Technische Universität München
TUM School of Life Sciences

Deep generative modeling of transcriptional dynamics and data-view agnostic inference of cellular state changes with single-cell omics data

Philipp Fabian Weiler

Vollständiger Abdruck der von der TUM School of Life Sciences der Technischen Universität München zur Erlangung eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz: Prof. Dr. Markus List

Prüfende der Dissertation:

1. Prof. Dr. Dr. Fabian J. Theis
2. Prof. Dr. Dana Pe'er
3. Prof. Dr. Jason Buenrostro

Die Dissertation wurde am 05.07.2024 bei der Technischen Universität München eingereicht und durch die TUM School of Life Sciences am 06.11.2024 angenommen.

TUM School of Life Sciences

Technische Universität München

Deep generative modeling of transcriptional dynamics and data-view agnostic inference of cellular state changes with single-cell omics data

Philipp Fabian Weiler

July 2024

HELMHOLTZ MUNICH

Acknowledgments

My work has only been possible thanks to the support - both scientific and non-scientific - and collaboration with many kind and intelligent colleagues and companions. I sincerely thank every one of you for the experiences that shaped both me as a person and researcher and my research journey. This brief section is too short to do justice to your impact and name every one of you explicitly. I stand on the shoulders of giants who paved the way for my work.

Thank you to my supervisor Fabian Theis for giving me the chance to pursue a doctoral degree in your research group. You continuously enabled me to learn, improve my research and expand my network in the single-cell community through conferences and research visits. The experiences in your lab allowed me to understand what my work life after the doctoral program should look like for me. I would also like to thank Dana Pe'er and Nicolas Battich for their roles on my Thesis Advisory Committee. Your support, opinions and suggestions have always been essential and very valuable to not get lost but to see the bigger picture, instead. Dana, thank you for taking a chance in me, always having an open ear, and helping me follow my passion.

Thank you to my collaborators I had the privilege to work with on different projects. I would like to explicitly thank Adam Gayoso and Nir Yosef, Donghoon Lee and Panos Roussos, and Marius Lange and Michal Klein. Thank you, Adam, for letting me learn from you and all your effort and time that went into publishing veloVI; thank you, Nir, for your support and your insight and guidance from our conversations. Thank you, Donghoon and Panos, for letting me be part of your research effort to understand Alzheimer's disease better. Thank you, Marius, for letting me take over the CellRank 2 project and co-leading it with me. Thank you, Mike, for all your contributions to making CellRank the software it is today.

Thank you to my friends and colleagues who have supported me both on a scientific and personal level in the last few years - your continuous support has been invaluable. Having people to talk to, set things into perspective, and take my mind off of work was crucial for my motivation and success. I would like to especially thank Amir Ali Moinfar, Anna Schaar, Laura Martens, Leon Hetzel, Lisa Sikkema, Fabiola Curion, Robert Gutgesell, Roshan Sharma and Sara Jimenez: Amir, Anna, Laura, Leon, Lisa, and Robert, thank you for taking on the journey towards a doctoral degree with me, all the nice memories and for becoming truly great friends. Robert, thank you for letting me tag along on bike rides and not dropping me. Faby, Roshan and Sara, thank you for your perspective, support and open ears. Thank you, Roshan, for making me feel at home and welcome in New York, and for all the nice memories we made when exploring New York by bike. I would also like to thank the ICB office for all their administrative and mental support: Thank you to Dani for making all administrative matters as simple as possible. Thank you, Sabine, for going above and beyond your responsibilities to help us out and always being there for us.

Finally, thank you to my family for always believing and supporting me. Thank you for giving me the chance and freedom to pursue my dreams and interests wherever they took me. The journey so far has been twisted and full of surprises, but I have finally found my place - I think.

Abstract

Single-cell genomics is revolutionizing the field of biology by continuously generating larger and more diverse datasets. As a result, studies reveal and investigate cellular heterogeneity, and elucidate the intricate mechanisms of cell differentiation both in health and disease. The fundamental challenge to recovering trajectories is the destructive nature of common single-cell sequencing assays: Experiments capture snapshot views of each cell instead of monitoring its change over time. However, sequencing protocols record a range of the underlying differentiation landscape, nonetheless, as biological processes unfold asynchronously. Additionally, common single-cell RNA sequencing (scRNA-seq) workflows detect mature and nascent messenger RNA (mRNA) transcripts, thereby providing alternative, directed information; sequencing at different time points adds temporal information for systems not in steady state.

Studying biological processes based on single-cell data usually requires reconstructing them computationally. Corresponding methods rely on different views to describe cellular state shifts: Cell-cell similarity and asynchronous state change allow assigning cells a pseudotime or stemness potential to order them along differentiation processes and induce a relative ranking to each other. Alternatively, optimal transport reconstructs cell trajectories across a sequence of time-resolved measurements. However, these approaches do not model the underlying mechanistic system directly to provide directed dynamic information. RNA velocity bridges this gap by modeling splicing dynamics based on unspliced and spliced mRNA counts. The recovered dynamics provide a vector field describing cellular state change mechanistically. Importantly, these methods are limited to specific data aspects, do not scale, or do not carry notions of uncertainty.

The first contribution of this dissertation is veloVI, a deep generative model for inferring RNA velocity and facilitating its analysis. Traditional RNA velocity methods make assumptions oftentimes violated, employ a restrictive inference scheme that does not easily generalize to more accurate models of splicing dynamics and do not scale. Additionally, these approaches do not quantify uncertainty of fits or model applicability, essential aspects to ensure accurate analyses and descriptions of biological systems. Conversely, veloVI provides such notions and introduces novel metrics to evaluate estimates and the applicability of RNA velocity; to compare different models and inference schemes consistently, I also present a quantitative evaluation scheme. The proposed model fits data better than competing approaches, is less sensitive to preprocessing choices, and extends naturally to more complex kinetic models. Similarly, a veloVI-specific evaluation pipeline gives actionable insight into the applicability of RNA velocity analysis for a given dataset.

Although RNA velocity has celebrated tremendous success, it suffers from experimental and conceptual limitations, rendering it not applicable to many datasets. Similarly, traditional trajectory inference focuses on gene expression alone, omitting other modalities such as valuable time point information available in emerging single-cell datasets; methods that incorporate these different data views do not generalize to other modalities, cannot be combined or do not scale. To overcome these limitations, I present CellRank 2

for unified trajectory inference and fate mapping as a second contribution. Under this framework, CellRank 2 incorporates pseudotime and stemness estimates independent of their quantification to infer cellular fate consistently. For datasets including time point information, I present an optimal transport-based trajectory inference scheme to combine inter with intra-time point information. Alternatively, metabolic labels introduce temporal information by characterizing newly transcribed mRNA molecules; I show how this information allows estimating cell-specific transcription and degradation rates. For each method, I describe corresponding analyses to validate their benefit.

Both contributions improve and generalize the inference of trajectories and cellular state changes. Their modular, scalable and versatile design facilitates the discovery of novel biology, guarantees further model improvements, and enables the incorporation of newly emerging data modalities.

Zusammenfassung

Die Einzelzellgenomik revolutioniert das Feld der Biologie, indem sie kontinuierlich größere und vielfältigere Datensätze erzeugt. Infolgedessen können Studien die zelluläre Heterogenität aufdecken und untersuchen und helfen die komplizierten Mechanismen der Zelldifferenzierung sowohl im gesunden als auch kranken Zustand zu beschreiben. Die grundlegende Herausforderung bei der Wiederherstellung von Trajektorien liegt in der destruktiven Natur gängiger Einzelzell-Sequenzierungsexperimenten: Die Experimente erfassen Momentaufnahmen jeder Zelle, anstatt ihre Veränderungen im Laufe der Zeit zu beobachten. Sequenzierungsprotokolle zeichnen jedoch nichtsdestotrotz einen Teil der zugrunde liegenden Differenzierungslandschaft auf, da sich biologische Prozesse asynchron entfalten. Darüber hinaus werden mit den üblichen Arbeitsabläufen für die Einzelzell-RNA-Sequenzierung (scRNA-seq) old und nascent Boten-RNA (mRNA)-Transkripte erfasst, wodurch alternative, gerichtete Informationen bereitgestellt werden; die Sequenzierung zu verschiedenen Zeitpunkten fügt zeitliche Informationen für Systeme hinzu, die sich nicht im Dauerzustand befinden.

Die Untersuchung biologischer Prozesse anhand von Einzelzelldaten erfordert in der Regel mittels computergestützter Rekonstruktion. Entsprechende Methoden stützen sich auf unterschiedliche Sichtweisen zur Beschreibung zellulärer Zustandsveränderungen: Zell-Zell-Ähnlichkeit und asynchrone Zustandsveränderungen erlauben es, Zellen eine pseudotime oder ein Potenzial zuzuordnen, um sie entlang von Differenzierungsprozessen zu ordnen und eine relative Rangfolge zueinander zu erstellen. Alternativ dazu rekonstruiert optimal transport die Zelltrajektorien über eine Sequenz von zeitaufgelösten Messungen. Diese Ansätze modellieren jedoch nicht direkt das zugrunde liegende mechanistische System, um gezielte dynamische Informationen zu liefern. Die RNA velocity überbrückt diese Lücke durch die Modellierung der Spleißdynamik auf der Grundlage der Anzahl der nicht gespleißten und gespleißten mRNA. Die so gewonnene Dynamik liefert ein Vektorfeld, das die Veränderung des Zellzustands mechanistisch beschreibt. Allerdings sind diese Methoden auf bestimmte Datenaspekte beschränkt, nicht skalierbar sind und beinhalten keine Unsicherheiten.

Der erste Beitrag dieser Dissertation ist veloVI, ein tiefes generatives Modell zur Ableitung der RNA velocity und zur Erleichterung ihrer Analyse. Herkömmliche Methoden zur Bestimmung der RNA velocity gehen von Annahmen aus, die häufig verletzt werden, verwenden ein restriktives Inferenzschema, das sich nicht ohne weiteres auf genauere Modelle der Spleißdynamik verallgemeinern lässt, und sind nicht skalierbar. Darüber hinaus quantifizieren diese Ansätze nicht die Unsicherheit der Anpassungen oder die Anwendbarkeit des Modells - wesentliche Aspekte, um genaue Analysen und Beschreibungen biologischer Systeme zu gewährleisten. Im Gegensatz dazu liefert veloVI solche Begriffe und führt neue Metriken ein, um Schätzungen und die Anwendbarkeit von RNA velolcity zu bewerten; um verschiedene Modelle und Inferenzschemata konsistent zu vergleichen, stelle ich auch ein quantitatives Bewertungsschema vor. Insgesamt beschreibt das vorgeschlagene Modell Daten besser als konkurrierende Ansätze, reagiert weniger empfindlich auf die Wahl der Vorverarbeitung und lässt sich natürlich auf komplexere kinetische Modelle ausweiten. Ebenso gibt eine veloVI-spezifische Bewertungspipeline

einen verwertbaren Einblick in die Anwendbarkeit der RNA-Geschwindigkeitsanalyse für einen bestimmten Datensatz.

Obwohl RNA velocity einen enormen Erfolg gefeiert hat, leidet sie unter experimentellen und konzeptionellen Einschränkungen, sodass sie auf viele Datensätze nicht anwendbar ist. In ähnlicher Weise konzentriert sich die herkömmliche Trajektorieninferenz allein auf die Genexpression und lässt andere Modalitäten wie wertvolle Zeitpunktinformationen, die in neu entstehenden Einzelzelldatensätzen verfügbar sind, außer Acht; Methoden, die diese verschiedenen Datenansichten einbeziehen, lassen sich nicht auf andere Modalitäten verallgemeinern, können nicht kombiniert werden oder sind nicht skalierbar. Um diese Einschränkungen zu überwinden, stelle ich als zweiten Beitrag CellRank 2 für eine einheitliche Trajektorieninferenz und Schicksalskartierung vor. In diesem Rahmen bezieht CellRank 2 pseudotime- und Potenzialschätzungen unabhängig von ihrer Quantifizierung ein, um das zelluläre Schicksal konsistent abzuleiten. Für Datensätze, die Zeitpunktinformationen enthalten, stelle ich ein optimal transport basiertes Trajektorieninferenzschema vor, um Inter- mit Intra-Zeitpunktinformationen zu kombinieren. Alternativ führen metabolische Markierungen zeitliche Informationen ein, indem sie neu transkribierte mRNA-Moleküle charakterisieren; ich zeige, wie diese Informationen die Schätzung zellspezifischer Transkriptions- und Abbauraten ermöglichen. Für jede Methode beschreibe ich entsprechende Analysen, um ihren Nutzen zu validieren.

Beide Beiträge verbessern und verallgemeinern die Inferenz von Trajektorien und zellulären Zustandsänderungen. Ihr modularer, skalierbarer und vielseitiger Aufbau erleichtert die Entdeckung neuartiger Biologie, garantiert weitere Modellverbesserungen und ermöglicht die Einbeziehung neu entstehender Datenmodalitäten.

Contents

1. Introduction

Multi-cellular organisms such as humans emerge from a single cell, proliferating and developing into trillions of cells in homeostasis[1]. These cells specialize in different tasks that define cell states, and each cell state results from different mechanisms such as molecular signaling and gene regulation during differentiation. The way these complementary cellular forms exactly arise is, however, largely unknown[2].

Similarly, dysregulation of gene expression in differentiation processes leads to diseases such as cancer, one of the most prevalent and deadliest diseases of our time; nearly 40 percent of men and women will be diagnosed with cancer during their lifetime[3]. Yet, the underlying biological processes leading to cancer are understood poorly, even though medical advances have led to substantially improved survival rates and prolonged life expectancy of patients. Similarly, the regulatory mechanisms leading to neurodegenerative disorders like dementia, Alzheimer's and Parkinson's disease, or inflammatory conditions such as inflammatory bowel disease or arthritis are largely unknown even though these diseases affect substantial portions of the population.

Understanding the biological processes underlying normal development, disease progression or reprogramming has been an active research field, its beginning dating back thousands of years[4,5]. Where original models were based on philosophy, scientific concepts took over, eventually leading to the discovery of the cell as an entity by Robert Hooke in 1665[6]. Since then, the focus has shifted to studying cellular state change to map genotypes to phenotypes.

Cells form the basis of biological processes like normal development and disease mechanisms and constitute larger structures such as tissues and organs. Although these cells perform a common task and influence each other through external stimuli, they are still physically distinguishable from each other. Similarly, large, cohesive cellular groups and structures emerge gradually through differentiation dynamics; since these processes are asynchronous, a single snapshot view contains different cell states, thereby enabling the study of the overall process. The discrete nature of cells allows studying them experimentally through dissociation, their asynchronous differentiation allows analyzing their state changes along dynamic processes from snapshot data. This publication-based dissertation focuses on modeling and describing the processes of cell differentiation and fate priming computationally.

1.1. Single-cell sequencing

Traditional approaches for studying dynamics in cell biology relied on bulk sequencing assays (Figure 1.1a): Pooled cell populations are sequenced, thereby revealing average features of cells; bulk RNA sequencing (RNA-seq)[7–10], for example, measures gene expression, and bulk ATAC-seq[11–14] chromatin accessibility. While bulk sequencing assays yield high feature expression at low measurement noise due to pooling, they lack sensitivity - rare but essential gene expression patterns of infrequent cell types are masked

by other, more abundant genes, for example. As such, data analysis based on bulk sequencing lacks resolution and may miss cellular heterogeneity.

In contrast to bulk RNA-seq experiments, bright-field[15] or fluorescence microscopy[16,17] protocols offer single-cell and even subcellular resolution[18–21] (Figure 1.1b). However, this approach has two limitations: Low throughput and low capture rate. As such, microscopy-based experiments cannot analyze large tissues or organs; the number of cells exceeds the protocols' processing capabilities, and the small number of targeted genes does not reveal the underlying cellular heterogeneity. Pre-selecting genes to target is an additional drawback as it complicates finding novel genes relevant to a given process.

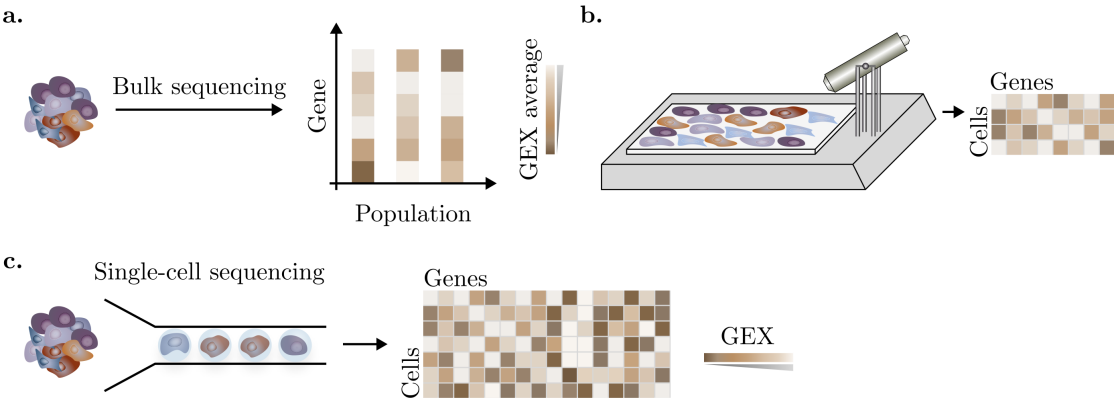

Figure 1.1.: Experimental techniques for characterizing and studying cells. **a.** Bulk sequencing records averaged population features such as gene expression (GEX) for bulk RNA sequencing[7–10]. **b.** Bright-field[15] or fluorescence microscopy[16,17] provides single-cell resolution but is low throughput. **c.** Single-cell sequencing measures cell-specific features like GEX for scRNA-seq at scale.

Biological tissues and processes consist of different sub-processes and states[22]; this variability manifests itself in distinct roles within the system[23,24] or causing alternative cellular fate[25], for example. To study this cellular heterogeneity masked by bulk sequencing assays at single-cell resolution and scale, single-cell sequencing protocols have been developed[26,27] (Figure 1.1c, Figure 1.2). The power of such methods stems from their ability to dissociate and isolate individual cells in droplets[28–30] or wells[31–37], associate cell-specific barcodes to each transcript, and sequence all transcripts together in a single experiment, using the barcodes to assign transcripts to individual cells. Single-cell RNA sequencing (scRNA-seq), for example, captures the transcriptome at single-cell resolution based on standard sequencing protocols[26]. Commercialization of protocols has simplified the experimental setup and decreased sequencing cost, leading to an increase in single-cell datasets both in size and abundance[38].

Similar to scRNA-seq, related omics protocols measure other quantities of a cell such as the DNA[39], DNA methylation patterns (scM&T-seq[40], scMT-seq[41], scNMT-seq[42]), surface proteins (CITE-seq[43]), chromatin accessibility (scATAC-seq[27]), histone modifications (scCUT&Tag[44]), and perturbation effects (Perturb-seq[45], ECCITE-seq[46], compressed Perturb-seq[47]) (Figure 1.2). Additionally, experimental protocols exist to measure multiple views in the same cell, such as gene expression and cell surface proteins (CITE-seq[43]), gene expression and intracellular proteins (SPARC[48]), or gene expression and chromatin accessibility[49–52]; newer technologies capture even more modalities from the same cell[53,54], and spatial assays measure cell features and location[55–58]. Alternative modalities identifiable in standard scRNA-seq experiments are nascent and mature mes-

senger RNA (mRNA) molecules, two quantities that offer directed, dynamic information by considering their biological causal dependency. Similarly, time-resolved sequencing data based on experimental time points or metabolic labels offer priors for studying cellular change.

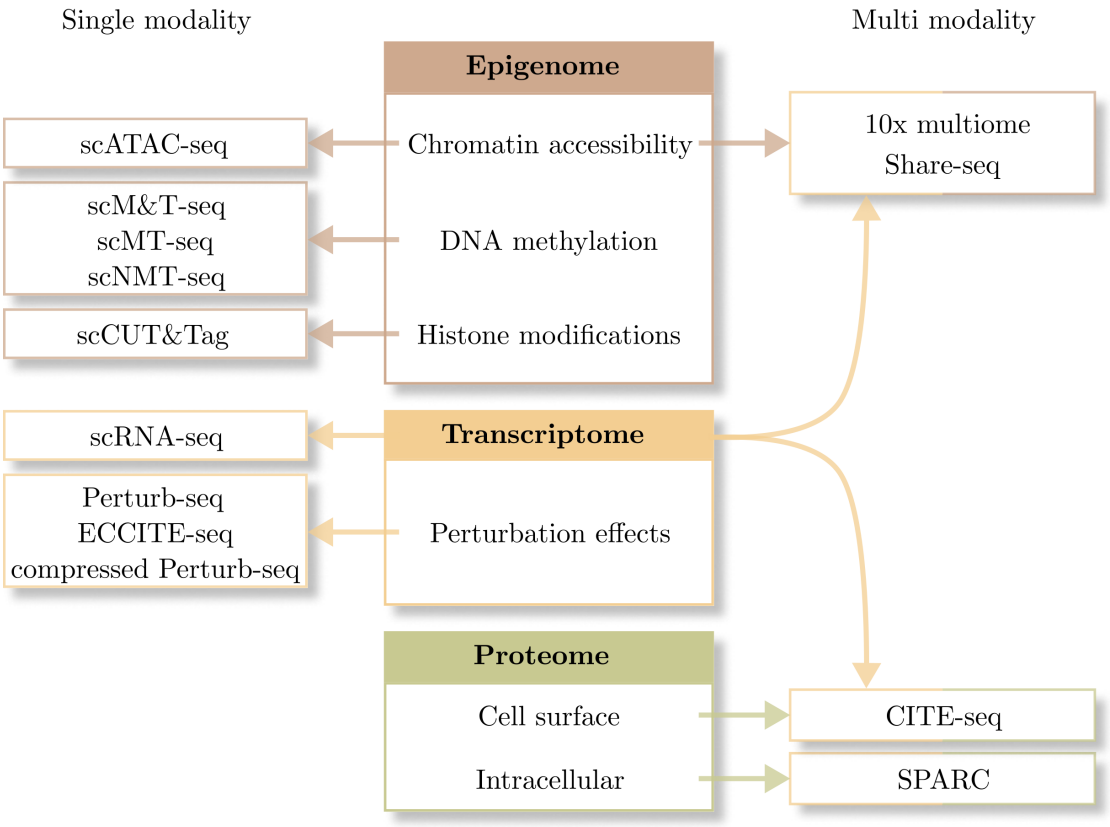

Figure 1.2.: Single-cell sequencing protocols record different cellular features. Single-cell sequencing experiments quantify epigenetic, transcriptomic, and proteomic data. The assays provide information about a single modality (left) or multiple ones (right).

1.2. Analysis of single-cell data

Statistical models describe observations in terms of their distribution, properties thereof or intrinsic differences. For bulk sequencing measurements, for example, generalized linear models or mixed effect models reveal differential expression[59,60]; similar approaches exist for single-cell observations[61,62]. Importantly, such approaches require prior annotation and formulated hypotheses, and increased dataset sizes require alternative machine-learning methods scaling to millions of data points. Modality-specific tools provide the necessary means to analyze measurements in a data-driven fashion to enable preprocessing and analysis in a consistent, reproducible framework: Scanpy[63] and Seurat[64–68] handle scRNA-seq measurements and Squidpy[69], Seurat[66,68] or Giotto[70] the transcriptome in spatial context; relatedly, muon[71], ArchR[72] or Signac[73] enable studying chromatin accessibility data, and muon[71] and Seurat[66,67] cell surface proteins.

Single-cell RNA sequencing measures the whole transcriptome at high throughput, yielding high dimensional datasets - both in the number of observations and features - but suffers from data sparsity. Conventional statistical and machine learning methods, there-

fore, oftentimes do not work out of the box. Thus, data-specific methods have been developed for analyzing scRNA-seq measurements[74]; example applications include reducing the dimension of the data[75–78] and interpreting corresponding latent dimensions[78–80], identifying differentially expressed genes, clustering cells with similar transcriptomic profiles[81,82], or integrating datasets while removing batch effects[76,83–86]. These applications focus on extracting statistically relevant information from sequencing data but do not recover changes along a dynamically unfolding process.

Reconstructing the cellular state change during differentiation processes from classical single-cell sequencing protocols is challenging as experiments are destructive by nature. Although recent experimental advances capture the transcriptome sequentially in the same cell, the techniques are experimentally demanding and do not yet scale[87–89]. Instead, computational methods recover the underlying dynamic process, leveraging the fact that biological processes unfold asynchronously. These methods focus on recovering the state change along differentiation trajectories - the field of trajectory inference (TI) - and quantifying cellular fate from single-cell data (Figure 1.3).

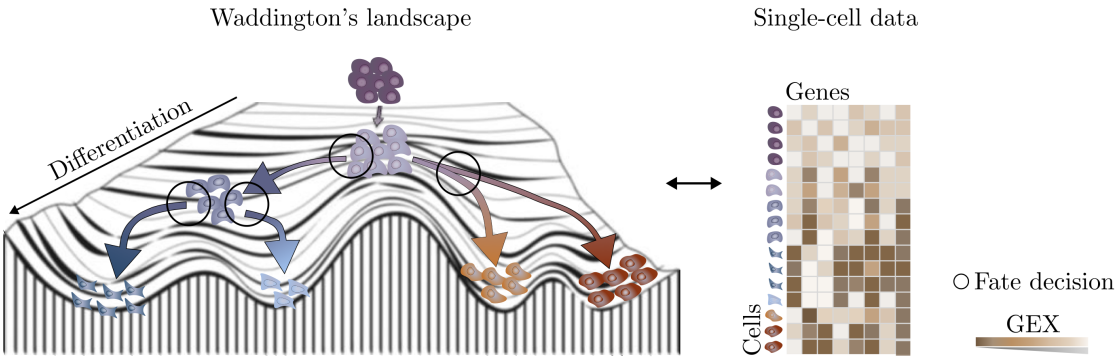

Figure 1.3.: Trajectory inference recovers cellular paths along the differentiation landscape. Waddington's landscape[90] models cell differentiation as a surface of changing potential: Less mature states are unstable and have high potential, whereas more differentiated cells exhibit smaller potential. Intuitively, this setup can be compared to a sphere rolling down a mountainous landscape. Trajectory inference methods align observations from single-cell sequencing experiments along this landscape to infer a differentiation order and positions of fate decisions.

Trajectory inference through pseudotime and stemness scores

A multitude of methods have been proposed for inferring trajectories, all assuming cells change gradually. Early methods mapped the discrete, observed cellular states aligned along the differentiation process onto a continuous domain and ranked cells relative to each other via a so-called pseudotime[91–97]. To map the differentiation direction, these methods rely on a pre-defined cell marking the beginning of the process. Following, popular methods estimate the distance traveled along the phenotypic manifold through diffusion[93] or model trajectories as random walks[97]. While such root states are apparent in well-characterized systems such as hematopoiesis[91,93,97], skeletal muscle differentiation[92,96], or the olfactory epithelium[96], they are not as easily, if at all, identifiable in more challenging settings such as disease or reprogramming.

Differentiation potentials can be used to infer cell trajectories, instead of assuming a pre-defined start state. Algorithms estimating such potentials rely on the paradigm of Waddington's landscape[90]: Less mature cells exhibit a high potential that decreases as

they differentiate into specialized cell types. Related methods estimate the entropic state of a cell[98] or assume a pre-defined gene expression structure along differentiation[99], for example. Although such frameworks circumvent the problem of selecting a root cell, they introduce additional assumptions that may not hold in every system, and they do not predict the future expression of a cell.

State change inference through temporal coupling

Relating temporally coupled cell states or views offers an alternative approach to model state changes. This modeling paradigm has two advantages over pseudotime and differentiation potentials: First, fundamental physical and biological processes define directionality by relating biological stages of mechanistic and temporal relationships. Second, this directionality naturally allows for predicting cell states at different time points.

RNA velocity[100,101] relates nascent and mature mRNA through a mechanistic model of splicing dynamics[102] to infer trajectories and predict the future expression of a cell. This concept is especially intriguing as standard scRNA-seq protocols capture unspliced and spliced mRNA[100,103–106]. However, the noisy, high-dimensional nature of the data is one of the challenges faced by RNA velocity; metabolically labeled mRNA provides complementary state estimates that can be linked through a mechanistic model in a similar fashion[107].

Different experimental time points offer an alternative way to estimate the direction of change. In this case, the challenge lies in matching cells at earlier time points with putative progenitors at later stages. Dedicated methods estimate population dynamics[108] or assume that earlier time points correspond to earlier stages of the biological process and match cells across time points with optimal transport (OT) to recover the underlying vector field[109–111]. Specifically, OT assigns each cell from a given time point its likely future states in the consecutive time point by minimizing an objective function to match most similar cell tuples (Appendix E). Here, typical challenges include choosing the spacing of experimental time points to capture small gradual changes or latent effects; epigenetic differences translating into transcriptomic dissimilarity are one possible latent variable[112].

