Learning cellular hierarchies through structured topic modeling

by

Patricia Er Ye

B.Sc., The University of British Columbia, 2021

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

The Faculty of Graduate and Postdoctoral Studies

(Bioinformatics)

THE UNIVERSITY OF BRITISH COLUMBIA
(Vancouver)

April 2024

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the thesis entitled:

Learning cellular hierarchies through structured topic modeling

submitted by **Patricia Er Ye** in partial fulfillment of the requirements for the degree of **Master of Science** in **Bioinformatics**.

Examining Committee:

Yongjin P. Park, Assistant Professor, Statistics and Pathology & Laboratory Medicine, UBC
Supervisor

Ramon Klein Geltink, Assistant Professor, Pathology & Laboratory Medicine, UBC
Co-Supervisor

Adi Steif, Assistant Professor, Medical Genetics, UBC
Supervisory Committee Member

Geoffrey Schiebinger, Assistant Professor, Mathematics, UBC
Supervisory Committee Member

Abstract

The human immune system relies on the function and balance of various immune cell subsets and their interactions. Immune cells undergo a series of differentiation steps following a lineage-tree structure stemming from hematopoietic stem cells to reach their mature cell state. During differentiation of immune cells in both homeostasis and pathological processes, many cellular features, including gene expression patterns, are shared by fully differentiated immune cell sub-types. The process of immune cell differentiation is complex and not fully understood. Additionally, aberrant function and balance plays a major contributing role in the pathogenesis of many immunological disorders, including systemic lupus erythematosus.

In this thesis, I propose `LaRCH`, a tree-structured neural topic model as a method to quantitatively characterize shared hierarchical features between cell subsets. In this model, single-cell gene expression profiles are represented by a mixture of topics consisting of latent features that follow an underlying tree structure, mirroring the dynamics of cellular differentiation.

I present findings of our model trained on simulated single-cell RNA sequencing based on cell-sorted bulk RNA-seq data and a scRNA-seq dataset of over 1.2 million cells from individuals with variable lupus disease phenotypes. The cellular topic profiles estimated by our model markedly improve cell type deconvolution accuracy over traditional methods. Trained model parameters of `LaRCH` illustrate cell-type specific transcriptomic differences between SLE phenotypes, revealing the contributions of multiple immune cell types in the manifestations of lupus. I also identify a number of candidate genes that may have implications in the driving mechanisms that contribute to lupus disease pathogenesis. Ultimately, `LaRCH` is able to capture the hierarchical context between immune cell subsets by simultaneously identifying shared and distinct latent features amongst cell subtypes within heterogeneous cell samples.

Lay Summary

The immune system is composed of many cells belonging to different immune cell types. These cells undergo complex, multi-step cellular processes to reach their mature state. Abnormalities in the function and balance of immune cells result in immune dysfunction, which can lead to immunological diseases. This thesis introduces a novel computational method to model the relationships between cell types involved in immune responses through their shared gene expression features.

An application of this model on a dataset pertaining to systemic lupus erythematosus, an autoimmune disorder with no known cure, provides insights into the genetic mechanisms of this disease.

Preface

This work has been completed under the supervision of Dr. Yongjin Park at the BC Cancer Research Centre and Dr. Ramon Klein Geltink at the BC Children's Hospital Research Institute. This project was conceptualized by myself, with the guidance of Dr. Park. All of the wet-lab experiments detailed in Chapter 1 were performed by me under the supervision of Dr. Klein Geltink and Annette Patterson. The algorithm design, methodology development, and implementation presented in Chapter 2 was carried out by myself, with the assistance of Dr. Yichen Zhang and guidance of Dr. Park. The analysis conducted in Chapter 3 was performed by me with the guidance of Dr. Park and Dr. Klein Geltink. Portions of this thesis are modified texts from a manuscript under preparation for submission, of which I am the first author.

The LATEX template used in this thesis is provided by Michael McNeil Forbes and is publicly available on GitHub.

Table of Contents

Abstract iii

Lay Summary iv

Preface v

Table of Contents vi

List of Tables ix

List of Figures x

List of Abbreviations xiv

Acknowledgements xvi

Dedication xvii

1 Introduction 1

1.1 Immune cell lineage 1

1.1.1 Drivers of immune cell differentiation 3

1.2 Pathogenesis of systemic lupus erythematosus 6

1.2.1 Role of T cells in SLE 7

1.3 Latent representations of biological data 8

1.3.1 Deep-learning methods for biological data 8

1.3.2 Methods to model cellular hierarchies 9

1.4 Thesis objectives 10

1.4.1 Summary of rationale 10

1.4.2 Thesis aims and outline . . . 10

2 Tree structured topic modeling . . . 12
2.1 LaRCH . . . 12
2.1.1 Data generative scheme . . . 14
2.1.2 Informing model parameters through Bayesian priors . . . 15
2.1.3 Variational inference . . . 17
2.1.4 Model implementation . . . 20
2.1.5 Code availability . . . 20
2.2 Modeling simulated data . . . 21
2.2.1 Bulk gene expression profile generation . . . 21
2.2.2 Data simulation scheme . . . 21
2.2.3 Resulting simulated data . . . 22
2.2.4 Cell type specific latent topic profiles . . . 24
2.2.5 Using latent features for downstream analysis . . . 27

3 Modeling Cellular Hierarchies... . . . 31
3.1 Single-cell RNA-seq data from a large SLE dataset . . . 31
3.1.1 Dataset details . . . 31
3.2 Deconvolution of immune cell subsets . . . 32
3.2.1 Cell clustering . . . 34
3.3 Disease dependent differences in gene expression . . . 37
3.4 Functional analysis of latent nodes . . . 41
3.4.1 Node-level gene clustering . . . 41
3.4.2 Gene set enrichment analysis of node-level gene embeddings . . . 41
3.5 Node-level marker genes . . . 45
3.5.1 Marker gene significance testing . . . 45
3.5.2 Marker genes of interest . . . 45

4 Conclusion . . . 50
4.1 Summary of findings . . . 50
4.2 Limitations . . . 51
4.3 Future directions . . . 51

References . . . 53

Appendices

A Supplementary tables and figures 64

B Materials and methods for *in vitro* experiments 71

B.1 CD4+ T cell isolation 71

B.2 CD4+ T cell subset polarization 71

B.3 Flow cytometry analysis 72

List of Tables

1.1 **Table summary of CD4+ polarization conditions** A meta-analysis of various studies with *in vitro* CD4+ subset polarization condition protocols. Included are the polarization conditions for Tregs and Th17s across 10 studies. Units are presented as described in the original studies. 5

A.1 **Table summary of CD4+ polarization conditions** A meta-analysis of various studies with *in vitro* CD4+ subset polarization condition protocols. Included are the polarization conditions for Th1s and Th2s across 10 studies. Units are presented as described in the original studies. 65

A.2 **Table outlining the number of significant genes for each latent node** Upregulated genes are identified as genes with an embedding value of $\hat{\beta} > 0$ to a significance level of $\alpha = 0.05$. Inversely, downregulated genes are those with an embedding values of $\hat{\beta} < 0$. 67

A.3 **Table outlining the number of unique significant genes for each latent node** Upregulated genes are identified as genes with an embedding value of $\hat{\beta} > 0$ to a significance level of $\alpha = 0.05$. Inversely, downregulated genes are those with an embedding values of $\hat{\beta} < 0$. Genes are considered unique to a node if it does not also appear as a significant gene in its parent and sibling node. 68

List of Figures

1.1 **Lineage Tree representation of immune cell differentiation** Immune cells are derived from HSCs. HSCs then develop into multipotent progenitors (MPPs) in preparation for further differentiation. MPPs differentiate into common lymphoid and myeloid progenitors (CLPs and CMPs, respectively) before further development into more specified immune cell types within the adaptive and innate immune systems. Cells that are included in peripheral blood mononuclear cells (PBMCs) are in bold. 2

1.2 ***In vitro* polarization of CD4+ T cells results in heterogeneous cell populations** Flow cytometry analysis performed on *in vitro* polarized CD4+ T cells in Th0, Th17, pathogenic Th17 (Th17p), and Treg polarization conditions. Flow plots show protein expression of Foxp3 and Rorγt —markers of Tregs and Th17, respectively —in mouse CD4+ T cells. 3

2.1 **Overview of the LaRCH method** scRNA-seq data is encoded in latent features using a VAE. Original input data is then reconstructed through a generalized linear decoder containing the underlying PBT structure. 13

2.2 **Simulated scRNA-seq datasets exhibit realistic distribution of expression patterns across various noise levels** Shown are the plots of first two PCs for generated immune cell scRNA-seq datasets from the simulation scheme. Datasets generated using various noise proportions $\rho = \{0.0, 0.25, 0.5, 0.75, 0.9, 1.0\}$ are shown and coloured by cell type. 23

2.3 **Topic proportions of simulated immune cell gene expression shows distinct latent profiles between cell types** Structure plot of estimated topic proportion θ values for each cell in the simulated data set. Cells are annotated below by simulated cell type. 24

2.4 **Latent topic profiles can be used to reconstruct tree-structured relationships between cell subsets** The top panel shows the reconstructed tree structure from average latent topic profiles calculated for each cell subset. Below is a heatmap of the mean latent topic proportion profiles across each immune cell subset. 25

2.5 **Louvain lustering on latent topic space recovers groups of cells corresponding to their simulated cell type** The top panel shows the simulated cells in tSNE space coloured by original simulated cell type. The bottom panel shows the clusters generated using louvain clustering on the latent topic representation of cells. For visualization, tSNE dimensions were generated from the latent topic space. 28

2.6 **Using the LaRCH latent space for clustering consistently outperforms clustering on PCA and flat latent topic space** NMI calculated from clustering results from each dimensionality reduction technique and the original simulated cell types across various noise levels. At each noise level, 10 datasets were simulated with different starting seeds. 29

3.1 **Distinct topic proportions of PBMCs in a scRNA-seq dataset of SLE and healthy individuals correspond to cell type labels** Structure plot of estimated θ values for a subsample of 10,000 cells in the data set. Cells are annotated below by labelled cell types obtained from Perez *et al.* [1] 32

3.2 **Tree-structure relationships between cell types are reconstructed using latent topic profiles of a real scRNA-seq dataset** The top panel shows the reconstructed tree structure based on average latent topic profile representations of cell types. Below is a heatmap of the latent topic proportion profiles across a subset of 10,000 cells from the dataset. 33

3.3 **Clustering on latent topic space recovers groups of immune cells belonging to related immune subsets** UMAP projection of scRNA-seq data are coloured by Louvain clusters derived from the latent topic space. UMAP projection values obtained from the original Perez *et al.* study [1] . 35

3.4 **Overlap exists between latent topic clusters and original cell labels** a) Heatmap depicting the similarity matrix between louvain clusters obtained from latent topic space and cell group labels. b) Heatmap depicting the similarity matrix between louvain clusters obtained from latent topic space and cell type labels. Cell clusters are represented along the columns of the heatmap while cell labels are represented along the rows. Cell group and cell types are described in [1]. Cell groups describe more general immune cell subsets while cell types describe specific immune cell subsets. 36

