Computational Methods for Single-Cell and Spatial Multimodal Data Integration

by

Chao Gao

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in the University of Michigan
2024

Doctoral Committee:

Associate Professor Joshua Welch, Chair
Professor Jun Li
Assistant Professor Jie Liu
Professor Kayvan Najarian
Professor Xiang Zhou

Chao Gao

gchao@umich.edu

ORCID iD: 0000-0002-9316-2185

Dedication

I would like to dedicate this dissertation to my wife, parents and grandparents.

Acknowledgments

In this incredible journey of my PhD studies, I owe a debt of gratitude to many people for their guidance and support.

First and foremost, I would like to express my gratitude to my advisor, Dr. Joshua Welch, whose unwavering support, insightful feedback, and invaluable mentorship have been fundamental to my research journey. His expertise and guidance have been pivotal in shaping both my academic development and this dissertation.

I am immensely thankful to my dissertation committee members, Dr. Jun Li, Dr. Jie Liu, Dr. Kayvan Najarian, and Dr. Xiang Zhou, for their substantial contributions, commitment, and wise counsel. Their perspectives and constructive feedback have greatly enriched my research over the years.

My sincere appreciation goes to the administrative team at the Department of Computational Medicine and Bioinformatics, particularly Julia Eussen, Kati Ellis, and Jane Wiesner, for their vital role in assisting students in the bioinformatics program. I also extend thanks to Ken Weiss, Paul Kopec, and Aaron Bookvich for their IT support. I am grateful to Dr. Margit Burmeister and Dr. Maureen Sartor, the co-directors of the PhD program, for their valuable advice throughout my training.

I am thankful to all my colleagues from Welch Lab for their collaborative spirit and support. I also express my heartfelt thanks to my friends in Ann Arbor; the joyful moments we shared are an integral part of my PhD experience. Special thanks to Yuzhong Yang and Irina Zhang for their lasting friendship.

My deepest gratitude goes to my family. To my wife, Ying Ma, for your endless love, understanding, and support. You fill my journey with joy and strength. To my parents and grandparents, whose unconditional love, sacrifices, and encouragement have been the bedrock of my strength and perseverance. To my uncles, aunts, cousins, nephews, and niece, for your support throughout my life. The love of my entire family has molded me into who I am today. This dissertation is a testament to your enduring love and belief in my potential.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

LIST OF ALGORITHMS

ALGORITHM

ABSTRACT

Advancements in sequencing technologies have revolutionized our ability to measure biomolecules. Single-cell single-omics sequencing allows for the examination of genome, transcriptome, epigenome at unprecedented resolution, providing a detailed view of cellular diversity and function. Furthermore, it addressed the limitations of bulk RNA sequencing that only profiles averaged gene expression across cells, masking the cellular heterogeneities. Following this, single-cell multimodal omics enables simultaneous analysis of multiple types of molecular measurements in the same cell. Such paired information has revealed genetic and epigenetic landscapes as well as their relationships. Further, spatial sequencing technologies provide molecular measurements with localization within tissues, adding an essential dimension to our understanding of biological complexity. They have assisted our research about how cells interact within spatial context, crucial for comprehending tissue organization, development, and disease pathology. In this dissertation, I propose three computational methods to address the challenges posed by each of these data types for identifying the heterogeneities within cell populations and tissue regions, advancing our knowledge of biological systems.

Integrating diverse single-cell unimodal datasets offers tremendous opportunities for unbiased, comprehensive, quantitative definition of cell identities. The published single-cell data integration approaches are not designed for integration of multiple modalities or not scalable to massive datasets. None of these methods can incorporate new data without recalculating from scratch. To this end, I develop an online learning algorithm to solve the integrative nonnegative matrix factorization (`Online iNMF`). For cell type inference, I apply `Online iNMF` to integrate large-scale, continually arriving single-cell datasets of diverse molecular modalities, including gene expression, chromatin accessibility, and DNA methylation. `Online iNMF` converges rapidly and decouples the peak memory usage from the size of the entire dataset. `Online iNMF` shows that the improved computational efficiency is not at the cost of dataset alignment and cluster preservation performance. `Online iNMF`'s ability to iteratively incorporate data is useful in building single-cell multi-omic atlases.

Single-cell multimodal epigenomic profiling simultaneously measures multiple histone modifications and chromatin accessibility in the same cells. Such parallel measurements provide opportunities to investigate how epigenomic modalities vary together across cell populations. I

propose ConvNet-VAE, a variational autoencoder comprising one-dimensional convolutional layers, for dimensionality reduction. After window-based genome binning, ConvNet-VAE leverages the multi-track and sequential nature of these data. I apply ConvNet-VAE to integrate histone modification marks and chromatin accessibility profiled from juvenile mouse brain and human bone marrow. Compared to multimodal VAEs with only fully connected layers, ConvNet-VAE can achieve better performance in dimensionality reduction and batch correction, while using significantly fewer parameters. The advantage of ConvNet-VAE increases with the number of modalities, making it a promising tool as the number of jointly profiled epigenomic modalities grows.

Multimodal spatial profiling has allowed for the simultaneous investigation of transcriptomics, proteomics, and epigenomics at the individual cell/bead/spot level in the tissue. I devise spaMVGAE, a multimodal variational autoencoder employing graph convolutional networks. By incorporating spatial location information, spaMVGAE adapts to various modalities and learns a joint low-dimensional embedding of cells/beads/spots for domain detection. I apply spaMVGAE to spatially resolved multimodal datasets from different biological contexts, such as breast cancer, mouse bone development, and adult mouse brain. spaMVGAE accurately detects regions of interest by capturing the heterogeneous and complex molecular makeup of the cells or tissue microenvironments. spaMVGAE scales to large datasets and carries out joint integration across multiple tissue sections.

CHAPTER 1

Introduction

1.1 Motivation and Research Objectives

Cells, the fundamental units of life, show remarkable diversity in shape and function, influenced by their location, developmental stages, external stimuli, and differences between healthy and diseased states. A multitude of biomedical research, such as comprehending brain functions, finding novel therapeutics, and studying the formation of complex tissues from a single cell, are all tied to understanding the variations in gene expression and epigenetic marks in these cells.

After the initial sequencing of the human genome, bulk DNA sequencing became a ubiquitous tool to profile transcriptomic or epigenomic information from the cells in entire tissues. For instance, RNA-seq assesses the overall gene expression levels in a bulk tissue (Stark et al. 2019), whereas the assay for transposase-accessible chromatin with sequencing (ATAC-Seq) determines chromatin accessibility across the genome, yielding a broad picture of the epigenetic landscape in cell populations (Grandi et al. 2022). These sequencing methods have helped researchers to understand basic cellular functions and disease mechanisms, leading to the development of targeted therapies, for example, for cancer (Hong et al. 2020). Over the past decade, researchers have continually driven the field forward with cumulative technological advances.

More recently, the field of biomedical sciences has been revolutionized by the advent of single-cell sequencing technologies, which have vastly expanded our understanding of the complexity and diversity of cellular processes. Unlike bulk sequencing, which analyzes a mixture of cells from a tissue sample and provides an averaged view of gene expression or epigenomic features across all cells in the sample, single-cell sequencing offers a more granular view of cell biology by isolating and analyzing the genetic material and other molecules from individual cells, allowing for the investigation of cellular heterogeneity and the identification of rare cell types that may be lost in bulk analyses. These single-cell technologies have become crucial tools in biomedical research, offering insights into the molecular mechanisms that underlie health and disease at an unprecedented resolution.

The current frontier of molecular assay development is molecular profiling within a spatial

context, literally introducing a new dimension into our understanding of the complexity of biological tissues. This advancement enables researchers to observe not only the molecular composition of cells but also their precise spatial arrangement and interactions within tissues, thus providing a more comprehensive picture of biological processes and disease pathogenesis. It allows for a more accurate and detailed investigation of tissue architecture and cellular microenvironments, potentially leading to the identification of novel biomarkers and therapeutic targets.

Given rapidly evolving sequencing protocols, the development of computational methods to analyze different data types is essential for advancements in biomedical research. The ideal methods should be able to tackle challenges such as integrating multiple data modalities through efficient joint dimensionality reduction for novel insights in molecular biology. To retain meaningful and significant patterns across biological conditions, sophisticated statistical and machine learning techniques are required. My work closely follows the frontier of molecular assay development and aims at devising tailored tools to extract latent structure of the data that can be used to answer a variety of biological questions. In this dissertation, I present these methods designed for three broad categories of sequencing technologies: single-cell single omics, single-cell multimodal omics, and spatially resolved multimodal data, and showcase their utilities in integrative analysis.

1.1.1 Single-cell single-omics

Next Generation Sequencing (NGS) technologies, characterized by high-throughput, cost-effectiveness, and wide application, have greatly accelerated the pace of genomic research (Van Dijk et al. 2014). NGS serves as the technological foundation for single-cell sequencing, an advanced application of NGS that focuses on analyzing the genetic material from individual cells. This finer resolution provides a more detailed and nuanced understanding of biology. The first single-cell whole-transcriptome sequencing (scRNA-seq) was introduced by Tang et al. (Tang et al. 2009). Since then, significant improvements in single-cell sequencing have enhanced its capabilities and applications. Advances in molecular assay technology have: 1) substantially improved the sensitivity and accuracy of molecule detection, increasing data quality; 2) increased throughput, enabling large-scale studies like organism-wide cell type maps; 3) reduced cost, leading to widespread adoption in research. Multiplexing capabilities, where multiple samples can be sequenced simultaneously in a single run, largely contributed to these improvements. They not only permit higher throughput and lower cost, but also reduces technical variability. The many sequencing methods that have been developed can be categorized into two primary groups. Microfluidic droplet-based transcriptomics encapsulates individual cells in tiny droplets, each containing a bead with unique DNA barcodes. The droplets are created at high throughput, allowing the analysis of thousands to millions of cells in a single experiment. Examples include Drop-seq (Macosko et al. 2015) and inDrop (Indexing Droplets) (Klein et al. 2015). On the other hand, micro-well plate-based transcriptomics approach

uses micro-well plates, where each well is designed to capture a single cell along with a uniquely barcoded bead, for instance, Smart-Seq (Ramsköld et al. 2012), Smart-Seq2 (Picelli et al. 2014), and Smart-Seq3 (Hagemann-Jensen et al. 2020).

Beyond scRNA-seq, a number of other single-cell sequencing technologies measure a single modality, each capturing different aspects of cellular biology at the single-cell level. Single-cell DNA Sequencing focuses on analyzing genomic DNA in single cells, which is used for detecting genomic variations such as single nucleotide polymorphisms (SNPs) (Dong et al. 2017), copy number variations (CNVs) (Mallory et al. 2020), and other mutations within individual cells. Single-nucleus ATAC-seq (snATAC-seq) is used to assess the chromatin accessibility landscape in individual cells (Cusanovich et al. 2015, Preissl et al. 2018). Another example sequencing protocol for epigenomic studies is single-cell bisulfite sequencing, such as snmC-seq and snmC-seq2 (Luo et al. 2017, 2018). These techniques are employed for investigate DNA methylation patterns, a crucial epigenetic modification that influences gene expression across single cells.

Named "Method of the Year (2013)" (Editorial 2014), single-cell single-omic technologies have expanded our ability to understand the complexity of biological systems in a highly detailed and nuanced manner. Each of them offers the potential for utilizing a specific cellular feature in cataloging cell types. More importantly, the combined use of these single-omics data allows scientists to reexamine traditional categorizations of cell types and states in a methodical, quantitative, and impartial manner. This quantitative approach to defining cell identity is poised to transform our comprehension of cell biology in various areas, including neuroscience and developmental biology. Additionally, creating a benchmark for the molecular states of healthy cells will facilitate investigations into the origins of cellular irregularities, potentially leading to the innovation of new, targeted treatments. To accomplish this objective, there is a need for an computational method that can assemble diverse molecular features from different batches of cells into a joint representation for cell identity inference.

A number of methods for integrating single-cell single-omics, such as Seurat v3 (Stuart et al. 2019) and Harmony (Korsunsky et al. 2019a), have emerged. However, these methods are not equipped to handle the integration of multiple data types or are unable to manage extremely large datasets. Additionally, these existing techniques lack the capability to add new data without having to start the calculations over from the beginning. I addressed these limitations by developing online integrative nonnegative matrix factorization (iNMF), an algorithm that allows scalable and iterative integration of single-cell datasets generated by different omics technologies, by extending the iNMF approach at the heart of the published LIGER method (Welch et al. 2019, Liu et al. 2020).

1.1.2 Single-cell multimodal omics

Single-cell multimodal omic experiments are more advanced techniques can simultaneously capture and analyze multiple types of molecular data from the same cell. These methods are groundbreaking as they provide a more comprehensive and integrated understanding of cellular function and state, making them "Method of the Year (2019)" (Teichmann and Efremova 2020). For example, G&T-seq (Macaulay et al. 2015) can measure both the genome and transcriptome, offering novel insights into the genome-transcriptome correlations. Later on, the simultaneous isolation of genomic DNA and total RNA (SIDR) (Han et al. 2018) was developed. This concurrent extraction can minimize sample loss and handling errors, potentially leading to more accurate and reliable correlations between genomic and transcriptomic data. TARGET-seq is designed for targeted sequencing of genomic regions, which allows for a more focused and in-depth analysis of particular genes or mutations of interest (Rodriguez-Meira et al. 2019). Moreover, CITE-seq allows for the concurrent detection of protein markers and RNA transcripts in individual cells (Stoeckius et al. 2017). It bridges the gap between genotype and phenotype, providing insights into how gene expression is translated into functional protein molecules. In 2017, researchers modified Nucleosome Occupancy and Methylome-sequencing (NOMe-seq) to measure chromatin accessibility and endogenous DNA methylation in single cells (scNOMe-seq) (Pott 2017). Afterwards, the plate-based Sci-CAR-seq (Cao et al. 2018) and the droplet-based SNARE-seq (Chen et al. 2019a) were introduced with improved scalability. The commercialized 10x Genomics Multiome platform also provides joint profiling of chromatin accessibility and gene expression.

Studying the epigenetic modifications of the genetic material has been a long-term goal in molecular biology. In 2019, single-cell chromatin immunoprecipitation followed by sequencing (scChIP-seq) (Grosselin et al. 2019), a single-omics approach, aimed to reveal chromatin landscapes in individual cells with high accuracy, and investigate cell populations by identifying discriminating chromatin features such as transcriptional permissive or repressive marks. Recently, researchers introduced the approach of nano-CUT&Tag (nano-CT) (Bartosovic and Castelo-Branco 2022) that enables simultaneous profiling of up to three different epigenomic features at the single-cell level with notably enhanced sensitivity and high sequencing depth per cell. Another method, single-cell nanobody-tethered transposition followed by sequencing (scNTT-seq) (Stuart et al. 2022), is also capable of measuring the genome-wide presence of multiple histone modifications at single-cell resolution.

Advanced statistical and computational methods are needed for single-cell multimodal omic data analysis to address questions such as how to perform cross-modal integration and cell type inference using these molecular measurements. Specifically for the study of epigenome, it's crucial to take into account the multi-track and ordered sequential nature of single-cell multimodal epigenomic data when developing the analytical tools. I tackled this challenge by devising the `ConvNet-VAE`,

a variational autoencoder framework that employs convolutional layers, for multimodal epigenomic data integration.

1.1.3 Spatially resolved multimodal data

Spatial transcriptomics, designated "Method of the Year (2020)", has been a revolutionary technology in the field of molecular biology that combines histological and transcriptomic information from tissue samples (Marx 2021). Traditional sequencing technologies, such as scRNA-seq, provide detailed information about gene expression level but lack spatial context. However, it is critical to recognize that the arrangement of cellular compartments, macro-structures, and the interactions between cells is crucial for the functioning of multicellular organisms (Baysoy et al. 2023). This has motivated researchers to develop sequencing technologies that reveal the spatial organization of gene expression in tissues.

The spatial transcriptomic technologies developed so far can be broadly categorized into three groups based on the technology applied. Imaging-based methods use fluorescent in situ hybridization (FISH) to visualize and quantify RNA molecules directly within tissue sections or cultured cells. Examples include multiplexed single-cell in situ RNA profiling by sequential hybridization (Lubeck et al. 2014) based upon the single-molecule FISH (smFISH) technique (Femino et al. 1998). The Multiplexed Error-Robust Fluorescence In Situ Hybridization (MERFISH) (Chen et al. 2015a) can simultaneously image a large number of RNA species with high accuracy. Researchers have demonstrated its utility by applying it to approximately 10 million cells and generating spatially resolved cell atlas of the whole mouse brain (Zhang et al. 2023a). Moreover, the latest 10x Genomics Xenium platform (Janesick et al. 2023) offers high-throughput high-resolution spatial mapping of RNA expression in tissues. Different from imaging-based methods, sequencing-based technologies directly profile biomolecules of interest within tissue samples using NGS. The researchers reported the development of "spatial transcriptomics" (Ståhl et al. 2016) in 2016, which they applied to produce spot-level (diameter: 100 μm) RNA-seq data, along with two-dimensional positional coordinates from the mouse brain and human breast cancer. Since then, technologies like Visium from 10x Genomics (diameter: 55 μm) and Slide-seq (bead size: 10 μm) (Rodriques et al. 2019) have pushed the boundaries further in terms of resolution and multiplexing capabilities. The third category is the laser capture microdissection (LCM)-based technologies. In this approach, the isolation of a region of interest in the tissue section is achieved by laser cutting. Chen et al. introduced geographical position sequencing (Geo-seq) (Chen et al. 2017), which combines laser capture microdissection (LCM) and scRNA-seq to carry out spatial profiling.

With the same scientific motivation as single-cell multimodal omics, the investigation of multiple data modalities within the same cell, bead, or spot using spatial sequencing is one of the latest frontiers in molecular assay technology. In practice, the scientists are presented with the choice to

utilize techniques like Hematoxylin & Eosin (H&E) or Nissl staining on either the sequenced tissue slice or adjacent sections. This approach provides valuable morphological details about cells, beads, or spots, which in turn enriches our comprehension of cell and tissue types, thereby augmenting our knowledge of cellular functions and tissue structures. Recently, cutting-edge technologies have surfaced that allow profiling of multiple molecular modalities in a single bead or spot. The spatial assay for transposase-accessible chromatin and RNA using sequencing (spatial ATAC-RNA-seq) (Zhang et al. 2023b) is able to simultaneously analyze chromatin accessibility and messenger RNA expression in two-dimensional grid of spatially barcoded tissue pixels. Additionally, Russell et al. developed the Slide-tags (Russell et al. 2023) technique, which labels nuclei with spatial barcodes, enabling direct application of any single-cell multimodal assay with the addition of spatial coordinates.

Tailored computational methods are required for integrating these spatially resolved multimodal data to distinguish cell populations within tissue sections. To this end, I developed `spaMVGAE`, a multimodal variational graph autoencoder, which efficiently incorporates spatial location information and multiple modalities to learn joint low-dimensional embeddings of spatially resolved measurements for spatial clustering.

1.2 Outline of Dissertation

The goal of this dissertation is to address these rising challenges associated with integrative analysis of large-scale single-cell and spatial multimodal omics data. The dissertation is outlined as follows.

Chapter 2 – Iterative single-cell multi-omic integration using online learning In this chapter, we propose `online iNMF`: an **Online** Learning algorithm to solve the **I**ntegrative **N**onnegative **M**atrix **F**actorization problem for integrating large, diverse, and continually arriving single-cell datasets. Our approach scales to arbitrarily large numbers of cells using fixed memory, iteratively incorporates new datasets as they are generated, and allows many users to simultaneously analyze a single copy of a large dataset by streaming it over the internet. Iterative data addition can also be used to map new data to a reference dataset. Comparisons with previous methods indicate that the improvements in efficiency do not sacrifice dataset alignment and cluster preservation performance. We demonstrate the effectiveness of online iNMF by integrating more than a million cells on a standard laptop, integrating large single-cell RNA-seq and spatial transcriptomic datasets, and iteratively constructing a single-cell multi-omic atlas of the mouse motor cortex.

Chapter 3 – Integrating single-cell multimodal epigenome data using 1D-convolutional neural networks In this work, we focus on the single-cell multimodal epigenomic profiling, which measures multiple histone modifications and chromatin accessibility within the same cell. Such

parallel measurements provide exciting new opportunities to investigate how epigenomic modalities vary together across cell types and states. A pivotal step in using this type of data is integrating the epigenomic modalities to learn a unified representation of each cell, but existing approaches are not designed to model the unique nature of this data type. Our key insight is to model single-cell multimodal epigenome data as a multi-channel sequential signal. Based on this insight, we developed `ConvNet-VAE`s, a novel framework that uses 1D-convolutional variational autoencoders (VAEs) for single-cell multimodal epigenomic data integration. We evaluated `ConvNet-VAE`s on nano-CT and scNTT-seq data generated from juvenile mouse brain and human bone marrow. We found that `ConvNet-VAE`s can perform dimension reduction and batch correction better than previous architectures while using significantly fewer parameters. Furthermore, the performance gap between convolutional and fully-connected architectures increases with the number of modalities, and deeper convolutional architectures can increase performance while the performance degrades for deeper fully-connected architectures. Our results indicate that convolutional autoencoders are a promising method for integrating current and future single-cell multimodal epigenomic datasets.

Chapter 4 – Integrating spatially resolved multimodal data using variational graph autoencoder Recent advancements in spatial profiling have allowed for the simultaneous investigation of transcriptomics, proteomics, and epigenomics at the individual cell/bead/spot level in the tissue. These technologies have been instrumental in revealing the heterogeneous and complex molecular makeup of the cells or tissue microenvironments. Deeper insights into the biological process can be gained by incorporating high-resolution image modalities. We present `spaMVGAE` for spatially informed multimodal integration, a multimodal variational graph autoencoder employing graph convolutional networks. It learns a joint embedding of cells/beads/spots by correlating molecular measurements (e.g., gene expression, chromatin accessibility), cell morphology (e.g., H&E histology), and spatial location information. The resulting low-dimensional embeddings can be used for diverse tasks such as domain detection. By applying `spaMVGAE` on spatially resolved multimodal datasets generated in a variety of biological contexts, we show that `spaMVGAE` can harness different sources of information and learn a refined representation of the observations by taking advantage of the spatial information, in a computationally efficient fashion.

Finally, I conclude the projects completed in this dissertation in Chapter 5 by summarizing the contributions and discussing the future directions.

CHAPTER 2

Iterative Single-Cell Multi-Omic Integration Using Online Learning

In this chapter, we propose `Online iNMF`: an **Online** Learning algorithm to solve the **I**ntegrative **N**onnegative **M**atrix **F**actorization problem for integrating large, diverse, and continually arriving single-cell datasets. Our approach scales to arbitrarily large numbers of cells using fixed memory, iteratively incorporates new datasets as they are generated, and allows many users to simultaneously analyze a single copy of a large dataset by streaming it over the internet. Iterative data addition can also be used to map new data to a reference dataset. Comparisons with previous methods indicate that the improvements in efficiency do not sacrifice dataset alignment and cluster preservation performance. We demonstrate the effectiveness of `Online iNMF` by integrating more than a million cells on a standard laptop, integrating large single-cell RNA-seq and spatial transcriptomic datasets, and iteratively constructing a single-cell multi-omic atlas of the mouse motor cortex.

2.1 Introduction

Cell types have long been qualitatively characterized by a combination of features such as morphology, presence or absence of cell surface proteins, and broad function (Ye and Sarkar 2018). Recently, high-throughput single-cell sequencing technologies have enabled researchers to profile multiple molecular modalities, including gene expression, chromatin accessibility and DNA methylation (Stuart et al. 2019). Integrating diverse single-cell datasets offers tremendous opportunities for unbiased, comprehensive, quantitative definition of discrete cell types and continuous cell states.

Several recent single-cell data integration approaches have been developed, including Seurat v3 and Harmony (Stuart and Satija 2019, Korsunsky et al. 2019a), but these approaches are not designed to integrate multiple modalities or do not scale to massive datasets. Furthermore, none of these existing methods can incorporate new data without recalculating from scratch.

We address these limitations by developing `Online iNMF`, an algorithm that allows scalable and iterative integration of single-cell datasets generated by different omics technologies. We extend the nonnegative matrix factorization approach at the heart of our recently published LIGER method

(Welch et al. 2019) to develop an online learning algorithm (Figure 2.1a). LIGER infers a set of latent factors ("metagenes") that represent the same biological signals in each dataset while also retaining the ways in which these signals differ across datasets; these shared and dataset-specific factors are then jointly used to identify cell types and states while also identifying and retaining cell-type-specific differences in the metagene features that define cell identities. In the present study, we combine LIGER with techniques for "online learning" (Mairal et al. 2010), in which calculations are performed iteratively and incrementally as new datasets become available. Note that online learning is a technical term that does not refer to the internet—an online learning algorithm is not necessarily a web tool, although internet applications with continually arriving data often benefit from such approaches. `Online iNMF` enables scalable and efficient data integration with fixed memory usage, as well as incorporating new data without recalculating from scratch.

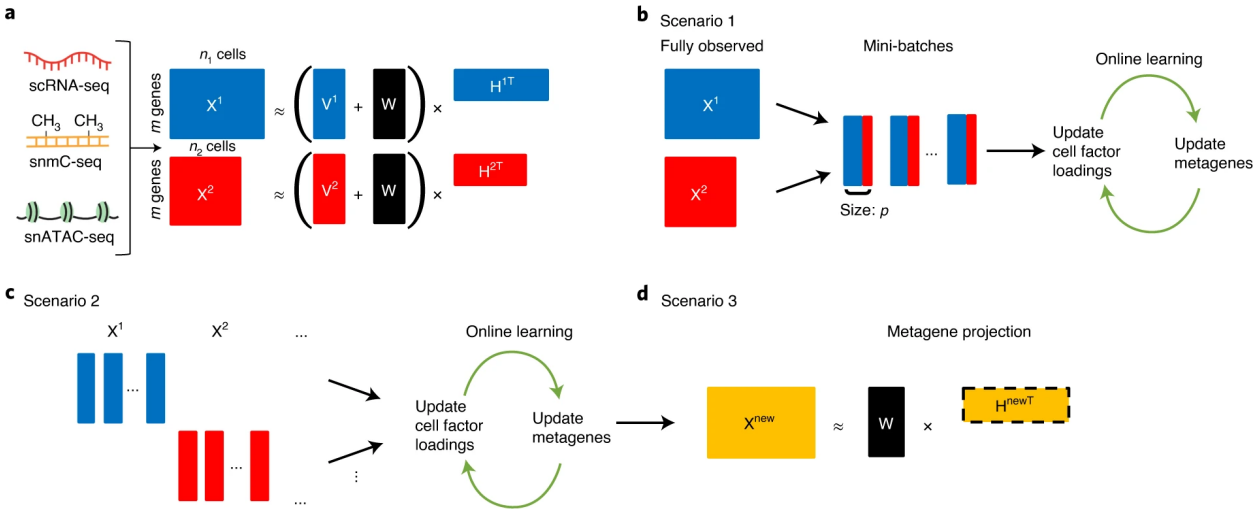


Figure 2.1: Overview of the `Online iNMF` algorithm. (a) Schematic of integrative nonnegative matrix factorization (`iNMF`): the input single-cell datasets are jointly decomposed into shared (W) and dataset-specific (V^i) metagenes and corresponding "metagene expression levels" or cell factor loadings H^i). These metagenes and cell factor loadings provide a quantitative definition of cell identity and how it varies across biological settings. **(b-d)**, Three different scenarios in which online learning can be used for single-cell data integration. **(b)** Scenario 1: the single-cell datasets are large but fully observed. `Online iNMF` processes the data in random mini-batches, enabling memory usage and/or disk storage independent of dataset size. Each cell may be used multiple times in different epochs of training to update the metagenes. **(c)** Scenario 2: the datasets arrive sequentially, and `Online iNMF` processes the datasets as they arrive, using each cell to update the metagenes exactly once. **(d)** Scenario 3: `Online iNMF` is performed as in scenario 1 or scenario 2 to learn W and V^i. Then cell factor loadings for the newly arriving dataset are calculated using the shared metagenes (W) learned from previously processed datasets. The new dataset is not used to update the metagenes.

2.2 `online iNMF`: An Online Learning Algorithm for Iterative Single-Cell Multi-Omic Integration

We developed an algorithm for `Online iNMF` inspired by the online nonnegative matrix factorization approach of (Mairal et al. 2010). `Online iNMF` provides two significant advantages: (1) integration of large single-cell multi-omic datasets by cycling through the data multiple times in small mini-batches and (2) integration of continually arriving datasets, where the entire dataset is not available at any point during training.

We envision using `Online iNMF` to integrate single-cell datasets in three different scenarios. In scenario 1, where the datasets are large and fully observed, the algorithm accesses mini-batches from all datasets at the same time and repeatedly updates the metagenes (W, V^i) and cell factor loadings (H^i). Each cell can be revisited throughout multiple epochs of training (Figure 2.1b). A key advantage of scenario 1 (compared to batch iNMF) is that only a single mini-batch needs to be in memory at a time. Scenario 1 even allows processing of large datasets without downloading them to disk, by streaming them over the internet. In scenario 2, the input datasets arrive sequentially, and the online algorithm uses each cell exactly once to update the metagenes, without revisiting data already seen (Figure 2.1c). The key advantage of scenario 2 is that the factorization is efficiently refined as new data arrives, without requiring expensive recalculation each time. A third scenario allows us to project new data into the latent space already learned, without using the new data to update the metagenes. In scenario 3, we first use `Online iNMF` to learn metagenes as in scenario 1 or scenario 2. Then, we use the shared metagenes (W) to calculate cell factor loadings for a new dataset, without using the new data to update the metagenes. Scenario 3 efficiently incorporates new data without changing the existing integration results, allowing users to query their data against a curated reference (Figure 2.1d).

2.3 Results

2.3.1 `Online iNMF` Converges Efficiently Without Loss of Accuracy Compared to Batch iNMF

In our first experiment, we evaluated the convergence performance of the `Online iNMF` algorithm on the adult mouse cortex dataset (Saunders et al. 2018), which comprises $156,167$ cells from the frontal cortex and $99,186$ cells from the posterior cortex. The `Online iNMF` algorithm converges much faster than previous batch iNMF algorithms on both the training set and a held-out test set (Figure 2.2a-b), converging to a significantly lower training iNMF objective in a fixed amount of time (Figure 2.2c). `Online iNMF` also shows superior performance on several other datasets from different biological contexts (Figure 2.3). Furthermore, the convergence behavior of

the online algorithm on both training and test sets is relatively insensitive to the mini-batch size (Figure 2.2d-e).

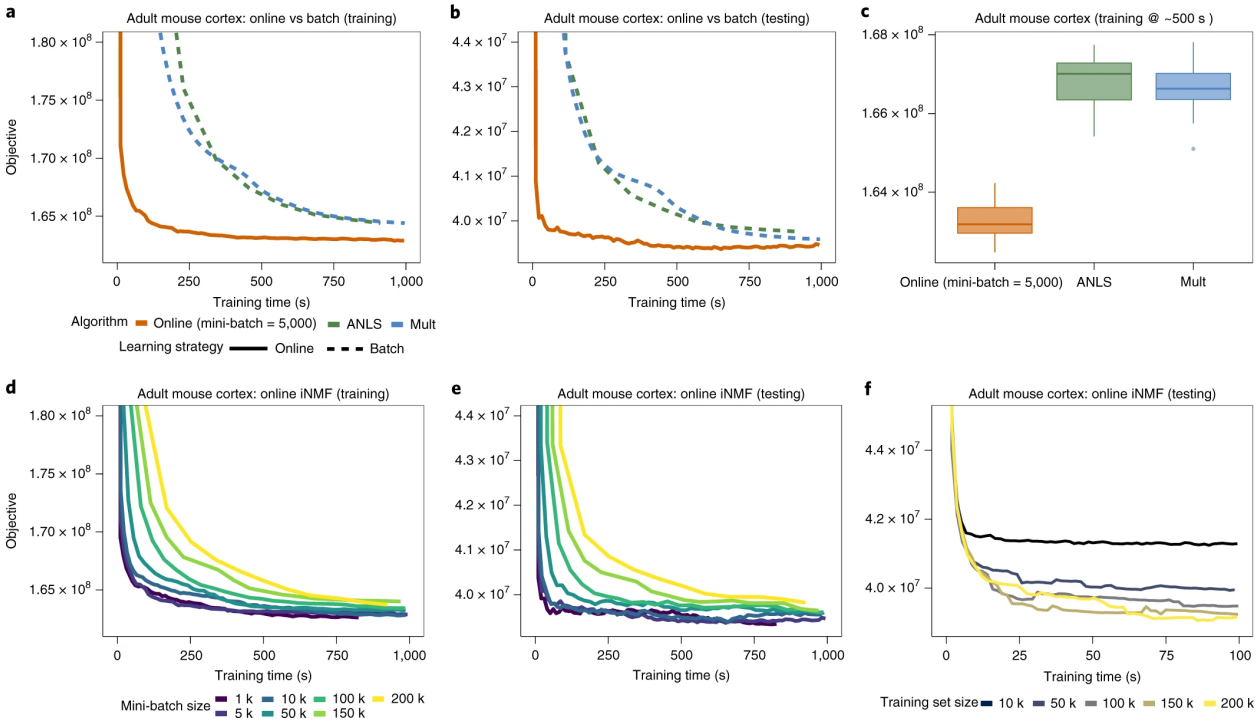


Figure 2.2: `Online iNMF` converges much faster than previously published batch algorithms. **(a,b)** The `Online iNMF` algorithm converges much more rapidly to a similar or better objective function value compared to the previously published batch methods–alternating nonnegative least squares (ANLS) and multiplicative updates (Mult)—on both training and testing sets. **(c)** Box plots comparing the objective function values achieved by applying online and batch iNMF algorithms on the mouse cortex data ($n = 255,353$) after a fixed amount of training time. Center line shows the median; box limits, upper and lower quartiles; whiskers, $1.5\times$ interquartile range; and points are outliers. **(d-e)** The convergence behavior of `Online iNMF` is nearly identical for mini-batch sizes from $1,000$ to $10,000$. **(f)** The `Online iNMF` algorithm becomes increasingly efficient (in terms of decrease in objective function value per unit time) as dataset size increases. The time required for the algorithm to converge does not significantly increase with growing dataset size once the dataset size exceeds $50,000$ cells.

Moreover, for a fixed test set, the runtime needed to reach convergence remains nearly constant once the total number of cells exceeds some minimum threshold (around $50,000$, in this case). (Figure 2.2f). This behavior likely occurs because, for a cell population of fixed complexity (for example, a tissue containing 12 cell types), only some fixed number of observations is required to effectively learn the metagenes. Thus, using the entire dataset to update the shared and data-specific metagenes at each iteration becomes increasingly inefficient as the dataset size exceeds the minimum threshold size needed to learn the metagenes. Conversely, the relative efficiency of `Online iNMF` compared to batch methods increases with dataset size.

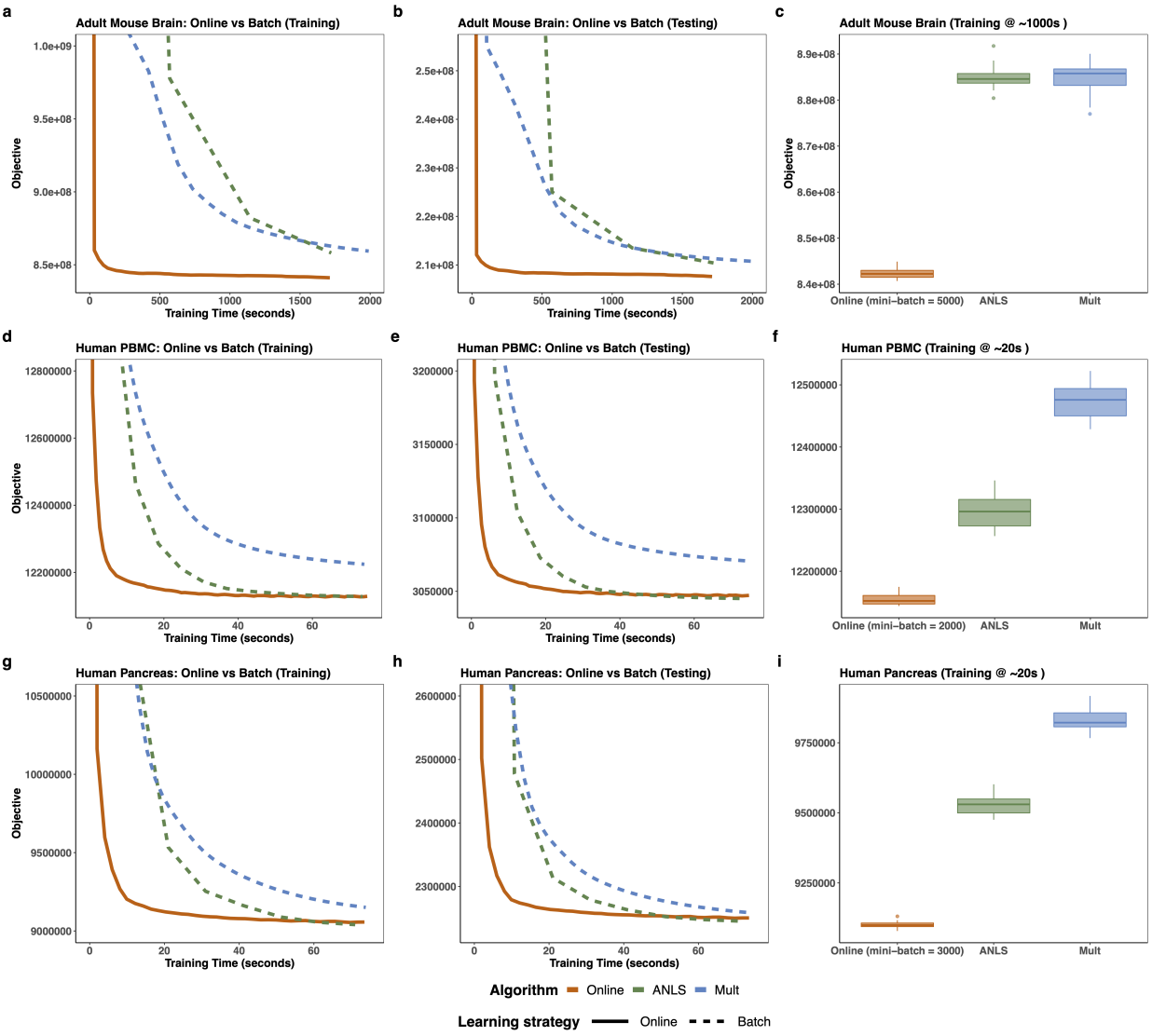


Figure 2.3: Convergence behavior for `Online iNMF` and batch iNMF algorithms on scRNA-seq data from the adult mouse brain, human PBMC and human pancreas. `Online iNMF` algorithm exhibits faster convergence and better objective minimization after a fixed amount of training time. The advantage of the online algorithm in convergence speed is more apparent for larger datasets. **(a-c)** Adult mouse brain ($n = 691,962$ cells, 9 individual datasets). **(d-f)** Human PBMCS ($n = 13,999$ cells, 2 individual datasets). **(g-i)** Human pancreas ($n = 14,890$ cells, 8 individual datasets). Center lines of box plots show the median; box limits, upper and lower quartiles; whiskers, $1.5\times$ interquartile range; and points are outliers.

Next we investigated whether `Online iNMF` yields similar dataset alignment and cluster preservation to our previously published alternating nonnegative least squares (ANLS) algorithm. (We refer to the ANLS algorithm as batch iNMF in subsequent discussions, to distinguish it from `Online iNMF`.) We applied both `Online iNMF` and batch iNMF to three scRNA-seq data collections, then visualized the factor loadings using UMAP plots (Figure 2.4). The `Online iNMF` algorithm yields visualizations that are qualitatively very similar to batch iNMF, suggesting nearly identical dataset alignment and accurate preservation of the original cluster structure for all three data collections.

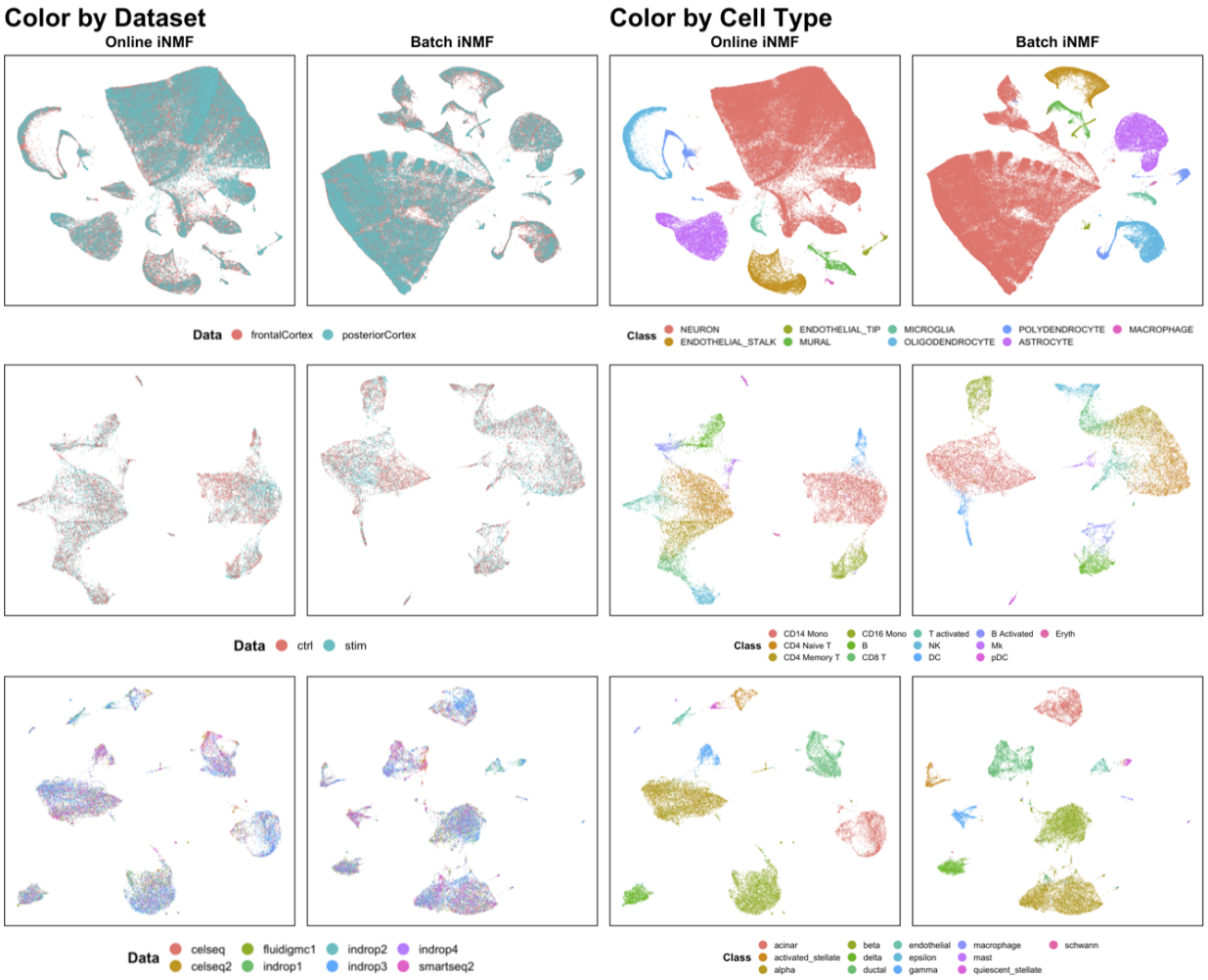


Figure 2.4: Online and batch iNMF yield highly similar UMAP visualizations. We performed `Online iNMF` and batch iNMF on data from mouse cortex ($n = 255,353$ cells), human PBMC ($n = 13,999$ cells), and human pancreas ($n = 14,890$ cells). `Online iNMF` and batch iNMF produce very similar visualizations, suggesting that the approaches give very similar dataset alignment and cluster preservation. We subsequently confirmed this qualitative observation using quantitative metrics.

2.3.2 `Online iNMF` Yields State-of-the-Art Single-Cell Data Integration Results Using Significantly Less Time and Memory

We next benchmarked `Online iNMF` (scenario 1) against batch iNMF (Welch et al. 2019) and two state-of-the-art single-cell data integration methods, Seurat v3 (Stuart et al. 2019) and Harmony (Korsunsky et al. 2019a). We selected these methods for comparison because a recent paper benchmarked 14 single-cell data integration methods and found that Harmony, Seurat, and LIGER consistently achieved the best dataset alignment and cluster preservation on a range of datasets (Tran et al. 2020).

To benchmark time and memory usage, we generated five datasets of increasing sizes (ranging from $10,000$ to $255,353$ cells in total) sampled from the same adult mouse frontal and posterior cortex data. Then we utilized them to compare the runtime and peak memory usage of `Online iNMF` (mini-batch size $= 5,000$) and the other methods (Figure 2.5a).

As expected, the runtime required for `Online iNMF` does not increase significantly as the dataset size grows, and the amount of memory needed for storing each minibatch is independent of the total number of cells. `Online iNMF` is also the fastest method overall, with Harmony the second fastest. Notably, the gap between Harmony and `Online iNMF` widens as the dataset size increases; on a dataset of 1.3 million cells from the mouse embryo, `Online iNMF` finishes dimension reduction in 25 minutes using 1.9 GB of RAM on a laptop, whereas Harmony requires 98 minutes and 109 GB of RAM on a large-memory server. Seurat and batch iNMF are significantly slower than `Online iNMF` and Harmony on the mouse cortex data, and the runtime of Seurat increases the most rapidly of any method.

Furthermore, the `Online iNMF` algorithm uses far less memory than any other approach, with memory usage primarily determined by mini-batch size, which is independent of the number of cells. Updating the factors with a mini-batch size of $5,000$ and $K = 40$ factors requires less than 500MB. In contrast, the memory requirements of batch iNMF, Harmony, and Seurat grow quickly with dataset size.

