TUM School of Life Sciences

Interpretable models of gene expression in single-cell immunology

David Sebastian Fischer

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

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitz: Prof. Dr. Dmitrij Frishman

Prüfer*innen der Dissertation:

1. Prof. Dr. Dr. Fabian Theis
2. Assoc. Prof. Nir Yosef, Ph.D.
3. Prof. Dr. Oliver Stegle

Die Dissertation wurde am 29.11.2021 bei der Technischen Universität München eingereicht und durch die TUM School of Life Sciences am 13.06.2022 angenommen.

Abstract

Single-cell omics data provide high-dimensional molecular characterisations of cells. These assays both extend previous efforts to understand cellular systems based on bulk assays and characterise tissue biology in a new and unbiased fashion. Accordingly, the mathematical methodology used to analyse these data sets is a mixture of traditional statistical models and strongly non-linear machine learning algorithms. These computational methods have been used to characterise a range of cellular systems, such as healthy tissues, developmental systems, and tissues in response to diseases and drugs. Among these application cases, immunology has stood out as a key application area of single-cell biology with discoveries such as immune cell states that are associated with disease outcomes. However, causal insights into the mechanisms underlying immune cell involvement in disease are often incomplete. In this thesis, I used mathematical models of single-cell omics data in T cell biology to attribute variance in cell-wise molecular states to tissue-level properties of the cellular systems, thus providing mechanistic insights into variation in single-cell data. First, I advanced current approaches to understand T cell maturation and antigen recognition by modelling T cells in the context of the population-level effects. Second, I proposed models of spatial dependencies of cells on proximal cells in their respective tissue niche. Third, I attributed heterogeneity of immune cells across samples to sample covariates. Fourth, I improved unsupervised learning approaches based on automated analysis to improve discovery of cellular phenomena. Fifth, I improved the characterisation of genetic and epigenetic variation in cells, which explains variation in measurements of different molecular views of cells, such as the transcriptome. Together, the presented approaches improved the characterisation of immune cells both individually and in the context of the tissue and, thus, advanced the state of the art in interpretable machine learning of single-cell immunology.

Zusammenfassung

Einzelzell-Omics-Experimente messen molekulare Charakterisierungen von Zellen in hoher Auflösung. Solche Experimente werden häufig benutzt um zellbiologische Theorien zu verfeinern, die auf Gewebsdurchschnittsmessungen basieren, und deshalb Phänomene auf der Größenordnung von Zellen oft nicht abbilden können. Zusätzlich bieten Einzelzellexperimente dank ihres großen Datenvolumens einen unvoreingenommenen Blickwinkel auf Zellsysteme und können deshalb benutzt werden, um neue Phänomene zu entdecken, wie beispielsweise mit Bezug auf die molekularen Zustände einzelner Zellen. Dementsprechend ist die mathematische Methodik, die zur Analyse dieser Datensätze verwendet wird, eine Mischung aus traditionellen statistischen Modellen und komplexem maschinellen Lernen. Diese Algorithmen werden benutzt, um zelluläre Systeme zu charakterisieren, von gesunden Geweben, über Systeme aus der Entwicklungsbiologie, bis hin zu Effekten von Krankheiten und Medikamenten auf Gewebe. Dies ist besonders relevant für die Einzelzellimmunologie, in der bereits eine Fülle molekularer Variationen zwischen Zellen mit Pathologien assoziiert werden konnte. Trotz dieser Datenfülle gibt es viele offene Fragen zu Krankheitsmechanismen. In meiner Dissertation habe ich mathematische Modelle für Einzelzell-Omics Daten von T-Zellen entwickelt und angewandt, um die Funktion dieser Zellen in ihrem Gewebeumfeld zu verstehen. Erstens habe ich die Modellierung der Reifung von T-Zellen und Antigenerkennung durch T-Zellen verbessert. Zweitens habe ich Modelle für räumliche Abhängigkeiten zwischen Zellen mit ihrer direkten Umgebung im Gewebe entwickelt. Drittens habe ich Variation von Immunzellen auf Charakteristiken der gemessenen Patienten bezogen. Viertens habe ich automatisierten Analysen optimiert, um die Entdeckung zellulärer Phänomene zu verbessern. Fünftens habe ich die Charakterisierung genetischer und epigenetischer Variation zwischen Zellen verbessert, um so die Variation anderer Molekülklassen der Zellen zu verstehen, wie zum Beispiel im Transkriptom. Die hier beschriebenen Algorithmen verbessern die Charakterisierung von Immunzellen und deren Interaktionen mit ihrem Gewebe und stellen neue Algorithmen des interpretierbaren maschinellen Lernens in der Einzelzellimmunolgie dar.

Acknowledgments

I would like to thank my supervisor, Prof. Fabian Theis, for academic input, freedom of project choice, professional development, and for the outstanding work environment that I experienced during my PhD at the Institute of computational biology. I would like to thank Prof. Julien Gagneur and Prof. Veit Hornung for their academic support as part of my Thesis Advisory Committee.

I would like to thank my collaborators for sharing my passion for the projects presented here and for contributing to the team efforts that lead to these manuscripts. I would like to thank Dr. Anna Sacher and Sabine Kunz for facilitating much of the administrative effort hidden in the projects presented here.

I would like to thank all members of the Machine Learning Group at the Institute of Computational Biology for collaborations, feedback and shared PhD experiences. Moreover, I would like to thank the members of the broader environment at the Institute of Computational Biology for collaborations and a rich social experience.

I would like to thank my friends from the Quantitative Biosciences Munich graduate school for all of the summer holidays, skiing trips and other activities that balanced my academic experience in Munich. In particular, I would like to mention David Brückner, Zhenya Edeleva, Alexandra Kühnlein, Joris Messelink, Kimbu Wade, and Lina Wendeler.

Ich bin meiner Familie sehr dankbar für ihre unglaubliche Unterstützung. Vielen Dank, Mama, Papa und Fabi.

Abbreviations

ATAC-seq: Assay for transposase-accessible chromatin using sequencing
CDR: Complementarity-determining region
Cas9: CRISPR-associated protein 9
cDNA: complementary DNA
CITE-seq: cellular indexing of transcriptomes and epitopes by sequencing
CRISPR: Clustered regularly interspaced short palindromic repeats
DNA: deoxyribonucleic acid
FACS: fluorescent activated cell sorting
FIM: fisher information matrix
FISH: fluorescence in situ hybridisation
IRLS: iteratively reweighted least squares
MLE: maximum likelihood estimate
MHC: major histocompatibility complex
NCEM: node-centric expression model
ODE: ordinary differential equation
PCA: principal component analysis
PCR: polymerase chain reaction
PDE: partial differential equation
pMHC: peptide-loaded MHC
RNA: ribonucleic acid
RNA-seq: RNA-sequencing
scRNA-seq: single-cell RNA-sequencing
TCR: T cell receptor
t-SNE: t-stochastic neighbour embedding
UMAP: uniform manifold approximation
UMI: unique molecular identifier

Chapter 1. General introduction

Interpretable modelling of single-cell biology is as a cell-centric approach to understanding biology (sec. 1.1), and depends on high-throughput omics measurements (sec. 1.2), and on analysis paradigms for these data modalities (sec. 1.3). Streamlining and automation across studies is an important topic in the context of single-cell data analyses and especially important with respect to interpretability in terms of generalisable covariates (sec. 1.4). Relaxing the cell-centric view, the tissue context of a cell provides a natural extension of cellular models through interpretable phenomena which explain cellular variation, such as communication events between cells (sec. 1.5). All of these topics are relevant to single-cell immunology because immune cell diversity can be characterised in individual cells but often requires the tissue context to be causally interpreted (sec. 1.6).

1.1 Cell biology

Cells are fundamental units of function based on which many biomedical phenomena can be explained. In multicellular organisms, they both represent functional entities in a tissue and are also compartmentalised in their biochemical processes because of lipid membranes. The cell is a useful building block in a bottom-up model of tissue and organism function because of its discrete nature as a unit of biological function: Phenotypes in diabetes can be understood based on pancreatic beta cell phenotypes (Bader et al. 2016; Sachs et al. 2020), certain tumour phenotypes can be understood based on cell states within the tumour microenvironment (Raza Ali et al. 2020), and inflammation in response to infection depends on specific immune cell states (Schulte-Schrepping et al. 2020). The cell's physically discrete nature makes its individual study feasible in many experimental setups: Many tissues can be dissociated into suspensions of cells that stay intact because they stay bound by the cell membrane. These suspensions allow for physical separation of cells into reaction chambers for individual measurement, such as wells or droplets in microfluidics set-ups (Macosko et al. 2015; Gierahn et al. 2017). Moreover, average molecular profiles across a single cell are often useful indicators of cellular phenotypes, even though there exist complex organelle sub-structures within cells. Thus, a surprisingly large proportion of theories in cell biology effectively treat cells as unstructured membrane-bound molecular bags which give rise to cellular function (Quake 2021): Similarity in average molecular profiles of a cell is often used as an indicator of functional similarity of cells (Eraslan et al. 2019), and average RNA expression of a gene is correlated with gene-specific phenotypes, such as cell–cell communication and receptor gene expression (Browaeys, Saelens, and Saeys 2020). Of course, sub-cellular resolution in molecular profiling is required to understand a multitude of cellular phenomena, such as organelle-specific enzyme activity differences caused by enzyme concentration differences, and the assembly of multi-protein complexes in particular cellular compartments. However, many recent advances in measurement technologies that yield average molecular profiles of single-cells have particularly accelerated progress on cell-level descriptions of cell biology. Most of these single-cell measurement technologies belong to the clade of omics measurements.

1.2 High-throughput biology: Omics

High-throughput experiments in cell biology yield complex descriptions of cellular systems, such as quantifications of changes in the transcriptome with 10,000s of detectable transcripts as a function of time, disease progression or treatment. These experiments improve our understanding of cell biological processes beyond previous state-of-the-art assays in molecular biology, which largely consisted of targeted probing of specific molecular hypotheses, such as through Western Blots or Northern Blots. Notably, RNA-seq extended the resolution of routine molecular biology experiments on the RNA level to the full transcriptome, ChIP-seq (Valouev et al. 2008) improved DNA binding assays to the full genome, DNaseI-seq and ATAC-seq (Buenrostro et al. 2013) enabled global chromatin accessibility profiling, and global chromosome organisation in terms of contact regions can now be established with Hi-C (Lieberman-Aiden et al. 2009) (Fig. 1). All of these assays depend on highly multiplexed sequencing of DNA reads in next generation sequencing (sec. 2.1.1). Since 2009 (Tang et al. 2009), many of these bulk sequencing protocols have been increasingly adapted for higher throughput in the observation dimension through single-cell protocols that measure molecular properties in individual cells rather than averages of populations of cells (bulk) only (Fig. 1).

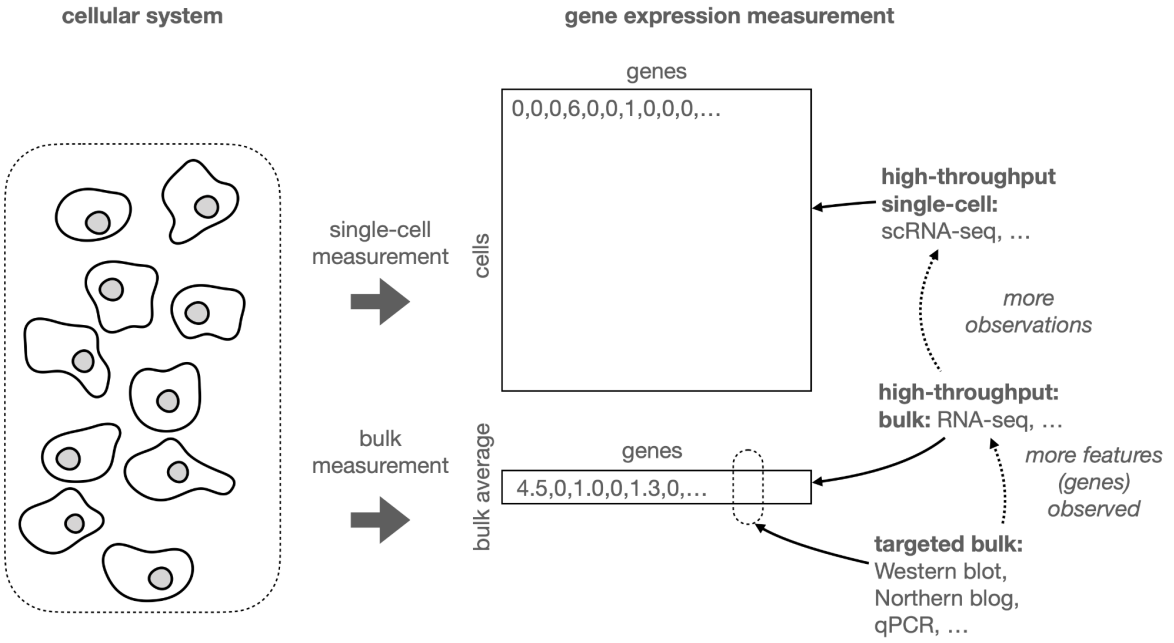

Figure 1: Molecular measurements of cells across phases of high-throughput biology. Targeted assays on bulk measurements create few data points per study: 10^0 to 10^2 data points (individual experiments). High-throughput bulk measurements create data points on the order of the assayed features, e.g. on the order of the number of genes of the assayed organism: around 10^4 data points. High-throughput single-cell measurements create data points on the order of assayed features multiplied with the number of cells that can be measured: frequently considerably more than 10^8 data points (10^4 genes x 10^4 cells).

1.3 Analysis of single-cell omics data

This drastic increase in both the observed feature space and the number of samples obtained for a cellular system drastically changed the set-up of cell biology studies and the analytic methods used to detect patterns in data, which my collaborators and I discussed in a dedicated review (Angerer et al. 2017). Traditionally, statistical models have been used to extract feature-wise patterns on bulk protocols: generalised linear models and mixed effect models have been used to analyse RNA-seq data in this paradigm (Love, Huber, and Anders 2014; Ritchie et al. 2015). Similar models centred on linear effects characterised by covariates have been used for single-cell data (Finak et al. 2015; Kharchenko, Silberstein, and Scadden 2014; Luecken and Theis 2019). However, these models are not sufficient as a stand-alone data exploration tool for data sets with many observations, as they heavily rely on *a priori* defined hypotheses about annotated components of variation in the data. In contrast, manifold learning techniques from unsupervised machine learning yield a fundamentally different view on large datasets and have proven to be better suited for data exploration in similar settings in other data-rich fields, such as in machine learning on images, speech, text and customer profiles (Fig. 2a,b). In particular, dimension reduction techniques can improve the discovery of latent structure of observations: Early on, principal component analysis, non-negative factorisation (Stein-O'Brien et al. 2019) and t-distributed stochastic neighbour embedding (Maaten and Hinton 2008) were used to explore single-cell data sets (Macosko et al. 2015). A key discovery in unsupervised learning on single-cell data was the presence of developmental trajectories in many biological systems (Trapnell et al. 2014; Haghverdi, Buettner, and Theis 2015; Haghverdi et al. 2016; Setty et al. 2016; Street et al. 2018; Wolf et al. 2019): These trajectories reconstruct sequential molecular states of a temporal process from snap-shot data and allow interpretation of a cellular process as a stereotypic developmental transition, which my collaborators and I discussed in a review (Tritschler et al. 2019). Developmental trajectories demonstrate how unsupervised characterisation of cellular data sets can improve models of cellular heterogeneity beyond what is possible based on linear models that attribute variance to sample-wise covariates. In a second stream of unsupervised analysis methods, cellular embeddings have been used to discover cellular states which are molecular configurations of cell types that are specific to certain sample conditions (Sachs et al.

2020). Third, these manifold learning techniques have enabled compositional analyses (M. Büttner et al., n.d.; Böttcher et al. 2021), which have been used to test if the distribution of cells over cell types (or states) is correlated with biologically relevant effects, such as diseases. As the complexity of the data sets increased, notions from semi-supervision and domain adaptation were increasingly used to inject prior knowledge into these unsupervised workflows: For example, semi-supervision has been used to carry over cell type annotations from pre-annotated reference data sets of the same data modality (Xu et al. 2021), and domain adaptation methods have been used to map annotation between samples (Haghverdi et al. 2018; Lopez et al. 2018; Polański et al. 2020; Lotfollahi et al. 2020). The growth of this zoo of exploratory machine learning methods has been fuelled by advances on the computational side, and by advances in single-cell measurement technology, which make new data modalities available.

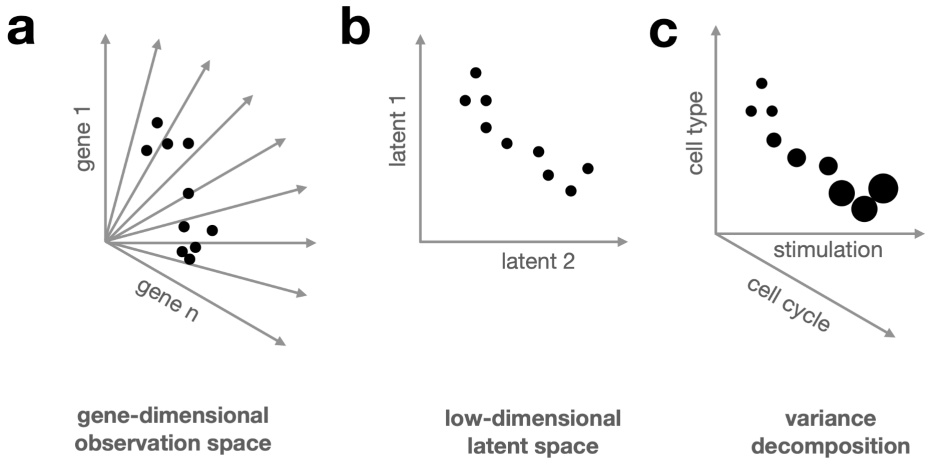

Figure 2: Learning representations of cells from molecular profiling. a) Gene expression assays, such as RNA and or protein quantification, measure cells in a gene-dimensional space if the profile is represented as a magnitude per gene, for example the number of observed unique reads mapped to a gene in scRNA-seq. **b)** This gene-dimensional space is often high-dimensional and structure in data sets can be discovered in low-dimensional projections of the data, such as from principal component analysis (PCA), t-stochastic neighbour embedding (t-SNE) (Maaten and Hinton 2008), or uniform manifold approximation (UMAP) (L. McInnes, Healy, and Melville 2018). **c)** In contrast to such embeddings dedicated to unsupervised learning, one can also project data sets to annotated axes of variation that correspond to interpretable cellular phenotypes.

Many common cell biological questions asked in the context of single-cell high-throughput assays can be reduced to flavours of variance attribution: Is the inter-conditional variation observed in a gene's mRNA distribution significant? This question is often framed as differential expression in generalised linear models (Love, Huber, and Anders 2014). Is the variance of global gene expression between two cell states large with respect to their intra-state variances? This question is often addressed in unsupervised cell type and state discovery workflows (Maren Büttner et al. 2019). Can the variance of a single-cell data set of a perturbed cellular system be partially attributed to the perturbation? Can axes of variation be annotated with known cellular or biochemical processes (Fig. 2c)? Notions of variance attribution often belong to the state-of-the art of interpretable modelling of cellular systems and give insights into relative scales of effects and highlight processes that could be targeted in therapeutic interventions. Hypotheses that are often posed in this context include: Is only a particular set of genes centred around a pathway affected by a disease? If so, molecular targets in this pathway could be explored in intervention design. Is a disease's effect limited to a few cell types? If so, those cell types could be selectively targeted. Importantly, these concepts stand in contrast to complete black-box representation learning, such as in purely predictive machine learning models. These black box models learn the full variance in the data without links to biologically meaningful labels.

1.4 Automation of single-cell analyses

The wide-spread use of high-throughput single-cell omics technologies has resulted in a rapid increase of available molecular profiles of single cells (Angerer et al. 2017). This increase in data availability both comes with a challenge of structuring available data for access and with an opportunity of increased automation of single-cell analyses through increasingly generalisable machine learning algorithms. The automation of analyses through pre-trained models requires community standards on model classes and frameworks to use and deploy model fits. Kipoi is such a framework that was suggested for the functional genomics community for DNA sequence-based models (Avsec et al. 2019). Such streamlining improves data and model re-usage, and reproducibility under the FAIR principles (Wilkinson et al. 2016), and is, therefore, also desirable in single-cell biology.

1.5 Cellular systems and tissue niches

Single-cell assays, such as scRNA-seq, single-cell ATAC-seq (Buenrostro et al. 2015) and CITE-seq (Stoeckius et al. 2017), are well suited to characterise heterogeneity of average molecular profiles of cells in a population. The discovery of molecular heterogeneity in cell atlases is core to a large fraction of single-cell biology papers published recently, such as in the context of the Human Cell Atlas (Regev et al. 2018). This molecular heterogeneity of cells is used to characterise cell types and cell states, and disease or treatment effects on gene expression (Han et al. 2018; Tabula Muris Consortium 2020; Litviňuková et al. 2020; Travaglini et al. 2020). A natural next step on biological length-scales is to leverage these cell biological insights in models of tissue biology. Tissue biology describes emergent tissue phenotypes (Fig. 3), which cannot be directly understood if cells are considered as independent building blocks of a tissue. Tissue properties emerge from cell–cell interactions and structured tissue architectures, which result in statistical dependencies between cells. Notably, cells are frequently modelled as independent and identically distributed (i.i.d.) in the single-cell modelling literature (Lopez et al. 2018; Eraslan et al. 2019), an assumption that is violated in the context of tissue biology. Thus, mathematical models of tissue biology may extend single-cell models by dependencies between cells. Cell–cell dependencies are readily motivated through biological mechanisms of tissue functions: First, tissue size is finite, thus imposing normalisation constraints on densities of cells across a molecular space. Second, cell–cell communication events give rise to specific gene expression signatures in signal receiving cells, thus creating a correlation of this signature in the receiver cell type with the presence of the sender cell type.