Cellular fate mapping

Trajectory inference reconstructs lineage relationships for single-cell data where lineage commitment is a gradual process: Pluripotent stem cells can differentiate into any lineage, and commitment to these lineages precedes a cascade of gene regulatory events like upregulation of GATA 1 for the erythroid lineage of hematopoiesis[113–115], for example. Genes regulating fate priming are known as lineage drivers, and computational methods disentangling such decisions assign each cell a fate probability based on cell statistics - a concept known as fate mapping.

Different approaches rely on different data views for cellular fate mapping: Slingshot[96] estimates pseudotime values and lineage weights simultaneously, Palantir[97] assigns lineage probabilities using Markov chains after pseudotime inference. Similarly, CellRank[116] combines RNA velocity estimates with Markov chain theory to model lineage priming. For time-resolved scRNA-seq data, Waddington OT[109] infers ancestor and progenitor populations from optimal transport with pushforward and pullback operators, respectively. Dynamo[107] uses the temporal information from metabolic labels to study fate decisions based on optimal transition paths derived from an estimated velocity field.

Overall, methods to map cellular fate exist but are tied to specific concepts or data views.

Gene regulatory inference and perturbation modeling

Gene regulatory networks (GRNs) govern cell differentiation and fate priming, and inferring the GRN of a process elucidates relevant genes and gene programs. Understanding how regulation changes from a healthy to a diseased state, for example, can provide putative drug targets for treatment or even prevention. The single-cell community has developed tools for inferring GRNs[117–119] and modeling perturbation effects[120–122]; although such models give different outputs, they share a common goal: Understanding how gene regulation shapes a given process. To this end, Pando[118], for example, infers GRNs by computationally connecting transcription factor-binding site parts with the gene expression of their target genes. CellOracle[123] describes GRN structures and relationships implicitly by perturbing transcription factors *in silico*. Alternatively, to model perturbation responses directly, scGen[120], CPA[121] and CellOT[122] rely on VAEs and OT and leverage data from single-cell perturbation experiments. However, similar to other single-cell methods, data sparsity and integrating multi-modal data hamper the efficacy of GRN inference and perturbation response prediction. Additionally, benchmarking GRN inference is challenging as ground-truth data does not exist.

1.3. Research question and scope of dissertation

Research question

This thesis addresses the problem of modeling splicing dynamics with deep generative neural networks, and inferring trajectories independent of a specific data view for fate mapping with single-cell data. These contributions are necessary since existing TI methods leave room for improvement even though the field of TI has celebrated great success:

1. RNA velocity has emerged as a bottom-up modeling approach for splicing dynamics to infer dynamic, directed state change. However, original schemes for inferring RNA velocity do not quantify estimation uncertainty, fail to provide a metric of applicability, and are not easily generalizable to more complex but accurate mechanistic models[124,125]. Thus, these methods lack interpretability and quantitative evaluation metrics, challenging their correct usage on real-world data, and cannot facilitate more accurate model descriptions.

2. Similarly, methods for analyzing different data modalities are tied to them and do not generalize to complementary and newly emerging data views. For example, CellRank focuses on RNA velocity[116], Waddington-OT on experimental time points[109], and dynamo on metabolic labeling data[107]. The single-cell field lacks a unified framework for mapping cellular fate in a modality-agnostic fashion as a result. Such a framework is essential, however, to include newly available modalities easily and derive cellular fate based on novel methods that model differentiation processes.

To address these open challenges, this dissertation discusses a two-fold research question:

Q1: How can RNA velocity be inferred in an uncertainty-aware fashion that enables increased interpretability and facilitates model flexibility?

Q2: How can different data views be used to map cellular fate in a unified fashion to rely on context-dependent, orthogonal information for improved trajectory inference?

I focused on these questions in published work presented in Sections 3.1 and 3.2. Specifically, to answer questions Q1 and Q2, I split my proposed solution into the following distinct steps, building upon and complementing each other:

Uncertainty-aware and scalable RNA velocity inference emerges as a crucial aspect with the growing size and complexity of scRNA-seq datasets. Specifically, traditional inference methods do not provide parameter uncertainty and rely on classical optimization routines, proving challenging when applied to datasets containing millions of cells. Recent advances in variational inference (VI) offer a framework for scalable, uncertainty-aware parameter inference (Appendix B). In this work, I show how to formulate the RNA velocity inference problem in a variational inference context (Section 2.4 and 3.1) to solve it with the help of deep learning architectures and answer Q1.

Metrics for assessing the applicability of RNA velocity are essential to ensure correct application of the method and elucidate model failure. The assumptions of original RNA velocity inference approaches are restrictive and often violated in real-world data and, therefore, necessitate such an evaluation scheme. Specifically, the used approaches require an explicit, structural dependency between unspliced and spliced counts; proposed frameworks, however, do not provide ways of quantifying if their modeling assumptions are valid for a given dataset. Addressing this lack of interpretability will facilitate the correct usage of RNA velocity. In this thesis, I present a permutation-based metric to evaluate if the required dependency between nascent and mature RNA is satisfied (Section 3.1); evaluating the distribution of this metric on positive and negative control cases allows assessing and comparing the quality of new datasets.

Flexibility of mechanistic models describing splicing dynamics is needed for more accurate descriptions of the underlying principles, but current state-of-the-art inference methods are tied to a single mechanistic model; this dynamic model is incomplete as it omits other aspects of the dynamical process. In this thesis, I discuss how the VI-based model generalizes naturally to more complex descriptions; I previously exemplified this feature with a splicing model that describes transcription as a time-dependent process[126] (Section 3.1).

Quantitative metrics for comparing models of splicing dynamics and inference approaches do not exist. Instead, comparisons to assess model fit predominantly rely on two approaches: Projecting high-dimensional velocity fields onto two-dimensional data representations, and high consistency between velocities of transcritomically similar cells. However, visual representations lack statistical and quantitative power and are sensitive to various parameters[100,125]; assuming similar velocities is valid in unidirectional processes but not during branching events such as fate priming. In this dissertation, I highlight a principled approach to compare models and methods based on the cell cycle for which the developmental direction is known a priori and experimental techniques establish a ground truth cell order[127,128] (Section 3.1).

Consistent trajectory inference and cellular fate mapping from pseudotime and stemness estimates offers an alternative approach to RNA velocity when violated assumptions prevent its applications. While RNA velocity inference may fail, alternative methods for recapitulating biological processes, such as pseudotime or stemness potentials, may work remarkably well[91,93,97,99]; but there is no consistent way of inferring trajectories and cellular fate based on these alternative quantities, and existing methods may not scale to atlas-sized datasets. To overcome these limitations, I propose a consistent, method-agnostic approach for bridging the gap between scalar pseudotime and stemness scores and cellular fate mapping and improved model formulations that scale to millions of cells (Sections 2.6 and 3.2).

Inter and intra-time point dynamics occur during asynchronous biological processes. Although methods that infer trajectories based on experimental time points exist, they omit intra-time point information. As a result, they map cellular change in a discrete fashion which makes studying gene expression change on a continuous domain impossible. In this thesis, I extend optimal transport-based trajectory inference to include intra-time point information, and infer initial and terminal states automatically; this procedure is consistent with the approach for pseudotime and developmental potential. Additionally, I outline how to infer a real-time-informed pseudotime that enables continuous fate mapping and models of gene expression change (Sections 2.6 and 3.2).

Modeling time-dependent rates of splicing dynamics is essential to an accurate description of the underlying biological process. Current models for inferring splicing mechanisms from metabolically labeled data employ and estimate cell-specific rates in a post-hoc fashion, though[107]. In addition, their method for inferring cellular fate and putative driver genes is deterministic, thereby ignoring model uncertainty and the stochastic nature of biology. As part of my consistent fate mapping framework, I present an estimation paradigm that infers time-dependent rates for each cell and recovers cellular fate and known lineage drivers more faithfully than competing approaches (Sections 2.4 and 3.2).

Scope of dissertation

To answer the posed research question and resolve existing challenges the single-cell community faces, I group my contributions into two: Deep generative modeling of transcriptional dynamics[126] and unified fate mapping in multiview single-cell data[129].

The first part of my contributions addresses RNA velocity inference (Q1):

Publication 1: ***Deep generative modeling of transcriptional dynamics for RNA velocity analysis in single cells***

Modeling splicing dynamics through deep generative modeling for RNA velocity analysis with VI is a novel, uncertainty-aware estimation of RNA velocity based on variational autoencoders (VAEs). The model improves parameter estimation compared to previous approaches, quantifies estimation uncertainties, and offers metrics for evaluating the applicability of RNA velocity inference for a given dataset. This publication attempts to answer Q1 and Section 3.1 presents it in greater detail; Section 2.4 discusses theoretical aspects of the corresponding deep learning model. I co-lead this study.

Additional publication 1: ***A Guide to Trajectory Inference and RNA Velocity***

With the rise of scRNA-seq data as a powerful experimental technique to study cellular heterogeneity and dynamic processes, dedicated computational methods have been developed. In this work, we reviewed the concepts of pseudotime and RNA velocity. I derived the mathematical principles underlying existing RNA velocity methods, highlighted and discussed their limitations and challenges, and showcased a typical RNA velocity analysis workflow. I am the lead author of this study.

Additional publication 2: ***Best practices for single-cell analysis across modalities***

Single-cell sequencing protocols have evolved to allow quantifying different data modalities such as transcriptomics, cell-surface proteins, or chromatin accessibility. As a result, many computational workflows and methods have been developed to analyze the corresponding data. In this work, we summarized independent benchmarks assessing existing approaches if such comparisons exist and reviewed and summarized existing methods otherwise. I contributed to the effort by discussing pseudotime and RNA velocity inference and complemented these theoretical aspects with best-practice tutorials. I am a supporting author of this study.

The second part of my contributions covers fate mapping with single-cell data in a unified and data-view agnostic fashion (Q2):

Publication 2: ***CellRank 2: unified fate mapping in multiview single-cell data***

CellRank 2 models cellular state changes probabilistically by inferring cell-cell transition probabilities via *kernels* and analyses them with *estimators*. This modular design makes CellRank 2 a data-agnostic framework, highlighted by using a pseudotime for hematopoiesis, stemness potential for embryoid body development, real-time information from experimental time points for mouse embryonic fibroblasts and pharyngeal endoderm development, and metabolic labeled RNA to study regulatory mechanisms in mouse intestinal organoids. The framework scales to atlas-sized datasets, generalizes to new, emerging data views, and outperforms competing methods. This publication presents a solution to research question Q2; Section 2.6 covers related methods, and Section 3.2 results. I co-lead this study.

Preprint 1: ***Plasticity of Human Microglia and Brain Perivascular Macrophages in Aging and Alzheimer's Disease***

Alzheimer's disease (AD) is a progressive brain disorder and the most common type of dementia. Yet, little is known about the mechanisms leading to the malfunctions of the underlying process, and there is no consistent taxonomy describing the heterogeneity and plasticity of microglia and perivascular macrophages, immune cells specific to the brain. To establish a common taxonomy and study the role of identified subtypes during AD progression, this work analyzed two independent, demographically diverse cohorts; the first included 157 donors, the second 1470. I contributed by inferring and analyzing a disease-stage-informed pseudotime relying on concepts developed for CellRank 2, and helped write the manuscript. I am a second author of this study.

Preprint 2: ***Modeling Single-Cell Dynamics Using Unbalanced Parameterized Monge Maps***

Optimal transport has emerged as a powerful tool for studying non-steady-state systems evolving over time. To study such systems, samples at different experimental time points are sequenced; OT matches cells across time points to describe the underlying process with couplings traditionally being inferred by solving a minimization problem with classical optimization routines. Recent advances in machine learning reformulate the problem to solve it with neural networks, instead, but fail to account for asymmetric shifts in cell state distributions, routinely present in biological processes. In this work, we proposed Unbalanced Parametrised Monge Maps to overcome this inherent limitation. I contributed by benchmarking different modeling approaches with CellRank 2's inference agnostic framework and writing the paper. I am a supporting author of this study.

The software developed for my contributions is open-source and, thus, available to the single-cell community. My contributions and the analysis of single-cell data in general rely on and are embedded in a greater collection of computational tools and data structures.

Additional publication 3: ***The scverse project provides a computational ecosystem for single-cell omics data analysis***

This correspondence presents the scverse ecosystem, a multi-institutional open-source software project. The work outlines how data storage and analysis for single-cell omics data is addressed. I contributed to this effort through software implementations for and maintenance of Python packages. I am a supporting author of this study.

2. Methods

The advancements in single-cell biology have been both experimental and methodological. On the experimental side, protocols have been proposed and optimized to measure different views of cells. Given this sparse and high-dimensional data, machine learning methods have been applied, adapted, and developed for use cases specific to single-cell biology.

This chapter is divided into three main parts: Section 2.1 discusses sequencing experiments relevant to data and methods presented in this thesis. Following, Section 2.2 outlines computational aspects of analyzing scRNA-seq data, followed by approaches for recovering how biological processes unfold: Section 2.3 discusses trajectory inference, Section 2.4 RNA velocity, and Section 2.5 methods for time-resolved scRNA-seq data.

2.1. Single-cell sequencing protocols

The ability to measure the transcriptional profile of individual cells at scale has enabled the study of cellular heterogeneity in health and disease. Different protocols offer complementary views of cells, including the transcriptome with scRNA-seq[26,28–37], chromatin accessibility with scATAC-seq[27], and cell surface proteins through CITE-seq[43]. Here, I focus on scRNA-seq and strategies for metabolically labeling newly transcribed mRNA; both types of experiments produce data analyzed with the computational methods I have developed.

Single-cell RNA-sequencing

Single-cell RNA-seq workflows can be grouped into three consecutive stages: (1) Cell suspension through dissociation and isolation of cells into droplets or wells, (2) library construction, and (3) sequencing (Figure 2.1a). Even though experimental workflows are optimized to capture the underlying biological process in its true form, some experimental artifacts may pertain. Dedicated computational methods mitigate experimental errors[130–132], such as empty droplets (wells) or doublets, *i.e.*, multiple cells captured in the same droplet (well).

Tissue dissociation forms the first step of scRNA-seq experiments as single-cell sequencing protocols operate on isolated cells, obtained through the dissociation of samples[133,134]. Isolating individual cells is an intricate process, leading to artifacts and reduced data quality if done improperly; possible ramifications include groups of cells sequenced together[130], dying cells[135,136], or extracellular debris[131,132]. Problem-specific analysis methods exist to remove such artifacts but entail data loss. Thus, proper tissue dissociation is key to ensuring optimal cell suspension.

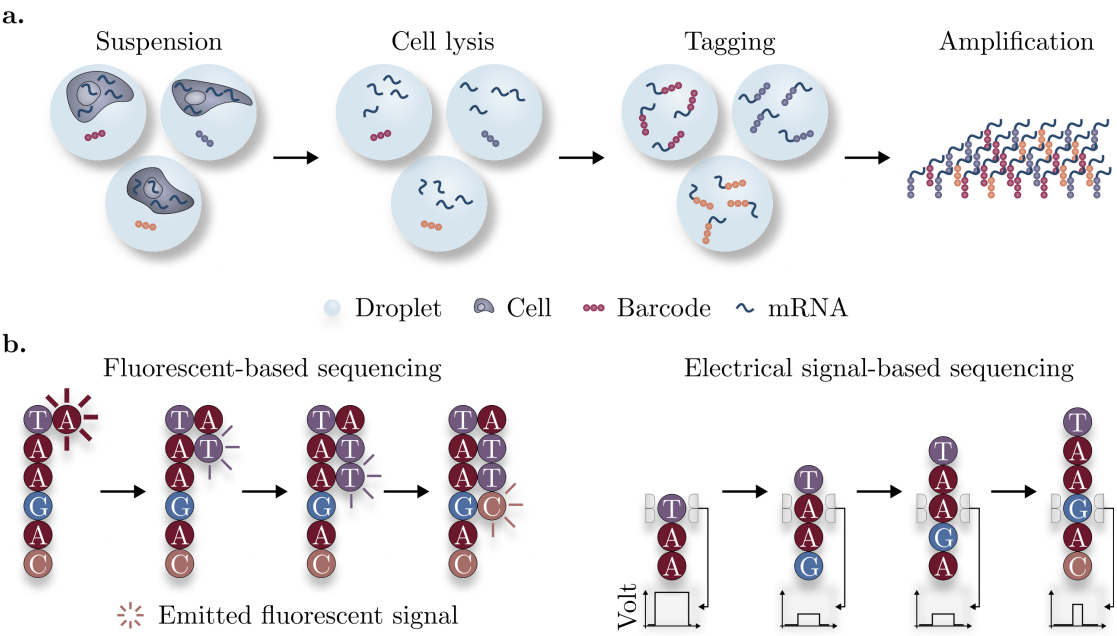

Figure 2.1.: Single-cell RNA-sequencing workflow. a. For library construction, assays load cells into droplets (wells), remove the cell membrane through lysis, and attach cellular barcodes to transcripts. Following, mRNA molecules are reverse transcribed into cDNA and amplified by PCR. **b.** NGS estimates gene abundance from fluorescent (left) or electrical signals (right).

Tissue dissociation is based on chemical[137–144] or mechanical approaches[145–147] or a mixture of the two[148,149] and grouped into two strategies - warm and cold tissue dissociation[150]. Warm tissue dissociation relies on enzymes operating optimally at warmer temperatures compared to cold tissue dissociation. Employing these methods too aggressively to maximize the number of cells for sequencing can trigger stress responses, reflected in altered gene expression patterns, however[151,152]. As such, different experimental setups require fine-tuning workflows for an optimal trade-off between isolation and yield.

Library construction follows cell suspension (Figure 2.1a). This stage of the experiment loads cells into droplets (wells) and collects RNA fragments in the so-called sequencing library. As a first step, cell lysis removes the cell membrane, making intra-cellular mRNA molecules accessible. Cellular barcodes mark these transcripts in each droplet (well) to assign captured transcripts to cells after sequencing; guaranteeing statistical power requires amplifying the molecules. The mRNA molecules themselves are unsuitable for amplification, however, as they are readily degraded by omnipresent RNases, for example[153,154]. Transcripts are, thus, reverse-transcribed into more stable complementary DNA (cDNA) that can be amplified with polymerase chain reaction (PCR)[155–157]; unique molecular indentifiers (UMIs) label captured molecules to mitigate cDNA amplification noise.

Next-generation sequencing (NGS) processes the amplified cDNA to quantify which and how many RNA molecules are present (Figure 2.1b). Compared to the labor-intensive Sanger sequencing[158,159], NGS has high throughput as parallel sequencing reactions are only spatially separated, not physically. The sequencing reactions are captured

either by fluorescent or electrical signals: Fluorescent technologies rely on detecting luminescent signals released during the correct addition of a nucleotide during sequencing by synthesis[39,160,161]. Alternatively, nanopore approaches detect changes in an electrical field induced by RNA molecules passing through nanopores[162,163]; changes in the ionic current match the nucleotide sequence passing through the biosensor.

Metabolic labeling of mRNA

Traditional single-cell sequencing protocols are destructive by nature, capturing the state of a cell only once instead of tracking its dynamic change over time. Sequencing samples at different experimental time points offers a promising solution to assess the cellular progression nonetheless, but, in most cases, experimental time points lie at least six hours apart. This setup thus struggles with capturing rapid transcriptomic changes on the time scale of minutes to hours since the median mRNA half-life in mammalian cells is relatively long[164,165]. Additionally, traditional scRNA-seq protocols also fail to characterize RNA processing and cannot differentiate the steps thereof. To overcome these limitations, methods for metabolically labeling newly synthesized mRNA molecules have been developed first for bulk[166–173], then for single-cell sequencing[127,174–176].

The fundamental principle of metabolic labeling lies in incorporating chemical tags into newly synthesized RNA; tagging is achieved by exposing cells to nucleoside analogues that are taken up, phosphorylated, and incorporated into nascent RNA. Single-cell labeling approaches convert the introduced nucleotides into a different organic molecule - guanine into adenine and uridine into cytosine - identified during sequencing; newly transcribed RNA, thus, characterizes itself by the presence of corresponding substitutes. Correctly identifying nascent RNA is challenging due to low incorporation frequency and sequencing errors; computational methods mitigate this effect by robustly estimating proportions of old and new RNA[172,177]. For a more in-depth discussion of metabolic labeling approaches, I refer to dedicated reviews[44,178].

Single-cell sequencing protocols yield high-dimensional, sparse representations of biological samples. Manual analysis of the data is, thus, impossible. Instead, computational frameworks provide the necessary tools.

2.2. Canonical analysis steps for single-cell RNA sequencing data

The rapid development and advancement of single-cell assays have been accompanied by computational advances to analyze the generated datasets[74]. The corresponding methods include both techniques known from traditional data analysis and adapted or newly developed strategies; these workflows address generic aspects like data preprocessing and problem-specific solutions for data integration or trajectory inference, for example. This section reviews data processing common to any scRNA-seq analysis, followed by sections focusing on approaches for specific problems. For a more complete overview of scRNA-seq data analysis, I refer to previously published reviews on best practices[63,179–181].

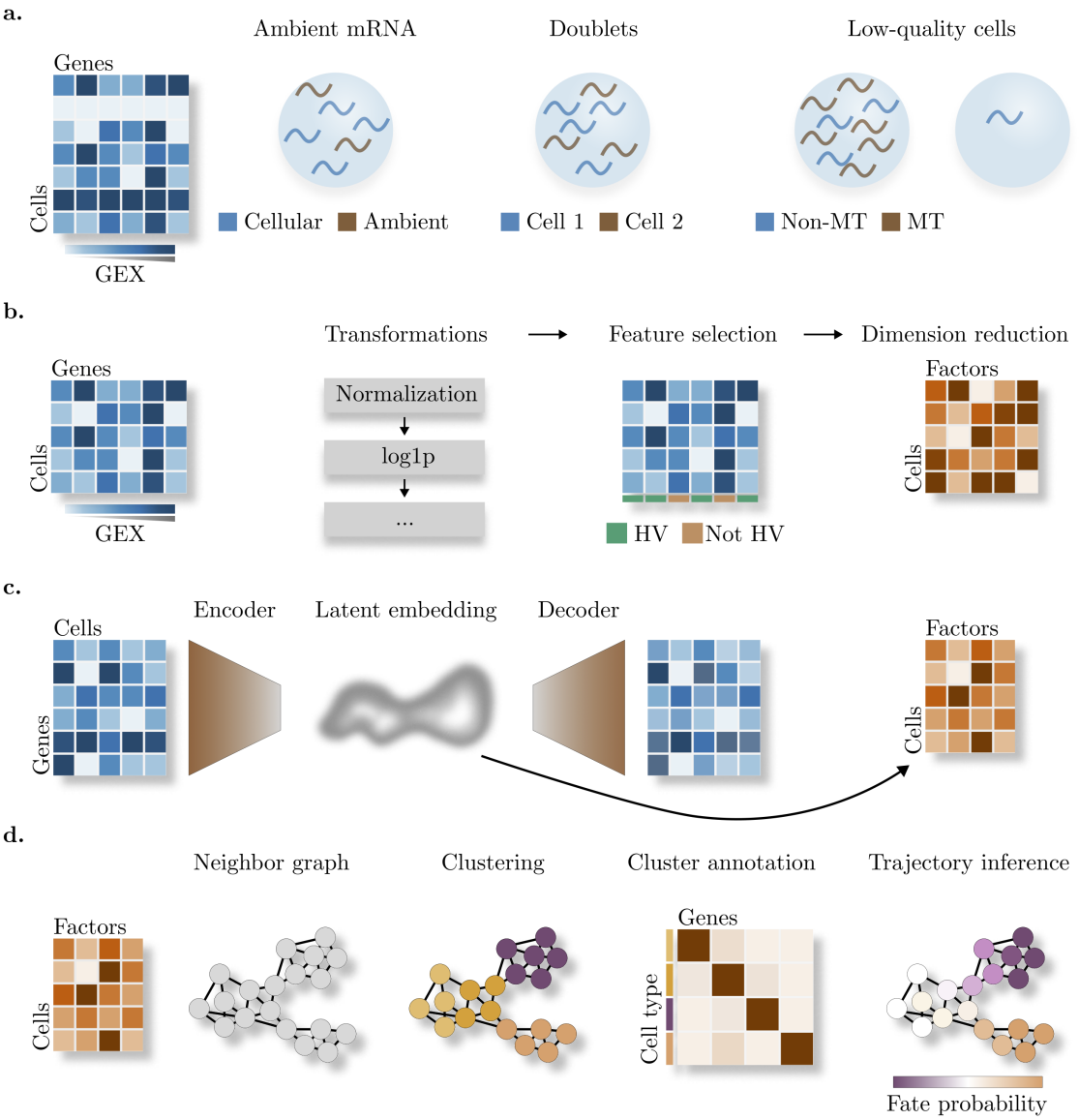

Figure 2.2.: scRNA-seq data processing workflow. a. Quality control identifies and removes or corrects measurements from the count matrix generated by scRNA-seq experiments. These steps include removing ambient RNA and detecting doublets or low-quality cells. **b.** Following quality control, data transformations prepare the compressing features into a lower dimensional space; common transformations include data normalization[182] and log1p transformation. Next, highly variable genes[64,183–187] (HV: highly variable) form the basis of dimensionality reduction techniques such as non-negative matrix factorization[78] or PCA[75]. **c.** Deep learning-based preprocessing workflows operate directly on the raw count matrix to recover a latent representation of the data. Autoencoders[76,77] are a common approach to project the data into a latent embedding via an encoder neural network; a decoder reconstructs the data to define a loss function for model training[76]. **d.** Nearest neighbor graphs approximate the phenotypic manifold and help reveal clusters[82] associated with cell types[66,85,188–191]. Similarly, the latent data representation and neighbor graph enable inferring cell trajectories and fate[192,193].

Data preprocessing

Data preprocessing is a common aspect of any machine learning pipeline. For scRNA-seq data, cells and genes are first filtered based on quality metrics (Figure 2.2a): Likely compromised cells are removed, ambient mRNA molecules estimated and removed, and doublets identified. A compromised cell membrane caused by cell death, for example, leads to a small number of detected genes, low count depth, and a high fraction of mitochondrial gene expression; quality metric-based filtering removes corresponding observations from further analyses. In the case of droplet-based sequencing assay, droplets contain ambient mRNA, inflating the estimate of cellular mRNA molecules present in the respective cell. Methods like SoupX[132] or DecontX[131] estimate and remove cell-free mRNA. Finally, doublets - multiple cells assigned the same barcode and sequenced - must be identified and removed. Dedicated benchmarks compare corresponding methods[194]. In the following, I refer to the data matrix obtained after quality control as the *count matrix*.

Following quality control, feature selection or transformation reduces the dimensionality of the count matrix to reveal the core information it contains; the fact that dynamical processes occur on a low-dimensional manifold justifies reducing the dimensionality without significant loss of information[195]. Dimensions are either reduced by successively applying transformations (Figure 2.2b) or in an end-to-end fashion (Figure 2.2c). Common transformations consist of count normalization[182], log-transformation, selecting highly variable genes[64,183–187], removing unwanted sources of variation and batch effects[83,196], and further dimensionality reduction through methods like principal component analysis (PCA)[75]. Part of the success of single-cell biology stems from implementing such preprocessing steps in open-source software libraries like Scanpy[63] or Seurat[64–68].

Compared to manually selected data transformations, end-to-end methods directly estimate a low-dimensional representation of the count matrix, possibly filtered for highly variable genes first; two popular examples of this modeling paradigm are DCA[77] and scVI[76], relying on autoencoders and variational inference to estimate a low-dimensional latent space. In brief, these approaches rely on non-linear function approximation with neural networks to map the high-dimensional count matrix into a low-dimensional embedding of latent factors. More recent methods extend these frameworks to interpretable latent space components[79,80]; similarly, other architectures address specific challenges and types of data, such as population-level single-cell atlases[85] like the human lung[23] or embryonic limb cell[197] atlas generated as part of the Human Cell Atlas Project[198]. Appendix B discusses the theory of variational inference and the scVI model in greater detail.

Data clustering and visualization

Following dimension reduction, data clustering and quantifying differentially expressed genes in these clusters allow for interpreting them as different cell states[66,85,188–191] (Figure 2.2d); similarity-based representations such as k-nearest neighbor (kNN) graphs form the basis for clustering: kNN graph construction first computes the k nearest neighbors of each cell, and symmetrized neighbor relations define the graph's adjacency matrix comprised of transformed cell-cell distances. An adaptive kernel transforms edge weights to account for large cell density changes along the phenotypic manifold[91,199,200].

