POLITECNICO DI TORINO

Master's Degree course in Data Science and Engineering

Master's Degree Thesis

scVEMO: Leveraging Single-cell Multiomics Data for Developmental Trajectory Reconstruction in the Embryonic Mouse Brain

Supervisors
Prof. Stefano Di Carlo
Prof. Alessandro Savino
Dr. Roberta Bardini
Ing. Lorenzo Martini

Candidate
Alessia Leclercq

Academic Year 2023-2024

Acknowledgements

Innanzitutto, vorrei ringraziare il gruppo SMILIES per avermi concesso l'opportunitá di lavorare su un progetto cosí entusiasmante. In particolare, ringrazio Roberta e Lorenzo che in questi mesi mi hanno seguita, guidata e consigliata.

Un ringraziamento speciale va a Matilda, che mi ha supportata dal mio primo giorno a Torino, come una sorella. Ringrazio anche Leonardo, Giovanni Pope, i ragazzi di via Giuria ed i compañeros che hanno contribuito a rendere questi anni decisamente piú leggeri. Infine ringrazio la mia famiglia per i sacrifici che hanno fatto e, soprattutto, per la loro pazienza. Mamma, papá: finalmente é finita!

Abstract

Single-cell sequencing has revolutionized the study of gene expression and its phenotypic consequences by enabling the simultaneous profiling of thousands of individual cells. Recent advancements in multimodal single-cell sequencing have further expanded the scope of these techniques, allowing for the integration of transcriptomics, epigenomics, proteomics, and other omic data to obtain a more comprehensive view of cellular states and dynamics. A specific application of multiomics single-cell sequencing is lineage tracing, which provides insights into the developmental process from pluripotent cell populations to fully differentiated states. This thesis proposes scVEMO, a multiomics-based approach to lineage tracing leveraging CellRank and the RNA-velocity estimation techniques, scVELO and its extension to changes in chromatin states, Multivelo. ScVEMO is validated on the Fresh Embryonic E18 Mouse Brain dataset provided by 10X Genomics. Building on the assumption that lineage commitment is a continuous process, where cells traverse a spectrum of intermediate states, scVEMO builds a K-Nearest Neighbor graph, connecting neighboring cells based on their joint scRNA-seq and scATAC-seq data profiles. Then it integrates gene expression, promoter peak counts, and RNA-velocity information to direct the graph and compute cell state transition probabilities. Finally, the CellRank framework is employed to simulate the system and identify terminal states. Particularly, CellRank uses the Generalized Perron Cluster Cluster Analysis (GPCCA) to coarse-grain the transition probability matrix into a set of macrostates, representing coarse-grained, metastable cellular states or phenotypes. The results from the random walk simulations, coupled with the identified macrostates, enable to compare the models and gain insights into how effectively each one is able to reconstruct biologically meaningful developmental lineages. The assessment then extends to examining cell fate probabilities, which are evaluated based on multilineage potential and the average probability of each cell cluster towards the identified terminal states. This additional investigation sheds light on the models' ability to capture cell fate commitment across various cellular populations. While the random walk simulations do not identify cell development at the granular cell-state level, results clearly demonstrate that the integration of the epigenomic profiles via scVEMO improves macrostate identification on the UMAP embedding. ScVEMO distinguishes an additional terminal state within the neuronal lineage, corresponding to the upper cortical layers. This enhanced reconstruction represents a significant improvement over the transcriptomics-only method, which solely recovers the deeper cortical layers, allowing for a more congruent lineage reconstruction with the existing biology literature. Additionally, multilineage potential investigations, assessed through the KL-divergence between the single cell fate probabilities and the average fate probability per lineage across all cells, show that scVEMO improves cell lineage commitment compared to the scRNA-seq approach. Together, the results demonstrate that multimodal data integration can yield to a robust and more informative lineage reconstruction compared to transcriptomics-only methods.

Contents

List of Figures 5

1 Introduction 11

2 Background 15

2.1 Technological Context 15

2.1.1 The DNA Structure 16

2.1.2 Next Generation Sequencing and Single Cell Sequencing Technologies 20

2.1.3 Transcriptional Profiling of Individual Cells through scRNA-seq . . 22

2.1.4 scATAC-seq: Mapping Chromatin Accessibility in Single Cells . . 24

2.1.5 Single Cell Multimodal Profiling: An Illustrative Example 26

2.2 Related Works 30

2.2.1 Lineage Tracing Methods and Algorithms 30

2.2.2 RNA-velocity: Mapping Transcription Temporal Dynamics 34

2.2.3 RNA-velocity: the Velocyto Framework 35

2.2.4 RNA-velocity: the scVELO Framework 37

2.2.5 RNA-velocity: the MultiVelo Framework 39

2.2.6 Notions on Graph Theory and Random Walks 40

2.2.7 Cell Trajectory Inference: the CellRank Framework 43

2.2.8 The Generalised Perron Cluster Cluster Analysis and its Application in CellRank 47

3 Methods 53

3.1 Dataset 53

3.2 Data Preprocessing 56

3.3 Transition Matrix Construction and Developmental Lineages Identification 59

3.4 Evaluation metrics 62

3.4.1 Markov Chains 62

3.4.2 Number of Macrostates 62

3.4.3 Terminal States and Developmental Lineages 63

3.4.4 Multilineage Potential 63

3.4.5 Preprocessing Parameters 65

4 Results 67

4.1 Markov Chain Simulations . . . 67

4.2 Number of Macrostates . . . 70

4.3 Developmental Lineages and Fate Probabilities . . . 73

4.4 Multilineage Potential . . . 76

4.5 Preprocessing Parameters . . . 77

5 Conclusions 99

Bibliography 103

A 109

B 127

List of Figures

2.1 Gene structure: **a.** Simplified gene structure, including the cis-regulatory elements like enhancers and promoters, as well as the open reading frame that contains the sequence of introns and exons where the genetic information is stored **b.** The pre-mRNA molecule that is produced during the transcription process. **c.** The mature mRNA molecule that results after RNA splicing or manipulation operations. The protein-coding region contains the expressed exons, which will then be translated into a biologically functional protein. A 5' cap and poly-A tail are added to the RNA transcript to distinguish mRNAs molecules from other RNA products, indicating respectively, the beginning and the end of the transcript [1]. . . . 19

2.2 Chromatin structure: **a.** DNA helicoidal structure; **b.** DNA is wrapped around histones represented in yellow. Chromatin remodelers use ATP to dissociate DNA from histones (nucleosome ejection, top right) or modify nucleosome position (nucleosome sliding, bottom right). Chromatin structure modifications enable ease of transcription. **c.** Heterochromatin and euchromatin representation. 20

2.3 Transcription process. **a.** Schematic of the transcription process: Transcription factors (TFs, shown in yellow) and RNA polymerase II (RNAPII, shown in blue) bind to the promoter region, specifically the TATA-box sequence (shown in green), which specifies the transcription start site. Transcription then proceeds as RNAPII moves along the open reading frame, synthesizing a new RNA molecule. **b.** Insights into the RNAPII molecule: The DNA strands are locally unwound, and ribonucleotide triphosphates (rNTPs) are added complementary to the 3' strand of the DNA to form the new RNA transcript. 21

2.4 PCR mechanism. A single PCR cycle consists of three steps: **(1)** Denaturation, the process of separating the DNA fragments strands through heat; **(2)** Annealing, the process that binds a short synthetic DNA primer to the complementary sequence; the primer acts as starting point of the DNA synthesis process; **(3)** Elongation, where the new DNA strand is synthesized complementary to the template, therefore doubling the initial DNA molecule. Each cycle repeats the denaturation, the annealing and the elongation steps, resulting in an exponential increase of the DNA copies. 23

2.5 scRNA-seq data. **a.** scRNA-seq pipeline and most common data analysis procedures. **b.** Single cell isolation techniques. **c.** Reverse transcription with poly(T) priming and poly(A) terminator. 25

2.6 Cell isolation and lysing pipeline overview from *Embryonic Mouse Brain for Single Cell Multiome ATAC + GEX, Document Number GC000366 Rev D, 10X Genomics, (2022, July 13).* 27

2.7 Overview of the *Chromium Single Cell Multiome ATAC + GEX protocol, Document Number CG000338 Rev. F, 10X Genomics, (2022, August 26).* **a.** Visualisation of GEMs encapsulation of nuclei with transposase enzymes and nuclei transposition into open chromatin regions. **b.** ATAC library (left) and GEX library (right) construction steps and components. 29

2.8 Overview of the CellRank framework. Upper left: the cell-cell transition probability matrix is coarse-grained using GPCCA into four macrostates which are divided into an initial, an intermediate, and two terminal states. Highlighted cells represent the 30 top likely cells to belong to the terminal state. Upper-right: cell fate probabilities mapped over UMAP embedding; each cell is colored according to the terminal state it is more likely to reach. Bottom left: normalized gene expression plot over the UMAP representation and lineage-specific gene expression trends along pseudotime. Bottom-right: heatmap representing gene expression along pseudotime for the top 50 genes correlating with terminal state B. Adapted from *Lange, Bergen, Klein et al., Cellrank for directed single-cell fate mapping, Nature Methods, 19:159-170, 2022, doi:10.1038/s41592-021-01346-6* Copyright . 33

2.9 Overview of the Palantir framework. Top: tSNE map of scRNA-seq epithelial enriched cells from the mouse colon colored according to cluster, cell pseudotime, and differentiation potential on the tSNE embedding. Bottom: Differentiation potential trends and gene expression trends along pseudotime. Adapted from *Setty, Kiseliovas, Levine et al., Characterization of cell fate probabilities in single-cell data with Palantir, Nature Biotechnology, 37:451-460, 2019, doi:10.1038/s41587-019-0068-4.* 34

2.10 Illustration of the STREAM framework. Left: flat tree point where each dot represents a cell colored according to cluster and lines represent branches. Center: subway map plot, the flat tree point is reordered according to a user-defined initial cell. Right: stream plot showing cell density along different trajectories. Note that both the subway map plot and the stream plot allow to visualise gene expression for a gene of interest. Adapted from *Chen, Albergante, et al., Single-cell trajectories reconstruction, exploration and mapping of omics data with STREAM, Nature Communications, 10, 2019, doi: 10.1038/s41467-019-09670-4* 35

2.11 Velocyto RNA-velocity representation. **a.** (top-left) Schematic representation of the compartmental model underlying the Velocyto rate equations. (bottom-left) The gene-specific phase portrait according to the Velocyto model: the dashed line represents the degradation rate γ, which separates the regions of gene induction (above γ) and gene repression (below γ); each cell, depicted as a red data point, can be plotted on this phase portrait, and the distance between the cell's position and the γ line corresponds to the RNA-velocity. (right) Visualization of the spliced and unspliced mRNA abundances as a function of the transcription rate α: this plot illustrates how changes in the transcriptional rate are reflected in the differential dynamics of these two RNA molecules. **b.** RNA-velocity vectors projection on the t-SNE embedding of the hindbrain of adolescent (P20) mice (left) and gene phase portraits (right). Adapted from *La Manno, G., Soldatov, R., Zeisel, A. et al. RNA-velocity of single cells. Nature 560, 494-498 (2018). `https://doi.org/10.1038/s41586-018-0414-6`.* 49

2.12 RNA-velocities derived for the pancreatic endocrinogenesis using the scVELO dynamical model (**a.**) and the Velocyto steady-state approach (**b.**). **c.** Phase portrait comparison for the *Cpe* gene recovered from the models: the steady state model incorrectly assigns *α-cells* to the repression phase. **d.** scVELO's latent time can better identify the cell's positioning in the biological development compared to pseudotime. Adapted from *Bergen, V., Lange, M., Peidli, S. et al. Generalizing RNA-velocity to transient cell states through dynamical modeling. Nat Biotechnol 38, 1408-1414 (2020). `https://doi.org/10.1038/s41587-020-0591-3`* 50

2.13 Comparison of the scVELO and MultiVelo frameworks over the mouse skin dataset. **a.** Velocity stream plot and latent time estimation using MultiVelo predict two developmental lineages. **b.** Velocity stream plot using scVELO: transcriptomic data only cannot correctly recover developmental lineages. **c.** Relative proportion of each type of kinetics across all fit genes. Adapted from *Li, Virgilio, Collins, et al., Multi-omic single-cell velocity models epigenome-transcriptome interactions and improves cell fate prediction, Nature Biotechnology, 41:387-398, 2023, doi:10.1038/s41587-022-01476-y.*. 51

2.14 Graph (left) and its condensation graph (right). Dashed circles highlight connected components, while numbers are used to label each connected component and the respective supernode. Edges are colored according to the supernodes they are connecting in both the graph and its condensation version. 51

2.15 Cellrank probabilities computations. A reference cell i is highlighted with its velocity vector v_i and neighbors. The displacement vector s_{ij} represents the difference in gene expression for neighboring cells. Probabilities are computed considering the angle α between each displacement vector and the velocity one. 52

2.16 Transition matrix representation of an uncoupled Markov Chain - up to a perturbation of the order of the objects. The block-diagonal matrix (left) and the representation of the associated disconnected graph (right) match the block color with the respective component. 52

3.1 UMAP visualisation of the 10X Fresh Embryonic E18 Mouse Brain. Cells are clustered according to cell labels provided by MultiVelo. 54

3.2 Data preprocessing **a.** Quality Metrics computed for the dataset. **b.** PCA explained variance and elbow plot for scRNA-seq data. **c.** SVD dimensions correlation with sequencing depth. 57

3.3 Transition Matrix computation for the scATAC-seq and Multiomics+scVELO models. 60

3.4 Transition Matrix computation for the Multiomics+MultiVelo model. . . 61

4.1 Examples of condensation graphs $H_{\mathcal{G}}$ in the case of a strongly connected graph (**a.**) and a disconnected graph (**b.**). 69

4.2 Random walk simulation for scRNA-seq model. Edges color becomes lighter (yellow) as the simulations reach the end 70

4.3 Random walk simulations for scATAC-seq model. Edges color becomes lighter (yellow) as the simulations reach the end 71

4.4 Random walk simulations for Multiomics+scVELO model. Edges color becomes lighter (yellow) as the simulations reach the end 71

4.5 Random walk simulations for Multiomics+MultiVelo model. Edges color becomes lighter (yellow) as the simulations reach the end 72

4.6 Top 30 likely cells according to the invariant probability distribution in Multiomics+scVELO (**a.**) and Multiomics+MultiVelo (**b.**) models. 79

4.7 Scatterplot evaluating GPCCA macrostate quality for the scRNA-seq model. The dotted vertical line corresponds to the optimal $minChi = 0$ value, while dashed horizontal lines correspond to the $crispess = \{0.6, 0.7\}$ optimality thresholds. 80

4.8 Scatterplot evaluating GPCCA macrostate quality for the scATAC-seq model. The dotted vertical line corresponds to the optimal $minChi = 0$ value, while dashed horizontal lines correspond to the $crispess = \{0.6, 0.7\}$ optimality thresholds. 80

4.9 Scatterplot evaluating GPCCA macrostate quality for the Multiomics+scVELO model. The dotted vertical line corresponds to the optimal $minChi = 0$ value, while dashed horizontal lines correspond to the $crispess = \{0.6, 0.7\}$ optimality thresholds. 81

4.10 Scatterplot evaluating GPCCA macrostate quality for the Multiomics+MultiVelo model. The dotted vertical line corresponds to the optimal $minChi = 0$ value, while dashed horizontal lines correspond to the $crispess = \{0.6, 0.7\}$ optimality thresholds. 81

4.11 Initial (a.) and terminal macrostates with cell fate probabilities (b.) for the scRNA-seq model with $n_s = 2$. 82

4.12 Initial (a.) and terminal macrostates with cell fate probabilities (b.) for the scATAC-seq model with $n_s = 2$. 83
4.13 Initial (a.) and terminal macrostates with cell fate probabilities (b.) for the Multiomics+scVELO model with $n_s = 2$. 84
4.14 Initial (a.) and terminal macrostates with cell fate probabilities (b.) for the Multiomics+MultiVelo model with $n_s = 2$. 85
4.15 Macrostate composition for $n_s = 2$ in the scRNA-seq model. 86
4.16 Macrostate composition for $n_s = 2$ in the scATAC-seq model. 86
4.17 Macrostate composition for $n_s = 2$ in the Multiomics+scVELO model. . . 87
4.18 Macrostate composition for $n_s = 2$ in the Multiomics+MultiVelo model. . 87
4.19 Average cell fate probabilities towards the "RG, Astro, OPC" (**a.**) and "Deeper Layer" (**b.**) terminal states in the $n_s = 2$ scenario. 88
4.20 Initial (a.) and terminal macrostates with cell fate probabilities (b.) for the scRNA-seq model with $n_s = 3$. 89
4.21 Initial (a.) and terminal macrostates with cell fate probabilities (b.) for the scATAC-seq model with $n_s = 3$. 90
4.22 Initial (a.) and terminal macrostates with cell fate probabilities (b.) for the Multiomics+scVELO model with $n_s = 3$. 91
4.23 Initial (a.) and terminal macrostates with cell fate probabilities (b.) for the Multiomics+MultiVelo model with $n_s = 3$. 92
4.24 Macrostate composition for $n_s = 3$ in the scRNA-seq model. 93
4.25 Macrostate composition for $n_s = 3$ in the scATAC-seq model. 93
4.26 Macrostate composition for $n_s = 3$ in the Multiomics+scVELO model. . . 94
4.27 Macrostate composition for $n_s = 3$ in the Multiomics+MultiVelo model. . 94
4.28 Average cell fate probabilities towards the "RG, Astro, OPC" (**a.**) and "Deeper Layer" (**b.**) terminal states in the $n_s = 3$ scenario. 95
4.29 Average cell fate probabilities towards the "Upper Layer" terminal states in the $n_s = 3$ scenario. 96
4.30 KL-divergence for $n_s = 2$ (**a.**) and $n_s = 3$ (**b.**) 97
4.31 Entropy for $n_s = 2$ (**a.**) and $n_s = 3$ (**b.**) 98

A.1 Linear model results for the Multiomics+scVELO model with $n_s = 2$ and Deeper Layer cluster 111
A.2 Linear model results for the Multiomics+scVELO model with $n_s = 2$ and Upper Layer cluster 112
A.3 Linear model results for the Multiomics+scVELO model with $n_s = 2$ and RG, Astro, OPC cluster 113
A.4 Linear model results for the Multiomics+scVELO model with $n_s = 2$ and IPC cluster 114
A.5 Linear model results for the Multiomics+scVELO model with $n_s = 3$ and Deeper Layer cluster 115
A.6 Linear model results for the Multiomics+scVELO model with $n_s = 3$ and Upper Layer cluster 116
A.7 Linear model results for the Multiomics+scVELO model with $n_s = 3$ and RG, Astro, OPC cluster 117

A.8 Linear model results for the Multiomics+scVELO model with $n_s = 3$ and IPC cluster . . . 118
A.9 Linear model results for the Multiomics+MultiVelo model with $n_s = 2$ and Deeper Layer cluster . . . 119
A.10 Linear model results for the Multiomics+MultiVelo model with $n_s = 2$ and Upper Layer cluster . . . 120
A.11 Linear model results for the Multiomics+MultiVelo model with $n_s = 2$ and RG, Astro, OPC cluster . . . 121
A.12 Linear model results for the Multiomics+MultiVelo model with $n_s = 2$ and IPC cluster . . . 122
A.13 Linear model results for the Multiomics+MultiVelo model with $n_s = 3$ and Deeper Layer cluster . . . 123
A.14 Linear model results for the Multiomics+MultiVelo model with $n_s = 3$ and Upper Layer cluster . . . 124
A.15 Linear model results for the Multiomics+MultiVelo model with $n_s = 3$ and RG, Astro, OPC cluster . . . 125
A.16 Linear model results for the Multiomics+MultiVelo model with $n_s = 3$ and IPC cluster . . . 126

Chapter 1

Introduction

The underlying basis for the phenotypic variations observed in biological systems can be traced back to the intricate patterns of gene expression [1]. The process of gene expression, whereby the genetic information encoded within DNA is transcribed into RNA and subsequently translated into functional proteins, is governed by a complex network of regulatory elements and mechanisms. This includes core promoter sequences that initiate transcription by RNA polymerase, transcription factors that bind to specific DNA motifs to either activate or repress gene expression, distal enhancer sequences that can amplify transcriptional activity, and dynamic chromatin remodeling events that alter DNA accessibility. The delicate balance and interplay between these diverse gene regulatory components is critical for orchestrating the precise spatiotemporal expression of genes during cellular development and differentiation, and tissue specification [2]. Furthermore, disruptions or dysregulation of these gene regulatory mechanisms can lead to profound phenotypic changes, underpinning the onset and progression of various diseases, such as cancer [3, 4], neurological disorders [4, 5], metabolic syndromes [6], and autoimmune conditions [7]. Therefore, elucidating the intricate details of gene regulation has been a central focus of molecular biology and genetics research, as it holds the key to understanding the genotype-to-phenotype relationship.

Researchers have gained invaluable insights into the regulation of gene expression through the rapid advancements in DNA sequencing technologies. The introduction of next-generation sequencing (NGS) platforms has enabled the high-throughput, cost-effective analysis of genetic information at an unprecedented scale [8, 9]. These powerful sequencing techniques have facilitated the deciphering of the precise sequence and organization of DNA, allowing scientists to identify and characterize the key regulatory elements that govern gene expression programs. Furthermore, the development of single-cell sequencing approaches has revolutionized the field by providing a means to investigate gene regulation at the level of individual cells [10–12]. The ability to simultaneously profile multiple omics layers, including the transcriptome, epigenome, and proteome, within individual cells has further expanded our understanding of the intricate interplay between genetic, epigenetic, and post-transcriptional mechanisms that ultimately shape cellular phenotypes.

In the era of big data and high-throughput sequencing, sophisticated data analysis pipelines have emerged to effectively manage, analyze, and interpret the vast amounts of multi-omics data generated [13–17]. These computational workflows and algorithms help researchers to identify key regulatory sequences, model gene regulatory networks, integrate multi-omics data to elucidate system-level mechanisms and develop machine learning-based approaches for predicting gene expression patterns and cell development. By bridging the gap between experimental biology and computational analysis, these data pipelines play a pivotal role in advancing our understanding of the complex gene regulatory landscapes that underlie organismal development, physiological processes, and disease pathogenesis.

Among the multiple core research topics involving gene expression and its regulation, lineage tracing has emerged as a powerful approach to elucidate the developmental origins and differentiation trajectories of various cell types. Lineage tracing refers to the techniques used to track the progeny of individual cells or cell populations, allowing researchers to map the ancestral relationships and developmental pathways that give rise to the diverse cell types present within complex tissues and organs [18, 19].

Multi-omics approaches have been proven to be extremely useful in deepening our research understanding of cell type development [2]. By integrating genomic, transcriptomic, epigenomic, and proteomic data from individual cells, scientists can gain a comprehensive, systems-level view of the dynamic changes in gene regulation that underpin the transition from pluripotent or multipotent progenitor cells to terminally differentiated cell types. The ability to simultaneously profile multiple molecular layers within the same cell has enabled the identification of key transcription factors and epigenetic modifications that act as critical regulators of lineage specification and commitment [11]. Moreover, the application of single-cell multi-omics has revealed the remarkable heterogeneity that exists within seemingly homogeneous cell populations, providing unprecedented insights into the stochastic and probabilistic nature of cell fate decisions [16, 17]. Among the various lineage tracing algorithms, two approaches stand out as being of particular interest for this project: the CellRank framework [16] and the RNA-velocity models [20, 21]. CellRank is a computational tool that reconstructs developmental lineages from scRNA-seq data, whereas RNA-velocity leverages the dynamics of RNA splicing to infer the future state of individual cells, providing insights into the directionality and kinetics of cellular transitions.

The current landscape of single-cell sequencing-based lineage tracing algorithms incorporating multiomics data to investigate developmental lineages and cell fate decisions is still relatively narrow in scope [22, 23].

This thesis project introduces scVEMO, a computational pipeline that aims to contribute to the field of multiomics-based lineage tracing. The pipeline extends the capabilities of the CellRank framework by leveraging a combination of single-cell RNA sequencing (scRNA-seq) and single-cell assay for transposase-accessible chromatin sequencing (scATAC-seq) data. Specifically, we will compare four distinct models: the first is the original CellRank model based on scRNA-seq data and RNA-velocity; the second model exploits the CellRank pipeline by combining scATAC-seq data with RNA-velocity, to examine whether epigenomic data, when combined with gene velocities, can also be used to identify developmental lineages. The third model is based on a transition matrix constructed from the combination of RNA-velocity with multiomics single-cell (RNA and ATAC) profiles. Finally, we will replace the computations of scVELO RNA-velocities with the more advanced MultiVelo approach [21] in the fourth model. For the purpose of identification, we will refer to these four models as the scRNA-seq model, the scATAC-seq model, the Multiomics+scVELO model, and the Multiomics+MultiVelo model, respectively.

Our primary goal is to investigate whether multiomics-based approaches, namely the Multiomics-based model and the MultiVelo-based model, lead to more accurate identification of terminal states and developmental lineages. We will run each model over a combination of preprocessing parameters to demonstrate that our findings hold independently from the specific preprocessing parameter choices.

ScVEMO is validated on the Fresh Embryonic E18 Mouse Brain (5k) dataset from 10x Genomics, which includes single-cell ATAC and gene expression profiles of the fresh cortex, hippocampus, and ventricular zone of the embryonic mouse brain at day 18 [24].

Chapter 2

Background

This chapter provides a comprehensive description of the technologies, models, and information necessary to fully understand the context within which this thesis project operates.

First, the chapter outlines the technological context, delving into the details of the various technologies employed to obtain reliable single-cell profiles from different modalities, including both the data generation and the data processing pipelines that enable cellular characterization.

Complementing the technological overview, the chapter then elaborates on the concept of lineage tracing. This section presents a deep dive into the scope and objectives of the research, providing the reader with a comprehensive understanding of the underlying biological questions and the state-of-the-art algorithms and techniques available for investigating cellular differentiation and fate determination.

Finally, the chapter delves into the mathematical foundations and theoretical models that form the basis of the project's methodologies. This includes a detailed description of the CellRank framework and the RNA-velocity models.

2.1 Technological Context

This section's goal is to provide a comprehensive understanding of the technological context underlying the analysis conducted within the scope of this thesis project. This includes sequencing techniques, as well as the data analysis processes used for single-cell multiomics-based research.

The section begins with a brief overview of the fundamentals of the DNA structure and the intricate interplay among the different elements that govern gene expression. This foundational knowledge highlights the importance of integrating multiple cellular profiles when studying complex biological phenomena.

Next, the section introduces the Next Generation Sequencing (NGS) and Single-Cell sequencing technologies. It focuses specifically on single-cell RNA sequencing (scRNA-seq) and single-cell assay for transposase-accessible chromatin sequencing (scATAC-seq), as well as the data preprocessing pipelines associated with these techniques.

Finally, the section presents an example of a protocol used for the simultaneous profiling of gene expression and epigenomic landscapes within individual cells. This example illustrates the multiomics library generation processes that might be employed to create a multimodal dataset, such as the Fresh Embryonic E18 Mouse Brain (5K) one, over which the scVEMO framework is validated.

2.1.1 The DNA Structure

This section provides a concise introduction to the fundamental biological processes that underpin the complexity of the genome and the intricate mechanisms governing cellular development. Beginning with an examination of the chemical structure and informational properties of deoxyribonucleic acid (DNA), the section establishes a foundational understanding of how the sequence of nucleotides within this macromolecule encodes the genetic instructions necessary for life. The discussion then transitions to an exploration of the higher-order packaging of DNA into chromatin - a dynamic nucleoprotein complex that facilitates the compaction and regulation of the genetic material within the cell nucleus. This organizational framework is of critical importance, as it serves to modulate the accessibility and transcriptional activity of the encoded genetic programs.

Building upon this structural knowledge, the section delves into the intricate mechanisms of gene regulation, highlighting the pivotal role played by cis-regulatory DNA elements, such as promoters and enhancers, in controlling the spatiotemporal expression of genes. These regulatory sequences act in concert with trans-acting factors, including transcription factors and epigenetic modulators, to orchestrate the precise activation or repression of genetic programs, thereby enabling the coordinated development and differentiation of cells. Finally, the section examines the process of transcription, whereby the information encoded within DNA is faithfully transcribed into ribonucleic acid (RNA) molecules. This essential biological process serves as the critical bridge between genotype and phenotype, as the resulting RNA transcripts direct the synthesis of functional proteins - the fundamental units of cellular structure and function. By providing this comprehensive understanding of DNA structure, chromatin organization, gene regulation, and transcriptional mechanisms, this section lays a robust theoretical foundation for the reader to contextualize the more advanced topics and research questions explored in the subsequent chapters of the thesis.

At the fundamental core of all biological life lies the deoxyribonucleic acid (DNA) molecule - the essential macromolecule that encodes the genetic information necessary for the development and function of living organisms. DNA consists of two complementary polynucleotide strands, each composed of a linear sequence of nucleotide subunits including the nitrogenous bases adenine, cytosine, guanine, and thymine. These strands are held together through hydrogen bonding, giving rise to the iconic double helix structure (Fig. 2.2a).

The genome, representing the complete set of genetic information carried by an organism, is contained within the nucleus of each somatic cell. During cellular division, the genome is faithfully replicated, ensuring the propagation of identical genetic material to daughter cells. The symmetry of the DNA double helix enables the reliable copying of the nucleotide sequence through various critical cellular processes, including DNA replication, transcription, DNA repair, genetic recombination, plasmid replication, and viral genome replication [1].

Genes, the fundamental units of heredity, are defined as specific DNA sequences that encode the instructions for the synthesis of functional biomolecules, predominantly proteins. The nucleotide sequence within a gene directly determines the amino acid sequence of the resulting protein, which in turn governs the protein's three-dimensional structure and, consequently, its biological function [1]. These differences in genetic information are the primary contributors to the remarkable biological diversity observed across species. By analyzing the nucleotide sequences of genes, researchers can gain valuable insights into the unique characteristics and evolutionary relationships of different organisms.

The structure of a gene typically includes several key regulatory elements that play crucial roles in its expression and transcriptional control (Fig. 2.1a). Promoters are DNA sequences located upstream of the transcription start site (TSS) that serve as binding sites for the RNA polymerase machinery, thereby facilitating the initiation of transcription [25]. Beyond the core promoter region, distal regulatory elements known as enhancers can significantly augment transcription ease, regardless of their orientation or distance from the gene. Enhancers provide binding sites for various transcription factors and co-regulatory proteins, which can enhance the recruitment and activity of the transcription machinery [25].

Transcription factors (TFs) are DNA-binding proteins that recognize and bind to specific sequences, such as those found in promoters and enhancers, and subsequently regulate the transcription of genes during RNA molecule formation. These trans-acting factors can act as either activators or repressors, respectively increasing or decreasing gene expression. Furthermore, transcription factors can recruit or interact with the RNA polymerase, chromatin remodeling complexes, and other regulatory proteins to modulate transcriptional activity [25–28].

The gene itself is composed of both coding regions, known as exons, and non-coding regions, referred to as introns [1]. The intricate interplay between promoters, enhancers, transcription factors, and the spliceosome machinery allows cells to precisely control the spatial, temporal, and quantitative expression of genes in response to various developmental, environmental, and physiological cues. Proper gene regulation is essential for maintaining cellular homeostasis, facilitating appropriate responses to changing conditions, and preventing the development of various diseases, such as cancer and genetic disorders [3, 4, 29].

The genetic information stored in DNA is maintained in the form of chromatin, which allows DNA to undergo the necessary folding to form chromosomes. Chromatin consists of nucleosomes, which are structures composed of approximately 147-145 DNA base pairs wrapped around proteins called histones, as illustrated in Fig. 2.2b.

While there have been conflicting observations in experimental studies about the exact consequences of DNA methylation [30], this epigenetic modification process plays a crucial role in regulating epigenetic changes that impact nucleosome positioning [31–33] and, consequently, gene expression and transcription [34]. For example, chromatin remodelers utilize adenosine triphosphate (ATP) energy to disrupt the interaction between histones and DNA, enabling remodeling processes such as nucleosome ejection and nucleosome sliding, which modify the structure and positioning of nucleosomes [31, 33]. Nucleosome ejection involves the complete removal of a nucleosome from the DNA, creating an open region that is accessible for transcription factors and other regulatory proteins, while nucleosome sliding refers to the process by which a nucleosome is moved along the DNA strand without being completely removed (Fig 2.2b). Additionally, pioneer transcription factors can also contribute to increased chromatin accessibility by binding to closed chromatin regions and facilitating the binding of other transcription factors [27]. In general, different levels of DNA methylation profiles have been associated with transcriptional ease [32]. For instance, low-expressed genomic regions often exhibit a relatively uniform methylation profile, while highly expressed ones tend to display reduced methylation towards the transcription termination site (TTS).

Moreover, depending on the degree of chromatin structure condensation, chromatin exists in two primary forms: euchromatin and heterochromatin [33]. Euchromatin corresponds to a more open and unfolded state of chromatin with widely spaced nucleosomes, which is associated with higher gene expression levels and facilitates transcription. The euchromatin state allows transcriptionally active gene bodies to be accessed, particularly by transcription factors (TFs) in promoter and enhancer regions, leading to productive transcription. In contrast, heterochromatin refers to a more condensed and closed chromatin form.

The core of the gene expression process is the transcription mechanism, whereby a single strand of DNA is copied into a complementary ribonucleic acid (RNA) molecule (Fig. 2.3). The transcription of genes is catalyzed by the enzyme RNA polymerase II. This key transcriptional machinery first binds to the promoter region of the target gene, specifically recognizing the TATA-box sequence element. The TATA-box serves as a critical transcription initiation site, providing the docking platform for RNAPII to begin synthesizing a complementary RNA transcript from the DNA template. This binding event unwinds the DNA double helix, exposing the template strand (initiation phase) [1]. The RNA polymerase then translocates along the DNA template, sequentially adding ribonucleotides complementary to the DNA sequence, thereby synthesizing the growing RNA chain in the 5' to 3' direction (elongation phase). Finally, the RNA polymerase reaches a termination signal on the DNA, which triggers the release of the completed RNA transcript. This primary RNA molecule, known as the precursor messenger RNA (pre-mRNA, Fig. 2.1b), undergoes further processing, such as splicing and modifications, before being exported from the nucleus for subsequent translation into functional proteins [1].

Finally, the RNA splicing process is essential for converting the primary RNA transcript (pre-mRNA) into a mature mRNA molecule (Fig. 2.1c). During this critical post-transcriptional modification, the spliceosome machinery excises the non-coding intronic sequences from the pre-mRNA, joining the remaining protein-coding exonic sequences to produce the final mRNA product [1]. Through the process of alternative splicing, multiple distinct mRNA isoforms can be generated from a single gene by selectively including or excluding different combinations of exons [35]. This allows a single genetic locus to encode for functionally diverse protein products, vastly expanding the coding capacity of the genome.

The precise mechanistic relationships and hierarchical coordination between the diverse factors in orchestrating transcriptional control have not been fully resolved. Enhancing our comprehensive understanding of such complex regulatory networks is therefore of paramount importance for illuminating the underlying principles of gene regulation. Emerging multiomics approaches have shown promise in beginning to unravel the intricate interdependencies between this myriad regulatory elements [2]. By integrating high-throughput datasets encompassing genomic, transcriptomic, proteomic, and epigenomic information, researchers have been able to elucidate previously obscure connections and causal relationships governing transcriptional programs. Elucidating the precise mechanisms by which enhancers, promoters, enzymes, transcription factors, and chromatin modifiers cooperate to fine-tune gene expression represents a critical frontier in molecular biology. Advancing this knowledge has broad implications, from enhancing our fundamental understanding of cellular function to informing the development of targeted therapeutic interventions for dysregulated transcriptional states underlying human disease.

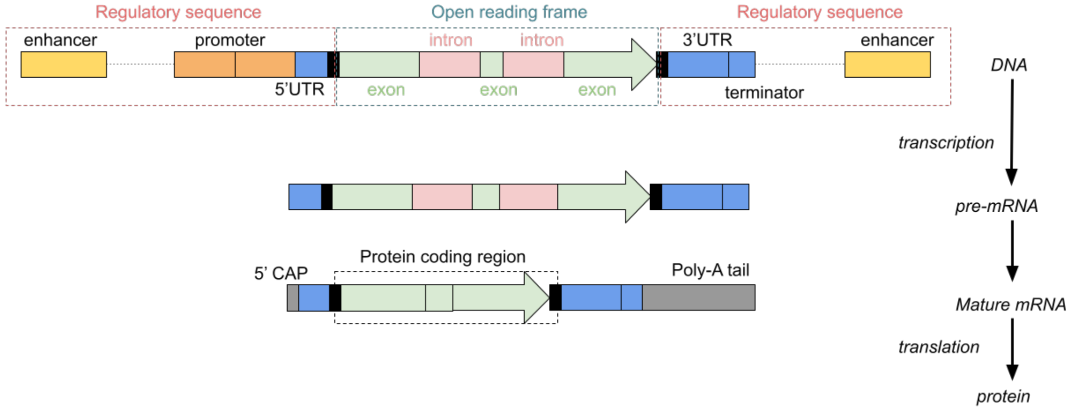


Figure 2.1: Gene structure: **a.** Simplified gene structure, including the cis-regulatory elements like enhancers and promoters, as well as the open reading frame that contains the sequence of introns and exons where the genetic information is stored **b.** The pre-mRNA molecule that is produced during the transcription process. **c.** The mature mRNA molecule that results after RNA splicing or manipulation operations. The protein-coding region contains the expressed exons, which will then be translated into a biologically functional protein. A 5' cap and poly-A tail are added to the RNA transcript to distinguish mRNAs molecules from other RNA products, indicating respectively, the beginning and the end of the transcript [1].

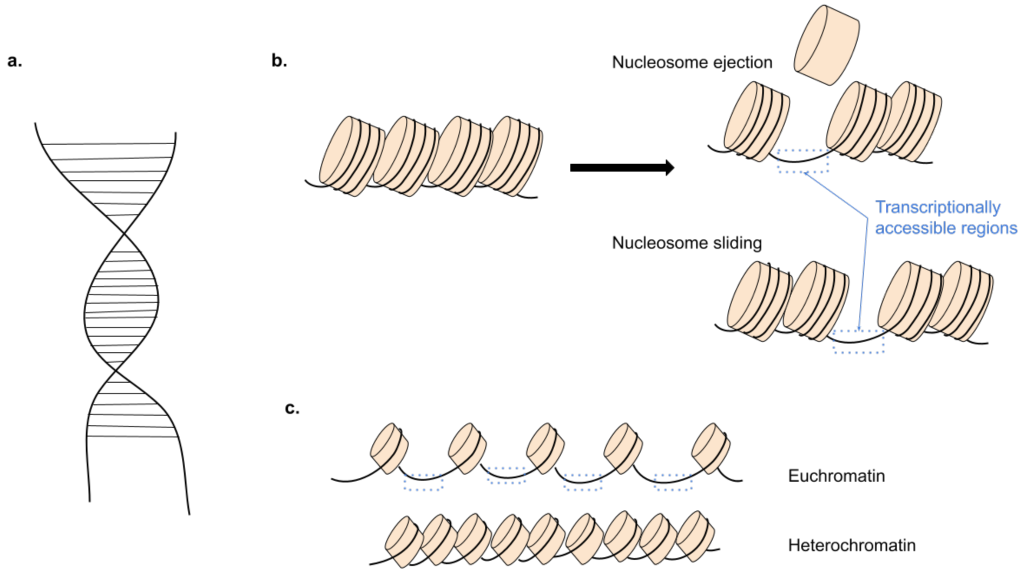


Figure 2.2: Chromatin structure: **a.** DNA helicoidal structure; **b.** DNA is wrapped around histones represented in yellow. Chromatin remodelers use ATP to dissociate DNA from histones (nucleosome ejection, top right) or modify nucleosome position (nucleosome sliding, bottom right). Chromatin structure modifications enable ease of transcription. **c.** Heterochromatin and euchromatin representation.

2.1.2 Next Generation Sequencing and Single Cell Sequencing Technologies

Understanding the phenotypic changes observed in a given biological system necessitates a comprehensive map of the underlying molecular mechanisms driving those changes. This can be achieved through DNA sequencing, which refers to the determination of the precise order of the four nucleobases that make up the DNA molecule. DNA sequencing techniques allow researchers to decipher the genetic code, providing valuable insights into the genetic architecture and potential drivers of the observed phenotypic variations. These sequencing approaches can be performed at various levels, from targeting single genes to analyzing entire chromosomes or even complete genomes.

The first DNA sequencing technologies, also referred to as Sanger Method, were introduced in the late 1970s, marking a significant milestone in the field of molecular biology and paving the way for a deeper understanding of genotype-phenotype relationships. The Sanger method [36] involved fragmenting and amplifying the DNA of interest, creating multiple copies of the template. During the sequencing process, nucleotides were incorporated complementary to each strand until a fluorescently labeled dideoxynucleotide (ddNTP) was added, serving as a chain terminator. This resulted in the generation of DNA fragments of varying lengths, each terminated by a ddNTP. These fragments were then separated by size using gel electrophoresis, allowing the precise sequence of

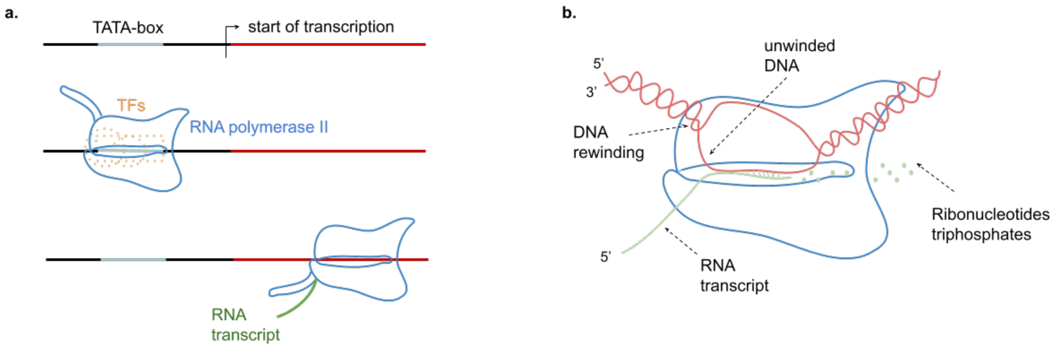


Figure 2.3: Transcription process. **a.** Schematic of the transcription process: Transcription factors (TFs, shown in yellow) and RNA polymerase II (RNAPII, shown in blue) bind to the promoter region, specifically the TATA-box sequence (shown in green), which specifies the transcription start site. Transcription then proceeds as RNAPII moves along the open reading frame, synthesizing a new RNA molecule. **b.** Insights into the RNAPII molecule: The DNA strands are locally unwound, and ribonucleotide triphosphates (rNTPs) are added complementary to the 3' strand of the DNA to form the new RNA transcript.

nucleotides in the original DNA template to be determined by the specific order of the fluorescent signals detected.

