequently, we conducted a GNN-based model that effectively models the spatiotemporal dynamic graph structure of our multi-stage cellular development dataset.

Training Data Construction. We constructed a graph as training and testing data based on Dataset 1 using the PAGA algorithm, where nodes represent cell populations [54] and edges denote connections inferred by PAGA. Moreover, as the model must not only capture the graph structure, i.e., the nodes and edges generated above, but also characterize node features, including gene expression profiles and cell distribution patterns, we define features for each node. Specifically, we first identified the top 20 differentially expressed genes (DEGs) for each cell population through the Wilcoxon rank-sum test [60]. This choice of 20 DEGs is a commonly used approach and has been demonstrated to distinguish between cell types effectively [69]. We first embedded these DEGs into 1024-dimensional semantic embeddings using the pre-trained model BioLinkBERT [64]. Simultaneously, we performed Principal Component Analysis (PCA) [41] on the spatial coordinates of cell populations, which can be represented as an $N \times 2$ matrix containing (x, y) coordinates for N cells, and reduced them to 1024-dimensional vectors enabling comprehensive analysis of cellular features and differentiation trajectories across samples.

Model Construction. Specifically, our model integrated two components: a two-layer Graph Attention Network (GAT) [55] module with multi-head attention for modelling inter-cell interactions and a two-layer Graph Convolutional Network (GCN) [30] module for global topological edge propagation. Through residual connections, we fused local attention patterns and global structural information, enabling complementary optimization of both modules. The model inputs a graph where each node has a 2048-dimensional feature embedding and outputs probabilistic predictions of edges. By testing different thresholds ranging from 0.5 to 0.9 on the test dataset, we found that the highest precision was achieved at a threshold of 0.75. Consequently, edges with probabilities greater than 0.75 are treated as existing, thereby constructing the predicted cell developmental trajectories.

Model Training and Predicting. To enhance model robustness during training, we generated negative sample edges in each training epoch and implemented batch shuffling to prevent overfitting. After model training, we used the graphs constructed in Section 4.1 that only consist of nodes without edges as input data and employed the optimized weight parameters from the training process to predict edges. All models are trained and evaluated using PyTorch on a 48-GB NVIDIA RTX 4090 GPU, for 100 epochs with a batch size of 8. The model has approximately 1,000,000 parameters and is optimized with AdamW [37] at an initial learning rate of 1×10^{-5} .

4.3 Cross-sample Developmental Path Refinement

Based on trajectory predictions generated by our proposed GNN-based model, we integrated subgraphs for each sample with predicted edges to construct a comprehensive directed graph $G_{pred} = (V, E_{pred})$. Each node $v \in V$ contains metadata such as cell type name, sample origin, temporal developmental stage, differentially expressed genes, and spatial coordinates of each single cell. The directed edges $e \in E_{pred}$ symbolize predicted developmental trajectories based on their origin src and terminus dst , which were filtered out under biological and temporal constraints to ensure that they represent cross-sample developmental trajectories: edges between different cell types $type(e_{src}) \neq type(e_{dst})$ and those directed from earlier to later stages $stage(e_{src}) < stage(e_{dst})$. Based on filtered edges, we merged redundant edges with identical source-target pairs by counting their occurrences, defining the merged

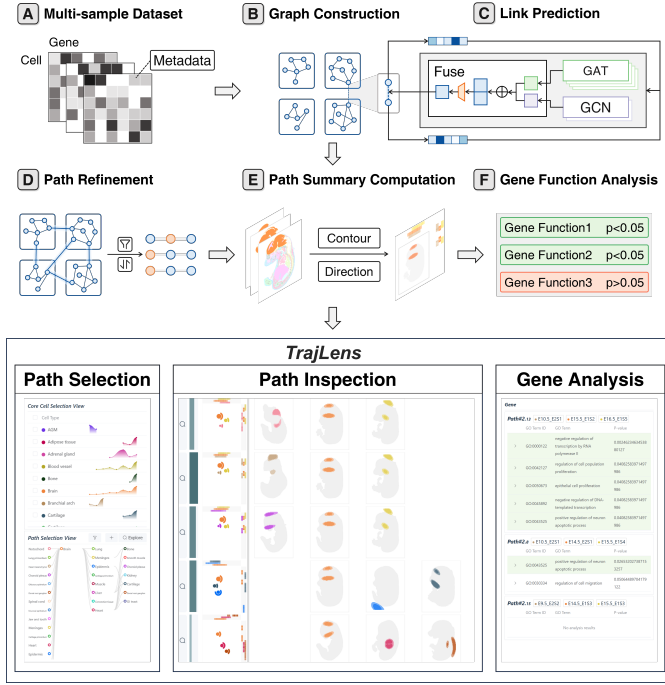


Fig. 3: The data processing and analyzing workflow of *TrajLens* is as follows: Each sample of our multi-sample scRNA-seq datasets is represented as a matrix containing cell and gene information with single-cell-level metadata describing cellular characteristics. We then abstracted the dataset into a unified graph structure, where each sample forms a subgraph collectively integrating into a large graph. Next, we conducted a GNN-based model to predict links between subgraphs, i.e., cross-sample cell developmental trajectories. Then, we identified high-frequency developmental paths composed of multiple cell types for user selection. We computed a summary overview for each selected developmental path and provided gene function analysis along specific paths.

edge weights as the count of original edges:

$$w(u \rightarrow v) = |\{e \in E_{filtered} \mid src(e) = u \wedge dst(e) = v\}|,$$

where w represents the weights of the merged edge, quantifying the total count of original edges that were combined.

Given the high density $\rho = \frac{|E|}{|V|(|V|-1)} \approx 50\%$ of the initial graph which was biologically suboptimal because the densely connected graph tends to contain substantial noise edges while obscuring critical biological signals [5]. To mitigate this, we retained only the top 15% most frequent paths, reducing the analysis complexity. The threshold selection aligned with typical biological practices in single-cell data analysis, retaining 5% to 15% of high-confidence features for core analysis [66]. For the user-selected key cell type, we conducted a breadth-first search (BFS) on G_{pred} to trace its ancestral and descendant cell populations along developmental trajectories. We then constructed a hierarchical tree rooted at the key cell type, with each node recording its topological distance from the root and its relationship as either an ancestor or descendant. Multiple cell types in the tree constitute a cross-sample cell developmental path.

4.4 Path Summary Computation

To help biologists understand the spatial distribution characteristics of cell developmental trajectories more intuitively, we generated detailed contour maps for each cell population along the developmental path to visualize their global spatial distribution patterns. We also provided a similarity metric between contour maps to identify cell populations with similar spatial distributions. Additionally, we derived cell developmental directions from spatial density gradients to reveal potential cellular evolutionary trends during cell development.

Contour map computation. With the spatial range of a specific sample, we created a 100×100 grid and used Gaussian Kernel Density Estimation to compute the probability density distribution of grid points based on the spatial distribution of cells within the sample. Then, we generated polygonal contours for predefined thresholds using the Marching Squares algorithm [36] and simplified these contours using the Douglas-Peucker algorithm [11]. The contour coordinates were projected onto the x - and y -axis and binned into 100 histogram intervals weighted by their occurrence frequency on these axes. To provide users with a spatial reference for cell contours, we utilized the Alpha Shape algorithm [12] to generate geometric shapes that capture the biological boundaries of cells samples and extract the x and y coordinates of these shapes.

