Perturbing the Genome: From Bench to Biophysics

Thesis by
Tara Chari

In Partial Fulfillment of the Requirements for the
Degree of
Doctor of Philosophy in Biological Engineering

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2024
Defended May 20, 2024

Tara Chari
ORCID: 0000-0002-6953-4313

ACKNOWLEDGEMENTS

The journey of this PhD has taken many twists and turns, and I'd like to thank all the people who made that journey not only possible, but enjoyable and memorable. First and foremost, to my advisor Lior Pachter, who was not only an understanding mentor, always willing to hear my ideas and provide feedback, but also enthusiastically encouraged my deviations and varied interests over the PhD. Similarly, thanks to my committee members Richard M. Murray, Lulu Qian, and David J. Anderson, who encouraged my diverse interests and whose feedback and discussion over the past 6 years I am grateful for.

I am also extremely grateful to have had as mentors Brandon Weissbourd and Jase Gehring, who led my initiation into the world of single-cell biology, and whose approaches to science and investigation inspired much of my work.

This work would also not have been possible without the open and engaging environment of the Pachter Lab and its members, many of whom I am grateful to have worked with and to be able to call friends. In particular, Gennady Gorin, Meichen Fang, and Maria Carilli, all of whom are co-authors in this thesis, and whose insights, discussion, and banter I've found invaluable. Also to Taleen Dilanyan, with whom I started in the wetlab during our Pachter Lab rotations, and to the rest of the lab (old and new) I've gotten to know, Anne Kil, Ángel Gálvez-Merchán, Nikhila Swarna, Catherine Felce, Kayla Jackson, Laura Luebbert, Joe Rich, Lambda Moses, Rebekah Loving Ngo, Delaney Sullivan, Kristján Eldjárn, Nadia Volovich, and Sina Booeshaghi.

Outside of the lab, I would also like to thank Fan Gao, in his capacity as the head of the Bioinformatics Resource Center, who helped us navigate the tricky world of non-model genomics. And Jamie Tijerna, in the Flow Cytometry Facility at Caltech, who spent many hours, and liters of seawater, with me in my quest to extract jellyfish neurons.

I am also grateful to Romain Lopez and Jan-Christian Hütter, who worked closely with me during my internship and were always willing to answer my questions and go to the whiteboard if needed. And to Aviv Regev and Kai Liu, who enabled my freedom to explore during the internship, and whose feedback and discussion I always looked forward to.

For helping form my passion for research, I also want to thank Sridhar Govindarajan

and Shogo Hamada, who helped me build my skills in the wetlab and at the computer, as I delved into the world of computational biology as a mere undergrad.

Thanks as well to the National Institutes of Health and the Tianqiao and Chrissy Chen Institute for Neuroscience, whose funding supported much of this work. And to Zdenek Sasek cartoons, which inspired the representations in this thesis.

Beyond the realm of research, I want to thank the ever-persistent book club, which has always provided me enough fiction to fill my free time, as well as Prof. Orcel and Prof. Merrill, whose French classes were always a welcome reprieve. To Kelly Kadlec, with whom I have gone from roommates, to Caltech gym regulars, to Pasadena natives, over the course of our PhDs. And to Prashant Bhat, whose conversation, be it scientific or otherwise, I always look forward to.

Over the entirety of college, into grad school, and beyond, I also want to thank Vaidehi Garg and Michelle Yang, who have always been there whether in person or online, to cry, laugh and overthink together.

To my family, Mallika, Amma, and Appa (and Howie and Luna too) who have always been there for me, to bring me back to my senses and to keep me going. And finally, to Ben, qui m'a accompagné tout au long de cette aventure, et avec qui j'espère en avoir beaucoup plus.

ABSTRACT

In single-cell genomics, we can simultaneously assay hundreds of thousands of cells, their molecular contents, and how they respond to perturbation, from genetic knockouts to environmental changes. This thesis focuses on how to merge experimental and computational techniques to generate and analyze large-scale perturbation data for high-resolution systems biology. Beginning at the bench, we demonstrate how combining large-scale cell atlas surveys with multi-condition experimentation can illuminate the diversity of cell types across whole organisms and cellular strategies in response to environmental changes and perturbations. We then investigate the limitations of current practice in exploratory analysis, and strategies for determining preservation or distortion of biological insight by these data transformation and dimensionality reduction techniques. To address these limitations, we demonstrate how stochastic biophysical models can rewrite the way we interpret complex perturbation data, taking greater advantage of the diverse molecular measurements to develop biological hypotheses about DNA and RNA regulation in cellular function, development, and disease.

PUBLISHED CONTENT AND CONTRIBUTIONS

The materials in the current thesis are largely drawn from the manuscripts listed below. Where relevant, the headings of sections credit the contributors for specific ideas, with the abbreviations as follows: L.P.: Lior Pachter, G.G.: Gennady Gorin, B.W.: Brandon Weissbourd, J.G.: Jase Gehring, A.F.: Anna Ferraioli, L.L.: Lucas Leclère, M.H.: Makenna Herl, S.C.: Sandra Chevalier, R.R.C.: Richard R. Copley, E.H.: Evelyn Houliston, D.J.A.: David J. Anderson, R.L.: Romain Lopez, J.C.H.: Jan-Christian Hütter, A.R.: Aviv Regev, K.L.: Kai Liu, and T.C.: Tara Chari.

Although not indicated in the list below, T.C. was co-first author on manuscript 4 with B.W., J.G., and A.F.

 Epigraph cartoons were hand-drawn by T.C.

[1] Maria Carilli, Gennady Gorin, Yongin Choi, Tara Chari, and Lior Pachter. Biophysical modeling with variational autoencoders for bimodal, single-cell RNA sequencing data. *bioRxiv*, May 2023.
T.C. participated in the design of analyses and visualizations, and contributed to writing of the manuscript.

[2] Tara Chari and Lior Pachter. The split senate. *APSA Preprints*, 2021.
T.C. designed and performed analysis and developed code, and T.C. participated in writing and editing the manuscript.

[3] Tara Chari and Lior Pachter. The specious art of single-cell genomics. *PLoS Comput. Biol.*, 19(8):e1011288, August 2023.
T.C. participated in conceptualizing the study, T.C. performed analysis and developed code, and T.C. participated in writing and editing the manuscript.

[4] Tara Chari, Brandon Weissbourd, Jase Gehring, Anna Ferraioli, Lucas Leclère, Makenna Herl, Fan Gao, Sandra Chevalier, Richard R Copley, Evelyn Houliston, David J Anderson, and Lior Pachter. Whole-animal multiplexed single-cell RNA-seq reveals transcriptional shifts across Clytia medusa cell types. *Sci Adv*, 7(48):eabh1683, November 2021.
T.C. participated in design and execution of experiments and performance of the single-cell experiments, T.C. performed whole-organism qPCR, T.C. participated in creation of scripts for processing the data and code for the analysis, T.C. developed the Google Colab notebooks, T.C. participated in analysis and interpretation of the data, and T.C. contributed to writing and editing the manuscript.

[5] Tara Chari, Gennady Gorin, and Lior Pachter. Biophysically interpretable inference of cell types from multimodal sequencing data. *bioRxiv*, September 2023.
T.C. participated in conceptualizing the study, designed and performed analysis and developed code, and participated in writing and editing the manuscript.

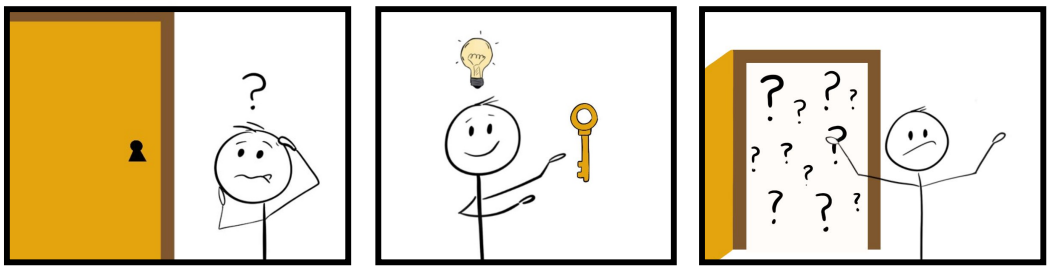
[6] Gennady Gorin, Meichen Fang, Tara Chari, and Lior Pachter. RNA velocity unraveled. *PLoS Comput. Biol.*, 18(9):e1010492, September 2022.
T.C. participated in the analysis, visualization, and writing of the manuscript.

[7] Gennady Gorin, Maria Carilli, Tara Chari, and Lior Pachter. Spectral neural approximations for models of transcriptional dynamics. *bioRxiv*, page 2022.06.16.496448, November 2023.
T.C. participated in the design of analyses and visualizations, and contributed to writing of the manuscript.

TABLE OF CONTENTS

Chapter 1

INTRODUCTION AND OUTLINE



The Journey of a PhD

This thesis summarizes work to develop experimental and computational methods for high-throughput perturbation biology. We demonstrate a novel protocol for whole-animal, multiplexed perturbation in marine organisms, to facilitate large-scale systems biology in non-standard model systems. To analyze and extract biological insight from these data, we develop mathematical and statistical tools to determine distortion of data by common dimensionality reduction and transformation techniques, and present alternative reduction and visualization approaches. To further aid hypothesis-driven biological investigation, we present methods which utilize biophysical and chemical kinetic relationships of the measurements in these genomics datasets to define the standard tasks of cell type categorization and perturbation response analysis through physically-interpretable parameters and cellular processes.

In the quest to understand the genome, to interpret the “language of life" [55], and the processes by which this language is translated into function and biological diversity, we have brought the ‘big data’ approach of the Information Age to genomics. In this thesis, we tackle both the experimental and computational challenges in ‘big data genomics’, particularly, how to investigate complex biological systems through perturbation.

With the development of single-cell RNA sequencing (scRNA-seq) technologies, we can generate measurements of gene production and regulation, e.g., messenger RNA (mRNA) expression, across tens of thousands to millions of cells, our ‘units of

life'. In combination with perturbation of these cells, from assaying various disease conditions to genetic or drug-based interventions [69, 92], we can not only study heterogeneous biological systems at high-resolution, but also probe changing states and their underlying mechanisms. This explosion in data production has also spurred an ever-increasing development of new methods and analyses to extract insight from these data types [273]. The parallel, rapid development of machine learning (ML) methods for large-scale, high-dimensional data [133], has also led to incorporation of these approaches in biological analysis pipelines. Given this large space of experimental data types and possible computational method pairings, we will not address all aspects of experiment and analysis in this thesis, but rather focus on how question-guided approaches to experimental design and method development can aid biological interpretation of perturbation data.

We begin this thesis on the experimental side of perturbation biology, Chapter 2, and how the combination of scRNA-seq surveys of heterogeneous cell populations with multiplexed experimentation, allows for a tractable, question-first approach to high-resolution, systems biology. In particular we demonstrate this on a developing model organism in Chapter 3, to simultaneously reveal the diversity of cell types across an organism and their cellular responses to environmental perturbation.

From this work, we then identify open challenges and questions that remain in common practice approaches, defined in Chapter 4, for exploratory analysis of multifaceted scRNA-seq data. To this end, we investigate how popular methods for embedding or representing such data in low dimensions (for data summary, analysis, and hypothesis-generation) are quantitatively limited in their preservation and representation of biological trends in Chapter 5.

Given these limitations in extracting biological insight from common representation learning methods, we present an alternative avenue to representation of high-throughput perturbation data in Chapter 6, with an eye towards scalability not only in dataset size but also in interpretability, i.e., how new biological measurements are incorporated into these representations. We use stochastic biophysical models of transcription for interpretation of perturbation data, rewriting the standard analysis tasks in scRNA-seq through this medium, and discussing both the limitations as well as promising extensions of this work, see Chapter 7 and 8, as the reach of high-throughput genomics continues to expand.

Overall, this thesis aims to demonstrate how to quantitatively realize otherwise potentially qualitative investigations of high-throughput perturbation biology, and

to fashion experimentalists with tools, as well as approaches to tool development, which illuminate the biology underlying their questions and hypotheses.

Chapter 2

PERTURBATION FOR DECIPHERING THE CELLULAR CODE

Technologies for experimental observation have given us some of the most fundamental discoveries in biology, such as the microscopes of Robert Hooke which in turn illuminated the world of 'cells' [117]. In a similar vein, the use of perturbation to create unique and interrogative experimental settings, has given us fundamental hypotheses about biological and genetic mechanisms, where the pea plant hybrids of Gregor Mendel [183], for example, informed an understanding of the inheritance of traits. A hundred or more years later we now have the ability to both observe and perturb complex biological systems at unprecedented scale, bringing together discovery of biological phenomena with mechanistic understanding. To this end, we begin this thesis by investigating how current technologies can simultaneously observe and perturb novel biological systems at-scale, allowing us to dive into the variety of genetic strategies and behaviors cells throughout an organism are employing. In Ch. 4, 5, and 6, we will then expand upon how current analytical techniques can better extract the plethora of information stored in these experimental assays, to produce new observations and new hypotheses.

To better understand the origins and capabilities of the genomics technologies available today, we begin in the 1950s, where the observations of Hooke and Mendel have paved the way for the illumination of how our cellular DNA and its 'genes' form the mechanisms of heredity [1]. The development of molecular cloning and DNA sequencing in the 1970s and 80s, empowered the movement towards understanding the genome, culminating in the Human Genome Project (HGP) in 1990, with the goal of sequencing and mapping the complete set of human genes [116]. The onset of the HGP brought with it developments in high-throughput DNA sequencing, that enabled large-scale studies of human genetic variations and cataloguing regulatory and non-regulatory elements of the genome. This inspired efforts like the ENCODE [78] and GTEx [165] consortiums, to not only map gene annotation to function, but to ascertain gene behavior or expression in various tissue contexts.

The development of microarrays in 1995 [217], unlocked the ability to assess thousands of genes on a single slide, but still required predefined probes and their gene targets. Following advances addressed these limitations, in the development of

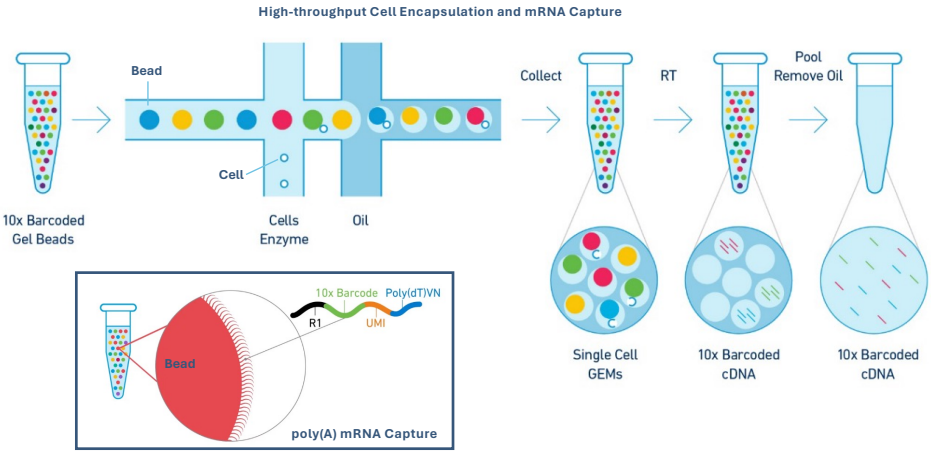


Figure 2.1: **Overview of 10x Single-Cell Profiling.** Diagram of how 10x beads and cells are encapsulated in droplets, where cells are lysed and mRNA captured by their poly(A) tails. mRNA is then reverse-transcribed (RT) to cDNA. Adapted from 10x Genomics.

next-generation sequencing (NGS) and 'RNA-seq' [186]. In particular, the approach of RNA-seq to capture any polyadenylated (poly(A)-containing) mRNAs from the pooled contents of cells, allowed for a more 'unbiased' sampling of these intermediate gene products. Such protocols are also denoted as 'bulk RNA-seq', as the mRNA molecules sampled come from the bulk (or aggregated) genomic contents of all the cells pooled. This ability to sample the spectrum of mRNA transcribed from the genome, the 'transcriptome', was then coupled to single cell isolation techniques, with technologies such as Smart-Seq [201]. This opened up the ability to perform transcriptome-wide exploratory investigation of individual cells' gene expression profiles.

In just the past 10 years, massive developments have scaled up such approaches, resulting in commercial platforms such as 10x Genomics (Fig. 2.1), which can easily isolate the mRNA contents of tens of thousands of cells at once for scRNA-seq. These high-throughput methods often utilize microfluidics to isolate the single cells [163], and can utilize Unique Molecular Identifiers (UMIs) to label the individual molecules captured from each cell (Fig. 2.1 'UMI' in subpanel). This allows for discrete quantification of molecule counts after subsequent sequencing and amplification steps. It should also be noted that most commercially available technologies convert captured mRNA to cDNA prior to sequencing, i.e., cDNA is a proxy for the mRNA molecules detected (Fig. 2.1). Improved cell capture and transcriptome coverage, have also enabled the production of scRNA-seq 'atlases' , where

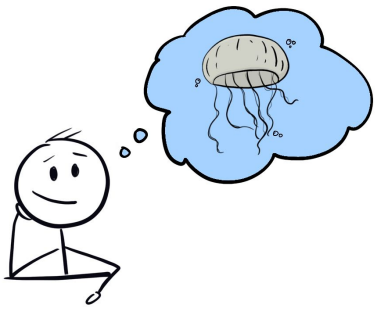
an ‘atlas’ denotes detailed delineations of the cell populations which comprise a heterogeneous system, such as a tissue sample, using the measured transcriptome- and genome-wide characteristics. In the spirit of the HGP mission, this has given rise to the Human Cell Atlas mission of building a reference map and resource of cellular diversity, to both catalogue this diversity and aid mechanistic insight [210].

However, to really tease apart and uncover the components driving regulation, production and processing of genes across cells, we need to combine study and observation of these systems with perturbation, be it disease conditions, drug combinations, or genetic interference. Perturbation enables inference of causality and directionality of interactions to, for example, construct models of gene signaling cascades which in turn control cellular development [9]. To this end, scRNA-seq techniques can now be combined with multi-condition experimentation, whereby cells from multiple samples, conditions, perturbations, etc. can be pooled together for sequencing but retain unique tags from which their original conditions can be decoded, referred to as ‘multiplexing’ of cells. In addition to multiplexed cells, there has also been a push towards simultaneous capture of multiple biological entities per cell, from mRNA and chromatin state information [205] to spatial distribution of cells and other imaging-based phenotypes [83]. In the 2020s, we thus enter an era of high-throughput, multimodal perturbation biology.

Looking back on the observations of Hooke and Mendel, this opens up a whole new world of exploration into the inner workings of cells, to ask questions about *how* their genes lead to the diversity of behaviors we observe, at the scale of the individual cell all the way up to their composite behaviors across an organism [163, 210].

Chapter 3

QUESTION-FIRST, WHOLE ORGANISM PERTURBATION



For the things we have to learn before we can do them, we learn by doing them.

ARISTOTLE

This chapter summarizes the contents of [49] by T.C.*, B.W.*, J.G*, A.F.*, L.L, M.H., F.G., S.C., R.C., E.H., D.J.A., L.P. * denotes co-first authorship. T.C., B.W., J.G., R.R.C., E.H., D.J.A., and L.P. conceived of experiments, T.C., B.W., and J.G. performed the single-cell experiments, T.C. performed whole-organism qPCR, T.C. and J.G. wrote scripts for processing the data and code for the analysis, T.C. developed the Google Colab notebooks, T.C., B.W., J.G., A.F., L.L., R.R.C., E.H., D.J.A analyzed and interpreted the data, and T.C., B.W., J.G., A.F., L.L., R.R.C., E.H., D.J.A., and L.P. contributed to writing and editing the manuscript.

In light of the advancements in high-throughput sequencing described in Chapter 2, we sought to merge the concepts of cell atlas surveys with multiplexed single-cell experimentation to take a 'question-first' approach to exploring the systems biology of whole organisms at single-cell resolution. In this way, we use the question of how this system reacts under some perturbation of interest, to guide which genes will be extracted for follow-up study and annotation, based on relevance of their activity (or expression) under the perturbation setting. This outlines an 'activity-based' approach, with less focus on prior annotation or functional understanding of gene, as compared to the efforts of the ENCODE and GTEx consortiums described previously.

Combining samples from multiple conditions and individual organisms can be

costly, and may be confounded by batch effects resulting from multiple distinct library preparations and sequencing runs [36, 248]. Thus the recent developments in scRNA-seq multiplexing technology expand the number of samples, individuals, or perturbations that can be incorporated within runs, facilitating well-controlled scRNA-seq experiments [92, 102, 178, 179, 233, 235]. Here we merge these techniques with high-throughput sequencing to demonstrate a powerful experimental paradigm on a planktonic model organism. We examine the medusa (free-swimming jellyfish) stage of the hydrozoan *Clytia hemisphaerica*, with dual motivations. Firstly, *Clytia* is a powerful, emerging model system spanning multiple fields, from evolutionary and developmental biology to regeneration and neuroscience [27, 131, 152, 153, 227, 234]. While previous work has characterized a number of cell types in the *Clytia* medusa [153], a whole-organism atlas of transcriptomic cell types has been lacking. Such an atlas is a critical resource for the *Clytia* community, and an important addition to the study of cell types across animal phylogeny.

Secondly, these emerging multiplexing techniques present new opportunities for systems-level studies of cell types and their changing states at unprecedented resolution in whole organisms. The *Clytia* medusa offers an appealing platform for pioneering such studies. It is small, transparent, and has simple tissues and organs, stem cell populations actively replenishing many cell types in mature animals, and remarkable regenerative capacity [13, 90, 131, 150, 152, 227]. Furthermore, the 1cm-diameter adult medusae used in this study contain on the order of 10^5 cells, making it possible to sample cells comprehensively across a whole animal in a cost-effective manner using current scRNA-seq technology (see Table S1, S2, Fig. S1 in [49]). In this work, we generate a cell atlas for the *Clytia* medusa while simultaneously performing a whole organism perturbation study, providing the first medusa single-cell dataset and an examination of changing cell states across the organism. Our approach also provides a proof-of-principle for perturbation studies in non-traditional model organisms, using multiplexing technology and a reproducible workflow with lessened reliance on functional annotation, from the experimental implementation to the data processing and analysis.

3.1 Experimental Paradigms for *Clytia* Perturbation

In this study, we compared control versus starved animals, as this strong, naturalistic stimulus was likely to cause significant, interpretable changes in transcription across multiple cell types. Laboratory-raised, young adult, female medusae were split into

two groups of five animals, one deprived of food for four days, and the second fed daily. We observed numerous phenotypic changes in starved animals, including a dramatic size reduction reflecting two- to threefold fewer cells [89] (see Fig. S2 in [49]), and a striking reduction in gonad size. Correspondingly, the number of eggs released per day decreased [12].

For scRNA-seq , single-cell suspensions were prepared from each whole medusa and individually labeled with unique ClickTag barcodes [92] using a sea-water compatible workflow. Briefly, animals were washed into hypertonic PBS-solution, in which single-cell suspensions of the organisms were made. Suspension were then spun down and re-suspended in methanol (with all suspensions steps performed on-ice). Methanol-fixed samples were labeled with two ClickTag barcodes, denoting each individual and each condition following the protocol in [92]. Labeled suspensions were then pooled and processed with the 10x Genomics v2.0 workflow and Illumina sequencing, allowing construction of a combined dataset across organisms and treatments, without requiring batch correction.

A second perturbation experiment was performed in the same manner, for validation and assessment of cell type diversity (and technical variation) observed in the first dataset, as well as extension of the multiplexed approach to investigate the existence of 'immediate early gene (IEG)'-like behaviors in *Clytia* [223], i.e., gene responses sensitive to more rapid (or subtle) gene perturbations than extreme starvation. We additionally demonstrated this experimental workflow with the newer 10x Genomics v3.0 platform. For this study, we exposed *Clytia* medusae to multiple transient, ionic stimuli and dissociated 1h later.

For selection of cells for downstream analysis, a separate cDNA library for the sequenced ClickTags was processed to determine the count of each ClickTag in the cells captured, selecting for cells with clear expression of two ClickTags (denoting individual and condition), without significant expression (or bleed-through) of other tags. For processing of the cells' cDNA library (obtained from the captured mRNA), all data was initially processed with the 10x Cell Ranger pipeline, however all samples were re-analyzed in a streamlined workflow using *kallisto|bustools* [29, 181]. The count matrices extracted from this workflow represented spliced mRNA (exon-containing or mature mRNA), though we will discuss the use of other mRNA types in Chapter 6. A total of 13,673 single-cell profiles derived from the ten individuals of the first perturbation study five control, five starved) passed quality control, with high concordance in cell type abundance and gene expression among

animals in the same treatment condition (see Fig. S5 in [49]). From this gene expression matrix, we 1) derived a *Clytia* medusa cell atlas, and 2) generated a high-resolution resource of the transcriptional impact of starvation across all observed cell types.

Though the majority of the results in the published work focus on the starvation perturbation dataset, with the short-term stimulation data, we likewise captured 18,921 single-cell profiles across twelve animals, with three animals each in KCl-treated, DI water-treated and SW (control seawater) conditions. We could thus not only compare the recovered cell types between datasets, but also identify candidate genes with IEG-like properties across many cell types, including neurons [223] (see Fig. S6, S8, Table S4 in [49]), by looking for genes with significantly different expression patterns in each condition. IEGs are valuable tools in neuroscience, to identify neurons that are active following a specific stimulus or behavior [223]. This methodology is thus able to detect transcriptional responses across diverse stimulus-response paradigms.

3.2 A *Clytia* Cell Atlas

To generate the cell atlas, we clustered the cells using the gene expression matrix across all starved and control individuals, extracting 36 cell types and their corresponding marker genes. Following standard practice, cells were selected that displayed higher total UMI counts (across genes detected) than the inflection point on the plot of total UMIs per cell vs. ranked cell barcodes (in descending order) (the 'knee-plot') [134, 262]. Cells were then scaled to have total cell counts of 10^4 (across genes) and log1p normalized [171]. For clustering, the Louvain algorithm for community detection [61] was used, after applying highly variable gene (HVG) selection and principal component analysis (PCA) to the reduce the data to 60 dimensions [134, 171]. These 36 clusters thus represented the extracted cell types, where marker genes for these cell types were selected as genes displaying significant differences in expression patterns between types, e.g., through the non-parametric Wilcoxon rank-sum test.

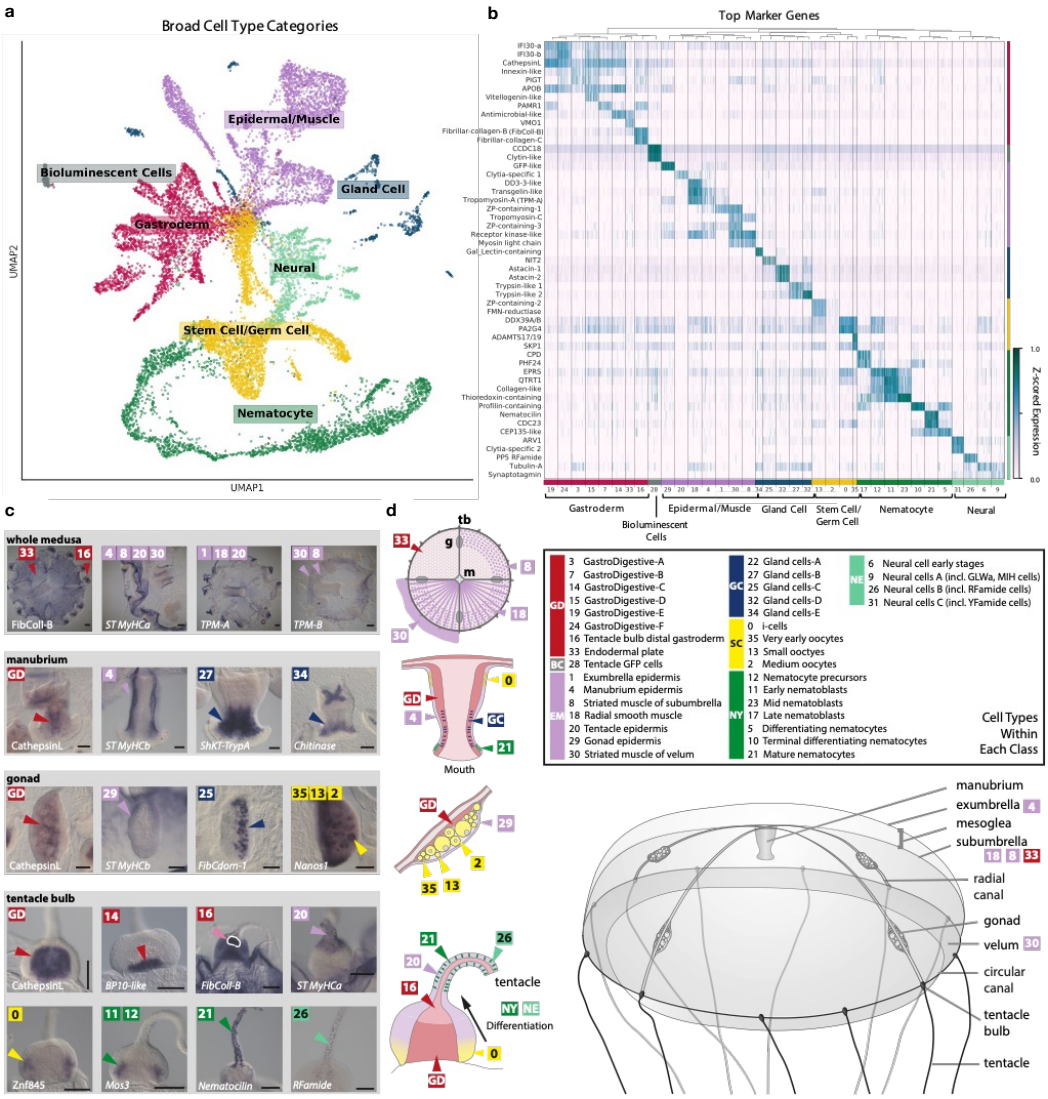


Figure 3.1: **The *Clytia* Medusa Cell Atlas.** **a)** 2D UMAP embedding of cells labeled by seven cell type classes. Class colors are retained in panels **b**, **c**, and **d**. **b)** Heatmap of top marker genes from the sequencing data with 36 Louvain clusters comprising the seven cell type classes. **c)** In situ hybridization patterns for a selection of cluster marker genes providing spatial location on the animal (comprehensive set in Fig. S14 in [49]). The label GD denotes general markers for GastroDigestive cell types. Scale bars: 100 µm. **d)** Schematics of Clytia medusa, manubrium, gonad, and tentacle bulb showing the main cell types. Abbreviations of cell class names GD: GastroDigestive, BC: bioluminescent cells, EM: epidermal /muscle, GC: gland cells, SC: stem cell/germ cell, NY: nematocytes, NE: neural cells. Adapted from [49].

We then generated a low-dimensional representation [180, 255] of these cell types (Fig. 3.1a) following standard procedure [134, 171], though we will discuss the quantitative properties of such visuals in Chapter 5. We grouped the cell types into seven broad classes (Fig. 3.1a) which correspond to the outer epidermis, the inner gastrodermis, and to likely derivatives of the multipotent interstitial stem cell population (i-cells). I-cells are a specific feature of hydrozoans, and are particularly well characterized in Hydra, where they generate neural cells, gland cells, and stinging cells (nematocytes), as well as germ cells [27, 110, 225]. Our dataset was derived from female medusae so it lacks male germ cells, and late stage oocytes are expected to be too large for capture by the dissociation procedure. The 36 cell types were concordant between the two separate multiplexed experiments ('Starvation' and 'Stimulation'). For some of them, cell type identity could be assigned on the basis of published information on gene expression in *Clytia* and/or of homologous genes in other animals, while for the others we performed in situ hybridization for selected marker genes. Previously known cell types apparent in our data included i-cells [151] and nematocytes at successive stages of differentiation [56, 65, 239], as well as oocytes [246], gonad epidermis, manubrium epidermis, and bioluminescent cells in the tentacles that each express specific endogenous green fluorescent proteins (GFPs) [86].

In situ hybridization for a selection of diagnostic muscle cell type genes allowed us to describe cell types making up the smooth and striated muscles, for instance, distinguishing the striated muscle cells lining the bell (subumbrella) and velum (Fig. 3.1c,d; see Fig. S14 in [49]) [150, 234]. Within known cell types, clustering revealed an unappreciated degree of cell heterogeneity, yielding novel subtypes. For example, eight cell types could be distinguished within the gastrodermis, six of which were designated gastro-digestive (GD A-F) on the basis of a largely shared set of marker genes (Fig. 3.1b), including enzymes associated with intracellular digestion, such as CathepsinL [234].

Digestive gland cells fell into five types expressing different mixtures of enzymes for extracellular digestion. These showed overlapping distributions in the mouth and stomach regions of the manubrium. Two subtypes of gland cells (type C and E) were also present within the gonad gastroderm. Four broad clusters corresponding to neural cells each appeared to represent mixed populations, and could be subdivided by further analyses to define 14 likely subpopulations of neurons (see below). Seven major clusters could be assigned identities as nematocytes (the sting-

ing cells of the jellyfish) at different developmental stages, where surprisingly the more mature nematocytes, later distinguished by in situ hybridizations, showed little enrichment of known nematocyte markers but highly conserved proteins of the actin-rich 'stereovilli' of vertebrate hair cells.

A remarkable feature of the *Clytia* medusa is that it constantly generates many cell types, notably neural cells and nematocytes from prominent i-cell pools in the tentacle bulb epidermis [65], as well as at other sites [151]. Within our dataset, we thus expected to be able to capture dynamic information relating to the development of i-cell derived cell types, similar to that extracted from *Hydra* polyp single-cell transcriptome data [225]. Unlike *Hydra*, we found no clear developmental connection between i-cells and gland cells, and little to no expression of markers of the common neuronal-gland cell precursors identified in *Hydra* [225], though potential connections appeared in the 2D embedding, leading to the investigations in Chapter 5.

To identify how far along the development pathway(s) the nematocyte and neural cells were, we used pseudotime analysis [105] to assign values to each cell along the trajectory from i-cell to 'mature' nematocyte or neuron (using the PCA-reduced data as input) . The purpose of this analysis was to uncover known as well as novel genes related to these development progressions. This revealed expression of genes not previously associated with nematocyte development (such as *Znf845* and *Mos3*), and the downregulation of known nematocyte-markers into the development of the mature cells, with this downregulation linked to expression of rare markers such as the M14 peptidase. We identified genes changing over the course of pseudotime through a random forest regression model, determining which genes were good predictors of a cell's pseudotime (grouping these pseudotime values, between 0 to 1, into quantiles).

The neurons in contrast, appeared to have a more clustered structure, of distinct, mature subpopulations, though in future more quantitative measures of model fit between discrete or continuous data representations such as presented in [81] could be used. We thus re-clustered the neural supergroup in Fig. 3.1b, selecting HVGs just over these cells (and performing the same PCA-reduction prior to clustering). Interestingly, we found clusters marked by distinct and specific expression of putative neuropeptide precursors, further validated with in situ hybridization. Several of these neuropeptides were previously regarded as unlikely candidate sequences, based on sequence homology. However in combination with specific expression or activity in

these neurons, we were able to re-identify these sequences as neuropeptide markers. Other forms of neurotransmission (e.g., chemical neurotransmission) were harder to identify in *Clytia*, based on expression patterns alone (also due to low sequence homology), but deeper sequencing and unbiased transcript capture (as well as protein readouts) may provide greater insight.

3.3 Cell State Shifts in Response to Starvation Across the Cell Atlas

To assess the transcriptional impact of starvation, we mapped individual cells to their corresponding control or starved labels. As there are around 60% fewer cells in a starved animal (see Fig. S2 in [49]), we first asked whether there were significantly different *numbers* of cells per cluster between control and starved conditions. We found that only one cluster had a significant difference (cluster 11, early nematoblasts in Fig. S5B in [49]), suggesting a nearly uniform reduction across cell types in the starved condition. Given the cell type resolution of the atlas, as determined operationally by clustering, we then asked how drastic the transcriptional changes incurred by perturbation were in comparison to the transcriptional differences defining the cell types, i.e., are the perturbation-induced changes encompassed within these cell type designations or are they larger in magnitude. We thus compared distances between control and starved cells within clusters to the distances between clusters (across all their cells). As a metric we used the L_1 distance, the sum of the absolute differences between centroid coordinates in PCA-reduced space . We found that the L_1 distances between control and starved cells within a cell type, versus between cell types (regardless of condition), formed nearly non-overlapping distributions (Fig. 3.2a). This suggests that, overall, in *Clytia* the transcriptional responses to starvation are defined by cell state shifts, and their cell type repertoire is well represented by the original clusters. We chose to use the L_1 distance metric as it tends to better retain relative distances in high dimensions, particularly in comparison to the commonly used Euclidean distance or other higher L-norms [2, 192], also discussed in Chapter 5.

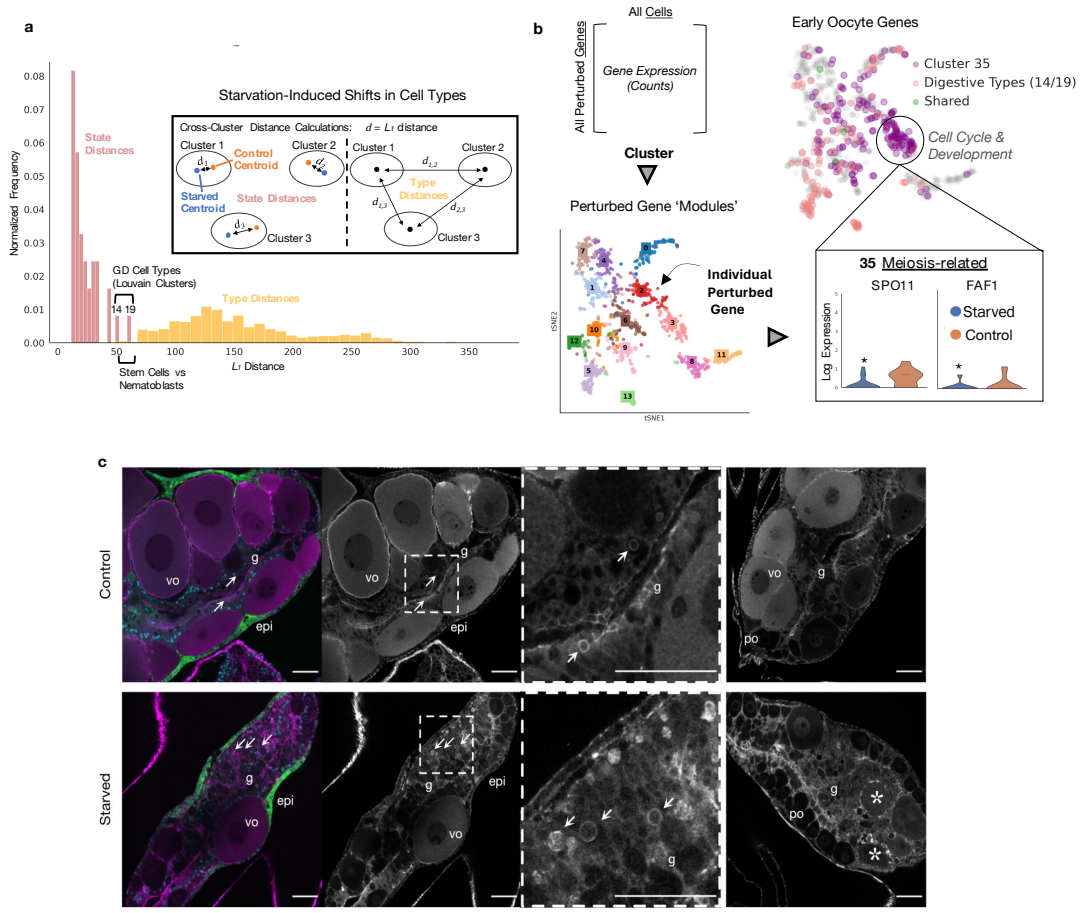


Figure 3.2: **Atlas-Wide Perturbation Analysis.** **a)** Histogram of L_1 distances between centroids of control and starved cells within cell types, versus pairwise L_1 distances between centroids of the 36 cell types. Clusters 14 and 19, with the largest internal distances, and clusters with smallest distances are denoted. Inset illustrates inter- and intra- L_1 calculations. **b)** Workflow for extracting 'perturbed' genes per cell type and clustering on genes to extract 'modules'. Visualization of perturbed genes (among the embedded modules) of early oocytes (cluster 35). Violin plots showing expression profiles for several perturbed genes in functional categories of interest. * p-value < 0.05 from non-parametric Wilcoxon test. Horizontal lines show quartiles and width of violins denote density of points. **c)** Confocal sections through gonads from control and starved medusae with cell morphology revealed by phalloidin staining of cell boundaries (magenta/grey). The first panel of each row shows co-staining of nuclei with Hoechst (blue) and endogenous GFP4 (green) in the outer epidermis (epi). Vitellogenic oocytes (vo) are largely absent during starvation, leaving a majority of pre-vitellogenic oocytes (po). The gastroderm (g) is heavily reorganized, with evidence of active phagocytosis (vesicles arrowed) and disintegrating oocytes (asterisks). The third panel in each row is a higher magnification of the boxed area in the second panel, and the fourth panel shows a second example gonad for each condition. Scale bars all represent 50µm. Adapted from [49].

However, the impact of starvation was variable across cell types, as reflected by the range of internal (state) distances (Fig. 3.2a). Starvation produced the largest perturbations in cells of the gastrovascular system, causing control-vs-starved distances large enough to overlap with the smallest inter-type distance, i.e., that between the stem cells and nematocyte precursors (Fig. 3.2a). This distinction between state shifts and type was also clearly visible in the lack of overlap between the distributions of inter- and intra-cluster distances within the second, stimulation experiment (see Fig. S6E in [49]). Though classification and distinction of cell state and type is a complex task [249], this analysis, based on relative distance in transcriptional space, provides a quantitative basis for delineation of type/state effects that may be useful in other contexts. We additionally validated the ability of this method to recapitulate the magnitude of state shifts in response to graded stimuli as well as state versus type distinctions, on two other published, multi-perturbation datasets (see Fig. S19 in [49]).

To characterize gene-level responses underlying these starvation-induced shifts, we then asked if responses are shared or unique across the cell types and compared the extent of the responses, in terms of gene quantity and expression level, across the atlas. For each cell type, we collected genes that were differentially expressed under starvation ('perturbed genes' Fig. 3.2b), and clustered perturbed genes into apparent 'gene modules' [250] by their patterns of co-expression across cells. This effectively uses the same clustering procedure as above, but on a transposed cell x gene matrix. We assigned putative functions to these gene modules through GO term enrichment, giving a global view of affected processes (see Fig. S20 in [49]). We found that certain gene modules were broadly shared across cell types, while others were almost entirely cell type-specific (see Fig. S20 in [49]). Striking examples include gene module 5, which is enriched in proteolytic genes (see Fig. S20 in [49]) and has shared expression across multiple GD cell types. In comparison, gene module 3 is largely composed of early oocyte gene expression (70%), and is enriched in cell cycle and developmental genes, which are commonly enriched in growing oocytes (Fig. 3.2b). Changes in expression of these genes likely reflect the processes of oocyte phagocytosis activated in the gonads of starving animals (see below).