3.5 **Cluster composition within samples is disease status dependent** a) Cell cluster proportion breakdown by SLE disease status shows cell composition within samples. b) Heatmap showing the median values of latent topic proportions across cells within each Louvain cluster. Clusters are shown along the rows of the heatmap. 37

3.6 **Individuals with identical disease status generally share latent topic patterns** Heatmap shows median latent topic profiles calculated for each individual included in the study. Rows represent each individual and are annotated by disease status. 38

3.7 **Disease status dependent differences exist in topic representation** Boxplots of median topic expressions per individual grouped by disease status. Topics with significant differences in median topic values between healthy and SLE individuals are shown. Significance between median expression values are calculated using Wilcoxon rank-sum tests between disease status individuals. 39

3.8 **Clustering of genes based on node-level embeddings show overarching gene programs** Heatmap shows number of significantly non-zero genes within each node that belong to gene clusters. Total numbers of non-zero genes per tree node are annotated above. 42

3.9 **Gene set enrichment analysis shows immune subset and disease relevant features described by latent node features** Heatmap shows the $-\log(p)$ enrichment significance of each gene set per node. Significance annotations are as follows: $***-p<0.0001, **-p<0.001, *-p<0.01$. . 43

3.10 **Canonical immune subset markers are captured in latent tree nodes of the topic model** Heatmap shows estimated node-wise gene embedding values $\hat{\beta}$ of genes previously identified as marker gene for immune cell subsets. 46

3.11 **Node-level gene embeddings reveal potential genes of interest with regards to SLE pathogenesis** Heatmap shows estimated node-wise gene embedding values $\hat{\beta}$ of genes with relevant immune function to SLE disease mechanisms. 48

A.1 **Simulated scRNA-seq datasets exhibit realistic distribution of expression patterns across various noise levels** Shown are the plots of third and fourth PCs for generated immune cell scRNA-seq datasets from the simulation scheme. Datasets generated using various noise proportions $\rho = \{0.0, 0.25, 0.5, 0.75, 0.9, 1.0\}$ are shown and coloured by cell type. 66

A.2 **Heatmap of top genes shows candidates for further analysis** Heatmap shows estimated gene embedding values $\hat{\beta}$ of top genes across all latent tree nodes. Genes are selected as the 80 gene with the greatest absolute $\hat{\beta}$ values amongst the aggregated pool of the top ten genes per latent tree node. . . . 69

A.3 **Enrichment analysis of gene sets related to immune function show cell function described by latent node embeddings** Heatmap shows the $-\log(p)$ enrichment significance of each gene set per node. Significance annotations are as follows: $***-p < 0.0001, **-p < 0.001, *-p < 0.01$. 70

List of Abbreviations

BALSAM	Bayesian Latent Topic Analysis with Sparse Association Matrix
DICE	Database of Immune Cell Expression
ETM	Embedded Topic Model
FACS	Fluorescence-activated Cell Sorting
GLM	Generalized Linear Model
GWAS	Genome-wide Association Study
HSC	Hematopoietic Stem Cell
HTS	High-throughput Screening
KL	Kullback-Leibler
LaRCH	Latent Representation of Cellular Hierarchies
LDA	Latent Dirichlet Allocation
lncRNA	Long Non-coding RNA
MPP	Multipotent Progenitors
NK	Natural Killer
NMF	Non-negative Matrix Factorization
NMI	Normalized Mutual Information
PBMC	Peripheral Blood Mononuclear Cells

PBT	Perfect Binary Tree
PC	Principal Component
scRNA-seq	Single-cell RNA sequencing
SLE	Systemic Lupus Erythematosus
Th	T Helper Cell
Th17p	Pathogenic T Helper 17
Treg	Regulatory T Cell
tSNE	t-distributed stochastic neighbor embedding
UMAP	Uniform Manifold Approximation and Projection
VAE	Variational Autoencoder

Acknowledgements

I would like to the deepest gratitude towards my supervisors, Dr. Yongjin Park and Dr. Ramon Klein Geltink. Thank you for your ongoing support and the freedom to pursue my research interests through this project. You both believed in me and my ideas, even when I did not. You have taught me how to conduct research with integrity and inspired me to follow my curiosities. I am so thankful to have worked with the two of you and I can't wait to use everything that I have learned from you in my future endeavors.

I would like to thank my committee members, Dr. Adi Steif and Dr. Geoffrey Schiebinger, for their support and valuable feedback on my work.

To the RKG lab, thank you for showing me how to pipette and letting me fool around with all your equipment. You are all such talented and ambitious scientists. I only wish that I had been able to spend more time in the lab with you.

To the Park lab, I am so grateful to have worked alongside people with such a wealth of knowledge and experience. I would like to thank Yichen Zhang in particular for providing much needed guidance on the conceptualization and implementation of my project.

Finally, I would like to thank my family and friends for their unconditional and unwavering love and support through this journey. To my mom, thank you for telling me not to become an engineer, this is way more fun. To the members of our mortgaged home, I don't know how I could have gotten through this without you guys by my side. You each inspire me in your own special ways and I love you.

Dedication

Dad, you exist in every piece of me.

Maggie, you are my reason why.

Chapter 1

Introduction

1.1 Immune cell lineage

The immune system performs a wide range of functions within the body, including defense against pathogens and tumours and maintaining homeostasis of various body systems. In order to perform its role, the immune system relies on the coordinated efforts of complex immune response mechanisms. These mechanisms fall within either the innate immune system or the adaptive immune system and are executed by specialized immune cells each with their own specific role and function. The innate immune system is the body's first line of defense against common microorganisms. Innate immune responses are able to rapidly identify molecular patterns common to microbes and toxins foreign to the host through hard-wired responses encoded within the genes of the host's germ line [2]. For pathogens unable to be fought by the innate immune system, the adaptive immune system takes over by utilizing antigen specific recognition to target specific pathogens. These antigen specific cells are committed to memory through long-lived dormant cells that persist following a pathogen's first offense allowing the adaptive immune system to protect against subsequent reinfections from the same pathogen [2, 3]. Innate immune cells are comprised of innate lymphocytes such as natural killer (NK) cells and myeloid cells, including eosinophils, neutrophils, basophils, monocytes, and mast cells while adaptive immune cells are primarily made up of B Lymphocytes and T lymphocytes [4].

There are over 80 different identified immune cell subsets each with specialized functions and characterized by specific cellular markers. Despite great heterogeneity amongst the various immune cell subsets, they are all commonly derived from hematopoietic stem cells (HSCs) in bone marrow. Through a multi-step differentiation process, immune cells commit to a cell lineage determined by various external and intrinsic factors including transcriptional reprogramming [5]. Since the differentiation process contains a number of sequential

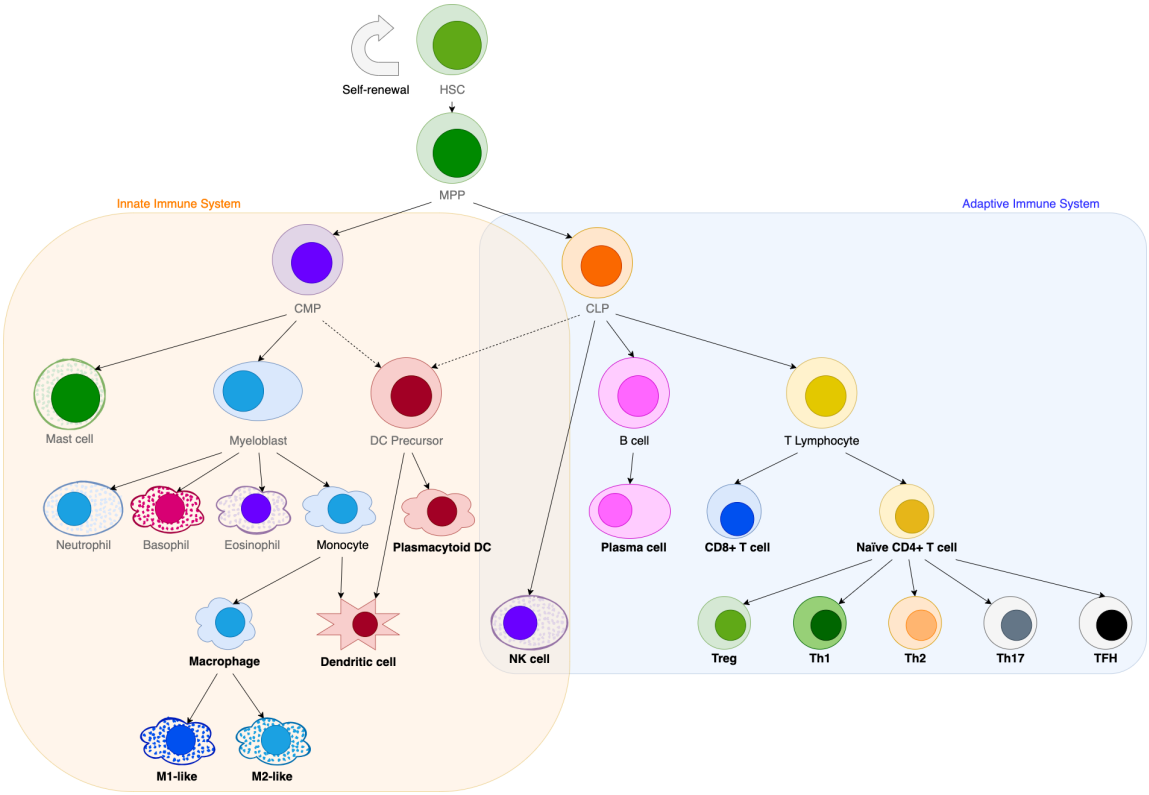

Figure 1.1: **Lineage Tree representation of immune cell differentiation** Immune cells are derived from HSCs. HSCs then develop into multipotent progenitors (MPPs) in preparation for further differentiation. MPPs differentiate into common lymphoid and myeloid progenitors (CLPs and CMPs, respectively) before further development into more specified immune cell types within the adaptive and innate immune systems. Cells that are included in peripheral blood mononuclear cells (PBMCs) are in bold.