Contour similarity computation. When calculating the similarity of given two distribution contours C_1 and C_2 from cell type pairs, we bidirectionally computed the minimum point-wise distances in both directions: $d_{C_1 \rightarrow C_2}$ and $d_{C_2 \rightarrow C_1}$. For each point p_i in contour C_1 , we computed the minimum distance to each point in contour C_2 : $d_{p_i \rightarrow C_2} = \min_{q_j \in C_2} \|p_i - q_j\|$. The overall distance $d_{C_1 \rightarrow C_2}$ is then obtained by taking the mean of all point-wise minima:

$$d_{C_1 \rightarrow C_2} = \frac{1}{|C_1|} \sum_{i=1}^{|C_1|} d_{p_i \rightarrow C_2}.$$

The symmetric spatial overlap D_{sym} is then defined as the reciprocal of the mean bidirectional distances:

$$D_{sym} = \frac{2}{d_{C_2 \rightarrow C_1} + d_{C_1 \rightarrow C_2}}.$$

Higher D_{sym} values indicate more significant similarity between the two contours, while lower values reflect distribution differences. This metric quantitatively compares spatial characteristics across cell types.

Cell developmental direction computation. Subsequently, we applied Principal Component Analysis (PCA) to identify the primary developmental directions of the contour map for each cell. Specifically, the input to PCA consisted of the set of two-dimensional coordinates (x, y) of each cell and the output provided the primary developmental direction of the cell's contour map, which was quantified with a range $[0, 2\pi]$ and calculate the stability ratio as follows:

$$R_{std} = \sqrt{\frac{\lambda_2}{\lambda_1 + \lambda_2}},$$

where λ_1 and λ_2 represent the eigenvalues of the primary and secondary components, respectively. The value R_{std} close to 0 indicates high stability in the developmental direction of a specific cell.

4.5 Gene Function Analysis

We analyzed gene functions of cell developmental trajectories using Gene Set Enrichment Analysis (GSEA) [51] combined with the Gene Ontology (GO) database [2, 53]. Specifically, we constructed a hierarchical network of GO terms based on the GO database using GOA-TOOLS [31], mapped genes based on their mapped genes to their functional terms via mouse Gene Association File (GAF) annotations [4]. For a specific developmental path, we used a statistical test method [31] to assess the enrichment of GO terms in the target gene set (highly expressed genes of specific cell types) relative to the background gene set (highly expressed genes of all samples), identifying overrepresented biological functions of these genes based on the GAF file.

5 SYSTEM

In this section, we introduce the visual designs of *TrajLens*. Within our analytical workflow, users first select core cell types based on temporal occurrences and population dynamics of cells in the *Core Cell Selection View* (T1). Subsequently, *TrajLens* presents other cell types predicted to have potential developmental trajectories related to the core cell types in the *Path Selection View*, where users can further select multiple cell combinations as candidate paths for comparative analysis (T2). Then,

users can evaluate the biological evolutionary relationships of these paths through multi-row visualization with cellular spatial distribution in the *Path Summary View* (T3). Additionally, they can delve into the trajectories constituting specific paths in the *Trajectory Inspection View* (T4) and utilize the *Gene Function View* to analyze the enriched gene functions presented in these trajectories (T5).

5.1 Cell Developmental Path Selection

Users determine paths for comparative analysis in two steps: first, selecting core cell types based on temporal occurrences and population dynamics (T1); then, selecting cell types predicted to have developmental relationships with the selected cells, forming multiple cell combinations as candidate paths for exploration (T2).

5.1.1 Core Cell Selection View

To help users select core cell types based on their temporal occurrences and quantitative changes during biological development, we propose the *Core Cell Selection View* (Fig.1(A)). The *Core Cell Selection View* combines a tabular layout with line charts, where each row displays a cell type's name along with its quantitative changes and occurrence points across developmental stages. The horizontal axis of each row's line chart represents different developmental stages, and the vertical axis represents the quantity of cells. In addition, each cell's color in the *Core Cell Selection View* corresponds to the standardized color-encoding defined in the original dataset, where cell types are consistently assigned specific colors based on their biological classification (e.g., muscle cells in red (●), bone cells in green (●)) to maintain consistency with prior studies.¹ Users can identify cells of interest by examining their presence and quantity changes over developmental stages. Cells existing along the shown developmental stages are typically selected as cells of interest for further investigation.

5.1.2 Path Selection View

Upon selecting a core cell type, users aim to explore sequences involving more than two types of cells relevant to the selected cell, which is defined to cell developmental paths in Section 3.2.

Description. When selecting candidate cell developmental paths, users typically prioritize those with high frequency. To address this requirement, we design the *Path Selection View* (Fig.1(B)), which supports filtering and visualizing high-frequency cell developmental paths while enabling intuitive selection and analysis of paths of interest. Given that the data involving the selected cell of interest and its neighboring cells forms a hierarchical structure (Section 4.3), where multiple developmental paths share common cell types and diverge at these shared cells, we adopt a hierarchical tree visualization for *Path Selection View* to present the relationships between paths and their shared nodes explicitly. The view places the core cell types as the root node (Fig.1(B1)), with adjacent columns presenting cells having predicted paths leading to it. Nodes in each column represent individual cell types, and links connect nodes across adjacent columns, indicating predicted cell developmental paths. To emphasize high-frequency paths, users can click the “Filter” button to exclude paths with frequencies less than a specific threshold, thereby focusing on more representative developmental paths. Additionally, the vertical order of nodes within each column is determined by the number of paths connected to them, which aligns with the users' requirement to explore cells that play pivotal roles in diverse developmental processes.

Interaction. When hovering over a node in the *Path Selection View*, users can access a line chart about the temporal occurrence and quantity changes of the corresponding cell type through a tooltip (Fig.1(B2)). Leveraging the tooltip, users can select multiple cells that emerge or disappear during development or show substantial quantity changes to form a developmental path. *TrajLens* will then automatically check for connections between the selected nodes. If the path is valid (i.e., all consecutive nodes are connected), it will be incorporated into the list for inspection in the following process. After selecting all desired paths,

users can click the “Explore” button, and our system will proceed with further in-depth analysis of these paths.

5.2 Cell Developmental Path Inspection

Biologists need to validate the biological evolutionary relationships underlying the selected developmental paths in Section 5.1.2 by assessing cellular spatial distributions across them (T3). To facilitate this task, we propose the *Path Inspection View*, featuring multi-row visualization for comparing spatial overlap and development trends among cell populations from multiple samples, along with an aggregated overview for intuitive comparison. Additionally, to refine the analysis of cell developmental trajectories constituting specific paths (T4), we introduce the *Trajectory Inspection View* for analyzing trajectories constituting specific paths, supporting the following gene function analysis.

5.2.1 Path Inspection View

The *Path Inspection View* consists of multi-row visualizations, where each row represents a selected developmental path, with the summary visualization for each row offering an overview of the path. Additionally, the similarity visualization is above the multi-row visualization to illustrate the similarity of node contour maps within a column.