To examine how *individual* perturbed genes are distributed across cell types, we visualized, for each cell type, how many perturbed genes it had, and how many of these genes are unique versus shared with other cell types (see Fig. 6E in [49]). We

found a large number of perturbed genes (72%) were cell type specific. For the most perturbed cell types, we examined whether the state shifts that we had observed were due to changes in a large or small number of genes, and how highly these genes were expressed. Consistent with the marked shrinkage of the gonads during starvation treatment, early oocytes contained the highest number of perturbed genes, which were spread across many gene modules (Fig. 3.2b).

In accordance with these distinct responses in GD cells and oocytes, comparison of the cellular organization of gonads from control and starved medusae revealed major reorganization of both the gastrodermis and the oocyte populations (Fig. 3.2c). Most strikingly, the population of mid-sized, growing oocytes, which progress daily through vitellogenesis in conditions of normal feeding [12], was largely depleted following starvation, leaving a majority of pre-vitellogenic oocytes (Fig. 3.2c). A sparse population of large oocytes in starved gonads likely results from growth of a minor subpopulation of oocytes fueled by recycling of somatic tissue and oocytes (disintegration and phagocytosis of smaller oocytes visible in Fig. 3.2c, asterisks). Consistently, GD cells in many parts of the gonad lost their regular epithelial organization and, despite the absence of any external food supply, showed evidence of active phagocytosis involving variably sized vesicles (arrows in Fig. 3.2c). Changes in organization and activity of the gonad gastrodermis were also evident from in situ hybridization images for the GD cell marker CathepsinL, while reduced expression was confirmed for a protease (ShKT-TrypA) expressed in gland cell types A and B positioned within the manubrium gastroderm, which is down-regulated during the starvation treatment. Shifts between gonad gastrodermis organization and transcriptional profiles induced by starvation thus accompany activation of tissue autodigestion programs, and likely also the mobilization of GD cells (termed MGD for Mobilizing Gastro Digestive cells [227]) from the gonad through the gastrovascular canal system, which has been observed both under conditions of starvation and during regeneration of the feeding organ.

3.3.1 Perturbation Responses to Short-Term Stimulation Across the Cell Atlas

In addition to the starvation data results, for this thesis we additionally highlight the results of the short-term stimulation for potential IEG discovery. As shown in Fig. 3.3a, each animal was given repeated bouts of stimulation over 30 minutes, with each stimulus administered every 2 minutes. 100 uL of each stimulant (150 mM KCl, DI water, or seawater (SW) as a control) was gently added just below (or just above for KCl) each medusa by pipette. Stimuli were chosen based on their ability

to reliably induce crumpling behavior, a protective response in which the bell is drawn in towards the mouth using the radial muscle [121].

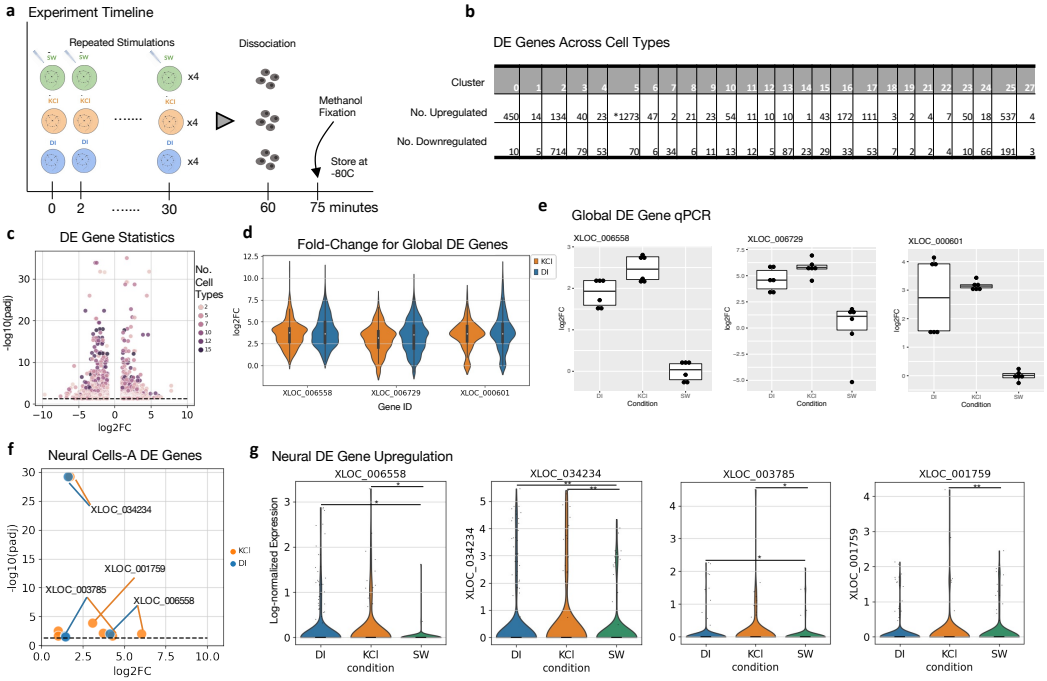


Figure 3.3: ***Clytia* Response to Ionic Stimuli a)** Diagram of the 'Stimulation' experiment. Four biological replicates (animals) used for each condition. SW denotes seawater (control), DI denotes deionized water, and KCl denotes potassium chloride. 30 minutes following the last stimulation, animals were dissociated and fixed in methanol. **b)** Summary table of numbers of up- and down-regulated genes in each cluster. *Denotes highest number of DE genes in 5, terminal differentiating nematocytes. **c)** Volcano plot of p-value and fold change for DE gene candidates. Dashed line denotes 0.05 alpha cutoff. Colors indicate the number of cell types a gene is found to be DE in. **d**) Fold change per condition across all cells for global (DE in many cell types), 'IEG' candidates. **e)** qPCR for DE gene candidates in d in both conditions. **f)** Volcano plot of upregulated DE genes in Neural Cells-9 (cluster 9), the majority of neural cells, colored by perturbation condition. Gene names denote selected candidates. **g)** Expression for cells in each condition of upregulated DE genes found in Neural Cells-9 (cluster 9) using the non-parametric, Wilcoxon test. P-values adjusted for multiple testing with Benjamini-Hochberg correction. * denotes p-value < 0.01, ** denotes p-value < 0.001. Adapted from [49].

We, similarly to the starvation analysis above, extracted differentially expressed or 'perturbed' genes in each cell type, with fold changes from the sequencing data (for perturbed genes represented in multiple cell types), and their subsequent qPCR fold changes shown in Fig. 3.3c-e. We then delved into the perturbed genes in the neural cell supergroup, pulling out candidate IEGs (i.e., genes with increased expression

within an hour time window) that demonstrated increased expression in both DI and KCl conditions, or only in the KCl condition (Fig. 3.3f,g). Further work to both characterize and implement such genes as cell activity markers could then be performed, given the genetically tractable features of the *Clytia* model [260].

3.4 Implications and Extensions of Whole-Animal Multiplexed scRNA-seq

The *Clytia* medusa single-cell atlas presented here is an important addition to the growing number of single-cell atlases across the animal tree of life. This provides the first cell-level transcriptomic characterization of a pelagic medusa stage, the most complex of the life cycle forms within the large and diverse phylum Cnidaria. Reflecting this complexity, we found greater cell type diversity in the *Clytia* medusa than in its polyp-only hydrozoan cousin *Hydra* [225]. The outer, epidermal body layer could be sub-divided into seven clusters encompassing all of the described *Clytia* muscle types, including two types of fast-contracting striated swimming muscle [150, 234]. Rich diversity was also uncovered in the inner gastroderm layer, which is elaborated in the medusa into distinct digestive compartments (mouth, stomach, gonad, and tentacle bulb) and also generates the thick mesoglea (jelly) characteristic of the medusa form. Our starvation experiment analyses revealed that these clusters were maintained operationally as distinct 'cell types' rather than 'cell states' between the two extreme conditions tested, but we cannot rule out that responses to other environmental or physiological perturbations may reveal plasticity between these clusters, for instance transdifferentiation between muscle and nerve cell types is well documented in hydrozoan medusae (overview in [150]).

In addition to the epithelial cell types of the epidermis and gastroderm, our single-cell atlas confirms the presence of an interstitial stem cell (i-cell) population in *Clytia* providing a similar set of somatic cell types to that described in *Hydra*, as well as the germ cells [151, 225]. In this medusa data we do not find strong evidence for direct progression from i-cells to gland cells, or for the shared neural-gland cell progenitors described in [225]. In contrast, our pseudotime analyses provide transcriptional signatures of the progressive stages of nematogenesis and neurogenesis from i-cells that will guide future studies of their developmental regulation. The large representation of nematogenic stages in this *Clytia* medusa scRNA dataset allowed us to link two distinct phases of nematocyte formation with extremely different transcriptional profiles. The initial phase covering nematocyst formation has been the focus of many studies [56, 64, 65, 239], but the terminal phase has been largely overlooked in previous transcriptomics studies, likely due to

the relatively low mRNA content [56, 225] and the extremely abrupt degradation of nematocyst-related mRNAs before the terminal phase [239]. We uncovered 14 mature neuronal subtypes in Clytia, which is similar to the number reported in *Hydra* and *Nematostella* [220, 225]. It is likely that further heterogeneity exists within these 14 subpopulations. Spatial expression analysis of neuropeptides that contributed to the signatures of one or more subpopulations revealed a wide variety of neuronal populations either associated with specific anatomical structures, such as the tentacles, nerve rings, and manubrium, or distributed across the medusa. How molecular cell type maps to function both within and across body parts, the roles of these peptides as primary transmitters and/or neuromodulators, and the uses, if any, of classical, small-molecule neurotransmission, remain unknown. Moving forward, with this cell atlas as the foundation, the ability to perform whole-organism, multiplexed scRNA-seq, in combination with emerging genetic tools and advantageous life history traits, makes *Clytia* a powerful, tractable platform for high-resolution systems biology.

This work further serves as a case study in using multiplexed single-cell transcriptomics to assess cellular responses to whole organism perturbations, and provides a guide for deployment in other organisms. The techniques for multiplexed experimentation that underlie this study are also well suited to large-scale perturbation studies (such as temperature, pH, or other environmental disturbances) in other marine organisms given the sea-water compatible workflow. Though the inclusion of multiple animals and conditions may currently limit the detection of very rare cell populations, as sequencing costs drop and cell throughput in scRNA-seq grows, this approach should become tractable for larger, more complex systems. The lack of library-induced batch effects demonstrates how large-scale experiments can be conducted without introduction (or minimizing introduction) of confounding factors from multiple experiments, which can be highly non-linear and difficult to account for [248]. The second perturbation dataset also demonstrates both how batch effect variability is reduced within multiplexed experiments (see Fig. S11 in [49]), by comparing the greater L_1 distances between control cells in PC-space of the two experiments, versus the smaller L_1 distances between control cells within each experiment.

By relying on expression, our strategy reduces the reliance on prior gene functional annotation, using specificity of expression to identify genes of interest, allowing for targeted annotation. This includes determination of strong diagnostic markers for

cell type definition, cell type specific and shared transcriptional responses to starvation, and 'modules' of co-expressed genes underlying these responses. The extent of these expression-based changes additionally highlights areas of the organism's biology that are strongly or uniquely affected by a perturbation. By applying simple and interpretable quantitative analyses to the various cell-type specific perturbation responses, we revealed the large-scale downregulation of gene expression in two GastroDigestive cell types and severe disruption of oocyte development under starvation. Together, this approach dramatically lowers the barriers for working with non-traditional models, and affords opportunities to match uniquely suited organisms to specific questions. Moving forward, the combination of scRNA-seq and other sequencing-based genomics techniques with multiplexing and annotation-agnostic analyses, could foster comprehensive high-resolution molecular studies of diverse organisms and their responses to numerous environmental perturbations.

Chapter 4

COMPUTATION FOR DECIPHERING PERTURBATION

In Chapter 3, we begin with a cell x gene matrix of molecule counts, and by the end, define an atlas over the populations of cells in this matrix, and extract perturbation responses or changes in expression, across those populations. As demonstrated in Chapter 3.2, there are several steps to processing this matrix just to obtain clusters (cell types) or compare expression between clusters. In this section we will review the standard data processing steps in scRNA-seq analysis, largely focused on methods implemented in the Python package scanpy [262] as much of the whole organism data processing utilized this workflow. We will not focus on the upstream processing of sequencing reads (e.g., FASTQ files) into count matrices, though the use of reference annotations and alignment algorithms in these steps are important in the interpretation of the produced counts [237]. We will begin with processing and transformation techniques applied to the count matrix, then touch on common approaches used to perform exploratory analysis of such datasets. Given these common techniques, we will then summarize analysis methods specifically for perturbation datasets and multimodal data, and their relation to the standard processing pipeline. In the proceeding Chapters, we will then address limitations of and alternatives to current practice.

4.1 Standard Count Processing

The count matrix, for mRNA expression, begins as a discrete count matrix of mRNA molecules per annotated gene or transcript. Current count matrices are actually comprised of multiple types of counts, e.g., nascent (intron-containing or unspliced) and mature (exon-containing or spliced) mRNA, thus there are in fact multiple biological measurements to consider (multiple matrices) within already published datasets [46, 237]. Though we focus only on spliced mRNA in Chapter 3, we will address the utilization of both mRNA modalities in Chapter 6.

First steps in scRNA-seq processing often include filtering out ‘cells’ that seem like empty droplets (for example in 10x Genomics) with spurious mRNA capture, and filtering for genes that have non-zero expression in the samples as well as for HVGs , i.e., genes that are likely variable as a function of the biological heterogeneity or cell types in the data [17]. This additionally reduces the number of genes from tens

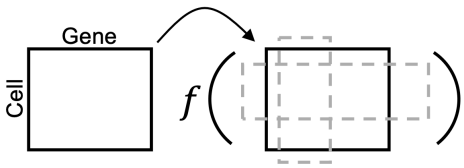


Figure 4.1: **Count Pre-processing.** Matrices are subset for cells and genes that meet expression criteria, and then transformed.

of thousands to a few thousand or hundreds of genes. Gene selection in common analysis packages [207], can require transformation of counts (into a continuous regime) prior to selection (see below).

Certain transformations are commonly applied to the count matrix, with the goal of removing noise and extracting biological signal. Counts are first 'depth-normalized', where for each cell their total counts are scaled to some common total value, then the counts are log-normalized (often with log1p) [5]. The depth-normalization assumes that sampling biases between cells are the same across all genes for the cell and result from only technical sources of variation (though it may be the case that such differences are due to 'true' biological heterogeneity).

Log-normalization represents a variance-stabilization transformation, that removes the relationships between mean and variance of genes (i.e., that higher mean genes have higher variance). ScRNA-seq counts often appear negative-binomial (displaying overdispersion) [5, 241], and the variance stabilizing transformation for a negative-binomial, as derived by Francis Anscombe, can be formulated as a similar log-transformation of the data [5]. The motivation for using such transforms is often for use with techniques such as PCA (see Chapter 4.2 below) and linear regression, i.e., approaches assuming homoscedasticity and/or where it is not desirable to have variance in data driven solely by high-expression genes [134]. There are other approaches to normalizing and transforming count data, however recent benchmarks have shown that many techniques create unintended effects on the counts (and resulting trends), thus there is a lack of consensus on the transformations to apply beyond those described here [5].

4.2 Computational Methods for Exploratory Analysis

Often the goal of generating scRNA-seq datasets is to come up with new questions and investigations of a biological system, which may or may not concord with some previous set of hypotheses or assumptions. This follows in the spirit of the ideas of

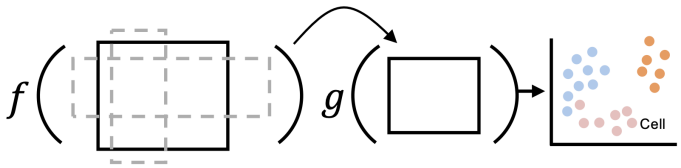


Figure 4.2: **Reduction of Data for Exploratory Analysis.** Diagram of how subset count matrices are then transformed through dimension reduction for analysis and visualization. Colors denote 'cell types'.

John W. Tukey, who coined the term 'exploratory analysis', where, as he put it, "[it] is not enough to look for what we anticipate. The greatest gains from data come from surprises" [251]. How well common practice actually follows the principles of exploratory analysis outlined by Tukey, is discussed in further detail in Chapter 5.

One of the most common techniques used to explore scRNA-seq data is principal component analysis (PCA) . PCA is used to reduce the dimensionality of the input data matrix (by selecting only the top principal components that capture more of the variance in the data), and ideally to remove noise [134, 139]. Count matrices are reduced to tens of dimensions (by PCA), then used as input into downstream analyses like unsupervised clustering, to extract populations of cells which display similar gene expression patterns (and potentially biological functions) [262]. Visualization of the transformed cells in 2 or 3D PCA-space, are also used to assess what biological properties the components are extracting [208]. Other techniques, such as nonnegative matrix factorization are also used methods to similarly represent the count data prior to downstream analysis [52].

As described in further detail in Chapter 5, several exploratory analyses often follow PCA reduction, to generate directions of investigation. These include low-dimension embeddings of the data, clustering, and pseudotime or trajectory inference, all often done in an unsupervised manner. To briefly summarize these areas, low-dimension embedding methods are used to find 2 or 3D representations of the data that display the relationships between the cells in the dataset [134, 139, 180]. Unsupervised clustering, usually through methods like K-Means or graph-based clustering methods like Louvain [61] or Leiden [247], is performed on the data to learn which cells cluster together based on expression similarities. And pseudotime or trajectory inference methods [214] are used to infer continuous relationships between cells, assuming cells are captured at different stages along some developmental timeline

or cellular process.

Each of these methods may or may not take in the count matrices in the same form. Most low-dimension embedding methods (particularly for visualization) use the normalized and PCA-reduced data, and similarly so for common clustering and trajectory inference approaches. More recent deep learning methods, based on variational autoencoders [137], which learn reduced, latent representations of the data for use with clustering methods (or other downstream analysis), can model the raw counts explicitly (as discrete count distributions) without normalization and PCA-reduction [166].

4.3 Methods for Analyzing Perturbation Data

With the advent of high-throughput perturbation data, there has also been development of analysis methods to extract insight from these data types in particular. Methods generally touch on two applications, modeling and prediction (though both can overlap within the same method) [125]

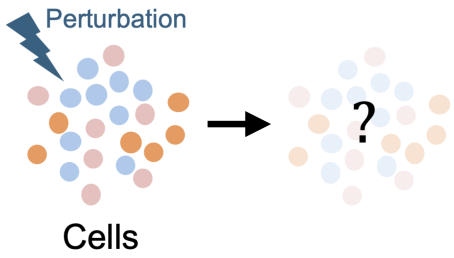


Figure 4.3: **Single-cell Perturbations.** Diagram of how perturbations are applied to single cells, where analysis addresses the interpretation of their effects (the '?').

More modeling-focused approaches, seek to understand how a perturbation affected a system through population-specific responses and interactions between features of the system (gene-gene interactions, for example). Causal inference approaches have begun to tackle the gene network inference problem using these high-throughput intervention (perturbation) data [31, 54]. Simpler linear regression models, such as in MIMOSCA [69], model the changes in gene expression as functions of the perturbations, specifically genetic interventions, to determine the contribution of each intervention to the final, observed expression patterns. Discerning populations of cells with differing perturbation responses is also of interest, following from the clustering task described above, with methods ranging from mixture model-based approaches to graph-based clustering techniques [34, 52].

The more prediction-focused approaches, aim to draw from the large datasets spanning several perturbation conditions, cell types, and potentially species, to predict responses to perturbation (e.g., change in gene expression) often using deep learning approaches [125, 170]. Data may also include protein information and phenotypic readouts of the cells, and thus the prediction may also involve predicting drug targets, perturbation interactions, and changes to chemical properties induced by perturbation [122, 125].

Each method likewise comes with its own requirements for normalization of the data prior to application, and potentially dimension reduction as well. The standard exploratory analysis techniques described above, are also often employed alongside these perturbation-specific methodologies, to visualize and assess the learned representations [127, 170].

4.4 Analysis Extensions to Multimodal Data

Though we will not delve into the details of all current techniques for data type (modality) or batch integration in this thesis, scRNA-seq count matrices may additionally come from different experiments or sequencing runs (not multiplexed together, for example), containing technical biases as a result, that potentially drown out the biological similarities between the samples [108, 253]. In addition, measurement of multiple biological entities, from nascent and mature mRNA to chromatin openness and protein counts, result in multiple data matrices to be used together in analyses.

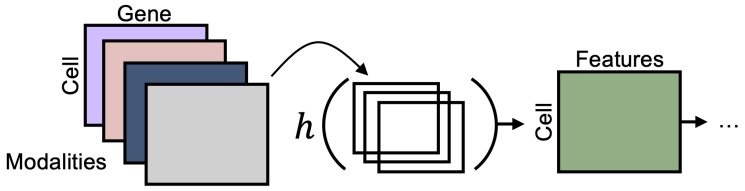


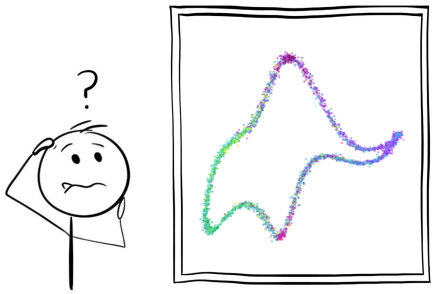
Figure 4.4: **Processing of Multimodal Data.** Diagram of modality-specific count matrices, transformed through some function or method, to produce a final, integrated representation.

There are thus numerous approaches to ‘integrate’ matrices prior to exploratory analysis [19, 108, 253]. Methods range from simply downsampling count matrices [253] across batches to merging nearest neighbor graphs across the modalities or batches [108] to deep learning approaches which learn shared latent representations

incorporating the modalities and batches present in a dataset [42, 91]. How well the integrated representations are suited to the possible downstream analysis tasks (such as clustering, differential expression, trajectory inference, etc.) has yet to be completely benchmarked, and current analyses demonstrate a tendency to remove biological variation (rather than or in addition to technical variation) during the integration process [19, 253]. We will discuss alternative approaches to this question of modality integration in Chapter 6, through biophysical modeling of technical and biological variation in molecule counts.

Chapter 5

DIMENSION REDUCTION FOR EXPLORATORY ANALYSIS



With four parameters I can fit an elephant,
and with five I can make him wiggle his trunk.
John von Neumann

The opening stages of an scRNA-seq perturbation dataset analysis often contain the preprocessing and exploratory analysis tasks of Chapter 4.1 and 4.2, as demonstrated in the whole organism analysis of Chapter 3.2. In particular, it has become increasingly common to utilize low-dimensional embedding methods to both open up exploratory analysis of such datasets (generate lines of inquiry) and to 'validate' results of other analysis tasks (such as clustering) [134, 145]. However, the properties of such methods, most commonly the UMAP [180] and t-SNE [255] methods for embedding and reduction of data, are not well defined, with previous works noting preservation, or lack thereof, of local and global relationships between embedded points [57, 139]. When examining a dataset such as the *Clytia* atlas through this lens, this raises questions of what observed patterns we can trust, how to quantitatively assess these patterns, and what the visualization overall is attempting to highlight from the data?

This initially prompted an investigation of how such embedding methods could be adapted to more explicitly preserve quantitative properties, such as distances between cell types in gene-space, in 2 or 3 dimensions. However, coupling more quantitative objectives with the objectives/algorithms of t-SNE or UMAP resulted in distortion of whatever quantitative preservation was obtained by the first objective. This then inspired an investigation into the properties of common practice, low-dimensional embedding for exploratory analysis and the extent to which biological

insight is preserved (or distorted).

5.1 The Specious Art of Single-Cell Genomics

This section summarizes the contents of [48] by T.C and L.P. T.C. and L.P. conceived of the study, T.C. performed analysis and developed code, T.C. and L.P. wrote and edited the manuscript.

Ostensibly, the goal of dimensionality reduction of high-dimensional genomics data is to filter noise, enable tractable computation, and facilitate exploratory data analysis (EDA) . The objectives of common techniques thus focus on preserving and extracting local and/or global structures from the data for biological inference [134, 139, 267]. Trial and error application of common techniques has resulted in a currently popular workflow combining initial dimensionality reduction to a few dozen dimensions, often using PCA , with further non-linear reduction to two dimensions using t-SNE [255] or UMAP [109, 134, 139, 180]. For single-cell genomics in particular, these embeddings are used extensively in qualitative and quantitative EDA tasks which fall into four main categories of applications (Fig. 5.1, 'Application'):

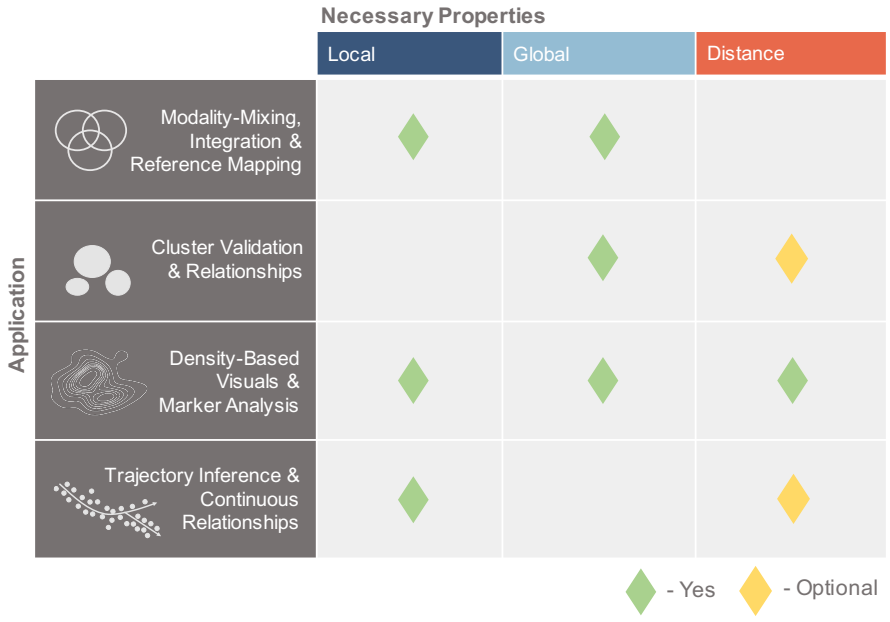


Figure 5.1: **Necessary Properties for Embedding Applications.** Application rows denote biological tasks, and columns denote which properties are necessary, i.e., key geometric properties whose preservation or representation is assumed in the task. Adapted from [48].

- Modality-Mixing, Integration, and Reference Mapping:
  Embeddings are used to visually assess the extent of integration, mixing, or similarities between cells from different batches [4, 73, 108] and to compare methods of integration/batch-correction [112]. For query dataset(s) mapped onto reference datasets/embeddings, visuals likewise provide an assessment of merged data similarities or differences [24, 132].

- Cluster Validation and Relationships:
  Visual applications range from assessing the existence of and relationships between predefined clusters, to inferring properties of the clusters (e.g., spread or heterogeneity) [3, 134, 139], and to generating the clusters themselves from the two-dimensional space (e.g., to define cell types or detect doublets) [66, 197, 267].

- Density-Based Visuals and Marker Analysis:
  Embeddings are used to justify or measure changes in cell populations between different conditions, by comparing contour locations and sizes in the density diagrams, as well as changes in intensity or spread of gene expression [16, 130, 230, 244, 271].

- Trajectory Inference and Continuous Relationships:
  Embedding applications range from implying or inferring local, continuous relationships between cells and assigning pseudotime coordinates [40, 146, 214, 250], to using the two-dimensional coordinates for explicit calculations of magnitude and direction of developmental progression [118, 146, 172].

Inherent in these applications are assumptions of preservation of local and global cell properties, as well as distances, delineated in Fig. 5.1. For each application, we demarcate which of these are the 'necessary' or key geometric properties that each task inherently assumes to be represented (and preserved). Based on previous works [3, 109, 140, 194] and the objective functions of UMAP and t-SNE [180, 255], 'local' is defined as nearest neighbor relationships, 'global' as neighbor relationships and properties of groups of cells (e.g., cell types), and 'distance' as Euclidean distance (L_2 norm) or Manhattan distance (L_1 norm) between points. Note that preservation of distance implies preservation of local and global properties. We utilize the L_2 norm as it is the default metric of UMAP/t-SNE. We also present results with the L_1 norm (see S1 Text in [48]), as L_1 is more suitable for measuring distance in high dimensions, particularly in comparison to other L_k norms [2, 23], and is commonly

applied to transcriptomic data [228, 254, 259], with comparable performance to the probabilistic Jensen-Shannon divergence in single-cell distance calculations [192].

Yet, despite the goals of these methods [109, 134, 267] to preserve local and/or global structure, there is little theory or empirical analysis to support these claims. For example, while the popular t-SNE and UMAP methods claim faithful representation of local and/or global structure in low dimensions [134, 139, 180], there is evidence they fail in this regard [57, 139], and theorems providing guarantees on the embeddings rely on numerous assumptions unlikely to hold in practice, and ignore the preprocessing by PCA prior to non-linear reduction [162].

Here we assess dimensionality reduction for single-cell gene expression, first investigating the preservation of the necessary properties comprising the columns of Fig. 5.1, then assessing the impact of these embeddings across the applications comprising the rows of Fig. 5.1.

5.1.1 Preservation of Local and Global Structure in 2D Embeddings

We begin with the columns of Fig. 5.1, and assess the preservation of these properties by two-dimensional embedding, as compared to the ambient space or higher-dimensional PCA space to which the ambient space is initially reduced prior to reduction to 2D.

'Ambient' space refers to the gene count matrix after HVG selection and log-normalization of the counts with scanpy [262]. We denote 'PCA-preprocessing' as the higher dimensional reduction of the ambient space by PCA, followed by a (non-linear) reduction to 2D (e.g., 'PCA-50D→UMAP') which mimics standard practice. Additionally, cell annotations or labels (such as cell type or condition) used in the following analyses were taken from the original studies. All count matrices used in this analysis contain the spliced mRNA counts only.

All PCA reduction was performed to 50 dimensions by default, unless otherwise noted. The t-SNE and UMAP algorithms were applied to the higher dimensional PCA embeddings with default settings. This sequence of dimension reduction by PCA first, prior to reduction to 2D by UMAP/t-SNE, is denoted as 'PCA-preprocessing'. The effect of a single parameter (n_neighbors) change is shown for UMAP embeddings in Fig. 5,6 and Fig. P,R-V in [48], but we did not adjust parameters beyond this. As per the discussion in [57], though slight changes in these aesthetic parameters can drastically impact low-dimensional embeddings, the choice of parameters for tuning is often informed by empirical observations/prior

knowledge leaving open the question of which metric(s) to use for determining 'optimal' parameters. Notably this tuning is also contradictory to the common use or desire of such techniques to produce 'unsupervised' representations of the data.

Local Preservation

Given the focus on preserving local nearest neighbors in the objectives of the UMAP and t-SNE methods, we first measured the recapitulation of nearest neighbors in 2D embeddings, as compared to the neighbors defined in ambient space. We used Euclidean (L_2) distance, the default for these non-linear reduction methods, to define each cell's 30 nearest neighbors and measured Jaccard distance (dissimilarity) between the neighbors in embedding and ambient space, defined as $1-\frac{|A \cap B|}{|A \cup B|}$ where A, B represent the sets of each cell's 30 nearest neighbors in the ambient and latent spaces, respectively. A Jaccard distance of 0 denotes completely overlapping sets, and 1 denotes completely non-overlapping sets of neighbors.

Several in vivo datasets were reduced to 2D, with PCA-preprocessing, including 10x Genomics and SMART-Seq assayed mouse ventromedial hypothalamus (VMH) neuron datasets [135], an ex-utero cultured mouse embryo dataset (at the E8.5 stage) and an ex- and in-utero mouse embryo dataset (at the E10.5 stage) from [4], and a mouse primary motor cortex (MOp) dataset [275]. We additionally reduced cell culture-derived datasets, with and without external perturbations, including mouse Embryonic Stem Cells (mESCs) treated in DMSO from [67] and multiplexed mouse Neural Stem Cells (NSCs) in 96 drug combination conditions (labeled '96-plex') [92] (see Table A in S1 Text of [48]).

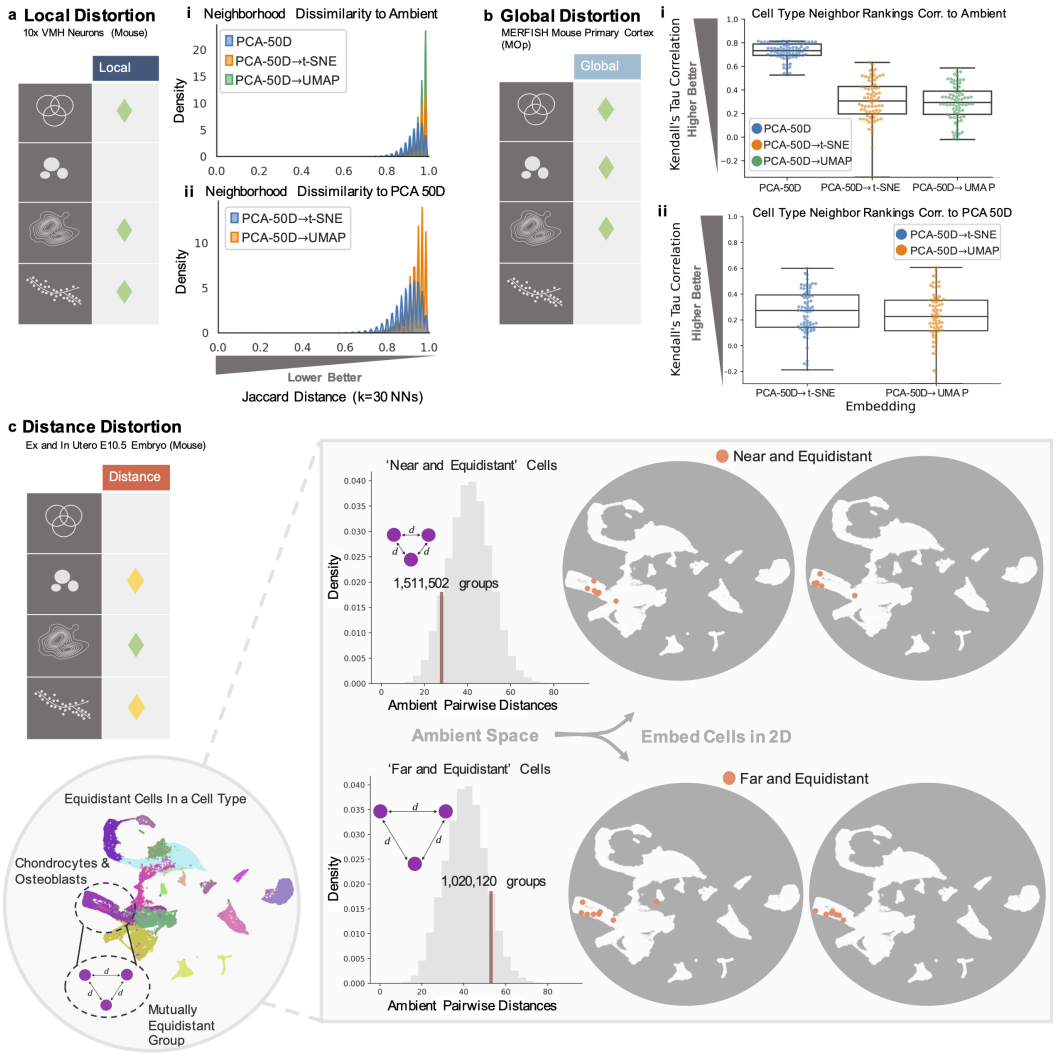


Figure 5.2: **Distortion of Necessary Properties in Embeddings. a)** i. Distribution of Jaccard distance of cell neighbors in PCA-preprocessed 2D embeddings and the relevant PCA space, as compared to ambient space. ii. Distribution of Jaccard distance of cell neighbors in PCA-preprocessed 2D embeddings, as compared to the higher dimensional PCA space. **b)** i. Boxplot of correlations of cell type neighbor rankings to ambient space for the PCA-preprocessed 2D embeddings and the relevant PCA space. ii. Boxplot of correlations of cell type neighbor rankings to the relevant higher dimensional PCA space for the PCA-preprocessed 2D embeddings. Embeddings generated n=3 times. **c)** Selection of equidistant groups with 'near' or 'far' distances in ambient space. UMAP embedding of the data in grey circles, with orange circles denoting all cells within the previously determined equidistant groups. Adapted from [48].

The 2D t-SNE/UMAP embeddings (e.g., 'PCA-50D→UMAP' in Fig. 5.2a) displayed large Jaccard distances with respect to the neighbors in ambient dimension, with an average consistently above 0.7 (70%). Interestingly, the embeddings of the more homogeneous mESCs dataset displayed relatively higher dissimilarity despite the small number of cells (see Fig. Bb and Bc in S1 Text in [48]). Poor neighborhood overlap was additionally retained, and often worsened, without PCA-preprocessing (i.e., direct reduction to 2D from ambient space). In some cases, the dissimilarity of neighbors was worse for two-dimensional PCA ('PCA-2D') as compared to t-SNE or UMAP reduction without PCA-preprocessing, consistent with other findings on the poor preservation of local neighborhoods by both PCA and the non-linear reduction methods [57, 139] (see Figs A and Bc in S1 Text in [48]). Similarly poor neighbor retention from the ambient space was observed in the higher dimensional PCA spaces as well ('PCA-50D' Fig. 5.2a i) [57], particularly for larger datasets. Even between the PCA-preprocessed 2D embeddings and their corresponding PCA space, Jaccard distances were consistently above 0.8 on average, regardless of the dimension of the initial PCA reduction (Fig. 5.2a ii).

Global Preservation

Turning to global relationships, we measured the preservation of the rankings of neighbors of cell 'types' rather than individual cells. Cell 'types' denote either author-provided cell type (Fig. 5.2b ii) or cell condition annotations. Rankings were constructed from average pairwise distances between the cells of the different types. For the same datasets as above, and a multiplexed dataset of human monocytes treated with 40 drugs [52], correlation of cell type neighbor rankings to that of the ambient space were low ($\leq$ 0.4) in PCA-preprocessed 2D embeddings , and at least 33% lower than those of the higher dimensional PCA spaces, with warped or even reversed correlations in comparison to the ambient (Fig. 5.2b i) or relevant PCA space (Fig. 5.2b ii, see Fig. Ca in S1 Text in [48]). These distortions were not specific to the distance measure used; we observed similar results when using the L_1 norm to determine cell type neighbors (see Fig. Cb in S1 Text in [48]). This is consistent with observations made in other studies [109, 140]. In general, correlation decreased over each step in the reduction process though there was not a clear trend related to other dataset properties (see Fig. Da and Ea in S1 Text in [48]). For analyses of recapitulation of cluster properties such as inferred heterogeneity or spread, see 'Clustering Validation and Relationships' and 'Embedding Properties are Arbitrary' below.

Distance Preservation

To examine distance preservation, we extracted groups of cells with quantitatively distinct relationships in the ambient space of the Seurat-integrated [108] ex- and in-utero mouse embryo dataset (at the E10.5 stage) [4], specifically equidistant groups of cells, where the distances between cells were all either equally small ('near') or large ('far') (Fig. 5.2c). This revealed upwards of 2.5 million such groups, with 3 to 8 cells in each (see Fig. Fa and Fe in S1 Text in [48]). However, once embedded into two dimensions, these quantitatively distinct groups of cells (orange dots on UMAPs, Fig. 5.2c) displayed the same dispersion patterns, violating distance preservation, and rendering these distinct, transcriptomic relationships indistinguishable.

This is not surprising, given previous theoretical work on the limits of distance preservation in low dimensions, particularly for equidistant points [20, 21, 174]. The Johnson-Lindenstrauss Lemma on the optimality of linear embedding [129, 147, 148] shows that preservation of pairwise distances with a margin of error of at most 20% for a modestly sized dataset of 10,000 cells would require at least 1,842 dimensions [60]. Distortion is inevitable: as shown in Theorem 1 below, given n points embedded in two dimensions, the distortion of the ratio of their maximum distance, D, to minimum distance, d ('max/min ratio'), grows as $O(\sqrt{n})$ [164].

Induced distortion has been investigated in the literature for various conformations and embedding of points, e.g., the minimum distortion bound for embedding an n-point spherical metric onto a line [20] (akin to pseudotime inference), and the number of dimensions required to embed a metric space into a low-dimension normed space (defined by some l-norm) [174]. However, investigation of the implication of these bounds in real datasets across the sciences has been limited. In this case, we focus on equidistant points, which can represent equally similar or dissimilar cells, and their distortion in two-dimensions to provide a more concrete realization of such bounds in the context of single-cell gene expression.

A trivial case is the result that no more than three points can be equidistant points in $\mathbb{R}^2$ (no more than $n + 1$ points in $\mathbb{R}^n$). This raises the question of how close to equidistant more than three points in $\mathbb{R}^2$ can be as even near-equality is impossible; specifically, a lower bound on the ratio between the maximum and minimum pairwise distances shows that distortion, which increases with the number of points, is inevitable.

A straightforward way to see this is via the two-dimensional isodiametric inequality

which states that among all shapes of a given diameter, the circle has the greatest area (for a simple proof see [164]). Formally, for any body in $\mathbb{R}^2$, the area A is bounded above by $\frac{\pi}{4}$ times the square of the diameter D (the supremum of distances between any pair of points), i.e.

$$A \leq \frac{\pi}{4} D^2. \tag{5.1}$$

Theorem 1 *Given $n \geq 3$ points in $\mathbb{R}^2$, let d be the minimum distance among all pairs of points, and D the maximum distance (i.e., the diameter). The ratio of D to d satisfies*

$$\frac{D}{d} \geq \sqrt{\frac{n-2}{2}}. \tag{5.2}$$

Proof: Let B be the set of points consisting of the convex hull of n points in $\mathbb{R}^2$, and let I denote the remaining points, with $|B| = k$ and $|I| = n - k$. Note that for each point in I, there exists a semi-circle of radius $\frac{d}{2}$ centered at the point that does not touch any other point, or extend beyond the convex hull of the points (Fig. 5.3). If we denote the sum of the areas of these semi-circles by A_I, we obtain

$$\begin{aligned} A_I &= \frac{1}{2}\left(\pi\left(\frac{d}{2}\right)^2\right)(n-k) \\ &= \frac{\pi d^2}{8}(n-k)\ . \end{aligned}$$

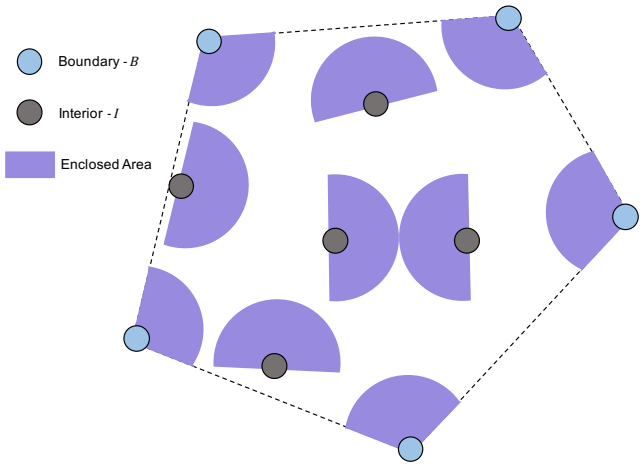


Figure 5.3: **Bounding the Area Enclosed by Points in Two-Dimensions.** Example of a set of 10 points showing the enclosed area for points in the I and B sets in the proof of Theorem 1. Adapted from [48].