Based on a low-dimensional representation and kNN graph, clustering methods identify groups of cells with similar profiles; the single-cell field commonly employs the Leiden[82] and Louvain[81] algorithms, where the former improves on the latter. The two methods

compare connections within a cluster to connections between clusters to optimize the so-called modularity metric. For more details, I refer to the work introducing the Leiden algorithm[82], for a comparison of clustering approaches to dedicated benchmarking papers[201–204]. Identified clusters can be mapped to biological quantities such as cell types by studying differentially expressed genes ranked by classical statistical tests[205–207] or more sophisticated computational methods[59,62,208,209].

The principal component or latent space is low-dimensional but usually has more than three dimensions; as such, the space cannot be visualized easily. For visualization purposes, the space is, thus, projected into two or three dimensions with manifold-learning techniques. Such methods aim at reducing dimensionality while preserving the higher-dimensional topology. Common approaches are diffusion maps[210–212], t-distributed stochastic neighbor embedding (t-SNE)[213], or uniform manifold approximation and projection (UMAP)[214].

2.3. Trajectory inference

Classical scRNA-seq protocols are destructive by nature and, thus, measure each cell only once instead of tracking its evolution over time. Despite this snapshot nature, cellular heterogeneity exists in single-cell experiments since differentiation processes unfold asynchronously. As such, a range of the underlying mechanism is captured. However, given a reference cell, its progenitor state, and their gene expression profiles, it is unclear if the reference cell precedes its progenitor or vice versa. Numerous methods have been developed to recover the direction of the biological process by aligning cells along a trajectory; the field of TI collects all such methods.

Early methods for recovering trajectories focused on unidirectional processes along a linear trajectory and assigned each observation a pseudotime based on expression similarity; the pseudotime of cells emerging at the beginning of the dynamics is small, and cells toward the end have high values. Pseudotimes form a continuous domain of the discrete, observed cellular states aligned along the differentiation process and rank cells relative to each other. Later methods extended the concept to more complex settings like branching. Approaches for constructing pseudotimes are based on clusters[82,215–217], neighbor graphs[91,93,199,218], manifold-learning[94,96,219], and probabilistic frameworks[97,220–223].

Cluster-based methods identify connections between clusters of cells, with connections based on similarity or a minimum spanning tree. Similarly, graph-based algorithms define and connect clusters through kNN graphs. Probabilistic Approximate Graph Abstraction (PAGA)[218], for example, identifies and connects Leiden clusters; RaceID[224], StemID[225] and SLICER[226] are alternative graph-based approaches. Both cluster and graph-based approaches quantify cluster similarity but do not align these clusters and the cells they contain along the differentiation trajectory. Instead, the constructed network may serve as the trajectory backbone.

To estimate the underlying trajectory on a cell level, manifold-learning-based techniques infer trajectories based on principle curves - a one-dimensional curve connecting higher-dimensional observations - or graphs. Slingshot[96], for example, defines the pseudotime of a cell as its orthogonal projection onto principle curves fitted to each branch of a minimum spanning tree[217].

Probabilistic modeling assigns transition probabilities between cell pairs, quantifying

how likely the reference cell precedes the other. This construction defines a discrete Markov chain, *i.e.*, random walk. Based on the constructed random walk, Wanderlust[91] and diffusion pseudotime (DPT)[93], for example, define pseudotimes as a scaled distance along the data manifold with respect to a pre-defined root cell: DPT considers the difference between consecutive states of the random walk, Wanderlust samples waypoint cells and iteratively refines the distance of the shortest paths between them. Alternatively, Palantir[97] models the trajectory as a Markov chain. Appendix C includes the relevant mathematical theory of discrete Markov chains.

Estimating developmental potential for TI is similar to pseudotime. CytoTRACE[99], for example, relies on a simple but robust assumption validated on 42 scRNA-seq datasets: Immature cells express more genes than their mature counterparts, biologically motivated by less developed cells regulating their chromatin less tightly (Figure 2.3). Constructing the CytoTRACE score is a three-step process: First, the algorithm computes the Pearson correlation between the number of genes expressed in each cell and each gene - the gene count signature GCS; further construction of the stemness score relies on the 200 genes with maximum GCS. In the second step, CytoTRACE smooths gene expression counts based on a nearest-neighbor graph by solving a non-negative least squares regression problem and simulating a diffusion process. The final step of the approach computes the developmental potential of each cell as the geometric mean of its smoothed counts over the subset of the 200 genes and scales it to the unit interval.

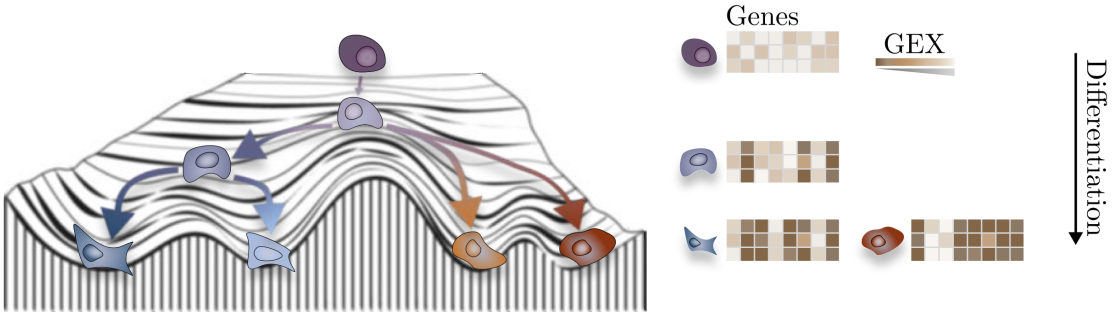

Figure 2.3.: CytoTRACE estimates stemeness potential from scRNA-seq data. CytoTRACE[99] assumes that cells express fewer genes as they differentiate into specialized cell types (GEX: gene expression). This metric aligns cells along the phenotypic manifold, visualized here by Waddington's landscape.

Dedicated reviews and related work compare pseudotime inference methods[227,228] and give a more in-depth introduction to trajectory inference and its limitations[192,193]. Although pseudotime and potential-based methods have celebrated great success, they neither describe nor predict cellular dynamics, such as splicing or mRNA turnover. As an alternative approach, RNA velocity has emerged as a putative bottom-up mechanistic modeling paradigm for estimating a vector field along the phenotypic manifold in a data-driven fashion.

2.4. RNA velocity

Similarity-based trajectory inference assigns cellular fate but does not recover directed dynamic information. According to the central dogma of molecular biology, genetic information flows unidirectionally: DNA is transcribed into nascent (unspliced) mRNA, followed by splicing into mature (spliced) mRNA and translation into proteins[229] (Figure 2.4a). Standard scRNA-seq experiments detect both nascent and mature mRNA

molecules through the presence and absence of introns, respectively[100,103–106]. Relating the two abundances via a gene-specific dynamical model, RNA velocity[100,101] - the time derivative of mature mRNA - models the transcriptomic change over the biological process.

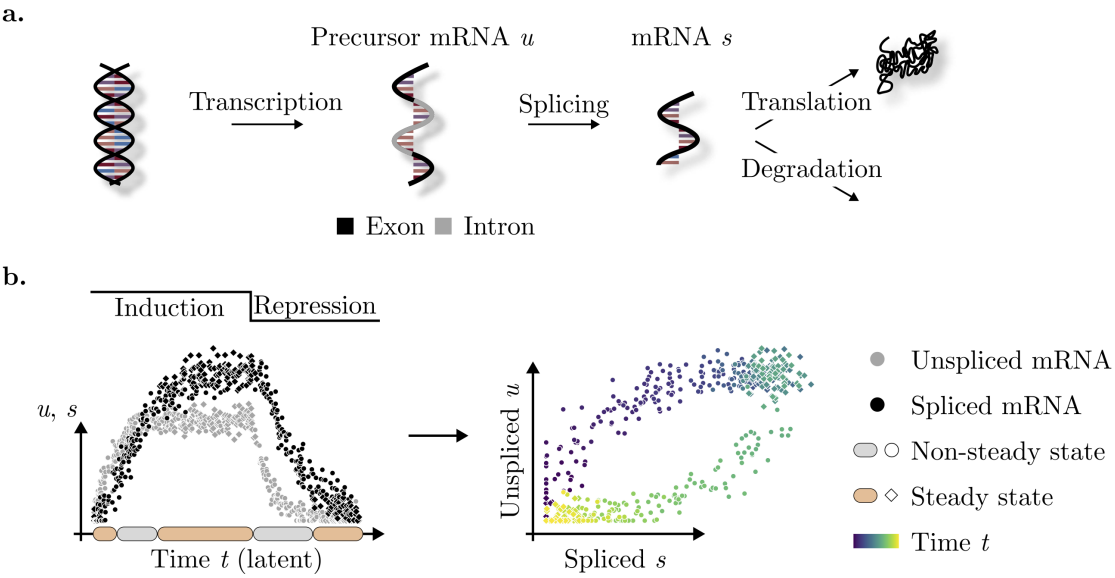

Figure 2.4.: Leveraging the central dogma of molecular biology with scRNA-seq data. a. DNA is transcribed into nascent (unspliced) mRNA u, followed by splicing into mature (spliced) mRNA s and translation into proteins[229]. **b.** Common scRNA-seq protocols produce snapshot data, lacking cell-specific time information. Instead of studying splicing dynamics based on its temporal progression (left), popular RNA velocity approaches[100,101] study the process via the phase space (u, s) (right).

Classical approaches for inferring parameters of dynamical systems rely on maximum likelihood (MLE) or maximum a posteriori (MAP) estimates of the likelihood and posterior distribution, respectively. However, sequencing data lacks observation-specific temporal information, rendering the MLE and MAP uncomputable. To infer the kinetic parameters underlying RNA velocity, nonetheless, the two most popular approaches - the *steady-state model*[100] and *EM model*[101] - investigate the problem in phase space (u, s) (Figure 2.4b).

The *steady-state* and *EM model* approximate splicing dynamics with a gene-specific ordinary differential equation (ODE), omitting gene interactions: Unspliced mRNA u is transcribed at rate α and spliced into spliced mRNA s at rate β; following, spliced mRNA is degraded at a rate γ. The corresponding dynamical system used for RNA velocity[100–102] inference approximates the chemical master equations (CME)

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}P_t(u=m,s=n) =& \alpha\left[P_t(u=m-1,s=n)-P_t(u=m,s=n)\right]+\\ &\beta\left[(m+1)P_t(u=m+1,s=n-1)-P_t(u=m,s=n)\right]+\\ &\gamma\left[(n+1)P_t(u=m,s=n+1)-P_t(u=m,s=n)\right]\end{aligned} \tag{2.1}$$

up to first order (Appendix D); $P_t(u=m,s=n)$ denotes the probability to observe $m \in \mathbb{N}_0$ unspliced and $n \in \mathbb{N}_0$ spliced molecules at time t. The *steady-state* and *EM model* are not count-based, necessitating the first-order approximation; they study the

process in a continuous and deterministic domain, instead. This view approximates (2.1) by first-order moments[102,230] to model splicing dynamics with

$$\begin{aligned} \dot{u} &= \alpha - \beta u \\ \dot{s} &= \beta u - \gamma s, \end{aligned} \tag{2.2}$$

with $\alpha > 0$ if the state is in induction, and $\alpha = 0$, otherwise.

The steady-state model

In addition to no gene-gene interactions, the *steady-state model* assumes (1) constant rate parameters, (2) a gene-shared unit splicing rate $\beta = 1$, and (3) that the steady-states of ODE (2.2) are observed. Under these assumptions, cells in equilibrium are located in the extreme quantiles of the systems' phase portrait (Figure 2.5a), and the dependence of unspliced and spliced mRNA is linear

$$u = \frac{\gamma}{\beta} s.$$

The *steady-state model* fits a regression line with slope γ^* to the extreme quantiles of the measured expression profiles and defines the RNA velocity v_j of an observation j with unspliced and spliced reads u_j and s_j, respectively, as the residual to this fit, *i.e.*,

$$v_j = u_j - \gamma^* s_j.$$

While the *steady-state model* successfully recapitulated the lineage tree of developing mouse hippocampus and human embryonic brain, its modeling assumptions are violated in many real-world datasets[100,101,124,231].

The EM model

Heterogeneous tissue samples or heterogeneous subpopulations with subpopulation-specific kinetics exist and violate the *steady-state model*'s assumptions of a common splicing rate for all genes. Similarly, the *steady-state model* estimates incorrect rates in transient systems that do not reach equilibrium. To overcome these limitations, the *EM model* estimates model parameters $\theta = (\alpha, \beta, \gamma)$ and latent time t and state k - induction, repression, or either steady state - for each cell (Figure 2.5b); the Python package *scVelo*[101] provides an implementation of the approach.

The *EM model* optimizes model parameters of (2.2). Although the analytic solution of (2.2)

$$\begin{aligned} u(t) &= e^{-\beta\tau} u_0 + \frac{\alpha}{\beta}(1 - e^{-\beta\tau}) \\ s(t) &= e^{-\gamma\tau} s_0 + \frac{\beta u_0 - \alpha}{\gamma - \beta}(e^{-\beta\tau} - e^{-\gamma\tau}) + \frac{\alpha}{\gamma}(1 - e^{-\gamma\tau}), \end{aligned} \tag{2.3}$$

with $\tau = t - t_0$ and initial states $u_0 = u(t_0)$ and $s_0 = s(t_0)$, exists, maximum likelihood or maximum a posteriori estimates are intractable as the model contains time and state

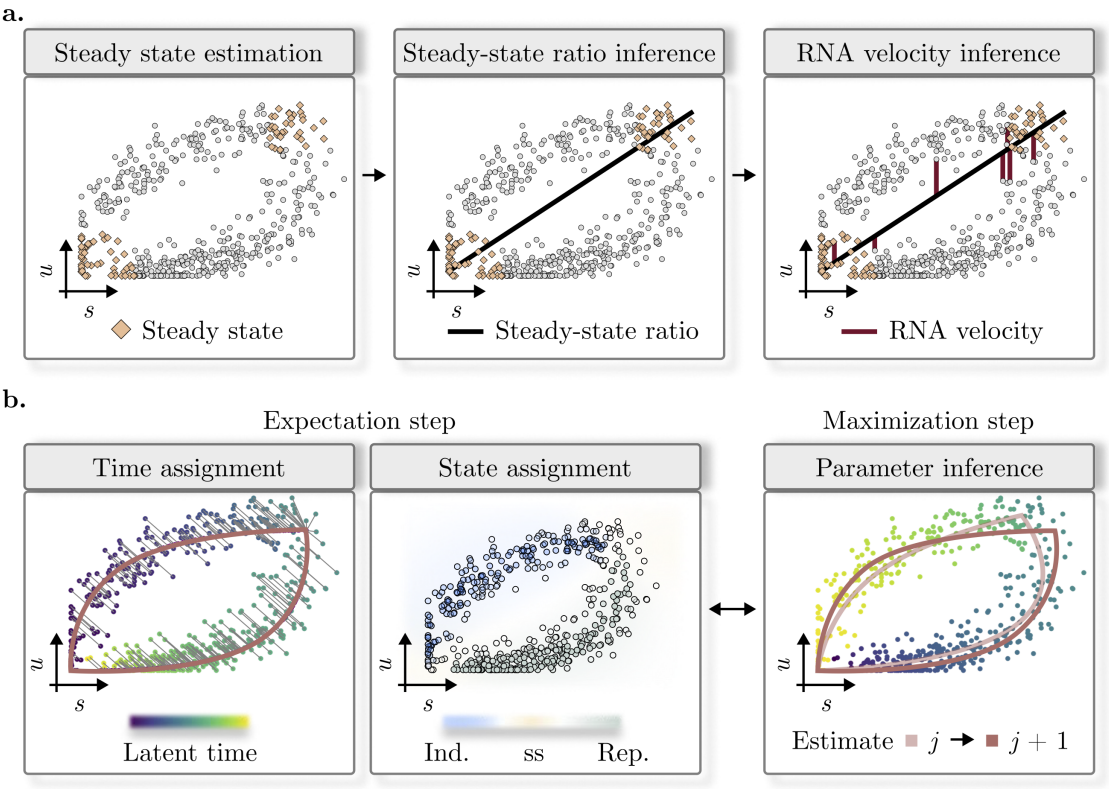

Figure 2.5.: Traditional RNA velocity inference methods. a. The *steady-state model*[100] estimates cells in steady state (left), infers the steady-state ratio via an extreme quantile linear regression fit (middle), and defines RNA velocity as the residual to the fitted line (right). **b.** The *EM model*[101] relies on less restrictive assumptions compared to the *steady-state model*, estimating the full set of model parameters and latent variables through an EM algorithm (Ind.: Induction; ss: steady state; Rep.: Repression).

as latent variables. Instead, an expectation-maximization (EM) optimization scheme[232] forms the basis of parameter inference: The M-step estimates observed parameters, and the E-step latent variables iteratively. The E-step updates the latent variables time t and state k of each observation. The latent time assignment minimizes the Euclidean distance between measured data and its estimate

$$t_j^{(n+1)} = \arg\min_t \left\| x_j - \hat{x}(t, \hat{\theta}^{(n)}) \right\|,$$

and the state assignment

$$k_j^* = \arg\min_k \left\| x_j - \hat{x}(t_j^{(n+1)}, \hat{\theta}^{(n)}) \right\|,$$

with observations $x = (u(t), s(t))$ and their estimates $\hat{x}(t|\hat{\theta})$, respectively; superscripts indicate the current step of the iterative optimization, the index the observation. The M-step then maximizes

$$\mathbb{E}_k[l(\theta|\mathcal{X}, t, k)] = \sum_{j=1}^{N_c} l(\theta|x_j, t_j).$$

scVelo implements the case of normal and Laplace distributed observations, where the gene-wise variance across all observations defines the corresponding variance; to decrease runtime, *scVelo* approximates the optimal latent time assignment, achieving a 30-fold speedup[101].

The *EM model* improves upon the *steady-state model* as it does not assume observed steady-states, and instead solves the full splicing dynamics by relying on the entire data distribution. However, the approach is still tied to a specific formulation of splicing kinetics that assumes constant rates and ignores gene-gene interactions. These assumptions may hinder its applicability in real-world datasets[101,231]. Similarly, the *EM model* does not provide uncertainty of estimated parameters.

Velocity variational inference

Variational inference (VI) approximates data distributions in a Bayesian setting[233] (Appendix B). As such, it carries notions of estimation uncertainty and, paired with recent advances in variational autoencoders, scales to large datasets[234,235]. The framework is, thus, suitable for RNA velocity to overcome limitations posed by the *EM model*, namely scalability, model flexibility, and parameter uncertainty. Such a model posits a generative process informed by splicing dynamics (2.2) but ties cell and gene-specific latent variables through the latent space (Figure 2.6).

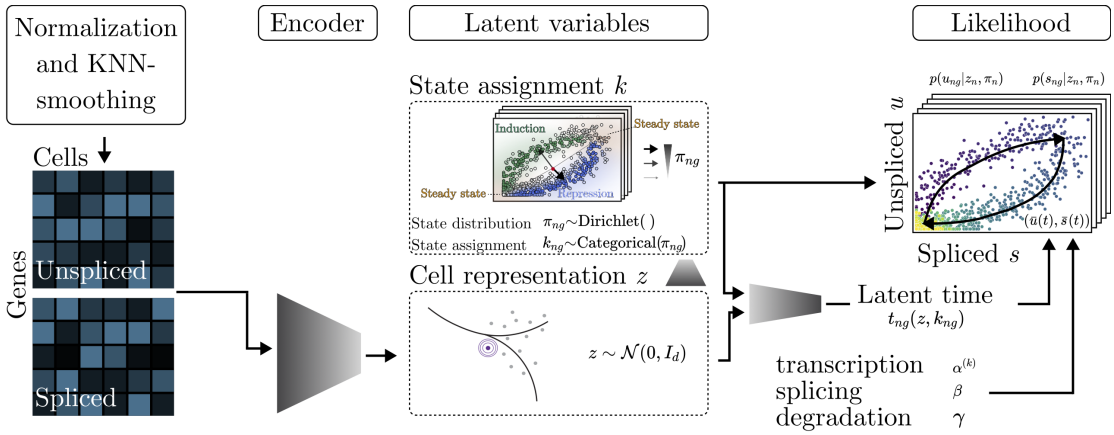

Figure 2.6.: veloVI infers RNA velocity with variational inference. veloVI encodes unspliced and spliced mRNA counts into a latent cell representation z, encoding cell-gene-specific transcriptional state k, based on neural networks. A corresponding decoder neural network provides cell-gene-specific latent time estimates based on the cell representation z and latent state k. Following, the model optimizes a likelihood function in an end-to-end fashion to infer state uncertainty, latent time, transcription, splicing, and degradation rates α, β and γ, respectively, of splicing dynamics. Figure adapted from the work introducing veloVI[126].

Velocity variational inference (veloVI)[126] assumes a latent state $z_n \sim \text{Normal}\,(0, I_{N_l})$ represents the latent state of a cell n, where I_{N_l} denotes the N_l-dimensional identity matrix. For each gene g, this latent state encodes a distribution over the possible cell states k with

$$\pi_{ng} \sim \text{Dirichlet}\,(0.25, 0.25, 0.25, 0.25)$$
$$k_{ng} \sim \text{Categorical}\,(\pi_{ng})$$

Here, $k = 1$ represents the induction phase, $k = 2$ its steady state, and $k = 3$ and $k = 4$ the repression phase and its equilibrium, respectively.

Latent state and cell representations encode cell-gene-specific latent times by two fully-connected neural networks: The first network $q_{\xi_1}^{(\text{ind})} : \mathbb{R}^{N_\text{l}} \to (0,1)^{N_\text{g}}$ is active during induction, $q_{\xi_2}^{(\text{rep})} : \mathbb{R}^{N_\text{l}} \to (0,1)^{N_\text{g}}$ during repression. Thus, $q_{\xi_1}^{(\text{ind})}$ encodes latent time as

$$t_{ng}^{(1)} = q_{\xi_1}^{(\text{ind})}(z_n)_g t_g^\text{s},$$

with switching time t_g^s, and $q_{\xi_2}^{(\text{rep})}$ with

$$t_{ng}^{(3)} = t_g^\text{s} + q_{\xi_2}^{(\text{rep})}(z_n)_g (t_\text{max} - t_g^\text{s}),$$

where t_max defines the maximum time of the process to guarantee parameter identifiability[101]. Finally, the observed data are sampled from normal distributions as

$$\begin{aligned} u_{ng}^{(\text{obs})} &\sim \text{Normal}\left(u^{(g)}(t_{ng}^{(k_{ng})}, k_{ng}), (c_k \sigma_g^u)^2\right) \\ s_{ng}^{(\text{obs})} &\sim \text{Normal}\left(s^{(g)}(t_{ng}^{(k_{ng})}, k_{ng}), (c_k \sigma_g^s)^2\right), \end{aligned}$$

with state-dependent scaling factors

$$c_k = \begin{cases} 1 & k \in \{1,2,3\} \\ 0.1 & k = 4, \end{cases}$$

reflecting the repression steady state corresponds to transcriptional inactivity. Importantly, the variance is not zero to model noise originating from sequencing workflows or preprocessing choices.

VAEs infer approximate posterior distributions with encoders q_ϕ with parameters ϕ. To guarantee tractable computation, veloVI factorizes the posterior according to

$$q_\phi(z, \pi \mid u, s) := \prod_{n=1}^{N_\text{s}} q_\phi(z_n \mid u_n, s_n) \prod_{g=1}^{N_\text{g}} q_\phi(\pi_{ng} \mid z_n). \tag{2.4}$$

Integrating over transcriptional states defines the likelihood of unspliced and spliced mRNA as a mixture of Gaussian distributions

$$\begin{aligned} p_\theta(u_{ng}^{(\text{obs})} \mid z_n, \pi_n) &= \sum_{k_{ng} \in \{1,2,3,4\}} \pi_{ngk_{ng}} \text{Normal}\left(u^{(g)}(t_{ng}^{(k_{ng})}, k_{ng}), (c_k \sigma_g^u)^2\right) \\ p_\theta(s_{ng}^{(\text{obs})} \mid z_n, \pi_n) &= \sum_{k_{ng} \in \{1,2,3,4\}} \pi_{ngk_{ng}} \text{Normal}\left(s^{(g)}(t_{ng}^{(k_{ng})}, k_{ng}), (c_k \sigma_g^s)^2\right), \end{aligned}$$

with model parameters θ including splicing parameters α, β, γ, switch time t^{s}, and network parameters ξ_1 and ξ_2. Parameters of the generative model minimize

$$\mathcal{L}_{\mathrm{velo}}(\theta, \phi; u, s) = -\mathrm{ELBO}\left[\theta, \phi; u, s\right] + \lambda \mathcal{L}_{\mathrm{switch}}(\theta; u, s),$$

with evidence lower bound ELBO and $\mathcal{L}_{\mathrm{switch}}$ biasing the switch from induction to repression towards the upper right part of the phase portrait. For this constraint, consider the median count u^* and s^* of unspliced and spliced observations above the 99th percentile, respectively; comparing these estimates to the fitted initial states $u_{\mathrm{r},0}^{(g)}$ and $s_{\mathrm{r},0}^{(g)}$ of the repression phase define $\mathcal{L}_{\mathrm{switch}}$ as

$$\mathcal{L}_{\mathrm{switch}}(\theta; u, s) = \sum_g \left(u_{\mathrm{r},0}^{(g)} - u_g^*\right)^2 + \left(s_{\mathrm{r},0}^{(g)} - s_g^*\right)^2. \tag{2.5}$$

To summarize, veloVI solves the RNA velocity inference problem in a Bayesian setting with VAEs. This modeling choice facilitates model flexibility as parameter optimization is not tied to a specific model formulation, scalability due to recent advances in deep learning, and uncertainty quantification through the Bayesian formulation. Additionally, veloVI can estimate RNA velocity for held-out data; comparing the original and shuffled expression patterns of a given gene, for example, reveals structural insight related to model applicability[126].

2.5. Analysis of time-resolved sequencing data

With decreasing sequencing costs and simplified experimental workflows, single-cell datasets have become more diverse; common dataset size has increased[38], and measuring samples at distinct time points for studying dynamic, non-steady state systems has become standard[109,197,236–238]. While this advancement offers a more complete view of the underlying processes, it poses new problems: Cell evolution needs to be mapped based on intra and inter-time point information. However, neither pseudotime nor RNA velocity inference explicitly incorporates the temporal information in their respective modeling paradigms.

Leveraging temporal information for pseudotime inference

Asynchronous differentiation results in similar cell states in consecutive time points; mature cells at an earlier stage correspond to less mature cells within a later stage. However, due to batch effects, classical neighbor graphs tend to neglect this information and focus on intra-time point similarity alone. To induce inter-time point connectivities, Harmony[239] identifies mutual nearest neighbors[196] - cells from different time points that are in each other's neighborhood - to build an augmented nearest neighbor graph; the augmented graph acts as input for classical pseudotime methods operating on graphs like Palantir[97]. However, Harmony does not overcome the limitations of pseudotime inference even though it incorporates temporal information into TI. Alternative methods try to overcome batch effects manifested in the kNN graph through relative information content (WNN[67]) or construct a graph that is balanced across all batches of the data; however, these approaches have not yet been tested and benchmarked in the context of pseudotime estimation.

Optimal transport infers cellular change across time points

Optimal transport (OT) provides a powerful tool for matching cells with their putative progenitor state at a later time point in a probabilistic fashion. The ground-truth cell distribution (Figure 2.7a) changes along experimental time but is unknown. Instead, scRNA-seq approximates it through discrete samples (Figure 2.7b). This paragraph focuses on briefly outlining the application of OT in the single-cell sequencing context, Appendix E provides the mathematical theory of OT more in-depth.

Given observations of cells at two consecutive experimental time points, OT matches the respective distributions probabilistically; Waddington-OT[109] (WOT) was among the first methods to apply unbalanced OT to scRNA-seq data: It associates each cell in an earlier time point t_1 with a set of putative future states in the later time point t_2, assigning each putative progenitor a probability (Figure 2.7c); coupling of consecutive time points and propagating state changes through matrix multiplication of the transport maps allows studying sequences of experimental time points.