Next, we quantified the dataset alignment and cluster preservation performance for `Online iNMF` and the other methods (Figure 2.5b-c). Following the benchmarking strategy used by Tran et al. (2020), we assessed both the alignment performance (measured using two metrics) and cluster preservation performance (measured using two metrics). Our results show that `Online iNMF` performs as well as or better than the state-of-the-art methods. The online and batch iNMF algorithms align the PBMC (Kang et al. 2018) and pancreas (Grün et al. 2016, Muraro et al. 2016, Lawlor et al. 2017, Baron et al. 2016, Segerstolpe et al. 2016) datasets equally well, beating Harmony and Seurat. Furthermore, the online algorithm achieves scores close to batch iNMF on both data collections, confirming that the gain in computational efficiency does not come at the cost of accuracy in data embedding. The difference between iNMF and the other methods is especially

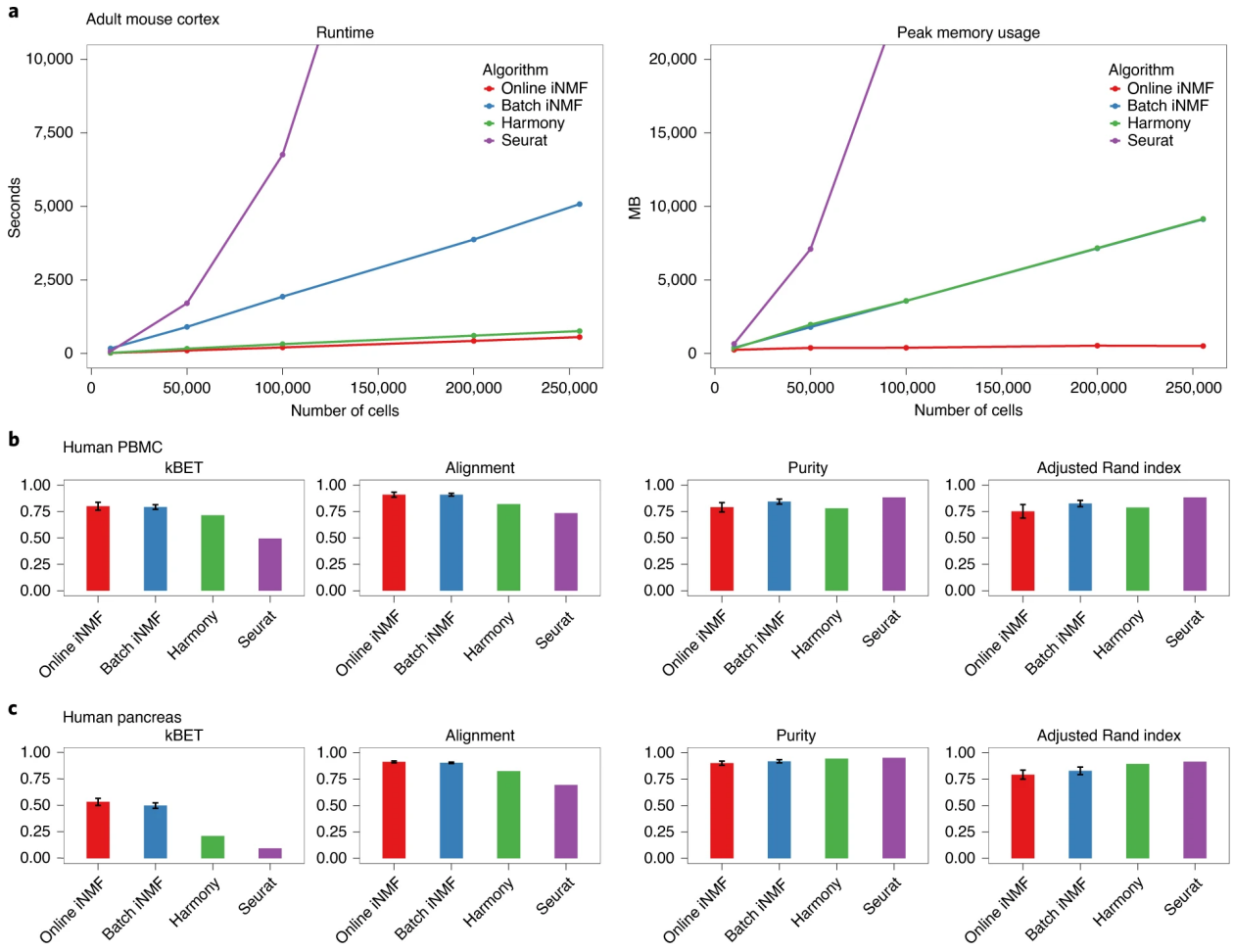


Figure 2.5: Benchmark of `Online iNMF`, batch iNMF, Harmony, and Seurat. The data are sampled from the adult mouse cortex ($n = 10,000, 50,000, 100,000, 200,000, 255,353$ cells, 2 individual datasets), human PBMC ($n = 13,999$ cells, 2 individual datasets) and human pancreas ($n = 14,890$ cells, 8 individual datasets). **(a)** The runtime and peak memory usage required for `Online iNMF`, batch iNMF, Harmony and Seurat to integrate the frontal and posterior cortex datasets. **(b,c)** Quantitative assessment of data integration and low-dimensional embedding carried out by four methods on the human PBMC and human pancreas datasets. Higher values are better for all 4 metrics. Error bars indicate standard deviation across 100 random initializations. The results from iNMF approaches (100 initializations each) are presented as mean values $\pm$ standard deviation, while Harmony and Seurat were only run once.

pronounced when comparing the values of kBET. We suspect that this difference occurs because our approach includes quantile normalization, which is stronger than the alignment strategies used by Harmony or Seurat. Consistent with our results, the benchmark of Tran et al. (2020) also included the pancreas dataset and found that LIGER (batch iNMF) gave substantially higher kBET values than competing methods (Tran et al. 2020). The online and batch iNMF algorithms produce comparable clustering results to the other approaches, although Harmony and Seurat give slightly higher cluster purity and adjusted rand index. This may be because the cluster labels we used for comparison are not real ground truth, but derived from PCA followed by clustering, which is more similar to the approaches used by Harmony and Seurat.

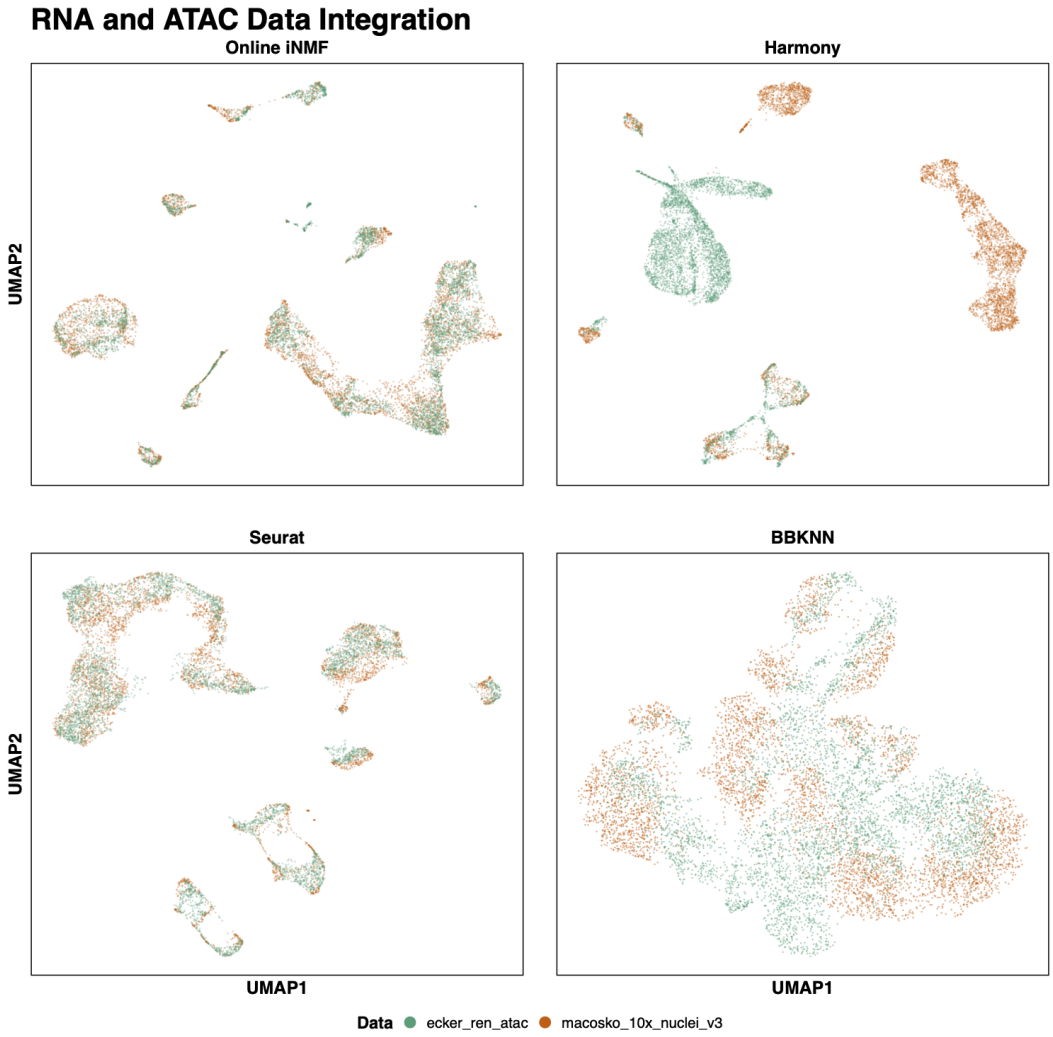


Figure 2.6: Benchmarking integration across data modalities (RNA+ATAC). $5,000$ cells from the snRNA-seq dataset and $5,000$ cells from the snATAC-seq dataset from MOP data collection were integrated using four different methods. The cells are exhibited in 2-dimensional UMAP space and colored by dataset.

We also compared the performance of `Online iNMF`, Seurat, Harmony and BBKNN when integrating two datasets of different modalities (Figure 2.6). Harmony and BBKNN showed inferior alignment, possibly because these approaches were not originally designed for multi-modal integration, unlike LIGER and Seurat. In contrast, both LIGER and Seurat produced UMAP visualizations indicating successful alignment of snRNA-seq and snATAC-seq data. Furthermore, the kBET and alignment metrics indicate that LIGER (alignment score $= 0.714$, kBET $= 0.574$) better integrates that datasets than either Seurat (alignment score $= 0.481$, kBET $= 0.231$) or Harmony (alignment score $= 0.113$, kBET $= 0.041$).

2.3.3 `Online iNMF` Rapidly Factorizes Large Datasets Using Fixed Memory

To demonstrate the scalability of our approach, we used `Online iNMF` (scenario 1) to analyze the scRNA-seq data of (Saunders et al. 2018), which contains $691,962$ cells sampled from nine regions (stored in nine individual datasets) spanning the entire mouse brain. Using `Online iNMF`, we factorized all of the datasets in 24 minutes on a MacBook Pro using about 1 GB of RAM. We note that the published analysis by Saunders et al. (2018) did not analyze all nine tissues simultaneously due to computational limitations, and that performing this analysis using our previous batch algorithm would have taken approximately 3.8 hours and 25 GB of RAM.

Cells within each class are well grouped together, and the distribution of neurons varies widely across regions, indicating neuronal subtypes specialized to different parts of the brain (Figure 2.7a). For example, neurogenic cells are identified predominantly in the hippocampus and striatum, consistent with reports of hippocampal and striatal neurogenesis in adult mammals Saunders et al. 2018, Toda et al. 2019, Ernst et al. 2014.

We used the factorization to group the cells into 40 clusters by assigning each cell to the factor on which it has the largest loading. We then examined differences in the regional proportions of each cell cluster. Neurons and oligodendrocytes show the most regional variation in composition, consistent with previous analyses (Zeisel et al. 2018). The total proportion of oligodendrocytes varies by region, but individual subtypes of oligodendrocytes are not region-specific, as expected. In contrast, individual subtypes of neurons are highly region-specific, reflecting diverse regional specializations in neuronal function (Figure 2.7b). We also investigated the biological properties of these cell factor loadings. Reassuringly, our cluster assignments largely represent subtypes within the broad cell classes and do not span class boundaries. As expected, neurons show by far the most diversity with eight subclusters. In contrast, ependymal cells, macrophages, microglia, and mitotic cells each correspond to only a single cluster (Figure 2.7c).

To further demonstrate the scalability of `Online iNMF`, we analyzed the mouse organogenesis cell atlas (MOCA) recently published by Cao et al. (2019).18. After filtering, MOCA contains $1,363,063$ cells from embryos between 9.5 to 13.5 days of gestation. We performed `Online`

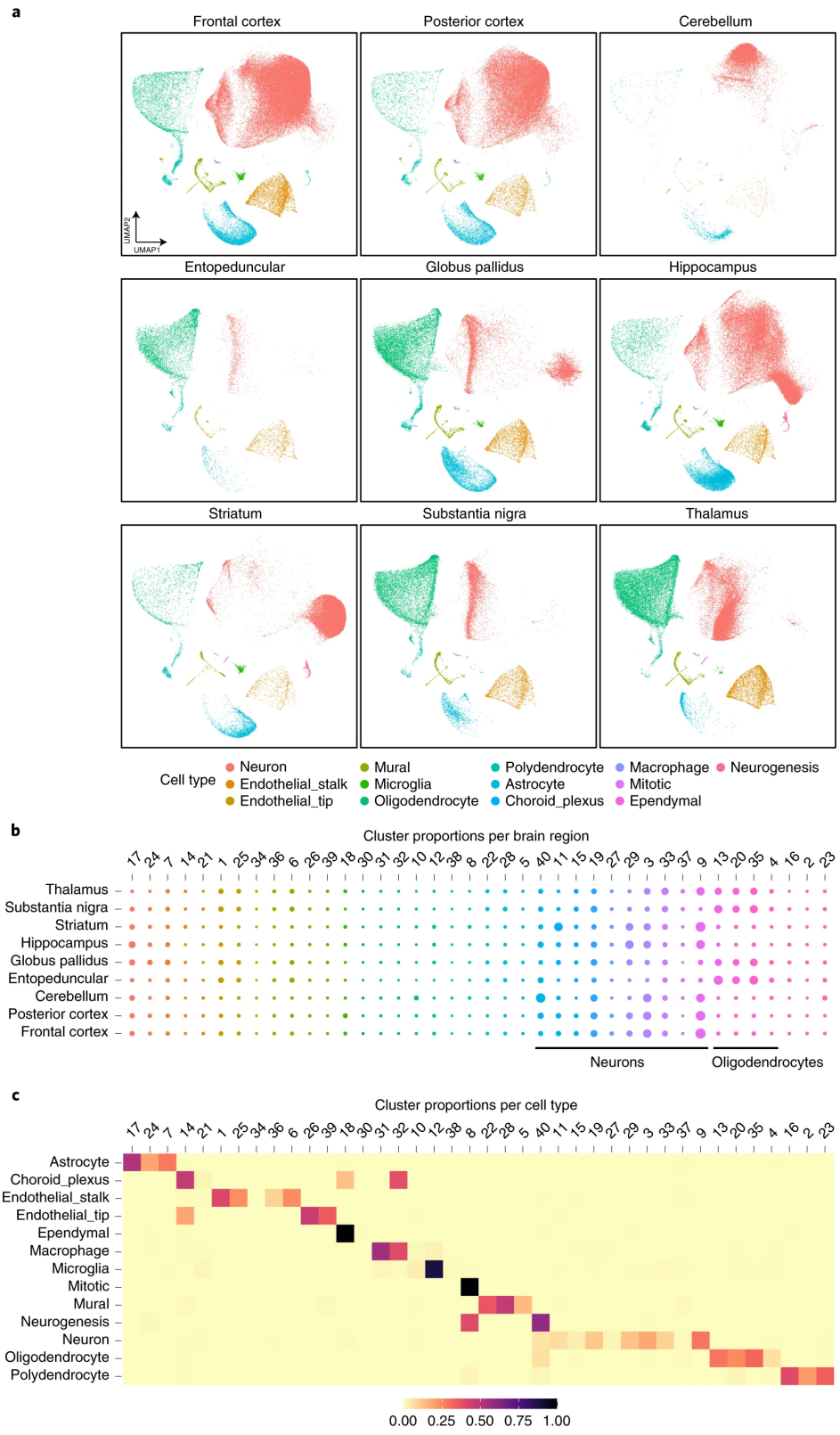


Figure 2.7: Joint analysis of nine regions of the adult mouse brain ($n = 691,962$ cells) using `Online iNMF`.

Joint analysis of nine regions of the adult mouse brain ($n = 691,962$ cells) using `Online iNMF`. **(a)** UMAP visualization of the iNMF factors learned for each brain region, colored by published cell class. **(b)** Dot plot showing the proportion of each of 40 clusters inferred from iNMF in each brain region. **(c)** Proportion of cells from each cluster in every cell type. The cells in each cluster mostly correspond to a single cell type.

`iNMF` on this dataset in 25 min using about 1.9 GB of RAM on a MacBook Pro. By comparison, we were not able to run Harmony on a laptop because of its high memory usage; running Harmony on a large-memory server required 98 minutes and 109 GB of RAM. Note that `Online iNMF`'s memory usage is higher for MOCA than for the mouse brain dataset primarily because of the higher value of K and a larger number of variable genes, not because of the number of cells. UMAP visualization shows that the cells from all five gestational ages are well aligned (Figure 2.17a), and the structure of 10 different developmental trajectories as defined by Cao et al. (2019) is also accurately preserved (Figure 2.17b).

Because `Online iNMF` processes only one mini-batch at a time, our approach allows processing datasets by streaming them over the internet instead of from disk. To demonstrate this capability, we created an HDF5 file containing the mouse cortex datasets ($n = 255,353$ cells), saved the file on a remote server, then read mini-batches directly over the internet. Processing the cortex dataset in this fashion took about 18 minutes, compared to around 6 minutes using local disk reads. This capability provides the unique advantage that many users can simultaneously analyze a single copy of a large cell atlas, without requiring each user to download and store the entire data collection.

2.3.4 `Online iNMF` Efficiently Integrates Large Single-Cell RNA and Spatial Transcriptomic Datasets

We next used `Online iNMF` to integrate single-cell RNA-seq and spatial transcriptomic datasets (Slide-seq and MERFISH). These spatial transcriptomic protocols provide spatial coordinates, but each has tradeoffs compared to scRNA-seq: Slide-seq may capture multiple cells on each barcoded bead and provides sparse transcriptome-wide measurements (Rodriques et al. 2019, Stickels et al.), and MERFISH measures only selected genes (Chen et al. 2015b). Integration with scRNA-seq data mitigates these limitations by incorporating deeper, transcriptome-wide data. Both spatial technologies can measure millions of cells, necessitating scalable methods for integration.

We used `Online iNMF` in scenario 3 to project Slide-seq data from mouse hippocampus ($59,858$ beads) onto a large single-cell RNA-seq dataset ($193,155$ cells) (Rodriques et al. 2019, Yao et al. 2021). Each Slide-seq bead may contain transcripts from more than one cell; thus, identifying H^i using W serves as a "deconvolution" operation in this case (Rodriques et al. 2019). The original Slide-seq paper performed a similar analysis using conventional nonnegative matrix factorization of single-cell RNA-seq data (Rodriques et al. 2019). Consistent with the published analysis, we

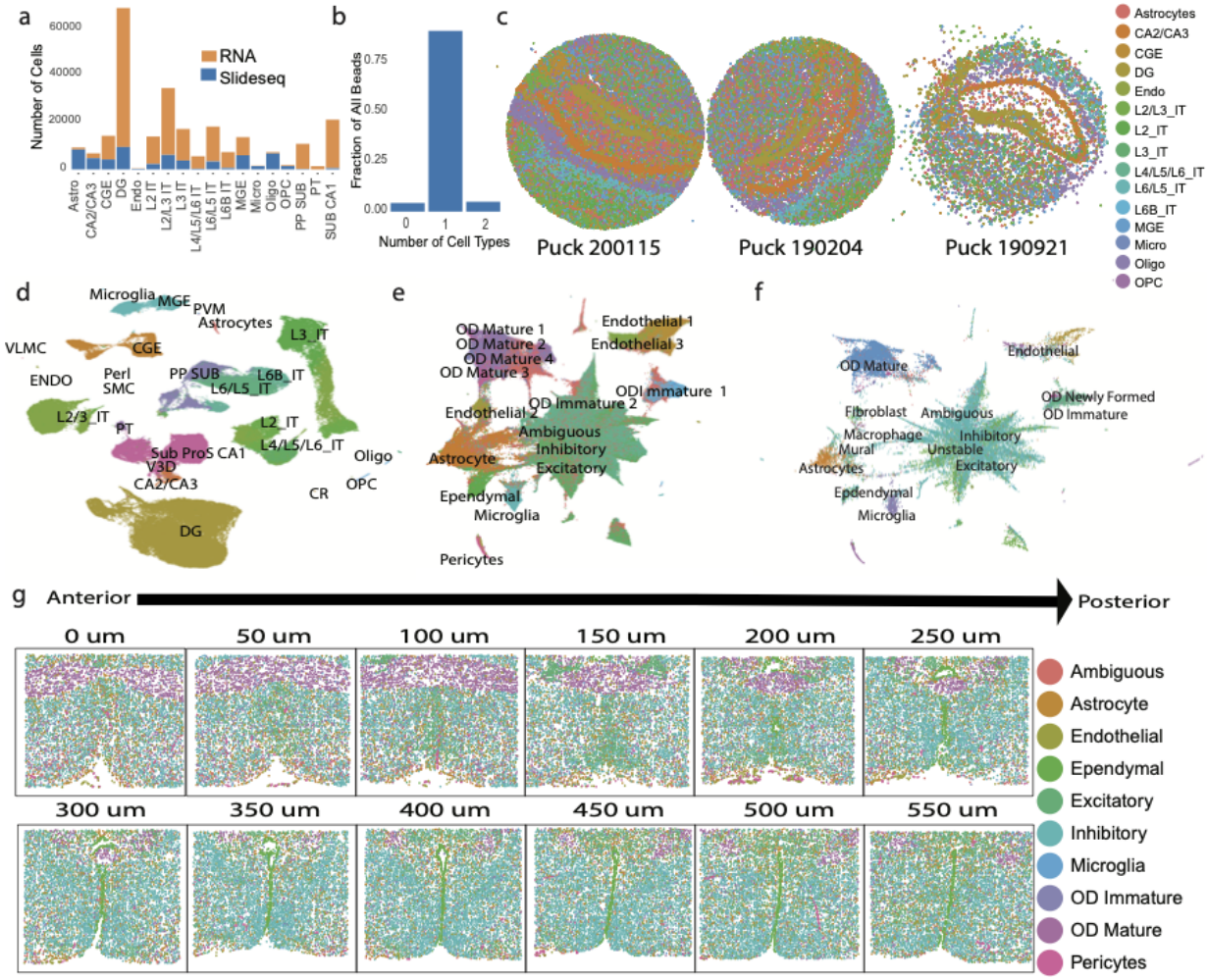
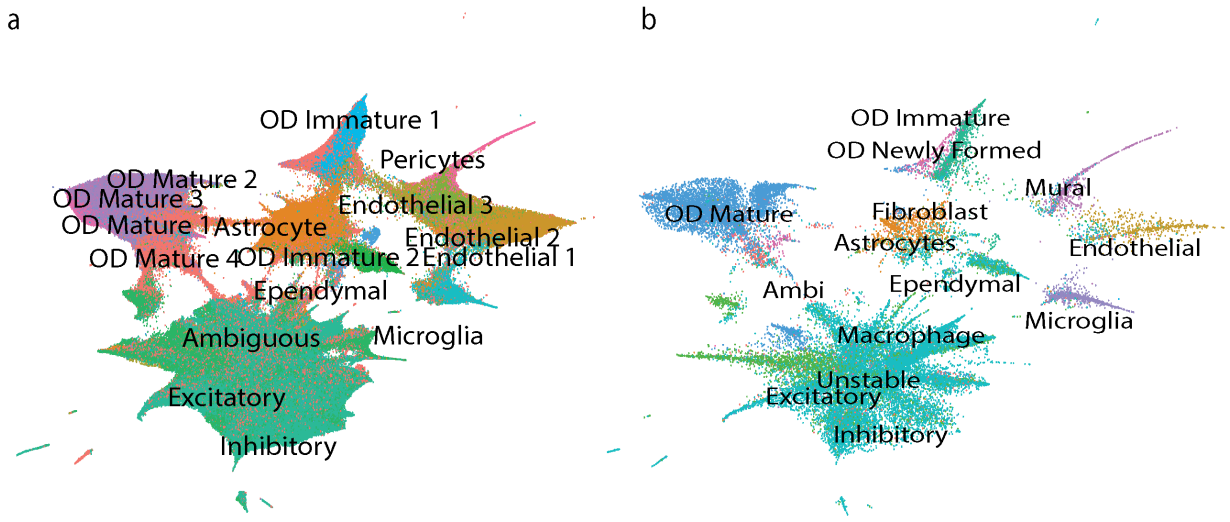


Figure 2.7: `Online iNMF` integrates large single-cell RNA-seq and spatial transcriptomic datasets. (a) The number of cells per cell type in scRNA-seq ($n = 193,155$ cells) and Slide-seq ($n = 59,858$ beads) datasets from mouse hippocampus. **(b)** Number of cell types assigned to each bead in the Slide-seq analysis. **(c)** Slide-seq beads colored by labels derived from projection onto scRNA-seq data using `Online iNMF` (scenario 3). The coordinates of each bead reflect its spatial position within the tissue. **(d)** UMAP plot of cell factor loadings (`Online iNMF`, scenario 1) for scRNA-seq data from mouse hippocampus. **(e)** UMAP plot of MERFISH cells from mouse hypothalamus ($n = 1,026,840$ cells), colored by published cluster assignments. The UMAP coordinates are derived from `Online iNMF` (scenario 3) integration of MERFISH and scRNA-seq data. **(f)** UMAP plot of scRNA-seq cells from mouse hypothalamus ($n = 31,250$ cells), colored by published cluster assignments. The UMAP coordinates are derived from `Online iNMF` (scenario 3) integration of MERFISH and scRNA-seq. **(g)** MERFISH slices, ordered from anterior to posterior, colored by labels derived from the `Online iNMF` integration. The coordinates of each cell reflect its spatial position within the tissue.

found that most Slide-seq beads contained a single dominant cell type, though a small number contained two cell types or no clear cell types (Figure 2.7b). Overall, the proportions of cell types were consistent across technologies, except that the scRNA-seq data contained fewer non-neurons, because the cells were experimentally enriched for neurons (Figure 2.7a). The spatial distributions of our annotated cell types reflect the known organization of the hippocampus, with Ammon's horn, dentate gyrus, white matter, part of the ventricles, and adjacent deep cortical layers clearly visible (Figure 2.7c). Thus, this integration reveals the spatial distributions of the clusters from the scRNA-seq data (Figure 2.7d).

We also used `Online iNMF` (scenario 1 and 3) to integrate MERFISH ($n = 1,026,840$ cells) and scRNA-seq ($n = 31,250$ cells) data from the preoptic region of mouse hypothalamus (Moffitt et al. 2018). Scenario 1 and scenario 3 gave very similar results (Figure 2.82). This integration analysis revealed the correspondence between scRNA-seq and MERFISH clusters (Figure 2.7e-f), which had been analyzed only separately in the original publication. The spatial distributions of our joint clusters accord well with the known structure of the hypothalamus (Figure 2.7g).



	Alignment	kBET	MERFISH purity	RNA Purity	MERFISH ARI	RNA ARI
Scenario 1	0.72	0.53	0.807	0.784	0.6	0.569
Scenario 3	0.74	0.56	0.802	0.781	0.64	0.57

Figure 2.8: Scenario 1 and scenario 3 achieve similar results on MERFISH data. The result of scenario 1 on the MERFISH dataset yielded similar MERFISH **(a)** and RNA **(b)** cluster placement as using scenario 3 (Figure 2.7). The performance of the different approaches showed remarkable similarity, as demonstrated by the alignment, kBET, cluster purity, and ARI scores shown in the table.

2.3.5 `Online iNMF` Enables Iterative Refinement of Single-Cell Multi-Omic Atlas from Mouse Motor Cortex

One of the most appealing properties of our online learning algorithm is the ability to incorporate new data points as they arrive. This capability is especially useful for large, distributed collaborative efforts to construct comprehensive cell atlases (Ecker et al. 2017, lcai@ caltech. edu 21 b Shendure Jay 9 Trapnell Cole 9 Lin Shin shinlin@ uw. edu 2 e Jackson Dana 9 et al. 2019, Regev et al. 2017). Such cell atlas projects involve multiple research groups asynchronously generating experimental data with constantly evolving protocols, making the ultimate cell type definition a moving target.

To demonstrate the utility of `Online iNMF` for iteratively refining cell type definitions, we used data generated by the BRAIN Initiative Cell Census Network (BICCN) (Yao et al. 2020). During a pilot phase starting in 2018, the BICCN generated single-cell datasets from a single region of mouse brain (primary motor cortex, MOp) spanning 4 modalities (single-cell RNA-seq, single-nucleus RNA-seq, single-nucleus ATAC-seq, single-nucleus methylcytosine-seq) and totaling $786,605$ cells.

Following scenario 2 (Figure 2.1c), we used `Online iNMF` to incorporate the MOp datasets in chronological order, refining the factorization with each additional dataset (Figure 2.9). Our approach successfully incorporated each new single-cell or single-nucleus RNA-seq dataset without revisiting previously processed cells, using each cell exactly once during the optimization process (Figure 2.9a). UMAP visualizations indicate that the structure of the datasets is iteratively refined with each successive dataset that is added. We jointly identified 15 cell types from the transcriptomic and epigenomic datasets (Figure 2.9d). Alignment and kBET metrics also indicate that the datasets are well aligned (Alignment score $= 0.786$, kBET $= 0.324$). To put these numbers in context, Seurat achieved scores of 0.481 and 0.231 on a simpler integration analysis of one scRNA-seq and one snATAC-seq dataset (Figure 2.6).

The results from performing this single-cell multi-omic integration are very similar whether the integration is performed iteratively (scenario 2), using all of the data at once (scenario 1), or by projecting the epigenomic data onto the transcriptomic data (scenario 3; 2.10). We also confirmed that scenario 2 is robust to the order of dataset arrival. To do this, we inspected the effect of random initializations and orderings of the input datasets on the iterative multi-omic integration (scenario 2). We integrated all eight datasets in their original order using 10 different initializations as well as five different orderings where each of the other sc/snRNA- seq datasets served as the first input. With our annotations as the reference, different orderings result in comparable variation in final cluster assignments compared to the variation from random initialization (average ARI $= 0.759$ from random input orders vs. 0.744 from random initializations).

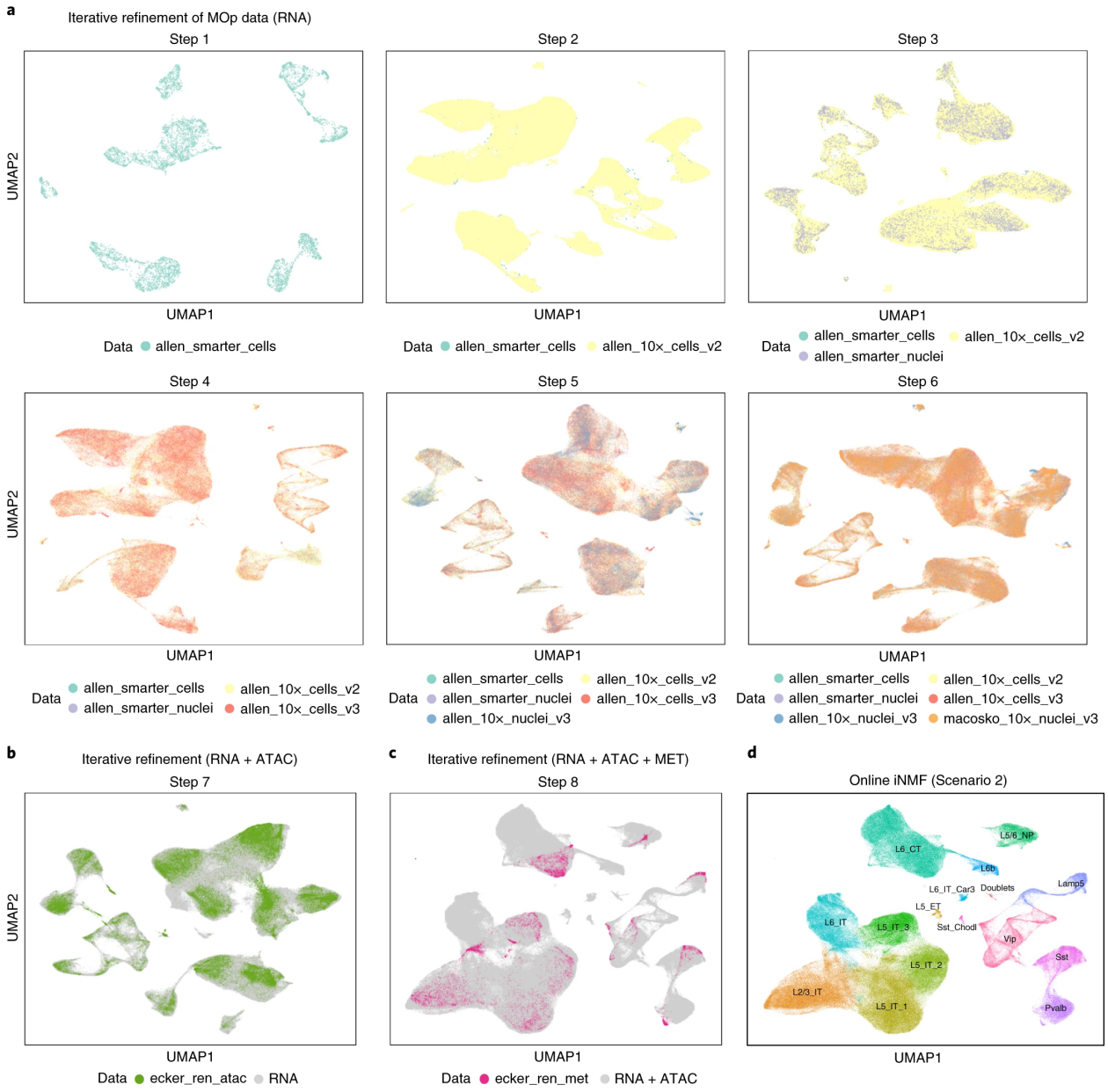


Figure 2.9: Iterative refinement of cell identity using multiple single-cell modalities from the mouse primary motor cortex. We integrated four scRNA-seq datasets, two snRNA-seq datasets, one snATAC-seq dataset and one snmC-seq dataset ($n = 408,885$ neurons). **(a)** Sequential integration of six scRNA-seq datasets (scenario 2). Each panel shows a UMAP plot using cell factors obtained after adding an additional dataset. **(b)** UMAP plot of cell factors obtained by adding snATAC-seq to the latent space learned from six RNA datasets in a (scenario 2). **(c)** UMAP plot of cell factors obtained by adding DNA methylation data (snmC-seq) to the latent space learned from the seven datasets shown in b (scenario 2). **(d)** Clusters obtained using the cell factor loadings of all eight aligned datasets. The clusters were named using marker genes from Tasic et al. (2016).

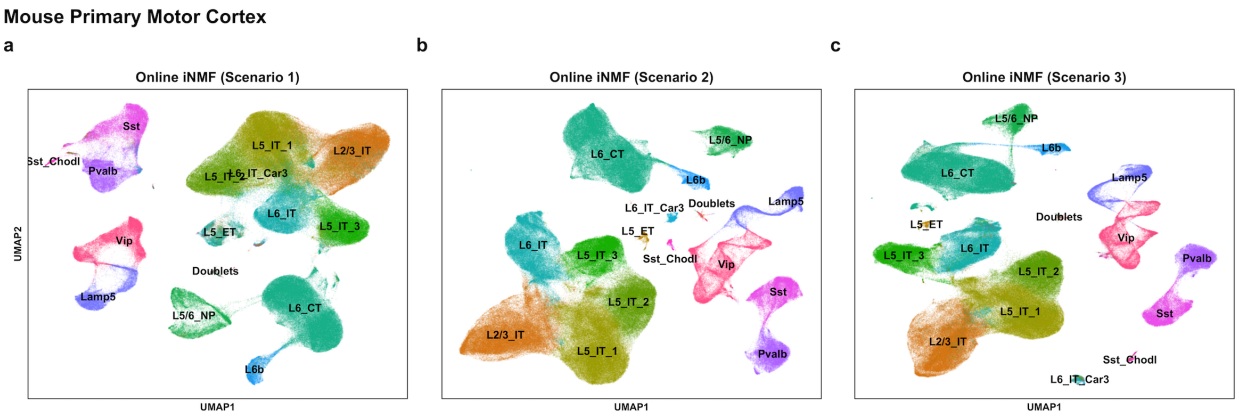


Figure 2.10: Performing `Online iNMF` in three scenarios produces similar results. These analyses were carried out separately to integrate 8 MOp datasets (scRNA-seq, snRNA-seq, snATAC-seq and snmC-seq, $n = 408,885$) using `Online iNMF` in scenario 1 **(a)**, scenario 2 **(b)**, and scenario 3 **(c)**. The results are visualized in UMAP coordinates and the cells are colored by the cell type annotations from Figure 2.9.

2.4 Discussion

By reading mini-batches from disk, `Online iNMF` not only converges faster than batch approaches, but also decouples memory usage from dataset size. The efficiency gains of `Online iNMF` will be even greater as the scale of single-cell datasets increases.

We envision `Online iNMF` enabling iterative single-cell data integration in three different scenarios. In scenario 1, when all single-cell datasets are currently available, the `Online iNMF` algorithm rapidly factorizes the single-cell data into metagenes and cell factor loadings using multiple epochs of training. In scenario 2, the online algorithm iteratively incorporates single-cell datasets as they arrive sequentially. We anticipate that scenario 2 will prove useful as researchers continually incorporate newly sequenced cells to build comprehensive cell atlases. Scenario 3 holds great promise for rapidly querying datasets against a large, curated reference atlas.

We anticipate that `Online iNMF` will become increasingly useful for integrating single-cell multi-omic datasets of growing scale from projects such as the BRAIN Initiative, Human Body Map, and Human Cell Atlas.

2.5 Methods

2.5.1 About `Online iNMF`

2.5.1.1 Utility of `Online iNMF`

In this study, we extend the online NMF approach of Mairal et al. (2010) to make it suitable for iNMF. `Online iNMF` provides two significant advantages: (1) integration of large multi-modal datasets by cycling through the data multiple times in small mini-batches and (2) integration of continually arriving datasets, where the entire dataset is not available at any point during training (Figure 2.1).

We envision using `Online iNMF` to integrate single-cell datasets in three different scenarios (Figure 2.1). We note that our `Online iNMF` approach is distinct from stochastic gradient descent (SGD), a general optimization technique that can be used for a range of objective functions. Instead of employing SGD, we have derived an online learning algorithm specifically tailored to the iNMF objective function. Our approach has two key advantages compared to SGD: (1) SGD requires choosing a data-dependent schedule of learning rates that vary over the whole learning process, while our approach does not involve a learning rate parameter at all and (2) we use optimization techniques that leverage the unique structure of the iNMF optimization problem, allowing theoretical convergence guarantees and fast empirical convergence. Mairal et al. (2010) explain this distinction in more detail.

2.5.1.2 Derivation of iNMF Updates

iNMF takes N single-cell multi-omic datasets $X^1, \ldots, X^N$ as input. After normalization, gene selection (m variable genes selected) and scaling, we have the preprocessed input data $X^i \in \mathbb{R}_+^{m \times n_i} (i = 1, ..., N)$ The goal is to find the shared and dataset-specific factors (metagenes) $W \in \mathbb{R}_+^{m \times K}$, $V^i \in \mathbb{R}_+^{m \times K}$ and $H^i \in \mathbb{R}_+^{n_i \times K} (i = 1, ..., N)$ that minimize the following empirical cost of the iNMF problem, given parameters K and λ.

$$\min_{\substack{W, V^i, H^i \geq 0 \\ i=1,\cdots,N}} \sum_{i=1}^{N} (\left\| X^i - (W + V^i) H^{i\top} \right\|_F^2 + \lambda \left\| V^i H^{i\top} \right\|_F^2) \tag{2.1}$$

For given W and V^i, we update H^i by numerically solving a nonnegative least squares problem:

$$H^i = \arg\min_{H \geq \mathbf{0}} \left\| \begin{pmatrix} W + V^i \\ \sqrt{\lambda} V^i \end{pmatrix} H^\top - \begin{pmatrix} X^i \\ \mathbf{0}^{m \times n_i} \end{pmatrix} \right\|_F^2 \tag{2.2}$$

We derived hierarchical alternating least squares (HALS) updates to calculate W and V^i, holding

the other two matrix blocks fixed:

$$W^*_{\cdot j} = \left[W_{\cdot j} + \frac{\sum_i (X^i H^i)_{\cdot j} - (W + V^i)(H^{i\top} H^i)_{\cdot j}}{\sum_i (H^{i\top} H^i)_{jj}} \right]_+$$
$$V^{i*}_{\cdot j} = \left[V^i_{\cdot j} + \frac{(X^i H^i)_{\cdot j} - (W + (1+\lambda) V^i)(H^{i\top} H^i)_{\cdot j}}{(1+\lambda)(H^{i\top} H^i)_{jj}} \right]_+ \tag{2.3}$$

See **Supplementary Note** (Available online) for detailed derivation of HALS updates.

2.5.1.3 Optimizing a Surrogate Function for iNMF

We developed an online learning algorithm for integrative nonnegative matrix factorization by adapting a previously published strategy for online dictionary learning (Mairal et al. 2010). The key innovation that makes it possible to perform online learning is to optimize a "surrogate function" that asymptotically converges to the same solution as the empirical iNMF cost. In the NMF problem with a sparsity penalty (e.g. L1 regularization), we want to find the nonnegative factors $W \in \mathbb{R}_+^{m \times K}$, $H \in \mathbb{R}_+^{n \times K}$ that optimally reconstruct the input $X \in \mathbb{R}_+^{m \times n}$ (n data points) by minimizing the following empirical cost function:

$$f_n(W) = \frac{1}{n} \sum_{s=1}^{n} \ell(\boldsymbol{x}_s, W) \tag{2.4}$$

$$\ell(\boldsymbol{x}_s, W) = \min_{h \geq 0} \sum_{s=1}^{N} (\left\| \boldsymbol{x_s} - W \boldsymbol{h_s}^\top \right\|_2^2 + \lambda \left\| \boldsymbol{h_s}^\top \right\|_1) \tag{2.5}$$

where $\boldsymbol{x_s}$ is the sth data point and h represents a row of H. The goal is to minimize the expected cost:

$$f(W) = \mathbb{E}_x[\ell(\boldsymbol{x}, W)] = \lim_{n \to \infty} f_n(W) \tag{2.6}$$

Assuming we randomly sample a data point $\boldsymbol{x}^{(t)}$ at the tth iteration, the original Mairal paper proved that the following surrogate function $\hat{f}_T(W)$ converges almost surely to $f_T(W)$ (and to a local minimum) as $T \to \infty$:

$$\hat{f}_t(W) = \frac{1}{T} \sum_{t=1}^{T} (\left\| \boldsymbol{x}^{(t)} - W \boldsymbol{h}^{(t)\top} \right\|_2^2 + \lambda \left\| \boldsymbol{h}^{(t)\top} \right\|_1) \tag{2.7}$$

where $\boldsymbol{x}^{(t)}$, W, $\boldsymbol{h}^{(t)}$ are nonnegative and T is the total number of iterations. Mairal et al. derived an online learning algorithm that performs NMF by updating h and W in an alternating fashion. They first solve for $\boldsymbol{h}^{(t)}$ using $W^{(t-1)}$ from the previous iteration and then obtain $W^{(t)}$ that minimizes the surrogate function. Intuitively, this strategy allows online learning because it expresses a formula

for incorporating a new observation $\boldsymbol{x}^{(t)}$ given the factorization result W and $\boldsymbol{h}$ for previously seen data points. Thus, we can iterate over the data points one-by-one or in small mini-batches.

In the proposed `Online iNMF` algorithm, we process the data in mini-batches, which improves convergence speed. Assuming we have data matrices $X^i \in \mathbb{R}_+^{m \times n_i} (i = 1, \ldots, N)$ and mini-batch $X_M^{(t)}$ of size p, where $X_M^{(t)}$ comprises data points $X_M^{i(t)}$ sampled from X^i, the empirical cost of iNMF is given by:

$$\min_{\substack{W, V^i, H^i \geq 0 \\ i=1,\cdots,N}} \frac{1}{\sum_i^N n_i} \sum_{i=1}^{N} (\left\| X^i - (W + V^i) H^{i\top} \right\|_F^2 + \lambda \left\| V^i H^{i\top} \right\|_F^2) \tag{2.8}$$

The corresponding surrogate function after the Tth iteration is:

$$\hat{f}_t(W, V^1, \cdots, V^N) = \frac{1}{T \times p} \sum_{t=1}^{T} \sum_{i=1}^{N} (\left\| X_M^{i(t)} - (W + V^i) H_M^{i(t)\top} \right\|_F^2 + \lambda \left\| V^i H_M^{i(t)\top} \right\|_F^2) \tag{2.9}$$

where subscript M indicates a sampled mini-batch. For a new mini-batch $X_M^{(t)}$, we first compute the corresponding cell factor loadings $H_M^{i(t)}$ for all input data using the shared $(W^{(t-1)})$ and dataset-specific $(V^{i(t-1)})$ factors from the last iteration. The authors of the original online learning paper employed the least angle regression algorithm (LARS) in their study. Here we use the ANLS update instead because it is highly efficient, designed specifically for NMF (rather than dictionary learning in general) and addresses the subproblem by running the solver exactly once within a single iteration of the `Online iNMF` algorithm. We also tried using a HALS update for $H_M^{i(t)}$, but found that convergence was slower (Figure 2.11). Upon acquiring $H_M^{i(t)}$, we utilize the HALS method to update the shared $W^{(t)}$ and $V^{i(t)}$, which is analogous to the updates used by Mairal et al. (2010) but derived specifically for iNMF. Because the updates for W and V^i depend on all of the previously seen data points and their cell factor loadings, a naive implementation would require storing all of the data and cell factor loadings in memory. However, the HALS updates depend on X^i and H^i only through the matrix products $H^{i\top} H^i$ and $X^i H^i$ (see Supplementary Note for details). These matrix products have only K^2 and mK elements respectively, allowing efficient storage, and can be computed incrementally with the incorporation of each newly sampled mini-batch $X_M^{i(t)}$ of size p_i:

$$\begin{aligned} A^{i(t)} &= A^{i(t-1)} + \frac{1}{p^i} H_M^{i(t)\top} H_M^{i(t)} \\ B^{i(t)} &= B^{i(t-1)} + \frac{1}{p^i} X_M^{i(t)} H_M^{i(t)} \end{aligned} \tag{2.10}$$

Note that, analogous to the mini-batch extension of the original online dictionary learning

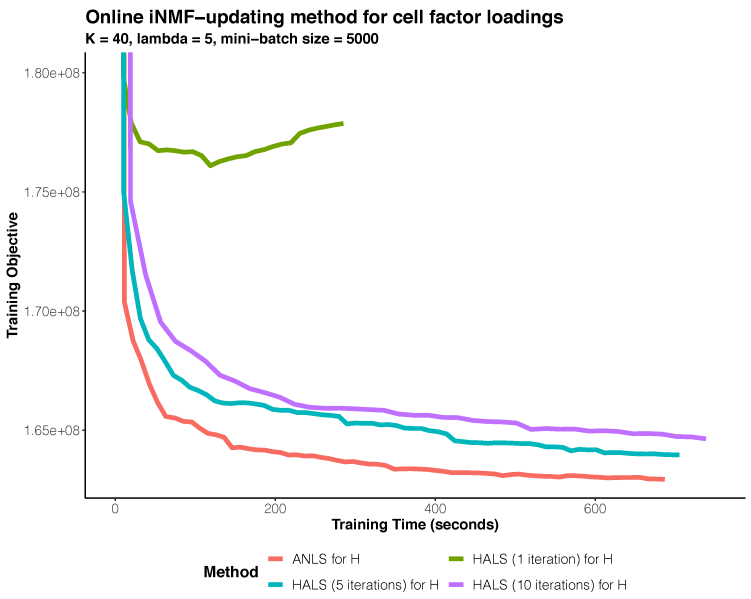


Figure 2.11: Comparison of methods for updating cell factor loadings (H). The training data are subsets (80%) of the adult mouse frontal ($n = 124,934$) and posterior cortex ($n = 79,349$) datasets. $1,111$ were selected variable genes for this analysis. ANLS for H clearly outperforms the other in minimizing the objective.

algorithm, we divide by p_i to average the inner products across all data points within each mini-batch.

2.5.1.4 Implementation of `Online iNMF`

Algorithm 2.1 summarizes our implementation of `Online iNMF`. We use our previous `Rcpp` implementation of the block principal pivoting algorithm5 to calculate the ANLS updates for $H_M^{i(t)}$. We implement the HALS updates for W and V^i using native R, since the updates require only matrix operations, which are highly optimized in R. Because the online algorithm does not require all of the data on each iteration (only a fixed-size mini-batch), we use the hdf5r package to load each mini-batch from disk on the fly. By creating HDF5 files with chunk size no larger than the mini-batch size, we achieve a time- and memory-efficient implementation that never loads more than a single mini-batch of the data from disk at once. In fact, we can go a step further and analyze datasets that are not stored on the same physical hard drive as the machine performing iNMF. We show that it is possible to analyze data by streaming over the internet without downloading the entire dataset onto the disk.