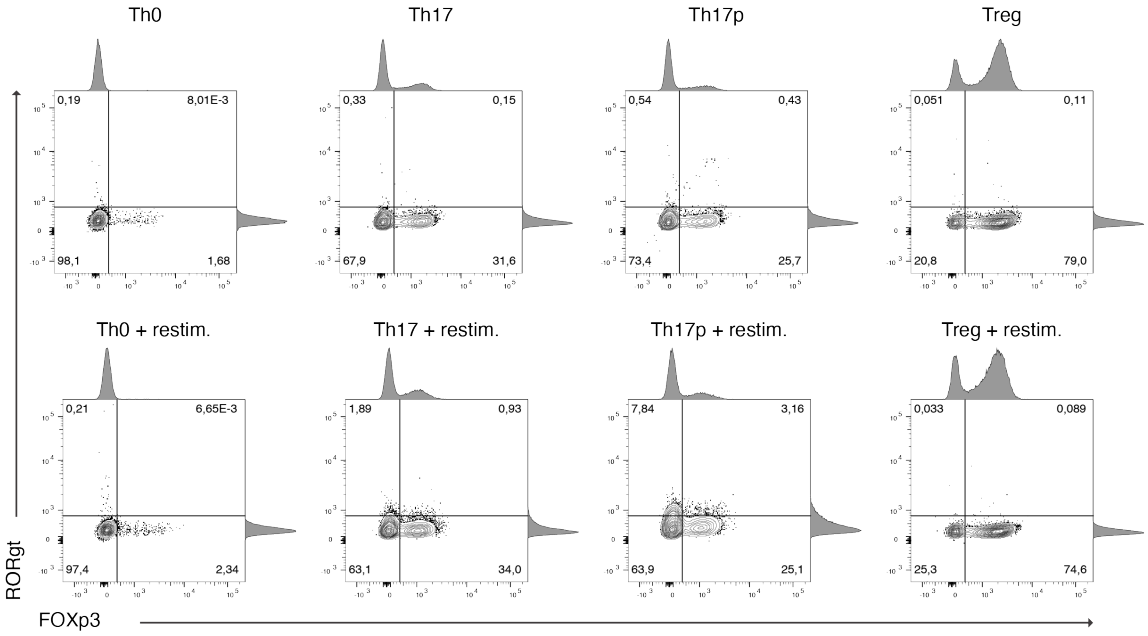

Figure 1.2: ***In vitro* polarization of CD4+ T cells results in heterogeneous cell populations** Flow cytometry analysis performed on *in vitro* polarized CD4+ T cells in Th0, Th17, pathogenic Th17 (Th17p), and Treg polarization conditions. Flow plots show protein expression of Foxp3 and Rorγt —markers of Tregs and Th17, respectively —in mouse CD4+ T cells.

specialization steps where the immune cells become increasingly distinct, the immune cell landscape is often referred to and depicted as a hierarchical lineage tree [4] (Fig. 1.1).

1.1.1 Drivers of immune cell differentiation

The process of immune cell differentiation is driven by cell signaling, a highly variable process influenced by a large number of factors including extracellular environment cytokine presence [6], epigenetic mechanisms [7], cellular metabolism [8, 9], transcription factors (TFs) [10], host and host environment, and microbiome [11].

The complexity of immune cell differentiation is best exemplified through CD4+ helper T cell polarization. Naïve CD4+ T cells further differentiate into helper T cell subsets marked by distinct gene expression signatures, including cytokine expression, and require the presence of specific extracellular cytokines to polarize [12]. In general, CD4+ T cell

populations are heterogeneous following polarization as shown from flow cytometry analysis performed as described in B.1 (Fig. 1.2). Protein expression levels of Foxp3 and Rorγt —regulatory markers of regulatory T cells (Tregs) and Th17 cells, respectively —are shown [13, 14]. Both Foxp3 and Rorγt are induced in the presence of Tgf-β and in the absence of a proinflammatory cytokine, such as IL-6, Foxp3 can inhibit Rorγt function to drive towards a Treg phenotype [15]. The fate of CD4+ T cells is therefore highly influenced by the balance between these two proteins. In the Th17 and pathogenic Th17 (Th17p) polarization conditions without additional restimulation, only a very small subset of cells express Rorγt, while a larger subset of cells express Foxp3, suggesting that cells are leaning towards a more suppressive phenotype rather than an effector Th17 phenotype. Cells expressing Rorγt and therefore exhibiting a Th17 phenotype are only seen upon restimulation, though the population of suppressive Foxp3+Rorγt- cells are still present in the same proportion (Fig.1.2).

Another factor that contributes to the uncertainty in the differentiation process of immune cells is plasticity between subsets, in the context of CD4+ T cells, it is the ability of a single cell to exhibit features of a number of T cell subsets simultaneously or over time [16].

A meta-analysis of a number of CD4+ polarization protocols summarized in Table 1.1 and A.1 found a lack of consensus between researchers on the necessary extracellular cytokines and their quantities for subset polarization, particularly for Th17 polarization. This demonstrates that CD4+ T cell subset differentiation and immune cell differentiation in general is not entirely understood and there exists the need to further investigate their driving mechanisms and ways to capture the heterogeneous nature of these cell populations in a hierarchical manner.

Treg									
Study	α CD3	α CD28	IL-2	TGF-β	TNF-α	α IFNγ	α IL-4	α IL-12	—
Field 2020 [17]	5 μg/mL	0.5 μg/mL	100 U/mL	5 ng/mL	—	10 μg/mL	10 μg/mL	10 μg/mL	—
Yang 2019 [18]	beads	beads	50 U/mL	2 ng/mL	"accordingly"	—	—	—	—
Xiao 2016 [19]	beads	beads	25 U/mL	3 ng/mL	—	—	—	—	—
Thomas 2012 [20]	1 μg/mL	1 μg/mL	20 U/mL	5 ng/mL	—	—	—	—	—
Jiang 2018 [21]	5 μg/mL	5 μg/mL	10 U/mL	2 ng/mL	—	—	—	—	—
Pham 2014 [22]	2 μg/mL	0.5 μg/mL	—	2 ng/mL	—	—	—	10 μg/mL	—
Wagner 2021 [23]	1 μg/mL	1 μg/mL	—	2 ng/mL	—	—	—	—	—
Puleston 2021 [24]	5 μg/mL	2 μg/mL	100 U/mL	10 ng/mL	—	4 μg/mL	4 μg/mL	—	—
Th17									
Study	α CD3	α CD28	IL-1β	IL-2	IL-6	IL-23	TGF-β	α IFNγ	α IL-4
Field 2020 [17]	5 μg/mL	0.5 μg/mL	10 μg/mL	100 U/mL	10 ng/mL	—	5 ng/mL	10 μg/mL	10 μg/mL
Yang 2019 [18]	1 μg/mL	1 μg/mL	—	—	10 ng/mL	—	2 ng/mL	5 μg/mL	5 μg/mL
Chen 2017 [25]	beads	beads	—	100 U/mL	20 ng/mL	—	20 ng/mL	—	—
Thomas 2012 [20]	1 μg/mL	1 μg/mL	—	—	10 ng/mL	10 ng/mL	1 ng/mL	20 μg/mL	20 μg/mL
Jiang 2018 [21]	5 μg/mL	5 μg/mL	10 ng/mL	—	20 ng/mL	50 ng/mL	0.5 ng/mL	—	—
Pham 2014 [22]	2 μg/mL	0.5 μg/mL	10 ng/mL	—	100 ng/mL	10 ng/mL	2 ng/mL	10 μg/mL	10 μg/mL
Wagner 2021 [23]	1 μg/mL	1 μg/mL	20 ng/mL	—	25 ng/mL	20 ng/mL	—	—	—
Gaublomme 2015 [26]	2 μg/mL	2 μg/mL	20 ng/mL	—	25 ng/mL	20 ng/mL	2 ng/mL	—	—
Puleston 2021 [24]	5 μg/mL	2 μg/mL	10 ng/mL	100 U/mL	5 ng/mL	—	10 ng/mL	10 μg/mL	10 μg/mL

Table 1.1: **Table summary of CD4+ polarization conditions** A meta-analysis of various studies with *in vitro* CD4+ subset polarization condition protocols. Included are the polarization conditions for Tregs and Th17s across 10 studies. Units are presented as described in the original studies.

1.2 Pathogenesis of systemic lupus erythematosus

Systemic lupus erythematosus (SLE) is a complex, chronic autoimmune disease that affects approximately 3.17 million adults worldwide [27, 28]. The disease presentation of SLE is highly heterogeneous between individuals and may affect any part of the body, but symptoms typically include fatigue, skin rashes, chronic inflammation, and fever. Many patients suffer from periods of disease exacerbations (flares) where worse or new symptoms are seen in one of more organ systems and a change or increase in treatment is required to suppress the overactive immune response [29]. Currently, there is no cure for SLE and therapeutic treatments of this disease severely affect the quality of life for patients.

Genetic factors play a crucial role in the development of SLE. About 90% of SLE cases are female, however this sex difference remains poorly understood [30, 31]. There is also a large discrepancy in the prevalence of SLE between races/ethnic backgrounds. East Asian and Hispanic/South American populations see the highest incidence rates of SLE [32]. The heritability of the disease is estimated to be around 44% [33], but there has been great difficulty in pinpointing any specific genetic cause of disease manifestation [34]. Currently, there are over 60 confirmed genetic risk loci associated with SLE [35]. Many of the genes containing these loci are involved with regulatory functions within the innate and adaptive immune system. Genetic variants are suggested to have a regulatory effect on disease pathology since the majority of the SNPs found to be associated with SLE fall within non-coding regions of the gene. Despite the identification of many genetic risk loci, they are only able to explain a small fraction of the heritability of SLE. There are no genetic factors that have been found to be shared between all SLE patients due to the heterogeneous nature of disease manifestation [34].

There are a number of environmental factors that have been shown to be significantly associated with the presence of SLE. These factors include exposures such as air pollution, heavy metals, and ultraviolet radiation. Certain lifestyle choices also contribute to the development of SLE including alcohol consumption, cigarette smoking, sleep patterns, socioeconomic factors and stress, and diet and microbiome [32].

Epigenetic factors, innate and adaptive immunity, and inflammation have also been found to play a role in SLE pathogenesis and exist in the interface between environmental exposures and genetic risk factors and provide potential mechanisms to explain the effect of environmental exposures on disease. [32, 35].

Several different risk factors are hypothesized to be associated with the development of SLE, but a direct cause of SLE remains unclear. The disease mechanism of SLE disease presentation is an area of ongoing investigation, but one known driver is the presence of autoreactive T and B cells, resulting in persistent inflammatory responses in the absence of pathogens. A better understanding of the cellular makeup of circulating immune cells in the body provides further insight into the causal mechanisms in SLE and other autoimmune disorders.

1.2.1 Role of T cells in SLE

SLE pathogenesis is characterized by the presence of autoantibodies leading to chronic inflammation causing damage to tissues and organs. These autoantibodies are the result of defects in cellular apoptotic debris clearance, pro-inflammatory expression signatures in peripheral lymphocytes, and dysfunctional peripheral tolerance mechanisms [27].

T cells contribute to SLE pathogenesis through the amplification of inflammation by secretion of pro-inflammatory cytokines. Additionally, the accumulation of autoreactive memory T cells results in sustaining disease over time.

CD4+ helper T cells signal to B cells through the production of cytokines. Abnormalities in the T cell receptor (TCR) leads to increased stimulation and results in aberrant activation of T cells [36]. SLE patients have an imbalance of CD4+ T cell subsets. In peripheral blood CD4+ T cell populations, pro-inflammatory Th17s are overabundant, while suppressive Tregs are under-represented. As a result, there is an increased prevalence of the pro-inflammatory signaling cytokines IL-6 and IL-17 [37]. Additionally, aberrant activation of PI3K/AKT/mTOR signal transduction leads to an increased accumulation of effector/memory T cells due to a resistance to activation-induced cell death [38]. Meanwhile, dysfunction in CD8+ T cells in SLE results in dampended cytotoxicity and results in increased risk of infection, potentially triggering autoimmunity [39].