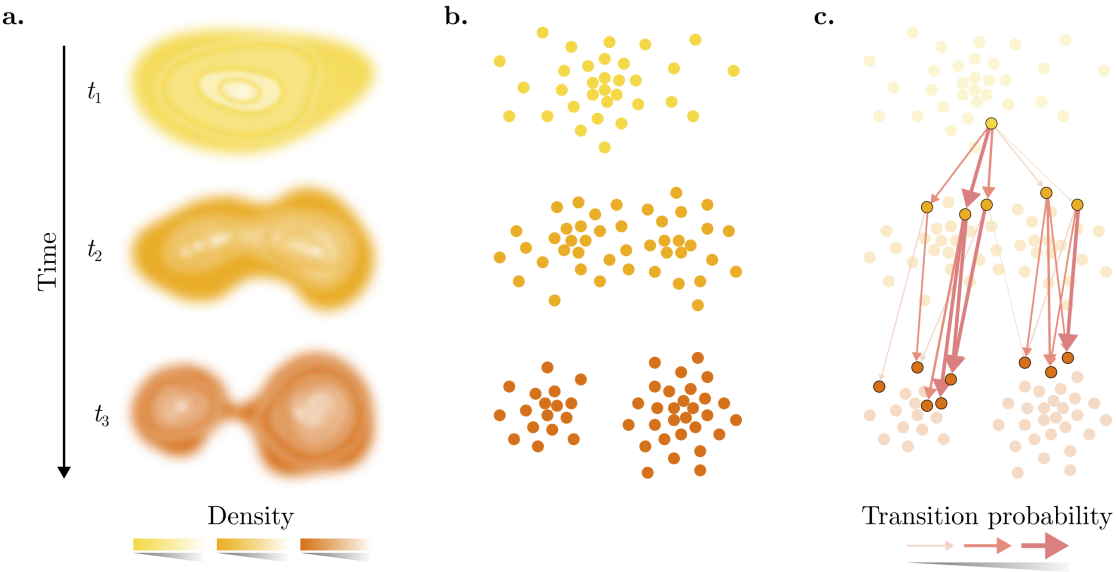

Figure 2.7.: Optimal transport models differentiation processes. **a.** The ground-truth distribution of cell states across different time points describes differentiation processes. **b.** scRNA-seq approximates these distributions with discrete samples. **c.** To connect cells at an earlier time point, OT matches a cell with its putative progenitor states in the consecutive stage probabilistically (Appendix E).

WOT relies on the Sinkhorn algorithm[240] to quantify probabilities of cellular state change. The assigned probabilities minimize

$$\langle C, T\rangle + \tau_1 \mathcal{D}_{\mathrm{KL}}\left[T1_{N_2} \parallel a\right] + \tau_2 \mathcal{D}_{\mathrm{KL}}\left[T^\top 1_{N_1} \parallel b\right] - \varepsilon \mathrm{H}(T), \tag{2.6}$$

with Kullback-Leibler divergence $\mathcal{D}_{\mathrm{KL}}$[241], cost matrix C, assignment matrix T, entropy H, regularization parameters $\tau_1, \tau_2, \varepsilon > 0$, marginal source and target distributions $a \in [0,1]^{N_1}$ and $b \in [0,1]^{N_2}$, and the constant, N_j-dimensional one vector 1_{N_j}, respectively; for WOT, the distance between the PCA representation of two observations represents the transport cost between them although other latent representations are possible as well[111].

WOT defines a uniform marginal distribution $b_j = \frac{1}{N_2}$, $j \in \{1, \dots N_2\}$, and accounts for cell proliferation and death at rates β and δ, respectively, from t_1 to t_2. An exponential growth and decline models the distribution shift as a birth-death process g; the normalized birth-death process defines a element-wise as

$$a_j = \frac{g(x_j)^{t_1 - t_2}}{\sum_{n=1}^{N_1} g(x_n)^{t_1 - t_2}}, \tag{2.7}$$

with $j \in \{1, \dots N_1\}$. WOT accounts for uncertainty in estimates of g by choosing marginal weights appropriately: A small τ_1 allows variation from the marginal on the source domain, and a large value τ_2 favors matching the marginal distribution b; by default, WOT sets $\tau_1 = 1$ and $\tau_2 = 50$.

Metabolic labeling for modeling RNA velocity

Metabolic labeling offers an alternative approach for introducing temporal resolution to single-cell sequencing data. Nucleotide analogues label newly transcribed mRNA molecules, offering temporally and mechanistically coupled modalities similar to nascent and mature mRNA for RNA velocity as consecutive steps of the central dogma of molecular biology[229]. Compared to discrete experimental time points, metabolic labeling can reveal biological mechanisms on shorter time scales and distinguish regulatory effects[178].

Dynamo[107] infers cellular dynamics through a mechanistic model similar to the splicing model (2.2) but based on metabolic labels. However, the method relies on a steady-state assumption, only uses a small subset of cells for parameter inference, and does not estimate cell-specific rates. CellRank 2[129] provides an alternative approach for estimating the velocity field underlying the biological dynamics.

CellRank 2's approach for kinetic rate estimation works with pulse and chase experiments[127]: Pulse experiments label n cell cultures at times t_j, $j \in \{1, \dots, n\}$ with $t_j < t_{j+1}$. Alternatively, chase experiments expose cells to nucleoside analogues long enough to guarantee that all transcripts are labeled before washing them out at times t_j, $j \in \{1, \dots, n\}$. Both types of experiments sequence cells at a single time t_f, defining the labeling time as $\tau_l^{(j)} = t_f - t_j$.

CellRank 2 estimates cell-specific transcription and degradation rates based on pulse and chase experiments. The method assumes that for each gene,

$$r(t) = r_0 e^{-\gamma t} + \frac{\alpha}{\gamma}\left(1 - e^{-\gamma t}\right)$$

describes the change in mRNA levels r with transcription rate α and degradation rate γ; the temporal relationship solves an ODE similar to (2.2). Assuming mRNA abundance changes according to the proposed model, pulse and chase-specific descriptions arise from the different experimental setups; the number of labeled transcripts r_l is 0 and r_0 for pulse and chase experiments, respectively. CellRank 2, thus, infers model parameters α, γ and r_0.

Inferring model parameters is a two-step process: First, for each gene g and cell j, the inference scheme includes the set of nearest neighbors $\mathcal{N}_g^{(k)}$ that contains 20 non-trivial

expression counts in g. Based on these observations, an optimization routine minimizes the quadratic loss

$$\ell(r_0^{(j,g)}, \alpha^{(j,g)}, \gamma^{(j,g)}) = \sum_k \sum_{j \in \mathcal{N}_g^{(k)}} [r_{l,j}(\tau_l^{(k)}) - \mathbb{1}(j \in \mathcal{C}) r_l^{(c)}(\tau_l^{(k)} | \alpha, \gamma, r_0) - \mathbb{1}(j \in \mathcal{P}) r_l^{(p)}(\tau_l^{(k)} | \alpha, \gamma)]^2.$$

Here, superscripts (p) and (c) indicate the type of experiment, and $\mathbb{1}$ denotes the indicator function. Importantly, at least two (three) labeling times guarantee parameter identifiability for pulse (chase) experiments. Compared to a previous approach[127], CellRank 2's framework focuses on total RNA, thereby avoiding the common pitfalls of identifying unspliced and spliced reads, and does not require a pseudotemporal ordering of cells.

2.6. Cellular fate mapping

Pseudotime, RNA velocity, and vector fields inferred on time-resolved sequencing data recover the underlying vector field of biological differentiation processes. However, such vector fields alone are merely descriptive instead of disentangling and quantifying the mechanisms driving cellular differentiation. Methods assigning cellular fate potential aim to infer biological mechanisms and associate key putative regulatory drivers; these methods identify initial and terminal states, assign cellular fate to define corresponding lineages, and associate potential drivers. For each data view, dedicated methods have been developed.

Cellular fate mapping with pseudotime-informed Markov chains

Pseudotime estimation ranks cells relative to each other along the differentiation landscape but does not assign cellular fate: To bridge this gap, Palantir[97] assumes that paths in kNN graphs correspond to possible differentiation paths and combines kNN graphs with pseudotime values; to approximate the phenotypic manifold accurately with a neighbor graph, an adaptive kernel[200] corrects weights based on the distance to the l-th neighbor - the scaling factor σ. For fate mapping, the algorithm relies on this neighbor graph and pseudotime estimates to (1) bias graph edges towards increased differentiation potential and (2) define transition probabilities between states, defining a Markov chain to infer terminal states and fate probabilities toward them (Figure 2.8a).

Neighbors with larger pseudotime values compared to a given observation pose likely future progenitors of that state. Given two observations j and k with pseudotimes t_j and t_k, respectively, Palantir prunes the edge e_{jk} if the difference between the pseudotime values exceeds the scaling factor σ_j. Following, the algorithm normalizes edge weights w_{jk} for each observation j over its neighborhood to define transition probabilities p_{jk}. These probabilities induce a Markov chain, where p_{jk} is the probability that cell j evolves into the state of cell k.

Terminal states do not differentiate further and random walks, thus, terminate in a perfect scenario; single-cell data is noisy and the Markov chain construction implicitly includes estimation uncertainty, however. Nonetheless, the biased neighbor graph skews the stationary distribution of the Markov chain toward states close to terminal by construction. Similarly, terminal states reside at the boundaries of the manifold based on

the definition of pseudotime. Palantir, thus, defines terminal states as the intersection of extrema in the stationary distribution and diffusion components; absorption probabilities toward terminal states define fate probabilities.

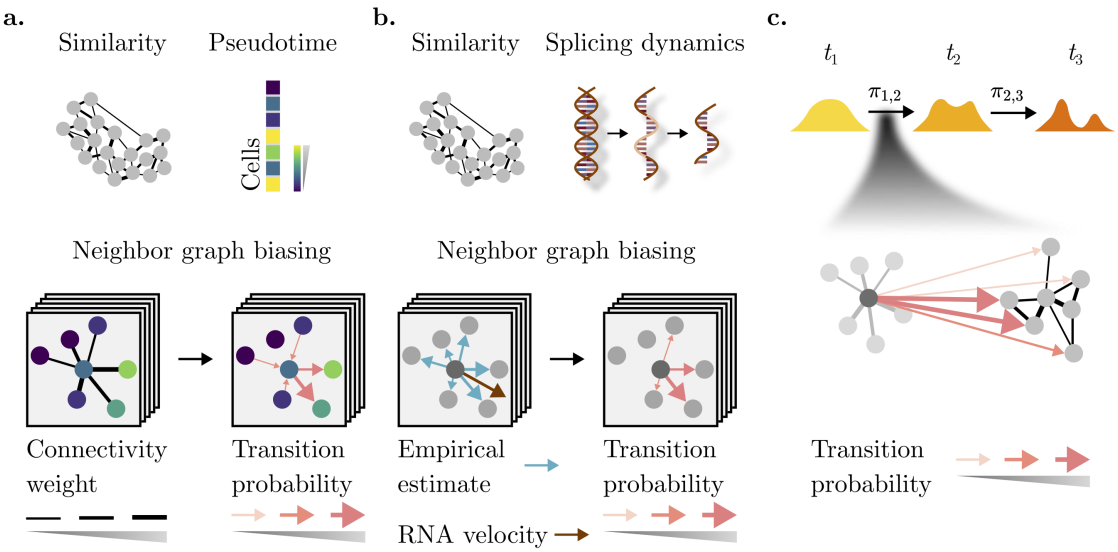

Figure 2.8.: Fate mapping for single-cell sequencing data. a. Palantir[97] biases the edges of a kNN cell-cell similarity graph towards increased pseudotime to assign transition probabilities that induce a Markov chain. **b.** Similarly, CellRank[116] relies on RNA velocity estimates to bias the kNN graph by comparing empirical velocity estimates to RNA velocity. **c.** OT-based approaches such as WOT[109] and moscot[111] assign transition probabilities to putative future cell states in consecutive stages, defined by experimental time, for example.

RNA velocity-based inference of directed, probabilistic state-change trajectories

Pseudotime inference requires a root state and does not offer directed information; RNA velocity overcomes these limitations by estimating a vector field without a pre-defined root cell. CellRank[116] combines the vector field with cell-cell similarity measures to recover initial and terminal states using Markov chains. Similar to Palantir, CellRank first estimates transition probabilities before inferring terminal states.

CellRank assigns transition probabilities by comparing RNA velocity estimates to state change estimates derived from a kNN (Figure 2.8b). For each observation j, the method first computes a displacement vector to each of its neighbors. Following, the transformed and normalized correlation between the RNA velocity of cell j and the putative state shifts quantifies how likely a given neighbor is the future state of cell j. This transition matrix T induces a Markov chain but is noisy, making it challenging to derive biological insight directly from it. Biological processes exhibit inherent structure, however, as cells differentiate from one state to another. CellRank makes use of this observation by coarse-graining the transition matrix.

The high-dimensional cell-cell transition matrix includes clusters of cells, so-called macrostates, recapitulating biological state changes. CellRank projects the original transition matrix T onto a coarse-grained representation with Generalized Perron Perron Cluster Analysis (GPCCA)[242–244]. GPCCA assigns each cell a soft macrostate membership by maximizing membership crispness[245]; an invariant subspace projection defines the transition probabilities between macrostates. CellRank either automatically defines terminal

states based on macrostate stability or manual selection, and assigns fate probabilities through absorption probabilities.

Optimal transport for mapping cell differentiation in time-resolved data

Non-steady state systems change over time and, thus, require samples from multiple time points to study the full system. Optimal transport provides a framework for matching likely ancestor-progenitor couples across consecutive time points (Figure 2.8c). These couplings inform cellular fate mapping with WOT[109]. For inference, WOT[109] assumes cellular states change in a Markovian fashion, similar to Palantir[97] and CellRank[116]. Optimal transport maps $\pi_{t_j,t_{j+1}}$ between consecutive time points, thus, define the long-range coupling $\gamma_{j,j+k}$ between arbitrary times t_j and t_{j+k} by successively applying the transport maps

$$\gamma_{j,j+k} = \pi_{j,j+1} \circ \pi_{j+1,j+2} \circ \cdots \circ \pi_{j+k-1,j+k}. \tag{2.8}$$

Pushing a cell set through $\gamma_{j,j+k}$ quantifies its descendants, pulling it through $\gamma_{j,j+k}$ its ancestors. T-test-based differential expression analysis of gene expression trends along trajectories identifies putative lineage drivers.

Geometric vector field properties for terminal state identification

Metabolic labels pose an alternative approach for adding temporal information in scRNA-seq experiments. Dynamo[107] offers a framework to recover the vector field along the phenotypic manifold and identify the initial and terminal states of the dynamical system. First, sparseVFC[246] approximates the continuous vector field; the method builds on reproducing kernel Hilbert spaces, characterized by kernel basis functions. In a second step, dynamo computes attractors of the recovered vector field to assign initial and terminal states. Putative drivers of each lineage are characterized as genes deviating most from the optimal transition path between two states - the least action path[247,248]; the deviation is defined as the squared error between gene expression at a sample time and the start of the trajectory.

Generalized fate mapping for multiview single-cell data

Following CellRank's modeling paradigm, CellRank 2[129] generalizes the concept to include pseudotime and cell potential estimates, real-time information, or metabolic labels. Importantly, the framework is modality agnostic, modular, and scales to atlas-sized datasets consisting of millions of cells; estimating cell-cell transition probabilities is decoupled from their analysis (Figure 2.9). As such, new data modalities such as lineage tracing[249], chromatin-derived metrics[250], and spatiotemporal measurements[111] extend the framework seamlessly.

The PseudotimeKernel generalizes pseudotime-based fate mapping following the Palantir approach - biasing similarity-based neighbor graphs towards increased pseudotime. Given the adjacency matrix, the PseudotimeKernel decreases edge weights if the pseudotime value of a reference cell j succeeds the estimate for a neighbor k. Similar to previous approaches, the procedure downweighs edges based on a hard threshold (Palantir[97]) or continuously (soft thresholding; VIA[251]); hard thresholding removes most edges towards

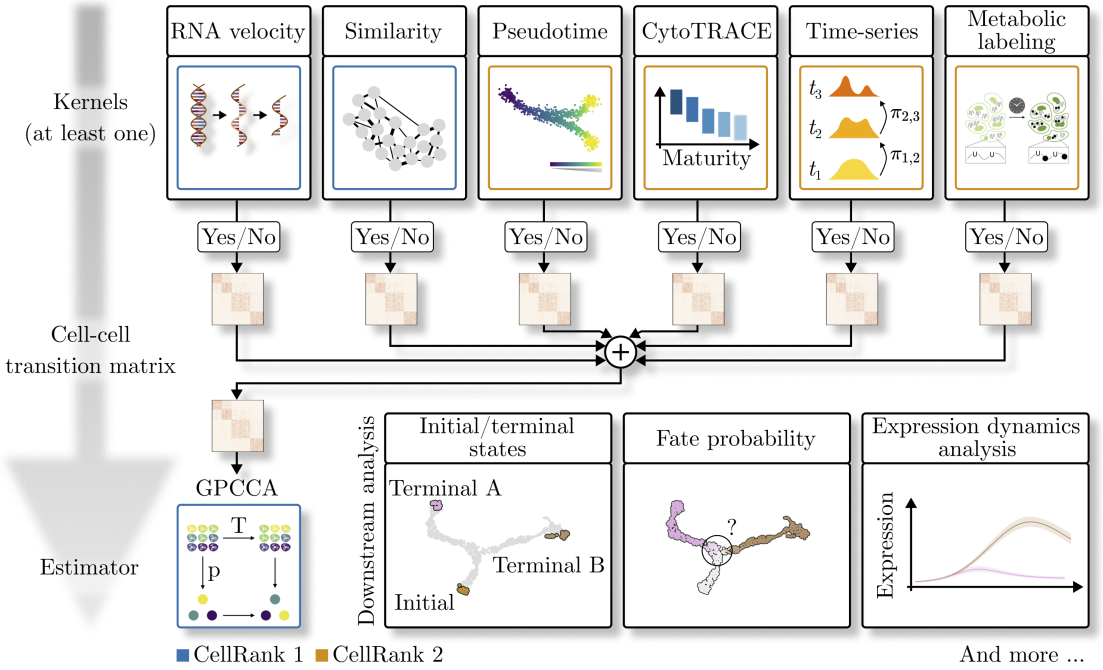

Figure 2.9.: CellRank 2 provides a framework to study single-cell fate decisions using Markov chains in a unified fashion. CellRank 2[129] decomposes the trajectory inference into cell-cell transition inference with *kernels* and analysis thereof with *estimators*. Estimators infer initial and terminal states, quantify fate probability and identify lineage-correlated genes; to incorporate multiple views, kernels can be combined. Blue coloring indicates features originally proposed in CellRank 1[116] and orange coloring new features. Figure reproduced and adapted from the original CellRank 2 publication[129].

neighbors in the pseudotemporal past of the reference cell, while soft thresholding updates edges based on weights

$$w(\Delta t) = \begin{cases} \frac{2}{\sqrt[\nu]{1+e^{b\Delta t}}}, & \Delta t < 0 \\ 1, & \Delta t \geq 0, \end{cases}$$

with difference in pseudotime Δt and parameters b and ν.

The CytoTRACEKernel estimates a stemness potential following the CytoTRACE[99] approach (Section 2.3) to skew neighbor graph edges similar to the PseudotimeKernel. Compared to the original method, CellRank 2 adapts the imputation step to yield comparable results but scale computation to millions of cells. The CytoTRACE score c assigns immature cells values close to 1, and mature observations values close to 0. To employ the biasing scheme of the PseudotimeKernel, CytoTRACEKernel, thus, transforms the score into a corresponding pseudotime via

$$p_{\text{cyt}} = 1 - c.$$

The RealTimeKernel combines inter with intra-time point connections to leverage the information from different experimental time points and asynchronous biological behavior contained in a single measurement. Considering time points j, OT couples observations between consecutive time points t_j and t_{j+1} with WOT[109] or moscot[111], and cell-cell similarity quantifies the dynamics within time point. The RealTimeKernel combines both sources into a single cell-cell transition matrix with block structure: The diagonal includes the similarity-based transitions, the first off-diagonal the OT estimates.

Unifying the within and across time point information includes row-normalization, weighting intra and inter-time point connections with a global factor, and thresholding transition probabilities to guarantee scalability; thresholding is necessary as entropic regularized OT yields dense transport maps although mostly negligible entries. By considering dynamics within and across the RealTimeKernel recovers biological insight more faithfully than approaches focusing on a single source of information[129].

Metabolic-labelling-based fate mapping is another feature included in CellRank 2, closely related to CellRank's[116] RNA velocity-based trajectory inference. Instead of estimating RNA velocity from unspliced and spliced mRNA abundances, CellRank 2 uses the velocity field that its routine for metabolic-labeling-informed RNA velocity estimation provides. Compared to dynamo[107], CellRank 2 does not rely on a deterministic framework to infer cellular trajectories.

Different methods for inferring cellular fate focus on alternative data views and require specific properties of sequencing data. However, each method first quantifies cell-cell transition probabilities, to then infer differentiation direction, terminal states, and regulatory mechanisms at branching points towards them. CellRank 2 unifies cellular fate mapping under this paradigm, making method-specific approaches applicable to more general settings.

3. Publication summary

This section summarizes the two main publications presented in this dissertation. Equal contribution is indicated by an asterisk.

3.1. Publication 1: Deep generative modeling of transcriptional dynamics for RNA velocity analysis in single cells

The paper titled *Deep generative modeling of transcriptional dynamics for RNA velocity analysis in single cells* has been published as an Article in *Nature Methods* in 2024. The full citation is

Adam Gayoso [*], **Philipp Weiler** [*], Mohammad Lotfollahi, Dominik Klein, Justin Hong, Aaron Streets, Fabian J. Theis, Nir Yosef Deep generative modeling of transcriptional dynamics for RNA velocity analysis in single cells. *Nat Methods* **21**, 50-59 (2024)

Contribution

I conceptualized the study with Adam Gayoso and Mohammad Lotfollahi. I contributed to conceptualizing the statistical model, its design and implementation, and quantifying model uncertainty. I designed and implemented model extensions to highlight model flexibility by fitting a model including time-dependent transcriptional rates. Similarly, I implemented analysis methods to evaluate model performance compared to other approaches and notions of ground truth. I performed the analyses to highlight veloVI's capabilities and better performance compared to existing methods with contributions from Adam Gayoso. I wrote the manuscript with Adam Gayoso, Mohammad Lotfollahi, Fabian Theis, and Nir Yosef.

Additional supplementary material

Additional supplementary material is available at the publisher's website (https://doi.org/10.1038/s41592-023-01994-w). Code to apply veloVI and reproduce the findings of the study are publicly available

- veloVI: https://github.com/YosefLab/velovi
- veloVI reproducibility: https://github.com/YosefLab/velovi_reproducibility

Summary

Single-cell RNA sequencing data has enabled studying the cellular heterogeneity of biological processes consistently. To recover the trajectories of the underlying mechanisms, the concept of RNA velocity relates nascent and mature mRNA molecules detected in standard scRNA-seq experiments via a mechanistic model (Section 2.4). However, the

most frequently used models do not carry notions of estimation uncertainties or means to assess if modeling assumptions are met; their inference framework is also tied to a specific splicing model, not easily generalizable if at all.

To address model uncertainty, my collaborators and I have reformulated RNA velocity inference in a deep learning-based setting using recent advances in variational inference (Section 2.4 and Appendix B). Our model veloVI makes two assumptions: (1) A process depending on transcription, splicing, and degradation rates, a latent time and transcriptional state generates nascent and mature mRNA for each gene, and (2) a latent, low dimensional cell representation ties the latent time of genes. My co-authors and I have modeled this process using VAEs: The encoder takes preprocessed unspliced and spliced counts as input, and outputs the cell representation that encodes the latent representation; these latent factors act as the input of a neural network parametrizing latent time. Stochastic gradient descent optimizes the likelihood of unspliced and spliced abundances as a function of the latent time, probabilistic state assignment, and kinetic rates.

Modeling the splicing dynamics with VAEs provides a posterior distribution over velocity estimates, quantifying intrinsic and extrinsic uncertainties of the estimates. The intrinsic uncertainty emerges from repeatedly sampling cell velocity, and extrapolating future cell states based on these samples quantifies extrinsic uncertainty. I have also relied on velocity estimates to define cell-cell transition probabilities and score genes by how well velocities and predicted future states align - the velocity coherence. I showed how these metrics relate to regions of putative fate decisions during pancreatic endocrinogenesis.

veloVI models splicing kinetics via neural networks and can, therefore, assess data not seen during model training. We have reasoned that model fits between the original cell order and a random permutation are comparable if the underlying system is in steady-state or the corresponding gene expression does not contain enough information for robust inference, due to low coverage or high noise, for example. I have applied this permutation score to five positive and four negative control cases, revealing distinct distributions of the gene-specific scores. This baseline comparison will aid in evaluating the applicability of future datasets for RNA velocity analysis.

My collaborators and I have devised veloVI in a model-agnostic way; the framework is, thus, easily extensible to more complex dynamics. I have highlighted this feature via splicing dynamics with time-dependent transcription, allowing for monotonically increasing or decreasing rates. I have validated the improved data fit through decreased mean squared errors on four datasets used in prior work on RNA velocity, and have shown how the model explains linear phase portraits better than the original formulation.

Benchmarking new approaches for RNA velocity routinely relies on visualizing velocity streams in low dimensions or on comparing velocity estimates in cell neighborhoods; the former lacks statistical power and is sensitive to feature selection, the latter assumes similar change of cells with similar gene expression profiles, an assumption violated during fate priming, for example. As an alternative approach, I have developed an evaluation scheme based on the cell cycle, a system for which a ground-truth cellular ordering exists; the given ordering gives a proxy for gene expression change.

To summarize, veloVI is an extensible modeling framework that outperforms previously proposed methods and provides metrics that can aid downstream data analyses.

3.2. Publication 2: CellRank 2: Unified fate mapping in multiview single-cell data

The paper titled *CellRank 2: unified fate mapping in multiview single-cell data* has been published as an Article in *Nature Methods* in 2024. The full citation is

Philipp Weiler [*], Marius Lange [*], Michal Klein, Dana Pe'er, Fabian Theis CellRank 2: unified fate mapping in multiview single-cell data. *Nat Methods* (2024)

Contribution

I conceptualized the study with Marius Lange and Fabian Theis. I designed and implemented the inference scheme for metabolic labeling data, and contributed to the implementation of CellRank 2. I performed all data analyses with input from all co-authors: I recovered the differentiation process of human hematopoiesis and mouse endodermal development, identified a putative progenitor state of medullary thymic epithelial cells, and delineated differentiation trajectories to pinpoint regulatory strategies in an intestinal organoid system; for each dataset, I compared the proposed techniques to an RNA velocity-based workflow. To compare alternative approaches consistently, I developed two metrics and applied them in my data analysis. I wrote the manuscript with Marius Lange, Fabian Theis and Dana Pe'er.

Additional supplementary material

Additional supplementary material is available at the publisher's website (https://doi.org/10.1038/s41592-024-02303-9). Code to apply CellRank 2 and reproduce the findings of the study are publicly available

- CellRank 2: https://github.com/theislab/cellrank
- Data relevant to the study: https://doi.org/10.6084/m9.figshare.c.6843633.v1
- CellRank 2 reproducibility: https://github.com/theislab/cellrank2_reproducibility

Summary

Trajectory inference methods have uncovered numerous biological insights but are typically limited to snapshot scRNA-seq data and cannot include relevant, orthogonal information such as experimental time points, multimodal measurements, RNA velocity, and metabolic labeling (Section 2.3). Although methods incorporating this information exist (Section 2.4 and 2.5), they are tied to a specific modality, rendering them inapplicable when modeling assumptions are violated.

My collaborators and I have developed CellRank 2 to unify fate mapping in a data view independent, robust, and scalable framework, scaling to millions of cells: *Kernels* quantify cell-cell transition probabilities based on which *estimators* identify initial and terminal states, estimate cellular fate, and perform other downstream tasks such as identifying putative lineage drivers (Section 2.6).

Pseudotime is a well-studied method ranking cells relative to each other along a differentiation process (Section 2.3). We have developed the *PseuodtimeKernel* to bias

edges of cell-cell similarity graphs towards increased differentiation potential, allowing us to consistently deduce initial and terminal states and fate probabilities. I applied the PseudotimeKernel to a dataset of human hematopoiesis where RNA velocity inference fails. This analysis included automatically inferring the system's initial and terminal states, and highlighting how CellRank 2 correctly identifies drivers of the plasmacytoid dendritic cell lineage by correlating fate probabilities with gene expression.

Although pseudotime is a powerful approach, its model requirements or assumptions may not always hold; a root state identifying the start of a differentiation process is not always known, for example. In such cases, stemness scores provide an alternative quantification metric (Section 2.3). CytoTRACE computes such a potential based on the assumption that less mature cells express more genes, an assumption validated on overall 42 datasets (Section 2.3). Crucially, the original algorithm failed to scale to large dataset size, though. To make the method applicable to such data nonetheless, we have developed a computationally efficient approach implemented in the CytoTRACEKernel that yields comparable results and scales to millions of cells (Section 2.6). I have studied human embryoid bodies using the CytoTRACEKernel to recover terminal states and fate decisions towards them; my analysis focused on endoderm development for which I recovered lineage drivers and activation patterns.