Furthermore, for each of the k points in B, there is a circle sector of radius $\frac{d}{2}$ spanning the interior angle of the convex hull at that point that does not touch any other point, or extend beyond the convex hull. Since the sum of the interior angles of a k-gon is $(k-2)\pi$, we find that the sum of the areas of the circle sectors, which we denote by A_B, is given by

$$\begin{aligned} A_B &= \pi\left(\frac{d}{2}\right)^2\left(\frac{(k-2)\pi}{2\pi}\right) \\ &= \frac{\pi d^2}{8}(k-2)\ . \end{aligned}$$

Summing A_I and A_B, we obtain a bound for the area enclosed by the n points:

$$\begin{aligned} A &\geq A_I + A_B \\ &= \frac{\pi d^2}{8}(k-2) + \frac{\pi d^2}{8}(n-k) \\ &= \frac{\pi d^2}{8}(n-2)\ . \end{aligned} \tag{5.3}$$

Combining the upper (5.1) and lower (5.3) bounds for the area A, we find that

$$\begin{aligned} \pi\frac{D^2}{4} &\geq \frac{\pi d^2}{8}(n-2) \\ \Rightarrow \frac{D}{d} &\geq \sqrt{\frac{n-2}{2}}\ . \end{aligned} \tag{5.4}$$

In practice, measuring these 'max/min ratios' in 2D embeddings, for the ex- and in-utero data (E10.5) as well as the 10x VMH neurons, revealed 4- to 200-fold increases in these ratios whether compared to the relevant PCA space or ambient space (with or without PCA-preprocessing). This was the case in groups of equidistant cells as well as groups of nearest neighbors (see Fig. F and G in S1 Text in [48]), and can result in trends such as displayed in Fig. 5.2c, with cells shot out across the embedding. For both datasets, we empirically verified the growth of this distortion with the number of cells considered in each equidistant group, i.e., as more cells are considered in 2D, the distortion grows (see Fig. H in S1 Text in [48]). Higher dimensional PCA spaces largely maintained similar max/min ratios to the ambient space (see Fig. G and H in S1 Text in [48]). However, we note that in low dimensions PCA embedding of equidistant points is tantamount to applying a random projection, similarly resulting in projected points displaying numerous mirages of structure or outliers (see Fig. I in S1 Text in [48]).

5.1.2 Distortion of Trends in Applications

Given the distortions of the necessary properties in Fig. 5.1, we then investigated their impact on each row or application, i.e., how in practice such embeddings affect the inferences and implications made in each application. Though each application is covered in depth in [48], we will focus here on the use of low-D embeddings for assessing dataset 'mixing' and cluster validation.

Modality-Mixing, Integration, and Reference Mapping

Malleability of 'structure' under low dimensional embedding is particularly apparent in the mixing properties of integrated, mapped, or batch-corrected datasets, where an integration procedure is accompanied by an embedding of the melded datasets (Fig. 5.4, see Fig. J in S1 Text in [48]) [4, 108]. This relies on preserving both local relationships (which cells are mixed) and global patterns (overall trends of mixing or non-mixing between datasets). For the integrated ex- and in-utero dataset (E10.5), we calculated the fraction of each cell's nearest neighbors with the same label as the cell, to compare whether embeddings accurately reflect the extent of mixing of ex- and in-utero cells by integration (Fig. 5.4a).

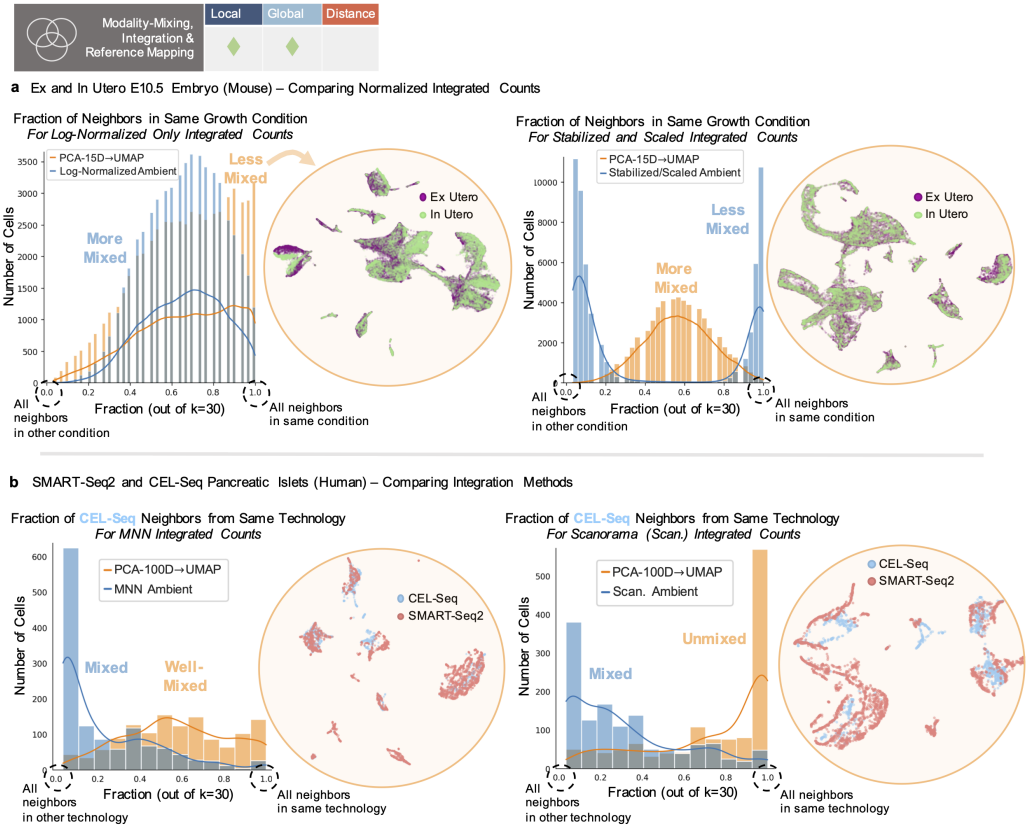


Figure 5.4: **Distortion of Mixing Patterns. a)** Left plot shows 'Log-normalized' ambient (blue) and 2D embedding (orange) distributions of mixing (fraction of cell neighbors in the same condition), where 1.0 is no mixing. Corresponding UMAP shown next to it. Right plot shows 'Variance-Stabilized and Scaled' ambient (blue) and 2D embedding (orange) distributions of mixing (fraction of cell neighbors in the same condition). Corresponding UMAP shown next to it. **b)** Left plot shows 'MNN Integrated' ambient (blue) and 2D embedding (orange) distributions of mixing (fraction of cell neighbors in the same condition) for CEL-Seq cells. Corresponding UMAP shown next to it. Right plot shows 'Scanorama Integrated' ambient (blue) and 2D embedding (orange) distributions of mixing (fraction of cell neighbors in the same condition) for CEL-Seq cells. Corresponding UMAP shown next to it. Adapted from [48].

The 'Log-Normalized' integrated, ambient data displayed a largely unimodal, well-mixed distribution of cells between conditions, while the distribution generated from embedding into two dimensions was shifted towards unmixed (left side, Fig. 5.4a). The 'Variance-Stabilized and Scaled' integrated, ambient data (a separate scaling procedure in the Seurat [108] package, performed after integration) displayed the opposite trend. The ambient data presented a bimodal distribution with completely unmixed cell populations, while the final embedding displayed a unimodal distribution of well-mixed cells from both conditions (right side, Fig. 5.4a).

Such mixing patterns are not only used to argue that different datasets are similar, but also to argue for the superiority of one integration method over another. To assess whether such inferences are legitimate, we merged the SMART-Seq2 and CEL-Seq pancreatic islet datasets utilized in [112] with one of two methods, MNN [106] or Scanorama [112]. Looking at the fraction of mixing of CEL-Seq cells in the merged ambient space reveals similar mixing by both methods (CEL-Seq cells 'mapped' to SMART-Seq2 cells) (ambient distributions, Fig. 5.4b). However the UMAP embeddings provide opposite pictures, with MNN appearing to result in a well-mixed distribution of CEL-Seq cells (left side, Fig. 5.4b) and Scanorama an unmixed distribution of cells (right side, Fig. 5.4b). In cases where batch correction largely fails (see Fig. Kb in S1 Text in [48]), the 'integrated' ambient spaces (by either method) are similar to the pre-integrated ambient space. However, reduction to 2D can enhance mixing for the 'integrated' spaces, but decrease mixing in the pre-integrated space. We found similar distortions when the L_1 norm was used, and with t-SNE as used in [112] (see Fig. Jb, Jc and Ka in S1 Text in [48]). Notably, the initial PCA reduction can drive the reversal or distortion of mixing trends, though removal of PCA-preprocessing does not alleviate this issue (see Fig. Jc and Ka in S1 Text in [48]). Thus, for a user, it is unclear what patterns of mixing are a result of the efficacy of the integration method, or arbitrary variation introduced by the dimensionality reduction procedure.

Cluster Validation and Relationships

Beyond the use of dimensionality reduction to 'validate' dataset merging, it is common to use two or three dimensional visuals to assess appearances of clusters. This can be to justify or directly generate cluster or cell type assignments [66, 134, 139, 197, 267], and to infer properties of clusters (their heterogeneity, separation, or similarity) [3, 109]. Such uses rely on retention of global relationships (Fig. 5.1), where local neighbors are less important compared to maintaining

group assignment or patterns of separation between groups. Distance preservation may also be necessary if conclusions are to be drawn on the extent of separation or locations of clusters (Fig. 5.1). However, across datasets of various sizes [135, 145] the prediction of a cell's label (cell type or condition) based on its neighbors is consistently worse in the 2D embedding space than in higher dimensional representations, even when labels are given as with supervised UMAP (UMAP Sup.) (Fig. 5.5a). Each dataset where cell type was predicted (the VMH neurons , the ex- and in-utero E10.5 embryos, and the developing mouse brain) additionally represented different methods of cell type assignment. The 96-plex NSCs provided an example of externally labeled cells, in this case by the cell's condition.

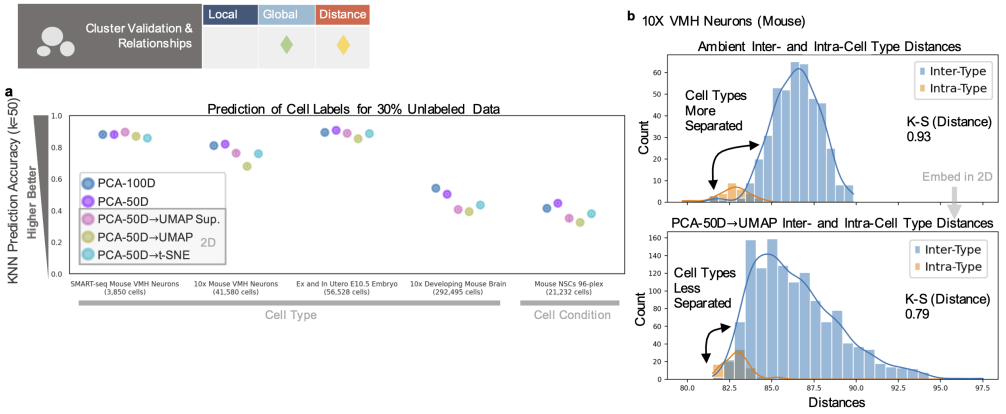


Figure 5.5: **Distortion in Cluster Validation and Relationships. a)** Prediction of cell label on 30% of the data, based on the labels of the 50 nearest neighbors. **b)** Distributions of cell type inter- and intra-type distances for the ambient or reduced space (bottom). K-S distance (the Kolmogorov–Smirnov statistic) shown as measure of separation, where higher values denote greater separation. Adapted from [48].

Additionally, by comparing the distribution of pairwise distances between cells of different cell 'types' ('inter-type') to the distribution of distances between cells within the same types ('intra-type'), we can measure how separated those distributions are, i.e., how separated or distinct cell types are from each other (Fig. 5.5b). 'Type' refers to either cell type or cell condition (see Fig Db in S1 Text in [48]) annotations. Though it may be desirable for the low dimensional visualizations to increase separability or clarify cell types as compared to the ambient space, such reduction can have the opposite effect (Fig. 5.5b), reducing the gap between inter-, intra-type distributions for some datasets and increasing the gap for others, whether using the L_2 or L_1 norm (see Fig. Db, Eb, N and O in S1 Text in [48]).

We found that cluster structures were additionally highly sensitive to the number of neighbors (perplexity for t-SNE) used in constructing non-linear embeddings, a commonly tuned parameter which can range from 1-10% or less of the data [109, 139], in line with other results on the effects of tuning [109, 140]. For the in-utero E10.5 dataset, common choices for this parameter result in different placements and overlaps of cell types, pushing progenitor populations away from their downstream cell states/types or incorrectly merging distinct, early stage populations (see Fig. P in S1 Text in [48]). Such inconsistencies have led to publication of incorrectly surmised differentiation trajectories from apparent relationships between cell types [10]. Even in a non-biological, machine learning (ML) , benchmark dataset [63], we found a muddling of cluster structures, with points belonging to different digits mixed within 'digit-specific' clusters (possibly hidden by order of points plotted), though high accuracy classification is possible in higher dimensions [37] (see Fig. Q in S1 Text in [48]). This reveals an assumption of distortion cancellation in interpreting such visuals, i.e., that relevant trends will pop out despite spurious distortion/noise, and a reliance on prior knowledge of ground truth labels (or expected trends) to determine how to interpret the 2D embedding and when tuning of the aesthetic parameters is sufficient.

Density-based Visuals and Marker Analysis

Density assessments of points in 2D embeddings are frequently used to quantitatively assess cell-cell relationships by directly relying on distances between the cells in two dimensions (Fig. 5.1). Common applications compare densities of cells in different conditions or batches, within a shared embedding space, to make statements on changes in population density or expression between groups [16, 130, 230, 267]. However, as demonstrated above, parameter tuning easily disrupts the placement of cells and clusters in such visuals, inherently affecting the generation of contours. Furthermore, using different numbers of neighbors for embedding generation can result in dramatic appearances of cell populations present in one condition but not the other (see circled numbers 1,4 in Fig. 5a and 5b in [48]), which can disappear when more or fewer neighbors are used, with those populations absorbed into overlapping contours. Likewise, densities of cell populations can appear of the same or different scale between conditions depending on the number of neighbors used in construction (see circled numbers 2,3,5,6 in Figs 5a and 5b in [48]), confounding the use of these visuals to make comparative statements.

Trajectory Inference and Continuous Relationships

Trajectory inference and pseudotime tasks, such as in RNA velocity [146] or Monocle [40, 250] workflows, focus on local, continuous relationships for inference and calculating pseudotime coordinates. Such tasks may also use distances between embedded points to construct the directions and magnitudes of arrows denoting inferred, developmental trajectories [146, 172] (Fig. 5.1).As shown with the standard velocyto workflow [146], using the neighbors of cells after reduction to two-dimensions to construct velocity arrows can result in erroneous trajectories, due to the arbitrary placement of cells under different parameter choices. Distortions can include loss of continuous relationships, trajectories in incorrect directions, or the addition of new pathways for development (see Fig. 6 in [48]). Distortions additionally occur due to upstream averaging over nearest neighbors in the inference procedure prior to the embedding procedure [96, 278]. Thus the resulting visual compounds distortions from embedding with these prior distortive effects.

In [48], we also demonstrate similar distortions of an underlying, continuous manifold by 2D reduction, using the Swiss-roll as a non-biological benchmark dataset for which we know the structure in three-dimensions, and moreover is a two-dimensional manifold. We demonstrate how the 3D Swiss-roll (constructed by rolling up the two-dimensional plane) loses its coherence when embedded in 2D with UMAP (see Fig. U in S1 Text in [48]). No embedding recapitulates the original plane [156] and depending on the number of neighbors used, distinct clusters or islands may appear, with a scrambling of local neighbors (made worse by increasing the tightness of the embedded roll). Thus knowledge of the true manifold is required to understand the disruption of continuity in these embeddings. Together, the use of such embeddings to imply or infer continuous relationships then becomes an arbitrary endeavor, with a user unable to trust seemingly dramatic connections or isolated populations, and likely to choose what seems most appealing or expected.

5.1.3 Embedding Properties are Arbitrary

To illustrate the indeterminate nature of two-dimensional UMAP and t-SNE embeddings, we developed an autoencoder framework to fit cells from any dataset to an arbitrary 2D shape, while preserving ambient cell-to-cell distances to an extent not much different than UMAP or t-SNE [37, 136, 242]. This essentially asks the question, what value or meaning do embeddings add, in comparison to a naive, arbitrary representation of the data? We found that it is possible to embed data in the shape

of a 'von Neumann elephant' [77, 175], in the spirit of this question of arbitrary representation, or a flower. Though it is unlikely scientists would present data in such forms, as shown below, they are quantitatively similar in terms of recapitulation of desired data properties in the ambient dimension, compared to UMAP or t-SNE embeddings. We call this method to produce customized embeddings 'Picasso', in homage to the eponymous artist's skill in imitating other artistic works.

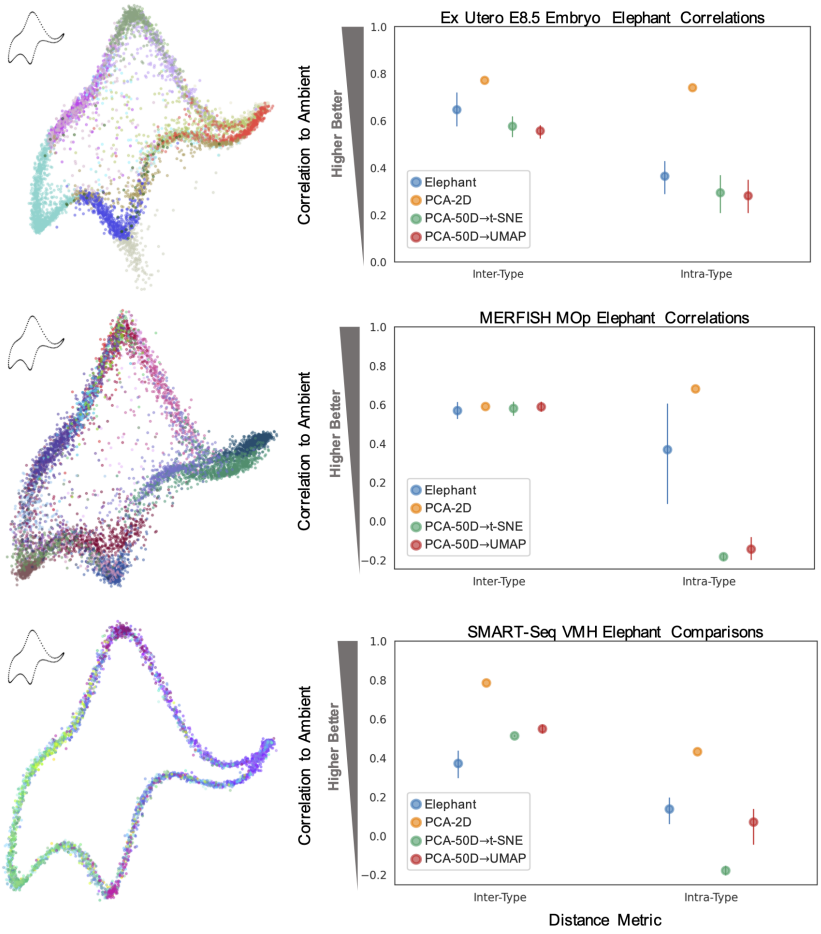


Figure 5.6: **Comparison of Embedding Properties.** Elephant shaped embeddings [77, 175] shown on the left, with corresponding correlations of data embeddings to ambient space shown in right-hand plots, for inter- and intra-type distance metrics. Metrics calculated over n=5 embeddings. Colors denote cell types, delineated in Fig W in S1 Text in [48]. Adapted from [48].

The autoencoder network used in Picasso is described in greater detail in [48], but the algorithm takes as input a centered/scaled count matrix $\mathbf{X} \in \mathbb{R}^{N \times G}$, N cells by G genes. The input is passed through two fully-connected layers of 128 nodes and D nodes, respectively, with D = 2 by default. We defined two loss functions: $L_{ShapeAware}$ and $L_{Reconstruction}$, which balance the fit of the input points to the desired

shape coordinates and reconstruction error in the decoder output as compared to the input. $\mathbf{C} \in \mathbb{R}^{P \times D}$ represents the coordinates comprising the desired shape, where D = 2 and P≥ N. The latent space $\mathbf{Z}$ is also limited to D = 2 dimensions. The pairwise distance matrix $\mathbf{D} \in \mathbb{R}^{N \times P}$ represents Euclidean distances between the cell coordinates in $\mathbf{Z}$ and shape coordinates $\mathbf{C}$ such that

$$d_{ij} = \|z_i - c_j\|_2.$$

Using $\mathbf{D}$, we define a Boolean, N × P adjacency matrix $\mathbf{A}$, where $\sum A_i = 1$. This matrix uniquely specifies an adjacent coordinate point for every cell, in a bipartite graph mapping the N cells to the P coordinates. $\mathbf{A}$ is determined by the linear_sum_assignment SciPy package, which assigns a shape coordinate to each cell by solving the minimization problem:

$$min \sum_i \sum_j d_{ij} a_{ij}$$

where $a_{ij} = 1$ iff row i is assigned to column j. Thus,

$$L_{ShapeAware} = \sum A \odot D.$$

Picasso performs this minimization to attempt to map cells to their closest, unique shape coordinates. The reconstruction loss is the L_2 norm of the difference between the reconstructed and input data:

$$L_{Reconstruction} = \|\hat{\mathbf{X}} - \mathbf{X}\|_2.$$

The total loss then incorporates both loss functions, balancing their contributions with f, a user-defined fraction weighting the effect of each term on the resulting embedding:

$$L = f * L_{ShapeAware} + (1 - f) * L_{Reconstruction}. \tag{5.5}$$

We compared correlations of inter- and intra-type distances between Picasso embeddings with those of t-SNE, UMAP and PCA, for the ex-utero (E8.5), MERFISH MOp , and SMART-Seq VMH neuron datasets [135]. These distances were constructed to represent trends often inferred from such visuals, where inter-type distances represent inter-cell-type relationships (or global relationships between clusters), and intra-type distances represent the variance or spread within the cell types. Each Picasso embedding demonstrated comparable performance to t-SNE and UMAP (Fig.

5.6), with cells of the same types distinctly grouped together in the arbitrary shapes. Picasso embeddings also improved upon t-SNE/UMAP intra-type correlations for all datasets (Fig. 5.6). Results were recapitulated for inter- and intra-distances calculated with the L_1 norm, and for trends between cells of different sexes (inter- and intra-sex distances) for the VMH neuron dataset (see Fig. W and X in S1 Text in [48]).

Thus, Picasso can quantitatively represent these visually inferred characteristics similarly to, or better than, the respective t-SNE/UMAP embeddings, while producing arbitrary shapes.

5.1.4 Limitations for Exploratory Data Analysis (EDA)

Although popular two-dimensional embeddings can reflect the broader strokes of the data such as cell type inter-distances, or highlight correlations between features [72], our findings highlight fundamental obstacles in reduction of high-dimensional data to 2D, the generation of multiple, possibly contradictory interpretations of the same data across applications, and the limited utility of these embeddings as EDA tools.

Though at the heart of EDA, as defined by statistician John W. Tukey [114, 251, 252], is the exploration of data through visualizations prior to confirmatory analysis, such visuals are intended to encompass robust or “resistant" analyses which extract (expected or unexpected) features of the data [251]. Thus the use of these 2D embeddings to reveal expected or unexpected properties is fraught by the fact that it is unclear which properties will be preserved or displayed, i.e., the purpose of the visual itself, where seemingly strong characteristics or patterns can be arbitrary distortions. Methods to show error or significance of cell placement on these visuals do not tackle the inherent limitations of such low dimension embedding: the lack of definition regarding which features are displayed and what is distortion to ignore [71, 194]. Prior analysis is required to determine ‘sufficient’ tuning of aesthetically oriented parameters and to define the purpose of the visual, undermining the use of such procedures as EDA tools. Together, this results in a user conducting two confounded exploratory analyses, that of the method properties and that of the data properties.

Another of the ‘guiding principles’ of EDA can be formulated as “analyses...before summaries" [251], where analyses are conducted to present particular features of the data, then collated as a summary. However, the use of such all-in-one visuals

begins from a place of summary rather than analysis, showing 'all points and all relationships' at once and attempting to approximate many properties. In general, the open-ended nature of these visuals and ability of parameter tuning to manipulate and create biological patterns demonstrate the ease with which such tools become confirmatory bias aids, and that such 2D spaces should be treated more as cartoon diagrams to be displayed post-analysis. However, in these cases conceptual graphics can be used instead which do not attempt to represent 'all points and all relationships' (to avoid over-interpretation), and higher-level diagrams which do not operate at the cell- or point-wise level [160, 263].

5.1.5 Assumptions and Incoherencies in the Dimensionality Reduction Process

The generation of the 2D embedding is additionally a multi-step process, demonstrated here as a preprocessing of the ambient data with a higher dimensional (linear) reduction by PCA , then a non-linear reduction to 2D by t-SNE/UMAP. Each step incurs some distortion of the data, where preservation of certain properties by one reduction can be lost by the next, as well as exaggeration of distorted patterns over the steps. However, this procedure is taken as a baseline [109, 140], and there is little discussion of the logic behind this coupling.

For example, though Euclidean (L_2) distance is the default metric for constructing neighborhood graphs in methods such as t-SNE and UMAP, this is not a requirement, and one might surmise that the non-linear methods instead learn other manifold-specific 'metrics' from cell neighborhoods by identifying 'biological geometries' (though this is not justified by the original authors [180, 255]). However, methods such as UMAP and t-SNE at their core rely on measuring distances locally, in concordance with common Euclidean analysis methods. This is the case for neighborhood graph construction as used for clustering [61], pseudotime and trajectory inference [105, 214], as well as non-linear embedding (e.g., UMAP/t-SNE) [180, 255]. Notably, the assumptions underlying the preprocessing of data by PCA may clash with the assumptions in extracting these other 'biological geometries' by non-linear dimensionality reduction, as PCA implicitly assumes Gaussian noise for data that lies in a Euclidean space. Embedding by PCA additionally reduces variance in the projected data, while methods such as UMAP add noise to embedded data (while removing biological signal) [94].

Utilizing these 2D visuals to infer structure of the underlying manifold then requires

knowledge of that manifold itself to interpret these outputs and distortions, a task confounded by noise present in biological data and the fact that common methods poorly recapitulate simple non-Euclidean manifolds (see Fig. U in S1 Text in [48]) [156]. And while PCA does impose assumptions of Euclidean geometry and Gaussian noise, the assumptions of heuristic, non-linear methods are more opaque and their results not easily falsifiable.

5.1.6 Alternative Methods and Analysis Approaches for Representation

From the findings in this study, we ultimately come to the conclusion that one should limit reliance on and blind application of such heuristic procedures, particularly across the range of applications in Fig. 5.1. Instead greater focus should be given to utilizing and developing an array of investigative and self-consistent analysis tools, which provide clearer interpretation of their goals and the biological features being assessed, present targeted low-dimensional embeddings and visuals displaying these features, and can easily be combined with statistical procedures to generate and falsify hypotheses.

With respect to the general task of preserving neighbor relationships (local or global) in an embedded space, it is possible to construct embedding spaces which more explicitly control and improve nearest-neighbor structure and retention (see Fig. Y and Z in S1 Text in [48]) [93, 265], as well as retention of desired metrics such as the intra-label metrics described above (see Fig. ZA in S1 Text in [48]). For example, if the goal of the visual representation is to display the data clusters, or other cell-wise annotations, we can replace the $L_{ShapeAware}$ loss in Picasso with $L_{LabelAware}$, where $L_{LabelAware}$ uses the Neighborhood Component Analysis (NCA) algorithm from [93]. This attempts to ensure that cells of the same label are represented together in the final embedding. For all cells a pairwise probability matrix $\mathbf{P} \in \mathbb{R}^{\mathrm{N}\times\mathrm{N}}$ is created where

$$p_{ij} = \frac{exp(-\|z_i - z_j\|^2)}{\sum_l exp(-\|z_i - z_l\|^2)}\ ,\ \sum_j p_{ij} = 1.$$

For discrete labeled data (e.g., cell type names) we defined $L_{Discrete}$ for all pairs of cells i, j where

$$L_{Discrete} = \sum_k \frac{\sum_{ij} p_{ij} \mathbb{1}_{ij}}{\sum_{ij} \mathbb{1}_{ij}} \text{ where } \mathbb{1}_{ij}(\mathbf{c}_k) := \begin{cases} 1 & \text{if } c_{k,i} = c_{k,j}\ , \\ 0 & \text{otherwise}\ . \end{cases}$$

Here C is the set containing label vectors for each class k, $C : \{\mathbf{c}_1, ..., \mathbf{c}_k\}$. Classes can be discrete or continuous, and multi-dimensional in the case of continuous classes (e.g., cell type, sex, condition, location).

Only the probabilities of cell pairs which are of the same label, for each class k, were summed and normalized to the total number of these cell pairs. For continuous classes of labels, such as spatial coordinates or pseudotime values, we used a separate loss function, $L_{Continuous}$. A probability weight matrix $\mathbf{W} \in \mathbb{R}^{\mathrm{N}\times\mathrm{N}}$ was generated for every pair of cells such that

$$w_{ij} = \frac{exp(-\|c_{k,i} - c_{k,j}\|^2)}{\sum_l exp(-\|c_{k,i} - c_{k,l}\|^2)}\ ,\ \sum_j w_{ij} = 1.$$

In place of the indicator function, the weights biased the masking of the original probability matrix $\mathbf{P}$ towards closer pairs of cells. Probabilities were also normalized to the maximum of the numerator (treating the weights $\mathbf{W}$ as constants):

$$L_{Continuous} = \sum_k \frac{\sum_{ij} w_{ij} p_{ij}}{\sum_i max(\mathbf{w}_i)}.$$

The $L_{LabelAware}$ is then defined as

$$L_{LabelAware} = L_{Discrete} + L_{Continuous}. \tag{5.6}$$

However, such optimizations require making an assumption regarding the appropriate distance/similarity metric, as is generally the case with the neighborhood-based analysis methods ubiquitous across the tasks in Fig. 5.1. Our analyses have focused on measuring distortions with respect to the L_1 metric, given its more desirable properties in higher dimensions than Euclidean (L_2) distance (see above), but other choices of distance or similarity metrics are possible and, whether in ambient or reduced space, can provide different interpretations of a dataset's properties [259]. To assess the suitability of different metrics across datasets, we used the 'relative contrast' ratio from [2] to measure the ability of an L_k norm to meaningfully delineate proximity between cells in high dimensions. We found that L_1 has higher contrast values than the L_2 norm across datasets (see Fig. 8 in [48]), suggesting preferential behavior in distinguishing cell relationships. How the various biological and technical features of each dataset drive or influence these contrast values is, however,

unexplored. There are other avenues for determining the relevance of a proximity metric, by assessing data properties such as 'hubness' (the presence of points with high proximity to many points in high dimensions) [85] and sparsity, discreteness, or continuity of the data structure [259], as well as the metric's biological interpretability in light of a given task or question. Thus, if such a metric is desired to represent relationships between cells, selection of the metric(s) should be carefully considered prior to downstream transformations and dimension reductions.

Across the applications in Fig. 5.1, there are existing methods and metrics, as well as opportunities for method development, which can provide more targeted alternatives in keeping with principles of EDA . For example, the assessment of multi-modal data integration and mixing can be directly calculated between cells, as shown by the metrics in this study, as well as by other metrics on mixing proportions and separation [73] or on the retention of true 'batch' differences (biological variation) [94, 253]. Such analyses can additionally be conducted in the ambient space, which minimizes the distortion/transformation of gene-related properties, useful for downstream experimentation.

For applications regarding clustering, clusters can be generated from higher dimensional embeddings if not from the ambient space itself [192], and given the central importance of marker gene expression in validating cluster assignment, existing tools such as heatmaps can directly display cluster results with the features (genes) which determined these groupings. Dimensionality reduction on the gene space can additionally be used to filter for genes or sets of genes best suited to separating the clusters [75, 182]. By targeting the objective of an embedding in such a manner, one can take advantage of prior knowledge/annotations and more directly determine the necessary dimensionality for a given question.

To assess heterogeneity within clusters or relationships between clusters, similarity metrics or distances can be calculated between the cells [259] and displayed with qualitative or quantitative visuals which preserve these metrics as a part of their objectives, including hierarchical relationship diagrams such as dendrograms and trees [120, 206], or graph-based network diagrams [79, 104]. Higher-level diagrams that do not seek to display all point-wise information can also be used to represent the results of other inter-cluster analyses [263, 276], better matching the resolution of the visual to the resolution of the analysis represented.

Such cluster-level visuals and metrics, as well as metrics on integration and higher dimensional distribution comparisons as presented here, can be used in lieu of anal-

yses based on contour plots generated from 2D coordinates. Regarding trajectories and continuous relationships, higher dimensions should be used to perform inference of differentiation trajectories [214, 263], and incorporation of probabilistic and biophysically-informed inference methods [74, 97, 160] offers falsifiable and interpretable approaches with targeted visualization alternatives. Such models additionally offer more interpretable handling of biological, as well as technical, noise [94], avoiding a smoothing over or removal of noise which could otherwise provide valuable biological signal.

Though it may seem appealing to produce visuals of 'all data and all relationships', common embedding practice distorts data in obscure ways, attempts to pack the capabilities of many different analyses into one space, and is easily manipulated. Given these limitations, and the distortions induced by earlier processing steps [5], it seems preferable to limit dimensionality reductions and ad hoc transformations, particularly when the space of interest can be treated directly, to utilize and develop targeted analyses for common questions that enable focused visuals, and collate these analyses to drive downstream, hypothesis-driven biological discovery.

5.2 Biological Visualizations for Quantitative Representation

The question of how to perform exploratory analysis through interpretable yet quantitative visual representations, extends beyond the realm of biology, into fields like the social sciences [3, 193]. As described above, there may already exist methods that better address a question from a quantitative perspective, especially in comparison to an all-in-one approach. Thus in the spirit of interpretable representation learning, and in contrast to the common induction of ML methods into the biology pipeline, we demonstrate an application of *biological* methods for low dimension visualization, specifically phylogenetic representation, and how its extension to other fields, such as political science, provide a quantitative representation of social structures that is easy to interpret and admits simple statistical tools for analysis.

This section summarizes the contents of [47] by T.C and L.P. L.P. conceived of the application of NNet to political structures. T.C. developed the analysis and code. T.C. and L.P. wrote and edited the manuscript.

5.2.1 Phylogenetic Methods for Hierarchical Representations

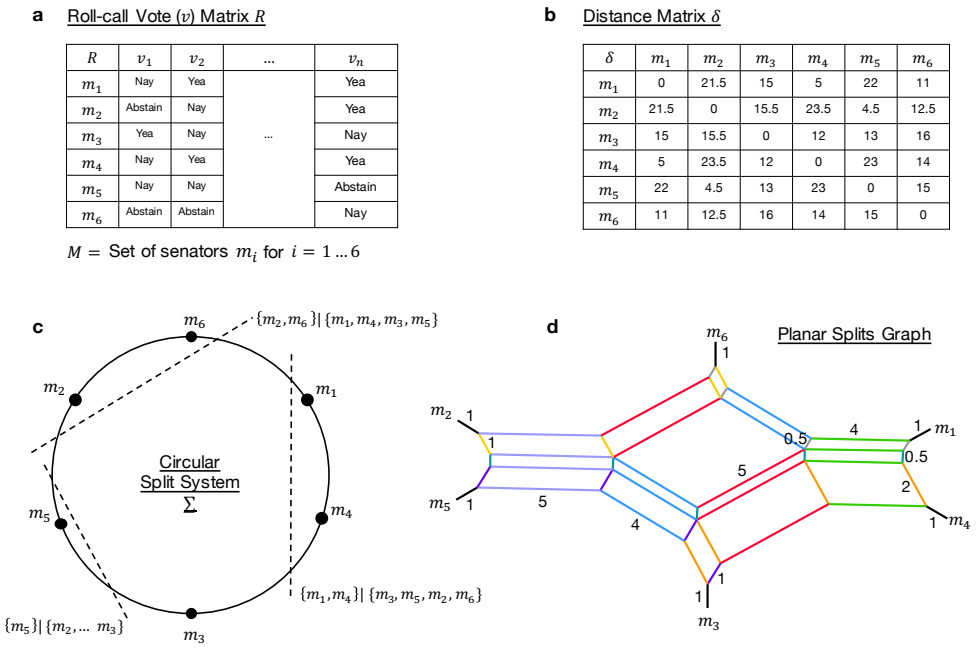
The field of phylogenetics seeks to understand, classify and quantify the evolutionary and connective relationships between biological entities such as groups of organisms. The 'phylogenetic tree', introduced in a drawing by Charles Darwin [59], is a convenient mathematical abstraction for representing such hierarchical evolutionary relationships. The observed traits of a species were used to construct groupings of organisms and outline possible developmental relationships and common ancestry. Thus from the trait-based features of the biological entities of interest, a structured model and visualization of the underlying relationships can be generated.

Over time, many methods have been developed to measure similarities and differences between phylogenetically-informative 'traits', such as DNA (or RNA) sequence information/characters, to characterize and quantify similarities between species or taxa [261]. Such approaches then enabled the use of phylogenetic tree methods based on the properties of distance matrices constructed between the taxa [209], in which pairwise distances are used to infer the topologies of phylogenetic trees. The most popular technique for distance matrix-based tree construction is neighbor-joining (NJ) , which was developed in 1987 by Naruya Saitou and Masatoshi Nei [215], and inspired the development of the Neighbor-Net (NNet) algorithm [32] that forms the basis for this work (see below). The algorithm uses a greedy, agglomerative clustering strategy to iteratively join pairs of nodes or taxa based on their pairwise distances. Essentially, nodes are joined which minimize the total length of the resulting tree. In terms of statistical consistency, NJ will yield the correct tree given sufficient amount of data from which to estimate accurate pairwise distances, and it has been shown that NJ will often be successful in recapitulating the underlying tree [184].

Given a pertinent set of informative features, a pairwise distance matrix can easily be constructed to fit user needs. Thus, methods for inferring phylogenetic trees, including NJ, have been adapted and extended over the years to a variety of fields. For example, in linguistics, NJ has been applied to distance matrices inferred from phonetic similarities, and it has been used to visualize and explore relationships between languages [62]. NJ has also been applied to cosmology, where distances constructed based on chemical composition delineate possible pathways for the origins of stellar populations [128].

5.2.2 The Neighbor-Net (NNet) Method

As previously described, phylogenetic trees can be represented using branching diagrams where taxa represent nodes in the tree connected by branches, and these can elucidate the hierarchical, evolutionary relationships between taxa. In some cases, however, relationships between taxa do not conform to tree structures, e.g., as a result of horizontal gene transfer or non-sexual movement of genetic information. Phylogenetic networks that generalize trees can be used to represent such non-conforming relationships, and are generally useful for highlighting conflicting signals in data.



R	v_1	v_2	...	v_n
m_1	Nay	Yea		Yea
m_2	Abstain	Nay		Yea
m_3	Yea	Nay	...	Nay
m_4	Nay	Yea		Yea
m_5	Nay	Nay		Abstain
m_6	Abstain	Abstain		Nay

δ	m_1	m_2	m_3	m_4	m_5	m_6
m_1	0	21.5	15	5	22	11
m_2	21.5	0	15.5	23.5	4.5	12.5
m_3	15	15.5	0	12	13	16
m_4	5	23.5	12	0	23	14
m_5	22	4.5	13	23	0	15
m_6	11	12.5	16	14	15	0

Figure 5.7: **Constructing Split Networks from Votes. a)** Matrix R of feature (vote) values for six elements (senators) m. **b)** L_1 pairwise distance matrix between all six elements. **c)** Circular split system with elements m shown in circular ordering, and splits defined as bi-sections of the circle. **d)** Splits graph representation of circular system in **c** with split weights denoted along splits. Parallel splits (representing the same split) denoted by the same colors. Figure adapted from [32, 155] and [47].

Here we focus on a specific type of phylogenetic network called a 'split-network', and on methods which use pairwise distance (dissimilarity) matrices between the informative features of the taxa to construct such networks. Split-networks in particular, seek to visualize these structures based on the splits in datasets, i.e., partitions of taxa into two distinct parts. From a collection of splits, also known as a 'split-systems', a graph is constructed that 'simultaneously represent[s] both

groupings in the data and evolutionary distances between taxa' [32].

NNet is a dissimilarity or distance-matrix based split network construction algorithm which determines a collection of weighted splits that can be realized as a split-network. In the linguistic setting it has been used with distances between languages estimated from phonemes [33, 76], and has been found to be useful because conflicting signals may otherwise be difficult to discern, e.g., where distinct characters may have been shared at different times, sometimes between spatially and temporally distant languages. Recently, NNet has also been used to analyze structure in single-cell gene expression data [276]. Though NNet has not previously been used for analysis of data from the political sciences, it is similar to commonly used MDS techniques in that its inputs are dissimilarity based measurements [30, 32, 212, 224]. However, in contrast to MDS-based embedding methods which focus on recovering ideal points for individuals in a low dimensional space [198], NNet additionally defines groupings ('coalitions') between individuals and represents the relative strengths of these relationships within its network construction.

We denote M to be a set of n elements $\{m_1, ...m_n\}$ representing members of the senate, and $\mathbf{R}$ to be an element $\times$ feature matrix (features represented as votes here) (Fig. 5.7a). A pairwise dissimilarity (distance) matrix δ is constructed (Fig. 5.7b) such that its entries form a function mapping $M \times M \rightarrow \mathbb{R}$ that satisfies

$$\delta_{i,j} = \delta_{j,i} \text{ and } \delta_{i,i} = 0.$$

We use the L_1 distance between the vote types of each member (0 Nay, 1 Yea, 0.5 Abstain)

A split $A|B$ is a bi-partition of the set of elements in M, where

$$A \cup B = M, A, B \neq \emptyset, \text{ and } A \cap B = \emptyset,$$

and a split-system is a collection of splits. In the example of a phylogenetic tree, each "branch divides the set of taxa up into a split, with the taxa on one side of the branch separated from the taxa on the other side" [32]. 'Compatible' splits denote splits which can be contained within a phylogenetic tree, however NNet can produce both compatible and incompatible splits, satisfying a weaker condition than compatibility [32, 155].

Given the defined distance matrix δ, NNet will generate a circular ordering of the elements $\pi = \{m_1, ...m_n\}$, where m_i and m_{i+1} are adjacent vertices on an n-cycle

C_n comprised of the elements of M, a split-system Σ (Fig. 5.7c), and the weights of the splits, λ, (represented by branch lengths in the planar graph) (Fig. 5.7d). Σ is a circular collection of splits, a generalization of compatible splits, where there is an ordering of the elements $\{m_1, ...m_n\}$ such that every split is of the form

$$\{m_i, m_{i+1}, ..., m_j\}|M - \{m_i, m_{i+1}, ..., m_j\}$$

for some i and j satisfying $1 \leq i \leq j \leq n$ [32]. Such splits always have a planar splits graph representation [119] (Fig. 5.7d). Further details of the algorithm can be found in [47]. Here we provide δ as the distance matrix input to the NNet implementation in SplitsTree4 [120], to calculate the ordering π and splits Σ for a graphical, non-hierarchical visualization, and to extract the split weights λ for the system.