For scenario 1, in which the mini-batch size p specifies the total number of cells to be processed per iteration across all datasets, we sample p^i cells from each dataset i, proportional to its full dataset size ($p_i = p \times n_i / \sum_i^N n_i$). Thus, each mini-batch in scenario 1 contains a representative

Algorithm 2.1: Online Learning for Integrative Nonnegative Matrix Factorization

Data: $X^i \in \mathbb{R}_+^{m \times n_i}, i = 1, \cdots, N$

Initialize $A^{i(0)} \in \mathbf{0}^{K \times K}$, $B^{i(0)} \in \mathbf{0}^{M \times K}$, $i = 1, \ldots, N$;

Initialize $W^{(0)}$ with random samples from a uniform distribution over $[0, 2]$;

Initialize $V^{i(0)}$ with random samples from X^i, $i = 1, \ldots, N$;

for $t = 1 \to T$ **do**

  for $i = 1 \to N$ **do**

    Sample a mini-batch $X_M^{i(t)}$ of size p^i from X^i, $i = 1, \cdots, N$;

    Compute $H_M^{i(t)}$ using ANLS, $i = 1, \cdots, N$;

    $H_M^{i(t)} = \arg\min_{H \geq \mathbf{0}} \left\| \begin{pmatrix} W^{(t-1)} + V^{i(t-1)} \\ \sqrt{\lambda} V^{i(t-1)} \end{pmatrix} H^\top - \begin{pmatrix} X_M^{i(t)} \\ \mathbf{0}_{m \times p^i} \end{pmatrix} \right\|_F^2$;

    Update $A^{i(t)}$ and $B^{i(t)}$ (remove old information older than 2 epochs);

    $A^{i(t)} \leftarrow \beta^{(t)} A^{i(t-1)} + \frac{1}{p^i} H_M^{i(t)\top} H_M^{i(t)}$;

    $B^{i(t)} \leftarrow \beta^{(t)} B^{i(t-1)} + \frac{1}{p^i} X_M^{i(t)\top} H_M^{i(t)}$;

  end

  Initialize $W^{(t)} = W^{(t-1)}$;

  for $j = 1 \to K$ **do**

    $W_{\cdot j}^{(t)} = \left[W_{\cdot j}^{(t)} + \frac{\sum_i B_{\cdot j}^{i(t)} - (W^{(t)} + V^{i(t-1)}) A_{\cdot j}^{i(t)}}{\sum_i A_{jj}^{i(t)}} \right]_+$;

  end

  Initialize $V^{i(t)} = V^{i(t-1)}, i = 1, \ldots, N$;

  for $j = 1 \to K$ **do**

    $V_{\cdot j}^{i(t)} = \left[V_{\cdot j}^{i(t)} + \frac{B_{\cdot j}^{i(t)} - (W^{(t)} + (1+\lambda) V^{i(t)}) A_{\cdot j}^{i(t)}}{(1+\lambda) A_{jj}^{i(t)}} \right]_+$;

  end

end

Compute $H^{i(T)}$ using ANLS, $i = 1, \ldots, N$;

return $W^{(T)}, V^{i(T)}, H^{i(T)}, i = 1, \ldots, N$.

sample of cells from all datasets. For scenario 2, in which only one dataset is available at a time, we sample the entire mini-batch from the current dataset. We also employ three heuristics that were used in the original online NMF paper: (1) we initialize the dataset-specific metagenes using K cells randomly sampled from the corresponding input data; (2) we downscale $A^{i(t-1)}$ and $B^{i(t-1)}$ when obtaining $A^{i(t)}$ and $B^{i(t)}$ using $H_M^{i(t)}$; and (3) we remove information older than two epochs from matrices $A^{i(t)}$ and $B^{i(t)}$ (only once at the start of a new epoch, exclusive to scenario 1 in practice). The intuition behind the second and third heuristics is as follows. By design, $A^{i(t)}$ and $B^{i(t)}$ carry all the $H_M^{i(t)\top} H_M^{i(t)}$ and $X_M^{i(t)} H_M^{i(t)}$ values respectively from t iterations. Each time when the same data points are revisited (assuming t iterations comprise multiple epochs), the accuracy of resulting cell factor loadings is improved because the metagene factors get refined during the implementation of the algorithm. Consequently, the variability in the quality of cell factor loadings is carried over to $A^{i(t)}$ and $B^{i(t)}$ by summing up matrix products shown above. Therefore, by downscaling $A^{i(t-1)}$ and $B^{i(t-1)}$(old information), the weight of the latest $H_M^{i(t)\top} H_M^{i(t)}$ and $X_M^{i(t)} H_M^{i(t)}$ increases. Mairal et al. (2010) observed faster convergence of online learning on small datasets by removing the matrix product involving the less-refined cell factor loadings and thus they adopted this heuristic in their online learning implementation. An example of applying heuristic (2) and (3) for $A^{i(t)}$ is shown in algorithm 2.2 (the same strategy applies to $B^{i(t)}$).

Algorithm 2.2: Example of Heuristic (2) and (3)

if *3rd epoch starts at tth iteration* ($t \geq 3$) **then**
  $A^{i(t-1)} \leftarrow A^{i(t-1)} - A^{i(t-2)}$ `// Remove old information`
  $\beta^{(t)} = \frac{t-2}{t-1}$;
  $A^{i(t)} \leftarrow \beta^{(t)} A^{i(t-1)} + \frac{1}{p^i} H_M^{i(t)\top} H_M^{i(t)}$ `// Downscale old information`
end

Additionally, we implemented dataset preprocessing—including library size normalization, variable gene selection, and gene scaling—using fixed-size mini-batches, so that preprocessing requires only a prespecified amount of memory.

2.5.2 Data Loading Methods and Overhead

To investigate whether loading data from disk causes significant overhead, we ran `Onine iNMF` (scenario 1) with 1,111 variable genes on the mouse cortex datasets stored either on disk or in memory. Then we implemented both approaches with different choices of mini-batch size ($n = 1,000, 5,000, 10,000, 50,000$) for 50 iterations, while keeping the other parameters the same ($K = 40$, $\lambda = 5$). The average runtime for 50 iterations for each setting is reported in the barplot. The standard deviation is displayed as error bars (Figure 2.12).

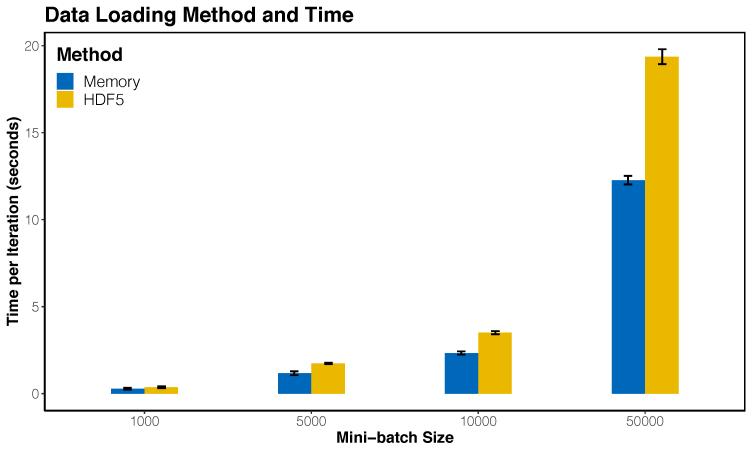


Figure 2.12: Reading mini-batches from disk adds minimal overhead. In this study, each chunk in HDF5 files stores $1,000$ samples (cells). Pulling data from the disk does not add significant overhead compared to loading the data from memory, as long as the mini-batch size is close to the specified chunk size. Mean time per iteration (processing one mini-batch) ($\pm$SD) of 50 iterations in each setting is displayed.

2.5.3 Quantile Normalization and Joint Clustering

We also implemented a much more efficient strategy for quantile normalization (See Algorithm 2.3) than our previously published approach (Welch et al. 2019). We found that, rather than performing time- and memory-intensive shared factor neighborhood clustering to identify joint clusters, we can perform the following steps: (1) assign each cell to the factor on which has the highest loading, giving a number of joint clusters equal to the number of metagene factors K. Note that one can center the cell factor loadings at first if the distribution of cell factor loadings from a given dataset significantly differs from the others (e.g. due to different data modalities); (2) for each input dataset, efficiently find approximate within-dataset nearest neighbors using the RANN package (k-nearest neighbors k = 20 and $\epsilon = 0.9$ by default) and then correct these maximum factor assignments by taking a majority vote among within-dataset nearest neighbors; and (3) perform quantile normalization on the refined joint clusters as before (Welch et al. 2019). By default, we choose the dataset with the largest number of cell samples as the reference dataset. Then, for cells from each of the joint clusters, we normalize the quantiles of the factor loadings for each metagene factor in the other datasets to match the quantiles of the factor loadings for the same metagene in the reference dataset. This strategy performs just as well as shared factor neighborhood clustering, but uses significantly less time and memory. Unless otherwise specified, we implemented quantile normalization with $k = 20$ (default) for k-nearest neighbors in analyses of both real and simulated datasets (note that k is denoted as Q in algorithm 2.3).

After performing quantile normalization, one can perform a second clustering step (e.g., Louvain

community detection) using the normalized cell factor loadings H^i (or unnormalized H^i if the data are aligned well even without quantile normalization).

2.5.4 Quantitative Metrics for Evaluating Alignment and Clustering

Alignment score, devised by Butler et al. (2018), measures the uniformity of mixing among samples from different datasets ($N \geq 2$) in the aligned latent space. High score (close to 1) implies the datasets share underlying cell types and are well integrated, while low score (close to 0) indicates the datasets do not share cognate populations and the samples are not aligned. In the manuscript, we report the alignment score calculated from the cell factor loading matrices H (dimension = number of metagenes K). We also employ the k-nearest neighbor batch-effect test (kBET) Büttner et al. (2019) to assess the data integration results on H. kBET first creates a k-nearest neighbor graph (we used $k = 20$ for all analyses in the paper), and then randomly samples $1,000$ cells to examine the batch label distribution in the cell's neighbourhood against the global batch label distribution, using a χ^2 -test (100 repeats) under the null hypothesis that input data batches are mixed well. If the datasets are well integrated, the local batch label distribution will be similar to the global batch label distribution and the statistical tests will not reject the null hypothesis, resulting in a low rejection rate for $1,000$ tested data points in each repeat. In our analyses, we took the median of the rejection rates from all repeats and subtracted it from 1 to report the overall acceptance rate. High acceptance rate indicates well-mixed datasets. To quantify clustering performance, we used the purity metric and the adjusted Rand index (ARI) (Hubert and Arabie 1985). Purity assesses the resulting clusters with respect to a reference clustering. To calculate purity, one can assign each cluster to the dominant class in the cluster and count the number of correctly assigned samples in it. Then the purity is calculated by taking the sum over all clusters and dividing by the total number of samples. ARI is another popular method to compare clustering results. It counts pairs of samples where two clustering results agree or disagree. ARI was built upon the Rand index (RI) (Rand 1971), and fixes the issues in practice suffered by RI such as narrow range and non-constant baseline. ARI lies between 0 (no match) and 1 (perfect match).

2.5.5 Integrative Analyses on Real Data

2.5.5.1 Study of Convergence Behavior of `Online iNMF`

To investigate the convergence behavior of `Online iNMF` (scenario 1), we utilized several strategies and datasets. The first experiment was conducted on the adult mouse frontal ($n = 156,167$) and posterior cortex ($n = 99,186$) datasets, generated by Saunders et al. (2018). We split both into training (80%) and testing sets (20%). Three methods were used for comparison: `Online iNMF` (mini-batch size $= 5,000$ cells), ANLS (batch iNMF) and multiplicative updates

Algorithm 2.3: Quantile Normalization

Data: $H^i \in \mathbb{R}_+^{n_i \times K}, i = 1, \ldots, N$

for $i = 1 \to N$ **do**
  for $j = 1 \to K$ **do**
    Scale $H^i_{\cdot j}$ (centering is optional) `// Cell (from` X^i`) loadings on` j`th metagene factor`
  end
end
Set X^R as reference dataset ($R = \arg\max_i n_i$);
for $i = 1 \to N$ **do**
  for $s = 1$ *to* n_i **do**
    $c^i_s = \arg\max_j H^i_{sj}$;
  end
end
for $i = 1 \to N$ **do**
  `// Cluster re-assignment of` $\boldsymbol{x}^i_s$
  for $s = 1 \to n_i$ **do**
    Identify Q nearest neighbors of $\boldsymbol{x}^i_s$;
    Obtain $c^i_{s(q)}, q = 1, \cdots, Q$;
    $c^{i*}_s = \arg\max_j \sum_{q=1}^{Q} \mathbb{I}[c^i_{s(q)} = j]$;
  end
end
for $j = 1 \to K$ **do**
  for $i = 1 \to N$ *(*$i \neq R$*)* **do**
    for $k = 1 \to K$ **do**
      Obtain $H^R_{j,k}$ `//  loadings of the cells (from` X^R`) in cluster` j `on` k`th metagene factor`
      Obtain $H^i_{j,k}$ `//  loadings of the cells (from` X^i`) in cluster` j `on` k`th metagene factor`
      Match the quantiles of $H^R_{j,k}$ and $H^i_{j,k}$;
    end
  end
end
return normalized H^i, $i = 1, \ldots, N$.

(Mult). With $1,111$ genes jointly selected from the input datasets, we tracked the training and testing objectives calculated based on the resulting factors (Figure 2.2a,b). In order to evaluate the testing objective, we calculated cell factor loadings for cells in the testing set using the metagene factors obtained from the training set. As the `Online iNMF` algorithm aims to minimize the expected cost, we expect the `Online iNMF` to converge more rapidly than batch methods on the testing set, which can be viewed as a surrogate of the expected cost. Mairal et al. (2010) took a similar approach to evaluate their `Online iNMF` algorithm. In the second experiment, we monitored the iNMF objective on the training set after 500 seconds and repeated 20 times with random initializations, in order to further demonstrate the efficiency of the algorithms (Figure 2.2c). For the third part of this study, we focused on the effect of the mini-batch size. We applied `Online iNMF` on the same training and testing cortex datasets, but with mini-batches of increasing size ($n = 1,000$, $5,000$, $10,000$, $50,000$, $100,000$, $150,000$, $200,000$). Similarly, we tracked the training and testing objectives until the algorithm converged (Figure 2.2d,e). Lastly, we implemented `Online iNMF` on multiple subsets of different sizes sampled from the training set (Figure 2.2f). At multiple time points throughout the training process, we used the learned metagenes to solve for the cell factor loadings on the testing set, and calculated the testing objective. We set the key parameters $K = 40$ and $\lambda = 5$ for all analyses discussed above.

We also carried out three additional analyses on different datasets to support our conclusions, where we looked at the trajectories of training/testing objectives as well as the minimization of the training objective within a given amount of time (Figure 2.3). The datasets and key parameters are listed as follows. 1) adult mouse brain (DropViz), 9 datasets (each corresponds to a brain region), $n = 691,962$, $K = 40$, $\lambda = 5$, mini-batch size $= 5,000$; 2) human PBMC (`SeuratData` package), $n = 13,999$, $2,000$ variable genes (selected through Seurat pipeline), $K = 20$, $\lambda = 5$, mini-batch size $= 2,000$; 3) human pancreas (`SeuratData` package), $n = 14,892$, $2,000$ variable genes (selected through `Seurat` pipeline), $K = 40$, $\lambda = 5$, mini-batch size = $3,000$.

2.5.5.2 Benchmark of Runtime and Peak Memory Usage

The benchmark study was carried out on the adult mouse frontal ($n = 156,167$) and posterior cortex ($n = 99,186$) datasets from the DropViz data collection (Saunders et al. 2018) (Figure 2.5a). We created four pairs of subsets of increasing sizes by sampling from each of the full datasets. Within each pair, the subset from the frontal cortex and the one from the posterior cortex held the same ratio as their full datasets ($61.2 : 38.8$). This resulted in five pairs of inputs ($n = 10,000$, $50,000$, $100,000$, $200,000$, $255,353$) for `Online iNMF` (scenario 1), batch iNMF, Harmony and Seurat v3. To ensure fair comparison, we preprocessed the data as suggested by each method. The preprocessing steps suggested by each method differ slightly as follows: (1) `Online iNMF` and batch iNMF normalize the gene expression measurements for each cell and then scale the

gene expression data without centering to zero mean, because iNMF expects nonnegative inputs. (2) Seurat log-transforms the normalized gene expression matrices. (3) Harmony log-transforms the normalized gene expression and scales each gene to unit variance, and centers to zero mean (Note that we ran Harmony using the `SeuratWrappers` package.).For fair comparison, we used the same set of $1,111$ variable genes for all approaches, the same number of dimensions of the latent space ($K = 40$) and the same penalty parameter $\lambda = 5$ for iNMF-based approaches. We ran `Online iNMF` for 5 epochs, the default setting. We also ran batch iNMF, Seurat and Harmony. During the benchmark, we measured runtime (using the `tictoc` package) and peak memory usage (`peakRAM` package) for factorization and alignment (quantile normalization included for online/batch iNMF). We did not include data preprocessing (normalization and scaling) in runtime and memory benchmarks.

2.5.5.3 Analysis of Human PBMC and Pancreas

We analyzed the human PBMC ($n = 13,999$ cells) and human pancreas ($n = 14,890$) datasets in several experimental settings. The human PBMC dataset consists of two batches, control (n = 6,548) and stimulated cells ($n = 7,451$). The human pancreas dataset comprises eight batches ($n = 638$, $1,937$, $1,004$, $2,285$, $1,724$, $3,605$, $1,303$, $2,394$) across five different technologies (SMARTSeq2, Fluidigm C1, CelSeq, CelSeq2, inDrops). In the first experiment (Figure 2.3b-c), we used these datasets to study the convergence behavior of the algorithms (discussed above). In the second experiment (Figure 2.4), we performed `Online iNMF` (scenario 1) on the PBMC with $1,778$ variable genes ($K = 20$, $\lambda = 5$, mini-batch size $= 2,000$, epochs $= 5$), and on the pancreatic islets with $2,051$ variable genes ($K = 40$, $\lambda = 5$, mini-batch size $= 3,000$, epochs $= 5$), followed by quantile normalization. We ran batch iNMF with the same variable genes, K, and λ until convergence. For the third experiment (Figure 2.5b-c), we used the human PBMC and pancreas to benchmark `Online iNMF` (scenario 1), along with batch iNMF, Harmony and Seurat, with respect to alignment and clustering performance. We used the top $2,000$ highly variable genes selected by Seurat for all algorithms. For online and batch iNMF, the analytical pipelines and the key parameters stayed the same as in the previous experiment. To account for the effect of random initialization, the iNMF-based analyses were repeated 100 times. For Harmony and Seurat, we ran the analyses once, with the number of dimensions for the latent space set to 20 and 40 respectively (matching the iNMF K). We also ran additional analyses on human PBMC to inspect the data reconstruction ability of `Online iNMF`, as well as the effect of λ on resulting data integration using `Online iNMF` (see Supplementary Note for details).

2.5.5.4 Analysis of Adult Mouse Brain

The adult mouse brain dataset (DropViz) comprises nine individual scRNA-seq datasets, each generated from a specific brain region. The brain regions assayed include frontal cortex ($n = 156,167$), posterior cortex ($n = 99,186$), cerebellum ($n = 26,139$), entopeduncular ($n = 19,214$), globus pallidus ($n = 66,318$), hippocampus ($n = 113,507$), striatum ($n = 77,454$), substantia nigra ($n = 44,416$) and thalamus ($n = 89,561$), totaling $691,962$ cells. We picked $1,111$ variable genes and integrated the frontal and posterior cortex datasets using `Online iNMF` (scenario 1) and batch iNMF. Then we obtained the UMAP coordinates from the quantile normalized cell factor loadings and colored the cells by datasets and published cell type labels. Although all $255,353$ cells from the cortex were used for factorization, $117,985$ of them were annotated by Saunders et al. (2018). and shown in the plot (Figure 2.4). Moreover, we integrated the data across all nine brain regions (Figure 2.7). We identified $1,914$ genes that are highly variable in at least one of the regions. Using these genes, we performed 3 epochs of `Online iNMF` (scenario 1) with mini-batch size of $5,000$, $K = 40$ and $\lambda = 5$. In this analysis, we found that quantile normalization was not necessary for these dataset–iNMF alone was sufficient for integration.

2.5.5.5 Analysis of Spatial Transcriptomic Data

In the Slide-seq analysis, we filtered the scRNA-seq data for low quality cells–labeled in the original annotation file–for a total of $193,155$ cells. We combined Pucks 190921, 191204, and 200115 from the Slide-seq data for a total of $59,858$ beads. We selected $16,655$ variable genes. We ran scenario 1 with $K = 30$ for 5 epochs, $\lambda = 5$ on the scRNA-seq data, then projected the Slide-seq data following scenario 3. After factorization, we performed quantile normalization and Louvain clustering. We then colored the Slide-seq beads with the new labels generated based on the marker genes. Because each Slide-seq bead may contain more than one cell, we used the cell factor loadings to estimate the proportion of each cell type on each bead. To do this, we annotated each iNMF factor to assign it to a cell type, as described in the original Slide-seq paper.The loading value of each metagene factor then indicates the cell proportions of the corresponding cell types on each bead. We excluded the beads with no clear cell type, and for those with two cell types contributing more than 35% to the factor loadings, we colored the beads by the one with the higher loading. In the second analysis, we used the MERFISH dataset ($n = 1,026,840$ cells) and scRNA-seq ($n = 31,250$) in scenario 1 and scenario 3. We used the 134 genes measured in the MERFISH dataset. We used $K = 30$ and $\lambda = 5$. Scenario 1 was run for 5 epochs, and the slides plotted for Figure 2.7g are from animal 1.

2.5.5.6 Analysis of Mouse Primary Motor Cortex

The mouse primary motor cortex (MOp) datasets were generated by the BRAIN Initiative Cell Census Network (BICCN). The eight datasets span four modalities (single-cell RNA-seq, single-nucleus RNA-seq, single-nucleus ATAC-seq, single-nucleus methylcytosine-seq) and include $786,605$ cells. For most of the analyses on MOp, we only used the neurons, $408,885$ in total, except for the analyses involving oligodendrocytes. These datasets are (in the chronological order they were generated) allen_smarter_cells ($n = 6,244$ neurons), allen_10x_cells_v2 ($n = 121,440$ neurons), allen_smarter_nuclei ($n = 5,911$ neurons), allen_10x_cells_v3 ($n = 69,727$ neurons), allen_10x_nuclei_v3 ($n = 39,706$ neurons), macosko_10x_nuclei_v3 ($n = 101,647$ neurons), ecker_ren_atac ($n = 54,844$ neurons), ecker_ren_met ($n = 9,366$ neurons). The RNA and ATAC datasets were preprocessed following the standard `LIGER` pipeline. We selected variable genes using the genes shared across all datasets. We preprocessed methylation data as described in the original `LIGER` paper (Welch et al. 2019). Briefly, we inverted the direction of gene-body mCH methylation (which is anticorrelated with gene expression) by taking the difference between the maximum of the matrix and each matrix element. The resulting gene-level methylation features are positively correlated with gene expression. Methylation data does not require library size normalization because its values are already ratios (the number of methylated nucleotides divided by the number of detected nucleotides). For iterative multi-omic integration using `Online iNMF` (scenario 2), we performed a single epoch of training (each cell participates in exactly one mini-batch). When adding a new dataset i ($1 \leq i \leq N$), we incorporated a new dataset-specific metagene V^i and randomly initialized it. We did not use the data previously seen to refine the metagenes after the initial single epoch per dataset. Then we re-computed the cell factor loadings for all datasets ($H^1, \ldots, H^N$) using the latest metagenes and quantile normalized them.

For integration of the entire MOp dataset ($N = 8$) in scenario 2 (Figure 2.9), we identified $4,783$ variable genes from the first input (i.e. allen_smarter_cells) and used a fixed mini-batch size of $5,000$ cells, $K = 30$, $\lambda = 1$. For integrating all MOp datasets in scenario 1 (Figure 2.10a), we applied the same parameter setting except for $\lambda = 5$. Moreover, we attempted another strategy, where we integrated the first six sc/snRNA-seq datasets sequentially in scenario 2 and then projected both epigenomic datasets (snATAC-seq and snmC-seq) into the learned latent space, followed by quantile normalization and Louvain clustering (Figure 2.10c). In order to benchmark the cross-modality data integration performance across algorithms (Figure 2.6), we randomly sampled $5,000$ cells from the snRNA-seq dataset (macosko_10x_nuclei_v3) and $5,000$ cells from the snATAC-seq dataset (ecker_ren_atac). We implemented data integration using `Online iNMF` (scenario 1, $3,717$ variable genes, $K = 30$, $\lambda = 5$) as well as Seurat v3, Harmony and BBKNN with the same set of genes and dimension $= 30$ for dimension reduction process. Unlike the other methods, BBKNN only outputs a graph, on which alignment score and kBET cannot be calculated. Therefore, in

the main text we only reported these metrics for `Online iNMF`, Seurat v3 and Harmony, which produce the latent coordinates. In addition, we tried calculating the alignment metrics on the UMAP coordinates. In this setting, `Online iNMF` is still the best (alignment score $= 0.816$, kBET $= 0.651$), followed by Seurat v3 (alignment score $= 0.747$, kBET $= 0.544$), BBKNN (alignment score $= 0.409$, kBET $= 0.218$) and Harmony (alignment score $= 0.139$, kBET $= 0.092$).

For other supplementary analyses, we retained or held out the cell types of interest and carried out `Online iNMF` in scenario 1, 2, and 3 as introduced in the supplementary notes (Figure 2.13, 2.14, 2.15). More specifically, for the analyses in scenario 2 reported in Figure 2.13a, we used $2,011$ and $1,997$ variable genes respectively. Similarly, for the analyses reported in Figure 2.13b, we selected $2,019$ variable genes for both. The other key parameters are $K = 30$, $\lambda = 1$, and $k = 200$ for quantile normalization. For the results displayed in Figure 2.14a, we used the same $2,111$ variable genes and set $K = 30$ for all approaches, while using $\lambda = 1$ for `Online iNMF` (scenario 2) and $\lambda = 5$ for `Online iNMF` (scenario 1) as well as batch iNMF. Upon completion of the factorization, we performed quantile normalization with $k = 2,000$.

In order to generate Figure 2.14b, we integrated two sc/snRNA-seq datasets with $2,045$ genes, $K = 30$, $\lambda = 5$ in scenario 1, and then projected the snATAC-seq dataset into the learned latent space. We quantile normalized the cell factor loadings with $k = 1,000$.

For the analysis shown in Figure 2.15, we used $2,210$ variable genes, $K = 30$, $\lambda = 5$ for the part done in scenario 1. After the last dataset was incorporated in scenario 3, we ran quantile normalization with $k = 200$.

As shown in Figure 2.16, we factorized snATAC-seq and snmC-seq data both alone and jointly with a snRNA-seq dataset using `Online iNMF` in scenario 1 ($2,008$ variable genes, $K = 30$, $\lambda = 5$). Then we run quantile normalization and louvain clustering following standard procedure.

2.5.5.7 Analysis of Mouse Organogenesis Cell Atlas

The mouse organogenesis cell atlas (MOCA) consists of $1,363,063$ cells from embryos between 9.5 to 13.5 days of gestation (e9.5, e10.5, e11.5, e12.5, e13.5). We first selected $2,557$ variable genes and then integrated the five MOCA datasets in scenario 1 with the following setting: mini-batch size $= 5,000$ cells, $K = 50$, $\lambda = 5$, epochs $= 1$. As the alignment was quite good without quantile normalization, the 3D UMAP coordinates were obtained from the unnormalized cell factor loadings. Lastly, we visualized the cells, colored by datasets (gestational age) and published developmental trajectory labels using the `rgl` package (Figure 2.17). We employed Harmony for this analysis with the same set of variable genes and dimensionality of the latent space (PCA).

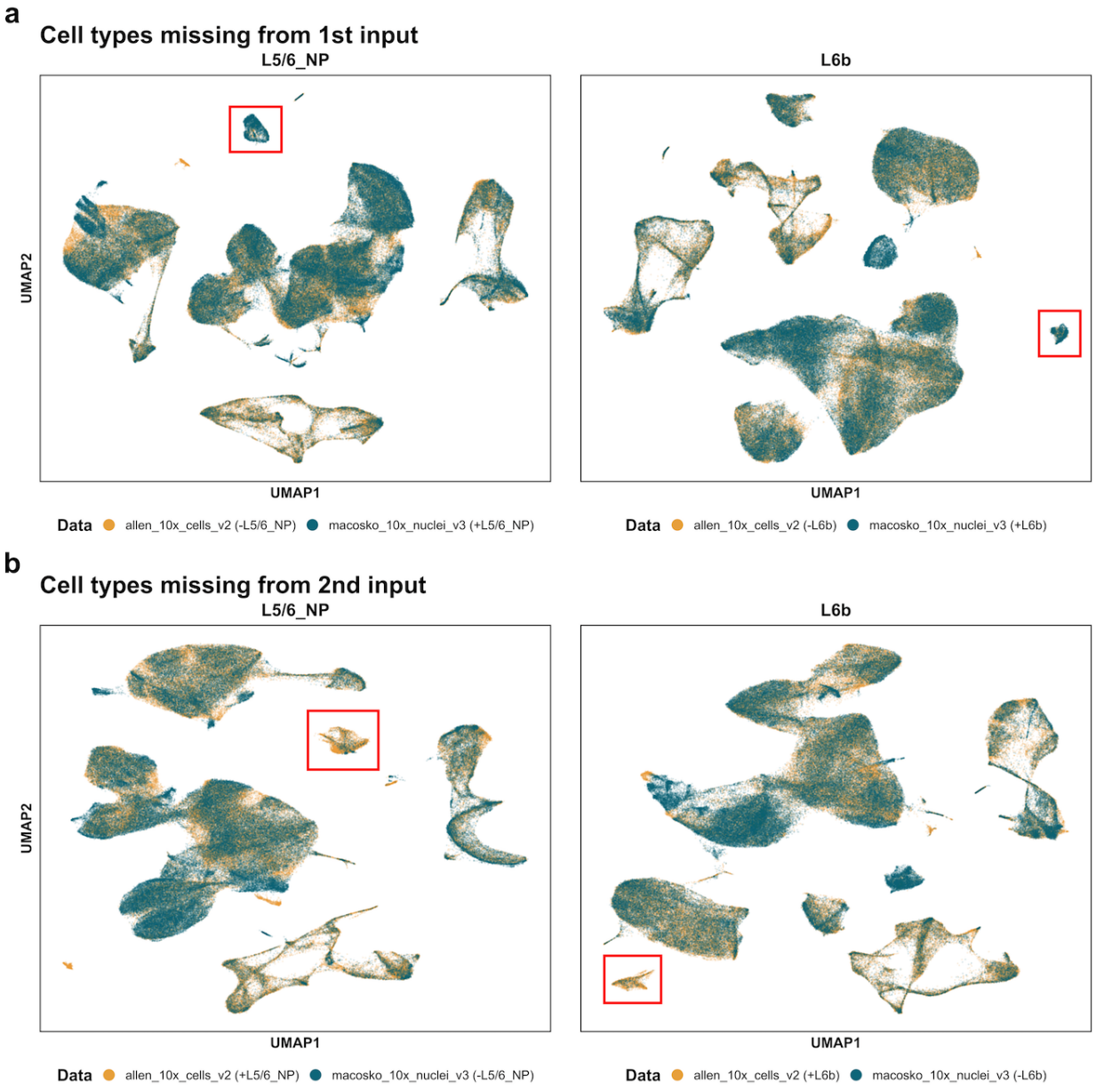


Figure 2.13: Performance of `Online iNMF` (scenario 2) with missing rare cell clusters (real data). The L5/6_NP and L6b cells missing from early- or late- arriving datasets are successfully identified. **(a)** The rare cell types were missing from the first input (allen_10x_cells_v2). **(b)** The rare cell types were missing from the second input (macosko_10x_nuclei_v3).

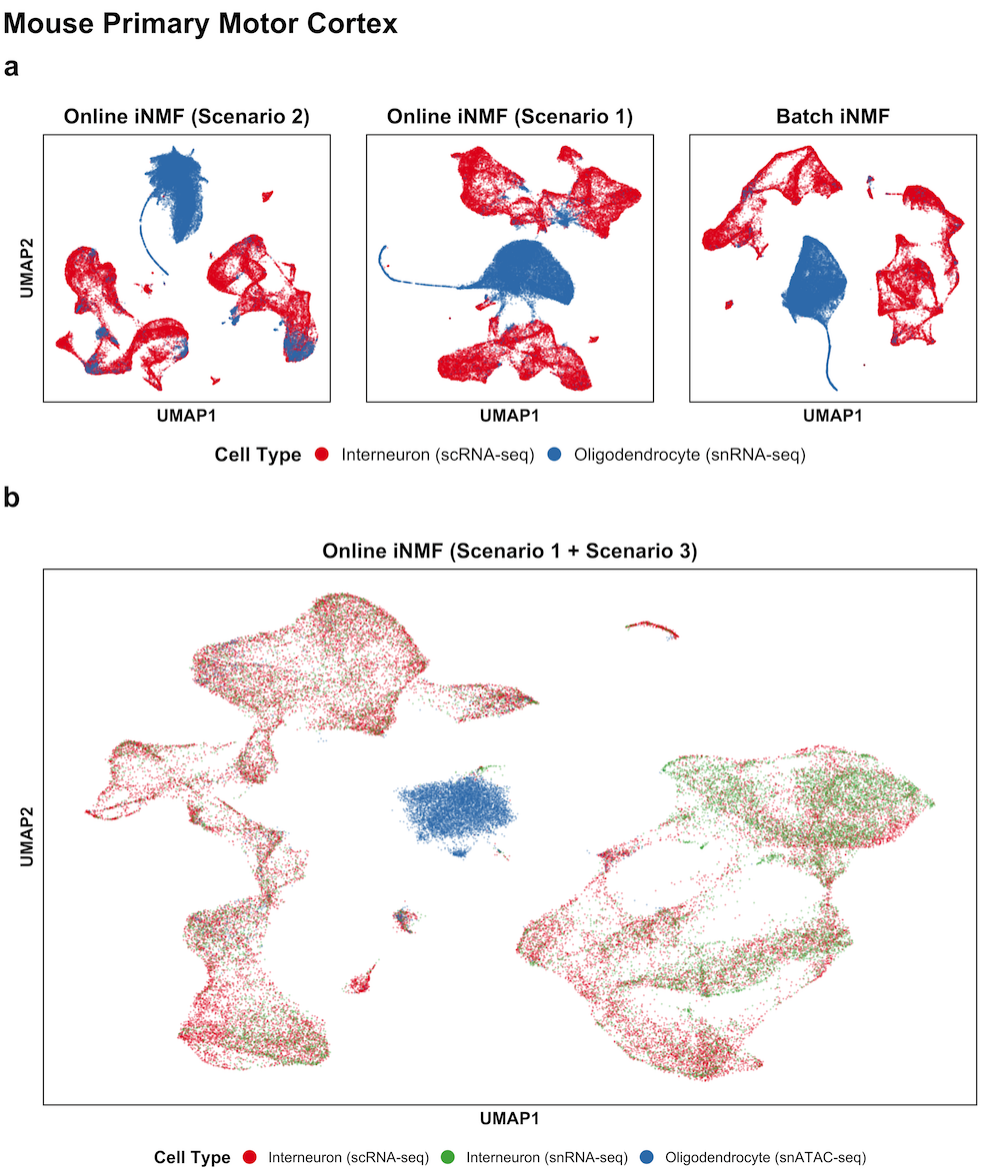


Figure 2.14: `Online iNMF` results in minimal spurious alignment for non-overlapping datasets (real data). (a) `Online iNMF` (scenario 1 & 2) and batch iNMF are utilized to integrate one dataset containing only interneurons (scRNA-seq, $n = 27,555$) and another containing only oligodendrocytes (snRNA-seq, $n = 21,404$) using 30 metagenes. **(b)** Projection of completely non-overlapping dataset into the existing latent space leads to minimal spurious alignment. An scRNA-seq dataset ($n = 27,555$) and a snRNA-seq dataset ($n = 15,255$) containing only interneurons are first integrated in scenario 1. Then an snATAC-seq dataset containing only oligodendrocytes ($n = 8,557$) is projected into this aligned latent space.

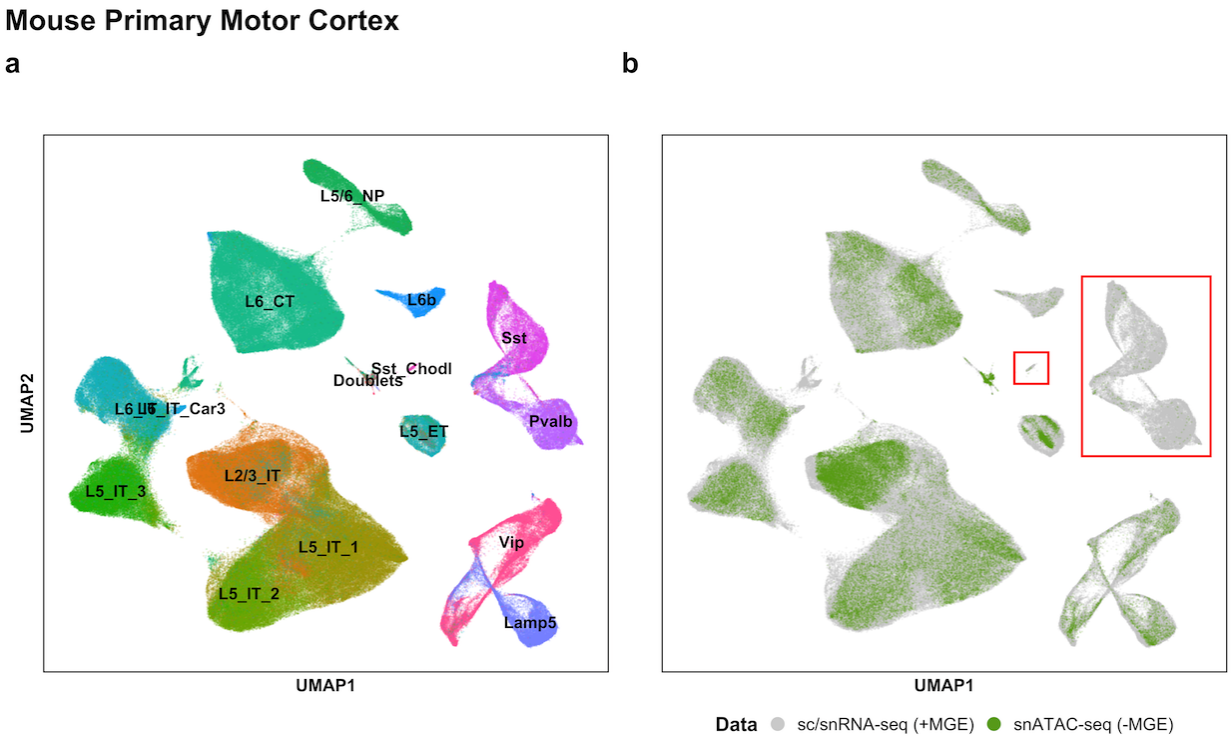


Figure 2.15: `Online iNMF` (scenario 3) leads to little spurious alignment when integrating partially-overlapping datasets. 6 sc/snRNA-seq datasets from the MOp ($n = 344,675$) were integrated using `Online iNMF` (scenario 1). Then an snATAC-seq dataset ($n = 49,167$) without MGEs (i.e. Pvalb, Sst and Chodl cells) was projected (scenario 3) into the atlas already built. **(a)** The UMAP visualization annotated by our cell class labels. **(b)** UMAP plot colored by dataset. Almost no cells from snATAC-seq data are observed in the clusters corresponding to Pvalb, Sst and Chodl cells (red boxes).

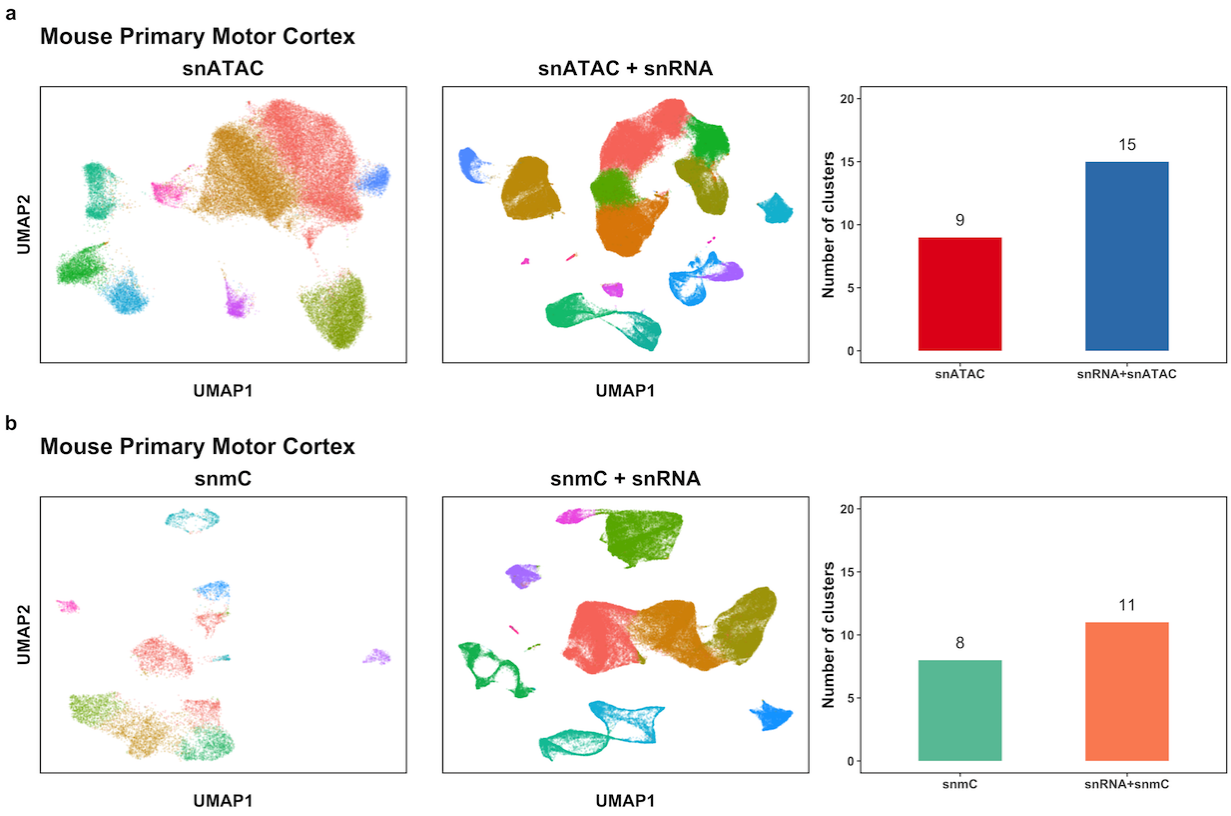


Figure 2.16: Integrating methylation or chromatin accessibility data with RNA data better separates clusters. **(a)** 6 more clusters are observed after joint analysis of snATAC-seq data ($n = 54,844$) and snRNA-seq data ($n = 101,647$) than analysis of snATAC-seq data alone. **(b)** 3 more clusters are obtained after incorporating the snRNA-seq data than investigating snmC-seq data ($n = 9,366$) alone.

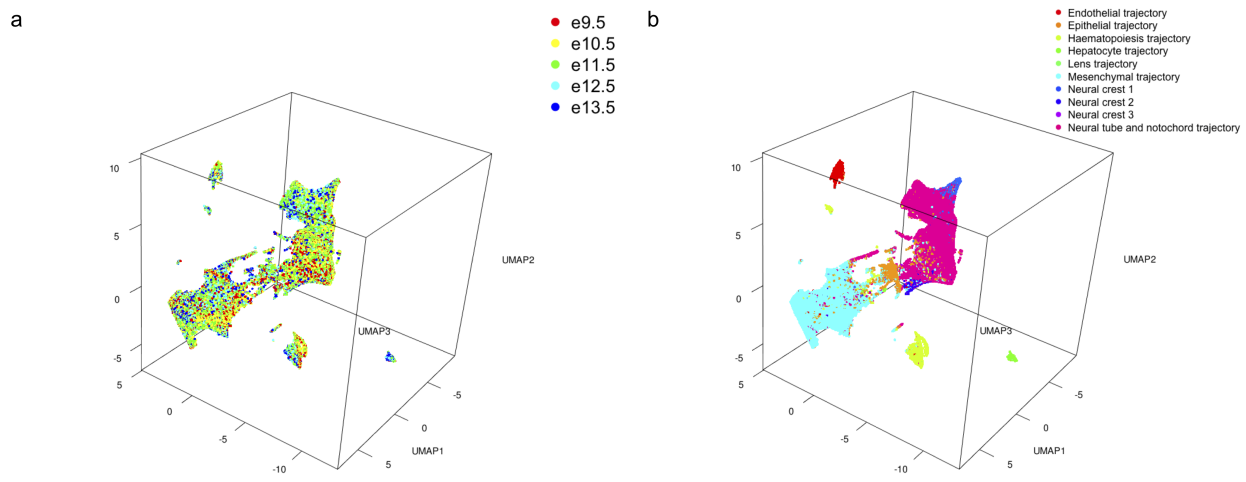


Figure 2.17: `Online iNMF` (scenario 1) efficiently factorizes the mouse organogenesis cell atlas (MOCA). The MOCA dataset consists of $1,363,063$ cells from embryos between 9.5 to 13.5 days of gestation. The `Online iNMF` analysis required 25 minutes and less than 2 GB of RAM on a MacBook Pro, compared to 98 minutes and 109 GB of RAM for Harmony, which could only be run on a large-memory server. **(a-b)** 3D UMAP plot of the `Online iNMF` results ($n = 200,000$ cells sampled for visualization), colored by dataset **(a)** and published developmental trajectory labels **(b)**.

2.5.6 Integrative Analyses on Simulated Data

2.5.6.1 Generating Simulated scRNA-seq Data

We employed the R package `Splatter` (Zappia et al. 2017) to simulate scRNA-seq datasets. Each dataset has $50,000$ cells and $10,000$ genes, separated into 6 batches and 8 cross-batch cell types (clusters). We adopted the settings from the recently reported benchmark study (Tran et al. 2020) while adjusting the proportion of each batch and cluster according to our needs. We determined the dataset compositions following one of these three strategies: 1) randomly sample the cluster proportions from the Dirichlet distribution for each simulation while keeping the batch sizes (also generated by Dirichlet distribution) in each simulation the same (Figure 2.18, 2.21); 2) randomly sample the batch sizes from the Dirichlet distribution for each simulation while keeping the cluster proportions (also generated by Dirichlet distribution) in each simulation the same (Figure 2.18); 3) use the same cell type and batch proportions to isolate the effect of differences in cell cluster membership across partially overlapping datasets (Figure 2.21, 2.19, 2.20).

2.5.6.2 Analysis of Simulated Data with Unbalanced Cell Clusters and Dataset Sizes

We generated the datasets for this analysis following the first and second data generation strategies described in the "Generating simulated scRNA-seq data" section, corresponding to the analysis of unbalanced cell clusters and datasets sizes respectively. To quantitatively measure the level of imbalance in each analysis, we computed the Shannon entropy (H) of both the cluster proportions ($H_{cluster}$) and batch sizes (H_{batch}) using the equation below, where P is a vector of n probabilities that add up to 1:

$$H(P) = -\sum_{i=1}^{n} p_i log_2(p_i)$$

Where $P = p_1, ..., p_n$, $0 < p_i < 1$, $\sum_{i=1}^{n} p_i = 1$. We then measured the performance of `Online iNMF` in scenario 1 and 2 (Table 2.1, 1st row). We also computed the spearman correlation between the evaluation metrics (Alignment, Purity, ARI, and kBET) and the entropy of cluster proportions and batch sizes (Figure 2.18e).

2.5.6.3 Analysis of Simulated Data with Missing Cell Clusters

We generated the datasets used in this analysis following the third data generation strategy described above. In this case, the cluster proportions and batch sizes were exactly the same for all 10 simulations, to isolate the effect of variable batch compositions. We then excluded 1-5 cell types from the first 5 batches to mimic the situations when the newly arriving data (Batch 6) share a varying number of common cell types with the reference data (Figure 2.19a). We applied `Online iNMF` in scenario 1 and 2 (Table 2.1, 2nd row) and visualized the evaluation metrics against the number of held-out cell types in line plots (Figure 2.19e). To test the performance of `Online iNMF` (scenario 3), we ran the pipeline again while treating the first 5 batches with missing cell types as the "reference data" and the last batch as the "projected data" (Table 2.1, third row). We plotted the results from the two evaluation metrics for `Online iNMF` on all cells, cells in the missing cell types, and cells in the shared cell types, along with the number of held-out cell types (Figure 2.20).