Description. The *Path Inspection View* employs a node-link diagram to represent cells within the selected paths with cellular spatial distribution (Fig.4(A)). Each row of this view corresponds to a selected path in Section 5.1.2 (Fig.4(A)), where nodes correspond to cell types in the path (Fig.4(A6)) and edges depict predicted developmental trajectories with their color intensity and height encoding the similarity between consecutive cells' spatial distribution (Fig.4(A8)). We visualize the overall distribution of each node through layered contour maps: the outer contour represents 98% core density coverage, while the inner contour shows 94% density (Fig.4(A7)), which enables researchers to simultaneously capture the most concentrated core and the broader distribution of cell populations. The color intensity of these contours is scaled to reflect the corresponding density coverage, providing a visual hierarchy of cell distribution. All sequences are spatially aligned based on the user-selected core cell to ensure cross-row consistency. To address the limitation that the *Path Inspection View* alone cannot provide an overview of a sequence's overall characteristics, we introduce the summary visualization of each row. Within each summary visualization, central dots represent the centroid of each cell population with their radius indicating the average cell count in samples (Fig.4(A4)). Fan-shaped arcs around the dots encode the primary developmental direction of the cell population through the central angle, and angular span reflects their directional variance computed in Section 4.4 (Fig.4(A5)). Additionally, we design a multi-projection visualization (Fig.4(A3)) that orthogonally projects the spatial distributions of each node onto the *x*- and *y*-axis and then stacks them sequentially. The innermost projections represent the leftmost node in the path, and subsequent projections expand outward to indicate the nodes' order in the path. Each row also features a bar chart where the width and color intensity indicates path frequency (Fig.4(A2)) and a button to the left to the view opens a *Trajectory Inspection View* in a new tab upon selecting and leads to the following analysis (Fig.4(A1)).

Interaction. Upon selecting a row by clicking the leftmost button (Fig.4(A1)), we present the similarity visualization for each node of the selected path and other nodes in the corresponding column by bar charts above the *Path Inspection View* (Fig.1(C1)). The vertical order of the bars matches the vertical order of the nodes in the column, facilitating users to identify and select additional paths similar to the selected ones.

Design alternative. During the iterative design process, we refined *Path Inspection View* based on user feedback. Initially, we integrated the *Path Selection View* and the *Path Inspection View* into a hierarchical tree visualization with contour maps for each node that presented cell type relationships and spatial distributions (Fig.4(C)). However, while effective for exploring cell relationships, this approach is limited in a detailed comparison of divergent branches. Furthermore, a node in this view may represent the same cell type from multiple paths, potentially originating from different sample sets, complicating the clear representation of a single node. Therefore, we separated the hierarchical

¹<https://db.cngb.org/stomics/mosta/spatial/>

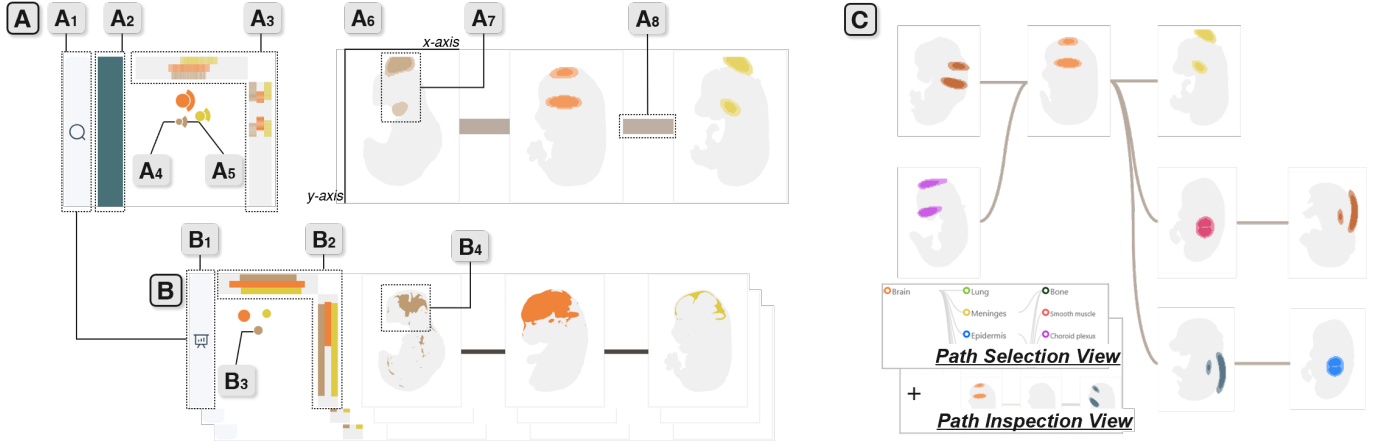


Fig. 4: The design of the multi-row visualization in Section 5.2: (A) A row in the *Path Inspection View* represents a cell developmental path defined in Section 3.2. (B) The *Trajectory Inspection View* supports the detailed inspection of the developmental trajectories that constitute the path in (A). (C) The design alternatives of this view integrates the *Path Selection View* and *Path Inspection View*.

tree from comparison views, allowing clearer comparison of different paths in the *Path Inspection View* and preserved the hierarchy-based exploration of cellular relationships in the *Path Selection View*.

5.2.2 Trajectory Inspection View

Selecting a row in *Path Inspection View* opens a new tab displaying *Trajectory Inspection View*, which, like *Path Inspection View*, includes multi-view and summary visualization components. Each node in the *Trajectory Inspection View* (Fig.4(B)) corresponds to the cell distribution map of a single sample (Fig.4(B4)).

Description. For summary visualization, we marked the centroids of cell populations with dots, where the dot positions represent spatial centroids and their sizes encode population counts (Fig.4(B3)). Additionally, we projected the spatial distributions of cell populations onto the *x*- and *y*-axis as density profiles (Fig.4(B2)).

Interaction. By clicking the button on the left side of the specific row (Fig.4(B1)), *TrajLens* will automatically conduct gene function analysis based on the gene set of this trajectory and present the analysis result in the *Gene Function View* in a table format.

5.3 Gene Function View

The *Gene Function View* presents the results of gene function analysis in table formats (T5) (Fig.1(D)). The results include the GO term ID, GO term description, and corresponding *p*-value, indicating the gene function's significance. To highlight significant functions, rows with *p*-values less than 0.05 are marked in green. When users expand a row, *Gene Function View* will present a list of all genes associated with the selected GO term, facilitating further analysis based on this gene list.

6 CASE STUDY

We conducted two case studies by recording the exploring process of two interviewed experts, E1 and E5, using our system, and then summarized their insights to demonstrate the practicality of our system.

6.1 Case Study I: Exploring Cross-sample Developmental Paths in Mouse Embryonic Development Dataset

As an expert in bioinformatics, E1 used *TrajLens* to explore the cross-sample cell development paths related to the brain cell type and investigate the gene functions underlying these paths.

Selecting the Core Cell Type (T1). Among the line charts of multiple cell types in *Core Cell Selection View* (Fig.1(A)), E1 identified distinct patterns of cellular dynamics throughout developmental stages in Dataset 1. Many cell types exhibited transient emergence, appearing only at specific developmental stages (e.g., AGM cells (●) transiently emerged at the first two embryonic days to generate stem cells), making them less informative for analysis. In contrast, the line chart depicting

the emergence of brain cells (●) exist throughout all developmental stages (Fig.1(A1)) in the *Core Cell Selection View* suggesting they may play a crucial role in nervous system formation. Consequently, E1 selected the brain cells as the core cell type for in-depth exploration.