The collection of weighted splits produced by NNet can be realized in 2D as a splits graph [32, 119]. A compatible collection of splits will be exactly represented as a tree, while incompatible splits are denoted as cycles/boxes in the diagram (Fig. 5.7c). These incompatible splits are thus represented as a collection of parallel edges, each with the same weight. Distance between members (taxa, etc.) is defined as the sum of the lengths of the paths/branches connecting them. Given that parallel edges have the same weight, the distance between two members thus becomes the sum of the split weights for the splits separating those members.

5.2.3 Circular Split Systems of the US Senate

For the current 116th Senate, the split network output of NNet , using the distance matrix δ generated from senate votes, is shown in Fig. 5.8a. In [47], we additionally generated the split network from Democrat (including Independent) and Republican senators separately. Note that the split network produced by running NNet run on a subset of a matrix will be the same as the restriction of the split network produced by running NNet on the full matrix.

Coalition Structures from NNet

We first verified that the generated split weights λ represented the same magnitudes of dissimilarity between pairs of senators as encapsulated in the input matrix δ. By Pearson correlation analysis of the pairwise distances calculated from λ, we found a correlation of 0.994. Having inferred the circular split-system representation and split weights for the 116th Senate, we next examined individual relationships and neighbors across all members (Fig. 5.8a). As expected, there is a split dividing

members of the two major parties. The split network also reveals member's nearest neighbors based on their voting behaviors, and noted 'mavericks' or 'centrists', such as Sen. Collins (Rep.) and Sen. Manchin (Dem.), stand out in their distant, centered placement relative to the rest of the senate [218] (Fig. 5.8a).

Beyond pairs of individuals, the senate-wide diagram highlights apparent coalitions within the greater senate structure, visible by clustering of individuals in the circular order, and in larger relative magnitudes of split weights (lengths) separating groups of individuals from the rest of the system (Fig. 5.8a, denoted 1 and 2). An interesting and notable example is the split of Democratic Primary candidates from the rest of the senate (Fig. 5.8a, denoted 1). Of the seven main incumbent senators to run in the 2020 Democratic Party presidential primary [35], five consistently cluster together in the senate-wide circular split-systems (Fig. 5.8a).

To verify whether this sequential ordering of these candidates was significant we used the Wald-Wolfowitz runs test [256] to determine the likelihood that this particular ordering was random (the null hypothesis). For this test, the circular ordering of senators can be represented as a linear ordering with senators that were Democratic Primary candidates represented as 0s and the other senators as 1s. To test for significant difference from the null hypothesis of a random ordering we found the probability of observing fewer than seven runs (at least five candidates clustered together) occurring in any ordering of the binarized senator representations. A run denotes a contiguous stretc.h of the ordering with senators from the same category (0 or 1). In both the Senate-wide and Democrat-only circular orderings, p-values were <0.001, revealing a statistically significant departure from randomness in the non-random ordering of these senate members.

The formulation of the split-network also facilitates mapping of the splits of interest back to the features (votes) that underlie that split. In this way, we can extract votes which contribute to particular splits of interest, i.e., we can apply the split to the original voting input R, and selected for features (votes) which characterize that split (where the votes of that individual or group of individuals differ from the other members). Thus, given the split of five Democratic Primary candidates, we traced back the split to the votes contributing to their unique voting pattern by first extracting votes where all candidates voted the same. Of these votes, we found a particular set in which a majority of the rest of the party did not vote in accordance with these senators (Fig. 5.8b), temporally clustered in the latter half of 2019 (Fig. 5.8b). These votes with the largest discrepancy were all abstentions by

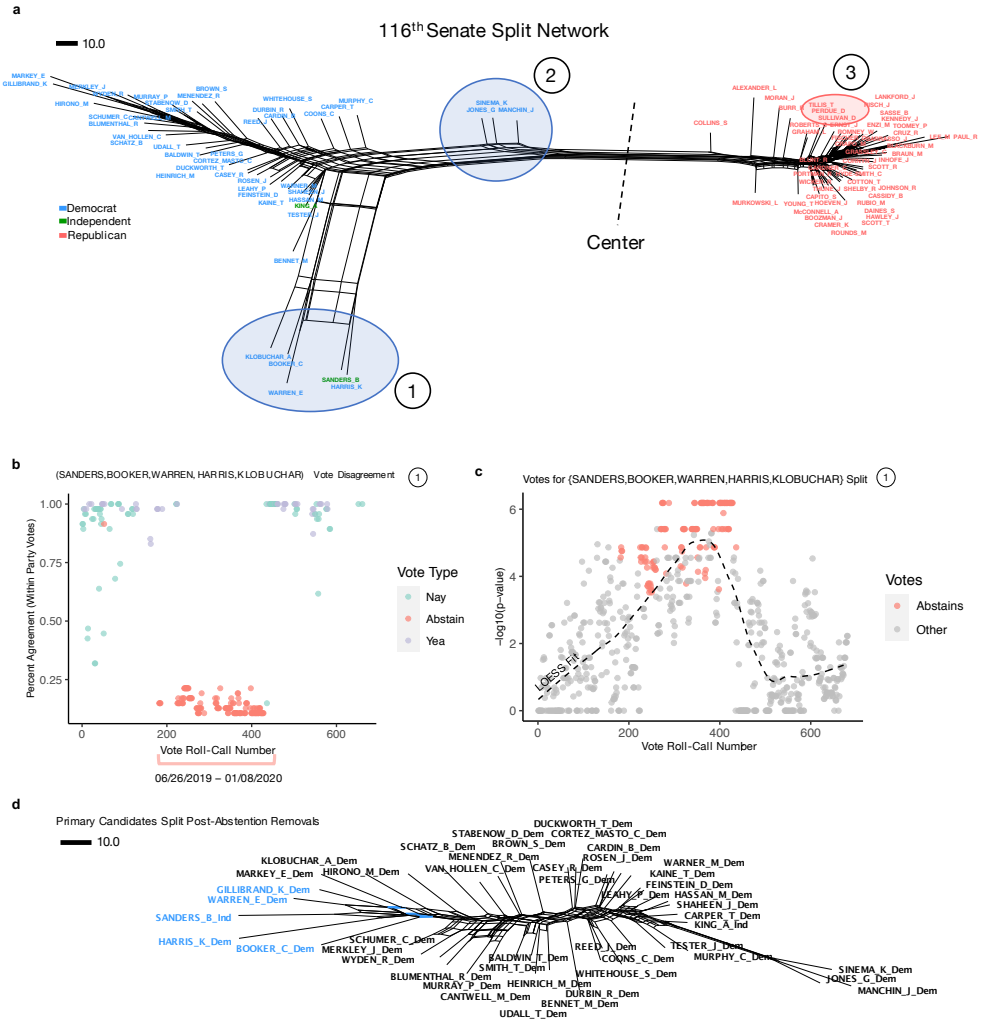


Figure 5.8: **116th Senate Split Network. a)** Splits graph representation of Senate-wide split network, for the 116th Congress. Republican, Democrat, and Independent members shown with colors. Nearest-neighbors and apparent 'coalitions' of members shown in circles. **b)** For roll-call votes where these five Primary candidates voted the same, percent disagreement within the party is shown (fraction of remaining members of the party who voted differently). Votes colored by the vote cast by the candidates. **c)** All roll-call votes ranked by p-value for the given split of the five Primary candidates. Abstentions with low agreement from **a** colored in ranking. Raw p-values reported here. LOESS (Local Regression) fit for p-value rankings over time (roll-call votes) shown by dashed line **d)** Splits graph for splits network of only Democrat (inclusive of Independent) votes after two iterative removals of low-agreement abstentions for clustered Primary candidates. Adapted from [47].

these senators, behavior which aligns with the previously noted trend of presidential candidates abstaining during campaign periods [28] (Fig. 5.8b).

For these (or any) splits of interest we can assign a statistical interpretation to how the votes contribute to the splits of interest by ranking them by p-value. To do this, we ranked votes by their likelihood of being associated with any given split of interest, defining p-values for a vote using the Fisher's exact test [84] for a 2×3 contingency tables between a split $\{A|B\}$ and counts of the vote types $\{0, 1/2, 1\}$. The table, shown below, denotes the counts for the intersections between members in $\{A, B\}$ and $\{C, D, E\}$, where C, D, and E are the sets of members whose votes were 0, $1/2$, and 1, respectively.

	C (0)	D (1/2)	E (1)
A	$\lvert A \cap C \rvert$	$\lvert A \cap D \rvert$	$\lvert A \cap E \rvert$
B	$\lvert B \cap C \rvert$	$\lvert B \cap D \rvert$	$\lvert B \cap E \rvert$

We used a two-sided Fisher's exact test to determine p-values for assessing how likely a more 'extreme' contingency table for (how far from random) a particular vote's table would be. We were thereby ranking, for a given split of senate members, the likelihood of the voting behaviors in each vote being associated with that split. For ranking purposes, we report the raw p-values of each vote.

For this particular split of the five Primary candidates we see that the ranking results (Fig. 5.8c) are concordant with the votes of low intra-party agreement (Fig. 5.8b). The clustered abstentions have the highest likelihoods of contributing to splits, among other Yea or Nay votes also contributing to this split. The p-value assignments also allow for investigation of distinct behaviors in the votes contributing to a split of interest. With the ranked votes we fit a LOESS (Local Regression) curve to the p-values (Fig. 5.8c, dashed line). This demonstrates the apparent temporal progression the contributing votes follow, with an upward trend in p-values leading to the abstention period, and a decrease in rankings following that time period (Fig. 5.8c). We then removed these clustered abstentions to discern who these five senate members vote similarly to outside of this abstention time period. After a second removal of low agreement abstentions, for the split of Senators Booker, Sanders, Warren and Harris, who remained clustered despite the initial removal, we see that this group remains split from the rest of the party by voting behavior, with Sen. Gillibrand (Fig. 5.8d).

Temporal Variation in Party-Specific Voting Agreement

From the split network in Fig. 5.8, we note a variety of structures within the enate and the individual parties, with particularly dense areas and sparse regions of individuals denoting areas of high or low voting agreement. To assess and visualize this agreement across senate members we denoted the 'center' (Fig. 5.8a) split to make relative quantifications of the spread of member's voting behaviors. This also provides a comparative metric for how 'left' or 'right' of center members are [124]. This assignment of distances from the center is not limited to the 116th Congress, and thus we explored the dynamics of this metric over time for all senates over the last 30 years (Fig. 5.9a).

By aggregating distances for each of the main parties, we visualized if or how the spread and magnitude of voting agreement within and between parties has changed over time as a product of their constituents. What we observed fits with previously reported trends of increasing partisanship in the Senate [18, 154, 185], at least within the last six years. This is demonstrated by upward shifts in the median party distances, i.e., increasing distances of each party's members from the center. The larger spread of center distances observed in the Democratic party in recent senates versus a tightening of the Republican distances also suggests differing levels of voting unification within each party [176]. This is also in contrast to earlier senates, where greater 'unification' (tighter distance distributions) in the Democratic party is demonstrated (Fig. 5.9a 101st, 102nd). Shifts in party-specific voting unification were further investigated by examining the distribution of center distances ranges (the difference between the highest and lowest distance) for each senate session with respect to the party in the senate majority (Fig. 5.9b). This revealed a significant difference in the range or spread of voting behaviors within each of the two main parties (Independents included with Democrat senators) when the opposing party was in the majority versus minority (Fig. 5.9b). Though there are many factors which can contribute to greater or lesser party unity [229], this suggests a relationship between voting behavior and the party's standing in the senate, possibly related to recent observations on the ability of the majority party to influence the legislative agenda of the chamber floor particularly when the party is ideologically cohesive [45].

We additionally investigated these agreement distributions at the level of their constituent members, as visualized for the 116th Senate (Fig. 5.9c). At the level of individual senators we can note the differences in magnitude of the center distances

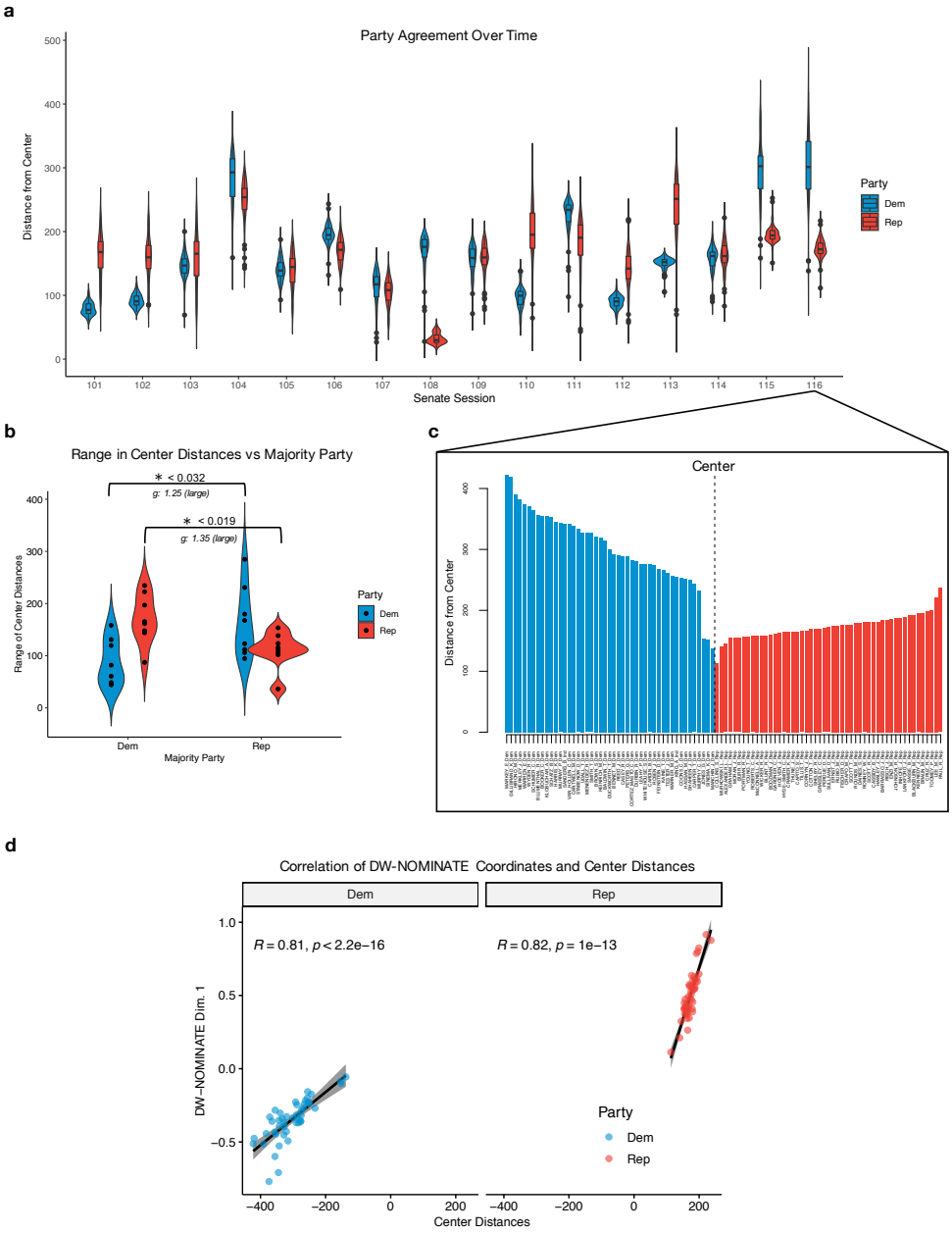


Figure 5.9: **Quantification of Voter Spread or Agreement.** **a)** Distribution of members distances from center shown for the 101st to 116th US Senates, within each of the main parties. **b)** Distribution of center distance ranges, calculated for each party per Senate session. Mann-Whitney U-test used to determine if party-specific ranges differ in sessions with either party in the majority. $n = 16$. * denotes p-value <0.05. g denotes Hedges' g measure of effect size. **c)** Center distances for the 116th Senate shown for each senator. **d)** Spearman correlation between DW-NOMINATE coordinates (first dimension) and center distances, by party, for senators in the 116th Senate. Center distances on opposing sides of the center split are negated. Adapted from [47].

among non-Republican senators versus Republican senators and place each senator within this greater distribution. These individual distances were then compared to the coordinates of the ideal points assigned to each senator by DW-NOMINATE, demonstrating a high correlation of ~ 0.8 to this benchmark methodology within each party (Fig. 5.9d) [44]. Here we utilized the first dimension of the DW-NOMINATE coordinates, as it tends to be the most interpretable and commonly utilized part of the embedding space [80]. This highlights the ability of NNet to not only replicate the spectrum of 'left' and 'right' within the senate, as the DW-NOMINATE coordinates reveal [44, 80], but also provide the structure of coalitions within which these preferences reside.

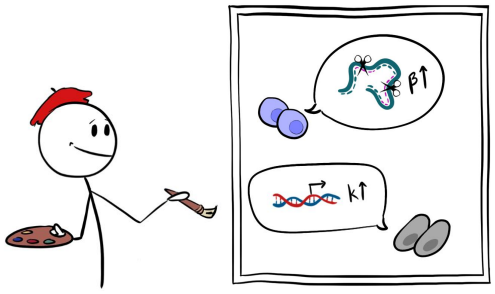
Our findings highlight the utility of the NNet-SplitsTree algorithms in creating representations and visualizations of voting data that facilitate exploratory analysis and facilitate identification of voting patterns that may not be readily apparent. This non-model-based approach minimizes assumptions on the structure of the input data, though the circular nature of the split-system can limit which relationships are accurately recapitulated in the visualization. However, as mentioned previously, these discrepancies can be utilized to detect members of the network who display more discordant or 'maverick' behavior. The analysis framework we have proposed is additionally limited to political relationships and structures visible at the level of voting behavior, and it is important to keep in mind that there may be other factors and behaviors which may influence the relationships between political members.

With the NNet-based approach, we determined relationships between pairs of senators within and across their respective parties, highlighting the impact of inter- versus intra-party voting behaviors on the stability of those relationships. The relative lengths of the generated splits additionally provided a quantitative visualization of both the strengths of these relationships and the level of divergence those shared behaviors represented. This gave rise to visible coalitions of senators within the greater split system, notably five of the 2020 Democratic Party presidential candidates and a separate group of Democrat 'centrists'. Utilizing the direct relationship of the defined splits to the input voting data, we recovered the contributing votes to each coalition and verified the existence of shared voting behaviors unique to the primary candidate coalition beyond the abstentions common during presidential bids. The split-system also provided an interpretable framework for the development of a statistical procedure to rank contributions of each vote to any given split of interest, thus connecting shared behaviors to their comprising features in a statistically

rigorous manner.

Chapter 6

BIOPHYSICAL REPRESENTATION OF PERTURBATION DATA



You cannot answer a question that you cannot ask, and you cannot ask a question that you have no words for.

JUDEA PEARL

The work in Chapter 5, highlights an important question in the analysis of high-throughput genomics data (or any high-dimensional data for that matter). What should our representations represent?

Rather than creating an all-in-one representation attempting to illuminate many relationships, it may be more promising, and interpretable, to create specific representations for specific questions, along the lines of the EDA discussion in Chapter 5.1.4. In particular, the questions resulting from perturbation data are often concerned with understanding how processes of DNA and RNA regulation lead to observed responses in the cell, whether that be changes in mRNA or protein expression, or morphological changes and the spatial distribution of cells. As described in Chapter 4.4, such data are becoming increasing multimodal, where within cells we can measure multiple biological entities simultaneously, from nascent and mature mRNA to chromatin 'openness' and protein expression. The idea of such simultaneous measurements, coupled to perturbation, is to illuminate not only different components of gene regulation, but how they work together to produce behaviors of interest, be it cell fate determination or development of tumorigenesis. However, treatment of multimodal data, with or without perturbations, is often limited to observational analysis only, i.e., at the level of expression and statistics on the counts. Thus we use changes in 'expression' as a proxy for effects on the transcriptional

process, though several different mechanistic explanations could give rise to such changes. Thus, if the goal of these increasingly high-resolution views into the cell is to probe intertwined processes of gene regulation, we may require tools with more specific grammar with which to describe these processes, and in turn, ask deeper questions.

6.1 Biophysical Inference of Multimodal Cell Types

We begin with the question of clustering, or determining which populations of cells exhibit similar transcriptional states, in a heterogeneous biological system. The clustering described and assessed in Chapter 3 and 5 only utilized one count matrix, but given multiple measurement matrices, what does it mean to cluster cells and how does one determine the appropriate balance of measurements in the clustering task? Below, we re-interpret the task of clustering in scRNA-seq analysis through the lens of stochastic, biophysical models of transcription, to provide instead, a kinetic representation of the data.

This section summarizes the contents of [50] by T.C, G.G., and L.P. The method was conceptualized by T.C. and G.G., T.C. developed the code and analysis. T.C., G.G., and L.P. wrote and edited the manuscript.

Determination of cell types is a central task in single-cell genomics analysis [111, 138], though the definition of 'type' and whether designations should be of a discrete or continuous nature is a matter of debate [70, 274], and is often dependent on the investigation or data properties [81, 159]. Here we focus on discrete categorizations, or clusters, of cells where common clustering methods include the Louvain [61] or Leiden [247] community detection algorithms (neighborhood graph-based algorithms), hierarchical clustering techniques [268], (finite-dimensional) mixture model approaches [38, 52], and marker gene-based analyses [7]. While these techniques are widely used, they are subject to heuristic tuning of hyperparameters [138], and/or assume Gaussian-distributed data in contrast with the sparse and discrete nature of single-cell data. Clustering results are also often derived after initial dimension reduction(s), and assessed after further dimension reduction and embedding to 2D [24, 180].

As we move towards more simultaneous measurements and modality types, delineating clusters with these methods becomes convoluted, particularly regarding the treatment of modality-specific features and variability. And though these multi-

modal experimental approaches provide opportunities for large-scale, mechanistic studies of the central dogma, i.e., to model the kinetics of these processes [101] and the roles of biological stochasticity in driving cellular heterogeneity [51, 68], such mechanistic studies are often limited to exploring cellular processes for only a handful of genes and/or for homogeneous systems [188, 200, 240]. Thus there is an opportunity and need for using multimodal, single-cell data to gain biophysical and mechanistic insight into heterogeneous cell types [51, 149, 266].

Specifically for scRNAseq data, standard clustering for both benchmarking and exploratory datasets [107, 238, 245, 270] is performed on a gene count matrix which is often constructed from spliced gene expression, or non-intron aligning reads that represent mature mRNA [113]. However, two modalities, Unspliced (**U**, intron-aligning) and Spliced (**S**) molecule counts, can be obtained from most scRNAseq datasets and allow us to tease apart the production and processing steps in the generation of mRNA [46]. For example, these counts are obtained by re-alignment of datasets to an intron-containing reference [237]. Uniquely, **U** and **S** are summed together by default to generate the count matrix in the 10x Genomics Cell Ranger 7.0.0 pipeline [113, 237], however this conflates the two measurements as the same biological entity, reduces interpretability, and is not the default in most other count generation pipelines [237].

Given these matrices of different measurement types, this immediately raises the question of which matrices, and what balance of matrices, are relevant for clustering? The choice of matrix has many implications for clustering methods, starting with the set of genes to be used. Perhaps unsurprisingly, different matrix choices result in distinct cluster assignments for the cells (see Fig. 1 'Assess Clusters' in [50]), necessitating either the arbitrary selection of a matrix or determination of consensus clusters across modalities, defined through some metric or heuristic.

Existing methods for determining such consensus or shared clusters largely ignore **U** counts, focusing on integrated clustering across **S** mRNA counts, protein, and chromatin accessibility modalities [91, 108, 161]. Methods that do integrate **U** counts build on standard RNA velocity pipelines [103], also described in Chapter 5.1.2, which rely on large numbers of arbitrary hyperparameters and ad hoc processing steps [96], and are often incompatible with known biophysics [96]. Outside of these approaches, multimodal clustering methods often utilize heuristics to balance the influence of modality-specific neighborhood-graphs or similarity matrices [108], which do not necessarily provide a consistent foundation for extension to new mea-

surements. Alternatively, deep learning and/or embedding approaches seek to find a common space which produces the partitions of the data into clusters [161], or which can then be clustered by any clustering method the user chooses (i.e., an arbitrary selection) [91]. Though here it is common for such methods to model the count data using discrete distributions, these methods model the modalities through independent observational distributions which obscures understanding of their innate causal relationships induced by underlying, transcriptional processes. Furthermore, it remains unclear how to justify or interpret the balance of modality-specific properties in such latent spaces [94, 253].

In light of these limitations, we propose meK-Means (mechanistic K-Means) as a method to cluster cells from multimodal single-cell data under a self-consistent, biophysical model. We demonstrate meK-Means on two modalities which describe mRNA production and processing: **U** and **S** gene count data, for which we utilize the Chemical Master Equation (CME) to formalize causal transcriptional processes in the cell and their governing rates [26, 226], as well as the technical sequencing process [95]. The CME is a natural framework to model the joint distribution of **U** and **S** counts at steady-state [26]. meK-Means presents a mixture model representation of this joint distribution, to learn clusters **Z** underlying the observed gene count data. A cluster is then inherently defined by the governing parameters of the cellular processes of interest, and represents shared transcriptional programs between genes, allowing for greater representation of gene-gene correlations as opposed to the standard, independent treatment of genes under the CME approach [100].

Thus meK-Means provides interpretable integration of modalities to learn cell clusters, and a basis for consistent extension to new measurements, as it inherently balances and unifies the modalities of interest through their underlying, biophysical relationships.

6.1.1 meK-Means Implementation and Interpretation

From simply aligning scRNAseq data to a reference transcriptome with intron and exon annotations, we can obtain two molecular measurements or 'modalities', Unspliced (**U**) and Spliced (**S**) count matrices which represent 'nascent' and 'mature' mRNA molecules, respectively [113, 237]. Each matrix is a cell x gene count matrix, and both are taken as input for clustering with meK-Means (Fig. 6.1a). meK-Means then models the joint distribution of these modalities using a biophysical model of

their transcriptional relationships, and infers which cells cluster together based on similar transcriptional kinetics (Fig. 6.1a). The output of meK-Means is effectively a cluster x gene x parameters matrix, where we learn a set of cluster-specific, biophysical parameters for each gene which describe the processes of transcription, splicing, and degradation (Fig. 6.1a). This model additionally includes the sampling of the molecules by the technical, sequencing process (resulting in the final observed count matrix). Thus this approach explicitly models biological and technical sources of variation in the data.

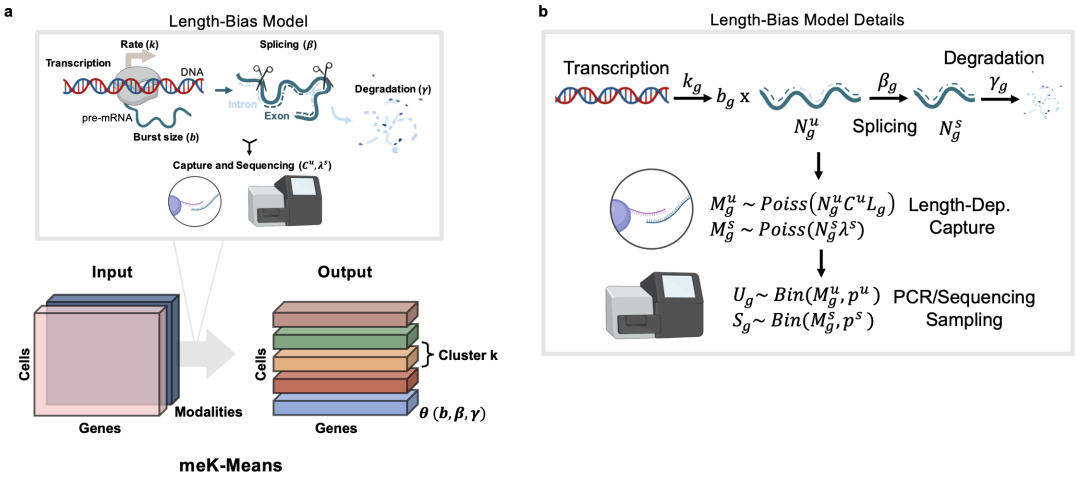


Figure 6.1: **meK-Means Clustering. a)** High-level diagram of Input and Output of meK-Means (from multimodal data to a matrix of cluster x gene x parameters). meK-Means fits data to a Length-Bias Model of transcription. **b)** Detailed description of the Length-Bias CME Model. Rates per gene g denoted. Model includes length-dependent technical sampling of the molecules produced by the transcription processes, which occurs during the sequencing process. This produces the final observed counts (i.e., U and S). Adapted from [50]. Created with BioRender.com.

The CME Model of Sequencing Data

The model of transcription underlying meK-Means uses the CME formalism described in detail here. The CME describes the probability of molecule counts X over time ($p(X,t)$), given some rates of transitions between states. States here are the possible molecular counts observed. It is particularly useful for modeling low-count (discrete) systems [26, 226] such as the sparse, low-count regime of scRNAseq data, and has been used extensively in the fluorescence transcriptomics literature over the past couple of decades to model the processes of transcription and protein synthesis in various cell systems [26, 187, 188, 226].

In our case, we define a CME that describes the probability of unspliced, U_g, and spliced, S_g, counts per gene g over time ($p(U_g, S_g, t)$). Here our rates of transition are the (gene-specific) kinetic rates describing the processes of transcription k_g (which produces unspliced molecules in bursts of size b_g [39]), splicing of unspliced molecules, β_G, and degradation of spliced molecules, γ_g (Fig. 6.1a,b). This model also includes the sampling of these unspliced and spliced molecules that happens external to the cellular transcription, during the sequencing process. Here we use the length-bias CME model (developed for 10x Genomics or poly(A)-capture scRNAseq methods) described in [95] where nascent pre-mRNA or unspliced molecules are captured in a length-dependent manner (as longer transcripts can contain more internal poly(A)s for internal priming/capture) (Fig. 6.1b). We will use parameters C^u, λ^s to denote the rate of capture of unspliced and spliced molecules (see below for more details).

Given that we are working with snapshot scRNAseq data here, we then study the behavior of $p(U_g, S_g, t)$, in the long-time limit (steady state), $p(U_g, S_g)$ as $t \to \infty$. At steady-state, certain gene parameters are not independently identifiable, thus we define relative splicing and degradation parameters, β_g/k_g and γ_g/k_g, where splicing and degradation rates, respectively, are relative to the transcription rate k_g [95]. b_g represents the mean of geometrically-distributed bursts of transcription [88]. The technical parameters, which are shared across genes, are shown in the $Poiss$ and Bin capture and sequencing sampling parameters in Fig. 6.1b, and are also not independently identifiable, thus we define net sampling rates λ^u, λ^s (where $\lambda^u = C^u L_g$, L_g represents length of the gene) which contain p^u, p^s. For simulation and meK-Means inference, these global technical parameters are set prior to inference of the physical parameters (i.e., we do not perform a grid search over these parameters during meK-Means inference).

We note that the bursty model of transcription is a limiting case of the two-state telegraph model in Fig. 6.2. Here the gene switches between an ON and OFF state, and while in the ON state produces transcripts at some rate. As we take $k_{ini}, k_{off} \to \infty$, the burst size b is defined as k_{ini}/k_{off}, and $k_{on} \to k$ as in Fig. 6.1.

Solving the CME Model

With this CME formulation we can define the steady-state probability generating function (PGF) form, H, of $p(U_g, S_g)$, and solve for $p(U_g, S_g; \boldsymbol{\theta}_g)$ where $\boldsymbol{\theta}_g = \left[b_g, \beta_g/k_g, \gamma_g/k_g, \lambda^u, \lambda^s\right]$ (the kinetic parameters of the model) as described in

$$\begin{aligned}
\text{OFF} &\xrightarrow{k_{on}} \text{ON} \\
\text{ON} &\xrightarrow{k_{off}} \text{OFF} \\
\text{ON} &\xrightarrow{k_{ini}} \text{ON} + N^u \\
N^u &\xrightarrow{\beta} N^s \\
N^s &\xrightarrow{\gamma} \emptyset
\end{aligned}$$

Figure 6.2: **Two-State Telegraph Model.** Definition of the two-state telegraph model of transcription, used to derive the bursty model of transcription.

[26, 95]:

$$p(U_g = \mathrm{u}, S_g = \mathrm{s}; \boldsymbol{\theta}_g) \approx \mathrm{iFFT}\, H\left(e^{\frac{-2\pi i \mathrm{u}}{W_g}}, e^{\frac{-2\pi i \mathrm{s}}{V_g}} \right) \tag{6.1}$$

where $\mathrm{u} = 0, .., W_g - 1, \mathrm{s} = 0, .., V_g - 1$ and W_g, V_g are sufficiently large, positive integers. This amounts to evaluating the PGF H around the complex unit circle and performing an inverse Fourier transform (denoted iFFT) to obtain the molecule count probabilities [26]. For a dataset $\mathbf{X}$, which comprises matrices $\mathbf{U}$ and $\mathbf{S} \in \mathbb{R}^{\mathrm{N}\times\mathrm{G}}$ (for N cells and G genes), W_g and V_g are defined as $W_g = \max(\mathbf{U}_g), V_g = \max(\mathbf{S}_g)$.

To obtain MLE (maximum likelihood) parameter estimates, $\hat{\boldsymbol{\theta}}_g$, given $\mathbf{X}$, the Kullback-Leibler Divergence (KLD) between the observed molecule count distribution and the distribution generated under the CME model is minimized :

$$\hat{\boldsymbol{\theta}}_g = \underset{\boldsymbol{\theta}_g}{\operatorname{argmin}}\, KLD(\widetilde{p}(\mathbf{U}_g, \mathbf{S}_g), p(\mathbf{U}_g, \mathbf{S}_g; \boldsymbol{\theta}_g))) \tag{6.2}$$

where $\widetilde{p}$ describes the observed histogram of counts. To optimize the global parameters λ^u, λ^s a grid search can be performed where Equation (6.2) is optimized for possible pairs of (λ^u, λ^s), and an optimal pair of (λ^u, λ^s) are chosen with the minimum KLD. This grid search is performed on the datasets prior to meK-Means inference, to obtain and set reasonable λ^u, λ^s values.

The moments of the CME model (Fig. 6.1b) are derived in [95]. For this study, we utilize the moment derivations for the unspliced and spliced mRNA count means:

$$\mu_u = \frac{\lambda^u k b}{\beta},\ \mu_s = \frac{\lambda^s k b}{\gamma}\ . \tag{6.3}$$

A general framework and tools for performing the numerical integration and parameter optimization for CME model inference with single-cell data are implemented

in the *Monod* package [94]. MeK-Means is integrated within the *Monod* package, utilizing these established workflows for solving such CME-based systems.

The meK-Means Algorithm

The meK-Means model introduces the latent variable Z to the CME model of transcription above, expanding the likelihood model of the data from $p(U, S|\boldsymbol{\theta})$ to $p(U, S, Z|\boldsymbol{\theta})$. Here Z can take any (integer) value from 1 to the user-defined K. Given that both the posterior, $p(Z|U, S, \boldsymbol{\theta})$, and parameters, $\boldsymbol{\theta}$, are unknown, we take an Expectation Maximization (EM)-based approach to optimize the Q function:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{\mathrm{t}}) = \mathbb{E}_{p(Z|U,S,\boldsymbol{\theta}^{\mathrm{t}})} \log\, p(U, S, Z|\boldsymbol{\theta}) \tag{6.4}$$

iterating between updating the posterior given parameter estimates $\boldsymbol{\theta}^{\mathrm{t}}$ (E-step), and determining the MLE parameter estimates which then maximize $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{\mathrm{t}})$ (M-step). See Algorithm 2 below.

To initialize the posterior, $p(\mathbf{Z}|\mathbf{U}, \mathbf{S}, \boldsymbol{\theta})$, given count matrices $\mathbf{U}, \mathbf{S}$, K-Means clustering was performed on the $\mathbf{U} + \mathbf{S}$ matrix (for the user-defined K) and for each cell n, $p(z_{\mathrm{n}} = \mathrm{k}_{\mathrm{n}}^{KMeans}|\mathbf{u}_{\mathrm{n}}, \mathbf{s}_{\mathrm{n}}) = 0.9$ where $\mathrm{k}_{\mathrm{n}}^{KMeans}$ is the cluster k assigned to cell n by K-Means. For all other k, $p(z_{\mathrm{n}} = \mathrm{k}_{\mathrm{n}}^{KMeans}|\mathbf{u}_{\mathrm{n}}, \mathbf{s}_{\mathrm{n}}) \sim Uniform[0, 1)$.

Since the numerical procedure for obtaining parameter estimates for the defined CME model requires a histogram over observed counts (see Equation (6.2)), we use hard assignment of cells to a single latent state or cluster k during the M-step, where the cell is assigned k such that

$$\mathrm{k} = \underset{\mathrm{k}}{\operatorname{argmax}}\; p(z_{\mathrm{n}} = \mathrm{k}|\mathbf{u}_{\mathrm{n}}, \mathbf{s}_{\mathrm{n}}; \boldsymbol{\theta}_{\mathrm{k}}^{\mathrm{t}}).$$

This is akin to the hard assignment of each observation (cell) to a cluster centroid (based on distance from the observation to that centroid) in K-Means clustering, hence the 'K-Means' in 'meK-Means'.

Note that in Equation (6.7), given that the hard assignment of k is determined by the maximum posterior value for a given cell, meK-Means can converge to a final number of assigned clusters less than the upper bound K set by a user.

6.1.2 meK-Means Benchmarking

Detailed benchmarking of the performance of meK-Means on simulated and real datasets, versus other common clustering approaches can be found in [50], but to

Algorithm 1: meK-Means

Data: $\mathbf{X} \in \mathbb{R}^{\mathrm{N}\times\mathrm{G}\times\mathrm{D}}$ with N cells, G genes, and D modalities. Here D = 2, where $\mathbf{X} = [\mathbf{U}, \mathbf{S}]$ and $\mathbf{U}, \mathbf{S} \in \mathbb{R}^{\mathrm{N}\times\mathrm{G}}$. User-defined K, for number of clusters.

Result: $\hat{\boldsymbol{\theta}}_{\mathrm{k}}$ and $\hat{\pi}_{\mathrm{k}}$ for k = 1, ..., K where $\hat{\boldsymbol{\theta}}_{\mathrm{k}} = \left[\hat{\boldsymbol{b}}, \frac{\hat{\boldsymbol{\beta}}}{\boldsymbol{k}}, \frac{\hat{\boldsymbol{\gamma}}}{\boldsymbol{k}}\right]_{\mathrm{k}}$, e.g., $\hat{\boldsymbol{b}}_{\mathrm{k}} = \left[\hat{b}_1, ..., \hat{b}_g\right]_{\mathrm{k}}$, and cluster assignments per cell $\hat{\mathbf{Z}} = [\hat{\mathrm{z}}_1, ..., \hat{\mathrm{z}}_{\mathrm{N}}]$ where $\hat{\mathrm{z}}_{\mathrm{n}} \in \{1, ..., \mathrm{K}\}$.

Initialize:

Mixing proportions $\boldsymbol{\pi}$ and $p(\mathbf{Z}|\mathbf{X};\boldsymbol{\theta})$. Set global parameters C^u, λ^s.

Optimize: $Q(\theta|\theta^{\mathrm{t}}) =$
$\sum_{\mathrm{k}}\sum_{\mathrm{n}}\left[p(z_{\mathrm{n}} = \mathrm{k}|\mathbf{x}_{\mathrm{n}};\boldsymbol{\theta}^{\mathrm{t}}_{\mathrm{k}})\log\, p(\mathbf{x}_{\mathrm{n}};\boldsymbol{\theta}_{\mathrm{k}})\right] + \sum_{\mathrm{k}}\sum_{\mathrm{n}}\left[p(z_{\mathrm{n}} = \mathrm{k}|\mathbf{x}_{\mathrm{n}};\boldsymbol{\theta}^{\mathrm{t}}_{\mathrm{k}})\log(\pi_{\mathrm{k}})\right]$.

for *t epochs* **do**

  if $t = 0$ **then**

    Do an M-step Update as in Equation (6.6) and Equation (6.7) to obtain $\boldsymbol{\pi}^0$ and $\boldsymbol{\theta}^0$ from initialized $p(\mathbf{Z}|\mathbf{X};\boldsymbol{\theta})$.

  end

  1. E-step Update

$$p(z_{\mathrm{n}} = \mathrm{k}|\mathbf{x}_{\mathrm{n}};\boldsymbol{\theta}^{\mathrm{t}}_{\mathrm{k}}) = \mathrm{softmax}(\log\, p(\mathbf{x}_{\mathrm{n}};\boldsymbol{\theta}^{\mathrm{t}}) + \log(\boldsymbol{\pi}^{\mathrm{t}}))_{\mathrm{k}} \tag{6.5}$$

  with $p(\mathbf{x}_{\mathrm{n}};\boldsymbol{\theta}^{\mathrm{t}})$ as in Equation (6.1).

  2. M-step Update

$$\hat{\pi}_{\mathrm{k}} = \frac{\sum_{\mathrm{n}} p(z_{\mathrm{n}} = \mathrm{k}|\mathbf{x}_{\mathrm{n}};\boldsymbol{\theta}^{\mathrm{t}}_{\mathrm{k}})}{\mathrm{n}} \text{ where } \sum_{\mathrm{k}} \pi_{\mathrm{k}} = 1. \tag{6.6}$$

$$\hat{\boldsymbol{\theta}}_{\mathrm{k}} = \underset{\boldsymbol{\theta}_{\mathrm{k}}}{\mathrm{argmax}} \sum_{\mathrm{n}} p(z_{\mathrm{n}} = \mathrm{k}|\mathbf{x}_{\mathrm{n}};\boldsymbol{\theta}^{\mathrm{t}}_{\mathrm{k}})\log\, p(\mathbf{x}_{\mathrm{n}};\boldsymbol{\theta}_{\mathrm{k}})$$

$$\text{where } p(z_{\mathrm{n}} = \mathrm{k}|\mathbf{x}_{\mathrm{n}};\boldsymbol{\theta}^{\mathrm{t}}_{\mathrm{k}}) = \begin{cases} 1 & \text{if k} = \underset{\mathrm{k}}{\mathrm{argmax}}\; p(z_{\mathrm{n}} = \mathrm{k}|\mathbf{x}_{\mathrm{n}};\boldsymbol{\theta}^{\mathrm{t}}_{\mathrm{k}}), \\ 0 & \text{otherwise.} \end{cases} \tag{6.7}$$

end

3. Final cluster assignment where $\hat{\mathbf{Z}} = [\hat{\mathrm{z}}_1, ..., \hat{\mathrm{z}}_{\mathrm{N}}]$ and $\hat{\mathrm{z}}_{\mathrm{n}} = \underset{\mathrm{k}}{\mathrm{argmax}}\; p(z_{\mathrm{n}} = \mathrm{k}|\mathbf{x}_{\mathrm{n}})$.

summarize, we first validated the performance of meK-Means to recapitulate cluster annotations, and ground truth parameters in the case of simulated data. We compared meK-Means results to several other clustering approaches beginning with standard methods such as Leiden [247, 262] and K-Means [173] clustering, as well as a combination of latent space (scVI [91]) or integrated nearest neighbor graph (WNN [108]) learning techniques with Leiden or K-Means as per suggested guidelines to combine reduced representations of the data with downstream clustering approaches [111]. We also compare results to scMDC [161] which dually learns a latent space representation and cluster partitions of the data, though this method was not able to run on all datasets. We also note that this is one of the few methods which explicitly combines integration of modalities and clustering. Most approaches currently tackle one or the other, without particular rationale for why or how to combine an integrated space with a clustering technique. For Leiden and K-Means we additionally run the algorithms with all possible input matrix options: $\mathbf{U}, \mathbf{S}, \mathbf{U}+\mathbf{S}, \mathbf{U} \oplus \mathbf{S}$. $\mathbf{U}+\mathbf{S}$ represents the summation of the individual $\mathbf{U}$ and $\mathbf{S}$ matrices. $\mathbf{U} \oplus \mathbf{S}$ represents the concatenation of the $\mathbf{U}$ and $\mathbf{S}$ matrices, or a more independent treatment of the modalities (as used for scVI). $(\mathbf{U}, \mathbf{S})$ denotes the treatment of each modality as its own matrix, by meK-Means and scMDC.