2.5.6.4 Analysis of Simulated Data with No Cell Types Shared Across All Datasets

We generated the datasets used in this analysis following the first data generation strategy described in previous section. Within each simulation, we excluded one different cluster in 5 batches and excluded the other three remaining clusters in the sixth batch to ensure that the intersection of cell types across all batches is the empty set (Figure 2.21a). To measure the performance of `Online iNMF` (scenario 1 and 2), we ran a number of regular LIGER analyses

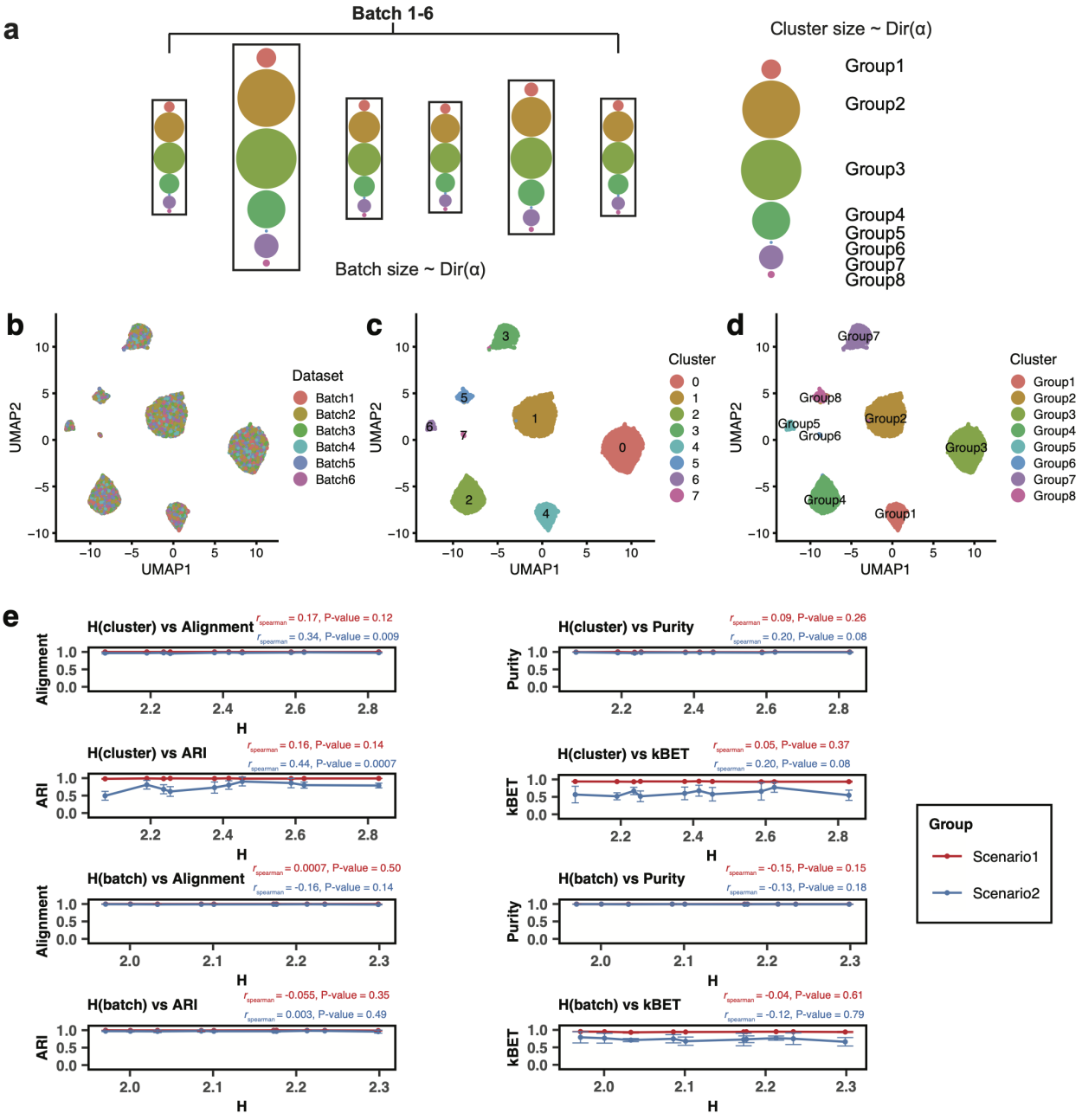


Figure 2.18: Performance of `Online iNMF` under unbalanced cell clusters and dataset sizes (simulations). **(a)** Schematic plot showing the composition of 8 clusters and 6 batches in each simulated dataset (with $10,000$ genes and $50,000$ cells). **(b-d)** UMAP representations of an example integration result plotted using batch labels **(b)**, LIGER cluster assignments **(c)**, and ground truth cluster labels **(d)**. **(e)** Line plots of four evaluation metric scores for `Online iNMF` (scenario 1 & 2) versus the Shannon entropy of cell type and batch size (larger H means more balanced composition). The data are presented as mean values $\pm$ standard deviation (5 random initializations for each simulated dataset, $n = 50,000$ cells in each simulated dataset). The p-value was obtained from one sided Spearman's rank correlation test without adjustment for multiple comparisons.

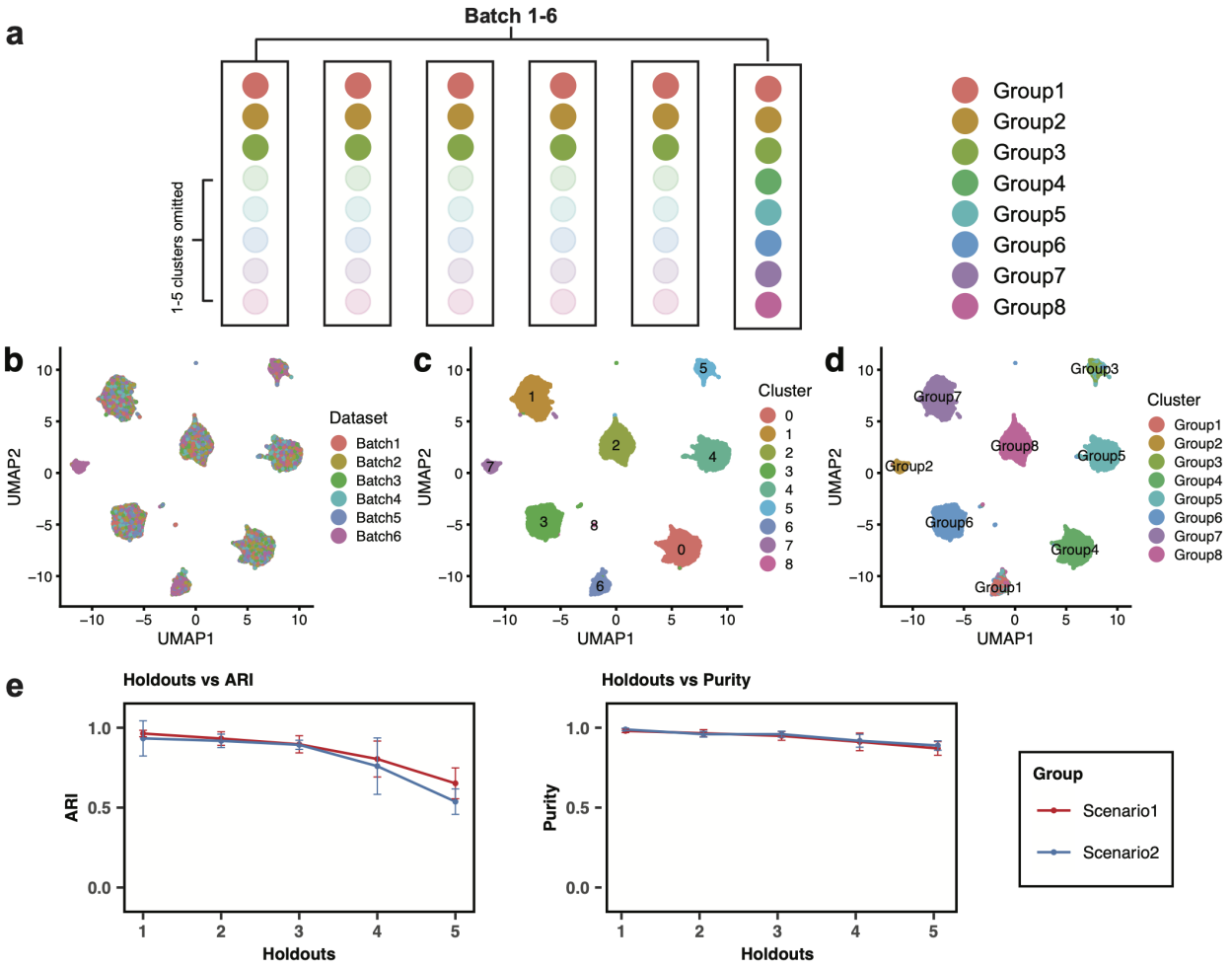


Figure 2.19: Performance of `Online iNMF` (scenario 1 and 2) with missing cell clusters (simulations). **(a)** Schematic plot showing the equal proportions of 8 clusters and 6 batches in each simulated dataset (with $10,000$ genes and $50,000$ cells) with 1-5 cell types excluded. **(b-d)** UMAP representations of an example integration result from scenario 1 from a simulation with three held-out cell types. The plots are colored using batch labels **(b)**, LIGER cluster assignments **(c)**, and ground truth cluster labels **(d)**. **(e)** Line plots of two evaluation metrics for `Online iNMF` (scenario 1 & 2) versus the number of cell types excluded. The data are presented as mean values ± standard deviation (10 random initializations for each simulated dataset, $n = 50,000$ cells in each simulated dataset before holding out any cell clusters).

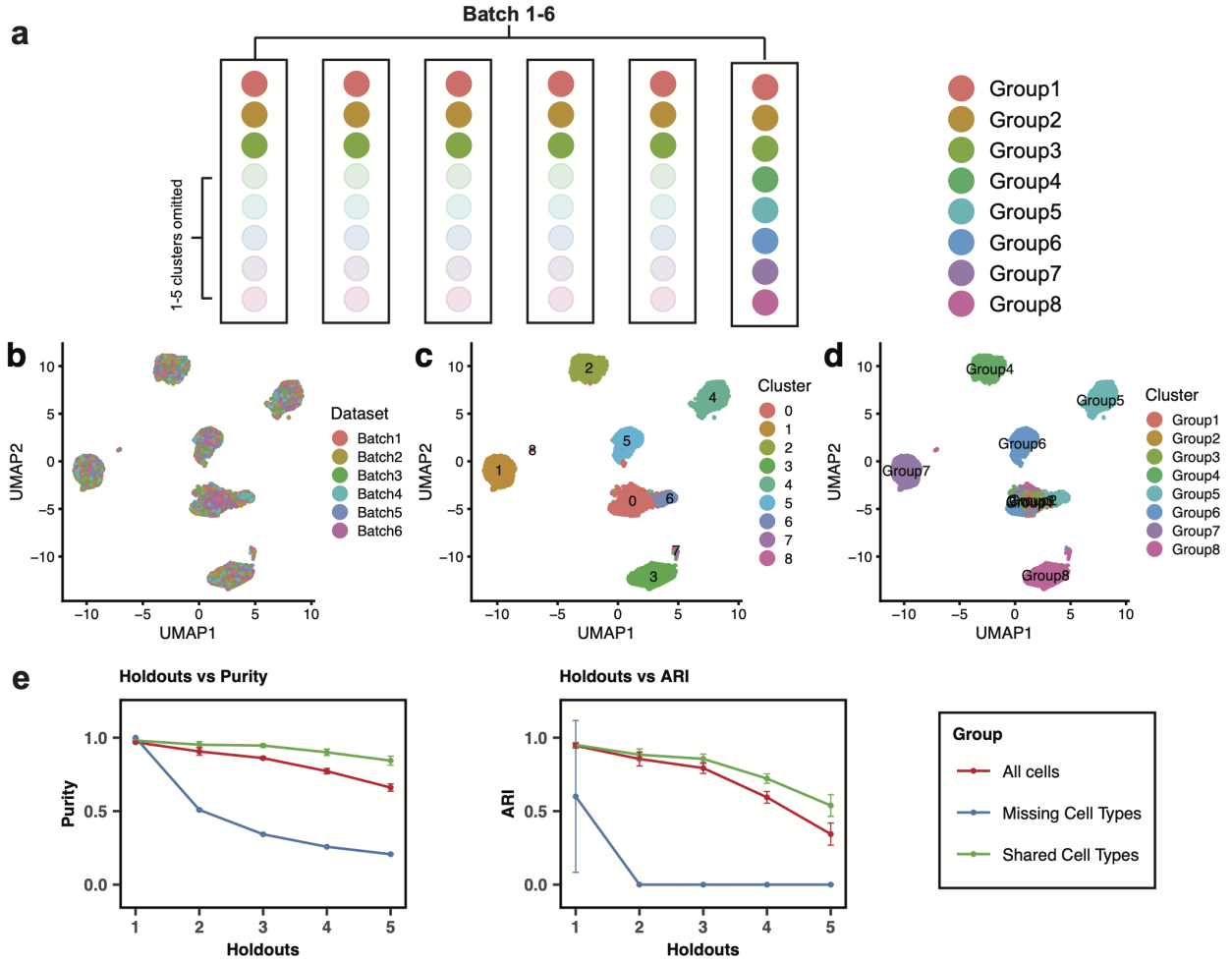


Figure 2.20: Performance of `Online iNMF` (scenario 3) with missing cell clusters (simulations). **(a)** Schematic plot showing the equal proportions of 8 clusters and 6 batches in each simulated dataset (with $10,000$ genes and $50,000$ cells) with 1-5 cell types excluded. **(b-d)** UMAP representations of an example integration result from scenario 3 from a simulation with 3 held-out cell types. The plots are colored using batch labels **(b)**, LIGER cluster assignments **(c)**, and ground truth cluster labels **(d)**. **(e)** Line plots of two evaluation metric scores for `Online iNMF` (scenario 3) on all cells (red line), cells in missing clusters (blue line), and cells in shared clusters (green line), versus the number of cell types excluded. The data are presented as mean values $\pm$ standard deviation (10 random initializations for each simulated dataset, $n = 50,000$ cells in each simulated dataset before holding out any cell clusters).

using mostly default parameters (Table 2.1, 4th row) and drew the boxplots using each evaluation metric calculated from 50 runs (Figure 2.21e).

2.5.6.5 Analysis of Simulated Data with Varying Number of Factors (K

The datasets used in this analysis were generated following the third data generation strategy described in the "Generating simulated scRNA-seq data" section, without any further subsetting or filtering. To measure the performance of `Online iNMF` (scenario 1 and 2) across a range of K values, we ran a number of analyses (Table 2.1, last row), and drew the line plots to show the relationship between each of the four evaluation metrics and values of K ranging from 10 to 40 (Figure 2.22).

Figure	Integration Strategy	K	λ	Mini-batch size	Variable Genes	# of iNMF Initializations	# of Simulations	# of total Runs
2.18	Scenario 1	20	5	5,000	$\sim$ 3,000 (*)	5	20	100
	Scenario 2	20	10	1,000	3,000 (**)	5	20	100
2.21	Scenario 1	20	5	5,000	$\sim$ 3,000 (*)	5	10	50
	Scenario 2	20	10	1,000	3,000 (**)	5	10	50
2.19	Scenario 1	20	5	5,000	$\sim$ 3,000 (*)	10	5	50
	Scenario 2	20	10	1,000	3,000 (**)	10	5	50
2.20	Scenario 3	20	5	5,000	$\sim$ 3,000 (***)	10	5	50
2.22	Scenario 1	10-40	5	5,000	$\sim$ 3,000 (*)	1	70	70
	Scenario 2	10-40	10	1,000	3,000 (**)	1	70	70

* Selected from all batches

** Selected from the first batch

*** Selected from batches with missing cell types

Table 2.1: Key parameter settings for integrated analysis on simulated data

2.6 Supplementary Note: Benchmarking Online iNMF Performance Across a Range of Conditions Using Real and Simulated Data

2.6.1 Benchmarking `Online iNMF` with Simulation Studies

We have demonstrated in the former sections the robust performance of `Online iNMF` on multiple real datasets including human PBMC, human pancreas, and mouse cortex. To provide additional theoretical understanding, we performed extensive simulations using the Splatter scRNA-seq simulator. We investigated the effects of different dataset orderings, relative dataset sizes, and cell type compositions. To give a clearer view, these results are separated into five figures and organized in a similar fashion, including a schematic plot of the simulation design (a, Figure 2.18,

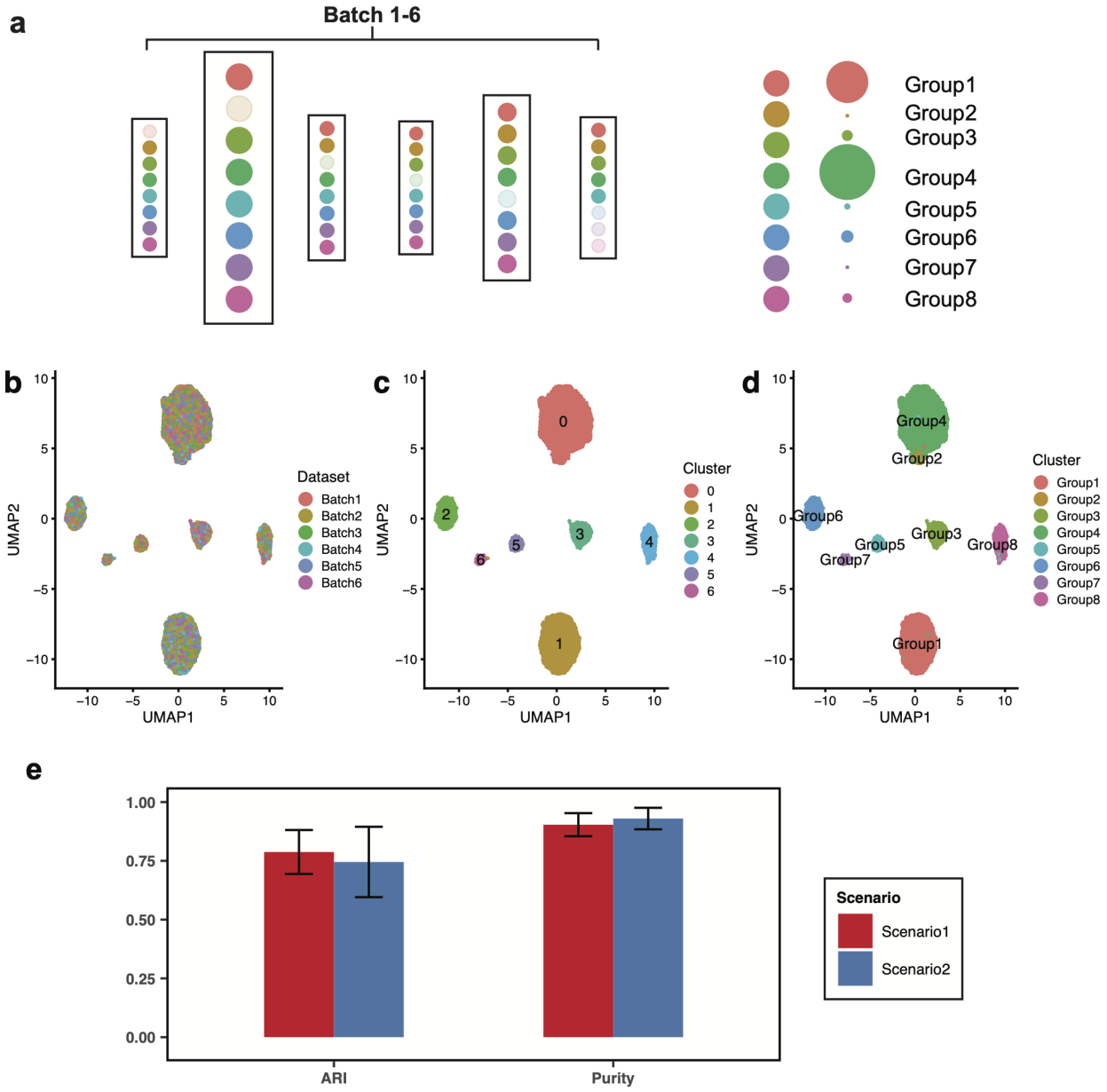


Figure 2.21: Performance of `Online iNMF` with no cell types shared across all datasets (simulations). (a) Schematic plot showing the composition of 8 clusters and 6 batches in each of ten simulated datasets; data were further filtered to make sure the intersection of all batches in each simulation is the empty set while the pairwise intersections of all batches are non-empty. **(b-d)** UMAP representations of an integration example result under scenarios 1 & 2 plotted using batch labels **(b)**, LIGER cluster assignments **(c)**, and ground truth cluster labels **(d)**. **(e)** Bar plot of the two evaluation metric scores for `Online iNMF` (scenario 1 & 2) simulated data. The data are presented as mean values $\pm$ standard deviation (50 runs in total for each metric, $n = 50,000$ cells in each simulated dataset).

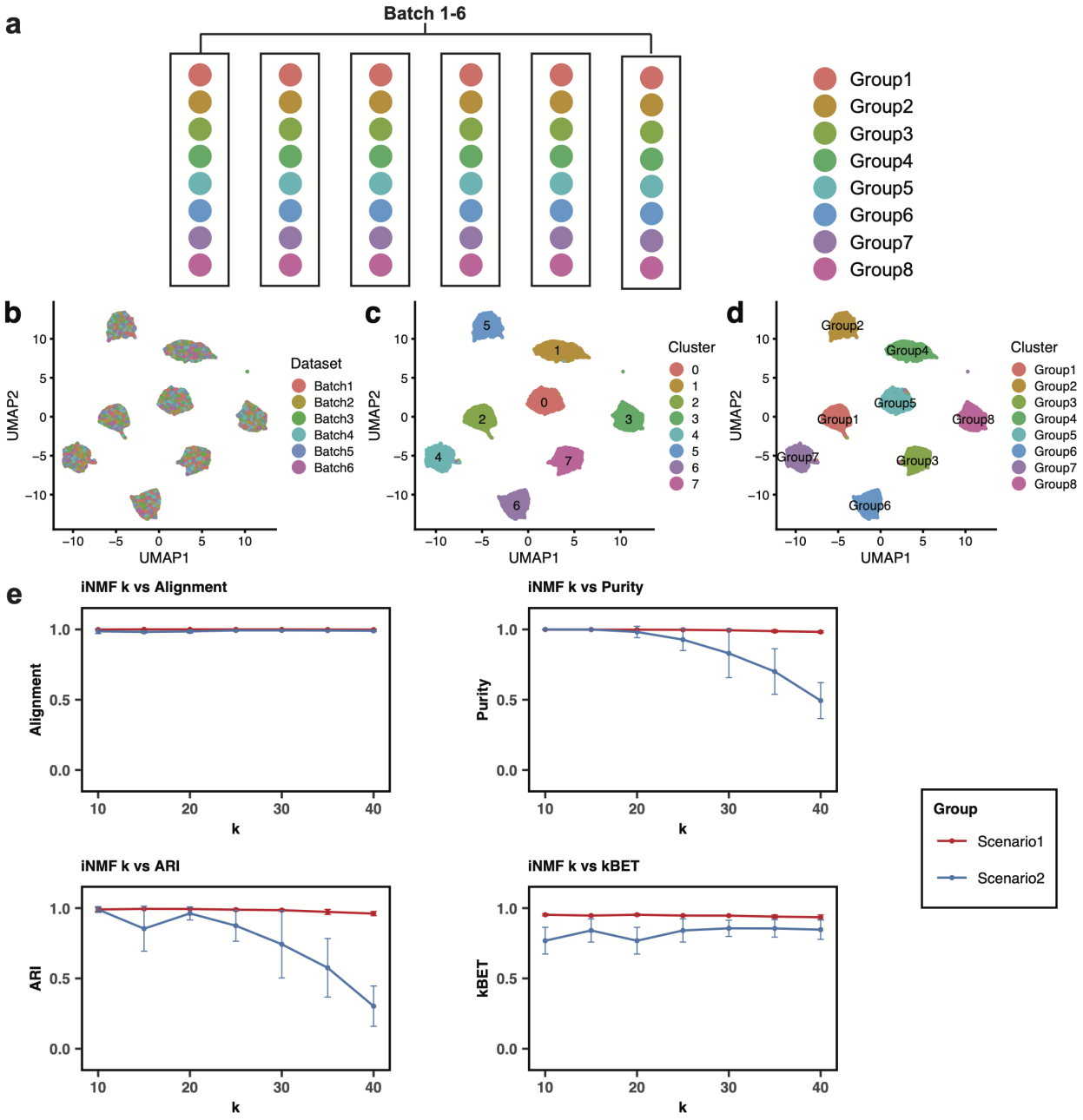


Figure 2.22: Performance of `Online iNMF` with varying K (number of metagenes) (simulations). (a) Schematic plot showing the composition of 8 clusters and 6 batches in each simulated dataset (with $10,000$ genes and $50,000$ cells). **(b-d)** UMAP representations of an example result plotted using batch labels **(b)**, LIGER cluster assignments **(c)**, and ground truth cluster labels **(d)**. **(e)** Line plots of four evaluation metric scores for `Online iNMF` (scenario 1 & 2) versus K varying from 10 to 40 incremented by 5. The data are presented as mean values $\pm$ standard deviation (10 random initializations for each K, $n = 50,000$ cells in each simulated dataset).

2.19, 2.20, 2.21, 2.22), UMAP plots of a representative integration result (b-d, Figure 2.18, 2.19, 2.20, 2.21, 2.22), and box or line plots of evaluation metrics (e, Figure 2.18, 2.19, 2.20, 2.21, 2.22).

Overall, we found that imbalanced cell cluster proportions and dataset sizes have very little effect on the results of `Online iNMF`, although scenario 2 is slightly more sensitive to imbalances in cell proportions than scenario 1. We computed the Spearman correlations between the Shannon entropy of batch and cluster sizes and the four evaluation metrics we employed (see Online Methods section for details). Most of the correlation p-values are much larger than 0.05, indicating that `Online iNMF` performance is not significantly affected by imbalances in dataset size or cluster proportions. The one exception is that for `Online iNMF` (scenario 2), there is a statistically significant correlation between cluster entropy and both adjusted rand index and alignment, indicating that scenario 2 is slightly more sensitive to imbalances in cell proportions than scenario 1. Furthermore, we used random dataset arrival orders in benchmarking scenario 2, and found that the relative order of small vs. large batches in scenario 2 makes little difference (Figure 2.18).

With missing cell clusters, the proportion of the missing cell types also has little effect on the results of iNMF (scenarios 1 and 2) (Figure 2.19). Under the same condition, we also performed simulations to test whether scenario 3 will force cells into the existing feature space. These results showed that `Online iNMF` does not cause spurious alignment, even if one or more cell types in the dataset to be projected are missing from the reference dataset. In the case of multiple cell types missing from the reference dataset, all of the new cell types cluster with each other (but not with the reference cells). This causes a decrease in overall Purity and ARI (Figure 2.20e, red line), but a much more gradual decrease in purity and ARI for the cell types shared between reference and query datasets (Figure 2.20e, green line). This behavior makes sense, because the shared metagenes (W) learned from the reference dataset cannot be expected to distinguish among multiple unseen cell types, which explains the poor evaluation metrics for the cells in the missing cell types (Figure 2.20e, blue line).

We also designed simulations in which no cell types occurred in every batch, but every pair of batches shared at least one cell type. This allowed us to test the performance of iNMF on data with complex biological compositions. Our results indicate that `Online iNMF` is quite robust to these situations and identifies every cluster clearly (Figure 2.21). It seems that the most important factor determining the difficulty of identifying a particular cluster is the total number of cells in the cluster observed across all datasets, independent of how those cells are distributed across datasets.

2.6.2 Reading Mini-Batches from Disk Adds Minimal Overhead

One highlight of the proposed `Online iNMF` algorithm is that it streams the mini-batches from the files on the disk without loading the entire data into the memory. Here we demonstrate that little overhead is added through this approach on the mouse frontal and posterior cortex scRNA-seq

datasets (details are discussed in 2.5). For a mini-batch size of $5,000$ cells, reading each mini-batch from disk does not require significant overhead (an average of less than 0.56 seconds per iteration over 50 iterations) (Figure 2.12).

2.6.3 Online iNMF Is Robust to Initialization and Input Ordering

The `Online iNMF` algorithm starts with randomly initialized metagene factors (W and V^i). Therefore, we inspected the effect of random initialization on the analyses on the MOp datasets by assessing the agreement, as measured by ARI, between the resulting cell clusters and our annotations generated in scenario 2 (Figure 2.9d). First, we performed `Online iNMF` (scenario 1) on all eight MOP datasets with 10 different random initializations, using the same variable genes that we used for the scenario 2 analysis in Figure 2.9. Based on the output cell clusters, the average ARI (vs. our annotations from scenario 2, shown in Figure 2.9d) is 0.725. Similarly, we ran `Online iNMF` (scenario 2) with 10 different initializations using the same set of genes (inputs ordered chronologically) and obtained an average ARI (vs annotations) of 0.744. These results indicate that `Online iNMF` scenario 1 and 2 are both robust to the effects of random initialization.

To investigate the effects of different dataset orders on scenario 2 results, we repeated the scenario 2 analysis using each of the other five sc/snRNA-seq datasets as the initial dataset. As with the analysis shown in Figure 2.9, we selected over $4,000$ variable genes from the first dataset and sequentially incorporated all remaining MOp datasets. We found that initiating the analysis with any of the six sc/snRNA-seq datasets leads to clusters in good agreement with our annotation (average ARI $= 0.759$), indicating that the results are robust to choice of starting dataset. We can even select genes from the snATAC-seq dataset, and use it as the first input, with slightly lower agreement (ARI $= 0.627$). If we instead use the snATAC-seq dataset as the starting dataset but use the genes selected from the first RNA dataset (SMARTer cells), the ARI is 0.752. Because the distribution of methylation is so different from gene expression, the statistical model for variable gene selection reported zero variable genes, and thus we were not able to select genes from the methylation data. Additionally, the results from scenario 1 and 2 are quite congruent (ARI $= 0.773$).

2.6.4 Integration with RNA Data Detects More Clusters from Epigenome

In the original LIGER paper, we showed that integrating single-cell methylation data with scRNA-seq data resolved more methylation clusters than using methylation data alone. Here we confirmed that this still holds true for `Online iNMF` on the mouse primary motor cortex (MOp) datasets: integrating methylation (snmC-seq) or chromatin accessibility data (snATAC-seq) with RNA data (snRNA-seq) better separates clusters compared to the epigenome data alone (Figure 2.16). In the first experiment (Figure 2.16a), we started by factorizing the snATAC-seq data ($n = 54,844$)

and obtained 9 clusters. After incorporating the snRNA-seq data ($n = 101,647$), the two datasets are well aligned. More importantly, we are able to observe 15 clusters, which implies the structure within the data is refined. Similarly, in the second experiment (Figure 2.16b), the "resolution" of snmC-seq data ($n = 9,366$) is also increased after being jointly analyzed with the same snRNA-seq data, where 3 additional clusters are detected.

2.6.5 `Online iNMF` Identifies Rare Cell Types Present in Only a Subset of the Datasets

We also looked into the detection of rare cell types in MOp data, L5/6_NP and L6b, in separate analyses (Figure 2.13). In the first experiment, we held out L5/6_NP and L6b cells from the first input (allen_10x_cells_v2, $n = 117,382$) in scenario 2. Next, we incorporated an snRNA-seq dataset (macosko_10x_nuclei_v3, $n = 101,647$) that includes L5/6_NP cells (3.3% of all cells). After Louvain clustering on the learned latent space, 96.4% of the L5/6_NP cells in the snRNA-seq dataset grouped together and formed a distinct cluster (highlighted with a red box). In the second experiment, we held out the L6b cells from the scRNA-seq dataset ($n = 119,183$) and subsequently incorporated the snRNA-seq dataset ($n = 101,647$), in which L6b cells make up 1.5% of all cells. L6b is rarer than L5/6_NP, which makes this task more challenging. Additionally, the L6b cluster is more continuous with the L6 CT cells, and the cluster boundary is somewhat unstable across different clustering runs. Nevertheless, 91.8% of the L6b cells formed a distinct cluster. Thus, these results indicate that `Online iNMF` in scenario 2 can still detect rare cell types in late arriving datasets. We observed very similar results if the rare cell type was missing from the first dataset (99.3% of L5/6_NP cells formed a distinct cluster, and 94.3% of L6b cells formed a distinct cluster). Consistent with our simulation results, these analyses suggest that the order of dataset arrival is not strongly influential in whether rare cell types are detected.

2.6.6 `Online iNMF` Robustly Integrates Datasets with Non-Overlapping or Partially-Overlapping Cell Types

First, we examined the performance of `Online iNMF` in integrating datasets of the same modality that do not share any common cell types. For this evaluation in scenario 2, we selected two datasets generated from MOp and only retained cells of dissimilar classes. The first input in scenario 2 (scRNA-seq, 10x v2) consists only of interneurons ($n = 27,555$), including medial ganglionic eminence (MGE)-derived cells and caudal ganglionic eminence (CGE)-derived cells. In contrast, the second input (snRNA-seq, 10x v3) only contains oligodendrocytes ($n = 21,404$). We also performed this analysis using `Online iNMF` (scenario 1) and batch iNMF for comparison. The results are visualized in 2-dimensional UMAP coordinates (Figure 2.14a). As expected, there is very little spurious alignment between the two cell classes when implementing online learning in

scenario 2. The corresponding alignment scores for `Online iNMF` (scenario 2), `Online iNMF` (scenario 1) and batch iNMF are 0.106, 0.034 and 0.027 respectively, while the kBET acceptance rates are 0.050, 0.014 and 0.002. Thus, all three approaches are quite comparable in their ability to avoid spurious alignment of the non-overlapping cell types. Moreover, 30 metagenes effectively capture the structure within the interneurons. We were also interested in how `Online iNMF` would perform in scenario 3 in a similar setting. In this experiment, we started by creating a curated atlas of interneurons using scRNA-seq dataset ($n = 27,555$) and a snRNA-seq dataset ($n = 15,255$) through scenario 1. Then we projected a snATAC-seq dataset, which only consists of oligodendrocytes, into this atlas ($n = 8,557$). As Figure 2.14b shows, the oligodendrocytes are clearly separated from interneurons, while the structure of interneurons are retained. This indicates that scenario 3 can still detect outliers even if the query sample has extra cell types, even across modalities.

Next, we investigated cases where the cell types in the input datasets partially overlap. As was discussed in the previous section, `Online iNMF` (scenario 2) performs well at identifying the rare cell types in partially-overlapping datasets. We anticipate that scenario 3 is most useful for projecting small and specialized samples onto a large and comprehensive atlas, so we investigated performance when the reference dataset contains more cell types than the query (Figure 2.15). We first integrated 6 sc/snRNA-seq datasets from the MOp ($n = 344,675$) using `Online iNMF` in scenario 1. Afterwards, we held out the MGEs (i.e. Pvalb, Sst and Chodl cells), which are approximately 10.4% of all the cells, from the snATAC-seq dataset. Then we projected this processed ATAC dataset ($n = 49,167$) into the established atlas. The UMAP visualization annotated by our cell class labels is exhibited for reference (Figure 2.15a) and it shows that different cell types are effectively identified. By coloring the cells by their data sources (Figure 2.15b), it can be observed that very few cells from snATAC-seq data are spuriously aligned to the clusters corresponding to Pvalb, Sst and Chodl cells (highlighted in the red boxes).

2.6.7 `Online iNMF` Achieves Accurate Data Reconstruction

Here we demonstrate `Online iNMF`'s capability of data reconstruction by providing a supplementary figure showing only the reconstruction portion of the objective (Figure 2.23). The resulting iNMF factors do indeed reconstruct the data comparably to PCA, batch iNMF, and regular NMF. In this experiment, we evaluated the performance of `Online iNMF` on reconstructing the human PBMC dataset ($n = 13,999$) along with batch iNMF, regular NMF and PCA, using $2,001$ variable genes. Next, we implemented the listed methods on the scaled data using the same setting ($K = 30$ for all methods and $\lambda = 5$ for iNMF-based methods). The metric for comparison is the mean squared error (MSE) between the scaled and the reconstructed gene expression matrices. We repeated the experiment 10 times for iNMF/NMF-based approaches to account for the effect of

random initialization and reported the average MSE. As is displayed in the plot, the performances of `Online iNMF` (mean MSE $= 0.831$), batch iNMF (mean MSE $= 0.830$) and batch NMF (mean MSE $= 0.830$) are quite similar, while PCA (mean MSE $= 0.825$) accomplishes this task slightly better.

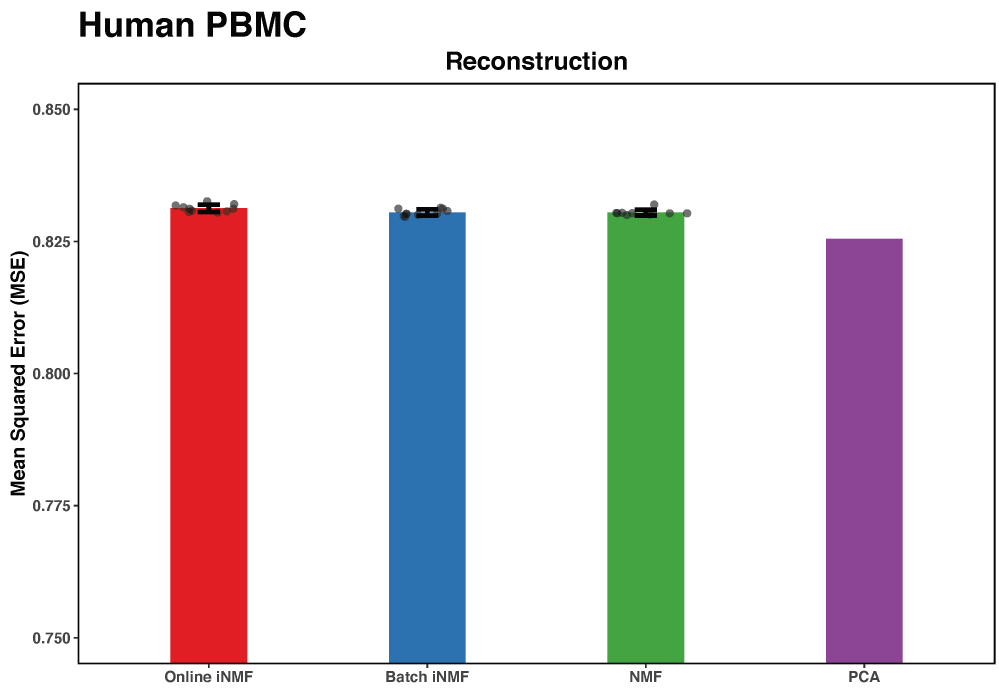


Figure 2.23: Comparison of data reconstruction among iNMF, NMF and PCA. Human PBMC dataset ($n = 13,999$) was used for this analysis. Average mean squared error (MSE) is shown and error bars indicate the standard deviation (10 random initializations for each method except for PCA, which is deterministic). Individual data points are shown for the NMF approaches.

2.6.8 Selection of Key Parameters (Knd λ

Selecting the dimensionality of the latent space is a perennial challenge in unsupervised data analysis. Due to the lack of ground truth, there is no way to pick the single best value for this parameter. In our previous paper, we described a heuristic for guiding the selection of K, by identifying an"elbow" in the plot of K vs. factor entropy. This is analogous to picking the number of eigenvectors for principal component analysis by inspecting a plot of the eigenvalue spectrum. In general, cell populations with a larger number of distinct cell types/states benefit from a larger K; for example, a sample of frontal cortex contains many more distinct subsets of cells than a sample of peripheral blood mononuclear cells. In practice, any K value between 20 and 40 usually gives reasonable results. Here we added analyses to demonstrate that `Online iNMF` performs

well across different choices of K on simulated data (Figure 2.22). The results show that, with any K in the range of 10 to 25, `Online iNMF` (scenario 1 or 2) successfully aligns the datasets and recovers the 8 true cell clusters.

We also examined the effect of regularization parameter λ on data alignment (Figure 2.24). To do so, we jointly analyzed the human PBMC datasets while varying λ and fixing K at 20. Similar to the original LIGER paper, an "elbow" shape was observed, which implies that the alignment quality remains robust for any $\lambda \geq 1$.

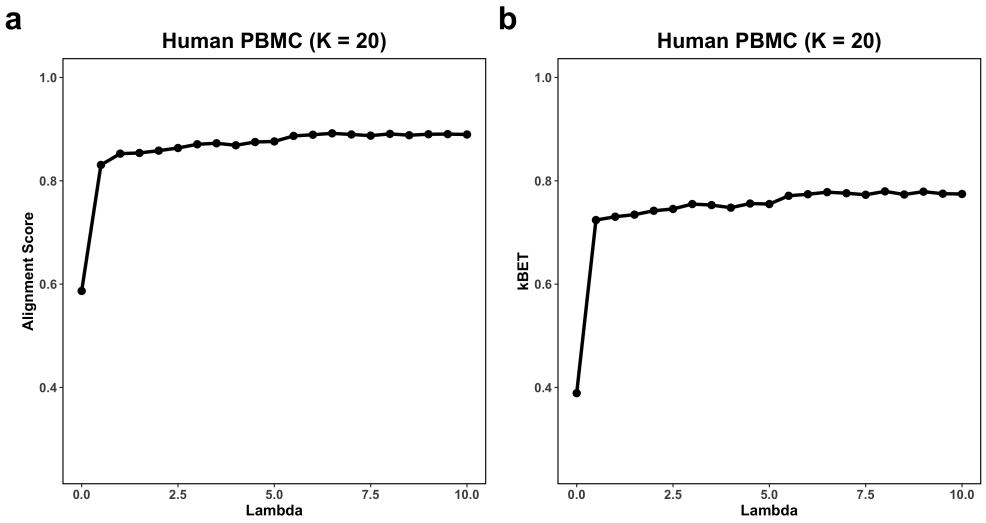


Figure 2.24: Selecting λ on human PBMC dataset. The human PBMC datasets ($n = 13,999$) were used to demonstrate the effect of λ on the data integration. Alignment score **(a)** and kBET **(b)** are reported to quantitatively assess dataset integration.

2.6.9 ANLS Outperforms HALS for Updating Cell Factor Loadings

We discovered an implementation detail that is crucial for achieving optimal `Online iNMF` performance: using ANLS to calculate H (cell factor loading) updates. Although in principle either HALS (hierarchical alternating least squares) or ANLS can be used to update $H_M^{i(t)}$ for each minibatch, we found empirically that the ANLS updates converge much faster than HALS updates (Figure 2.11). This may be because ANLS gives an optimal solution for all of the cell factor loadings (rows of H) simultaneously, while HALS updates are only optimal for one individual column of H at a time, requiring multiple iterations. We note also that Mairal et al. (2010) opted for least angle regression (LARS), which is directly analogous to our ANLS update, rather than a HALS-like update for H in their implementation.

CHAPTER 3

Integrating Single-Cell Multimodal Epigenome Data Using 1D-Convolutional Neural Networks

Recent experimental developments enable single-cell multimodal epigenomic profiling, which measures multiple histone modifications and chromatin accessibility within the same cell. Such parallel measurements provide exciting new opportunities to investigate how epigenomic modalities vary together across cell types and states. A pivotal step in using this type of data is integrating the epigenomic modalities to learn a unified representation of each cell, but existing approaches are not designed to model the unique nature of this data type. Our key insight is to model single-cell multimodal epigenome data as a multi-channel sequential signal. Based on this insight, we developed `ConvNet-VAE`s, a novel framework that uses 1D-convolutional variational autoencoders (VAEs) for single-cell multimodal epigenomic data integration. We evaluated `ConvNet-VAE`s on nano-CT and scNTT-seq data generated from juvenile mouse brain and human bone marrow. We found that `ConvNet-VAE`s can perform dimension reduction and batch correction better than previous architectures while using significantly fewer parameters. Furthermore, the performance gap between convolutional and fully-connected architectures increases with the number of modalities, and deeper convolutional architectures can increase performance while performance degrades for deeper fully-connected architectures. Our results indicate that the convolutional autoencoders are a promising method for integrating current and future single-cell multimodal epigenomic datasets.

3.1 Introduction

Single-cell sequencing technologies have revolutionized our understanding of cellular heterogeneity and the complexity of biological systems. Recently, single-cell multimodal chromatin profiling has emerged as an exciting new experimental approach to investigate the cellular epigenetic landscape. Two independent studies fused nanobodies (nb) to a transposase enzyme (Tn5) and used these nb-Tn5 conjugates to detect up to three epigenome layers (histone modification or chromatin accessibility) within the same cell (Bartosovic and Castelo-Branco 2022, Stuart et al. 2022). The nano-CT and scNTT-seq technology can in principle be used to detect transcription factor binding

as well, though this has not yet been demonstrated. These multimodal datasets provide simultaneous measurements of multiple epigenomic layers within individual cells, offering unprecedented opportunities to unravel how histone modifications and chromatin states drive cellular diversity. For example, one can use this type of data to investigate how different histone modifications at the same genomic locus combine to activate or repress transcription of nearby genes. However, the structure of single-cell multimodal epigenomic data is unique compared to other single-cell data types: each modality is a one-dimensional genomic track, and the total measurement for a cell consists of multiple one-dimensional tracks measured at the same genomic positions. This is quite different from any other type of single-cell data–such as scRNA-seq, snATAC-seq, CITE-seq, or 10X multiome–in which the space of features is most naturally represented in terms of genes or discrete peaks.

A number of computational approaches have been designed to perform joint dimension reduction on single-cell multimodal data types such as CITE-seq and 10X multiome. For example, the `Seurat` weighted nearest neighbor algorithm, the multi-omic factor analysis (`MOFA+`), and the `multiVI` perform linear or nonlinear dimension reduction on single-cell multimodal datasets that can be represented as genes and peaks (Hao et al. 2021, Argelaguet et al. 2020, Ashuach et al. 2023). Approaches based on variational autoencoders (VAEs) are especially powerful for learning joint representations from single-cell multimodal data. VAEs are unsupervised probabilistic deep learning models that excel at distilling compact and meaningful representations of complex data, as evidenced by their successful applications in single-cell RNA-sequencing (scRNA-seq) data integration (Lopez et al. 2018). For multimodal problems, a VAE based on the concept of the Product of Experts (PoE) was introduced (Wu and Goodman 2018). This method factorizes the joint distribution over the latent variables into a product of conditional distributions, each representing the output of a modality-specific "expert" model. Each expert is comprised of a encoder and a decoder, designed to model a specific data modality. Beyond their initial applications in image transformation and machine translation, such VAEs have been adapted for multimodal single-cell sequencing data. For example, `Cobolt` and `multiVI` use multimodal VAEs to integrate paired measurements, such as gene expression and chromatin accessibility (peaks), and learn a unified cell embedding for cell clustering and visualization (Gong et al. 2021, Ashuach et al. 2023). Although multimodal VAEs can in principle use any type of neural network layers, single-cell multimodal VAEs have only used fully-connected layers due to the unordered nature of gene features. Thus, we refer to these previous approaches as `FC-VAE`s.

However, directly applying such approaches to single-cell multimodal epigenomic data has several disadvantages. First, it requires calling peaks separately on each epigenomic layer, which results in extremely high-dimensional data because each epigenomic modality is measured across the whole genome. The number of peaks per modality usually exceeds 10^5, and the peaks often do

not overlap across modalities, further increasing the number of peaks as the number of modalities per cell increases. Second, by using a peak-centric feature representation, previous approaches neglect the ordered sequential nature of single-cell epigenomic data, in which the epigenomic state of a particular locus shares strong conditional dependence with the states of loci immediately before and after it in linear genome order. Finally, using genes and peaks neglects the multi-track nature of single-cell multimodal epigenomic data, removing the crucial information of shared genome position across modalities. This third limitation is especially problematic because it prevents integration algorithms from learning the relationship among different epigenome modalities at a given position within a single cell, which is one of the key motivations for performing single-cell multimodal epigenomic measurement in the first place.

One-dimensional convolutional neural networks (1D-CNNs) have shown success in the analysis of sequential data, especially when the spatial or temporal relationships within the data are crucial (Kiranyaz et al. 2021). In particular, deep learning models using 1D CNN layers have been widely used in analysis of bulk RNA-seq and bulk epigenome data. Such networks have been trained on bulk data from cell lines and tissues to predict transcriptional and epigenetic profiles from DNA sequence (Kelley et al. 2018, Chen et al. 2022). Recently, Yuan et al. extended this line of work to single-cell ATAC-seq data: `scBasset` (Yuan and Kelley 2022) takes DNA sequences as input and utilizes CNNs to predict chromatin accessibility in single cells. However, to our knowledge, only `FC-VAE`s have been used to perform dimension reduction and integration of single-cell data.