In recent years, the field of DNA sequencing has experienced a transformative shift with the advent of next-generation sequencing (NGS) technologies. These advanced sequencing platforms have dramatically increased the throughput and efficiency of genetic analysis, allowing researchers to generate vast amounts of genomic data at significantly lower costs compared to earlier sequencing methods [9]. Rather than the slow and labor-intensive Sanger sequencing approach, NGS technologies employ parallel processing to enable the simultaneous sequencing of huge amounts of DNA fragments.

Next-generation sequencing (NGS) technologies differ in several key aspects, including the number of DNA base pairs (bp) that can be read in a single sequence, the number of individual DNA fragments that can be sequenced in a single run, and the overall quantity of sequence data generated. Despite these differences, NGS approaches generally involve an initial sample preparation step. This typically consists of fragmenting the DNA samples and amplifying the resulting templates, often using polymerase chain reaction (PCR) techniques [8, 9], which is shown and described in Fig. 2.4.

Following sample preparation, the actual sequencing process is carried out. Even though the specific sequencing methodology can vary, NGS platforms are broadly classified into two main categories - second-generation and third-generation sequencing. Second-generation sequencing focuses on generating large numbers of relatively short sequence reads. In contrast, third-generation sequencing technologies have been developed to produce much longer sequence reads, sometimes exceeding thousands of base pairs, and can often perform sequencing in real-time as the DNA fragments are being synthesized, also without prior amplification steps [8].

After the sequencing step, an alignment process is performed to match the generated sequence reads to a reference DNA template. Alternatively, a de novo assembly approach can be used to reconstruct the target genomic sequences without the need for a pre-existing reference. The choice between reference-based alignment and de novo assembly depends on the specific research objectives and the availability of well-annotated reference genomes for the organism or system under investigation.

Building upon the advancements in NGS, single-cell sequencing techniques have emerged as a powerful tool for studying the genetic diversity and heterogeneity within complex biological samples [9]. Single-cell sequencing refers to the process of isolating and sequencing the nucleic acids (DNA or RNA) from individual cells, allowing for the precise characterization of gene expression patterns, genetic variations, and other molecular profiles at the single-cell level. However, the applications of single-cell technologies have expanded beyond just nucleic acid analysis and innovations in single-cell sequencing have therefore enabled the exploration of other cellular features and molecules, including the epigenome, transcriptome, proteome, and metabolome [2]. These approaches, known as multi-omic single-cell, provide a more comprehensive understanding of cellular identity, function, and interactions within heterogeneous cell populations. For example, single-cell epigenomics can reveal the chromatin accessibility and DNA methylation patterns of individual cells, while single-cell proteomics can quantify the abundance and post-translational modifications of proteins. By isolating and profiling individual cells, researchers can now uncover the remarkable diversity and complexity that exists within tissues, organs, and entire organisms. Single-cell multiomics techniques proved to be extremely accurate in applications such as cell subpopulation identification [10] [12], the reconstruction of cell hierarchy and developmental lineages [10,12,16,17,22,23], pseudotime reconstruction [12,17,22] and inference of gene regulatory networks [10–12].

2.1.3 Transcriptional Profiling of Individual Cells through scRNA-seq

Single-cell RNA sequencing (scRNA-seq) enables the dissection of transcriptional heterogeneity at the single-cell level, unlike traditional bulk RNA sequencing approaches that provide averaged, population-level insights [12, 37].

The scRNA-seq workflow begins with the isolation of individual cells, which can be achieved through a variety of techniques [10, 12, 37]. While labor-intensive methods such as laser capture microdissection and micromanipulation can be employed when sample sizes are limited, these approaches rely on visual identification and morphological assessment of cells [10, 37]. More time-efficient and high-throughput techniques, such as fluorescence-activated cell sorting (FACS) and microfluidic encapsulation, leverage fluorescence detection or light scattering properties to identify and isolate individual cells, but these methods typically require larger sample sizes [10, 37].

Once the single cells have been captured, their nuclei are lysed, and the released mRNA molecules are then reverse-transcribed into complementary DNA (cDNA). This reverse transcription step often relies on poly(T) priming, which can introduce technical noise due to biases in the efficiency of reverse transcription [10, 37]. To obtain the second

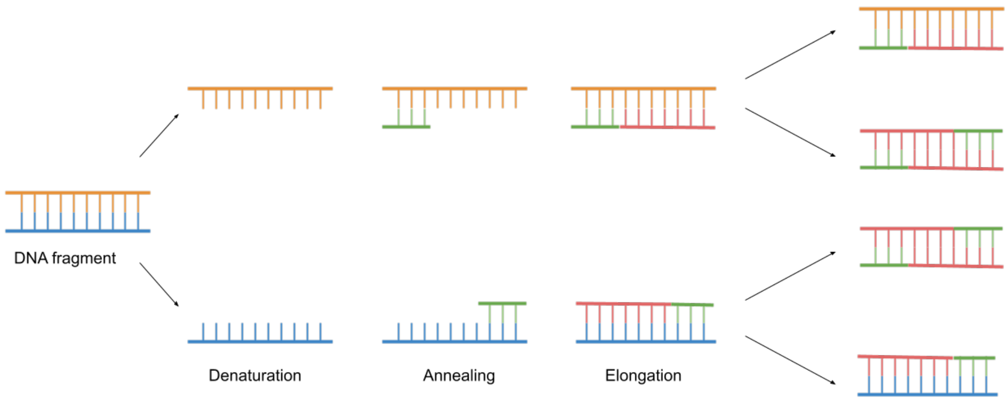


Figure 2.4: PCR mechanism. A single PCR cycle consists of three steps: **(1)** Denaturation, the process of separating the DNA fragments strands through heat; **(2)** Annealing, the process that binds a short synthetic DNA primer to the complementary sequence; the primer acts as starting point of the DNA synthesis process; **(3)** Elongation, where the new DNA strand is synthesized complementary to the template, therefore doubling the initial DNA molecule. Each cycle repeats the denaturation, the annealing and the elongation steps, resulting in an exponential increase of the DNA copies.

strand of cDNA, two main approaches are commonly used: poly(A) tailing and template-switching mechanisms. In the poly(A) tailing method, a poly(A) sequence is added to the 3' end of the first-strand cDNA, which then serves as a priming site for the synthesis of the second strand. Alternatively, the template-switching approach takes advantage of the terminal transferase activity of reverse transcriptase, which can add a short stretch of nucleotides to the 5' end of the first-strand cDNA. This allows for the subsequent annealing of a template-switching oligonucleotide, which then primes the synthesis of the second strand.

The amplification of the resulting double-stranded cDNA library can then be achieved through traditional PCR or in vitro transcription. This latter approach involves the use of RNA polymerase to generate multiple RNA copies from a single DNA template, thereby increasing the overall transcript representation [37].

Before sequencing, the samples often undergo a multiplexing step, where unique molecular identifiers (UMIs), hence short DNA barcodes, are introduced during the reverse transcription process. These UMIs uniquely tag each mRNA molecule within a cell, allowing for the accurate quantification of transcript levels and the elimination of PCR duplicates during data analysis [37]. Finally, the prepared and multiplexed libraries are subjected to high-throughput sequencing, generating millions of short reads that can be computationally mapped to the reference genome, enabling the quantification of gene expression levels in each individual cell.

Despite the transformative potential of single-cell RNA sequencing, the raw data generated through this approach can be inherently noisy and prone to technical biases [10, 12, 37]. Consequently, a robust preprocessing pipeline is essential to improve the quality and reliability of the scRNA-seq datasets prior to downstream analysis. The most used scRNA-seq data preprocessing and analysis platforms are Seurat [14] and Scanpy [15].

One key indicator of cell quality is the total number of reads (or percentage of reads) detected per cell. Cells with abnormally low or high read counts may represent dead cells or doublets (instances where multiple cells were captured together), respectively. Additionally, the proportion of reads mapping to the mitochondrial genome can provide valuable insights into cell health. A high percentage of mitochondrial reads may suggest that the cell is under stress or undergoing apoptosis, and such cells are often excluded from downstream analyzes [12]. Also, genes with insufficient sequencing depth, as indicated by low read counts, are also commonly discarded. Retaining only genes with adequate coverage helps to enhance the overall signal-to-noise ratio in the dataset. Furthermore, when scRNA-seq libraries are constructed using diverse experimental protocols or across multiple batches, systematic technical variations can be introduced. In such cases, specialized batch correction algorithms can be employed to integrate the data while accounting for these unwanted sources of variability [12].

Normalization is a critical subsequent step, aiming to adjust for biases stemming from differences in sequencing depth, as well as technical noise arising from dropouts and other artifacts [10,12]. By applying appropriate normalization strategies, both within and across samples, the data can be rendered more amenable for downstream comparative analyses and interpretations.

Finally, to tackle the inherent high dimensionality of scRNA-seq data, dimensionality reduction techniques are often leveraged. These include the identification of highly variable genes (HVGs) and the application of principal component analysis (PCA), which can facilitate feature selection while preserving the key properties of the system. Complementary visualization methods, such as t-SNE and UMAP, further enable the exploration and interpretation of the underlying cellular landscapes [12].

An illustration of the full scRNA-seq pipeline can be found in Fig. 2.5.

2.1.4 scATAC-seq: Mapping Chromatin Accessibility in Single Cells

Complementing the insights gained from single-cell RNA sequencing, the field of single-cell ATAC sequencing (scATAC-seq) has opened new avenues for the study of chromatin accessibility at the individual cell level.

Much like the scRNA-seq workflow, the scATAC-seq pipeline begins with the isolation and lysis of single cells, followed by the generation of a sequencing library. The cell isolation and lysis strategies can be broadly categorized into two main approaches: (1) split-and-pool combinatorial cellular indexing, which leverages 96-well plates and fluorescence-activated cell sorting (FACS) to uniquely barcode each cell, and (2) microfluidics-based methods, which offer higher sequencing depth per cell but can process fewer cells simultaneously [11]. Several variations of these techniques have been developed, including

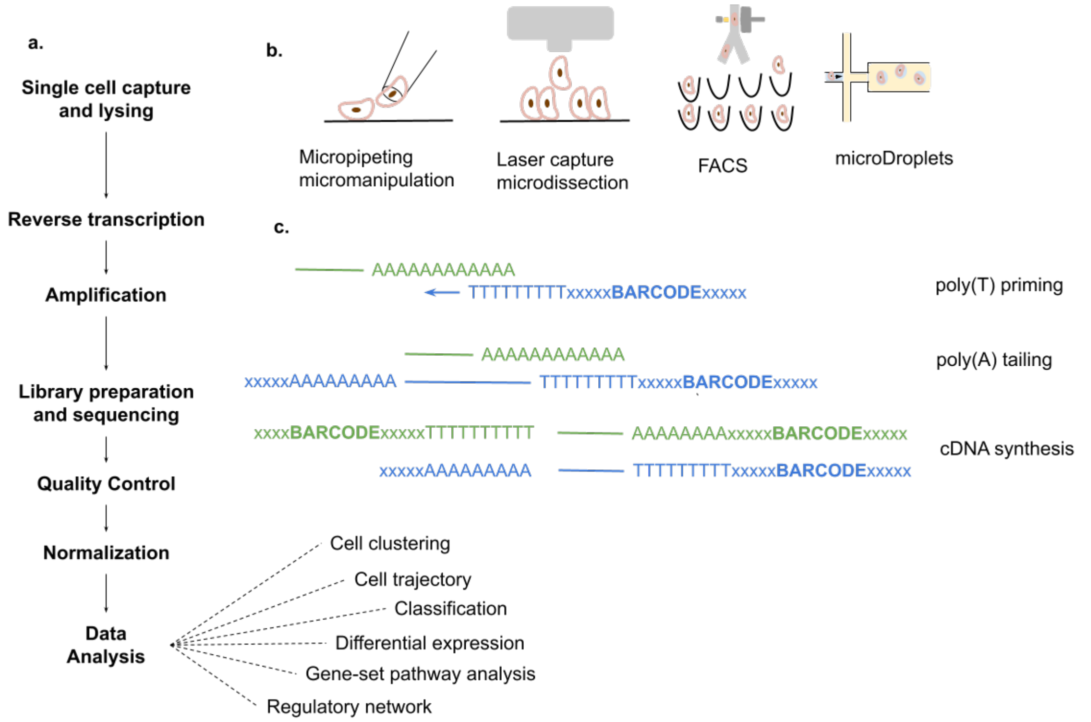


Figure 2.5: scRNA-seq data. **a.** scRNA-seq pipeline and most common data analysis procedures. **b.** Single cell isolation techniques. **c.** Reverse transcription with poly(T) priming and poly(A) terminator.

the 10X Genomics Chromium system [38], T-ATAC [39], plate-ATAC [40], and scip-ATAC [41].

A demultiplexing step is performed to deconvolute the cell-specific barcodes prior to sequencing. The resulting data is then subjected to quality control, where metrics such as barcode read depth, which can help identify low-quality cells or potential doublets mirroring the approaches used in scRNA-seq preprocessing [11]. More specific metrics such as the ratio of reads mapping to promoter regions or transcription start sites can be exploited as well.

The core output of scATAC-seq is a cell-by-peak matrix, where the peaks represent regions of open chromatin that are accessible to transcription factors and other regulatory elements. These peaks can be annotated to genomic features, often using methods like MACS2 [42].

It is important to note that scATAC-seq data inherently exhibits increased sparsity compared to scRNA-seq. While in scRNA-seq between 10-20% of the overall detectable information is actually sequenced [12, 37], with scATAC-seq this ratio decreases to 1-10% [11]. To cope with scATAC-seq data inherent sparsity, text mining approaches, such as term frequency-inverse document frequency (TF-IDF), have proven useful in highlighting the most informative peaks [11]. Consequently, the dimensionality reduction approaches leveraged in this scenario can be instrumental in mitigating the effects of technical noise and batch effects while preserving the biologically relevant variance. For dimensionality reduction, Singular Value Decomposition (SVD) is often exploited. The combination of TF-IDF and SVD is also known as Latent Semantic Indexing (LSI) and enables the effective handling of both high dimensionality and sparsity in the scATAC-seq data [11]. Complementary visualization techniques, like t-SNE and UMAP, can then be applied to the reduced-dimensionality data to aid in the exploration and interpretation of the chromatin accessibility landscape.

The resulting cell-by-peak matrix can be then exploited for a wide range of downstream analyses, such as cell type identification, the study of chromatin accessibility dynamics, TF-motif-based hypothesis generation, and enhancer-driven investigations [11].

2.1.5 Single Cell Multimodal Profiling: An Illustrative Example

The integration of complementary single-cell sequencing modalities has emerged as a powerful strategy to gain a more comprehensive understanding of cellular systems. Algorithms that leverage the combined information from scATAC-seq and scRNA-seq datasets have been shown to yield more accurate and robust results compared to the analysis of individual data types [2, 11]. Furthermore, the development of advanced multiomics sequencing approaches has enabled the simultaneous profiling of both the epigenome and transcriptome from the same single cells, providing an even richer perspective on the regulatory mechanisms underlying cellular identity and function [43].

In this section, as an example of a simultaneous multiomics sequencing technique, we will illustrate the 10X Genomics Chromium Single Cell Multiome ATAC + Gene Expression protocol [44], which has been used to obtain the cellular profiles of the Fresh Embryonic E18 Mouse Brain (5K) dataset [24] employed in this thesis project. The protocol overview for single-cell isolation and lysis is illustrated in Fig. 2.6. Specific information and further details can be found in the demonstrated protocol [45].

The single-cell isolation and lysis protocol prepares nuclei for the 10X Genomics Chromium GEM protocol [44]. This microfluidics-based platform allows for the construction of both single-cell RNA sequencing (scRNA-seq) and single-cell ATAC sequencing (scATAC-seq) libraries from the same cell sample. The key steps of the 10X Chromium Single Cell Multiome ATAC + GEX protocol employed for the dataset construction are listed below and illustrated in Fig. 2.7, while explicit details can be found in [44]. Generally, the protocol follows:

1. A transposase enzyme enters the cell nucleus and fragments DNA in open chromatin regions. Adapters are added to the DNA fragments' ends.

2. The transposed DNA fragments are then attached to GEMs (Gel Beads-in-Emulsion) containing poly(dT) and Spacer sequences (Fig. 2.7). Poly(dT) can hybridize to the poly-adenylated mRNA molecules enabling complementary DNA (cDNA) synthesis. cDNA will be further processed and sequenced to determine the gene expression profile of each individual cell. The Spacer sequence allows unique barcodes to bind the transposed DNA fragments representing regions of open chromatin. GEM refers to the encapsulation of individual gel beads, transposed nuclei, and other necessary

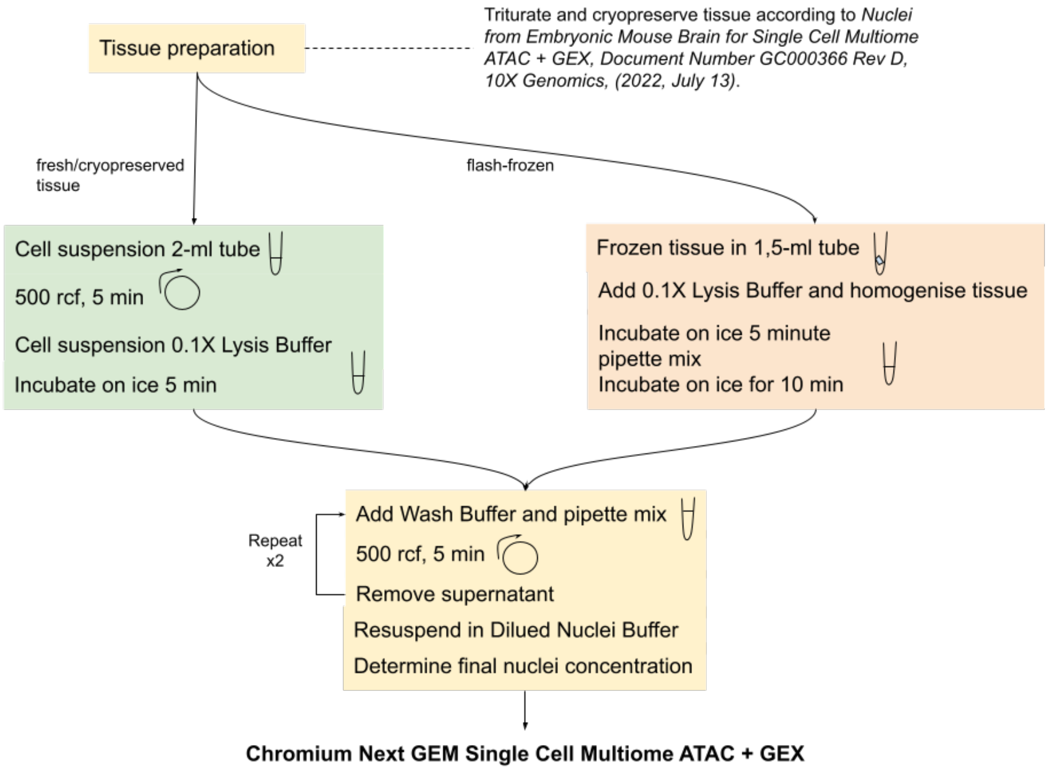

Tissue preparation
Triturate and cryopreserve tissue according to *Nuclei from Embryonic Mouse Brain for Single Cell Multiome ATAC + GEX, Document Number GC000366 Rev D, 10X Genomics, (2022, July 13).*
fresh/cryopreserved tissue
flash-frozen
Cell suspension 2-ml tube
500 rcf, 5 min
Cell suspension 0.1X Lysis Buffer
Incubate on ice 5 min
Frozen tissue in 1,5-ml tube
Add 0.1X Lysis Buffer and homogenise tissue
Incubate on ice 5 minute
pipette mix
Incubate on ice for 10 min
Repeat x2
Add Wash Buffer and pipette mix
500 rcf, 5 min
Remove supernatant
Resuspend in Dilued Nuclei Buffer
Determine final nuclei concentration
Chromium Next GEM Single Cell Multiome ATAC + GEX

Figure 2.6: Cell isolation and lysing pipeline overview from *Embryonic Mouse Brain for Single Cell Multiome ATAC + GEX, Document Number GC000366 Rev D, 10X Genomics, (2022, July 13).*

components within tiny droplets of oil. Such a process happens within the microfluidic Chromium Next GEM Chips. To achieve single-cell resolution, the nuclei are delivered at a limiting dilution, hence nuclei concentration is deliberately reduced in the input mixture to ensure that the majority of the generated GEMs (around 90-99%) do not contain any nuclei. GEMs, where a single nucleus is encapsulated along with the gel bead, provide the desired single-cell resolution for downstream analysis. The remaining GEMs are discarded during subsequent steps. Upon GEM generation, the gel bead is dissolved, releasing the components.

3. Pre-amplification using PCR of both cDNA and DNA.

4. ATAC library construction. The P7 sequence, an Illumina sequencing adapter, is added to the transposed DNA fragments. This P7 sequence is used in conjunction with the Illumina P5 sequence (previously added in the GEMs step) during Illumina bridge amplification.

5. Barcoded, full-length pre-amplified cDNA is amplified via PCR to generate sufficient mass for gene expression library construction.

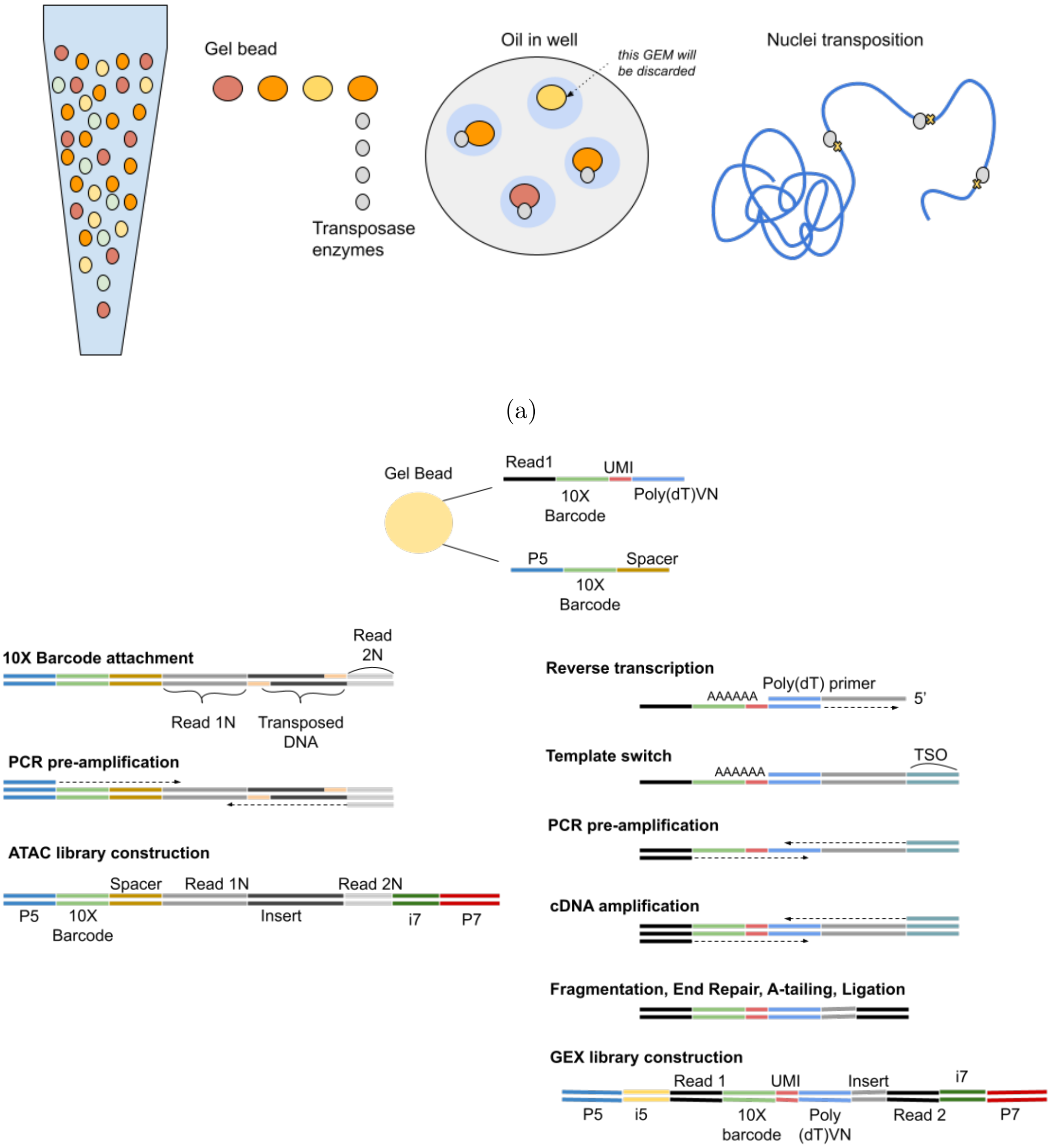
6. Gene Expression library construction. First cDNA samples are fragmented during the enzymatic fragmentation process. Then, cDNA fragments within a desired length are isolated during the size selection process. This step involves removing cDNA fragments that are either too small or too large, resulting in a narrower size distribution that is ideal for downstream processing. Subsequently, P5, P7, i7, and i5 sample indexes, along with the TruSeq Read 2 primer sequence, are added through multiple steps: End Repair, A-tailing, Adaptor Ligation, and PCR. End Repair involves repairing the ends of the cDNA fragments to ensure they have blunt ends, which are compatible with subsequent ligation steps. A-tailing is the addition of an adenine (A) nucleotide to the 3' end of the repaired cDNA fragments. This A-tailing process prepares the fragments for the ligation of adapters. Adaptor Ligation involves the attachment of specific adapters to the A-tailed cDNA fragments. These adapters contain the necessary sequences for subsequent sequencing and indexing steps. PCR is then performed to amplify the cDNA fragments with the attached adapters and incorporate the Illumina specific P5 and P7 primers.

7. Finally, sequencing is performed using The Illumina Novaseq 6000 v1 Kit. The Chromium Single Cell Multiome ATAC library consists of double-stranded DNA with standard Illumina paired-end constructs that begin with the P5 sequence and end with the P7 sequence. When coupled with i5 and i7, these sequences are crucial for demultiplexing and identifying the individual cells in the ATAC library. On the other hand, the Chromium Single Cell Multiome Gene Expression library consists of cDNA inserts with standard Illumina paired-end constructs that also begin with the P5 sequence and end with the P7 sequence. In this case, the combination of P5, P7, TruSeq Read 1 sequences and the 10X barcode UMI allow for cell identification and accurate quantification of gene expression at the single-cell level (Fig. 2.7b).

After library construction, the 10X Genomics CellRanger ARC 2.0.0 [46] pipeline is employed for preprocessing purposes. This workflow demultiplexes the Illumina BCL (Base Call) separating the sequencing reads into distinct GEX and ATAC FASTQ files. This crucial step ensures the proper deconvolution of the individual cells and their corresponding molecular signatures.

Next, the Cell Ranger ARC workflow generates single-cell feature count matrices for both the transcriptomic and epigenomic modalities. These matrices provide a quantitative representation of gene expression levels and chromatin accessibility profiles at the single-cell level, respectively.

Building upon these comprehensive datasets, the data analysis pipeline then also produces a suite of summary statistics, feature linkage analyses, unsupervised clustering, and dimensionality reductions. These powerful computational approaches enable the identification of distinct cell populations, the characterization of their unique molecular signatures, and the exploration of the intricate relationships between the epigenome and transcriptome.



(a)

(b)

Figure 2.7: Overview of the *Chromium Single Cell Multiome ATAC + GEX protocol, Document Number CG000338 Rev. F, 10X Genomics, (2022, August 26).* **a.** Visualisation of GEMs encapsulation of nuclei with transposase enzymes and nuclei transposition into open chromatin regions. **b.** ATAC library (left) and GEX library (right) construction steps and components.

2.2 Related Works

This section establishes the theoretical background underlying this thesis project. First, it introduces the field of lineage tracing, enabling the reader to understand the broader context within which the project stands and its related works.

Then, this section provides a comprehensive introduction to the mathematical background employed in this project. It delves into the key theoretical frameworks and models that form the basis of scVEMO, including the concept of RNA-velocity and the CellRank framework which lay at the basis of our model. The section also describes key principles underlying random walks which will be useful for model evaluation. Finally, the section elucidates the Generalized Perron Cluster Cluster Analysis (GPCCA) method, which enables the identification and characterization of the system's fully-differentiated states. By providing this comprehensive overview of the system model and the theoretical frameworks, this section lays the foundation for the specific methodologies, experiments, and findings of the research.

2.2.1 Lineage Tracing Methods and Algorithms

Lineage tracing refers to the process of tracking the developmental origins and differentiation trajectories of individual cells or cell populations within a tissue or organism [16–19, 22, 47, 48]. This powerful technique provides invaluable insights into the dynamic gene regulatory mechanisms that orchestrate cellular fate decisions and tissue patterning during development, regeneration, and disease. As a core topic of this research, we will introduce the state-of-the-art techniques and methods currently used for lineage tracing.

According to the comprehensive reviews by [18], we can broadly categorize lineage tracing approaches into two main groups: prospective and retrospective lineage tracing techniques. Additionally, the rapid advances in single-cell genomics have given rise to a suite of computational algorithms that leverage high-throughput single-cell data to reconstruct developmental lineages [19, 47]. In the following sections, we will delve into the details of some lineage tracing methodologies, discussing their underlying principles, experimental workflows, and computational frameworks. This comprehensive overview of the lineage tracing landscape will serve as a solid foundation for the experimental design and data analysis components of the thesis work.

Prospective lineage tracing [18, 19] involves techniques that leverage the integration and tracking of genetic barcodes within target cells and their progeny. These strategies utilize the introduction of synthetic DNA sequences, such as random DNA oligonucleotides or viral genetic elements, that serve as heritable markers within the target cells. As these cells divide and differentiate, the unique barcodes are passed on to their descendants, allowing researchers to reconstruct the lineage relationships between different cell types and infer the underlying gene regulatory mechanisms. Some approaches exploit the integration of DNA transposons [18, 49]. Another widely used approach is the Cre-LoxP system, where the site-specific recombinase Cre induces the expression of a reporter gene (e.g., fluorescent proteins) in a cell-type-specific manner, marking the lineage of those cells and their progeny [18, 50]. More recently, CRISPR-based lineage tracing leverages the ability of the Cas9 endonuclease to introduce targeted DNA modifications, such as small insertions or deletions, within the genome of target cells [18,19]. As these cells divide, the unique genetic scars are passed on to their descendants, creating a heritable barcode that can be used to reconstruct the lineage relationships between cells. Collectively, these prospective lineage tracing techniques have become invaluable tools for dissecting the complex cellular hierarchies and developmental trajectories underlying diverse biological systems. However, such strategies can be challenging to implement when working with large and complex organisms. In such cases, researchers can leverage a second category of lineage tracing approaches, known as retrospective lineage tracing.

Retrospective lineage tracing [18, 19] techniques exploit naturally occurring spontaneous mutations that can be inherited by daughter cells. Copy Number Variants (CNVs) are large-scale structural variations in the genome, such as deletions, duplications, or amplifications of DNA segments, that can be detected and tracked across cells. Single Nucleotide Variations (SNVs), on the other hand, refer to single base pair changes that accumulate in the genome over successive cell divisions. Microsatellite repeats are short, tandem repeats of DNA sequences that exhibit high mutation rates, making them suitable as lineage-specific markers. Finally, LINE-1 elements are transposable genetic elements that can undergo random insertions into the genome, creating unique integration sites that can be leveraged for lineage tracing.

Single-cell sequencing has emerged as a powerful tool for inferring developing lineages by leveraging the transcriptional and epigenomic profiles of individual cells. The underlying premise is that cells with similar transcriptomes are likely to be found in close proximity within the differentiation trajectories, as trascriptomics is strictly related to cell functioning and identification [18,47]. Furthermore, transcription factors are known to play a crucial role in shaping the genomic landscape and orchestrating the precise spatiotemporal patterns of gene expression that define cellular identity and fate [47]. A growing number of single-cell data-based lineage tracing algorithms have been developed to reconstruct developmental lineages, ordering cells along pseudotime and identifying developmental branches.

A well-established set of single-cell lineage tracing techniques exploits the underlying manifold structure of the sc-RNA sequencing data to approximate cell state transitions and reconstruct developmental trajectories [16,17]. These approaches begin by constructing a nearest neighbor (NN) graph, where the vertices represent individual cell states, and the edges connect the most similar cell states based on their transcriptomic profiles. The CellRank algorithm [16] further enhances this graph-based representation by incorporating RNA-velocity information. RNA-velocity is a measure of the rate of change in gene expression, which can be used to infer the directionality of cellular transitions. By leveraging the correlation between each cell's RNA-velocity and the differences in neighboring cell transcriptomes, CellRank can direct the NN graph and compute cell-cell transition probabilities. In contrast, the PALANTIR framework [17] employs diffusion maps [51] and the projection of the data onto the top diffusion components to compute the pseudotime ordering of the cells. This pseudotime information is then used to direct and weight the NN graph, leading to the construction of the cell-cell transition probability matrix. Both CellRank and PALANTIR utilize the obtained transition probability matrices to simulate the developmental system using Markov Chains (MC). However, they differ in the strategies for identifying the initial and terminal states within the system. PALANTIR identifies the absorbing states of the MC, which correspond to the terminal cell fates. CellRank, on the other hand, exploits a Generalized Perron Cluster Cluster Analysis (GPCCA) [52] to perform spectral clustering on the MC transition probability matrix, summarizing the cell-cell system into broader macrostates. These macrostates can then be further divided into initial, terminal, and intermediate states of the biological developmental process. Ultimately, both CellRank and PALANTIR output developmental lineages, cell-fate transition probabilities, and the associated gene expression patterns along specific trajectories, offering researchers a comprehensive understanding of the complex cellular differentiation dynamics (Fig. 2.8 2.9).

In addition, another set of single-cell lineage tracing algorithms exploits the integration of scRNA-seq and scATAC-seq data to provide a more comprehensive visualization of developmental lineages in a branching tree structure [22, 23]. One such framework is STREAM [22], which first projects the combined single-cell sequencing data onto a lower-dimensional space using the Modified Locally Linear Embedding (MLLE) method [53]. It then infers the cellular trajectories by applying an optimized version of the Elastic Principal Graph (ElPiGraph) algorithm. In contrast, the MIRA framework [23] takes a different approach, employing a variational autoencoder to perform topic modeling on the integrated scRNA-seq and scATAC-seq data. It then constructs a k-nearest neighbor (KNN) graph based on the identified accessibility and expression topics, which is used to build a developmental tree structure. Furthermore, MIRA leverages regulatory potential (RP) modeling to understand the regulatory influence of chromatin accessibility on gene expression patterns along the branches of the reconstructed lineage tree. Both STREAM and MIRA result in tree-structured visualizations that summarize the developmental lineages, including branching points, pseudotime ordering, cell type density, and the underlying regulatory mechanisms driving the cellular transitions along the different branches (Fig. 2.10).

Finally, there exists a category of single-cell techniques that aim to identify the biological development of a system by explicitly leveraging the concept of RNA-velocity [20, 21, 48]. RNA-velocity refers to the rate of change in a cell's gene expression, which can be inferred from the ratio of unspliced to spliced mRNA transcripts detected via single-cell RNA sequencing. This temporal information provides a powerful means to infer the directionality of cellular transitions. The pioneering velocyto model [48] and its subsequent refinement, scVELO [20], exploit a system of two ordinary differential equations (ODEs) to describe the evolution of spliced and unspliced RNA. These models incorporate gene-specific parameters for transcription, splicing, and degradation rates. While the original velocyto approach assumes a common splicing rate across all genes and relies on a steady-state approximation, scVELO overcomes these limitations by introducing gene-specific splicing rates and cell state parameters resulting in a more generalized framework that can handle transient systems. The MultiVelo model [21] further expands the ODE system

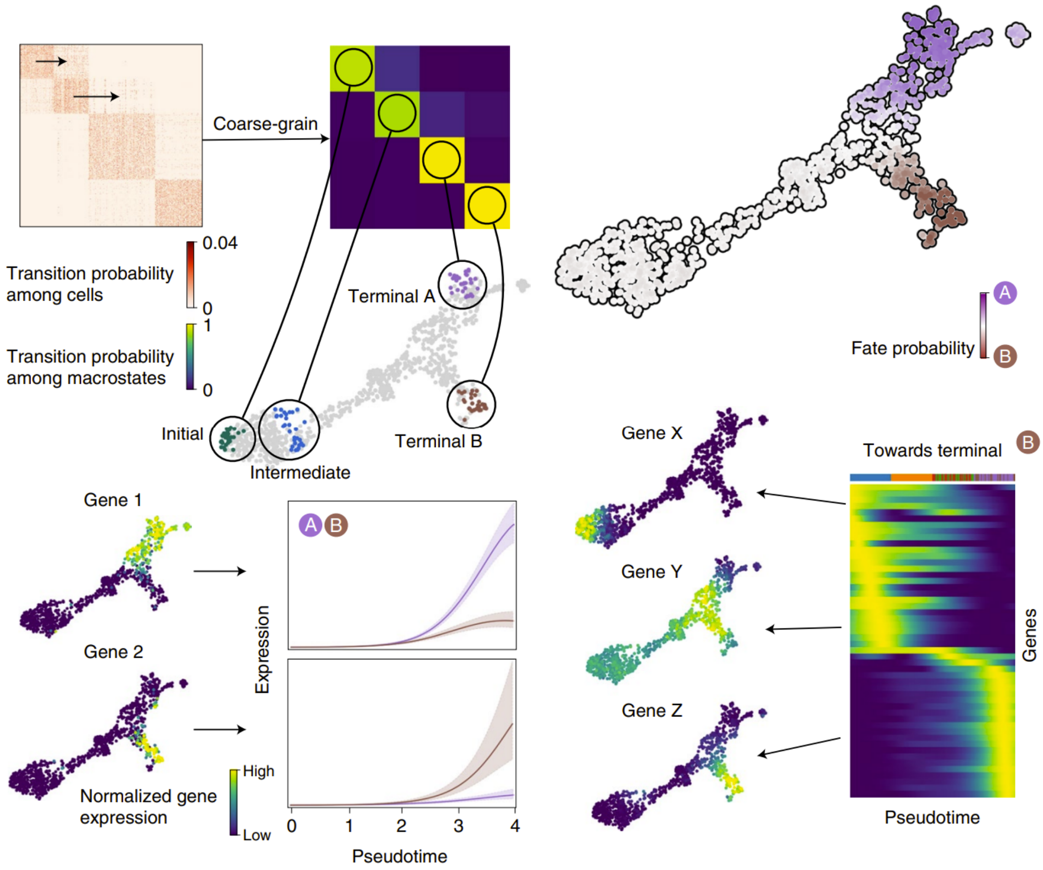


Figure 2.8: Overview of the CellRank framework. Upper left: the cell-cell transition probability matrix is coarse-grained using GPCCA into four macrostates which are divided into an initial, an intermediate, and two terminal states. Highlighted cells represent the 30 top likely cells to belong to the terminal state. Upper-right: cell fate probabilities mapped over UMAP embedding; each cell is colored according to the terminal state it is more likely to reach. Bottom left: normalized gene expression plot over the UMAP representation and lineage-specific gene expression trends along pseudotime. Bottom-right: heatmap representing gene expression along pseudotime for the top 50 genes correlating with terminal state B. Adapted from *Lange, Bergen, Klein et al., Cellrank for directed single-cell fate mapping, Nature Methods, 19:159-170, 2022, doi:10.1038/s41592-021-01346-6* Copyright

to include a third equation describing the rate of chromatin accessibility changes underlying the regulation of gene splicing and transcription. These models employ expectation-maximization (EM) approaches to estimate the relevant parameters from the single-cell data. The output of these RNA-velocity-based methods includes gene-specific phase portraits as well as lower-dimensional visualizations (UMAP, tSNE) that incorporate the inferred directionality of cellular trajectories (Fig. 2.13a and Fig. 2.12a,b). However, it is

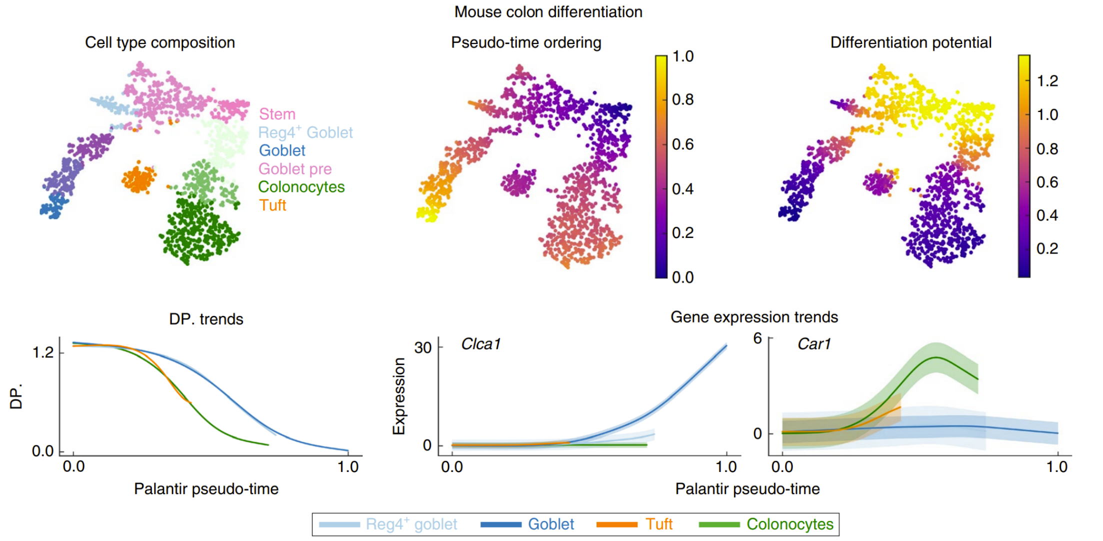


Figure 2.9: Overview of the Palantir framework. Top: tSNE map of scRNA-seq epithelial enriched cells from the mouse colon colored according to cluster, cell pseudotime, and differentiation potential on the tSNE embedding. Bottom: Differentiation potential trends and gene expression trends along pseudotime. Adapted from *Setty, Kiseliovas, Levine et al., Characterization of cell fate probabilities in single-cell data with Palantir, Nature Biotechnology, 37:451-460, 2019, doi:10.1038/s41587-019-0068-4*.

important to note that RNA-velocity estimates are extremely uncertain [16], and these techniques are often best utilized in combination with other single-cell lineage tracing algorithms to provide a more comprehensive and robust understanding of the underlying developmental processes.

2.2.2 RNA-velocity: Mapping Transcription Temporal Dynamics

RNA-velocity is a computational technique that leverages the differential abundance of spliced and unspliced RNA transcripts within scRNA-seq data to infer the future transcriptional state of each cell. By exploiting the temporal dynamics of gene expression, RNA-velocity suggests the directionality of cellular differentiation and developmental trajectories.

The underlying principle of RNA-velocity is the observation that unspliced, pre-mRNA sequences are indicative of recently activated genes, while spliced, mature mRNA molecules represent the cumulative gene expression history of a cell. By quantifying the relative abundance of these two RNA species, it is possible to estimate the rate of change (velocity) in a cell's transcriptional program.