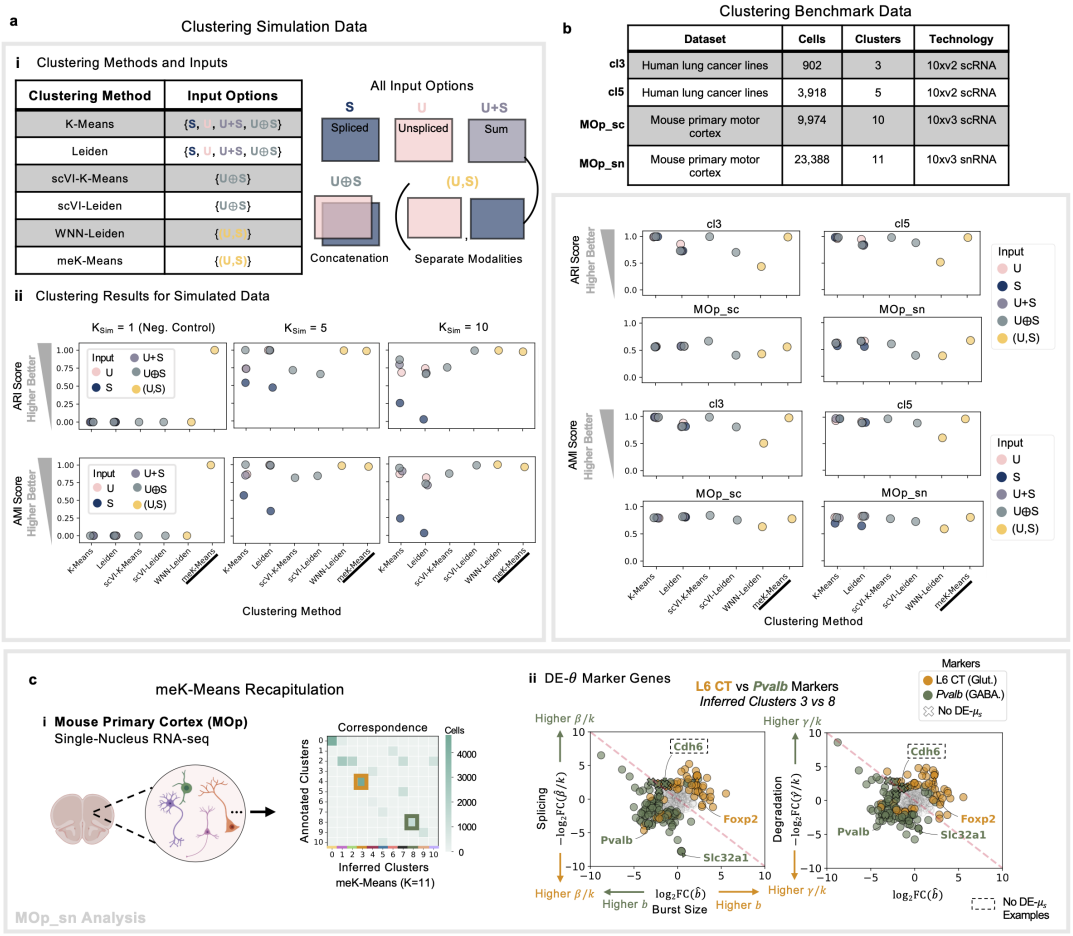


Figure 6.3: **meK-Means Benchmark Performance. a)** i. Table of all clustering methods tested with possible data input options. ii. ARI and AMI scores for each method versus true clusters, across possible inputs for the three simulations (with 1, 5, or 10 simulated clusters). For methods with a K hyperparameter, the same K as the data was used, and the default Leiden res parameter otherwise. **b)** Table of datasets used for benchmarking with relevant properties listed. ARI and AMI scores for all clustering methods across all possible inputs (see Supplementary Fig. 3 in [50] for scMDC results), as compared to the annotated cell types. For methods with a K hyperparameter, the same K as the data was used, and the default Leiden res parameter otherwise. **c)** i. Correspondence of meK-Means inferred clusters to the MOp_sn annotations. Values denote cell counts in the cluster overlaps. ii. 'DE' or 'Differentially Expressed' genes at the parameter-level (θ) shown between the clusters corresponding to L6 CT (Glutamatergic) and *Pvalb* (GABAergic) neurons. Genes in dashed box denote genes where DE is detected at the parameter-level but not at the observed, mean S-level. (Left) Splicing rate vs. burst size shown for genes. (Right) Degradation rate versus burst size shown. Adapted from [50]. Created with BioRender.com.

To generate simulation data we used the CME model of transcription described above to simulate K_{Sim} clusters, where a set of genes for each of the K_{Sim} clusters were perturbed, either increasing burst size or decreasing (relative) splicing rate for each gene chosen. By modulating burst size and splicing rate we induce changes in both unspliced counts and spliced counts, and by decreasing splicing (thus potentially spliced counts) we can test the detection limits of the clustering methods. These rate parameters for each gene, per cluster, define a probability distribution over unspliced (U) and spliced (S) molecule counts from which a dataset $\mathbf{X}$ can be sampled (see Equation (6.1)), containing $\mathbf{U}$ and $\mathbf{S}$ cell-by-gene count matrices. We tested three simulation datasets with $K_{Sim} = 1, 5, 10$, with 5000 cells each over a range of cluster sizes, where $K_{Sim} = 1$ represents a negative control dataset without cluster partitions. All the clustering methods were run on these datasets, with the Adjusted Rand Index (ARI) and Adjusted Mutual Index (AMI) scores used to assess each method's cluster assignments versus the ground truth assignments. For ARI, 1.0 denotes overlapping assignments and 0.0 represents poor or random assignments. For AMI, 1.0 denotes identical assignments and 0.0 represents when the mutual information between assignments is the same as the value expected due to chance. Overall, meK-Means performed as well on both metrics as the other best performing method in each simulation case, while no method besides meK-Means was able to determine that there was only one cluster for $K_{Sim} = 1$ (Fig. 6.3a ii), (resulting in varying, non-overlapping numbers of clusters from these methods). In addition to clustering assignment performance, meK-Means was also able to recover the biophysical parameters for the genes across the clusters, with high correlation (see [50]). Spearman and Pearson correlation are denoted by ρ and r, respectively.

Moving beyond simulation, we then tested the performance of meK-Means and all other clustering methods on benchmark biological datasets which had 'ground truth' clusters assignments [270] or cluster assignments that were obtained from several experimental paradigms (e.g., RNAseq and spatial transcriptomics [268]). We tested the methods on two scMixology datasets [270] which were a mix of 3 or 5 human lung adenocarcinoma lines ('3cl' and '5cl'), and two Allen Institute Mouse Primary Motor Cortex (MOp) datasets, an scRNAseq dataset ('MOp_sc') and an snRNAseq dataset (single-nucleus RNAseq,'MOp_sn') ([268] see Methods) (Fig. 6.3b). Datasets were filtered for HVGs as per standard scanpy procedure [262], and these genes were filtered for genes that were overdispersed and had sufficient U and S counts. This left 466, 357, 682, 359 genes for the cl3, cl5, MOp_sc, MOp_sn datasets, respectively. In future, sequencing technologies that provide

more unbiased capture of both nascent and mature mRNA will improve the number of genes capable of being used for such biophysical modeling [11, 264]. Again, meK-Means performed at least as well as the other best-performing method for each dataset, while the methods that performed worse on the real datasets did not necessarily correspond to the worse performers on the simulated dataset (i.e., there was no clear rationale as to which of the other methods would consistently perform better or worse).

In addition to cluster assignment, we analyzed the recapitulation of the MOp_sn data by meK-Means . For example, for the inferred cluster 8, the inferred mean unspliced and spliced counts across genes was highly correlated with the observed means in the cluster (see [5]). With meK-Means clustering results we can also look for 'DE-θ' genes, or genes where there is differential expression, a $\log_2$FC (fold change) > 2, in at least one parameter $b, \beta/k, \gamma/k$ between two clusters. Since the parameters describe the full, joint distribution of counts, parameter-level FCs may not be discernible FCs ($\log_2$FC > 1) at the level of mean spliced expression, which mimics the standard approach for differential expression [15]. Thus, DE-θ genes may not be DE-μ_s, 'differentially expressed in mean, spliced counts μ_s'. Expanding DE to the parameter level, also expands the definition of a gene being a 'marker' of one cluster versus another. For example, if we are interested in increased splicing between cell populations, a gene with a higher splicing rate (β/k) in cluster 1 versus cluster 2 would be a cluster 1 marker. Likewise, a gene with a lower degradation rate (γ/k) in cluster 1 versus cluster 2 could be denoted as a cluster 2 marker, if we are interested in increased mRNA stability. The definition of a marker gene is then broader, and does not necessarily agree with the more standard definition of marker gene as increased, spliced gene expression.

For these biological datasets, we thus denote a DE-θ gene as a cluster's marker when burst size is increased or both splicing and degradation are decreased (i.e., there is increased burst frequency or transcription rate k), as both suggest increased mRNA production. If neither is the case, an increase in splicing or decrease in degradation (increased mRNA stability) denote a marker. However, these parameter-level definitions of marker genes are flexible and likely to be task- or investigation-dependent.

We then extracted DE-θ genes between a glutamatergic (Glut.) and GABAergic (GABA.) cluster (Fig. 6.3c ii). Between the L6 CT (Glut.) and *Pvalb* (GABA.) neuron clusters we recovered known markers such as *Pvalb, Slc32a1,* and *Foxp2*

[268] (Fig. 6.3c ii). Additionally, we found DE-θ genes that did not have detectable FC if only considering spliced expression counts (they were not DE-μ_s), such as for the cadherin gene *Cdh6*, whose differential expression patterns across brain regions help specify developmental compartments and neuronal circuitries [123, 211] (Fig. 6.3c ii). Thus meK-Means is able to consistently cluster the benchmark data comparably to the other best performing clustering methods as well uncover kinetic differences between clusters which underlie known cell type markers and define novel markers for further investigation.

6.1.3 Biological Discovery with meK-Means

In addition to investigating the ability of meK-Means to define cluster markers through kinetic differences, we also sought to demonstrate the use of meK-Means for exploratory analysis and to develop hypotheses for how transcriptional dynamics define novel cell populations. For example, in [177] the authors were interested in understanding how transcription and splicing affect the maturation of germ cells in mice. However most inferences about transcription and splicing dynamics were made at the level of the observed, expression counts, e.g., where higher spliced expression implies greater transcription and a higher unspliced to spliced ratio implies less splicing. With meK-Means, we clustered testicular germ cells from early and later stages of development (E11.5 and E13.5 cells) to more explicitly model the changing dynamics between the populations and simultaneously identify heterogeneity within these stages, i.e., cluster of cells in different states of maturation and thus kinetic regulation [177].

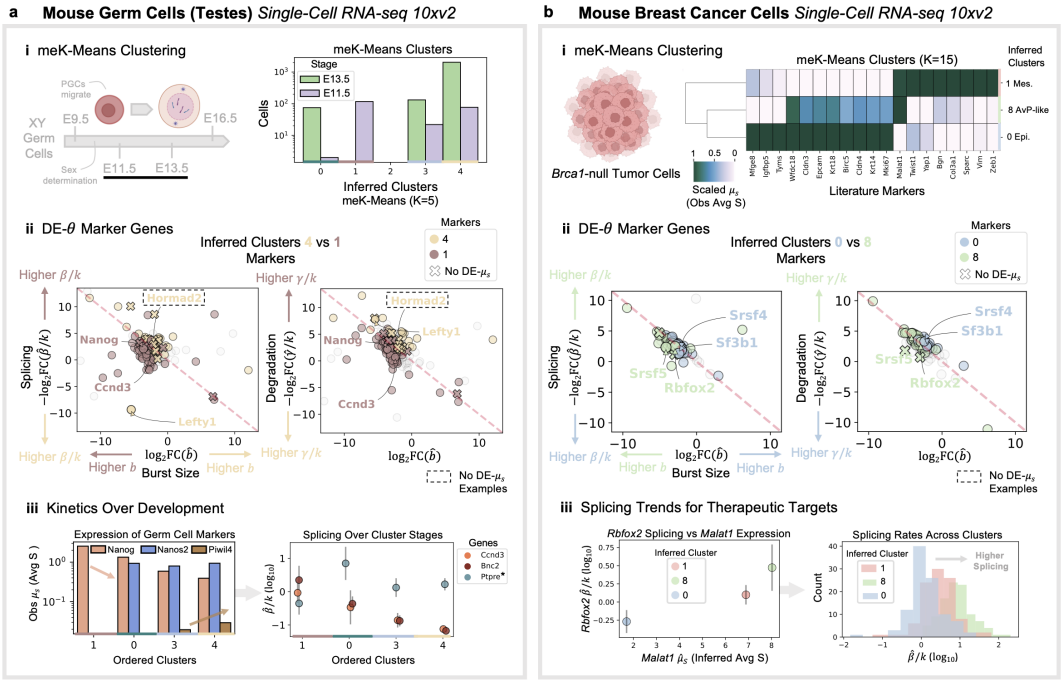


Figure 6.4: **meK-Means for Biological Discovery.** **a)** meK-Means results for mouse germ cells dataset. i. meK-Means clustering results for cells from both E11.5 and E13.5 stages. Barplot shows distribution of stages among the inferred clusters. ii. 'DE' or 'Differentially Expressed' genes at the parameter-level (θ) shown between inferred clusters 4 and 1. Genes in dashed box denote genes where DE is detected at the parameter-level but not at the observed, mean S-level. (Left) Splicing rate vs. burst size shown for genes. (Right) Degradation rate versus burst size shown. iii. Expression of germ cell maturation markers over the clusters, ordered by increasing 'maturity'. Splicing rate for genes shown across ordered clusters. Asterisk denotes gene trends not discussed in the original study. Error bars denote the 99% C.I. **b)** meK-Means results for mouse breast cancer dataset. i. meK-Means clustering result of *Brca1*-null tumor cells, shown on right in a heatmap, with expression of published markers for breast cancer populations (Epithelial Epi., Alveolar Progenitor-like AvP, Mesenchymal Mes.). ii. 'DE' genes at the parameter-level (θ) shown between inferred clusters 0 and 8. Genes in dashed box denote genes where DE is detected at the parameter-level but not at the observed, mean S-level. (Left) Splicing rate vs. burst size shown for genes. (Right) Degradation rate versus burst size shown. iii. (Left) Splicing rate of *Rbfox2* vs. mean *Malat1* expression (both inferred from biophysical parameters) in each cluster. Error bars denote the 99% C.I. (Right) Histograms of splicing rates in each cluster. Adapted from [50]. Created with BioRender.com.

We found that four clusters of cells emerged from our analysis, containing differing proportions of early and later stage cells (Fig. 6.4a i). Note that given the likelihood-based approach of meK-Means , we can use the Akaike Information Criterion (AIC) as a measure of model quality and appropriateness, to choose between meK-Means results at several values of K. In comparing the cluster with the largest number of E13.5 cells to the cluster of solely E11.5 cells (inferred clusters 4 vs. 1), our DE-θ analysis accordingly extracted the known pluripotency marker *Nanog* and cell cycle control-related *Ccnd3* gene as early stage markers, which are both downregulated in maturing testicular germ cells [177] (Fig. 6.4a ii). *Lefty1*, a gene associated with axis specification, displayed lowered burst-size but increased splicing and lowered degradation in the E13.5-dominated cluster, while *Hormad2*, a gene implicated in several processes regulating mammalian meiosis [142], displayed both decreased splicing and degradation (i.e, likely increased transcription rate k) in this same cluster, where fold change was not discernible through spliced expression analysis only (Fig. 6.4a iii).

Delving deeper into the assigned clusters, we noted that the clusters demonstrated graded expression of germ cell maturation markers, specifically displaying decreasing expression of the pluripotency marker *Nanog* and increasing expression of male-specific markers in gametogenesis, *Nanos2* and *Piwil4*, in more mature populations (Fig. 6.4a iii left). In accordance with these expression patterns, spermatogenesis-related genes such as *Bnc2* and cell-cycle-related *Ccnd3* both showed decreasing splicing rates over these cluster stages (Fig. 6.4a iii right), matching previous literature on increased intronic counts, and intro-retention, of these genes during spermatogenesis [177]. We also searched for genes whose splicing dynamics showed other interesting dynamics over the cluster stages, as with the phosphatase *Ptpre*, with roles in signal transduction and cell differentiation [157], which displayed increased splicing across the clusters (Fig. 6.4a iii right). We note that standard error bars can be calculated for the parameter estimates, as displayed in Fig. 6.4a iii, calculated from the square root of the diagonals of the inverse Fisher Information Matrix (FIM) . The FIM is calculated as the Hessian matrix of the KLDs between the observed count histograms and the distributions induced by the final inferred parameters. Thus meK-Means was able to quantitatively resolve the initial analysis in [177], to both identify relevant dynamics and novel states in the maturation of these germ cells, and move beyond spliced gene expression analysis.

In addition to revealing how transcriptional dynamics regulate normal development,

such kinetics are particularly relevant to understanding the progression and resistance of cancer cell populations, and for identifying therapeutic targets to, for example, mitigate aberrant splicing behaviors in tumor cells [14, 141]. Recent work by [269] used scRNAseq to profile tumor cell populations in several mouse cancer models. Using the mammary tumor sequencing data from *Brca1*$^{F/F}$ *Trp53*$^{F/F}$ *Krt14-Cre*, or '*Brca1*-null' mice), we applied meK-Means to interrogate the cell populations in these tumor samples that mimic basal-like breast cancers.

meK-Means clustering settled on three clusters within the *Brca1*-null tumor sample, an epithelial cluster, 0 (Epi.), an alveolar progenitor-like population, 8 (AvP-like), and a mesenchymal cluster, 1 (Mes.), corresponding to observations in [269] (Fig. 6.4b i, Supplementary Fig. 7). Through DE-θ analysis we found several splicing factors whose dynamics differed between the clusters. For example, splicing of *Rbfox2*, which can delineate mesenchymal cell states in the epithelial-mesenchymal transition (EMT) and increase "metastatic potential" [141] marked both the AvP-like and mesenchymal clusters in comparison to the epithelial cluster 0 (Fig. 6.4b ii). *Srsf4*, which in combination with drug therapies like cisplatin has been used to induce cancer cell apoptosis [141], and *Sf3b1*, whose knockdown can diminish tumorigenesis in MYC hyperactivated breast cancers [141] both demonstrated decreased degradation in the epithelial cluster versus the AvP-like cluster (Fig. 6.4b ii).

Interestingly, we found that splicing rates for *Rbfox2* increase in accordance with increasing *Malat1* expression across the clusters (Fig. 6.4b iii left). Additionally, histograms of all gene splicing rates in each cluster demonstrated a similar trend, with histograms shifted towards higher splicing rates in the clusters with greater *Malat1* expression (Fig. 6.4b iii right). These findings allowed us to not only parse the transcriptional dynamics underlying cancer cell populations in different stages of EMT, but also to develop hypotheses about how potential therapeutic targets, such as *Malat1*, may be affecting downstream genes and regulators, e.g., by altering *Rbfox2* splicing, and to design experiments to investigate or target those relationships. Together, by coupling our definition of clusters to the cellular processes underlying molecular measurements we can more explicitly investigate what components of DNA and RNA regulation are driving these clusters or states of interest.

6.1.4 Extending the meK-Means Framework

Through meK-Means we develop a scalable and interpretable methodology for defining clusters in single-cell, multimodal datasets, coupling the definition of 'cluster' to the governing parameters of the underlying, cellular processes which produce the joint distribution of the molecular measurements we observe. In the current workflow, preprocessing of data once the raw counts are obtained is limited to the selection of genes for inference. For most datasets we filter the standard HVGs selected by scanpy to ensure a minimum threshold of counts in both modalities and overdispersed behavior (in accordance with a bursty model of transcription). However, with multimodal data, common practice of selecting HVGs from spliced gene expression may not be the most informative approach, particularly if there is variability in another modality. In future one could learn which genes to retain for discerning clusters during inference, akin to determining which genes contribute most to a particular task or objective function such as separation of cell types [53, 75].

In future, integration of the meK-Means model with other machine learning (ML) techniques could enable simultaneous selection or filtering of relevant gene features as well as tuning the hyperparameter K. For example, recent work utilizing ML approaches to perform inference with CME-based models [43, 236], highlights how integration with meK-Means could enable simultaneous fitting of an 'optimal' K (as in [161]), cell size (read-depth) effects, and technical sampling parameters. Building off these approaches, one could in principle, also retain single-cell *and* single-gene resolution [43]. This would generalize the finite-dimensional representation of the cells in meK-Means to continuous representations of cells, where each cell is a nondeterministic function of a latent representation [43]. ML techniques and frameworks can also be implemented to improve runtime and scalability of meK-Means [43]. MeK-Means' runtime remains between 5-10 minutes per dataset for data spanning three orders of magnitude (100 cells to 100k cells) (see [50]), though solving the analytical solution to the CME model requires storing an array of size Ω (where Ω is a finite subdomain determined by maximum molecular counts observed for a gene) and a time complexity of $O(\Omega \log \Omega)$ [99]. Thus extension of this work with ML integration would further improve runtime capabilities and GPU compatibility [43, 99].

The model underlying meK-Means additionally assumes steady-state behavior of the cells. However, we note that setting K very large (i.e., as $K \to \infty$) can be seen

as approximating a continuous model (infinite mixture model) of the data. This is conceptually similar to the model in Chronocell [81], which models cells as being distributed along cellular time (trajectory inference), through a biophysical model of transcription. Thus, extensions of the biophysical model underlying meK-Means along these lines, would allow users to explore more continuous properties of the data while incorporating bursty transcription and technical sequencing effects.

As it stands, meK-Means results can also inherently be combined with a host of statistical tools to analyze the inferred parameters and model power. For example, the FIM calculated from the inferred models can be used to assess uncertainty in parameter estimates, as well as information content of parameters, to optimize downstream experiment design [87, 143]. Additionally, Chi-squared goodness-of-fit testing is used to reject genes with poor parameter inference, and was used prior to any DE analysis here.

6.2 Stochastic Modeling of Biophysical Responses to Perturbation

This section summarizes unpublished work by T.C, G.G., and L.P. The study was conceptualized by T.C., T.C. developed the code and analysis, and T.C., G.G., and L.P. wrote and edited the manuscript.

How do the principles of meK-Means, the reworking of classical scRNA-seq analyses through stochastic models of biology, then extend to multi-condition, multi-sample experimentation which can cover hundreds of genome-wide perturbations?

Though such datasets are promising for unraveling how the processes of DNA/RNA regulation produce the observed cellular responses, most tools for analysis of large-scale perturbation datasets focus on observational effects only, e.g., changes in expression, using only mature mRNA information, rather than the generative, transcriptional processes themselves [52, 69, 127, 169]. Statements about changes in transcription dynamics, for example, are then implied through changes in the counts. Deep learning approaches also focus on prediction of expression patterns [125, 169], even when considering multiple modalities [122], where physical interpretation of the learned parameters and relationships between the measurements can be hard to extract. These tools and approaches also often require several transformations of the data, to remove noise or enhance biological signal, which can themselves incur distortion and opaque interpretation [6, 48, 253]. Mechanistic approaches and investigations of transcription are often limited to smaller or more homogeneous systems [22, 87, 272], or assess modalities, and their corresponding dynamics,

independently [216].

Thus we demonstrate here, how extension of stochastic models of transcription to these noisy, high-throughput perturbation datasets alternatively defines common perturbation analyses through the underlying, biophysical processes of DNA/RNA regulation. Using just the unspliced and spliced count modalities, we can uncover condition-specific kinetics, predict regulation of transcription kinetics in combined perturbation settings, and define novel cell states induced by perturbation. With this approach, we can generate hypotheses about *how* perturbations affect the transcription and processing of RNA, for downstream investigation and experimentation.

6.2.1 Kinetic Effects of Perturbation on Transcription

For our analysis, we take as input scRNA-seq perturbation datasets with unspliced and spliced gene count matrices as described above, which represent nascent and mature mRNA molecule counts [113, 237]. The datasets in this analysis encompass both drug-based perturbations, A549 lung cancer cells under Dexamethasone (DEX) treatment [41] and mouse NSCs under 96 different drug combinations [92], as well as genetic intervention assays, specifically (single and dual) CRISPRa [191] and CRISPRi [204] perturbations in K562 cells (leukemia cell line).

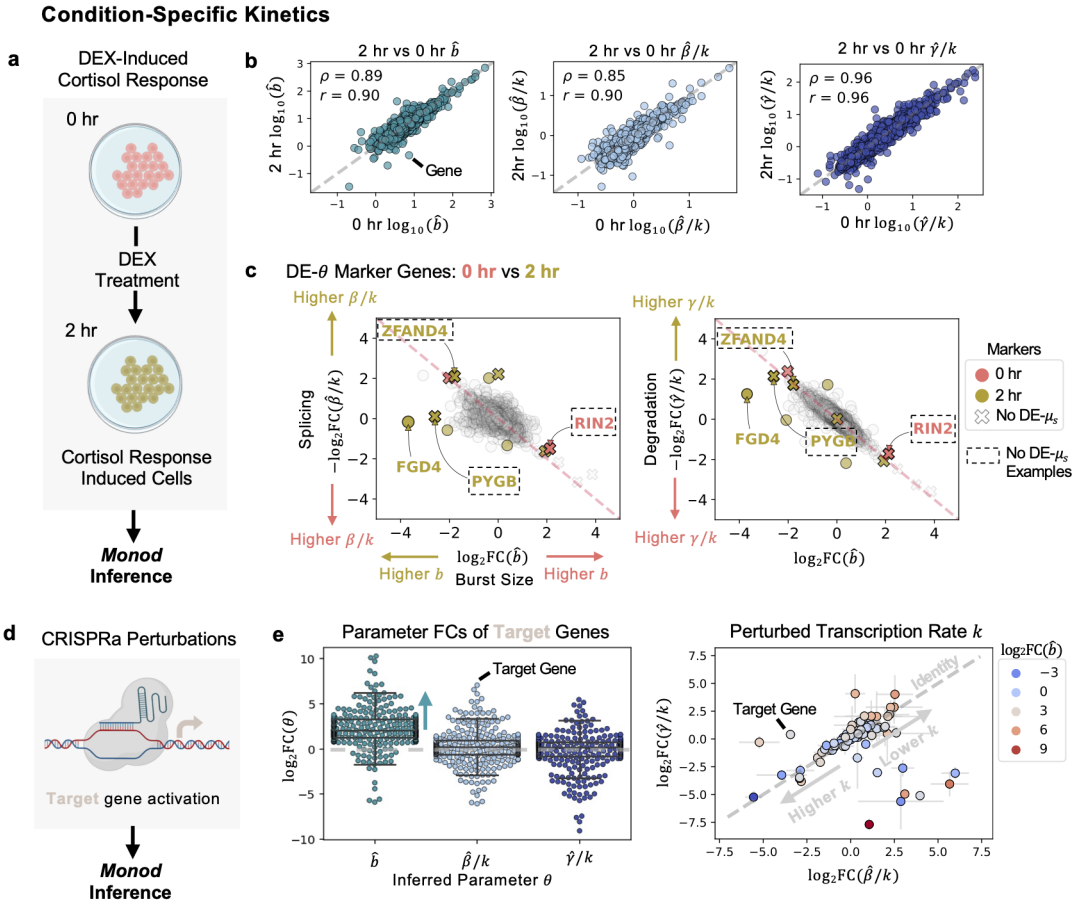


Figure 6.5: **Perturbation Condition-Specific Kinetics. a)** Diagram of DEX-treated A549 cells, at 0 and 2 hours. Parameter inference for data done using *Monod*. **b)** Inferred parameters, for burst size, splicing, and degradation compared between cells at 0 and 2 hours of treatment. Spearman and Pearson correlation are denoted by ρ and r, respectively. **c)** 'DE' or 'Differentially Expressed' genes at the parameter level (θ) shown between the clusters corresponding to 0 and 2 hour cells. Genes in dashed box denote genes where DE is detected at the parameter level but not at the observed, mean S-level. (Left) Splicing rate vs. burst size shown for genes. (Right) Degradation rate versus burst size shown. Grey genes denote ambiguous markers, or non-significant FCs. **d)** Diagram of CRISPRa perturbation used for *Monod* parameter inference. **e)** (Left) fold change (FC) of inferred parameters for the target (activated) genes. FCs shown as compared to all controls in the study. (Right) FCs for degradation versus splicing parameters for the target genes. Error bars denote standard deviation of FC across all controls. Genes shaded by burst size FC. Created with BioRender.com.

With the *Monod* package [94], described above, for parameter inference of CME models from single-cell data, we can then fit the biological (and technical) parameters which define the biophysical model of the sequencing data in Chapter 6.1.1, for the individual conditions in these datasets. This does not use the meK-Means model, as we are simply fitting parameters across *all* cells in a condition. This produces gene-specific parameters for each perturbation condition of interest, $\hat{b}, \hat{\beta}/k, \hat{\gamma}/k$ where the inferred estimate of a parameter θ is $\hat{\theta}$.

To better understand the *how* behind a perturbation's effect on gene expression, e.g., mRNA production, we used *Monod* to fit the biophysical parameters of the model in Fig. 6.1b on 3,000 genes (which contained a minimum number of unspliced and spliced counts) for the DEX-treated A549 cells at 0 hours of treatment and 2 hours of treatment. The single-cell indexing and labeling technique sci-fate was used to generate these data [41]. We thus extracted the burst size, splicing, and degradation parameters across genes (Fig. 6.5b) at each treatment time. All unspliced and spliced mRNA counts, which combine the 4sU labeled and unlabeled counts captured in this experiment, were used for parameter inference. As described in the original study [41], we found that the degradation rates between the two conditions displayed high correlations (Fig. 6.5b). However, we could additionally assess these correlations relative to the correlations between burst sizes and splicing rates of the two conditions, highlighting greater differences in these parameters as opposed to degradation rates (Fig. 6.5b). All three parameters were also fit under the same model of transcription, as opposed to fitting separate kinetic models for the parameter(s) of interest [41].

The original study noted a potential difficulty in using whole transcriptome information (labeled and unlabeled transcripts) to assess differences in the cells between the 0 hour and 2 hour condition [41]. It appeared that standard HVG selection and dimensionality reduction of the data resulted in an inability to separate the 0 hour and 2 hour responses, without solely focusing on the newly transcribed (labeled) mRNA or combining the individual principal components (PCs) of the labeled and unlabeled data. However, by fully describing the joint distribution of the unspliced and spliced counts through biophysical parameters, we have the ability to not only differentiate the perturbation responses of the 0 hour and 2 hour conditions through DE-θ genes, but also detect these differences for genes that do not necessarily show discernible FCs when comparing mean expression between conditions, i.e., that are not DE-μ_s, as is done in classical DE analyses (Fig. 6.5c). For example, we can

detect changes in the burst size of the cortisol response marker *FGD4* [41, 203] as well as the changes in burst size and splicing or degradation, for the gene *ZFAND4* (a prognostic and metastasis marker) [144], the glycogen phosphorylase gene *PYGB* [199] induced in other stress conditions, and the *RIN2* GTPase gene involved in endocytosis and membrane trafficking [243], which did not display FCs at the mean spliced expression-level (Fig. 6.5c).

To further validate and explore our kinetic realizations in various perturbation settings, we fit biophysical parameters for each genetic intervention condition (with greater than 50 cells) in the CRISPRa (activation) Perturb-seq study [191], which encompassed single and dual gRNA conditions (Fig. 6.5d). For this dataset, samples were generated with the 10x Genomics v2 protocol. The CRISPRa mechanism increases transcriptional output by potentially recruiting and stabilizing components of the transcription preinitiation complex [196].

In each activation condition, we found that burst sizes of the corresponding target genes were increased (with, on average, $\log_2$FCs greater than 2) (Fig. 6.5e left), mirroring burst size observations discussed in the Narta activation protocol, which recruited artificial transcription factors to the transcription site [158]. In comparison, average splicing and degradation rate FCs were near zero (Fig. 6.5e left). However, in cases where both the splicing and degradation rate FCs are in the same direction (sign), we can infer potential changes to the denominator of these relative rates, i.e., in the transcription rate k (Fig. 6.5e right) [94]. This reveals genes where transcription rate or burst frequency is altered, suggesting different strategies for the gene's transcriptional regulation (Fig. 6.5e right) [58, 190, 258].

By taking advantage of the full joint distribution between the modalities in these datasets, we can not only quantitatively realize the kinetic effects of a perturbation, incorporating transcriptional bursting and splicing dynamics, but also detect changes in these kinetic parameters not discernible in more standard, expression-based analysis.

6.2.2 Predictive Models of Combinatorial Perturbations

Many methods for analysis of perturbation data additionally focus on predicting the effects of perturbations in novel settings [125, 169, 170]. This can aid in simulation of responses, and minimizing experimental efforts for downstream investigations. Often these predicted changes or effects are defined as changes in spliced expression, a proxy for changes in transcription. However, different underlying mechanisms may

contribute to these observed changes. Previous work has described potential models of how perturbations or regulatory inputs in combination can impact transcription kinetics [219], and in turn downstream expression [216, 219], but application of such models is not generally extended to single-cell genomics perturbation data.

Additionally, it is non-trivial to use predicted spliced and unspliced counts to predict underlying, mechanistic effects. If tools predict changes in mean expression [69], this does not provide enough distributional information to infer dynamics. When counts are modeled more explicitly, the distributions parametrizing the observed counts from multiple modalities are independent, ignoring causal relationships and making physical interpretation of the learned parameters difficult [122, 169].

However, by using the inferred parameters from *Monod* in single-perturbation conditions, we can define models at the level of the kinetic parameters to predict the parameters in dual conditions (i.e., where both perturbations are present). The parameters of the single-perturbations then inherently describe their full joint distributions of unspliced and spliced count, and extend this description to the dual-perturbation setting. This also extends previous investigations, which assess the additive and multiplicative properties of mean spliced expression under perturbation [216], to the behavior of the governing rates which produce those observed behaviors. Given the inferred parameters in the single-guide conditions in the CRISPRa Perturb-seq study, and the inferred parameters in the control conditions, we can test the ability of simple models of additive and multiplicative behavior to recapitulate the kinetic parameters in dual-guide conditions. Specifically, we assessed how well multiplicative and additive models of the changes in burst size and transcription rate describe the observed changes in parameters in the combined conditions (Fig. 6.6a). We define the prediction models below:

Definition of Predictive Models

For the predictive models of kinetics in combined perturbation conditions, we focused on additive and multiplicative models of transcript production, i.e., to describe the changes in b or k. As described in [216, 219], multiplicative changes at both the level of mean expression FC and transcription/forward rate can result from the 'one-step recruitment' model [219] of transcription regulation, where the combined interventions behave in an additive manner to alter the free energy of the system, resulting in multiplicative effects at the level of the rates and observed expression FCs (Fig. 6.6a,c).

For this study, the multiplicative models of the parameters in the combined conditions (relative to control) are then:

$$b_{1,2}/b_{ctrl} = (b_1/b_{ctrl}) \times (b_2/b_{ctrl}) \tag{6.8}$$

and

$$k_{1,2}/k_{ctrl} = (k_1/k_{ctrl}) \times (k_2/k_{ctrl}) \tag{6.9}$$

where 1 and 2 denote the single-perturbation conditions, and '1,2' denotes the combined condition. Parameters denoted with 'ctrl' represent the control condition's parameters. These models also assume that changes in β/k represent changes in k, i.e., β/k and γ/k change together. Given that our biophysical model does not have a separate k parameter, we model these changes through β/k (denoted below as β' for convenience). Thus we rewrite Equation 6.9 as:

$$\beta'_{1,2}/\beta'_{ctrl} = (\beta'_1/\beta'_{ctrl}) \times (\beta'_2/\beta'_{ctrl}). \tag{6.10}$$

As also described in [216, 219], additive effects are also observed at the level of expression and can be derived from independent effects of the perturbations to, in parallel, catalyze or reduce the forward rate [219] (Fig. 6.6a,c).

For this study, the additive models are then:

$$b_{1,2}/b_{ctrl} = (b_{ctrl} + \Delta_1 + \Delta_2)/b_{ctrl}$$

where $b_1 = b_{ctrl} + \Delta_1$ and $b_2 = b_{ctrl} + \Delta_2$. This can be rewritten as:

$$b_{1,2}/b_{ctrl} = (b_1/b_{ctrl}) + (b_2/b_{ctrl}) - 1. \tag{6.11}$$

Likewise for k:

$$k_{1,2}/k_{ctrl} = (k_1/k_{ctrl}) + (k_2/k_{ctrl}) - 1. \tag{6.12}$$

To obtain a formulation in terms of β':

$$\begin{aligned} k_{1,2}/\beta &= (k_1 + k_2 - k_{ctrl})/\beta \\ \beta'_{1,2} &= \frac{1}{(1/\beta'_1) + (1/\beta'_2) - (1/\beta'_{ctrl})} \\ \beta'_{1,2}/\beta'_{ctrl} &= \frac{1}{\beta'_{ctrl}/\beta'_1 + \beta'_{ctrl}/\beta'_2 - 1}. \end{aligned} \tag{6.13}$$

Currently, these predictive models do not incorporate multiple steps in transcription production [22] as this uses a simplified version of the two-state telegraph model, where in the bursty-limit there is one forward rate k or k_{on} (inactive to active state transition) and the burst size b represents a ratio between the k_{ini} rate (rate of production) and k_{off} rate (active to inactive transition), as $k_{ini}, k_{off} \to \infty$ [95].

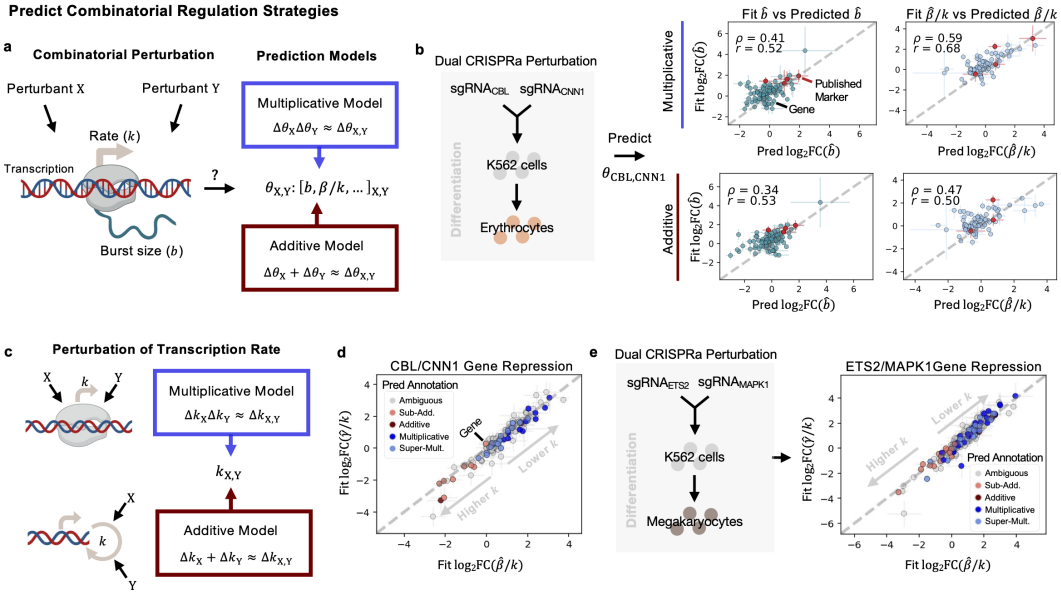


Figure 6.6: **Prediction of Combinatorial Regulation Strategies.** **a)** Diagram of potential effects of perturbants on transcription. Multiplicative and additive prediction models predict parameters in combined perturbation conditions from single perturbation conditions. **b)** Predicted parameters for the dual CRISPRa condition, for genes CBL and CNN1. Predictions from both models shown for burst size and splicing, correlated against the 'Fit' FC, from the inferred parameters in the combined condition. Spearman and Pearson correlation are denoted by ρ and r, respectively. **c)** Diagram of potential models of perturbant effects on transcription rate k, in a single-step transcription model. **d)** FCs of the inferred degradation rate versus splicing rate, in the combined CBL/CNN1 condition. Error bars denote standard deviation of FC across all controls. Genes colored by the best-fitting predictive model. **e)** FCs of the inferred degradation rate versus splicing rate, in the combined ETS2/MAPK1 condition. Error bars denote standard deviation of FC across all controls. Genes colored by the best-fitting predictive model. Created with BioRender.com.

Assessment of Combined Perturbation Predictions

The genes tested in these predictive models were selected in a similar fashion to the genes selected in the CRISPRa data study, where a random forest regression model was used to select genes that separated well the single-guides and dual-guide conditions from the control condition [191]. However, we did not include the dual condition in regression-based selection of genes, as we treat this condition as unseen. For these genes, we see positive correlations, > 0.5, for at least one models' predictions as compared to the observed FCs for the inferred ('Fit') parameters, across the dual conditions described in the [191], where Fig. 6.6b displays the correlation across all genes tested, of the predicted burst size and relative splicing rates in the dual condition where genes *CBL* and *CNN1* were activated. Genes in red denote 'Published Markers' or genes described in the original text as markers of the combined condition [191]. We then tested these predictive models on other data types, including dual CRISPRi conditions [204] and dual drug conditions (where low/mid/high ranges of concentrations of EGF, BMP4, or retinoic acid, RA, were added to NSCs) [92]. We found similarly positive correlations between the predicted parameter changes and observed changes, and could distinguish when, for example, the additive model better described changes in burst size than the multiplicative model, and vice versa. The positive correlations of the predicted parameters were also higher than the correlations produced by a negative control model, where parameters from single, control guide conditions replaced the single-guide conditions in the predictive models (representing how well random noise could predict the fit parameter changes).

Interpreting what these additive or multiplicative changes in kinetic parameters mean in terms of transcriptional regulation strategies being employed by the cell can be difficult, particularly if changes induced by the individual conditions are in different directions. Thus we focused interpretation on scenarios where the transcription rate k was likely affected (splicing and degradation rate FCs were in the same direction/of the same sign in both single-conditions). Given our bursty transcription model's assumption of the single forward rate k, additive effects on the transcription rate in the combined condition can be framed as deriving from parallel and independent effects of individual interventions to catalyze the reaction rate, k (see Eqn. 6.12) [219]. In the multiplicative case, additive effects of the interventions that affect recruitment and binding energy (e.g., of RNA pol II) combine to produce multiplicative effects at the level of the rate k [219] (see Eqn. 6.9) (Fig. 6.6c).