The significant role that T cells play in SLE pathogenesis, coupled with the lack of understanding behind the genetic drivers of SLE and the complex mechanisms of immune cell differentiation inspires the need to further investigate genomic mechanisms in immune cell subsets in patients with SLE.

1.3 Latent representations of biological data

With the recent advancements of high-throughput screening (HTS) technologies, specifically single-cell RNA sequencing (scRNA-seq), knowledge of the molecular profiles of specific cell types has expanded. Since scRNA-seq data is high-dimensional, often capturing expression of tens of thousands of genes across thousands of cells, scalable and interpretable computational methods are necessary to perform sophisticated analysis of the data in order to uncover biological insights. One branch of these methods focuses on representing scRNA-seq data in a lower dimension latent space in a meaningful and interpretable way. Latent representations of biological data allows for further downstream analysis, such as cellular deconvolution using unsupervised clustering techniques [40].

A key challenge in latent embedding of biological data is the inability to produce biologically interpretable results. Therefore extracting meaning from the data requires backfilling of information through manual annotation and analysis [41].

1.3.1 Deep-learning methods for biological data

As datasets become increasingly large in the number of samples as well as dimensionality, there has been a shift towards deep-learning methods which are scalable and out-perform linear methods such as the iterative Harmony algorithm [42] and non-negative matrix factorization (NMF) based methods such as LIGER [43]. Deep-learning based approaches for scRNA-seq most commonly use autoencoders, a type of artificial neural network, as a means to project high-dimensional data onto a low-dimensional latent space [44]. A number of recent methods utilize a variational autoencoder (VAE) [45], a Bayesian method that relies on stochastic variational inference to efficiently scale to large datasets. These methods include single-cell variational inference (`scVI`) [46] and single-cell VAE (`scVAE`) [47], the latter of which extends the VAE framework by using a Gaussian-mixture model as the prior distribution for latent variables over a Gaussian prior to learn cell clusters through a categorical latent variable.

Single-cell Embedded Topic Model (`scETM`) [48] and Bayesian latent topic analysis with sparse association matrix (`BALSAM`) [49] are methods that make use of an embedded topic model (ETM) framework [50], an extension of VAE. In these methods, neural networks are used to encode scRNA-seq data into a latent topic space, while a generalized linear model (GLM) is used to decode latent embeddings back to the original gene expression space. As

a result, these methods are able to simultaneously learn the encoder parameters and topic specific gene embeddings to create a highly interpretable latent space.

1.3.2 Methods to model cellular hierarchies

It has been suggested to use a tree-based reference structure to represent cell types as opposed to an atlas or periodic table-like classification system in order to better capture relationships between cell types [51]. Therefore, a method that can embed scRNA-seq data in an interpretable latent space while learning a hierarchical tree-structure of data has the potential effectively identify cell clusters while providing insights into cellular relationships.

Hierarchical extensions to the topic modeling framework have been proposed where latent topics exist as nodes that lie in a tree structure. These methods include hierarchical latent Dirichlet allocation (`hLDA`) [52] and tree-structured neural topic model (`TSNTM`) [53]. `hLDA` uses a latent Dirichlet allocation (LDA) approach to represent topics in a hierarchical structure informed by a nested Chinese restaurant process prior. `TSNTM` makes use of autoencoding variational Bayes to efficiently construct a tree-structured latent topic space. Since the latent topic tree structures in these methods exist in an infinite tree space, meaning each node can have infinite child nodes, these models rely on Markov Chain Monte Carlo (MCMC) algorithms in order to simulate models over super-exponential space. Despite theoretical guarantees, these methods were not suitable for the purposes of scRNA-seq data due to the size of the data and the computational complexity associated with learning the topic tree structure.

Another approach to model hierarchical representations of scRNA-seq data is to do *post-hoc* construction of the tree-structured relationships between cells. `treeArches` [54] aims to construct and extend a cell reference atlas and corresponding hierarchical classifier, relying on prior information from an existing cell atlas. `cellTree` infers tree-structured relationship hierarchies between individual cells based upon LDA latent representations and minimum spanning trees to construct the tree structure. These methods construct their relevant cellular tree representations after estimating their latent representations rather than learning the cellular hierarchies simultaneously with their latent representation. Additionally, `cellTree` only constructs the direct hierarchical relationships between individual cells, differing from the other modeling approaches which aim to define relationships between cell features through topics in order to represent hierarchical relationships.

1.4 Thesis objectives

1.4.1 Summary of rationale

Immune cell differentiation is a complex biological process that remains an area for further investigation. The various immune cell subsets originate from a common HSC progenitor and follow a series of branching differentiation steps to reach maturity. As immune cells develop, they become more highly specific in cell features and function, resulting in greater dissimilarity with other immune subsets. The process results in a lineage tree structure between immune cell subsets where cells share hierarchical relationships with one another. Tree-structured latent modeling of HTS data, specifically scRNA-seq, has the potential to uncover insights into the cellular differentiation process, hierarchical relationships between immune cell subsets, and their shared and differing cellular features. Though a number of statistical latent models of scRNA-seq data exist, an efficient and effective tree-structured model for this purpose has yet to be shown.

Applications of such a model may also provide biological insights on the pathogenesis of disease driven by immune dysregulation such as SLE. SLE is a complex autoimmune disease that is characterized by chronic autoinflammation in one or many organ systems. One mechanism driving this autoinflammation is aberrant function and imbalance of peripheral blood lymphocytes. Though SLE is a heritable disease and through genome-wide association studies (GWAS) a number of risk loci have been identified, genetic mechanisms driving disease presentation are not well understood.

1.4.2 Thesis aims and outline

In the work presented in this thesis, there are two main research aims that are addressed.

First, I aim to develop a tree-structured topic model to represent immune cell subsets and their hierarchical relationships through shared and differing cellular features. In order for this model to be effective for these purposes and provide biological insight, I bear in mind the efficiency and interpretability of methods used. The second aim of this thesis is to apply the developed model to a scRNA-seq dataset from individuals with varying SLE disease conditions to show pathological differences in immune cell composition and identify specific genomic features that may play a role in SLE pathogenesis.

In Chapter 2, I propose `LaRCH` (Latent Representation of Cellular Hierarchies), a neural

topic model with a underlying tree-structure representation of latent topics. This model is based on an ETM framework which allows for highly interpretable topic specific gene embeddings. Autoencoding variational Bayes is used to train the model to ensure computational efficiency. To demonstrate the efficacy of this model, I train LaRCH on a realistic simulated scRNA-seq dataset of immune cell subsets. Using this model I am able to reconstruct cellular hierarchies from their latent topic representations and perform further downstream analysis.

Chapter 3 of this thesis shows the application of the LaRCH model on a large scRNA-seq dataset of PBMC samples from individuals with varying SLE disease statuses. I demonstrate the ability for LaRCH to deconvolve immune cell subtypes. Then through disease stratified analysis, I identify disease dependent differences in cell-type composition and cellular features. The interpretable model features of LaRCH are used to assign biological meaning to the latent topic space and identify potential genetic drivers of SLE disease mechanisms that inspire further experimental investigation.

Chapter 2

Tree structured topic modeling

To simultaneously learn a latent representation of cellular gene expression data and their hierarchical relationships, a tree-structured topic model is introduced. In this chapter, details of this model, referred to as a Latent Representation of Cellular Hierarchies (LaRCH), are shown along with the results of a trained model on a simulated single-cell gene expression dataset.

2.1 LaRCH: Latent Representation of Cellular Hierarchies

LaRCH uses an ETM [50] approach built on a VAE [45, 55, 56] framework borrowed from neural network architectures with additional tree-node layer that aggregates sparse gene expression effects. Within a typical ETM, cells are represented as a mixture of latent topics each describing a gene expression profile of raw scRNA-seq count data. LaRCH introduces an additional layer of latent nodes that lie within a PBT structure (Fig. 2.1). As a result, cells are viewed as a mixture of latent topics, each corresponding to a summation of tree nodes along a path from the root node to a given leaf node of the underlying tree. From this structure, it is inferred that topics that lie in close proximity on the tree share a number of properties while those further apart are also more distant in their features. As a more specific example, two topics corresponding to paths terminating at sibling leaf nodes would share all features captured in their common parent nodes, only diverging at the last branching point in their paths. Cells within these two topics would be said to share a large number of gene expression features represented by their shared parent nodes in their topic paths.

The LaRCH model is comprised of an encoder and decoder component. The encoder transforms raw single-cell gene expression count data into the latent topic space through a neural net. In this model, rather than choosing the number of latent topics to represent

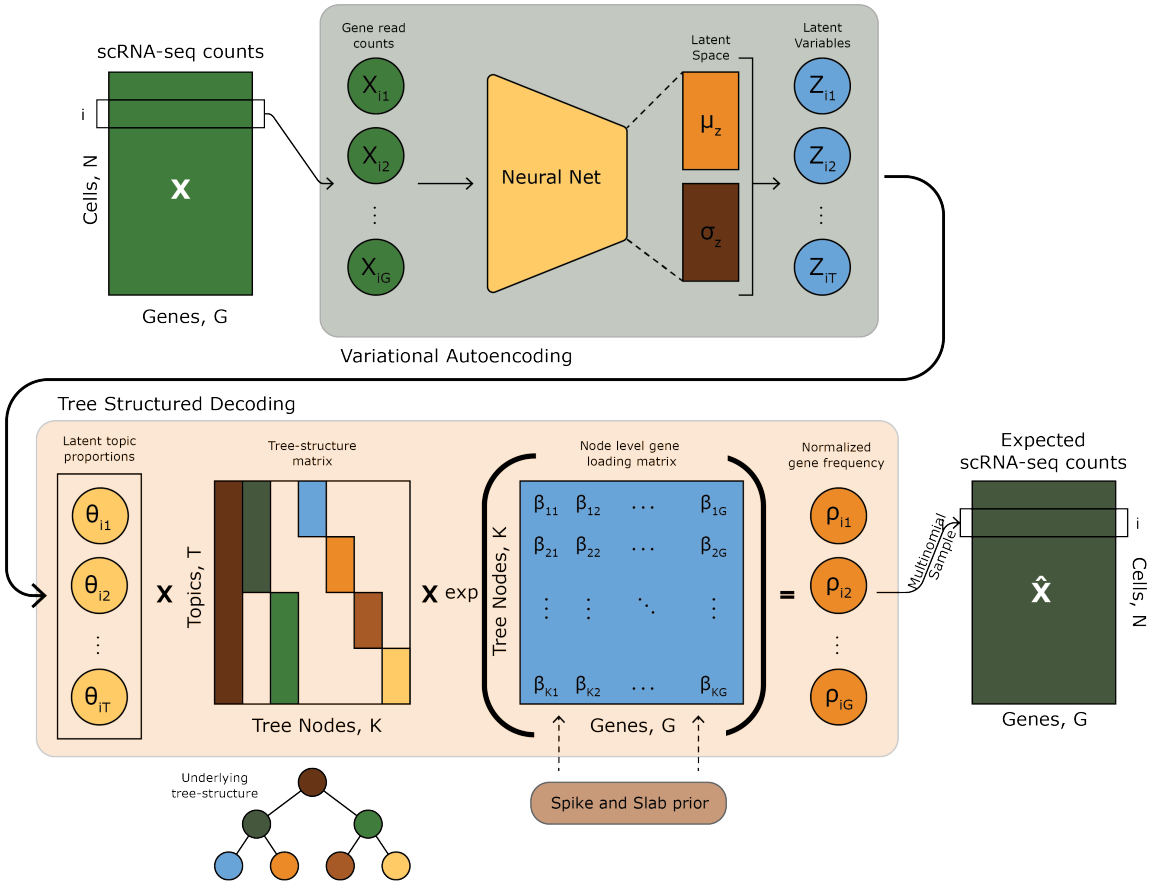

Figure 2.1: **Overview of the LaRCH method** scRNA-seq data is encoded in latent features using a VAE. Original input data is then reconstructed through a generalized linear decoder containing the underlying PBT structure.

the data, the depth of the latent-node tree D is pre-determined, therefore the resulting number of total latent tree nodes is $2^D - 1$, and the latent topic space that the gene count data is transformed to is of dimension 2^{D-1}. This latent representation is then converted to a topic proportion representation using the softmax function.