Throughout Section 2.2.3, 2.2.4 and 2.2.5, we will explore the application of three major frameworks for RNA-velocity estimation: velocyto [48], scVELO [20], and MultiVelo [21].

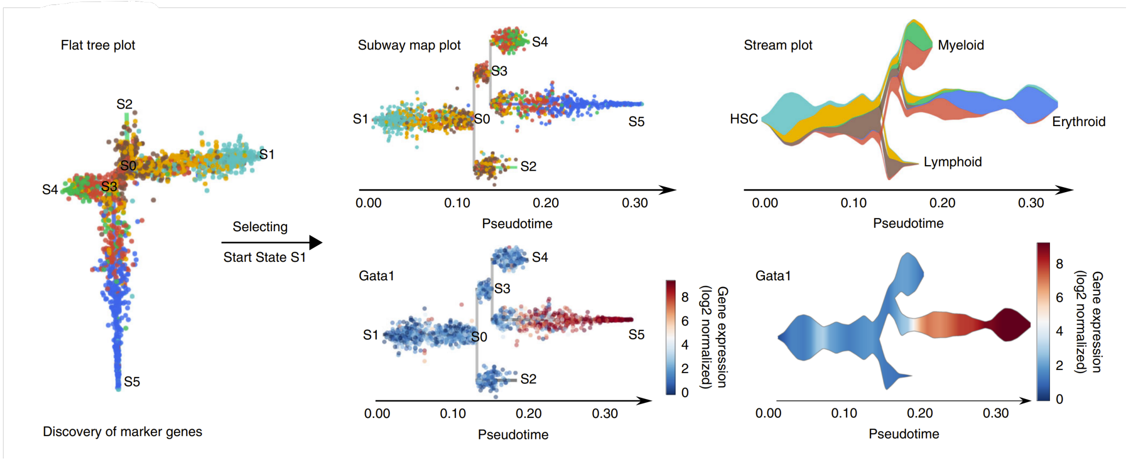


Figure 2.10: Illustration of the STREAM framework. Left: flat tree point where each dot represents a cell colored according to cluster and lines represent branches. Center: subway map plot, the flat tree point is reordered according to a user-defined initial cell. Right: stream plot showing cell density along different trajectories. Note that both the subway map plot and the stream plot allow to visualise gene expression for a gene of interest. Adapted from *Chen, Albergante, et al., Single-cell trajectories reconstruction, exploration and mapping of omics data with STREAM, Nature Communications, 10, 2019, doi: 10.1038/s41467-019-09670-4*

2.2.3 RNA-velocity: the Velocyto Framework

The foundational framework for RNA-velocity analysis is the groundbreaking Velocyto model [48], which is also known as the steady-state approach. This model relies on the key assumption that the full gene splicing dynamics, including the induction, repression, and steady-state phases, can be observed within the single-cell transcriptomic data.

At the core of the Velocyto model is a set of gene-specific, deterministic, and continuous-valued rate equations that describe the time evolution of the expected number of spliced and unspliced mRNA molecules. Specifically, for each gene, the model follows the system of equations:

$$\begin{aligned}\frac{du(t)}{dt} &= \alpha(t) - \beta(t)u(t)\\ \frac{ds(t)}{dt} &= \beta(t)u(t) - \gamma(t)s(t)\end{aligned} \tag{2.1}$$

In these equations, $\alpha(t)$, $\beta(t)$, and $\gamma(t)$ represent the gene-specific transcription, splicing, and degradation rates, respectively, while $u(t)$ and $s(t)$ denote the abundances of unspliced and spliced mRNA molecules at time t. Constant values for the gene-specific transcription and degradation rates are assumed, such that $\alpha(t) = \alpha$, $\gamma(t) = \gamma$, while the splicing rate $\beta(t) = \beta = 1$ is constant and shared across all genes.

The analytical solution to the system of equations Eq. 2.1 can be derived as follows:

$$
\begin{aligned}
u(t) &= \alpha(1-e^{-t}) + u_0 e^{-t} \\
s(t) &= \frac{e^{-t(1+\gamma)}\left[e^{t(1+\gamma)}\alpha(\gamma-1) + e^{t\gamma}(u_0-\alpha)\gamma + e^{t}\left(\alpha - \gamma(s_0+u_0+s_0\gamma)\right)\right]}{\gamma(\gamma-1)}
\end{aligned}
\tag{2.2}
$$

Here, u_0 and s_0 represent the initial abundances of unspliced and spliced mRNA, respectively, at time $t = 0$. The analytical solutions derived for the Velocyto model's rate equations provide a powerful means of extrapolating mRNA abundances to future time points. However, this approach requires the estimation of the gene-specific parameters α and γ.

The Velocyto model leverages a steady-state assumption, where the rate of change in spliced mRNA abundance is considered constant over time, i.e., $\frac{ds(t)}{dt} = 0$ for all t. Under this assumption, the model parameters can be derived as follows:

$$
\begin{aligned}
\gamma &= \frac{u}{s} \\
\alpha &= u
\end{aligned}
\tag{2.3}
$$

The Velocyto framework provides a straightforward approach to estimating the degradation rate γ for a given gene. This is achieved by fitting a linear model, where the size-normalized unspliced mRNA abundance $\hat{u}$ is regressed against the size-normalized spliced mRNA abundance $\hat{s}$ across all cells. In this formulation, the slope of the linear regression line directly corresponds to the degradation rate γ. Graphically, this parameter can be visualized on the gene-specific phase portrait, as depicted in Fig. 2.11.

While the Velocyto model provides a straightforward approach to estimating the gene-specific degradation rate γ, the transcription rate α cannot be determined. This poses a challenge when trying to extrapolate the future abundances of spliced mRNA molecules, $s(t)$, using the model's analytical solutions.

To address this issue, the framework proposes two alternative approaches:

1. The constant velocity assumption: RNA-velocity is considered constant over time, $\frac{ds(t)}{dt} = v$, leading to a simplified expression for the spliced mRNA abundance:

$$
s(t) = s_0 + vt \tag{2.4}
$$

2. The constant unspliced molecules assumption: $u(t) = u_0$. Under this assumption the analytical solution for $s(t)$ becomes:

$$
s(t) = s_0 e^{-\gamma t} - \frac{u_0}{\gamma}(1 - e^{-\gamma t}) \tag{2.5}
$$

The Velocyto framework offers two primary visualization techniques that enable the interpretation of RNA-velocity. The first approach involves projecting the estimated velocity vectors onto low-dimensional embeddings of the single-cell data, such as t-SNE or UMAP plots as shown in Fig. 2.11b. The resulting streamlines represent the inferred directionality of the transcriptional dynamics within the cellular state space. By visualizing the velocity vectors over the embedding, we can identify regions of the state space where cells are actively transitioning between transcriptional states, indicating areas of dynamic cellular differentiation.

The second visualization technique involves the construction of gene-specific phase portraits (Fig. 2.11b), which plot the unspliced versus spliced mRNA abundances for each gene. The degradation rate slope separates the regions of gene induction and gene repression. By projecting individual cells onto these phase portraits, we can examine the distance between the cell's position and the degradation rate line, which corresponds to the RNA-velocity for that cell and gene. This approach enables the identification of genes that are actively being induced or repressed within specific cellular subpopulations.

2.2.4 RNA-velocity: the scVELO Framework

The scVELO framework [20] aims to address two key limitations of the original Velocyto approach. Firstly, Velocyto relies on the steady-state assumption, which may not always hold true, particularly in the context of dynamic, heterogeneous cellular populations. The authors of scVELO recognized the need for RNA-velocity analysis to generalize to transient states and diverse cellular subpopulations. Secondly, Velocyto assumes a common splicing rate shared across all genes, whereas the scVELO model introduces gene-specific reaction rates to better capture the heterogeneity in transcriptional regulation. To address these limitations, the scVELO framework introduces a more comprehensive set of parameters, including not only the gene-specific reaction rates, but also additional variables describing the positioning of cells within the developing biological system.

The scVELO gene-specific deterministic rate equations now capture the dependency of the transcription rate on the cell's internal transcriptional state:

$$\begin{aligned} \frac{du(t)}{dt} &= \alpha^{(k)}(t) - \beta u(t) \\ \frac{ds(t)}{dt} &= \beta u(t) - \gamma s(t) \end{aligned} \tag{2.6}$$

In this formulation, the parameters $\alpha^{(k)}(t)$, β, and γ still represent the gene-specific transcription, splicing, and degradation rates, respectively. However, the transcription rate α is now dependent on the cell's state, as encoded by the parameter k. This cell state-dependent transcription rate $\alpha^{(k)}(t)$ allows the scVELO model to capture gene up- and down-regulation dynamics that are influenced by the specific transcriptional program of the cell. The parameter k represents either the induction, repression, and their related steady states. Consequently, the analytical solutions for the mRNA abundances also become k dependent:

$$\begin{aligned} u(t) &= u_0^{(k)} e^{-\beta\tau} + \frac{\alpha^{(k)}}{\beta}(1 - e^{-\beta\tau}) \\ s(t) &= s_0^{(k)} e^{-\gamma\tau} + \frac{\alpha^{(k)}}{\gamma}(1 - e^{-\gamma\tau}) + \frac{\alpha^{(k)} - \beta u_0^{(k)}}{\gamma - \beta}(e^{-\gamma\tau} - e^{-\beta\tau})\ \tau = t - t_0^{(k)} \end{aligned} \tag{2.7}$$

In this expanded analytical solution, the cell state k impacts both the transcription rate $\alpha^{(k)}$, the initial mRNA abundances $u_0^{(k)}$ and $s_0^{(k)}$, as well as the time point of switching between states $t_0^{(k)}$. By incorporating this cell state-dependent formulation, the scVELO model is able to more accurately infer the spliced mRNA abundance $s(t)$.

The scVELO framework employs the Expectation-Maximization (EM) algorithm [54] to simultaneously infer the key parameters of the model, including the gene-specific reaction rates, the cell-specific transcriptional states, and the switching time points between states. Given the size-normalized observed mRNA abundances, u^{obs} and s^{obs}, the EM approach aims to estimate the phase trajectory χ that best approximates the observations. This is achieved by minimizing the gene-specific negative log-likelihood function, which quantifies the goodness-of-fit between the estimated phase trajectory $x(\theta)$ and the observed data x^{obs}, where $\theta = \left(\alpha^{(k)}, \beta, \gamma\right)$:

$$l(\theta) = \frac{1}{2} log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\frac{1}{n}\sum_{i}^{n} ||x_i^{obs} - x_{t_i}(\theta)||^2 \tag{2.8}$$

In this formulation, σ^2 represents the variance of the normally distributed residuals between the observed and the estimated phase trajectory.

The EM algorithm consists of two iterative steps:

- E-step: given the current estimate of the model parameters θ, which parameterize the phase trajectory $x(\theta)$, the E-step assigns a latent time t_i to each observed data point x_i^{obs}. This is done by finding the time point on the trajectory that minimizes the distance between the observed data and the estimated phase trajectory. Additionally, the E-step assigns state likelihoods to each cell, based on their proximity to the different regions of the phase portrait.

- M-step: the algorithm updates the model parameters θ and the switching time points t_0 to better fit the observed mRNA abundance data via Nelder-Mead method [55].

By addressing the limitations of the original Velocyto model and introducing cell parameters and gene-specific splicing rates, the scVELO framework provides a more comprehensive and biologically realistic approach to RNA-velocity analysis. As illustrated in Fig. 2.12, the velocity stream plots and gene phase portraits obtained using the scVELO model showcase its enhanced capabilities compared to the Velocyto-based approach. For the pancreatic endocrinogenesis dataset [56], the velocity vectors computed by scVELO are able to accurately capture the cycling population of endocrine progenitors, a feature that the original Velocyto model was unable to achieve (Fig. 2.12a,b). Furthermore, the scVELO framework offers improved gene-specific phase portraits that enable better assignment of α-cells to induction and repression states, as compared to the Velocyto model (Fig. 2.12c). Finally, the scVELO model also provides enhanced latent time estimates that better position cells along the developmental trajectory compared to pseudotime (Fig. 2.12d).

2.2.5 RNA-velocity: the MultiVelo Framework

Building upon the advancements of the scVELO model, the MultiVelo [21] framework introduces further enhancements to capture the role of epigenomic changes in gene expression regulation. The MultiVelo model extends the gene-specific rate equations by incorporating a term to describe the dynamics of chromatin accessibility:

$$\begin{aligned}\frac{dc(t)}{dt} &= k_c\alpha_c - \alpha_c c(t)\\ \frac{du(t)}{dt} &= \alpha^{(k)}c(t) - \beta u(t)\\ \frac{ds(t)}{dt} &= \beta u(t) - \gamma s(t)\end{aligned} \tag{2.9}$$

In this formulation, the variable c represents the time-varying levels of chromatin accessibility. Specifically, it is the sum of accessibility at the promoter and linked peaks for a gene, which is later normalized to mimic a value for c approaching 1 in the opening and 0 in the closing phases. Such a variable is coupled with the rate of chromatin opening and closing, α_c, and the chromatin state (closing and opening) indicator, $k_c = \{0,1\}$. This "chromatin velocity" term, $\frac{dc(t)}{dt}$, is then integrated into the mRNA abundance dynamics, capturing the interplay between epigenomic changes and transcriptional regulation. The parameters α, k, β, and γ maintain their interpretations from the scVELO model, where the $k = \{0,1\}$ values represent, respectively, repression and induction. By considering the combinations of k and k_c, the MultiVelo framework enables the representation of multiple biologically feasible developmental cell states.

The analytical solutions for the chromatin accessibility, unspliced, and spliced mRNA abundances are provided in Equation 2.10. These solutions incorporate the initial states, c_0, $u_0^{(k)}$, and $s_0^{(k)}$, as well as the time points t_0 at which the cellular states change.

$$\begin{aligned}c(t) &= c_0 e^{-\alpha_c\tau} + k_c(1 - e^{-\alpha_c\tau})\\ u(t) &= u_0^{(k)} e^{-\beta\tau} - \frac{\alpha^{(k)}k_c}{\beta}(1 - e^{-\beta\tau}) + \frac{(k_c - c_0)\alpha^{(k)}}{\beta - \alpha_c}(e^{-\beta\tau} - e^{-\alpha_c\tau})\\ s(t) &= s_0^{(k)} e^{-\gamma\tau} + \frac{\alpha^{(k)}k_c}{\gamma}(1 - e^{-\gamma\tau}) + \frac{\beta}{\gamma - \beta}\left(\frac{\alpha^{(k)}k_c}{\beta} - u_0^{(k)} - \frac{(k_c - c_0)\alpha^{(k)}}{\beta - \alpha_c}\right)\\ &\quad (e^{-\gamma\tau} - e^{-\beta\tau}) + \frac{\beta}{\gamma - \alpha_c}\frac{(k_c - c_0)\alpha^{(k)}}{\beta - \alpha_c}(e^{-\gamma\tau} - e^{-\alpha_c\tau})\end{aligned} \tag{2.10}$$

Similar to the scVELO framework, the MultiVelo approach aims to estimate a trajectory $x_i = (c_i, u_i, s_i)$ that best approximates the observed multiomics data $x^{obs} = (c^{obs}, u^{obs}, s^{obs})$. This is achieved by minimizing the following negative log-likelihood function:

$$l(\theta) = \frac{3}{2}log(2\pi\sigma^2) + \frac{1}{2n\sigma^2}\sum_i^n ||x_i - x^{obs}||^2 \tag{2.11}$$

In this formulation, σ^2 represents the variance of the normally distributed residuals between the estimated trajectory and the observed measurements across the three dimensions: chromatin accessibility (c), unspliced mRNA (u), and spliced mRNA (s).

The MultiVelo framework employs the Expectation-Maximization (EM) algorithm to infer the parameters of the underlying ordinary differential equations and the cell-specific latent times t. However, the estimation of the latent times differs from the scVELO approach: instead of directly assigning the latent time to each cell, MultiVelo computes the (c, u, s) values of the ODEs solution at several uniformly distributed "anchor" time points. The cell is then assigned to the anchor with the shortest distance to the cell's observed multiomics measurements at each iteration of the EM algorithm.

Building upon the foundations of the scVELO model and incorporating additional multiomics measurements, the MultiVelo approach offers several key advantages over previous RNA-velocity methods. First and foremost, the original Velocyto and the scVELO models were limited to analyzing transcriptional kinetics alone, without considering the underlying epigenetic changes that play a crucial role in gene expression regulation. By bridging this gap, the MultiVelo approach offers enhanced insights into the diverse developmental states and lineage trajectories that emerge during cellular development (Fig. 2.13a,b). Furthermore, the MultiVelo framework introduces a novel latent time estimation strategy that offers several benefits as it can better accommodate the inherent heterogeneity and asynchrony within complex biological systems (Fig. 2.13a). Finally, the rigorous mathematical framework underlying the MultiVelo model, including the analytical solutions for the chromatin accessibility, unspliced, and spliced mRNA abundances, also provides a more robust means of inferring the underlying parameters. This enhanced parameterization and modeling capability enables the MultiVelo approach to analyze multiple state cellular configurations and identify the most appropriate model for each gene (Fig. 2.13c).

2.2.6 Notions on Graph Theory and Random Walks

To study the dynamic behavior of the biological system under investigation, scVEMO conducts a series of random walk simulations based on the computed cell-to-cell transition probability matrix P. The outcomes of these stochastic simulations are heavily influenced by the topological properties of the network encoded within the transition matrix P. Recognizing the importance of understanding the network structure for interpreting the random walk dynamics, the current section delves into a detailed analysis of the properties of this underlying network. This network-level understanding is crucial for interpreting the results of the random walk simulations.

This analysis follows the comprehensive framework outlined in the work by Fagnani and Como [57].

A transition probability matrix $P \in \mathbb{R}^{N \times N}$ is a non-negative square matrix satisfying the following properties:

$$0 \leq p_{ij} \leq 1 \text{ and } \sum_{i} p_{ij} = 1 \tag{2.12}$$

In other words, the matrix P represents a set of transition probabilities, where each entry p_{ij} describes the likelihood of transitioning from state i to state j, and the rows of P sum to 1, ensuring the probabilities are properly normalized.

This transition probability matrix P can be directly used to construct an associated weighted directed graph $\mathcal{G} = \{V, \varepsilon, W\}$. The set of nodes V corresponds to the states in P, while the edges ε and their associated weights W correspond to the entries of P, where the weight of each edge is the transition probability.

An important topological property of the graph $\mathcal{G}$ is its connectivity. Specifically, a graph is said to be strongly connected if, for any two nodes i and j in the graph, there exists a path (a sequence of directed edges) that connects i to j. This strong connectivity property ensures that all states within the system are accessible from one another through state transitions.

Additionally, the concept of periodicity is also crucial in the analysis of the graph structure. The period $per_{\mathcal{G}}(i)$ of a node i is defined as the greatest common divisor of the lengths of all circuits (closed paths) starting and ending at that node. If the period of all nodes in the graph is 1, the graph is considered to be aperiodic.

The topological properties of strong connectivity and aperiodicity, are fundamental for the existence and structure of the dominant eigenvalue and eigenvectors associated with the transition probability matrix P. This relationship is formally captured by Theorem 1.

Theorem 1 *Let $\mathcal{G} = \{V, \varepsilon, W\}$ a graph with strictly positive out-degree for every node i. Then, there exists a positive dominant eigenvalue λ_W with associated non-negative right eigenvector $x = \lambda_W^{-1} W x$ and left eigenvector $x = \lambda_W^{-1} W' y$.*

Specifically, given a non-negative and row-normalized matrix P, the dominant eigenvalue λ_w has specific properties. First, $\lambda_w = 1$ and, in the case of a strongly connected and aperiodic graph, λ_w is both algebraically and geometrically simple. Consequently, there is a unique dominant eigenvalue $\pi : \mathbf{1}'\pi = 1 \textit{ and } P'\pi = \pi$ and its corresponding eigenspace has dimension 1. Finally, all the other eigenvalues λ of P satisfy the $\lambda < 1$ relationship. The non-negative vector π is referred to as the "invariant probability distribution" of the graph $\mathcal{G}$ and its uniqueness is ensured when the graph is strongly-connected and aperiodic.

Generally, to understand the number of invariant probability distributions of a graph $\mathcal{G}$ one could also study its condensation graph $H_{\mathcal{G}}$. The condensation graph $H_{\mathcal{G}}$ is a directed aperiodic graph (DAG). It is constructed by collapsing the nodes of $\mathcal{G}$ into supernodes, where each supernode represents a strongly connected component of the original graph. The edges in $H_{\mathcal{G}}$ are defined such that there exists a link from one supernode to another if there is at least one edge in the original graph $\mathcal{G}$ that points from a node in the first connected component to a node in the second connected component (Fig. 2.14).

By analyzing the structure of the condensation graph $H_{\mathcal{G}}$, we can gain valuable insights into the number of distinct invariant probability distributions associated with the original graph $\mathcal{G}$. Theorem 2 establishes a direct link between the structure of the condensation graph $H_{\mathcal{G}}$ and the properties of the invariant probability distributions associated with the original graph $\mathcal{G}$. By analyzing the number of sinks in $H_{\mathcal{G}}$, i.e., supernodes with out-degree equal to zero, researchers can identify the number of distinct invariant probability distributions in $\mathcal{G}$ and understand their support.

Theorem 2 *Let $\mathcal{G} = \{V, \varepsilon, W\}$ be a graph. Then,*

- *Any convex combination of invariant probability distributions of $\mathcal{G}$ is and invariant probability distribution of $\mathcal{G}$.*

- *For every sink in $H_{\mathcal{G}}$ of $\mathcal{G}$ there exists an invariant probability distribution supported on the connected component corresponding to such a sink. Such invariant probability distributions are referred to as extremal.*
- *Every invariant probability distribution can be obtained as a convex combination of the extremal invariant probability distributions.*
- *If $s_{\mathcal{G}} = 1$, then $\mathcal{G}$ has a unique invariant probability distribution π whose support coincides with the connected component of such a sink.*

Stochastic complex systems are usually simulated using random walks. A discrete-time stochastic process $X(t),\ t = 0{,}1,\ldots$ with state space Ω is such that for any couple of states $(i, j) \in \Omega$:

$$\mathbb{P}\left(X_{t+1} = j | X(0) = i_0, X(1) = i_1, \ldots, X(t-1) = i_{t-1}, X(t) = i_t\right) = \mathbb{P}_{i_t,j} \tag{2.13}$$

Equation 2.13 states that the future state of the system $X(t+1)$ only depends on the current system state and it is independent from the past, a property known as "memoryless" or "Markov" property. Furthermore, the probability of moving to another state is described by the underlying transition probability matrix P. A process satisfying Equation 2.13 is known as discrete-time Markov chain (MC) with probability matrix $P \in [0{,}1]^{N\times N}$.

To get insights about the behavior of the MC one needs to specify both the underlying transition probability matrix P and an initial probability distribution $\hat{\pi}(0) \in [0{,}1]^N$, indicating the chain probabilities of starting in a specific state. Then, the trajectory follows:

$$\hat{\pi}'(t) = \hat{\pi}'(0)P,\ P(t) = P^t\ \forall t \geq 1 \tag{2.14}$$

In Equation 2.14, $\hat{\pi} \in [0{,}1]^N$ is the marginal probability distribution of $X(t)$ and it describes the probability that the random walk will end in a specific state at time t. When the chain is run infinitely long, $\hat{\pi}$ is known as "stationary distribution". Interestingly, when the underlying transition probability matrix is irreducible - hence the associated graph $\mathcal{G}$ is strongly connected and aperiodic - the MC stationary distribution aligns with the invariant probability distribution of the underlying transition matrix P, as stated in Theorem 3. Consequently, it is of interest to determine if such an invariant probability distribution exists, whether it is unique, and identify its support to ascertain the most likely endpoint of the Markov Chain.

Theorem 3 (Convergence in Probability) *Let P be an irreducible and aperiodic stochastic matrix and $\pi = P'\pi$ its normalized invariant probability distribution. Let $\hat{\pi}(t),\ t = 0{,}1,\ldots$ be the probability distribution vectors of a Markov Chain with transition probability P. Then,*

$$\lim_{t\to\infty} \hat{\pi}(t) = \pi \tag{2.15}$$

For any initial probability distribution $\pi(0)$.

Within the MC framework, states can be classified as either recurrent or transient. A state i is considered recurrent if, when the Markov chain starts from that state, it will return to i with probability 1. Importantly, when the Markov chain begins in a recurrent state, it will remain within the set of recurrent states indefinitely. In contrast, transient states are those from which the Markov chain will eventually depart and never return. While the recurrent states determine the stationary, long-term behavior of the Markov chain, the transient ones play a crucial role in influencing the short-term, transient dynamics of the cellular system.

Finally, the concept of absorption probability is central to both the Markov Chain simulations and CellRank. Consider a subset of states $\mathcal{S}$ within the overall state space Ω of the Markov chain. For any initial state $i \in \Omega$, the absorption probability in state $s \in \mathcal{S}$ is defined as:

$$H_{i,s} = \mathbb{P}_i(X(T_{\mathcal{S}}) = s) = \mathbb{P}_i(T_{\mathcal{S}} = T_s) \tag{2.16}$$

The absorption probability $H_{i,s}$ represents the likelihood that, starting from the initial state i, the Markov chain will first hit the state $s \in \mathcal{S}$ before reaching any other state in $\mathcal{S} \setminus s$. In Equation 2.16, T_s represents the hitting time of state s, which is the first time $t \geq 0$ that the Markov chain $X(t)$ reaches state s. Additionally, $T_{\mathcal{S}}$ denotes the hitting time of the subset $\mathcal{S}$, defined as the minimum of the hitting times of the individual states $s \in \mathcal{S}$:

$$T_{\mathcal{S}} := inf\{t \geq 0 : \ X(t) \in \mathcal{S}\} = \min_{s \in \mathcal{S}}\{T_s\} \tag{2.17}$$

2.2.7 Cell Trajectory Inference: the CellRank Framework

The CellRank framework [16] leverages single-cell RNA-sequencing data to perform trajectory inference, with the goal of automatically detecting the initial, intermediate, and terminal states of the underlying biological system. This algorithm, then, performs a soft probability assignment of cells to each of the terminal states, providing insights into the likelihood of a particular cell to develop into a specific differentiated cell type.

While RNA-velocity platforms [21, 48] are valuable tools for reconstructing developmental lineages, they also embody intrinsic limitations. Authors of CellRank recognized, for example, the high dependence of velocity computations on the presence of intron-rich sequences, the use of a single set of gene-specific parameters across all cells, and the limited interpretability of the velocity vectors themselves.

To address these challenges, CellRank combines RNA-velocity estimations with a similarity-based trajectory inference approach. By integrating the dynamic information from RNA-velocity with transcriptomic similarity and the topological constraints described by a k-nearest neighbor (KNN) graph, CellRank is able to model cell-state transitions via a Markov Chain. This approach allows CellRank to better cope with the inherent stochasticity in cellular differentiation processes, while providing a more robust and interpretable means of tracing lineage relationships.

CellRank assumes that the state transitions between cellular profiles are gradual, with each state being transcriptomically similar to the previous one. This reflects the continuous nature of cellular differentiation processes. Then, it also assumes that the set of cellular profiles spans the entire trajectory of state changes. This ensures that the framework can adequately capture the full spectrum of developmental progression. Finally, state transitions are modeled as a Markov Chain, which relies on the memoryless property.

The CellRank algorithm requires two key inputs: a cell-by-gene expression matrix $X \in \mathbb{R}^{N \times G}$, and a matrix $V \in \mathbb{R}^{N \times G}$ representing a vector field, such as the RNA-velocity estimates. It then operates in three distinct steps. First, it computes a cell-state transition probability matrix P to model the stochastic state transitions as a Markov Chain. Next, CellRank coarse-grains the transition matrix P into a set of initial, terminal, and intermediate macrostates, and computes the macrostate transition probabilities into a matrix P_c. It then assigns each cell a soft probability of belonging to each macrostate, captured in the membership matrix χ. Finally, CellRank leverages the computed macrostate transition probabilities to determine the fate probabilities of each cell towards the terminal macrostates and returns a fate matrix F that encapsulates the likelihood of each cell to develop into the various differentiated cell types.

At the core of the CellRank framework is the construction of a k-nearest neighbor (KNN) graph, which serves to limit the set of possible cellular transitions to only those between nearest neighbors. This is a crucial step, as it ensures that the state transitions modeled by CellRank are consistent with the inherent topological structure of the single-cell transcriptomic data. To construct the KNN graph, CellRank first projects the high-dimensional scRNA-seq data onto the first L principal components, in line with the dimensionality reduction techniques commonly employed in scRNA-seq data analysis. Next, for each cell i, CellRank computes the Euclidean distance to its K nearest neighbors. The resulting adjacency matrix A is then symmetrized, such that neighboring cells i and j are only considered as such if i is a neighbor of j and vice versa. The symmetrization process however implies that each cell i now has a variable number of neighbors K_i. Importantly, this means that $K_i \geq K$, where K is the original number of nearest neighbors specified in the KNN graph construction. The associated KNN graph is now undirected and symmetric.

The CellRank framework leverages the information contained in the field matrix V to direct the graph and compute cell-to-cell transition probabilities. The core idea is to assign higher transition probabilities to those neighboring cells whose direction of transcriptomic change, as encoded in the displacement vector $s_{ik} = x_k - x_i$, best aligns with the direction of the velocity vector v_i associated with the reference cell i. This alignment between the velocity vector and the state-change vectors captures the degree to which the RNA-velocity estimates can predict the observed transcriptomic changes between the neighboring cellular profiles. Specifically, for each cell i with gene expression profile $x_i \in \mathbb{R}^G$ and velocity vector $v_i \in \mathbb{R}^G$, CellRank computes the Pearson correlation $c_i \in [0,1]^{K_i}$ between v_i and the set of state-change vectors $\{s_{ik}\}$ for all K_i neighboring cells (Fig. 2.15). These cell-neighbor correlations provide a measure of how well the velocity vector can predict the transcriptomic changes in the neighboring state profiles.

To transform these correlations into transition probabilities $p_i \in [0,1]^{K_i}$, CellRank employs the softmax function, as shown in Equation 2.18. This operation ensures that the transition probabilities are non-negative and sum to unity.

$$p_{ik} = \frac{e^{\sigma c_{ik}}}{\sum_l e^{\sigma c_{il}}} \tag{2.18}$$

Here, σ is a scalar controlling how centered the distribution is around the state-change transition with maximum correlation and can be computed using the heuristic in Equation 2.19. The probabilities are collected within a transition probability matrix $P \in [0,1]^{N\times N}$.

$$\sigma = \frac{1}{median\left(\{|c_{ik}|\ \forall(i,k)\}\right)} \tag{2.19}$$

The Cellrank framework leverages the Generalised Perron Cluster Cluster Analysis [52] to coarse-grain the obtained transition probability matrix P and project it onto a lower dimensional space. Such a step enables CellRank to characterize the complex cellular dynamics and lineage relationships at a more interpretable macrostate level. By uncovering initial and terminal macrostates, CellRank enables the identification of the starting and ending cellular types and the developmental lineages within an evolving system (Fig. 2.8a).

It is important to distinguish between the terms "cluster" and "macrostate" as they are employed in the subsequent discussion. The GPCCA method identifies coarse-grained, metastable cellular states, referred to as "macrostates" within the context of this thesis. This term is used to emphasize that these macrostates represent distinct, higher-level cellular phenotypes and configurations rather than groupings of individual cells. Conversely, the term "cluster" is typically associated with a hard assignment of data points to distinct groups. In the context of this thesis project, we refer to clusters to indicate the result from a community detection algorithm, such as the Louvain [58] or the Leiden [59] methods, which identify cells that are densely connected within the KNN. Notably, the macrostates identified by the GPCCA analysis can be composed of cells belonging to different clusters.

Details about the GPCCA method and its implementation in CellRank can be found in Section 2.2.8.

The coarse-grained transition matrix P_c obtained through the GPCCA analysis is then used to automatically infer the initial and terminal states within the system. This is accomplished by leveraging the stability index (SI), which is defined as the self-transition probability $P_{c_{mm}}$ for each macrostate m. Notably, the terminal macrostates are first identified as those having a very high stability index ($SI \geq 0.96$), indicating that they are less likely to transition to other macrostates. Such a high degree of self-transition probability suggests that they represent stable, endpoint cellular phenotypes. Conversely, the CellRank framework determines the initial macrostates by computing the invariant probability distribution π_c of the coarse-grained transition matrix P_c. This invariant distribution π_c represents the long-term, stationary probabilities of the system occupying each macrostate. The initial macrostates are characterized by having low values in the π_c vector. This is because the initial states are less likely to be visited multiple times in the Markov chain dynamics, as the system tends to transition towards the terminal ones. All macrostates which are not classified as either initial or terminal are known as intermediate.

Building upon the GPCCA-derived membership matrix χ and the identified terminal macrostates, the CellRank framework computes the likelihood of each individual cell transitioning towards each of the terminal macrostates. For each terminal macrostate $t \in \{1, \ldots, n_t\}$, CellRank first identifies a set of f cells that are strongly assigned to that macrostate according to χ. These cells are considered to be the representatives of the terminal macrostate t and are grouped into the terminal index set R_t. All remaining cells are then collected into a disjoint transient index set T. For each cell in T, CellRank computes a cell fate probability vector $f_i \in \mathbb{R}^{n_t}$, where each element $f_{i,t}$ represents the probability of that cell transitioning towards the corresponding terminal macrostate t. Finally, these cell fate probabilities f_i vectors form a valid probability distribution, satisfying the partition of unity and non-negativity properties. The complete set of cell fate probabilities is accumulated into a fate matrix $F \in \mathbb{R}^{N \times n_t}$.

The computation of the cell fate probabilities within the CellRank framework relies on the concept of absorption probabilities in Markov Chains. As previously discussed, the Markov Chain states can be categorized as either recurrent or transient states. Upon a suitable permutation of the cell barcodes, the transition matrix P can be represented in the following block form:

$$\begin{pmatrix} \tilde{P} & 0 \\ S & Q \end{pmatrix} \tag{2.20}$$

In this representation, the submatrix $\tilde{P}$ represents the transition probabilities within the set of recurrent states, hence cells in $\{R_t\}$, while Q represents the transition probabilities within the set of transient states, i.e., cells in T. The submatrix S captures the transitions from the transient states to the recurrent states, while it is not possible to move from the recurrent to the transient ones.

To compute fate probabilities towards terminal states, CellRank computes absorption probabilities towards cells in $\{R_t\}$. CellRank converts the terminal index set into absorption states by removing out-going edges from cells belonging to $\{R_t\}$ in the graph $\mathcal{G}$ underlying P. Absorption probabilities are then computed as follows:

$$A = (I - Q)^{-1} S \tag{2.21}$$

To retrieve probabilities towards a specific R_t, CellRank sums the absorption probabilities towards the representing individual cells.

The key strength of the CellRank framework is its ability to provide intuitive visualizations that bring complex cellular dynamics and lineage relationships to life. One particularly powerful visualization is the overlay of the computed cell fate probabilities and terminal macrostates onto the UMAP embedding of the single-cell data. The visualization presents each cell as a point, with the color of the point encoding the cell's fate probability vector f_i, revealing the developmental trajectories and branching points within the cellular landscape. Regions of the UMAP corresponding to high probability for a given terminal macrostate are clearly delineated, providing an intuitive understanding of the lineage relationships and fate decisions governing the system. Overlaid on top of this fate probability map, CellRank also highlights the location of the terminal macrostates identified through the GPCCA analysis. These terminal states serve as the attractors or end-points of the developmental trajectories, as illustrated in Fig. 2.8a.

2.2.8 The Generalised Perron Cluster Cluster Analysis and its Application in CellRank

The Generalised Perron Cluster Cluster Analysis (GPCCA) [52] is a spectral clustering method that leverages the eigenvectors of a row-stochastic matrix to perform dimensionality reduction. In the context of this work, the input to GPCCA is the row-stochastic CellRank transition probability matrix P, which encodes the similarities and transition probabilities between the individual cells. The algorithm identifies n_s macrostates, which represent the coarse-grained, metastable cellular states or phenotypes. GPCCA delivers a fuzzy clustering of the data, where each cell is assigned a probability of belonging to each of the identified macrostates.

The fuzzy spectral clustering problem consists of separating N data objects $o_1, \dots, o_N$ into different n_s macrostates $M_1, \dots, M_{n_s}$ according to their pairwise similarities s_{ij}. This problem can be represented by a matrix $\chi \in \mathbb{R}^{N \times n_s}$, also known as membership matrix, satisfying:

$$0 \leq \chi_{ij} \leq 1, \ \sum_{j=1}^{n_s} \chi_{ij} = 1 \ \forall i = 1, \dots, N \tag{2.22}$$

In χ, each value χ_{ij} can be interpreted as the membership of the i-th data object belonging to the j-th macrostate.

The underlying assumption in GPCCA is that any stochastic matrix can be viewed as a small perturbation of an uncoupled Markov Chain. In this latter case, there exists a unique linear transformation of the matrix eigenvectors that can be used to represent the spectral clustering. Specifically, in the uncoupled Markov Chain scenario, the underlying transition probability matrix is block diagonal. This implies that the associated graph is composed by multiple disconnected components. Intuitively, in the uncoupled MC scenario, the macrostates identified by GPCCA would correspond directly to the disconnected components in the graph. Furthermore, the data objects (in this case, the individual cells) would be automatically assigned to the graph component to which they belong. (Fig. 2.16).

While the real-world transition probability matrix P derived from the single-cell data is unlikely to be precisely an uncoupled Markov chain (Equation 2.20), the GPCCA method leverages this underlying assumption to find a suitable fuzzy clustering. Even in the case of a general stochastic matrix, GPCCA is able to identify a membership matrix χ that satisfies an optimality criterion, providing a meaningful representation of the coarse-grained, metastable cellular states.

The identification of macrostates within the GPCCA framework is typically performed by projecting the transition probability matrix P into its coarse-grained version $P_c \in \mathbb{R}^{n_s \times n_s}$. This projection is achieved by leveraging an invariant subspace of P. However, a significant challenge arises in the case of non-reversible matrices, such as the CellRank transition probability matrix P, where there exist complex-valued eigenvalues. In such situations, since the traditional eigenvector decomposition approach is not directly applicable, the GPCCA method utilizes the Schur decomposition [60]:

$$P = QRQ^T \tag{2.23}$$

Here Q is a unitary matrix whose columns are the Schur vectors. Instead, R is an upper quasi-triangular matrix whose diagonal elements correspond to the eigenvalues of P - real eigenvalues lie on the 1×1 blocks while conjugate complex pairs on the 2×2 blocks. The key insight is that the real Schur vectors and the Schur vectors associated to complex conjugate eigenvalues - only when such vectors are coupled - span the same invariant subspace as the eigenvectors of the transition probability matrix P [52].

The matrix $\tilde{Q} \in \mathbb{R}^{N \times n_s}$ is obtained by selecting n_s columns from Q. Then, the membership and the projected transition matrices are obtained as in Equation 2.24.

$$\begin{aligned} \chi &= \tilde{Q}A \\ P_c &= (\chi^T D \chi)^{-1}(\chi^T D P \chi) \end{aligned} \tag{2.24}$$

Where $A \in \mathbb{R}^{n_s \times n_s}$ is a non-singular matrix, and D is a diagonal matrix such that $\tilde{Q}^T D \tilde{Q} = I$. Typically, these diagonal entries are chosen according to some distribution of the cellular states of interest, which in the case of CellRank is the uniform distribution. The goal of the GPCCA analysis is now to find a transformation matrix A that satisfies the partition of unity and positivity conditions of the membership matrix χ, as expressed in Equation 2.22. As previously mentioned, there exists a set of feasible transformation matrices $\mathcal{F}_A$ that satisfy these conditions. To identify the optimal transformation matrix A, the GPCCA approach employed in CellRank resolves the following optimization problem:

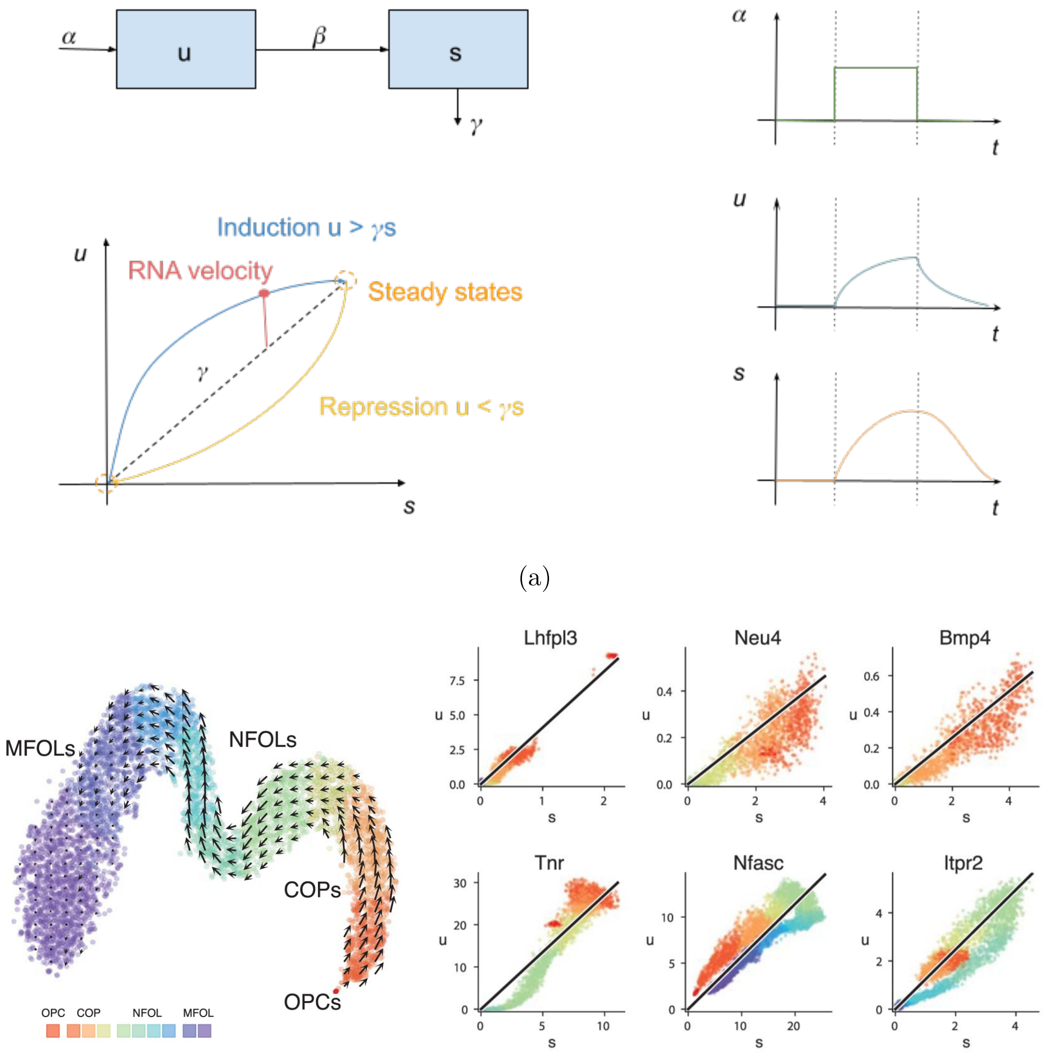
$$\begin{aligned} \min \quad & f_{n_s}(A) = n_s - trace(\tilde{D}^{-1} \chi^T D \chi) \\ s.t. \quad & \chi_{ij} \geq 0 \ \forall i \in \{1, \ldots, N\} \ \forall j \in \{1, \ldots, n_s\}, \\ & \sum_{j=1}^{n_s} \chi_{ij} = 1 \ \forall i \in \{1, \ldots, N\} \end{aligned} \tag{2.25}$$

Where

$$\tilde{D}^{-1} = diag\left(\frac{1}{\sum_j (\chi^T D \chi)_{1j}}, \ldots, \frac{1}{\sum_j (\chi^T D \chi)_{n_s j}}\right)$$

The GPCCA optimization problem aims to make the identified macrostates as crisp or distinct as possible. In other words, the goal is to find a transformation matrix A that leads to a membership matrix χ that is as similar as possible to an indicator matrix. An indicator matrix is a binary matrix where each row represents a data point (in this case, a cell) and each column represents a macrostate. Each entry in the indicator matrix is either 0 or 1, indicating whether the corresponding data point belongs to the respective macrostate or not. By seeking a membership matrix χ that is as similar to an indicator matrix as possible, the GPCCA optimization process attempts to achieve an unambiguous assignment of cells to the identified macrostates and provide a more interpretable representation of the distinct cellular states.