We then delved into two dual-guide conditions from the CRISPRa Perturb-seq study, where the individual guides demonstrated transcriptional effects in the same direction, likely inducing differentiation towards erythrocyte (Fig. 6.6d) or megakaryocyte (Fig. 6.6e) lineages [191], and selected for repressed genes, i.e., where there was a negative FC in unspliced counts. We assigned the most representative predictive model to the observed changes in the parameters of the dual-conditions, where a model was assigned to an observed parameter FC if the predicted FC fell within the 95% C.I. of the observed FC in the combined condition (constructed from the standard deviation of the FCs calculated in comparison to the individual control conditions) [216]. If both predicted FCs (from the multiplicative and additive models) fell into this range, the FC was denoted as 'Ambiguous'. Observed FCs larger, or smaller, than both predictions were denoted as 'Super-Multiplicative' or 'Sub-Additive', respectively.

Though for many genes it was ambiguous whether the additive or multiplicative model fit better [216], among the genes where we could discern more model-specific behavior, we found a dominance of the multiplicative and super-multiplicative predictions when the transcription rate was lowered (Fig. 6.6d,e). This suggests use of a more recruitment-based strategy as described above, potentially with non-independent effects of the interventions [219], to affect repression of genes in these conditions. Thus, even with simpler models of how perturbations act in combination, by approaching the prediction task from the level of the kinetic parameters, we can expand previous works assaying multiplicative and additive behaviors at the expression-level, and provide hypotheses of regulation strategies employed by the cell.

6.2.3 Uncovering Perturbed Populations with Distinct Kinetics

In addition to prediction of kinetic parameters across modalities, such as unspliced and spliced counts, it is a non-trivial problem to discover and define populations of cells demonstrating distinct perturbation responses given multiple molecular measurements. As discussed in Chapter 6.1, standard approaches to clustering scRNA-seq data do not use multiple modalities at once [52, 247], use heuristics to map between clusters or neighborhoods if given individual modalities [108], or use non-physically-interpretable deep-learning methods to integrate the modalities for clustering [161]. This then results in multiple, potentially arbitrary choices for a user to make when deciding how to combine the modality-specific matrices for clustering or determining which method's clustering results to proceed with [50].

To this end, we applied the meK-Means clustering algorithm of Chapter 6.1.1 to simultaneously learn populations or clusters of cells in heterogeneous perturbation conditions [50].

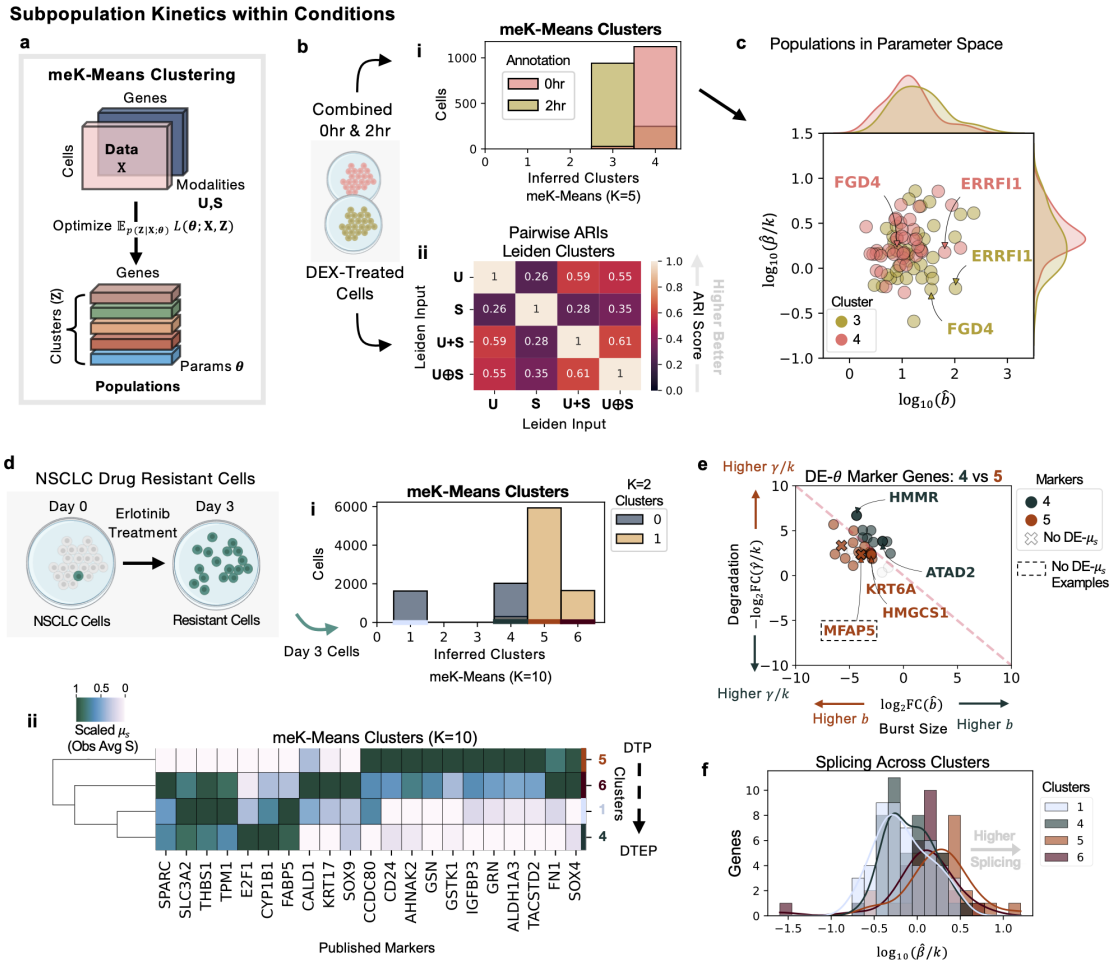


Figure 6.7: **Inference of Subpopulation Kinetics within Conditions. a)** Diagram of the meK-Means algorithm for clustering **b)** Diagram of combined 0 and 2 hour DEX-treated cells, then passed to meK-Means or Leiden for clustering. i. Barplot of cluster assignments from meK-Means shown, where K=5, and distribution of 0 and 2 hours cells between them. ii. Pairwise ARI scores between the Leiden clustering results given various input matrix options. **c)** Plot of inferred splicing versus burst size parameters for the inferred clusters 3 and 4. Density plots of the respective parameter distributions shown per cluster (top and side of plot). **d)** Diagram of drug resistant NSCLCs after 3 days of erlotinib treatment. Day 3 cells passed to meK-Means for clustering. i. Barplot of cluster assignments from meK-Means, where K=10, and the distribution of the two populations of cells from the meK-Means K=2 clustering between them. ii. Hierarchical dendrogram plot of meK-Means inferred clusters based on mean expression (scaled across columns) of marker genes from the literature. **e)** 'DE' or 'Differentially Expressed' genes at the parameter level (θ) shown between the inferred clusters 4 and 5 . Genes in dashed box denote genes where DE is detected at the parameter level but not at the observed, mean S-level. Degradation rate versus burst size shown. Grey genes denote ambiguous markers, or non-significant FCs. **f)** Histograms of splicing rates shown for all genes across the inferred clusters. Created with BioRender.com.

We first demonstrated the use of meK-Means on the DEX-treated A549 cells, combining cells from 0 and 2 hours of treatment (Fig. 6.7b). As described in the original study [41], it was difficult to separate these two populations using whole transcriptome information and standard scRNA-seq processing pipelines. However, given a previously published list of genes potentially implicated in the cortisol response [203], also filtered for sufficient unspliced and spliced counts as well as overdispersion [50], meK-Means was clearly able to separate the treated cells (including 2 hour treated cells with lesser responses, i.e, more 0 hour-like properties) (Fig. 6.7b i). Given that only 45 genes remained after filtering, this demonstrates the importance of quality over quantity in gene selection, and potential pitfalls of standard HVG selection in cases like this, or other more ad hoc procedures.

Effectively, meK-Means clusters cells in biophysical parameter-space, as shown in Fig. 6.7c, where parameters of genes with validated transcriptional changes in the cortisol response genes (e.g., *FGD4* and *ERRFI1*) [41, 203] stand out from each other between the two populations. The meK-Means-inferred parameters for these genes also corresponded near-identically to the parameters previously inferred from the conditions separately.

We then applied meK-Means to cells without clear treatment partitions, where PC9 cells, an EGFR-mutant non-small cell lung cancer (NSCLC) cell line, were treated with erlotinib, a common first-line treatment for NSCLC, and sequenced after three days (following the 10x Genomics v3 protocol) [8]. At Day 3, multiple drug-resistant populations of cells had developed [8]. In the development and persistence of drug-resistant cancer cells, the kinetics of splicing as well as transcription are particularly relevant to how these cells acquire resistance and proliferate [257]. The Day 3 cells were thus clustered with meK-Means, using genes from both 'classical' HVG selection [262] and genes from the literature potentially marking resistance development, again filtered for overdispersed behavior and minimum unspliced and spliced counts (Fig. 6.7d). From this, we found four clusters of cells (Fig. 6.7d i), belonging to two larger populations of cells as also described in the original study [8]. These four populations spanned drug-tolerant persister (DTP) and drug-tolerant expanded persister (DTEP) states described in the study [8], which can persist and proliferate in the presence of drug treatment. For example, populations representing more DTP-like states demonstrated greater expression of the resistance marker *TACSTD2*, while DTEP-like states expressed the marker *CYP1B1* at higher-levels (Fig. 6.7d ii, Fig. 1g in [8]).

We then extracted DE-θ genes between the inferred populations, specifically more DTP or more DTEP clusters, e.g., 4 and 5. Markers included the microfibril-associated gene *MFAP5*, which did not display high FCs at the mean spliced-level, but did display differing burst size, splicing, and degradation rates between the populations (Fig. 6.7e). DE-θ genes also included genes with differential behavior between the DTP and DTEP states in the original study, such as downregulation of *KRT6A* (associated with epithelial development) [8] and *HMGCS1* (associated with cholesterol metabolism) [8] in the DTEP cluster (4) (Fig. 6.7e). We additionally found reduced degradation of the *HMMR* gene between the populations, a prognostic marker gene in several other human cancers [222] (Fig. 6.7e). Since splicing dynamics in resistant populations are also of interest, we examined the distribution of splicing rates in each population. This revealed increased splicing rates overall in the DTP-like populations as compared to the DTEP population (Fig. 6.7f), suggesting potentially more aberrant splicing behavior in these populations [257].

This biophysical approach to clustering perturbation data, brings together the count modalities under a self-consistent model of the biology, and highlights not only which cells demonstrate similar perturbation responses, but also which components in the transcription processing pipeline define those shared responses.

6.2.4 Limits and Extension of Stochastic Models

Overall, this application of stochastic biophysical models to high-throughput genomics data, demonstrates an alternative avenue for how we analyze, interpret, and develop hypotheses from large-scale perturbation. The *Monod* CME inference package and the meK-Means clustering algorithm, enable this analysis for noisy and discrete single-cell data, and extract physically-interpretable parameters from the data as well as demonstrate methods that can be consistently extended to new measurements. This approach to analysis, additionally removes several transformation and preprocessing steps in standard practice (beyond cell and gene selection), which even across the most popular packages for scRNA-seq analysis, are not standardized in their implementations, can result in opaque distortions, and arbitrary choices for the user to make [207].

As discussed above in Chapter 6.1.4, the biophysical models utilized in *Monod* and meK-Means focus on a relatively simple model of bursty transcription, and assume the effects of cell size (read depth) are negligible. These methods additionally utilize CME models of biological systems where analytical solutions are available, solved

through the *Monod* framework. However, recent developments in combining ML with biophysical models of transcription, for parameter inference [43, 236], suggest promising extensions of this work to simultaneously incorporate cell size effects on transcription and extend inference to more complex biophysical models, without analytical solutions, that can be simulated. Recent work to incorporate the effects of cell cycle in CME-based models of transcription, for high-throughput data, could also be used to both model such dynamics explicitly and determine if the data falls within regimes that require such considerations [126].

Along these lines, though meK-Means does represent shared relationships between genes in order to cluster the cells, it does not learn more direct interactions between genes and how they may effect the biophysical parameters of the system. However, integration of such approaches with statistical and learning-based approaches for causal inference from perturbation data [54, 168, 231], described further in Chapter 7, would merge the learned interactions with their effects on transcriptional dynamics.

Yet, as described in Chapter 6.1.4, this biophysical approach to perturbation analysis naturally inherits the use of several statistical tools, from rejection testing to model selection criteria, to enable a user to reject hypotheses in an interpretable manner. These underlying physical models additionally mirror the physics of other inter-modality relationships, such as between mRNA and protein expression [26], and chromatin state and mRNA transcription [82]. Thus, as perturbation genomics data become increasingly complex and multimodal, this work demonstrates a paradigm that aims for scalability not just in dataset size but also in interpretation, with methods that can extend to new measurements and modalities and provide physically-interpretable insights into the cellular processes governing our molecular measurements.

6.3 Moments for Time-Dependent Biophysical Modeling

In the previous sections, we focus on steady-state models of transcription in our analysis, particularly as many publicly available perturbation datasets are generated many hours up to days after the perturbation(s) have been applied i.e., enough time for the system to reach a 'steady-state'. However with technologies like sci-fate, we can gain temporal resolution of these transcriptional processes (sampling cells within a smaller time window). In these cases, we can directly fit the time-dependent version of the CME model [226] described in Chapter 6.1.1, with a 'true'

time defined by the experimental start time.

Below we solve for the moments of the length-biased CME model, which allow for future extensions of the above work to time-dependent investigation of perturbation within the current CME inference frameworks, such as *Monod*. This should enable fitting of all four biophysical parameters (including transcription rate k).

As defined in Fig. 6.1b, N^u, N^s are random variables that represent the 'biological' unspliced, spliced counts, and U, S the final sequenced unspliced and spliced counts. B is a random variable representing burst size, as denoted in [226]. Note that $\mathbb{E}[B] = b, \mathbb{E}[B^2] = b(2b+1)$ assuming mean geometric burst sizes b [39]. We assume initial conditions of $\mathbb{E}[N^u](0), \mathbb{E}[N^s](0) = 0$. This is the case if, for example, we are modeling labeled mRNA with labeling starting at timepoint 0 (i.e., prior to the start time, no mRNA was being sampled or labeled for capture). All notation follows the model definitions in Chapter 6.1.1.

We first derive and provide the first and second moments for the nascent counts N^u, and the first moment for N^s, integrating over N given the initial conditions above. For more details on the differential equations defined here see 'Export Processes Enhance mRNA Autocorrelation Times' in [226].

$$\begin{aligned} \frac{\mathrm{d}\mathbb{E}[N^u]}{\mathrm{d}t} &= kb - \beta\mathbb{E}[N^u] \\ \mathbb{E}[N^u](t) &= \frac{kb}{\beta}(1 - e^{-\beta t}) \end{aligned} \tag{6.14}$$

$$\begin{aligned} \frac{\mathrm{d}\mathbb{E}[N^s]}{\mathrm{d}t} &= \beta\mathbb{E}[N^u] - \gamma\mathbb{E}[N^s] \\ \mathbb{E}[N^s](t) &= \begin{cases} \frac{kb}{\gamma} + (-\frac{kb}{\gamma} - kbt)e^{-\gamma t} & \beta = \gamma \\ \frac{kb}{\gamma}(1 - e^{-\gamma t}) + \frac{kb}{\beta - \gamma}(e^{-\beta t} - e^{-\gamma t}) & \beta \neq \gamma \end{cases} \end{aligned} \tag{6.15}$$

$$\begin{aligned} \frac{\mathrm{d}\mathbb{E}[(N^u)^2]}{\mathrm{d}t} &= k\mathbb{E}[B^2] + \beta\mathbb{E}[N^u] + 2kb\mathbb{E}[N^u] - 2\beta\mathbb{E}[(N^u)^2] \\ \mathbb{E}[(N^u)^2](t) &= -2\frac{k^2b^2}{\beta^2}e^{-\beta t} + \frac{k^2b^2}{\beta^2} + \frac{kb^2}{\beta} - \frac{kb}{\beta}e^{-\beta t} + \frac{kb}{\beta} + \frac{kb^2}{\beta}(\frac{k}{\beta} - 1)e^{-2\beta t} \end{aligned} \tag{6.16}$$

$$\begin{aligned}\mathbb{V}[N^u](t) &= \mathbb{E}[(N^u)^2] - (\mathbb{E}[N^u])^2 \\ &= \frac{kb^2}{\beta} - \frac{kb}{\beta}e^{-\beta t} + \frac{kb}{\beta} - \frac{kb^2}{\beta}e^{-2\beta t}\end{aligned} \tag{6.17}$$

The first and second moments for the sequenced counts U, S, after technical sampling, at some time t, can be derived from the relations in [95]:

$$\begin{aligned}\mathbb{E}[U](t) &= \lambda^u \mathbb{E}[N^u](t) \\ \mathbb{E}[S](t) &= \lambda^s \mathbb{E}[N^s](t) \\ \mathbb{V}[U](t) &= (\lambda^u)^2 \mathbb{V}[N^u](t) + \lambda^u \mathbb{E}[N^u](t) \\ \mathbb{V}[S](t) &= (\lambda^s)^2 \mathbb{V}[N^s](t) + \lambda^s \mathbb{E}[N^s](t)\end{aligned}$$

At this point we have the moments $\mathbb{E}[N^u](t), \mathbb{E}[N^s](t), \mathbb{V}[N^u](t)$, but not $\mathbb{V}[N^s](t)$ (i.e., four equations for the four unknowns $\theta = k, b, \beta, \gamma$). Given some value of t, we could then solve for the parameters in terms of the moments to initialize estimates for inference, in *Monod*, for example. However the forms of $\mathbb{E}[(N^s)^2](t)$ and $\mathbb{E}[N^u, N^s](t)$ are difficult to solve analytically, thus it may be more prudent to initialize one parameter with a biologically reasonable estimate given the experimental paradigm, then recover the remaining moment-based estimates of the other parameters. For example, given timecourse data, a simple exponential decay model can be fit to approximate β (decrease in unspliced mRNA over time), then used to calculate the moments-based estimates for b, k, γ.

Chapter 7

CAUSAL INFERENCE FOR NOISY PERTURBATION DATA

As noted in the Chapter 6.1, meK-Means clustering does extend the standard, independent treatment of genes in CME model inference to incorporate greater representation of gene-gene correlations. However, there is not an explicit construction or representation of interactions between genes (for example, how one gene influences the transcription of another) or joint regulation of genes. But, perturbation (or 'intervention') data can help illuminate the causality and directionality of such interactions [231]. However, current models for causal inference of gene networks generally ignore noise in our measurements or assume a single source of noise. How can existing structures for causal inference, under perturbation, be extended to noisy, transcriptional systems with intrinsic and extrinsic sources of noise as described in Chapter 6.1.1? To understand the capacity of causal inference in such systems, we begin with a simpler, non-kinetic model of transcription and sequencing, and investigate the performance of popular causal inference methods extended to noisy regimes.

This chapter summarizes unpublished work by T.C under the supervision of R.L., J.C.H, A.R, and K.L. at Genentech. The study was conceptualized by T.C., R.L., and J.C.H, T.C. developed the code and analysis, with feedback from R.L, J.C.H., A.R., and K.L, and T.C. wrote and edited the report below.

7.1 Causal Inference for Intervention Data

Determining causal relationships between variables, whether to predict or simulate outcomes, or to understand how interventions perturb particular relationships is an essential task for biological investigation. This can range from probing protein-protein interaction networks to predicting disease outcomes from patient risk factors [232, 279]. Often, this investigation amounts to learning the graphical, conditional relationships between variables e.g., directed acyclic graphs (DAGs) . Until recently, algorithms for causal inference and DAG learning were dominated by combinatorial search algorithms, dependent on pairwise conditional independence testing/statistics and score-based, greedy graph search methods [31, 213, 232, 277]. However, with the recent increase in high throughput genomics, perturbation datasets, these combinatorial methods become less tractable, and/or do not utilize the interventions

in these datasets [31, 277].

The development of the NO-TEARS [277] algorithm, bypassed the necessity for a combinatorial search, utilizing a continuous optimization approach. Differentiable Causal Discovery from Interventional Data (DCDI) [31] in turn, builds upon this work, expanding the probabilistic representation of the data for generative modeling purposes, and learning nonlinear relationships between parental and child nodes in the DAG. From this, Differentiable Causal Discovery of Factor Graphs (DCD-FG) [167] improves the scalability and utility of intervention-based DAG discovery by representing the nodes (genes) as low rank factors (e.g., 'gene modules').

Beyond the work to improve continuous optimization DAG inference techniques for large-scale genomics datasets, other causal inference techniques focus on developing the biological interpretability of DAG inference by incorporating the discrete nature of the molecular counts [195]. Additionally, the previously described algorithms are observational models (treating the observed data as causal, Fig. 7.1 leftmost model). Given that biological processes themselves are intrinsically noisy, measured data is often a proxy for the underlying, causal factors, and that sequencing/measurement noise is prevalent in many biological disciplines [213], measurement noise models and latent causal models provide an avenue for treating these various sources of noise and separating observational versus latent ('intrinsic') dependencies [213, 232] (Fig. 7.1 middle and right models). Again, however, these algorithms are largely relegated to combinatorial search methods and/or do not incorporate intervention data.

Here we aim to develop scalable, biologically relevant noise models for causal inference from intervention data, by building upon the aforementioned continuous optimization, DAG-learning approaches to incorporate more complex and relevant noise models of the data. Using black-box stochastic variational inference (SVI) [115], we hope to develop latent causal inference models which can incorporate measurement noise, as well as latent/shared confounders or intrinsic states, and can easily expand the observational model class to various continuous and discrete distributions. The goal would be to use such models to improve prediction of unseen perturbations, determination of gene targets for follow-up experimentation, and incorporation of other data modalities, as well as imperfect intervention states.

In this study, we begin by implementing discrete observational causal inference models and develop latent causal models with measurement noise utilizing SVI, extending the existing NO-TEARS, DCDI, and DCD-FG approaches. We then investigate the ensuing challenges in fitting these complex noise models. To this

end, we assess the potential limitations of the variational lower bound provided by the ELBO loss (in the implementation below) towards resolving DAG structures for latent, measurement noise models by designing and testing a diverse set of simulated datasets under various inference model assumptions. We then provide suggestions for followup investigation and model design to minimize performance variability and identify data properties to which each model is better or worse-suited.

7.1.1 Background for NO-Tears, DCDI, and DCD-FG Inference Algorithms

Briefly, the models developed here build off of the NO-TEARS, DCDI and DCD-FG algorithms. However, we can define all these models through the DCD-FG framework in [167]. In [167] data is represented as a factor-directed acyclic graph (f-DAG) , given an input data matrix $X \in \mathbb{R}^{d \times n}$ for d features (genes) and n observations (cells). The f-DAG is represented as $G_f = (V, F, E)$ with $V = \{v_1 ... v_d\}$ feature vertices, $F = \{f_1 ... f_m\}$ factor vertices, and the edge set E that links vertices (of different factor types). The number of factors m is set by the user.

Factors can be thought of as gene 'modules', or groups of similarly acting genes. G_f induces two half-square graphs $G_f^2[V]$ and $G_f^2[F]$, representing the graphs of distance two between pairs of vertices or factors, respectively. The adjacency matrix for the graph G_f can be represented by the adjacency matrix of just the half-square graph $G = G_f^2[V]$, where $\mathcal{A}(G) = \mathbf{U} \diamond \mathbf{V}$ and $\diamond$ is the Boolean matrix product. $\mathbf{U} \in \{0, 1\}^{d \times m}$, representing edges from features to factors, and $\mathbf{V} \in \{0, 1\}^{m \times d}$ representing edges from factors to features. $\mathbf{UV}$ additionally represents the weighted adjacency matrix for G.

Utilizing these lower rank representations of G_f, acyclicity need only be enforced on either $\mathbf{UV}$ or $\mathbf{VU}$. This is currently enforced either by penalizing the trace (Tr) of $\exp\{\mathbb{E}[\mathbf{UV}]\}$ or the spectral radius of $\mathbb{E}[\mathbf{UV}]$. Together, the algorithm presented in [167], optimizes the objective below, comprised of the likelihood model and the acyclicity constraints:

$$\max_{\Phi,\Theta} \mathcal{S}(\Phi, \Theta) - \gamma_t C(\mathbb{E}[\mathbf{M}(\Phi)]) - \frac{\mu_t}{2}(C(\mathbb{E}[\mathbf{M}(\Phi)])^2, \tag{7.1}$$

where C refers to the acyclicity penalty, and

$$\mathcal{S}(\Phi, \Theta) = \mathbb{E}_{\mathbf{M}' \sim \mathbf{M}(\mathbf{\Phi})} \left[\sum_{k=1}^{K} \mathbb{E}_{X \sim P_{\text{data}}^{(k)}} \sum_{j \notin \mathcal{I}_k} \log p_\Theta^j(X_j; \mathbf{M}'_\mathbf{j}, X_{-j}) \right] - \lambda ||\mathbb{E}[\mathbf{M}(\Phi)]||_1. \tag{7.2}$$

$\mathbf{M}(\Phi) = [\mathbf{U}(\Phi), \mathbf{V}(\Phi)]$ represents the distribution over f-DAGs parametrized by Φ. Θ represents the conditional distribution parameters, and $P^{(k)}_{\text{data}}$ is the distribution of data points under the intervention regime k. $\mathcal{I}_k$ denotes the nodes intervened on in regime k.

The density model $p^j_\Theta(X_j|X_{-j})$ is specified through the factor and variable nodes, specifically through a deterministic, non-linear function of the genes within a factor (Eqn. 7.3) and a probabilistic representation of the final observed counts (Eqn. 7.4).

$$h_f = \text{MLP}(\mathbf{U}_{:,\mathbf{f}} \circ X; \Theta_f) \text{ for } f \in F \tag{7.3}$$

and

$$X_j \sim \text{ Normal } (\alpha_j^\top (\mathbf{V}_{:,\mathbf{j}} \circ h) + \beta_j, \sigma_j^2). \tag{7.4}$$

MLP represents a multilayer perceptron which can approximate an arbitrary, non-linear relationship.

The NO-TEARS and DCDI algorithms discussed in Ch. 7.1, can be represented by simplified versions of the objective in Eqn. 7.1 by removing the factor-dependence of the density model in Eqn. 7.4 such that the distribution $p^j_\Theta(X_j|X_{-j})$ is defined where $X_j \sim \text{ Normal } (\alpha_j^\top (f(\mathbf{M}^T X)) + \beta_j, \sigma_j^2)$. For NO-TEARS $f = I$ and for DCDI $f = \text{MLP}$, i.e., the data is a linear or non-linear function of the learned DAG.

7.2 SVI Extensions for Causal Inference on Noisy Systems

In the following sections and results we tested a range of 3 types of inference models building off the NO-TEARS, DCDI , and DCD-FG models, with details of each model in the sections below. Overall, we constructed (1) observational models (without latent variables) shown on the left in Fig. 7.1, as baseline models for comparison. These observational models represent the NO-TEARS or DCDI model objectives, with the addition that P can be Poisson, to test discrete observation distributions. (2) We extended the NO-TEARS and DCDI observational models to incorporate latent variables Z which represent noisy biological counts, that are sampled and measured as the observed counts X (Fig. 7.1 middle). (3) We extended the DCD-FG model to incorporate latent variables Z into the factor representations, such that the factors (or gene modules) produce noisy gene counts Z which are then sampled and measured as the observed counts X (Fig. 7.1 right).

For the purposes of this thesis we will focus on the results from the observational and latent noise models *without factors*, as the challenges observed in these models extended to the factor-level models, and are easier to parse through these simpler model descriptions.

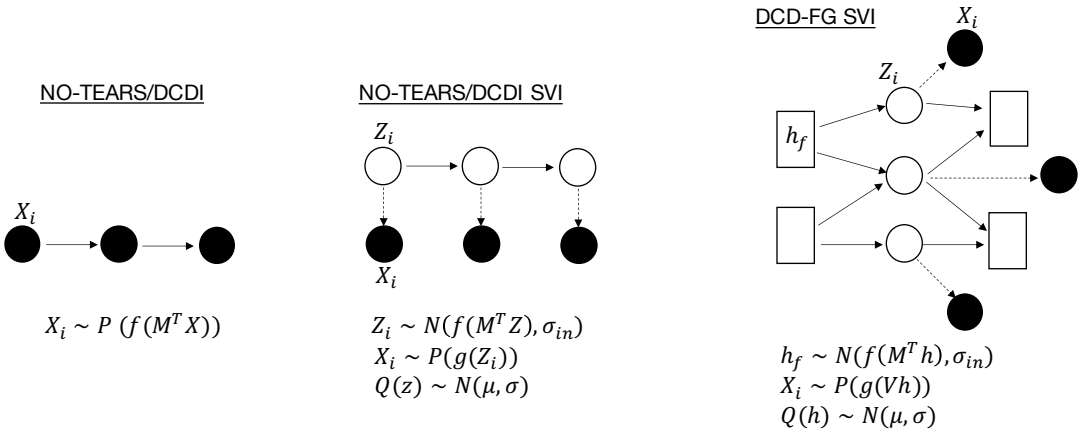


Figure 7.1: **Causal Model Descriptions.** Diagrams of the observational and latent noise models implemented (without and with factors).

7.2.1 Observational Model Description

We modified the objectives of the NO-TEARS/DCDI models to incorporate discrete models of the counts data, where $X_j \sim$ Poisson $(e^{(\alpha_j^\top (f(\mathbf{M}^T X)) + \beta_j)})$. Thus the output of $f(\mathbf{M}^T X)$ provides the λ for the Poisson.

7.2.2 NO-TEARS/DCDI SVI Model Description and Implementation

We implement SVI as a black box inference method to fit the measurement noise model below. The model below is shown for a Gaussian measurement noise model, however the sampled observations do not need to be Gaussian, as described above.

The model is defined as

$$Z_j \sim \text{Normal}(f(\mathbf{M}^\top Z), \sigma_{in}), \tag{7.5}$$

$$X_j \sim \text{Normal}(Z_j, \sigma_{ext}), \tag{7.6}$$

and

$$q(Z|X) \sim \text{Normal}(\mu, \sigma), \tag{7.7}$$

We denote the NO-TEARS or DCDI SVI models as those where $f = I$ or MLP , respectively. We assume a diagonal covariance structure for the Normal models,

where the diagonals are the denoted σ values. For Poisson SVI models, $X_j \sim \text{Poisson}(exp(Z_j + l))$.

The objective for optimization then becomes:

$$\mathcal{S}(\Phi, \Theta, \phi) = \mathbb{E}_{\mathbf{M}' \sim \mathbf{M}(\mathbf{\Phi})} \left[\sum_{k=1}^{K} \mathbb{E}_{q^{(k)}(Z|X)} \sum_{j} [\log p^j_{\Theta}(X_j, Z_j; \mathbf{M}'_{\mathbf{j}}, Z_{-j}) - \log q^j_{\phi}(Z_j|X_j)] \right] - \lambda ||\mathbb{E}[\mathbf{M}(\Phi)]||_1. \tag{7.8}$$

The ELBO loss within (7.8) (bracketed difference between the log-likelihood and log-posterior) is expanded as

$$\sum_{j \notin \mathcal{I}_k} [\log p^j_{\Theta}(X_j, Z_j; \mathbf{M}'_{\mathbf{j}}, Z_{-j}) - \log q^j_{\phi}(Z_j|X_j)] + \sum_{j \in \mathcal{I}_k} [\log p^j_{\Theta}(X_j, Z_j) - \log q^j_{\phi}(Z_j|X_j)]$$

where q represents our approximation to the true posterior. We assume perfect interventions, where intervened nodes ($j \in \mathcal{I}_k$) have no conditional dependencies. All models were fit with σ_{in} and σ_{ext} held fixed.

The inference algorithm for fitting the SVI latent noise models is shown in Algorithm Box 2. The full loss, including the augmented Lagrangian terms following from [31, 167] are additionally included. Within each minibatch, prior to Step 3, we additionally tested including extra training loops in which only the minibatch's variational parameters are updated via gradient descent with respect to the other parameters (i.e., the DAG parameters and bias terms are held fixed) (denoted as 'Train' models in results below), prior to the full parameter update following Step 3. We also note that the mean parameters for $q(Z|X)$ are initialized at the data X. The same algorithm is used to fit the observational models, without any latent variables.

'Full' Gaussian NO-TEARS SVI Implementations as Baseline

As a comparison to the SVI model above, we also implemented a Gaussian noise model with full covariance matrices (denoted as 'Full') to be fit, where Z is sampled as $Z = \mu + L\epsilon$ where μ from $q(Z|X)$ is learned and L is the Cholesky decomposition of the (learned) covariance matrix. This was to test if including a full covariance matrix improved DAG inference/performance.

We additionally tested model performance using the exact posterior form for the Gaussian measurement noise model (denoted as 'Post.') in Eqn. 7.5, Eqn. 7.7 where $q(Z|X) = p(Z|X)$. Thus

$$\mu = \Sigma_{in}(\Sigma_{in} + I)^{-1} X^T, \; \Sigma = \Sigma_{in} - \Sigma_{in}(\Sigma_{in} + I)^{-1} \Sigma_{in}^T$$

Algorithm 2: Latent Augmented Lagrangian with SVI

Data: $X \in \mathbb{R}^{n \times d}$

Result: Variational distribution $q_\phi(Z|X) = \prod_{j=1}^{d} q_\phi(Z_j|X_j)$, Weighted DAG $\mathbf{M}(\Phi)$, Distributional parameters θ, ϕ, Φ

Initialization:

$q_\phi(Z|X) = Normal(\mu_q, \sigma_q)$ where $\mu_q \in \mathbb{R}^{n \times d}, \sigma_q \in \mathbb{R}^d$

Create m minibatches from X of size $b \times d$

while $C(\mathbb{E}[\mathbf{M}(\Phi_t^*)]) > 10^{-8}$ *or* $\mu_t < 10^{32}$ **do**

  for i *in* $1...m$ **do**

    1. Sample Z^i where $Z^i = \mu_q^i + \sigma_q^i * \epsilon$ and $\epsilon \sim Normal(0, I)$

    2. Calculate loss

$$L = \mathbb{E}_{\mathbf{M}' \sim \mathbf{M}(\mathbf{\Phi})} \Big[\sum_{k=1}^{K} \mathbb{E}_{q^{(k)}(Z|X)} \sum_{j \notin \mathcal{I}_k} [\log p_\Theta^j(X_j^i|Z_j^i) p_\Theta^j(Z_j^i|\mathbf{M}'_\mathbf{j}, Z_{-j}^i) - \log q_{\phi^i}^j(Z_j^i|X_j^i)]$$
$$+ \sum_{j \in \mathcal{I}_k} [\log p_\Theta^j(X_j^i|Z_j^i) - \log q_{\phi^i}^j(Z_j^i|X_j^i)] \Big] - \lambda ||\mathbb{E}[\mathbf{M}']||_1 - \gamma_t C(\mathbb{E}[\mathbf{M}']) - \frac{\mu_t}{2}(C(\mathbb{E}[\mathbf{M}'])^2$$

    3. Calculate gradients and update θ, ϕ^i, Φ

  end

  if $(\theta_t^*, \phi_t^*, \Phi_t^*)$ *converged (stationary point found)* **then**

    Update Lagrangian parameters $\rightarrow \gamma_{t+1}, \mu_{t+1}$

  end

end

for $p(Z|X)$, where $\Sigma_{in} = (I - \mathbf{M}^T)^{-1}(I - \mathbf{M}^T)^{-T}$. $\mathbf{M}$ is the only learned parameter in this model. $f = I$ for the 'Full' and 'Post.' models.

7.2.3 DCD-FG (Factor) SVI Model Definition

For completeness we include the model definition for the extension of the DCD-FG factor-based model to a latent noise model. The intrinsic/extrinsic noise model for factor-graphs, which utilizes non-deterministic, low-rank factor nodes, is defined as:

$$h_f \sim \text{Normal}(f(\mathbf{M}^T h), \sigma_{in}), \tag{7.9}$$

$$X_j \sim \text{Normal}(\mathbf{V}^T h, \sigma_{ext}), \tag{7.10}$$

and

$$q(h|X) \sim \text{Normal}(\mu, \sigma), \tag{7.11}$$

The ELBO loss is thus defined with respect to $p_\Theta^j(X_j, h; \mathbf{M}'_\mathbf{j}, h_{Pa(j)})$, where $X_j \sim Poisson(\mathbf{V}^T h)$ for the Poisson Factor SVI model.

7.3 Causal Inference Performance on Noisy Simulations

For all simulations, models, and results in this section we limit analysis to systems where $f = I$, i.e., where observations are a linear function of the DAG/their parent nodes. Thus all models shown below are NO-TEARS observational or latent noise SVI models.

7.3.1 Noisy Simulations

Utilizing the DCDI simulation code base, we generated noisy, gene count datasets akin to the measurement noise model formulation of [213]. (1) Nodes without parents and intervened nodes were initialized from a Normal$(0, 1)$. (2) Coefficients/weights for the DAG M are sampled from $U[-2, -0.5] \cup U[0.5, 2]$ to increase the signal-to-noise ratio. (3) The intrinsic latent variables (Z) are sampled from a Normal$(f(\mathbf{M}^T Z), \sigma_{in})$, given $\mathbf{M}$. (4) To simulate Gaussian sampling, observations X_j are sampled from a Normal$(g(Z_j), \sigma_{ext})$. Here $f, g = I$. For Poisson sampling, observations X_j are sampled from Poisson$(exp(Z_j + l))$ where l is $\in U[1, 3]$.

All model results shown include results for each inference model over 10 sampled DAGs, and over a hyperparameter grid for sparsity penalty λ (see Algorithm 2) at values $\{.001, .01, .1, 1, 10\}$. For each DAG the 'best' result is selected by lowest Val (validation data) NLL or ELBO loss. SVI models were also run with learning_rates at $\{.00001, .0001, .001\}$, as stochastic gradient descent (SGD) with higher learning rates often ran into nans in the gradient updates. Non-SVI models were run with a learning_rate of .001.

To determine the accuracy of DAG recovery from any model we measured several metrics relative to the ground truth DAG. Between the predicted and true binary DAG matrix we assessed the FDR (false discovery rate, where 0.0 is ideal), F1 score (combining precision and recall capabilities for edges predicted, where 1.0 is ideal), and SHD (structural hamming distance) i.e., every edge that would need to be removed/added/flipped to obtain the true DAG. For some noise models we additionally calculated the Frobenius norm ('fro') of the difference of the weighted true and predicted DAGs as well as the MAE (mean absolute error) of the predicted means for the validation data and the true data values ('Val MAE').

Low variance Poisson Simulation Results:

Poisson gene count simulations were generated as described above in (3) and (4), with $\sigma_{in} = .01$ for 10 nodes with expected outdegree of 1. A dataset of 50k observations with 100 single and double interventions were used.

The baseline models were NO-TEARS and NO-TEARS Pois (NO-TEARS observational model with Poisson counts X), for comparison with the NO-TEARS Pois SVI model. We found that vanilla SGD optimization seemed to improve model performance (not shown), thus the results below utilize SGD only. However, even with increased inner-training rounds ('Train', as described above), the SVI model is unable to reach the level of performance of either baseline model (Fig. 7.2a).

Low variance Gaussian Simulation Results:

To remove any model fitting limitations introduced by the Poisson distribution, we fit the NO-TEARS and NO-TEARS SVI models to data simulated from a Gaussian model, as described above in (3) and (4). These simulations were generated with $\sigma_{in} = \sigma_{ext} = .01$ for 10 nodes with expected outdegree of 1. A dataset of 50k observations with 100 single and double interventions was generated. As shown in Fig. 7.2b, we again observed a decreased accuracy with respect to DAG recovery for the SVI model, even if training rounds within each parameter update step were increased (i.e., where the variational parameters where updated for several rounds while holding the DAG fixed).

High Noise Gaussian Simulation Results:

To clarify whether facets of the simulation were leading to the better performance of baseline methods, or the inconsistencies in the SVI models, we simulated a Gaussian measurement noise model with greater variance among the sampled observations and utilized the known latent values Z (from simulation) to perform sanity checks on the implemented SVI model.

We simulated Gaussian data as described above in (3) and (4), with $\sigma_{in} = .01$ and $\sigma_{ext} = 1$ for 10 nodes with expected outdegree of 1. 5k observations with 100 single and double interventions were used.

In addition to the baseline NO-TEARS model and our NO-TEARS SVI (latent) model, we tested providing the true latent values Z from simulation as μ for $q(Z|X)$, to determine if in a more ideal case the underlying DAG could at least be recovered (denoted as 'Give Z' models in Fig. 7.2c).

We also tested increased inner training rounds for the SVI model (as described before as 'Train'), or implementing a warmstart procedure ('Warm' models) where for 50 initial epochs only the parameter $\mathbf{M}$ (the DAG) is updated.

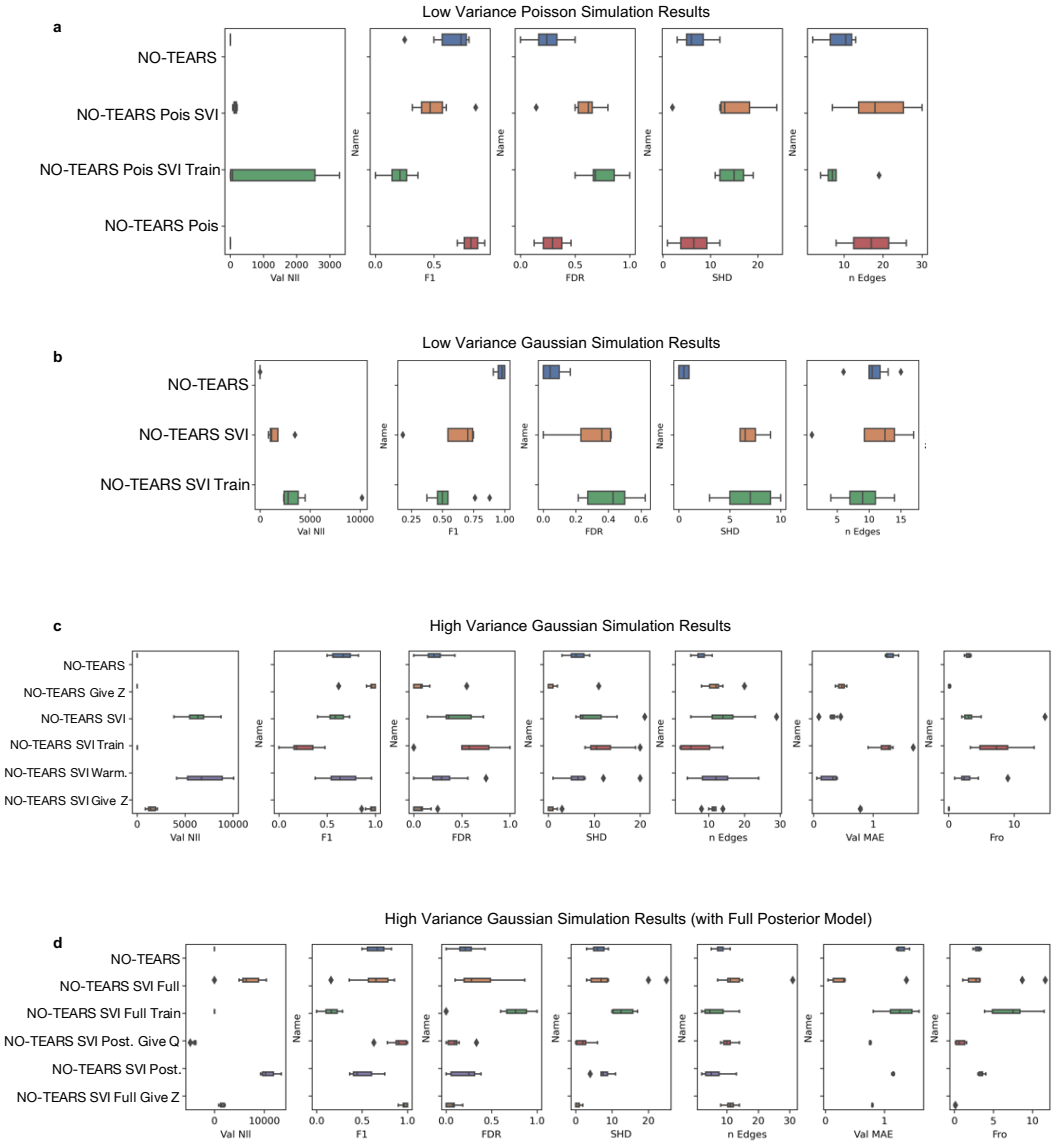


Figure 7.2: **Performance of SVI Models on Simulated Data.** **a)** Results of the NO-TEARS Poisson SVI model versus baseline models. Val/nll refers to full ELBO loss for SVI **b)** Results of the NO-TEARS SVI model, with Gaussian observations, versus baseline models. **c)** Results of the NO-TEARS SVI model, with higher variance Gaussian observations, versus baseline models. 'Give Z' models refer to models trained with $\mu = Z$ for $q(Z|X)$. **d)** Results of NO-TEARS SVI models with full Gaussian variational posteriors ('NO-TEARS SVI Full') and the true *form* of the posterior, $q(Z|X) = p(Z|X)$ ('NO-TEARS SVI Post.'). 'Give Q' refers to models trained with the true $q(Z|X) = p(Z|X)$ where $p(Z|X)$ is also determined from the true DAG.