Selecting Cross-sample Cell Developmental Paths (T2). After selecting the brain cells as the core cell type, *Path Selection View* presented all associated cell types hierarchically, including direct precursors and successors as well as indirectly linked cells (Fig.1(B)). E1 filtered out paths with frequencies less than 40 to reduce analysis complexity, as low-frequency paths were less likely to be involved in critical developmental processes. Based on their ranking in the column and temporal occurrence and quantity changes shown in tooltip line charts (Fig.4(B2)), E1 selected cells such as the Notochord (●), Choroid plexus (●), and Dorsal root ganglion (●), etc. Their tooltip line charts spanning only specific stages suggest these cells may differentiate into other cell types or originate from them and are more likely to participate in brain-related developmental processes. Some highly ranked cells in their columns, such as Cartilage (●), were not selected due to their persistent presence without significant changes. This approach focused on dynamically changing cell types with potential critical functions.

Inspecting Cross-sample Cell Developmental Paths (T3). E1 analyzed paths formed by the selected cell types based on frequency, as indicated by the bar chart width to the left of each row in the *Path Inspection View*. Path#2 showed the highest frequency (Fig.4(C3)), indicating its prevalence and suggesting it might represent a key developmental path in embryonic development. Additionally, Path#2 exhibited optimal alignment in cell spatial distribution across nodes, with the contour projections of its three cell types showing significant overlap on both the *x*- and *y*-axis in the summary visualization (Fig.4(C4)). Moreover, the similarity of spatial distribution in Path#2 is further supported by the color intensity and height of the links between each pair of nodes (Fig.4(C5)). The path also presented consistent directionality, as seen from the uniform direction of arcs surrounding each dot in the summary visualization (Fig.4(C4)). Based on these observations, E1 chose Path#2 for further investigation. Upon clicking the left-most button of Path#2, a new tab opened in the *Path Inspection View*, displaying the list of cross-sample trajectories constituting Path#2 in *Trajectory Inspection View* (Fig.4(C2)). Moreover, E1 noticed that the cell types of Path#3 showed significant similarity to those of Path#2 by the width and color intensity bar charts on the top of each column (Fig.4(C1)), suggesting potentially related developmental processes.

Gene Function Analysis (T4, T5). When exploring the cross-sample trajectories (Fig.4(C2)), E1 found that Path#2.13 exhibited the highest alignment in cellular distribution, as indicated by the most consistent contour projections on the summary visualization (Fig.4(C2)). Therefore, Path#2.13 was selected for gene function analysis. The results highlighted significant gene functions in neuronal apoptosis (e.g.,

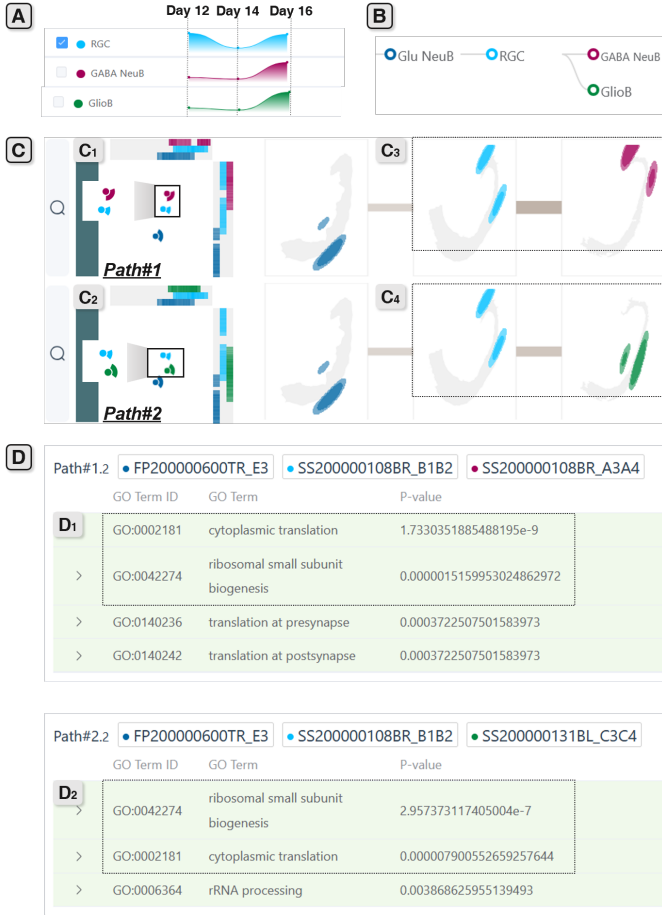


Fig. 5: Identifying cell developmental branches in Case Study II. (A) The *Core Cell Selection View* presents RGC (●), GABA NeuB (●), and Gliob (●) across three stages: Day 12 to Day 16. The quantity of RGC began to decrease at D14, while the quantities of GABA NeuB and Gliob increased from D14. (B) The *Path Selection View* shows the predicted links related to RGC, indicating a potential branch relationship between these three cell types. (C) The *Path Inspection View* presents cell distribution along Path#1 and Path#2, confirming their spatial proximity. (D) The *Gene Function View* reveals similar gene expression profiles for the two paths, confirming their branching evolutionary relationship.

GO:0000122) and cell proliferation control (e.g., GO:0042127), suggesting its potential role in brain specialization (Fig.1(D)). Conversely, Path#2.8 and Path#2.15 had few or even no enriched gene functions, indicating Path#2.8 represents an early phase of brain development while Path#2.15 resides closer to the developmental starting point.

In this case, E1 explored cross-sample cell developmental paths associated with brain cells and identified paths that exhibit biological evolutionary relationships through spatiotemporal patterns.

6.2 Case Study II: Identifying Cross-sample Cell Developmental Branch Patterns in Mouse Midbrain Dataset

E5, a medical expert who specializes in neurology and frequently uses bioinformatics tools for scRNA-seq data analysis, utilized *TrajLens* to study mouse neural development in Dataset 2.

Identifying Cell Developmental Branches (T1, T2). Firstly, E5 observed the quantity of RGC (●) decreased at Day 14 while GABA NeuB (●) and Gliob (●) increased from Day 14 to Day 16 (Fig.5(A)) through corresponding line charts. Selecting RGC, E5 found multiple predicted links from RGC to GABA NeuB and Gliob in *Path Selection View* (Fig.5(B)), suggesting RGC may differentiate into GABA NeuB and Gliob simultaneously. Moreover, E5 observed similar contour maps (Fig.5(C3, C4)) and similar direction of arcs surrounding dots corresponding to GABA NeuB and Gliob (Fig.5(C1, C2)), which repre-

sents the similar developmental directions among these two cell types and supported RGC differentiation into GABA NeuB and Gliob.

Validating Cell Developmental Branches (T3, T5). To test this hypothesis, he conducted gene function analysis in *Gene Function View* on two trajectories of these paths Path#1.2 and Path#2.2. The results revealed that the gene functions expressed in both paths were highly consistent, primarily focusing on cytoplasmic translation (GO:0002181) and ribosomal small subunit biogenesis (GO:0042274) (Fig.5(D1, D2)). These gene functions are closely associated with the rapid proliferation of neural cells, thereby further supporting the hypothesis that RGC differentiates into both GABA NeuB and Gliob.

In summary, E5 discovered that a cell type RGC can differentiate into two distinct cell types, GABA NeuB, and Gliob, during the midbrain development of mouse embryos. He further verified the biological significance of this developmental path in terms of gene function, indicating that differentiation patterns are not limited to a one-to-one relationship but extend to multiple cell types.

7 EVALUATION

In our evaluation process, we first conducted a multi-metric quantitative analysis to assess the performance of our proposed GNN-based predictive model and then validated the practical usefulness and effectiveness of *TrajLens* through qualitative expert interviews.