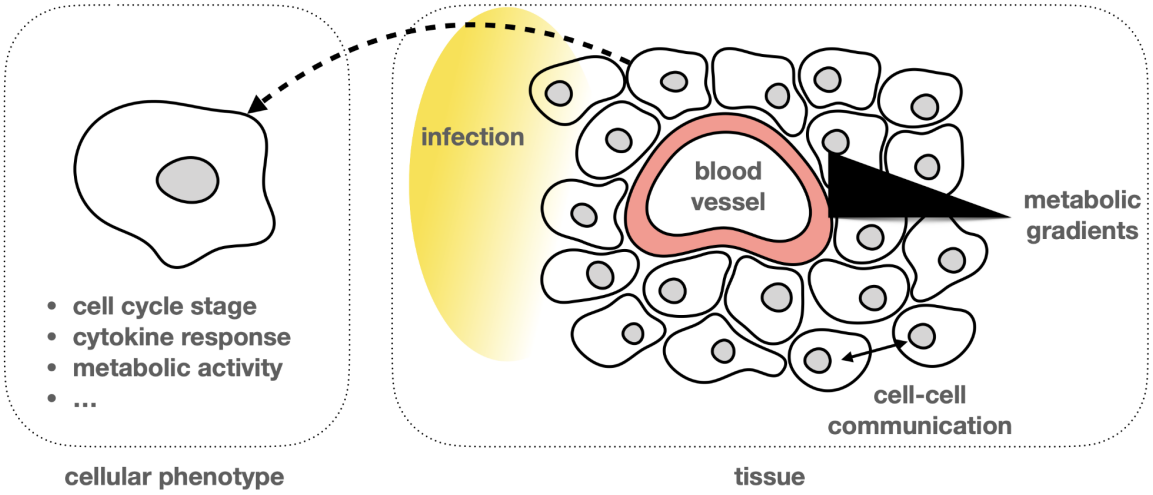

Figure 3: The tissue context is a determinant of cellular phenotypes. Tissue biological effects such as inflammation, metabolite availability, and cell–cell communication explain variation in cellular phenotypes.

Normalisation constraints on tissues are relevant for domain translation models of single-cell states. In an approximation that ignores spatial tissue architecture, one may think of a biological process that modifies a tissue as a transformation of the cell state distribution over the tissue. The transformed distribution is normalised and thus introduces conditional dependencies between domain maps of individual cells. In traditional machine learning, such domain maps can be learned with optimal transport algorithms and style-transfer through generative adversarial models, for example. In population dynamics, the transformation is learned as a parameterisation of a differential equation model. Such dynamic models usually require temporally resolved single-cell measurement unless steady-state systems are studied and specifically modelled. Inference in steady-state systems has been shown to suffer from non-identifiability (Weinreb et al. 2018), thus compromising the interpretability of this approach. In contrast to steady-state models, time-resolved models of cell populations have been fit based on discrete cell states characterised by high-throughput assays (Kafri et al. 2013; Kuritz et al. 2017), such as from flow cytometry or mass cytometry (Bandura et al. 2009). Probability distributions on a discrete state space in time are covered by an extensive literature of ordinary differential equation modelling. However, the first high-dimensional measurements of cellular development with scRNA-seq showed that many developmental processes are better approximated as continuous processes which can be captured as pseudotemporal orderings along a stereotypic developmental trajectory of a cell through molecular space (Trapnell et al. 2014; Haghverdi et al. 2016; Setty et al. 2016). In extension to such one dimensional descriptions of a developmental process, lineage branching events can be identified in dimension reductions such as t-SNE, UMAP, diffusion maps (Haghverdi, Buettner, and Theis 2015) and in a single-cell graph coarsening (Wolf et al. 2019). These continuous description of cell states in development suffer from two core shortcomings: First, multiple key properties of a cellular system, such as death and proliferation events, cannot be easily identified in such pseudotemporal orderings or diffusion maps: Apoptotic states are usually not observed because of the associated physical degradation of the cells. Proliferation events can be assigned to cell states via gene expression profiles (Macosko et al. 2015) as states with an increased fraction of cells in S, G2 or M phase (Macosko et al. 2015). However, the relative extent of proliferation to death or developmental flux cannot usually be determined from these observations. Second, dimension reductions and pseudotemoral orderings are fundamentally not able to infer the directionality of a process. Instead, directionality of lineages is usually determined via prior knowledge of the molecular identity of progenitor cell states, or, more recently, via RNA velocities (La Manno et al. 2018; Bergen et al. 2020). These two shortcomings of unsupervised analysis on molecular states, non-identifiability in direction and birth and death events, are part of the non-identifiability also encountered in dynamic models fit on steady state data (Weinreb et al. 2018). This equivalence is intuitive considering that traditional unsupervised molecular state analysis via dimension reductions is effectively unaware of time covariates. These limitations can be addressed with differential equation modelling on time series scRNA-seq data.

Cell–cell interactions have been studied as cell–cell communication based on cognate ligand and receptor gene expression in putatively interacting cell types (Browaeys, Saelens, and Saeys 2020; Efremova et al. 2020). In spatial molecular profiling data, spatial proximity between pairs of putative interacting cells can be used as a prior for cell–cell communication. Many spatial transcriptomics protocols do not give single-cell resolution but yield data comparable to small bulk samples at defined spots in the tissue (Asp, Bergenstråhle, and Lundeberg 2020). There are, however, methods with high enough resolution to distinguish individual cells in a slice of a tissue. Common protocols include approaches based on antibody-based labelling of proteins *in situ*, with advanced methods measuring 10s of proteins (Goltsev et al. 2018), imaging mass cytometry, which can measure on the order of 100 proteins *in situ* (Giesen et al. 2014), and variations of fluorescence in situ hybridisation (FISH) which can measure up to 100s to 1000s of transcript species: seq-FISH+ (Eng et al. 2019) and MERFISH (Xia et al. 2019). In all of these cases, the processed data consists of a 2D spatial grid with a third dimension of gene expression measurement. This data can be segmented into cells, and the molecular state of the cell as well as its position in the image can be recorded to give a single-cell data set that contains molecular characterisations of cells as well as the additional spatial location of the cell. Previous work on spatial transcriptomics has addressed gene function prediction in the context of ligand and receptor genes (Yuan and Bar-Joseph 2020), neighbourhood enrichment based on cell type labels (Dries et al., n.d.), and matrix factorisation techniques that distinguish intracellular and extracellular components of variation (Arnol et al. 2019). Antigen-recognition by immune cells is a specific example for cell–cell interactions in a tissue niche (sec. 1.6).

1.6 Single-cell immunology

Immunology is the study of the immune system, a multi-faceted, cross-tissue mechanism with which organisms mitigate infections and other aberrations from a healthy tissue state. Broadly, the immune system of vertebrates can be divided into an innate and an adaptive part. The innate immune system deals with broad molecular signatures of pathology and is often involved in the primary response to an infection or wound, for example. The adaptive immune system is centred on an antigen-specific response via B cells and T cells and facilitates recognition and targeted destruction of infectious agents and infected cells. Both, innate and adaptive immunity, elicit immune responses to infectious diseases, such as SARS-CoV2, but can also contribute to autoimmune diseases if dysregulated (I. B. McInnes and Gravallese 2021). Thus, innate and adaptive immunity are centrally involved in pathologies that impose a large burden on our societies. Indeed, the immune system is a key target of therapy development, with prominent examples such as immune checkpoint targeting (Wolchok 2021) and genetically modified T cells (Sterner and Sterner 2021). In adaptive immunity, antigen recognition through T cells is a key molecular event in disease mitigation and is facilitated by a receptor protein complex, the T cell receptor (TCR).

A TCR is a heterodimeric protein which consists predominantly of an alpha and a beta chain which are genetically highly diverse between T cells in a single organism. This genetic diversity stems from mutation events during cellular maturation. Both alpha and beta chains consist of C (constant), V (variable), D (diversity, only in beta chain) and J (joining) segments, which exist as multiple variants in different loci in each cell. During T cell maturation, protein-coding genes are assembled by genetic recombination between these segments. These recombination events happen sequentially for both chains (Yui and Rothenberg 2014). T cells first develop functional beta-chain during beta-selection and then develop alpha chains that yield functional heterodimers with the beta chain. The functional heterodimer is acquired in the double positive stage of maturation, which is followed by positive and negative selection. Positive selection removes cells from the population that have receptors that cannot bind to major histocompatibility complex (MHC) on antigen presenting cells. Negative selection removes cells that bind to auto-epitopes, which could otherwise trigger autoimmunity. The specificity of a TCR to an antigen is largely determined by small variable regions on both chains, named complementarity determining regions (CDRs), which obtain their variability from the genetic recombination events in cellular development. Molecularly, the unique specificity of the receptor is derived from variability in the physico-chemical properties of the TCR binding surface that is exposed to antigens bound on a MHC of an antigen presenting cell. The total number of unique TCR sequences among the full set of 10^{12} T cells in a human is estimated to lie around 10^{7} (Arstila et al. 1999). This large diversity underlies core mechanisms of adaptive immunity which rely on recognition of new pathogen-derived antigens by existing TCRs. Accordingly, the dependency of T cell phenotypes on antigen recognition events has been a longstanding focus of research in adaptive immunity.

CDR3 chains can be measured with techniques such as nucleotide sequencing in mRNA or genomic DNA assays. Recently, it has become possible to reconstruct CDR3 sequences in single-cell mRNA capture assays (De Simone, Rossetti, and Pagani 2018). These two modalities can be routinely combined with surface protein quantification through oligomer-labelled antibodies in CITE-seq (Stoeckius et al. 2017) to improve identification of T cell states (Mimitou et al. 2019). The genetic variability across T cells measured with CDR3 chain sequences is a powerful predictor of T cell phenotypes during antigen exposure. The relationship between CDR3 and antigen sequences in specific TCR-antigen pairs has been addressed in many statistical learning studies: Originally, pairs of specific TCR and antigen-loaded MHC were identified in bulk assays and collected in databases with 1000s to 10,000s of validated pairs (Bagaev et al. 2020; Shugay et al. 2018; Vita et al. 2019). Antigen-specificity has been largely addressed as a supervised learning problem in mathematical modelling studies. There are, however, also studies based on explicit protein structure modelling (Flower et al. 2010). The prediction of T cell specificity has largely been performed with models of antigen binding as a function of TCR sequence. Antigen-binding was treated both as a binary variable (Glanville et al. 2017), the presence of a binding event, or as a sequence variable, the sequence of a binding antigen (Montemurro et al. 2021). The sequence representation of antigens could allow models to generalise to unseen antigens and is, therefore, an attractive avenue for global models of antigen specificity.

Two phases along the T cell life cycle are key to understanding their function: First, T cell maturation, and second, antigen recognition through mature T cells. T cell maturation is characterised by proliferation and selection events and is often described in phases that map maturation progression to particular gene expression markers. However, this functional insight into the system of developing T cells had not yet been contextualised with scRNA-seq measurements which yield high-resolution molecular characterisation of cell states. Yet, this characterisation of the full transcriptome often helps with understanding gene regulatory circuits that underlie maturation and selection events in development. Antigen recognition through T cells happens in the context of antigen presenting cells and is usually accompanied by cytokine signalling, thus affecting cell states of many proximal cells, not only the antigen-specific ones (Fig. 4a). Antigen recognition events are often given in samples from acute-phase patients of infectious diseases after a sufficient delay between initial infection and measurement. However, the full set of cellular events that accompany antigen recognition is still not completely understood and many aspects of the cellular interactions and their impact on disease resolution, pathogen persistence and disease time course are current focus points of research. Here, single-cell omics assays can uncover components of molecular variation of T cells that are associated with antigen recognition events (Fig. 4b).

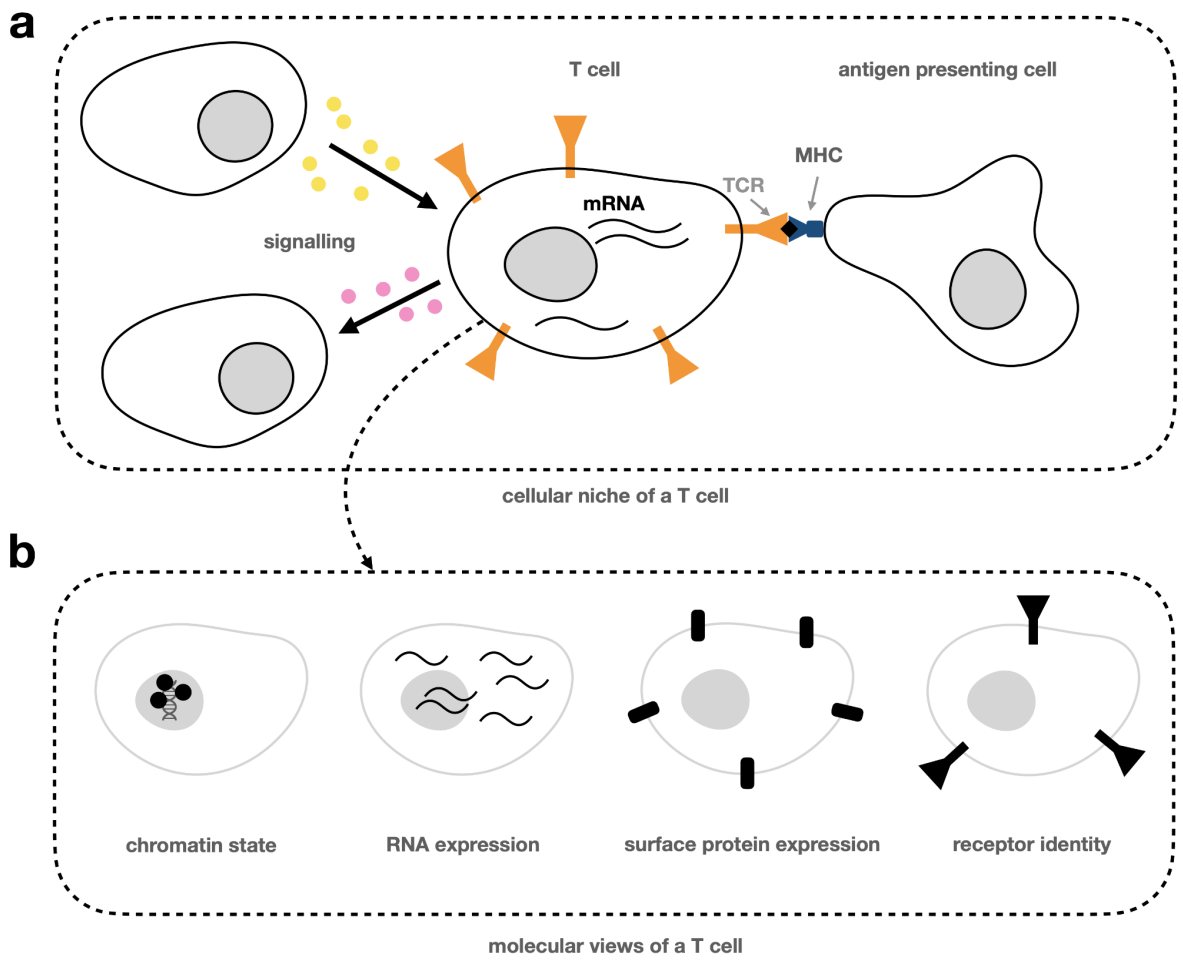

Figure 4: Multi-view measurements of T cells. a) T cells interact with their environment through signalling which is facilitated by cytokines, for example, and perform antigen recognition on antigen presenting cells using T cell receptors. **b)** The state of a single T cell can be interrogated with multiple molecular single-cell assays, including scRNA-seq identifying transient transcriptomic cell states, CITE-seq for surface protein expression quantification identifying markers used for cell state classification on FACS data, variable-region sequencing for T cell receptors identifying specificity to antigens, and scATAC-seq identifying chromatin states. Various combinations of these assays exist.

1.7 Aims of this thesis

In this thesis, I investigated the dependency of T cell states on their tissue context with new algorithmic approaches tailored to single-cell and bulk omics data. Many properties of T cells are not well understood, even though measurements of T cells are abundant and heterogenous. I considered this research question in the context of specific experimental systems that assay particular characteristics of T cells and improved the interpretability of models used for inference in these specific cases. I structured this endeavour into five project areas (Fig. 5): The first group is centred on annotating molecular heterogeneity of T cells measured with scRNA-seq with labels of tissue function, advancing beyond cell-centric descriptions of heterogeneity. The remaining project areas represent basic cell biology research that addresses conceptual bottlenecks in single-cell immunology and related disciplines: In the second part, I consider spatial dependencies of cells on their tissue niche. The third project group is centred on statistical models for variance attribution to sample covariates. The fourth project area is centred on automation and reproducibility of single-cell data analyses. In the fifth part, I propose computational methods and infrastructure for variance attribution on epigenetic features.

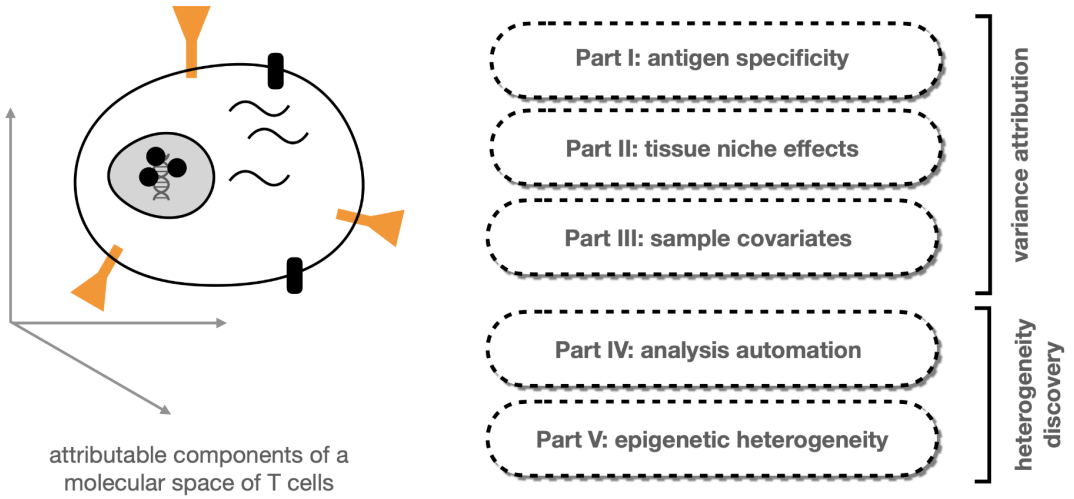

Figure 5: Discovering and attributing variance in molecular cell states of T cells. In this thesis, I developed and used methods to attribute transcriptomic variance of T cells based on antigen specificity (I), the spatial tissue niche (II), and sample or patient covariates (III). I also developed methods to improve the discovery of heterogeneity on T cells by improving analysis throughput through automation (IV), and by improving single-cell epigenetic and genetic analyses (V).

In the first part of this thesis, I integrated population-level phenotypes with high-resolution molecular profiles of individual T cells. This integration results in an annotation of cell states with functional labels derived from the tissue phenotypes. At the same time, this approach extends single-cell immunology towards tissue biological principles. Specifically, I addressed three such integration settings:

- **Publication 1 "Inferring Population Dynamics from Single-Cell RNA-Sequencing Time Series Data." (sec. 3.1, sec. 5.1):** We mapped the population-level phenotype of total population size onto a cell distribution over the molecular space observed with single-cell RNA-seq to map proliferation and selection events along T cell maturation to specific molecular states. This approach allowed us to annotate *in vivo* T cell heterogeneity in the thymus with birth- and death rates, and a developmental potential that describes directed development. We developed a model that reconciles interpretable domain translation in time with cell number normalisation constraints that are specific to growing organs, thus accounting for the tissue biology centred on total functional T cell output of this system. I was the lead author in this study.
- **Publication 2 "Predicting Antigen Specificity of Single T Cells Based on TCR CDR3 Regions." (sec. 3.2, sec. 5.1):** We proposed a method to predict antigen-specificity of single T cells from single-cell measurements to attribute *in vivo* T cell heterogeneity to antigen recognition. The proposed

method addresses the interaction between antigens presented in a tissue and T cells to which these antigens are presented. I was the lead author in this study.

- **Publication 3 "Single-cell RNA sequencing reveals ex vivo signatures of SARS-CoV-2-reactive T cells through 'reverse phenotyping'." (sec. 3.3, sec. 5.1):** We proposed a method to reduce the amount of experimental confounding during the inference of the molecular state of antigen-reactive T cells in patient samples to couple antigen recognition events more directly to *in vivo* T cell heterogeneity. Our method highlights the relevance of conditional statistical dependencies between individual immune cell's responses to infection to understand the overall system behaviour. I was the lead author in this study.
- **Additional publication 9 "Concepts and Limitations for Learning Developmental Trajectories from Single Cell Genomics." (sec. 5.2):** I contributed to a review on molecular trajectory estimation from scRNA-seq data which covers trajectory learning trends relevant for population dynamics described in (Publication 1). I was a supporting author in this study.