It is important to note that the resulting coarse-grained transition matrix $P_c \in \mathbb{R}^{n_s \times n_s}$ obtained through the GPCCA optimization process does not represent a Markov process anymore. In other words, P_c is not a transition probability matrix itself. Unlike the original transition probability matrix P, the coarse-grained matrix P_c may contain negative values. When the optimization process is unable to achieve a clear separation of the macrostates, the resulting P_c matrix will exhibit such negative values.



(a)

(b)

Figure 2.11: Velocyto RNA-velocity representation. **a.** (top-left) Schematic representation of the compartmental model underlying the Velocyto rate equations. (bottom-left) The gene-specific phase portrait according to the Velocyto model: the dashed line represents the degradation rate γ, which separates the regions of gene induction (above γ) and gene repression (below γ); each cell, depicted as a red data point, can be plotted on this phase portrait, and the distance between the cell's position and the γ line corresponds to the RNA-velocity. (right) Visualization of the spliced and unspliced mRNA abundances as a function of the transcription rate α: this plot illustrates how changes in the transcriptional rate are reflected in the differential dynamics of these two RNA molecules. **b.** RNA-velocity vectors projection on the t-SNE embedding of the hindbrain of adolescent (P20) mice (left) and gene phase portraits (right). Adapted from *La Manno, G., Soldatov, R., Zeisel, A. et al. RNA-velocity of single cells. Nature 560, 494-498 (2018). https://doi.org/10.1038/s41586-018-0414-6.*

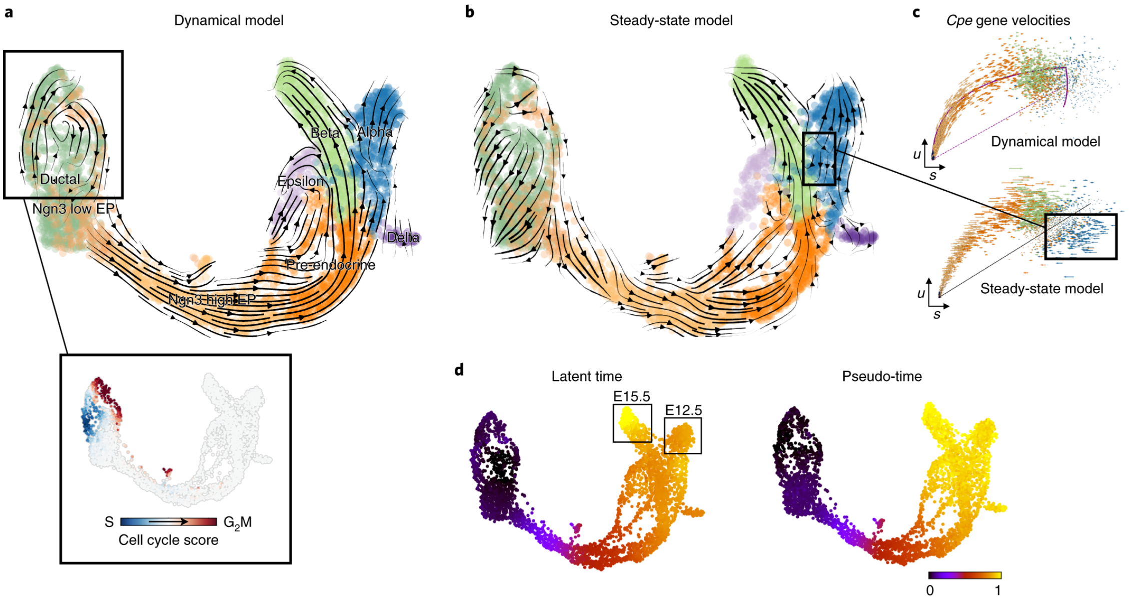


Figure 2.12: RNA-velocities derived for the pancreatic endocrinogenesis using the scVELO dynamical model (**a.**) and the Velocyto steady-state approach (**b.**). **c.** Phase portrait comparison for the *Cpe* gene recovered from the models: the steady state model incorrectly assigns α-*cells* to the repression phase. **d.** scVELO's latent time can better identify the cell's positioning in the biological development compared to pseudotime. Adapted from *Bergen, V., Lange, M., Peidli, S. et al. Generalizing RNA-velocity to transient cell states through dynamical modeling. Nat Biotechnol 38, 1408-1414 (2020).* https://doi.org/10.1038/s41587-020-0591-3

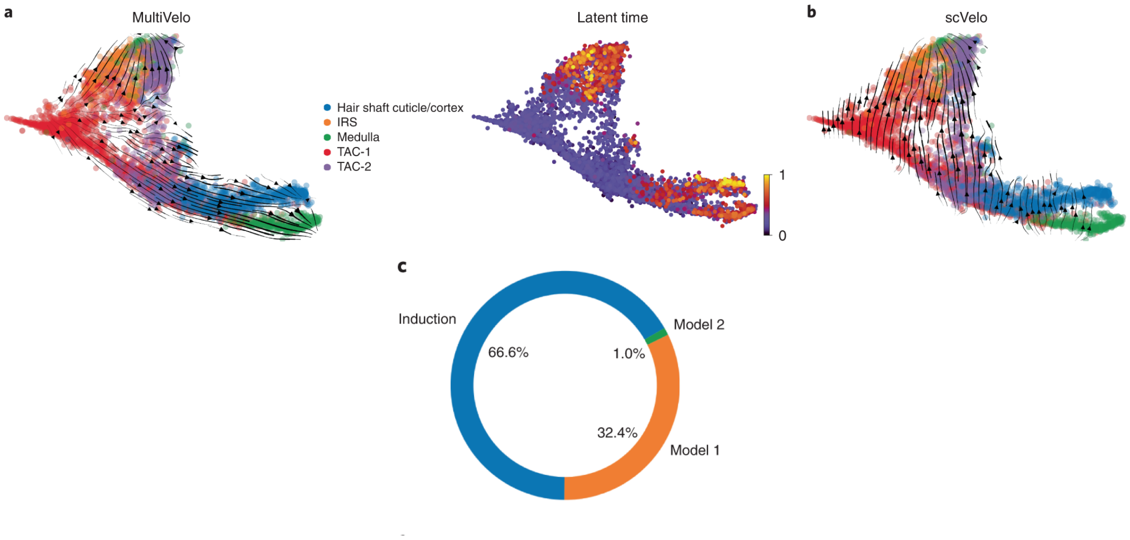


Figure 2.13: Comparison of the scVELO and MultiVelo frameworks over the mouse skin dataset. **a.** Velocity stream plot and latent time estimation using MultiVelo predict two developmental lineages. **b.** Velocity stream plot using scVELO: transcriptomic data only cannot correctly recover developmental lineages. **c.** Relative proportion of each type of kinetics across all fit genes. Adapted from *Li, Virgilio, Collins, et al., Multi-omic single-cell velocity models epigenome-transcriptome interactions and improves cell fate prediction, Nature Biotechnology, 41:387-398, 2023, doi:10.1038/s41587-022-01476-y.*.

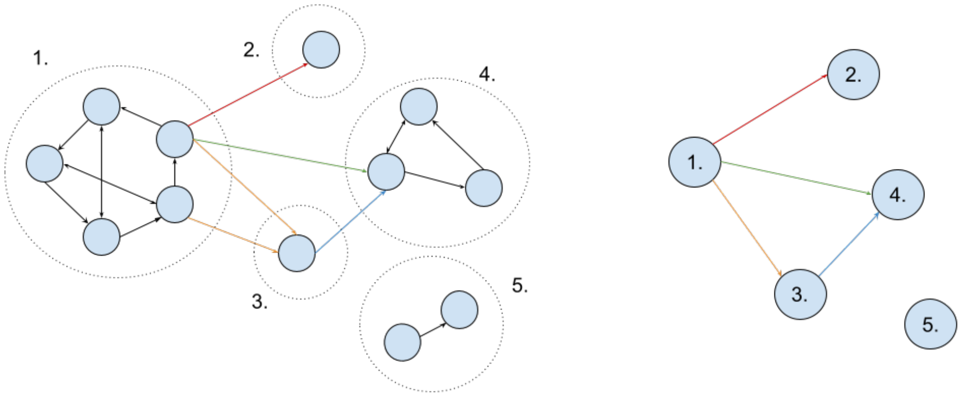


Figure 2.14: Graph (left) and its condensation graph (right). Dashed circles highlight connected components, while numbers are used to label each connected component and the respective supernode. Edges are colored according to the supernodes they are connecting in both the graph and its condensation version.

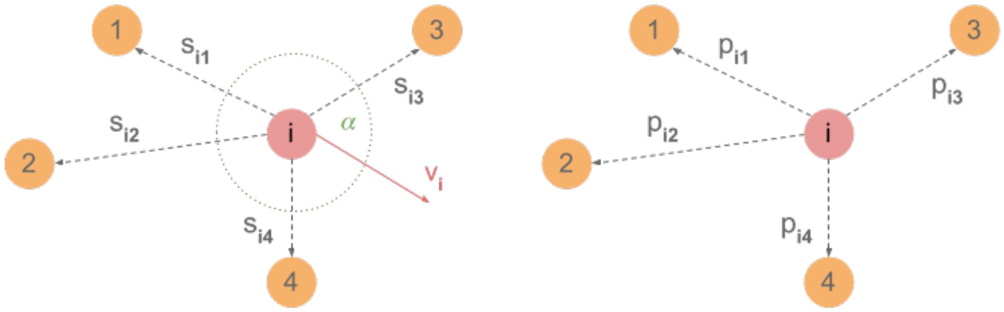


Figure 2.15: Cellrank probabilities computations. A reference cell i is highlighted with its velocity vector v_i and neighbors. The displacement vector s_{ij} represents the difference in gene expression for neighboring cells. Probabilities are computed considering the angle α between each displacement vector and the velocity one.

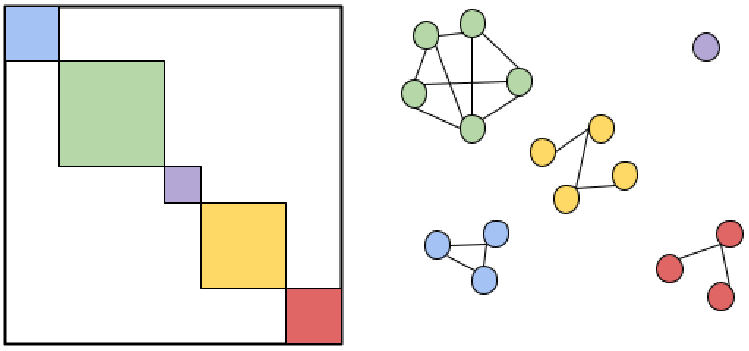


Figure 2.16: Transition matrix representation of an uncoupled Markov Chain - up to a perturbation of the order of the objects. The block-diagonal matrix (left) and the representation of the associated disconnected graph (right) match the block color with the respective component.

Chapter 3

Methods

This chapter presents the various models and computational approaches employed in our analysis of the 10X Genomics dataset on the embryonic mouse brain.

The baseline model is based on scRNA-seq data alone. This serves as a reference to assess the performance of the multimodal approaches. Next, we construct a model based solely on scATAC-seq data. This allows us to investigate whether it is possible to recover meaningful lineages only from the highly sparse epigenomic data. Finally, we build two multiomics models that leverage both epigenomics and transcriptomics data, albeit with different strategies. The Multiomics+scVELO model correlates both the peak count matrix and the gene expression one with the RNA velocities computed using the scVELO framework. Instead, the Multiomics+MultiVelo model employs a gene activity matrix and computes the velocities using the MultiVelo framework.

The chapter first provides a brief overview of the 10X Genomics dataset, describing the key contents of the files that will be leveraged throughout our analysis. Next, it delves into the preprocessing pipelines used to perform quality control and compute the RNA-velocity estimates for each of the models under investigation. Building upon the preprocessed data, it then introduces the methodologies used to compute the transition matrices used as input for the CellRank model.

Finally, this chapter provides an overview of the model evaluation metrics employed to assess the performance and robustness of the various approaches.

3.1 Dataset

The 10X Fresh Embryonic E18 Mouse Brain (5k) dataset [24] consists of paired ATAC and GEX cellular profiles. This single-cell sequencing dataset comprises cells sampled from the fresh cortex, hippocampus, and ventricular zone of the embryonic mouse brain at day 18. Cells are extracted from a combination of fresh, cryopreserved, and flash-frozen tissue samples. Nuclei isolation is performed using the Embryonic Mouse Brain pipeline for the Single Cell Multiome ATAC+GEX [44] and the ATAC and RNA libraries preparation step is carried out using the 10X Genomics Chromium Single Cell Multiome ATAC+GEX protocol [38], as described in previous sections.

The processed libraries are then analyzed using CellRanger-ARC 2.0.0 [46]. The resulting dataset contains 4878 different cells, around ten thousand linked genes, and 56,000 linked peaks.

To leverage this multimodal dataset, we employ the feature-barcode matrix provided in HDF5 format. This matrix contains the count matrix entries, the features identifiers, and cell barcodes, as well as the feature type to distinguish genes from peaks. Additionally, we use the peak annotation TSV file, which provides a map from peaks to genes, with the peaks categorized as promoter (±1000 base pairs from any TSS), distal (within 200kb from the closest TSS but not in the promoter region), or intragenic.

In addition, the spliced and unspliced mRNA counts required for RNA velocity estimations are also provided by the MultiVelo framework [21]. These velocity-related counts are computed from the dataset using the Velocyto Command Line Interface (CLI) [48].

Furthermore, MultiVelo authors also provide cell type annotations, which identify 12 distinct cell clusters within the embryonic mouse brain sample. A visualization of the clusters is provided in Fig. 3.1.

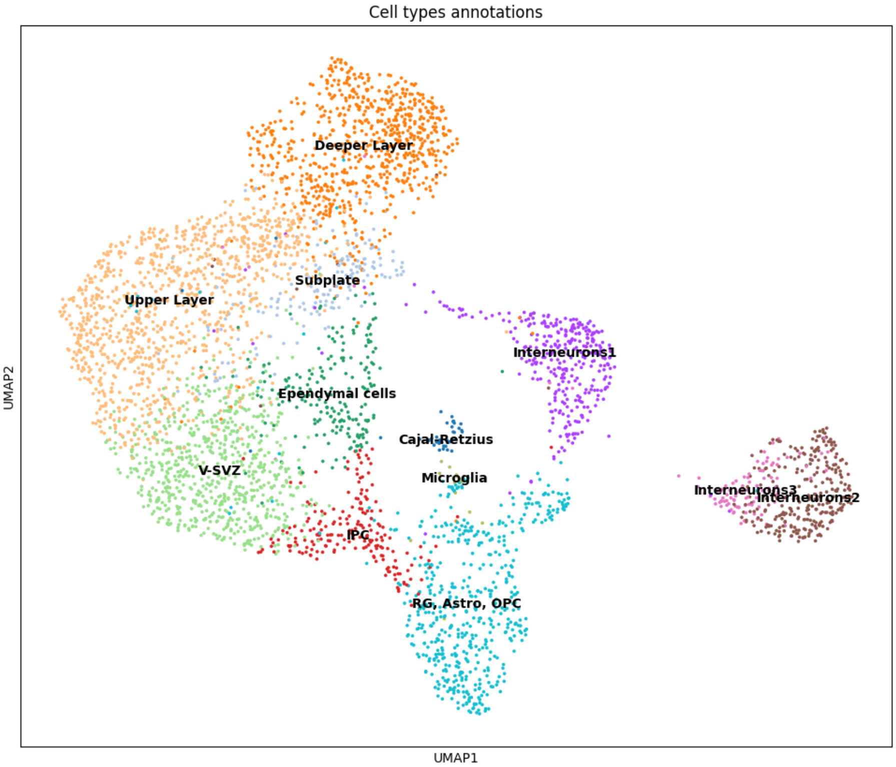


Figure 3.1: UMAP visualisation of the 10X Fresh Embryonic E18 Mouse Brain. Cells are clustered according to cell labels provided by MultiVelo.

Clusters have been validated by investigating the expression of known marker genes [61–64] for each cluster as listed in Table 3.1. Gene expressions are plotted in Appendix B.

Table 3.1: Marker genes per cluster obtained from literature

Cajal-Retzius	*Reln*
Interneurons	*Dlx1, Gad1, Lhx6*
Microglia	*Trem2*
OPCs	*Olig2, Pdgfra*
Astrocytes	*Aldoc, Slc1a3*
Radial Glia	*Vim*
IPCs	*Eomes*
V-SVZ	*Sema3c*
Deeper Layer	*Fezf2, Rorb*
Upper Layer	*Satb2, Inbha*

To provide the necessary context about the embryonic mouse brain development, we present results on the complex process of mouse brain development [61–64]. This literature review will later help us determine the expected terminal states and differentiated cell types that our models should be able to recover. The process of embryonic development begins with the gastrulation process (around E7), where the three primary germ layers - ectoderm, endoderm, and mesoderm - are established [61,64]. These germ layers serve as the building blocks for the development of specialized tissues and organs, including the brain and nervous system. The ectoderm - the outermost layer - gives rise to the epidermis, nervous system (both brain and spinal cord), sensory organs, and other derivatives. The endoderm forms the digestive tract and associated organs, the respiratory system, and parts of the urinary and reproductive systems. The mesoderm or middle layer germ, on the other hand, contributes to the musculoskeletal system, cardiovascular tissues, and urinary system.

During the subsequent neurulation stage (between E8 and E10), the formation of the neural crest and neural tube occurs [61,64]. The neural crest is a temporary structure that forms between the neural tube and the ectoderm, giving rise to multipotent cells that differentiate into diverse tissues, such as the meninges, sympathetic and parasympathetic nervous systems, and some non-neuronal brain cells like Schwann cells. The neural tube, formed from the ectoderm, serves as the precursor of the central nervous system, hence the brain and the spinal cord. Radial glia cells (RG) emerge from the neural tube as proliferating cells and act as precursors for all neural cell types.

As development progresses, around E14, the radial glia cells lose their proliferative capacity and transform, giving rise to glial lineage [64]. The glial cells (GCs) are non-neuronal cells that provide support and insulation for neurons, as well as contribute to the functioning of the nervous system. The glial cell population further differentiates into astrocytes (cells that provide structural support to neurons and contribute to processes such as repair of the neural tissue and synaptic functioning) and oligodendrocytes (OPCs, cells involved in the electrical signaling processes).

In addition, the radial glia cells also serve as precursor cells for neurons. The neurogenesis process happens between E10 and E18 [64]. RG cells proliferate into neuronal intermediate progenitor cells (nIPCs) [62], which subsequently transition and form intermediate structures such as the ventricular zone, the subventricular zone, and the subplate [61, 64]. The subplate is a transient layer of cells that acts as an intermediary zone, involved in early neuronal connections, neuronal maturation, and cortical circuit formation. The ventricular zone (VZ) is a germinal zone considered a primary source of neurons during neurogenesis, the process of creating neurons. The subventricular zone (SVZ) contains progenitor cells derived from radial glia in the VZ. Even after the second postnatal week, RG cells in the V-SVZ continue to generate new granule cells, albeit at a reduced rate [62].

Finally, the neurons divide and form the cerebral cortex [63]. The cerebral cortex consists of six layers, which can be categorized into three groups: Layer I, Upper Layers, and Deeper Layers. Layer I primarily consists of Cajal-Retzius cells and represents the first cortical layer to form. Layers V-VI (Deeper Layers) are generated before Layers II-IV (Upper Layers) due to the inside-out pattern of cortical development. Neurons that are born earliest occupy the deepest layers, while later-born neurons migrate past them and occupy more superficial layers [63]

3.2 Data Preprocessing

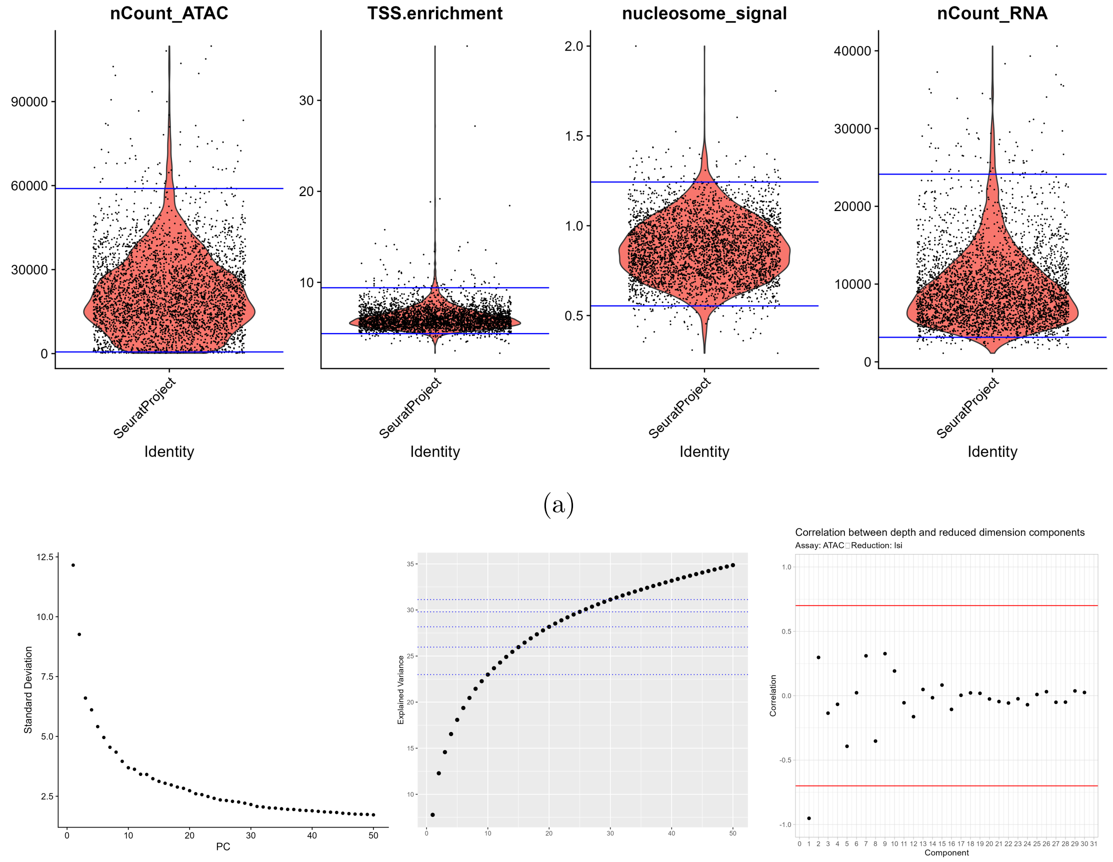
Data preprocessing is performed using Seurat 5.0.0 [14] Signac 1.12.0 [13] and Scanpy 1.9.6 [15]. The SeuratDisk 0.0.0.9021 package has been employed to perform conversions between the SeuratObject and the corresponding H5 file.

We compute the standard quality control metrics on the dataset and perform filtering as recommended in standard single-cell sequencing pipelines [13–15]. For the scRNA-seq data, we keep genes with shared counts (i.e., counts in both the spliced and unspliced layers) greater than 10, while for the scATAC-seq data, we retain peaks that are found in at least 10 cells. This selective filtering allows us to focus our analysis on the most informative features within the dataset.

Next, we integrate the cell annotations and remove the "Cajal-Retzius", "Microglia", and the three interneurons clusters. This allows us to concentrate our analysis on the developmental cells, which are the primary focus of our research.

We compute additional standard quality control metrics, as illustrated in Fig. 3.2a. This includes examining the gene and peak counts per cell to identify and remove any potential duplicates or dead cells. The nucleosome signal and TSS enrichment score are employed to filter the epigenomic data. The nucleosome signal provides insights into the degree of chromatin accessibility, with a high nucleosome signal potentially indicating technical artifacts or low cell lysis. Signac computes the ratio of mononucleosome (147-291 base pairs) and nucleosome-free (<147 base pairs) fragments to identify the nucleosome signal for each barcode. The TSS enrichment score, on the other hand, reflects the concentration of ATAC-seq signal around transcription start sites, which is an indicator of successful transcription. To compute the TSS enrichment, the reads distribution around a TSS (±2000 base pairs) is collected and normalized by taking the average read depth in the 100 base pairs at each of the end flanks of the distribution (for a total of 200 base pairs of averaged data) and then calculating the fold change at each position over that average read depth.

We filter barcodes within the 2nd and 98th percentiles for each metric.



(a)

(b)

(c)

Figure 3.2: Data preprocessing **a.** Quality Metrics computed for the dataset. **b.** PCA explained variance and elbow plot for scRNA-seq data. **c.** SVD dimensions correlation with sequencing depth.

After the initial quality control and filtering steps, we normalize the scRNA-seq data. This process involves log-transforming the data, as shown in Equation 3.1, followed by scaling and it helps remove any potential batch effects present in the dataset. Next, we extract the 2,000 most variable features from the normalized scRNA-seq data. This feature selection step is crucial for capturing the most informative and biologically relevant genes within the dataset. To reduce the dimensionality of the scRNA-seq data, we perform Principal Component Analysis (PCA). As depicted in Fig. 3.2b, the elbow is reached between 10 and 30 principal components (PCs). Within this range, the explained variance reaches approximately 33%, which is considered sufficient given the highly sparse nature of the scRNA-seq data.

$$log\left(\frac{X_{ij}}{\sum_i X_{ij}} * median(\sum_i X_{ij}) + 1\right) \tag{3.1}$$

To handle the highly dimensional and variable nature of the scATAC-seq data, instead, we employ the Latent Semantic Indexing (LSI) technique, which is a natural language processing (NLP) method typically used to identify patterns and relationships between terms and concepts in unstructured text collections. The LSI approach first applies the Term-Frequency-Inverse-Document-Frequency (TF-IDF) weighting scheme to the peak-count matrix. The term frequency statistic, $tf(p_i, c_j)$, represents the relative frequency of the i-th peak in the j-th cell. The inverse document frequency, on the other hand, measures how much information the peak provides in terms of its rarity across the entire collection of barcodes, C. The rationale behind using TF-IDF is that a peak has higher significance for a cell when its term frequency is high, and its inverse document frequency is low, meaning the peak is not commonly found across the whole collection. By applying this weighting scheme, TF-IDF aims to identify peaks that are specific to each cell, while assigning lower weights to common peaks in the collection, as shown in Equation 3.2.

$$\begin{aligned} tf(p_i, c_j) &= \frac{counts_{ij}}{\sum_i counts_{ij}} \\ idf(p_i, C) &= \frac{\sum_j counts_{ij}}{N} \\ tfidf(p_i, c_j, C) &= log(1 + tf(p_i, c_j) * idf(p_i, C) * 10^4) \end{aligned} \tag{3.2}$$

Following data transformation, LSI employs Singular Value Decomposition (SVD) to determine patterns in the relationship between peaks and cells and perform dimensionality reduction. As illustrated in Fig. 3.2c, we examine the correlation between the sequencing depth and the reduced components. Based on this analysis, we decide to discard the first dimension, as it appears to capture sequencing depth-related variations rather than biologically relevant information [13].

To compute the RNA-velocities for our multimodal analysis, we leveraged scVELO 0.3.1 [20] and integrated it with our preprocessing workflows in Scanpy and Seurat.

For the scRNA-seq data, we first use Scanpy to compute the K-Nearest Neighbor (KNN) graph and the UMAP embedding of the cells. We then normalize the spliced and unspliced RNA count matrices and compute the first-order moments -i.e. mean gene expressions - for each cell across its neighborhood. These moments are then used by scVELO to recover the full splicing kinetics for the genes and infer the RNA velocities in all models except the MultiVelo-based one.

For the Multiomics+scVELO and the scATAC-seq models we use, respectively, Seurat `FindNeighbors` and `FindMultiModalNeighbors` functions to compute the KNN graph. We then leverage Seurat UMAP embedding implementation.

The Multiomics+MultiVelo model employs the preprocessing pipeline developed by MultiVelo 0.1.3 [21]. We use Seurat to compute the multimodal KNN and UMAP embedding, and then exploit scVELO to get the first-order moments over the RNA data. Additionally, we use the MultiVelo TFIDF procedure to transform the peak count matrix. The framework aggregates the promoter and enhancer peaks to genes based on the peak annotation TSV file from the CellRanger ARC pipeline and constructs a gene activity matrix that summarizes chromatin accessibility information for the genes of interest. The `aggregate_peaks_10x` function collects distal putative enhancer peaks with a correlation greater than or equal to 0.5 with promoter accessibility or gene expression either annotated to the same gene or within 10kb of that gene. Then the function annotates these distal putative enhancers to the promoter peaks for the corresponding genes. Finally, we normalize and smooth this new gene activity matrix representation and computed RNA-velocities using the MultiVelo framework.

A further distinction between the multiomics models, beyond the employed RNA-velocity framework, lies in the construction of the nearest neighbor graph. In the case of the Multiomics+scVELO approach, we utilize the shared nearest neighbor (WSNN) graph computed by the Seurat's `FindMultiModalNeighbors` function, while the Multiomics+MultiVelo model relies on the standard KNN graph, as suggested in [21]. Unlike KNN, SNN considers the effect of shared nearest neighbors around each node, capturing local density and connectivity.

To ensure the reliability and robustness of our velocity estimation models, we explore various preprocessing parameter configurations. This allows us to study the sensitivity of the scVELO framework to different hyperparameter settings. Specifically, we investigate the impact of the size of the local neighborhood, which determines the number of neighbors in the KNN graph. We refer to this parameter as K in the following sections. Additionally, we explore the effects of including different numbers of principal components and latent semantic indexing dimensions in the dimensionality reduction steps. We denote these parameters as PC and LSI, respectively. The specific parameter values explored in our robustness study are reported in Table 3.2.

Table 3.2: Preprocessing parameters

K	*10,20,30,50,60,80*
PC	*10,15,20,25,30*
LSI	*10,15,20,25,30*

3.3 Transition Matrix Construction and Developmental Lineages Identification

After the preprocessing steps, each model is associated with a feature count matrix or a gene activity matrix (in the case of the Multiomics+MultiVelo model) and an RNA-velocity matrix, computed either via scVELO or MultiVelo. It is important to note that the scATAC-seq model alone cannot provide a velocity matrix, and therefore needs to be coupled with a scRNA-seq dataset to obtain the necessary RNA-velocities. The pipeline proceeds with the computation of a transition matrix for cell-state transitions and the identification of the macrostates and cell fate probabilities using CellRank 2.0.2 [16].

The scRNA-seq transition matrix is been computed through the standard CellRank `VelocityKernel` class. Such a class performs computations as shown in Fig. 2.15.

The scATAC-seq model only consists of a transformed matrix $X_P \in \mathbb{R}^{N \times n_{peaks}}$ and it requires RNA-velocities from an additional scRNA-seq embedding, specifically the $K = 30, PC = 30, LSI = 10$ model. The peak count matrix is first subset, keeping only promoter peaks to highly variable genes. Then, columns corresponding to peaks that are promoters to multiple genes are repeated within the matrix, as illustrated in Fig. 3.3. Such a matrix manipulation step is performed as the correlation operation requires the RNA-velocity and the displacement vectors to have the same length. Then, we compute the displacement vectors between neighboring cells via the expanded peak count matrix. Finally, we correlate each promoter peak with the velocity of the corresponding gene for the same cell, using a similar approach to the original CellRank framework. Correlations are then transformed in probabilities via softmax.

The Multiomics+scVELO model consists of two input matrices: the gene expression matrix for highly variable genes, $X_G \in \mathbb{R}^{N \times n_{genes}}$, and the peak count matrix, $X_P \in \mathbb{R}^{N \times n_{peaks}}$. Similar to the scATAC-seq case, the ATAC matrix is subset to retain only the promoter peaks corresponding to the highly variable genes and then expanded for those peaks that are associated with multiple genes. In this multimodal setting, the RNA-velocity vector is correlated with both the gene expression and the peak counts for each cell. This integrated approach leverages the complementary information from the transcriptomic and epigenomic data to compute the transition matrix.

The process of creating the transition matrices for the scATAC-seq and Multiomics+scVELO models is illustrated in Fig. 3.3.

The CellRank framework provides the `PrecomputedKernel` class, which takes a user-supplied transition matrix and the annotated data object as input, and offers an interface to perform the CellRank computations, such as random walk simulations and the spectral clustering with GPCCA.

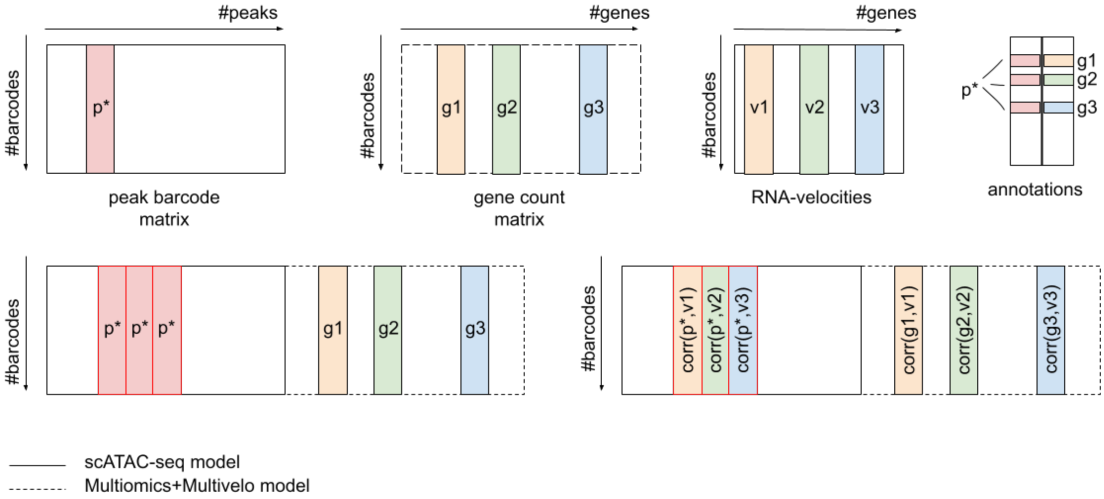


Figure 3.3: Transition Matrix computation for the scATAC-seq and Multiomics+scVELO models.

Finally, in the Multiomics+MultiVelo model, we utilize two input matrices: the gene activity matrix $X_A \in \mathbb{R}^{N \times n_{genes}}$ and the gene count matrix $X_G \in \mathbb{R}^{N \times n_{genes}}$, which contains the gene expression values. To compute the transition matrix, we correlate the neighboring cells' gene activities with the chromatin velocities, as well as the gene counts with the RNA-velocities. Both the chromatin and RNA-velocities are derived using the MultiVelo framework. After the individual correlation matrices are obtained, we combine them into a unified representation (Fig. 3.4) and apply a softmax transformation to convert the values into probabilities.

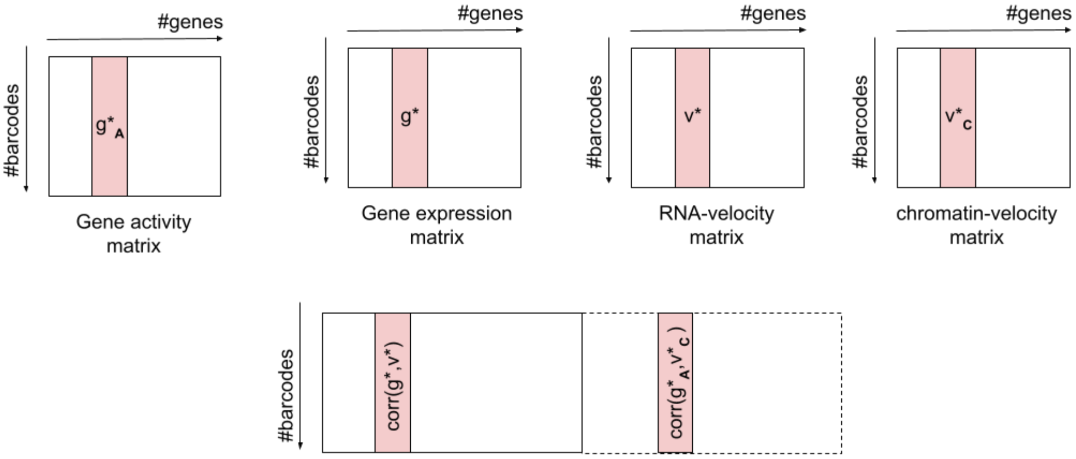


Figure 3.4: Transition Matrix computation for the Multiomics+MultiVelo model.

Following transition matrix construction, we employ the CellRank framework to simulate the system at the cell-state level using a random walk. CellRank provides the `plot_random_walks` function for both the `VelocityKernel` and `PrecomputedKernel` classes, which simulates a Markov Chain using the transition matrix and plots the results over the UMAP embedding. Such a function requires the user to provide an initial or terminal set of barcodes for the simulation. In our analysis, we randomly selected 250 barcodes from the "IPC", "V-SVZ", and "RG, Astro, OPC" clusters as the initial set and the initial probability distribution for the walk is the uniform over such a set. Each model, along with its parameter configurations, is simulated 200 times, with the stopping condition set to reach 25% of the total number of barcodes.

We then employ the Generalised Perron Cluster Cluster Analysis to coarse-grain the transition matrices through Schur decomposition. To this end, Cellrank provides a `GPPCA` class and its `compute_schur` method. This operation allows us to identify the 10 dominant real and complex conjugate eigenvalues. For the corresponding number of macrostates (n_s), we compute and plot the initial and terminal states, via the CellRank `GPCCA` class methods `compute_macrostates`, `predict_terminal_states`, `plot_terminal_states` and the corresponding initial macrostates functions. Here, we allow CellRank to identify overlapping terminal and initial macrostates, since we anticipate the models to recover not only the initial macrostate, but also a terminal macrostate within the "RG, Astro, OPC" cluster. We also retain the GPCCA minChi and crispness values for each model and parameter configuration to evaluate the quality of the spectral clustering.

Next, we compute the cell fate probabilities and determine the multilineage potential metrics. The CellRank method `compute_fate_probabilities` implements an iterative procedure to solve the linear problem described in Equation 2.21. Given the system complexity, we utilized petsc4py 3.20.5 [65] linear solver, as recommended by the CellRank framework. Furthermore, we employed the Incomplete LU (ILU) preconditioner in our implementation. This is because the convergence of the petsc4py solver typically depends on the spectrum of the input matrix. The use of preconditioning techniques, can alter the spectrum of the matrix and thereby accelerate the convergence rate of the iterative methods [65–68].

Finally, we retrieve multilineage potential metrics through the `compute_lineage_degree` function from CellRank specifying both the KL-divergence and entropy methods.

3.4 Evaluation metrics

This section outlines the specific strategies and metrics employed to evaluate models performance. Initially, it examines the transition matrix to gain insights into the behavior of each model at the cell-state level. Subsequently, it introduces GPCCA evaluation metrics that aid in determining the optimal number of macrostates. Lastly, it elucidates the metrics about multi-lineage potential, the identification of terminal states, and cell fate probabilities.

3.4.1 Markov Chains

In the previous sections, we highlighted how the behavior of a random walk is influenced by the topological properties of the graph $\mathcal{G}$ associated with the transition probability matrix P. Specifically, we will focus on two crucial properties: connectivity and aperiodicity. Establishing these properties is essential for ensuring the existence of an invariant probability distribution for the transition matrix, as per the theoretical results presented in Theorem 2.

To further assess the number of extremal invariant distributions π and their support, we construct the condensation graph $H_{\mathcal{G}}$ of the original graph $\mathcal{G}$. By analyzing the number of sink components in $H_{\mathcal{G}}$, we can gain insights into the structure of the invariant probability distribution of P. By thoroughly examining the connectivity, aperiodicity, and condensation structure of the graph $\mathcal{G}$, we can ensure that the Markov chain model satisfies the theoretical requirements for the existence and uniqueness of π. This, in turn, will enable us to confidently interpret the long-term cell fate probabilities and developmental trajectories predicted by the Markov chain runs.

Graph analysis has been performed using networkx 3.2.1 [69].

3.4.2 Number of Macrostates

The number of macrostates n_s for the spectral clustering using GPCCA is a crucial hyperparameter that can be tuned using various methods, as discussed in the literature [16, 52]. Some of the most commonly employed approaches for determining the optimal number of macrostates include the eigengap heuristic, crispness, and minChi criterion.

These methods provide insights into the macrostates sharpness as well as their overlap in composition- i.e., in terms of clusters.

The crispness (ϵ) metric measures the optimality of the solution for the GPCCA optimization problem, as presented in Equation 2.25. A larger value of ϵ corresponds to a smaller overlap between the macrostates, indicating a crisper assignment of cells in the membership matrix χ. Therefore, an optimal choice for n_s would maximize the following crispness criterion:

$$\frac{n_s - f_{n_s}}{n_s} = \frac{\text{trace}(\tilde{D}^{-}1\chi^{T}D\chi)}{n_s} \tag{3.3}$$

Complementary to the crispness metric, the minChi criterion suggests that selecting the number of macrostates n_s associated with a minChi value close to 0 leads to a crisper decomposition of the dynamics. The minChi value is computed as:

$$\text{minChi} = \min_i \min_j \chi_{i,j} \tag{3.4}$$

Therefore, the tuning of n_s involves first selecting the candidates with a minChi value close to 0, and then prioritizing the higher values of the crispness ϵ metric to identify the optimal number of macrostates.