7.1 Model Evaluation

We conducted ablation studies to validate the effectiveness of each module in our model and performed comparative experiments to establish its superiority over existing methods. We also tested the time consumption of our model in completing common tasks to prove its efficiency and hyperparameter settings to prove the rationality of our hyperparameter choices, with details provided in Supplement Materials.

Experiment setup. In our ablation study, we isolated two key modules of our model: GAT and Fusion. Specifically, we replaced the GAT with a GCN to assess the benefits of attention-based aggregation and replaced the Fusion module into a weighted summation module to examine its impact on feature integration. To demonstrate the superiority of our proposed model in our specific application domain, we compared its performance against standard GAT [55] and GraphSAGE [20] models. We utilized the Mouse-Embryo Dataset used in prior analyses and two multi-sample scRNA-seq datasets, Zebrafish-Embryo [14] and Schiebinger2019 [45]. Additionally, we utilized two widely adopted GNN benchmarks, CoraFull [46] and CiteSeer [19], to ensure the generalizability of our model. Details about the datasets is provided in Supplement Materials. We computed the accuracy (ACC) metric and conducted five repeated experiments to obtain the standard deviation.

Evaluation Criteria. For the Mouse-Embryo dataset, we collaborated with domain experts to annotate 118 labels for cross-sample cell developmental trajectories, which served as the reference background. In the case of the Zebrafish-Embryo and Schiebinger2019 datasets, we adopted their inherent differentiation trajectories as ground truth. As for the CiteSeer and CoraFull datasets, their original graph structures served as ground truth. Then, we separately apply our model to edge-less graphs extracted from datasets to predict missing edges.

Results. Table 1 presents the evaluation results of our proposed model on five datasets and averaged over five independent runs, demonstrating its superior performance. Notably, the model achieves an accuracy of 88.56 ± 1.46 on the Mouse-Embryo dataset, aligning well with expert expectations for accuracy. Ablation studies confirm that the full model outperforms its ablated variants, while comparative experiments show it surpasses baseline models. Additionally, on the CiteSeer and CoraFull datasets, our model exhibits robust performance, highlighting its strong generalization across large-scale datasets. In summary, our model's accuracy meets user requirements across diverse datasets while demonstrating robustness and effectiveness.

7.2 Expert Interview

To validate the usefulness and effectiveness of *TrajLens*, we conducted a semi-structured interview with six domain experts over 70 minutes. These experts included the four E1-E6 introduced in Section 3.3 and two

Table 1: The evaluation results of our proposed model on five datasets and 5 runs.

Dataset	Our Model	Our Model - GAT	Our Model - Fusion	GraphSAGE [20]	GAT [55]
Mouse-Embryo [8]	88.56 ± 1.46	85.72 ± 0.69	87.83 ± 0.46	88.42 ± 0.18	87.73 ± 0.91
Zebrafish-Embryo [14]	76.91 ± 0.25	75.12 ± 0.64	75.81 ± 0.42	74.42 ± 0.80	<u>76.62 ± 0.33</u>
Schiebinger2019 [45]	71.98 ± 1.49	70.75 ± 0.54	71.16 ± 1.16	70.26 ± 1.59	70.99 ± 1.81
CiteSeer [19]	81.53 ± 0.36	80.59 ± 1.12	<u>80.87 ± 0.54</u>	79.39 ± 1.38	80.16 ± 1.63
CoraFull [46]	88.40 ± 0.37	87.51 ± 1.38	<u>88.23 ± 0.88</u>	87.50 ± 1.45	87.42 ± 1.46

Bold indicates the best performance; underlined indicates the second-best.

additional experts E7, E8 from different institutions who had experience using tools to construct cell developmental trajectories based on scRNA-seq data. We first introduced the background of our research (15 minutes), followed by an introduction to the dataset used and the usage step of our system (15 minutes). Then, we asked the experts to complete the exploration process for both Dataset 1 and Dataset 2, with each task requiring at least 20 minutes. Finally, we had a semi-structured interviews with them to collect their feedback (20 minutes).

Usefulness. E1 to E4 acknowledged that the data processing and workflow of *TrajLens* align with the conventions of biology field. Our workflow follows standardized biological analysis practices for constructing cell developmental trajectories, analyzing cell developmental paths, and identifying enriched gene functions. Our innovation lies in introducing a graph-based prediction model that can capture and preserve the non-linear structure and dynamics of scRNA-seq samples. As E2 noted: “*GNN propagates messages through edges to capture both local cell characteristics and global inter-sample patterns*”. Moreover, by integrating cellular features, inter-cellular relationships, and cross-sample heterogeneity, our proposed model effectively bridges local details with a global perspective, thereby significantly enhancing predictive accuracy. Additionally, our system enables biologists to construct developmental trajectories and interpret their biological relationships through multi-dimensional evidence, such as gene expression patterns, cell-cell links, and cellular spatial distribution dynamics.

Visualization and Analytical Strengths. All experts praised the visualization design of *TrajLens* as a key strength. The *Path Selection View*, which extracts high-frequency paths into a hierarchical tree structure, was highlighted by E4 as: “*This structure improves the readability of complex trajectory networks by presenting cell relationships with frequency and distance to the core cell. Moreover, it also displays characteristics of entire high-frequency paths, helping me understand developmental relationships and their significance.*” The contour map abstraction and multi-row layout significantly enhance cross-sample analysis by enabling multi-dimensional comparisons of cell populations in a path. As highlighted by E5: “*This approach facilitates the identification of cell populations related to specific tissues while supporting both horizontal and vertical comparisons.*” Specifically, horizontal comparisons reveal cellular changes across different developmental paths, while vertical comparisons highlight variations in cell distributions at corresponding positions within diverse trajectories. “*This dual-axis analytical framework provides a comprehensive view of cellular dynamics and differentiation patterns*”, noted by E7. The summary visualization for each row, achieved through the projection of cell states onto x- and y-axis and the summary of distribution maps, enables users to compare cellular patterns across paths intuitively. Such an overview-to-detail analytical framework provides a comprehensive understanding of cellular dynamics and differentiation patterns.

Suggestion. E5 suggested incorporating a pseudo-time ordering algorithm [57] to calculate the direction of cellular development, thereby identifying the developmental direction of specific cell types more clearly. This would facilitate users in assessing the continuity of developmental directions among multiple cells. Additionally, E3 suggested that the system should support the selection of multiple core cell types for comparative analysis, thereby revealing interaction patterns associated with these different core cell types. E2 proposed that the system can directly provide differential gene expression analysis results for each path, enabling rapid identification of key differentially expressed genes and supporting hypothesis validation in exploration processes.

8 DISCUSSION

In this section, we discuss the implications of our designing process, the stability of our *Path Inspection View*, the generalizability of our proposed workflow, and the limitations and future work of our system.

Design Implications. Our workflow abstracts complex biological subjects into computable structures for analysis. First, cell developmental trajectories are presented as graphs, where nodes represent specific cell populations and edges represent cell developmental trajectories, which enables the prediction of cross-sample cell developmental trajectories based on GNN. Second, spatial distributions of cells from multiple samples are simplified into contour maps highlighting key features such as density peaks, tissue boundaries, and overlapping regions while facilitating cross-sample comparisons. These abstractions convert biological data into computable formats such as graphs and maps, providing an intuitive analysis workflow and visualization that facilitates users’ understanding of these complex datasets.