In the second part of the thesis, I discuss how the spatial context of a cell in a tissue can be used to relate molecular cell states to cell–cell communication and tissue niche:

- **Preprint 1 "Learning cell communication from spatial graphs of cells." (sec. 3.7, sec. 5.1):** We studied graph neural networks on spatial graphs of cells as a means to induce spatial proximity priors in cell–cell interaction inference. Among other cell biological systems, we studied the dependency of T cells on their cellular niche in solid tissues. I was the lead author in this study.
- **Additional publication 6 "Graph Representation Learning for Single Cell biology." (sec. 5.2):** We reviewed the broader context of graphs in single-cell biology, including spatial cell graphs. This study contextualises the usage of spatial proximity graphs of cells in the broader context of graph representation learning, thus connecting this project area to the machine learning and other computational biology communities. I was a second author in this study.
- **Additional publication 8 "Squidpy: A Scalable Framework for Spatial Single Cell Analysis." (sec. 5.2):** We designed a python framework for basic analyses on spatial molecular profiling data. This framework serves as a basic toolbox to access and analyse spatially-resolved single-cell data and serves as a building block for graph-based modelling projects (Preprint 1). I was a supporting author in this study.
- **Additional publication 10 "Spatial components of molecular tissue biology." (sec. 5.2):** We reviewed variance attribution concepts in spatial molecular profiling data. This study contextualises the description of cells as a function of their tissue niche in the field of single-cell genomics (Preprint 1). I was a co-first author in this study.

In the third part of this thesis, I discuss modelling approaches centred on statistical models that are related to generalised linear models. These models can extend the molecular characterisation T cells by patient covariates and thus yield mechanistic insight into cell states:

- **Publication 4 "Impulse Model-Based Differential Expression Analysis of Time Course Sequencing Data" (sec. 3.4, sec. 5.1):** Temporally-resolved gene expression profiles are often encountered in time course experiments on stimulated cell systems, such as in the model system of infectious disease or inflammation. We improved the modelling of non-linear temporal gene expression trends on bulk RNA-seq data and showed that many gene expression trajectories along stimulation processes can be characterised by simple parametric functions. I was the lead author in this study.
- **Additional publication 7 "Single-Cell Meta-Analysis of SARS-CoV-2 Entry Genes across Tissues and Demographics." (sec. 5.2):** Between patient variation results in significantly different disease outcomes for different patient groups. Methods to account for patient covariates can be used to relate cellular heterogeneity to these patient strata. I studied challenges of generalised linear models that account for patient covariates on a large cell atlas in the context of SARS-CoV-2 infection. I was a supporting author in this study.
- **Additional preprint 1 "Ultra-High Sensitivity Mass Spectrometry Quantifies Single-Cell Proteome Changes upon Perturbation." (sec. 5.3):** The description of cells on the proteomic rather than on the RNA level may yield descriptions of cells that are closer to the cell function. Mass-spectrometry-based cell measurements are fundamentally different from next-generation sequencing-based measurements

and may need to be addressed with different statistical tools. I contributed to a study of distributional characteristics of a first generation of mass spectrometry-based single-cell proteomics data in relation to single-cell RNA-seq data. The proposed methodology serves as a proof-of-concept for the analysis and measurement technology which may be extended to immunology-inspired case studies in the future. I was a supporting author in this study.

In the fourth part of this thesis, I discussed automation strategies related to exploratory single-cell analysis workflows centred around a data and model zoo.

- **Publication 5: "Sfaira Accelerates Data and Model Reuse in Single Cell Genomics." (sec. 3.5, sec. 5.1):** Automated single-cell data analyses and model benchmarks require streamlined data sets and streamlined model classes. We addressed this need with a python framework for data streamlining and computational model streamlining. Moreover, we designed a modelling approach to account for ontology-structured cell type labels in label-based models. The framework can be used for single-cell data sets from any tissue. I was the lead author in this study.
- **Additional publication 2 "Single Cells Make Big Data: New Challenges and Opportunities in Transcriptomics" (sec. 5.2):** We reviewed big data challenges in single-cell genomics. The conclusions drawn hold for the broader single-cell genomics community but also for single-cell immunology. I was a supporting author in this study.

In the fifth part of this thesis, I studies the utility of genetic and epigenetic states as descriptors of cellular heterogeneity:

- **Publication 6 "MPRAnalyze: Statistical Framework for Massively Parallel Reporter Assays." (sec. 3.6, sec. 5.1):** We extended generalised linear models to a new model class that can model coupled DNA and RNA abundance observations in massively parallel reporter assays (MPRA) which can be used to attribute transcription rates to sequence variation in regulatory elements close to the promotor. MPRA can be used to understand genetic regulation, developmental and stimulation processes, and may be useful to understand the consequences of genetic variations between patients in the future. I was a co-first author in this study.
- **Additional publication 3 "Learning Tn5 Sequence Bias from ATAC-Seq on Naked Chromatin." (sec. 5.2):** Chromatin accessibility is usually assayed through ATAC-seq in single-cell experiments, making use of the Tn5 transposase. We studied Tn5 sequence bias, which is a confounding source of variation in ATAC-seq data and, therefore, relevant to understand different chromatin accessibility across genomic sites. I was a second author in this study.
- **Additional publication 4 "EpiScanpy: Integrated Single-Cell Epigenomic Analysis." (sec. 5.2):** Epigenetic variation can be described in high-throughput in single cells on the level of chromatin accessibility and methylation, for example. We designed a python framework for single-cell epigenetic data analyses. Such epigenetic characteristics will be valuable to describe long-lived chromatin states of immune cells, for example. I was a second author in this study.

Chapter 2. General methods

This methods section is divided into experimental and computational methods. The experimental methods describe molecular principles underlying the datasets discussed in this thesis. The computational methods describe analytic concepts used to model data in this thesis.

2.1 Cell biological methods

The vast majority of datasets discussed in this thesis rely on a next-generation sequencing read-out which is described in (sec. 2.1.1). These assays differ in their nucleic acid capture strategies and are, for example, specific to single-cell capture (sec. 2.1.2, 2.1.3) or plasmid and transcript co-capture (sec. 2.1.4).

2.1.1 Next-generation sequencing

Next-generation sequencing is a term specific to nucleotide sequencing and presents a ground-breaking advance over previous low-throughput sequencing methods (Goodwin, McPherson, and McCombie 2016), such as Sanger sequencing (Sanger, Nicklen, and Coulson 1977). In brief, next-generation sequencing protocols allow for a massively parallelised characterisation of short DNA sequences in terms of their base sequence. This parallel setup allows for many short reads to be handled in a single experiment. The sequencing process is based on one of two core physicochemical mechanisms: First, nucleotides can be recognised via fluorescent signals in a cyclic strand extension process based on reversibly blocked and fluorescently-labelled nucleotides. This technique is employed in clonal bridge amplification which is used commonly in sequencing. Second, nucleotides can be recognised based on their electrical properties in a nanopore which are assayed in Oxford nanopore sequencing (Jain et al. 2016). Most of the protocols discussed here produce short DNA reads that are usually sequenced in clonal bridge amplification machines to yield reads that are then aligned against a reference transcriptome or genome to characterise the origin of the DNA sequence.

2.1.2 Single-cell RNA-sequencing

Single-cell RNA-sequencing (scRNA-seq) is the measurement of abundances of RNA species in individual cells. Protocols can be largely distinguished based on two core characteristics: the method with which cells are compartmentalised in distinct reaction chambers and the RNA capture mechanism (Fig. 6). Early scRNA-seq protocols largely relied on well-based separation of cells which were either based on manual or robotic pipetting of individual cells into wells (Tang et al. 2009; Ramsköld et al. 2012). Here, the wells could be treated in parallel with reaction cocktails thereby increasing the throughput in comparison to a naive set-up in which each cell is handled as an entirely individual experiment. These early scRNA-seq protocols were enabled by the biochemical optimisation of reverse transcription that enabled scRNA-seq on the low input RNA amounts presented by cells (Tang et al. 2009). These complementary DNA (cDNA) generation methods classify as “full length” protocols which yield reads across the entire transcript and stand in contrast to 3’- and 5’-capture protocols which yield read spectra heavily biased to one end of the transcript. Full-length protocols can be used to study transcript diversity, such as splicing variants, but cannot be controlled by unique molecular identifiers (UMIs) and are not strictly necessary to simply count RNA molecules in a sample. Drop-seq (Macosko et al. 2015) and inDrop (Macosko et al. 2015; Zilionis et al. 2017) introduced the usage of microfluidics to separate cells into droplets in a reverse emulsion in oil, using droplets as reaction chambers. These microfluidics protocols were later commercialised and translated to other sequencing-based assays in single-cells. Microfluidics-based set-ups facilitated an increase in throughput from 100s of cells per experiment to 10,000s per experiment and up to 1,000,000s cells per study. Microwell-based separation protocols constitute a third pillar of physical separation of cells (Gierahn et al. 2017; Han et al. 2018). These methods have comparable throughput to microfluidics methods because they do not require targeted placement of cells into wells but rely on stochastic separation of cells over many small wells. Both microfluidics and micro-well based separation of cells rely on transcript capture of lysed cells in a reaction volume by a bead that transfers bead- and molecule specific barcodes to a cDNA upon reverse transcription in the reaction volume. These bead- and molecule-specific barcodes allow both demultiplexing of the RNA-seq results into the individual cells (beads) and deduplication of reads originating from a single cDNA via molecular barcodes, thus improving accuracy of transcript count estimates. Full length

RNA-seq protocols cannot be trivially performed in these parallelised reaction volume setups because the demultiplexing depends on the transfer of barcodes from beads to cDNAs, and barcodes are carried by beads that transfer barcodes only at the time of transcript capture, and, therefore, once per transcript. This difference to well-based setups is one of the core reasons for the parallel evolutions of full length protocols, especially SMART-seq1-4 (Ramsköld et al. 2012; Picelli et al. 2013; Hagemann-Jensen et al. 2020), and bead-based transcript-end capture assays. In yet another approach to transcript capture and barcoding, sci-RNA-seq replaces the physical isolation of cells or nuclei by combinatorial indexing of reads in permeabilised cells or nuclei (Cao et al. 2017). Finally, spatial transcriptomics approaches extend this paradigm of measuring cells in dissociated tissues to measuring cells *in situ* (A. Rao et al. 2021).

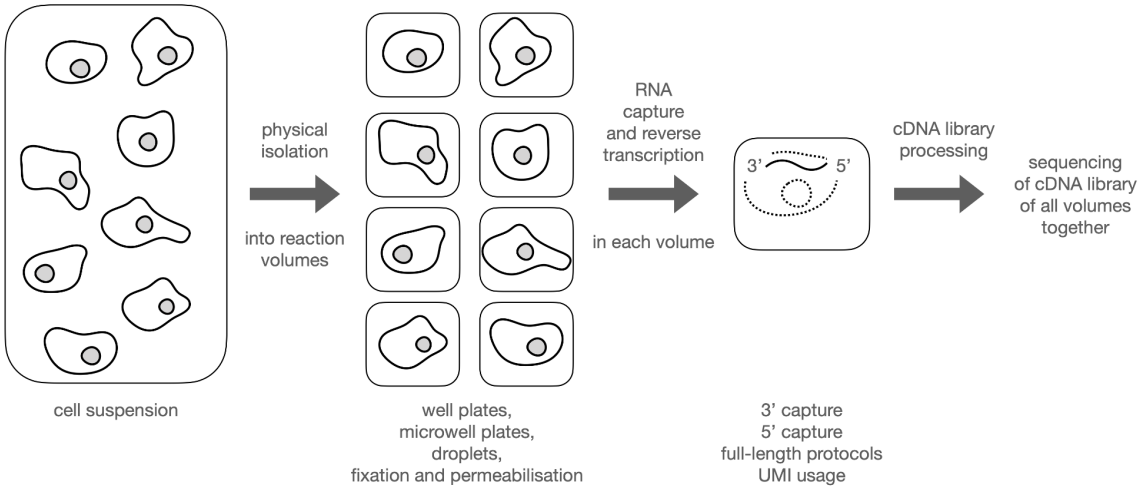

Fig. 6: Technical variations in scRNA-seq protocols. Protocols vary in the isolation mechanism with which cells are separated into physically distinct reaction volumes and in the RNA capture mechanism within these volumes. Further differences occur in the chemical treatment of each reaction volume and in cDNA library processing.

It is worth noting that while bead-based transcript capture assays are designed for transcript counting, they do resolve splicing state differences in some cases even though their strong read bias to one end of the transcript. This splicing state observation gave rise to the field of RNA velocities (La Manno et al. 2018; Bergen et al. 2020).

2.1.3 Single-cell ATAC-seq

Assay for transposase-accessible chromatin using sequencing (Buenrostro et al. 2013) (ATAC-seq) is a transposase-base chromatin profiling method and provides an alternative to DNase-seq (Song and Crawford 2010) for this purpose. DNase-seq is a digestion-based chromatin profiling method in which DNA is first digested with DNaseI, DNA strand breaks are then ligated to primers and the library then sequenced to identify the digestion sites. The accumulation of such sites in hypersensitive regions provides a measure for chromatin accessibility, such as the absence of nucleosomes. In ATAC-seq, this notion was leveraged in an assay that measures hyperactive Tn5 transposase integration instead of enzymatic digestion. This transposase is able to cleave DNA and ligate primer to the DNA strand breaks in a single-step, “tagmentation” (Adey et al. 2010). ATAC-seq can be used to assess chromatin accessibility, transcription factor footprinting and nucleosome positioning. Single-cell ATAC-seq was achieved via microfluidics-based single-cell compartmentalisation (Buenrostro et al. 2013). To date, scATAC-seq is the only single-cell open chromatin profiling method and the only epigenetic method which is routinely used in molecular characterisation of single-cells.

2.1.4 Massively parallel reporter assays

Massively parallel reporter assays (Inoue and Ahituv 2015) (MPRA) were developed to quantify the effect of sequence variation in regulatory DNA sequences on gene expression. Often, one is constrained to rely on naturally occurring genetic variability to study such variation. The MPRA *in vitro* system allows for targeted perturbations of the dependence of gene activation on these genetic variants in high-throughput: First, a reporter

gene with a minimal promoter is cloned into plasmid library with barcoded enhancer elements with variable regulatory sequences such that the barcode is included into the reporter gene transcript. The sequencing diversity in the resulting plasmid library is assayed via bulk DNA-seq and an aliquot of the library is transfected into a cell culture which is then assayed in bulk via RNA-seq. One can then use these two bulk DNA and RNA abundance measurements of barcodes to assess the relative abundance differences of barcodes between DNA and RNA measurements, which corresponds to the effect of the regulatory sequence on transcription in this controlled system. The barcodes can be coupled to explicit regulatory sequences via DNA sequencing to give mechanistic insights into differential transcription regulation.

2.2 Computational methods

Inference in mathematical models depends on parameter estimation and cost functions which are very general concepts that underpin much of computational biology. Differential equation models provide many unique challenges to inference which I discussed in a separate section. Finally, uncertainty estimation and the design of priors are frequently encountered in mechanistic models that are common in computational biology.

2.2.1 Parameter estimation in statistical learning

Parameter estimation is the pursuit of finding parameter estimates for a given model such that a cost function is minimised. Usually, the cost function contains regularisation terms and a deviation measure between model predictions and observed data. A popular theoretical framework for parameter estimation in frequentist statistics is maximum likelihood estimation, in which the objective function is directly optimised. In other frameworks, such as variational inference (Blei, Kucukelbir, and McAuliffe 2017), the objective function cannot always be directly evaluated and, therefore, can also not directly be optimised and a lower bound to the objective is used instead. Both direct and lower-bound-based techniques use gradients of a cost function with respect to the model parameters to relay the deviation of the model prediction from the training data to the parameter updates. This paradigm of gradient-based optimisation requires continuously differentiable models.

In maximum likelihood estimation methods (Hastie, Tibshirani, and Friedman 2013), the cost function is a likelihood function and quantifies deviation of model estimates to the observed data, subject to assumptions on the distribution of residuals. Depending on the model, one can either derive the maximum likelihood estimates (MLE) as an analytic expression of the extremum of the likelihood function or obtain local MLEs through iterative parameter updates. If the likelihood function is convex, these local MLEs are also global MLEs. The following example illustrates these differences: A linear model, a Poisson generalised linear model (Agresti 2015) and a non-linear multi-layer perceptron (Hastie, Tibshirani, and Friedman 2013) can all be fit to a given data set using maximum likelihood estimation. The linear model and the generalised linear model both are convex optimisation problems, the multi-layer perceptron is not necessarily convex. Additionally, we can derive an analytic description for the MLE of the linear model (eq. M.1), the least squares estimator (eq. M.2), whereas this is not possible for the Poisson generalised linear model (Agresti 2015):

$$y = X\theta \quad (M.1)$$

$$\hat{\theta} = (X^T X)^{-1} X^T y \quad (M.2)$$

Here, θ are parameters of the model, X the model inputs, and y the labels. Accordingly, both generalised linear model and multi-layer perceptron have to be fitted with iterative parameter updates. As the cost function of the generalised linear model is convex, the local MLE can be obtained through iterative optimisation is also the global MLE and insensitive to initialisation. In contrast to this, the optimisation of multi-layer perceptron is sensitive to initialisation, a problem that is subject to many practical considerations concerning neural network training, such as weight initialisation and training schedules (Glorot and Bengio 2010; Klambauer et al. 2017). Additionally, generalised linear models can also be optimised more efficiently than multi-layer perceptrons because of parameter updates that exploit a quadratic approximation of the likelihood function. Such updates from the class of Newton-Rhapson or quasi-Newton methods require the full hessian matrix or an approximation thereof (Broyden 1970; Fletcher 1970; Goldfarb 1970; Shanno 1970). The computational complexity of hessians is quadratic in the number of parameters and therefore usually infeasible for neural networks with many parameters. If estimation using Newton-Rhapson or quasi-Newton type methods becomes infeasible, one usually resorts to methods that only require a linear local approximation of the cost function through the Jacobian. Again,

this Jacobian may be computed in closed form or with finite differences. A core strength of deep learning software frameworks like tensorflow (“TensorFlow” n.d.) and pytorch (“PyTorch” n.d.) is that they use auto-differentiation to define the analytic Jacobian of a model based on the forward pass described in code. This auto-differentiation accelerates modelling cycles drastically and lies at the heart of the revolution of deep learning models. Standard gradient descent optimisers have been developed to more advanced optimisers such as ADAM (Kingma and Ba 2014), that also have linear complexity in the number of parameters of the system but improve convergence compared to gradient descent better on rugged loss functions with saddle points that are frequently encountered in neural networks.

Gradient-based learning on large data sets imposes the additional challenge that the complexity of gradient evaluations is linear in the number of observations, which leads to excessively large parameter update times for many applications. Here, stochastic optimisation is often used to reduce the number of gradient evaluations per parameter update without biasing the resulting parameter estimators (Hastie, Tibshirani, and Friedman 2013). In stochastic optimisation, a compromise between the number of updates until convergence and the time required to compute a single update is defined. Importantly, with large data sets, gradient evaluations across observations can be sufficiently correlated such that stochastic methods yield overall faster convergence than deterministic optimisation methods. Additionally, parallelised gradient evaluation is often not possible in memory for large models with many parameters which results in further computational advantages of stochastic optimisers. In deep learning, mini-batched stochastic optimisation is a common form of stochastic optimisation with a range of 10s to 100s of observations used for each gradient evaluation.

2.2.2 Cost functions for optimisation

Cost functions for optimisation quantify deviation between model predictions and observed data and thus give rise to parameter optimisation problems. The exact weighting of deviations depends on the prior knowledge on the data modelled. Classification and regression cost functions are umbrella terms that are used for a wide range of maximum likelihood problems for models that predict either categorical or continuous outcomes (Hastie, Tibshirani, and Friedman 2013).

2.2.3.1 Classification problems

A classification problem is characterised by a one-hot encoded label that represents a probability mass distribution over a categorical label space (Hastie, Tibshirani, and Friedman 2013). Classification models typically predict a probability mass distribution over this label space. In the case of a log multinomial likelihood with one trial per observation, the cross entropy cost function is the negative log likelihood function (eq. M.3):

$$\log L(\theta; x) = \sum_{n}^{N} \log p_{nk} \quad (M.3)$$

This log likelihood over N observation and K label classes, where p_{nk} is the predicted probability mass for observation the true class k of observation n. I used such a cross entropy cost function for categorical antigen specificity prediction (Publication 2) and basic cell type prediction (Publication 5), for example. In sfaira (Publication 5), we adapted the cross-entropy cost function to a label space in which labels can be related to each other through a directed acyclic graph, a cell type ontology (eq. M.4):

$$\log L(\theta; x) = \sum_{n}^{N} \log(\sum_{k \in K^+} p_{nk}) \quad (M.4)$$

This likelihood function is designed for a model that predicts a probability mass distribution over leaf nodes of a directed acyclic graph provided by an ontology. This log likelihood over N observation, is based on the sum of probability mass values assigned to any of the labelled classes K^+. The set of labelled classes is the set of leaf nodes that correspond to a particular observed node in the directed acyclic graph. Therefore, a model fit for a given observation is evaluated based on the predicted probability mass of all admissible leaf nodes that match the observations. If all observations are assigned to a single class, which is equivalent to saying that all observations are assigned to leaf nodes of the ontology, the likelihood function becomes a standard cross-entropy function of a multi-class classification problem.

Classification models can be evaluated with metrics that derive from a confusion matrix of the prediction problem at a defined classification score threshold, such as accuracy or F1 score, or score-independent metrics, such as the area under the curve of the receiver operator characteristic curve.