Experimental time points provide valuable information in non-equilibrium systems with successively emerging cell types. OT has previously been applied in the form of Waddington OT (WOT) to match cells with putative progenitors across time points (Section 2.5 and Appendix E). This approach, however, neglects valuable intra-time point transitions. The RealTimeKernel incorporates both intra and inter-time point information via transcriptomic similarity and OT, respectively, into a single cell-cell transition matrix (Section 2.6). I have validated the importance of considering intra-time point information on datasets of mouse embryonic fibroblasts (MEF) and pharyngeal endoderm development; without intra-time point transitions, not all terminal states of the MEF system were identified. I have recovered the terminal states present in the pharyngeal endoderm development dataset and analyzed cell maturation into medullary thymic epithelial cells (mTECs) in detail. This analysis has revealed a cluster of putative progenitor states and recovered known lineage drivers more accurately than classical WOT.

In addition to ignoring dynamics within a time point, WOT studies fate priming on a discrete domain as it is constrained to distinct time points. The RealTimeKernel, however, enables a continuous view. To make use of this potential, I have developed an inference scheme for a real-time-informed pseudotime and highlighted its utility on the MEF data.

Metabolic labels are an alternative experimental way to generate time-resolved single-cell data (Section 2.1). I have devised a computational method to estimate cell and gene-specific kinetic rates of splicing dynamics (Section 2.4) and have applied it to study mouse intestinal organoids; I have used the estimated velocity similarly to the way CellRank 1 employed RNA velocity via the VelocityKernel. The developed approach identified all terminal states and has ranked known lineage drivers better than an existing approach estimating velocities from metabolic labels, or classical RNA velocity. My analysis also pinpointed regulatory mechanisms - cooperative and destructive - governing fate priming.

To compare different approaches in general, I have developed two metrics: Terminal state identification (TSI) and cross-boundary correctness (CBC). The TSI score quantifies how accurately a given kernel recovers terminal states compared to an optimal

scenario. For each CellRank 2 kernel, I have used the TSI score to show that the used data view leads to better results than relying on RNA velocity. The CBC score is applicable if a differentiation order is known a priori and quantifies how well a given kernel recapitulates these known state transitions; correlating extrapolated state changes with an empirical estimate thereof defines the metric. I have computed the CBC score for the PseudotimeKernel and VelocityKernel and have shown that the pseudotime approach consistently yields statistically better results.

4. Discussion and outlook

Single-cell sequencing protocols have evolved significantly over the last decade, leading to larger and more complex datasets; corresponding computational advances make use of these technological developments to recover cellular trajectories and fate, for example. Although existing methods have proven powerful in many settings, they leave room for improvement: Approaches focus on specific data aspects and do not generalize to newly emerging data modalities, or include restrictive modeling paradigms. This dissertation focused on facilitating RNA velocity inference and unifying fate mapping in a data-view agnostic fashion.

4.1. Discussion

RNA velocity inference through variational inference

RNA velocity has celebrated great success since its introduction[100] and is readily applied to existing scRNA-seq datasets as it does not require changes to experimental workflows. However, existing models make restrictive modeling assumptions, fail to quantify inference uncertainty, and do not easily generalize to more complex but accurate descriptions of splicing dynamics. To overcome these limitations, I developed veloVI, a deep generative model for inferring RNA velocity.

Compared to previous approaches[100,101], veloVI does not employ classical optimization techniques but relies on variational inference and neural network architectures to estimate RNA velocity, instead. This alternative model formulation entails estimation uncertainty, decoupling inference from a specific dynamic model, and reduced sensitivity to data preprocessing[100,103–106,252]. I have shown how veloVI performs favorably to established methods[100,101] and have constructed metrics facilitating downstream data analyses: Intrinsic and extrinsic uncertainty characterize estimates on the level of cells, velocity coherence on a gene level, and the permutation score genes and entire datasets.

Modeling splicing dynamics is an ongoing challenge, with emerging data modalities and more complex models leading to more faithful descriptions of the underlying kinetics - a consistent, statistically significant benchmarking pipeline for model assessment does not exist, though. Instead, model comparisons traditionally rely on comparing projections of high-dimensional vector fields onto low-dimensional data representations, or comparing velocity estimates in cell neighborhoods. I have developed an alternative approach that scores RNA velocity estimates on cell cycle data, a well-studied, unidirectional system for which fluorescent ubiquitination-based cell-cycle indicators define ground-truth cell orderings experimentally[127,128].

While developing veloVI during my dissertation, competing approaches for RNA velocity inference have been proposed: VeloAE[253] uses an autoencoder architecture with graph convolutional networks (GCNs) in the encoder and an attention mechanism in the decoder but defines RNA velocity in the latent embedding; the estimates, therefore,

lack mechanistic interpretation and interpretable links between latent factors and genes. DeepVelo[254], another GCN-based model, includes RNA velocity aspects but relies on pseudotime estimates instead of modeling time explicitly. Other approaches employ neural ODEs[255] or focus on the cell cycle[256,257]. VeloVAE, a conceptually similar approach to veloVI, does not provide metrics aiding RNA velocity analysis; the alternative framework also estimates cell-specific latent times, thereby aggregating simultaneously occurring but orthogonal processes such as the cell cycle and differentiation. Similarly, improvements introduced by veloVAE still result in spurious state transitions with low state uncertainties. The work introducing veloVI[126] includes an exhaustive comparison of veloVI with alternative approaches.

veloVI facilitates RNA velocity analysis through robust, scalable and uncertainty-aware inference, model flexibility, and evaluation metrics. As such, the framework improves upon previous work. However, the current modeling approach still includes restrictive modeling assumptions. First, the default dynamical system assumes constant rates of transcription, splicing and degradation. Second, the model omits gene-gene interactions. Thus, RNA velocity inference fails when these assumptions do not hold approximately. Conversely, alternative methods for estimating cell differentiation may work, motivating a general fate-mapping approach to leverage different vector fields inferred.

General fate mapping framework for single-cell data

Trajectory inference methods reconstruct biological processes from single-cell sequencing data. Dedicated algorithms assign pseudotime or stemness potentials, infer RNA velocity, or leverage temporal information from experimental time points or metabolic labels. Importantly, methods for mapping cell fate are tied to a specific modality and do not easily generalize if at all. Consequently, new data views and modalities require adapting existing methods or conceptualizing new ones.

I have developed CellRank 2 to recover cellular trajectories and fate priming consistently in a unified framework extensible to newly emerging data views. CellRank 2's framework consists of two main parts, separating the inference of cell-cell transitions from their analysis: *Kernels* estimate transition probabilities, and *estimators* use this information to recover biological insight like initial and terminal states or fate probabilities.

Different data aspects allow quantifying cell transitions: The PseudotimeKernel and CytoTRACEKernel bias edges of cell similarity graphs towards increased pseudotime and differentiation potential, respectively, for example. In the accompanying publication of this cumulative thesis[129], I showed how the PseudotimeKernel outperforms an RNA-velocity-based workflow when recovering the lineages of the human hematopoietic system. Similarly, the CytoTRACEKernel improves upon a previously proposed method[99] to scale data analyses to millions of cells. I used the CytoTRACEKernel to study human embryoid body development and the maturation of the endoderm in particular.

Single-cell sequencing assays provide snapshot data of typically asynchronous systems; experimental time points and metabolic labels add an essential temporal component when studying systems not in homeostasis. The RealTimeKernel combines inter-time-point connections via OT with intra-time-point transitions based on cell-cell similarity - a modeling choice that leads to improved performance on a dataset of mouse embryonic fibroblasts[129]; the transition matrix of the RealTimeKernel can further function as the basis for constructing a real-time-informed pseudotime[129]. To compare the proposed ap-

proach to a classical OT-based workflow, I studied medullary thymic epithelial cell maturation. In brief, the CellRank 2 pipeline identified a cluster of putative progenitor states and recovered known lineage drivers and transcription factors more faithfully[129].

Metabolic labels offer an alternative experimental option for adding temporal information to single-cell measurements. I have devised an inference scheme to estimate a vector field based on the metabolic labeling information and applied it to murine intestinal development. Compared to competing approaches, the new method inferred all terminal states and ranked known lineage drivers consistently higher[129]. Recent experimental advances stress the importance of reliable estimation based on metabolically labeled mRNA; improvements include throughput[175,176,258,259], applications to in vivo systems[260,261] and paired alternative modalities[261,262]. Methods developed in parallel to my approach, do not estimate transcription rates and assume constant degradation rates across all cells[258] or do so through post-processing steps and rely on deterministic downstream analyses[263].

Different kernels make different assumptions and are, as a consequence, potentially applicable to different datasets. To compare kernel performance, I have conceived the terminal state identification (TSI) and cross-boundary correctness[264] (CBC) metric; the TSI metric quantifies how well a kernel identifies terminal states compared to an optimal identification scheme, and the CBC score how accurately a kernel recapitulates a priori known cell state transitions. To use different data views at the same time, kernels can be combined via a global weighting. Such combinations enable harvesting the power of complementary views.

CellRank 2 is a robust, modular and scalable framework, extensible by different data views. Existing studies relied on these features to incorporate spatio-temporal[111] and lineage-tracing information[249], or combined kernels to study the developmental processes in epicardioids[265] and to reveal the developmental history during human cortical gyrification[250]. Although CellRank 2-based analyses have recapitulated known and recovered novel biology, the identification of putative driver genes is correlation-based but not causal. Additionally, CellRank 2's approach for resolving terminal states does not reveal the transition paths themselves or the speed of transitions. Future iterations of the CellRank framework will have to address these shortcomings to help describe biological processes more robustly and in greater detail.

4.2. Outlook

Multi-modal RNA velocity

Splicing kinetics neither start with unregulated transcription nor end with spliced mRNA. Instead, chromatin accessibility and gene regulation dictate transcription, and spliced mRNA translates proteins[229]; spatial context and molecular signaling affect the underlying dynamics similarly. However, the RNA velocity model[102] employed by veloVI and related approaches does not consider these additional modalities. A multimodal model is essential to a more complete and accurate description of the dynamics regulating mRNA levels.

Traditional scRNA-seq protocols capture only the transcriptome but more advanced assays include additional information relevant to transcriptional dynamics. Chromatin accessibility, for example, poses an additional, orthogonal view to gene expression, measurable in well-established protocols[49–52], building on the Assay for Transposase Ac-

cessible Chromatin (ATAC). However, these assays do not yield count data. Instead, fragments quantify chromatin accessibility, entailing additional challenges: The dimension of ATAC data is even higher, more sparse and does not translate into interpretable and quantifiable data like gene expression counts; dedicated computational approaches embed RNA and ATAC information into a joint low-dimensional latent space[266–268], construct low-dimensional factors[72,269], or compute metacell aggregates[270]. Defining the distribution generating ATAC data is also not as straightforward as for the gene expression case that relied on biophysical arguments[271–273]. The new data modality, thus, does not directly integrate into the existing modeling framework.

Despite the challenges posed by ATAC data, several computational methods have incorporated the data view into trajectory inference. Pseudotime inference based on neighbor graphs from shared embeddings naturally includes the new modality; I exemplified this approach in my analysis of human hematopoiesis with CellRank 2's PseudotimeKernel[129]. Alternatively, domains of regulatory chromatin (DORCs) define putative future RNA states, thereby offering estimates of cellular change[52,274]; however, this approach relies on pseudobulking and difference vectors derived from low-dimensional data representations.

Two models - to the best of my knowledge - have integrated chromatin information into RNA velocity inference. MultiVelo[275] generalizes the *EM model* to a three-dimensional ODE, modeling splicing kinetics based on chromatin accessibility, and unspliced and spliced RNA; the sum of accessibility at the promoter and linked peaks for a gene defines chromatin accessibility, an estimation not yet validated. The model suffers from similar limitations as the *EM model*, though, such as qualitatively different phase portraits in simulated data compared to real-world examples. Additionally, MultiVelo assumes that chromatin accessibility changes translate into gene expression changes, a simplistic causal relationship between regulatory dynamics. Gene regulation is more complex as multiple, interlinked features confound the influence of chromatin states; scKinetics[276] attempts to model this more accurate depiction of gene regulation to infer gene regulatory relations and cell velocities simultaneously.

Chromatin accessibility is part of the onset of the central dogma of molecular biology, proteins form part of the end. Recent technological advances allow the capturing of related information through sequencing. CITE-Seq[43], for example, captures cell surface proteins in parallel to gene expression. Different approaches tried to include the protein information into the splicing model[277,278] but faced limitations: In previous work, I showed how a straight-forward extension of the steady-state model and additional estimation of protein acceleration leads to spurious transitions - B to CD4+ T cells in cord blood mononuclear cells - caused by excessive data imputation, generating phase portraits artificially[278]. Another challenge is the shift in data distribution between the different modalities. Measurements of cell surface proteins exhibit different properties compared to scRNA-seq measurements and, thus, require different data preprocessing; example steps include centered log-ratio transformation[43] or denoised and scaled by background[279], resulting in different phase portraits[278]. Additionally, these transformations are non-linear and, consequently, require non-trivial reformulations of the dynamical systems.

It is challenging to include cell surface proteins in translation dynamics in general as they may bind to cells long before mRNA transcription sets in; intracellular proteins are likely to correlate better with the mRNA level and production cycle. The development of assays profiling intracellular proteins at the single-cell level either in isolation or in combination with gene expression[48,280,281] may provide the relevant data for a

multimodal model describing translation dynamics but do not eliminate all problems: Assays for paired measurements are technically challenging and, thus, low throughput, and single-cell proteomics measurements yield mass spectrometry instead of count data. Consequently, models including intra-cellular proteins will have to account for varying preprocessing steps, potential data shifts resulting from varying sensitivity and combining mass spectrometry with count measurements in a coherent manner.

Measuring all major quantities comprising splicing kinetics is not yet possible. However, experimental workflows measuring more than two views exist[53,54]. Developing such technologies further will ultimately provide the necessary data but the discussed challenges, *i.e.*, sensitivity, preprocessing, and data representation, still require computational solutions.

Alternative extensions of RNA velocity to additional modalities include metabolic labels and spatial context. Others and I have proposed approaches to include metabolic labeling information[107,258,263], but leave room for improvement: Deep generative modeling will facilitate model uncertainty and scalability, aligning cells along the differentiation process requires latent time inference, and gene regulation inference necessitates more accurate models. Zman-seq[282] provides similar information to metabolic labeling assays by recording transcriptional dynamics through fluorescent pulse labels. These labels approximate the time circulating immune cells have been exposed to a tissue. As such, the given data is similar to the pulse-chase data I relied on for estimating RNA velocity from metabolic labels.

Spatial dependencies and molecular signaling may offer alternative information and priors for modeling gene-gene interactions; spatial assays have matured, reaching subcellular resolution[55–58,283], and methods for estimating communication events computationally exist[284–288]. Although spatial measurements do not distinguish between unspliced and spliced counts, they provide spatial proximity estimates or RNA velocity based on nuclear export[283]. Importantly, spatial data can provide different flavors of RNA velocity: In steady-state systems such as the intestine, cellular change occurs along a spatial dimension that implicitly encodes time; immune infiltration may provide similar information. Modeling non-steady state systems such as normal development will require spatio-temporal resolution to describe the underlying processes accurately, entailing more complex processing steps including batch correction or sample alignment.

To conclude, technological advances will facilitate multi-modal RNA velocity models through orthogonal information. However, these model extensions are not straightforward as they require alternative preprocessing pipelines and corresponding model and inference adaptations, for example.

Characterization of cell state transitions

Terminal cell states define lineages but multiple paths along the phenotypic manifold can lead to a given terminal state, such as the gut tube in murine systems[236,238,239]. CellRank 2's kernels quantify cell-cell transition probabilities, and its GPCCAEstimator infers terminal states and assigns fate probabilities. The estimator does not resolve probable paths towards these terminal states, however. Transition path theory[289–291] may offer a powerful additional step to quantify the rate at which transition paths occur between states identified with the existing estimator.

The speed at which cell state transitions occur is another aspect not captured by the

GPCCAEstimator - the estimator only assigns fate probabilities to observations, not how stable the observation or its cell state is even though rare and fast transitions are relevant in numerous biological systems[91,109,292–294]. Quantifying how transitory paths evolve will give insight into putative fate choice and dysregulation, both relevant information for designing intervention treatments. Combining cell density estimates from Mellon[295], for example, with fate probabilities may provide the relevant information, or help identify and characterize terminal states.

Gene expression change within and across lineages is related to the rate of transitory events and the speed at which they occur. A statistical framework to identify or relate gene expression profiles along lineages with CellRank 2 does not exist, however. Instead, putative driver genes emerge from correlating gene expression with fate probabilities; analyses to identify regulatory relationships are manual and based on visual inspection of expression cascades. For a more consistent and streamlined approach, different methods identify genes differentially expressed across lineages[296–298].

GPfates[296] employs Gaussian mixture models to test differential expression, and BEAM[297] checks if expression changes coincide with branching points; GPfates' fits can potentially yield more powerful insight if combined with fate probabilities assigned by CellRank 2, and fate priming can be characterized better by comparing putative driver genes identified through BEAM to positions and fate probabilities of cells along the differentiation process. However, GPfates and BEAM are not able to identify which parts along trajectories are differentially expressed and cannot identify or characterize patterns. TradeSeq[298] attempts to overcome these limitations by identifying differential expression patterns within and across lineages. The algorithm defines statistical tests that rely on pseudotime values and lineage weights assigned by classical TI methods. Instead, CellRank 2's fate probabilities are an alternative soft assignment to lineages based on different data views. Additionally, more sophisticated tests for comparing gene pattern relationships can pinpoint putative gene regulatory events. Finally, moving beyond patterns in gene expression by incorporating multi-modal dynamics and coupling them with gene patterns is an unanswered but critical problem to be solved.

Improved inference of cell transitions

CellRank 2 unifies fate mapping into a modular, robust, and scalable framework for multiview single-cell data. So far, pseudotime, developmental potential, experimental time points, and metabolic labels inform inference of state change; related work incorporates an alternative flavor of OT for biological systems in equilibrium[299], lineage tracing[249], and spatio-temporal relations[111]. Different views can provide orthogonal information or characterize state transitions that warrant further investigation.

Combining different kernels incorporates alternative views and can improve numerical stability but transition matrices are combined globally, ignoring local nuances. However, a kernel may provide robust and accurate estimates for a subset of a differentiation process but estimates with less confidence or fail in other regions. Aggregating kernels globally ignores such shortcomings and local differences. Instead, combinations based on local weights - cell-wise or for each cell neighborhood or state, for example - can accommodate the advantages of individual views. How to choose weights is an open question: Confidence-based weighting is not a valid option, for instance, as predictors can be confident but wrong; low confidence can instigate more in-depth analyses of the corresponding regions, though.

Prior knowledge is a natural way to inform kernel importance locally; transitions should be silenced with near-zero weights if a kernel violates known biology in a subset of the data, for example. The CBC score connects prior knowledge with cell-cell transitions to quantify how accurately kernels recapitulate known state transitions. These state change-specific scores can, consequently, help define local weights; normalized CBC scores as weights are a straightforward option. If only terminal states but not individual state transitions are known a priori, the TSI score could guide kernel importance, instead; failure to identify a given terminal state should result in downweighing of the corresponding lineage in the kernel, for example.

Permutation-based tests similar to veloVI's permutation score[126] may offer an alternative weighting scheme. Inferred transitions on permuted data reveal the dominating data feature - noise or structure. If transitions under a given inference approach do not change under permutation, the underlying structure is noisy. A putative test reshuffles counts for each cell independently, followed by inference and comparison of the original and permutation-based transitions or vector field. For generative cases, goodness-of-fit metrics can define test scores as exemplified by veloVI's permutation score[126], and correlation otherwise, for example.

Local kernel weighting is only one possible avenue for better estimates of state change. Incorporating additional views is necessary to cover more aspects of biological processes. Important modalities include spatial organization[197,300] or signaling, epigenetic traits including chromatin accessibility[237] and methylation patterns[301,302], or metabolites. Different concepts have attempted to incorporate some of these views but leave room for improvement: Moscot[111] includes spatio-temporal mapping but does not model niche development and omits the sub-cellular resolution of common in-situ sequencing protocols. The same framework includes chromatin accessibility into OT by relying on concatenated latent representations of gene expression and scATAC data. Other approaches extend velocity inference[275,276] or derive mitotic age estimates from scATAC data[250].

Finally, experimental and computational advances will offer data relevant to accurate fate mapping. Lineage tracing[303,304], for example, provides ground-truth state evolution, recoverable after sequencing[305]. Algorithms already incorporate lineage tracing into TI[110,249] but are limited to experimental techniques not applicable to human studies. Mitochondrial lineage tracing[306] provides an intriguing alternative approach in systems where inducing artificial genetic mutations is impossible. Experimental techniques profiling mitochondrial DNA alongside additional modalities exist[53,307,308], but not every system may experience sufficient mutations and the single-cell field lacks computational approaches to reconstruct the corresponding lineage trees automatically.

Spatial omics data

The tissue environment of a cell is an essential factor and driver of its fate, but the TI field traditionally leverages data from classical single-cell assays. Consequently, the spatial organization of cells is lost as the protocols work on dissociated data. Spatial assays[55–58] bridge this gap, but many open challenges remain before TI can incorporate spatial information robustly.

Cell segmentation is a fundamental challenge of single-cell resolved spatial protocols. Although several tools exist[309–311], the lack of ground-truth annotations makes a quantitative benchmark challenging. The shortcomings of existing tools include that they do not work equally well in every tissue and assign many transcripts incorrectly, leading

to noisy expression profiles. As a result, the segmentation step is laborious and, as a result, time-consuming, and downstream analyses are easily confounded by data noise. Improved experimental workflows promise better segmentation through membrane staining. Importantly, staining will not solve the segmentation problem as (1) it is imperfect, (2) the number of stains is limited, making not all cell types identifiable and (3) cells can overlap in space and include multiple nuclei or none at all. Nonetheless, membrane staining will provide an important prior for computational segmentation tools that will have to scale to the increasingly high resolution and size of spatially resolved expression data.

Well-segmented measurements of spatial assays will allow studying biological processes across multiple scales - from subcellular to tissue and organ. The subcellular patterning will elucidate cell-cell variability and identify functionally relevant gene signatures in space, related to cell-cell communication, for example. So far, corresponding features have only been handcrafted[312,313], however, and not described in a data-driven manner through recent advances in deep learning, for example. Relatedly, at the level of cells, spatial distribution and clustering into niches define cell interactions. Methods for describing cellular niches exist[314,315], but the field lacks a proper definition of a niche, adaptable to different contexts; instead, current notions use a fixed number of neighboring cells in space or all cells within a given distance to a reference cell. Describing cellular organization beyond local neighborhoods will improve our understanding of the functions of tissues and organs and the roles of individual cells within and across them. Overall, spatially resolved cell profiles will help describe cross-scale dependencies, interactions and regulation to ultimately describe and compare trajectories within and across scales.

The future of spatial single-cell data will also include experimental advances. Current single-cell resolved high-throughput experiments at spatial resolution focus on the transcriptome. Future advancements will include epigenomic and proteomic information which is currently only possible at scale in dissociated samples. Such additional sources of information will paint a clearer picture of cellular interplay, regulation, and organization within and across scales.

Multimodal data is not limited to cellular features but includes hematoxylin and eosin (H&E) staining of tissues, corresponding annotations, or metadata, including donor information, for example. Incorporating different modalities such as text, audio and video recordings into a single foundational model has recently sparked great enthusiasm in society by generating one modality given another[316,317]. While these models solve tasks such as question-answer queries or image generation, their concept may be generalizable to single-cell biology: In the context of spatial genomics, different modalities include the position of individual molecules, cellular features such as gene expression or pathology slides, to name but a few; predicting gene expression from H&E stained and annotated images is a straightforward application of a foundation model for single-cell data. It is, however, unclear how applicable foundation models are to problem-specific questions in biology. Such models require vast amounts of data for training, but biology is oftentimes driven by rare cell populations. Additionally, robustness and uncertainty quantification are critical in the medical domain, and model hallucination needs to be prevented. But, several data properties present in current genomics data cause this model failure: The data is incomplete and noisy, and it includes inherent biases as datasets from human samples are skewed in their distribution of ethnicity[318]. Both the design and evaluation of future models need to consider these domain-specific challenges before the approaches can be fine-tuned for trajectory inference or cellular fate mapping.

Human cell atlases

The Human Cell Atlas Project aims at defining all cell types of the human body[198]. Existing atlases focus on organs in equilibrium such as the lung[23], brain[24] and heart[319,320] or developing systems like T cell development in the thymus[321,322] or embryonic limb development[197]. The first atlases were defined based on dissociated data, but recent examples consider cellular identity in their spatial context[197,322,323]. However, these studies still rely mostly on spot-based spatial transcriptomics for which in-silico approaches estimate single-cell resolution. Future iterations of existing and new atlases will need to include spatial organization based on *in situ* sequencing and emerging modalities in their analyses and definitions.

Single-cell atlases provide a reference for new datasets to annotate cell states, for example; several tools exist for this task[67,84,85,324]. One key challenge is the lack of Common Coordinate Frameworks (CCFs) and consistent cell type annotations. So far, a CCF exists only for the brain, made partially possible by the organ's highly structured anatomy[24,325]; for less organized organs, defining a CCF will most likely be more challenging.

Atlases of developing systems[2] need to define cellular identity and trajectories, where trajectory annotation includes both its identity and the position of a cell along it. Consequently, such atlases provide an additional layer of information compared to their counterparts for fully developed organs. However, it is currently not clear how to map trajectories onto query datasets - a computational challenge similar to but different from mapping cell states; compared to cell states, mapping entire trajectories needs to consider cell state and position along the process, and assigning the correct trajectory itself, for instance. Putative approaches will describe trajectories in a generative framework that would allow annotating held-out data and mapping it onto the existing reference. Similar to cell typing, confounding factors such as batch effects, for example, will have to be taken care of.

To conclude, this dissertation presented veloVI, a deep generative model for inferring RNA velocity, and CellRank 2, a framework for unified fate mapping in multiview single-cell data. veloVI facilitates RNA velocity analysis through improved and more robust inference, and actionable metrics for interpretable data analyses. The generative formulation can function as a blueprint for future methods inferring cellular state changes and, thus, potentially aid in mapping trajectories onto emerging developmental atlases. Similarly, the model is flexible enough to infer more complex models of splicing dynamics including gene-gene interaction or additional modalities, for instance. CellRank 2 generalizes Markov chain-based TI to arbitrary views, again facilitating data analyses. The proposed algorithm allows utilizing orthogonal data views accompanied by different advantages; the modular, scalable and extensible design further enables rapid development and integration of inference methods based on new modalities. The flexibility and scalability will allow the seamless incorporation of emerging data modalities and ensure applicability to large datasets such as cell atlases and spatial datasets.

References

1. Hatton, I. A., Galbraith, E. D., Merleau, N. S. C., Miettinen, T. P., Smith, B. M. & Shander, J. A. The human cell count and size distribution. *Proceedings of the National Academy of Sciences* (2023).
2. Haniffa, M., Taylor, D., Linnarsson, S., Aronow, B. J., Bader, G. D., Barker, R. A., Camara, P. G., Camp, J. G., Chédotal, A., *et al.* A roadmap for the Human Developmental Cell Atlas. *Nature* (2021).
3. Sasieni, P. D., Shelton, J., Ormiston-Smith, N., Thomson, C. S. & Silcocks, P. B. What is the lifetime risk of developing cancer?: the effect of adjusting for multiple primaries. *British Journal of Cancer* (2011).
4. Platt, A. in *Complete Works of Aristotle, Volume 1* (Princeton University Press, 1985).
5. Lennox, J. in *The Stanford Encyclopedia of Philosophy* (ed Zalta, E. N.) Fall 2021 (Metaphysics Research Lab, Stanford University, 2021).
6. Hooke, R. & Allestry, J. *Micrographia: Or Some Physiological Descriptions Of Minute Bodies Made By Magnifying Glasses* (London: Printed for James Allestry, Printer to the Royal Society, 1667).
7. Emrich, S. J., Barbazuk, W. B., Li, L. & Schnable, P. S. Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Research* (2006).
8. Perou, C. M., Sørlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., *et al.* Molecular portraits of human breast tumours. *Nature* (2000).
9. Shapiro, E., Biezuner, T. & Linnarsson, S. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Reviews Genetics* (2013).
10. Stegle, O., Teichmann, S. A. & Marioni, J. C. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics* (2015).
11. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods* (2013).
12. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Current Protocols in Molecular Biology* (2015).
13. Corces, M. R., Trevino, A. E., Hamilton, E. G., Greenside, P. G., Sinnott-Armstrong, N. A., Vesuna, S., Satpathy, A. T., Rubin, A. J., Montine, K. S., *et al.* An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nature Methods* (2017).
14. Grandi, F. C., Modi, H., Kampman, L. & Corces, M. R. Chromatin accessibility profiling by ATAC-seq. *Nature Protocols* (2022).
15. Wang, G. & Fang, N. in *Methods in Enzymology* (Elsevier, 2012).
16. Schwarzacher, T., Leitch, A. R., Bennett, M. D. & Heslop-Harrison, J. S. In Situ Localization of Parental Genomes in a Wide Hybrid. *Annals of Botany* (1989).
17. Lichtman, J. W. & Conchello, J.-A. Fluorescence microscopy. *Nature Methods* (2005).
18. Gall, J. G. & Pardue, M. L. FORMATION AND DETECTION OF RNA-DNA HYBRID MOLECULES IN CYTOLOGICAL PREPARATIONS. *Proceedings of the National Academy of Sciences* (1969).