3.4.3 Terminal States and Developmental Lineages

The primary objective of performing spectral clustering with GPCCA is to identify the initial and terminal macrostates while obtaining cell-specific probabilities of reaching each terminal state through a soft assignment procedure. To assess the biological relevance of the developed models, it is essential to examine whether the identified terminal states and their cellular composition align with the current biological understanding of embryonic mouse brain development. Based on the literature findings in Section 3.1, we expect the models to identify one terminal macrostate composed primarily of cells from the "RG, Astro, OPC" cluster, representing the glial cell lineage and at least one terminal macrostate associated with the neuronal developmental lineage, i.e., cells from the "Deeper Layer" and "Upper Layer" clusters. We expect also models to recover an initial state consisting of cells from the "RG, Astro, OPC" cluster. By assessing the alignment between the model outputs and the well-established biological underpinnings of embryonic mouse brain development, we can validate the ability of the computational framework to recover biologically meaningful terminal cellular states and developmental lineages.

3.4.4 Multilineage Potential

Multilineage potential refers to the inherent ability of a single cell or progenitor cell to undergo differentiation and give rise to multiple distinct cell lineages or cell types. It suggests that a particular cell possesses the capacity to develop into diverse cellular identities, providing valuable insights into the developmental potential of various cell populations and the hierarchical relationships that exist among them. In the context of this work, the computed cell fate probabilities are leveraged as a tool to evaluate the multilineage potential of the identified cellular populations.

CellRank [16], offers two methods for estimating multilineage potential. The first approach utilizes entropy calculations based on fate probabilities. In information theory, entropy refers to the average level of uncertainty or information associated with the potential outcomes of a random variable. In the case of a discrete probability random variable X, with a sample space Ω, the entropy is computed as:

$$H(X) := -\sum_{x\in\Omega} p(x)\log(p(x)) \tag{3.5}$$

A higher entropy value for X indicates a greater degree of uncertainty regarding the possible outcomes of the experiment. For instance, when flipping a fair coin, the outcome is more uncertain compared to an unfair coin, as the fair coin is expected to result in a more balanced distribution of outcomes. Consequently, a fair coin is associated with higher entropy than an unfair one.

Entropy, also known as Differentiation Potential (DP) in [17], serves as a measure of cell plasticity, reflecting the cell's capacity for differentiation. In the context of developmental lineage identification, DP is computed for each cell based on fate probabilities towards terminal states and it is then correlated with pseudo-time, revealing that early progenitor cells possess higher entropy (DP) compared to terminally differentiated cells. In CellRank [16], the entropy of a cell, denoted as $\mathcal{S}_i$, is utilized to quantify the extent to which the cell's fate probabilities f_i deviate from a uniform distribution. Cells exhibiting higher values of $\mathcal{S}_i$ indicate a reduced level of commitment to a particular fate, implying a higher potential for exploring alternative differentiation paths.

The second method utilized to estimate multilineage potential involves the use of Kullback-Leibler (KL) divergence. The KL divergence, also known as relative entropy and denoted as $D_{KL}(P||Q)$, is a statistical measure that quantifies the dissimilarity between a probability distribution of interest P and a reference probability distribution Q. Specifically, when given two probability distributions P and Q defined on the same sample space Ω, $D_{KL}(P||Q)$ calculates the expected logarithmic difference between P and Q. The KL divergence can be interpreted in various ways. In the context of machine learning, $D_{KL}(P||Q)$ represents the information gain obtained when using P instead of Q. In statistics, it corresponds to the expected value of the logarithm of the ratio of the likelihoods of the two distributions. In Bayesian statistics, the KL divergence measures the information gained by updating prior beliefs, represented by the prior probability distribution Q, with the posterior distribution P.

The Kullback-Leibler (KL) divergence, despite being used to measure the difference between two distributions and often interpreted as a measure of distance, does not meet the criteria to be considered a metric due to its inherent asymmetry. For discrete probability distributions, the relative entropy is defined as:

$$\begin{gathered} D_{KL}(P||Q) = \sum_{x\in\Omega} P(x) log\left(\frac{P(x)}{Q(x)}\right) \\ D_{KL}(Q||P) = -\sum_{x\in\Omega} P(x) log\left(\frac{Q(x)}{P(x)}\right) \end{gathered} \tag{3.6}$$

A relative entropy of 0 indicates that the two distributions being compared are identical.

In CellRank, the Kullback-Leibler (KL) divergence (also referred to as priming degree [70]) is employed to capture information regarding cell commitment and differentiation towards specific lineages by comparing the fate probabilities f_i of a cell i with the average fate probability per lineage across cells denoted as $\bar{f}$. This measure allows for assessing the extent of lineage priming, where a higher degree of priming indicates a stronger commitment of a cell to a particular lineage.

$$D_{KL}(f_i||\bar{f}) = \sum_{j \in \{terminal\ \ states\}} f_{ij} * log\left(\frac{f_{ij}}{\bar{f}_j}\right) \tag{3.7}$$

The use of both $\mathcal{S}_i$ and $D_{KL}(f_i||\bar{f})$ allows for the comparison of different models in terms of multilineage potential. This is because lower values of $D_{KL}(f_i||\bar{f})$ and higher values of $\mathcal{S}_i$ indicate cells with lower levels of commitment. By considering these measures together, it becomes possible to assess and compare the extent of cell commitment across different models.

To identify cell commitment during the development process, we therefore conducted a comparison between the different models using both entropy and KL divergence. Results are compared for models sharing the same size of the neighborhood K, the number of principal components PC, and the number of terminal macrostates n_s.

It is important to note that in scenarios where the initial cells are anticipated to exhibit a clear bias towards a specific fate direction, author in [16] suggests employing KL divergence instead of entropy. This recommendation is based on the observation that KL divergence increases monotonically as cells progress towards terminal states. In contrast, entropy reaches its maximum at the point where the initial and terminal states converge closest to a uniform distribution.

3.4.5 Preprocessing Parameters

In this analysis, we investigate the impact of the size of the local neighborhood K, the number of principal components PC, and the number of SVD dimensions LSI on the models' results. We aim to assess the robustness of the models and understand how these preprocessing parameters influence the key outcomes.

First, we identify whether there are any variations in the models' ability to recover the developmental lineages, terminal and initial macrostates, and macrostate compositions. Coherent results across different parameter settings would indicate the models' robustness to these variations.

Next, we explore the impact of the parameters on the multilineage potential. To do this, we fit linear models to the KL-divergence values for each cell cluster and the different n_s values. The linear model is defined as follows:

$$\begin{aligned} KL = &\beta_0 + \beta_1 * K + \beta_2 * PC + \beta_3 * LSI + \beta_4 * K * PC + \\ &\beta_5 * K * LSI + \beta_6 * PC * LSI + \beta_7 * K * PC * LSI \end{aligned} \tag{3.8}$$

The coefficient β_0 represents the baseline, corresponding to the average value of KL-divergence when $K = 20$, $PC = 10$, and $LSI = 10$. This serves as a reference point to understand the changes in KL-divergence as the parameter values vary. The β_1, β_2, and β_3 coefficient vectors represent the changes in mean KL-divergence compared to the baseline values when a single parameter changes, while keeping the other parameters constant. For example, $\beta_{1,1}$ is the mean change in KL-divergence when $K = 30$ is used instead of $K = 20$, keeping PC and LSI constant (10,10). The additional coefficient vectors (β_4 to β_7) represent the added KL-divergence means due to the interaction of two or three changing parameters with respect to the baseline. Therefore, to explain the effect of $K = 30$, $PC = 30$ and $LSI = 10$ on multilineage potential, then one should compute $\beta_0 + \beta_{1,1} + \beta_{2,1} + \beta_{4,1}$.

The presence of many categorical combinations, however, makes single coefficients and the associated p-values difficult to interpret. To determine these interactions' impacts, we use ANOVA to investigate the following hypotheses:

$$\begin{aligned} H_0 &: \beta_i = 0,\ i = 4{,}5{,}6{,}7 \\ H_A &: \beta_i \neq 0,\ i = 4{,}5{,}6{,}7 \end{aligned} \tag{3.9}$$

The ANOVA test shows a $p - value < 0.05$ when interactions are significant. In such cases, we further evaluate if any trend exists in the multilineage potential as the parameter combinations vary.

Chapter 4

Results

This chapter presents the key findings and comparisons across the various models investigated in this study. The primary focus is on the performance of the multiomics-based approaches in relation to the transcriptomic-only model. Additionally, the chapter provides an analysis of the scATAC-seq model's behavior in comparison to the scRNA-seq framework.

At first, the chapter examines the cell-state level behavior of the different models. This analysis aims to determine whether the scVEMO approach can effectively simulate the developing system under investigation. Moving forward, it delves into the results of the GPCCA analysis. Here, the goal is to study whether the models accurately identify the developmental lineages, as well as the initial and terminal states. We will consider the models that can recover the two distinct developmental lineages and the differentiation between the upper and deeper cortical layers as the well-performing ones.

Finally, this chapter provides an evaluation of the cell fate commitment. It first introduces an analysis of the cluster-specific average fate commitment towards the terminal states for each model. However, since the performance of this metric is heavily influenced by the identified macrostates, models that recover different states are unlikely to be directly comparable. To address this challenge, we will leverage the multilineage potential analysis, which has been successfully employed in the Palantir framework. We expect the well-performing models to exhibit high potential values (i.e., low lineage commitment) for the progenitor clusters.

Through this comprehensive analysis, this chapter aims to provide a thorough understanding of the strengths and limitations of the multiomics-based approaches in comparison to the transcriptomic-only and scATAC-seq models.

4.1 Markov Chain Simulations

To understand the behavior of the random walks, we first investigate the properties of the graph $\mathcal{G}$ associated with each transition matrix P. We identify the connectivity and aperiodicity topological properties of the graph associated with the transition probability matrix P, as well as the condensation graph $H_{\mathcal{G}}$, for each model and parameter configuration. Finally, we run the Markov chain on P.

All scRNA-seq models exhibit transition matrices P that correspond to an aperiodic and strongly connected graph $\mathcal{G}$. The associated condensation graphs $H_{\mathcal{G}}$ consist of a single sink component as shown in Fig. 4.1a. Consequently, the transition matrices possess a single invariant probability distribution π supported on the entire embedding, as highlighted in Theorem 2.

scATAC-seq transition matrices also yield to an aperiodic and strongly connected graph. However, there is one exception in the $K = 10, LSI = 20$ model where the graph is not connected, and two sink components are present (Fig. 4.1b.).

Instead, the Multiomics + scVELO models transition matrices result in a single sink in $H_{\mathcal{G}}$ and an aperiodic graph only for $K > 10$. When $K = 10$, regardless of the PC and LSI parameters, the resulting graphs exhibit two disconnected components. This observation suggests that the neighborhood size is too small to recover a connected system. On the other hand, the Multiomics+MultiVelo models consistently produce strongly connected and aperiodic graphs across all parameter configurations. This result can be attributed to the less restrictive nature of the KNN approach, which allows for the formation of more densely connected graphs compared to the SNN one used in the Multiomics+scVELO framework.

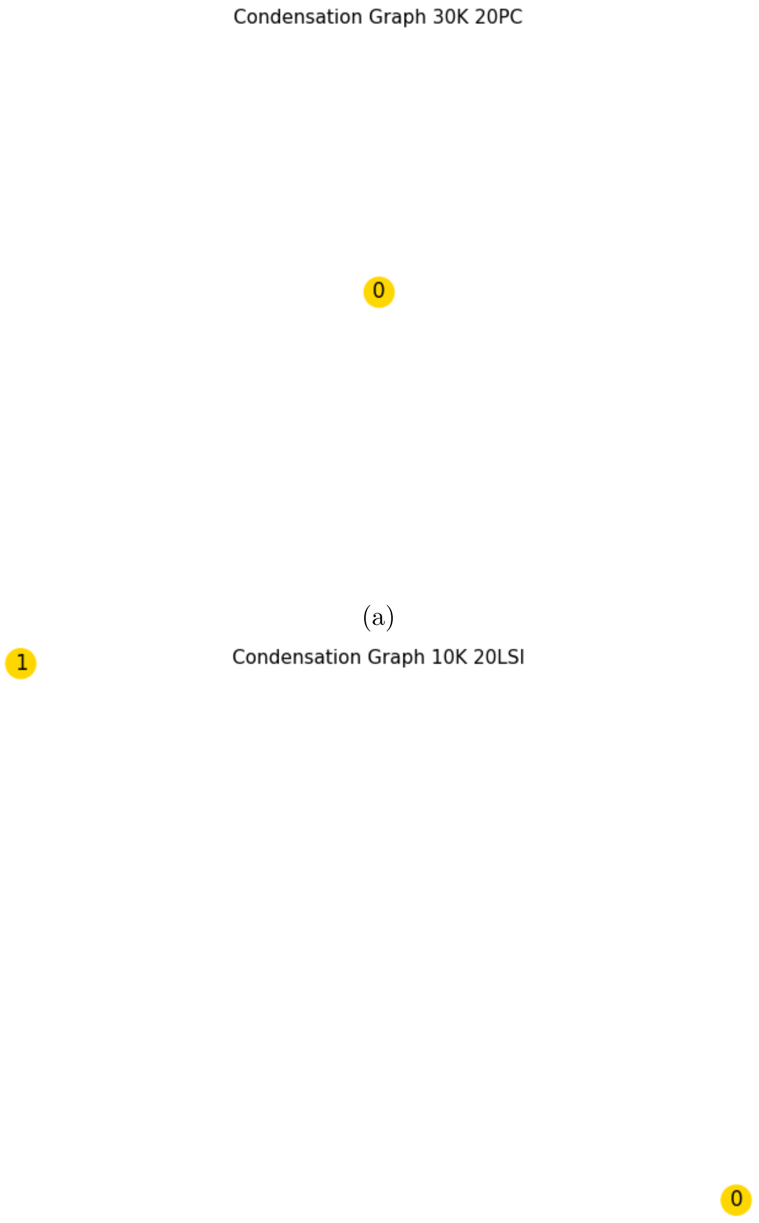
As discussed in previous sections, the Convergence in Probability Theorem (Theorem 3) allows us to predict the expected behavior of a random walk using the invariant probability distribution of the transition matrix, π. However, this requires strong connectivity and aperiodicity to be satisfied. When these conditions are met, the long-term behavior of the Markov chain $\hat{\pi}$ is guaranteed to converge to the invariant probability distribution π. This means that the random walk will eventually settle into a stable pattern that reflects the underlying dynamics of the system.

On the other hand, if the aperiodicity constraint is not satisfied, the nodes in the graph $\mathcal{G}$ may form a grid-like structure. In such cases, the random walk can exhibit periodic behavior, repeatedly visiting the same set of nodes cyclically. Furthermore, the dominant eigenmodes of the system will correspond to these periodic patterns, causing the random walk to recognize these cycles as metastable states and remain trapped within them. However, such a scenario never occurs in our simulations.

Lastly, if the graph $\mathcal{G}$ is disconnected, the random walk is confined to the starting component, and its long-term behavior is determined solely by the initial probability distribution $\hat{\pi}(0)$. This means that the random walk is unable to explore the full state space of the system, limiting the conclusions that can be drawn from the analysis.

In Fig. 4.2-4.5 random walks for the four models of interest are shown. In the plots, black-color coded dots represent starting cells, while the yellow ones coincide with the simulations' ending points. The edges' colors tend to yellow as the walk reaches its end. This visual encoding provides information about the progression towards the terminal states within the lineage reconstruction framework.

The majority of the scRNA-seq simulations end in the "Deeper Layer" cluster, with a few reaching cells in the "RG, Astro, OPC" one. These results underscore the ability of the CellRank framework to effectively capture the developmental trajectory of the neuronal lineage. In contrast, the gliogenic lineage, which is a crucial component of the developing system under investigation although minor, is not as comprehensively recovered by the



(a)

(b)

Figure 4.1: Examples of condensation graphs $H_{\mathcal{G}}$ in the case of a strongly connected graph (**a.**) and a disconnected graph (**b.**).

scRNA-seq model.

Conversely, simulations of the scATAC-seq and the two multiomics models result in cells scattered throughout the embeddings. For such cases, we investigate the 30 most likely cells on each embedding, highlighting the barcodes associated with higher invariant probability distribution values (Fig. 4.6). Interestingly, these top likely cells span the entire UMAP. This outcome implies that the transition matrices, which correlate RNA velocities with genes and/or promoter peak frequencies, fail to accurately capture the developmental trajectories at the cell-state level. Such a result could be attributed to two potential factors:

- The correlation operation between neighboring displacement vectors and cell-specific RNA velocities may not be aligning the graph with the expected developmental lineages.
- The correlation operation between neighboring displacement vectors and cell-specific RNA velocities tends to assign high probability values also to cell-state transitions that are opposite to the anticipated direction of development.

Despite the models are not able to simulate the embryonic mouse brain dynamic at the single-cell level using P, we will later see that the incorporation of the cellular epigenomic profiles improves the identification of the macrostates.

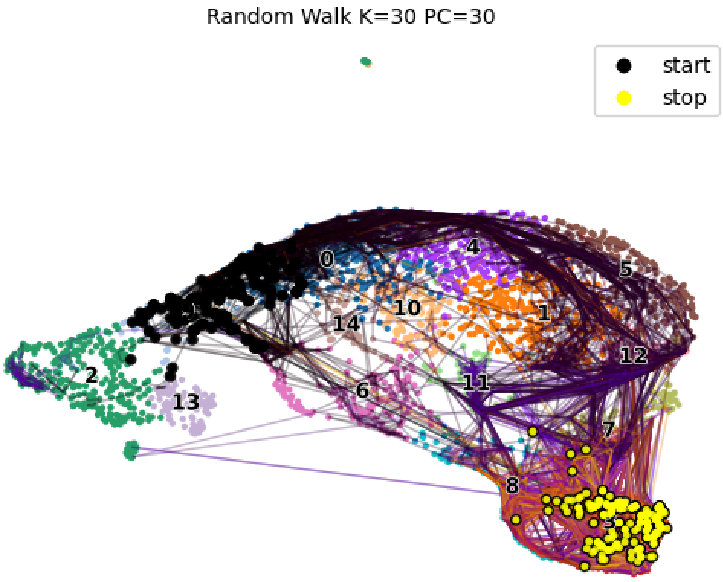


Figure 4.2: Random walk simulation for scRNA-seq model. Edges color becomes lighter (yellow) as the simulations reach the end

4.2 Number of Macrostates

In this work, we tune the number of macrostates (n_s) used in the GPCCA (Generalized Perron Cluster Cluster Analysis) framework by leveraging a combination of two key metrics: minChi and crispness. Optimal values of n_s are those where minChi is 0 and have higher crispness values, indicating that the macrostates are well-separated.

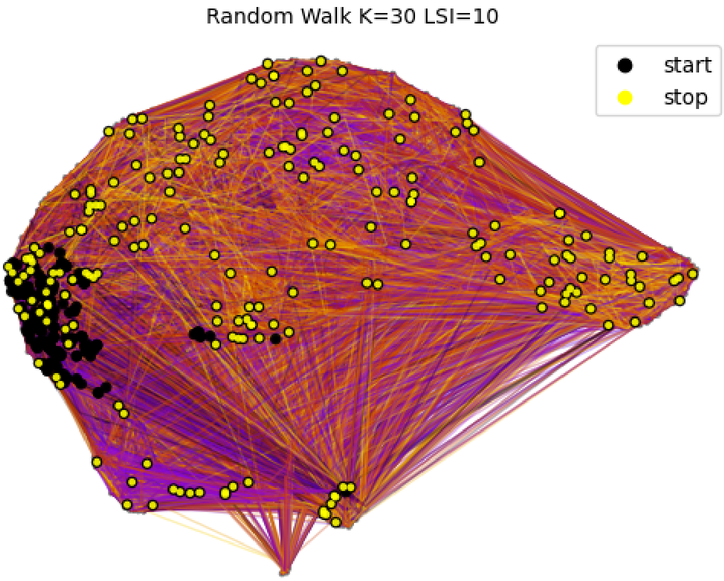


Figure 4.3: Random walk simulations for scATAC-seq model. Edges color becomes lighter (yellow) as the simulations reach the end

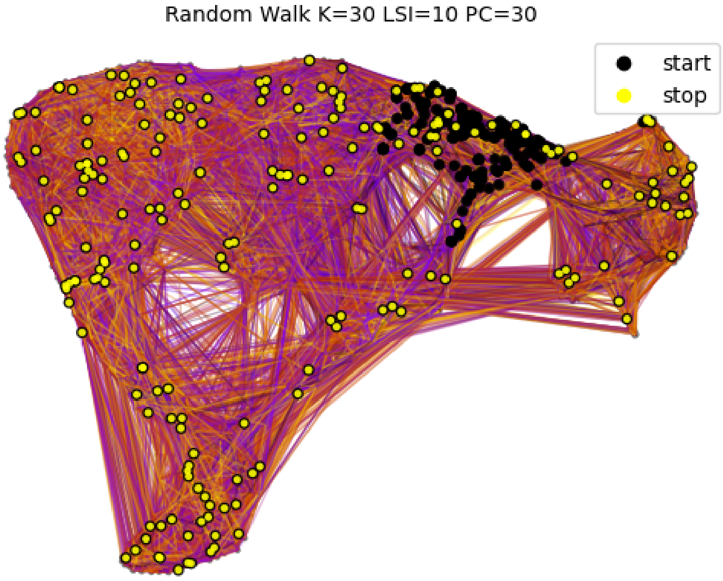


Figure 4.4: Random walk simulations for Multiomics+scVELO model. Edges color becomes lighter (yellow) as the simulations reach the end

Figure 4.7-4.10 illustrate the results of our investigation, where each data point represents a combination of preprocessing parameters, color-coded based on the number

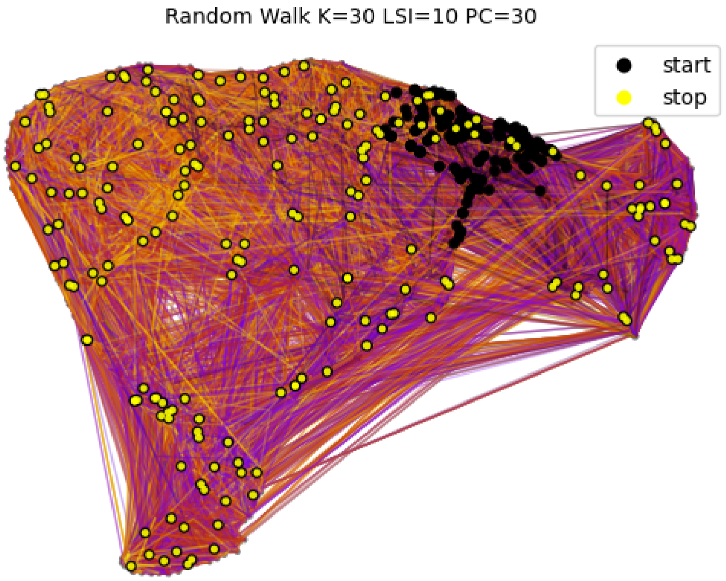


Figure 4.5: Random walk simulations for Multiomics+MultiVelo model. Edges color becomes lighter (yellow) as the simulations reach the end

of macrostates. We have values of n_s ranging from 2 to 10, as we expect at least two developmental lineages to be present - the glial and the neuronal-related ones.

The analysis of the minChi metric shows that all the considered numbers of macrostates behave correctly, with all values being 0.

When examining the crispness metric, we set two thresholds: 0.6 and 0.7. The multiomics models with $n_s = 2$ and $n_s = 3$ deliver crispness values above the 0.6 threshold, with only the models with $n_s = 2$ surpassing the more stringent 0.7 threshold. The Multiomics+scVELO models provide in general higher crispness values than the Multiomics+MultiVelo ones. For the scRNA-seq data, we also observe good solutions for some $n_s = 4$. Conversely, the scATAC-seq models provide lower crispness values overall, suggesting that the identification of developmental lineages based only on the cell epigenomics profiles might not actually provide accurate results.

Based on these findings, we will consider the models with $n_s = \{2,3\}$ for further analysis. This result aligns with our expectations regarding the number of terminal states in the system. We anticipate that the initial and one terminal state should coincide with cells in the "RG, Astro, OPC" cluster, representing the glial cell lineage. Furthermore, we expect at least one terminal state to be recovered in the neuronal lineage, specifically composed of cells from the "Deeper Layer" and "Upper Layer" clusters.

Concordantly, the case of $n_s \geq 4$ leads to a series of undesirable properties:

- Some macrostates within the same embedding exhibit smaller sizes. In certain cases, it is not even possible to identify a macrostate with more than the minimum size of 6 cells, leading to limitations in determining reliable cell fate probabilities.

- The composition of the macrostates becomes highly heterogeneous, with many macrostates not being uniquely associated with a single cell type and significant overlap occurring in terms of the underlying cell clusters.
- The models are unable to accurately identify the developmental lineages of interest.
- A considerable number of entries in the coarse-grained transition matrix are negative, indicating bad problem conditioning.

Consequently, we focus our subsequent analyses exclusively on the results obtained with 2 and 3 macrostates, as these configurations exhibit more desirable properties and align better with our expectations regarding the number of terminal states in the system.

4.3 Developmental Lineages and Fate Probabilities

Based on the literature review presented in Section 3.1, we expect the models to recover two key developmental lineages. The first anticipated lineage is the glial trajectory, which encodes the transition from radial glia cells into more differentiated glial cell types, such as astrocytes and oligodendrocyte precursor cells. We therefore expect to observe a lineage that is primarily composed of cells from the "RG, Astro, OPC" cluster. The second expected lineage is the neuronal development trajectory, which should include the neural intermediate progenitor cells (IPC), ventricular, subventricular zone (V-SVZ), and subplate cells, eventually terminating in the cerebral cortex clusters. Superior model predictions will further distinguish between the upper and deeper cortical layers, as they correspond to early-born and later-born neurons, respectively. In addition to the identification of these developmental lineages, we will investigate the macrostate composition for each initial and terminal state. This analysis will provide further insights into the accurate identification of the macrostates and their biological relevance, with heterogeneous composition being an indicator of worse model performances.

We begin by investigating the case where the number of macrostates, n_s, is set to 2. The results of this analysis are illustrated in Fig. 4.11-4.14.

Interestingly, all of the models considered in this study correctly recover the single initial "RG, Astro, OPC" macrostate. This suggests that the various computational techniques can unanimously identify the starting point of the developmental process corresponding to the multipotent radial glia cells.

Moving forward, the models also correctly identify two terminal developmental lineages, corresponding to the "Deeper Layer" and "RG, Astro, OPC" macrostates. From the biological perspective, these recovered lineages encode the expected trajectories of cell differentiation, with RG transitioning to IPCs and giving rise to neurons and the glial products, astrocytes, and oligodendrocytes, arising from the radial glia as well.

To further validate these findings, we examine the macrostate composition, as shown in Fig. 4.15-4.18. The cells belonging to the corresponding homonymous clusters are indeed the predominant contributors to the identified macrostates, providing additional confidence in the biological relevance of the results.

In Fig.4.11-4.14 cells are color-coded according to the cell-fate probabilities. Interestingly, the scRNA-seq model exhibits higher probability values compared to the multiomics-based models, particularly within the neuronal lineage, while the scATAC-seq model presents the lowest probability values overall. To further investigate these cell fate probabilities, we conducted a comparative analysis across the four models. This involved examining the average probability for each cell cluster to transition towards the two terminal states, along with the corresponding 95% confidence intervals. The results are plotted in Fig. 4.19. The resulting confidence intervals are narrow, indicating a sufficient number of samples to yield reliable estimates for the average fate probabilities. The desired model behavior entails assigning a greater average fate probability to cells within the "RG, Astro, OPC" cluster, thereby facilitating movement towards the corresponding terminal state. Conversely, all remaining clusters should exhibit a higher average probability towards the "Deeper Layer" macrostate.

Upon examination, it becomes evident that for the "Deeper Layer" terminal state (Fig. 4.19b), the scRNA-seq embedding exhibits higher average fates for all clusters, compared to the multiomics approach. Furthermore, when considering the "RG, Astro, OPC" terminal state (Fig. 4.19a), all multiomics average probabilities surpass the corresponding scRNA-seq values, with the highest average probabilities associated with the "RG, Astro, OPC" cluster. However, a closer inspection of the scRNA-seq model with $n_s = 2$ reveals an interesting observation. The "Upper Layer" cluster displays a strong average fate probability to transition towards the deeper cortical macrostate, which is not expected, as the cortical layers II-IV are formed later in development than the layers V-VI. Having fixed the terminal states, the absorption probabilities computation is subject to the partition of unity and non-negativity constraints. As a result, all cells can either transition towards the "Deeper Layer" or the "RG, Astro, OPC" clusters, regardless of the potential presence of other terminal states. The scRNA-seq model assigns then a strong probability to transition towards the "Deeper Layer" terminal state to most clusters. Such a distribution results from the strongly committed random walks towards the corresponding cluster. Conversely, the multiomics-based Markov Chain simulations are not as engaged to the "Deeper Layer" cluster as the baseline model, a result that might indicate the presence of additional terminal states within the same lineage. Such a result explains the lower cell-fate transition probabilities in the multimodal approaches, as well as the results in the $n_s = 3$ case, where cells show increased cell-fate probability, although towards the "Upper Layer" terminal state.

The scATAC-seq-based model stands out as the weakest performer in terms of average cell fate probabilities compared to the other approaches. When examining the "RG, Astro, OPC" terminal state, the scATAC-seq model exhibits the highest average probabilities across all neuronal-related clusters. Conversely, when considering the "Deeper Layer" terminal state, the scATAC-seq approach demonstrates the lowest average cluster probabilities. Interestingly, for the "RG, Astro, OPC" cluster itself, the scATAC-seq model performs comparably to the Multiomics+MultiVelo framework.

More interesting results emerge when considering the case of $n_s = 3$ terminal states. The scRNA-seq model identifies two "RG, Astro, OPC" macrostates and a single "Deeper Layer" macrostate (Fig. 4.24). The lineage reconstruction plot displayed in Fig. 4.20 illustrates that the scRNA-seq-based model recovers the same terminal states and developmental lineages as in the $n_s = 2$ scenario. In contrast, the multiomics models uncover an "Upper Layer" terminal state in addition to the "RG, Astro, OPC" and "Deeper Layer" ones (Fig. 4.22,4.23).

Regarding cell fate probabilities, the outcomes mirror findings obtained in the scenario with $n_s = 2$. Considering the "Deeper Layer" terminal state, the multiomics models consistently exhibit lower average fate probabilities compared to the scRNA-seq-only approach, regardless of the specific cell cluster (Fig. 4.28b). When examining the "RG, Astro, OPC" terminal state (Fig. 4.28a), the average fate probabilities for the "RG, Astro, OPC" cluster in the multiomics models exceed the scRNA-seq. However, average fate probabilities for all other clusters are lower in the scRNA-seq models compared to the multiomics.

Interestingly, compared to the $n_s = 2$ scenario, the cell fate probabilities towards the "RG, Astro, OPC" terminal state for neuronal-related clusters decreased in the multiomics models. This result suggests that the higher number of macrostates has further boosted the performance of the multiomics models.

In the $n_s = 3$ scenario, the scATAC-seq model recovers a macrostate composed mostly of "V-SVZ" cells, in addition to the "Deeper Layer" and the "RG, Astro, OPC" ones. Consistent with the $n_s = 2$ case, the scATAC-seq model presents the lowest performances for the average cell fate probabilities across all clusters.

When comparing the Multiomics+scVELO and Multiomics+MultiVelo approaches, several key insights emerge. In both $n_s = 2$ and $n_s = 3$ scenarios, the average cell fate probabilities towards the "RG, Astro, OPC" and "Deeper Layer" macrostates are generally comparable between the two models. However, the Multiomics+scVELO framework exhibits a slight advantage when considering the "RG, Astro, OPC" cluster, providing results more aligned with the expected fate behaviors. Turning to the "Upper Layer" terminal state, generally, the scVELO-related model yields superior average cell fate probabilities compared to the MultiVelo-based approach. Finally, the "Upper Layer" macrostate in the MultiVelo case shows a higher percentage of "V-SVZ" cells (Fig. 4.27). While the two approaches generally produce comparable results, the scVELO model demonstrates enhanced performance in capturing the behavior of specific cell clusters, particularly those associated with the "Upper Layer" macrostate (Fig. 4.29). This highlights the importance of carefully evaluating model performance across diverse cellular subpopulations to fully appreciate the strengths and tradeoffs of the different multimodal integration strategies.

The identification of the "Upper Layer" terminal state can be considered an improvement of the multiomics models over the scRNA-seq approach. Additionally, the accurate identification of biologically relevant states in the multiomics scenario indicates that, even though the transition probability matrix correlating RNA-velocities and promoter peak frequencies may not fully capture the system's development, it can still be effectively utilized for identifying terminal states. The discrepancy in the scATAC-seq model's behavior across different cell types highlights the limitations of relying solely on chromatin accessibility data for comprehensive lineage reconstruction. This underscores the value of integrating multimodal information, such as transcriptomic and epigenomic profiles, to achieve more robust and reliable predictions of cellular developmental trajectories.

4.4 Multilineage Potential

A significant factor that impacts the performance of the multiomics-based models in terms of average cluster fate probabilities is the identification of the "Upper Layer" terminal state. The presence of different macrostates in the $n_s = 3$ scenario makes it challenging to interpret the model comparisons indeed. To gain deeper insights into cell fate probabilities and accurately evaluate developmental lineages, it is important to examine the multilineage potential of the cells. Multilineage potential refers to the capacity of a single cell or progenitor cell to give rise to multiple distinct cell lineages or cell types. This metric has been previously exploited in other lineage tracing frameworks [16, 17].

This approach utilizes the cell fate probabilities to calculate the KL-divergence and entropy measures, as highlighted in the previous sections. The underlying premise is that progenitor cells are expected to exhibit low commitment (or higher potential), while terminally differentiated clusters should display high fate commitment (or lower potential). Importantly, lower commitment levels correspond to higher entropy and lower KL-divergence values.

In this study, we will specifically compare the average KL-divergence and entropy, along with the corresponding 95% confidence intervals, for the following cell clusters: "RG, Astro, OPC", "Deeper Layer", "Upper Layer", and "IPC". Results are illustrated in Fig. 4.30 and Fig. 4.31.

The multiomics-based models demonstrate their ability to accurately recover the multilineage potential of the cellular subpopulations. When examining the KL-divergence values, the multiomics-based approaches correctly exhibit higher values compared to the scRNA-seq model for the "Deeper Layer" and "RG, Astro, OPC" clusters, across both the $n_s = 2$ and $n_s = 3$ configurations. This trend is further corroborated by the entropy results. The multiomics-based models present lower entropy for the "Deeper Layer" cluster, while showing lower or comparable entropy for the "RG, Astro, OPC" one. These findings indicate that the multiomics-based frameworks can better capture the fate commitment of these cell populations, with the terminally differentiated "Deeper Layer" cells exhibiting lower multilineage potential compared to the more progenitor-like "RG, Astro, OPC" cluster.

Interestingly, all the models retrieve higher commitment compared to the scRNA-seq case, as evidenced by lower entropy and higher KL-divergence values, for the "IPC" cluster. While we might expect high multipotency levels in this cluster, representing pluripotent cells, the literature suggests that such cells are only capable of developing towards the neuronal lineage and cannot differentiate into astrocytes or oligodendrocytes. This commitment of the "IPC" cluster to the neuronal lineage might justify the multilineage potential metrics behavior.

When considering the "Upper Layer" cluster, an additional insight emerges. In the $n_s = 2$ scenario, we do not expect the fate probabilities to drive these cells towards a specific lineage, as the "Upper Layer" terminal state is not recovered by the models. However, the $n_s = 3$ configuration highlights how the integration of epigenomic profiles in the multiomics-based approaches yields better results in macrostate identification, which in turn increases the lineage commitment of the "Upper Layer" cluster (Fig. 4.30b).

Notably, the scATAC-seq model also provides the worst performance in identifying multilineage potential. With $n_s = 2$, the scATAC-seq model fails to recover the commitment levels of the "RG, Astro, OPC" cluster, while in the $n_s = 3$ scenario, it performs comparably to the scRNA-seq case. Regarding the deeper cortical layers cluster, the scATAC-seq model slightly outperforms the scRNA-seq one for $n_s = 2$, while for $n_s = 3$, it recovers better performances than the baseline. Lastly, the scATAC-seq model exhibits similar performances to the transcription-only-based approach when considering the "IPC" cluster.

4.5 Preprocessing Parameters

In this section, we present the evaluation of the models' behavior in response to variations in preprocessing parameter settings.

For the $n_s = 2$ scenario, all the models exhibit robustness to the parameter variations, consistently recovering the two developmental lineages associated with the "RG, Astro, OPC" and "Deeper Layer" macrostates. However, when considering the $n_s = 3$ case, the models display divergent behaviors. The scRNA-seq model remains confined to the same lineages as in the $n_s = 2$ setting, while the multiomics-based approaches continue to identify the "Upper Layer" terminal state. This terminal state typically comprises a composition of subventricular zone cells and upper layer neurons, although its specific composition tends to improve and exclude the V-SVZ cells as PC and K are increased. Interestingly, the Multiomics+scVELO model outperforms the Multiomics+MultiVelo approach in accurately capturing the upper layer terminal state.

Another improvement observed in the multiomics models compared to the scRNA-seq approach is the identification of the initial states in the $n_s = 3$ scenario. The transcriptomic-based model does not always recover an initial state in the "RG, Astro, OPC" cluster, sometimes identifying the initial developmental point with ependymal or nIPCs cells instead. In contrast, the multiomics-based models exhibit robustness in accurately identifying an initial macrostate in the "RG, Astro, OPC" cluster.

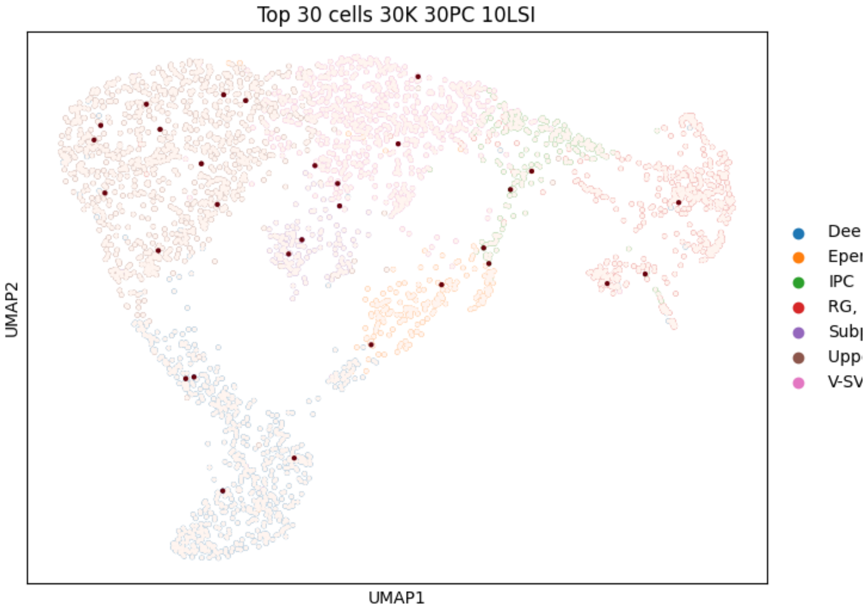
In contrast, the scATAC-seq model struggles to perform GPCCA and identify macrostates for high values of K. These results further emphasize that relying solely on epigenomic data, likely due to its inherent sparsity, cannot reliably recover the cellular development trajectories.

To further examine the influence of parameter configurations on the multilineage potential, we perform the ANOVA model presented in Equation 3.8. The results of this analysis, reported in Appendix A, indicate that the parameter interactions are statistically significant for the Multiomics+scVELO model. Therefore, we investigate eventual trends in the multilineage potential by plotting the results of the linear model (Appendix A). In the scenario with $n_s = 2$ macrostates, different parameter configurations display comparable behaviors. However, in the $n_s = 3$ case, we observe a positive trend associated with increasing values of the K parameter, especially for the "RG, Astro, OPC" and "Deeper Layer" clusters. Additionally, the identification of a terminal "Upper Layer" macrostate improves the KL-divergence for all parameter combinations compared to the $n_s = 2$ scenario, especially when $PC > 10$.

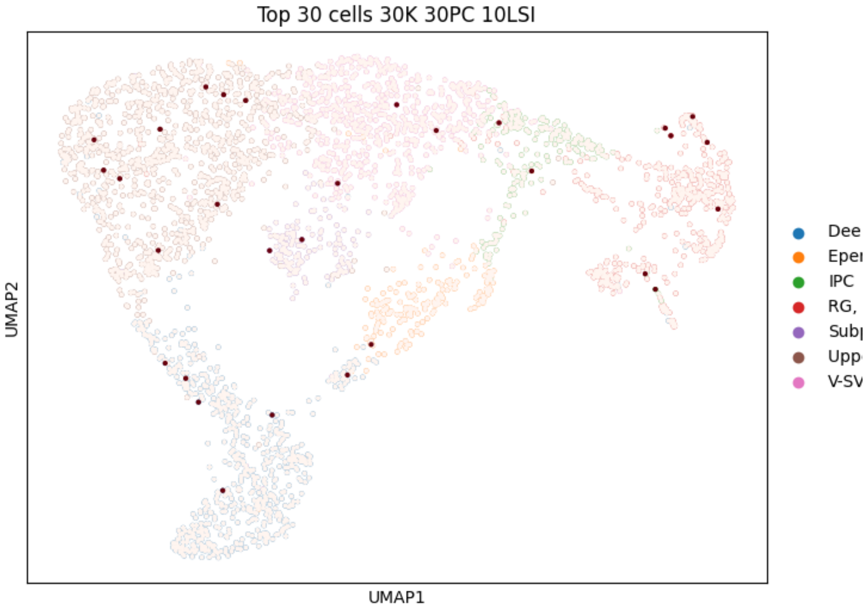
Similar findings concern the Multiomics+MultiVelo model. When considering the

$n_s = 2$ macrostate scenario, the parameter interactions are non or little significant for the "Deeper Layer" and "IPC" clusters. Even in the "RG, Astro, OPC" cluster, where a negative trend with increasing K and $PC > 10$ is observed, the confidence intervals for the different parameter settings overlap. This indicates that the model maintains a relatively consistent performance across the tested parameter configurations, without exhibiting drastic changes in the KL-divergence metric. As in the previous model, the analysis for $n_s = 3$ shows a positive trend in KL-divergence with increasing K values, despite its magnitude being smaller compared to the Multiomics+scVELO model.

Although there are statistically significant interactions between the parameters, the F-statistic and the difference between the residuals in the ANOVA are relatively small, indicating that the interactions may not have a substantial practical impact. The models' behaviors do not significantly change between different parameter configurations, and the confidence intervals in the plots are overlapping. The most appreciable impact is observed with the K parameter, ultimately improving the multilineage potential identification in the clusters, thus enhancing the overall model performance.



(a)



(b)

Figure 4.6: Top 30 likely cells according to the invariant probability distribution in Multiomics+scVELO (**a.**) and Multiomics+MultiVelo (**b.**) models.