2.2.3.1 Regression problems

The categorical antigen specificity prediction problem on labels derived from peptide-MHC counts can be extended to the full count labels on specificity (Publication 2). The new labels are on a positive support and thus require cost functions that deal with count structured (Agresti 2015) or continuous real numbered labels. Such supervision problems can be addressed with regression cost functions. In the case of peptide-MHC counts, we chose a mean squared error on log-transformed count predictions to mitigate count data-induced heteroscedasticity in the data (eq. M.5):

$$\log L(\theta) = \sum_{n}^{N} (\log f(x_n, \theta) - \log y_n)^2 \quad (M.5)$$

Shown is the mean squared log error of a model *f* that predicts observation *y* in the positive domain based on the input *x* and the parameters θ.

In next-generation sequencing modelling scenarios, one is often confronted with reconstruction cost functions which are special cases of regression cost functions that make use of domain-specific noise models (Lopez et al. 2018; Eraslan et al. 2019). A noise model is a probability density or probability mass function that describes an observation distribution, often based on a few distributional moments that can be parameterised. Reconstructing models cover scenarios that predict cell-wise gene expression states based on cellular covariates in supervised models, or cell embeddings in latent variable models, for example. Many commonly used models parametrise a location and a scale term and use exponential family probability distributions.

A core modelling framework for the statistical analysis of next-generation sequencing data is the generalised linear model (Ritchie et al. 2015; Love, Huber, and Anders 2014): generalised linear models decompose the data variance into different effects encoded in covariates. Generalised linear models are commonly defined to be based on exponential family noise models (Hastie, Tibshirani, and Friedman 2013; Agresti 2015), for which parameters can be efficiently estimated by using iterative updates that directly optimise the likelihood with an algorithm called iteratively re-weighted least squares (IRLS) (Hastie, Tibshirani, and Friedman 2013). Exponential family probability distributions include functions such as normal, log-normal, beta and Poisson distributions, and are characterised by having a density that can be rewritten in the format shown in (eq. M.6):

$$f_\theta(y) = \exp(\frac{y\theta - b(\theta)}{\phi} + c(y, \theta)) \quad (M.6)$$

The probability density (or mass) function of a canonical exponential family distribution can be rewritten into the above canonical form. θ are parameters of the model, y the model inputs, ϕ a dispersion term relating to variance, and b and c are functions .

In IRLS, the optimal parameter update under a local quadratic approximation of the likelihood function at a given point in the parameter space can be solved as a weighted least-squares problem, where the weights depend on the current parameter estimate, thus requiring iterative re-weighting across updates (Agresti 2015). The negative binomial distribution (eq. M.7) is a special case of exponential family generalised linear models because it is only an exponential family distribution if the dispersion parameter is fixed.

$$L(x; r, p) = \binom{x + r - 1}{x} (1 - p)^r p^x \quad (M.7)$$

The negative binomial distribution describes the probability mass of an urn experiment outcome of drawing x successes before r failures occur, where the success probability is p.

In next-generation sequencing modelling, a re-parameterisation of the negative binomial density which explicitly includes the expectation of the modelled count distribution as a parameter is very common (eq. M.8) (Love, Huber, and Anders 2014; Lopez et al. 2018; Eraslan et al. 2019):

$$L(x;r,p) = \frac{\Gamma(x+r)}{x!\Gamma(r)}(\frac{r}{r+m})^r(\frac{m}{r+m})^x \quad (M.8)$$

This re-parameterisation of the standard negative binomial distribution (eq. M.6) is often used in noise models as it explicitly contains the expectation of x as a parameter m. The second parameter r can be interpreted as an over-dispersion parameter which quantifies the deviation of the mean-variance trend of the distribution with respect to a Poisson distribution.

With the scale model fixed, the location model parameters can be updated with IRLS as a convex problem for the negative binomial model. However, the complete model estimation including dispersion parameter is no longer a convex problem. In the R package MASS (Venables and Ripley 2002), negative binomial generalised linear models are fit using a coordinate ascent scheme with alternating IRLS updates conditioned on a dispersion estimate and line search updates on the dispersion parameters.

A second class of non-exponential family probability distributions that has been important for RNA-seq data modelling consists of bi-modal noise models (Agresti 2015): In MAST (Finak et al. 2015), a hurdle model was proposed, a two-component Poisson was proposed in scde (Kharchenko, Silberstein, and Scadden 2014), and multiple other publications were centred around zero-inflated negative binomial distributions (Risso et al. 2017): These bimodal noise models (eq. M.9) were largely proposed to deal with an apparent bi-modality of single-cell data. The lower mode was typically close to zero and was described as “drop-out” (Risso et al. 2017), a technical artefact from the measurement.

$$L^{ZINB}(x;\mu,\phi,\pi) = \pi I_{x=0} + (1-\pi)L^{NB}(x;\mu,\phi) \quad (M.9)$$

The zero-inflated negative binomial likelihood L^{ZINB} is a mixture model of a point density *I* at zero and a negative binomial likelihood L^{NB} with a mixing parameter π (eq. M.8). Since its first reports, multiple studies have established that this drop-out was an artefact that likely originated in the lack of control of PCR bias during library amplification in early non-UMI protocols (Svensson 2020). Recently, almost all single-cell experiments have been performed with UMI, thus controlling for PCR bias, and there has not been evidence for drop-out effects in these newer protocols (Svensson 2020).

Regression models can be evaluated based on the likelihood value to take the residual weighting into account that was used for training, but are often also evaluated based on metrics that are more naive with respect to residual weighting, such as mean squared error or mean squared log error, correlation scores, or explained variance scores.

2.2.3 Differential equation models for single-cell population dynamics

Differential equation models are ubiquitous in computational biology and model the temporal evolution of systems such as biochemical reactions, populations of cells in tissues and populations of organisms in ecology (Holmes et al. 1994; Fröhlich et al. 2017). Cellular development can be described with population dynamics models to model transitions of cells between developmental stages over time. This type of model naturally lends itself to a differential equation model that models a probability distribution over time and a state space. Often, developmental stages are defined as discrete molecular states, such as FACS bins. A population dynamics model then describes the flux of cells between connected compartments over time as a system of coupled ordinary differential equations (ODEs) (eq. M.10):

$$\frac{d}{dt}X = AX \quad (M.10)$$

The temporal derivative of the discrete states X depends on the states X themselves. This model is called ordinary because the states are discrete. Here, this dependency is a linear dependency via the weight matrix A, the ordinary differential system is, therefore, called linear.

If one defines a continuous state space, such an ODE model can be replaced by a partial differential equation (PDE) model (eq. M.11):

$$\frac{d}{dt}c = D\frac{d^2}{ds^2}c \quad (M.11)$$

In a PDE, the partial derivatives of the differential equation model depend on each other. Shown here is a classical PDE model, Fick's second law, which models the temporal derivative of concentration c of a substance as a function of the second spatial s derivative of the concentration profile and the diffusion coefficient D.

In a few cases, the temporal evolution of a differential equation system can be expressed in closed form. If this is not the case, one needs to approximate the local temporal behaviour through the temporal derivative encoded in the equation system at a given state. However, this interpolation in time requires the continuous space used in PDEs to be discretised. One can distinguish three main paradigms for discretising space in forward simulations of PDEs: Finite differences, finite volumes and finite elements (Fröhlich et al. 2021). The finite difference method approximates the partial derivative of a differential equation at a given position in state space (eq. M.12) and iteratively uses a linear extrapolation in state space using this partial derivative in a first order Taylor approximation of the function to model the state of the system across the state space based on an initial condition.

$$\frac{d}{dx}f(x=a) = \lim_{h\to 0}\frac{f(x=a+h)-f(x=a)}{h} \quad (M.12)$$

The finite volumes method extends this point-grid approximation of the state space to a discretisation into volumes which are treated as compartments of an ordinary differential equation. The PDE is then simulated based on this system of ordinary differential equations.

The approximation of the temporal behaviour of a differential equation system is necessary both for prediction of unseen states and for parameter estimation: Parameters of differential equation models can be fit with maximum likelihood estimation and with Bayesian estimation schemes (Stapor et al. 2018).

In the pseudodynamics model (Publication 1), directed development, asynchronicity of maturation and proliferation or death events are core hallmarks of the cellular process which are captured and mapped to the cell state space. In this PDE model, a drift term represents directed development, a diffusion term models the effects of asynchronous development on the variance of the distribution and a reaction term with a birth-death rate represents proliferation and death events (eq. 13):

$$\frac{d}{dt}u(s,t) = \frac{d}{ds}\left(D(s,t)\frac{d}{ds}u(s,t)\right) - \frac{d}{ds}\left(v(s,t)u(s,t)\right) + g(s,t)u(s,t) \quad (M.13)$$

In this single-branch pseudodynamics PDE model for cellular development, u is a probability density function of a population of cells over a molecular space s and time t, D is a diffusion coefficient, v is a drift coefficient, and g is a birth-death coefficient. This equation reproduces (Fischer et al. 2019) Online Methods eq. 1.

We used a no flux boundary condition at both ends of the *s* domain, additionally enforcing zero drift at the right boundary (eq. M.14, M.15):

$$\left(D(s,t)\frac{d}{ds}u(s,t) - v(s,t)u(s,t)\right)\Big|_{s=0} = 0 \quad (M.14)$$

$$\frac{d}{ds}u(s,t)\Big|_{s=s_{max}} = 0 \quad (M.15)$$

Note that the molecular domain *s* was padded beyond the observed molecular states in all cases. Accordingly, the zero drift constraint did not directly affect observed states. These equations reproduce (Fischer et al. 2019) Online Methods eq. 2,3.

To account for developmental branching events, we coupled multiple PDEs for each branch through probability mass exchange terms at the ends of the modelled cell state intervals (eq. M.16, M.17):

$$\frac{d}{dt}u_1(s,t) = \frac{d}{ds}\left(D_1(s,t)\frac{d}{ds}u_1(s,t)\right) - \frac{d}{ds}\left(v_1(s,t)u_1(s,t)\right) + g_1(s,t)u_1(s,t) - T_1(s)\left(\delta_{12}u_1(s,t) - \delta_{21}u_2(s,t)\right) \quad (M.16)$$

$$\frac{d}{dt}u_2(s,t) = \frac{d}{ds}\left(D_2(s,t)\frac{d}{ds}u_2(s,t)\right) - \frac{d}{ds}\left(v_2(s,t)u_2(s,t)\right) + g_2(s,t)u_2(s,t) - T_2(s)\left(\delta_{21}u_2(s,t) - \delta_{12}u_1(s,t)\right) \quad (M.17)$$

This model for two branches is a system of two coupled PDEs which describe the two branches as two probability density functions on separate 1D domains, u_1 and u_2. On each branch, the population behaves as described through (eq. M.13). The coupling is induced through probability mass exchange terms on a defined region *T(s)* in the molecular state space, the "branching region", and probability mass exchange rates δ_{ij} from branch *i* to *j*. Extensions for more branches can be defined by the introduction of further branching regions. These equations reproduce (Fischer et al. 2019) Online Methods eq. 4,5.

Additionally, we also modelled the total population size to increase identifiability of the system (eq. M.18):

$$N(t) = \sum_{b\in B}\int_{s=0}^{s=s_{max}} u_b(s,t) \quad (M.18)$$

The total population size *N* of the cellular system is the sum of the integrals over the molecular space *s* of all branches *B*. This equation reproduces (Fischer et al. 2019) Online Methods eq. 10.

We defined a likelihood function using the modelled and observed probability density functions over molecular space at the observed time points and the modelled and observed population sizes at the observed time points. (eq. 19):

$$\log L(\theta) = \left(\sum_{b\in B}\sum_{t\in T^{cdf}} \log L(\theta; \mathrm{ecdf}_{S_{b,t}}(s))\right) + \left(\sum_{t\in T^N} \log L(\theta; \bar{N}_t, \sigma_t^N)\right) + \left(\sum_{b\in B\setminus b_{max}}\sum_{t\in T^{cdf}} \log L(\theta; \omega_{b,t}, \sigma_{b,t}^{\omega})\right) \quad (M.19)$$

The overall likelihood decomposes into three terms: First, the likelihood function for the fit of the modelled probability density is based on the empirical cumulative density functions of predicted and observed density at time points at which the population distribution was observed T^{cdf}. Second, the likelihood of the predicted population size is evaluated based on the mean and the standard deviation of the population sizes observed at the set of time points at which the population size was observed T^N. Third, the likelihood of the population size distribution over branches B is evaluated at T^{cdf}. This equation reproduces (Fischer et al. 2019) Online Methods eq. 11. We defined a loss function to optimise the parameters of the differential equation system, diffusion, drift and growth-rate coefficients, using the likelihood function and additional constraints on the parameters. In this study, we used a spline model for the dependency of diffusion, drift and growth-rate coefficients on the molecular space and imposed smoothness constraints on these splines in the loss function.

2.2.4 Uncertainty estimation: From linear models to neural networks

In pure frequentist statistics, one is only interested in finding the maximum likelihood estimator of the model parameters, a point estimator. However, point estimators are liable to identifiability issues in a model. In machine learning, identifiability issues often materialise as overfitting: The identity of the point estimator and the prediction change drastically with small changes in the training data. This phenomenon can be largely mapped to moments of the posterior of the model parameters, the core tool of statistical uncertainty estimation in a Bayesian approach: The posterior describes a distribution over parameter estimate values that incorporates both their likelihood under the data and prior assumptions. The posterior can be described through Bayes' theorem (eq. M.20):

$$p(\theta; D) = \frac{p(D;\theta)p(\theta)}{p(D)} \quad (M.20)$$

The posterior $p(\theta; D)$ is the probability of the model parameters θ given the data D, and is based on the likelihood $p(D; \theta)$, the prior $p(\theta)$ and the evidence $p(D)$. For a unimodal posterior, the variance of the distribution indicates the range of probable parameter estimates, so that one can interpret a higher posterior variance as a higher

uncertainty in the parameter estimate. Indeed, such a wide posterior yields strongly fluctuating parameter estimates if the training data is re-sampled and no further prior constraints are enforced on the parameter.

A popular framework that uses the notion of posterior width for uncertainty estimation is the Wald test (Wald 1943). In exceptional cases, one can find a closed form description of the posterior distribution of the parameters, and thus derive all required notions of uncertainty from characteristics of this distribution. However, in most cases this is not possible and one has to resort to approximations of the posterior distribution, such as through sampling, marginal-wise approximation, or parametric approximation.

First, one can sample an approximate posterior distribution by sub-sampling the training data and computing parameters estimates on each sample. In machine learning, this is done in some applications of cross-validation and is primarily deployed to quantify uncertainty in the model output rather than in the model parameters. In statistics, this re-sampling-based uncertainty estimation is called bootstrap. While simple to implement, this sub-sampling based methodology is expensive as it requires new model fits for each sub-sample, and rests on the assumption that the variation in sub-samples is representative of the variation in truly repeated sampling, which is often not given in real world data settings. A different approach to posterior sampling is implicit in optimisation schemes that sample the posterior, such as Markov-chain Monte Carlo methods (Ballnus et al. 2017). Here, samples from a chain in steady state can be considered samples from the posterior. Monte Carlo methods are however computationally expensive and liable to set-up issues in the chains.

Second, one can focus on the parameter-wise marginals and sample the posterior or likelihood function on a grid, which is usually much too expensive on the full multivariate posterior. An approach in this class of methods is called likelihood-profiling (Stapor et al. 2018): Here, target values on the grid of the profiled marginal are fixed and the remaining parameters optimised to yield an optimal cost function value for this grid point. While effective in describing the marginal posterior, these methods are computationally intensive because of the separate parameter estimation run for each grid point, liable to issues in choice of the grid and limited by their focus on the marginals of the full posterior.

Third, one can sometimes derive a parametric approximation to the full posterior. If a hessian can be computed, one can derive a parameter covariance matrix via the fisher information matrix (eq. 21).

$$\Sigma_\theta \geq I^{-1}(X,\theta) = \left(- E[H(X,\theta)] \right)^{-1} \quad (M.21)$$

A lower bound on the parameter estimate covariance matrix Σ can be derived based on the Fisher information matrix I (Cramér-Rao bound), which can be expressed based on the expected Hessian H (C. R. Rao 1945). This approach has conceptual overlaps to variational inference, in which a posterior is approximated during model fitting through a variational posterior, which is often chosen from the set of exponential family probability distributions.

This parameter covariance matrix together with the vector of maximum likelihood point estimates yield estimates of the first two moments of the multivariate posterior. These two moments are frequently used to approximate the full posterior with a multivariate gaussian distribution, which is the maximum entropy choice of distribution in this setting. In generalised linear models, this posterior approximation is used in the Wald test (eq. M.22, M.23):

$$W = \frac{(\hat{\theta} - \theta_0)^3}{var(\hat{\theta})} \quad (M.22)$$

$$W \overset{H_0}{\sim} \chi^2(k=1) \quad (M.23)$$

The Wald statistic W is the squared deviation between the parameter estimate and the null hypothesis; which is usually $\theta_0 = 0$. The test statistic depends on the deviation between estimate and null hypothesis and the variance of the parameter posterior, which can be derived from the fisher information matrix for generalised linear models (eq. 13). W is χ^2-distributed with a distribution with one degree of freedom ($k = 1$) if one parameter is tested. The distributional assumption on W under the null hypothesis allows for the derivation of p-values for a given model.

A common theme in uncertainty management is model regularisation to reduce overfitting and to reduce uncertainty. Lasso, Ridge regression and elastic nets are common variants of linear models with L1 and/or L2 penalties on the parameters (Zou and Hastie 2005). Similarly, L1 and L2 penalties are frequently enforced on parameters in neural networks. Input or hidden layer drop-out is another important mechanism of regularisation in neural networks (Hinton et al. 2012). Especially with large sample sizes, one is also confronted with the concept of data-intrinsic regularisation: With large enough data sets, the posterior variance becomes sufficiently small to avoid overfitting even without strong priors. While this can be directly assayed via Wald tests in generalised linear models, this is much harder to diagnose in neural networks: Here, overfitting is usually diagnosed based on held-out data but its management is hard. Often, one considers the training of the network as a trajectory through parameter space along which models are increasingly overfitted. Accordingly, one can attempt to find a model estimate that compromises training data fit and generalisability. For this purpose, one can co-evaluate the model on held-out validation data to diagnose overfitting during training.

2.2.5 Representations of prior knowledge on feature correlation patterns in neural networks

A strong trend in the deep learning literature has been the optimisation of network layer architectures to the underlying data type. Examples of such layers include convolutions, recurrent units, and graph convolutions.

Convolutional neural networks revolutionised computer vision through their flexible representation of local pixel correlations (Krizhevsky, Sutskever, and Hinton 2012), which correspond to recurring image objects across length scales, from edges to faces. Sequential data has been a second area of focus of deep learning model development and was previously approached with dedicated models, too, such as hidden Markov models, for example (Baum and Petrie 1966). Sequential data are often characterised by frequent local correlations and sparse distant correlations. Recurrent neural networks (Rumelhart, Hinton, and Williams 1986) capture local correlation in a layer structure element called “cell” that is repeated across sequence positions and, therefore, allows learning and prediction on varying sequence lengths. Similarly, convolutional networks can also be used for learning on sequence data. However, both suffer from difficulties with representing long-range correlations. The recurrent neural network cell was extended to a long-short term memory (Hochreiter and Schmidhuber 1997) (LSTM) cell to explicitly represent these long-range correlations. Convolutional networks were adapted to long sequences through dilated convolutions which increase the field of vision of a latent unit exponentially in the depth of the network, rather than linearly as in standard convolutional networks. Sequence-based neural networks are used for representation learning on text, sound and protein sequences, for example. Recently, representation learning on text has received especially large interest by the Deep Learning community. One advance in this field are transformer networks (Vaswani et al. 2017). Transformers have been very successful at text generation tasks and are scaled to more than one billion parameters (Radford et al. 2019).

Graphs are a very flexible mathematical representation of expected correlation structures and have recently gained much interest in the Deep Learning community as intuitive representations of spatial, spatiotemporal and contact networks (Kipf and Welling 2016; Duvenaud et al. 2015; Bruna et al. 2013). Graph-neural networks forward propagate activations in graphs and can be mapped to convolutional neural networks (graphs of pixels) and recurrent neural networks (graphs of sequence elements). The self-attention concept was also translated to node embedding in graph-neural network, in graph attention networks (Veličković et al. 2017).

Layer structure choice is a crucial step in many Deep Learning projects because advanced layer architectures are often more parameter-efficient in encoding particular correlation structures than densely connected layers are. Layers reflect prior knowledge by constraining the function space accessible to the neural network.

Chapter 3. Publication summaries

This section contains one-page summaries of the seven main publications and preprints presented in this thesis.

3.1 Publication 1: Inferring Population Dynamics from Single-Cell RNA-Sequencing Time Series Data.

The paper “Inferring Population Dynamics from Single-Cell RNA-Sequencing Time Series Data.” was published in 2019 in the journal Nature Biotechnology as an article. The full citation is:

Fischer, David S.*, Anna K. Fiedler*, Eric M. Kernfeld, Ryan M. J. Genga, Aimée Bastidas-Ponce, Mostafa Bakhti, Heiko Lickert, Jan Hasenauer, Rene Maehr, and Fabian J. Theis. 2019. “Inferring Population Dynamics from Single-Cell RNA-Sequencing Time Series Data.” *Nature Biotechnology* 37 (4): 461–68, https://doi.org/10.1038/s41587-019-0088-0.

Contribution:
I performed the unsupervised analysis of the single-cell data that yielded the developmental trajectories and branching events and which yielded the hypothesis of the quasi steady state in T cell maturation. I performed model interpretation with respect to the quasi-steady state discovered in T cell development and cell intrinsic and extrinsic effects in beta cell maturation. I contributed to experiment design I defined the modelling scenario tailored to intrinsic and extrinsic predictors of cell state of pancreatic beta cells. I wrote the manuscript with Anna K. Fiedler, Jan Hasenauer, and Fabian J. Theis with assistance from all other authors.