The greater observational noise demonstrated the limitations of the baseline, observational NO-TEARS model which displays worse DAG recovery due to the increased noise in the input data. Additionally, given Z the SVI models can recover the true DAG (Fig. 7.2c).

However, we also observed the same trend from the previous results in these high noise simulations, where the other SVI models performed worse than the baseline NO-TEARS observational model (Fig. 7.2c). We additionally observed that the ELBO loss (denoted 'Val NLL') can be effectively reduced/minimized with increased inner-training rounds ('Train' model), but the DAG recovery is then poorer than SVI models without this extra training (Fig. 7.2c). A similar trend was additionally observed if KL annealing was implemented, where a lower weight is placed on the KL term in the loss function. This resulted in better DAG recovery but minimal movement away from the variational parameter initializations, i.e., when means of $q(Z|X)$ remained fixed around the data X we better learn the DAG (not shown). This was also noted with the warmstart procedure ('Warm' models). Thus though a low ELBO loss is possible with the optimization procedure, the resulting variational parameters and corresponding DAG may not be good fits to the observed data/data likelihood. And improved DAG recovery for the SVI models came at the expense of learning the variational parameters.

High Noise Gaussian Simulations with Full Gaussian and Posterior Inference:

In an attempt to combat this 'ELBO bias' we allowed the model to fit a full Gaussian covariance matrix for $q(Z|X)$, and implemented the true posterior form for $p(Z|X)$ as defined in Chapter 7.1.1, to determine if this level of structure is then necessary to effectively recover the true DAG with the latent SVI models (i.e., is there a fundamental limitation with the variational ELBO approach)?

We utilized the exact same simulated data as above, fitting NO-TEARS SVI models with full, learned covariance matrices ('Full' models), or given the exact posterior form for $q(Z|X)$, i.e., $p(Z|X)$ ('Post.' models). The 'Give Q' model represents a similar control for the posterior model, where $q(Z|X) = p(Z|X)$ for the *true* DAG $\mathbf{M}$. It does appear that the 'ELBO bias' observed previously, does not occur with the exact posterior models i.e., higher ELBO losses accordingly denote worse DAG recovery (Fig. 7.2d). Yet, the DAG recovery itself is poorer than the DAG recovery with the NO-TEARS SVI Full or standard NO-TEARS SVI models (Fig. 7.2d). This is a surprising result, as the NO-TEARS SVI Post. model should have an "easier" optimization task, where given the posterior form for $q(Z|X)$ and thus deterministic

formulas for the mean and covariance, $\mathbf{M}$ is the only parameter to optimize.

Given that the exact posterior form includes multiple matrix inversions and subtractions, the numerical instabilities that seem to result during gradient updates/calculations may be affecting the ability of this model to effectively optimize for $\mathbf{M}$. To currently address these numerical issues, posterior mean and covariance calculations are calculated without saved gradients and I is added to the diagonal of the posterior covariance.

7.4 Limitations and Future Directions for Noisy, Causal Inference

Overall, the investigations here reveal a hidden complexity in fitting both unknown variational parameters and causal relationships in tandem, with this naive SVI approach. The gap between the ELBO loss described here and optimizing directly the full likelihood of the data may allow for too much variability in this stochastic learning procedure, and thus result in worse recovery of the gene-gene interaction DAGs. This raises two questions and potential avenues of future investigation/optimization: (1) how fundamental to the model is this 'gap' in optimization (at which steps is this variability incurred/most detrimental), and (2) what properties of data (simulated or real) potentially affect the performance of these causal inference algorithms?

To parse and diagnose the difficulties encapsulated in (1), it is likely that further dissection of the components of the optimization procedure and model is required. In particular, it is important to investigate further the behavior of the posterior model/inference procedure, beginning with the initialization of the posterior. Though we currently set the posterior means to the observed data, it may be more fruitful to use random initializations, or to initialize all means to observational data without interventions. It will also likely be informative to take a fully deterministic approach (utilizing the MLE for the intrinsic/extrinsic noise models described here) to assess if in this setting it is possible to recover the DAG well. Another potential source of variability is in our handling of interventions. Currently, though intervened nodes are simulated from a Normal$(0, I)$ distribution, the objective function in Eqn. 7.1 does not model this distribution, but incorporating this information, through the prior distributions of the intervened nodes, may improve model fit. Likewise, removing this variability, and setting interventions to constant terms, may also improve learning abilities.

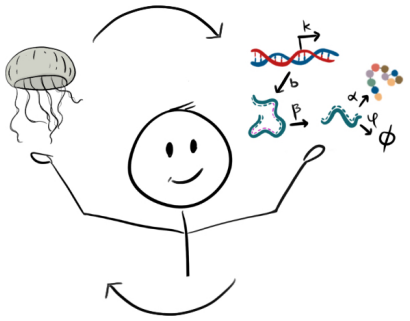
Beyond the internal limitations of the algorithms tested, there is also the more general question of (2): for all these algorithms (from NO-TEARS to the SVI models), in

what situations is each algorithm more or less appropriate, and actually beneficial in terms of the information gained about the system? Are there particular properties of these simulated datasets that were better suited to the simpler NO-TEARS model? In this study, we focus on equal intrinsic and extrinsic noise settings, or settings with greater extrinsic noise, and DAGs with sparse outdegrees and linear relationships between parent and child nodes. But it is possible these SVI/latent noise models may confer greater advantage in settings with greater intrinsic noise, for example. Thus an interesting avenue of investigation lies in determining the DAG properties that better suite more simplistic causal models versus more complex, latent variable models of the data. In this vein, it may also be more worthwhile to utilize simplistic approaches to learn gene-gene interactions which could more easily be integrated with the biophysical models of these counts, as described in Ch. 6, potentially updating this graph based on the fit of the (joint) count distributions generated from the biophysical parameters to the observed count distributions.

Together, using the results of this study we can attempt to form a deeper understanding of the limiting complexities of stochastic, latent noise models of high-throughput genomics perturbation data, which can in turn help illuminate other ways of developing scalable approaches for learning gene-gene interactions while incorporating multiple sources of noise present in our datasets.

Chapter 8

DISCUSSION AND CONCLUSION



I am made and remade continually. Different people draw different words from me.

VIRGINIA WOOLF

Through this thesis, we have grappled with the intricacies of multiplexed and multi-faceted perturbation biology, the pitfalls and limitations of dimensionality reduction for exploratory analysis, and the potential for stochastic biophysical models to rewrite how we extract insight from these complex data types. The work presented suggests that in exploratory analysis of high-throughput, perturbation data, there is not necessarily one solution or representation to our biological questions, rather there are different representations to be constructed from the data which are better (and worse) suited to particular lines of inquiry. This means, that though we present biophysical paradigms for reinterpretation of classical analysis tasks for more interpretable data representation, several directions of development remain for improving both experimental and computational methods, and particularly in interlinking the two throughout the scientific process. Thus as we look to the future of multimodal perturbation biology, what challenges can be addressed which improve experimental design and data collection alongside physically-interpretable data representation?

In Chapter 3, multiplexed and multi-condition experimentation enabled the discovery of novel cell types as well as their potential strategies in response to perturbation. However, only exon-containing (spliced) mRNA was used for analysis. Many common scRNA-seq platforms alongside 10x Genomics, utilize poly(A) capture of mRNA (unless more targeted sequencing is desired), meaning that the capture of

mRNA in intermediate or earlier states of processing is largely accidental. As described in Chapter 6.1, this means that many genes do not have enough counts of nascent and mature mRNA, for example, to allow for modeling of their transcriptional kinetics.

To enable improved resolution of the transcriptional process, RNA-seq technologies that offer more 'unbiased' capture of transcripts provide a promising alternative, such as the long-read and single-molecule resolution techniques provided by Pacific Biosciences and Oxford Nanopore [11], which do not require the shearing of mRNA into short sequences or extra amplifications of the molecules. Likewise the use of random-primed sequencing data (as opposed to the standard oligo-dT priming) as in technologies like LR-Split-seq [202], lessens the 3' bias of transcripts assessed (as the poly(A) tail is found at the 3' end of processed mRNA).

scRNA-seq data additionally loses the spatial context of the cells and their transcripts, though promising developments in high-resolution spatial perturbation assays [83], in situ sequencing [221], and measurements of 3D genome organization and cellular compartmentalization [25] expand the experimental toolkit and, in turn, the underlying biophysics and regulatory interactions to be explored. These orthogonal experimental assays also promise high-throughput methods for better understanding the overlap (and differences) between molecular count interpretations from scRNA-seq and imaging/fluorescence-based readouts.

In Chapters 5 and 6 we highlight the lack of biological interpretability of common dimension reduction techniques for data representation, and instead present biophysically-grounded approaches to common perturbation analysis tasks, such as clustering, differential expression, and mechanistic interpretation of perturbation interactions. However, we also note the limitations with the current biophysical approaches, particularly in light of potential extensions to the alternative experimental readouts described above. The biophysical models described here are limited to those with analytical solutions and sequential, memory-less behaviors. But in combination with ML approaches to extend parameter inference to more complex biophysical scenarios [43, 236], and with temporal resolution (see Chapter 6.3) [22, 87, 266], models representing alternative mechanistic hypotheses could be fit at-scale and with greater parameter resolution. The biophysical paradigm additionally offers a self-consistent starting point for modeling the technical biases inherent in non-poly(A)-based sequencing techniques, as well as for simulating and exploring the regimes in which particular data types may or may not distinguish mechanistic

hypotheses [95, 98].

With meK-Means we were able to expand representation of gene-gene correlations, as compared to the independent treatment of genes in the standard CME-fitting procedure. Yet, there remains no explicit form of the regulatory interactions between genes and their products in this framework. As highlighted in Chapter 7, we can take inspiration from continuous optimization [167] and statistical inference [231] techniques for causal inference, to unpack regulation and feedback between genes through their effects on the kinetics of DNA/RNA regulation. This could be represented through learning the underlying graphs of interactions between genes in connection to their resulting kinetic rates, i.e., we learn the governing rates as a function of these regulatory graphs. To limit the combinatorial search space, prior knowledge of gene-gene interactions can be used, following common strategies in the knowledge graph literature [189].

Given the exploratory capabilities of scRNA-seq, across the technologies described above, it is also important to integrate biophysical analyses (and the accompanying statistical tools) into the experimental design loop to improve hypothesis selection and efficient use of experimental resources to gain new insight and develop new models. Previous work in fluorescence transcriptomics, has demonstrated use of Fisher information criteria to optimize experimental design and detection of environmental fluctuations [87, 143] given timecourse data. By combining such information-based analysis with our biophysical models, we can then optimize parameters, such as timepoints collected and sequencing coverage required, to recover dynamic information and disease hallmarks including dysregulation of mRNA splicing and decay. The biological insight extracted from these follow-up experiments, can then feed back into the development and update of the biophysical mechanisms and representations.

Overall, this thesis aims to highlight how experimental and computational methods can take advantage of the richness of high-throughput perturbation data, particularly through the construction of methods from first-principles, i.e., rooted in knowledge of the biological questions at hand. Though many challenges remain, to scale such approaches and integrate new measurements and biological entities, we demonstrate how question-guided investigation can bring together the domains of biology, physics, and mathematics, to take fuller advantage of the many facets of the cell we continue to uncover.

Bibliography

[1] Jill Adams and Others. Sequencing human genome: the contributions of Francis Collins and Craig Venter. *Nature Education*, 1(1):133, 2008.

[2] Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. On the surprising behavior of distance metrics in high dimensional space. In Jan Van den Bussche and Victor Vianu, editors, *Database Theory — ICDT 2001*, pages 420–434, Berlin, Heidelberg, 2001. Springer Berlin Heidelberg. ISBN 978-3-540-44503-6.

[3] Akshay Agrawal, Alnur Ali, and Stephen Boyd. Minimum-Distortion embedding. *arXiv*, March 2021.

[4] Alejandro Aguilera-Castrejon, Bernardo Oldak, Tom Shani, Nadir Ghanem, Chen Itzkovich, Sharon Slomovich, Shadi Tarazi, Jonathan Bayerl, Valeriya Chugaeva, Muneef Ayyash, Shahd Ashouokhi, Daoud Sheban, Nir Livnat, Lior Lasman, Sergey Viukov, Mirie Zerbib, Yoseph Addadi, Yoach Rais, Saifeng Cheng, Yonatan Stelzer, Hadas Keren-Shaul, Raanan Shlomo, Rada Massarwa, Noa Novershtern, Itay Maza, and Jacob H Hanna. Ex utero mouse embryogenesis from pre-gastrulation to late organogenesis. *Nature*, 593(7857):119–124, May 2021.

[5] Constantin Ahlmann-Eltze and Wolfgang Huber. Comparison of transformations for single-cell RNA-seq data. *Nature Methods*, 20(5):665–672, May 2023. ISSN 1548-7105. doi: 10.1038/s41592-023-01814-1. URL https://doi.org/10.1038/s41592-023-01814-1.

[6] Constantin Ahlmann-Eltze and Wolfgang Huber. Comparison of transformations for single-cell RNA-seq data. *Nat. Methods*, 20(5):665–672, May 2023.

[7] Sara Aibar, Carmen Bravo González-Blas, Thomas Moerman, Vân Anh Huynh-Thu, Hana Imrichova, Gert Hulselmans, Florian Rambow, Jean-Christophe Marine, Pierre Geurts, Jan Aerts, Joost van den Oord, Zeynep Kalender Atak, Jasper Wouters, and Stein Aerts. SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods*, 14(11): 1083–1086, November 2017.

[8] Alexandre F Aissa, Abul B M M K Islam, Majd M Ariss, Cammille C Go, Alexandra E Rader, Ryan D Conrardy, Alexa M Gajda, Carlota Rubio-Perez, Klara Valyi-Nagy, Mary Pasquinelli, Lawrence E Feldman, Stefan J Green, Nuria Lopez-Bigas, Maxim V Frolov, and Elizaveta V Benevolenskaya. Single-cell transcriptional changes associated with drug tolerance and response to combination therapies in cancer. *Nat. Commun.*, 12(1):1628, March 2021.

[9] Réka Albert. Network inference, analysis, and modeling in systems biology. *Plant Cell*, 19(11):3327–3338, November 2007.

[10] José Alquicira-Hernandez, Joseph E Powell, and Tri Giang Phan. No evidence that plasmablasts transdifferentiate into developing neutrophils in severe COVID-19 disease. *Clin. Transl. Immunology*, 10(7):e1308, June 2021.

[11] Shanika L Amarasinghe, Shian Su, Xueyi Dong, Luke Zappia, Matthew E Ritchie, and Quentin Gouil. Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.*, 21(1):30, February 2020.

[12] Aldine Amiel and Evelyn Houliston. Three distinct RNA localization mechanisms contribute to oocyte polarity establishment in the cnidarian clytia hemisphaerica. *Dev. Biol.*, 327(1):191–203, March 2009.

[13] Aldine Amiel, Patrick Chang, Tsuyoshi Momose, and Evelyn Houliston. Clytia hemisphaerica: a cnidarian model for studying oogenesis. *Oogenesis: the universal process. Chichester: John Wiley & Sons*, pages 81–102, 2010.

[14] Nicola Amodio, Lavinia Raimondi, Giada Juli, Maria Angelica Stamato, Daniele Caracciolo, Pierosandro Tagliaferri, and Pierfrancesco Tassone. MALAT1: a druggable long non-coding RNA for targeted anti-cancer approaches. *J. Hematol. Oncol.*, 11(1):63, May 2018.

[15] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biol.*, 11(10):R106, October 2010.

[16] Massimo Andreatta, Jesus Corria-Osorio, Sören Müller, Rafael Cubas, George Coukos, and Santiago J Carmona. Interpretation of T cell states from single-cell transcriptomics data using reference atlases. *Nat. Commun.*, 12(1):2965, May 2021.

[17] Tallulah S Andrews, Vladimir Yu Kiselev, Davis McCarthy, and Martin Hemberg. Tutorial: guidelines for the computational analysis of single-cell RNA sequencing data. *Nat. Protoc.*, 16(1):1–9, January 2021.

[18] Clio Andris, David Lee, Marcus J Hamilton, Mauro Martino, Christian E Gunning, and John Armistead Selden. The rise of partisanship and supercooperators in the U.S. House of Representatives. *PLoS One*, 10(4):e0123507, April 2015.

[19] Sindri Emmanúel Antonsson and Páll Melsted. Batch correction methods used in single cell RNA-sequencing analyses are often poorly calibrated. March 2024.

[20] Mihai Badoiu, Kedar Dhamdhere, Anupam Gupta, Yuri Rabinovich, Harald Räcke, Ramamoorthi Ravi, and Anastasios Sidiropoulos. Approximation algorithms for low-distortion embeddings into low-dimensional spaces. In *SODA*, volume 5, pages 119–128. Citeseer, 2005.

[21] Martin Balko, Attila Pór, Manfred Scheucher, Konrad Swanepoel, and Pavel Valtr. Almost-Equidistant sets. *Graphs Combin.*, 36(3):729–754, May 2020.

[22] Caroline R Bartman, Nicole Hamagami, Cheryl A Keller, Belinda Giardine, Ross C Hardison, Gerd A Blobel, and Arjun Raj. Transcriptional burst initiation and polymerase pause release are key control points of transcriptional regulation. *Mol. Cell*, 73(3):519–532.e4, February 2019.

[23] T Batu, L Fortnow, R Rubinfeld, W D Smith, and P White. Testing that distributions are close. In *Proceedings 41st Annual Symposium on Foundations of Computer Science*, pages 259–269. ieeexplore.ieee.org, November 2000.

[24] Etienne Becht, Leland McInnes, John Healy, Charles-Antoine Dutertre, Immanuel W H Kwok, Lai Guan Ng, Florent Ginhoux, and Evan W Newell. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.*, December 2018.

[25] Prashant Bhat, Amy Chow, Benjamin Emert, Olivia Ettlin, Sofia A Quinodoz, Yodai Takei, Wesley Huang, Mario R Blanco, and Mitchell Guttman. 3D genome organization around nuclear speckles drives mRNA splicing efficiency. *bioRxiv*, January 2023.

[26] Pavol Bokes, John R King, Andrew T A Wood, and Matthew Loose. Exact and approximate distributions of protein and mRNA levels in the low-copy regime of gene expression. *J. Math. Biol.*, 64(5):829–854, April 2012.

[27] Thomas C G Bosch, Alexander Klimovich, Tomislav Domazet-Lošo, Stefan Gründer, Thomas W Holstein, Gáspár Jékely, David J Miller, Andrea P Murillo-Rincon, Fabian Rentzsch, Gemma S Richards, Katja Schröder, Ulrich Technau, and Rafael Yuste. Back to the basics: Cnidarians start to fire. *Trends Neurosci.*, 40(2):92–105, February 2017.

[28] C Boudreaux, R M Coats, and B Walia. Voting and abstaining in the US senate: Mr. Downs goes to Washington. *Voting and Abstaining in the US*, 2012.

[29] Nicolas L Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Erratum: Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.*, 34(8): 888, August 2016.

[30] Timothy J Brazill and Bernard Grofman. Factor analysis versus multidimensional scaling: binary choice roll-call voting and the US Supreme Court. *Soc. Networks*, 24(3):201–229, July 2002.

[31] Brouillard, Lachapelle, Lacoste, and others. Differentiable causal discovery from interventional data. *Adv. Eng. Educ.*, 2020.

[32] David Bryant and Vincent Moulton. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Mol. Biol. Evol.*, 21(2):255–265, February 2004.

[33] David Bryant, Flavia Filimon, and Russell D Gray. Untangling our past: Languages, trees, splits and networks. *The evolution of cultural diversity: A phylogenetic approach*, pages 67–84, 2005.

[34] Daniel B Burkhardt, Jay S Stanley, 3rd, Alexander Tong, Ana Luisa Perdigoto, Scott A Gigante, Kevan C Herold, Guy Wolf, Antonio J Giraldez, David van Dijk, and Smita Krishnaswamy. Quantifying the effect of experimental perturbations at single-cell resolution. *Nat. Biotechnol.*, 39(5):619–629, May 2021.

[35] Alexander Burns, Matt Flegenheimer, Jasmine C Lee, Lisa Lerer, and Jonathan Martin. Who's running for president in 2020? *The New York Times*, January 2019.

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

[37] Adam Byerly, Tatiana Kalganova, and Ian Dear. No routing needed between capsules. *Neurocomputing*, 463:545–553, November 2021.

[38] Biao Cai, Jingfei Zhang, and Will Wei Sun. Jointly modeling and clustering tensors in high dimensions. April 2021.

[39] Long Cai, Nir Friedman, and X Sunney Xie. Stochastic protein expression in individual cells at the single molecule level. *Nature*, 440(7082):358–362, March 2006.

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

[41] Junyue Cao, Wei Zhou, Frank Steemers, Cole Trapnell, and Jay Shendure. Sci-fate characterizes the dynamics of gene expression in single cells. *Nat. Biotechnol.*, 38(8):980–988, August 2020.

[42] Zhi-Jie Cao and Ge Gao. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nat. Biotechnol.*, 40(10): 1458–1466, October 2022.

[43] Maria Carilli, Gennady Gorin, Yongin Choi, Tara Chari, and Lior Pachter. Biophysical modeling with variational autoencoders for bimodal, single-cell RNA sequencing data. *bioRxiv*, May 2023.

[44] Royce Carroll, Jeffrey B Lewis, James Lo, Keith T Poole, and Howard Rosenthal. Measuring bias and uncertainty in DW-NOMINATE ideal point estimates via the parametric bootstrap. *Polit. Anal.*, 17(3):261–275, 2009.

[45] J L Carson, A J Madonna, and others. Regulating the floor: Tabling motions in the US Senate, 1865-1946. *Am. Polit. Q.*, 2016.

[46] John T Chamberlin, Younghee Lee, Gabor T Marth, and Aaron R Quinlan. Differences in molecular sampling and data processing explain variation among single-cell and single-nucleus RNA-seq experiments. July 2023.

[47] Tara Chari and Lior Pachter. The split senate. 2021.

[48] Tara Chari and Lior Pachter. The specious art of single-cell genomics. *PLoS Comput. Biol.*, 19(8):e1011288, August 2023.

[49] Tara Chari, Brandon Weissbourd, Jase Gehring, Anna Ferraioli, Lucas Leclère, Makenna Herl, Fan Gao, Sandra Chevalier, Richard R Copley, Evelyn Houliston, David J Anderson, and Lior Pachter. Whole-animal multiplexed single-cell RNA-seq reveals transcriptional shifts across clytia medusa cell types. *Sci Adv*, 7(48):eabh1683, November 2021.

[50] Tara Chari, Gennady Gorin, and Lior Pachter. Biophysically interpretable inference of cell types from multimodal sequencing data. *bioRxiv*, September 2023.

[51] Po-Ta Chen, Benjamin Zoller, Michal Levo, and Thomas Gregor. Common bursting relationships underlie eukaryotic transcription dynamics. *ArXiv*, April 2023.

[52] Sisi Chen, Paul Rivaud, Jong H Park, Tiffany Tsou, Emeric Charles, John R Haliburton, Flavia Pichiorri, and Matt Thomson. Dissecting heterogeneous cell populations across drug and disease conditions with PopAlign. *Proc. Natl. Acad. Sci. U. S. A.*, 117(46):28784–28794, November 2020.

[53] Xiaoqiao Chen, Sisi Chen, and Matt Thomson. Minimal gene set discovery in single-cell mRNA-seq datasets with ActiveSVM. *Nature Computational Science*, 2(6):387–398, June 2022.

[54] Mathieu Chevalley, Yusuf Roohani, Arash Mehrjou, Jure Leskovec, and Patrick Schwab. CausalBench: A large-scale benchmark for network inference from single-cell perturbation data. October 2022.

[55] Francis Collins. *The Language of Life: DNA and the Revolution in Personalised Medicine*. Profile Books, December 2010.

[56] Thomas Condamine, Muriel Jager, Lucas Leclère, Corinne Blugeon, Sophie Lemoine, Richard R Copley, and Michaël Manuel. Molecular characterisation of a cellular conveyor belt in clytia medusae. *Dev. Biol.*, 456(2):212–225, December 2019.

[57] Shamus M. Cooley, Timothy Hamilton, Samuel D. Aragones, J. Christian J. Ray, and Eric J. Deeds. A novel metric reveals previously unrecognized distortion in dimensionality reduction of scrna-seq data. *bioRxiv*, 2022. doi: 10.1101/689851. URL https://www.biorxiv.org/content/early/2022/01/09/689851.

[58] Roy D Dar, Brandon S Razooky, Abhyudai Singh, Thomas V Trimeloni, James M McCollum, Chris D Cox, Michael L Simpson, and Leor S Weinberger. Transcriptional burst frequency and burst size are equally modulated across the human genome. *Proc. Natl. Acad. Sci. U. S. A.*, 109(43):17454–17459, October 2012.

[59] Charles Darwin and Leonard Kebler. *On the origin of species by means of natural selection, or, The preservation of favoured races in the struggle for life*. London: J. Murray, 1859.

[60] Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Struct. Algorithms*, 22(1):60–65, January 2003.

[61] P De Meo, E Ferrara, G Fiumara, and A Provetti. Generalized Louvain method for community detection in large networks. In *2011 11th International Conference on Intelligent Systems Design and Applications*, pages 88–93. ieeexplore.ieee.org, November 2011.

[62] Antonella Delmestri and Nello Cristianini. Linguistic phylogenetic inference by PAM-like matrices. *J. Quant. Linguist.*, 19(2):95–120, May 2012.

[63] Li Deng. The MNIST database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Process. Mag.*, 29(6): 141–142, November 2012.

[64] Elsa Denker, Eric Bapteste, Hervé Le Guyader, Michaël Manuel, and Nicolas Rabet. Horizontal gene transfer and the evolution of cnidarian stinging cells. *Curr. Biol.*, 18(18):R858–9, September 2008.

[65] Elsa Denker, Michaël Manuel, Lucas Leclère, Hervé Le Guyader, and Nicolas Rabet. Ordered progression of nematogenesis from stem cells through differentiation stages in the tentacle bulb of clytia hemisphaerica (hydrozoa, cnidaria). *Dev. Biol.*, 315(1):99–113, March 2008.

[66] Erica A K DePasquale, Daniel J Schnell, Pieter-Jan Van Camp, Íñigo Valiente-Alandí, Burns C Blaxall, H Leighton Grimes, Harinder Singh, and Nathan Salomonis. DoubletDecon: Deconvoluting doublets from Single-Cell RNA-Sequencing data. *Cell Rep.*, 29(6):1718–1727.e8, November 2019.

[67] Ravi V Desai, Xinyue Chen, Benjamin Martin, Sonali Chaturvedi, Dong Woo Hwang, Weihan Li, Chen Yu, Sheng Ding, Matt Thomson, Robert H Singer,

Robert A Coleman, Maike M K Hansen, and Leor S Weinberger. A DNA repair pathway can regulate transcriptional noise to promote cell fate transitions. *Science*, 373(6557), August 2021.

[68] Ravi V Desai, Xinyue Chen, Benjamin Martin, Sonali Chaturvedi, Dong Woo Hwang, Weihan Li, Chen Yu, Sheng Ding, Matt Thomson, Robert H Singer, Robert A Coleman, Maike M K Hansen, and Leor S Weinberger. A DNA repair pathway can regulate transcriptional noise to promote cell fate transitions. *Science*, 373(6557), August 2021.

[69] Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P Fulco, Livnat Jerby-Arnon, Nemanja D Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, Britt Adamson, Thomas M Norman, Eric S Lander, Jonathan S Weissman, Nir Friedman, and Aviv Regev. Perturb-Seq: Dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell*, 167(7):1853–1866.e17, December 2016.

[70] Silvia Domcke and Jay Shendure. A reference cell tree will serve science better than a reference cell atlas. *Cell*, 186(6):1103–1114, March 2023.

[71] Xiaoru Dong and Rhonda Bacher. Data-driven assessment of dimension reduction quality for single-cell omics data. *Patterns Prejudice*, 3(3):100465, March 2022.

[72] Michael W Dorrity, Lauren M Saunders, Christine Queitsch, Stanley Fields, and Cole Trapnell. Dimensionality reduction by UMAP to visualize physical and genetic interactions. *Nat. Commun.*, 11(1):1–6, March 2020.

[73] Jinzhuang Dou, Shaoheng Liang, Vakul Mohanty, Qi Miao, Yuefan Huang, Qingnan Liang, Xuesen Cheng, Sangbae Kim, Jongsu Choi, Yumei Li, Li Li, May Daher, Rafet Basar, Katayoun Rezvani, Rui Chen, and Ken Chen. Biorder multimodal integration of single-cell data. *Genome Biol.*, 23(1):112, May 2022.

[74] Jin-Hong Du, Ming Gao, and Jingshu Wang. Model-based trajectory inference for single-cell RNA sequencing using deep learning with a mixture prior. *bioRxiv*, 2020. doi: 10.1101/2020.12.26.424452. URL `https://www.biorxiv.org/content/early/2020/12/27/2020.12.26.424452`.

[75] Bianca Dumitrascu, Soledad Villar, Dustin G Mixon, and Barbara E Engelhardt. Optimal marker gene selection for cell type discrimination in single cell analyses. *Nat. Commun.*, 12(1):1186, February 2021.

[76] Michael Dunn, Angela Terrill, Ger Reesink, Robert A Foley, and Stephen C Levinson. Structural phylogenetics and the reconstruction of ancient language history. *Science*, 309(5743):2072–2075, September 2005.

[77] Freeman Dyson. A meeting with Enrico Fermi. *Nature*, 427(6972):297, January 2004.

[78] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, September 2012.

[79] Sacha Epskamp, Angélique O J Cramer, Lourens J Waldorp, Verena D Schmittmann, and Denny Borsboom. qgraph: Network visualizations of relationships in psychometric data. *J. Stat. Softw.*, 48:1–18, 2012.

[80] Phil Everson, Rick Valelly, Arjun Vishwanath, and Jim Wiseman. NOMINATE and American political development: A primer. *Studies in American Political Development*, 30(2):97–115, October 2016.

[81] Meichen Fang, Gennady Gorin, and Lior Pachter. Trajectory inference from single-cell genomics data with a process time model. January 2024.

[82] Catherine Felce, Gennady Gorin, and Lior Pachter. A biophysical model for ATAC-seq data analysis. January 2024.

[83] David Feldman, Luke Funk, Anna Le, Rebecca J Carlson, Michael D Leiken, Funien Tsai, Brian Soong, Avtar Singh, and Paul C Blainey. Pooled genetic perturbation screens with image-based phenotypes. *Nat. Protoc.*, 17(2):476–512, February 2022.

[84] R A Fisher. On the interpretation of χ 2 from contingency tables, and the calculation of P. *J. R. Stat. Soc.*, 85(1):87, January 1922.

[85] Arthur Flexer and Dominik Schnitzer. Choosing l^p norms in high-dimensional spaces based on hub analysis. *Neurocomputing*, 169:281–287, 2015. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2014.11.084. URL https://www.sciencedirect.com/science/article/pii/S0925231215004336. Learning for Visual Semantic Understanding in Big Data ESANN 2014 Industrial Data Processing and Analysis.

[86] Cécile Fourrage, Karl Swann, Jose Raul Gonzalez Garcia, Anthony K Campbell, and Evelyn Houliston. An endogenous green fluorescent protein–photoprotein pair in clytia hemisphaerica eggs shows co-targeting to mitochondria and efficient bioluminescence energy transfer. *Open Biol.*, 4(4):130206, 2014.

[87] Zachary R Fox, Gregor Neuert, and Brian Munsky. Optimal design of Single-Cell experiments within temporally fluctuating environments. *Complexity*, 2020, June 2020.

[88] N Friedman, L Cai, and X S Xie. Stochasticity in gene expression as observed by single-molecule experiments in live cells. *Isr. J. Chem.*, 2009.

[89] Sosuke Fujita, Erina Kuranaga, and Yu-Ichiro Nakajima. Cell proliferation controls body size growth, tentacle morphogenesis, and regeneration in hydrozoan jellyfish cladonema pacificum. *PeerJ*, 7:e7579, August 2019.

[90] Brigitte Galliot and Volker Schmid. Cnidarians as a model system for understanding evolution and regeneration. *Int. J. Dev. Biol.*, 46(1):39–48, January 2002.

[91] Adam Gayoso, Zoë Steier, Romain Lopez, Jeffrey Regier, Kristopher L Nazor, Aaron Streets, and Nir Yosef. Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nat. Methods*, 18(3):272–282, March 2021.

[92] Jase Gehring, Jong Hwee Park, Sisi Chen, Matthew Thomson, and Lior Pachter. Highly multiplexed single-cell RNA-seq by DNA oligonucleotide tagging of cellular proteins. *Nat. Biotechnol.*, 38(1):35–38, January 2020.

[93] Jacob Goldberger, Sam Roweis, Geoff Hinton, and Ruslan Salakhutdinov. Neighbourhood components analysis. In *Proceedings of the 17th International Conference on Neural Information Processing Systems*, NIPS'04, pages 513–520, Cambridge, MA, USA, December 2004. MIT Press.

[94] Gennady Gorin and Lior Pachter. Monod: mechanistic analysis of single-cell RNA sequencing count data. *bioRxiv*, 2022. doi: 10.1101/2022.06.11.495771. URL https://www.biorxiv.org/content/early/2022/06/12/2022.06.11.495771.

[95] Gennady Gorin and Lior Pachter. Length biases in single-cell RNA sequencing of pre-mRNA. *Biophys Rep (N Y)*, 3(1):100097, March 2023.

[96] Gennady Gorin, Meichen Fang, Tara Chari, and Lior Pachter. RNA velocity unraveled. *PLoS Comput. Biol.*, 18(9):e1010492, September 2022.

[97] Gennady Gorin, John J. Vastola, Meichen Fang, and Lior Pachter. Interpretable and tractable models of transcriptional noise for the rational design of single-molecule quantification experiments. *Nature Communications*, 13 (1):7620, Dec 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-34857-7. URL https://doi.org/10.1038/s41467-022-34857-7.

[98] Gennady Gorin, John J Vastola, Meichen Fang, and Lior Pachter. Interpretable and tractable models of transcriptional noise for the rational design of single-molecule quantification experiments. *Nat. Commun.*, 13(1):7620, December 2022.

[99] Gennady Gorin, Maria Carilli, Tara Chari, and Lior Pachter. Spectral neural approximations for models of transcriptional dynamics. November 2023.

[100] Gennady Gorin, John J Vastola, and Lior Pachter. Studying stochastic systems biology of the cell with single-cell genomics data. *bioRxiv*, May 2023.

[101] R Grima and P M Esmenjaud. Systematic biases in transcriptional parameters inferred from single-cell snapshot data. *bioRxiv*, 2023.

[102] Chuner Guo, Wenjun Kong, Kenji Kamimoto, Guillermo C Rivera-Gonzalez, Xue Yang, Yuhei Kirita, and Samantha A Morris. CellTag indexing: genetic barcode-based sample multiplexing for single-cell genomics. *Genome Biol.*, 20(1):90, May 2019.

[103] R Gupta and M Claassen. Factorial state-space modelling for kinetic clustering and lineage inference. August 2023.

[104] Aric Hagberg, Pieter Swart, and Daniel S Chult. Exploring network structure, dynamics, and function using networkx. Technical Report LA-UR-08-05495; LA-UR-08-5495, Los Alamos National Lab. (LANL), Los Alamos, NM (United States), January 2008.

[105] Laleh Haghverdi, Maren Büttner, F Alexander Wolf, Florian Buettner, and Fabian J Theis. Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods*, 13(10):845–848, October 2016.

[106] Laleh Haghverdi, Aaron T L Lun, Michael D Morgan, and John C Marioni. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.*, 36(5):421–427, June 2018.

[107] Jichang Han, Alexandre Gallerand, Emma C Erlich, Beth A Helmink, Iris Mair, Xin Li, Shaina R Eckhouse, Francesca M Dimou, Baddr A Shakhsheer, Hannah M Phelps, Mandy M Chan, Rachel L Mintz, Daniel D Lee, Joel D Schilling, Conor M Finlay, Judith E Allen, Claudia V Jakubzick, Kathryn J Else, Emily J Onufer, Nan Zhang, and Gwendalyn J Randolph. Human serous cavity macrophages and dendritic cells possess counterparts in the mouse with a distinct distribution between species. *Nat. Immunol.*, 25(1): 155–165, January 2024.

[108] Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M Mauck, 3rd, Shiwei Zheng, Andrew Butler, Maddie J Lee, Aaron J Wilk, Charlotte Darby, Michael Zager, Paul Hoffman, Marlon Stoeckius, Efthymia Papalexi, Eleni P Mimitou, Jaison Jain, Avi Srivastava, Tim Stuart, Lamar M Fleming, Bertrand Yeung, Angela J Rogers, Juliana M McElrath, Catherine A Blish, Raphael Gottardo, Peter Smibert, and Rahul Satija. Integrated analysis of multimodal single-cell data. *Cell*, May 2021.

[109] Cody N Heiser and Ken S Lau. A quantitative framework for evaluating Single-Cell data structure preservation by dimensionality reduction techniques. *Cell Rep.*, 31(5):107576, May 2020.

[110] Georg Hemmrich, Konstantin Khalturin, Anna-Marei Boehm, Malte Puchert, Friederike Anton-Erxleben, Jörg Wittlieb, Ulrich C Klostermeier, Philip Rosenstiel, Hans-Heinrich Oberg, Tomislav Domazet-Lošo, Toshimi Sugimoto, Hitoshi Niwa, and Thomas C G Bosch. Molecular signatures of the three stem cell lineages in hydra and the emergence of stem cell function at

the base of multicellularity. *Mol. Biol. Evol.*, 29(11):3267–3280, November 2012.

[111] Lukas Heumos, Anna C Schaar, Christopher Lance, Anastasia Litinetskaya, Felix Drost, Luke Zappia, Malte D Lücken, Daniel C Strobl, Juan Henao, Fabiola Curion, Single-cell Best Practices Consortium, Herbert B Schiller, and Fabian J Theis. Best practices for single-cell analysis across modalities. *Nat. Rev. Genet.*, 24(8):550–572, August 2023.

[112] Brian Hie, Bryan Bryson, and Bonnie Berger. Efficient integration of heterogeneous single-cell transcriptomes using scanorama. *Nat. Biotechnol.*, 37(6): 685–691, June 2019.

[113] Kristján Eldjárn Hjörleifsson, Delaney K Sullivan, Guillaume Holley, Páll Melsted, and Lior Pachter. Accurate quantification of single-nucleus and single-cell RNA-seq transcripts. December 2022.

[114] David C Hoaglin. John w. tukey and data analysis. *Stat. Sci.*, 18(3):311–318, 2003.

[115] M Hoffman, D Blei, Chong Wang, and J Paisley. Stochastic variational inference. *J. Mach. Learn. Res.*, 14:1303–1347, June 2012.

[116] Leroy Hood and Lee Rowen. The Human Genome Project: Big science transforms biology and medicine. *Genome Med.*, 5(9):79, September 2013.

[117] Robert Hooke. *Micrographia: Or Some Physiological Descriptions of Minute Bodies Made by Magnifying Glasses. With Observations and Inquiries Thereupon. By R. Hooke, Fellow of the Royal Society*. Jo. Martyn, 1667.

[118] Yuqiong Hu, Xiaoye Wang, Boqiang Hu, Yunuo Mao, Yidong Chen, Liying Yan, Jun Yong, Ji Dong, Yuan Wei, Wei Wang, Lu Wen, Jie Qiao, and Fuchou Tang. Dissecting the transcriptome landscape of the human fetal neural retina and retinal pigment epithelium by single-cell RNA-seq analysis. *PLoS Biol.*, 17(7):e3000365, July 2019.

[119] D H Huson. SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics*, 14(1):68–73, 1998.

[120] Daniel H Huson, Tobias Kloepper, and David Bryant. SplitsTree 4.0-computation of phylogenetic trees and networks. *Bioinformatics*, 14:68–73, 2008.

[121] Libbie H Hyman. OBSERVATIONS AND EXPERIMENTS ON THE PHYSIOLOGY OF MEDUSAE, 1940.

[122] Kemal Inecik, Andreas Uhlmann, Mohammad Lotfollahi, and Fabian Theis. MultiCPA: Multimodal compositional perturbation autoencoder. July 2022.

[123] Yukiko U Inoue, Junko Asami, and Takayoshi Inoue. Cadherin-6 gene regulatory patterns in the postnatal mouse brain. *Mol. Cell. Neurosci.*, 39(1): 95–104, September 2008.

[124] Detlef Jahn. Conceptualizing left and right in comparative politics: Towards a deductive approach. *Party Politics*, 17(6):745–765, November 2011.

[125] Yuge Ji, Mohammad Lotfollahi, F Alexander Wolf, and Fabian J Theis. Machine learning for perturbational single-cell omics. *Cell Syst*, 12(6):522–537, June 2021.

[126] Chen Jia and Ramon Grima. Coupling gene expression dynamics to cell size dynamics and cell cycle events: Exact and approximate solutions of the extended telegraph model. *iScience*, 26(1):105746, January 2023.