19. RUDKIN, G. T. & STOLLAR, B. D. High resolution detection of DNA-RNA hybrids in situ by indirect immunofluorescence. *Nature* (1977).

20. Singer, R. H. & Ward, D. C. Actin gene expression visualized in chicken muscle tissue culture by using in situ hybridization with a biotinated nucleotide analog. *Proceedings of the National Academy of Sciences* (1982).

21. Moses, L. & Pachter, L. Museum of spatial transcriptomics. *Nature Methods* (2022).

22. Altschuler, S. J. & Wu, L. F. Cellular Heterogeneity: Do Differences Make a Difference? *Cell* (2010).

23. Sikkema, L., Ramírez-Suástegui, C., Strobl, D. C., Gillett, T. E., Zappia, L., Madissoon, E., Markov, N. S., Zaragosi, L.-E., Ji, Y., *et al.* An integrated cell atlas of the lung in health and disease. *Nature Medicine* (2023).

24. Siletti, K., Hodge, R., Mossi Albiach, A., Lee, K. W., Ding, S.-L., Hu, L., Lönnerberg, P., Bakken, T., Casper, T., *et al.* Transcriptomic diversity of cell types across the adult human brain. *Science* (2023).

25. Goyal, Y., Busch, G. T., Pillai, M., Li, J., Boe, R. H., Grody, E. I., Chelvanambi, M., Dardani, I. P., Emert, B., *et al.* Diverse clonal fates emerge upon drug treatment of homogeneous cancer cells. *Nature* (2023).

26. Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B. B., *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods* (2009).

27. Buenrostro, J. D., Wu, B., Litzenburger, U. M., Ruff, D., Gonzales, M. L., Snyder, M. P., Chang, H. Y. & Greenleaf, W. J. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* (2015).

28. Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* (2015).

29. Zilionis, R., Nainys, J., Veres, A., Savova, V., Zemmour, D., Klein, A. M. & Mazutis, L. Single-cell barcoding and sequencing using droplet microfluidics. *Nature Protocols* (2016).

30. Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., *et al.* Massively parallel digital transcriptional profiling of single cells. *Nature Communications* (2017).

31. Ramsköld, D., Luo, S., Wang, Y.-C., Li, R., Deng, Q., Faridani, O. R., Daniels, G. A., Khrebtukova, I., Loring, J. F., *et al.* Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nature Biotechnology* (2012).

32. Picelli, S., Björklund, Å. K., Faridani, O. R., Sagasser, S., Winberg, G. & Sandberg, R. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature Methods* (2013).

33. Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., *et al.* Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types. *Science* (2014).

34. Bose, S., Wan, Z., Carr, A., Rizvi, A. H., Vieira, G., Pe'er, D. & Sims, P. A. Scalable microfluidics for single-cell RNA printing and sequencing. *Genome Biology* (2015).

35. Gierahn, T. M., Wadsworth, M. H., Hughes, T. K., Bryson, B. D., Butler, A., Satija, R., Fortune, S., Love, J. C. & Shalek, A. K. Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nature Methods* (2017).

36. Keren-Shaul, H., Kenigsberg, E., Jaitin, D. A., David, E., Paul, F., Tanay, A. & Amit, I. MARS-seq2.0: an experimental and analytical pipeline for indexed sorting combined with single-cell RNA sequencing. *Nature Protocols* (2019).

37. Hagemann-Jensen, M., Ziegenhain, C., Chen, P., Ramsköld, D., Hendriks, G.-J., Larsson, A. J. M., Faridani, O. R. & Sandberg, R. Single-cell RNA counting at allele and isoform resolution using Smart-seq3. *Nature Biotechnology* (2020).

38. Svensson, V., da Veiga Beltrame, E. & Pachter, L. A curated database reveals trends in single-cell transcriptomics. *Database* (2020).

39. Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y.-J., *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* (2005).

40. Angermueller, C., Clark, S. J., Lee, H. J., Macaulay, I. C., Teng, M. J., Hu, T. X., Krueger, F., Smallwood, S. A., Ponting, C. P., *et al.* Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nature Methods* (2016).

41. Hu, Y., Huang, K., An, Q., Du, G., Hu, G., Xue, J., Zhu, X., Wang, C.-Y., Xue, Z., *et al.* Simultaneous profiling of transcriptome and DNA methylome from a single cell. *Genome Biology* (2016).

42. Clark, S. J., Argelaguet, R., Kapourani, C.-A., Stubbs, T. M., Lee, H. J., Alda-Catalinas, C., Krueger, F., Sanguinetti, G., Kelsey, G., *et al.* scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nature Communications* (2018).

43. Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P. K., Swerdlow, H., Satija, R. & Smibert, P. Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods* (2017).

44. Bartosovic, M., Kabbe, M. & Castelo-Branco, G. Single-cell CUT&Tag profiles histone modifications and transcription factors in complex tissues. *Nature Biotechnology* (2021).

45. Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C. P., Jerby-Arnon, L., Marjanovic, N. D., Dionne, D., Burks, T., *et al.* Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* (2016).

46. Mimitou, E. P., Cheng, A., Montalbano, A., Hao, S., Stoeckius, M., Legut, M., Roush, T., Herrera, A., Papalexi, E., *et al.* Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells. *Nature Methods* (2019).

47. Yao, D., Binan, L., Bezney, J., Simonton, B., Freedman, J., Frangieh, C. J., Dey, K., Geiger-Schuller, K., Eraslan, B., *et al.* Scalable genetic screening for regulatory circuits using compressed Perturb-seq. *Nature Biotechnology* (2023).

48. Reimegård, J., Tarbier, M., Danielsson, M., Schuster, J., Baskaran, S., Panagiotou, S., Dahl, N., Friedländer, M. R. & Gallant, C. J. A combined approach for single-cell mRNA and intracellular protein expression analysis. *Communications Biology* (2021).

49. Cao, J., Cusanovich, D. A., Ramani, V., Aghamirzaie, D., Pliner, H. A., Hill, A. J., Daza, R. M., McFaline-Figueroa, J. L., Packer, J. S., *et al.* Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* (2018).

50. Zhu, C., Yu, M., Huang, H., Juric, I., Abnousi, A., Hu, R., Lucero, J., Behrens, M. M., Hu, M., *et al.* An ultra high-throughput method for single-cell joint analysis of open chromatin and transcriptome. *Nature Structural & Molecular Biology* (2019).

51. Chen, S., Lake, B. B. & Zhang, K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nature Biotechnology* (2019).

52. Ma, S., Zhang, B., LaFave, L. M., Earl, A. S., Chiang, Z., Hu, Y., Ding, J., Brack, A., Kartha, V. K., *et al.* Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell* (2020).

53. Mimitou, E. P., Lareau, C. A., Chen, K. Y., Zorzetto-Fernandes, A. L., Hao, Y., Takeshima, Y., Luo, W., Huang, T.-S., Yeung, B. Z., *et al.* Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. *Nature Biotechnology* (2021).

54. Swanson, E., Lord, C., Reading, J., Heubeck, A. T., Genge, P. C., Thomson, Z., Weiss, M. D., Li, X.-j., Savage, A. K., *et al.* Simultaneous trimodal single-cell measurement of transcripts, epitopes, and chromatin accessibility using TEA-seq. *eLife* (2021).

55. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* (2015).

56. He, S., Bhatt, R., Brown, C., Brown, E. A., Buhr, D. L., Chantranuvatana, K., Danaher, P., Dunaway, D., Garrison, R. G., *et al.* High-plex imaging of RNA and proteins at subcellular resolution in fixed tissue by spatial molecular imaging. *Nature Biotechnology* (2022).

57. Chen, A., Liao, S., Cheng, M., Ma, K., Wu, L., Lai, Y., Qiu, X., Yang, J., Xu, J., *et al.* Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. *Cell* (2022).

58. Janesick, A., Shelansky, R., Gottscho, A. D., Wagner, F., Williams, S. R., Rouault, M., Beliakoff, G., Morrison, C. A., Oliveira, M. F., *et al.* High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis. *Nature Communications* (2023).

59. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* (2014).

60. Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W. & Smyth, G. K. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* (2015).

61. Kharchenko, P. V., Silberstein, L. & Scadden, D. T. Bayesian approach to single-cell differential expression analysis. *Nature Methods* (2014).

62. Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K., Slichter, C. K., Miller, H. W., McElrath, M. J., *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology* (2015).

63. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology* (2018).

64. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology* (2015).

65. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology* (2018).

66. Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., Hao, Y., Stoeckius, M., Smibert, P., *et al.* Comprehensive Integration of Single-Cell Data. *Cell* (2019).

67. Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., *et al.* Integrated analysis of multimodal single-cell data. *Cell* (2021).

68. Hao, Y., Stuart, T., Kowalski, M. H., Choudhary, S., Hoffman, P., Hartman, A., Srivastava, A., Molla, G., Madad, S., *et al.* Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nature Biotechnology* (2023).

69. Palla, G., Spitzer, H., Klein, M., Fischer, D., Schaar, A. C., Kuemmerle, L. B., Rybakov, S., Ibarra, I. L., Holmberg, O., *et al.* Squidpy: a scalable framework for spatial omics analysis. *Nature Methods* (2022).

70. Dries, R., Zhu, Q., Dong, R., Eng, C.-H. L., Li, H., Liu, K., Fu, Y., Zhao, T., Sarkar, A., *et al.* Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome Biology* (2021).

71. Bredikhin, D., Kats, I. & Stegle, O. MUON: multimodal omics analysis framework. *Genome Biology* (2022).

72. Granja, J. M., Corces, M. R., Pierce, S. E., Bagdatli, S. T., Choudhry, H., Chang, H. Y. & Greenleaf, W. J. ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nature Genetics* (2021).

73. Stuart, T., Srivastava, A., Madad, S., Lareau, C. A. & Satija, R. Single-cell chromatin state analysis with Signac. *Nature Methods* (2021).

74. Zappia, L. & Theis, F. J. Over 1000 tools reveal trends in the single-cell RNA-seq analysis landscape. *Genome Biology* (2021).

75. Pearson, K. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* (1901).

76. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nature Methods* (2018).

77. Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S. & Theis, F. J. Single-cell RNA-seq denoising using a deep count autoencoder. *Nature Communications* (2019).

78. Kotliar, D., Veres, A., Nagy, M. A., Tabrizi, S., Hodis, E., Melton, D. A. & Sabeti, P. C. Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. *eLife* (2019).

79. Lotfollahi, M., Rybakov, S., Hrovatin, K., Hediyeh-zadeh, S., Talavera-López, C., Misharin, A. V. & Theis, F. J. Biologically informed deep learning to query gene programs in single-cell atlases. *Nature Cell Biology* (2023).

80. Kunes, R. Z., Walle, T., Land, M., Nawy, T. & Pe'er, D. Supervised discovery of interpretable gene programs from single-cell data. *Nature Biotechnology* (2023).

81. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* (2008).

82. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports* (2019).

83. Polański, K., Young, M. D., Miao, Z., Meyer, K. B., Teichmann, S. A. & Park, J.-E. BBKNN: fast batch alignment of single cell transcriptomes. *Bioinformatics* (ed Berger, B.) (2019).

84. Lotfollahi, M., Naghipourfar, M., Luecken, M. D., Khajavi, M., Büttner, M., Wagenstetter, M., Avsec, Ž., Gayoso, A., Yosef, N., *et al.* Mapping single-cell data to reference atlases by transfer learning. *Nature Biotechnology* (2021).

85. De Donno, C., Hediyeh-Zadeh, S., Moinfar, A. A., Wagenstetter, M., Zappia, L., Lotfollahi, M. & Theis, F. J. Population-level integration of single-cell datasets enables multi-scale analysis across samples. *Nature Methods* (2023).

86. Luecken, M. D., Büttner, M., Chaichoompu, K., Danese, A., Interlandi, M., Mueller, M. F., Strobl, D. C., Zappia, L., Dugas, M., *et al.* Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods* (2021).

87. Cadwell, C. R., Palasantza, A., Jiang, X., Berens, P., Deng, Q., Yilmaz, M., Reimer, J., Shen, S., Bethge, M., *et al.* Electrophysiological, transcriptomic and morphologic profiling of single neurons using Patch-seq. *Nature Biotechnology* (2015).

88. Guillaume-Gentil, O., Grindberg, R. V., Kooger, R., Dorwling-Carter, L., Martinez, V., Ossola, D., Pilhofer, M., Zambelli, T. & Vorholt, J. A. Tunable Single-Cell Extraction for Molecular Analyses. *Cell* (2016).

89. Chen, W., Guillaume-Gentil, O., Rainer, P. Y., Gäbelein, C. G., Saelens, W., Gardeux, V., Klaeger, A., Dainese, R., Zachara, M., *et al.* Live-seq enables temporal transcriptomic recording of single cells. *Nature* (2022).

90. Waddington, C. *Principles of Development and Differentiation* (Macmillan, 1966).

91. Bendall, S. C., Davis, K. L., Amir, E.-a. D., Tadmor, M. D., Simonds, E. F., Chen, T. J., Shenfeld, D. K., Nolan, G. P. & Pe'er, D. Single-Cell Trajectory Detection Uncovers Progression and Regulatory Coordination in Human B Cell Development. *Cell* (2014).

92. Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S., *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology* (2014).

93. Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nature Methods* (2016).

94. Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H. A. & Trapnell, C. Reversed graph embedding resolves complex single-cell trajectories. *Nature Methods* (2017).

95. Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D. M., Hill, A. J., Zhang, F., Mundlos, S., Christiansen, L., *et al.* The single-cell transcriptional landscape of mammalian organogenesis. *Nature* (2019).

96. Street, K., Risso, D., Fletcher, R. B., Das, D., Ngai, J., Yosef, N., Purdom, E. & Dudoit, S. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* (2018).

97. Setty, M., Kiseliovas, V., Levine, J., Gayoso, A., Mazutis, L. & Pe'er, D. Characterization of cell fate probabilities in single-cell data with Palantir. *Nature Biotechnology* (2019).

98. Teschendorff, A. E. & Enver, T. Single-cell entropy for accurate estimation of differentiation potency from a cell's transcriptome. *Nature Communications* (2017).

99. Gulati, G. S., Sikandar, S. S., Wesche, D. J., Manjunath, A., Bharadwaj, A., Berger, M. J., Ilagan, F., Kuo, A. H., Hsieh, R. W., *et al.* Single-cell transcriptional diversity is a hallmark of developmental potential. *Science* (2020).

100. La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M. E., Lönnerberg, P., *et al.* RNA velocity of single cells. *Nature* (2018).

101. Bergen, V., Lange, M., Peidli, S., Wolf, F. A. & Theis, F. J. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nature Biotechnology* (2020).

102. Zeisel, A., Köstler, W. J., Molotski, N., Tsai, J. M., Krauthgamer, R., Jacob-Hirsch, J., Rechavi, G., Soen, Y., Jung, S., *et al.* Coupled pre-mRNA and mRNA dynamics unveil operational strategies underlying transcriptional responses to stimuli. *Molecular Systems Biology* (2011).

103. Petukhov, V., Guo, J., Baryawno, N., Severe, N., Scadden, D. T., Samsonova, M. G. & Kharchenko, P. V. dropEst: pipeline for accurate estimation of molecular counts in droplet-based single-cell RNA-seq experiments. *Genome Biology* (2018).

104. Srivastava, A., Malik, L., Smith, T., Sudbery, I. & Patro, R. Alevin efficiently estimates accurate gene abundances from dscRNA-seq data. *Genome Biology* (2019).

105. Melsted, P., Booeshaghi, A. S., Liu, L., Gao, F., Lu, L., Min, K. H., da Veiga Beltrame, E., Hjörleifsson, K. E., Gehring, J., *et al.* Modular, efficient and constant-memory single-cell RNA-seq preprocessing. *Nature Biotechnology* (2021).

106. He, D., Zakeri, M., Sarkar, H., Soneson, C., Srivastava, A. & Patro, R. Alevin-fry unlocks rapid, accurate and memory-frugal quantification of single-cell RNA-seq data. *Nature Methods* (2022).

107. Qiu, X., Zhang, Y., Martin-Rufino, J. D., Weng, C., Hosseinzadeh, S., Yang, D., Pogson, A. N., Hein, M. Y., Hoi (Joseph) Min, K., *et al.* Mapping transcriptomic vector fields of single cells. *Cell* (2022).

108. Fischer, D. S., Fiedler, A. K., Kernfeld, E. M., Genga, R. M. J., Bastidas-Ponce, A., Bakhti, M., Lickert, H., Hasenauer, J., Maehr, R., *et al.* Inferring population dynamics from single-cell RNA-sequencing time series data. *Nature Biotechnology* (2019).

109. Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., Gould, J., Liu, S., Lin, S., *et al.* Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming. *Cell* (2019).

110. Forrow, A. & Schiebinger, G. LineageOT is a unified framework for lineage tracing and trajectory inference. *Nature Communications* (2021).

111. Klein, D., Palla, G., Lange, M., Klein, M., Piran, Z., Gander, M., Meng-Papaxanthos, L., Sterr, M., Bastidas-Ponce, A., *et al.* Mapping cells through time and space with moscot (2023).

112. Weinreb, C. & Klein, A. M. Lineage reconstruction from clonal correlations. *Proceedings of the National Academy of Sciences* (2020).

113. Pevny, L., Simon, M. C., Robertson, E., Klein, W. H., Tsai, S.-F., D'Agati, V., Orkin, S. H. & Costantini, F. Erythroid differentiation in chimaeric mice blocked by a targeted mutation in the gene for transcription factor GATA-1. *Nature* (1991).

114. Weiss, M. J. & Orkin, S. H. Transcription factor GATA-1 permits survival and maturation of erythroid precursors by preventing apoptosis. *Proceedings of the National Academy of Sciences* (1995).

115. Fujiwara, Y., Browne, C. P., Cunniff, K., Goff, S. C. & Orkin, S. H. Arrested development of embryonic red cell precursors in mouse embryos lacking transcription factor GATA-1. *Proceedings of the National Academy of Sciences* (1996).

116. Lange, M., Bergen, V., Klein, M., Setty, M., Reuter, B., Bakhti, M., Lickert, H., Ansari, M., Schniering, J., *et al.* CellRank for directed single-cell fate mapping. *Nature Methods* (2022).

117. Aibar, S., González-Blas, C. B., Moerman, T., Huynh-Thu, V. A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.-C., Geurts, P., *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nature Methods* (2017).

118. Fleck, J. S., Jansen, S. M. J., Wollny, D., Zenk, F., Seimiya, M., Jain, A., Okamoto, R., Santel, M., He, Z., *et al.* Inferring and perturbing cell fate regulomes in human brain organoids. *Nature* (2022).

119. Bravo González-Blas, C., De Winter, S., Hulselmans, G., Hecker, N., Matetovici, I., Christiaens, V., Poovathingal, S., Wouters, J., Aibar, S., *et al.* SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks. *Nature Methods* (2023).

120. Lotfollahi, M., Wolf, F. A. & Theis, F. J. scGen predicts single-cell perturbation responses. *Nature Methods* (2019).

121. Lotfollahi, M., Klimovskaia Susmelj, A., De Donno, C., Hetzel, L., Ji, Y., Ibarra, I. L., Srivatsan, S. R., Naghipourfar, M., Daza, R. M., *et al.* Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology* (2023).

122. Bunne, C., Stark, S. G., Gut, G., del Castillo, J. S., Levesque, M., Lehmann, K.-V., Pelkmans, L., Krause, A. & Rätsch, G. Learning single-cell perturbation responses using neural optimal transport. *Nature Methods* (2023).

123. Kamimoto, K., Stringa, B., Hoffmann, C. M., Jindal, K., Solnica-Krezel, L. & Morris, S. A. Dissecting cell identity via network inference and in silico gene perturbation. *Nature* (2023).

124. Bergen, V., Soldatov, R. A., Kharchenko, P. V. & Theis, F. J. RNA velocity—current challenges and future perspectives. *Molecular Systems Biology* (2021).

125. Gorin, G., Fang, M., Chari, T. & Pachter, L. RNA velocity unraveled. *PLOS Computational Biology* (ed Nie, Q.) (2022).

126. Gayoso, A., Weiler, P., Lotfollahi, M., Klein, D., Hong, J., Streets, A., Theis, F. J. & Yosef, N. Deep generative modeling of transcriptional dynamics for RNA velocity analysis in single cells. *Nature Methods* (2023).

127. Battich, N., Beumer, J., de Barbanson, B., Krenning, L., Baron, C. S., Tanenbaum, M. E., Clevers, H. & van Oudenaarden, A. Sequencing metabolically labeled transcripts in single cells reveals mRNA turnover strategies. *Science* (2020).

128. Mahdessian, D., Cesnik, A. J., Gnann, C., Danielsson, F., Stenström, L., Arif, M., Zhang, C., Le, T., Johansson, F., *et al.* Spatiotemporal dissection of the cell cycle with single-cell proteogenomics. *Nature* (2021).

129. Weiler, P., Lange, M., Klein, M., Pe'er, D. & Theis, F. CellRank 2: unified fate mapping in multiview single-cell data. *Nature Methods* (2024).

130. Wolock, S. L., Lopez, R. & Klein, A. M. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Systems* (2019).

131. Yang, S., Corbett, S. E., Koga, Y., Wang, Z., Johnson, W. E., Yajima, M. & Campbell, J. D. Decontamination of ambient RNA in single-cell RNA-seq with DecontX. *Genome Biology* (2020).

132. Young, M. D. & Behjati, S. SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. *GigaScience* (2020).

133. Potter, A. S. & Steven Potter, S. in *Kidney Organogenesis* (Springer New York, 2019).

134. Vieira Braga, F. A. & Miragaia, R. J. in *Single Cell Methods* (Springer New York, 2019).

135. Edinger, A. L. & Thompson, C. B. Death by design: apoptosis, necrosis and autophagy. *Current Opinion in Cell Biology* (2004).

136. Vachon, P. H. in *Colorectal Cancer* (Springer New York, 2018).

137. Tuttle, J. B., Suszkiw, J. B. & Ard, M. Long-term survival and development of dissociated parasympathetic neurons in culture. *Brain Research* (1980).

138. Sawamura, Y., Abe, H., Aida, T., Hosokawa, M. & Kobayashi, H. Isolation and in vitro growth of glioma-infiltrating lymphocytes, and an analysis of their surface phenotypes. *Journal of Neurosurgery* (1988).

139. Singh, S. K., Hawkins, C., Clarke, I. D., Squire, J. A., Bayani, J., Hide, T., Henkelman, R. M., Cusimano, M. D. & Dirks, P. B. Identification of human brain tumour initiating cells. *Nature* (2004).

140. Panchision, D. M., Chen, H.-L., Pistollato, F., Papini, D., Ni, H.-T. & Hawley, T. S. Optimized Flow Cytometric Analysis of Central Nervous System Tissue Reveals Novel Functional Relationships Among Cells Expressing CD133, CD15, and CD24. *Stem Cells* (2007).

141. Yu, G., Floyd, Z. E., Wu, X., Halvorsen, Y.-D. C. & Gimble, J. M. in *Methods in Molecular Biology* (Humana Press, 2010).

142. Yan, Y., Xu, Y., Gao, Y.-Y., Zong, Z.-H., Zhang, Q., Li, C. & Wang, H.-Q. Implication of 14-3-3ε and 14-3-3θ/τ in proteasome inhibition-induced apoptosis of glioma cells. *Cancer Science* (2012).

143. Barkauskas, C. E., Cronce, M. J., Rackley, C. R., Bowie, E. J., Keene, D. R., Stripp, B. R., Randell, S. H., Noble, P. W. & Hogan, B. L. Type 2 alveolar cells are stem cells in adult lung. *Journal of Clinical Investigation* (2013).

144. Patel, A. P., Tirosh, I., Trombetta, J. J., Shalek, A. K., Gillespie, S. M., Wakimoto, H., Cahill, D. P., Nahed, B. V., Curry, W. T., *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* (2014).

145. Hayashida, Y., Partida, G. J. & Ishida, A. T. Dissociation of retinal ganglion cells without enzymes. *Journal of Neuroscience Methods* (2004).

146. Vorobjev, V. S. Vibrodissociation of sliced mammalian nervous tissue. *Journal of Neuroscience Methods* (1991).

147. Daoust, R. Localization of deoxyribonuclease in tissue sections. *Experimental Cell Research* (1957).

148. Gomez, G. G. & Kruse, C. A. in *Cancer Cell Culture* (Humana Press).

149. Biotec, M. *Neural Tissue Dissociation Kit* ().

150. Denisenko, E., Guo, B. B., Jones, M., Hou, R., de Kock, L., Lassmann, T., Poppe, D., Clément, O., Simmons, R. K., *et al.* Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows. *Genome Biology* (2020).

151. Van den Brink, S. C., Sage, F., Vértesy, Á., Spanjaard, B., Peterson-Maduro, J., Baron, C. S., Robin, C. & van Oudenaarden, A. Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. *Nature Methods* (2017).

152. Adam, M., Potter, A. S. & Potter, S. S. Psychrophilic proteases dramatically reduce single cell RNA-seq artifacts: A molecular atlas of kidney development. *Development* (2017).

153. Fordyce, S. L., Kampmann, M.-L., van Doorn, N. L. & Gilbert, M. T. P. Long-term RNA persistence in postmortem contexts. *Investigative Genetics* (2013).

154. Chanfreau, G. F. in *The Enzymes* (Elsevier, 2017).

155. Clark, D. P. & Pazdernik, N. J. in *Molecular Biology* (Elsevier, 2013).

156. Ziegenhain, C., Vieth, B., Parekh, S., Reinius, B., Guillaumet-Adkins, A., Smets, M., Leonhardt, H., Heyn, H., Hellmann, I., *et al.* Comparative Analysis of Single-Cell RNA Sequencing Methods. *Molecular Cell* (2017).

157. Ziegenhain, C., Vieth, B., Parekh, S., Hellmann, I. & Enard, W. Quantitative single-cell transcriptomics. *Briefings in Functional Genomics* (2018).

158. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences* (1977).

159. Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, J. C., Hutchison, C. A., Slocombe, P. M. & Smith, M. Nucleotide sequence of bacteriophage φX174 DNA. *Nature* (1977).

160. Ronaghi, M., Uhlén, M. & Nyrén, P. A Sequencing Method Based on Real-Time Pyrophosphate. *Science* (1998).

161. Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* (2008).

162. Deamer, D., Akeson, M. & Branton, D. Three decades of nanopore sequencing. *Nature Biotechnology* (2016).

163. Jain, M., Olsen, H. E., Paten, B. & Akeson, M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology* (2016).

164. Yang, E., van Nimwegen, E., Zavolan, M., Rajewsky, N., Schroeder, M., Magnasco, M. & Darnell, J. E. Decay Rates of Human mRNAs: Correlation With Functional Characteristics and Sequence Attributes. *Genome Research* (2003).

165. Dölken, L., Ruzsics, Z., Rädle, B., Friedel, C. C., Zimmer, R., Mages, J., Hoffmann, R., Dickinson, P., Forster, T., *et al.* High-resolution gene expression profiling for simultaneous kinetic parameter analysis of RNA synthesis and decay. *RNA* (2008).

166. Cleary, M. D., Meiering, C. D., Jan, E., Guymon, R. & Boothroyd, J. C. Biosynthetic labeling of RNA with uracil phosphoribosyltransferase allows cell-specific microarray analysis of mRNA synthesis and decay. *Nature Biotechnology* (2005).

167. Miller, M. R., Robinson, K. J., Cleary, M. D. & Doe, C. Q. TU-tagging: cell type-specific RNA isolation from intact complex tissues. *Nature Methods* (2009).