Where LaRCH differs from a typical ETM lies in the decoder. The decoder contains a gene embedding matrix β for each tree node as model parameters. Gene embedding values represent relative gene expression frequencies for a given tree node. Additionally, the tree-structured relationships at the topic level are represented mathematically using a structural matrix. The learned gene embedding matrix, structural matrix, and inferred topic proportions are passed to a GLM to estimate the cell specific gene frequencies for each data point. These gene frequencies are then used to compute likelihood values for expected gene counts.

2.1.1 Data generative scheme

To apply the LaRCH topic model, each cell is treated as a document, the set of genes in the dataset as the vocabulary of size G, and each scRNA-seq read as a token of a gene from the vocabulary. Each cell i of N cells is then represented as a mixture of T latent topics in our model.

In a classic ETM, each topic has a corresponding distribution over the gene space $\beta \in \mathbb{R}^{T \times G}$ [50]. This is altered by introducing an additional latent tree node layer to the embedding process. K latent nodes lie in a PBT structure of depth D, where $K = 2^D - 1$. Each of the T latent topics, where $T = 2^{D-1}$, corresponds to a path from the root node to a leaf node of the PBT. Each tree node now has an embedding over the gene space $\beta \in \mathbb{R}^{K \times V}$ and each topic also has an embedding made up of the tree nodes along its corresponding path $\bar{\beta} = A \exp(\beta)$, where A is a $T \times K$ matrix that captures the PBT structure. From this, two topics with similar paths in the tree-structured node will share gene embeddings while the paths are identical, then differ once the paths split.

The formal data-generating process for each cell i, where $i = 1, \ldots, N$, in the scRNA-seq dataset is:

1. Draw latent topic proportion θ_i for cell i from:

$$\theta_i = \text{softmax}(z_i) = \frac{\exp(z_{it})}{\sum_{t=1}^{T} \exp(z_{it})}, z_i \sim \mathcal{N}(0, \mathbf{I}) \tag{2.1}$$

2. Determine the categorical distribution ρ_i of genes for cell i based on θ_i:

$$\tilde{\rho}_i = \theta_i A \exp(\beta), \rho_i \sim \text{Dirichlet}(\tilde{\rho}_{i1}, \tilde{\rho}_{i2}, \ldots, \tilde{\rho}_{iG}) \tag{2.2}$$

3. Draw gene g for each read of a cell from a multinomial distribution to give total gene counts over R reads:

$$X_i|\rho_i \sim \text{Multinomial}(R, \rho_i) \tag{2.3}$$

   This gives the probability distribution:

$$p(X_i|\rho_i) \propto \prod_{g \in G} \rho_{ig}^{X_{ig}}$$

   Where $\sum_g \rho_{ig} = 1$ for all $i \in 1, \ldots, N$ and $\rho_{ig} \geq 0$ for all cells i and genes $g \in [G]$

Gene expression counts are said to follow independent multinomial probabilities (a bag of words assumption). Compared to other deep learning methods based on Poisson, Negative Binomial, and Gaussian distributions, a multinomial likelihood better preserves scale-invariant properties mitigating batch effects on the trained model.

Here z_i is the $1 \times T$ latent topic embedding of cell i, θ_i is the $1 \times T$ topic mixture of cell i where $\sum_{t=1}^{T} \theta_{i,t} = 1$, β is a $K \times G$ gene embedding matrix for latent nodes, A is a $T \times K$ matrix representing the tree-structure of nodes, ρ_i is the $1 \times G$ cell specific distribution of gene reads, and X_i is the $1 \times G$ count vector of genes.

2.1.2 Informing model parameters through Bayesian priors

A Dirichlet prior to inform multinomial sampling

A Dirichlet prior is introduced to inform the categorical distribution of genes for each given cell, $\rho_i \sim \text{Dir}(\tilde{\rho}_i)$. This allows for direct transformation from the parameters $\tilde{\rho}$ to a distribution vector from which to sample multinomial values. The Dirichlet parameters

are formulated as a GLM,

$$\tilde{\rho}_{ig} = \sum_{t=1}^{T} \theta_{it} \sum_{k=1}^{K} A_{tk} \exp(\beta_{kg}) - Db_g$$

This GLM captures the additive effects of nodes within each topic and the proportional additive effects of topics within each cell on the relative gene expression profile.

An additional gene-specific bias parameter b_g is used to represent invariant effects across all latent nodes. Since this parameter is present for each node, at the topic level, the additive effect of this parameter is Db_g. This is a free parameter without any specified prior distribution.

Since the multinomial and Dirichlet distribution share a conjugate relationship, the composite variable ρ can be integrated out to obtain the posterior predictive likelihood,

$$\begin{aligned} p(x_i|\theta_i, \{\beta_{kg}\}, \{b_g\}) &= \int p(x_i|\rho_i) p(\rho_i|\theta_i, \{\beta_{kg}\}, \{b_g\}) d\rho_i \\ &= \frac{\Gamma(\sum_g \tilde{\rho}_{ig}) \Gamma(\sum_g X_{ig} + 1)}{\Gamma(\sum_g \tilde{\rho}_{ig} + X_{ig})} \prod_g \frac{\Gamma(X_{ig} + \tilde{\rho}_{ig})}{\Gamma(\tilde{\rho}_{ig}) \Gamma(X_{ig} + 1)} \\ &\propto \frac{\Gamma(\sum_g X_{ig} + 1) \prod_g \Gamma(X_{ig} + \tilde{\rho}_{ig})}{\prod_g \Gamma(X_{ig} + 1) \, \Gamma(\sum_g \tilde{\rho}_{ig} + X_{ig})}, \end{aligned}$$

where $\tilde{\rho}_{ig} = \sum_{t=1}^{T} \theta_{it} \sum_{k=1}^{K} A_{tk} \exp(\beta_{kg}) - Db_g$, and $\Gamma(\cdot)$ is the Euler's gamma function.

Using a Bayesian prior to induce model sparsity

Since scRNA-seq data is typically sparse with only a few non-zero features and shared expression of many highly expressed genes in general, the trained model used to represent this data should reflect this. To induce sparsity and increase interpretability in the gene embedding model parameters, β, a Bayesian spike-and-slab distribution prior [57] is used to inform their values. Within the spike-and-slab prior distribution, it is assumed that the majority of gene-embedding values β_{kg} are statistically zero with some probability $1 - \pi$. This concentration around zero is called the "spike" component of the distribution. The "slab" portion of the distribution is explained by a normal distribution centred around 0

with variance τ.

$$\beta_{kg} \sim \pi\mathcal{N}(0,\tau) + (1-\pi)\delta_0(\beta_{kg}) \tag{2.4}$$

2.1.3 Variational inference

Since exact inference of the posterior probability of latent topics $p(\theta_i|x_i)$ is computationally intractable in high dimensional spaces, stochastic variational inference is used as a scalable approach to finding approximate distributions [45].

The variable estimation scheme as described in Kingma *et al.* is used[45]. In this reparameterization technique, latent variable inference is cast into an optimization problem in a deep belief network. This can then be solved using back-propagation steps with respect to model parameters.

Variational inference is used to estimate two sets of parameters, the cell-specific latent topic proportions $q(\theta_n)$ and the node-specific gene-embedding parameters $q(\beta_{kg})$. For each set of parameters, the minimized Kullbeck-Leibler (KL) divergence is used to approximate the actual data likelihood probability. This is equivalent to maximizing the evidence lower bound (ELBO) for the total data likelihood, $\mathcal{L}$.

$$\ln p(\mathbf{X}) \geq \mathbb{E}_q\left[\ln \frac{p(\mathbf{X},\Theta,\beta)}{q(\Theta,\beta|\phi,\xi)}\right] \triangleq \mathcal{L} \tag{2.5}$$

$$\sum_{i\in[N]}\int_{\theta,\beta} p(x_i,\theta_i,\beta)d\theta_i d\beta_i \geq \mathbb{E}_q\left[\ln \frac{p(\mathbf{X}|\Theta,\beta)p(\theta_i)p(\beta|\pi,\tau)}{q(\Theta|\phi)q(\beta|\xi)}\right] \tag{2.6}$$

$$= \mathbb{E}_q\left[\sum_{i\in[N]} \ln p(x_i|\theta_i,\beta)\right] + \mathbb{E}_q\left[\sum_{i\in[N]} \ln \frac{p(\theta_i)p(\beta|\pi,\tau)}{q(\theta_i|\phi)q(\beta|\xi)}\right] \tag{2.7}$$

$$= \mathbb{E}_q\left[\sum_{i\in[N]} \ln p(x_i|\theta_i,\beta)\right] + \sum_{i\in[N]}\left[\mathbb{E}_q\left[\ln \frac{p(\theta_i)}{q(\theta_i|\phi)}\right] + \mathbb{E}_q\left[\ln \frac{p(\beta|\pi,\tau)}{q(\beta|\xi)}\right]\right] \tag{2.8}$$

$$= \mathbb{E}_q\left[\sum_{i\in[N]} \ln p(x_i|\theta_i,\beta)\right] - \sum_{i\in[N]}\left[\mathrm{D}_{\mathrm{KL}}(q_\theta \parallel p_\theta) + \mathrm{D}_{\mathrm{KL}}(q_\beta \parallel p_\beta)\right] \tag{2.9}$$

Where ϕ and ξ are used to denote all parameters of the latent state or parameter distributions, π and τ are the probability of inclusion and variance parameters, respectively.

Stochastic gradient steps with respect to the variational parameters ϕ and ξ are taken in the variational inference algorithm, optimizing the ELBO objective (2.7) in order to find approximate posterior distributions. Variational inference is implemented in `Pytorch` using `torch.autograd` [58, 59] to calculate gradients.