[127] Jialong Jiang, Sisi Chen, Tiffany Tsou, Christopher S McGinnis, Tahmineh Khazaei, Qin Zhu, Jong H Park, Inna-Marie Strazhnik, John Hanna, Eric D Chow, David A Sivak, Zev J Gartner, and Matt Thomson. D-SPIN constructs gene regulatory network models from multiplexed scRNA-seq data revealing organizing principles of cellular perturbation response. *bioRxiv*, May 2023.

[128] Paula Jofré, Payel Das, Jaume Bertranpetit, and Robert Foley. Cosmic phylogeny: reconstructing the chemical history of the solar neighbourhood with an evolutionary tree. *Mon. Not. R. Astron. Soc.*, 467(1):1140–1153, February 2017.

[129] William B Johnson and Joram Lindenstrauss. Extensions of lipschitz mappings into a hilbert space 26. *Contemp. Math.*, 26, 1984.

[130] Kenji Kamimoto, Blerta Stringa, Christy M. Hoffmann, Kunal Jindal, Lilianna Solnica-Krezel, and Samantha A. Morris. Dissecting cell identity via network inference and in silico gene perturbation. *Nature*, 614(7949):742–751, Feb 2023. ISSN 1476-4687. doi: 10.1038/s41586-022-05688-9. URL `https://doi.org/10.1038/s41586-022-05688-9`.

[131] Zach Kamran, Katie Zellner, Harry Kyriazes, Christine M Kraus, Jean-Baptiste Reynier, and Jocelyn E Malamy. In vivo imaging of epithelial wound healing in the cnidarian clytia hemisphaerica demonstrates early evolution of purse string and cell crawling closure mechanisms. *BMC Dev. Biol.*, 17(1): 1–14, 2017.

[132] Joyce B Kang, Aparna Nathan, Kathryn Weinand, Fan Zhang, Nghia Millard, Laurie Rumker, D Branch Moody, Ilya Korsunsky, and Soumya Raychaudhuri. Efficient and precise single-cell reference atlas mapping with symphony. *Nat. Commun.*, 12(1):5890, October 2021.

[133] Akshaya Karthikeyan and U Deva Priyakumar. Artificial intelligence: machine learning for chemical sciences. *J. Chem. Sci.*, 134(1):2, 2022.

[134] Peter V Kharchenko. The triumphs and limitations of computational methods for scRNA-seq. *Nat. Methods*, 18(7):723–732, July 2021.

[135] Dong-Wook Kim, Zizhen Yao, Lucas T Graybuck, Tae Kyung Kim, Thuc Nghi Nguyen, Kimberly A Smith, Olivia Fong, Lynn Yi, Noushin Koulena, Nico Pierson, Sheel Shah, Liching Lo, Allan-Hermann Pool, Yuki Oka, Lior Pachter, Long Cai, Bosiljka Tasic, Hongkui Zeng, and David J Anderson. Multimodal analysis of cell types in a hypothalamic node controlling social behavior. *Cell*, 179(3):713–728.e17, October 2019.

[136] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv*, December 2014.

[137] Diederik P Kingma, Tim Salimans, and M Welling. Variational dropout and the local reparameterization trick. *Adv. Neural Inf. Process. Syst.*, pages 2575–2583, June 2015.

[138] Vladimir Yu Kiselev, Tallulah S Andrews, and Martin Hemberg. Publisher correction: Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat. Rev. Genet.*, 20(5):310, May 2019.

[139] Dmitry Kobak and Philipp Berens. The art of using t-SNE for single-cell transcriptomics. *Nat. Commun.*, 10(1):5416, November 2019.

[140] Dmitry Kobak and George C Linderman. Initialization is critical for preserving global data structure in both t-SNE and UMAP. *Nat. Biotechnol.*, 39(2): 156–157, February 2021.

[141] Esmee Koedoot, Liesanne Wolters, Bob van de Water, and Sylvia E Le Dévédec. Splicing regulatory factors in breast cancer hallmarks and disease progression. *Oncotarget*, 10(57):6021–6037, October 2019.

[142] Hiroshi Kogo, Makiko Tsutsumi, Hidehito Inagaki, Tamae Ohye, Hiroshi Kiyonari, and Hiroki Kurahashi. HORMAD2 is essential for synapsis surveillance during meiotic prophase via the recruitment of ATR activity. *Genes Cells*, 17(11):897–912, November 2012.

[143] C Kreutz, J Timmer, W Dubitzky, O Wolkenhauer, and others. Optimal experiment design, fisher information. *Encyclopedia of Systems*, 2013.

[144] Miyako Kurihara-Shimomura, Tomonori Sasahira, Hiroshi Nakamura, Chie Nakashima, Hiroki Kuniyasu, and Tadaaki Kirita. Zinc finger AN1-type containing 4 is a novel marker for predicting metastasis and poor prognosis in oral squamous cell carcinoma. *J. Clin. Pathol.*, 71(5):436–441, May 2018.

[145] G La Manno, K Siletti, A Furlan, D Gyllborg, E Vinsland, and others. Molecular architecture of the developing mouse brain. *BioRxiv*, 2020.

[146] Gioele La Manno, Ruslan Soldatov, Amit Zeisel, Emelie Braun, Hannah Hochgerner, Viktor Petukhov, Katja Lidschreiber, Maria E Kastriti, Peter Lönnerberg, Alessandro Furlan, Jean Fan, Lars E Borm, Zehua Liu, David van Bruggen, Jimin Guo, Xiaoling He, Roger Barker, Erik Sundström, Gonçalo Castelo-Branco, Patrick Cramer, Igor Adameyko, Sten Linnarsson, and Peter V Kharchenko. RNA velocity of single cells. *Nature*, 560(7719):494–498, August 2018.

[147] Kasper Green Larsen and Jelani Nelson. The Johnson-Lindenstrauss Lemma is optimal for linear dimensionality reduction. *arXiv*, November 2014.

[148] Kasper Green Larsen and Jelani Nelson. Optimality of the Johnson-Lindenstrauss lemma. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 633–638. ieeexplore.ieee.org, October 2017.

[149] Anton J M Larsson, Per Johnsson, Michael Hagemann-Jensen, Leonard Hartmanis, Omid R Faridani, Björn Reinius, Åsa Segerstolpe, Chloe M Rivera, Bing Ren, and Rickard Sandberg. Genomic encoding of transcriptional burst kinetics. *Nature*, 565(7738):251–254, January 2019.

[150] Lucas Leclère and Eric Röttinger. Diversity of cnidarian muscles: Function, anatomy, development and regeneration. *Front Cell Dev Biol*, 4:157, 2016.

[151] Lucas Leclère, Muriel Jager, Carine Barreau, Patrick Chang, Hervé Le Guyader, Michaël Manuel, and Evelyn Houliston. Maternally localized germ plasm mRNAs and germ cell/stem cell formation in the cnidarian clytia. *Dev. Biol.*, 364(2):236–248, April 2012.

[152] Lucas Leclère, Richard R Copley, Tsuyoshi Momose, and Evelyn Houliston. Hydrozoan insights in animal development and evolution. *Curr. Opin. Genet. Dev.*, 39:157–167, August 2016.

[153] Lucas Leclère, Coralie Horin, Sandra Chevalier, Pascal Lapébie, Philippe Dru, Sophie Peron, Muriel Jager, Thomas Condamine, Karen Pottin, Séverine Romano, Julia Steger, Chiara Sinigaglia, Carine Barreau, Gonzalo Quiroga Artigas, Antonella Ruggiero, Cécile Fourrage, Johanna E M Kraus, Julie Poulain, Jean-Marc Aury, Patrick Wincker, Eric Quéinnec, Ulrich Technau, Michaël Manuel, Tsuyoshi Momose, Evelyn Houliston, and Richard R Copley. The genome of the jellyfish clytia hemisphaerica and the evolution of the cnidarian life-cycle. *Nature Ecology & Evolution*, 3(5):801–810, May 2019.

[154] Frances E Lee. The 115th congress and questions of party unity in a polarized era. *J. Polit.*, 80(4):1464–1473, October 2018.

[155] Dan Levy and Lior Pachter. The neighbor-net algorithm. *Adv. Appl. Math.*, 47(2):240–258, August 2011.

[156] Stan Z Li, Zelin Zang, and Lirong Wu. Deep manifold computing and visualization. *arXiv e-prints*, pages arXiv–2010, 2020.

[157] Jinping Liang, Jun Shi, Na Wang, Hui Zhao, and Jianmin Sun. Tuning the protein phosphorylation by receptor type protein tyrosine phosphatase epsilon (PTPRE) in normal and cancer cells. *J. Cancer*, 10(1):105–111, January 2019.

[158] Ying Liang, Haiyue Xu, Tao Cheng, Yujuan Fu, Hanwei Huang, Wenchang Qian, Junyan Wang, Yuenan Zhou, Pengxu Qian, Yafei Yin, Pengfei Xu, Wei Zou, and Baohui Chen. Gene activation guided by nascent RNA-bound transcription factors. *Nat. Commun.*, 13(1):7329, November 2022.

[159] Hong Seo Lim and Peng Qiu. Quantifying the clusterness and trajectoriness of single-cell RNA-seq data. *PLoS Comput. Biol.*, 20(2):e1011866, February 2024.

[160] Chieh Lin and Ziv Bar-Joseph. Continuous-state HMMs for modeling time-series single-cell RNA-Seq data. *Bioinformatics*, 35(22):4707–4715, April 2019.

[161] Xiang Lin, Tian Tian, Zhi Wei, and Hakon Hakonarson. Clustering of single-cell multi-omics data with a multimodal deep learning method. *Nat. Commun.*, 13(1):7705, December 2022.

[162] George C Linderman and Stefan Steinerberger. Clustering with t-SNE, provably. *SIAM Journal on Mathematics of Data Science*, 1(2):313–332, January 2019.

[163] Sten Linnarsson. Single-cell biology meeting marks rebirth of an old science. *Genome Biol.*, 14(4):305, April 2013.

[164] John Edensor Littlewood. *Littlewood's Miscellany*. Cambridge University Press, October 1986.

[165] John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, Barbara Foster, Mike Moser, Ellen Karasik, Bryan Gillard, Kimberley Ramsey, Susan Sullivan, Jason Bridge, Magazine, Harold, John Syron, Johnelle Fleming, Laura Siminoff, Heather Traino, Maghboeba Mosavel, Laura Barker, Scott Jewell, Dan Rohrer, Dan Maxim, Dana Filkins, Philip Harbach, Eddie Cortadillo, Bree Berghuis, Lisa Turner, Eric Hudson, Kristin Feenstra, Leslie Sobin, James Robb, Phillip Branton, Greg Korzeniewski, Charles Shive, David Tabor, Liqun Qi, Kevin Groch, Sreenath Nampally, Steve Buia, Angela Zimmerman, Anna Smith, Robin Burges, Karna Robinson, Kim Valentino, Deborah Bradbury, Mark Cosentino, Norma Diaz-Mayoral, Mary Kennedy, Theresa Engel, Penelope Williams, Kenyon Erickson, Kristin Ardlie, Wendy Winckler, Gad Getz, David DeLuca, Daniel

MacArthur, Manolis Kellis, Alexander Thomson, Taylor Young, Ellen Gelfand, Molly Donovan, Yan Meng, George Grant, Deborah Mash, Yvonne Marcus, Margaret Basile, Jun Liu, Jun Zhu, Zhidong Tu, Nancy J Cox, Dan L Nicolae, Eric R Gamazon, Hae Kyung Im, Anuar Konkashbaev, Jonathan Pritchard, Matthew Stevens, Timothèe Flutre, Xiaoquan Wen, Emmanouil T Dermitzakis, Tuuli Lappalainen, Roderic Guigo, Jean Monlong, Michael Sammeth, Daphne Koller, Alexis Battle, Sara Mostafavi, Mark McCarthy, Manual Rivas, Julian Maller, Ivan Rusyn, Andrew Nobel, Fred Wright, Andrey Shabalin, Mike Feolo, Nataliya Sharopova, Anne Sturcke, Justin Paschal, James M Anderson, Elizabeth L Wilder, Leslie K Derr, Eric D Green, Jeffery P Struewing, Gary Temple, Simona Volpi, Joy T Boyer, Elizabeth J Thomson, Mark S Guyer, Cathy Ng, Assya Abdallah, Deborah Colantuoni, Thomas R Insel, Susan E Koester, A Roger Little, Patrick K Bender, Thomas Lehner, Yin Yao, Carolyn C Compton, Jimmie B Vaught, Sherilyn Sawyer, Nicole C Lockhart, Joanne Demchok, and Helen F Moore. The Genotype-Tissue expression (GTEx) project. *Nat. Genet.*, 45(6):580–585, May 2013.

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

[167] Romain Lopez, Jan-Christian Hütter, Jonathan K Pritchard, and Aviv Regev. Large-Scale differentiable causal discovery of factor graphs. June 2022.

[168] Romain Lopez, Natasa Tagasovska, Stephen Ra, Kyunghyun Cho, Jonathan Pritchard, and Aviv Regev. Learning causal representations of single cells via sparse mechanism shift modeling. In Mihaela van der Schaar, Cheng Zhang, and Dominik Janzing, editors, *Proceedings of the Second Conference on Causal Learning and Reasoning*, volume 213 of *Proceedings of Machine Learning Research*, pages 662–691. PMLR, 2023.

[169] Mohammad Lotfollahi, F Alexander Wolf, and Fabian J Theis. scgen predicts single-cell perturbation responses. *Nat. Methods*, 16(8):715–721, August 2019.

[170] Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Leon Hetzel, Yuge Ji, Ignacio L Ibarra, Sanjay R Srivatsan, Mohsen Naghipourfar, Riza M Daza, Beth Martin, Jay Shendure, José L McFaline-Figueroa, Pierre Boyeau, F Alexander Wolf, Nafissa Yakubova, Stephan Günnemann, Cole Trapnell, David Lopez-Paz, and Fabian J Theis. Predicting cellular responses to complex perturbations in high-throughput screens. *Mol. Syst. Biol.*, page e11517, May 2023.

[171] Malte D Luecken and Fabian J Theis. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.*, 15(6):e8746, June 2019.

[172] Sai Ma, Bing Zhang, Lindsay M LaFave, Andrew S Earl, Zachary Chiang, Yan Hu, Jiarui Ding, Alison Brack, Vinay K Kartha, Tristan Tay, Travis

Law, Caleb Lareau, Ya-Chieh Hsu, Aviv Regev, and Jason D Buenrostro. Chromatin potential identified by shared Single-Cell profiling of RNA and chromatin. *Cell*, 183(4):1103–1116.e20, November 2020.

[173] James MacQueen and Others. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. books.google.com, 1967.

[174] Jiří Matoušek. On the distortion required for embedding finite metric spaces into normed spaces. *Israel J. Math.*, 93(1):333–344, December 1996.

[175] Jürgen Mayer, Khaled Khairy, and Jonathon Howard. Drawing an elephant with four complex parameters. *Am. J. Phys.*, 78(6):648–649, June 2010.

[176] William G Mayer. *The Divided Democrats: Ideological Unity, Party Reform, And Presidential Elections*. Routledge, February 2018.

[177] Chloé Mayère, Yasmine Neirijnck, Pauline Sararols, Chris M Rands, Isabelle Stévant, Françoise Kühne, Anne-Amandine Chassot, Marie-Christine Chaboissier, Emmanouil T Dermitzakis, and Serge Nef. Single-cell transcriptomics reveal temporal dynamics of critical regulators of germ cell fate during mouse sex determination. *FASEB J.*, 35(4):e21452, April 2021.

[178] James M McFarland, Brenton R Paolella, Allison Warren, Kathryn Geiger-Schuller, Tsukasa Shibue, Michael Rothberg, Olena Kuksenko, William N Colgan, Andrew Jones, Emily Chambers, Danielle Dionne, Samantha Bender, Brian M Wolpin, Mahmoud Ghandi, Itay Tirosh, Orit Rozenblatt-Rosen, Jennifer A Roth, Todd R Golub, Aviv Regev, Andrew J Aguirre, Francisca Vazquez, and Aviad Tsherniak. Multiplexed single-cell transcriptional response profiling to define cancer vulnerabilities and therapeutic mechanism of action. *Nat. Commun.*, 11(1):1–15, August 2020.

[179] Christopher S McGinnis, David M Patterson, Juliane Winkler, Daniel N Conrad, Marco Y Hein, Vasudha Srivastava, Jennifer L Hu, Lyndsay M Murrow, Jonathan S Weissman, Zena Werb, Eric D Chow, and Zev J Gartner. MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. *Nat. Methods*, 16(7):619–626, July 2019.

[180] Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv*, February 2018.

[181] Páll Melsted, A Sina Booeshaghi, Fan Gao, Eduardo Beltrame, Lambda Lu, Kristján Eldjárn Hjörleifsson, Jase Gehring, and Lior Pachter. Modular and efficient pre-processing of single-cell RNA-seq. July 2019.

[182] Samuel Melton and Sharad Ramanathan. Discovering a sparse set of pairwise discriminating features in high-dimensional data. *Bioinformatics*, 37(2):202–212, April 2021.

[183] Gregor Mendel. *Versuche über Pflanzenhybriden: Zwei abhandlungen. (1865 und 1869.)*. W. Engelmann, 1901.

[184] Radu Mihaescu, Dan Levy, and Lior Pachter. Why Neighbor-Joining works. *Algorithmica*, 54(1):1–24, May 2009.

[185] James Moody and Peter J Mucha. Portrait of political party polarization1. *Network Science*, 1(1):119–121, April 2013.

[186] Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, 5(7):621–628, July 2008.

[187] Brian Munsky and Mustafa Khammash. The finite state projection algorithm for the solution of the chemical master equation. *J. Chem. Phys.*, 124(4): 044104, January 2006.

[188] Brian Munsky, Zachary Fox, and Gregor Neuert. Integrating single-molecule experiments and discrete stochastic models to understand heterogeneous gene transcription dynamics. *Methods*, 85:12–21, September 2015.

[189] David N Nicholson and Casey S Greene. Constructing knowledge graphs and their biomedical applications. *Comput. Struct. Biotechnol. J.*, 18:1414–1428, June 2020.

[190] Damien Nicolas, Benjamin Zoller, David M Suter, and Felix Naef. Modulation of transcriptional burst frequency by histone acetylation. *Proc. Natl. Acad. Sci. U. S. A.*, 115(27):7153–7158, July 2018.

[191] Thomas M Norman, Max A Horlbeck, Joseph M Replogle, Alex Y Ge, Albert Xu, Marco Jost, Luke A Gilbert, and Jonathan S Weissman. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science*, August 2019.

[192] Vasilis Ntranos, Govinda M Kamath, Jesse M Zhang, Lior Pachter, and David N Tse. Fast and accurate single-cell RNA-seq analysis by clustering of transcript-compatibility counts. *Genome Biol.*, 17(1):112, May 2016.

[193] R Nunes Moni da Silva, A Spritzer, and C Dal Sasso Freitas. Visualization of roll call data for supporting analyses of political profiles. In *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 150–157. ieeexplore.ieee.org, October 2018.

[194] Svetlana Ovchinnikova and Simon Anders. Exploring dimension-reduced embeddings with sleepwalk. *Genome Res.*, 30(5):749–756, May 2020.

[195] Park and Park. High-Dimensional Poisson structural equation model learning via l1-regularized regression. *J. Mach. Learn. Res.*, 2019.

[196] Pablo Perez-Pinera, David G Ousterout, Jonathan M Brunger, Alicia M Farin, Katherine A Glass, Farshid Guilak, Gregory E Crawford, Alexander J Hartemink, and Charles A Gersbach. Synergistic and tunable human gene activation by combinations of synthetic transcription factors. *Nat. Methods*, 10(3):239–242, March 2013.

[197] Azam Peyvandipour, Adib Shafi, Nafiseh Saberian, and Sorin Draghici. Identification of cell types from single cell data using stable clustering. *Sci. Rep.*, 10(1):12349, July 2020.

[198] Keith T Poole and Howard Rosenthal. A spatial model for legislative roll call analysis. *Am. J. Pol. Sci.*, 29(2):357–384, 1985.

[199] Radek Pudil, Martina Vasatová, Juraj Lenco, Milos Tichý, Vít Rehácek, Alena Fucíková, Jan M Horácek, Jan Vojácek, Miloslav Pleskot, Jirí Stulík, and Vladimír Palicka. Plasma glycogen phosphorylase BB is associated with pulmonary artery wedge pressure and left ventricle mass index in patients with hypertrophic cardiomyopathy. *Clin. Chem. Lab. Med.*, 48(8):1193–1195, August 2010.

[200] Michal Rabani, Joshua Z Levin, Lin Fan, Xian Adiconis, Raktima Raychowdhury, Manuel Garber, Andreas Gnirke, Chad Nusbaum, Nir Hacohen, Nir Friedman, and Others. Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. *Nat. Biotechnol.*, 29(5):436–442, 2011.

[201] Daniel Ramsköld, Shujun Luo, Yu-Chieh Wang, Robin Li, Qiaolin Deng, Omid R Faridani, Gregory A Daniels, Irina Khrebtukova, Jeanne F Loring, Louise C Laurent, Gary P Schroth, and Rickard Sandberg. Author correction: Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.*, 38(3):374, March 2020.

[202] Elisabeth Rebboah, Fairlie Reese, Katherine Williams, Gabriela Balderrama-Gutierrez, Cassandra McGill, Diane Trout, Isaryhia Rodriguez, Heidi Liang, Barbara J Wold, and Ali Mortazavi. Mapping and modeling the genomic basis of differential RNA isoform expression at single-cell resolution with LR-Split-seq. *Genome Biol.*, 22(1):286, October 2021.

[203] Timothy E Reddy, Florencia Pauli, Rebekka O Sprouse, Norma F Neff, Kimberly M Newberry, Michael J Garabedian, and Richard M Myers. Genomic determination of the glucocorticoid response reveals unexpected mechanisms of gene regulation. *Genome Res.*, 19(12):2163–2171, December 2009.

[204] Joseph M Replogle, Thomas M Norman, Albert Xu, Jeffrey A Hussmann, Jin Chen, J Zachery Cogan, Elliott J Meer, Jessica M Terry, Daniel P Riordan, Niranjan Srinivas, Ian T Fiddes, Joseph G Arthur, Luigi J Alvarado, Katherine A Pfeiffer, Tarjei S Mikkelsen, Jonathan S Weissman, and Britt Adamson.

Combinatorial single-cell CRISPR screens by direct guide RNA capture and targeted sequencing. *Nat. Biotechnol.*, 38(8):954–961, August 2020.

[205] Miguel Reyes, Kianna Billman, Nir Hacohen, and Paul C Blainey. Simultaneous profiling of gene expression and chromatin accessibility in single cells. *Adv Biosyst*, 3(11), November 2019.

[206] Bruno Ribeiro-Gonçalves, Alexandre P Francisco, Cátia Vaz, Mário Ramirez, and João André Carriço. PHYLOViZ online: web-based tool for visualization, phylogenetic inference, analysis and sharing of minimum spanning trees. *Nucleic Acids Res.*, 44(W1):W246–51, July 2016.

[207] Joseph M Rich, Lambda Moses, Pétur Helgi Einarsson, Kayla Jackson, Laura Luebbert, A Sina Booeshaghi, Sindri Antonsson, Delaney K Sullivan, Nicolas Bray, Páll Melsted, and Lior Pachter. The impact of package selection and versioning on single-cell RNA-seq analysis. *bioRxiv*, April 2024.

[208] Markus Ringnér. What is principal component analysis? *Nat. Biotechnol.*, 26(3):303–304, March 2008.

[209] Jeffrey Rizzo and Eric C Rouchka. Review of phylogenetic tree construction. *University of Louisville Bioinformatics Laboratory Technical Report Series*, 1:1–7, 2007.

[210] Orit Rozenblatt-Rosen, Michael J T Stubbington, Aviv Regev, and Sarah A Teichmann. The human cell atlas: from vision to reality. *Nature*, 550(7677): 451–453, October 2017.

[211] Jonna Saarimäki-Vire, Annamari Alitalo, and Juha Partanen. Analysis of cdh22 expression and function in the developing mouse brain. *Dev. Dyn.*, 240(8):1989–2001, August 2011.

[212] Jose Manuel Sabucedo and Constantino Arce. Types of political participation: A multidimensional analysis. *Eur. J. Polit. Res.*, 20(1):93–102, July 1991.

[213] Basil Saeed, Anastasiya Belyaeva, Yuhao Wang, and Caroline Uhler. Anchored causal inference in the presence of measurement error. In Jonas Peters and David Sontag, editors, *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pages 619–628. PMLR, 2020.

[214] Wouter Saelens, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.*, 37 (5):547–554, May 2019.

[215] N Saitou and M Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, 4(4):406–425, July 1987.

[216] Eric M Sanford, Benjamin L Emert, Allison Coté, and Arjun Raj. Gene regulation gravitates toward either addition or multiplication when combining the effects of two signals. *Elife*, 9, December 2020.

[217] M Schena, D Shalon, R W Davis, and P O Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–470, October 1995.

[218] David Schoch and Ulrik Brandes. Legislators' roll-call voting behavior increasingly corresponds to intervals in the political spectrum. *Sci. Rep.*, 10 (1):17369, October 2020.

[219] Clarissa Scholes, Angela H DePace, and Álvaro Sánchez. Combinatorial gene regulation through kinetic control of the transcription cycle. *Cell Syst*, 4(1):97–108.e9, January 2017.

[220] Arnau Sebé-Pedrós, Baptiste Saudemont, Elad Chomsky, Flora Plessier, Marie-Pierre Mailhé, Justine Renno, Yann Loe-Mie, Aviezer Lifshitz, Zohar Mukamel, Sandrine Schmutz, Sophie Novault, Patrick R H Steinmetz, François Spitz, Amos Tanay, and Heather Marlow. Cnidarian cell type diversity and regulation revealed by Whole-Organism Single-Cell RNA-Seq. *Cell*, 173(6):1520–1534.e20, May 2018.

[221] Sheel Shah, Yodai Takei, Wen Zhou, Eric Lubeck, Jina Yun, Chee-Huat Linus Eng, Noushin Koulena, Christopher Cronin, Christoph Karp, Eric J Liaw, Mina Amin, and Long Cai. Dynamics and spatial genomics of the nascent transcriptome by intron seqFISH. *Cell*, 174(2):363–376.e16, July 2018.

[222] Junyi Shang, Xiaoju Zhang, Guangjie Hou, and Yong Qi. HMMR potential as a diagnostic and prognostic biomarker of cancer—speculation based on a pan-cancer analysis. *Frontiers in Surgery*, 9, 2023.

[223] M Sheng and M E Greenberg. The regulation and function of c-fos and other immediate early genes in the nervous system. *Neuron*, 4(4):477–485, April 1990.

[224] R Shikiar. THE PERCEPTION OF POLITICIANS AND POLITICAL ISSUES : A MULTIDIMENSIONAL SCALING APPROACH. *Multivariate Behav. Res.*, 9(4):461–477, October 1974.

[225] Stefan Siebert, Jeffrey A Farrell, Jack F Cazet, Yashodara Abeykoon, Abby S Primack, Christine E Schnitzler, and Celina E Juliano. Stem cell differentiation trajectories in hydra resolved at single-cell resolution. *Science*, 365 (6451), July 2019.

[226] Abhyudai Singh and Pavol Bokes. Consequences of mRNA transport on stochastic variability in protein levels. *Biophys. J.*, 103(5):1087–1096, September 2012.

[227] Chiara Sinigaglia, Sophie Peron, Jeanne Eichelbrenner, Sandra Chevalier, Julia Steger, Carine Barreau, Evelyn Houliston, and Lucas Leclère. Pattern regulation in a regenerating jellyfish. *Elife*, 9, September 2020.

[228] Michael A Skinnider, Jordan W Squair, and Leonard J Foster. Evaluating measures of association for single-cell transcriptomics. *Nat. Methods*, 16(5): 381–386, May 2019.

[229] James M Snyder and Tim Groseclose. Estimating party influence in congressional Roll-Call voting. *Am. J. Pol. Sci.*, 44(2):193–211, 2000.

[230] No-Joon Song, Carter Allen, Anna E Vilgelm, Brian P Riesenberg, Kevin P Weller, Kelsi Reynolds, Karthik B Chakravarthy, Amrendra Kumar, Aastha Khatiwada, Zequn Sun, Anjun Ma, Yuzhou Chang, Mohamed Yusuf, Anqi Li, Cong Zeng, John P Evans, Donna Bucci, Manuja Gunasena, Menglin Xu, Namal P M Liyanage, Chelsea Bolyard, Maria Velegraki, Shan-Lu Liu, Qin Ma, Martin Devenport, Yang Liu, Pan Zheng, Carlos D Malvestutto, Dongjun Chung, and Zihai Li. Treatment with soluble CD24 attenuates COVID-19-associated systemic immunopathology. *J. Hematol. Oncol.*, 15(1):5, January 2022.

[231] Chandler Squires, Anna Seigal, Salil Bhate, and Caroline Uhler. Linear causal disentanglement via interventions. November 2022.

[232] Chandler Squires, Annie Yun, Eshaan Nichani, Raj Agrawal, and Caroline Uhler. Causal structure discovery between clusters of nodes induced by latent factors. In Bernhard Schölkopf, Caroline Uhler, and Kun Zhang, editors, *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pages 669–687. PMLR, 2022.

[233] Sanjay R Srivatsan, José L McFaline-Figueroa, Vijay Ramani, Lauren Saunders, Junyue Cao, Jonathan Packer, Hannah A Pliner, Dana L Jackson, Riza M Daza, Lena Christiansen, Fan Zhang, Frank Steemers, Jay Shendure, and Cole Trapnell. Massively multiplex chemical transcriptomics at single-cell resolution. *Science*, 367(6473):45–51, January 2020.

[234] Patrick R H Steinmetz, Johanna E M Kraus, Claire Larroux, Jörg U Hammel, Annette Amon-Hassenzahl, Evelyn Houliston, Gert Wörheide, Michael Nickel, Bernard M Degnan, and Ulrich Technau. Independent evolution of striated muscles in cnidarians and bilaterians. *Nature*, 487(7406):231–234, July 2012.

[235] Marlon Stoeckius, Marlon Stoeckius, and Peter Smibert. CITE-seq. *Protoc. Exch.*, July 2017.

[236] Augustinas Sukys, Kaan Öcal, and Ramon Grima. Approximating solutions of the chemical master equation using neural networks. *iScience*, 25(9): 105010, September 2022.

[237] Delaney K Sullivan, Kyung Hoi Joseph Min, Kristján Eldjárn Hjörleifsson, Laura Luebbert, Guillaume Holley, Lambda Moses, Johan Gustafsson, Nicolas L Bray, Harold Pimentel, A Sina Booeshaghi, Páll Melsted, and Lior Pachter. kallisto, bustools, and kb-python for quantifying bulk, single-cell, and single-nucleus RNA-seq. *bioRxiv*, January 2024.

[238] Guoqiang Sun, Kuan Li, Jiale Ping, Liyun Zhao, Chao Cui, Junping Wu, Lixin Xie, Shuai Ma, Yan Fan, Weiqi Zhang, and Others. A single-cell transcriptomic atlas of the lungs of patients with pulmonary tuberculosis. 2024.

[239] Kartik Sunagar, Yaara Y Columbus-Shenkar, Arie Fridrich, Nadya Gutkovich, Reuven Aharoni, and Yehu Moran. Cell type-specific expression profiling unravels the development and evolution of stinging cells in sea anemone. *BMC Biol.*, 16(1):108, September 2018.

[240] David M Suter, Nacho Molina, David Gatfield, Kim Schneider, Ueli Schibler, and Felix Naef. Mammalian genes are transcribed with widely different bursting kinetics. *Science*, 332(6028):472–474, April 2011.

[241] Valentine Svensson. Droplet scRNA-seq is not zero-inflated. *Nat. Biotechnol.*, 38(2):147–150, February 2020.

[242] Valentine Svensson, Adam Gayoso, Nir Yosef, and Lior Pachter. Interpretable factor models of single-cell RNA-seq via variational autoencoders. *Bioinformatics*, 36(11):3418–3421, June 2020.

[243] Delfien Syx, Fransiska Malfait, Lut Van Laer, Jan Hellemans, Trinh Hermanns-Lê, Andy Willaert, Abdelmajid Benmansour, Anne De Paepe, and Alain Verloes. The RIN2 syndrome: a new autosomal recessive connective tissue disorder caused by deficiency of ras and rab interactor 2 (RIN2). *Hum. Genet.*, 128(1):79–88, July 2010.

[244] Peter A Szabo, Hanna Mendes Levitin, Michelle Miron, Mark E Snyder, Takashi Senda, Jinzhou Yuan, Yim Ling Cheng, Erin C Bush, Pranay Dogra, Puspa Thapa, Donna L Farber, and Peter A Sims. Single-cell transcriptomics of human T cells reveals tissue and activation signatures in health and disease. *Nat. Commun.*, 10(1):4706, October 2019.

[245] Tabula Muris Consortium, Overall coordination, Logistical coordination, Organ collection and processing, Library preparation and sequencing, Computational data analysis, Cell type annotation, Writing group, Supplemental text writing group, and Principal investigators. Single-cell transcriptomics of 20 mouse organs creates a tabula muris. *Nature*, 562(7727):367–372, October 2018.

[246] Noriyo Takeda, Yota Kon, Gonzalo Quiroga Artigas, Pascal Lapébie, Carine Barreau, Osamu Koizumi, Takeo Kishimoto, Kazunori Tachibana, Evelyn

Houliston, and Ryusaku Deguchi. Identification of jellyfish neuropeptides that act directly as oocyte maturation-inducing hormones. *Development*, 145 (2), January 2018.

[247] V A Traag, L Waltman, and N J van Eck. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.*, 9(1):5233, March 2019.

[248] Hoa Thi Nhu Tran, Kok Siong Ang, Marion Chevrier, Xiaomeng Zhang, Nicole Yee Shin Lee, Michelle Goh, and Jinmiao Chen. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol.*, 21(1):12, January 2020.

[249] Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms. *Nat. Biotechnol.*, 28 (5):511, 2010.

[250] Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J Lennon, Kenneth J Livak, Tarjei S Mikkelsen, and John L Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.*, 32(4):381–386, April 2014.

[251] John W Tukey. Exploratory data analysis as part of a larger whole. In *Proceedings of the 18th conference on design of experiments in army research and development i. Washington, dc*, volume 1010. apps.dtic.mil, 1972.

[252] John W Tukey. We need both exploratory and confirmatory. *Am. Stat.*, 34(1): 23–25, February 1980.

[253] Scott R Tyler, Supinda Bunyavanich, and Eric E Schadt. Pmd uncovers widespread cell-state erasure by scrnaseq batch correction methods. *bioRxiv*, 2021. doi: 10.1101/2021.11.15.468733. URL `https://www.biorxiv.org/content/early/2021/11/19/2021.11.15.468733`.

[254] Oana Ursu, James T Neal, Emily Shea, Pratiksha I Thakore, Livnat Jerby-Arnon, Lan Nguyen, Danielle Dionne, Celeste Diaz, Julia Bauman, Mariam Mounir Mosaad, Christian Fagre, April Lo, Maria McSharry, Andrew O Giacomelli, Seav Huong Ly, Orit Rozenblatt-Rosen, William C Hahn, Andrew J Aguirre, Alice H Berger, Aviv Regev, and Jesse S Boehm. Massively parallel phenotyping of coding variants in cancer with perturb-seq. *Nat. Biotechnol.*, 40(6):896–905, June 2022.

[255] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *J. Mach. Learn. Res.*, 9(86):2579–2605, 2008.

[256] A Wald and J Wolfowitz. On a test whether two samples are from the same population. *Ann. Math. Stat.*, 11(2):147–162, 1940.

[257] Bi-Dar Wang and Norman H Lee. Aberrant RNA splicing in cancer and drug resistance. *Cancers*, 10(11), November 2018.

[258] Yaolai Wang, Jiaming Qi, Jie Shao, and Xu-Qing Tang. Signaling mechanism of transcriptional bursting: A technical Resolution-Independent study. *Biology*, 9(10), October 2020.

[259] Ebony Rose Watson, Ariane Mora, Atefeh Taherian Fard, and Jessica Cara Mar. How does the structure of data impact cell–cell similarity? evaluating how structural properties influence the performance of proximity metrics in single cell RNA-seq data. *Brief. Bioinform.*, 23(6):bbac387, 2022.

[260] Brandon Weissbourd, Tsuyoshi Momose, Aditya Nair, Ann Kennedy, Bridgett Hunt, and David J Anderson. A genetically tractable jellyfish model for systems and evolutionary neuroscience. *Cell*, 184(24):5854–5868.e20, November 2021.

[261] Carl R Woese and George E Fox. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proc. Natl. Acad. Sci. U. S. A.*, 74(11): 5088–5090, November 1977.

[262] F Alexander Wolf, Philipp Angerer, and Fabian J Theis. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.*, 19(1):15, February 2018.

[263] F Alexander Wolf, Fiona K Hamey, Mireya Plass, Jordi Solana, Joakim S Dahlin, Berthold Göttgens, Nikolaus Rajewsky, Lukas Simon, and Fabian J Theis. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol.*, 20 (1):59, March 2019.

[264] Yuguang Xiong, Magali Soumillon, Jie Wu, Jens Hansen, Bin Hu, Johan G C van Hasselt, Gomathi Jayaraman, Ryan Lim, Mehdi Bouhaddou, Loren Ornelas, Jim Bochicchio, Lindsay Lenaeus, Jennifer Stocksdale, Jaehee Shim, Emilda Gomez, Dhruv Sareen, Clive Svendsen, Leslie M Thompson, Milind Mahajan, Ravi Iyengar, Eric A Sobie, Evren U Azeloglu, and Marc R Birtwistle. A comparison of mRNA sequencing with random primed and 3'-directed libraries. *Sci. Rep.*, 7(1):14626, November 2017.

[265] Chenling Xu, Romain Lopez, Edouard Mehlman, Jeffrey Regier, Michael I Jordan, and Nir Yosef. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Mol. Syst. Biol.*, 17 (1):e9620, January 2021.

[266] Zihan Xu, Andras Sziraki, Jasper Lee, Wei Zhou, and Junyue Cao. PerturbSci-Kinetics: Dissecting key regulators of transcriptome kinetics through scalable single-cell RNA profiling of pooled CRISPR screens. January 2023.

[267] Yang Yang, Hongjian Sun, Yu Zhang, Tiefu Zhang, Jialei Gong, Yunbo Wei, Yong-Gang Duan, Minglei Shu, Yuchen Yang, Di Wu, and Di Yu. Dimensionality reduction by UMAP reinforces sample heterogeneity analysis in bulk transcriptomic data. *Cell Rep.*, 36(4):109442, July 2021.

[268] Zizhen Yao, Hanqing Liu, Fangming Xie, Stephan Fischer, Ricky S Adkins, Andrew I Aldridge, Seth A Ament, Anna Bartlett, M Margarita Behrens, Koen Van den Berge, Darren Bertagnolli, Hector Roux de Bézieux, Tommaso Biancalani, A Sina Booeshaghi, Héctor Corrada Bravo, Tamara Casper, Carlo Colantuoni, Jonathan Crabtree, Heather Creasy, Kirsten Crichton, Megan Crow, Nick Dee, Elizabeth L Dougherty, Wayne I Doyle, Sandrine Dudoit, Rongxin Fang, Victor Felix, Olivia Fong, Michelle Giglio, Jeff Goldy, Mike Hawrylycz, Brian R Herb, Ronna Hertzano, Xiaomeng Hou, Qiwen Hu, Jayaram Kancherla, Matthew Kroll, Kanan Lathia, Yang Eric Li, Jacinta D Lucero, Chongyuan Luo, Anup Mahurkar, Delissa McMillen, Naeem M Nadaf, Joseph R Nery, Thuc Nghi Nguyen, Sheng-Yong Niu, Vasilis Ntranos, Joshua Orvis, Julia K Osteen, Thanh Pham, Antonio Pinto-Duarte, Olivier Poirion, Sebastian Preissl, Elizabeth Purdom, Christine Rimorin, Davide Risso, Angeline C Rivkin, Kimberly Smith, Kelly Street, Josef Sulc, Valentine Svensson, Michael Tieu, Amy Torkelson, Herman Tung, Eeshit Dhaval Vaishnav, Charles R Vanderburg, Cindy van Velthoven, Xinxin Wang, Owen R White, Z Josh Huang, Peter V Kharchenko, Lior Pachter, John Ngai, Aviv Regev, Bosiljka Tasic, Joshua D Welch, Jesse Gillis, Evan Z Macosko, Bing Ren, Joseph R Ecker, Hongkui Zeng, and Eran A Mukamel. A transcriptomic and epigenomic cell atlas of the mouse primary motor cortex. *Nature*, 598 (7879):103–110, October 2021.

[269] Syn Kok Yeo, Xiaoting Zhu, Takako Okamoto, Mingang Hao, Cailian Wang, Peixin Lu, Long Jason Lu, and Jun-Lin Guan. Single-cell RNA-sequencing reveals distinct patterns of cell state heterogeneity in mouse models of breast cancer. *Elife*, 9, August 2020.

[270] Yue You, Luyi Tian, Shian Su, Xueyi Dong, Jafar S Jabbari, Peter F Hickey, and Matthew E Ritchie. Benchmarking UMI-based single-cell RNA-seq preprocessing workflows. *Genome Biol.*, 22(1):339, December 2021.

[271] Hengshi Yu and Joshua D Welch. PerturbNet predicts single-cell responses to unseen chemical and genetic perturbations. July 2022.

[272] Bo Yuan, Ciyue Shen, Augustin Luna, Anil Korkut, Debora S Marks, John Ingraham, and Chris Sander. CellBox: Interpretable machine learning for perturbation biology with application to the design of cancer combination therapy. *Cell Systems*, 12(2):128–140.e4, February 2021.

[273] Luke Zappia and Fabian J Theis. Over 1000 tools reveal trends in the single-cell RNA-seq analysis landscape. *Genome Biol.*, 22(1):301, October 2021.

[274] Hongkui Zeng. What is a cell type and how to define it? *Cell*, 185(15): 2739–2755, July 2022.

[275] Meng Zhang, Stephen W. Eichhorn, Brian Zingg, Zizhen Yao, Hongkui Zeng, Hongwei Dong, and Xiaowei Zhuang. Molecular, spatial and projection diversity of neurons in primary motor cortex revealed by in situ single-cell transcriptomics. *bioRxiv*, 2020. doi: 10.1101/2020.06.04.105700. URL https://www.biorxiv.org/content/early/2020/06/05/2020.06.04.105700.

[276] Tengjiao Zhang, Yichi Xu, Kaoru Imai, Teng Fei, Guilin Wang, Bo Dong, Tianwei Yu, Yutaka Satou, Weiyang Shi, and Zhirong Bao. A single-cell analysis of the molecular lineage of chordate embryogenesis. *Sci Adv*, 6(45), November 2020.

[277] Zheng, Aragam, Ravikumar, and others. Dags with no tears: Continuous optimization for structure learning. *Adv. Neural Inf. Process. Syst.*, 2018.

[278] Shijie C Zheng, Genevieve Stein-O'Brien, Leandros Boukas, Loyal A Goff, and Kasper D Hansen. Pumping the brakes on RNA velocity – understanding and interpreting RNA velocity estimates. June 2022.

[279] Wujuan Zhong, Li Dong, Taylor B Poston, Toni Darville, Cassandra N Spracklen, Di Wu, Karen L Mohlke, Yun Li, Quefeng Li, and Xiaojing Zheng. Inferring regulatory networks from mixed observational data using directed acyclic graphs. *Front. Genet.*, 11:8, February 2020.

INDEX