Scalability. For analyzing cross-sample cell developmental trajectories, we used the Mouse-Embryo dataset, which spans a few developmental developmental stages. Due to its temporal resolution, predicted paths of the dataset are typically four steps or shorter. To address more complex scRNA-seq datasets, we collected two large-scale datasets Zebrafish-Embryo and Schiebinger2019 to evaluate the model’s scalability. Additionally, we validated the model’s generalizability using widely used GNN benchmark datasets CiteSeer and CoraFull. *TrajLens* currently implements multi-row visualization for analyzing cross-sample cell developmental trajectories. Considering longer or more paths, *TrajLens* allows users to explore by scrolling horizontally or vertically. Additionally, our system supports sorting multiple developmental paths by their frequency, enabling users to quickly identify high-frequency paths and improve analysis efficiency.

Limitations and Future Work. Currently, the limited publicly accessible multi-sample scRNA-seq datasets restrict our ability to explore various relationships across samples. To address this, we can incorporate heterogeneous samples from diverse origins to enhance the robustness of our system, such as those describing the same subject but originating from different datasets (e.g., samples from different laboratories or batches of the same disease). In addition, we identified that the cell contour estimation based on Gaussian Kernel Density Estimation generated density estimates that exceeded the actual distribution boundaries of the samples, especially in sparse regions or at the edges. This limitation disrupts users’ manual comparison of contour overlaps across samples, leading to misinterpretations of cell distribution overlaps. To mitigate this, we plan to implement overlay scatter plots of raw cell distributions as reference layers to help users distinguish between “true” and “estimated” contour regions.

9 CONCLUSION

We introduce *TrajLens*, a visual analytics system with a GNN-based model to assist biologists in predicting and exploring cross-sample cell developmental trajectories for multi-sample scRNA-seq datasets. Our proposed multi-row visualization with contour maps facilitates the validation of biological evolutionary relationships of candidate paths through the overview of cellular spatial distributions. To assess the performance of our model and the effectiveness of our system, we conducted a quantitative evaluation of our model’s performance and performed qualitative validation of our system’s effectiveness through two case studies and interviews with eight experts.

ACKNOWLEDGMENTS

This work was supported by National Natural Science Foundation of China (Nos.62172289).

SUPPLEMENTAL MATERIALS

All supplemental materials are available on GitHub at <https://github.com/imwangqp/TrajLens-Supplement-Materials/blob/main/supplement.pdf>, including dataset descriptions and extended results.

REFERENCES

- [1] T. S. Andrews, V. Y. Kiselev, D. McCarthy, and M. Hemberg. Tutorial: Guidelines for the computational analysis of single-cell rna sequencing data. *Nature Protocols*, 16(1):1–9, 2021. doi: 10.1038/s41596-020-00409-w 3
- [2] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, et al. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000. doi: 10.1038/75556 5
- [3] L. Atta, A. Sahoo, and J. Fan. Veloviz: Rna velocity-informed embeddings for visualizing cellular trajectories. *Bioinformatics*, 38(2):391–396, 2022. doi: 10.1093/bioinformatics/btab653 2
- [4] R. M. Baldarelli, C. L. Smith, M. Ringwald, J. E. Richardson, C. J. Bult, and M. G. I. Group. Mouse genome informatics: an integrated knowledge-base system for the laboratory mouse. *Genetics*, 227(1):iyae031, 2024. doi: 10.1093/genetics/iyae031 5
- [5] A.-L. Barabási and Z. N. Oltvai. Network biology: understanding the cell’s functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2004. doi: 10.1038/nrg1272 5
- [6] J. Candia and L. Ferrucci. Assessment of gene set enrichment analysis using curated rna-seq-based benchmarks. *PLOS ONE*, 19(5):e0302696, 2024. doi: 10.1371/journal.pone.0302696 4
- [7] J. Cao, M. Spielmann, X. Qiu, X. Huang, D. M. Ibrahim, A. J. Hill, F. Zhang, S. Mundlos, L. Christiansen, F. J. Steemers, et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, 566(7745):496–502, 2019. doi: 10.1038/s41586-019-0969-x 2
- [8] A. Chen, S. Liao, M. Cheng, K. Ma, L. Wu, Y. Lai, X. Qiu, J. Yang, J. Xu, S. Hao, et al. Spatiotemporal transcriptomic atlas of mouse organogenesis using dna nanoball-patterned arrays. *Cell*, 185(10):1777–1792.e21, 2022. doi: 10.1016/j.cell.2022.04.003 9
- [9] F. CRICK. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970. doi: 10.1038/227561a0 3
- [10] D. Dermadi, M. Bscheider, K. Bjegovic, N. H. Lazarus, A. Szade, H. Hadeiba, and E. C. Butcher. Exploration of cell development pathways through high-dimensional single cell analysis in trajectory space. *iScience*, 23(2):100842, 2020. doi: 10.1016/j.isci.2020.100842 4
- [11] D. H. Douglas and T. K. Peucker. *Algorithms for the Reduction of the Number of Points Required to Represent a Digitized Line or its Caricature*, chap. 2, pp. 15–28. John Wiley & Sons, Ltd, 2011. doi: 10.1002/9780470669488.ch2 5
- [12] H. Edelsbrunner, D. Kirkpatrick, and R. Seidel. On the shape of a set of points in the plane. *IEEE Transactions on Information Theory*, 29(4):551–559, 1983. doi: 10.1109/TIT.1983.1056714 5