Additional supplementary material:
Additional supplementary material is available at the publisher’s website (https://doi.org/10.1038/s41587-019-0088-0). All code published in the context of this project can be found on Github (https://github.com/theislab/pseudodynamics).

Summary:

Time-resolved models of developmental processes measured with scRNA-seq would require models or cellular distributions in a continuous state space in time; a modelling paradigm that can be captured by PDEs. We proposed a PDE model for cellular development in a continuous molecular cell state space. This state space is a dimension reduction of a high-dimensional gene space. This approach overcomes the limitations inherent in discrete state models used in ordinary differential equation models for flow cytometry data and the identifiability issues related to steady-state observations (sec. 1.3). We used this new mathematical model, pseudodynamics, to gain insights in two cellular systems with a strong temporal component to development: Embryonic T cell maturation and pancreatic beta cell maturation.

My collaborators and I developed the pseudodynamics model for directed cellular development, asynchronicity of cellular development and proliferation or death events during cellular development. We mapped these hallmarks of cellular processes onto a molecular cell state space. In this PDE model, a drift term represents directed development, a diffusion term to model effects of asynchronous development on the variance of the distribution and a reaction term with a birth-death rate represents proliferation and death events (sec. 2.2.3). To account for developmental branching events, we coupled multiple such PDEs for each branch through probability mass exchange terms at the ends of the modelled cell state intervals. Additionally, we also modelled the total population size to increase identifiability of the system. We defined a likelihood function using the modelled and observed probability density functions over molecular space at the observed time points and the modelled and observed population sizes at the observed time points. We defined a loss function to optimise the parameters of the differential equation system, diffusion, drift and growth-rate coefficients, using the likelihood function and additional constraints on the parameters.

My collaborators sampled the thymus from mouse embryos between ages E12.5 and P0 with Drop-seq to capture the developmental trajectory of T cells in embryonic mice. I performed the unsupervised analysis of this scRNA-seq time course data of developing T cells in the mouse thymus based on previous cell type assignments by my collaborators. I recovered the full maturation trajectory from progenitors up to double-positive T-cells, including a branching event to NK-T cells in an unsupervised analysis of the scRNA-seq data. I defined a maturation progression coordinate as a 1D compression of the transcriptomic state using diffusion pseudotime and set out to model the probability density of the population in the 1D transcriptomic space (cell state) as a function of time. Notably, this 1D cell state space was the first continuous interpolation of the discrete developmental stages usually defined for T cell maturation, a significant advance in molecular characterisation of this system. I discovered a quasi steady state in T cell maturation at later time points and found that the stationary distribution co-localises both with transcriptomic markers of beta-selection, and with the cell states on which negative selection acts. Therefore, I hypothesised that a source-sink system to be the cause of the quasi steady-state. I validated the annotation of the developmental trajectory with the beta-selection event by integrating the wild type samples with samples from Rag1/Rag2 knockout mice that have T cells incapable of passing beta-selection. T cells from these knockout mice were indeed delayed in developmental progress compared to wild type T cells from matched time points. I then designed a validation experiment based on Rag1/2 knockout mice with my collaborators. Indeed, we could show that T cells from these mice arrest at a transcriptomic state that corresponds to the proposed beta selection cell state. My collaborators and I designed a probabilistic model of the position of beta selection on the developmental trajectory based on these knockout samples. We concluded that T cell development is a bi-phasic process from a dynamic point of view and quantified birth-death rates along the process, mapping transcriptomic states to proliferation after beta-selection and to death caused by apoptosis induced by positive and negative selection on T cells with fully developed T cell receptors.

3.2 Publication 2: Predicting antigen specificity of single T cells based on TCR CDR3 regions.

The paper “Predicting antigen specificity of single T cells based on TCR CDR3 regions.” was published in 2020 in the journal Molecular Systems Biology as an Article. The full citation is:

Fischer, David S., Yihan Wu, Benjamin Schubert, and Fabian J. Theis. 2020. “Predicting Antigen Specificity of Single T Cells Based on TCR CDR3 Regions.” *Molecular Systems Biology* 16 (8): e9416, https://doi.org/10.15252/msb.20199416.

Contribution:
I designed the models in this study, performed analysis, led code development, and wrote the manuscript with the other authors.

Additional supplementary material:
Additional supplementary material is available at the publisher’s website (https://doi.org/10.15252/msb.20199416). All code published in the context of this project can be found on Github (https://github.com/theislab/tcellmatch).

Summary:
We studied the prediction of single T cell specificity to antigens as quantified by binding of oligonucleotide-labelled and pMHC multimers, such as dextramers, focussing on data modalities that are specific to the single-cell measurement setting. I identified dextramer counts coupled to CDR3 sequence measurements in the same cells as data that allow for the fitting of supervised models that predict antigen specificity, measured by dextramer counts per cell, as a function of the cells CDR3 sequences. Here, cells are observations, in contrast to previous models that used specificity measurements from bulk experiments. This approach drastically improves the effective experimental throughput that can be used by models: We were able to train models on the order of 1e5 observations from a single study, in contrast to about 1e4 observations which were previously curated from the entire literature (Shugay et al. 2018; Vita et al. 2019). Moreover, I identified the transcriptomic state and surface protein state as an additional confounder of dextramer binding in individual cells, thus extending this supervised model class by information not available from bulk assays. I designed models capable of using this transcriptomic information and an ablation study to show its relevance for the prediction problem. My collaborators and I then implemented neural networks that perform the prediction task and performed a benchmarking study.

First, we found that models trained on both alpha- and beta-chain sequences were only weakly more predictive for antigen binding than models only fit on beta-chain sequences. This finding can be rationalised based on the co-evolution of alpha- and beta-chain sequence during T cell maturation. Second, we did not find evidence that models could generalise well to unseen antigens given the currently available data. We hypothesise that this is due to the extremely sparse sample of the antigen space and the confounding of binding events by HLA types. Here, larger and more targeted data collection efforts are required to build models that can predict binding of TCRs to unseen antigens.

Next, we established a machine learning model to use single-cell measurements of pMHC oligomers as a readout of TCR-antigen binding events (Bentzen et al. 2016). The opportunities of this data modality to advance TCR specificity modelling are manifold: First, the specificity of a single TCR (cell) to multiple antigens can be assayed in a single cell by treating a T cell sample with a mixture of different pMHC identities. Second, the binding assay is multiplexed via cells and yields much larger observation sets than previously possible with bulk assays (Bagaev et al. 2020). Third, molecular states of cells can be assayed in parallel to binding events and can confound unspecific binding events. Fourth, the pMHC concentration bound to cells may be used as a quantitative read-out for binding, yielding data beyond binary binding observations often reported in databases. We addressed all four points in new models trained on a large single T cell data with RNA, surface proteins, CDR3 and pMHC binding measured for around 100,000 cells from four patients. We could indeed show in ablation studies that the molecular state of cells conveys additional predictive power to CDR3-based models of pMHC binding. Moreover, we could also show that pMHC binding can be predicted in a quantitative fashion.

The dependency of binding events on transient molecular states and constant genetic states, that we established in this study, narrow the gap from the bioinformatics field of TCR-antigen binding prediction to contemporary fields of T cell research: It is widely accepted the molecular states are core determinants of antigen-induced T cell behaviour in many physiological settings, e.g. in T cell exhaustion (Blank et al. 2019). By allowing for the integrating of cellular molecular states with TCR-antigen binding events, we provide a holistic understanding of T cell action *in vivo*.

3.3 Publication 3: Single-Cell RNA Sequencing Reveals in Vivo Signatures of SARS-CoV-2-Reactive T Cells through ‘reverse Phenotyping.’

The paper “Single-Cell RNA Sequencing Reveals in Vivo Signatures of SARS-CoV-2-Reactive T Cells through ‘reverse Phenotyping.’” was published in 2021 in the journal Nature Communications as an article. The full citation is:

Fischer, David S.*, Meshal Ansari*, Karolin I. Wagner* et al. 2021. Single-cell RNA sequencing reveals ex vivo signatures of SARS-CoV-2-reactive T cells through ‘reverse phenotyping’. *Nature Communications* 12, 4515, https://doi.org/10.1038/s41467-021-24730-4.

Contribution:
I designed and performed all analysis related to the stimulation assay in this study and contributed to manuscript writing and to the integrative analysis with published data sets.

Additional supplementary material:
Additional supplementary material is available at the publisher’s website (https://www.nature.com/articles/s41467-021-24730-4#Sec43).

Summary:

CDR3 measurements in single cells do not just serve the purpose of investigating TCR sequence-dependent antigen specificity, but also allow for a grouping of cells into clonotypes: groups of cells derived from the same ancestor T cell. In T cells, these lineage relations can be established based on common or similar TCR sequences (Sturm et al. 2020). We hypothesised that we can use clonotype assignments to track large clones across short term *ex vivo* stimulations. This type of analysis requires the identification of virus-specific T cells among other T cells that may also be affected by the disease or co-infections. Previously, virus-specific T cells were identified based on staining with MHC multimers loaded with virus epitopes, or, based on *ex vivo* stimulation phenotypes. Both are subject to fundamental limitations, including the restriction to particular epitopes and confounding of *in vivo* transcriptomic states with *ex vivo* stimulation gene expression signatures, respectively. We addressed these limitations using a new coupled experimental and algorithmic approach: reverse phenotyping.

My collaborators sampled peripheral blood mononuclear cells of two patients, stimulated cells *ex vivo* with SARS-CoV2 spike protein, isolated T cells using FACS and performed scRNA-seq with additional CDR3 sequence capture on T cells from before and after stimulation. Additionally, they performed scRNA-seq on T cells from tracheal aspirate samples from the same patient cohort. First, I characterised the transcriptomic state of T cells specific to virus spike protein before re-stimulation based on interferon gene expression. I performed unsupervised scRNA-seq analysis on the T cells, identifying CD4 and CD8 T cells, T cell phenotypes and activation states. I then identified clonotypes based on CDR3 sequences and coupled clonotypes across conditions within patients, thus establishing maps between T cell phenotypes in stimulated and unstimulated conditions. I used these maps to identify virus-specific cells in the unstimulated condition and used this classification to build gene signatures characteristic of specific cells. I identified distinct transcriptomic signatures of virus-specific clonotypes, characterising clonotype heterogeneity both in the unstimulated condition and with respect to their response to stimulation.

My collaborators used and validated these gene expression signatures on tracheal aspirate samples from our study and on T cells from other COVID-19 cohorts. My collaborators then validated the TCRs that I predicted to be spike protein-specific in an additional screen using orthotopic TCR replacement: My collaborators used a genetic engineering system to replace the innate TCRs in control T cells from healthy donors with the putative spike-specific TCRs. The transgenic T cells produced significantly more cytokines associated with TCR signalling activity when stimulated with virus antigens then non-transgenic control cells, thus corroborating the validity of the reverse phenotyping approach. As a second core result of this project, these validated SARS-CoV2-specific TCRs can now be used in *in vitro* models of SARS-CoV2 immunity. In summary, we improved the characterisation of the T cell response to SARS-CoV2 infection and presented a generalisable framework to identify phenotypes of disease-specific T cells during an infection.

The reverse phenotyping method discussed here has applications in many stimulation assays in T cell systems. This method improves on unconstrained distribution matching between patients, such as with optimal transport and CycleGANs (Zhu et al. 2017), by establishing explicitly observed domain transitions.

3.4 Publication 4: Impulse Model-Based Differential Expression Analysis of Time Course Sequencing Data.

The paper “Impulse Model-Based Differential Expression Analysis of Time Course Sequencing Data.” was published in 2018 in the journal Nucleic Acids Research as an Article. The full citation is:

Fischer, David S., Fabian J. Theis, and Nir Yosef. 2018. “Impulse Model-Based Differential Expression Analysis of Time Course Sequencing Data.” *Nucleic Acids Research* 46 (20): e119, https://doi.org/10.1093/nar/gky675.

Contribution:
I designed the maximum likelihood approach to fitting the ImpulseDE2 model to data and designed statistical hypothesis tests to the relevant questions in time series data. I implemented ImpulseDE2 as an R package. I performed the benchmarking and wrote the manuscript with the other authors.

Additional supplementary material:
Additional supplementary material is available at the publisher’s website (https://academic.oup.com/nar/article/46/20/e119/5068248#124731012). All code published in the context of this project can be found on GitHub (https://github.com/YosefLab/ImpulseDE2).

Summary:

One of the core challenges of analysing bulk RNA-seq lies in the proper treatment of technical variation (“noise”) and biological variation. As bulk RNA-seq became much more accessible, also densely sampled time series bulk measurements became more common (Graveley et al. 2010; Broadbent et al. 2015; Jovanovic et al. 2015; Baran-Gale, Purvis, and Sethupathy 2016; Sykes et al. 2016). Time series measurements are interesting from a noise modelling point of view as biological variation can be captured as a smooth trend across time, which allows for a disentanglement of the mean trajectory and of technical variation, even without replicating samples. In previous work, Sander et al. had modelled bulk RNA time series with a double-sigmoid (“impulse”) model (eq. R.1):

$$\mu(t) = \frac{1}{h_1}\left(h_0 + \frac{h_1 - h_0}{1+^{\beta(t-t_1)}}\right) * \left(h_2 + \frac{h_1 - h_2}{1+^{\beta(t-t_2)}}\right) \quad (R.1)$$

The mean μ is modelled as a function of time t via three amplitude parameters h_0, h_1 and h_2, two switch-point parameters t_1 and t_2 and a slope parameter β. ImpulseDE (Sander, Schultze, and Yosef 2017) could recover smooth temporal trends but was very slow to fit and was tailored to microarray data and did, therefore, not account for the strong mean-variance relationship usually encountered in RNA-seq data. Indeed, other differential expression frameworks, which were not centred on time series analysis, already used generalised linear models with negative binomial noise models at this time (Love, Huber, and Anders 2014). I defined a maximum likelihood approach to fitting the impulse model and defined the model likelihood as a negative binomial function with mean $\mu(t)$, scaled by a size factor, and a pre-defined dispersion parameter.

I defined the temporal dependency of the mean of the expression distribution that is necessary to parameterise the negative binomial likelihood through the impulse model. I modelled the dispersion parameter as a constant over time, similar to how the dispersion model was previously treated in generalised linear models on RNA-seq data (Love, Huber, and Anders 2014). I solved the issue of regularising the dispersion estimate of the negative binomial noise model by pre-fitting this parameter with a regularised generalised linear model from DESeq2 (Love, Huber, and Anders 2014). Secondly, I designed a model selection mechanism based on similar methods used for differential expression analysis with generalised linear models: I introduced a likelihood ratio test that is capable of assigning significance to non-constant gene expression trajectories in time trends. Here, the null model has a constant mean in time, which I showed to be a nested model of the impulse model. The likelihood ratio test yields p-values instead of empirical p-values which were previously used in ImpulseDE. These p-values are much faster to compute than their empirical counterparts because they do not require repeated model fits to sample the null distribution of the test statistic. I designed a similar likelihood ratio test for non-constant condition effects on gene expression trajectories in time, covering the most common experimental scenarios in this field. I could show in benchmarking of the ImpulseDE2 model that ImpulseDE2 outperforms ImpulseDE both in fitting speed and accuracy of detecting differentially expressed genes. Moreover, the prior implicit in the parametric form of the mean trajectory described by the impulse model did indeed give additional statistical power in detecting differentially expressed genes if many, more than six, time points were sampled, as our model uses less degrees of freedom (six) for a similar fit as a generalised linear model that is based on a categorical description of time (number of time points). In this benchmarking we also investigated spline-based models RNA-seq measurements of gene expression in time with generalised linear models which had not been discussed much in the literature before. We found that these spline-based generalised linear models were a good alternative to reduce the number of parameters used in time series models. A core advantage of the impulse model compared to these splined based models was that we could stratify differentially expressed genes into permanently changed and transiently regulated genes based on nested likelihood ratio tests.

3.5 Publication 5: Sfaira Accelerates Data and Model Reuse in Single Cell Genomics.

The paper “Sfaira Accelerates Data and Model Reuse in Single Cell Genomics.’” was published in 2021 in the journal Genome Biology as a Software article. The full citation is:

Fischer, David S.*, Leander Dony*, Martin König, Abdul Moeed, Luke Zappia, Lukas Heumos, Sophie Tritschler, Olle Holmberg, Hananeh Aliee, and Fabian J. Theis. 2021. “Sfaira accelerates data and model reuse in single cell genomics.” *Genome Biology* 22, 248. https://doi.org/10.1186/s13059-021-02452-6.

Contribution:
I designed the sfaira project and the overall software architecture, was lead developer of the project and developed the aggregate cross-entropy solution to ontology-based cell type classification. I wrote the manuscript with the other authors.

Additional supplementary material:
Additional supplementary material is available at the publisher’s website (https://doi.org/10.1101/2020.12.16.419036). All code published in the context of this project can be found on GitHub (https://github.com/theislab/sfaira, https://github.com/theislab/sfaira_benchmarks).

Summary:
Many unsupervised analyses start with alignment output summaries (Luecken and Theis 2019), which consist of an UMI count matrix for most scRNA-seq protocols. There are strong format differences in these count matrices between studies. Additionally, there are often cell- or dataset-wise meta-data, such as cell type labels or tissue of origin. These meta-data are heterogeneous in terms of storage format and differ in naming conventions: T cell subtypes are not always called the same but are largely labelled as free text. This happens even though there are extensive ontology efforts which aim at structuring such metadata (Diehl et al. 2016). This lack of streamlined data is a major bottleneck for many single-cell analysis projects. First, re-analysis of published data is slowed down, even though more frequent re-analyses would be both beneficial to the scientific community by providing reproduction of published analyses, and would enhance new results as they would be contextualised in the context of more public reference data sets. Second, advanced automation of single-cell data analyses depends on standardised data formats and may greatly accelerate single-cell analysis projects in the future. Because of this lack of data streamlining, the input and output to machine learning models used on scRNA-seq data is typically also not streamlined. We addressed this lack of standardisation with a software and algorithm set implemented in the software package sfaira.

I planned the software architecture required to assemble data loaders in a community-driven way and to allow for decentralised model deployment and contribution. My collaborators and I then implemented this software, yielding a data zoo and a model zoo. My collaborators and I leveraged this data repository to build an entirely automated model deployment pipeline for scRNA-seq: We divided the data set collection up into data sets per organism and organ (anatomic structure) to train and deploy models across these partitions. I defined two modelling settings required for basic interpretation of scRNA-seq data: First, embedding models yield a dimension reduction which can be visualised with dimension reduction methods such as UMAP (L. McInnes, Healy, and Melville 2018). Second, a cell-type model predicts a cell type for each cell. Taken together, both yield a dimension reduction with cell type annotation without the need for any data processing. My collaborators and I showed that these embedding models and cell type predictor models work stably across organs in data set hold-out experiments. In this process, we identified cell type annotation granularity as a key issue with training supervised models across studies: A particular group of cells would be annotated coarsely as T cell in one study, while the corresponding set of cells would be annotated as fine grained T cell subtypes in another study. These differences originate from data set size, data quality and study focus and are difficult to mitigate retrospectively without major re-analysis efforts. I developed an aggregate cross entropy loss function and aggregate accuracy score which can be used in these settings, which generalise cell type prediction models from cell type lists (categorical label space) that were used before to ontologies (directed acyclic graphs) (sec. 2.2.3.1). In this aggregate cross entropy loss function, probability mass is propagated from finer labels to coarser labels in the directed acyclic graphs provided by a cell type ontology to train on data sets with different levels of annotation granularity at the same time.

My collaborators and I showed that these embedding and cell type classifier models can be meaningfully deployed in a scRNA-seq analysis setting. This level of automation advanced the state-of-the-art beyond what was done in previous publications which often relied on integrated data sets from different studies or preprocessed gene feature spaces for a particular cell type prediction problem.

3.6 Publication 6: MPRAnalyze: Statistical Framework for Massively Parallel Reporter Assays.

The paper “MPRAnalyze: Statistical Framework for Massively Parallel Reporter Assays.” was published in 2019 in the journal Genome Biology as a Method article. The full citation is:

Ashuach, Tal*, **David S. Fischer***, Anat Kreimer, Nadav Ahituv, Fabian J. Theis, and Nir Yosef. 2019. “MPRAnalyze: Statistical Framework for Massively Parallel Reporter Assays.” *Genome Biology* 20 (1): 183, https://doi.org/10.1186/s13059-019-1787-z.

Contribution:
I designed the initial models, initial R package and proof-of-concept application. I contributed to further development and manuscript writing.

Additional supplementary material:
Additional supplementary material is not suitable for printing and is available at the publisher’s website (https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1787-z#Sec26). All code published in the context of this project can be found on GitHub (https://github.com/YosefLab/MPRAnalyze).

Summary:
MPRAs characterise variability in gene expression as a function of the DNA sequence in proximal regulatory elements in a plasmid model system. In contrast to genome wide association studies, they do not rely on natural variability but allow for the screening of arbitrary sequence motif libraries. Moreover, they are multiplexed in the sequences that are assayed and give a readout for a full library of sequence motifs in a single bulk experiment. This multiplexing results in a confounding of the output of reporter genes that quantify promotor activity with the number of plasmids in the cell culture, which may vary between sequence motifs due to concentration difference after clonal amplification, for example. In MPRA, the plasmid concentration is sampled alongside the reporter gene RNA concentration, thus allowing for this confounding to be resolved. Therefore, a statistical treatment of many common hypotheses revolving around differences in promotor activity between assayed sequences needs to account for the plasmid concentration.