168. Hida, N., Aboukilila, M. Y., Burow, D. A., Paul, R., Greenberg, M. M., Fazio, M., Beasley, S., Spitale, R. C. & Cleary, M. D. EC-tagging allows cell type-specific RNA analysis. *Nucleic Acids Research* (2017).

169. Riml, C., Amort, T., Rieder, D., Gasser, C., Lusser, A. & Micura, R. Osmium-Mediated Transformation of 4-Thiouridine to Cytidine as Key To Study RNA Dynamics by Sequencing. *Angewandte Chemie International Edition* (2017).

170. Kofoed, R. H., Betzer, C., Lykke-Andersen, S., Molska, E. & Jensen, P. H. Investigation of RNA Synthesis Using 5-Bromouridine Labelling and Immunoprecipitation. *Journal of Visualized Experiments* (2018).

171. Schofield, J. A., Duffy, E. E., Kiefer, L., Sullivan, M. C. & Simon, M. D. TimeLapse-seq: adding a temporal dimension to RNA sequencing through nucleoside recoding. *Nature Methods* (2018).

172. Neumann, T., Herzog, V. A., Muhar, M., von Haeseler, A., Zuber, J., Ameres, S. L. & Rescheneder, P. Quantification of experimentally induced nucleotide conversions in high-throughput sequencing datasets. *BMC Bioinformatics* (2019).

173. Kawata, K., Wakida, H., Yamada, T., Taniue, K., Han, H., Seki, M., Suzuki, Y. & Akimitsu, N. Metabolic labeling of RNA using multiple ribonucleoside analogs enables the simultaneous evaluation of RNA synthesis and degradation rates. *Genome Research* (2020).

174. Erhard, F., Baptista, M. A. P., Krammer, T., Hennig, T., Lange, M., Arampatzi, P., Jürges, C. S., Theis, F. J., Saliba, A.-E., *et al.* scSLAM-seq reveals core features of transcription dynamics in single cells. *Nature* (2019).

175. Cao, J., Zhou, W., Steemers, F., Trapnell, C. & Shendure, J. Sci-fate characterizes the dynamics of gene expression in single cells. *Nature Biotechnology* (2020).

176. Qiu, Q., Hu, P., Qiu, X., Govek, K. W., Cámara, P. G. & Wu, H. Massively parallel and time-resolved RNA sequencing in single cells with scNT-seq. *Nature Methods* (2020).

177. Jürges, C., Dölken, L. & Erhard, F. Dissecting newly transcribed and old RNA using GRAND-SLAM. *Bioinformatics* (2018).

178. Erhard, F., Saliba, A.-E., Lusser, A., Toussaint, C., Hennig, T., Prusty, B. K., Kirschenbaum, D., Abadie, K., Miska, E. A., *et al.* Time-resolved single-cell RNA-seq using metabolic RNA labelling. *Nature Reviews Methods Primers* (2022).

179. Amezquita, R. A., Lun, A. T. L., Becht, E., Carey, V. J., Carpp, L. N., Geistlinger, L., Marini, F., Rue-Albrecht, K., Risso, D., *et al.* Orchestrating single-cell analysis with Bioconductor. *Nature Methods* (2019).

180. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology* (2019).

181. Heumos, L., Schaar, A. C., Lance, C., Litinetskaya, A., Drost, F., Zappia, L., Lücken, M. D., Strobl, D. C., Henao, J., *et al.* Best practices for single-cell analysis across modalities. *Nature Reviews Genetics* (2023).

182. Ahlmann-Eltze, C. & Huber, W. Comparison of transformations for single-cell RNA-seq data. *Nature Methods* (2023).

183. Brennecke, P., Anders, S., Kim, J. K., Kołodziejczyk, A. A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S. A., *et al.* Accounting for technical noise in single-cell RNA-seq experiments. *Nature Methods* (2013).

184. Vallejos, C. A., Marioni, J. C. & Richardson, S. BASiCS: Bayesian Analysis of Single-Cell Sequencing Data. *PLOS Computational Biology* (ed Morris, Q.) (2015).

185. Buettner, F., Natarajan, K. N., Casale, F. P., Proserpio, V., Scialdone, A., Theis, F. J., Teichmann, S. A., Marioni, J. C. & Stegle, O. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature Biotechnology* (2015).

186. Lun, A. T., McCarthy, D. J. & Marioni, J. C. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Research* (2016).

187. Chen, H.-I. H., Jin, Y., Huang, Y. & Chen, Y. Detection of high variability in gene expression from single-cell RNA-seq profiling. *BMC Genomics* (2016).

188. Brbić, M., Zitnik, M., Wang, S., Pisco, A. O., Altman, R. B., Darmanis, S. & Leskovec, J. MARS: discovering novel cell types across heterogeneous single-cell experiments. *Nature Methods* (2020).

189. Xu, C., Lopez, R., Mehlman, E., Regier, J., Jordan, M. I. & Yosef, N. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Molecular Systems Biology* (2021).

190. Domínguez Conde, C., Xu, C., Jarvis, L. B., Rainbow, D. B., Wells, S. B., Gomes, T., Howlett, S. K., Suchanek, O., Polanski, K., *et al.* Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science* (2022).

191. Xu, C., Prete, M., Webb, S., Jardine, L., Stewart, B. J., Hoo, R., He, P., Meyer, K. B. & Teichmann, S. A. Automatic cell-type harmonization and integration across Human Cell Atlas datasets. *Cell* (2023).

192. Weinreb, C., Wolock, S., Tusi, B. K., Socolovsky, M. & Klein, A. M. Fundamental limits on dynamic inference from single-cell snapshots. *Proceedings of the National Academy of Sciences* (2018).

193. Tritschler, S., Büttner, M., Fischer, D. S., Lange, M., Bergen, V., Lickert, H. & Theis, F. J. Concepts and limitations for learning developmental trajectories from single cell genomics. *Development* (eds Klein, A. & Treutlein, B.) (2019).

194. Xi, N. M. & Li, J. J. Benchmarking Computational Doublet-Detection Methods for Single-Cell RNA Sequencing Data. *Cell Systems* (2021).

195. Wagner, A., Regev, A. & Yosef, N. Revealing the vectors of cellular identity with single-cell genomics. *Nature Biotechnology* (2016).

196. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology* (2018).

197. Zhang, B., He, P., Lawrence, J. E. G., Wang, S., Tuck, E., Williams, B. A., Roberts, K., Kleshchevnikov, V., Mamanova, L., *et al.* A human embryonic limb cell atlas resolved in space and time. *Nature* (2023).

198. Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., *et al.* The Human Cell Atlas. *eLife* (2017).

199. Setty, M., Tadmor, M. D., Reich-Zeliger, S., Angel, O., Salame, T. M., Kathail, P., Choi, K., Bendall, S., Friedman, N., *et al.* Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nature Biotechnology* (2016).

200. Van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A. J., Burdziak, C., Moon, K. R., Chaffer, C. L., *et al.* Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell* (2018).

201. Weber, L. M. & Robinson, M. D. Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data. *Cytometry Part A* (2016).

202. Duò, A., Robinson, M. D. & Soneson, C. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Research* (2018).

203. Freytag, S., Tian, L., Lönnstedt, I., Ng, M. & Bahlo, M. Comparison of clustering tools in R for medium-sized 10x Genomics single-cell RNA-sequencing data. *F1000Research* (2018).

204. Kiselev, V. Y., Andrews, T. S. & Hemberg, M. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nature Reviews Genetics* (2019).

205. Student. The Probable Error of a Mean. *Biometrika* (1908).

206. Wilcoxon, F. Individual Comparisons by Ranking Methods. *Biometrics Bulletin* (1945).

207. Mann, H. B. & Whitney, D. R. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics* (1947).

208. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biology* (2010).

209. McDavid, A., Finak, G., Chattopadyay, P. K., Dominguez, M., Lamoreaux, L., Ma, S. S., Roederer, M. & Gottardo, R. Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments. *Bioinformatics* (2012).

210. Coifman, R. R., Lafon, S., Lee, A. B., Maggioni, M., Nadler, B., Warner, F. & Zucker, S. W. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the National Academy of Sciences* (2005).

211. Coifman, R. R. & Lafon, S. Diffusion maps. *Applied and Computational Harmonic Analysis* (2006).

212. Haghverdi, L., Buettner, F. & Theis, F. J. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics* (2015).

213. Van der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research* (2008).

214. McInnes, L., Healy, J. & Melville, J. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction* 2018.

215. Lloyd, S. Least squares quantization in PCM. *IEEE Transactions on Information Theory* (1982).

216. MacQueen, J. B. *Some Methods for Classification and Analysis of MultiVariate Observations* in *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability* (eds Cam, L. M. L. & Neyman, J.) (University of California Press, 1967).

217. Pettie, S. & Ramachandran, V. An optimal minimum spanning tree algorithm. *Journal of the ACM* (2002).

218. Wolf, F. A., Hamey, F. K., Plass, M., Solana, J., Dahlin, J. S., Göttgens, B., Rajewsky, N., Simon, L. & Theis, F. J. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biology* (2019).

219. Mukherjee, S., Heath, L., Preuss, C., Jayadev, S., Garden, G. A., Greenwood, A. K., Sieberts, S. K., De Jager, P. L., Ertekin-Taner, N., *et al.* Molecular estimation of neurodegeneration pseudotime in older brains. *Nature Communications* (2020).

220. Strauß, M. E., Reid, J. E. & Wernisch, L. GPseudoRank: a permutation sampler for single cell orderings. *Bioinformatics* (ed Berger, B.) (2018).

221. Campbell, K. R. & Yau, C. A descriptive marker gene approach to single-cell pseudotime inference. *Bioinformatics* (ed Birol, I.) (2018).

222. Ahmed, S., Rattray, M. & Boukouvalas, A. GrandPrix: scaling up the Bayesian GPLVM for single-cell data. *Bioinformatics* (ed Stegle, O.) (2018).

223. Lin, C. & Bar-Joseph, Z. Continuous-state HMMs for modeling time-series single-cell RNA-Seq data. *Bioinformatics* (ed Kelso, J.) (2019).

224. Grün, D., Lyubimova, A., Kester, L., Wiebrands, K., Basak, O., Sasaki, N., Clevers, H. & van Oudenaarden, A. Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* (2015).

225. Grün, D., Muraro, M. J., Boisset, J.-C., Wiebrands, K., Lyubimova, A., Dharmadhikari, G., van den Born, M., van Es, J., Jansen, E., *et al.* De Novo Prediction of Stem Cell Identity using Single-Cell Transcriptome Data. *Cell Stem Cell* (2016).

226. Welch, J. D., Hartemink, A. J. & Prins, J. F. SLICER: inferring branched, nonlinear cellular trajectories from single cell RNA-seq data. *Genome Biology* (2016).

227. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nature Biotechnology* (2019).

228. Deconinck, L., Cannoodt, R., Saelens, W., Deplancke, B. & Saeys, Y. Recent advances in trajectory inference from single-cell omics data. *Current Opinion in Systems Biology* (2021).

229. Crick, F. H. On protein synthesis. *Symp Soc Exp Biol* (1958).

230. Li, T. On the Mathematics of RNA Velocity I: Theoretical Analysis. *CSIAM Transactions on Applied Mathematics* (2021).

231. Barile, M., Imaz-Rosshandler, I., Inzani, I., Ghazanfar, S., Nichols, J., Marioni, J. C., Guibentif, C. & Göttgens, B. Coordinated changes in gene expression kinetics underlie both mouse and human erythroid maturation. *Genome Biology* (2021).

232. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning* (Springer New York, 2009).

233. Blei, D. M., Kucukelbir, A. & McAuliffe, J. D. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association* (2017).

234. Kingma, D. P. & Welling, M. *Auto-Encoding Variational Bayes* 2013.

235. Kingma, D. P. & Ba, J. *Adam: A Method for Stochastic Optimization* 2014.

236. Pijuan-Sala, B., Griffiths, J. A., Guibentif, C., Hiscock, T. W., Jawaid, W., Calero-Nieto, F. J., Mulas, C., Ibarra-Soria, X., Tyser, R. C. V., *et al.* A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* (2019).

237. Argelaguet, R., Lohoff, T., Li, J. G., Nakhuda, A., Drage, D., Krueger, F., Velten, L., Clark, S. J. & Reik, W. Decoding gene regulation in the mouse embryo using single-cell multi-omics (2022).

238. Qiu, C., Martin, B. K., Welsh, I. C., Daza, R. M., Le, T.-M., Huang, X., Nichols, E. K., Taylor, M. L., Fulton, O., *et al.* A single-cell time-lapse of mouse prenatal development from gastrula to birth. *Nature* (2024).

239. Nowotschin, S., Setty, M., Kuo, Y.-Y., Liu, V., Garg, V., Sharma, R., Simon, C. S., Saiz, N., Gardner, R., *et al.* The emergent landscape of the mouse gut endoderm at single-cell resolution. *Nature* (2019).

240. Cuturi, M. *Sinkhorn Distances: Lightspeed Computation of Optimal Transportation Distances* 2013. arXiv: 1306.0895 [stat.ML].

241. Kullback, S. & Leibler, R. A. On Information and Sufficiency. *The Annals of Mathematical Statistics* (1951).

242. Reuter, B., Weber, M., Fackeldey, K., Röblitz, S. & Garcia, M. E. Generalized Markov State Modeling Method for Nonequilibrium Biomolecular Dynamics: Exemplified on Amyloid β Conformational Dynamics Driven by an Oscillating Electric Field. *Journal of Chemical Theory and Computation* (2018).

243. Reuter, B., Fackeldey, K. & Weber, M. Generalized Markov modeling of nonreversible molecular kinetics. *The Journal of Chemical Physics* (2019).

244. Reuter, B. *Generalisierte Markov-Modellierung: Modellierung irreversibler β-Amyloid-Peptid-Dynamik unter Mikrowelleneinfluss* (Springer Fachmedien Wiesbaden, 2020).

245. Röblitz, S. & Weber, M. Fuzzy spectral clustering by PCCA+: application to Markov state models and data classification. *Advances in Data Analysis and Classification* (2013).

246. Ma, J., Zhao, J., Tian, J., Bai, X. & Tu, Z. Regularized vector field learning with sparse approximation for mismatch removal. *Pattern Recognition* (2013).

247. Wang, J., Zhang, K., Xu, L. & Wang, E. Quantifying the Waddington landscape and biological paths for development and differentiation. *Proceedings of the National Academy of Sciences* (2011).

248. Perez-Carrasco, R., Guerrero, P., Briscoe, J. & Page, K. M. Intrinsic Noise Profoundly Alters the Dynamics and Steady State of Morphogen-Controlled Bistable Genetic Switches. *PLOS Computational Biology* (ed Tusscher, T.) (2016).

249. Lange, M., Piran, Z., Klein, M., Spanjaard, B., Klein, D., Junker, J. P., Theis, F. J. & Nitzan, M. Mapping lineage-traced cells across time points with moslin (2023).

250. Xiao, Y., Jin, W., Ju, L., Fu, J., Wang, G., Yu, M., Chen, F., Qian, K., Wang, X., *et al.* Tracking single-cell evolution using clock-like chromatin accessibility loci. *Nature Biotechnology* (2024).

251. Stassen, S. V., Yip, G. G. K., Wong, K. K. Y., Ho, J. W. K. & Tsia, K. K. Generalized and scalable trajectory inference in single-cell omics data with VIA. *Nature Communications* (2021).

252. Soneson, C., Srivastava, A., Patro, R. & Stadler, M. B. Preprocessing choices affect RNA velocity results for droplet scRNA-seq data. *PLOS Computational Biology* (ed Li, M.) (2021).

253. Qiao, C. & Huang, Y. Representation learning of RNA velocity reveals robust cell transitions. *Proceedings of the National Academy of Sciences* (2021).

254. Cui, H., Maan, H., Vladoiu, M. C., Zhang, J., Taylor, M. D. & Wang, B. DeepVelo: deep learning extends RNA velocity to multi-lineage systems with cell-specific kinetics. *Genome Biology* (2024).

255. Chen, Z., King, W. C., Hwang, A., Gerstein, M. & Zhang, J. DeepVelo: Single-cell transcriptomic deep velocity field learning with neural ordinary differential equations. *Science Advances* (2022).

256. Riba, A., Oravecz, A., Durik, M., Jiménez, S., Alunni, V., Cerciat, M., Jung, M., Keime, C., Keyes, W. M., *et al.* Cell cycle gene regulation dynamics revealed by RNA velocity and deep-learning. *Nature Communications* (2022).

257. Lederer, A. R., Leonardi, M., Talamanca, L., Herrera, A., Droin, C., Khven, I., Carvalho, H. J., Valente, A., Mantes, A. D., *et al.* Statistical inference with a manifold-constrained RNA velocity model uncovers cell cycle speed modulations (2024).

258. Maizels, R. J., Snell, D. M. & Briscoe, J. Deep dynamical modelling of developmental trajectories with temporal transcriptomics (2023).

259. De Jonghe, J., Kaminski, T. S., Morse, D. B., Tabaka, M., Ellermann, A. L., Kohler, T. N., Amadei, G., Handford, C. E., Findlay, G. M., *et al.* spinDrop: a droplet microfluidic platform to maximise single-cell sequencing information content. *Nature Communications* (2023).

260. Van't Sant, L. J., White, J. J., Hoeijmakers, J. H. J., Vermeij, W. P. & Jaarsma, D. In vivo 5-ethynyluridine (EU) labelling detects reduced transcription in Purkinje cell degeneration mouse mutants, but can itself induce neurodegeneration. *Acta Neuropathologica Communications* (2021).

261. Mitic, N., Neuschulz, A., Spanjaard, B., Schneider, J., Fresmann, N., Novoselc, K. T., Strunk, T., Münster, L., Olivares-Chauvet, P., *et al.* Dissecting the spatiotemporal diversity of adult neural stem cells. *Molecular Systems Biology* (2024).

262. Ren, J., Zhou, H., Zeng, H., Wang, C. K., Huang, J., Qiu, X., Sui, X., Li, Q., Wu, X., *et al.* Spatiotemporally resolved transcriptomics reveals the subcellular RNA kinetic landscape. *Nature Methods* (2023).

263. Peng, Q., Qiu, X. & Li, T. Storm: Incorporating transient stochastic dynamics to infer the RNA velocity with metabolic labeling information (2023).

264. Gao, M., Qiao, C. & Huang, Y. UniTVelo: temporally unified RNA velocity reinforces single-cell trajectory inference. *Nature Communications* (2022).

265. Meier, A. B., Zawada, D., De Angelis, M. T., Martens, L. D., Santamaria, G., Zengerle, S., Nowak-Imialek, M., Kornherr, J., Zhang, F., *et al.* Epicardioid single-cell genomics uncovers principles of human epicardium biology in heart development and disease. *Nature Biotechnology* (2023).

266. Ashuach, T., Gabitto, M. I., Koodli, R. V., Saldi, G.-A., Jordan, M. I. & Yosef, N. MultiVI: deep generative model for the integration of multimodal data. *Nature Methods* (2023).

267. Lotfollahi, M., Litinetskaya, A. & Theis, F. J. Multigrate: single-cell multi-omic data integration (2022).

268. Cao, Z.-J. & Gao, G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nature Biotechnology* (2022).

269. Bravo González-Blas, C., Minnoye, L., Papasokrati, D., Aibar, S., Hulselmans, G., Christiaens, V., Davie, K., Wouters, J. & Aerts, S. cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. *Nature Methods* (2019).

270. Persad, S., Choo, Z.-N., Dien, C., Sohail, N., Masilionis, I., Chaligné, R., Nawy, T., Brown, C. C., Sharma, R., *et al.* SEACells infers transcriptional and epigenomic cellular states from single-cell genomics data. *Nature Biotechnology* (2023).

271. Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y. & Tyagi, S. Stochastic mRNA Synthesis in Mammalian Cells. *PLoS Biology* (ed Schibler, U.) (2006).

272. Grün, D., Kester, L. & van Oudenaarden, A. Validation of noise models for single-cell transcriptomics. *Nature Methods* (2014).

273. Martens, L. D., Fischer, D. S., Yépez, V. A., Theis, F. J. & Gagneur, J. Modeling fragment counts improves single-cell ATAC-seq analysis. *Nature Methods* (2023).

274. Mitra, S., Malik, R., Wong, W., Rahman, A., Hartemink, A. J., Pritykin, Y., Dey, K. K. & Leslie, C. S. Single-cell multi-ome regression models identify functional and disease-associated enhancers and enable chromatin potential analysis. *Nature Genetics* (2024).

275. Li, C., Virgilio, M. C., Collins, K. L. & Welch, J. D. Multi-omic single-cell velocity models epigenome-transcriptome interactions and improves cell fate prediction. *Nature Biotechnology* (2022).

276. Burdziak, C., Zhao, C. J., Haviv, D., Alonso-Curbelo, D., Lowe, S. W. & Pe'er, D. scKINETICS: inference of regulatory velocity with single-cell transcriptomics data. *Bioinformatics* (2023).

277. Gorin, G., Svensson, V. & Pachter, L. Protein velocity and acceleration from single-cell multiomics experiments. *Genome Biology* (2020).

278. Weiler, P., Maddocks, J. H. & Theis, F. *Protein Velocity in Single Cells using Multi-Omics Modelling* PhD thesis (Master's thesis, EPFL, TU Munich, 2021).

279. Mulè, M. P., Martins, A. J. & Tsang, J. S. Normalizing and denoising protein expression data from droplet-based single cell profiling. *Nature Communications* (2022).

280. Mund, A., Coscia, F., Kriston, A., Hollandi, R., Kovács, F., Brunner, A.-D., Migh, E., Schweizer, L., Santos, A., *et al.* Deep Visual Proteomics defines single-cell identity and heterogeneity. *Nature Biotechnology* (2022).

281. Brunner, A.-D., Thielert, M., Vasilopoulou, C., Ammar, C., Coscia, F., Mund, A., Hoerning, O. B., Bache, N., Apalategui, A., *et al.* Ultra-high sensitivity mass spectrometry quantifies single-cell proteome changes upon perturbation. *Molecular Systems Biology* (2022).

282. Kirschenbaum, D., Xie, K., Ingelfinger, F., Katzenelenbogen, Y., Abadie, K., Look, T., Sheban, F., Phan, T. S., Li, B., *et al.* Time-resolved single-cell transcriptomics defines immune trajectories in glioblastoma. *Cell* (2024).

283. Xia, C., Fan, J., Emanuel, G., Hao, J. & Zhuang, X. Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression. *Proceedings of the National Academy of Sciences* (2019).

284. Hou, R., Denisenko, E., Ong, H. T., Ramilowski, J. A. & Forrest, A. R. R. Predicting cell-to-cell communication networks using NATMI. *Nature Communications* (2020).

285. Efremova, M., Vento-Tormo, M., Teichmann, S. A. & Vento-Tormo, R. CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes. *Nature Protocols* (2020).

286. Jin, S., Guerrero-Juarez, C. F., Zhang, L., Chang, I., Ramos, R., Kuan, C.-H., Myung, P., Plikus, M. V. & Nie, Q. Inference and analysis of cell-cell communication using CellChat. *Nature Communications* (2021).

287. Fischer, D. S., Schaar, A. C. & Theis, F. J. Modeling intercellular communication in tissues using spatial graphs of cells. *Nature Biotechnology* (2022).

288. Dimitrov, D., Türei, D., Garrido-Rodriguez, M., Burmedi, P. L., Nagai, J. S., Boys, C., Ramirez Flores, R. O., Kim, H., Szalai, B., *et al.* Comparison of methods and resources for cell-cell communication inference from single-cell RNA-Seq data. *Nature Communications* (2022).

289. Weinan, E. & Vanden-Eijnden, E. Towards a Theory of Transition Paths. *Journal of Statistical Physics* (2006).

290. Metzner, P., Schütte, C. & Vanden-Eijnden, E. Transition Path Theory for Markov Jump Processes. *Multiscale Modeling & Simulation* (2009).

291. Zhou, P., Wang, S., Li, T. & Nie, Q. Dissecting transition cells from single-cell transcriptome data through multiscale stochastic dynamics. *Nature Communications* (2021).

292. Yang, D., Jones, M. G., Naranjo, S., Rideout, W. M., Min, K. H. (, Ho, R., Wu, W., Replogle, J. M., Page, J. L., *et al.* Lineage tracing reveals the phylodynamics, plasticity, and paths of tumor evolution. *Cell* (2022).

293. Remsik, J., Tong, X., Kunes, R. Z., Li, M. J., Osman, A., Chabot, K., Sener, U. T., Wilcox, J. A., Isakov, D., *et al.* Leptomeningeal anti-tumor immunity follows unique signaling principles (2023).

294. Gazestani, V., Kamath, T., Nadaf, N. M., Dougalis, A., Burris, S., Rooney, B., Junkkari, A., Vanderburg, C., Pelkonen, A., *et al.* Early Alzheimer's disease pathology in human cortex involves transient cell states. *Cell* (2023).

295. Otto, D. J., Jordan, C., Dury, B., Dien, C. & Setty, M. Quantifying cell-state densities in single-cell phenotypic landscapes using Mellon. *Nature Methods* (2024).

296. Lönnberg, T., Svensson, V., James, K. R., Fernandez-Ruiz, D., Sebina, I., Montandon, R., Soon, M. S. F., Fogg, L. G., Nair, A. S., *et al.* Single-cell RNA-seq and computational analysis using temporal mixture modeling resolves T_H1/T_{FH} fate bifurcation in malaria. *Science Immunology* (2017).

297. Qiu, X., Hill, A., Packer, J., Lin, D., Ma, Y.-A. & Trapnell, C. Single-cell mRNA quantification and differential analysis with Census. *Nature Methods* (2017).

298. Van den Berge, K., Roux de Bézieux, H., Street, K., Saelens, W., Cannoodt, R., Saeys, Y., Dudoit, S. & Clement, L. Trajectory-based differential expression analysis for single-cell sequencing data. *Nature Communications* (2020).

299. Zhang, S., Afanassiev, A., Greenstreet, L., Matsumoto, T. & Schiebinger, G. Optimal transport analysis reveals trajectories in steady-state systems. *PLOS Computational Biology* (ed Spiegler, A.) (2021).

300. Espinosa-Carrasco, G., Scrivo, A., Zumbo, P., Dave, A., Betel, D., Hellmann, M., Burt, B. M., Lee, H.-S. & Schietinger, A. Intratumoral immune triads are required for adoptive T cell therapy-mediated elimination of solid tumors (2023).

301. Argelaguet, R., Clark, S. J., Mohammed, H., Stapel, L. C., Krueger, C., Kapourani, C.-A., Imaz-Rosshandler, I., Lohoff, T., Xiang, Y., *et al.* Multi-omics profiling of mouse gastrulation at single-cell resolution. *Nature* (2019).

302. Clark, S. J., Argelaguet, R., Lohoff, T., Krueger, F., Drage, D., Göttgens, B., Marioni, J. C., Nichols, J. & Reik, W. Single-cell multi-omics profiling links dynamic DNA methylation to cell fate decisions during mouse early organogenesis. *Genome Biology* (2022).

303. Woodworth, M. B., Girskis, K. M. & Walsh, C. A. Building a lineage from single cells: genetic techniques for cell lineage tracking. *Nature Reviews Genetics* (2017).

304. Wagner, D. E. & Klein, A. M. Lineage tracing meets single-cell omics: opportunities and challenges. *Nature Reviews Genetics* (2020).

305. Jones, M. G., Khodaverdian, A., Quinn, J. J., Chan, M. M., Hussmann, J. A., Wang, R., Xu, C., Weissman, J. S. & Yosef, N. Inference of single-cell phylogenies from lineage tracing data using Cassiopeia. *Genome Biology* (2020).

306. Ludwig, L. S., Lareau, C. A., Ulirsch, J. C., Christian, E., Muus, C., Li, L. H., Pelka, K., Ge, W., Oren, Y., *et al.* Lineage Tracing in Humans Enabled by Mitochondrial Mutations and Single-Cell Genomics. *Cell* (2019).

307. Lareau, C. A., Ludwig, L. S., Muus, C., Gohil, S. H., Zhao, T., Chiang, Z., Pelka, K., Verboon, J. M., Luo, W., *et al.* Massively parallel single-cell mitochondrial DNA genotyping and chromatin profiling. *Nature Biotechnology* (2020).

308. Lareau, C. A., Liu, V., Muus, C., Praktiknjo, S. D., Nitsch, L., Kautz, P., Sandor, K., Yin, Y., Gutierrez, J. C., *et al.* Mitochondrial single-cell ATAC-seq for high-throughput multiomic detection of mitochondrial genotypes and chromatin accessibility. *Nature Protocols* (2023).

309. Qian, X., Harris, K. D., Hauling, T., Nicoloutsopoulos, D., Muñoz-Manchado, A. B., Skene, N., Hjerling-Leffler, J. & Nilsson, M. Probabilistic cell typing enables fine mapping of closely related cell types in situ. *Nature Methods* (2019).