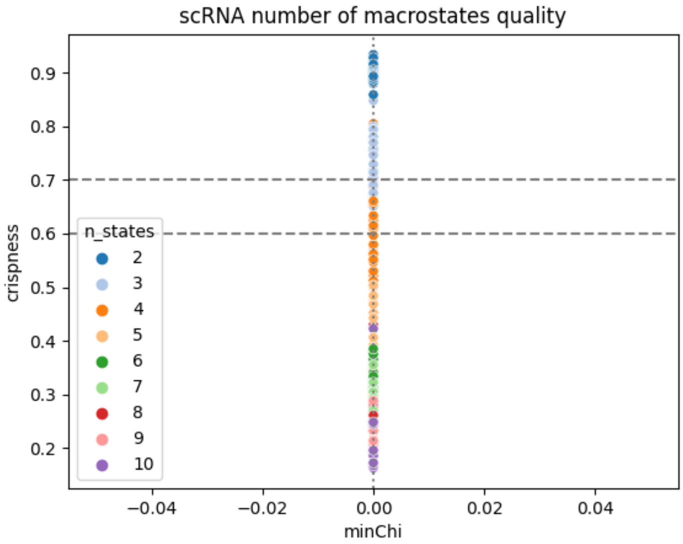


Figure 4.7: Scatterplot evaluating GPCCA macrostate quality for the scRNA-seq model. The dotted vertical line corresponds to the optimal *minChi* $= 0$ value, while dashed horizontal lines correspond to the *crispess* $= \{0.6, 0.7\}$ optimality thresholds.

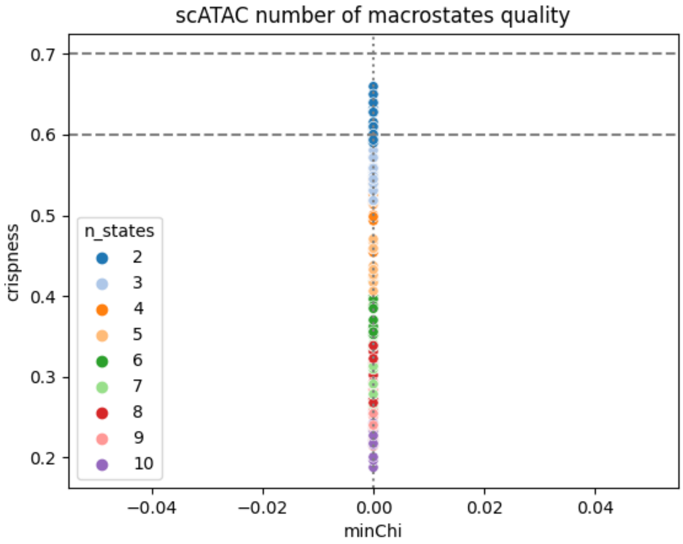


Figure 4.8: Scatterplot evaluating GPCCA macrostate quality for the scATAC-seq model. The dotted vertical line corresponds to the optimal *minChi* $= 0$ value, while dashed horizontal lines correspond to the *crispess* $= \{0.6, 0.7\}$ optimality thresholds.

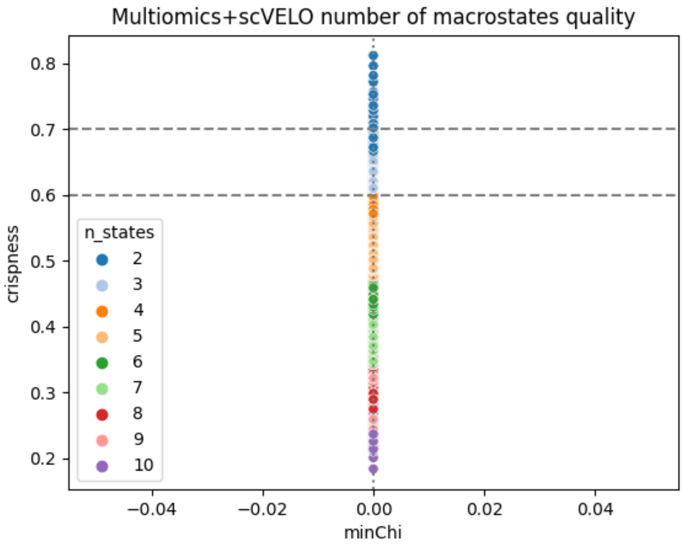


Figure 4.9: Scatterplot evaluating GPCCA macrostate quality for the Multiomics+scVELO model. The dotted vertical line corresponds to the optimal *minChi* = 0 value, while dashed horizontal lines correspond to the *crispess* = $\{0.6, 0.7\}$ optimality thresholds.

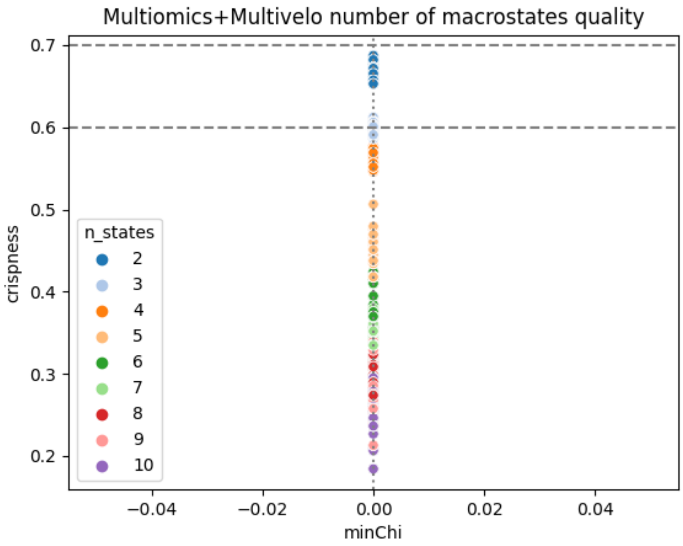
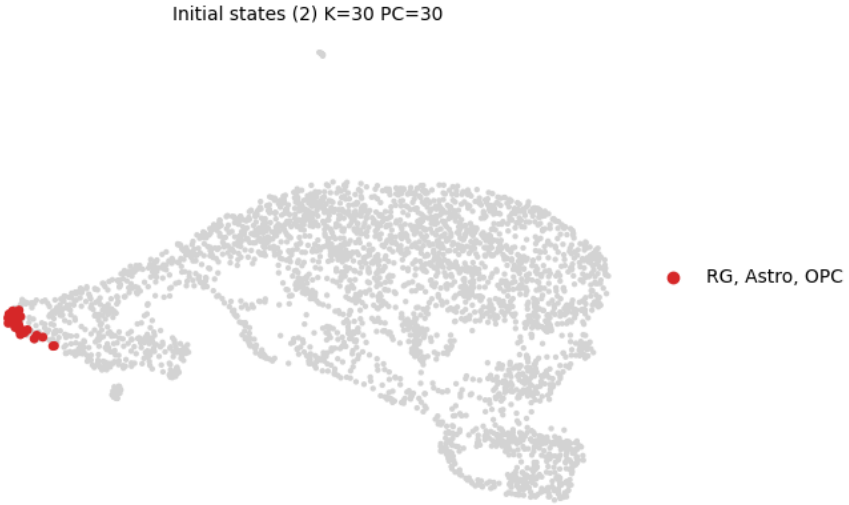
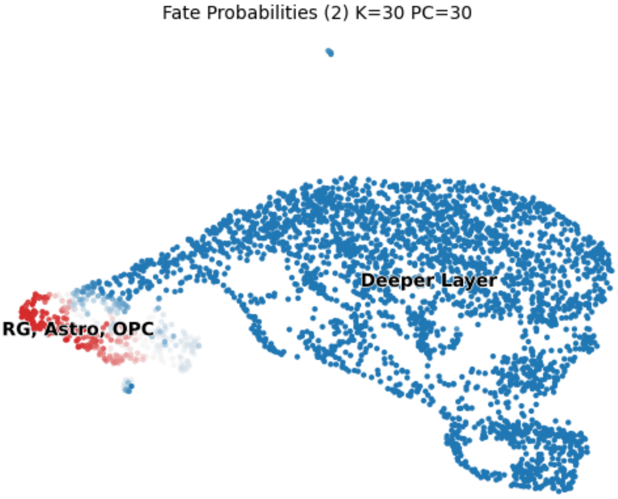


Figure 4.10: Scatterplot evaluating GPCCA macrostate quality for the Multiomics+MultiVelo model. The dotted vertical line corresponds to the optimal *minChi* = 0 value, while dashed horizontal lines correspond to the *crispess* = $\{0.6, 0.7\}$ optimality thresholds.

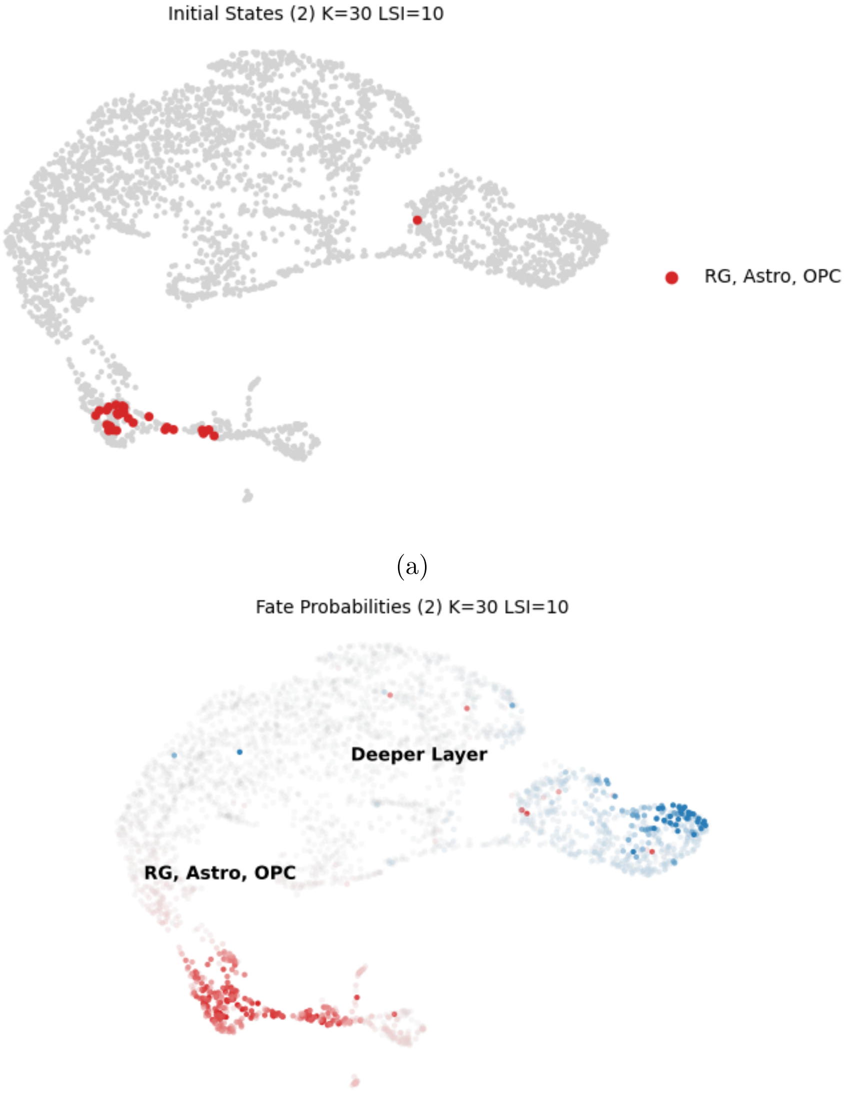


(a)



(b)

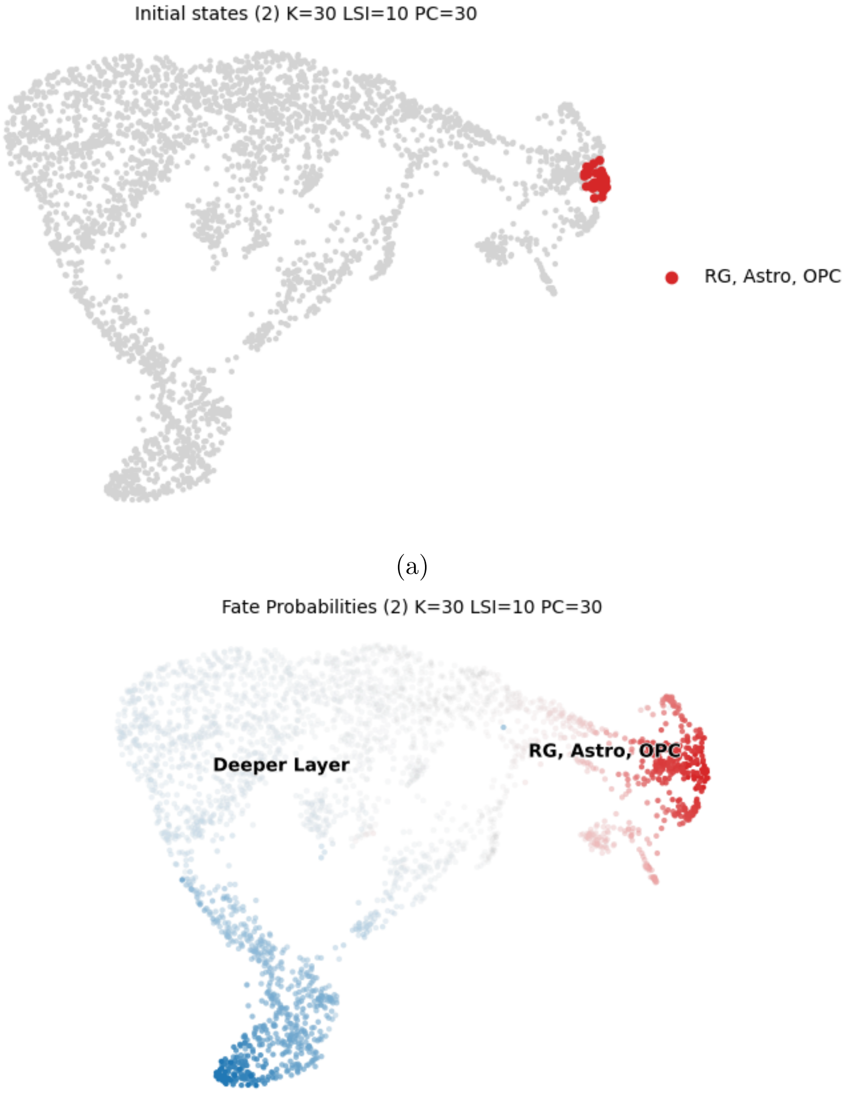
Figure 4.11: Initial (a.) and terminal macrostates with cell fate probabilities (b.) for the scRNA-seq model with $n_s = 2$.



(a)

(b)

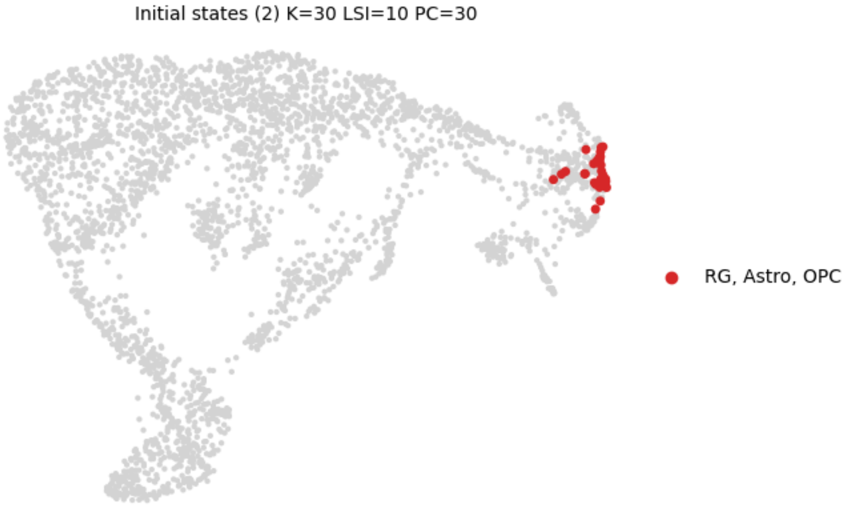
Figure 4.12: Initial (a.) and terminal macrostates with cell fate probabilities (b.) for the scATAC-seq model with $n_s = 2$.



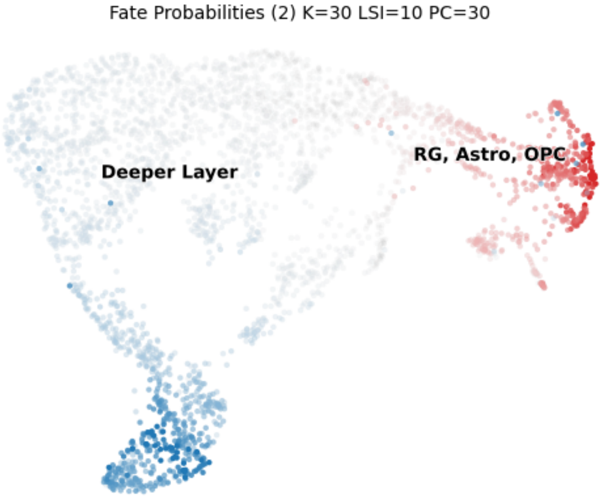
(a)

(b)

Figure 4.13: Initial (a.) and terminal macrostates with cell fate probabilities (b.) for the Multiomics+scVELO model with $n_s = 2$.



(a)



(b)

Figure 4.14: Initial (a.) and terminal macrostates with cell fate probabilities (b.) for the Multiomics+MultiVelo model with $n_s = 2$.

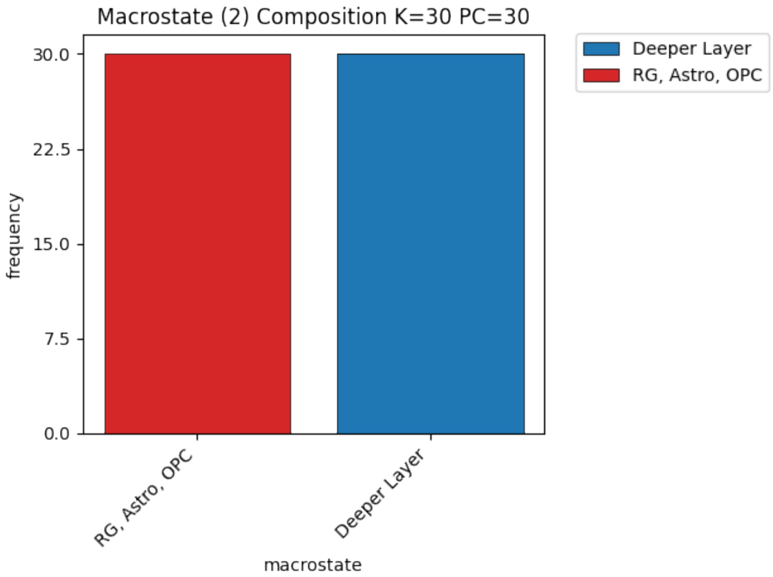


Figure 4.15: Macrostate composition for $n_s = 2$ in the scRNA-seq model.

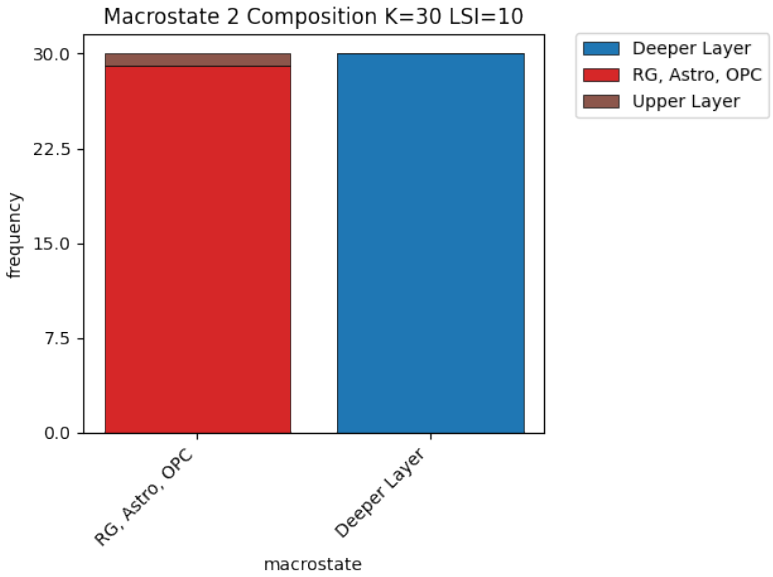


Figure 4.16: Macrostate composition for $n_s = 2$ in the scATAC-seq model.

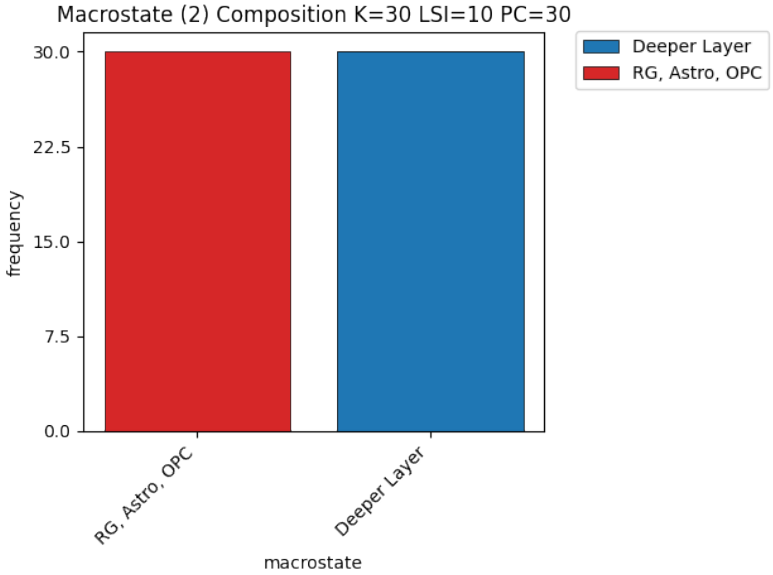


Figure 4.17: Macrostate composition for $n_s = 2$ in the Multiomics+scVELO model.

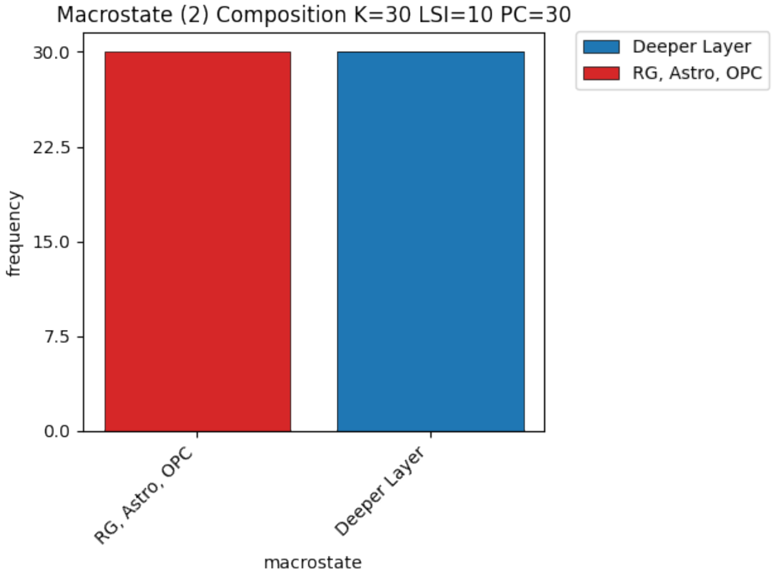
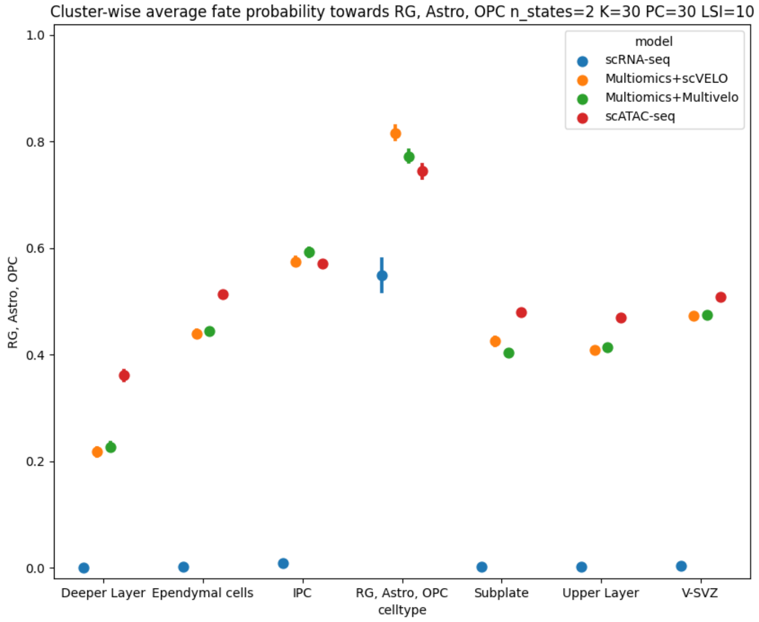
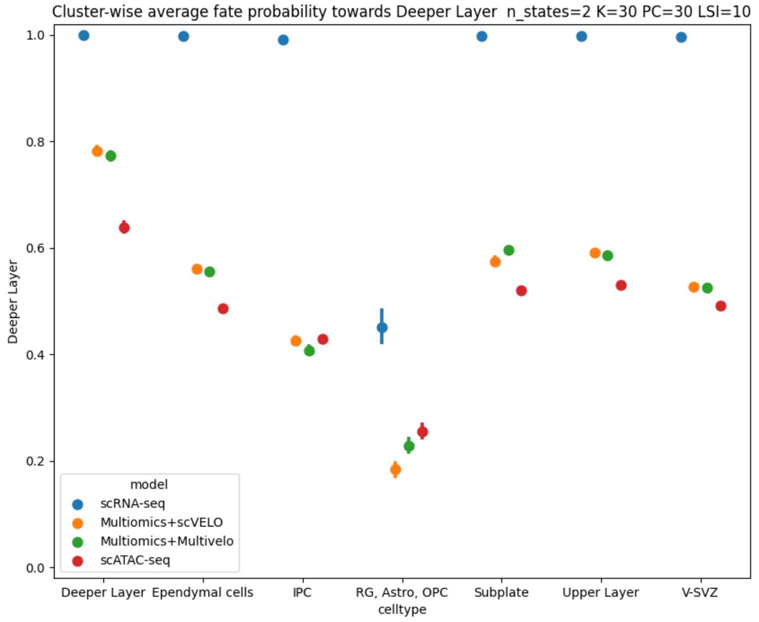


Figure 4.18: Macrostate composition for $n_s = 2$ in the Multiomics+MultiVelo model.

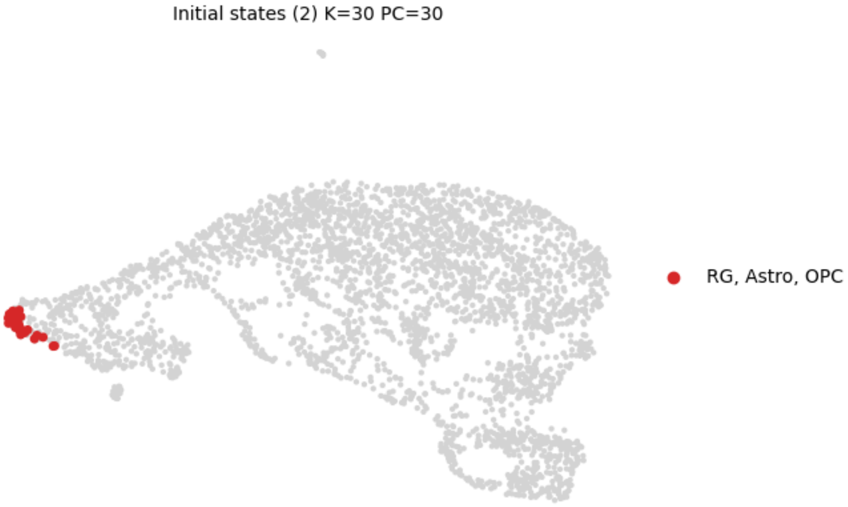


(a)

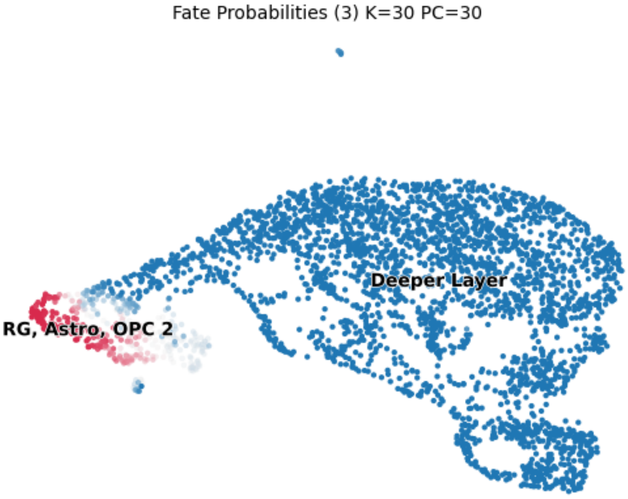


(b)

Figure 4.19: Average cell fate probabilities towards the "RG, Astro, OPC" (**a.**) and "Deeper Layer" (**b.**) terminal states in the $n_s = 2$ scenario.

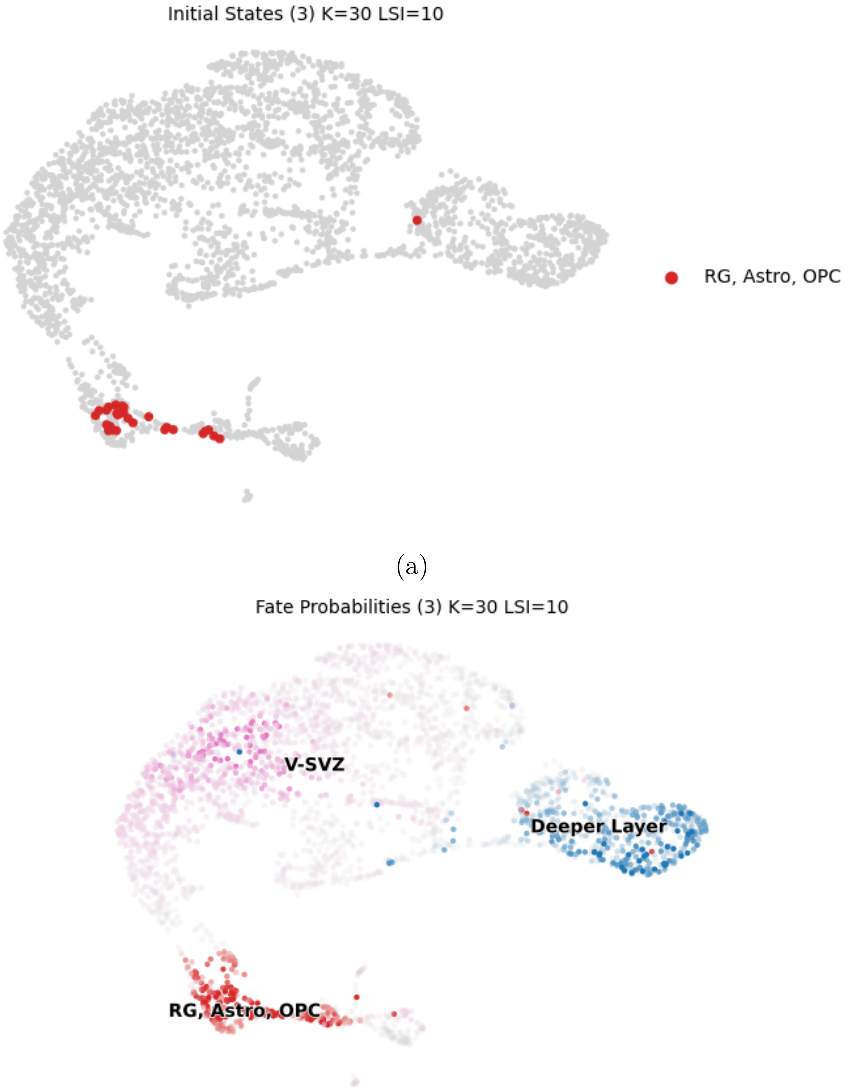


(a)



(b)

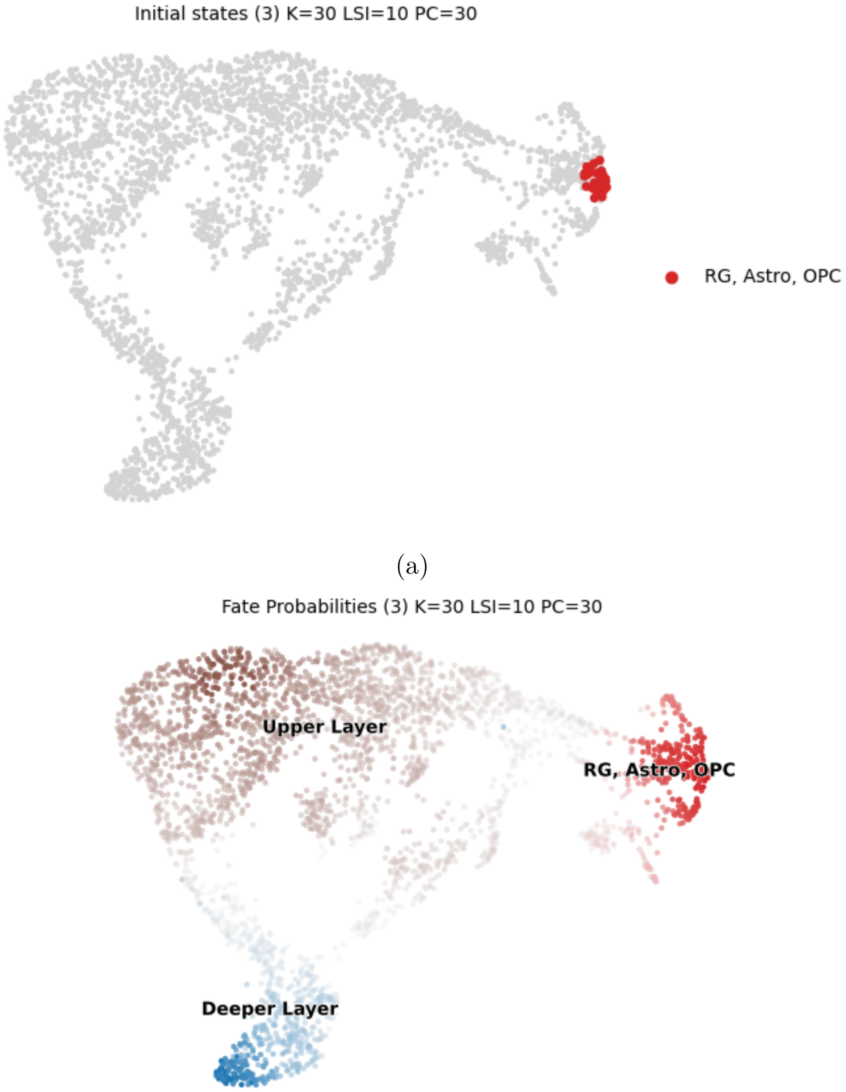
Figure 4.20: Initial (a.) and terminal macrostates with cell fate probabilities (b.) for the scRNA-seq model with $n_s = 3$.



(a)

(b)

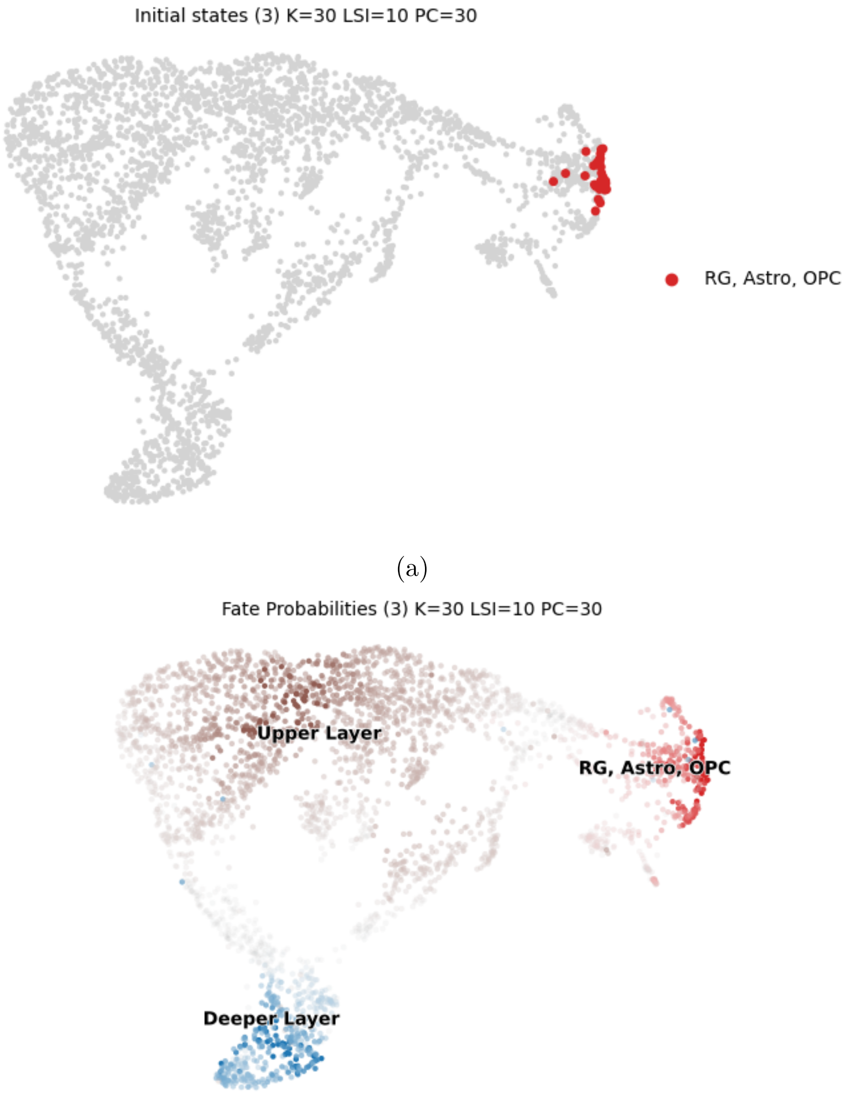
Figure 4.21: Initial (a.) and terminal macrostates with cell fate probabilities (b.) for the scATAC-seq model with $n_s = 3$.



(a)

(b)

Figure 4.22: Initial (a.) and terminal macrostates with cell fate probabilities (b.) for the Multiomics+scVELO model with $n_s = 3$.



(a)

(b)

Figure 4.23: Initial (a.) and terminal macrostates with cell fate probabilities (b.) for the Multiomics+MultiVelo model with $n_s = 3$.

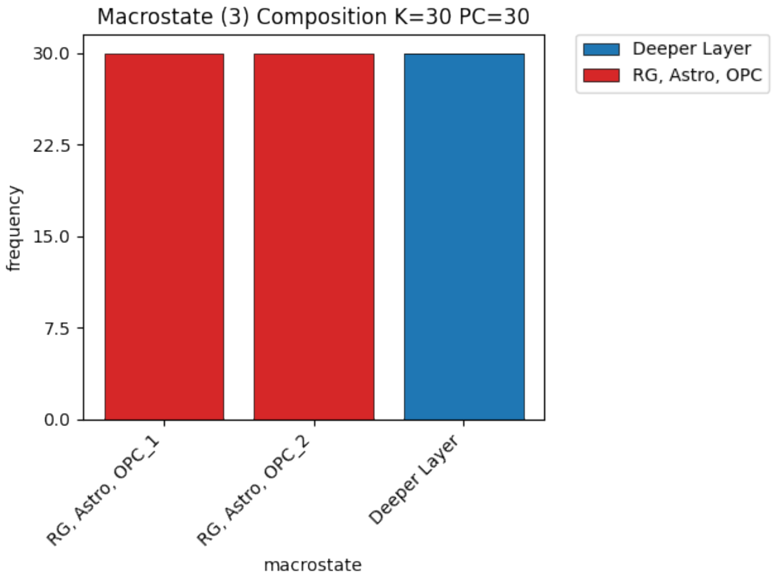


Figure 4.24: Macrostate composition for $n_s = 3$ in the scRNA-seq model.

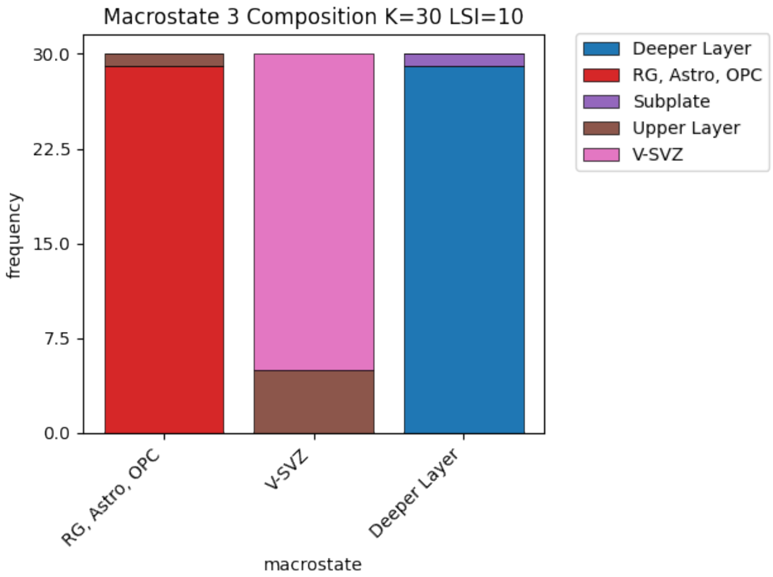


Figure 4.25: Macrostate composition for $n_s = 3$ in the scATAC-seq model.

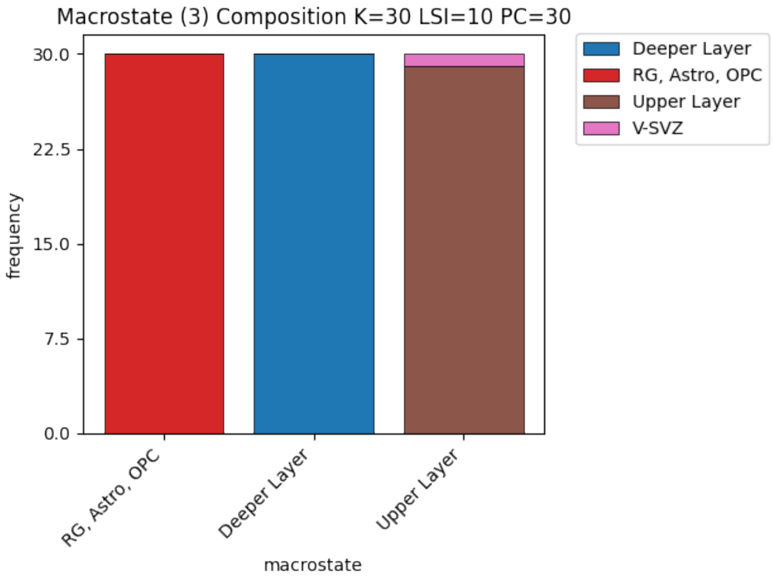


Figure 4.26: Macrostate composition for $n_s = 3$ in the Multiomics+scVELO model.