Here, we present a novel 1D convolutional variational autoencoder framework (`ConvNet-VAE`s) tailored for integrating single-cell multimodal epigenomic data. We model single-cell multimodal epigenomic data as a multi-channel sequential signal. A key innovation of our method is that, by performing convolution over ordered feature space, it adopts a more appropriate inference bias than VAEs with only the fully-connected layers that are suitable for unordered features. Our approach combines two streams of work: 1D CNNs for bulk genomic data and VAEs for dimension reduction of single-cell data. Importantly, our method is fundamentally different from this previous work in several key aspects: (1) we utilize a window-based genome binning strategy on the multimodal profiles from single cells and model the fragment count in each bin; (2) we use 1D convolutional layers that operate over different epigenetic modalities instead of nucleotide bases; and (3) unlike the previous multimodal VAEs, `ConvNet-VAE`s consists of only one encoder-decoder pair. We show that `ConvNet-VAE`s can leverage the strengths of both VAEs and convolution. They effectively reduce data dimensionality and extract local genomic features with a more economical parameter usage compared to that of `FC-VAE`s.

3.2 `ConvNet-VAE`: 1D-convolutional neural networks for single-cell multimodal epigenomics integration

We introduce `ConvNet-VAE`, a novel approach designed to efficiently learn biologically meaningful low-dimensional cell representations from high-throughput single-cell multimodal epigenomic data. This framework capitalizes on recent advancements in chromatin profiling technologies which permit parallel measurements of histone modifications (e.g., H3K27ac, H3K27me3) and chromatin states at single-cell resolution (Bartosovic and Castelo-Branco 2022, Stuart et al. 2022). The sequenced fragments over the genome are obtained from each individual cell (Figure 3.1a).

Because single-cell multimodal epigenomic experiments measure different features over the same sequential domain (i.e., the genome), we reasoned that the data is most naturally represented as a multi-channel 1D sequential signal. This is a quite different approach than previous single-cell multimodal neural networks, which treat each modality as if it measured completely unrelated features (e.g., distinct genes or peak locations for each modality). Additionally, previous approaches often use a separate encoder and decoder network for each modality, while ours uses a single encoder and a single decoder that operate on multi-channel signals. By operating on this multi-channel representation of the data, we introduce an appropriate inductive bias that significantly reduces the number of parameters and enforces statistical dependence among neighboring genomic locations within a modality and across modalities at a given genomic locus.

`ConvNet-VAE` is a convolutional variational autoencoder based upon a Bayesian generative model (Figure 2.1b). To apply 1D-convolutional filters (Conv1D), the input multimodal data are transformed into 3-dimensional arrays (cell $\times$ modality $\times$ bin), following window-based genome binning at 10 kilobase resolution (Chen et al. 2019b) (Figure 2.1a). The encoder efficiently extracts latent factors, which are then mapped back to the input feature space by the decoder network. We use a discrete data likelihood (Poisson distribution) to directly model the observed raw counts.

We also extended `ConvNet-VAE`s to incorporate conditional information such as experimental batches, allowing batch correction using conditional VAEs, which has proven an effective strategy for scRNA-seq data (Lopez et al. 2018). In our model, the categorical variables (e.g., batch information) are one-hot encoded and then concatenated with the flattened convolutional layer outputs, instead of being combined directly to the multimodal fragment count data over the sorted genomic bins. We incorporated the conditional information in this way because, unlike fully-connected layers, convolutional layers are more naturally suited to accommodate sequential data, rather than one-hot encodings. Thus, we found it more natural to inject the batch information after the convolutional layers.

In the following sections, we showcase the effectiveness and superiority of `ConvNet-VAE`s by

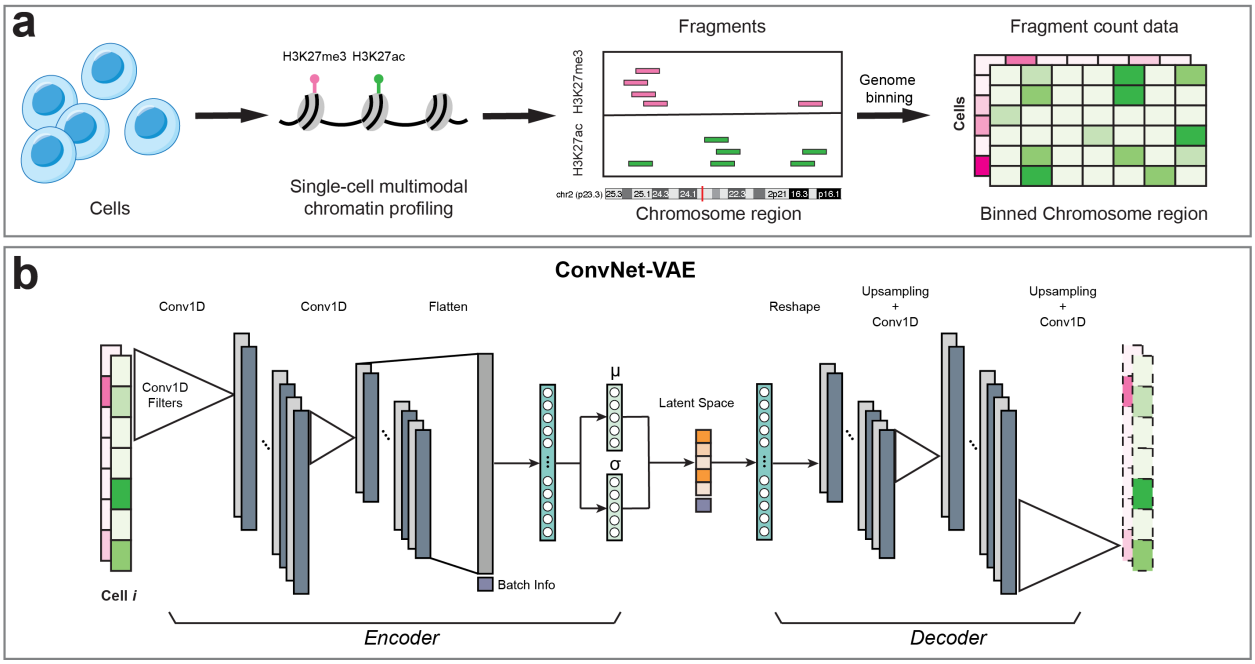


Figure 3.1: Overview of ConvNet-VAE. (**a**) For each cell, the fragments of each measured epigenomic modalities are acquired by multimodal single-cell epigenome profiling (e.g., H3K27ac + H3K27me3). Followed by genome binning, we obtain the fragment count data with dimension cell $\times$ modality $\times$ Bin. (**b**) `ConvNet-VAE` applies 1D-convolution and learns low-dimensional representations of the cells from the binned multimodal fragment count of input.

evaluating them on real data and comparing with FC−VAEs.

3.3 Results

3.3.1 ConvNet−VAEs learn cell representations using fewer parameters

A key advantage of ConvNet−VAEs is the proper inductive bias induced by convolution, which should result in considerable parameter savings. This advantage should increase with the number of modalities per cell: The number of peaks per modality usually exceeds 10^5, and the peaks often do not overlap across modalities.

To investigate the advantage of ConvNet−VAEs on real data, we analyzed a recently published single-cell bimodal dataset from juvenile mouse brains generated by the nano-CUT&Tag (nano-CT) technology (Bartosovic and Castelo-Branco 2022). After preprocessing, the dataset consists of $11,981$ cells from 4 experimental batches with H3K27ac and H3K27me3 modalities. We extracted the top $25,000$ bins identified across both modalities as the input feature set. We then separately examined the effects of (1) kernel size and stride and (2) number of convolutional layers on number of parameters and performance of ConvNet−VAE models. When examining the effects of kernel size and stride, we used architectures with a single convolutional layer and varied kernel size (K) and stride (S) from 11 to 51, with $K = S$ in each case. Second, we examined the effects of varying the number of convolutional layers from 1 to 3, while keeping a fixed kernel size of 11 and stride of 3. (Note that we used a smaller $S = 3$ with multiple convolutional layers to avoid the output dimensionality being too small.) We compared all models against FC−VAEs. To ensure a fair comparison, we ran all models through 5-fold cross-validation, with 300 training epochs.

Single-Conv1D-layer ConvNet−VAEs do indeed require fewer trainable parameters than FC−VAEs in this setting. For example, ConvNet−VAE (K51, S51) only uses 20% of the parameters that are needed for FC−VAEs, while ConvNet−VAE (K31, S31) uses 33%. As shown in Figure 3.2c, as the number of convolutional layers increases, ConvNet−VAEs uses fewer parameters. According to the UMAP visualization (colored by the published labels) of the cell embeddings obtained by the selected models, ConvNet−VAEs from varying K, S, and number of layers result in qualitatively similar embeddings compared to the FC−VAE (Figure 3.2a).

ConvNet−VAEs took slightly longer to complete the training (Figure 3.2d). The most compact ConvNet−VAE (K51, S51) led to a 2.5% decrease in average marginal log-likelihood on the validation sets (Figure 3.2e), but the (K11, S11) model achieved comparable or better marginal likelihood using 1M fewer parameters than the FC−VAE. Increasing the number of convolutional layers or stride resulted in worse marginal likelihood. However, the models with slightly worse marginal likelihood still excelled in learning low-dimensional cell representations that could reproduce the published cluster assignments (Figure 3.2f). The Adjusted Rand Index (ARI) first

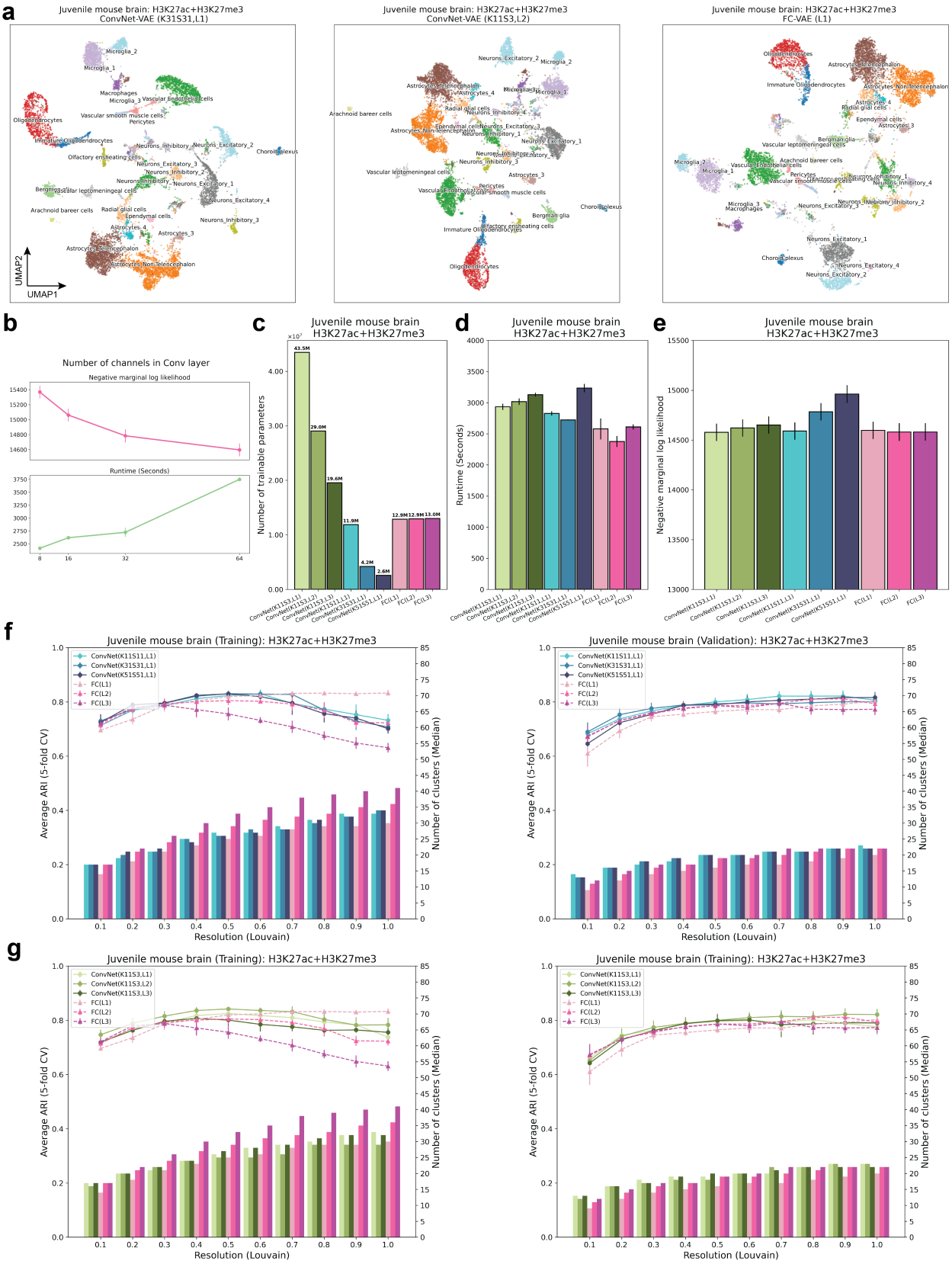


Figure 3.2: `ConvNet-VAE`s integrate single-cell bimodal epigenomic profiling data from mouse brain.

ConvNet-VAEs integrate single-cell bimodal epigenomic profiling data from mouse brain. (**a**) UMAP visualization of cell embeddings from ConvNet-VAEs (Left, Middle) and FC-VAE (Right). (**b**) For single-Conv1D-layer ConvNet-VAE, more channels in the convolutional layer lead to larger marginal log likelihood of the validation set at the cost of longer runtime in training, according to the result from 5-fold cross-validation. (**c**) The number of trainable parameters depends on the number of Conv1D layers and stride. ConvNet-VAEs from Group 1 (Blue) require fewer parameters than FC-VAEs (Pink) while those from Group 2 (Green) need more parameters. (**d**) Average training time is reported for each model. Error bars indicate standard deviation across 5-fold cross-validation. (**e**) Average negative marginal log likelihood of validation set estimated through importance sampling. Lower value implies larger marginal log likelihood. Error bars indicate standard deviation across 5-fold cross-validation. (**f**) Comparisons between ConvNet-VAEs with single Conv1D layer (Group 1) and FC-VAEs in terms of the quality of cell embeddings (training set: left; validation set: right). The bars show the median number of clusters obtained by the Louvain algorithm from 5 splits in cross-validation over a range of resolutions. The corresponding average Adjust Rand Index (ARI) is calculated by comparing to the published cell type labels, displayed as a line plot. Error bars indicate standard deviation across 5-fold cross-validation. (**g**) Comparisons between ConvNet-VAEs with multiple Conv1D layer (Group 2) and FC-VAEs in terms of cell embeddings' quality (training set: Left; validation set: Right), exhibited in the same way as (**f**).

increased as more cell clusters were identified by the Louvain algorithm at a higher clustering resolution, then decreased due to potential over-clustering. ConvNet-VAE (K51, S51) achieved the same highest ARI of 0.83 (average over 5 random runs of the Louvain clustering) as single-layer FC-VAE did on the training sets, and beat FC-VAEs with ARI of $0.82(\pm 0.01)$. Similarly configured ConvNet-VAE with smaller kernels and stride displayed a comparable pattern in cluster counts and ARI scores. The two-layer ConvNet-VAE performed slightly better in terms of ARI than the one-layer, while one fully-connected layer performed the best, with each additional layer worsening performance. In summary, this first set of tests indicates that ConvNet-VAE can achieve similar or better performance compared with FC-VAE using fewer parameters.

3.3.2 ConvNet-VAEs show a larger advantage with increasing number of modalities per cell

Because our approach treats each modality as a different channel along a shared sequential domain, we expect the advantage of our approach to increase with the number of modalities profiled per cell. To investigate this, we expanded the analysis by incorporating a 3rd modality, chromatin accessibility, which was measured alongside H3K27ac and H3K27me3 by the developers of nano-CT using assay for transposase-accessible chromatin (ATAC-seq) (Bartosovic and Castelo-Branco 2022). A total of $4,434$ cells from 2 experimental batches have ATAC, H3K27ac and H3K27me3 profiles (three modalities per cell). As in the previous section, we selected the $25,000$ bins with the

highest counts across modalities and generated a $4,434 \times 3 \times 25,000$ input for `ConvNet-VAE`s.

Through qualitative evaluation in the UMAP space, the single-Conv1D-layer `ConvNet-VAE` model with a large kernel and stride (K51, S51) results in more compact cell clusters than single-layer `FC-VAE` (Figure 3.2a), while requiring 87% fewer trainable parameters (Figure 3.2b). This efficiency remained notable even with a smaller kernel and stride (K11, S11), with a 39% reduction in parameters. The gap in runtime between `ConvNet-VAE`s and `FC-VAE`s also becomes narrower. For instance, single-Conv1D-layer `ConvNet-VAE` (K51, S51) takes 14% more time than the single-layer `FC-VAE`s to finish 300 training epochs, a decrease from the 25% longer runtime seen in the bimodal analysis (Figure 3.2c). There was no statistical difference in the marginal log-likelihoods across all investigated VAE variants (Figure 3.2d), implying equivalent capabilities in modeling the data distribution.

The advantage of `ConvNet-VAE`s becomes even more apparent when evaluating the quality of the cell embeddings (Figure 3.2e,f). On both training and validation sets, the single-Conv1D-layer `ConvNet-VAE`s lead in clustering accuracy (highest ARI: $0.78(\pm 0.02)$ at resolution 1.0 for training $0.69(\pm 0.01)$ at resolution 1.1 for validation), as compared to the `FC-VAE`s' highest ARI of $0.74(\pm 0.01)$ at resolution 0.6 for training and $0.67(\pm 0.03)$ at resolution 1.1 for validation (Figure 3.2e). This superiority is further supported by the performance of the multi-Conv1D-layer VAEs, which are top performers at almost all clustering resolutions (Figure 3.2f). For example, 2-Conv1D-layer `ConvNet-VAE` (K11, S3) stands out by producing an ARI of $0.81(\pm 0.01)$ and $0.72(\pm 0.01)$ on the training and validation sets respectively. Interestingly, unlike `FC-VAE`s, where additional layers usually lead to lower quality of the cell latent factors in the training data, `ConvNet-VAE`s can actually benefit from extra convolutional layers (Figure 3.2g, 3.2f). Furthermore, the `ConvNet-VAE`s are more effective than `FC-VAE`s as more modalities are added, as evidenced by the collective results from bimodal and trimodal integrative analyses.

3.3.3 `ConvNet-VAE`s allow for improved batch-effect correction

Single-cell data are often generated from different experiments, leading to batch effects that stem from technical rather than biological differences. Therefore, correcting for these effects is essential for clustering and visualization to accurately reflect the underlying biology. A number of methods have been introduced to address this problem in single-cell uni-modal data (Welch et al. 2019, Gao et al. 2021, Stuart et al. 2019, Korsunsky et al. 2019a). The same challenge occurs with these single-cell multimodal epigenomics datasets (Figure 3.2a, 3.2c). Without removing batch effects, the cells with bimodal and trimodal measurements from different datasets are poorly aligned, resulting in clusters that separate by dataset rather than underlying biological cell type.

Here, we selected `ConvNet-VAE`s with single and multiple Conv1D layers to demonstrate their capacity to remove batch-associated technical variation. There are four different batches in

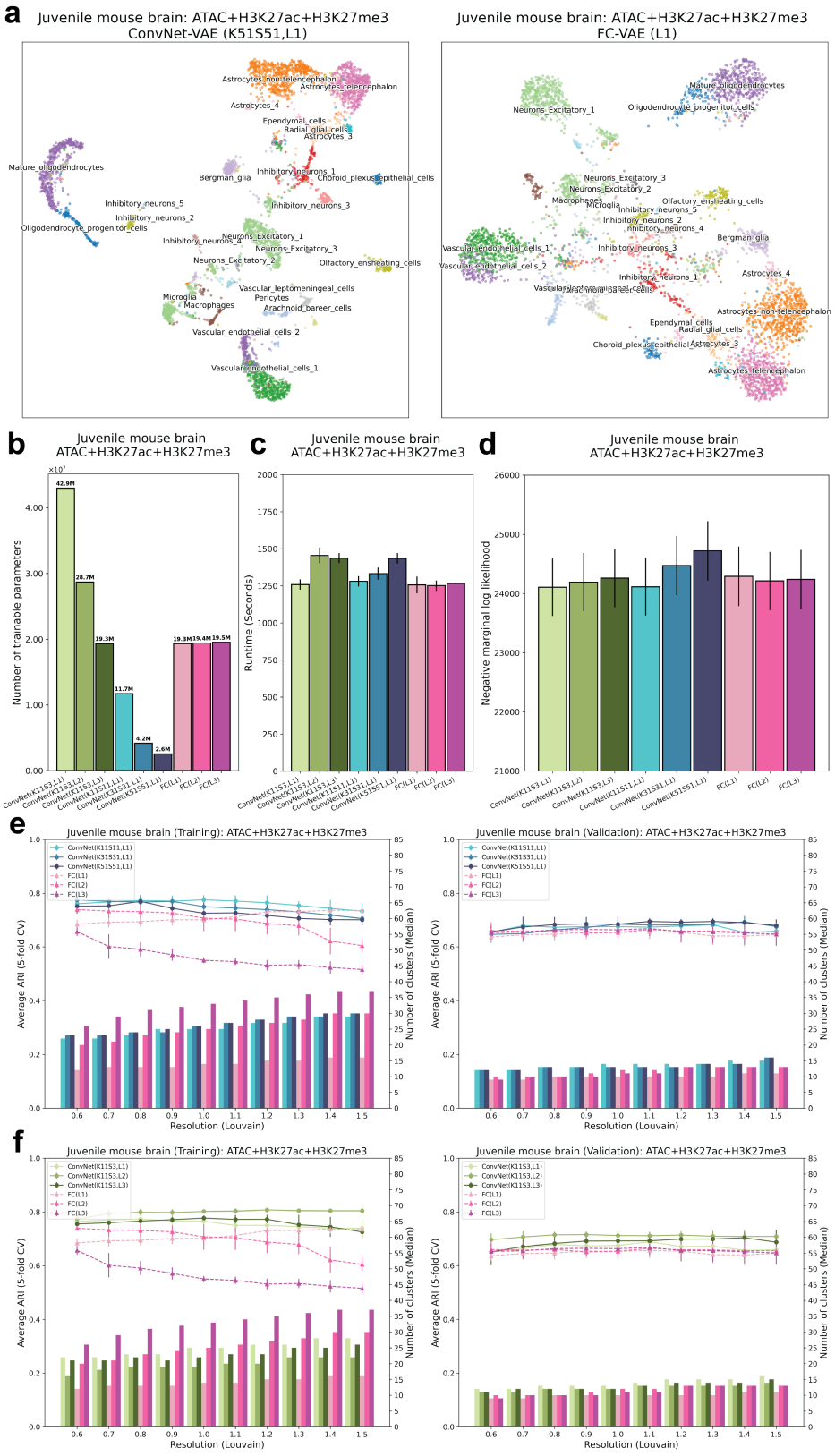


Figure 3.2: `ConvNet-VAE`s integrate single-cell trimodal epigenomic profiling data from mouse brain.

ConvNet-VAEs integrate single-cell trimodal epigenomic profiling data from mouse brain. (**a**) UMAP visualization of cell embeddings from ConvNet-VAEs (Left) and FC-VAE (Right). (**b**) The number of trainable parameters of ConvNet-VAEs from Group 1 (Blue), Group 2 (Green), and FC-VAEs (Pink). (**c**) Average training time is reported for each model. (**d**) Average negative marginal log likelihood of validation set estimated through importance sampling. Lower value implies larger marginal log likelihood. (**e,f**) Comparisons between ConvNet-VAEs and FC-VAEs in terms of the quality of cell embeddings (training set: left; validation set: right). The bars show the median number of clusters obtained by the Louvain algorithm from 5 splits in cross-validation over a range of resolutions. The corresponding average Adjust Rand Index (ARI) is calculated by comparing to the published cell type labels, displayed as a line plot. Error bars indicate standard deviation across 5-fold cross-validation.

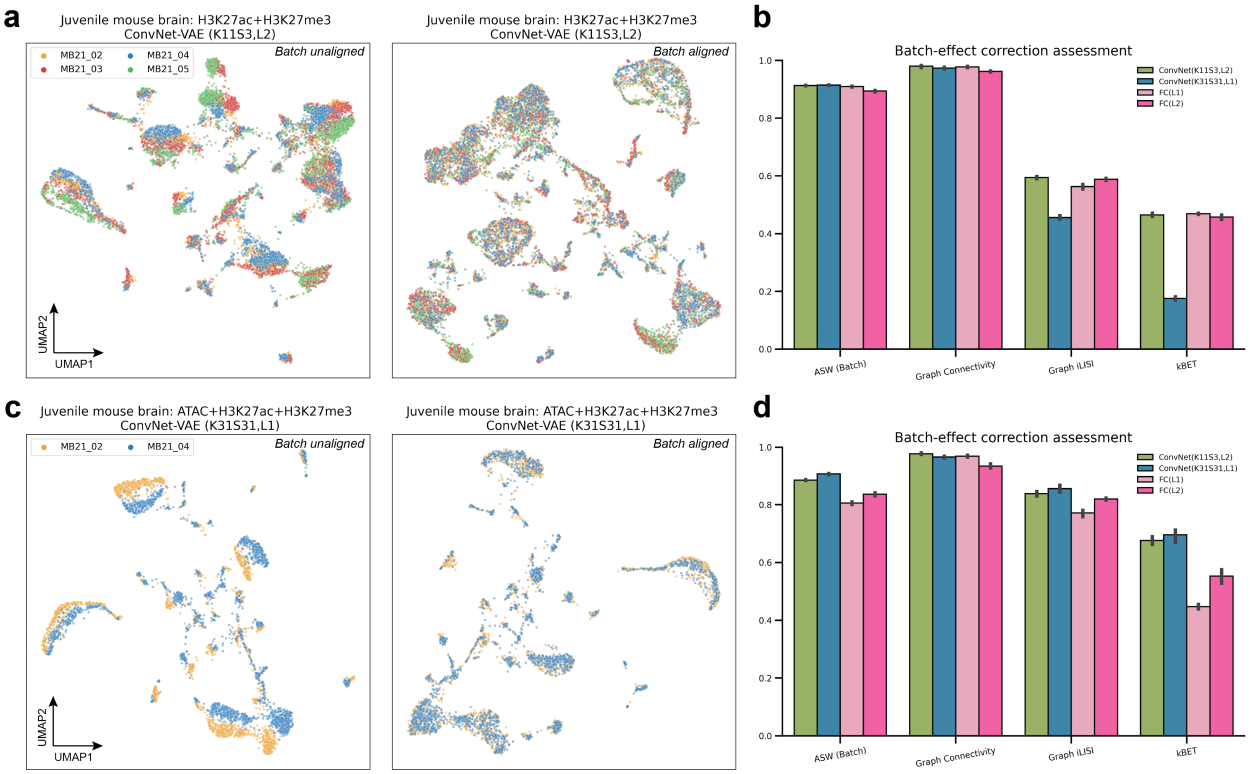


Figure 3.2: Benchmark of ConvNet-VAEs on batch-effect removal. VAE models were applied on the entire tested datsets without training/validation splitting. (**a**, **c**) UMAP visualizations of cell embeddings from selected models on the bi- and trimodal data, before and after alignment. (**b**, **f**) Quantitative comparison between ConvNet-VAEs and FC-VAEs based on four different metrics. The bars show average scores with standard deviation from 5 random runs.

the single-cell bimodal juvenile mouse brain data with measurement of H3K27ac and H3K27me3. When we apply the ConvNet-VAE with 2 convolutional layers (K11, S3) with batch information, the cells are well mixed in each cluster (Figure 3.2a). In the trimodal setting (simultaneous profiling of chromatin accessibility, H3K27ac, and H3K27me3), the cells from two replicates are clearly separated as shown. Based on the quality of batch mixing, the architecture with a single Conv1D-layer (K31, S31) successfully aligned these cells from both batches (Fig. 3.2c).

Beyond the qualitative evaluation, we further carried out quantitative assessment of batch-effect removal. To do this, we calculated four metrics: Average silhouette width (ASW, Batch), Graph Connectivity, Graph iLISI, and kBET (Luecken et al. 2022). In comparison to FC-VAEs, both selected ConvNet-VAEs showed similar or better performance in terms of ASW (Batch) and Graph connectivity, while the ConvNet-VAE with a single convolutional layer (K31, S31) is less favored with respect to Graph iLISI and kBET (Fig. 3.2b). Encouragingly, the selected ConvNet-VAEs excelled across all metrics when more modalities were involved (3.2d). The improvements in ASW (batch) and k-BET were particularly significant. These results align with the results from the previous section, indicating that ConvNet-VAEs show greater advantage over FC-VAEs as the number of epigenomic modalities increases.

3.3.4 ConvNet-VAEs integrate histone modifications from scNTT-seq data

In addition to nano-CT, Stuart et al. (Stuart et al. 2022) developed nanobody-tethered transposition followed by sequencing (scNTT-seq), enabling genome-wide measurement of multiple histone modifications at single-cell resolution. In this part, we showcase the adaptability and consistent performance of ConvNet-VAEs when applied to multimodal data obtained through varied sequencing methods.

Toward this goal, we integrated single-cell bimodal (H3K27ac and H3K27me3) epigenomic data profiled from bone marrow mononuclear cells (BMMCs) of healthy human donors ($N = 5,236$). According to the UMAP plots, H3K27ac itself doesn't carry sufficient information to distinguish different cell types, whereas H3K27me3 provides sufficient information to identify the major cell types. Combining both modalities with the selected single-layer ConvNet-VAE (K11, S11), we achieved more compact cell clusters (Figure 3.3a).

Although training ConvNet-VAEs with multiple convolutional layers, or those with larger kernels and strides, might require additional time compared to FC-VAEs (Figure 3.3c), the ConvNet-VAEs display comparable or superior performance in estimating the marginal log-likelihood for the validation data, making them preferable as generative models (Figure 3.3d). More strikingly, the proposed ConvNet-VAEs outperform the FC-VAEs on the training and validation sets by a large margin, when comparing ARI as the measure of the effectiveness of dimension reduction (Figure 3.3e,f).

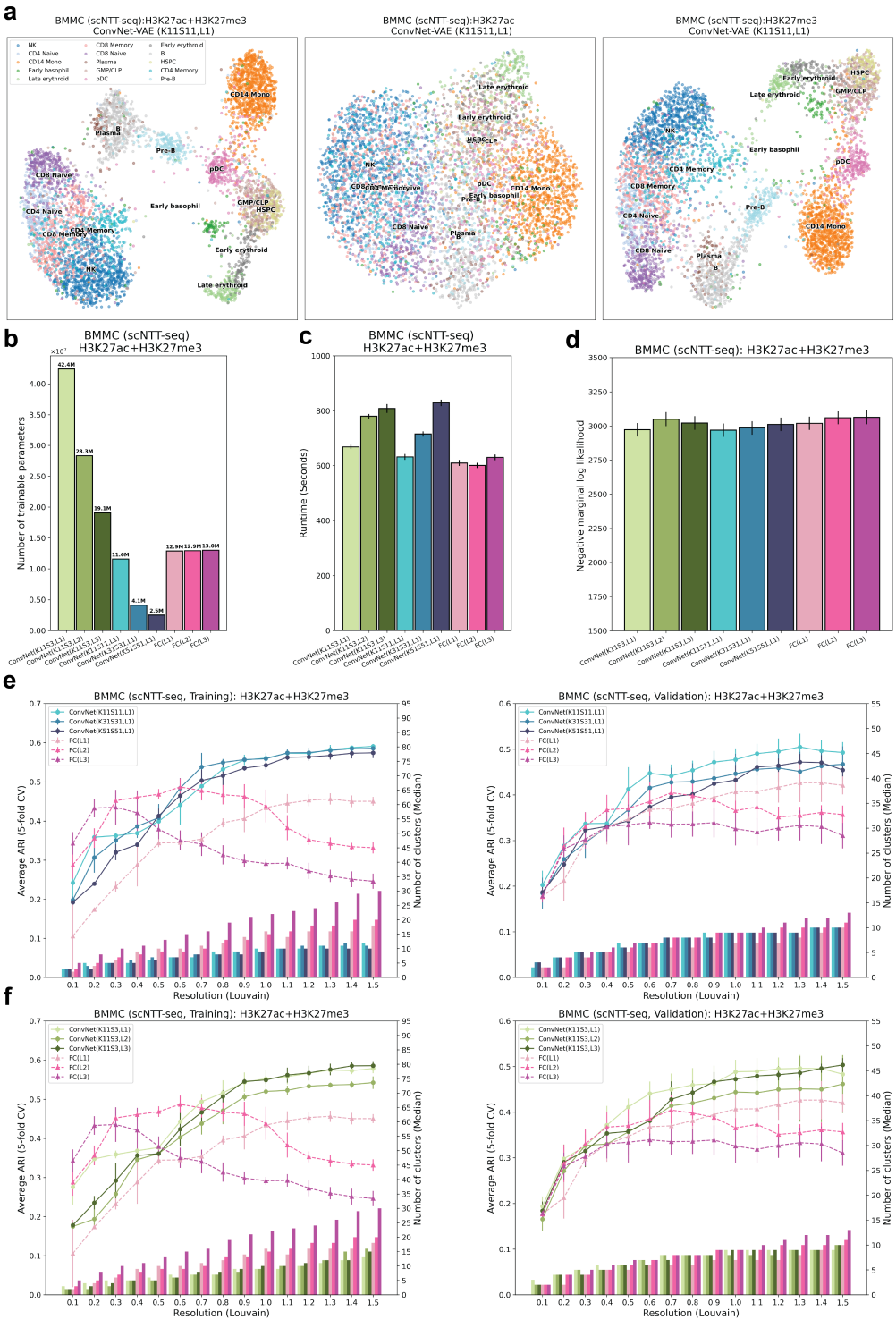


Figure 3.3: `ConvNet-VAE`s effectively integrate scNTT-seq data.

ConvNet-VAEs effectively integrate scNTT-seq data. (**a**) UMAP visualization of cell embeddings from ConvNet-VAE (single Conv1D layer, Kernel size 11, stride 11) on BMMCs data containing H3K27ac and H3K27me3 (Left), H3K27ac (Middle), and H3K27me3 (Right). (**b**) The number of trainable parameters of ConvNet-VAEs from Group 1 (Blue), Group 2 (Green), and FC-VAEs (Pink). (**c**) Average training time is reported for each model. (**d**) Average negative marginal log likelihood of validation set estimated through importance sampling. Lower value implies larger marginal log likelihood. (**e,f**) Comparisons between ConvNet-VAEs and FC-VAEs in terms of the quality of cell embeddings (training set: left; validation set: right). The bars show the median number of clusters obtained by the Louvain algorithm from 5 splits in cross-validation over a range of resolutions. The corresponding average Adjust Rand Index (ARI) is calculated by comparing to the published cell type labels, displayed as a line plot. Error bars indicate standard deviation across 5-fold cross-validation.

In order to examine whether the convolutional layers are able to exploit the sequential relationships among genomic locations, we randomly shuffled genomic bins from this BMMC dataset and re-analyzed it with ConvNet-VAEs. The decline in ARI upon bin shuffling confirmed that the convolutional layers are indeed sensitive to local epigenomic patterns. In terms of the marginal log-likelihood, the negative effect brought by bin shuffling becomes more apparent when larger kernel size is used (Supplementary Figure 3.3). All these observations underscore the ability of 1D-convolutional layers to capture spatial dependencies in the tested single-cell multimodal epigenomic data.

Moreover, we investigated the applicability of ConvNet-VAEs on unimodal single-cell data. In analyses of PBMCs gene expression, PBMC ATAC (peaks), as well as the mouse organogenesis ATAC (peaks), single-Conv1D-layer ConvNet-VAEs perform on par with FC-VAEs. ConvNet-VAEs lead the performance in reducing the dimension of the large-scale mouse cortex and hippocampus transcriptomic profile (Supplementary Figures 3.4,3.4,3.4,3.4).

3.4 Discussion

In this study, we proposed the ConvNet-VAE framework, specifically designed to model single-cell multimodal epigenomic data. This model comprises 1D-convolutional layers and hence takes multi-channel binned fragment counts as input. The encoder network within this framework learns low-dimensional representations of cells that facilitate cell type inference following clustering. We validated ConvNet-VAEs' utilities through integrative analyses of bimodal (H3K27ac + H3K27me3) and trimodal juvenile mouse brain data (ATAC + H3K27ac + H3K27me3), as well as bimodal data from human bone marrow mononuclear cells (H3K27ac + H3K27me3).

As demonstrated by the results, ConvNet-VAEs are able to extract information about chromatin states and histone modifications, accurately capture the data distribution, and correct for batch effects. The 1D-convolution layers are capable of capturing the spatial relationships among sequentially

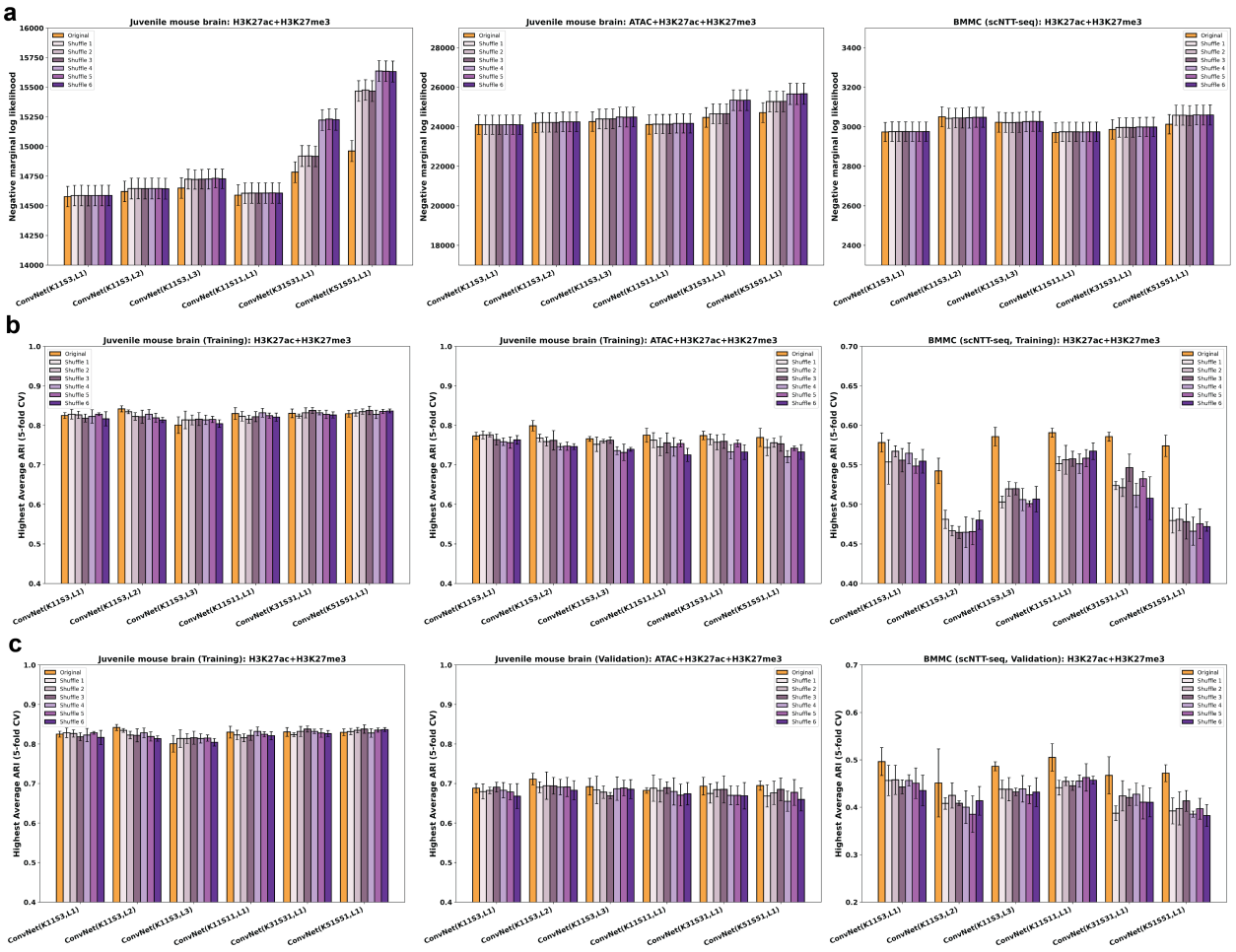


Figure 3.3: Performance of `ConvNet-VAE`s after shuffling genomic bins. Side-by-side comparison between the results from the ordered bins and shuffled bins. Bimodal juvenile mouse brain (Left column), trimodal junvile mouse brain (Middle column), BMMCs (Right column). (**a**) Comparison of the marginal log likelihood (validation set). (**b**) The highest average ARI that each model can achieve on the training sets over a range of clustering resolution. (**c**) The highest average ARI that each model can achieve on the validation sets over a range of clustering resolution. Error bars indicate the standard deviation from 5-fold cross-validation.

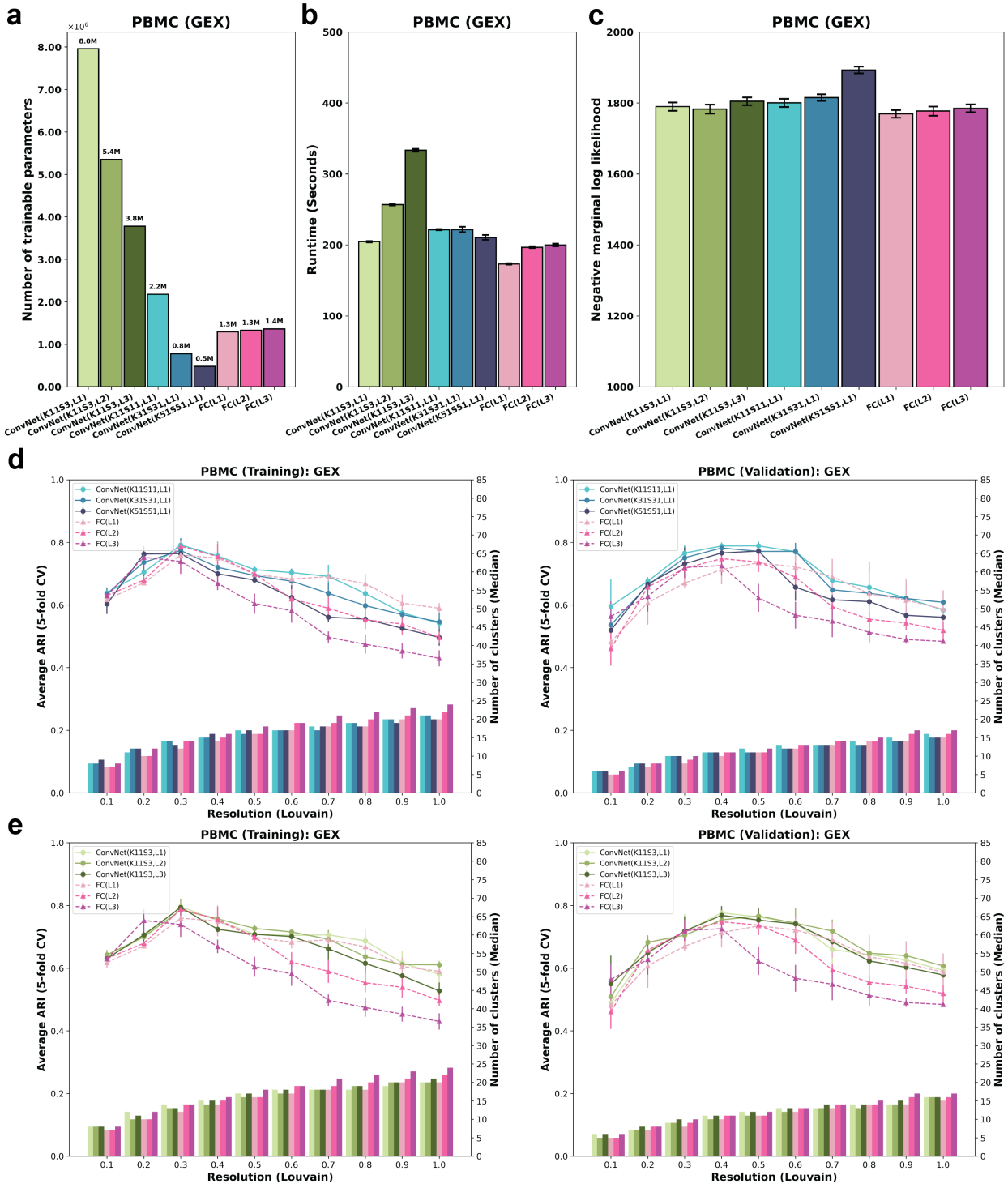


Figure 3.4: Evaluation of `ConvNet-VAE`s on PBMCs (gene expression).

Evaluation of `ConvNet-VAE`s on PBMCs (gene expression). (**a**) The number of trainable parameters of `ConvNet-VAE`s from Group 1 (Blue), Group 2 (Green), and `FC-VAE`s (Pink). (**b**) Average training time is reported for each model. Error bars indicate standard deviation across 5-fold cross-validation. (**c**) Average negative marginal log likelihood of validation set estimated through importance sampling. (**d,e**) Comparisons between `ConvNet-VAE`s and `FC-VAE`s on cell embeddings' quality. The bars show the median number of clusters obtained by the Louvain algorithm from 5 splits in cross-validation over a range of resolutions. The corresponding average Adjust Rand Index (ARI) is calculated by comparing to the published cell type labels (line plot). Error bars indicate standard deviation across 5-fold cross-validation.

arranged genomic bins. In qualitative and quantitative benchmarking with `FC-VAE`s, which solely utilize fully connected layers, `ConvNet-VAE`s show effectiveness by achieving on-par or enhanced performance using far fewer parameters. Unlike `FC-VAE`s, `ConvNet-VAE`s can also benefit from including more layers (Conv1D) in the model architecture. Notably, the advantage of `ConvNet-VAE`s over `FC-VAE`s becomes more evident when jointly analyzing three modalities instead of two.

Nevertheless, the `ConvNet-VAE`s presented in this report are not without limitations. Due to the use of convolutional filters, they require that all modalities share the same feature space (i.e. an identical set of bins). Moreover, there is potential to further refine model performance by optimizing parameters like kernel size and stride length.

To summarize, the `ConvNet-VAE` framework stands out for its performance in integrating single-cell multimodal epignomic data. We anticipate that the utility of our approach will become more promising as the number of modalities and cells in single-cell multimodal epigenomic datasets increases in the future.