I translated common mathematical concepts from statistical modelling of RNA-seq data to coupled plasmid DNA and reporter gene RNA read counts in MPRAs (Inoue and Ahituv 2015). In a MPRA, one measures the transcription rate induced by a promotor on a plasmid vector in cell culture by measuring both plasmid abundance and reporter gene abundance. Both observations are noisy and depend on each other. Previously, this data had been analysed by estimating a translation rate as the ratio of RNA and DNA counts (“ratio method”) (Myint et al. 2019). The “ratio method” suffers from noise in RNA and DNA observations and requires an extension to a linear model to correct for confounding variables (Myint et al. 2019). I proposed a model in which both a latent RNA and DNA concentration and their dependence in a sample are modelled, thus accounting for technical noise. For this purpose, I defined the observed RNA distribution as a convolution of Poisson technical noise over a gamma-distributed latent RNA amount random variable. The latent RNA amount is the scaled gamma-distributed DNA amount. Here, the scaling factor α is the translation rate. This convolution can be expressed in closed-form as a negative binomial distribution, making this a very efficient model for maximum-likelihood based methods (eq. R.2, R.3):

$$D \sim \mathrm{Gamma}(k, b) \quad (R.2)$$
$$R \sim \mathrm{Pois}(\lambda = \alpha * D) \quad (R.3)$$

Here, the DNA amount *D* is gamma distributed and the RNA amount *R* is Poisson distributed with the modelled DNA amount scaled by a constant α as a constant.

I also introduced a relaxations of this model that do not couple the full latent DNA distribution to the RNA observations but only their mean (eq. R.4, R.5):

$$\log D = \log D^{latent} + log(s_D) = X_D\beta + log(s_D) \quad (R.4)$$
$$\log R = X_D\beta + X_R\gamma + log(s_R) \quad (R.5)$$

Shown is a linear model for the log amount of DNA *D* and RNA *R* with a design matrix for DNA amount variation X_D and a design matrix for gene expression rate variation X_R with parameters β and γ. s_D and s_R are size factors of the DNA and RNA samples, respectively, and adjust the relative amount by the sequencing depth of the library. In this model, the latent DNA amount D^{latent} appears as a point estimator $X_R\beta$ in the RNA model.

My collaborators compared different MPRAnalyze models and previously used methods with respect to their deviation from the “ratio-method”, specificity and sensitivity on control experiments, performance on simulation and downstream analysis of transcription factor binding site enrichments in active screened sequences. My collaborators and I found that MPRAnalyze is a robust computation method that caters to all common analysis scenarios related to MPRA data. Importantly, MPRAnalyze is the first method to properly account for the count statistics inherent in MPRA data.

3.7 Preprint 1: Learning cell communication from spatial graphs of cells.

The paper Learning cell communication from spatial graphs of cells." was published in 2021 as a preprint on bioarxiv. The full citation is:

Fischer, David S.*, Anna C. Schaar*, Fabian J. Theis. 2021. "Learning cell communication from spatial graphs of cells." Preprint available on bioarxiv by *Cold Spring Harbor Laboratory*, https://doi.org/10.1101/2021.07.11.451750.

Contribution:
I designed the project and models, performed analyses with Anna C. Schaar, and wrote the manuscript with the other authors. I developed the software with Anna C. Schaar.

Additional supplementary material:
Additional supplementary material is available at the publisher's website (https://www.biorxiv.org/content/10.1101/2021.07.11.451750v1.supplementary-material). All code published in the context of this project can be found on GitHub (https://github.com/theislab/ncem, https://github.com/theislab/ncem_benchmarks).

Summary:
Spatial molecular profiling data with sub-cellular resolution characterise cell types and molecular states of cells in their tissue context. Experimental protocols capable of producing sub-cellular resolution include MERFISH (Xia et al. 2019), seq-FISH (Eng et al. 2019), and CODEX (Goltsev et al. 2018), for example. Data from these protocols is often analysed based on cellular or nuclear segmentation, which give rise to spatial graphs in which nodes are molecular vectors describing individual cells and edges represent spatial proximity. Previously, neighbourhood enrichment analysis was used to identify pairs of cell types that co-localise (Dries et al., n.d.; Palla et al., n.d.), variance decomposition approaches have been used to decompose gene expression variation in space (Arnol et al. 2019; Tanevski et al., n.d.), and ligand-receptor gene function has been identified based on the spatial gene distribution (Yuan and Bar-Joseph 2020). On the other hand, ligand and receptor gene expression is used in scRNA-seq to identify pairs of communicating cell types (Browaeys, Saelens, and Saeys 2020; Efremova et al. 2020).

I proposed to describe statistical dependencies between cells in spatial graphs with a graph neural network that receives cell types in the input and predicts gene expression in the output. We named this model node centric expression models (NCEMs). NCEMs leverage the spatial context of a cell to learn its exact molecular state, without confounding this inference with the cell type frequency distribution. I showed that this graph neural network can be framed both as a linear model in which individual interactions between cell types can be tested with Wald tests, and as a nonlinear graph neural network that can capture higher order dependencies, but is less interpretable. My collaborators and I showed in spatial ablation studies that these models identify statistical dependencies between cells on biochemically reasonable length scales in the tissue, at distances slightly larger than average cell radii. Furthermore, we showed that the identified statistical dependencies between pairs of cell types are interpretable with respect to niches in the tissue. Finally, we showed current limitations in leveraging this framework for deep variational inference with conditional variational autoencoders. NCEMs provide a principled statistical backbone for modelling statistical dependencies between cells in spatial molecular profiling data.

Chapter 4 General discussion and outlook

Immune cells can be characterised with multi-modal single-cell measurements and a zoo of mathematical models is used to infer mechanistic insights on immune system processes from this high-dimensional data. Here, I presented multiple novel approaches to understand single-cell data and discussed exemplary insights on T cell maturation and antigen recognition. These findings are contextualised within computational biology and single-cell biology in the discussion section. Individually, progress on each of the presented approaches depends on the specific measurement technology required and on related algorithmic advances, as outlined in the outlook.

4.1 General discussion

The findings described in this thesis have implications for mathematical models of single-cell data, and for single-cell immunology.

4.1.1 Bottom-up and top-down cell biology

The algorithms discussed in this study can be put into the context of other modelling approaches that are often classified as bottom-up or top-down in systems biology: bottom-up models describe system properties based on defined events, latent states are usually directly interpretable as species or other physical properties of the system. On the other hand, top-down modelling is primarily centred on describing variation in a system, often with unsupervised, or predictive angles. Traditionally, cell biology has been dominated by reductionist approaches to cellular systems (Van Regenmortel 2004): Famous success stories of reductionism in cell biology include the discovery of DNA and the functional characterisation of many proteins. It is noticeable that many of these discoveries were not sparked by full bottom-up models of system behaviour which are often deployed in physics or chemistry. Reductionism is a valuable tool in this context as it allows the isolation of a particular property of a complex system in a bottom-up model if the hypothesis is phrased carefully enough. For example, individual genes can be identified as necessary or sufficient for a particular process based on studies centred on those genes. However, reductionism suffers from biases that relate to the class of hypotheses that can be tested and the questions that are asked, which are typically inspired by the current state of knowledge. Because of this bias, reductionism can slow discovery of novel phenomena.

On the top-down modelling side, black-box models from statistical learning can be used to learn representation of systems and may be predictive, but often suffer from strongly reduced interpretability compared to bottom-up models. Moreover, the limited mechanistic constraints imposed on these blackbox models in cell biology may result in very limited out-of-domain performance, which is key to most generalisation tasks that are relevant in cell biology. The approaches discussed in this thesis leverage carefully designed biological priors to model axes of variation in high-throughput cell biological experiments, with a focus on single-cell and bulk omics experiments. I advanced the available analytic methods both by defining new models and by translating modelling paradigms from other machine learning disciplines to cell biology. Instead of attempting to infer all parameters of a global dynamic or causal statistical model of a cell with molecular entities as states in a bottom-up setting, variance attribution correlates components of variation with known biological effects and enables the disentanglement of cause and effects in particular cases in which biological priors are available: The response of a T cell to antigen challenge presents correlation of condition with state change. From molecular immunology, we know the pathways that are activated in response to T cell receptor signalling and can therefore identify selected cell state changes as antigen-induced. In the spatial setting, we can attribute molecular variation within cell types to niche composition. One can consider these classes of variance attributing models as a middle-ground between black-box top-down models and complete bottom-up models, that both fit the data formats currently available in these branches of cell biology and the type of hypothesis often considered in single-cell studies.

As the data complexity increases in the future, variance attributing models may be pushed further towards bottom-up models as increasingly fine grained biological priors can be represented. On the other hand, interpretability mechanisms for black box models may improve in the future and may yield a competitive alternative for extracting actionable biological hypotheses. Likely, combinations of these modelling approaches will remain relevant and their relative added value will remain application-dependent. However, as machine

learning models evolve rapidly at the movement, there is reason to believe that black-box modelling and variance attribution are promising approaches for the near future. All of these advances depend crucially on well structured and annotated training data, a challenge that I addressed with the sfaira toolbox.

The diversity of the biological priors discussed here shows that models have to be carefully tailored to the biological hypothesis at hand and the modalities that can be obtained for the given system. Both cell biological questions and available data change rapidly, making the potential insights gained through variance attribution and its best reflection in a mathematical model a moving target. This insight highlights the need for a continuous rephrasing of biological questions in the language of data acquisitions set-ups and analytical frameworks, the core pillars of computational biology. New insights into cell biology are not limited to hypotheses defined in reductionist study designs anymore, but are often complemented by global insights into systems based on statistical modelling and machine learning.

4.1.2 Single-cell immunology

Molecular heterogeneity of T cells can be described based on single-cell data alone. However, T cell function is heavily dependent on the tissue context *in vivo*. Thus, models of T cells that take their environment into account will likely yield more meaningful representations in the future. In this thesis, I improved models used to characterise T cells based on antigen binding. First, I addressed a bottleneck in antigen-binding measurements through predictive modelling of antigen binding. Second, I demonstrated how T cell phenotypes in ex vivo antigen stimulation can be disentangled into antigen-specific and by-stander responses using a novel combined experimental and computational approach (reverse phenotyping). These approaches improve the characterisation of antigen-mediated activation of T cells and therefore improve the characterisation of potential targets of therapies centred on T cells.

From a basic T cell biology point of view, understanding the compromises defined by the thymus between reducing the production of auto-reactive T cells and maintaining a large enough T cell repertoire capable of quickly identifying any pathogen may improve the design of therapeutic immune system modulation. In this thesis, I proposed a trajectory model of T cell development which may be queried for molecular hypothesis on T cell selection. Moreover, our population dynamic description of this cellular system quantifies T cell selection pressure. Research that extends this approach may quantify tolerable levels of auto-reactive T cells in the future which may be used to tune T cell therapies or may be used to recreate T cell maturation *ex vivo*, thus increasing the biotechnological tool set to produce T cells for therapies. Lastly, T cell maturation is a key phase in the life of a T cell and many regulatory insights gained on T cell function during this phase may likely be transferred to phenotypes of mature T cells.

4.2 Outlook

Future progress in each of the five project areas depends on particular measurement technologies and algorithm development for particular data modalities and questions.

4.2.1 Annotating molecular heterogeneity of T cells

A shift from considering cells individually to modelling populations of cells will yield new insights into cellular systems. Differential equations in time and molecular space are a strong model class candidate for these settings but sufficient time resolution is still relatively rarely sampled because of the high costs of the experiments. Distribution matching (domain translation) machine learning methods may provide population transport maps in the molecular space even with two time points and may be an alternative to differential equation modelling. However, interpretable use cases for single-cell data that can be leveraged for hypothesis generation are still rare. RNA velocity inference via spliced and unspliced read counting in scRNA-seq opens a separate avenue for inference of directional development, even in non-temporal snapshot data (La Manno et al. 2018; Bergen et al. 2020). Observation-wise gradient observations from velocities have already been used as constraints for a distribution matching model in TrajectoryNet (Tong et al. 2020). Such combined models are a promising avenue for modelling both temporal data and splicing states. Similar distribution models of cell populations may also be

useful to model *ex vivo* stimulations, where specific domain coupling observed through reverse phenotyping can be used to constrain the domain transition.

The study of T cell activation and specificity has received much interest in the field of biotechnology because of the relevance of T cells to engineering immune responses. First, immune responses can be engineered through perturbation of cell signalling with drugs, including the important example of PD1-mediated signalling inhibitors (Lei et al. 2020). Second, immune responses can be induced through engineered T cells, such as CAR-T cells (Kuwana et al. 1987). High quality out-of-sample predictions on antigen-TCR binding will likely require more data and a new generation of high-throughput assays that cover larger sets of antigens. However, a major step towards increased throughput has already been taken with multiplexing of T cell specificity measurements in single cells, as discussed in this thesis. Therefore, data-driven advances in this field may become available in the near future. From a modelling point-of-view, amino acid sequence-based models for TCR sequences are part of a category of the rapidly evolving sequence-based deep learning models. Most famously, sequence-based deep learning has been scaled to models with billions of parameters in the transformer model class, such as in GPT-3 (Brown et al. 2020), and has recently made leaps in protein domain structure prediction, shown by AlphaFold (Senior et al. 2020) and its successor AlphaFold2 (Jumper et al. 2021). TCR CDR3 sequences are much smaller than most protein sequences tackled in AlphaFold2, and are indeed only a single loop in each chain of the larger TCR dimer. However, loops are often worse defined in tertiary structures of proteins because of their lack of secondary peptide structures, thus presenting different challenges to those addressed in full protein structure models. Still, there are likely insights and modelling concepts that can be translated from protein structure modelling to TCR-antigen binding event modelling.

4.2.2 Modelling spatial single-cell data using spatial graphs

Models of spatial single-cell data are still in their infancy, with both the experimental capabilities and the modelling frameworks developing very quickly. The scientific journal Nature Methods elected “spatial transcriptomics” as its “Method of the Year” in 2020 (“Method of the Year 2020: Spatially Resolved Transcriptomics” 2021), showing the great potential believed to lie in this field. In this thesis, I studied cell interactions as a core area of biological insights that can be gained from spatial data. Further abstraction of cell and tissue biological questions are likely required to represent the full complexity of emergence in tissues beyond niche phenotypes studied here.

The work presented here characterises overlaps between spatial molecular profiling and graph representation learning approaches (Kipf and Welling 2016). Graphs are very flexible mathematical formalisms that can encode a variety of prior knowledge. Still, further development of graph kernels to reflect the full host of prior knowledge available in cell biology is required. Building on the NCEMs presented here, one may consider constraining edges between cells further by matching ligand and receptor expression (Efremova et al. 2020; Browaeys, Saelens, and Saeys 2020), for example.

High-resolution spatial cellular profiling may also increasingly question the model of a cell as an unstructured bag of molecules, which is typically used for cell-wise average molecular data from dissociated cells. High-resolution assays may be used to describe cells with organelle resolution. These cellular representations with increased spatial resolution may both yield mechanistic insights into cellular function but may also uncover heterogeneity between cells that is lost during averaging. Thus, spatial molecular profiling will continue to provide opportunities for cellular representation learning.

4.2.3 Statistical modelling of gene expression data

Mass-spectrometry-based single-cell proteomics is clearly an interesting avenue for statistical modelling of gene expression data in the future as many aspects of the data generating process are not as well understood as for scRNA-seq data yet. In scRNA-seq data, models of donor and study variability will likely receive further attention. In the study on SARS-CoV2 up-take receptor expression, we found that strong donor and study variations can result in generalised linear models that overfit. Overfitting in generalised linear models on scRNA-seq data was recently also discussed with respect to dispersion modelling (Hafemeister and Satija 2019). Random effects for donor covariates in mixed effect models are a mathematical formalism to constrain these coefficients, but are

also not necessarily free of overfitting and priors are not trivial to use by non-statistician analysts of single-cell data. Moreover, mixed effect models usually require much more computational resources to fit and can be more numerically insatiable, resulting in decreased usability compared to generalised linear models. Another problem with single-cell differential expression analysis as a data exploration tool is that with the increase in number of cells per experiment, even small effects become significant if cells are treated as observations. Accordingly, the selection of small gene sets for downstream investigation is often not only powered by p-values anymore, but also by effect sizes or by prior knowledge. Still, variations of differential expression analysis will likely remain a bread-and-butter analysis option for single-cell data also in the future because of its high interpretability.

Basic statistical models may be increasingly used in the context of multi-modal single-cell experiments to discover causal gene relationships in the future. MPRA data were one of the first experimental set-ups in which RNA was paired with a causally related second modality, plasmid DNA content. Paired single-cell RNA-ATAC measurements may allow for inference of dependencies between enhancer accessibility and gene expression, for example. First steps into this direction were taken in studies that relate the expression of genes controlled by the same transcription factors to each other (Kamimoto, Hoffmann, and Morris 2020) and may now be supplemented by increasingly mechanistic models exploiting paired RNA-ATAC measurements.

4.2.4 Automation of exploratory analysis of single-cell data and building a machine learning community around single-cell data

The automation of unsupervised analysis of single-cell data would empower experimentalists to interact better with the data and would therefore ease integration of biological domain-specific prior knowledge into these analyses. We took a first step into this direction with sfaira. Sfaira also opens exciting avenues in learning complex and potentially interpretable representations of cells by enabling model training on very diverse single-cell data collection for the first time. Efforts on collecting large image datasets enabled models with multi-purpose convolutional layer stacks which then significantly advanced the image-based deep learning field (Krizhevsky, Sutskever, and Hinton 2012), highlighting the relevance of such large structured databases for advances in machine learning. Finally, structured data collections and modelling interfaces reduce barriers for machine learning and statistics researchers to work on single-cell data and will facilitate future interdisciplinary research.

4.2.5 Modelling single-cell epigenetic data

Joint scRNA-seq and scATAC-seq measurements are arguably the most commonly deployed single-cell epigenetic assay and are part of the group of methods which was awarded Method of the Year 2019 by Nature Methods (“Method of the Year 2019: Single-Cell Multimodal Omics” 2020). Not only does this experiment provide unprecedented descriptive information for unsupervised characterisation of cells, it also allows causal models for chromatin regulation of RNA to be built. The syntax of transcription factor binding sites on the genome has recently been modelled with increasingly complex Deep Learning models of DNA sequence (Avsec, Agarwal, et al. 2021; Avsec, Weilert, et al. 2021). Often, MPRA can be used to validate predictions in these settings. Our model of Tn5 specificity lays a basis of a proper treatment of ATAC-data in this context and may give rise to such models of chromatin opening that achieve higher orders of abstraction.

Chapter 5. Publication record

This section contains citations of all publications which I was involved in as part of this thesis.
** Authors contributed equally*

5.1 Publications presented in thesis

The publications listed here and my respective contributions are summarised in Chapter 3.

Publication 1:
Fischer, David S.*, Anna K. Fiedler*, Eric M. Kernfeld, Ryan M. J. Genga, Aimée Bastidas-Ponce, Mostafa Bakhti, Heiko Lickert, Jan Hasenauer, Rene Maehr, and Fabian J. Theis. 2019. “Inferring Population Dynamics from Single-Cell RNA-Sequencing Time Series Data.” *Nature Biotechnology* 37 (4): 461–68, https://doi.org/10.1038/s41587-019-0088-0.

Publication 2:
Fischer, David S., Yihan Wu, Benjamin Schubert, and Fabian J. Theis. 2020. “Predicting Antigen Specificity of Single T Cells Based on TCR CDR3 Regions.” *Molecular Systems Biology* 16 (8): e9416, https://doi.org/10.15252/msb.20199416.

Publication 3:
Fischer, David S.*, Meshal Ansari*, Karolin I. Wagner* et al. 2021. Single-cell RNA sequencing reveals ex vivo signatures of SARS-CoV-2-reactive T cells through ‘reverse phenotyping’. *Nature Communications* 12, 4515. https://doi.org/10.1038/s41467-021-24730-4.

Publication 4:
Fischer, David S., Fabian J. Theis, and Nir Yosef. 2018. “Impulse Model-Based Differential Expression Analysis of Time Course Sequencing Data.” *Nucleic Acids Research* 46 (20): e119, https://doi.org/10.1093/nar/gky675.

Publication 5:
Fischer, David S.*, Leander Dony*, Martin König, Abdul Moeed, Luke Zappia, Lukas Heumos, Sophie Tritschler, Olle Holmberg, Hananeh Aliee, and Fabian J. Theis. 2021. “Sfaira accelerates data and model reuse in single cell genomics.” *Genome Biology* 22, 248. https://doi.org/10.1186/s13059-021-02452-6.

Publication 6:
Ashuach, Tal*, **David S. Fischer***, Anat Kreimer, Nadav Ahituv, Fabian J. Theis, and Nir Yosef. 2019. “MPRAnalyze: Statistical Framework for Massively Parallel Reporter Assays.” *Genome Biology* 20 (1): 183, https://doi.org/10.1186/s13059-019-1787-z.

Preprint 1:
Fischer, David S.*, Anna C. Schaar*, Fabian J. Theis. 2021. “Learning cell communication from spatial graphs of cells.” Preprint available on bioarxiv by *Cold Spring Harbor Laboratory*, https://doi.org/10.1101/2021.07.11.451750.

5.2 Additional publications

Additional publication 1:
Angerer, Philipp, **David S. Fischer**, Fabian J. Theis, Antonio Scialdone, and Carsten Marr. 2020. “Automatic Identification of Relevant Genes from Low-Dimensional Embeddings of Single Cell RNAseq Data.” *Bioinformatics*, Volume 36, Issue 15, Pages 4291–4295, https://doi.org/10.1093/bioinformatics/btaa198.
Contribution: I contributed to the definition of the gradient-based gene relevance metric and contributed to manuscript writing.