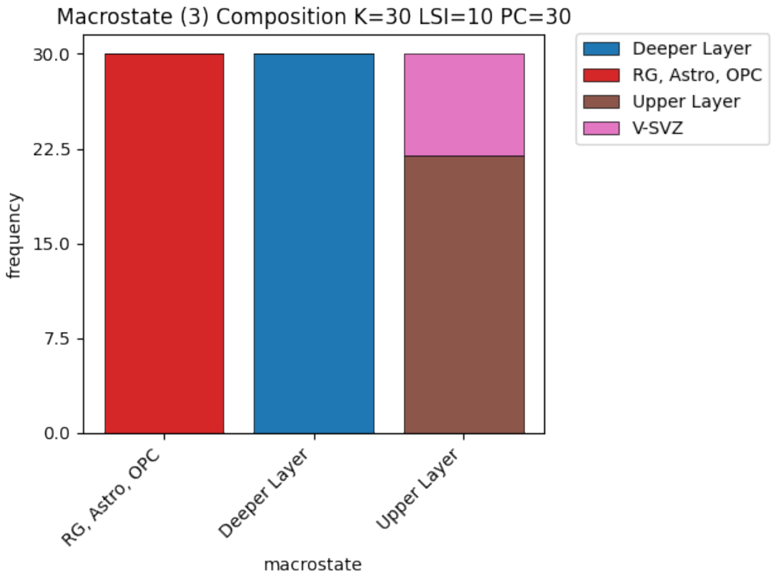
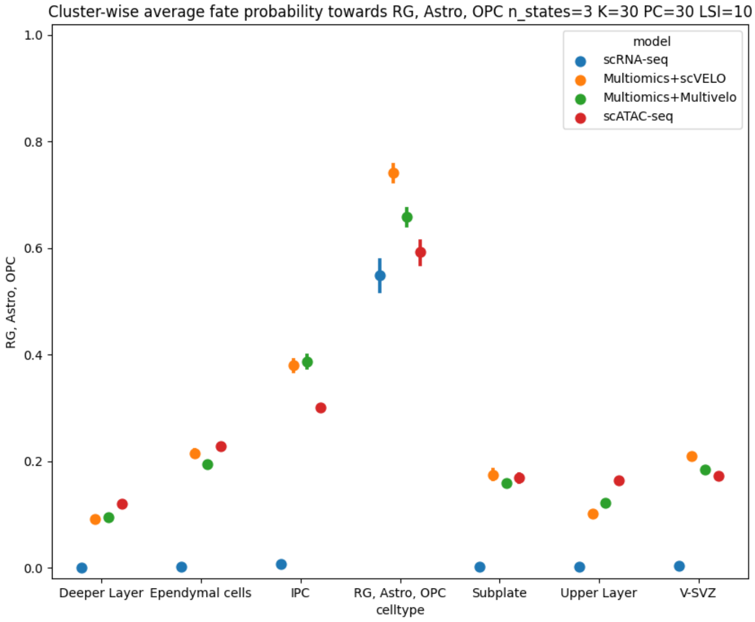
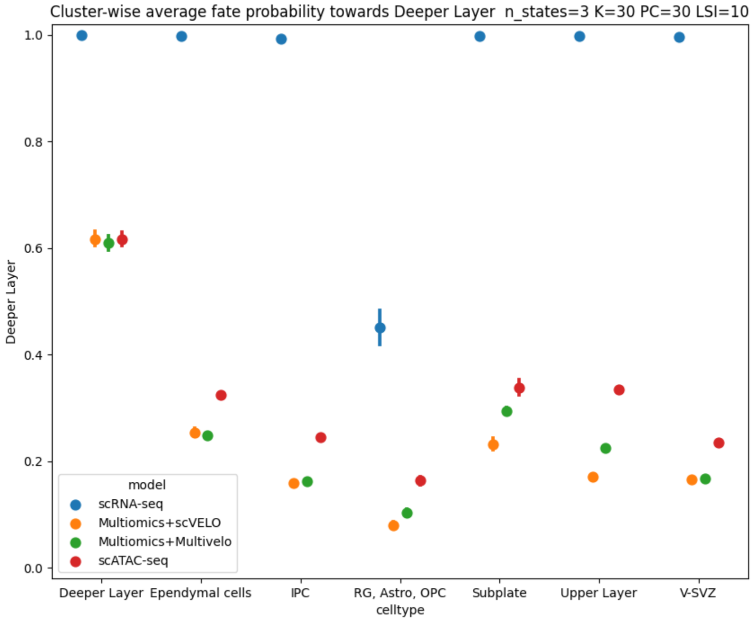


Figure 4.27: Macrostate composition for $n_s = 3$ in the Multiomics+MultiVelo model.



(a)



(b)

Figure 4.28: Average cell fate probabilities towards the "RG, Astro, OPC" (**a.**) and "Deeper Layer" (**b.**) terminal states in the $n_s = 3$ scenario.

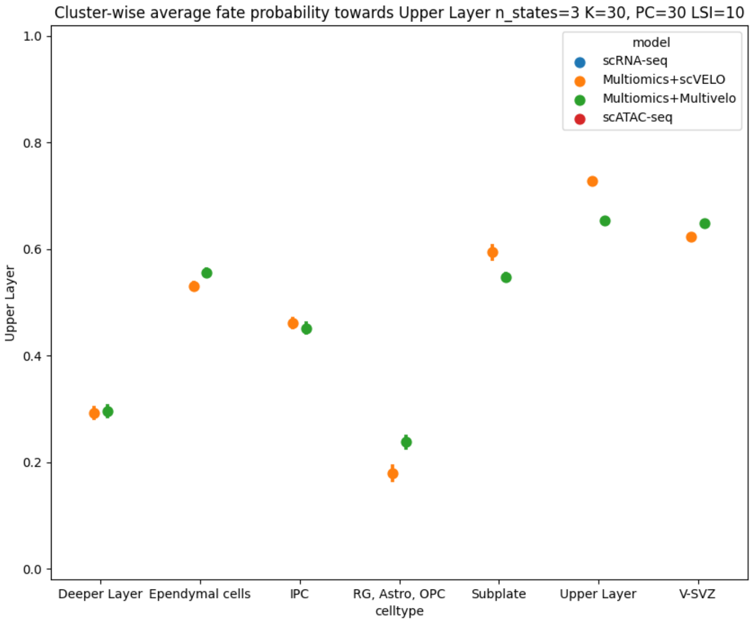
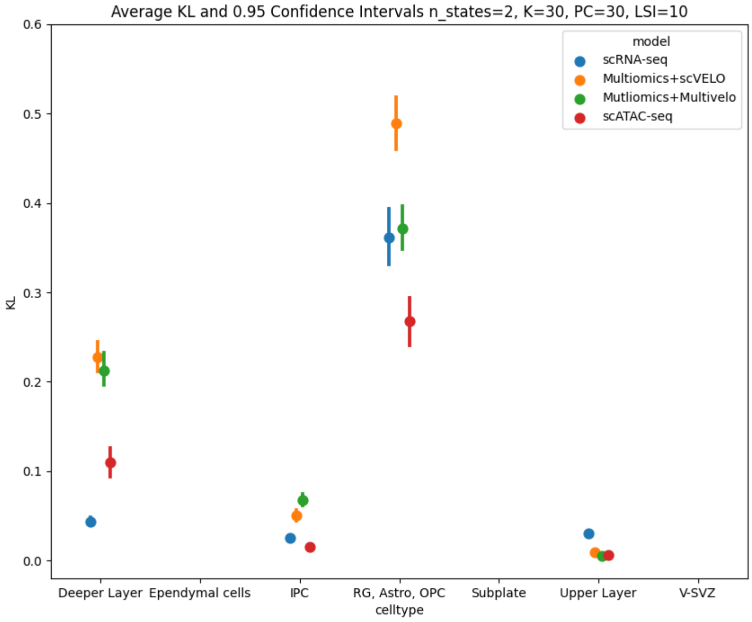
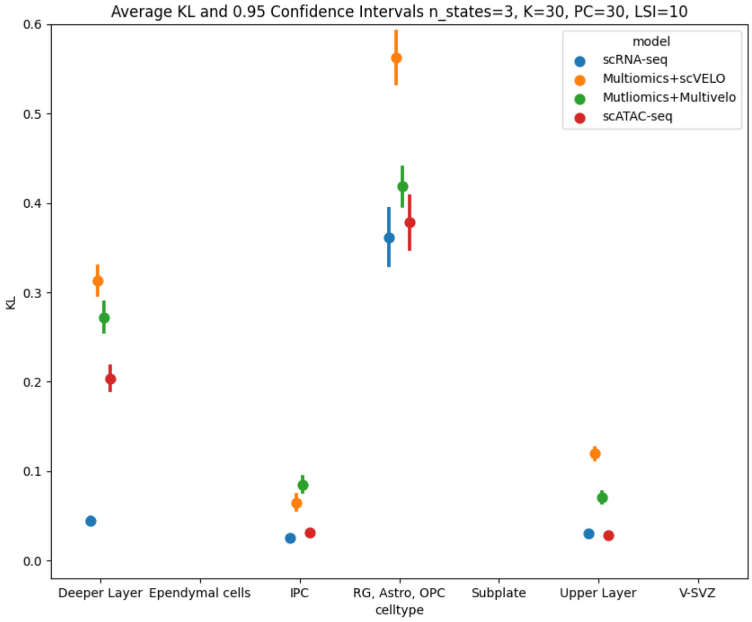


Figure 4.29: Average cell fate probabilities towards the "Upper Layer" terminal states in the $n_s = 3$ scenario.

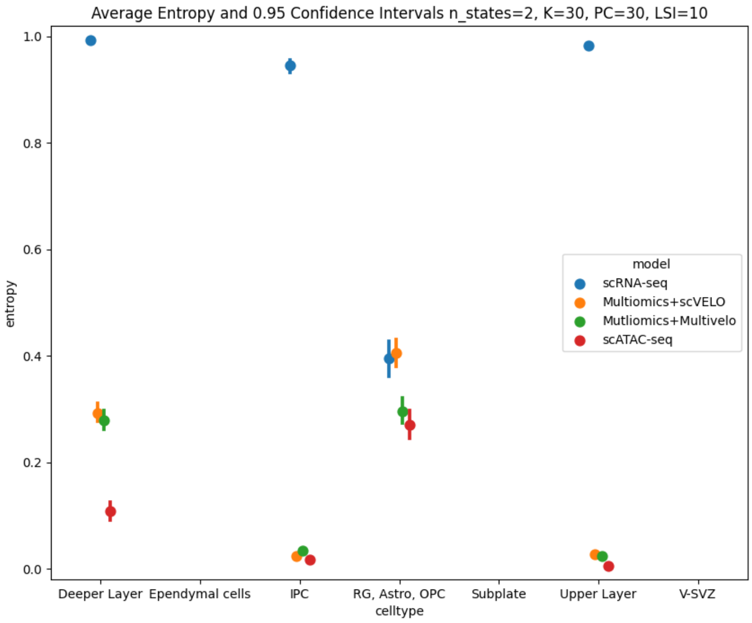


(a)

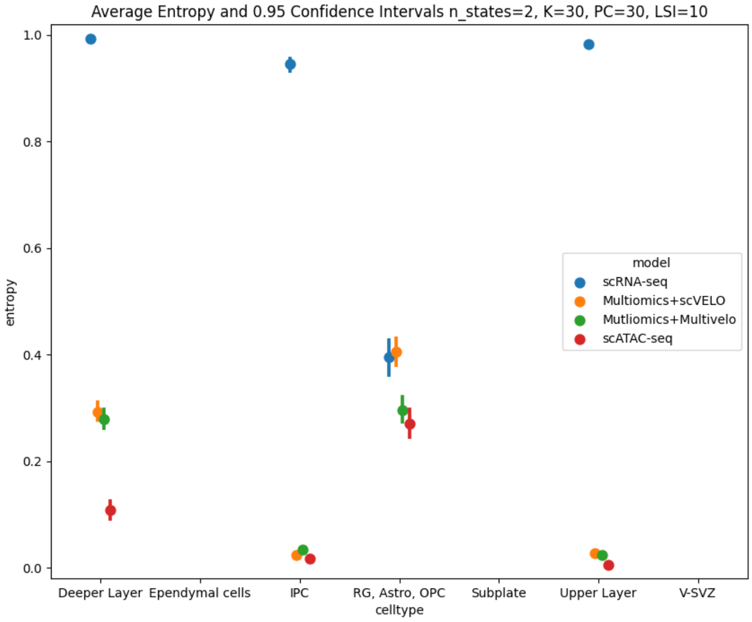


(b)

Figure 4.30: KL-divergence for $n_s = 2$ (**a.**) and $n_s = 3$ (**b.**)



(a)



(b)

Figure 4.31: Entropy for $n_s = 2$ (**a.**) and $n_s = 3$ (**b.**)

Chapter 5

Conclusions

The primary goal of this project is to understand whether the integration of epigenomic data can lead to better identification of developmental lineages compared to using transcriptomic data alone. To this end, this work presents scVEMO, a comprehensive pipeline that reconstructs developmental lineages from multimodal data. ScVEMO constructs transition matrices describing the system's evolution over time using various modeling approaches, including scATAC-seq data combined with RNA-velocity, as well as two multimodal frameworks integrating both epigenomic and transcriptomic information. The pipeline is validated on the Embryonic E18 Mouse Brain (5K) dataset from 10X Genomics, with the expectation that the system would develop from radial glia (RG) cells into two distinct lineages: one representing gliogenesis products (astrocytes and oligodendrocytes) and the other representing the neuronal developmental lineage culminating in the six cortical layers.

In the first stage of the analysis, scVEMO simulates the system at the granular cell-state level, where individual barcodes represent cellular states and the connections between them reflect cell-state transitions. The goal is to understand whether the integration of scRNA-seq, scATAC-seq, and RNA-velocity can capture the developmental continuum, where cells traverse a spectrum of intermediate states during differentiation. However, our models are unable to recover the system's development at this granular cell-state level. This suggests that the correlation of RNA-velocity with epigenomic data alone does not improve the computation of transition probabilities compared to the transcriptomics-only-based approaches.

We then investigate the models' capacity to identify the developmental lineages, including the initial and terminal states. For this purpose, we employ the CellRank framework and its spectral clustering implementation to reduce the high-dimensional transition matrix into a set of biologically meaningful macrostates, representing coarse-grained cellular states. CellRank also computes cell-fate probabilities towards each macrostate. Our analysis focuses on two key aspects:

- The identified macrostates correspond to terminally differentiated cell clusters?
- The computed fate probabilities are meaningful, i.e., whether the different cell clusters differentiate towards the correct lineages?

We expect our models to identify at least one terminal macrostate in the neuronal lineage. The multiomics models perform on par with the transcriptomics-only approach in recovering a macrostate associated with cells from the deeper cortical layers, astrocytes, and OPCs. However, with an increased number of macrostates, the multiomics models outperform the scRNA-seq-only approach by also recovering an additional terminal state composed of cells from the upper cortical layers. This aligns with the biological understanding that the deeper layers V-VI form before the upper layers II-IV during corticogenesis.

The multilineage potential analysis further demonstrates the reliability of the multimodal approaches in computing fate probabilities towards the identified terminal states. These models provide higher average lineage commitment for the differentiated clusters compared to the scRNA-seq-only approach. Additionally, the multimodal models yield slightly higher multilineage potential values for the intermediate "IPC" progenitor cluster, however, within the investigated system such progenitor cells are limited to the neuronal lineage and cannot contribute to other lineages, such as glia.

The insights gained from this project suggest that the integration of epigenomic data, when combined with transcriptomic information and RNA-velocity, can enhance the identification of biologically relevant macrostates and improve the computation of cell-fate probabilities during developmental lineage reconstruction. However, the limitations observed in the granular cell-state level simulations indicate that further methodological advancements may be necessary to fully harness the potential of multimodal data integration for this purpose.

In this project, we primarily focused on the role of promoter peaks in the integration of epigenomic and transcriptomic data for developmental lineage reconstruction. While promoter regions play a crucial role in gene regulation, gene expression is ultimately governed by a complex interplay of various regulatory elements beyond just the promoter. The regulation of gene expression involves the coordinated activity of diverse genomic features, including enhancers, silencers, insulators, and other distal regulatory sequences, in addition to the proximal promoter regions. These regulatory elements can act in concert to fine-tune the spatiotemporal patterns of gene expression during cellular differentiation and development. By primarily considering promoter peaks in our models, we may have overlooked the potential contributions of these other regulatory elements in shaping the transcriptional landscape and the underlying lineage dynamics. The inclusion of a broader range of epigenomic features, beyond just promoter regions, could lead to a more comprehensive representation of the gene regulatory mechanisms governing cellular differentiation.

Finally, the CellRank framework employed in our analysis utilizes a deterministic approach to reduce the high-dimensional transition matrix into a set of biologically meaningful macrostates. This deterministic implementation provides a straightforward and interpretable means of identifying the key developmental lineages and their corresponding terminal states. However, the stochastic implementation of CellRank could provide additional insights and potentially enhance the robustness of our lineage reconstruction results, as it could help capture the probabilistic nature of cellular transitions and account for the heterogeneity within the system. The exploration of alternative lineage branching scenarios could provide a more comprehensive assessment of the inherent flexibility and plasticity of the cellular differentiation process. By leveraging the stochastic formulation, the lineage reconstruction process can account for the noise and variability present in the RNA-velocity estimates. As we continue to refine and expand the multimodal lineage reconstruction methodologies, the incorporation of stochastic RNA-velocity modeling should be a key area of focus.

Bibliography

[1] B. Alberts, A. Johnson, J. Lewis, and other, *Molecular Biology of the Cell*. New York: Garland Science, 4th ed., 2002. Available From: https://www.ncbi.nlm.nih.gov/books/NBK21054/.

[2] K. Vandereyken, A. Sifrim, B. Thienpont, and V. T., "Methods and applications for single-cell and spatial multi-omics," *Nature Reviews Genetics*, vol. 24, pp. 494–515, Aug. 2023. doi:10.1038/s41576-023-00580-2.

[3] A. Feinberg and A. Levchenko, "Epigenetics as a mediator of plasticity in cancer," *Science*, vol. 379, Feb 2023. doi:10.1126/science.aaw3835.

[4] T. Cooper, L. Wan, and G. Dreyfuss, "RNA and disease," *Cell*, vol. 136, pp. 777–93, Feb 2009. doi:10.1016/j.cell.2009.02.011.

[5] U. Langeh and S. Singh, "Targeting S100B Protein as a Surrogate Biomarker and its Role in Various Neurological Disorders," *Current Neuropharmacology*, vol. 19, no. 2, pp. 265–277, 2021. doi:10.2174/1570159X18666200729100427.

[6] E. Kassi, P. Pervanidou, G. Kaltsas, and G. Chrosous, "Metabolic syndrome: definitions and controversies," *BMC Medicine*, May 2011. doi:10.1186/1741-7015-9-48.

[7] L. Wang, F. Wang, and M. Gershwin, "Human autoimmune diseases: a comprehensive update," *Jornal of Internal Medicine*, vol. 278, pp. 369–95, Oct 2015. doi:10.1111/joim.12395.

[8] C. Pareek, R. Smoczynski, and A. Tretyn, "Sequencing technologies and genome sequencing," *Applied Genetics*, vol. 52, pp. 413–35, Nov. 2011. doi:10.1007/s13353-011-0057-x.

[9] R. Kishore, M. Cowley, and R. Davis, "Next-Generation Sequencing and Emerging Technologies," *Seminars in Thrombosis and Hemostasis*, vol. 47, pp. 661–673, Oct 2019. doi:10.1055/s-0039-1688446.

[10] B. Hwang, J. Lee, and D. Bang, "Single-cell RNA sequencing technologies and bioinformatics pipelines," *Experimental & molecular medicine*, vol. 50, no. 8, pp. 1–14, 2018. doi:10.1038/s12276-018-0071-8.

[11] B. Seungbyn and L. Insuk, "Single-cell ATAC sequencing analysis: From data preprocessing to hypothesis generation," *Computational and Structural Biotechnology Journal*, vol. 18, pp. 1429–1439, 2020. url:https://www.sciencedirect.com/science/article/pii/S2001037020303019.

[12] G. Chen, B. Ning, and T. Shi, "Single-Cell RNA-Seq Technologies and Related Computational Data Analysis," *Frontiers in Genetics*, vol. 10, 2019. https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2019.00317.

[13] T. Stuart, A. Srivastava, S. Madad, C. Lareau, and R. Satija, "Single-cell chromatin state analysis with Signac," *Nature Methods*, 2021. doi:10.1038/s41592-021-01282-5.

[14] Y. Hao, T. Stuart, *et al.*, "Dictionary learning for integrative, multimodal and scalable single-cell analysis," *Nature Biotechnology*, 2023. doi:10.1038/s41587-023-01767-y.

[15] F. Wolf, P. Angerer, and F. Theis, "SCANPY: large-scale single-cell gene expression data analysis," *Genome Biology*, Feb 2018. doi:10.1186/s13059-017-1382-0.

[16] M. Lange, V. Bergen, M. Klein, *et al.*, "Cellrank for directed single-cell fate mapping," *Nature Methods*, vol. 19, pp. 159–170, 2022.

[17] M. Setty, V. Kiseliovas, J. Levine, *et al.*, "Characterization of cell fate probabilities in single-cell data with Palantir," *Nat. Biotech.*, vol. 37, pp. 451–460, 2019. doi:10.1038/s41587-019-0068-4.

[18] C. Baron and A. van Oudenaarden, "Unravelling cellular relationships during development and regeneration using genetic lineage tracing," *Nature Review Molecular Cell Biology*, vol. 20, pp. 753–765, 2019. url: https://doi.org/10.1038/s41580-019-0186-3.

[19] D. Wagner and A. Klein, "Lineage tracing meets single-cell omics: opportunities and challenges," *Nature Review Genetics*, vol. 21, pp. 410–427, 2020. doi:10.1038/s41576-020-0223-2.

[20] V. Bergen, M. Lange, S. Peidli, F. Wolf, and F. Theis, "Generalizing RNA velocity to transient cell states through dynamical modeling," *Nature Biotechnology*, vol. 38, pp. 1408–1414, Aug 2020. doi:10.1038/s41587-020-0591-3.

[21] C. Li, M. Virgilio, K. Collins, *et al.*, "Multi-omic single-cell velocity models epigenom-transcriptome interactions and improves cell fate prediction," *Nature Biotechnology*, vol. 41, pp. 387–398, 2023. doi:10.1038/s41587-022-01476-y.

[22] H. Chen, L. Albergante, and other, "Single-cell trajectories reconstruction, exploration and mapping of omics data with STREAM," *Nature Communications*, vol. 10, 2019. doi:10.1038/s41467-019-09670-4.

[23] A. Lynch, C. Theodoris, H. Long, and other, "MIRA: joint regulatory modeling of multimodal expression and chromatin accessibility in single cells," *Nature Methods*, vol. 19, pp. 1097–1108, 2022. doi:10.1038/s41592-022-01595-z.

[24] Fresh Embryonic E18 Mouse Brain (5k) (v1), Single Cell Multiome ATAC + Gene Expression Dataset by Cell Ranger ARC 2.0.0, 10X Genomics, (2021, May 3).

[25] A. Field and K. Adelman, "Evaluating Enhancer Function and Transcription," *Annual Review of Biochemistry*, vol. 89, pp. 213–234, Jun 2020. doi:10.1146/annurev-biochem-011420-095916.

[26] P. Farnham, "Insights from genomic profiling of transcription factors," *Nature Reviews Genetics*, vol. 10, no. 9, pp. 605–616, 2009. doi:10.1038/nrg2636.

[27] P. Weidemüller, M. Kholmatov, E. Petsalaki, and J. Zaugg, "Transcription factors: Bridge between cell signaling and gene regulation," *Proteomics*, vol. 21, pp. 23–24, Dec 2021. doi:10.1002/pmic.202000034.

[28] K. Struhl, "Fundamentally different logic of gene regulation in eukaryotes and prokaryotes," *Cell*, vol. 98, pp. 1–4, Jul 1999. doi:10.1016/S0092-8674(00)80599-1.

[29] S. Li, F. Garrett-Bakelman, S. Chung, *et al.*, “Distinct evolution and dynamics of epigenetic and genetic heterogeneity in acute myeloid leukemia,” *Nature Medicine*, vol. 22, pp. 792–9, Jul 2016. doi:10.1038/nm.4125.

[30] L. Shuxiang, P. Yunhui, and A. Panchenko, “DNA methylation: Precise modulation of chromatin structure and dynamics,” *Current Opinion in Structural Biology*, vol. 75, p. 102430, 2022. doi:10.1016/j.sbi.2022.102430.

[31] C. Jiang and B. Pugh, “Nucleosome positioning and gene regulation: advances through genomics,” *Nature Reviews Genetics*, vol. 10, no. 3, pp. 161–172, 2009. doi:10.1038/nrg2522.

[32] D. Buitrago, M. Labrador, J. P. Arcon, *et al.*, “Impact of DNA methylation on 3D genome structure,” *Nature Communications*, vol. 12, no. 3243, 2021. doi:10.1038/s41467-021-23142-8.

[33] O. Morrison and J. Thakur, “Molecular Complexes at Euchromatin, Heterochromatin and Centromeric Chromatin,” *International Journal of Molecular Sciences*, vol. 22, no. 13, p. 6922, 2021. doi:10.3390/ijms22136922.

[34] T. Saldi, M. Cortazar, R. Sheridan, and D. Bentley, “Coupling of RNA Polymerase II Transcription Elongation with Pre-mRNA Splicing,” *Journal of Molecular Biology*, vol. 428, pp. 2623–2635, Jun 2016. doi:10.1016/j.jmb.2016.04.017.

[35] J. Ule and B. Blencowe, “Alternative Splicing Regulatory Networks: Functions, Mechanisms, and Evolution,” *Molecular Cell*, vol. 76, pp. 329–345, Oct. 2019. doi:10.1016/j.molcel.2019.09.017.

[36] F. Sanger, S. Nicklen, and A. Coulson, “DNA sequencing with chain-terminating inhibitors,” *Proceedings of the National Academy of Sciences*, vol. 74, no. 12, pp. 5463–5467, 1977. doi:10.1073/pnas.74.12.5463.

[37] A. Kolodziejczyk, K. Jong, V. Svensson, M. J.C., and S. Teichmann, “The Technology and Biology of Single-Cell RNA Sequencing,” *Molecular Cell*, vol. 58, pp. 610–620, 2015. doi:10.1016/j.molcel.2015.04.005.

[38] Chromium Single Cell V(D)J Reagent Kits with Feature Barcoding technology for Cell Surface Protein, Document Number CG000186 Rev A, 10x Genomics, (2019, July 25).

[39] A. Satpathy, N. Saligrama, J. Buenrostro, *et al.*, “Transcript-indexed ATAC-seq for precision immune profiling,” *Nature Medicine*, vol. 24, pp. 580–590, 2018. doi:10.1038/s41591-018-0008-8.

[40] X. Chen, R. Miragaia, K. Natarajan, *et al.*, “A rapid and robust method for single cell chromatin accessibility profiling,” *Nature Communications*, vol. 9, 2018. doi:10.1038/s41467-018-07771-0.

[41] R. M. Mulqueen, B. A. DeRosa, C. A. Thornton, *et al.*, “Improved single-cell ATAC-seq reveals chromatin dynamics of in vitro corticogenesis,” *bioRxiv*, 2019. doi:10.1101/637256.

[42] Y. Zhang, T. Liu, C. Meyer, and other, “Model-based Analysis of ChIP-Seq (MACS),” *Genome Biology*, vol. 9, 2008. doi:10.1186/gb-2008-9-9-r137.

[43] S. Chen, L. B.B, and K. Zhang, “High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell,” *Nature Biolotechnology*, vol. 37, pp. 1452–1457, Dec 2019. doi:10.1038/s41587-019-0290-0.

[44] Chromium Next GEM Single Cell Multiome ATAC + Gene Expression, Document Number CG000338 Rev F, 10X Genomics, (2022, August 26).

[45] Nuclei from Embryonic Mouse Brain for Single Cell Multiome ATAC + GEX Sequencing, Document Number GC000338 Rev D, 10X Genomics, (2022, July 13).

[46] X. Genomics, "Cell Ranger ARC 2.0.0." `https://support.10xgenomics.com/single-cell-vdj/software/overview/welcome`. (2023, July 6).

[47] L. Haghverdi and L. Ludwig, "Single-cell multi-omics and lineage tracing to dissect cell fate decision-making," *Stem Cell Reports*, vol. 18, no. 1, pp. 13–25, 2023. doi:10.1016/j.stemcr.2022.12.003.

[48] G. La Manno, R. Soldatov, and Z. A., "RNA velocity of single cells," *Nature*, vol. 560, pp. 494–498, 2018. doi:10.1038/s41586-018-0414-6.

[49] J. Sun, A. Ramos, B. Chapman, *et al.*, "Clonal dynamics of native haematopoiesis," *Nature*, vol. 514, pp. 322–327, 2014. doi:10.1038/nature13824.

[50] B. Sauer, "Inducible Gene Targeting in Mice Using the Cre/loxSystem," *Methods*, vol. 14, no. 4, pp. 381–392, 1998. doi:10.1006/meth.1998.0593.

[51] R. Coifman, S. Lafon, A. Lee, S. Zucker, and other, "Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps," *Proceedings of the National Academy of Sciences*, vol. 102, no. 21, pp. 7426–7431, 2005. doi:10.1073/pnas.0500334102.

[52] S. Röblitz and M. Weber, "Fuzzy spectral clustering by PCCA+: application to Markov state models and data classification," *Advances in Data Analysis and Classification*, vol. 7, pp. 147–179, 2013. doi:10.1007/s11634-013-0134-6.

[53] B. Schölkopf, J. Platt, and T. Hofmann, *MLLE: Modified Locally Linear Embedding Using Multiple Weights*. The MIT Press, 2007.

[54] A. Dempster, L. N.M., and D. Rubin, "Maximum Likelihood from Incomplete Data via the EM algorithm," *Journal of the Royal Statistical Society. Series (B) Methodological*, vol. 39, no. 1, pp. 1–38, 1977.

[55] J. Nelder and R. Mead, "A Simplex Method for Function Minimization," *Computer Journal*, vol. 7, pp. 308–313, 1965. url: https://api.semanticscholar.org/CorpusID:2208295.

[56] A. Bastidas-Ponce, S. Tritschler, L. Dony, *et al.*, "Comprehensive single cell mRNA profiling reveals a detailed roadmap for pancreatic endocrinogenesis," *Development*, vol. 146, 06 2019.

[57] F. Fagnani, G. Como, Lecture notes on Network Dynamics, 2022, Politecnico di Torino.

[58] D. Blondel, J. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, oct 2008. doi:10.1088/1742-5468/2008/10/P10008.

[59] V. A. Traag, L. Waltman, and N. J. Van Eck, "From Louvain to Leiden: guaranteeing well-connected communities," *Scientific Reports*, vol. 9, no. 1, 2019. doi:10.1038/s41598-019-41695-z.

[60] D. S. Mackey, N. Mackey, C. Mehl, and V. Mehrmann, "Structured Polynomial Eigenvalue Problems: Good Vibrations from Good Linearizations," *SIAM Journal on Matrix Analysis and Applications*, vol. 28, no. 4, pp. 1029–1051, 2006.

doi:10.1137/050628362.

[61] Edgar *et al.*, “LifeMap Discovery: The Embryonic Development, Stem Cells, and Regenerative Medicine Research Portal,” *PLoS ONE*, vol. 8, no. 7, 2013. doi:10.1371/journal.pone.0066629.

[62] H. Hochgerner, A. Zeisel, P. Lönnerberg, and S. Linnarsson, “Conserved properties of dentate gyrus neurogenesis across postnatal development revealed by single-cell RNA sequencing,” *Nature Neuroscience*, vol. 21, no. 2, pp. 290–299, 2018. doi:10.1038/s41593-017-0056-2.

[63] L. Loo, J. Simon, L. Xing, and other, “Single-cell transcriptomic analysis of mouse neocortical development,” *Nature Communications*, vol. 10, no. 1, p. 134, 2019. doi:10.1038/s41467-018-08079-9.

[64] G. La Manno, K. Siletti, A. Furlan, and other, “Molecular architecture of the developing mouse brain,” *Nature*, vol. 596, no. 7870, pp. 92–96, 2021. doi:10.1038/s41586-021-03775-x.

[65] L. Dalcin, R. Paz, P. Kler, and A. Cosimo, “Parallel distributed computing using Python,” *Advances in Water Resources*, vol. 34, no. 9, pp. 1124–1139, 2011. doi:10.1016/j.advwatres.2011.04.013.

[66] Chan and V. D. Vorst, *Approximate and Incomplete Factorizations*, pp. 167–2022. Dordrecht: Springer Netherlands, 1997. doi:10.1007/978-94-011-5412-3_6.

[67] T. Dupont, R. Kendall, and H. Rachford, Jr., “An Approximate Factorization Procedure for Solving Self-Adjoint Elliptic Difference Equations,” *SIAM Journal on Numerical Analysis*, vol. 5, no. 3, pp. 559–573, 1968. doi:10.1137/0705045.

[68] K. Mustapha, “An implicit finite-difference time-stepping method for a sub-diffusion equation, with spatial discretization by finite elements,” *IMA Journal of Numerical Analysis*, vol. 30, 2011. doi:10.1093/imanum/drp057.

[69] NetworkX- Network analysis in Python. 2014-2024, url: https://networkx.org/.

[70] L. Velten, S. F. Haas, S. Raffel, and other, “Human haematopoietic stem cell lineage commitment is a continuous process,” *Nature Cell Biology*, vol. 19, pp. 271–281, Apr 2017. doi:10.1038/ncb3493.

Appendix A

Table A.1: Multiomics+scVELO: ANOVA results for KL-divergence on "Deeper Layer" cluster.

	$n_s = 2$				$n_s = 3$			
	Res.	***DF***	***F-score***	***P-val***	***Res.***	***DF***	***F-score***	***P-val***
Additive	*78362*				*78362*			
Interactions	*78250*	*112*	*46.875*	***	*78250*	*112*	*3.685*	***

Table A.2: Multiomics+scVELO: ANOVA results for KL-divergence on "IPC" cluster.

	$n_s = 2$				$n_s = 3$			
	Res.	***DF***	***F-score***	***P-val***	***Res.***	***DF***	***F-score***	***P-val***
Additive	*22862*				*22862*			
Interactions	*22750*	*112*	*1.8683*	***	*22750*	*112*	*1.9114*	***

Table A.3: Multiomics+scVELO: ANOVA results for KL-divergence on "RG, Astro, OPC" cluster.

	$n_s = 2$				$n_s = 3$			
	Res.	***DF***	***F-score***	***P-val***	***Res.***	***DF***	***F-score***	***P-val***
Additive	54862				54862			
Interactions	54750	112	16.325	***	54750	112	11.511	***

Table A.4: Multiomics+scVELO: ANOVA results for KL-divergence on "Upper Layer" cluster.

	$n_s = 2$				$n_s = 3$			
	Res.	***DF***	***F-score***	***P-val***	***Res.***	***DF***	***F-score***	***P-val***
Additive	119987				119987			
Interactions	119875	112	16.325	***	119875	112	215.78	***

Table A.5: Multiomics+MultiVelo: ANOVA results for KL-divergence on "Upper Layer" cluster.

	$n_s = 2$				$n_s = 3$			
	Res.	***DF***	***F-score***	***P-val***	***Res.***	***DF***	***F-score***	***P-val***
Additive	116147				116147			
Interactions	116039	108	3.0831	***	116039	108	2.7219	***

Table A.6: Multiomics+MultiVelo: ANOVA results for KL-divergence on "Deeper Layer" cluster.

	$n_s = 2$				$n_s = 3$			
	Res.	***DF***	***F-score***	***P-val***	***Res.***	***DF***	***F-score***	***P-val***
Additive	75854				75854			
Interactions	75746	108	5.8101	***	75746	108	9.39	***

Table A.7: Multiomics+MultiVelo: ANOVA results for KL-divergence on "IPC" cluster.

	$n_s = 2$				$n_s = 3$			
	Res.	***DF***	***F-score***	***P-val***	***Res.***	***DF***	***F-score***	***P-val***
Additive	22130				22130			
Interactions	22022	108	1.296	*	22022	108	1.569	***

Table A.8: Multiomics+MultiVelo: ANOVA results for KL-divergence on "RG, Astro, OPC" cluster.

	$n_s = 2$				$n_s = 3$			
	Res.	***DF***	***F-score***	***P-val***	***Res.***	***DF***	***F-score***	***P-val***
Additive	53106				53106			
Interactions	52998	108	7.845	***	52998	108	3.748	***

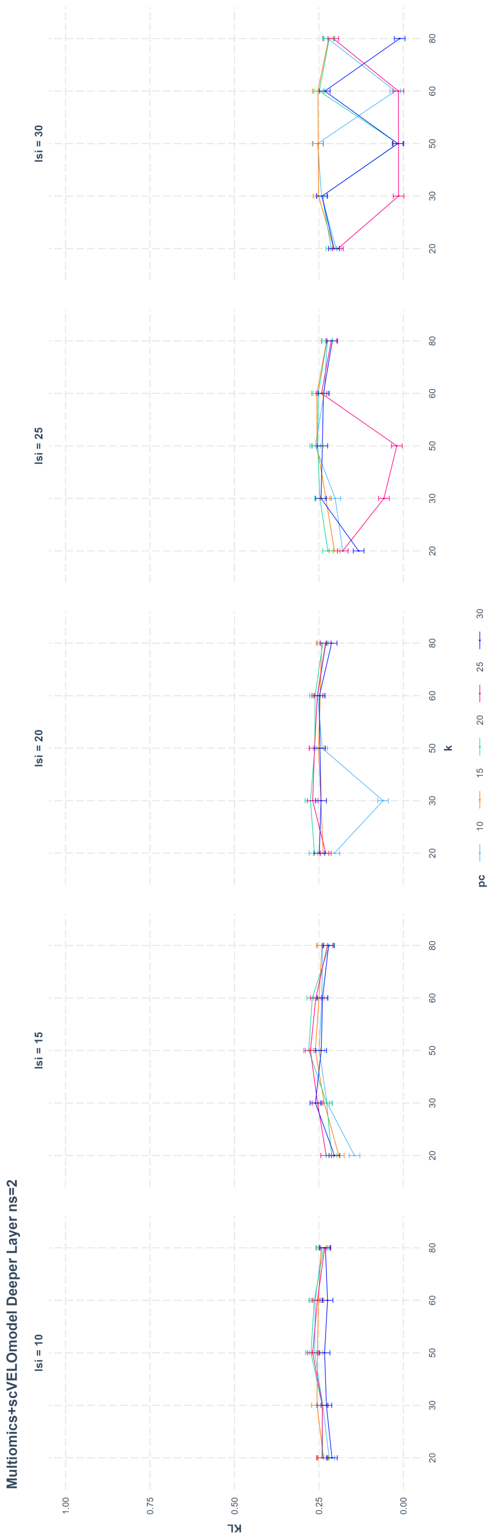


Figure A.1: Linear model results for the Multiomics+scVELO model with $n_s = 2$ and Deeper Layer cluster

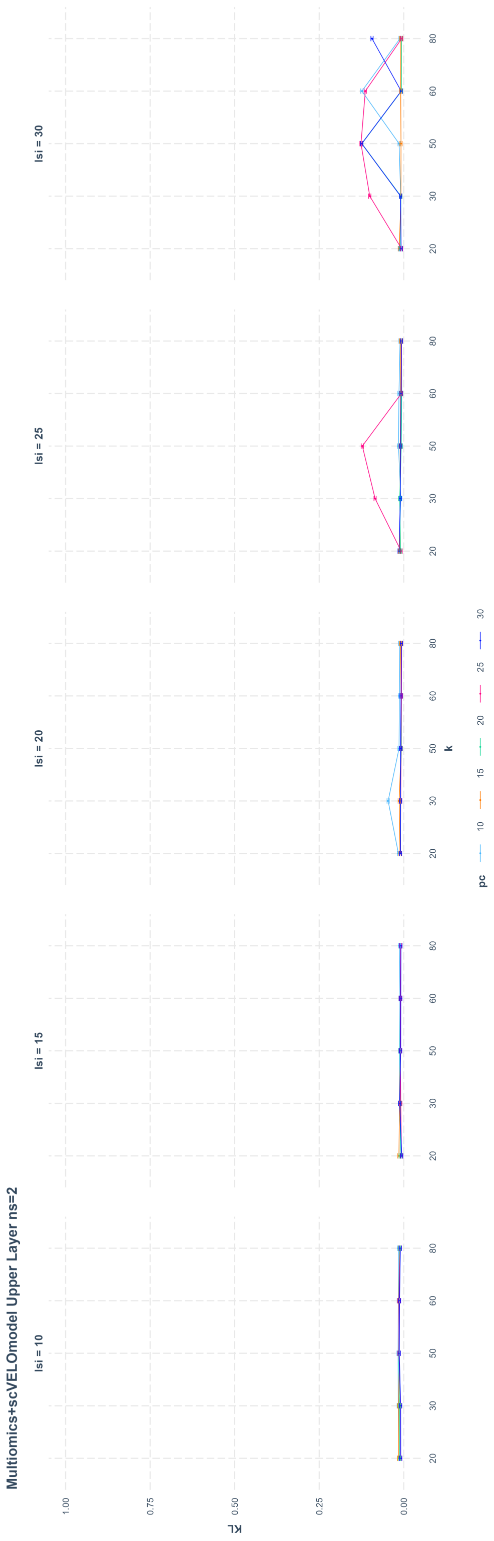


Figure A.2: Linear model results for the Multiomics+scVELO model with $n_s = 2$ and Upper Layer cluster

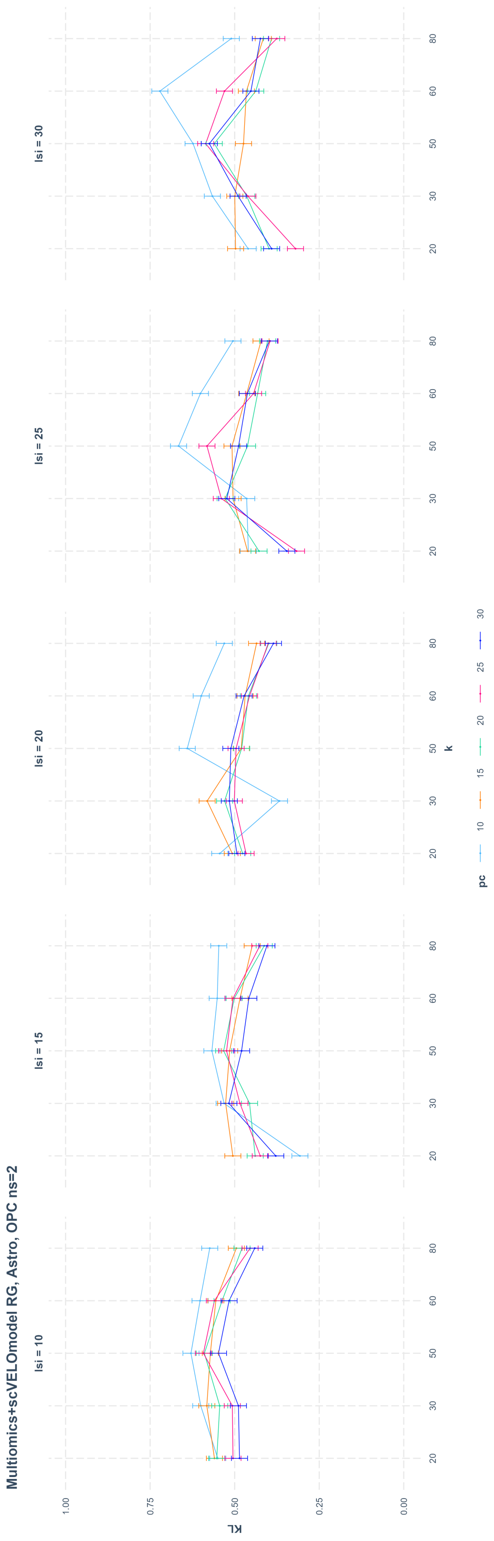


Figure A.3: Linear model results for the Multiomics+scVELO model with $n_s = 2$ and RG, Astro, OPC cluster