Approximation of latent topic proportions θ

Latent topic proportions, θ cannot be exactly evaluated, so are instead approximated by summing over the posterior predictive log-likelihood using sampled instances of $\theta^{(s)}$ and $\beta^{(s)}$ for each minibatch sample $s \in [S]$

$$\mathbb{E}_q \left[\sum_{i \in [N]} \ln p(x_i | \theta_i, \beta) \right] \approx \frac{1}{S} \sum_{s \in [S]} \ln p(x_s | \theta^{(s)}, \beta^{(s)}) \tag{2.10}$$

The mean μ and variance σ functions for latent variable inference are parameterized in the encoder model which takes the original high-dimensional scRNA-seq data x. $\theta^{(s)}$ is then posteriorly sampled using the reparameterization trick of the Logistic Normal distribution as follows:

1. Sample random error $\epsilon_s \sim \mathcal{N}(0, 1)$
2. Reparameterize latent topic variables $z_s \leftarrow \mu(x_s) + \sigma(x_s) \circ \epsilon_s$
3. Transform latent variables to latent topic proportions: $\theta_t^{(s)} \leftarrow \frac{\exp(Z_{st})}{\sum_{j=1}^{T} \exp(Z_{sj})}$

Assuming $z_s \sim \mathcal{N}(0, I)$ *a priori* for all s, the corresponding variational parameters $\phi \equiv (\mu, \sigma)$ is used to derive the KL divergence between the prior and variational distributions as the second term of (2.9):

$$\mathbb{E}_q \left[\ln \frac{q(Z_{st} | \phi)}{p(Z_{st})} \right] = \mathrm{D}_{\mathrm{KL}}(q_\theta \parallel p_\theta) = \sum_{t=1}^{T} \frac{1}{2} [\mu_{st}^2 + \sigma_{st}^2 - \ln \sigma_{st}^2 - 1] \tag{2.11}$$

Approximation of global spike-and-slab parameters β

Fully-factored spike-and-slab distributions are used as variational distributions for node-specific gene embedding parameters, β_{kg}, to derive the final term of (2.9), the negative KL

loss of the global parameters, β.

When β_{kg} is 'on', denoted by a latent indicator variable $h_{kg} = 1$ with probability α_{kg}, β_{tg} is parameterized by a Gaussian distribution:

$$q(\beta_{kg}|h_t g = 1) = \mathcal{N}\left(\mu_{kg}^{\beta}, \nu_{kg}^{\beta}\right)$$

with probability $\alpha_{kg} \triangleq p(h_{kg} = 1)$; otherwise, β_{kg} is set to zero:

$$q(\beta_{kg}|h_{kg} = 0) = \delta_0(\beta_{kg})$$

with probability $1 - \alpha_{kg}$.

From the variational parameters $\xi \equiv (\alpha, \mu, \nu)$, the variational distribution is characterized as:

$$q(\beta_{kg}) = \prod_{g \in G} \alpha_{kg} \mathcal{N}(\mu_{kg}, \nu_{kg}) \tag{2.12}$$

with

$$\mathbb{E}_q[\beta_{kg}] = \alpha_{kg}\mu_{kg} \tag{2.13}$$

$$\mathbb{V}_q[\beta_{kg}] = \alpha_{kg}(1 - \alpha_{kg})\mu_{kg}^2 + \alpha_{kg}\nu_{kg} \tag{2.14}$$

This give the full reparameterization of β as follows:

1. Sample random error $\epsilon \sim \mathcal{N}(0, 1)$
2. Reparameterize node-specific gene embedding values $\beta_{kg} \leftarrow \mathbb{E}_q[\beta_{kg}] + \mathbb{V}_q[\beta_{kg}]^{\frac{1}{2}} * \epsilon - b_g$

Assuming $h_{kg} = 1$ with probability π_0 and $\beta|h = 1 \sim \mathcal{N}(0, \tau_0)$ *a priori*, the KL loss is derived as follows:

$$-D_{KL}(q||p) = \frac{\alpha_{kg}}{2}\left[1 + \ln\frac{\nu_{kg}}{\tau_0} - \frac{1}{\tau_0}(\mu_{kg}^2 + \nu_{kg})\right] + \tag{2.15}$$

$$\left[\alpha_{kg}\ln\frac{\pi_0}{\alpha_{kg}} + (1 - \alpha_{kg}\ln\left(\frac{1-\pi_0}{1-\alpha_{kg}}\right)\right] \tag{2.16}$$

2.1.4 Model implementation

LaRCH is implemented in Python, using the `PyTorch` machine learning library [58]. The model makes use of the `torch.autograd` function to perform stochastic variaitonal inference [59].

2.1.5 Code availability

Code containing the most up-to-date version of the `LaRCH` package, documentation, data simulation scripts of the simulation scheme presented in 2.2.2, and a model pipeline can be found at `https://github.com/causalpathlab/LaRCH`.

2.2 Modeling simulated data from bulk RNA-seq expression data

Simulated single-cell gene expression data of immune cell subtypes was generated based on bulk-sequenced gene expression profiles from the Database of Immune Cell Expression [60]. This generated data was then used to train an instance of the LaRCH model in order to test its efficacy.

2.2.1 Bulk gene expression profile generation

The DICE database bulk gene expression profiles are measured from human PBMCs obtained from leukapheresis samples. Immune cell types of interest were isolated from PBMC samples prior to total RNA isolation using fluorescence-activated cell sorting (FACS) based on fluorescent antibody staining either directly or following pre-enrichment using human B cell or memory CD4+ T cell isolation kits for B cell and CD4+ T cell samples, respectively [60]. FACS-sorted naive CD4+ and CD8+ T cells underwent an additional *ex vivo* CD3/CD28 activation step for their respective activation conditions.

Total RNA purified from FACS-sorted cell samples was then bulk sequenced and mapped against the hg19 reference genome and the GENCODE annotation v19 as the gene reference model. The expression profiles used in the data simulation scheme are expressed in transcripts per million (TPM) units. A full outline of the sample processing and RNA sequencing steps for this data can be found in Schmiedel et al. [60].

2.2.2 Data simulation scheme

The simulation scheme used to generate realistic single-cell gene expression data for N cells is as follows:

1. Determine the desired noise proportion ρ and total transcript count R. ρ represents the proportion of the gene expression distribution for a given 'cell' that is explained by a null distribution, π_0, common to all cell types. $1-\rho$ then represents the proportion of the gene expression distribution for a given 'cell' that is explained by a cell-type specific distribution $\hat{\pi}_t$.
2. Determine the null and cell type-specific gene expression distributions, π_0 and $\hat{\pi}$,

respectively:

$$\pi_0 \sim \text{Dirichlet}(1, \ldots, 1), \hat{\pi}_{tg} = \frac{\text{bulk}_{tg}}{\sum_g \text{bulk}_{tg}} \tag{2.17}$$

Where bulk_{tg} is the bulk expression profile for cell type t at gene g.

3. Determine a cell type distribution across the simulated data sample, m:

$$m \sim \text{Dirichlet}(1, \ldots, 1)$$

4. For a cell i, where $i = 1, \ldots, N$:
   (a) Randomly sample cell type t_i according to $t_i \sim \text{Categorical}(m)$
   (b) Gene expression distribution for cell type t_i:

$$\pi_{t_i} = (1 - \rho) * \hat{\pi}_{t_i} + \rho * \pi_0 \tag{2.18}$$

   (c) Sample gene transcript counts X_i:

$$X_i \sim \text{Multinomial}(R, \pi_t)$$

2.2.3 Resulting simulated data

Pseudo scRNA-seq datasets of $N = 2000$ cells and total transcript count $R = 2500$ with varying noise proportions ρ are obtained from the simulation scheme (Fig. 2.2). Total gene count value R is chosen to approximate UMI count values seen in real scRNA-seq data [61]. As expected, as ρ is increased, simulated gene expression profiles of immune cell types become less distinct from one another and at $\rho = 1$, the simulated cell population is entirely homogeneous.

This simulation scheme is able to generate realistic scRNA-seq data that captures the similarities and differences between the given cell types. From the principal component (PC) visualization in the first two PCs, it is seen that classical and non-classical monocytes are distinct from the simulated lymphocyte populations in the first PC (Fig. 2.2). Similarly, activated T cell populations are separated along the second PC. The population of inactivated T cells (CD4+ and CD8+ T cells) are clustered together in the first two PCs

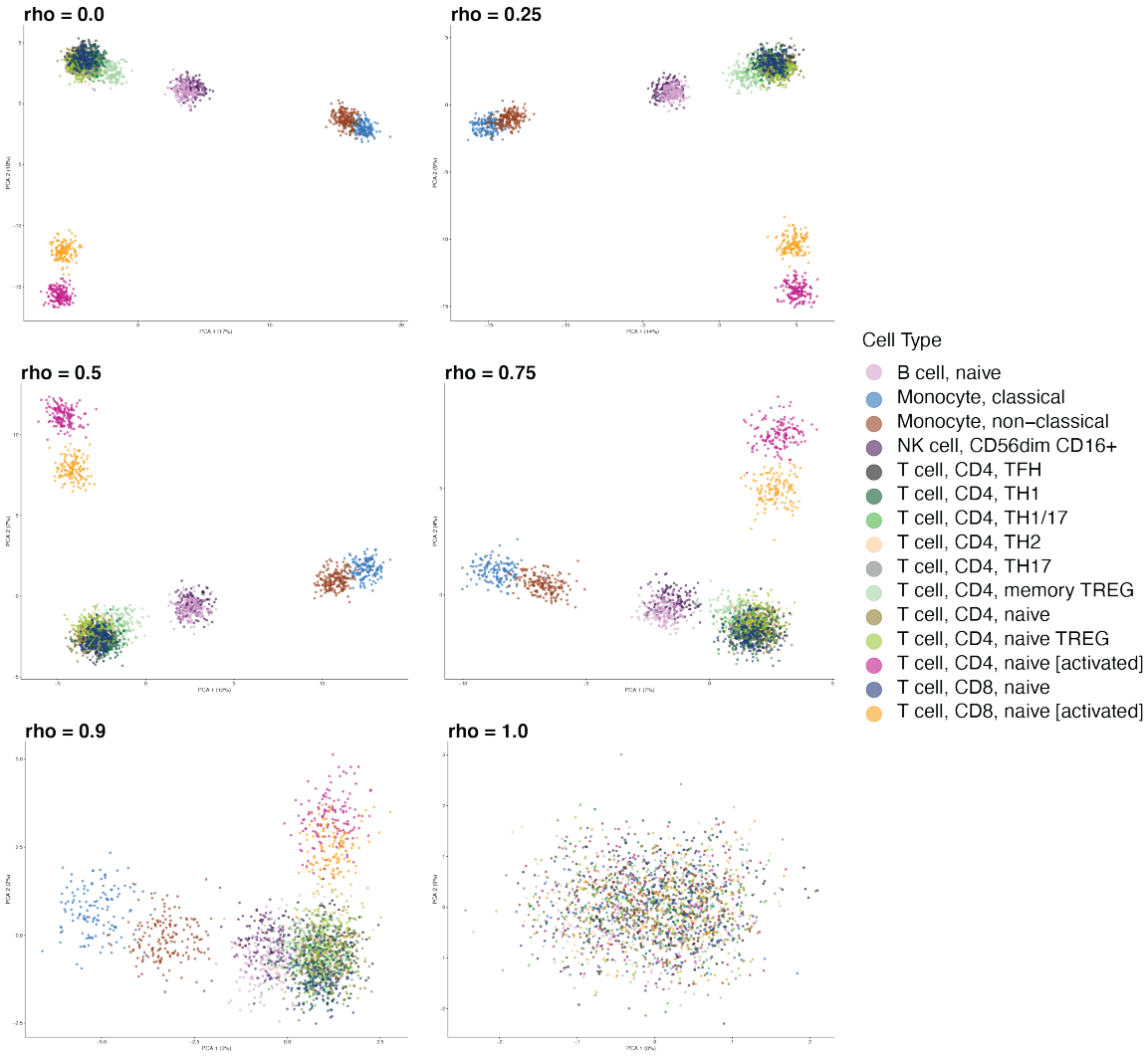

Figure 2.2: **Simulated scRNA-seq datasets exhibit realistic distribution of expression patterns across various noise levels** Shown are the plots of first two PCs for generated immune cell scRNA-seq datasets from the simulation scheme. Datasets generated using various noise proportions $\rho = \{0.0, 0.25, 0.5, 0.75, 0.9, 1.0\}$ are shown and coloured by cell type.