3.5 Methods

3.5.1 Generative probabilistic model of epigenomic data

We modeled the count data of a given feature (e.g., a histone modification such as H3K27ac) by using a Poisson distribution. Consider multimodal single-cell data comprised of M modalities from B different experimental batches, with a total of N cells. All modalities share the same set of features G (e.g., binned genomic regions). We represent cell i with a latent factor $\mathbf{z}^i$ sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I})$, characterized by batch information $\mathbf{b}^i$, and a modality-specific library size factor l_m^i. We model the generative process of the count x_{mg}^i of the molecular feature g within modality m ($m \in \{1, 2, \ldots, M\}$) as follows:

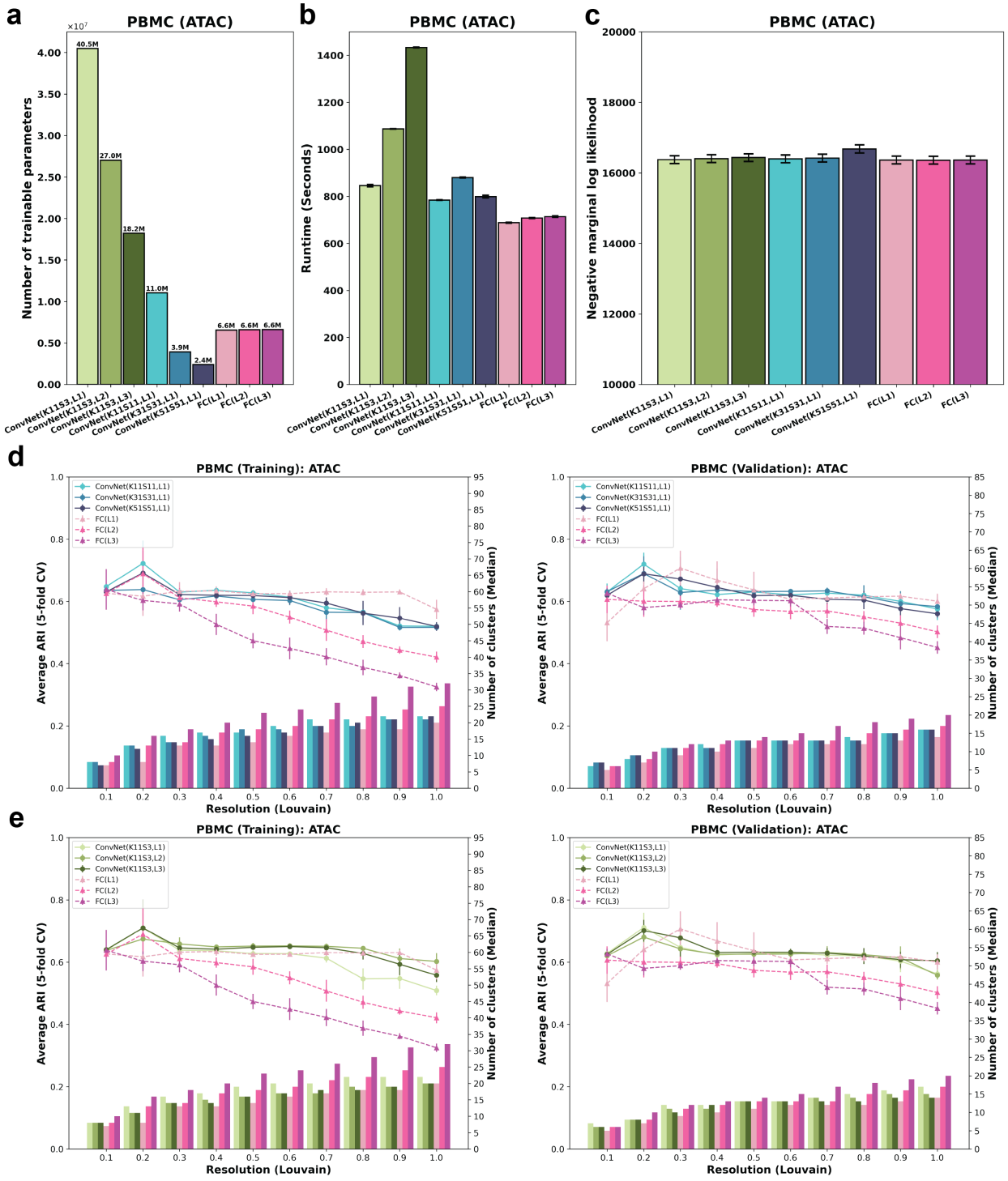


Figure 3.4: Evaluation of `ConvNet-VAE`s on PBMCs (ATAC peaks).

Evaluation of `ConvNet-VAEs` on PBMCs (ATAC peaks). (**a**) The number of trainable parameters of `ConvNet-VAEs` from Group 1 (Blue), Group 2 (Green), and `FC-VAEs` (Pink). (**b**) Average training time is reported for each model. Error bars indicate standard deviation across 5-fold cross-validation. (**c**) Average negative marginal log likelihood of validation set estimated through importance sampling. (**d,e**) Comparisons between `ConvNet-VAEs` and `FC-VAEs` on cell embeddings' quality. The bars show the median number of clusters obtained by the Louvain algorithm from 5 splits in cross-validation over a range of resolutions. The corresponding average Adjust Rand Index (ARI) is calculated by comparing to the published cell type labels (line plot). Error bars indicate standard deviation across 5-fold cross-validation.

$$
\begin{aligned}
\rho_{mg}^{i} &= f^{Dec}(\mathbf{z}^{i}, \mathbf{b}^{i}) \\
w_{mg}^{i} &= \mathrm{softmax}(\rho_{mg}^{i}) \\
\lambda_{mg}^{i} &= w_{mg}^{i} l_{m}^{i} \\
x_{mg}^{i} &\sim \mathrm{Poisson}(\lambda_{mg}^{i})
\end{aligned}
$$

Here, $\mathbf{x}_{mg}^{i} \in \mathbb{N}_0$ represents the count data, $\mathbf{z}^{i} \in \mathbb{R}^{D}$ is the latent representation of each cell in a D-dimensional space, with D selected according to the complexity of the data. The modality-specific library size factor is denoted as $l_{m}^{i} \in \mathbb{N}_0$. $\mathbf{b}^{i}$ is a B-dimensional one-hot encoded vector containing batch information. The function f^{Dec} denotes the decoder neural network, which consists of convolutional layers and/or fully connected layers. Through the application of a softmax activation function in the final layer, the decoder network maps the latent factors and batch label of cell i to the original feature space. In this study, we also implemented negative binomial (NB) distribution in the models, which is able accommodate overdispersion in the data by including an extra parameter for dispersion. In our experiments spanning three single-cell multimodal datasets, we observed that `ConvNet-VAEs` employing negative binomial distribution exhibited negligible differences compared to those utilizing Poisson distributions (Supplementary Figure 3.4). Under this distributional assumption, `ConvNet-VAEs` also maintain their edge over `FC-VAEs` (Supplementary Figures 3.5, 3.6, 3.7). With these observations, our studies focus on Poisson-based modeling in the rest of this report.

3.5.2 Multimodal variational autoencdoers

3.5.2.1 Variational autoencoders (VAEs)

As previously described, we consider the observed feature vector $\mathbf{x}_m$ of a cell derived from hidden variable $\mathbf{z}$, from batch $\mathbf{b}$. Researchers have harnessed the VAE framework for efficient approximation of the posterior distribution for $\mathbf{z}$ (Kingma and Welling 2013). VAEs, as deep

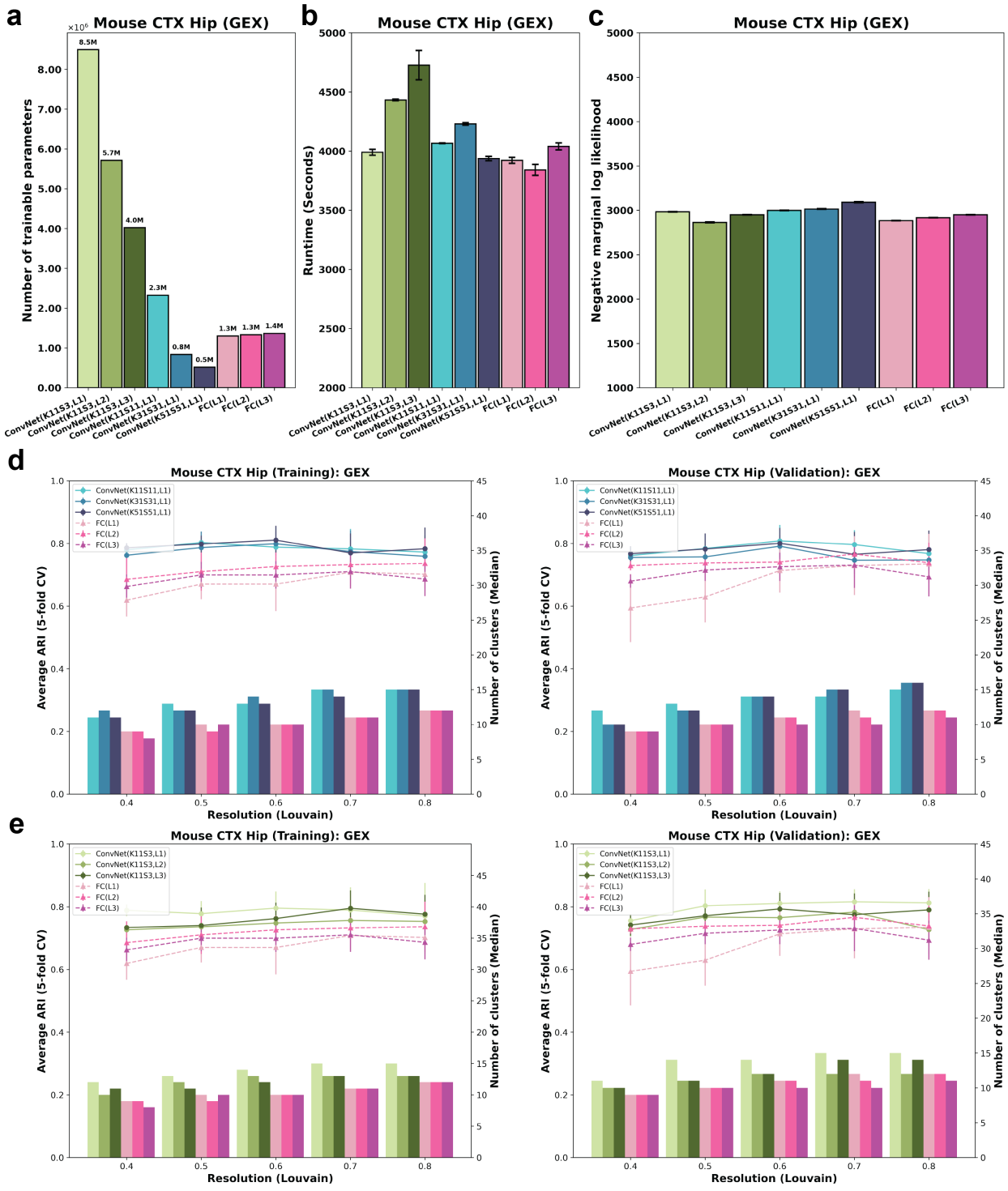


Figure 3.4: Evaluation of `ConvNet-VAE`s on mouse cortex and hippocampus (gene expression).

Evaluation of `ConvNet-VAE`s on mouse cortex and hippocampus (gene expression). (**a**) The number of trainable parameters of `ConvNet-VAE`s from Group 1 (Blue), Group 2 (Green), and `FC-VAE`s (Pink). (**b**) Average training time is reported for each model. Error bars indicate standard deviation across 5-fold cross-validation. (**c**) Average negative marginal log likelihood of validation set estimated through importance sampling. (**d,e**) Comparisons between `ConvNet-VAE`s and `FC-VAE`s on cell embeddings' quality. The bars show the median number of clusters obtained by the Louvain algorithm from 5 splits in cross-validation over a range of resolutions. The corresponding average Adjust Rand Index (ARI) is calculated by comparing to the published cell type labels (line plot). Error bars indicate standard deviation across 5-fold cross-validation.

generative models, exploit neural networks for variational inference, facilitating representation learning from high-dimensional data. The functionality is crucial for single-cell data integration and subsequent cell type identification (Lopez et al. 2018). Typically, VAEs are trained to optimize the evidence lower bound (ELBO) using stochastic gradient methods. In a unimodal scenario where $M = 1$, the ELBO for a feature vector $\mathbf{x}_1$ is defined as follows:

$$\text{ELBO}(\mathbf{x}_1) \triangleq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}_1,\mathbf{b})}[\log p_\theta(\mathbf{x}_1|\mathbf{z},\mathbf{b})] - D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}_1,\mathbf{b}) \parallel p(\mathbf{z})), \tag{3.1}$$

where $q_\phi(\mathbf{z}|\mathbf{x}_1,\mathbf{b})$ and $p_\theta(\mathbf{x}_i|\mathbf{z},\mathbf{b})p(\mathbf{z})$ are the inference model (parameterized by ϕ) and generative model (parameterized by θ) respectively. To address the challenges in modeling multimodal single-cell data, we introduce multimodal VAEs in the next sections.

3.5.2.2 Convolutional variational autoencoders with 1D-convolutional layers (`ConvNet-VAE`)

Convolutional neural networks (CNNs) effectively perform tasks such as data compression and classification by learning representations of the input (for example, 1D for signals or sequences, 2D for images) (LeCun et al. 2015). In the context of 1D-CNNs, `Conv1D` filters work on the 1D input sequences and move in one direction. We introduce `ConvNet-VAE`, a variational autoencoder architecture that utilizes 1D-convolutional layers to model and integrate single-cell multimodal epigenomic data. By incorporating `Conv1D` layers, `ConvNet-VAE` efficiently embeds high-dimensional multimodal epigenomic features of the cells into a low-dimensional space suitable for clustering tasks. For compatibility with 1D-CNN, we treat the fragment count of different modalities along the binned genome as 1D sequence with multiple channels, where each channel corresponds to a different modality. Given N cells, then we have $\{\mathbf{X}^i\}_{i=1}^N$, and $\mathbf{X}^i \in \mathbb{N}_0^{M\times G}$. For instance, in a bimodal setting, $\mathbf{x}^i_{1\cdot}$ denotes the first channel, and $\mathbf{x}^i_{2\cdot}$ the second. The ELBO is formulated as below.

$$\text{ELBO}(\mathbf{X}) \triangleq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{X},\mathbf{b})}[\log p_\theta(\mathbf{X}|\mathbf{z},\mathbf{b})] - D_{KL}(q_\phi(\mathbf{z}|\mathbf{X},\mathbf{b}) \parallel p(\mathbf{z})) \tag{3.2}$$

Note that we assume different modalities $\mathbf{x}_m$ (channels of $\mathbf{X}$) are conditionally independent on $\mathbf{z}$ and $\mathbf{b}$ for tasks involving multiple modalities.

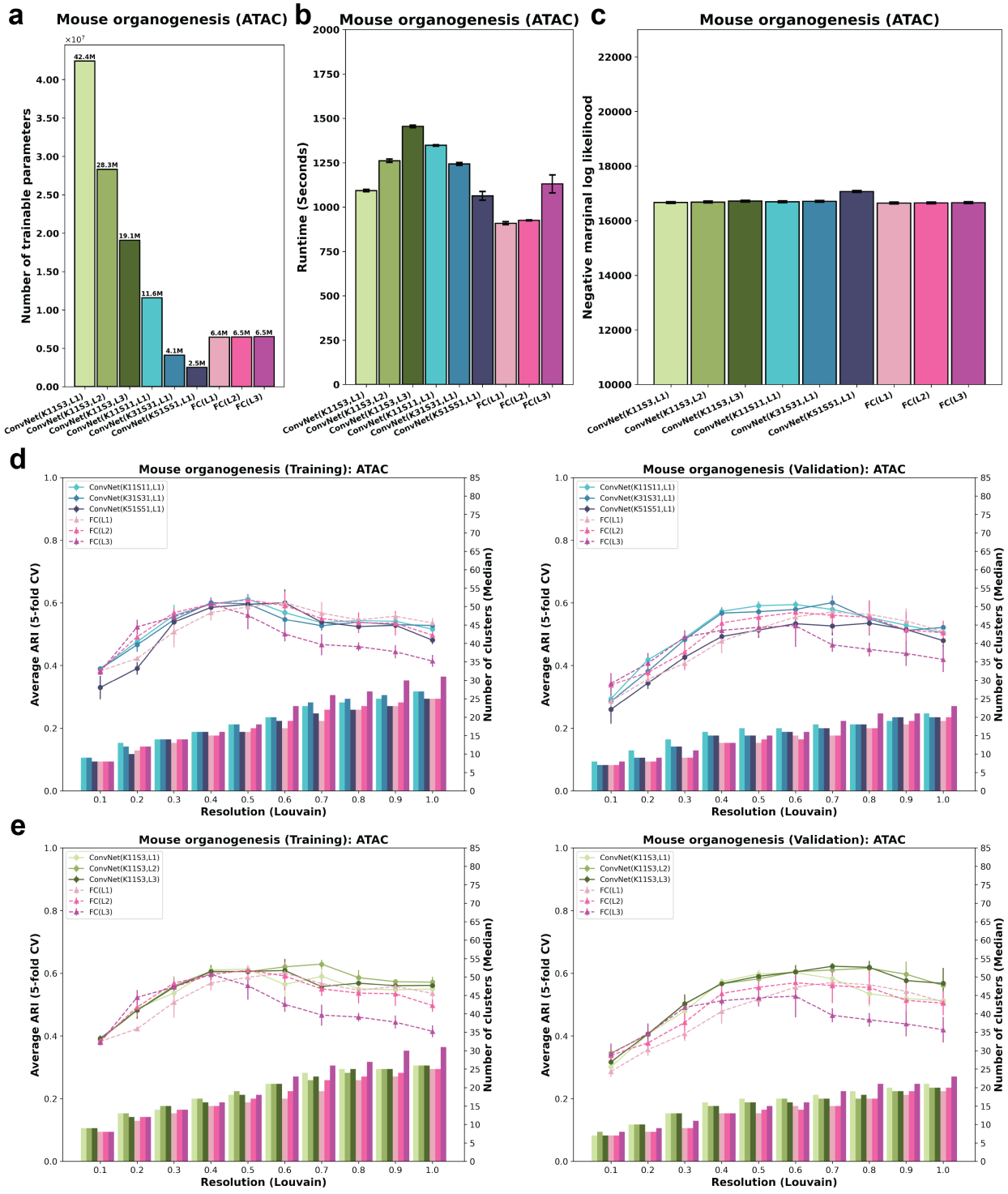


Figure 3.4: Evaluation of `ConvNet-VAE`s on mouse organogenesis (ATAC peaks).

Evaluation of `ConvNet-VAE`s on mouse organogenesis (ATAC peaks). (**a**) The number of trainable parameters of `ConvNet-VAE`s from Group 1 (Blue), Group 2 (Green), and `FC-VAE`s (Pink). (**b**) Average training time is reported for each model. Error bars indicate standard deviation across 5-fold cross-validation. (**c**) Average negative marginal log likelihood of validation set estimated through importance sampling. (**d,e**) Comparisons between `ConvNet-VAE`s and `FC-VAE`s on cell embeddings' quality. The bars show the median number of clusters obtained by the Louvain algorithm from 5 splits in cross-validation over a range of resolutions. The corresponding average Adjust Rand Index (ARI) is calculated by comparing to the published cell type labels (line plot). Error bars indicate standard deviation across 5-fold cross-validation.

The architecture of `ConvNet-VAE` is depicted in Figure 3.1. This research focuses on two main configurations of `ConvNet-VAE` models. The first group of models comprises a single convolutional layer with varying sizes of kernel (K) and stride (S). The second group features multiple convolutional layers with constant kernel size and stride. By experimenting single-Conv1D-layer `ConvNet-VAE` (with a kernel size of 31 and stride of S31) on the bimodal juvenile mouse brain dataset, we notice an increase in the marginal log-likelihood of validation data when more kernels (output channels) are applied. However, there is a disproportionately large increase in computational time compared to the gains in capturing the data distribution when the kernel count is doubled from 32 to 64 (Figure 3.2b). Therefore, for single-Conv1D-layer `ConvNet-VAE`s, we set the kernel count to 32. In the case of models incorporating a second or third convolutional layer, the output channels are set to 64 and 128, as is commonly done in CNN architectures. Complete specifications are provided in Table 3.1, including the number of feature channels produced by the convolutional layers (indicated in the parentheses). The final Conv1D layer in the decoder produces an output with a channel count that matches the number of data modalities.

In general, Conv1D and FC layers are followed by Batch Normalization (1D), ReLU activation, and Dropout layers. We perform softmax activation on the output from the last decoding Conv1D layer, without any other transformation. FC(μ,σ) as well as the FC layer in the decoder are linear layers. The pooling layer is replaced by applying a large stride (≥ 3). For enhanced numerical stability, each input channel—representing a different modality—undergoes log transformation ($\log(\mathbf{x}+\mathbf{1})$).

3.5.2.3 Variational autoencoders with fully connected layers (`FC-VAE`)

In order to demonstrate the advantage of `ConvNet-VAE`, we include `FC-VAE` for benchmark analyses. To address the problem of learning joint representations of multiple modalities, the idea of product-of-experts (PoE) was introduced by Wu and Goodman in 2018 (Wu and Goodman 2018). We adapted the PoE approach for our specific task of multimodal inference within the context of single-cell epigenomics. Consistent with the settings described in the previous sections, we establish the following joint posterior by assuming conditional independence between $p(\mathbf{x}_m|\mathbf{z},\mathbf{b})$,

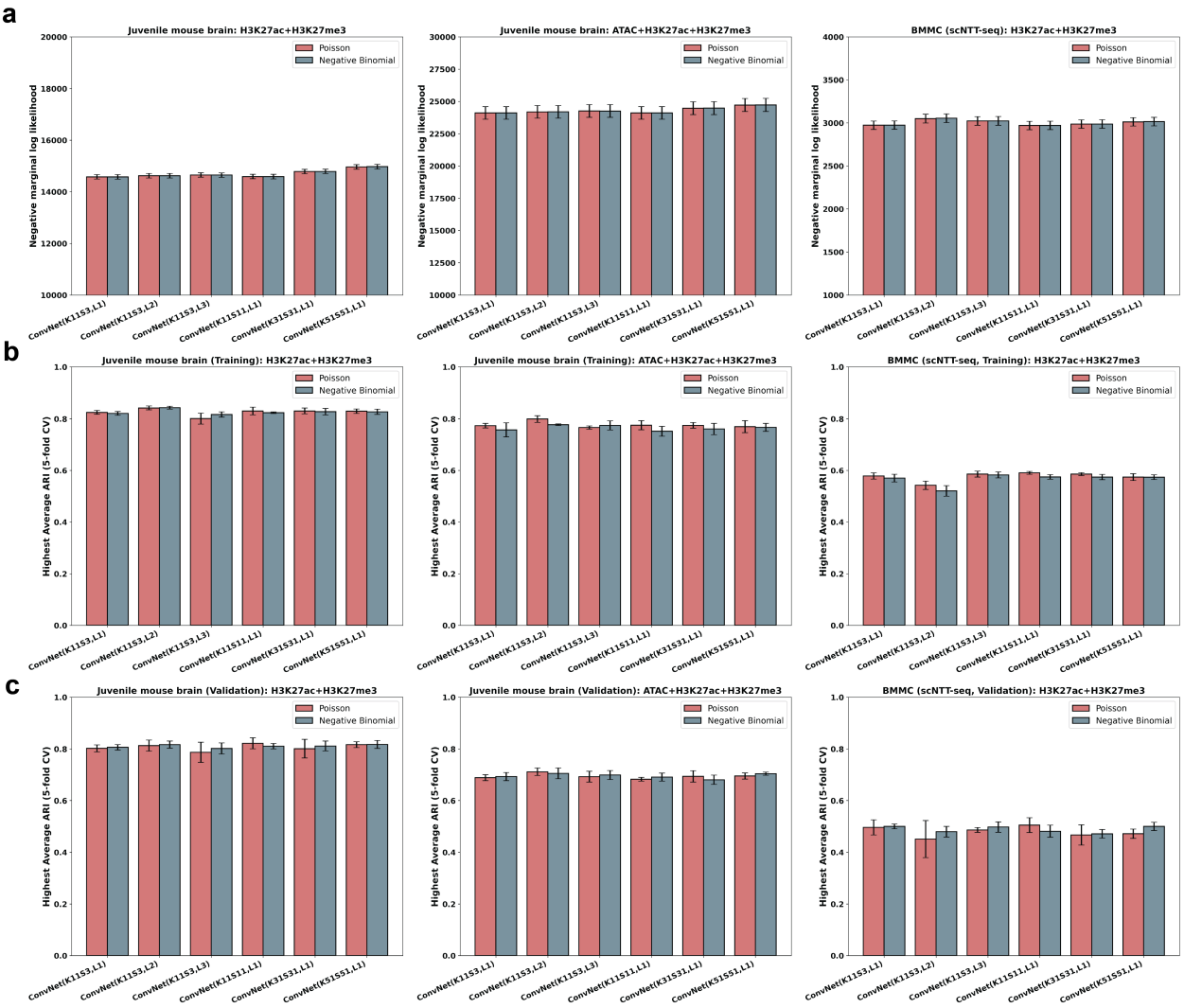


Figure 3.4: Models with Poisson and negative binomial distributions lead to comparable performance on studied datasets. Bimodal junvile mouse brain (Left column), trimodal junvile mouse brain (Middle column), BMMCs (Right column). (**a**) Comparison of the marginal log likelihood (validation set) from `ConvNet-VAE`s under Poisson and negative binomial distributional assumption. (**b**) The highest average ARI that each model can achieve on the training sets over a range of clustering resolution. (**c**) The highest average ARI that each model can achieve on the validation sets over a range of clustering resolution. All error bars indicate the standard deviation from 5-fold cross-validation.

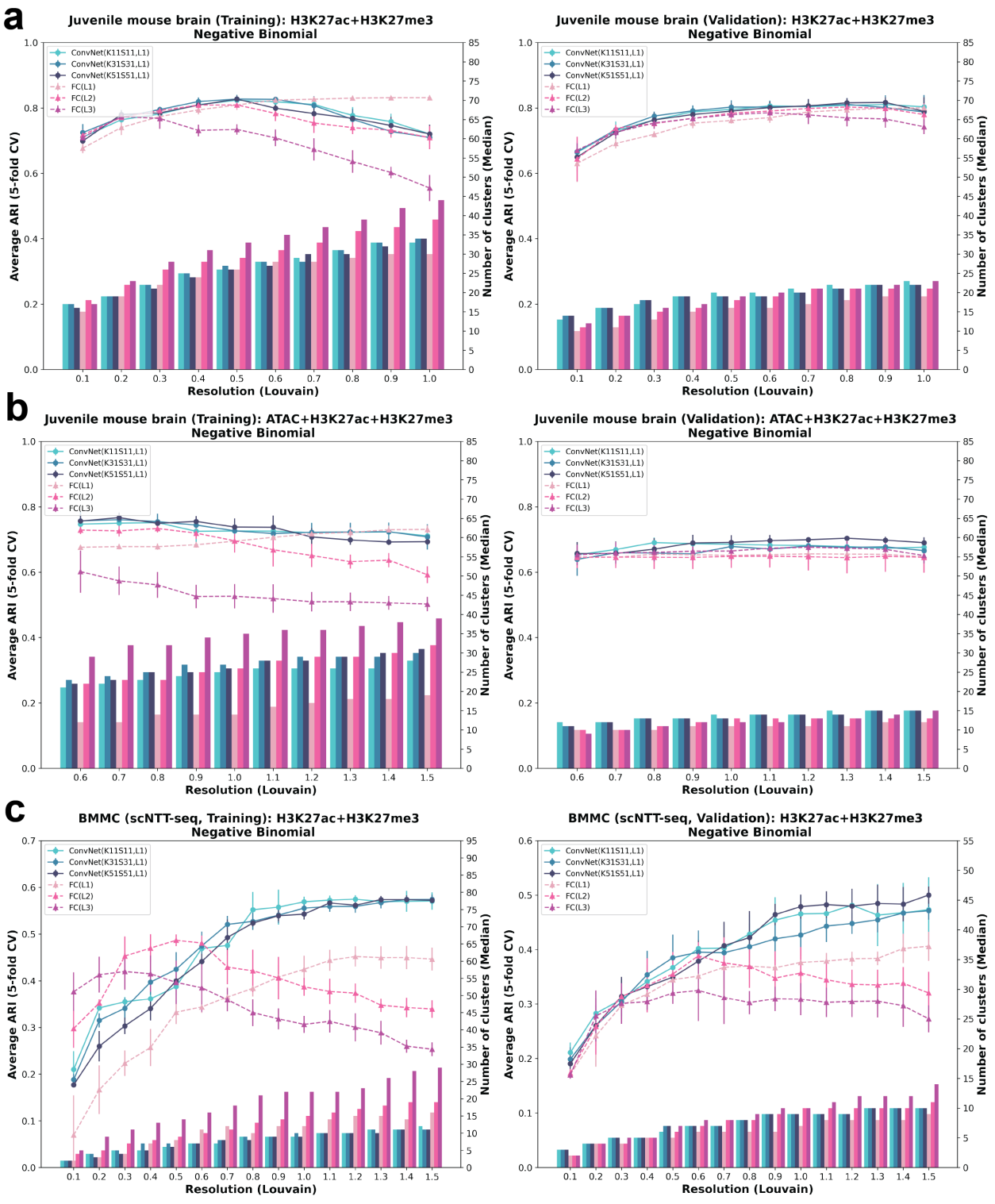


Figure 3.5: Evaluation of 1-Conv1D-layer `ConvNet-VAE`s with negative binomial distribution: ARI. Comparison between `ConvNet-VAE`s (Group 1) using negative binomial modeling and `FC-VAE`s on the quality of cell embeddings, evaluated by ARI. (**a**) Bimodal junvile mouse brain. (**b**) Trimodal junvile mouse brain. (**c**) BMMCs. Error bars indicate the standard deviation from 5-fold cross-validation.

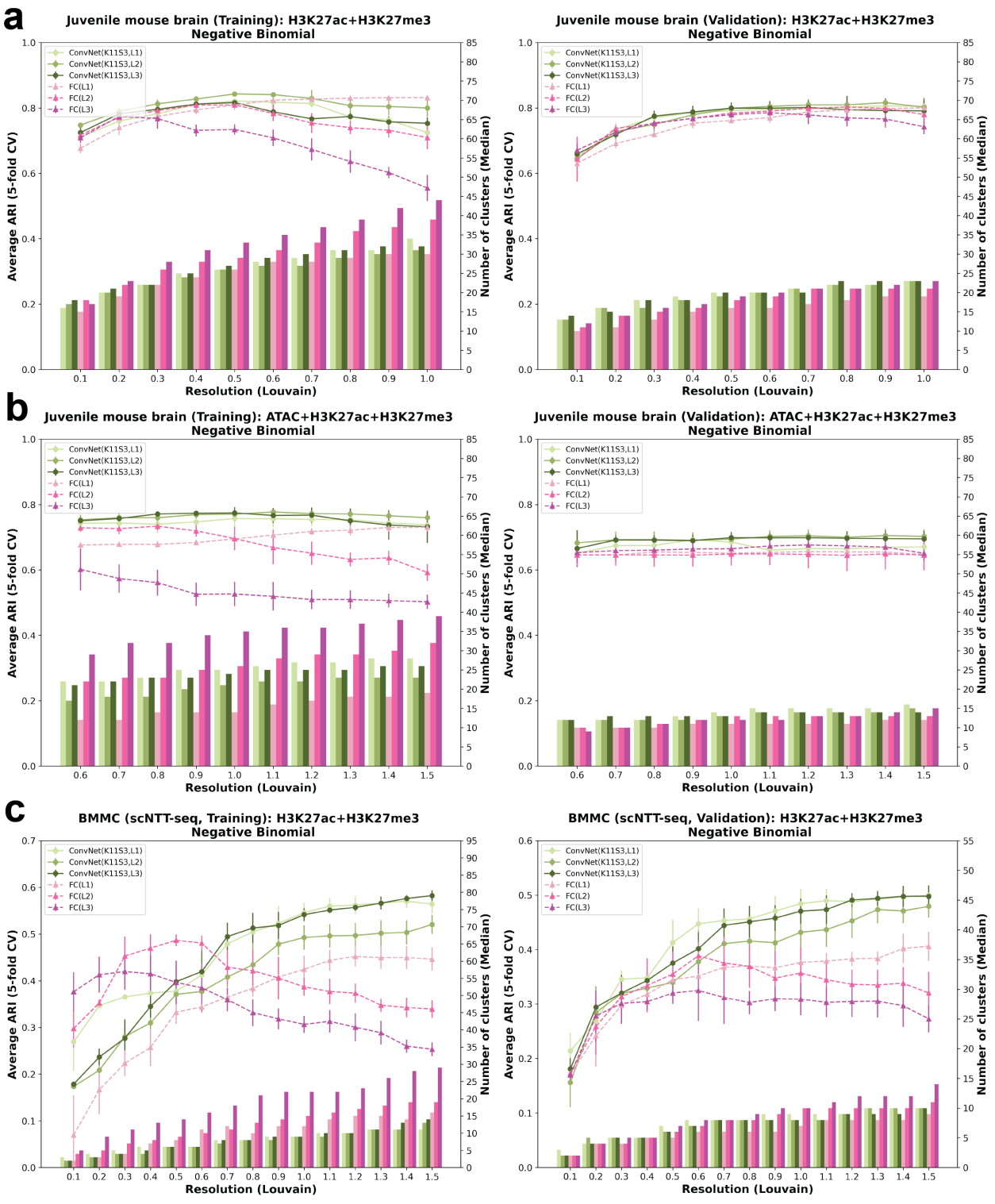


Figure 3.6: Evaluation of multi-Conv1D-layer `ConvNet-VAE`s with negative binomial distribution: ARI. Comparison between `ConvNet-VAE`s (Group 2) using negative binomial modeling and `FC-VAE`s on the quality of cell embeddings, evaluated by ARI. (**a**) Bimodal junvile mouse brain. (**b**) Trimodal junvile mouse brain. (**c**) BMMCs. Error bars indicate the standard deviation from 5-fold cross-validation.

			ConvNet-VAE Model Architecture								
Grp.	Kernel(K)	Stride(S)	Encoder Layers					Decoder Layers			
1	11	11	Conv1D(32)			FC	FC(μ,σ)	FC			Conv1D
	31	31	Conv1D(32)			FC	FC(μ,σ)	FC			Conv1D
	51	51	Conv1D(32)			FC	FC(μ,σ)	FC			Conv1D
2	11	3	Conv1D(32)			FC	FC(μ,σ)	FC			Conv1D
	11	3	Conv1D(32)	Conv1D(64)		FC	FC(μ,σ)	FC		Conv1D(32)	Conv1D
	11	3	Conv1D(32)	Conv1D(64)	Conv1D(128)	FC	FC(μ,σ)	FC	Conv1D(64)	Conv1D(32)	Conv1D

Table 3.1: ConvNet-VAE Model Architecture

Two groups of ConvNet-VAEs are employed in this study. In Group 1, the encoders and decoders of ConvNet-VAEs comprise only one 1D-convolutional layer, with increasing size of the kernel (K) and stride (S) of filters (from 11 to 51). In the second group, the ConvNet-VAEs contain an increasing number of 1D-convolutional layers (from 1 to 3), with fixed kernel size and stride.

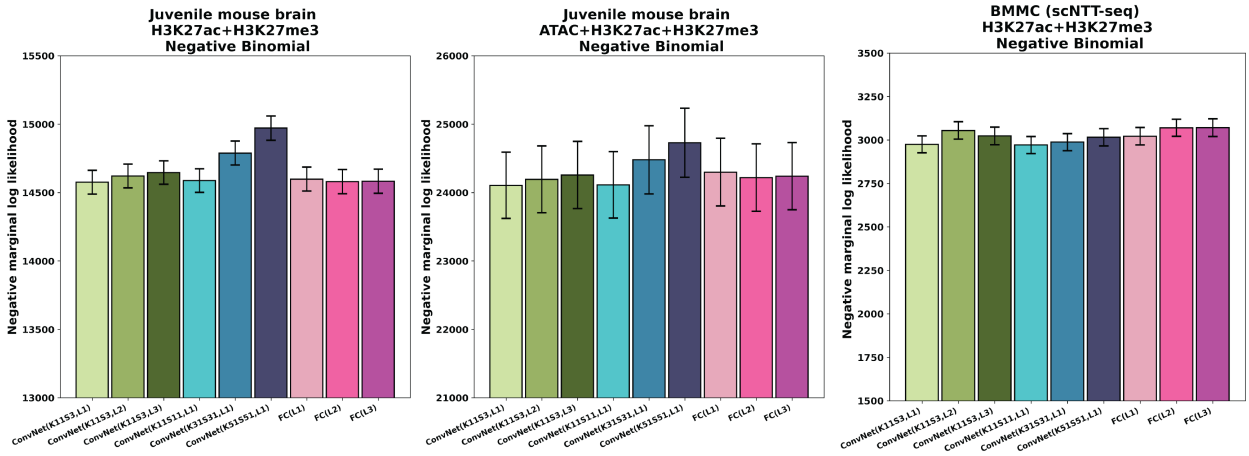


Figure 3.7: Evaluation of `ConvNet-VAE`s with negative binomial: Marginal log likelihood (Validation). Comparison of the marginal log likelihood of the validation set between `ConvNet-VAE`s (negative binomial modeling) and `FC-VAE`s.

$$p(\mathbf{z}|\mathbf{X}, \mathbf{b}) = p(\mathbf{z}|\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M, \mathbf{b}) \propto \frac{\prod_{m=1}^{M} p(\mathbf{z}|\mathbf{x}_m, \mathbf{b})}{\prod_{m=1}^{M-1} p(\mathbf{z})} \tag{3.3}$$

We further approximate the true single-modality posterior $p(\mathbf{z}|\mathbf{x}_m, \mathbf{b})$ using $q(\mathbf{z}|\mathbf{x}_m, \mathbf{b})$ (a Gaussian "expert") learned from modality-specific neural networks (parameterized by ϕ_m).

$$q(\mathbf{z}|\mathbf{x}_m, \mathbf{b}) \equiv q_{\phi_m}(\mathbf{z}|\mathbf{x}_m, \mathbf{b})p(\mathbf{z}) \tag{3.4}$$

The product of Gaussian experts is still Gaussian distributed (Cao and Fleet 2014). Assuming that the m-th expert outputs μ_m and V_m and setting $T_m \equiv V_m^{-1}$, we can define the product Gaussian of $\mathbf{z}$ with the following parameters:

$$\mu_{PoE} = \frac{\sum_m \mu_m T_m}{(\sum_m T_m)^{-1}}$$
$$\Sigma_{PoE} = (\sum_m T_m)^{-1}$$

We configured `FC-VAE`s under three different settings with their architectures detailed in Table 3.2. Each expert model is comprised of two fully connected (FC) layers in the encoder and an additional two FC layers in the decoder, similar to that employed in `scVI` (Lopez et al. 2018). An example model architecture is shown in Figure 3.8. Like `ConvNet-VAE`s, `FC-VAE`s apply Batch Normalization (1D), ReLU activation, and Dropout layers in the FC layers, except for FC(μ,σ)

FC−VAE Model Architecture									
	Layers*	Encoder Layers				Decoder Layers			
For each expert	1	FC			FC(μ,σ)			FC	FC
	2	FC	FC		FC(μ,σ)		FC	FC	FC
	3	FC	FC	FC	FC(μ,σ)	FC	FC	FC	FC

* Number of encoding layers (excluding FC(μ,σ))

Table 3.2: FC−VAE Model Architecture

and the final FC decoding layer. The models are trained to optimize the ELBO defined as 3.5. In addition to the comparable architectures of FC−VAEs and ConvNet−VAEs, we use exactly the same dimension (D) for the latent space, dropout rate, training/validation data splits, training scheme, and parameters for clustering, to ensure fair comparison (detailed in 3.5.7).

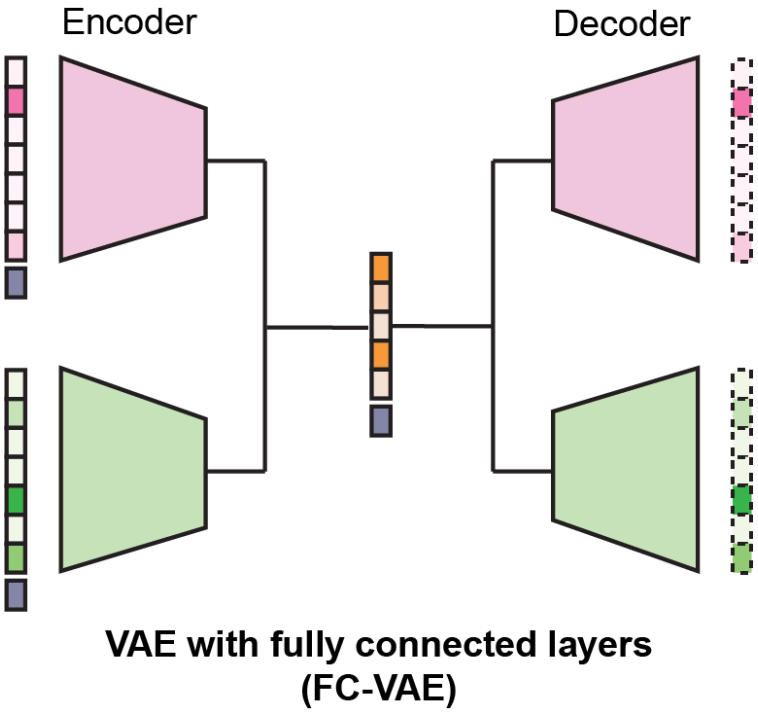


Figure 3.8: Architecture of FC−VAE (bimodal). An brief illustration of the architecturea of a bimodal FC−VAE based on Product of Experts (PoE). Each expert corresponds to a modality. It easily extends to additional modalities by adding encoder-decoder pairs.

$$\text{ELBO}(\mathbf{x}_1,\ldots,\mathbf{x}_M) \triangleq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}_1,\ldots,\mathbf{x}_M,\mathbf{b})}[\sum_{\mathbf{x}_m\in\mathbf{X}} \log p_\theta(\mathbf{x}_m|\mathbf{z},\mathbf{b})] - D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}_1,\ldots,\mathbf{x}_M,\mathbf{b}) \parallel p(\mathbf{z})]) \tag{3.5}$$

3.5.3 Evaluation on batch-effect correction

To evaluate model efficacy in data integration, we utilized four distinct metrics: ASW (Batch), graph connectivity, graph iLISI, and kBET. These metrics, introduced by Luecken et al., are tailored to assess batch-effect removal (Luecken et al. 2022).

Average silhouette width (ASW) quantifies the separation of clusters. ASW (Batch) measures batch mixing, ranging from 0 to 1, where 1 indicates perfect mixing. Graph connectivity investigates how well the cells with same identity are connected in the k-nearest neighbor (kNN) graph built from integrated data. A graph connectivity score of 1 implies good integration, where all cells with same label are connected in the kNN graph. Korsunsky et al. employed integration Local Inverse Simpson's Index (iLISI) to measure the batch distribution using local neighbors chosen on a pre-defined perplexity (Korsunsky et al. 2019a). As an extension, Graph iLISI is able to take graph-based integration outputs and higher score represents better data integration. k-nearest neighbor batch-effect test, known as kBET, starts by constructing a kNN graph, and then examines the batch label distribution in the cell's neighbourhood against the global batch label distribution through random sampling (Büttner et al. 2019). The detailed descriptions of these metrics are available in their original publications.

In this benchmark analysis, we trained VAE models using the entire dataset and 5 different random initializations. We computed these metrics with default settings using the resulting cell embeddings. All metrics reported in this study are average scores across 5 runs.

3.5.4 Evaluation of VAEs' ability to capture data distribution

To benchmark Bayesian probabilistic models in a uni-modal setting ($\mathbf{x}_1$), a popular strategy is to compare the marginal likelihood. A VAE model that is better at capturing the data distribution and generating samples is expected achieve a higher marginal log-likelihood $\log p(\mathbf{x}_1)$ on the test set. Similarly, here we used joint conditional log-likelihood $\log p(\mathbf{x}_1, \mathbf{x}_2)$ and $\log p(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$ as the evaluation metrics in the multi-modal settings, to compare the quality of the tested deep generative models. These marginal log-likelihoods (marginal with respect to latent variable $\mathbf{z}$) can be approximated through importance sampling (Owen and Zhou 2000, Wu and Goodman 2018). Assuming test data $\mathbf{x}_1$, $\mathbf{x}_2$, as well as the latent representation $\mathbf{z}$ from a given sample i in the bimodal setting, hence we have

$$\log p(\mathbf{x}_1, \mathbf{x}_2|\mathbf{b}) \approx \log \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}_1,\mathbf{x}_2,\mathbf{b})}\left[\frac{p_\theta(\mathbf{x}_1, \mathbf{x}_2|\mathbf{z}, \mathbf{b})p(\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x}_1, \mathbf{x}_2, \mathbf{b})}\right] \tag{3.6}$$

The RHS of 3.6 can be estimated by Eq. (3.7):

$$\log \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}_1,\mathbf{x}_2,\mathbf{b})}\left[\frac{p_\theta(\mathbf{x}_1,\mathbf{x}_2|\mathbf{z})p(\mathbf{z},\mathbf{b})}{q_\phi(\mathbf{z}|\mathbf{x}_1,\mathbf{x}_2,\mathbf{b})}\right] \approx \log \frac{1}{N_S}\sum_{s=1}^{N_s}\frac{p_\theta(\mathbf{x}_1|\mathbf{z}_s,\mathbf{b})p_\theta(\mathbf{x}_2|\mathbf{z}_s,\mathbf{b})p_{\mathcal{N}(0,1)}(\mathbf{z}_s)}{q_{\mathcal{N}(\mu(\mathbf{x}_1,\mathbf{x}_2,\mathbf{b}),\sigma(\mathbf{x}_1,\mathbf{x}_2,\mathbf{b}))}(\mathbf{z}_s)}, \quad (3.7)$$

where the samples z_s are randomly drawn from the importance distribution $\mathcal{N}(\mu(\mathbf{x}_1,\mathbf{x}_2,\mathbf{b}),$ $\sigma(\mathbf{x}_1,\mathbf{x}_2,\mathbf{b}))$ defined by the output from the inference networks. N_s is the number of importance samples. $p_\theta(x_1|z_s)$ and $p_\theta(x_2|z_s)$ are calculated with data distribution obtained by the decoders. We estimated the mean joint log-likelihood of the validation set of 100 importance samples ($N_s = 100$) on all datasets and reported the average values over 5-fold cross-validation.

3.5.5 Evaluation of the cell representations learned by VAEs

We applied the Louvain community detection algorithm (Waltman and Van Eck 2013) to the low-dimensional representations of cells generated by the models on the training and validation sets. Resolution is a parameter that influences the number of identified clusters–a higher value yields more clusters. By running the Louvain clustering over a range of resolution values, we then compare the resulting clusters against the published cell type annotations using the Adjusted Rand Index (ARI) (Hubert and Arabie 1985) as a measure of how well the learned representations capture the underlying structure of the data.

3.5.6 Data pre-processing

3.5.6.1 Juvenile mouse brain

Bartosovic et al. recently developed nano-CUT&Tag technology, enabling multimodal chromatin profiling at single-cell resolution (Bartosovic and Castelo-Branco 2022). The authors succeeded in measuring up to three modalities, ATAC, H3K27ac and H3K27me3, simultaneously within individual cells from the mouse brains (19-day old). Starting with the fragment data of each modality, we used `Signac` to segmented the genome into windows, resulting in a count matrix (fragment count in each genomic bin) with the dimension of cell by bin (Stuart et al. 2021). Fang et al. showed that bin size ranging from 1kb to 10kb performed similarly in their benchmark studies(Fang et al. 2021). Therefore, we set the bin width to be 10kb to reduce the input dimension for this analysis. We excluded the bins that overlap with the regions in the ENCODE mouse genome (mm10) blacklist(Amemiya et al. 2019). We retained the cells with authors' annotation for the analysis. After filtering, H3K27ac and H3k27me3 were measured in total $N = 11,981$ cells (4 biological replicates: $N_1 = 2,117$, $N_2 = 2,479$, $N_3 = 2,392$, $N_4 = 4,993$), and $4,434$ of them (2 biological replicates: $N_1 = 2,084$, $N_3 = 2,350$) have additional ATAC measurements. We further selected the $25,000$ bins with the largest counts jointly from all of the modalities (i.e. the union of

bins of highest fragment count of each modality) to reduce sparsity. For bimodality data, we used 28 cell type labels generated on the H3K27ac (similar to labels generated on H3K27me3) for model performance evaluation. For the dataset encompassing three modalities, we utilized 26 cell classes from WNN analysis conducted by the authors. For convolutional neural networks, we constructed 3-dimensional input arrays (cell $\times$ modality $\times$ bin).

3.5.6.2 Human bone marrow mononuclear cells (BMMCs)

Stuart et al. collected bone marrow mononuclear cells from healthy human donors, and jointly profiled H3K27ac and H3K27me3 using single-cell Nanobody-tethered transposition followed by sequencing (scNTT-seq) technology(Stuart et al. 2022). We downloaded the processed R object from Zenodo (https://zenodo.org/record/7102159), which contains $N = 5,236$ cells with top $71,253$ bins from H3K27ac modality and top $43,170$ bins from H3K27me3 (bin size $= 1$ kb) used for the original analysis, where 15 different cell types were identified through WNN workflow on aggregated bin data by the authors. For our analysis, we obtained the genome bin features from fragment files and selected top $25,000$ bins (bin size $= 10$ kb) using the same strategy described above.