Additional publication 2:

Angerer, Philipp, Lukas Simon, Sophie Tritschler, F. Alexander Wolf, **David Fischer**, and Fabian J. Theis. 2017. “Single Cells Make Big Data: New Challenges and Opportunities in Transcriptomics.” *Current Opinion in Systems Biology*, https://doi.org/10.1016/j.coisb.2017.07.004.
Contribution: I contributed to the definition of opportunities and challenges introduced in this manuscript and contributed to manuscript writing.

Additional publication 3:
Ansari, Meshal, **David S. Fischer**, and Fabian J. Theis. 2020. “Learning Tn5 Sequence Bias from ATAC-Seq on Naked Chromatin.” *In Artificial Neural Networks and Machine Learning – ICANN 2020*, 105–14. Springer International Publishing.
Contribution: I designed this project and supervised the master thesis in which this project was majorly performed and contributed to manuscript writing.

Additional publication 4:
Danese, A., Maria L. Richter, Kridsadakorn Chaichoompu, **David S. Fischer**, Fabian J. Theis & Maria Colomé-Tatché. EpiScanpy: integrated single-cell epigenomic analysis. *Nature Communications* 12, 5228 (2021). https://doi.org/10.1038/s41467-021-25131-3.
Contribution: I contributed through supervising a master thesis in which we demonstrated the utility of unsupervised machine learning techniques, such as PCA and t-SNE, on scATAC-seq data. I contributed to manuscript writing.

Additional publication 5:
Dony, Leander, Martin König, **D. Fischer**, and Fabian J. Theis. 2020. “Variational Autoencoders with Flexible Priors Enable Robust Distribution Learning on Single-Cell RNA Sequencing Data.” *In ICML 2020 Workshop on Computational Biology (WCB) Proceedings Paper.* Vol. 37.
Contribution: I co-supervised the Master Thesis on this project and contributed to manuscript writing.

Additional publication 6:
Hetzel, Leon, **David S. Fischer**, Stephan Günnemann, and Fabian J. Theis. 2021. “Graph Representation Learning for Single Cell Biology.” *Current Opinion in Systems Biology*, May. https://doi.org/10.1016/j.coisb.2021.05.008.
Contribution: I contributed to manuscript writing, in particular to defining cell biological use cases of graph learning.

Additional publication 7:
Muus, Christoph, Malte D. Luecken, Gökcen Eraslan, Lisa Sikkema, Avinash Waghray, Graham Heimberg, Yoshihiko Kobayashi, **et al.** 2021. “Single-Cell Meta-Analysis of SARS-CoV-2 Entry Genes across Tissues and Demographics.” *Nature Medicine* 27 (3): 546–59, https://doi.org/10.1038/s41591-020-01227-z.
Contribution: In this study, I co-performed the initial generalised linear model analyses, identified underfitting and overfitting cases.

Additional publication 8:
Palla, Giovanni, Hannah Spitzer, Michal Klein, **David S. Fischer**, Anna Christina Schaar, Louis Benedikt Kuemmerle, Sergei Rybakov, et al. 2020. “Squidpy: A Scalable Framework for Spatial Single Cell Analysis.” *Nature Methods (in press)*. Preprint available on bioarxiv by Cold Spring Harbor Laboratory, https://doi.org/10.1101/2021.02.19.431994.
Contribution: I contributed to planning the project, contributed to code development and manuscript writing.

Additional publication 9:
Tritschler, Sophie, Maren Büttner, **David S. Fischer**, Marius Lange, Volker Bergen, Heiko Lickert, and Fabian J. Theis. 2019. “Concepts and Limitations for Learning Developmental Trajectories from Single Cell Genomics.” *Development* 146 (12), https://doi.org/10.1242/dev.170506.
Contribution: I contributed to manuscript writing.

Additional publication 10:

Palla, Giovanni*, **David S. Fischer***, Aviv Regev, Fabian J. Theis. “Spatial components of molecular tissue biology.”. *Nature Biotechnology (in press).*
Contribution: I co-designed the variance decomposition paradigm used in this review and wrote the manuscript with the other authors.

5.3 Additional manuscripts available as preprint

Additional preprint 1:
Brunner, Andreas-David, Marvin Thielert, Catherine Vasilopoulou, Constantin Ammar, Fabian Coscia, Andreas Mund, Ole Bjeld Horning, Nicolai Bache, Amalia Apalategui, Markus Lubeck, Oliver Raether, Melvin Park, Sabrina Richter, **David S. Fischer**, Fabian J Theis, Florian Meier, Matthias Mann. 2020. “Ultra-High Sensitivity Mass Spectrometry Quantifies Single-Cell Proteome Changes upon Perturbation.” Preprint available on bioarxiv by *Cold Spring Harbor Laboratory*, https://doi.org/10.1101/2020.12.22.423933.
Contribution: I contributed to the design of the overall scRNA-seq analysis strategy and the comparative RNA-protein analysis and contributed to the interpretation and detection of zero-inflation.

5.5 Oral presentations

- Single-cell Omics Germany 2019: Conference talk ”diffxpy: Scalable differential expression analysis for single-cell RNA-seq data”.

5.6 Poster presentations

- Single cell genomics 2019: Conference poster ”T-cell receptor binding affinity prediction”.
- Single cell genomics 2018: Conference poster ”diffxpy: Fast and scalable differential expression analysis”.
- Single cell genomics 2017: Conference poster “Beyond pseudotime: Following T-cell maturation in single-cell RNA-seq (“pseudodynamics”)”.

Chapter 6. References

Adey, Andrew, Hilary G. Morrison, Asan, Xu Xun, Jacob O. Kitzman, Emily H. Turner, Bethany Stackhouse, et al. 2010. “Rapid, Low-Input, Low-Bias Construction of Shotgun Fragment Libraries by High-Density in Vitro Transposition.” *Genome Biology* 11 (12): R119.

Agresti, Alan. 2015. *Foundations of Linear and Generalized Linear Models*. John Wiley & Sons.

Angerer, Philipp, Lukas Simon, Sophie Tritschler, F. Alexander Wolf, David Fischer, and Fabian J. Theis. 2017. “Single Cells Make Big Data: New Challenges and Opportunities in Transcriptomics.” *Current Opinion in Systems Biology*. https://doi.org/10.1016/j.coisb.2017.07.004.

Arnol, Damien, Denis Schapiro, Bernd Bodenmiller, Julio Saez-Rodriguez, and Oliver Stegle. 2019. “Modeling Cell-Cell Interactions from Spatial Molecular Data with Spatial Variance Component Analysis.” *Cell Reports* 29 (1): 202–11.e6.

Arstila, T. P., A. Casrouge, V. Baron, J. Even, J. Kanellopoulos, and P. Kourilsky. 1999. “A Direct Estimate of the Human Alphabeta T Cell Receptor Diversity.” *Science* 286 (5441): 958–61.

Asp, Michaela, Joseph Bergenstråhle, and Joakim Lundeberg. 2020. “Spatially Resolved Transcriptomes—Next Generation Tools for Tissue Exploration.” *BioEssays*. https://doi.org/10.1002/bies.201900221.

Avsec, Žiga, Vikram Agarwal, Daniel Visentin, Joseph R. Ledsam, Agnieszka Grabska-Barwinska, Kyle R. Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R. Kelley. 2021. “Effective Gene Expression Prediction from Sequence by

Integrating Long-Range Interactions.” *Nature Methods* 18 (10): 1196–1203.
Avsec, Žiga, Roman Kreuzhuber, Johnny Israeli, Nancy Xu, Jun Cheng, Avanti Shrikumar, Abhimanyu Banerjee, et al. 2019. “The Kipoi Repository Accelerates Community Exchange and Reuse of Predictive Models for Genomics.” *Nature Biotechnology* 37 (6): 592–600.
Avsec, Žiga, Melanie Weilert, Avanti Shrikumar, Sabrina Krueger, Amr Alexandari, Khyati Dalal, Robin Fropf, et al. 2021. “Base-Resolution Models of Transcription-Factor Binding Reveal Soft Motif Syntax.” *Nature Genetics* 53 (3): 354–66.
Bader, Erik, Adriana Migliorini, Moritz Gegg, Noah Moruzzi, Jantje Gerdes, Sara S. Roscioni, Mostafa Bakhti, et al. 2016. “Identification of Proliferative and Mature β-Cells in the Islets of Langerhans.” *Nature* 535 (7612): 430–34.
Bagaev, Dmitry V., Renske M. A. Vroomans, Jerome Samir, Ulrik Stervbo, Cristina Rius, Garry Dolton, Alexander Greenshields-Watson, et al. 2020. “VDJdb in 2019: Database Extension, New Analysis Infrastructure and a T-Cell Receptor Motif Compendium.” *Nucleic Acids Research* 48 (D1): D1057–62.
Ballnus, Benjamin, Sabine Hug, Kathrin Hatz, Linus Görlitz, Jan Hasenauer, and Fabian J. Theis. 2017. “Comprehensive Benchmarking of Markov Chain Monte Carlo Methods for Dynamical Systems.” *BMC Systems Biology* 11 (1): 63.
Bandura, Dmitry R., Vladimir I. Baranov, Olga I. Ornatsky, Alexei Antonov, Robert Kinach, Xudong Lou, Serguei Pavlov, Sergey Vorobiev, John E. Dick, and Scott D. Tanner. 2009. “Mass Cytometry: Technique for Real Time Single Cell Multitarget Immunoassay Based on Inductively Coupled Plasma Time-of-Flight Mass Spectrometry.” *Analytical Chemistry* 81 (16): 6813–22.
Baran-Gale, Jeanette, Jeremy E. Purvis, and Praveen Sethupathy. 2016. “An Integrative Transcriptomics Approach Identifies miR-503 as a Candidate Master Regulator of the Estrogen Response in MCF-7 Breast Cancer Cells.” *RNA* 22 (10): 1592–1603.
Baum, Leonard E., and Ted Petrie. 1966. “Statistical Inference for Probabilistic Functions of Finite State Markov Chains.” *Annals of Mathematical Statistics* 37 (6): 1554–63.
Bentzen, Amalie Kai, Andrea Marion Marquard, Rikke Lyngaa, Sunil Kumar Saini, Sofie Ramskov, Marco Donia, Lina Such, et al. 2016. “Large-Scale Detection of Antigen-Specific T Cells Using Peptide-MHC-I Multimers Labeled with DNA Barcodes.” *Nature Biotechnology* 34 (10): 1037–45.
Bergen, Volker, Marius Lange, Stefan Peidli, F. Alexander Wolf, and Fabian J. Theis. 2020. “Generalizing RNA Velocity to Transient Cell States through Dynamical Modeling.” *Nature Biotechnology* 38 (12): 1408–14.
Blank, Christian U., W. Nicholas Haining, Werner Held, Patrick G. Hogan, Axel Kallies, Enrico Lugli, Rachel C. Lynn, et al. 2019. “Defining ‘T Cell Exhaustion.’” *Nature Reviews. Immunology* 19 (11): 665–74.
Blei, David M., Alp Kucukelbir, and Jon D. McAuliffe. 2017. “Variational Inference: A Review for Statisticians.” *Journal of the American Statistical Association*. https://doi.org/10.1080/01621459.2017.1285773.
Böttcher, Anika, Maren Büttner, Sophie Tritschler, Michael Sterr, Alexandra Aliluev, Lena Oppenländer, Ingo Burtscher, et al. 2021. “Non-Canonical Wnt/PCP Signalling Regulates Intestinal Stem Cell Lineage Priming towards Enteroendocrine and Paneth Cell Fates.” *Nature Cell Biology* 23 (1): 23–31.
Broadbent, Kate M., Jill C. Broadbent, Ulf Ribacke, Dyann Wirth, John L. Rinn, and Pardis C. Sabeti. 2015. “Strand-Specific RNA Sequencing in Plasmodium Falciparum Malaria Identifies Developmentally Regulated Long Non-Coding RNA and Circular RNA.” *BMC Genomics* 16 (June): 454.
Browaeys, Robin, Wouter Saelens, and Yvan Saeys. 2020. “NicheNet: Modeling Intercellular Communication by Linking Ligands to Target Genes.” *Nature Methods* 17 (2): 159–62.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. “Language Models Are Few-Shot Learners.” *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2005.14165.

Broyden, C. G. 1970. “The Convergence of a Class of Double-Rank Minimization Algorithms 1. General Considerations.” *IMA Journal of Applied Mathematics*. https://doi.org/10.1093/imamat/6.1.76.

Bruna, Joan, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. 2013. “Spectral Networks and Locally Connected Networks on Graphs.” *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1312.6203.

Buenrostro, Jason D., Paul G. Giresi, Lisa C. Zaba, Howard Y. Chang, and William J. Greenleaf. 2013. “Transposition of Native Chromatin for Fast and Sensitive Epigenomic Profiling of Open Chromatin, DNA-Binding Proteins and Nucleosome Position.” *Nature Methods* 10 (12): 1213–18.

Buenrostro, Jason D., Beijing Wu, Ulrike M. Litzenburger, Dave Ruff, Michael L. Gonzales, Michael P. Snyder, Howard Y. Chang, and William J. Greenleaf. 2015. “Single-Cell Chromatin Accessibility Reveals Principles of Regulatory Variation.” *Nature* 523 (7561): 486–90.

Büttner, Maren, Zhichao Miao, F. Alexander Wolf, Sarah A. Teichmann, and Fabian J. Theis. 2019. “A Test Metric for Assessing Single-Cell RNA-Seq Batch Correction.” *Nature Methods* 16 (1): 43–49.

Büttner, M., J. Ostner, C. L. Müller, F. J. Theis, and B. Schubert. n.d. “scCODA: A Bayesian Model for Compositional Single-Cell Data Analysis.” https://doi.org/10.1101/2020.12.14.422688.

Cao, Junyue, Jonathan S. Packer, Vijay Ramani, Darren A. Cusanovich, Chau Huynh, Riza Daza, Xiaojie Qiu, et al. 2017. “Comprehensive Single-Cell Transcriptional Profiling of a Multicellular Organism.” *Science* 357 (6352): 661–67.

De Simone, Marco, Grazisa Rossetti, and Massimiliano Pagani. 2018. “Single Cell T Cell Receptor Sequencing: Techniques and Future Challenges.” *Frontiers in Immunology* 9 (July): 1638.

Diehl, Alexander D., Terrence F. Meehan, Yvonne M. Bradford, Matthew H. Brush, Wasila M. Dahdul, David S. Dougall, Yongqun He, et al. 2016. “The Cell Ontology 2016: Enhanced Content, Modularization, and Ontology Interoperability.” *Journal of Biomedical Semantics* 7 (1): 44.

Dries, Ruben, Qian Zhu, Rui Dong, Chee-Huat Linus Eng, Huipeng Li, Kan Liu, Yuntian Fu, et al. n.d. “Giotto, a Toolbox for Integrative Analysis and Visualization of Spatial Expression Data.” https://doi.org/10.1101/701680.

Duvenaud, David, Dougal Maclaurin, Jorge Aguilera-Iparraguirre, Rafael Gómez-Bombarelli, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P. Adams. 2015. “Convolutional Networks on Graphs for Learning Molecular Fingerprints.” *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1509.09292.

Efremova, Mirjana, Miquel Vento-Tormo, Sarah A. Teichmann, and Roser Vento-Tormo. 2020. “CellPhoneDB: Inferring Cell–cell Communication from Combined Expression of Multi-Subunit Ligand–receptor Complexes.” *Nature Protocols* 15 (4): 1484–1506.

Eng, Chee-Huat Linus, Michael Lawson, Qian Zhu, Ruben Dries, Noushin Koulena, Yodai Takei, Jina Yun, et al. 2019. “Transcriptome-Scale Super-Resolved Imaging in Tissues by RNA seqFISH.” *Nature* 568 (7751): 235–39.

Eraslan, Gökcen, Lukas M. Simon, Maria Mircea, Nikola S. Mueller, and Fabian J. Theis. 2019. “Single-Cell RNA-Seq Denoising Using a Deep Count Autoencoder.” *Nature Communications* 10 (1): 390.

Finak, Greg, Andrew McDavid, Masanao Yajima, Jingyuan Deng, Vivian Gersuk, Alex K. Shalek, Chloe K. Slichter, et al. 2015. “MAST: A Flexible Statistical Framework for

Assessing Transcriptional Changes and Characterizing Heterogeneity in Single-Cell RNA Sequencing Data." *Genome Biology* 16 (December): 278.

Fischer, David S., Anna K. Fiedler, Eric M. Kernfeld, Ryan M. J. Genga, Aimée Bastidas-Ponce, Mostafa Bakhti, Heiko Lickert, Jan Hasenauer, Rene Maehr, and Fabian J. Theis. 2019. "Inferring Population Dynamics from Single-Cell RNA-Sequencing Time Series Data." *Nature Biotechnology* 37 (4): 461–68.

Fletcher, R. 1970. "A New Approach to Variable Metric Algorithms." *The Computer Journal*. https://doi.org/10.1093/comjnl/13.3.317.

Flower, Darren R., Kanchan Phadwal, Isabel K. Macdonald, Peter V. Coveney, Matthew N. Davies, and Shunzhou Wan. 2010. "T-Cell Epitope Prediction and Immune Complex Simulation Using Molecular Dynamics: State of the Art and Persisting Challenges." *Immunome Research* 6 Suppl 2 (November): S4.

Fröhlich, Fabian, Barbara Kaltenbacher, Fabian J. Theis, and Jan Hasenauer. 2017. "Scalable Parameter Estimation for Genome-Scale Biochemical Reaction Networks." *PLoS Computational Biology* 13 (1): e1005331.

Fröhlich, Fabian, Daniel Weindl, Yannik Schälte, Dilan Pathirana, Łukasz Paszkowski, Glenn Terje Lines, Paul Stapor, and Jan Hasenauer. 2021. "AMICI: High-Performance Sensitivity Analysis for Large Ordinary Differential Equation Models." *Bioinformatics* , April. https://doi.org/10.1093/bioinformatics/btab227.

Gierahn, Todd M., Marc H. Wadsworth 2nd, Travis K. Hughes, Bryan D. Bryson, Andrew Butler, Rahul Satija, Sarah Fortune, J. Christopher Love, and Alex K. Shalek. 2017. "Seq-Well: Portable, Low-Cost RNA Sequencing of Single Cells at High Throughput." *Nature Methods* 14 (4): 395–98.

Giesen, Charlotte, Hao A. O. Wang, Denis Schapiro, Nevena Zivanovic, Andrea Jacobs, Bodo Hattendorf, Peter J. Schüffler, et al. 2014. "Highly Multiplexed Imaging of Tumor Tissues with Subcellular Resolution by Mass Cytometry." *Nature Methods* 11 (4): 417–22.

Glanville, Jacob, Huang Huang, Allison Nau, Olivia Hatton, Lisa E. Wagar, Florian Rubelt, Xuhuai Ji, et al. 2017. "Identifying Specificity Groups in the T Cell Receptor Repertoire." *Nature* 547 (7661): 94–98.

Glorot, Xavier, and Yoshua Bengio. 2010. "Understanding the Difficulty of Training Deep Feedforward Neural Networks." In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, edited by Yee Whye Teh and Mike Titterington, 9:249–56. Proceedings of Machine Learning Research. Chia Laguna Resort, Sardinia, Italy: PMLR.

Goldfarb, Donald. 1970. "A Family of Variable-Metric Methods Derived by Variational Means." *Mathematics of Computation*. https://doi.org/10.1090/s0025-5718-1970-0258249-6.

Goltsev, Yury, Nikolay Samusik, Julia Kennedy-Darling, Salil Bhate, Matthew Hale, Gustavo Vazquez, Sarah Black, and Garry P. Nolan. 2018. "Deep Profiling of Mouse Splenic Architecture with CODEX Multiplexed Imaging." *Cell* 174 (4): 968–81.e15.

Goodwin, Sara, John D. McPherson, and W. Richard McCombie. 2016. "Coming of Age: Ten Years of next-Generation Sequencing Technologies." *Nature Reviews. Genetics* 17 (6): 333–51.

Graveley, Brenton R., Angela N. Brooks, Joseph W. Carlson, Michael O. Duff, Jane M. Landolin, Li Yang, Carlo G. Artieri, et al. 2010. "The Developmental Transcriptome of Drosophila Melanogaster." *Nature* 471 (7339): 473–79.

Hafemeister, Christoph, and Rahul Satija. 2019. "Normalization and Variance Stabilization of Single-Cell RNA-Seq Data Using Regularized Negative Binomial Regression." *Genome Biology* 20 (1): 296.

Hagemann-Jensen, Michael, Christoph Ziegenhain, Ping Chen, Daniel Ramsköld, Gert-Jan

Hendriks, Anton J. M. Larsson, Omid R. Faridani, and Rickard Sandberg. 2020. “Single-Cell RNA Counting at Allele and Isoform Resolution Using Smart-seq3.” *Nature Biotechnology* 38 (6): 708–14.

Haghverdi, Laleh, Florian Buettner, and Fabian J. Theis. 2015. “Diffusion Maps for High-Dimensional Single-Cell Analysis of Differentiation Data.” *Bioinformatics* 31 (18): 2989–98.

Haghverdi, Laleh, Maren Büttner, F. Alexander Wolf, Florian Buettner, and Fabian J. Theis. 2016. “Diffusion Pseudotime Robustly Reconstructs Lineage Branching.” *Nature Methods* 13 (10): 845–48.

Haghverdi, Laleh, Aaron T. L. Lun, Michael D. Morgan, and John C. Marioni. 2018. “Batch Effects in Single-Cell RNA-Sequencing Data Are Corrected by Matching Mutual Nearest Neighbors.” *Nature Biotechnology* 36 (5): 421–27.