- [13] C. F., S. K. M., Q. H., G. N., and W. Q. Polyphony: an interactive transfer learning framework for single-cell data analysis. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):591–601, 2023. doi: 10.1109/TVCG.2022.3209408 2
- [14] J. A. Farrell, Y. Wang, S. J. Riesenfeld, K. Shekhar, A. Regev, and A. F. Schier. Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science*, 360(6392):eaar3131, 2018. doi: 10.1126/science.aar3131 8, 9
- [15] R. M. Fisher, J. Z. Shik, and J. J. Boomsma. The evolution of multicellular complexity: the role of relatedness and environmental constraints. *Proceedings of the Royal Society B: Biological Sciences*, 287(1931):20192963, 2020. doi: 10.1098/rspb.2019.2963 3
- [16] L. Fouqueau and J. Polechová. Eco-evolutionary dynamics in changing environments: integrating theory with data. *Journal of Evolutionary Biology*, 37(6):579–587, 2024. doi: 10.1093/jeb/voae067 2
- [17] V. G., F. N., and P. J. Morevis. A visual summary for spatiotemporal moving regions. *IEEE Transactions on Visualization and Computer Graphics*, 30(4):1927–1941, 2024. doi: 10.1109/TVCG.2023.3250166 2
- [18] x. Genomics. Cell ranger, 2025. 2
- [19] C. L. Giles, K. D. Bollacker, and S. Lawrence. CiteSeer: an automatic citation indexing system. In *Proceedings of the Third ACM Conference on Digital Libraries*, p. 89–98. Association for Computing Machinery, 1998. doi: 10.1145/276675.276685 8, 9
- [20] W. Hamilton, Z. Ying, and J. Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, pp. 1024–1034. Curran Associates, Inc., 2017. 8, 9
- [21] J. Han, M. Kamber, and J. Pei. 13 - data mining trends and research frontiers. In *Data Mining (Third Edition)*, pp. 585–631. Morgan Kaufmann, 2012. 2
- [22] X. Han, B. Wang, C. Situ, Y. Qi, H. Zhu, Y. Li, and X. Guo. scapgnn: A graph neural network-based framework for active pathway and gene module inference from single-cell multi-omics data. *PLOS Biology*, 21(11):e3002369, 2023. doi: 10.1371/journal.pbio.3002369 4
- [23] Y. Hao, T. Stuart, M. H. Kowalski, S. Choudhary, P. Hoffman, A. Hartman, A. Srivastava, G. Molla, S. Madad, C. Fernandez-Granda, and R. Satija. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nature Biotechnology*, 42(2):293–304, 2024. doi: 10.1038/s41587-023-01767-y 2
- [24] T. Hashimshony, N. Senderovich, G. Avital, A. Klochendler, Y. de Leeuw, L. Anavy, D. Gennert, S. Li, K. J. Livak, O. Rozenblatt-Rosen, et al. Cel-seq2: Sensitive highly-multiplexed single-cell rna-seq. *Genome Biology*, 17(1):77, 2016. doi: 10.1186/s13059-016-0938-8 3
- [25] J. Hong, A. Trubuil, and T. Isenberg. Lineage: An interactive visual system for plant cell lineage assignments based on correctable machine learning. *Computer Graphics Forum*, 41(3):195–207, 2022. doi: 10.1111/cgf.14533 2
- [26] C. Jahn, M. Ibrahim, J. Busch, Q. Lin, H. Manchanda, H. Mohr, D. Plischke, H. G. Roider, and A. Steffen. Cellenium—a scalable and interactive visual analytics app for exploring multimodal single-cell data. *Bioinformatics*, 39(6):btad349, 2023. doi: 10.1093/bioinformatics/btad349 2
- [27] S. Jin, M. V. Plikus, and Q. Nie. Cellchat for systematic analysis of cell-cell communication from single-cell transcriptomics. *Nature Protocols*, 20(1):180–219, 2025. doi: 10.1038/s41596-024-01045-4 2
- [28] J. B. Kang, A. Nathan, K. Weinand, F. Zhang, N. Millard, L. Rumker, D. B. Moody, I. Korsunsky, and S. Raychaudhuri. Efficient and precise single-cell reference atlas mapping with symphony. *Nature Communications*, 12(1):5890, 2021. doi: 10.1038/s41467-021-25957-x 1
- [29] M. S. Keller, I. Gold, C. McCallum, T. Manz, P. V. Kharchenko, and N. Gehlenborg. Vitessce: integrative visualization of multimodal and spatially resolved single-cell data. *Nature Methods*, 22(1):63–67, 2025. doi: 10.1038/s41592-024-02436-x 2
- [30] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations*. OpenReview.net, 2017. 4
- [31] D. V. Klopfenstein, L. Zhang, B. S. Pedersen, F. Ramirez, A. Warwick Vesztrocy, A. Naldi, C. J. Mungall, J. M. Yunes, O. Botvinnik, M. Weigel, et al. Goatools: A python library for gene ontology analyses. *Scientific Reports*, 8(1):10872, 2018. doi: 10.1038/s41598-018-28948-z 5
- [32] I. Korsunsky, N. Millard, J. Fan, K. Slowikowski, F. Zhang, K. Wei, Y. Baglaenko, M. Brenner, P.-r. Loh, and S. Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with harmony. *Nature Methods*, 16(12):1289–1296, 2019. doi: 10.1038/s41592-019-0619-0 4
- [33] M. P. Kumar, J. Du, G. Lagoudas, Y. Jiao, A. Sawyer, D. C. Drummond, D. A. Lauffenburger, and A. Raue. Analysis of single-cell rna-seq identifies cell-cell communication associated with tumor characteristics. *Cell Reports*, 25(6):1458–1468.e4, 2018. doi: 10.1016/j.celrep.2018.10.047 3
- [34] Y.-H. Kuo, D. Liu, and K.-L. Ma. Spreadline: Visualizing egocentric dynamic influence. *IEEE Transactions on Visualization and Computer Graphics*, 31(1):1050–1060, 2025. doi: 10.1109/TVCG.2024.3456373 2
- [35] G. La Manno, R. Soldatov, A. Zeisel, E. Braun, H. Hochgerner, V. Petukhov, K. Lidschreiber, M. E. Kastrii, P. Lönnerberg, A. Furlan, et al. Rna velocity of single cells. *Nature*, 560(7719):494–498, 2018. doi: 10.1038/s41586-018-0414-6 2
- [36] W. E. Lorensen and H. E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques*, p. 163–169. Association for Computing Machinery, 1987. doi: 10.1145/37401.37422 5
- [37] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations*. OpenReview.net, 2019. 4
- [38] M. M., M. T., D. A., and P. H. Multeesum: A tool for comparative spatial