3.5.6.3 Human peripheral blood mononuclear cells (PBMCs)

The PBMC sample was obtained from a healthy female donor ($N = 11,909$ before quality control). The dataset was generated by 10x Genomics using single Cell Multiome ATAC + Gene Expression (publicly available on 10x Genomics website). For each cell, $36,601$ genes and $106,056$ peaks were profiled in parallel. We followed the Weighted-Nearest Neighbor (WNN) workflow (Seurat V4) to generate the cell type labels through joint analysis of the transcriptomics (RNA-seq) and chromatin accessibility (ATAC-seq) profiles. We kept the cells ($N = 11,402$) that meet the specified criteria for quality control (number of ATAC-seq counts $\in [5,000, 70,000]$; number of RNA-seq counts $\in [1,000, 25,000]$, $> 20\%$ mitochondrial counts). Top $5,000$ genes and top $25,460$ peaks were selected for WNN analysis. As a result, the Louvain algorithm (resolution $= 0.25$) led to 15 clusters, which were further used for method benchmark after cell type annotation. For ATAC peak data, the read (fragment end) count are converted to the fragment count, as suggested by Martens et al. (estimated fragment count $=$ (odd read count $+ 1)/2$) (Martens et al. 2023).

3.5.6.4 Mouse cortex and hippocampus

Yao et al. sequenced approximately 1.3 million cells in the adult mouse cortex and hippocampus regions and obtained their transcriptomic profiles, leading to a thorough assortment of glutamatergic and GABAergic neuron types (Yao et al. 2021). For this study, we used the single-cell transcriptomic

data generated by 10x Genomics Chromium platform (version 2 chemistry). We downloaded the processed data ($N = 1,169,213$) from the Neuroscience Multi-omic (NeMO) Data Archive as part of the BRAIN Initiative Cell Census Network. Out of genes measured in total, we selected $5,000$ highly variable genes from the normalized dataset using LIGER pipeline (Welch et al. 2019, Gao et al. 2021, Lu and Welch 2022). For evaluation on the investigated methods, we used 42 cell classes and subclasses annotated by the authors following Tasic et al.'s work (Tasic et al. 2018).

3.5.6.5 Mouse organogenesis

Argelaguet et al. investigated the mouse early organogenesis by simultaneously profiling gene expression and chromatin accessibility in the same nuclei (10x Multiome) from mouse embryos between 7.5 to 8.75 days (E7.5-8.75) of gastrulation(Argelaguet et al. 2022). Specifically, we selected the E7.5, E8, E8.5 and E8.75 embryos ATAC-seq datasets ($N = 68,804$), and preprocessed the fragment files following the ArchR pipeline provided by the authors(Granja et al. 2021). After excluding the cells identified as low-quality or doublets, we obtained the peak count matrix comprising $191,407$ peaks from $N = 41,705$ cells. We further selected the $25,000$ peak features with the highest total number of counts across all cells for analysis. Fragment count was estimated using read count following the same approach described above.

3.5.7 Experiments

We benchmarked the selected models through 5-fold cross-validation over a variety of datasets. For model training, we used a mini-batch size of 128, Adam optimizer (learning rate $= 0.001$). Each fully connected layer has 128 hidden units. The architecture incorporated Batch Normalization and ReLU activation functions in the majority of layers, alongside a dropout rate of 0.2 to prevent overfitting. We performed the Louvain algorithm (k-nearest neighbors: $k = 20$) on the latent cell embeddings for clustering. The algorithm was run with 5 random starts unless stated otherwise. The cluster assignment with the best quality was recorded. For the juvenile mouse brain data, the dimension of the latent space D was set to 30 and all models underwent 300 epochs of training. For the BMMCs data, we used $D = 30$ and 200 training epochs. For single-cell unimodal datasets: $D = 20$ and 200 training epochs were employed for PBMCs data; $D = 30$ and 15 training epochs for mouse cortex and hippocampus data; $D = 30$ and 50 training epochs for mouse organogenesis data.

3.5.8 Model implementation

All reported VAE models were implemented in Pytorch 1.10.1 and Python 3.8, trained with 2.9 GHz Intel Xeon Gold 6226R and NVIDIA A40 GPU.

CHAPTER 4

Integrating Spatially Resolved Multimodal Data Using Variational Graph Autoencoder

Recent advancements in spatial profiling have allowed for the simultaneous investigation of transcriptomics, proteomics, and epigenomics at the individual cell/bead/spot level in tissue. These technologies have been instrumental in revealing the heterogeneous and complex molecular makeup of the cells or tissue microenvironments. Deeper insights into the biological process can be gained by incorporating high-resolution image modalities. For spatially informed multimodal integration, we present `spaMVGAE`, a multimodal variational graph autoencoder that employs graph convolutional networks. It learns a joint embedding of cells/beads/spots by correlating molecular measurements (e.g., gene expression, chromatin accessibility), cell morphology (e.g., Hematoxylin and Eosin (H&E) histology), as well as spatial location information. The resulting low-dimensional embeddings can be used for diverse tasks such as domain detection. By applying `spaMVGAE` on spatially resolved multimodal datasets generated in a variety of biological contexts, we show that `spaMVGAE` can harness different sources of information and learn a refined representation of the observations by taking advantage of the spatial information, in a computationally efficient fashion.

4.1 Introduction

Single-cell multimodal omics technologies have gained their popularity by having the ability to investigate multiple types of molecular information, including transcriptomics, proteomics, and epigenomics, from an individual cell. Examples are 10x single-cell Multiome ATAC + Gene Expression (RNA + ATAC), CITE-seq (RNA + Protein) (Stoeckius et al. 2017), nanobody-based single-cell CUT&Tag (ATAC + H3K27ac + H3K27me3) (Bartosovic and Castelo-Branco 2023). These omics modalities collectively unravel the heterogeneous nature of cells. Lately, the development of spatial transcriptomics is revolutionary, as it provides a comprehensive view of gene expression within the spatial context of the tissues. (Rodriques et al. 2019, Stickels et al. 2021, Baysoy et al. 2023). Moreover, the researchers have the option to apply techniques such as Hematoxylin & Eosin (H&E) staining on the sequenced tissue slice or the adjacent ones (Janesick et al. 2023). Such morphologi-

cal information of the cells/beads/spots provides an additional aspect of the cell types and tissue types, enhancing our understanding of cellular function and tissue architecture. Most recently, the technologies that profile multiple molecular modalities in the same bead or spot have emerged. For instance, Zhang et al. (Zhang et al. 2023b) developed the spatial assay for transposase-accessible chromatin and RNA using sequencing (spatial ATAC-RNA-seq), which measures chromatin accessibility and messenger RNA expression in parallel in up to $10,000$ barcoded pixels. Russell et al. (Russell et al. 2023) introduced the Slide-tags approach that can label nuclei with spatial barcodes, and applied it to multi-omics sequencing (Slide-tags multiome).

While these spatially resolved multimodal datasets offer tremendous opportunity in making new biological discoveries, there are several challenges that we need to take into consideration when devising a tailored computational method for integrative analysis. The method needs to (1) be able to jointly embed 2, 3 or more modalities into the shared latent space for downstream tasks, given that the sequencing technology is rapidly evolving; (2) be scalable to high-throughput dataset, as the state-of-the-art technologies like Xenium can produce datasets comprising hundreds of thousands of observations (Janesick et al. 2023); (3) be capable of integrating multiple tissue sections; (4) properly incorporate the spatial location information. A few computational tools have been developed to integrate this type of data, however they fail to meet all the requirements. Hu et al. proposed a graph convolutional network (GCN)-based method, `spaGCN` (Hu et al. 2021), for spatial domain detection. It utilizes gene expression, location, and histology to detect clusters in an iterative fashion. However, `spaGCN` extracts imaging features by deriving a weighted sum of the RGB values. It doesn't generate a joint embedding of different modalities. Additionally, this method suffers from scalability issues. Bao et al. developed a multimodal autoencoder, `MUSE`, to integrate both transcriptomic profile and cell morphologies for spatial clustering (Bao et al. 2022). Yet, `MUSE` doesn't explicitly incorporate the spatial location information. It requires initial cell labels for model training. `MUSE` also has difficulty dealing with large-scale datasets and joint analysis of multiple tissue sections. Recently published `CellCharter` (Varrone et al. 2023) relies on `scVI` to perform dimensionality reduction on the transcriptomic and epigenomic profiles separately before feature aggregation among neighboring cells for spatial clustering. `SpatialGlue` (Long et al. 2023), a VAE framework, utilizes GCN layers for incorporating spatial information and the attention mechanism for multimodal fuision. However, it requires the dimensionality of input to be reduced through principal component analysis (PCA). Its adaptability to multi-section data has not been validated.

To address the computational needs for this multimodal integration problem, we developed `spaMVGAE`, a scalable variational graph autoencoder that can infer joint latent space of the cells/beads/spots from multiple spatial data modalities (e.g., RNA, ATAC, cell morphology) multiple tissue sections. `spaMVGAE` models the molecular measurements in an end-to-end manner without

additional dimension reduction procedure on the input. It takes advantage of spatial information through the implementation of GCN. `spaMVGAE` shows its applicability and utility on a variety of spatial datasets, including slide-seq, spatial ATAC-RNA-seq, 10x Genomics Xenium.

4.2 Multimodal variational graph autoencoder for spatially resolved multimodal data

The recent technological advancement has allowed for simultaneous measurement of multiple modalities, including transcriptome, epigenome, with spatial context from individual cells, beads, or spots, based on the platform. In addition to the multi-omics data, morphological information can also be acquired by methods such as H&E staining (Figure 4.1a). To effectively integrate these data modalities for identification of meaningful spatial domains, we present `spaMVGAE`, a scalable Bayesian variational inference framework that is able to jointly model high-throughput spatially resolved multimodal data.

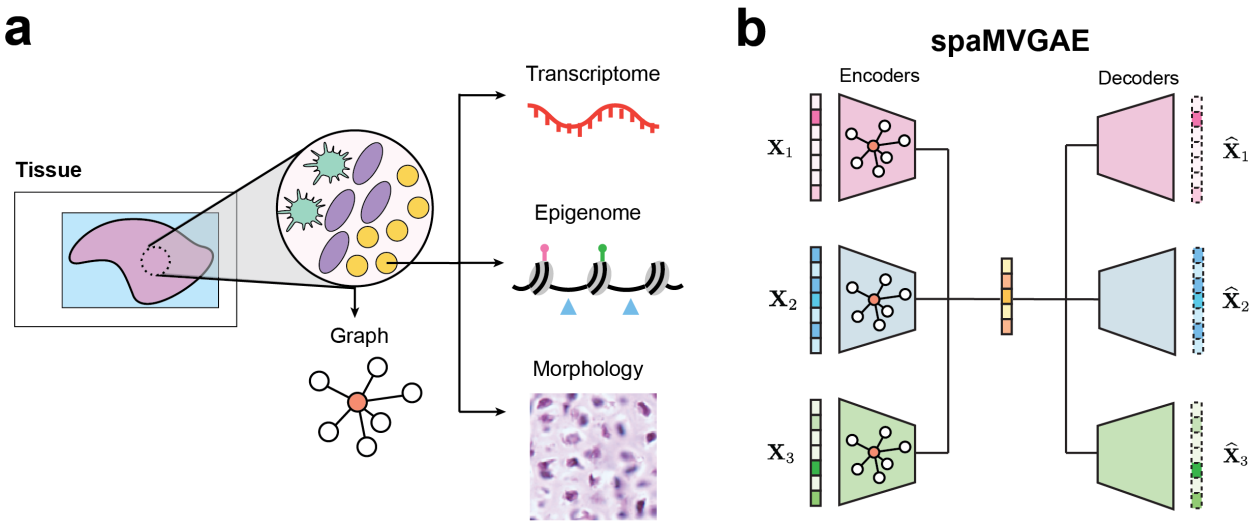


Figure 4.1: Overview of `spaMVGAE`. (**a**) In spatially resolved data, multiple modalities can be acquired for each cell/bead/spot (e.g., gene expression, chromatin accessibility, H&E histology). A K-nearest neighbor graph is built based on the 2-dimensional coordinates. (**b**) An illustration of `spaMVGAE` architecture. It consists of modality-specific encoder-decoder pairs. The spatial graph is shared across all graph convolutional encoders. The joint latent space is inferred through the Product of Experts (PoE).

The core of the `spaMVGAE` is a multimodal variational graph autoencoder (Figure 4.1b). The joint distribution of multi-omic profile and morphological information is captured through the implementation of the concept of product of experts (PoE) (Wu and Goodman 2018). `spaMVGAE` correlates heterogeneous modalities and provides a unified view of the data. The joint latent space enables tasks such as spatial domain detection and trajectory inference. `spaMVGAE` also offers the flexibility to accommodate the growing variety of modalities produced by spatial profiling.

For the model input, we select the top highly variable features from molecular measurements, and then generate cell/bead/spot by feature count matrices. As for the images, we obtain the feature vectors from pre-trained image classification models (Bao et al. 2022). In `spaMVGAE`, each encoder (inference) network is tailored to a specific modality. The encoding process is accomplished by the graph convolutional networks (GCN) (Kipf and Welling 2016a), which is able to capture the similarities between the observations in the local regions. To incorporate the spatial location information, we build a KNN graph for each cell/bead/spot and other cells/beads/spots in close proximity defined by the Euclidean distance. The resulting adjacency matrix, together with the feature matrices, serves as input to the GCN layers. Our results show that the aggregation of information from the cell/bead/spot neighborhood is crucial for learning their low-dimensional representation. Moreover, `spaMVGAE` can easily handle the multimodal datasets from multiple tissue sections by concatenating the adjacency matrices and feature matrices (See Section 4.5).

In the subsequent sections, we demonstrate the performance of `spaMVGAE` on multimodal integration using a variety of real datasets, in comparison to the selected baseline models.

4.3 Results

4.3.1 `spaMVGAE` enables domain detection on HER2 breast cancer data

To demonstrate the power of `spaMVGAE` in spatially informed multimodal integration, we first applied it for domain detection in the HER2 (human epidermal growth factor receptor 2)-positive breast tumor tissue section (Figure 4.2a) (Andersson et al. 2021). The data were generated using the Spatial Transcriptomics (ST) technology (Ståhl et al. 2016). By taking the gene expression data, image features extracted from cropped H&E image tiles, and the spatial graph, `spaMVGAE` was able to reduce the dimensionality of the input data. We performed Leiden clustering (Traag et al. 2019) on these spot embeddings for clustering so that domains of interest can be detected. By comparing against the manual annotation from an experienced pathologist (Figure 4.2a,b), we can tell that `spaMVGAE` indeed captures the structure of the tumor tissue. It successfully identifies the cancer in situ, invasive cancer, adipose tissue, and regions with immune infiltration (Figure 4.2c).

Next, we benchmarked `spaMVGAE` against a VGAE that only uses gene expression and the published `MUSE` (multi-modal structured embedding) model. All methods use the same set of highly variable genes (top $2,000$). For fair comparison, `spaMVGAE` and `MUSE` also take as input the same image feature matrix generated by a pre-trained convolutional neural network model. We ran each method with 20 different random model initializations, and obtained the Leiden cluster assignments from the resulting spot embeddings over a range of clustering resolutions. For each run, we only record the highest adjusted Rand index (ARI) (calculated against the published labels) over all the tested resolutions, and report the average highest ARI across these 20 runs. As shown in Figure 4.2d,

spaMVGAE takes advantage of both transcriptomic and morphological information for accurate spatial dimension reduction, outperforming the VGAE ($p < 0.0004$). This observation highlights that the additional image modality indeed improves the model's capability in domain detection. Notably, spaMVGAE beats MUSE by a large margin in terms of average ARI ($p <$ 2E-21).

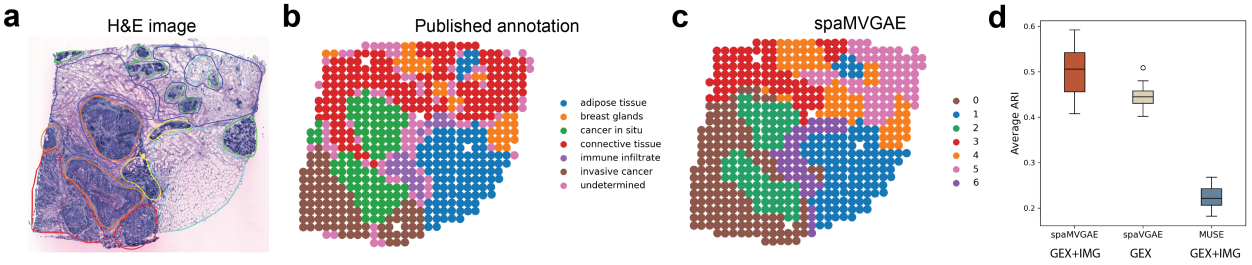


Figure 4.2: spaMVGAE achieves accurate domain detection. **(a)** H&E staining image of HER2-positive breast tumor tissue. **(b)** Spots are plotted using 2-dimensional coordinates and colored based on the manual annotation. **(c)** Spots are colored by Leiden cluster assignments from spaMVGAE output. **(d)** Benchmark between spaMVGAE, spaVGAE, and MUSE. The boxplot shows ARI averaged over 20 runs with different random model initializations. Note that the spots with undetermined identity are excluded from ARI calculation.

4.3.2 spaMVGAE identifies multi-layer growth plate structure in a knockout mouse

Understanding the process of bone development is crucial for unraveling mechanisms of bone diseases. Growth plates are the multi-layered cellular template near the ends of the long bones, which can be divided into three zones: resting, proliferating, and hypertrophic chondrocytes. One of the key signaling pathways involves Indian hedgehog (Ihh)/parathyroid hormone-related protein (PTHrP) negative-feedback loop (Kronenberg 2003). PTHrP promotes proliferation of chondrocytes. In order to further elucidate the role of PTHrP in bone development, Dr. Orikasa and Dr. Ono from the University of Texas Health Science Center utilized PTHrP-mCherry knock-in reporter mice. In such mice, the *Pthrp* allele is modified to express a red fluorescent protein in place of a functional PTHrP protein. They extracted fresh frozen sections of leg tissue from a $Pthrp^{mCherry/mCherry}$ (PTHrP-KO) mouse at embryonic day (E) 18.5, and obtained spatial transcriptomic profiles using CurioSeeker (v1.0) (boxed region) (Figure 4.3a).

In addition to the expression data of the top $2,000$ variable genes, we also registered the H&E image of the neighboring tissue section to the sequenced beads and cropped a 100×100-pixel image centered around each bead (Figure 4.4b). 2048-dimensional feature vectors were generated from the resulting image tiles. We applied spaMVGAE to a total of $50,465$ beads that passed quality control, which effectively learned the low-dimensional embeddings of the beads jointly from transcriptome and morphology. The Leiden algorithm (Traag et al. 2019) led to 18 clusters using the bead embedding. In particular, three of them are identified in the growth plate (enclosed

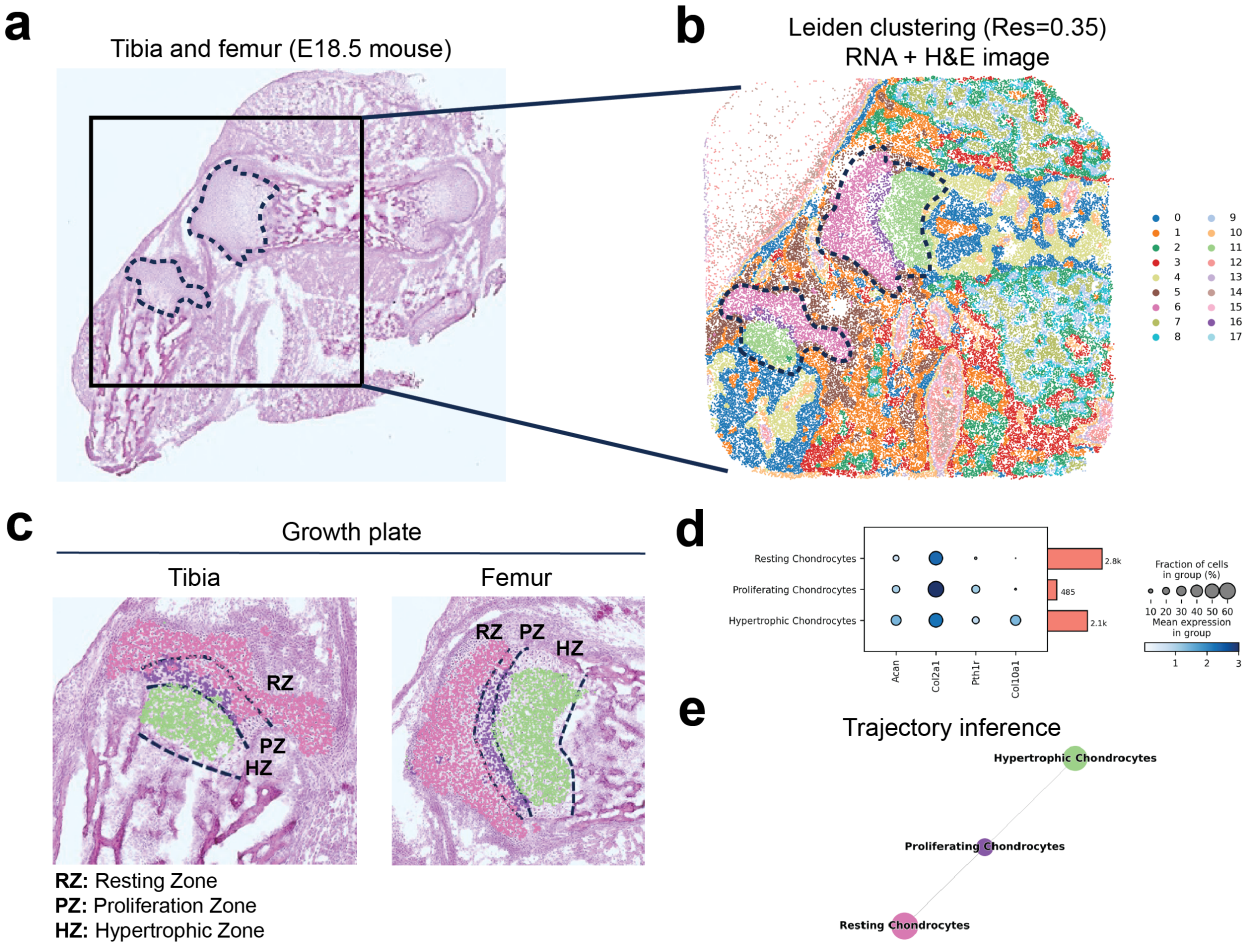


Figure 4.3: `spaMVGAE` characterizes multi-layer structure in growth plate of mouse bone. (**a**) H&E histology of mouse leg tissue from the adjacent tissue section. The spatial transcriptomic profiling was applied to the the boxed region. The dashed line highlights the growth plates in tibia and femur. (**b**) `spaMVGAE` embedded the beads using gene expression, morphology, and the spatial information. The beads are plotted with original spatial coordinates, colored by cluster assignment through the Leiden algorithm. (**c**) The clusters in growth plates are superimposed on the corresponding regions in H&E images. The growth plate can be divided into three zones, resting zone (RZ), proliferation zone (PZ), and hypertrophic zone (HZ). The dashed lines represent manual annotations based on histology. (**d**) The dot plot shows the expression level of the selected marker genes. The size of the dot represents the fraction of cells expressing a specific gene. Darker color indicates higher mean expression in the cluster. (**e**) Trajectory analysis of three clusters of interest.

by dashed lines), the region of interest (Figure 4.3b), corresponding to resting chondrocytes (RC), proliferating chondrocytes (PC), and hypertrophic chondrocytes (HC) respectively. In contrast, a similar model without incorporating the location information of the beads failed such task (Figure 4.4c). We noticed the unimodal version of `spaMVGAE` (H&E images only) was able to achieve similar results, with inferior stratification of the proliferation zone (Figure 4.4d).

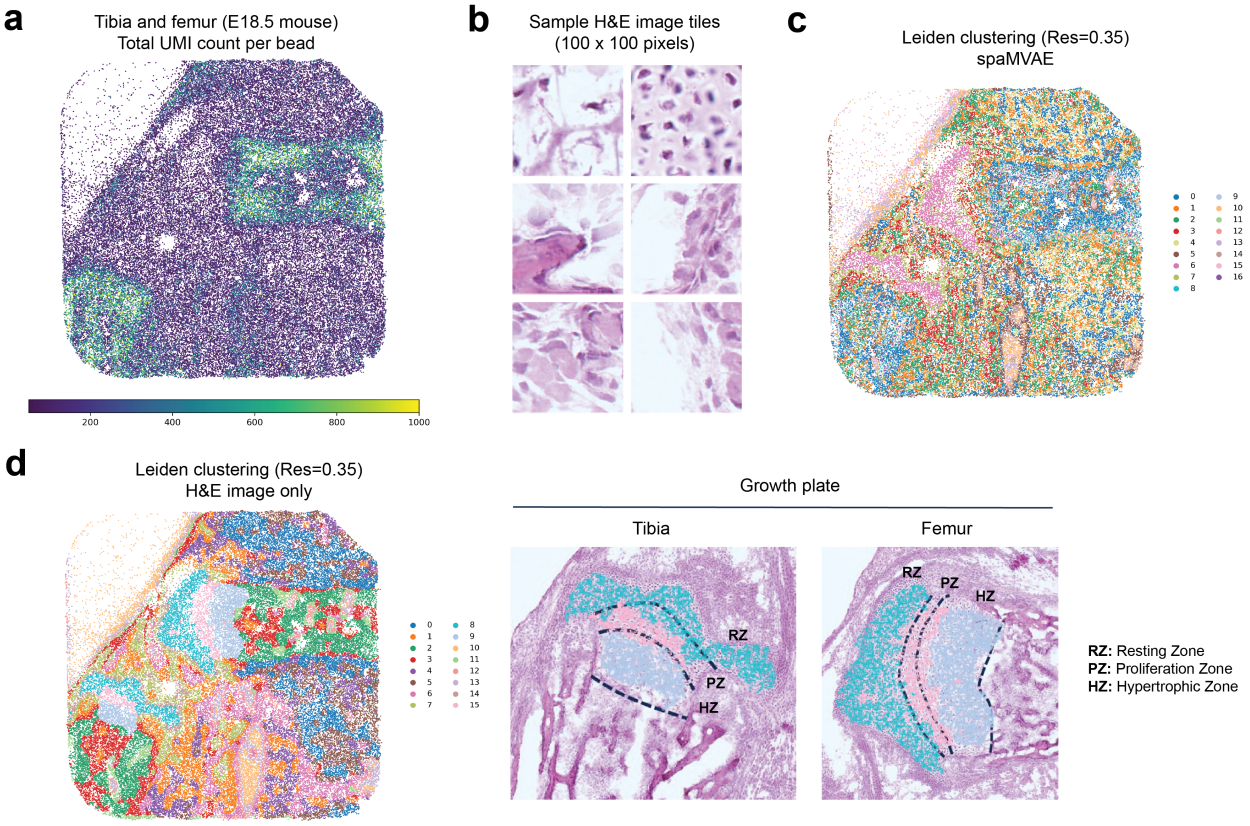


Figure 4.4: Analyses of PTHrP-KO mouse data (slide-seq). (**a**) Spatial plot showing the total UMI count for each bead. (**b**) Example H&E staining image tiles used for feature extraction. (**c**) Leiden clustering result from multimodal VAE `spaMVAE` that doesn't incorporate neighborhood graph. (**d**) Leiden clustering result from VGAE using H&E image information only. The clusters of interest are superimposed on the growth plate regions in the H&E image.

The identified cellular layers in both tibia and femur agree well with the manual annotation (dashed lines) from our collaborators (Figure 4.3c). Notably, the proliferation zone is much smaller due to PTHrP knockout. As expected, the PCs show the highest mean expression of Col2a1 as compared to the other populations. Consistent with previous studies, Col10a1, a specific marker of chondrocytes undergoing hypertrophy, is most abundant in the hypertrophic zone (Yang et al. 2014). Furthermore, we can see the PCs in the knockout mouse express the most Pth1r (Figure 4.3d). This is because the missing of PTHrP causes a breakage of the PTHrP-Ihh negative feedback loop. In this case, PCs try to express more PTH1R to try to compensate for the loss of PTHrP. In

addition, we performed PAGA trajectory inference (Wolf et al. 2019). The properly ordered nodes representing three chondrocyte populations in the growth plate match the underlying biology, further indicating that spaMVGAE embeds the beads properly using transcriptomics, morphology, as well as the spatial location information (Figure 4.3e).

4.3.3 spaMVGAE allows for integration of spatial transcriptome and epigenome of mouse brain

Next, we applied spaMVGAE on the spatial chromatin accessibility (ATAC) and transcriptome (RNA) co-sequencing dataset ($9,215$ pixels) of P22 mouse brain. The profiled region is marked on the Nissl-stained image obtained from the adjacent tissue section, and the corresponding brain structures are shown in the image registered to the Allen Mouse Brain Atlas (Figure 4.5a) for reference. With $3,000$ highly variable genes and $6,000$ highly variable peaks, neither the RNA modality nor the ATAC modality is able to properly identify all the brain structures by itself (Figure 4.5b). For instance, the RNA clusters poorly stratify the layers in the cortex (Cluster 0, 2, and 5), partly due to the fact that the pixels from Cluster 2 have lower quality, with the lowest transcript count among the three (Figure 4.6b). In contrast, spaMVGAE jointly modeled both modalities and achieved pixel embeddings of higher quality, leading to smoother spatial clustering and more accurate identification of spatial domains (Figure 4.5c). We further show that the incorporation of spatial location information is critical, as multimodal VAE without GCN failed to achieve the same performance on learning meaningful representations of these pixels (Figure 4.6c).

The resulting clusters agree well with the anatomical annotations. For example, Cluster 1 and 4 match the striatum. Cluster 6 depicts corpus callosum and anterior commissure. Cluster 11 corresponds to the lateral ventricle. Notably, genes closely related to the major island of Calleja (islm) are highly expressed in Cluster 12, such as Rreb1 (a marker of striatal projection neurons, $p < 2E$-25) and Isl1 (a gene essential for proper formation of D1 and D2 neurons, $p < 6E$-17) (Figure 4.5d) (Stanley et al. 2020). Moreover, the pixels in Cluster 3 stand out with high expression of genes associated with neurons in the cortex region (Figure 4.5e). Specifically, Cux1 ($p < 3E$-8) and Cux2 ($p < 2E$-10) are markers of pyramidal neurons, a major population of nerve cells in the cerebral cortex (Matho et al. 2021). This observation aligns well with the significant expression of Slc30a3 ($p = 0.02$), reported as a marker of excitatory neurons in Layer 2/3 and 4/5 isocortex (Yao et al. 2021). Furthermore, we also identified differentially expressed markers of excitatory neurons (Tle4: $p = 1.7E$-4; Nxph3: $p = 2E$-6) from Layer 6b of isocortex (CTX) in Cluster 10 (Yao et al. 2021), as well as Slc17a7 ($p = 0.02$), a marker of glutamatergic neurons (Yao et al. 2023), in Cluster 0 and 2 covering piriform area (PIR) and anterior cingulate area (ACA), and L6a CTX (Figure 4.5f, Figure 4.6d).

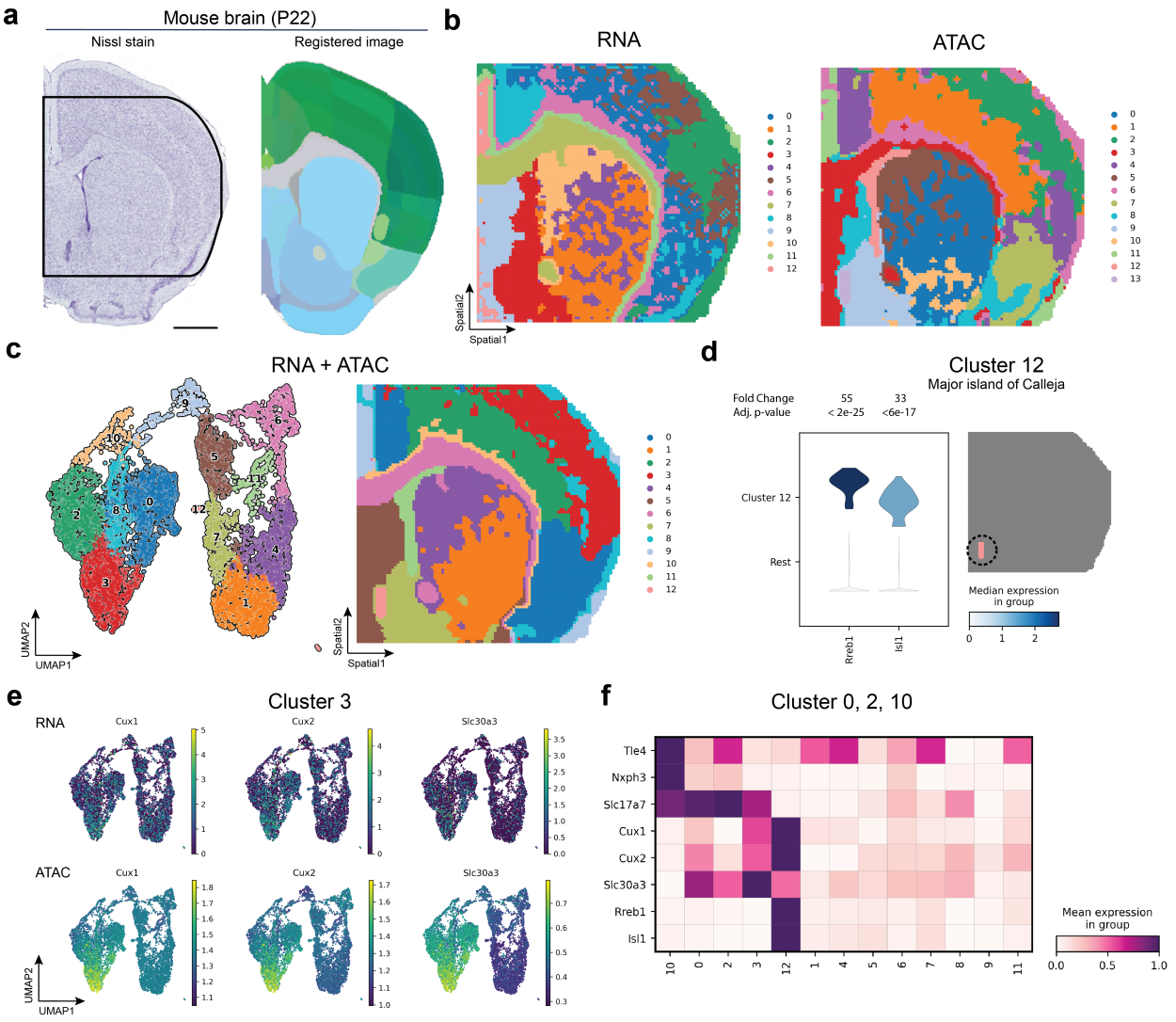
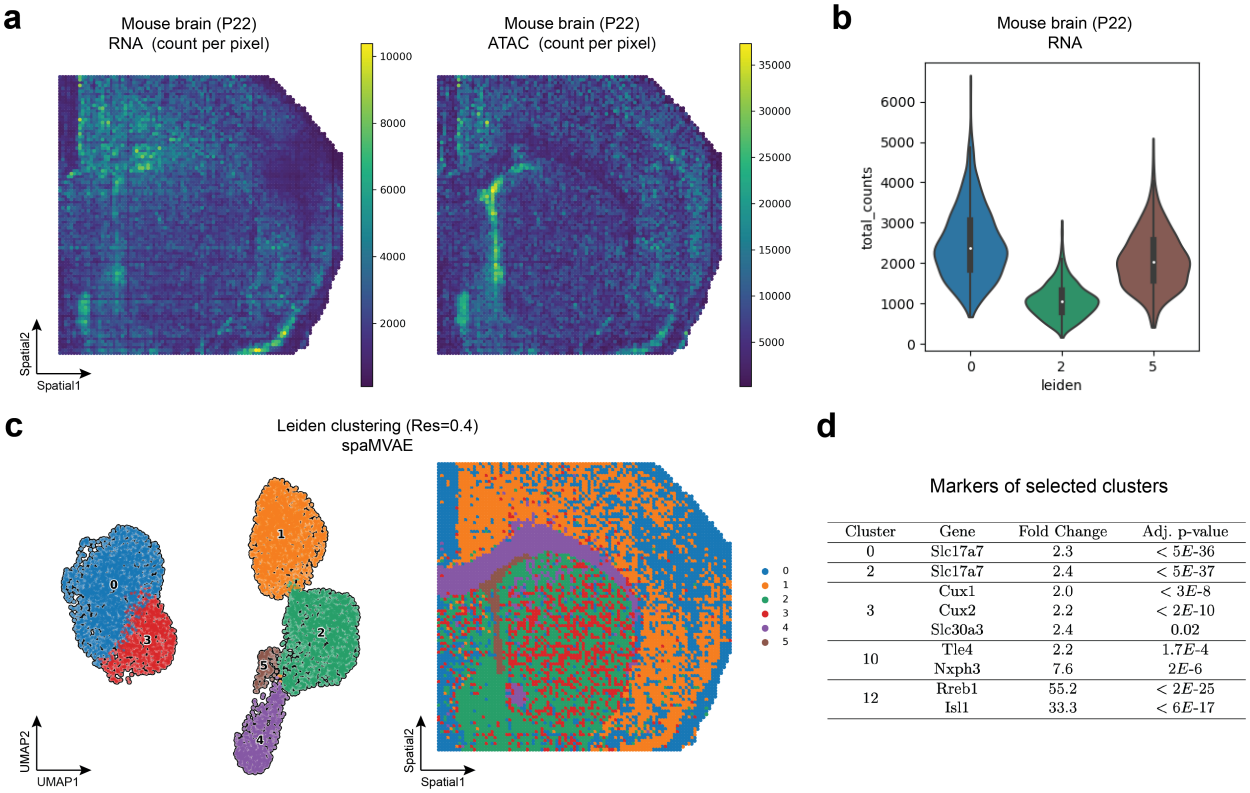


Figure 4.5: `spaMVGAE` identifies the mouse brain structures in coronal section (**a**) (Left) Nissl staining of the adjacent coronal section. (Right) Mapped brain structure by registering the Nissl stain image to the Allen Mouse Brain Atlas. (**b**) Pixels colored by Leiden cluster assignments resulted from VGAE using gene expression and chromatin accessibility respectively. (**c**) Joint modeling through `spaMVGAE`. The clusters are visualized by UMAP (Left) and original spatial coordinates (Right). (**d**) Violin plots of the expression of selected genes in Cluster 12 and the other clusters. (**e**) Marker gene expression of Cluster 3 in UMAP space. (**f**) Heatmap of marker genes associated with Cluster 0, 2, and 10. For each gene, the mean expression is scaled by subtracting the minimum and dividing by its maximum.



Cluster	Gene	Fold Change	Adj. p-value
0	Slc17a7	2.3	$< 5E$-36
2	Slc17a7	2.4	$< 5E$-37
3	Cux1	2.0	$< 3E$-8
	Cux2	2.2	$< 2E$-10
	Slc30a3	2.4	0.02
10	Tle4	2.2	$1.7E$-4
	Nxph3	7.6	$2E$-6
12	Rreb1	55.2	$< 2E$-25
	Isl1	33.3	$< 6E$-17

Figure 4.6: Analyses of mouse (P22) brain data. (**a**) The distribution of UMI count (Left) and fragment count (Right) per pixel in the tissue section. (**b**) Violin plot showing the total UMI counts for Cluster 0, 2, and 5. (**c**) Result from multimodal VAE without using spatial location information. (Left) Leiden clusters visualized in 2-dimensional UMAP space. (Right) Pixels are plotted with their spatial coordinates and colored by cluster identity. (**d**) Table of marker genes associated with clusters of interest. The fold change and adjusted p-value are reported for each gene.

4.3.4 `spaMVGAE` efficiently integrates multi-section human breast cancer data

Finally, we applied `spaMVGAE` to a high resolution spatial transcriptomics dataset generated by 10x Xenium platform. This data consists of gene expression measurement on 313 genes across two tissue sections, with DAPI imaging information available. Breast cancer is a highly heterogeneous and multifaceted disease, characterized by pronounced intratumoral and intertumoral variability in both histological and molecular attributes. In the 10x Xenium data, `spaMVGAE` jointly models both tissue sections, incorporating gene expression measurements and DAPI imaging information as modalities. Regardless of the inclusion of DAPI imaging information, `spaMVGAE` successfully identifies clusters, including ductal carcinoma in situ (DCIS) clusters and invasive ductal carcinoma (IDC) clusters, demonstrating its robustness (Figure 4.7a). However, with the incorporation of DAPI imaging information, `spaMVGAE` can further elucidate intratumoral heterogeneity within DCIS regions (Cluster 2 and Cluster 9), a feature that is not captured when modeling only gene expression across multiple sections (Figure 4.7a,b).

To investigate intratumor heterogeneity in the DCIS regions, we conducted differential expression (DE) analysis and gene set enrichment analysis (GSEA) between Cluster 2 and Cluster 9, as identified by `spaMVGAE` when DAPI imaging information was incorporated. Our DE analysis revealed 254 DE genes, with the top DE genes highlighted in the volcano plot (Figure 4.7c). The top 10 up-regulated genes in the DCIS cluster (Cluster 9) include ERBB2 (Swain et al. 2023), S100A4 (Fei et al. 2017), CTTN (Moon et al. 2023), SERHL2 (Paul et al. 2023), KRT8 (Scott et al. 2020), SCD (Kubota and Espenshade 2022), ENAH (Di Modugno et al. 2006), TENT5C (Kazazian et al. 2020), CCND1 (Valla et al. 2022), and RUNX1 (van Bragt et al. 2014), all of which have been previously identified as breast cancer-related marker genes in previous studies (Figure 4.7c). Notably, S100A4 has been associated with poor prognosis in breast cancer patients (Fei et al. 2017), and CCND1, located on chromosome 11q13, is linked to high histopathological grade, high proliferation, and the Luminal B subtype of breast cancer (Valla et al. 2022). Moreover, our GSEA analysis identified the top 10 gene sets related to breast cancer (Figure 4.7d). These include "CLIMENT BREAST CANCER COPY Number UP," "WP ERBB SIGNALING PATHWAY," and "NIKOLSKY BREAST CANCER 11Q12 Q14 AMPLICON," all of which are associated with breast cancer-related gene sets and pathways. These results suggest differences in the functional properties of tumors between Cluster 2 and Cluster 9.

In addition to the intratumor heterogeneity identified by `spaMVGAE` when DAPI imaging information was utilized, we observed that by including imaging information, the identified clusters exhibited greater consistency with important marker gene expression patterns. For instance, Cluster 11, identified by `spaMVGAE` with DAPI information, which is situated in the outer layer of the DCIS region, is consistent with the marker genes SERPINA3 and KRT14 (Figure 4.7e). Specifically, SERPINA3 is a marker gene highly expressed in myeloid cells, epithelial cells, and dendritic cells,

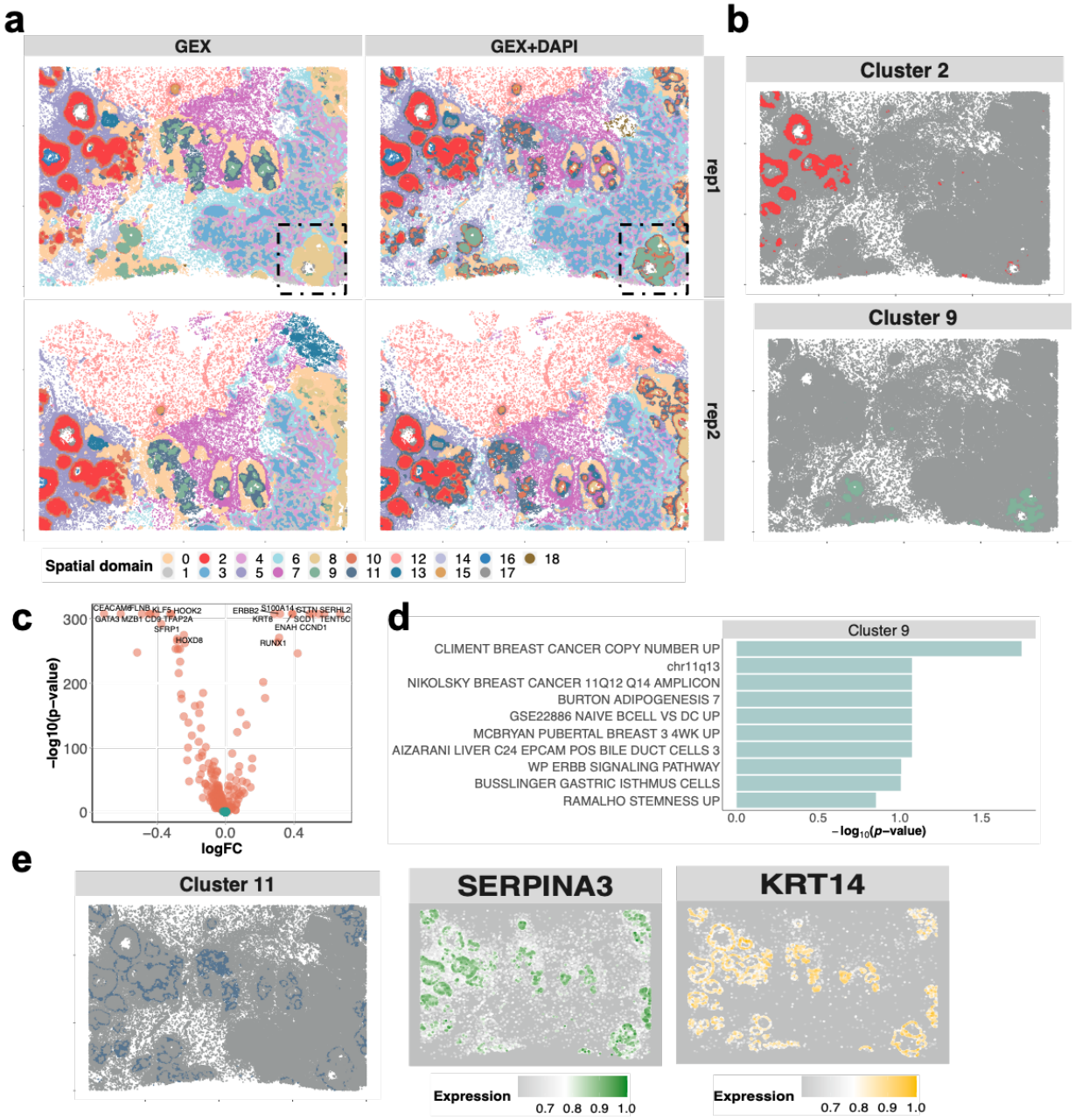


Figure 4.7: `spaMVGAE` characterizes intratumoral and inter-tumoral heterogeneity in 10x Xenium breast cancer data. **(a)** Clusters identified by `spaMVGAE` with gene expression only and with DAPI imaging information added. **(b)** Two DCIS related clusters identified by `spaMVGAE` only when adding the DAPI imaging information. **(c)** Volcano plot displays the DE genes between Cluster 2 and 9. **(d)** GSEA analysis identified the enriched gene sets. **(e)** `spaMVGAE` identified Cluster 11, which is consistent with the immune cells and basal epithelial cells marker genes SERPINA3, and KRT14.

which characterize the important immune environment near the DCIS region (Yu et al. 2023). KRT14 is a marker gene that characterizes basal epithelial cells and has been found to be markedly more highly expressed in DCIS of non-basal-like subtypes compared to their invasive counterparts, particularly in the basal-like subtype (Bergholtz et al. 2020). These results underscore the benefits of spaMVGAE in jointly modeling both gene expression and imaging information across multiple sections.