Han, Xiaoping, Renying Wang, Yincong Zhou, Lijiang Fei, Huiyu Sun, Shujing Lai, Assieh Saadatpour, et al. 2018. “Mapping the Mouse Cell Atlas by Microwell-Seq.” *Cell* 173 (5): 1307.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2013. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media.

Hinton, Geoffrey E., Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R. Salakhutdinov. 2012. “Improving Neural Networks by Preventing Co-Adaptation of Feature Detectors.” *arXiv [cs.NE]*. arXiv. http://arxiv.org/abs/1207.0580.

Hochreiter, S., and J. Schmidhuber. 1997. “Long Short-Term Memory.” *Neural Computation* 9 (8): 1735–80.

Holmes, E. E., M. A. Lewis, J. E. Banks, and R. R. Veit. 1994. “Partial Differential Equations in Ecology: Spatial Interactions and Population Dynamics.” *Ecology*. https://doi.org/10.2307/1939378.

Inoue, Fumitaka, and Nadav Ahituv. 2015. “Decoding Enhancers Using Massively Parallel Reporter Assays.” *Genomics* 106 (3): 159–64.

Jain, Miten, Hugh E. Olsen, Benedict Paten, and Mark Akeson. 2016. “The Oxford Nanopore MinION: Delivery of Nanopore Sequencing to the Genomics Community.” *Genome Biology* 17 (1): 239.

Jovanovic, Marko, Michael S. Rooney, Philipp Mertins, Dariusz Przybylski, Nicolas Chevrier, Rahul Satija, Edwin H. Rodriguez, et al. 2015. “Immunogenetics. Dynamic Profiling of the Protein Life Cycle in Response to Pathogens.” *Science* 347 (6226): 1259038.

Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, et al. 2021. “Highly Accurate Protein Structure Prediction with AlphaFold.” *Nature* 596 (7873): 583–89.

Kafri, Ran, Jason Levy, Miriam B. Ginzberg, Seungeun Oh, Galit Lahav, and Marc W. Kirschner. 2013. “Dynamics Extracted from Fixed Cells Reveal Feedback Linking Cell Growth to Cell Cycle.” *Nature* 494 (7438): 480–83.

Kamimoto, K., C. M. Hoffmann, and S. A. Morris. 2020. “CellOracle: Dissecting Cell Identity via Network Inference and in Silico Gene Perturbation.” *bioRxiv*. https://www.biorxiv.org/content/10.1101/2020.02.17.947416v1.abstract.

Kharchenko, Peter V., Lev Silberstein, and David T. Scadden. 2014. “Bayesian Approach to Single-Cell Differential Expression Analysis.” *Nature Methods* 11 (7): 740–42.

Kingma, Diederik P., and Jimmy Ba. 2014. “Adam: A Method for Stochastic Optimization.” *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1412.6980.

Kipf, Thomas N., and Max Welling. 2016. “Semi-Supervised Classification with Graph Convolutional Networks.” *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1609.02907.

Klambauer, Günter, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. 2017. “Self-Normalizing Neural Networks.” In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 972–81.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. 2012. “ImageNet Classification with Deep Convolutional Neural Networks.” In *Advances in Neural Information Processing Systems*, edited by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger. Vol. 25. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

Kuritz, K., D. Stöhr, N. Pollak, and F. Allgöwer. 2017. “On the Relationship between Cell Cycle Analysis with Ergodic Principles and Age-Structured Cell Population Models.” *Journal of Theoretical Biology* 414 (February): 91–102.

Kuwana, Yoshihisa, Yoshihiro Asakura, Naoko Utsunomiya, Mamoru Nakanishi, Yohji Arata, Seiga Itoh, Fumihiko Nagase, and Yoshikazu Kurosawa. 1987. “Expression of Chimeric Receptor Composed of Immunoglobulin-Derived V Resions and T-Cell Receptor-Derived C Regions.” *Biochemical and Biophysical Research Communications* 149 (3): 960–68.

La Manno, Gioele, Ruslan Soldatov, Amit Zeisel, Emelie Braun, Hannah Hochgerner, Viktor Petukhov, Katja Lidschreiber, et al. 2018. “RNA Velocity of Single Cells.” *Nature* 560 (7719): 494–98.

Lei, Qingyang, Dan Wang, Kai Sun, Liping Wang, and Yi Zhang. 2020. “Resistance Mechanisms of Anti-PD1/PDL1 Therapy in Solid Tumors.” *Frontiers in Cell and Developmental Biology* 8 (July): 672.

Lieberman-Aiden, Erez, Nynke L. van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, et al. 2009. “Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome.” *Science* 326 (5950): 289–93.

Litviňuková, Monika, Carlos Talavera-López, Henrike Maatz, Daniel Reichart, Catherine L. Worth, Eric L. Lindberg, Masatoshi Kanda, et al. 2020. “Cells of the Adult Human Heart.” *Nature* 588 (7838): 466–72.

Lopez, Romain, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. 2018. “Deep Generative Modeling for Single-Cell Transcriptomics.” *Nature Methods* 15 (12): 1053–58.

Lotfollahi, M., M. Naghipourfar, M. D. Luecken, and M. Khajavi. 2020. “Query to Reference Single-Cell Integration with Transfer Learning.” *bioRxiv*. https://www.biorxiv.org/content/10.1101/2020.07.16.205997v1.abstract.

Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. “Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2.” *Genome Biology* 15 (12): 550.

Luecken, Malte D., and Fabian J. Theis. 2019. “Current Best Practices in Single-Cell RNA-Seq Analysis: A Tutorial.” *Molecular Systems Biology* 15 (6): e8746.

Maaten, Laurens van der, and Geoffrey Hinton. 2008. “Visualizing Data Using T-SNE.” *Journal of Machine Learning Research: JMLR* 9 (Nov): 2579–2605.

Macosko, Evan Z., Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, et al. 2015. “Highly Parallel Genome-Wide Expression Profiling of Individual Cells Using Nanoliter Droplets.” *Cell* 161 (5): 1202–14.

McInnes, Iain B., and Ellen M. Gravallese. 2021. “Immune-Mediated Inflammatory Disease Therapeutics: Past, Present and Future.” *Nature Reviews. Immunology* 21 (10): 680–86.

McInnes, Leland, John Healy, and James Melville. 2018. “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.” *arXiv [stat.ML]*. arXiv. http://arxiv.org/abs/1802.03426.

“Method of the Year 2019: Single-Cell Multimodal Omics.” 2020. *Nature Methods* 17 (1): 1.

“Method of the Year 2020: Spatially Resolved Transcriptomics.” 2021. *Nature Methods* 18 (1): 1.

Mimitou, Eleni P., Anthony Cheng, Antonino Montalbano, Stephanie Hao, Marlon Stoeckius, Mateusz Legut, Timothy Roush, et al. 2019. “Multiplexed Detection of Proteins, Transcriptomes, Clonotypes and CRISPR Perturbations in Single Cells.” *Nature Methods* 16 (5): 409–12.

Montemurro, Alessandro, Viktoria Schuster, Helle Rus Povlsen, Amalie Kai Bentzen, Vanessa Jurtz, William D. Chronister, Austin Crinklaw, et al. 2021. “NetTCR-2.0 Enables Accurate Prediction of TCR-Peptide Binding by Using Paired TCRα and β Sequence Data.” *Communications Biology* 4 (1): 1060.

Myint, Leslie, Dimitrios G. Avramopoulos, Loyal A. Goff, and Kasper D. Hansen. 2019. “Linear Models Enable Powerful Differential Activity Analysis in Massively Parallel Reporter Assays.” *BMC Genomics* 20 (1): 209.

Palla, Giovanni, Hannah Spitzer, Michal Klein, David Fischer, Anna Christina Schaar, Louis Benedikt Kuemmerle, Sergei Rybakov, et al. n.d. “Squidpy: A Scalable Framework for Spatial Single Cell Analysis.” https://doi.org/10.1101/2021.02.19.431994.

Picelli, Simone, Åsa K. Björklund, Omid R. Faridani, Sven Sagasser, Gösta Winberg, and Rickard Sandberg. 2013. “Smart-seq2 for Sensitive Full-Length Transcriptome Profiling in Single Cells.” *Nature Methods* 10 (11): 1096–98.

Polański, Krzysztof, Matthew D. Young, Zhichao Miao, Kerstin B. Meyer, Sarah A. Teichmann, and Jong-Eun Park. 2020. “BBKNN: Fast Batch Alignment of Single Cell Transcriptomes.” *Bioinformatics* 36 (3): 964–65.

“PyTorch.” n.d. Accessed April 25, 2021. https://pytorch.org/.

Quake, Stephen R. 2021. “The Cell as a Bag of RNA.” *Trends in Genetics: TIG* 37 (12): 1064–68.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. “Language Models Are Unsupervised Multitask Learners.” *OpenAI Blog* 1 (8): 9.

Ramsköld, Daniel, Shujun Luo, Yu-Chieh Wang, Robin Li, Qiaolin Deng, Omid R. Faridani, Gregory A. Daniels, et al. 2012. “Full-Length mRNA-Seq from Single-Cell Levels of RNA and Individual Circulating Tumor Cells.” *Nature Biotechnology* 30 (8): 777–82.

Rao, Anjali, Dalia Barkley, Gustavo S. França, and Itai Yanai. 2021. “Exploring Tissue Architecture Using Spatial Transcriptomics.” *Nature* 596 (7871): 211–20.

Rao, C. Radhakrishna. 1945. “Information and the Accuracy Attainable in the Estimation of Statistical Parameters.” *Reson. J. Sci. Educ* 20: 78–90.

Raza Ali, H., Hartland W. Jackson, Vito R. T. Zanotelli, Esther Danenberg, Jana R. Fischer, Helen Bardwell, Elena Provenzano, et al. 2020. “Imaging Mass Cytometry and Multiplatform Genomics Define the Phenogenomic Landscape of Breast Cancer.” *Nature Cancer* 1 (2): 163–75.

Regev, Aviv, Sarah Teichmann, Orit Rozenblatt-Rosen, Michael Stubbington, Kristin Ardlie, Ido Amit, Paola Arlotta, et al. 2018. “The Human Cell Atlas White Paper.” *arXiv [q-bio.TO]*. arXiv. http://arxiv.org/abs/1810.05192.

Risso, D., F. Perraudeau, S. Gribkova, S. Dudoit, and J. P. Vert. 2017. “ZINB-WaVE: A General and Flexible Method for Signal Extraction from Single-Cell RNA-Seq Data.” *BioRxiv*. https://www.biorxiv.org/content/10.1101/125112v2.abstract.

Ritchie, Matthew E., Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. 2015. “Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies.” *Nucleic Acids Research* 43 (7): e47.

Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. 1986. “Learning Representations by Back-Propagating Errors.” *Nature* 323 (6088): 533–36.

Sachs, Stephan, Aimée Bastidas-Ponce, Sophie Tritschler, Mostafa Bakhti, Anika Böttcher, Miguel A. Sánchez-Garrido, Marta Tarquis-Medina, et al. 2020. “Targeted Pharmacological Therapy Restores β-Cell Function for Diabetes Remission.” *Nature Metabolism* 2 (2): 192–209.

Sander, Jil, Joachim L. Schultze, and Nir Yosef. 2017. “ImpulseDE: Detection of Differentially Expressed Genes in Time Series Data Using Impulse Models.” *Bioinformatics* 33 (5): 757–59.

Sanger, F., S. Nicklen, and A. R. Coulson. 1977. “DNA Sequencing with Chain-Terminating Inhibitors.” *Proceedings of the National Academy of Sciences of the United States of America* 74 (12): 5463–67.

Schulte-Schrepping, Jonas, Nico Reusch, Daniela Paclik, Kevin Baßler, Stephan Schlickeiser, Bowen Zhang, Benjamin Krämer, et al. 2020. “Severe COVID-19 Is Marked by a Dysregulated Myeloid Cell Compartment.” *Cell* 182 (6): 1419–40.e23.

Senior, Andrew W., Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, et al. 2020. “Improved Protein Structure Prediction Using Potentials from Deep Learning.” *Nature* 577 (7792): 706–10.

Setty, Manu, Michelle D. Tadmor, Shlomit Reich-Zeliger, Omer Angel, Tomer Meir Salame, Pooja Kathail, Kristy Choi, Sean Bendall, Nir Friedman, and Dana Pe’er. 2016. “Wishbone Identifies Bifurcating Developmental Trajectories from Single-Cell Data.” *Nature Biotechnology* 34 (6): 637–45.

Shanno, D. F. 1970. “Conditioning of Quasi-Newton Methods for Function Minimization.” *Mathematics of Computation*. https://doi.org/10.1090/s0025-5718-1970-0274029-x.

Shugay, Mikhail, Dmitriy V. Bagaev, Ivan V. Zvyagin, Renske M. Vroomans, Jeremy Chase Crawford, Garry Dolton, Ekaterina A. Komech, et al. 2018. “VDJdb: A Curated Database of T-Cell Receptor Sequences with Known Antigen Specificity.” *Nucleic Acids Research* 46 (D1): D419–27.

Song, Lingyun, and Gregory E. Crawford. 2010. “DNase-Seq: A High-Resolution Technique for Mapping Active Gene Regulatory Elements across the Genome from Mammalian Cells.” *Cold Spring Harbor Protocols* 2010 (2): db.prot5384.

Stapor, Paul, Daniel Weindl, Benjamin Ballnus, Sabine Hug, Carolin Loos, Anna Fiedler, Sabrina Krause, et al. 2018. “PESTO: Parameter EStimation TOolbox.” *Bioinformatics* 34 (4): 705–7.

Stein-O’Brien, Genevieve L., Brian S. Clark, Thomas Sherman, Cristina Zibetti, Qiwen Hu, Rachel Sealfon, Sheng Liu, et al. 2019. “Decomposing Cell Identity for Transfer Learning across Cellular Measurements, Platforms, Tissues, and Species.” *Cell Systems* 8 (5): 395–411.e8.

Sterner, Robert C., and Rosalie M. Sterner. 2021. “CAR-T Cell Therapy: Current Limitations and Potential Strategies.” *Blood Cancer Journal* 11 (4): 69.

Stoeckius, Marlon, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K. Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert. 2017. “Simultaneous Epitope and Transcriptome Measurement in Single Cells.” *Nature Methods* 14 (9): 865–68.

Street, Kelly, Davide Risso, Russell B. Fletcher, Diya Das, John Ngai, Nir Yosef, Elizabeth Purdom, and Sandrine Dudoit. 2018. “Slingshot: Cell Lineage and Pseudotime Inference for Single-Cell Transcriptomics.” *BMC Genomics* 19 (1): 477.

Sturm, Gregor, Tamas Szabo, Georgios Fotakis, Marlene Haider, Dietmar Rieder, Zlatko Trajanoski, and Francesca Finotello. 2020. “Scirpy: A Scanpy Extension for Analyzing Single-Cell T-Cell Receptor-Sequencing Data.” *Bioinformatics* 36 (18): 4817–18.

Svensson, Valentine. 2020. “Droplet scRNA-Seq Is Not Zero-Inflated.” *Nature Biotechnology* 38 (2): 147–50.

Sykes, David B., Youmna S. Kfoury, François E. Mercier, Mathias J. Wawer, Jason M. Law, Mark K. Haynes, Timothy A. Lewis, et al. 2016. “Inhibition of Dihydroorotate Dehydrogenase Overcomes Differentiation Blockade in Acute Myeloid Leukemia.” *Cell* 167 (1): 171–86.e15.

Tabula Muris Consortium. 2020. “A Single-Cell Transcriptomic Atlas Characterizes Ageing

Tissues in the Mouse." *Nature* 583 (7817): 590–95.
Tanevski, Jovan, Attila Gabor, Ricardo Ramirez Flores, Denis Schapiro, and Julio Saez-Rodriguez. n.d. "Explainable Multi-View Framework for Dissecting Intercellular Signaling from Highly Multiplexed Spatial Data." https://doi.org/10.21203/rs.3.rs-735362/v1.
Tang, Fuchou, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, et al. 2009. "mRNA-Seq Whole-Transcriptome Analysis of a Single Cell." *Nature Methods* 6 (5): 377–82.
"TensorFlow." n.d. Accessed April 25, 2021. https://www.tensorflow.org/.
Tong, Alexander, Jessie Huang, Guy Wolf, David Van Dijk, and Smita Krishnaswamy. 2020. "Trajectorynet: A Dynamic Optimal Transport Network for Modeling Cellular Dynamics." In *International Conference on Machine Learning*, 9526–36. PMLR.
Trapnell, Cole, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J. Lennon, Kenneth J. Livak, Tarjei S. Mikkelsen, and John L. Rinn. 2014. "The Dynamics and Regulators of Cell Fate Decisions Are Revealed by Pseudotemporal Ordering of Single Cells." *Nature Biotechnology* 32 (4): 381–86.
Travaglini, Kyle J., Ahmad N. Nabhan, Lolita Penland, Rahul Sinha, Astrid Gillich, Rene V. Sit, Stephen Chang, et al. 2020. "A Molecular Cell Atlas of the Human Lung from Single-Cell RNA Sequencing." *Nature* 587 (7835): 619–25.
Tritschler, Sophie, Maren Büttner, David S. Fischer, Marius Lange, Volker Bergen, Heiko Lickert, and Fabian J. Theis. 2019. "Concepts and Limitations for Learning Developmental Trajectories from Single Cell Genomics." *Development* 146 (12). https://doi.org/10.1242/dev.170506.
Valouev, Anton, David S. Johnson, Andreas Sundquist, Catherine Medina, Elizabeth Anton, Serafim Batzoglou, Richard M. Myers, and Arend Sidow. 2008. "Genome-Wide Analysis of Transcription Factor Binding Sites Based on ChIP-Seq Data." *Nature Methods* 5 (9): 829–34.
Van Regenmortel, Marc H. V. 2004. "Reductionism and Complexity in Molecular Biology: Scientists Now Have the Tools to Unravel Biological Complexity and Overcome the Limitations of Reductionism." *EMBO Reports* 5 (11): 1016–20.
Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Ł. Ukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need." In *Advances in Neural Information Processing Systems 30*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, 5998–6008. Curran Associates, Inc.
Veličković, Petar, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2017. "Graph Attention Networks." *arXiv [stat.ML]*. arXiv. http://arxiv.org/abs/1710.10903.
Venables, W. N., and B. D. Ripley. 2002. "Modern Applied Statistics with S." *Statistics and Computing*. https://doi.org/10.1007/978-0-387-21706-2.
Vita, Randi, Swapnil Mahajan, James A. Overton, Sandeep Kumar Dhanda, Sheridan Martini, Jason R. Cantrell, Daniel K. Wheeler, Alessandro Sette, and Bjoern Peters. 2019. "The Immune Epitope Database (IEDB): 2018 Update." *Nucleic Acids Research* 47 (D1): D339–43.
Wald, Abraham. 1943. *Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations Is Large*.
Weinreb, Caleb, Samuel Wolock, Betsabeh K. Tusi, Merav Socolovsky, and Allon M. Klein. 2018. "Fundamental Limits on Dynamic Inference from Single-Cell Snapshots." *Proceedings of the National Academy of Sciences of the United States of America* 115 (10): E2467–76.
Wilkinson, Mark D., Michel Dumontier, I. Jsbrand Jan Aalbersberg, Gabrielle Appleton,

Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. “The FAIR Guiding Principles for Scientific Data Management and Stewardship.” *Scientific Data* 3 (March): 160018.

Wolchok, Jedd D. 2021. “Checkpoint Blockade: The End of the Beginning.” *Nature Reviews. Immunology* 21 (10): 621.

Wolf, F. Alexander, Fiona K. Hamey, Mireya Plass, Jordi Solana, Joakim S. Dahlin, Berthold Göttgens, Nikolaus Rajewsky, Lukas Simon, and Fabian J. Theis. 2019. “PAGA: Graph Abstraction Reconciles Clustering with Trajectory Inference through a Topology Preserving Map of Single Cells.” *Genome Biology* 20 (1): 59.

Xia, Chenglong, Jean Fan, George Emanuel, Junjie Hao, and Xiaowei Zhuang. 2019. “Spatial Transcriptome Profiling by MERFISH Reveals Subcellular RNA Compartmentalization and Cell Cycle-Dependent Gene Expression.” *Proceedings of the National Academy of Sciences of the United States of America* 116 (39): 19490–99.

Xu, Chenling, Romain Lopez, Edouard Mehlman, Jeffrey Regier, Michael I. Jordan, and Nir Yosef. 2021. “Probabilistic Harmonization and Annotation of Single-Cell Transcriptomics Data with Deep Generative Models.” *Molecular Systems Biology* 17 (1): e9620.

Yuan, Ye, and Ziv Bar-Joseph. 2020. “GCNG: Graph Convolutional Networks for Inferring Gene Interaction from Spatial Transcriptomics Data.” *Genome Biology* 21 (1): 300.

Yui, Mary A., and Ellen V. Rothenberg. 2014. “Developmental Gene Networks: A Triathlon on the Course to T Cell Identity.” *Nature Reviews Immunology*. https://doi.org/10.1038/nri3702.

Zhu, Jun-Yan, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. “Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks.” In *Proceedings of the IEEE International Conference on Computer Vision*, 2223–32.

Zilionis, Rapolas, Juozas Nainys, Adrian Veres, Virginia Savova, David Zemmour, Allon M. Klein, and Linas Mazutis. 2017. “Single-Cell Barcoding and Sequencing Using Droplet Microfluidics.” *Nature Protocols* 12 (1): 44–73.

Zou, Hui, and Trevor Hastie. 2005. “Regularization and Variable Selection via the Elastic Net.” *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 67 (2): 301–20.