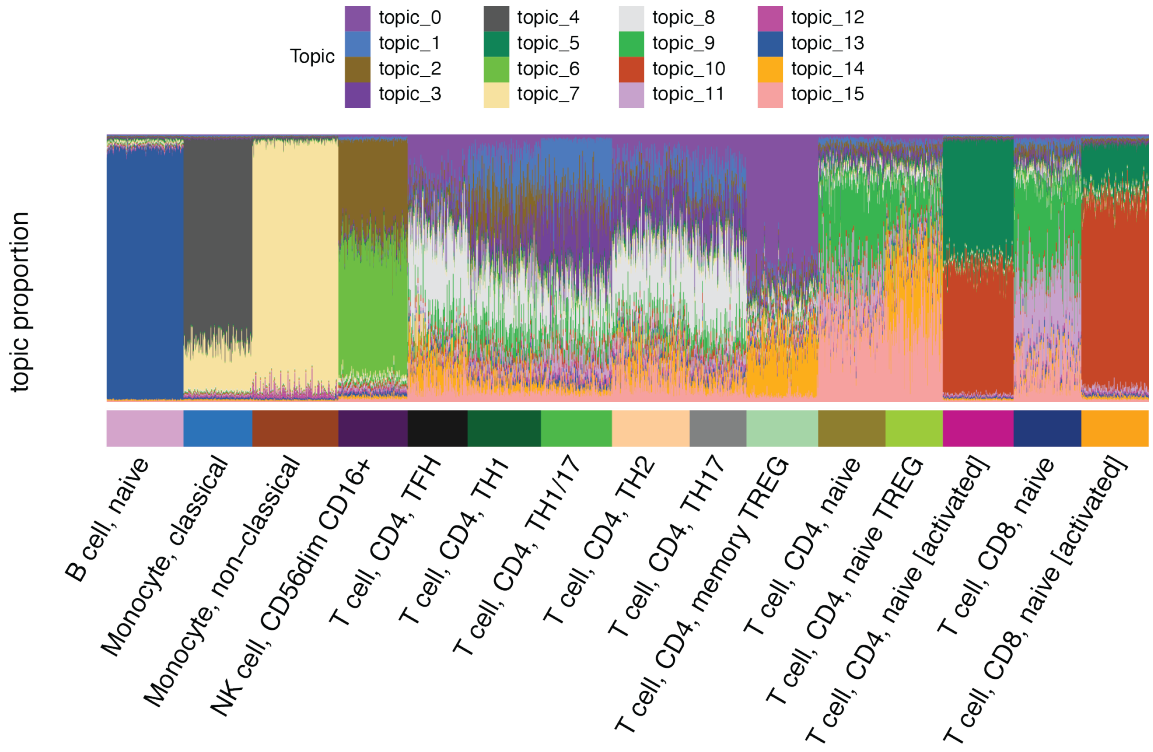

Figure 2.3: **Topic proportions of simulated immune cell gene expression shows distinct latent profiles between cell types** Structure plot of estimated topic proportion θ values for each cell in the simulated data set. Cells are annotated below by simulated cell type.

suggesting only minor differences in the gene expression profiles of these cells. In the third and fourth PCs, B cells and NK cells distinguish themselves from T cell subsets and each other while T cell subsets begin to separate (Fig. A.1).

For analysis, a simulated dataset with parameters $\rho = 0.5$ is used. At this ρ value, cell types are still preserved and can be well distinguished while maintaining a level of randomness that accounts for the noise seen in real scRNA-seq samples.

2.2.4 Cell type specific latent topic profiles

The LaRCH model was trained on a simulated immune cell scRNA-seq dataset with simulation parameters:

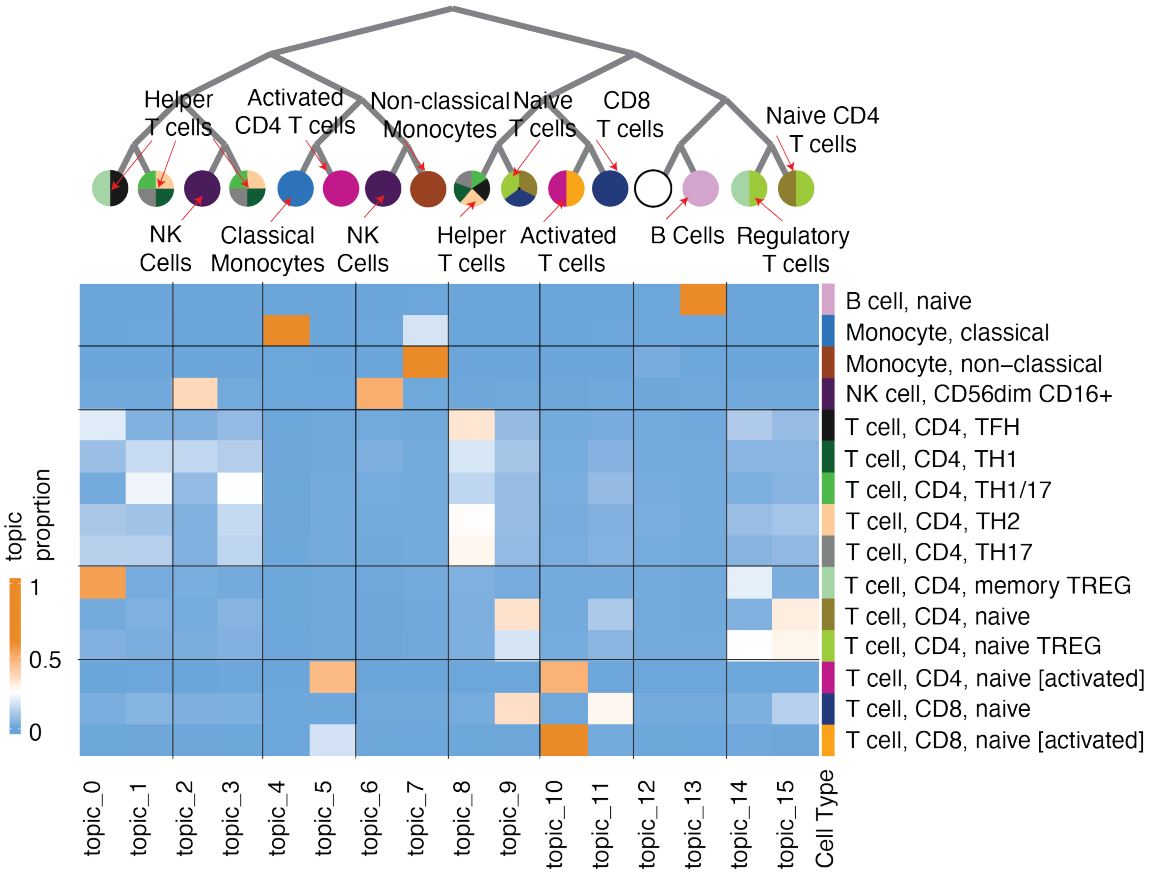

Figure 2.4: **Latent topic profiles can be used to reconstruct tree-structured relationships between cell subsets** The top panel shows the reconstructed tree structure from average latent topic profiles calculated for each cell subset. Below is a heatmap of the mean latent topic proportion profiles across each immune cell subset.

- Number of cells $N = 2000$
- Noise proportion $\rho = 0.5$
- Read depth $R = 2500$

and model parameters:

- tree depth $D = 5$
- prior inclusion probability $\pi_0 = 0.1$
- prior normal variance $\tau_0 = 1$
- batch size $|s| = 128$ for $s \in [S]$
- learning rate $lr = 0.01$
- local KL weight term $kl\_weight = 1.0$
- beta KL weight term $kl\_weight\_beta = 1.0$

The model is able to recover the simulated cell types and represent each cell subset with a unique topic profile (Fig. 2.3). Biologically similar cell types share topics. This is seen most prominantly amongst helper CD4+ T cell subsets which contain mixtures of topic 1, 3, and 8 (Fig. 2.4). This suggests that the these cell subsets share overlapping expression features. The minor differences in proportions of each topic for each cell subset demonstrates that each helper T cell subset expresses the genetic features captured by each topic to varying levels. Topic 14 is highly represented in regulatory CD4+ T cell subsets, suggesting a regulatory signature being captured in the leaf node within this topic. Similarly, topic 15 is seen in naïve CD4+ T cells, possibly capturing a naïve signature.

Using the topic average latent topic representations of cell subsets, the underlying tree structure is reconstructed qualitatively (Fig. 2.4). Topics with high representation of particular subsets are labeled as a marker topic for said subset. The reconstructed cell hierarchy shares some noticeable similarities to the lineage tree representation of immune cell subsets (Fig. 1.1).

The latter half of the tree representation contains only cells within the adaptive immune system. Monocytes, which branch from the other subsets at the earliest point in the differentiation process, are solely contained within topics 4 and 7 which share 3 nodes along

their paths. Activated CD4+ T cells share a path with classical monocytes, suggesting some activation features captured in their shared nodes.

Similarly, NK cells, which are involved in both innate and adaptive immune function, are represented in topic 7, sharing a node path with non-classical monocytes and it is likely that innate immune response gene signatures are captured within their shared paths, especially in nodes 4 and 10 which are not shared by any other subsets. NK cells are also represented by topic 2 which shares a node path with topic 3 which represents a number of helper T cell subsets. The shared nodes between these topics (nodes 1, 3, 8) likely contain gene embeddings related to adaptive immune response.

Topics 8 through 11 solely represent T cell subsets, capturing overlapping T lymphocyte gene expression patterns. More distinct cell types are contained within unique branches of the tree, such as B cells which are the only cell type represented in the branch containing topics 12 and 13.

Topic 14 captures features of regulatory T cell subsets which topic 15 captures those of naïve CD4+ T cells. This suggests that the nodes that constitute the shared path between these two topics (nodes 0, 2, 6, and 14) contain embeddings of genes related to CD4+ T cell function that are shared between all these T cell subsets, while node 29, the leaf node of topic 14 contains gene embeddings related to regulatory function and node 30, the leaf node of topic 15 is related to naïve expression profiles.