- and temporal gene expression data. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):908–917, 2010. doi: [10.1109/TVCG.2010.1372](https://doi.org/10.1109/TVCG.2010.1372)
- [39] M. S. Misty, M. A.-A. Hoque, and S. A. Mukul. Assessment of urban green space dynamics in dhaka south city corporation of bangladesh using geospatial techniques. *Land*, 13(9), 2024. doi: [10.3390/land130914262](https://doi.org/10.3390/land130914262)
- [40] X. Pan and X. Zhang. Studying temporal dynamics of single cells: Expression, lineage and regulatory networks. *Biophysical Reviews*, 16(1):57–67, 2024. doi: [10.1007/s12551-023-01090-52](https://doi.org/10.1007/s12551-023-01090-52)
- [41] K. Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901. 4
- [42] S. Rahaman, S. Jahangir, M. S. Haque, R. Chen, and P. Kumar. Spatio-temporal changes of green spaces and their impact on urban environment of mumbai, india. *Environment, Development and Sustainability*, 23(4):6481–6501, 2021. doi: [10.1007/s10668-020-00882-z2](https://doi.org/10.1007/s10668-020-00882-z2)
- [43] W. S., K. R., J. N. A., G. G., J. J., C. R. C., H. J., M. J., L. B. M., J. T., S. S., K. S. P., and P. H. Visinity: Visual spatial neighborhood analysis for multiplexed tissue imaging data. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):106–116, 2023. doi: [10.1109/TVCG.2022.32093782](https://doi.org/10.1109/TVCG.2022.32093782)
- [44] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009. doi: [10.1109/TNN.2008.20056054](https://doi.org/10.1109/TNN.2008.20056054)
- [45] G. Schiebinger, J. Shu, M. Tabaka, B. Cleary, V. Subramanian, A. Solomon, J. Gould, S. Liu, S. Lin, P. Berube, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943.e22, 2019. doi: [10.1016/j.cell.2019.01.0068,9](https://doi.org/10.1016/j.cell.2019.01.0068,9)
- [46] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Gallagher, and T. Eliassi Rad. Collective classification in network data. *AI Magazine*, 29(3):93–106, 2008. doi: [10.1609/aimag.v29i3.21578,9](https://doi.org/10.1609/aimag.v29i3.21578,9)
- [47] S. Singh, A. Praveen, N. Dudha, V. K. Sharma, and P. Bhadrecha. Single-cell transcriptomics: a new frontier in plant biotechnology research. *Plant Cell Reports*, 43(12):294, 2024. doi: [10.1007/s00299-024-03383-91](https://doi.org/10.1007/s00299-024-03383-91)
- [48] A. Sountoulidis, S. Marco Salas, E. Braun, C. Avenel, J. Bergenstr hle, J. Theelke, M. Vicari, P. Czarnecki, A. Lontos, X. Abalo, et al. A topographic atlas defines developmental origins of cell heterogeneity in the human embryonic lung. *Nature Cell Biology*, 25(2):351–365, 2023. doi: [10.1038/s41556-022-01064-x3](https://doi.org/10.1038/s41556-022-01064-x3)
- [49] S. V. Stassen, G. G. K. Yip, K. K. Y. Wong, J. W. K. Ho, and K. K. Tsia. Generalized and scalable trajectory inference in single-cell omics data with via. *Nature Communications*, 12(1):5528, 2021. doi: [10.1038/s41467-021-25773-33](https://doi.org/10.1038/s41467-021-25773-33)
- [50] A. Subramanian, M. Alperovich, Y. Yang, and B. Li. Biology-inspired data-driven quality control for scientific discovery in single-cell transcriptomics. *Genome Biology*, 23(1):267, 2022. doi: [10.1186/s13059-022-02820-w2](https://doi.org/10.1186/s13059-022-02820-w2)
- [51] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005. doi: [10.1073/pnas.05065801025](https://doi.org/10.1073/pnas.05065801025)
- [52] F. Tang, C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, J. Bodeau, B. B. Tuch, A. Siddiqui, K. Lao, and M. A. Surani. Mra-seq whole-transcriptome analysis of a single cell. *Nature Methods*, 6(5):377–382, 2009. doi: [10.1038/nmeth.13153](https://doi.org/10.1038/nmeth.13153)
- [53] G. O. C. The, S. A. Aleksander, J. Balhoff, S. Carbon, J. M. Cherry, H. J. Drabkin, D. Ebert, M. Feuermann, P. Gaudet, N. L. Harris, et al. The gene ontology knowledgebase in 2023. *Genetics*, 224(1):iyad031, 2023. doi: [10.1093/genetics/iyad0314,5](https://doi.org/10.1093/genetics/iyad0314,5)
- [54] V. A. Traag, L. Waltman, and N. J. van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1):5233, 2019. doi: [10.1038/s41598-019-41695-z4](https://doi.org/10.1038/s41598-019-41695-z4)
- [55] P. Veli kovi , G. Cucurull, A. Casanova, A. Romero, P. Li , and Y. Bengio. Graph attention networks. In *6th International Conference on Learning Representations*. OpenReview.net, 2018. 4, 8, 9
- [56] G. U. N. Wallner, L. Wang, and C. Dormann. Visualizing the spatio-temporal evolution of gameplay using storyline visualization: A study with league of legends. *Proceedings of the ACM on Human-Computer Interaction*, 7(CHI PLAY), 2023. doi: [10.1145/36110582](https://doi.org/10.1145/36110582)
- [57] J. Wei, T. Zhou, X. Zhang, and T. Tian. Dtfow: Inference and visualization of single-cell pseudotime trajectory using diffusion propagation. *Genomics*, 19(2):306–318, 2021. doi: [10.1016/j.gpb.2020.08.0031,2,3,9](https://doi.org/10.1016/j.gpb.2020.08.0031,2,3,9)
- [58] X. Wei, S. Fu, H. Li, Y. Liu, S. Wang, W. Feng, Y. Yang, X. Liu, Y.-Y. Zeng, M. Cheng, et al. Single-cell stereo-seq reveals induced progenitor cells involved in axolotl brain regeneration. *Science*, 377(6610):eabp9444, 2022. doi: [10.1126/science.abp94443](https://doi.org/10.1126/science.abp94443)
- [59] D. Weng, C. Zheng, Z. Deng, M. Ma, J. Bao, Y. Zheng, M. Xu, and Y. Wu. Towards better bus networks: A visual analytics approach. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):817–827, 2021. doi: [10.1109/TVCG.2020.30304582](https://doi.org/10.1109/TVCG.2020.30304582)
- [60] F. A. Wolf, P. Angerer, and F. J. Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):15, 2018. doi: [10.1186/s13059-017-1382-04](https://doi.org/10.1186/s13059-017-1382-04)
- [61] F. A. Wolf, F. K. Hamey, M. Plass, J. Solana, J. S. Dahlin, B. G ttgens, N. Rajewsky, L. Simon, and F. J. Theis. Paga: Graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biology*, 20(1):59, 2019. doi: [10.1186/s13059-019-1663-x2,3](https://doi.org/10.1186/s13059-019-1663-x2,3)
- [62] J. Xie, S. Ruan, M. Tu, Z. Yuan, J. Hu, H. Li, and S. Li. Clustering single-cell rna sequencing data via iterative smoothing and self-supervised discriminative embedding. *Oncogene*, 43(29):2279–2292, 2024. doi: [10.1038/s41388-024-03074-53](https://doi.org/10.1038/s41388-024-03074-53)
- [63] Y. Yang, Z. Liu, Y. Wei, S. He, A. Gu, Z. Li, J. Li, Z. Xu, and B. Cen. Single-cell multi-omics analysis reveals candidate therapeutic drugs and key transcription factor specifically for the mesenchymal subtype of glioblastoma. *Cell & Bioscience*, 14(1):151, 2024. doi: [10.1186/s13578-024-01332-31](https://doi.org/10.1186/s13578-024-01332-31)
- [64] Yasunaga, Michihiro and Leskovec, Jure and Liang, Percy. LinkBERT: Pretraining language models with document links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8003–8016. Association for Computational Linguistics, 2022. doi: [10.18653/v1/2022.acl-long.5514](https://doi.org/10.18653/v1/2022.acl-long.5514)
- [65] H. Yin, H. Duo, S. Li, D. Qin, L. Xie, Y. Xiao, J. Sun, J. Tao, X. Zhang, Y. Li, et al. Unlocking biological insights from differentially expressed genes: Concepts, methods, and future perspectives. *Journal of Advanced Research*, 2024. doi: [10.1016/j.jare.2024.12.0044](https://doi.org/10.1016/j.jare.2024.12.0044)
- [66] S. H. Yip, P. C. Sham, and J. Wang. Evaluation of tools for highly variable gene discovery from single-cell rna-seq data. *Briefings in Bioinformatics*, 20(4):1583–1589, 2019. doi: [10.1093/bib/bby0115](https://doi.org/10.1093/bib/bby0115)
- [67] D. Yu, O. Ian, L. Jie, Y. Xiaoru, and N. Q. Vinh. User-centered visual explorer of in-process comparison in spatiotemporal space. *Journal of Visualization*, 26(2):403–421, 2023. doi: [10.1007/s12650-022-00882-32](https://doi.org/10.1007/s12650-022-00882-32)
- [68] D. Z., C. S., S. T., D. D., T. T., X. M., W. D., and W. Y. Visualizing large-scale spatial time series with geochron. *IEEE Transactions on Visualization and Computer Graphics*, 30(1):1194–1204, 2024. doi: [10.1109/TVCG.2023.33271622](https://doi.org/10.1109/TVCG.2023.33271622)
- [69] M. Zhang, S. Liu, Z. Miao, F. Han, R. Gottardo, and W. Sun. Ideas: individual level differential expression analysis for single-cell rna-seq data. *Genome Biology*, 23(1):33, 2022. doi: [10.1186/s13059-022-02605-14](https://doi.org/10.1186/s13059-022-02605-14)
- [70] G. X. Y. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu, et al. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8(1):14049, 2017. doi: [10.1038/ncomms140494](https://doi.org/10.1038/ncomms140494)