4.4 Discussion

In this work, we proposed spaMVGAE, a multimodal graph variational autoencoder, to address the challenges presented in integrating spatially resolved multimodal data generated by the latest spatial sequencing technologies. As an end-to-end probabilistic method, spaMVGAE is able to directly model the transcripts (gene expression) and the fragments (chromatin accessibilty) as count data, without additional treatment such as principal component analysis. The spatial location information is incorporated by graph convolutional layers during data encoding. spaMVGAE effectively extracts the biologically meaningful low-dimensional embeddings of the cells/beads/spots that can be used for spatial clustering to gain biological insight of the data.

spaMVGAE manages data produced by different sequencing platforms. The versatility of spaMVGAE is demonstrated through its application on spatially resolved data of various combinations of modalities and biological contexts. To summarize, spaMVGAE makes use of transcriptomic and H&E image features of the spots in HER2-positive human breast tumor tissue, showing the superior capability of domain detection. Next, by incorporating the histology information in addition to the bead-level gene expression data, spaMVGAE successfully stratifies the multi-layer structure in the growth plate in tibia and femur from PTHrP-Knockout mouse that is consistent with the underlying biology of bone development. Furthermore, spaMVGAE integrates the spatially resolved transcriptome and accessibility of chromatin from barcoded pixels, and captures the brain structure within P22 mouse brain with better accuracy than using single modalities. Lastly, we showcase that spaMVGAE is scalable to large datasets by integrating the transcriptomic profile and DAPI images from two adjacent human breast cancer tissue sections. Beyond the utility of spaMVGAE on these real datasets, we can potentially increase its performance by making additional modifications. One improvement can be made is the incorporation of the subgraph-based training scheme (Zeng et al. 2019, Bai et al. 2021), which further boosts the model's scalability in rapidly growing size of the data. Another direction to explore is the addition of modality-specific weight, which allows the model to prioritize more informative modalities.

To conclude, spaMVGAE framework is efficient in spatially informed multimodal integration,

without restrictions on the number of modalities and data types. We envision that `spaMVGAE` will be increasingly useful as more modalities will be available in spatially resolved data thanks to the rapid technological development in this field.

4.5 Methods

4.5.1 Spatial multimodal variational graph autoencoder

We propose spatial multimodal variational graph autoencoder (`spaMVGAE`) to integrate spatially resolved multimodal datasets to obtain joint low-dimensional representation of the cells (Note that they can also be beads, spots depending on the sequencing technologies being used). In this framework, we employ the product of experts (PoE) (Wu and Goodman 2018) for joint posterior inference on the latent space. Then we are able to detect meaningful spatial domains by running clustering methods, such as the Leiden algorithm (Traag et al. 2019). With the assumption that the biological state of a cell is influenced by its immediate environment in the tissue, we want to take into account the information from the neighboring cells during the embedding process. In order to capture this local context, we built a graph for the cells and their surrounding neighbors, and implemented the graph convolutional network (GCN) to aggregate their features.

Consider a total of N cells from T tissue sections, the spatially resolved multimodal dataset comprises M modalities ($\mathbf{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_M\}$), for instance, gene expression, peak, histology. For the ith cell, its feature vector $\mathbf{x}_m$ for each modality contains G_m profiled features. In general, we employed a Poisson distribution to model the count data of gene expression or chromatin accessibility (peaks). Meanwhile, we used mean squared error (MSE) to quantify the reconstruction loss of imaging features. For integration of the transcriptomic profile and morphological features, we applied MSE for both modalities.

4.5.1.1 Graph construction

We first construct a neighborhood graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ for the cells using their spatial coordinates, where $\mathcal{V}$ represents the nodes (i.e., the cell and its K nearest neighbors (KNN) defined by Euclidean distance) and $\mathcal{E}$ denotes the edges connecting these nodes. The choice of K depends on the dataset and sequencing technologies. Finally, we obtain the adjacency matrix $\mathbf{A}$ of $\mathcal{G}$ depicting the spatial relationship among all the cells in the tissue. In dealing with multiple tissue sections ($\mathbf{X}^1, \ldots, \mathbf{X}^T$), we built graphs separately for them and stacked the resulting adjacency matrices along the diagonal:

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}^1 & & \\ & \ddots & \\ & & \mathbf{A}^T \end{bmatrix} \tag{4.1}$$

4.5.1.2 Multimodal variational graph autoencoder

The variational graph autoencoder was proposed for unsupervised learning on data with graph structure (Kipf and Welling 2016b). To extend it for joint cell embeddings from multiple modalities, we adopt the concept of product-of-experts (PoE), introduced by Wu and Goodman for multimodal unsupervised learning (Wu and Goodman 2018). In our setting, each expert is a variational graph autoencoder trained on data from a specific modality ($\mathbf{X}_m$), where graph $\mathcal{G}$ is constructed using the spatial location of the cells, and shared across different modalities. In the multimodal VGAE, we modeled the latent space (Z) of the cells as a Gaussian distribution, and formulated the joint posterior with the assumption of conditional independence between $p(\mathbf{X}_m|\mathbf{Z}, \mathbf{A})$. Then we have

$$p(\mathbf{Z}|\mathbf{X}, \mathbf{A}) = p(\mathbf{Z}|\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_M, \mathbf{A}) \tag{4.2}$$

$$= \frac{p(\mathbf{Z})}{p(\mathbf{X}_1, \ldots, \mathbf{X}_M)} \prod_{m=1}^{M} \frac{p(\mathbf{Z}|\mathbf{X}_m)p(\mathbf{X}_m)}{p(\mathbf{Z})} \tag{4.3}$$

$$\propto \frac{\prod_{m=1}^{M} p(\mathbf{Z}|\mathbf{X}_m, \mathbf{A})}{\prod_{m=1}^{M-1} p(\mathbf{Z})} \tag{4.4}$$

Given each modality and the shared graph information, its true posterior $p(\mathbf{Z}|\mathbf{X}_m, \mathbf{A})$ can be approximated by $q_{\phi_m}(\mathbf{Z}|\mathbf{X}_m, \mathbf{A})$ through the graph convolutional encoder (parameterized by ϕ_m) of the forementioned "expert" model. The decoder, parameterized by θ_m, reconstructs each modality by $p_{\theta_m}(\mathbf{X}_m|\mathbf{Z})$ using fully connected layers. This differs from the original VGAE which learns to reconstruct the adjacency matrix of the input graph. The proposed multimodal VGAE is trained to optimize the evidence lower bound (ELBO) below:

$$\text{ELBO}(\mathbf{X}_1, \ldots, \mathbf{X}_M) \triangleq \mathbb{E}_{q_\phi(\mathbf{Z}|\mathbf{X}_1,\ldots,\mathbf{X}_M,\mathbf{A})}[\sum_{\mathbf{X}_m \in \mathbf{X}} \log p_\theta(\mathbf{X}_m|\mathbf{Z}, \mathbf{A})] - D_{KL}(q_\phi(\mathbf{Z}|\mathbf{X}_1, \ldots, \mathbf{X}_M, \mathbf{A}) \parallel p(\mathbf{Z})]) \tag{4.5}$$

To demonstrate the advantage of leveraging the cells' spatial locations, we also configured a multimodal variational autoencoder (`MVAE`) for benchmark analyses. It is based on the same PoE approach, but without utilizing the graph $\mathcal{G}$.

Encoder Layers			Decoder Layers	
GCNConv	GCNConv	GCNConv(μ, σ)	FC	FC

Table 4.1: Architecture of modality-specific encoder-decoder

4.5.1.3 Model setting and implementation

Given the feature matrix from a measured modality, the encoder infers the modality-specific latent distribution through graph convolutional layers, whereas the decoder reconstructs the input via fully connected (FC) layers. The architecture of a encoder-decoder pair is detailed below 4.1.

The PoE approach is employed to obtain the joint distribution of the latent space of all data modalities. For model training, we input the entire dataset and the complete graph. The Adam optimizer (learning rate $= 0.0001$) was applied. The GCN layers ($GCN_l, l \in \{1, \ldots, L\}$, $L = 2$ by default) in the encoder output 128-dimensional feature maps ($\mathbf{H}_l$), except for $GCN_\mu(\mathbf{H}_L, \mathbf{A})$ and $GCN_\sigma(\mathbf{H}_L, \mathbf{A})$ which depends on the choice of the dimension (D) of cell/bead/spot embedding. The FC layers have 128 hidden units. Non-linearity was modeled through the ReLU activation function. Dropout layers (dropout rate $= 0.2$) were applied to prevent overfitting. Batch Normalization was only applied to FC-layers.

All `spaMVGAE` models were implemented in Pytorch 1.10.1 and Python 3.8, trained with a 2.9 GHz Intel Xeon Gold 6226R and an NVIDIA A40 GPU. Graph convolutional netoworks were implemented with PyTorch Geometric 2.2.0. MUSE was run with TensorFlow 2.7.0.

4.5.2 Downstream analyses

4.5.2.1 Clustering

We used the `scanpy` pipeline with default settings to perform the Leiden algorithm on the low-dimensional cell/bead/spot embeddings for domain detection (Wolf et al. 2018). The resolution for clustering varies across different experiments.

4.5.2.2 Differential expression analysis

To identify differentially expressed genes across cell clusters, we utilized non-parametric Wilcoxon rank sum test (Mann and Whitney 1947, Korsunsky et al. 2019b) and reported the adjusted p-values corrected by Benjamini-Hochberg method (Benjamini and Hochberg 1995).

4.5.2.3 Trajectory inference

Following Leiden clustering, we applied the partition-based graph abstraction (`PAGA`) (Wolf et al. 2019) via `scanpy` to investigate the developmental trajectories of individual cells/beads/spots.

In the output graph, each node represents cell cluster, and the edges show the transitions between the identified cell populations.

4.5.3 Data pre-processing

4.5.3.1 Human breast cancer (spatial transcriptomics and histology)

Andersson et al. applied the Spatial Transcriptomics technology to measure spot-level gene expression in HER2-positive tumors tissues from eight donors (patient A-H). For this analysis, we focused the data generated from patient H. In addition to the expression profile of $15,029$ genes, we obtained the H&E image tile centered around each of the 613 spot, with each side measuring 150 pixels. These image tiles were then used as input to the pre-trained Inception v3, which outputs the 2048-dimensional feature vectors (Bao et al. 2022). Top $2,000$ variable genes were used for integration. For this analysis, we use $K = 10$ to build the KNN graph. The `spaMVGAE` was trained for $5,000$ epoachs, with latent dimension(D) set to 30. MSE loss was applied.

4.5.3.2 Mouse (E18.5) bone development (spatial transcriptomics and histology)

Our collaborators Shion Orikasa and Noriaki Ono, from the Department of Diagnostic and Biomedical Sciences, University of Texas Health Science Center at Houston, extracted the tissue from the leg of PTHrP-KO mouse at embryonic day (E) 18.5. Spatial transcriptomics data were generated using CurioSeeker (v1.0) technology, which is based on slide-seq. In this dataset, UMI count of $28,514$ genes were profiled from $69,138$ beads. In the quality control step, we kept $50,465$ beads with at least 50 UMI count and no more than $1,000$ (Figure 4.4a). We excluded genes expressed in fewer than 3 beads. After log normalization, $2,000$ highly variable genes were selected using `Seurat` for integration. In addition, H&E image was obtained from the adjacent tissue section, then registered to the beads. We cropped the image tile centered around each bead, with each side measuring 100 pixels. Then we resized these histology image tiles to 299×299 pixels and extracted 2048-dimensional feature vectors using pre-trained Inception v3. We built KNN graph with $K = 10$, and used it to train `spaMVGAE` with the dimension of the latent space (D) set to 30, and 600 epochs. MSE loss was applied for both data modalities.

4.5.3.3 Mouse (P22) brain (spatial transcriptomics and epigenomics)

Zhang et al. jointly profiled chromatin accessibility and transcriptome of P22 mouse brain coronal sections by applying the latest spatial ATAC-RNA-seq (Zhang et al. 2023b). This technology allows for sequencing up to $10,000$ pixels per tissue section. The published data object consists of $9,215$ pixels, with $22,914$ genes and $121,068$ peaks profiled. In addition, the author generated $24,027$ ATAC features that are associated with genes. The median number of transcripts and

fragments per pixel is $2,168$ and $6,653$ respectively (Figure 4.6a). For joint modeling, we used the same set of highly variable genes reported by the authors, and selected top $6,000$ variable peaks. To incorporate the spatial information, we built KNN graph with $K = 4$, and used it to train `spaMVGAE` with the dimension of the latent space (D) set to 30, and 1000 epochs. Poisson distributional assumption was applied for both transcript and peak count data.

4.5.3.4 Multi-section human breast cancer (high-throughput spatial transcriptomics and morphology)

Janesick et al. from 10x Genomics applied their latest image-based Xenium technology and profiled 313 genes in the human breast cancer tissues (Janesick et al. 2023). After removing cells with fewer than 100 transcripts, there are $124,945$ in tissue section 1 (Replicate 1), and $90,424$ in tissue section 2 (Replicate 2). The segmented grayscale DAPI image of each cell was first resized to 299×299, then stacked three times so that they could used to extract the 2048-dimensional feature vector via Inception v3 (requires RGB image input). We used all genes for integrative analyses. We used all 313 genes and the image features for multi-section multimodal integration. The spatial information was taken into account through the construction of KNN graph with $K = 10$. The `spaMVGAE` was trained for 1000 epochs, with the dimension of the latent space (D) set to 50 and MSE loss.

CHAPTER 5

Conclusion

This dissertation has focused on developing efficient computational methods to tackle the challenges in integrative analysis of single-cell unimodal omics, single-cell multimodal omics, and spatially resolved multimodal data. The goal is to extract a biologically meaningful, low-dimensional representation of cells from single-cell data and cells/beads/spots from spatial sequencing data. These embeddings capture cellular heterogeneities in transcriptome, epigenome, and morphology. They can be used for cell type inference and spatial domain detection, thereby enhancing our understanding in cell and disease biology. The contribution of the completed studies can be summarized below.

In chapter 2, I focused on the integration of high-throughput single-cell unimodal datasets, each measuring the same or different molecular modalities. For this purpose, I developed `Online iNMF`, which is able to iteratively refines the metagene factors by reading mini-batches from disk. `Online iNMF` stands out for its speedier convergence compared to other batch algorithms and its ability to keep memory usage independent of the dataset size. `Online iNMF` accomplishes iterative integration of single-cell data across three scenarios. In the first scenario, where all single-cell datasets are already available, `Online iNMF` rapidly factorizes the single-cell data into metagenes and cell factor loadings through multiple training epochs. The second scenario involves `Online iNMF` iteratively incorporating single-cell datasets as they sequentially arrive. This approach is expected to be particularly beneficial for researchers who regularly add new sequenced cells to comprehensive cell atlases. The third scenario is promising for querying datasets with a vast, well-established reference atlas. The performance of `Online iNMF` is validated on simulated data and real datasets generated from mouse brain, human PBMC, and human pancreas. The increasing utility of `Online iNMF` is anticipated in the integration of large-scale single-cell multi-omic datasets from major projects like the BRAIN Initiative, Human Body Map, and Human Cell Atlas.

In chapter 3, my work centers around the integration of single-cell multimodal epignomic data. I take the window-based genome binning approach and count the fragments (measurement of histone modification or chromatin accessibility) in each bin for each modality. To model the

count data jointly for all modalities, I introduce the `ConvNet-VAE` framework, characterized by its use of 1D-convolutional layers specially designed to accommodate the multi-track and sequential nature of the binned multimodal epigenomic data. `ConvNet-VAE` successfully integrates bi-modal (H3K27ac + H3K27me3) and tri-modal juvenile mouse brain data (ATAC + H3K27ac + H3K27me3), as well as bi-modal data from human bone marrow mononuclear cells (H3K27ac + H3K27me3). `ConvNet-VAE` shows proficiency in extracting joint embeddings from chromatin states and histone modifications, precisely capturing the data distribution, and correcting for batch-effects. The advantage of `ConvNet-VAE` is more pronounced when analyzing three modalities, rather than just two. `ConvNet-VAE` does have certain limitations. The necessity for all modalities to have a uniform feature space due to the convolutional filters is one such limitation. Additionally, there's room for enhancing model efficacy by fine-tuning parameters like kernel size and stride length. Most recently, researchers have attempted to profile up to 6 epigenomic modalities in single cells (Lochs et al. 2023). As the scale and complexity of such datasets grow, we expect the relevance and utility of this approach to increase significantly.

In chapter 4, I concentrated on the spatially informed integration of multimodal data. I devise `spaMVGAE`, a multimodal graph variational autoencoder, specifically designed to tackle the complexities of integrating spatially resolved multimodal data generated by cutting-edge spatial sequencing technologies. `spaMVGAE` is adaptable to data from various sequencing platforms and demonstrates its versatility across a range of modalities (e.g., gene expression, chromatin accessibility, morphology) and biological contexts (e.g., breast tumor mouse brain, mouse bone). `spaMVGAE` is capable of multi-section integration and scalable to hundreds of thousands of cell. Potential improvments to `spaMVGAE` include adopting a subgraph-based training scheme for further increasing scalability with growing data sizes, and incorporating modality-specific weights to prioritize more informative modalities.

In summary, I have developed three computational methods for integrating single-cell and spatial multimodal data. These methods effectively capture the biological information from each modality and learn joint cell embeddings for clustering, enabling the identification of cell populations and tissue structures. I believe that these developed tools will assist researchers in analyzing complex biomedical data, thereby facilitating biological discovery.

BIBLIOGRAPHY

Rory Stark, Marta Grzelak, and James Hadfield. Rna sequencing: the teenage years. *Nature Reviews Genetics*, 20(11):631–656, 2019.

Fiorella C Grandi, Hailey Modi, Lucas Kampman, and M Ryan Corces. Chromatin accessibility profiling by atac-seq. *Nature protocols*, 17(6):1518–1552, 2022.

Mingye Hong, Shuang Tao, Ling Zhang, Li-Ting Diao, Xuanmei Huang, Shaohui Huang, Shu-Juan Xie, Zhen-Dong Xiao, and Hua Zhang. Rna sequencing: new technologies and applications in cancer research. *Journal of hematology & oncology*, 13(1):1–16, 2020.

Erwin L Van Dijk, Hélène Auger, Yan Jaszczyszyn, and Claude Thermes. Ten years of next-generation sequencing technology. *Trends in genetics*, 30(9):418–426, 2014.

Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B Tuch, Asim Siddiqui, et al. mrna-seq whole-transcriptome analysis of a single cell. *Nature methods*, 6(5):377–382, 2009.

Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

Allon M Klein, Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Victor Li, Leonid Peshkin, David A Weitz, and Marc W Kirschner. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201, 2015.

Daniel Ramsköld, Shujun Luo, Yu-Chieh Wang, Robin Li, Qiaolin Deng, Omid R Faridani, Gregory A Daniels, Irina Khrebtukova, Jeanne F Loring, Louise C Laurent, et al. Full-length mrna-seq from single-cell levels of rna and individual circulating tumor cells. *Nature biotechnology*, 30(8):777–782, 2012.

Simone Picelli, Omid R Faridani, Åsa K Björklund, Gösta Winberg, Sven Sagasser, and Rickard Sandberg. Full-length rna-seq from single cells using smart-seq2. *Nature protocols*, 9(1):171–181, 2014.

Michael Hagemann-Jensen, Christoph Ziegenhain, Ping Chen, Daniel Ramsköld, Gert-Jan Hendriks, Anton JM Larsson, Omid R Faridani, and Rickard Sandberg. Single-cell rna counting at allele and isoform resolution using smart-seq3. *Nature Biotechnology*, 38(6):708–714, 2020.

Xiao Dong, Lei Zhang, Brandon Milholland, Moonsook Lee, Alexander Y Maslov, Tao Wang, and Jan Vijg. Accurate identification of single-nucleotide variants in whole-genome-amplified single cells. *Nature methods*, 14(5):491–493, 2017.

Xian F Mallory, Mohammadamin Edrisi, Nicholas Navin, and Luay Nakhleh. Methods for copy number aberration detection from single-cell dna-sequencing data. *Genome biology*, 21(1):1–22, 2020.

Darren A Cusanovich, Riza Daza, Andrew Adey, Hannah A Pliner, Lena Christiansen, Kevin L Gunderson, Frank J Steemers, Cole Trapnell, and Jay Shendure. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*, 348(6237):910–914, 2015.

Sebastian Preissl, Rongxin Fang, Hui Huang, Yuan Zhao, Ramya Raviram, David U Gorkin, Yanxiao Zhang, Brandon C Sos, Veena Afzal, Diane E Dickel, et al. Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation. *Nature neuroscience*, 21(3):432–439, 2018.

Chongyuan Luo, Christopher L Keown, Laurie Kurihara, Jingtian Zhou, Yupeng He, Junhao Li, Rosa Castanon, Jacinta Lucero, Joseph R Nery, Justin P Sandoval, et al. Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. *Science*, 357(6351):600–604, 2017.

Chongyuan Luo, Angeline Rivkin, Jingtian Zhou, Justin P Sandoval, Laurie Kurihara, Jacinta Lucero, Rosa Castanon, Joseph R Nery, António Pinto-Duarte, Brian Bui, et al. Robust single-cell dna methylome profiling with snmc-seq2. *Nature communications*, 9(1):3824, 2018.

N Editorial. Method of the year 2013. *Nat. Methods*, 11(1):1, 2014.

Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902, 2019.

Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with harmony. *Nature methods*, 16(12):1289–1296, 2019a.

Joshua D Welch, Velina Kozareva, Ashley Ferreira, Charles Vanderburg, Carly Martin, and Evan Z Macosko. Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell*, 177(7): 1873–1887, 2019.

Jialin Liu, Chao Gao, Joshua Sodicoff, Velina Kozareva, Evan Z Macosko, and Joshua D Welch. Jointly defining cell types from multiple single-cell datasets using liger. *Nature protocols*, 15(11):3632–3662, 2020.

Sarah Teichmann and Mirjana Efremova. Method of the year 2019: single-cell multimodal omics. *Nat. Methods*, 17(1):2020, 2020.

Iain C Macaulay, Wilfried Haerty, Parveen Kumar, Yang I Li, Tim Xiaoming Hu, Mabel J Teng, Mubeen Goolam, Nathalie Saurat, Paul Coupland, Lesley M Shirley, et al. G&t-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nature methods*, 12(6):519–522, 2015.

Kyung Yeon Han, Kyu-Tae Kim, Je-Gun Joung, Dae-Soon Son, Yeon Jeong Kim, Areum Jo, Hyo-Jeong Jeon, Hui-Sung Moon, Chang Eun Yoo, Woosung Chung, et al. Sidr: simultaneous isolation and parallel sequencing of genomic dna and total rna from single cells. *Genome research*, 28(1):75–87, 2018.

Alba Rodriguez-Meira, Gemma Buck, Sally-Ann Clark, Benjamin J Povinelli, Veronica Alcolea, Eleni Louka, Simon McGowan, Angela Hamblin, Nikolaos Sousos, Nikolaos Barkas, et al. Unravelling intratumoral heterogeneity through high-sensitivity single-cell mutational analysis and parallel rna sequencing. *Molecular cell*, 73(6):1292–1305, 2019.

Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert. Simultaneous epitope and transcriptome measurement in single cells. *Nature methods*, 14(9):865–868, 2017.

Sebastian Pott. Simultaneous measurement of chromatin accessibility, dna methylation, and nucleosome phasing in single cells. *elife*, 6:e23203, 2017.

Junyue Cao, Darren A Cusanovich, Vijay Ramani, Delasa Aghamirzaie, Hannah A Pliner, Andrew J Hill, Riza M Daza, Jose L McFaline-Figueroa, Jonathan S Packer, Lena Christiansen, et al. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science*, 361(6409):1380–1385, 2018.

Song Chen, Blue B Lake, and Kun Zhang. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nature biotechnology*, 37(12):1452–1457, 2019a.

Kevin Grosselin, Adeline Durand, Justine Marsolier, Adeline Poitou, Elisabetta Marangoni, Fariba Nemati, Ahmed Dahmani, Sonia Lameiras, Fabien Reyal, Olivia Frenoy, et al. High-throughput single-cell chip-seq identifies heterogeneity of chromatin states in breast cancer. *Nature genetics*, 51(6):1060–1066, 2019.

Marek Bartosovic and Goncalo Castelo-Branco. Multimodal chromatin profiling using nanobody-based single-cell cut&tag. *Nature Biotechnology*, pages 1–12, 2022.

Tim Stuart, Stephanie Hao, Bingjie Zhang, Levan Mekerishvili, Dan A Landau, Silas Maniatis, Rahul Satija, and Ivan Raimondi. Nanobody-tethered transposition enables multifactorial chromatin profiling at single-cell resolution. *Nature Biotechnology*, pages 1–7, 2022.

Vivien Marx. Method of the year: spatially resolved transcriptomics. *Nature methods*, 18(1):9–14, 2021.

Alev Baysoy, Zhiliang Bai, Rahul Satija, and Rong Fan. The technological landscape and applications of single-cell multi-omics. *Nature Reviews Molecular Cell Biology*, pages 1–19, 2023.

Eric Lubeck, Ahmet F Coskun, Timur Zhiyentayev, Mubhij Ahmad, and Long Cai. Single-cell in situ rna profiling by sequential hybridization. *Nature methods*, 11(4):360–361, 2014.

Andrea M Femino, Fredric S Fay, Kevin Fogarty, and Robert H Singer. Visualization of single rna transcripts in situ. *Science*, 280(5363):585–590, 1998.

Kok Hao Chen, Alistair N Boettiger, Jeffrey R Moffitt, Siyuan Wang, and Xiaowei Zhuang. Spatially resolved, highly multiplexed rna profiling in single cells. *Science*, 348(6233):aaa6090, 2015a.

Meng Zhang, Xingjie Pan, Won Jung, Aaron Halpern, Stephen W Eichhorn, Zhiyun Lei, Limor Cohen, Kimberly A Smith, Bosiljka Tasic, Zizhen Yao, et al. A molecularly defined and spatially resolved cell atlas of the whole mouse brain. *bioRxiv*, pages 2023–03, 2023a.

Amanda Janesick, Robert Shelansky, Andrew D Gottscho, Florian Wagner, Stephen R Williams, Morgane Rouault, Ghezal Beliakoff, Carolyn A Morrison, Michelli F Oliveira, Jordan T Sicherman, et al. High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis. *Nature Communications*, 14(1):8353, 2023.

Patrik L Ståhl, Fredrik Salmén, Sanja Vickovic, Anna Lundmark, José Fernández Navarro, Jens Magnusson, Stefania Giacomello, Michaela Asp, Jakub O Westholm, Mikael Huss, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82, 2016.

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

Jun Chen, Shengbao Suo, Patrick PL Tam, Jing-Dong J Han, Guangdun Peng, and Naihe Jing. Spatial transcriptomic analysis of cryosectioned tissue samples with geo-seq. *Nature protocols*, 12(3):566–580, 2017.

Di Zhang, Yanxiang Deng, Petra Kukanja, Eneritz Agirre, Marek Bartosovic, Mingze Dong, Cong Ma, Sai Ma, Graham Su, Shuozhen Bao, et al. Spatial epigenome–transcriptome co-profiling of mammalian tissues. *Nature*, 616(7955):113–122, 2023b.

Andrew JC Russell, Jackson A Weir, Naeem M Nadaf, Matthew Shabet, Vipin Kumar, Sandeep Kambhampati, Ruth Raichur, Giovanni J Marrero, Sophia Liu, Karol S Balderrama, et al. Slide-tags enables single-nucleus barcoding for multimodal spatial genomics. *Nature*, pages 1–9, 2023.

Zi Ye and Casim A Sarkar. Towards a quantitative understanding of cell identity. *Trends in cell biology*, 28 (12):1030–1048, 2018.

Tim Stuart and Rahul Satija. Integrative single-cell analysis. *Nature reviews genetics*, 20(5):257–272, 2019.

Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11(1), 2010.

Arpiar Saunders, Evan Z Macosko, Alec Wysoker, Melissa Goldman, Fenna M Krienen, Heather de Rivera, Elizabeth Bien, Matthew Baum, Laura Bortolin, Shuyu Wang, et al. Molecular diversity and specializations among the cells of the adult mouse brain. *Cell*, 174(4):1015–1030, 2018.

Hoa Thi Nhu Tran, Kok Siong Ang, Marion Chevrier, Xiaomeng Zhang, Nicole Yee Shin Lee, Michelle Goh, and Jinmiao Chen. A benchmark of batch-effect correction methods for single-cell rna sequencing data. *Genome biology*, 21:1–32, 2020.

Hyun Min Kang, Meena Subramaniam, Sasha Targ, Michelle Nguyen, Lenka Maliskova, Elizabeth McCarthy, Eunice Wan, Simon Wong, Lauren Byrnes, Cristina M Lanata, et al. Multiplexed droplet single-cell rna-sequencing using natural genetic variation. *Nature biotechnology*, 36(1):89–94, 2018.

Dominic Grün, Mauro J Muraro, Jean-Charles Boisset, Kay Wiebrands, Anna Lyubimova, Gitanjali Dharmadhikari, Maaike van den Born, Johan Van Es, Erik Jansen, Hans Clevers, et al. De novo prediction of stem cell identity using single-cell transcriptome data. *Cell stem cell*, 19(2):266–277, 2016.

Mauro J Muraro, Gitanjali Dharmadhikari, Dominic Grün, Nathalie Groen, Tim Dielen, Erik Jansen, Leon Van Gurp, Marten A Engelse, Francoise Carlotti, Eelco Jp De Koning, et al. A single-cell transcriptome atlas of the human pancreas. *Cell systems*, 3(4):385–394, 2016.

Nathan Lawlor, Joshy George, Mohan Bolisetty, Romy Kursawe, Lili Sun, V Sivakamasundari, Ina Kycia, Paul Robson, and Michael L Stitzel. Single-cell transcriptomes identify human islet cell signatures and reveal cell-type–specific expression changes in type 2 diabetes. *Genome research*, 27(2):208–222, 2017.

Maayan Baron, Adrian Veres, Samuel L Wolock, Aubrey L Faust, Renaud Gaujoux, Amedeo Vetere, Jennifer Hyoje Ryu, Bridget K Wagner, Shai S Shen-Orr, Allon M Klein, et al. A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure. *Cell systems*, 3 (4):346–360, 2016.

Åsa Segerstolpe, Athanasia Palasantza, Pernilla Eliasson, Eva-Marie Andersson, Anne-Christine Andréasson, Xiaoyan Sun, Simone Picelli, Alan Sabirsh, Maryam Clausen, Magnus K Bjursell, et al. Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. *Cell metabolism*, 24(4): 593–607, 2016.

Tomohisa Toda, Sarah L Parylak, Sara B Linker, and Fred H Gage. The role of adult hippocampal neurogenesis in brain health and disease. *Molecular psychiatry*, 24(1):67–87, 2019.

Aurélie Ernst, Kanar Alkass, Samuel Bernard, Mehran Salehpour, Shira Perl, John Tisdale, Göran Possnert, Henrik Druid, and Jonas Frisén. Neurogenesis in the striatum of the adult human brain. *Cell*, 156(5): 1072–1083, 2014.

Amit Zeisel, Hannah Hochgerner, Peter Lönnerberg, Anna Johnsson, Fatima Memic, Job Van Der Zwan, Martin Häring, Emelie Braun, Lars E Borm, Gioele La Manno, et al. Molecular architecture of the mouse nervous system. *Cell*, 174(4):999–1014, 2018.

Junyue Cao, Malte Spielmann, Xiaojie Qiu, Xingfan Huang, Daniel M Ibrahim, Andrew J Hill, Fan Zhang, Stefan Mundlos, Lena Christiansen, Frank J Steemers, et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, 566(7745):496–502, 2019.

RR Stickels et al. Sensitive spatial genome wide expression profiling at cellular resolution. biorxiv (2020).

KH Chen, AN Boettiger, JR Moffitt, S Wang, and X Zhuang. Rna imaging. spatially resolved, highly multiplexed rna profiling in single cells. science 348, aaa6090, 2015b.

Zizhen Yao, Cindy TJ van Velthoven, Thuc Nghi Nguyen, Jeff Goldy, Adriana E Sedeno-Cortes, Fahimeh Baftizadeh, Darren Bertagnolli, Tamara Casper, Megan Chiang, Kirsten Crichton, et al. A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation. *Cell*, 184(12):3222–3241, 2021.

Jeffrey R Moffitt, Dhananjay Bambah-Mukku, Stephen W Eichhorn, Eric Vaughn, Karthik Shekhar, Julio D Perez, Nimrod D Rubinstein, Junjie Hao, Aviv Regev, Catherine Dulac, et al. Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science*, 362(6416):eaau5324, 2018.

Joseph R Ecker, Daniel H Geschwind, Arnold R Kriegstein, John Ngai, Pavel Osten, Damon Polioudakis, Aviv Regev, Nenad Sestan, Ian R Wickersham, and Hongkui Zeng. The brain initiative cell census consortium: lessons learned toward generating a comprehensive brain cell atlas. *Neuron*, 96(3):542–557, 2017.

Caltech-UW TMC Cai Long lcai@ caltech. edu 21 b Shendure Jay 9 Trapnell Cole 9 Lin Shin shinlin@ uw. edu 2 e Jackson Dana 9, UCSD TMC Zhang Kun kzhang@ bioeng. ucsd. edu 15 b Sun Xin 15 Jain Sanjay 24 Hagood James 25 Pryhuber Gloria 26 Kharchenko Peter 8, California Institute of Technology TTD Cai Long lcai@ caltech. edu 21 b Yuan Guo-Cheng 35 Zhu Qian 35 Dries Ruben 35, Harvard TTD Yin Peng peng_yin@ hms. harvard. edu 36 37 b Saka Sinem K. 36 37 Kishi Jocelyn Y. 36 37 Wang Yu 36 37 Goldaracena Isabel 36 37, Purdue TTD Laskin Julia jlaskin@ purdue. edu 10 b Ye DongHye 10 38 Burnum-Johnson Kristin E. 39 Piehowski Paul D. 39 Ansong Charles 39 Zhu Ying 39, Stanford TTD Harbury Pehr harbury@ stanford. edu 11 b Desai Tushar 40 Mulye Jay 11 Chou Peter 11 Nagendran Monica 40, et al. The human body at cellular resolution: the nih human biomolecular atlas program. *Nature*, 574(7777):187–192, 2019.

Aviv Regev, Sarah A Teichmann, Eric S Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, et al. The human cell atlas. *elife*, 6: e27041, 2017.

Zizhen Yao, Hanqing Liu, Fangming Xie, Stephan Fischer, A Sina Booeshaghi, Ricky S Adkins, Andrew I Aldridge, Seth A Ament, Antonio Pinto-Duarte, Anna Bartlett, et al. An integrated transcriptomic and epigenomic atlas of mouse primary motor cortex cell types. *Biorxiv*, pages 2020–02, 2020.

Bosiljka Tasic, Vilas Menon, Thuc Nghi Nguyen, Tae Kyung Kim, Tim Jarsky, Zizhen Yao, Boaz Levi, Lucas T Gray, Staci A Sorensen, Tim Dolbeare, et al. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nature neuroscience*, 19(2):335–346, 2016.

Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology*, 36(5): 411–420, 2018.

Maren Büttner, Zhichao Miao, F Alexander Wolf, Sarah A Teichmann, and Fabian J Theis. A test metric for assessing single-cell rna-seq batch correction. *Nature methods*, 16(1):43–49, 2019.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2:193–218, 1985.

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

Luke Zappia, Belinda Phipson, and Alicia Oshlack. Splatter: simulation of single-cell rna sequencing data. *Genome biology*, 18(1):174, 2017.

Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M Mauck, Shiwei Zheng, Andrew Butler, Maddie J Lee, Aaron J Wilk, Charlotte Darby, Michael Zager, et al. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587, 2021.

Ricard Argelaguet, Damien Arnol, Danila Bredikhin, Yonatan Deloro, Britta Velten, John C Marioni, and

Oliver Stegle. Mofa+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome biology*, 21(1):1–17, 2020.

Tal Ashuach, Mariano I Gabitto, Rohan V Koodli, Giuseppe-Antonio Saldi, Michael I Jordan, and Nir Yosef. Multivi: deep generative model for the integration of multimodal data. *Nature Methods*, 20(8): 1222–1231, 2023.

Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058, 2018.

Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. *Advances in neural information processing systems*, 31, 2018.

Boying Gong, Yun Zhou, and Elizabeth Purdom. Cobolt: integrative analysis of multimodal single-cell sequencing data. *Genome biology*, 22(1):1–21, 2021.

Serkan Kiranyaz, Onur Avci, Osama Abdeljaber, Turker Ince, Moncef Gabbouj, and Daniel J Inman. 1d convolutional neural networks and applications: A survey. *Mechanical systems and signal processing*, 151:107398, 2021.

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

Kathleen M Chen, Aaron K Wong, Olga G Troyanskaya, and Jian Zhou. A sequence-based global map of regulatory activity for deciphering human genetics. *Nature genetics*, 54(7):940–949, 2022.

Han Yuan and David R Kelley. scbasset: sequence-based modeling of single-cell atac-seq using convolutional neural networks. *Nature Methods*, 19(9):1088–1096, 2022.

Huidong Chen, Caleb Lareau, Tommaso Andreani, Michael E Vinyard, Sara P Garcia, Kendell Clement, Miguel A Andrade-Navarro, Jason D Buenrostro, and Luca Pinello. Assessment of computational methods for the analysis of single-cell atac-seq data. *Genome biology*, 20(1):1–25, 2019b.

Chao Gao, Jialin Liu, April R Kriebel, Sebastian Preissl, Chongyuan Luo, Rosa Castanon, Justin Sandoval, Angeline Rivkin, Joseph R Nery, Margarita M Behrens, et al. Iterative single-cell multi-omic integration using online learning. *Nature biotechnology*, 39(8):1000–1007, 2021.

Malte D Luecken, Maren Büttner, Kridsadakorn Chaichoompu, Anna Danese, Marta Interlandi, Michaela F Müller, Daniel C Strobl, Luke Zappia, Martin Dugas, Maria Colomé-Tatché, et al. Benchmarking atlas-level data integration in single-cell genomics. *Nature methods*, 19(1):41–50, 2022.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

Yanshuai Cao and David J Fleet. Generalized product of experts for automatic and principled fusion of gaussian process predictions. *arXiv preprint arXiv:1410.7827*, 2014.

Art Owen and Yi Zhou. Safe and effective importance sampling. *Journal of the American Statistical Association*, 95(449):135–143, 2000.

Ludo Waltman and Nees Jan Van Eck. A smart local moving algorithm for large-scale modularity-based community detection. *The European physical journal B*, 86:1–14, 2013.

Tim Stuart, Avi Srivastava, Shaista Madad, Caleb A Lareau, and Rahul Satija. Single-cell chromatin state analysis with signac. *Nature methods*, 18(11):1333–1341, 2021.

Rongxin Fang, Sebastian Preissl, Yang Li, Xiaomeng Hou, Jacinta Lucero, Xinxin Wang, Amir Motamedi, Andrew K Shiau, Xinzhu Zhou, Fangming Xie, et al. Comprehensive analysis of single cell atac-seq data with snapatac. *Nature communications*, 12(1):1337, 2021.

Haley M Amemiya, Anshul Kundaje, and Alan P Boyle. The encode blacklist: identification of problematic regions of the genome. *Scientific reports*, 9(1):9354, 2019.

Laura D Martens, David S Fischer, Vicente A Yépez, Fabian J Theis, and Julien Gagneur. Modeling fragment counts improves single-cell atac-seq analysis. *Nature Methods*, pages 1–4, 2023.

Lu Lu and Joshua D Welch. Pyliger: scalable single-cell multi-omic data integration in python. *Bioinformatics*, 38(10):2946–2948, 2022.

Bosiljka Tasic, Zizhen Yao, Lucas T Graybuck, Kimberly A Smith, Thuc Nghi Nguyen, Darren Bertagnolli, Jeff Goldy, Emma Garren, Michael N Economo, Sarada Viswanathan, et al. Shared and distinct transcriptomic cell types across neocortical areas. *Nature*, 563(7729):72–78, 2018.

Ricard Argelaguet, Tim Lohoff, Jingyu Gavin Li, Asif Nakhuda, Deborah Drage, Felix Krueger, Lars Velten, Stephen J Clark, and Wolf Reik. Decoding gene regulation in the mouse embryo using single-cell multi-omics. *bioRxiv*, pages 2022–06, 2022.

Jeffrey M Granja, M Ryan Corces, Sarah E Pierce, S Tansu Bagdatli, Hani Choudhry, Howard Y Chang, and William J Greenleaf. Archr is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nature genetics*, 53(3):403–411, 2021.

Marek Bartosovic and Gonçalo Castelo-Branco. Multimodal chromatin profiling using nanobody-based single-cell cut&tag. *Nature Biotechnology*, 41(6):794–805, 2023.

Robert R Stickels, Evan Murray, Pawan Kumar, Jilong Li, Jamie L Marshall, Daniela J Di Bella, Paola Arlotta, Evan Z Macosko, and Fei Chen. Highly sensitive spatial transcriptomics at near-cellular resolution with slide-seqv2. *Nature biotechnology*, 39(3):313–319, 2021.

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

Feng Bao, Yue Deng, Sen Wan, Susan Q Shen, Bo Wang, Qionghai Dai, Steven J Altschuler, and Lani F Wu. Integrative spatial analysis of cell morphologies and transcriptional states with muse. *Nature biotechnology*, 40(8):1200–1209, 2022.

Marco Varrone, Daniele Tavernari, Albert Santamaria-Martínez, Logan A Walsh, and Giovanni Ciriello. Cellcharter reveals spatial cell niches associated with tissue remodeling and cell plasticity. *Nature Genetics*, pages 1–11, 2023.

Yahui Long, Kok Siong Ang, Sha Liao, Raman Sethi, Yang Heng, Chengwei Zhong, Hang Xu, Nazihah Husna, Min Jian, Lai Guan Ng, et al. Integrated analysis of spatial multi-omics with spatialglue. *bioRxiv*, pages 2023–04, 2023.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016a.

Alma Andersson, Ludvig Larsson, Linnea Stenbeck, Fredrik Salmén, Anna Ehinger, Sunny Z Wu, Ghamdan Al-Eryani, Daniel Roden, Alex Swarbrick, Åke Borg, et al. Spatial deconvolution of her2-positive breast cancer delineates tumor-associated cell type interactions. *Nature communications*, 12(1):6012, 2021.

Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):5233, 2019.

Henry M Kronenberg. Developmental regulation of the growth plate. *Nature*, 423(6937):332–336, 2003.

Liu Yang, Kwok Yeung Tsang, Hoi Ching Tang, Danny Chan, and Kathryn SE Cheah. Hypertrophic chondrocytes can become osteoblasts and osteocytes in endochondral bone formation. *Proceedings of the National Academy of Sciences*, 111(33):12097–12102, 2014.

F Alexander Wolf, Fiona K Hamey, Mireya Plass, Jordi Solana, Joakim S Dahlin, Berthold Göttgens, Nikolaus Rajewsky, Lukas Simon, and Fabian J Theis. Paga: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome biology*, 20:1–9, 2019.

Geoffrey Stanley, Ozgun Gokce, Robert C Malenka, Thomas C Südhof, and Stephen R Quake. Continuous and discrete neuron types of the adult murine striatum. *Neuron*, 105(4):688–699, 2020.

Katherine S Matho, Dhananjay Huilgol, William Galbavy, Miao He, Gukhan Kim, Xu An, Jiangteng Lu, Priscilla Wu, Daniela J Di Bella, Ashwin S Shetty, et al. Genetic dissection of the glutamatergic neuron system in cerebral cortex. *Nature*, 598(7879):182–187, 2021.

Zizhen Yao, Cindy TJ van Velthoven, Michael Kunst, Meng Zhang, Delissa McMillen, Changkyu Lee, Won Jung, Jeff Goldy, Aliya Abdelhak, Matthew Aitken, et al. A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain. *Nature*, 624(7991):317–332, 2023.

Sandra M Swain, Mythili Shastry, and Erika Hamilton. Targeting her2-positive breast cancer: Advances and future directions. *Nature Reviews Drug Discovery*, 22(2):101–126, 2023.

Fei Fei, Jie Qu, Mingqing Zhang, Yuwei Li, and Shiwu Zhang. S100a4 in cancer progression and metastasis: A systematic review. *Oncotarget*, 8(42):73219, 2017.

So-Jeong Moon, Hyung-Jun Choi, Young-Hyeon Kye, Ga-Young Jeong, Hyung-Yong Kim, Jae-Kyung Myung, and Gu Kong. Cttn overexpression confers cancer stem cell-like properties and trastuzumab resistance via dkk-1/wnt signaling in her2 positive breast cancer. *Cancers*, 15(4):1168, 2023.

Evan D Paul, Barbora Huraiova, Natalia Valkova, Natalia Birknerova, Daniela Gabrisova, Sona Gubova, Helena Ignacakova, Tomas Ondris, Silvia Bendikova, Jarmila Bila, et al. Multiplexed rna-fish-guided laser capture microdissection rna sequencing improves breast cancer molecular subtyping, prognostic classification, and predicts response to antibody drug conjugates. *medRxiv*, pages 2023–12, 2023.

Madeleine KD Scott, Maneesha Limaye, Steven Schaffert, Robert West, Michael G Ozawa, Pauline Chu, Viswam S Nair, Albert C Koong, and Purvesh Khatri. A multi-scale integrated analysis identifies krt8 as a pan-cancer early biomarker. In *BIOCOMPUTING 2021: Proceedings of the Pacific Symposium*, pages 297–308. World Scientific, 2020.

Casie S Kubota and Peter J Espenshade. Targeting stearoyl-coa desaturase in solid tumors. *Cancer research*, 82(9):1682–1688, 2022.

Francesca Di Modugno, Marcella Mottolese, Anna Di Benedetto, Andrea Conidi, Flavia Novelli, Letizia Perracchio, Irene Venturo, Claudio Botti, Elke Jager, Angela Santoni, et al. The cytoskeleton regulatory protein hmena (enah) is overexpressed in human benign breast lesions with high risk of transformation and human epidermal growth factor receptor-2–positive/hormonal receptor–negative tumors. *Clinical Cancer Research*, 12(5):1470–1478, 2006.

Karineh Kazazian, Yosr Haffani, Deanna Ng, Chae Min Michelle Lee, Wendy Johnston, Minji Kim, Roland Xu, Karina Pacholzyk, Francis Si-Wah Zih, Julie Tan, et al. Fam46c/tent5c functions as a tumor suppressor through inhibition of plk4 activity. *Communications biology*, 3(1):448, 2020.

Marit Valla, Elise Klæstad, Borgny Ytterhus, and Anna M Bofin. Ccnd1 amplification in breast cancer-associations with proliferation, histopathological grade, molecular subtype and prognosis. *Journal of Mammary Gland Biology and Neoplasia*, 27(1):67–77, 2022.

Maaike PA van Bragt, Xin Hu, Ying Xie, and Zhe Li. Runx1, a transcription factor mutated in breast cancer, controls the fate of er-positive mammary luminal cells. *Elife*, 3:e03881, 2014.

Qiyi Yu, Tianyuan Xie, Yidong Zhang, Tianyue Pan, Yongmei Tan, Hai Qin, and Simin Yan. Exploration of serpina family functions and prognostic value in breast cancer based on transcriptome and in vitro analysis. *Environmental Toxicology*, 2023.

Helga Bergholtz, Tonje G Lien, David M Swanson, Arnoldo Frigessi, Maria Grazia Daidone, Jörg Tost, Fredrik Wärnberg, and Therese Sørlie. Contrasting dcis and invasive breast cancer by subtype suggests basal-like dcis as distinct lesions. *NPJ Breast Cancer*, 6(1):26, 2020.

Hanqing Zeng, Hongkuan Zhou, Ajitesh Srivastava, Rajgopal Kannan, and Viktor Prasanna. Graphsaint: Graph sampling based inductive learning method. *arXiv preprint arXiv:1907.04931*, 2019.

Jiyang Bai, Yuxiang Ren, and Jiawei Zhang. Ripple walk training: A subgraph-based training framework for large and deep graph neural network. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021.

Thomas N Kipf and Max Welling. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*, 2016b.

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

Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60, 1947.

Ilya Korsunsky, Aparna Nathan, Nghia Millard, and Soumya Raychaudhuri. Presto scales wilcoxon and auroc analyses to millions of observations. *BioRxiv*, page 653253, 2019b.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

Silke JA Lochs, Robin H van der Weide, Kim L de Luca, Tessy Korthout, Ramada E van Beek, Hiroshi Kimura, and Jop Kind. Combinatorial single-cell profiling of major chromatin types with mabid. *Nature Methods*, pages 1–11, 2023.