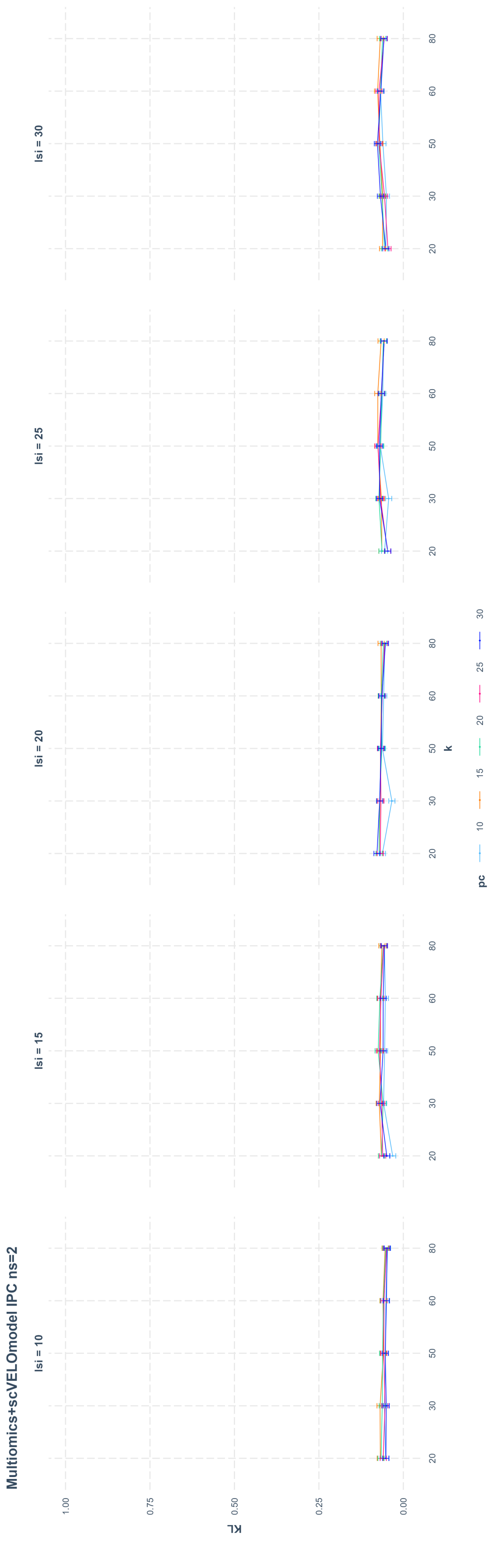


Figure A.4: Linear model results for the Multiomics+scVELO model with $n_s = 2$ and IPC cluster

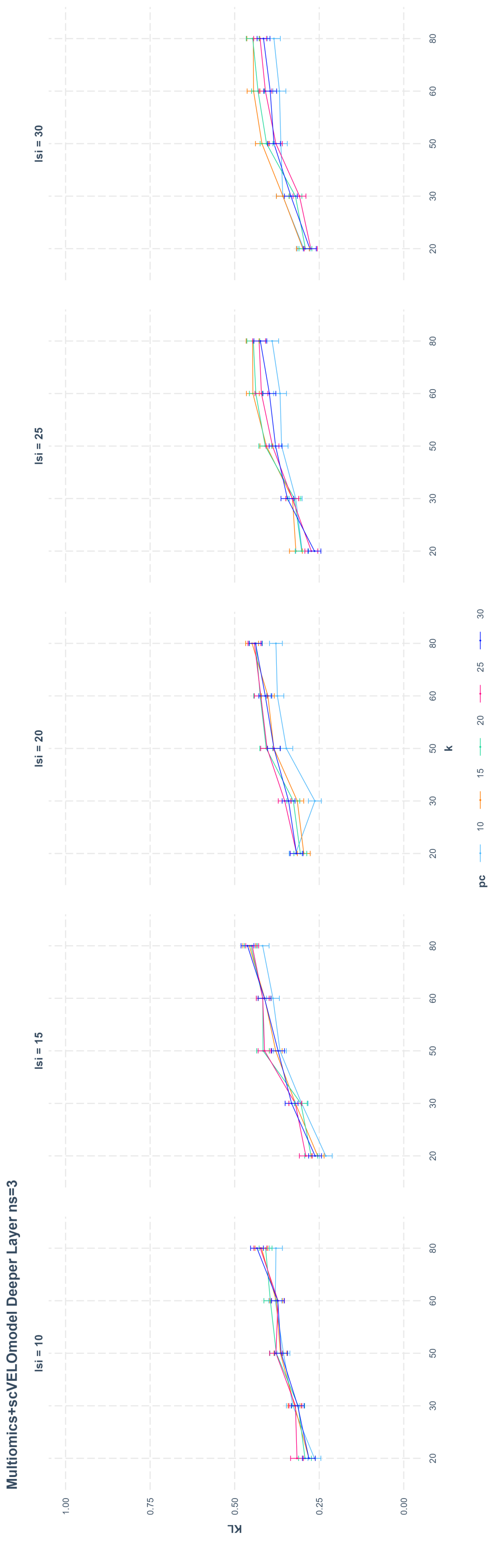


Figure A.5: Linear model results for the Multiomics+scVELO model with $n_s = 3$ and Deeper Layer cluster

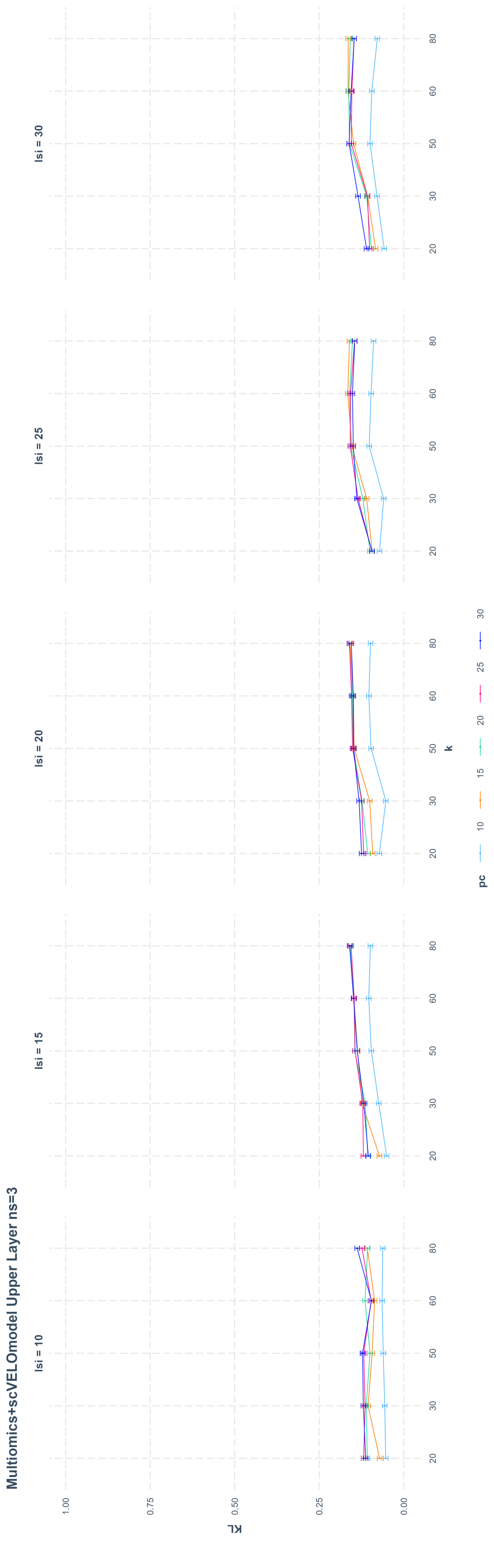


Figure A.6: Linear model results for the Multiomics+scVELO model with $n_s = 3$ and Upper Layer cluster

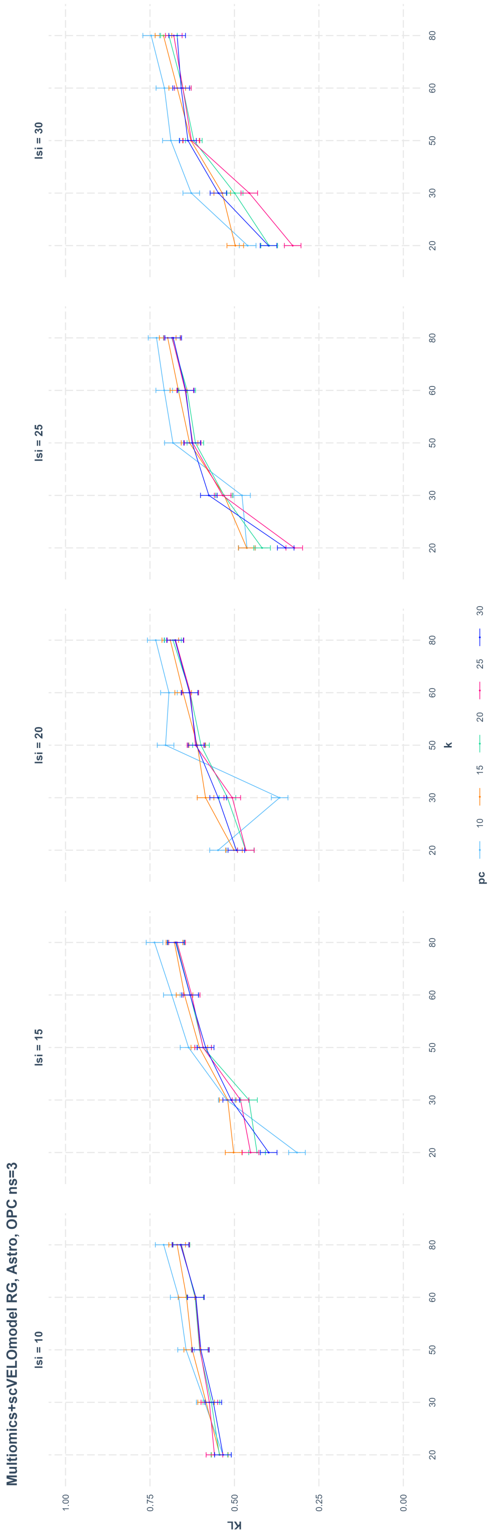


Figure A.7: Linear model results for the Multiomics+scVELO model with $n_s = 3$ and RG, Astro, OPC cluster

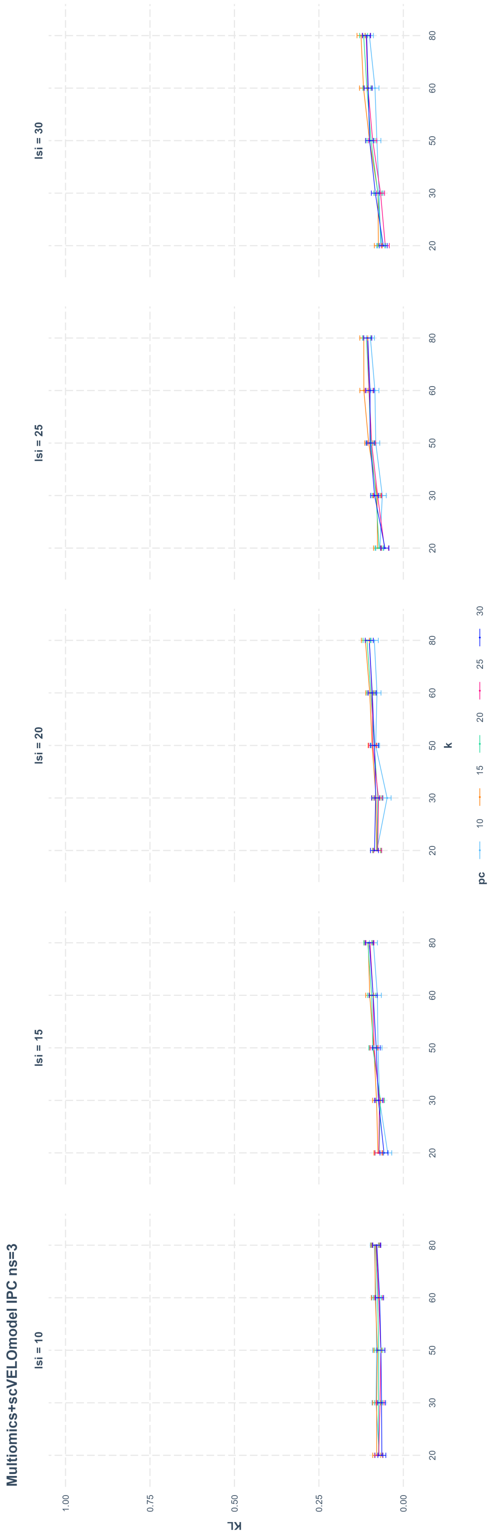


Figure A.8: Linear model results for the Multiomics+scVELO model with $n_s = 3$ and IPC cluster

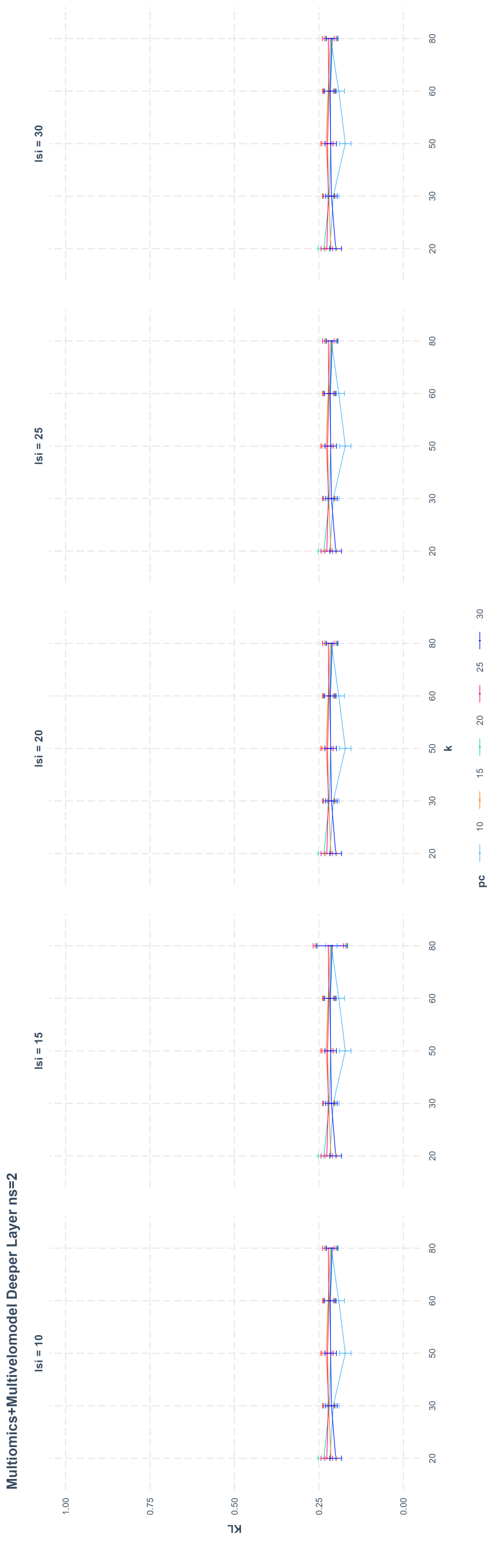


Figure A.9: Linear model results for the Multiomics+MultiVelo model with $n_s = 2$ and Deeper Layer cluster

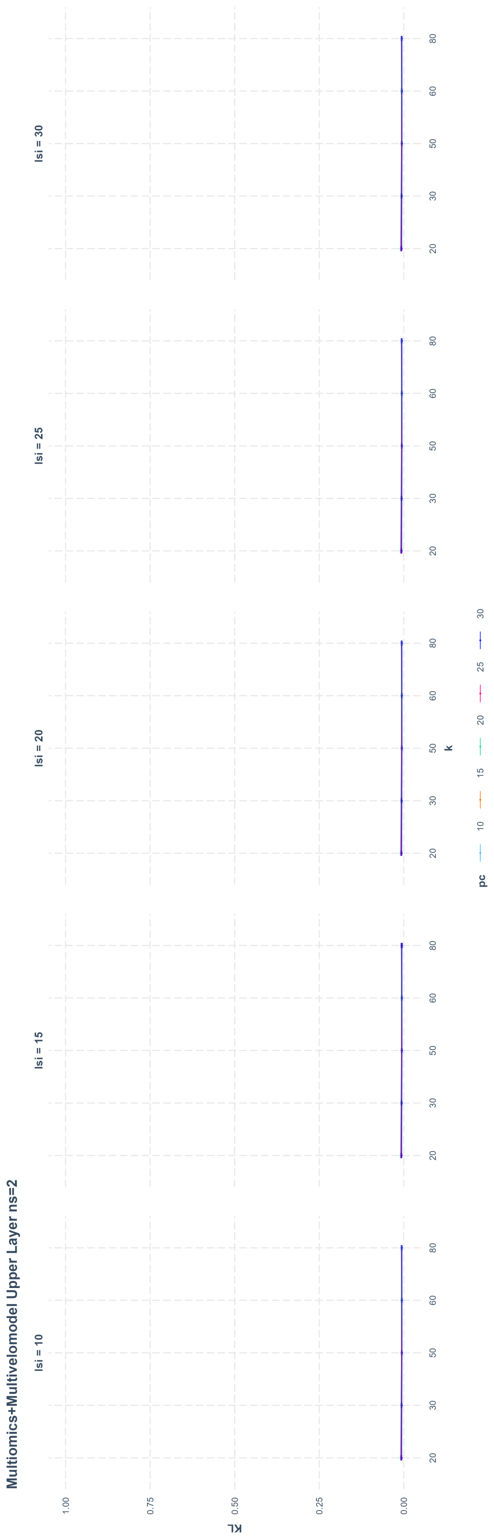


Figure A.10: Linear model results for the Multiomics+MultiVelo model with $n_s = 2$ and Upper Layer cluster

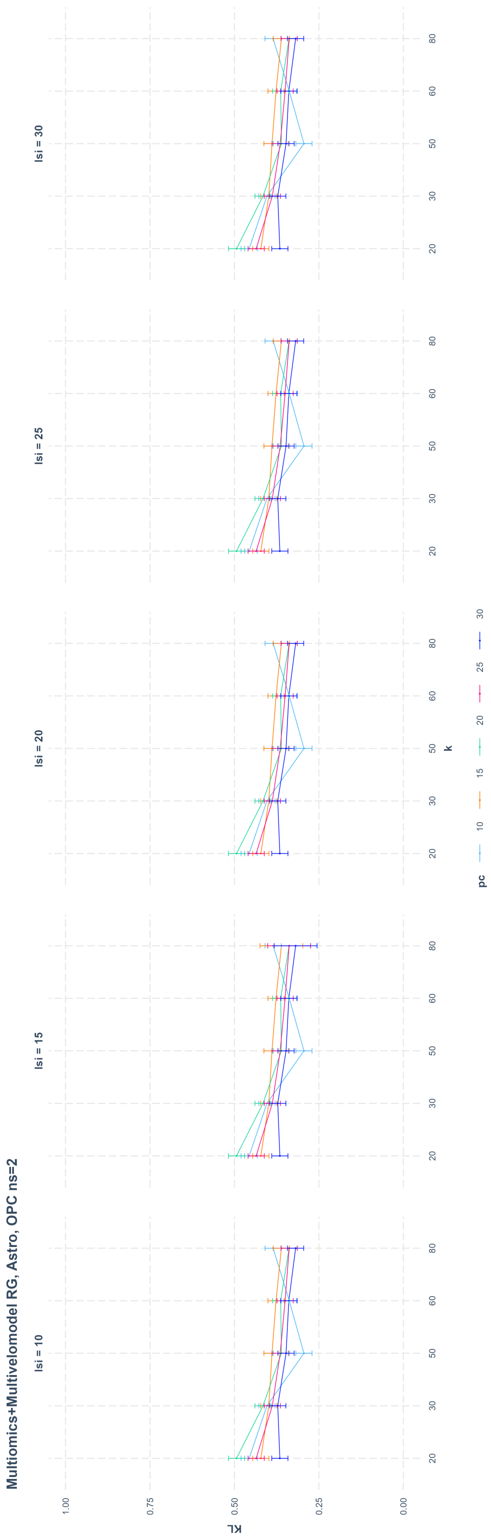


Figure A.11: Linear model results for the Multiomics+MultiVelo model with $n_s = 2$ and RG, Astro, OPC cluster

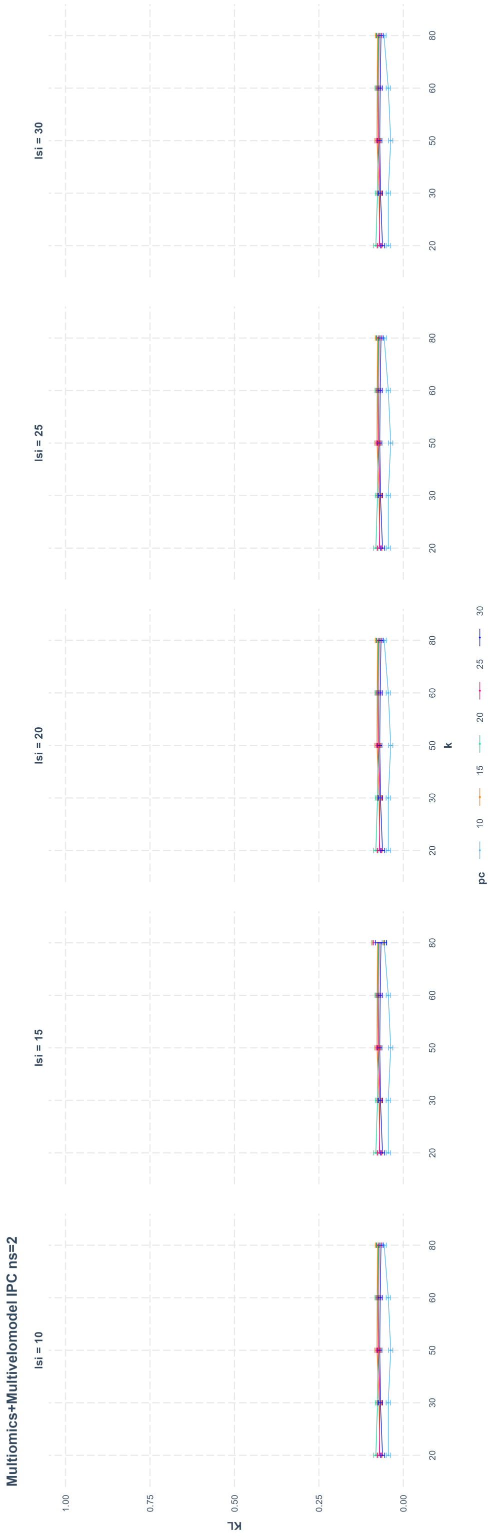


Figure A.12: Linear model results for the Multiomics+MultiVelo model with $n_s = 2$ and IPC cluster

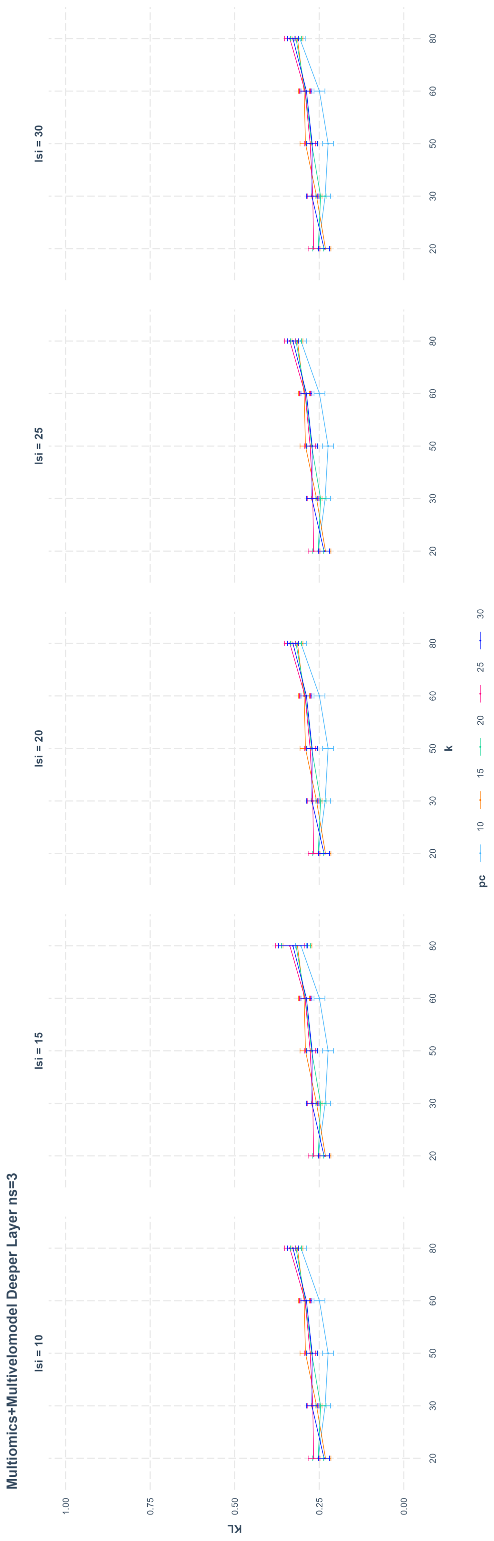


Figure A.13: Linear model results for the Multiomics+MultiVelo model with $n_s = 3$ and Deeper Layer cluster

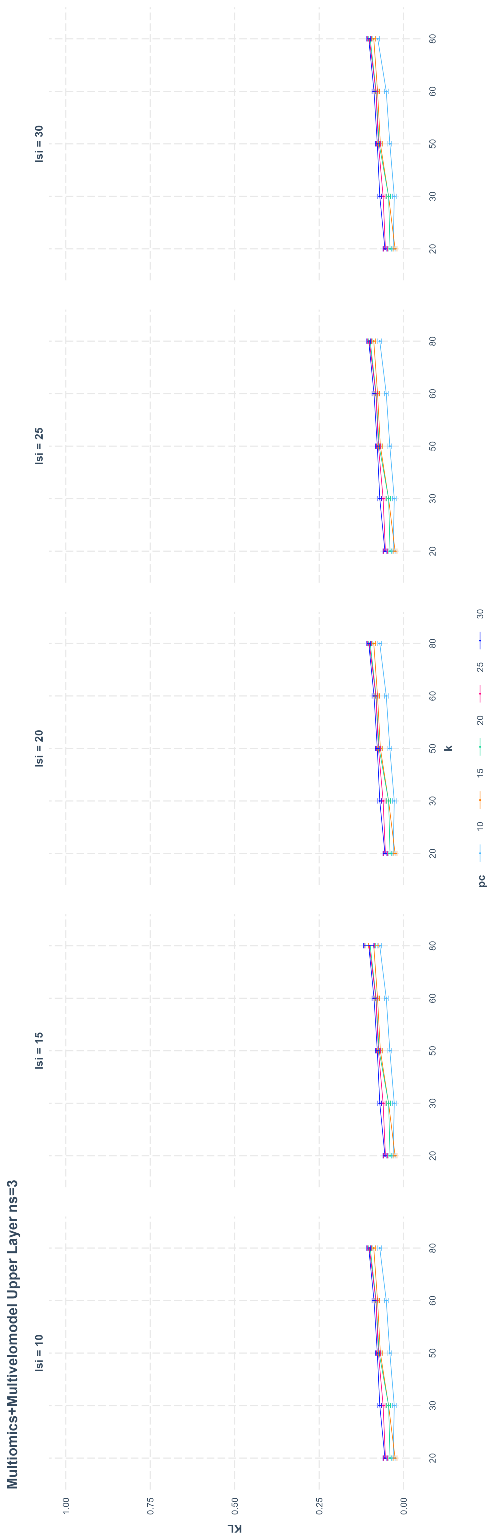


Figure A.14: Linear model results for the Multiomics+MultiVelo model with $n_s = 3$ and Upper Layer cluster

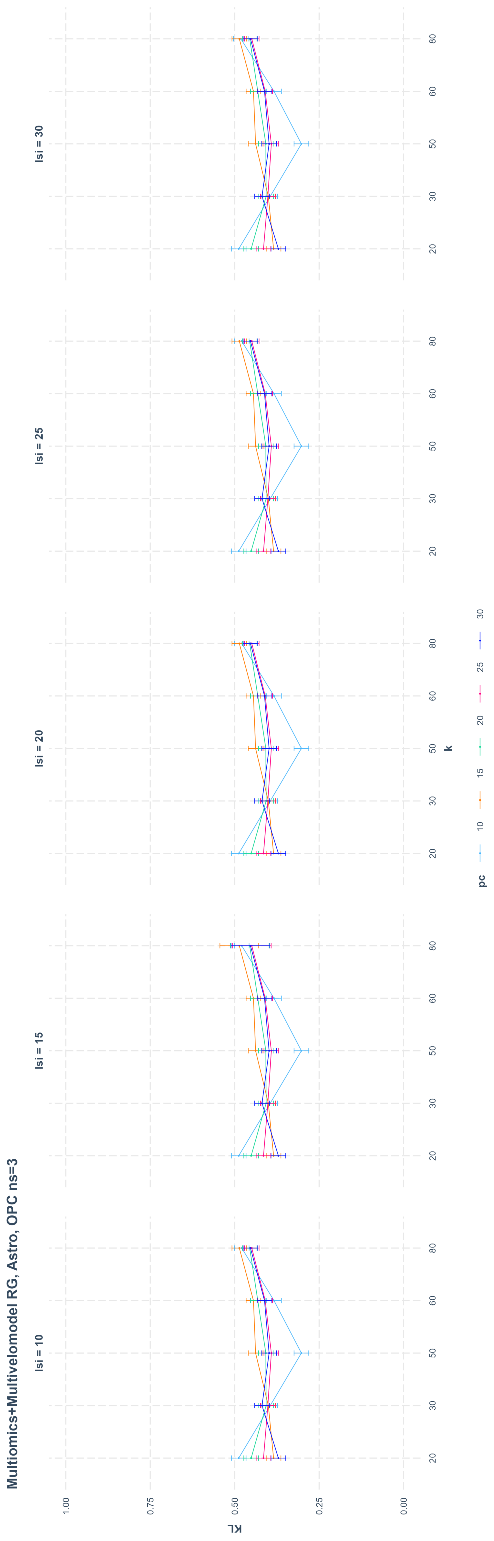


Figure A.15: Linear model results for the Multiomics+MultiVelo model with $n_s = 3$ and RG, Astro, OPC cluster

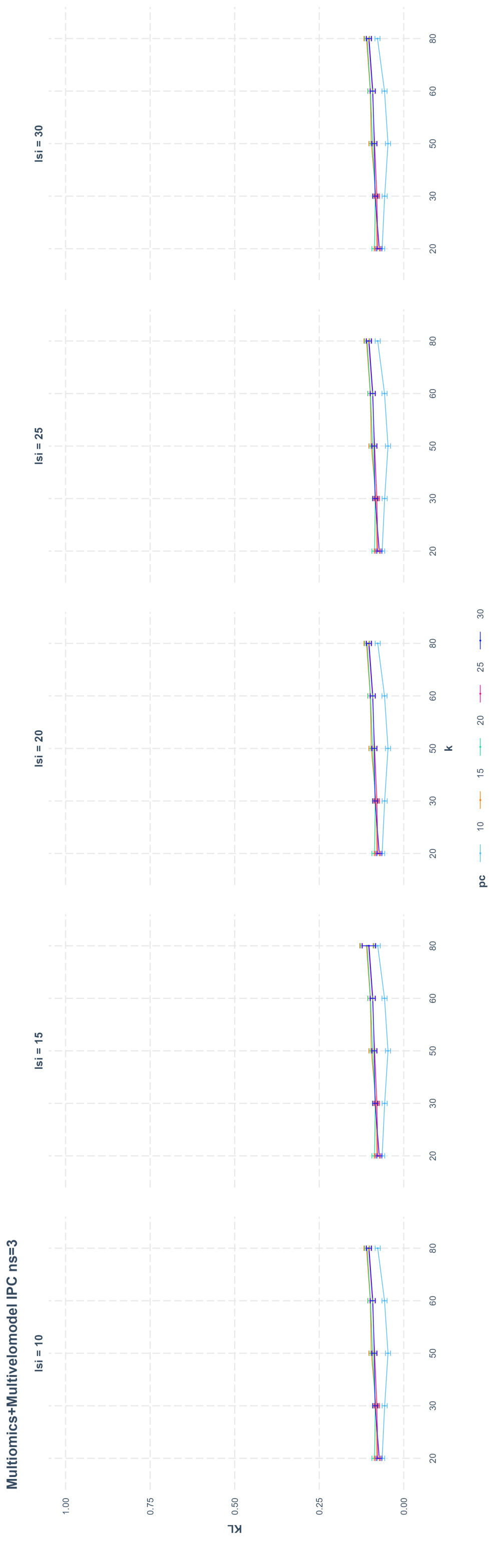
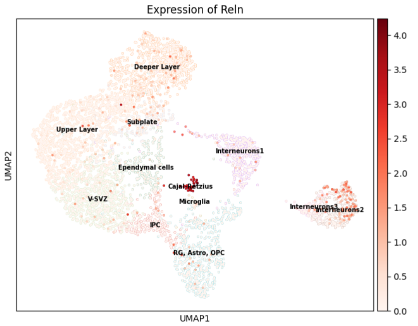
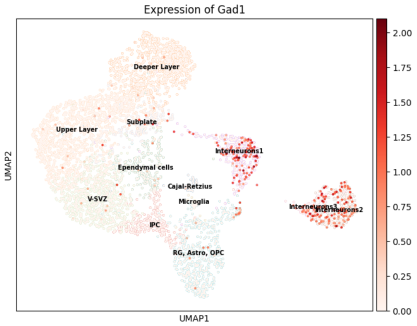
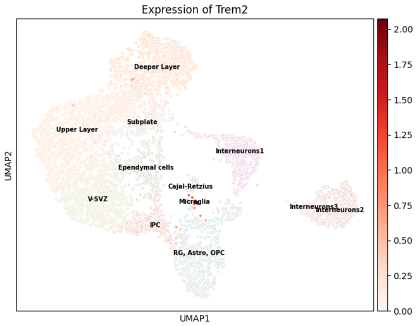


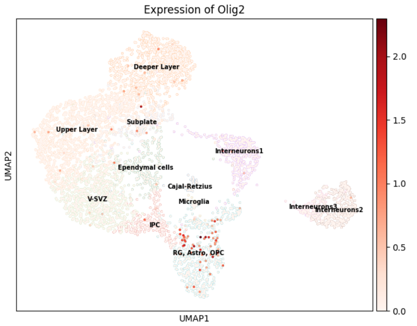
Figure A.16: Linear model results for the Multiomics+MultiVelo model with $n_s = 3$ and IPC cluster

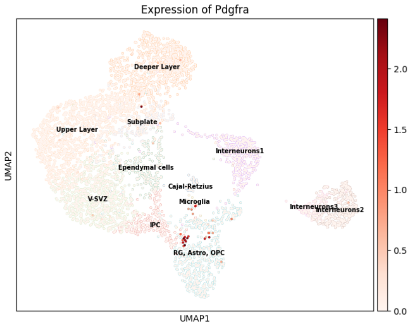
Appendix B

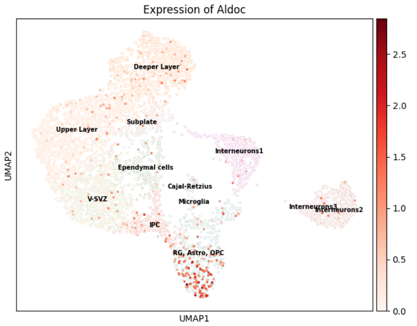

Expression of Reln

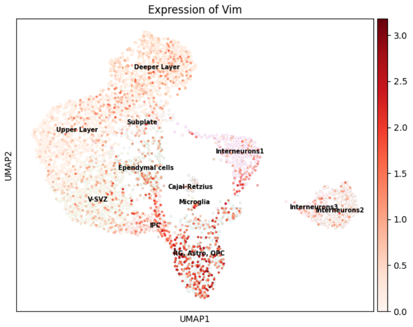

Expression of Gad1

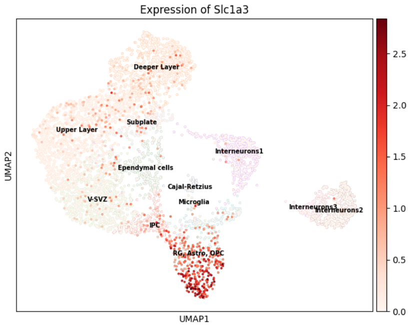

Expression of Trem2

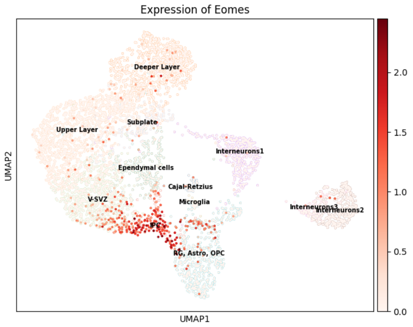

Expression of Olig2

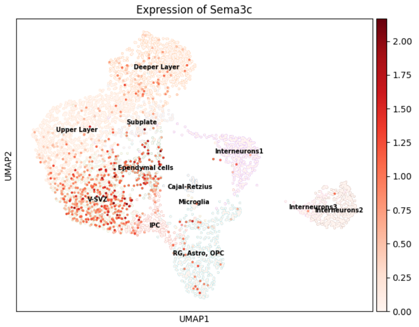

Expression of Pdgfra

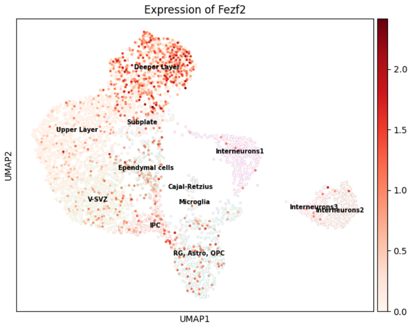

Expression of Aldoc

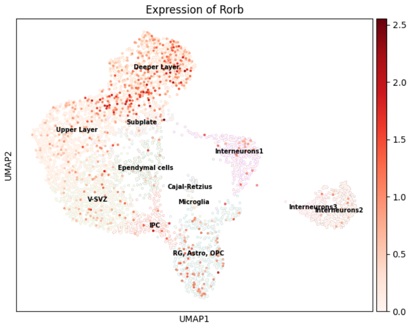

Expression of Vim
Deeper Layer
Upper Layer
Subplate
Interneurons1
Ependymal cells
Cajal-Retzius
V-SVZ
Microglia
Interneurons3
Interneurons2
IPC
RG, Astro, OPC
UMAP1
UMAP2
3.0
2.5
2.0
1.5
1.0
0.5
0.0

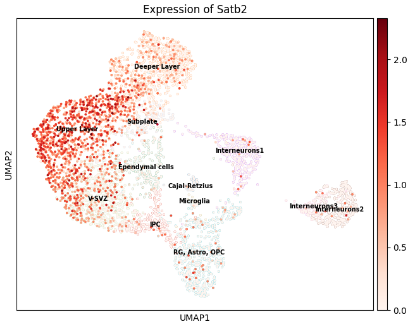

Expression of Slc1a3
Deeper Layer
Upper Layer
Subplate
Interneurons1
Ependymal cells
Cajal-Retzius
V-SVZ
Microglia
Interneurons3
Interneurons2
IPC
RG, Astro, OPC
UMAP1
UMAP2
2.5
2.0
1.5
1.0
0.5
0.0

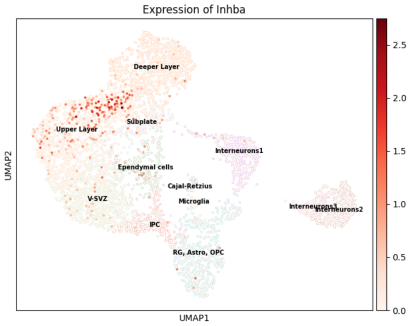

Expression of Eomes
Deeper Layer
Upper Layer
Subplate
Interneurons1
Ependymal cells
Cajal-Retzius
V-SVZ
Microglia
Interneurons3
Interneurons2
IPC
RG, Astro, OPC
UMAP1
UMAP2
2.0
1.5
1.0
0.5
0.0


Expression of Sema3c
Deeper Layer
Upper Layer
Subplate
Interneurons1
Ependymal cells
Cajal-Retzius
V-SVZ
Microglia
Interneurons3
Interneurons2
IPC
RG, Astro, OPC
UMAP1
UMAP2
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00


Expression of Fezf2
Deeper Layer
Upper Layer
Subplate
Interneurons1
Ependymal cells
Cajal-Retzius
V-SVZ
Microglia
Interneurons3
Interneurons2
IPC
RG, Astro, OPC
UMAP1
UMAP2
2.0
1.5
1.0
0.5
0.0


Expression of Rorb
Deeper Layer
Upper Layer
Subplate
Interneurons1
Ependymal cells
Cajal-Retzius
V-SVZ
Microglia
Interneurons3
Interneurons2
IPC
RG, Astro, OPC
UMAP1
UMAP2
2.5
2.0
1.5
1.0
0.5
0.0


Expression of Satb2
Deeper Layer
Upper Layer
Subplate
Interneurons1
Ependymal cells
Cajal-Retzius
V-SVZ
Microglia
Interneurons3
Interneurons2
IPC
RG, Astro, OPC
UMAP1
UMAP2
2.0
1.5
1.0
0.5
0.0


Expression of Inhba
Deeper Layer
Upper Layer
Subplate
Interneurons1
Ependymal cells
Cajal-Retzius
V-SVZ
Microglia
Interneurons3
Interneurons2
IPC
RG, Astro, OPC
UMAP1
UMAP2
2.5
2.0
1.5
1.0
0.5
0.0