310. He, Y., Tang, X., Huang, J., Ren, J., Zhou, H., Chen, K., Liu, A., Shi, H., Lin, Z., *et al.* ClusterMap for multi-scale clustering analysis of spatial gene expression. *Nature Communications* (2021).

311. Petukhov, V., Xu, R. J., Soldatov, R. A., Cadinu, P., Khodosevich, K., Moffitt, J. R. & Kharchenko, P. V. Cell segmentation in imaging-based spatial transcriptomics. *Nature Biotechnology* (2021).

312. Battich, N., Stoeger, T. & Pelkmans, L. Image-based transcriptomics in thousands of single human cells at single-molecule resolution. *Nature Methods* (2013).

313. Mah, C. K., Ahmed, N., Lopez, N. A., Lam, D. C., Pong, A., Monell, A., Kern, C., Han, Y., Prasad, G., *et al.* Bento: a toolkit for subcellular analysis of spatial transcriptomics data. *Genome Biology* (2024).

314. Haviv, D., Remšík, J., Gatie, M., Snopkowski, C., Takizawa, M., Pereira, N., Bashkin, J., Jovanovich, S., Nawy, T., *et al.* The covariance environment defines cellular niches for spatial inference. *Nature Biotechnology* (2024).

315. Schaar, A. C., Tejada-Lapuerta, A., Palla, G., Gutgesell, R., Halle, L., Minaeva, M., Vornholz, L., Dony, L., Drummer, F., *et al.* Nicheformer: a foundation model for single-cell and spatial omics (2024).

316. OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., *et al. GPT-4 Technical Report* 2023.

317. Gemini Team, Georgiev, P., Lei, V. I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., *et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context* 2024.

318. Kosaji, N., Zehra, B., Nassir, N., Tambi, R., Orszulak, A. R., Lim, E. T., Berdiev, B. K., Woodbury-Smith, M. & Uddin, M. Lack of ethnic diversity in single-cell transcriptomics hinders cell type detection and precision medicine inclusivity. *Med* (2023).

319. Litviňuková, M., Talavera-López, C., Maatz, H., Reichart, D., Worth, C. L., Lindberg, E. L., Kanda, M., Polanski, K., Heinig, M., *et al.* Cells of the adult human heart. *Nature* (2020).

320. Kanemaru, K., Cranley, J., Muraro, D., Miranda, A. M. A., Ho, S. Y., Wilbrey-Clark, A., Patrick Pett, J., Polanski, K., Richardson, L., *et al.* Spatially resolved multiomics of human cardiac niches. *Nature* (2023).

321. Park, J.-E., Botting, R. A., Domínguez Conde, C., Popescu, D.-M., Lavaert, M., Kunz, D. J., Goh, I., Stephenson, E., Ragazzini, R., *et al.* A cell atlas of human thymic development defines T cell repertoire formation. *Science* (2020).

322. Yayon, N., Kedlian, V. R., Boehme, L., Suo, C., Wachter, B., Beuschel, R. T., Amsalem, O., Polanski, K., Koplev, S., *et al.* A spatial human thymus cell atlas mapped to a continuous tissue axis (2023).

323. Sountoulidis, A., Marco Salas, S., Braun, E., Avenel, C., Bergenstråhle, J., Theelke, J., Vicari, M., Czarnewski, P., Liontos, A., *et al.* A topographic atlas defines developmental origins of cell heterogeneity in the human embryonic lung. *Nature Cell Biology* (2023).

324. Kang, J. B., Nathan, A., Weinand, K., Zhang, F., Millard, N., Rumker, L., Moody, D. B., Korsunsky, I. & Raychaudhuri, S. Efficient and precise single-cell reference atlas mapping with Symphony. *Nature Communications* (2021).

325. Hawrylycz, M. J., Lein, E. S., Guillozet-Bongaarts, A. L., Shen, E. H., Ng, L., Miller, J. A., van de Lagemaat, L. N., Smith, K. A., Ebbert, A., *et al.* An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature* (2012).

Appendices

A. Acronyms

cDNA complementary DNA. 12

GRN gene regulatory network. 6

kNN k-nearest neighbor. 15, 16, 23, 26, 27

mRNA messenger RNA. ii, iii, 2, 5, 11–13, 15, 17–19, 22, 25, 30, 32, 39, 40, 71

NGS next-generation sequencing. 12

ODE ordinary differential equation. 18, 19, 25, 38, 40

OT optimal transport. ii, iii, 5, 6, 8, 10, 24, 27, 28, 30, 35, 38, 39, 42, 43, 72–74

PCA principal component analysis. 14, 15, 24

PCR polymerase chain reaction. 12

scATAC-seq single-cell Assay for Transposase-Accessible Chromatin using sequencing. 2, 11

scRNA-seq single-cell RNA sequencing. ii, 2–5, 7, 9, 11, 13–18, 24, 28, 32, 34, 37, 39, 40, 67

TI trajectory inference. ii, iii, 4–8, 11, 13, 16, 17, 23, 29, 30, 34, 38, 40, 42, 43, 45

UMI unique molecular indentifier. 12

VAE variational autoencoder. 6, 8, 22, 23, 32, 65, 66

veloVI velocity variational inference. i, ii, 21–23, 31–33, 37–39, 43, 45

VI variational inference. 7, 8, 15, 21, 32, 37, 65, 66

B. Variational inference

Statistical models describe data in terms of its distribution or model parameters. Classical approaches for parameter inference rely on maximum likelihood (MLE) or maximum a posteriori (MAP) estimates if all parameters are observed; if latent variables exist, the Expectation-Maximization algorithm[232] infers parameters, instead. However, none of these approaches directly approximates the distribution that generates the observations, a limitation that Markov chain Monte Carlo (MCMC) algorithms and variational inference overcome.

MCMC methods approximate posterior distributions by constructing a Markov chain on latent variables that (1) is ergodic and (2) has the posterior as its stationary distribution; sampling from the Markov chain, thus, provides samples from the posterior to approximate it. Classical MCMC algorithms are the Metropolis-Hastings algorithm[326,327] or Gibbs sampling[328]; extensions and improvements of these traditional approaches exist and have solved many problems[329–331]. However, MCMC routines do not scale to large datasets of current Machine Learning settings and are computationally too expensive for complex statistical models.

Variational inference offers an alternative approach to overcome the limitations of MCMC. The following paragraphs summarize relevant aspects of VI and VAEs to understand their application to single-cell data; dedicated reviews introduce and discuss the material in greater detail[233,332].

B.1. Variational inference

VI considers the joint distribution $p(x, z)$ of latent and observed variables z and x, respectively. Following the Bayesian modeling paradigm, a prior on z links the data likelihood to model observed values. Whereas MCMC relies on sampling, VI optimizes the Kulback-Leibler (KL) divergence[241]

$$q^* = \arg\min_{q \in \mathcal{Q}} \mathcal{D}_{\mathrm{KL}} \left[q(z) \parallel p(z|x) \right] \tag{B.1}$$

to identify the best approximating distribution q^* from a set of putative candidates $\mathcal{Q}$. Importantly, candidate distributions are complex enough to model $p(z|x)$ but simple enough for fast optimization. Solving (B.1) provides estimates of z and enables data generation but $p(z|x)$ is, in general, intractable: Bayes' theorem relates the posterior to the joint probability and evidence $p(x)$

$$p(z|x) = \frac{p(x, z)}{p(x)}.$$

As the evidence

$$p(x) = \int p(x|z)dz$$

is intractable in general, the posterior suffers from the same challenge. To solve the VI problem nonetheless, approximate inference solves a related problem - maximizing the evidence lower bound (ELBO)

$$\text{ELBO}\left[q\right] = \mathbb{E}[\log(p(x|z))] - \mathcal{D}_{\text{KL}}\left[q(z) \parallel p(z)\right].$$

B.2. Variational autoencoders

VAEs solve the approximate inference problem by coupling two neural networks: the encoder and the decoder. The encoder approximates the intractable posterior, and the decoder generates samples from latent factors. Formally, the encoder $q_\phi^{(\text{e})}$ is a neural network with parameters ϕ that takes observations x as input and approximates the conditional distribution $q(z|x)$ of the latent variables. Similarly, the decoder $q_\theta^{(\text{d})}$ models the likelihood $p(z|x)$ through a neural network with parameters θ and input z.

The ELBO is differentiable w.r.t. model parameters if latent variables are continuous and the encoder and decoder are differentiable, a property true for standard neural networks. However, differentiating w.r.t. ϕ is intractable since z is stochastic. The reparameterization trick resolves this problem through a change of variables[234]: In brief, a differentiable transformation through a random variable independent of parameters ϕ and observations x describes z instead of $q(z|x)$; this change of variables allows approximating derivatives through Monte Carlo estimators.

B.3. Variational inference for single-cell data

The single-cell field applies VAEs to reduce data dimensionality[76,77,79], remove batch effects[76,83–85], or enable joint analysis of multiview data[266–268,333], with most methods building upon the scVI framework[76]. This section briefly recapitulates this modeling paradigm as an application of VAEs but omits the dropout included in the original model formulation.

Single-cell RNA-seq data follows a negative binomial distribution[271,272]. To model this distribution, scVI relies on two facts:

1. The negative binomial is the mixture of Gamma-Poisson distributions.
2. The assumption that a
   a) latent representation $z_n \sim \text{Normal}\left(0, 1\right)$
   b) one-hot-encoded batch $s_n \in \{0, 1\}^{N_\text{b}}$, and
   c) batch-specific scaling factors $\ell_n|s_n \sim \text{Lognormal}\left(\ell_\mu^\top s_n, \ell_{\sigma^2}^\top s_n\right)$, with empirical mean $\ell_\mu \in \mathbb{R}^{N_\text{b}}$ and variance $\ell_{\sigma^2} \in \mathbb{R}_+^{N_\text{b}}$ of the batch-specific log-library size

   generate observations $x_n \in \mathbb{R}^{N_\text{g}}$.

To generate scRNA-seq data, the decoder network $q_\theta^{(\mathrm{d})} : \mathbb{R}^{N_\mathrm{l}} \times \{0,1\}^{N_\mathrm{l}} \to \Delta_{N_\mathrm{g}}$ maps the latent representation and batch encoding back to gene expression space, where Δ_{N_g} denotes the N_g-dimensional probability simplex; mapping to Δ_{N_g} allows interpreting the output as observation-specific gene frequencies. Following, the decoder output defines the generating distribution

$$\begin{aligned} w_{ng}|z_n, \ell_n, s_n &\sim \mathrm{Gamma}\left(\ell_n q_\theta^{(\mathrm{d})}(z_n, s_n)_g, \alpha\right) \\ x_{ng}|w_{ng} &\sim \mathrm{Poisson}\left(w_{ng}\right), \end{aligned}$$

with the shape α of the Gamma distribution estimated during inference. Integrating out w yields the negative binomial distribution

$$x_{ng}|z_n, \ell_n, s_n \sim \mathrm{NB}\left(\ell_n q_\theta^{(\mathrm{d})}(z_n, s_n)_g, \alpha_g\right).$$

To guarantee feasible inference, scVI assumes the posterior decomposes into a product of the latent representation - a mean-field approximation - and scaling factors approximated by the encoder $q_\phi^{(\mathrm{e})}$

$$q_\phi^{(\mathrm{e})}(z_n, l_n|x_n, s_n) = q_\phi^{(\mathrm{e})}(z_n|x_n, s_n) q_\phi^{(\mathrm{e})}(l_n|x_n, s_n),$$

Mini-batched stochastic-gradient descent optimizes the ELBO for parameters α, ϕ, and θ[235].

References

326. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics* (1953).

327. Hastings, W. K. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* (1970).

328. Geman, S. & Geman, D. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (1984).

329. Newman, M. E. & Barkema, G. T. *Monte Carlo methods in statistical physics* (Clarendon Press, 1999).

330. Herschlag, G., Kang, H. S., Luo, J., Graves, C. V., Bangia, S., Ravier, R. & Mattingly, J. C. Quantifying Gerrymandering in North Carolina. *Statistics and Public Policy* (2020).

331. Karras, C., Karras, A., Avlonitis, M. & Sioutas, S. in *Artificial Intelligence Applications and Innovations. AIAI 2022 IFIP WG 12.5 International Workshops* (Springer International Publishing, 2022).

332. Kingma, D. P. & Welling, M. An Introduction to Variational Autoencoders. *Foundations and Trends® in Machine Learning* (2019).

333. Gayoso, A., Steier, Z., Lopez, R., Regier, J., Nazor, K. L., Streets, A. & Yosef, N. Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nature Methods* (2021).

C. Markov chains

Markov chains describe random processes unfolding in time; discrete Markov chains consider discrete time steps, and continuous Markov chains study random processes on a continuous scale. The theory of Markov chains is well developed and a lot of in-depth literature exists[334–336]. This appendix briefly introduces the core concepts of Markov chains relevant for their application to the single-cell field; the theoretical aspects follow dedicated literature[334,337].

C.1. Definition and basic properties

Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a distribution μ_0, finite set $\mathcal{S}$, and random variables $X_n : \Omega \mapsto \mathcal{S}$. A random process $(X_n)_{n\geq 0}$ is a Markov chain with initial distribution μ_0 if

(i) X_0 has distribution μ_0

(ii) X_n is Markovian, *i.e.*,

$$\mathbb{P}(X_n = s|X_{n-1}, \ldots, X_0) = \mathbb{P}(X_n = s|X_{jn-1}) \;\; \forall n > 0, s \in \mathcal{S}$$

The distribution μ_0 is the initial distribution of the Markov chain, and $\mathcal{S}$ is the state-space with N_s possible states. The homogeneous transition matrix $P \in [0,1]^{N_\mathrm{s}\times N_\mathrm{s}}$

$$P_{jk} = \mathbb{P}(X_{n+1} = j|X_n = k)$$

and initial distribution μ_0 uniquely define the Markov chain; P is homogeneous if it is constant with respect to n. The transition matrix also defines the k-step probability as

$$\mathbb{P}(X_{n+k} = j|X_n = k) = P^k_{jk}$$

as a result of the Chapman-Kolmogorov equation[336].

C.2. Absorption probabilities

Markov chain model evolutionary processes that may terminate in specific states - the absorbing states; in the context of single-cell datasets, for instance, absorbing states correspond to terminal states. To quantify the probability of reaching such absorbing states, this section first formally defines them before introducing hitting times to characterize the probability of reaching absorbing states.

Absorbing states form a specific subset of Markov chains. Breaking down problems into smaller subproblems or subsets that are easy to solve and analyze but elucidate the entire problem is a common technique; communicating classes form such subproblems for Markov chains. Intuitively, a communication class is a collection of states for which a path to any other state in the set exists. For a formal definition, consider states $s_1, s_2 \in \mathcal{S}$. State s_1 leads to state $s_2 \in \mathcal{S}$, denoted by $s_1 \to s_2$, if

$$\exists n \geq 0 : \ \mathbb{P}(X_n = s_2 | X_0 = s_1).$$

If both states lead to each other, s_1 and s_2 communicate, denoted by $s_1 \leftrightarrow s_2$. Note that

- $s_1 \leftrightarrow s_1$, *i.e.*, $\leftrightarrow$ is reflexive,
- $s_1 \leftrightarrow s_2 \iff s_2 \leftrightarrow s_1$, *i.e.*, $\leftrightarrow$ is symmetric,
- $s_1 \leftrightarrow s_2 \wedge s_2 \leftrightarrow s_3 \implies s_1 \leftrightarrow s_3$ for any $s_3 \in \mathcal{S}$, *i.e.*, $\leftrightarrow$ is transitive.

The relation $\leftrightarrow$ is, thus, an equivalence relation and provides a disjoint partition - the communicating classes - of the state-space $\mathcal{S}$; if $\mathcal{S}$ contains a single class, the Markov chain is irreducible. A class C is closed if there are no transitions to states outside the class, *i.e.*,

$$s_1 \in C \wedge s_1 \to s_2 \implies s_2 \in C.$$

The state s_1 is absorbing if $\{s_1\}$ is a closed class.

Hitting times define the first time a random walk induced by a Markov chain reaches a given set of states. For a formal definition, let $M \subseteq \mathcal{S}$; the hitting time H^M is the random variable

$$\begin{aligned} H^M : \Omega &\to \mathbb{N} \cup \{\infty\} \\ \omega &\mapsto \inf_{n \geq 0} \{X_n \in M\}. \end{aligned}$$

Hitting times characterize classes or cell states in general: If a Markov chain starts in a class C and its corresponding hitting time is finite for infinitely many n, the class is recurrent. If the hitting time is non-finite, it is transient.

Absorption probabilities quantify how likely the Markov chain finishes in a closed class $\mathcal{R}$ in finite time: For a transient initial state s, $\mathbb{P}(H^{\mathcal{R}} < \infty | X_0 = s)$ defines the absorption probability a_s in $\mathcal{R}$; similarly, for a transient set $\mathcal{T}$, the absorption probability is

$$a_{\mathcal{T}} = \sum_{s \in \mathcal{T}} a_s = \sum_{s \in \mathcal{T}} \mathbb{P}(H^{\mathcal{R}} < \infty | X_0 = s) = \sum_{s \in \mathcal{T}} \sum_{s_{\mathrm{R}} \in \mathcal{R}} \mathbb{P}(H^{\{s_{\mathrm{R}}\}} < \infty | X_0 = s).$$

The transition matrix P describes state changes within transient and recurrent sets, denoted by T and R, respectively, and transition probabilities from the union of transient

to recurrent classes Q; these sub-matrices decompose P into a lower triangular block structure

$$P = \begin{pmatrix} T & 0 \\ Q & R \end{pmatrix}$$

and allows for computing absorption probabilities efficiently. Consider the induced random walk starting in state s_0. The expected number of times the process reaches state s_2 is

$$\mathbb{E}_n[X_n = s_2 | X_0 = s_0] = (I - R)^{-1}_{s_0 s_2}.$$

If s_2 is recurrent,

$$a_{s_2} = \left((I - R)^{-1} Q\right)_{s_0 s_2}$$

is the probability that s_2 is the first recurrent state visited[337].

C.3. Long-term behavior

Stationary distributions can characterize the long-term behavior of Markov chains: Intuitively, a distribution is stationary if it does not change under additional steps of the Markov chain. Formally, a distribution $\pi \in \mathbb{R}_+^{N_s}$ is stationary if

$$\pi^\top P = \pi.$$

Stationary distributions exist and are unique for irreducible Markov chains[334,338]; the non-negative left eigenvector with eigenvalue 1 defines a stationary distribution.

References

334. Norris, J. R. *Markov Chains* (Cambridge University Press, 1997).

335. Liggett, T. *Continuous Time Markov Processes* (American Mathematical Society, 2010).

336. Durrett, R. *Probability: Theory and Examples* (Cambridge University Press, 2019).

337. Tolver, A. *An introduction to Markov chains* (University of Copenhagen, Copenhagen, 2016).

338. Perron, O. Zur Theorie der Matrices. *Mathematische Annalen* (1907).

D. RNA velocity

D.1. The chemical master equation of splicing dynamics

Splicing dynamics constitutes distinct steps where different molecules are either produced, transformed, or degraded. The process is, thus, similar to chemical reactions that are commonly modeled by chemical master equations (CMEs). The CME describing splicing kinetics emerges as the limiting process of the probability $P_{t+\Delta t}(u = m, s = n)$ to observe $m \in \mathbb{N}_0$ unspliced and $n \in \mathbb{N}_0$ spliced molecules at time $t + \Delta t$. This state is reached if at time t the system includes

- the same number of unspliced and spliced molecules and no transcription, splicing, or degradation occurs in the time interval $[t, t + \Delta t]$,
- $u = m + 1$ and $s = n - 1$ molecules, and a single pre-mRNA is transcribed in $[t, t + \Delta t]$,
- $u = m$ and $s = n+1$ transcripts of which one mature mRNA degrades in $[t, t+\Delta t]$.

Considering the probabilities at which these events happen, the state $P_{t+\Delta t}(u = m, s = n)$ is described by

$$\begin{aligned} P_{t+\Delta t}(u = m, s = n) =& P_t(u = m, s = n)(1 - \alpha\Delta t)(1 - \beta\Delta t)^m(1 - \gamma\Delta t)^n + \\ & P_t(u = m + 1, s = n - 1)\beta\Delta t + P_t(u = m, s = n + 1)\gamma\Delta t. \end{aligned}$$

Omitting higher order terms and considering the limiting process $\Delta t \to 0$ yields the CME

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}P_t(u = m, s = n) =& \alpha\left[P_t(u = m - 1, s = n) - P_t(u = m, s = n)\right] + \\ & \beta\left[(m + 1)P_t(u = m + 1, s = n - 1) - P_t(u = m, s = n)\right] + \\ & \gamma\left[(n + 1)P_t(u = m, s = n + 1) - P_t(u = m, s = n)\right]. \end{aligned} \tag{D.1}$$

E. Optimal transport

In its original formulation, optimal transport described how to move a pile of soil into a pre-defined shape with minimal work[339]; generalizing this problem led to optimal couplings of distributions[340]. Throughout this appendix, consider sample spaces $\mathcal{X} = \{x_1, x_2, \ldots, x_{N_\mathrm{s}}\}$ and $\mathcal{Y} = \{y_1, y_2, \ldots, y_{N_\mathrm{t}}\}$ with probability measures

$$\alpha = \sum_{j=1}^{N_\mathrm{s}} a_j \delta_{x_j}$$
$$\beta = \sum_{j=1}^{\mathcal{Y}} b_j \delta_{y_j},$$

with the Dirac delta function δ, respectively; $\mathcal{X}$ will always denote the *source domain*, and $\mathcal{Y}$ the *target domain*. The cost function $c : \mathcal{X} \to \mathcal{Y}$ defines the cost associated with mapping data from the source to the target domain; the cost matrix $C \in \mathbb{R}^{N_\mathrm{s} \times N_\mathrm{t}}$ with $C_{jk} = c(x_j, y_k)$ collects all costs.

E.1. The Monge problem

The Monge problem seeks to optimally assign each source $x \in \mathcal{X}$ a unique target $y \in \mathcal{Y}$ while preserving mass. If such a coupling exists, the solution is a map $T : \mathcal{X} \to \mathcal{Y}$ satisfying

$$\min_T \left\{ \sum_j c(x_j, T(x_j) \mid T_\sharp \alpha = \beta) \right\},$$

with the push-forward operator[341] $T_\sharp : \mathcal{M}(\mathcal{X}) \to \mathcal{M}(\mathcal{Y})$; $\mathcal{M}(\mathcal{X})$ denotes the set of Radon measures on $\mathcal{X}$[336].

The Monge problem describes the situation of reshaping a pile of soil or distributing products to buyers. For single-cell data, the assignment of cells at an earlier time to cells at a later stage is not unique, however: A cell observed at a time point t_1 may divide into daughter cells observed at a later time point t_2; similarly, cells may experience apoptosis. To apply OT nonetheless, the single-cell field relies on the Kantorovich relaxation.

E.2. The Kantorovich relaxation

The Kantorovich relaxation eases the requirement that source points be mapped to a single target point. Instead, source points are distributed over targets in a probabilistic fashion. Possible couplings $P \in \mathbb{R}^{N_\mathrm{s} \times N_\mathrm{t}}$ form the set

$$U(a,b) = \{P \in \mathbb{R}^{N_\mathrm{s} \times N_\mathrm{t}} \mid P1_{N_\mathrm{s}} = a \ \wedge \ P^\top 1_{N_\mathrm{t}} = b\},$$

with $1_N = (1, 1, \ldots, 1) \in \mathbb{R}^N$. The solution $\mathcal{L}_c(\alpha, \beta)$ of the Kantorovich formulation solves the optimization problem

$$\min_{P \in U(a,b)} \langle C, P \rangle = \min_{P \in U(a,b)} \sum_{j=1}^{N_\mathrm{s}} \sum_{k=1}^{N_\mathrm{t}} C_{jk} P_{jk}. \tag{E.1}$$

Solving problem (E.1) is non-trivial as

(i) equation (E.1) is linear

(ii) $U(a,b)$ imposes $N_\mathrm{s} + N_\mathrm{t}$ equality constraints

(iii) $U(a,b)$ is bounded,

i.e. (E.1) is a convex linear program[342,343], making the solution, in general, not unique. Additionally, optimization strategies devised for solving (E.1) do not scale to large domain dimensions.

E.3. The Sinkhorn algorithm

To solve (E.1) efficiently, the problem can be entropically regularized, yielding a unique global optimum $\mathcal{L}_c^\varepsilon(\alpha, \beta)$[240]; the solution $\mathcal{L}_c^\varepsilon(\alpha, \beta)$ of

$$\min_{P \in U(a,b)} \langle C, P \rangle - \varepsilon \underbrace{\sum_{j=1}^{N_\mathrm{s}} \sum_{k=1}^{N_\mathrm{t}} P_{jk} \log(P_{jk} - 1)}_{:= -\mathrm{H}(P)} \tag{E.2}$$

approximates $\mathcal{L}_c(\alpha, \beta)$ as $\mathcal{L}_c^\varepsilon(\alpha, \beta) \to \mathcal{L}_c(\alpha, \beta)$ for $\varepsilon \to 0$.

The Sinkhorn algorithm efficiently computes the solution $\mathcal{L}_c^\varepsilon(\alpha, \beta)$ of (E.2): Consider optimization variables $u \in \mathbb{R}_+^{N_\mathrm{s}}$ and $v \in \mathbb{R}_+^{N_\mathrm{t}}$, and the Gibbs kernel

$$\begin{aligned} K : \mathbb{R}^{n \times m} &\to \mathbb{R}^{n \times m} \\ k_{jl} &\mapsto e^{\frac{k_{jl}}{\varepsilon}}, \quad (j,l) \in [\![n]\!] \times [\![m]\!], \end{aligned}$$

with $[\![n]\!] \coloneqq \{1, 2, \ldots n\}$. With these variables, the unique solution $\mathcal{L}_c^\varepsilon(\alpha, \beta)$ of (E.2) can be written as[342]

$$P = \mathrm{diag}(u) K \mathrm{diag}(v). \tag{E.3}$$

Based on this formulation, an iterative scheme defines the solution computationally by alternatively scaling rows and columns of a solution candidate: As a feasible solution, (E.3) lies in $U(a,b)$, *i.e.*,

$$\operatorname{diag}(u)Kv = a$$
$$\operatorname{diag}(v)K^\top u = b,$$

where $\operatorname{diag}(u)$ is an $N_\mathrm{s} \times N_\mathrm{s}$ matrix with u on its diagonal and zeros entries otherwise - a problem known as the matrix scaling problem. The matrix scaling problem can be solved by first fixing v and updating u to satisfy the first equality, followed by fixing u and updating v to satisfy the second equation. Taken together, the Sinkhorn algorithm is, thus, given by

Algorithm 1: Sinkhorn algorithm

Data:
- Cost matrix $C \in \mathbb{R}^{N_\mathrm{s} \times N_\mathrm{t}}$
- Source marginal $a \in \mathbb{R}^{N_\mathrm{s}}$
- Target marginal $b \in \mathbb{R}^{N_\mathrm{t}}$
- Regularization parameter $\varepsilon > 0$

Result: Solution matrix P of entropically regularized Kantorovich OT problem (E.2)

$K \leftarrow \exp\{\frac{C}{\varepsilon}\}$
$\ell \leftarrow 0$
$u^{(\ell)} \leftarrow 1_{N_\mathrm{s}}$
$v^{(\ell)} \leftarrow 1_{N_\mathrm{t}}$
while *not converged* **do**
  $u_j^{(\ell+1)} \leftarrow \frac{a_j}{\sum_{n=1}^{N_\mathrm{t}} K_{jn} v_n^{(\ell)}} \quad j \in [\![N_\mathrm{s}]\!]$
  $v_n^{(\ell+1)} \leftarrow \frac{b_n}{\sum_{j=1}^{N_\mathrm{s}} K_{nj} u_j} \quad n \in [\![N_\mathrm{t}]\!]$
  $\ell \leftarrow \ell + 1$
end
$P \leftarrow \operatorname{diag}(u) K \operatorname{diag}(v)$
return P

References

339. Monge, G. Mémoire sur la théorie des déblais et des remblais. *Mem. Math. Phys. Acad. Royale Sci.* (1781).

340. Kantorovic, L. *On the translocation of masses* in *CR (Doklady) Acad. Sci. URSS (NS)* (1942).

341. Tao, T. *An Introduction to Measure Theory* (American Mathematical Society, 2011).

342. Peyré, G. & Cuturi, M. Computational Optimal Transport (2018).

343. Brualdi, R. A. *Combinatorial Matrix Classes* (Cambridge University Press, 2006).