2.2.5 Using latent features in place of other dimensionality reduction techniques for downstream analysis

To test the efficacy of the LaRCH model and its ability to perform cell type deconvolution from single-cell gene expression data, clustering was performed using the latent topic features estimated from simulated data.

The Louvain method of community detection [62] was used to extract cell type clusters from a K-nearest neighbour graph ($k = 10$ was used) constructed using the latent topic features to calculate distance between data points. In general, the resulting cell clusters separate the various simulated cell types, missing only some granularity between closely related cell types (Fig. 2.5). In this case, the clustering done is unable to distinguish between certain subsets of T cells. The helper T cell subsets of T follicular helpers, Th17s, and Th2s are all contained within cluster 3 from figure 2.5, it is seen that these cells lie closely in the t-

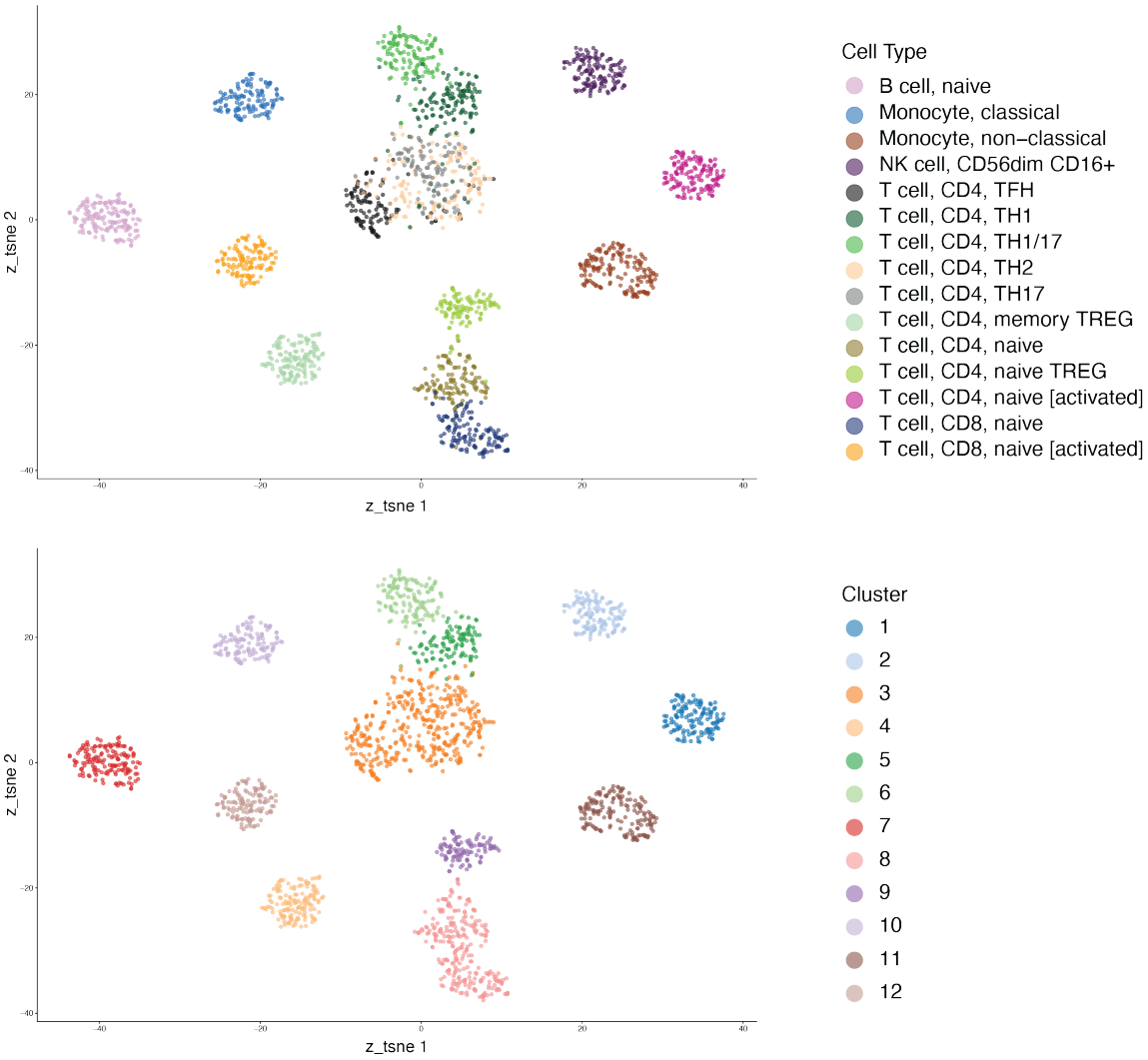

Figure 2.5: **Louvain lustering on latent topic space recovers groups of cells corresponding to their simulated cell type** The top panel shows the simulated cells in tSNE space coloured by original simulated cell type. The bottom panel shows the clusters generated using louvain clustering on the latent topic representation of cells. For visualization, tSNE dimensions were generated from the latent topic space.

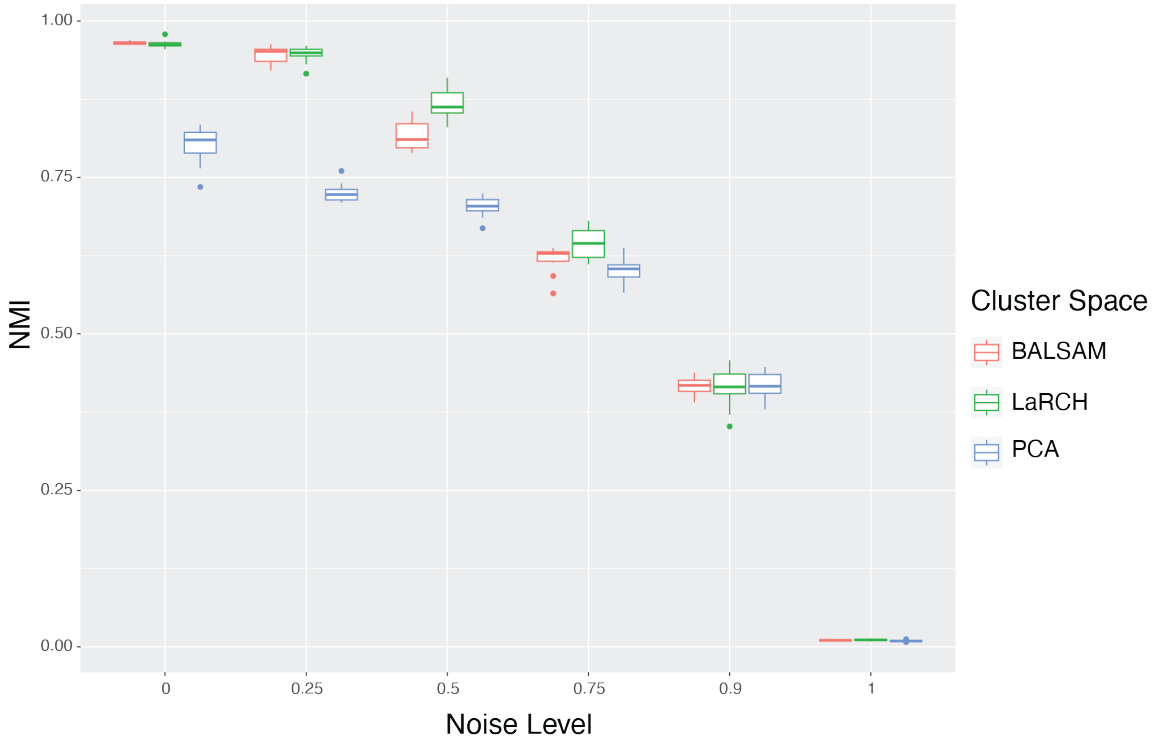

Figure 2.6: **Using the LaRCH latent space for clustering consistently outperforms clustering on PCA and flat latent topic space** NMI calculated from clustering results from each dimensionality reduction technique and the original simulated cell types across various noise levels. At each noise level, 10 datasets were simulated with different starting seeds.

distributed stochastic neighbor embedding (tSNE) reduced dimension space, showing that the simulated transcriptional profiles of these cell are very similar. Additionally, cluster 8 contains the naïve subsets of both CD4+ and CD8+ cell types. The grouping of these cell types within the latent feature space suggests an overlap in overall gene expression profile and cellular function between cell types. Changing Louvain clustering parameters ($k = 5$) creates a more granular clustering that more accurately separates individual cell types, but this also runs the risk of over-clustering.

Clustering results from latent topic features as the clustering space were compared to clusters obtained from the first 50 PCs as well as BALSAM [49] with 16 latent topics,

the same number of latent topics as a LaRCH model with depth $D = 5$. BALSAM is a comparable ETM that utilizes the same encoder-decoder scheme as LaRCH without the underlying tree structure. At each noise level ($\rho = 0$, 0.25, 0.5, 0.75, 0.9, 1), 10 datasets were simulated with different starting seeds. The Louvain method of community detection [62] was then used to detect clusters from the K-nearest neighbour graphs ($k = 10$ was used) constructed from each reduced dimension space. The normalized mutual information (NMI) was calculated from each clustering result and original simulated cell types to compare clustering accuracy (Fig. 2.6).

At all noise proportions, clustering from the LaRCH latent feature space outperforms PCA clusters. At low noise proportions, BALSAM and LaRCH are comparable in NMI while at mid-level noise proportions ($rho = 0.5$ and $\rho = 0.75$), LaRCH performs slightly better than BALSAM, showing that the underlying tree-structure provides additional insight on cell type features to the latent topics that are not captures in a flat model. At $rho = 0.9$, where the simluated data is primarily noise and no longer realistic in its representation of a real scRNA-seq dataset, the NMI drops below 0.5 and all clustering methods struggle.

Chapter 3

Modeling Cellular Hierarchies in a Real scRNA-seq Dataset

As a practical application of the LaRCH model, the model is trained on a real single-cell RNA-seq dataset used to study the cell-type specific molecular and genetic associations of SLE [1]. In this section, details of the scRNA-seq data are described and the resulting model trained on the dataset is used to explore various aspects of the disease pathogenesis of SLE.

3.1 Single-cell RNA-seq data from a large SLE dataset

The scRNA-seq dataset used to train the model is from a 2022 paper from Perez *et al.* [1]. This dataset consists of single-cell transcriptome data of 1.2 million PBMCS from 99 healthy control individuals and 162 individuals with SLE.

3.1.1 Dataset details

PBMC samples used in this analysis were collection from SLE cases in the California Lupus Epidemiological Study (CLUES) cohort. Matched healthy control samples were collected from the UCSF Rheumatology Clinic and the Immune Variaton Project (ImmVar) [1].

Single-cell gene expression profiles from pooled antibody-stained and unstained PBMCs was generated using 10x Genomics' Chromium Single Cell 3' V2 chemistry and processed with the 10x Cell Ranger pipeline. Filtered scRNA-seq read count data was obtained directly from accession number GSE174188 in the Gene Expression Omnibus (GEO). Additionally, patient metadata, cell type annotation, and UMAP projections from the original study were used for the data analysis in this chapter.

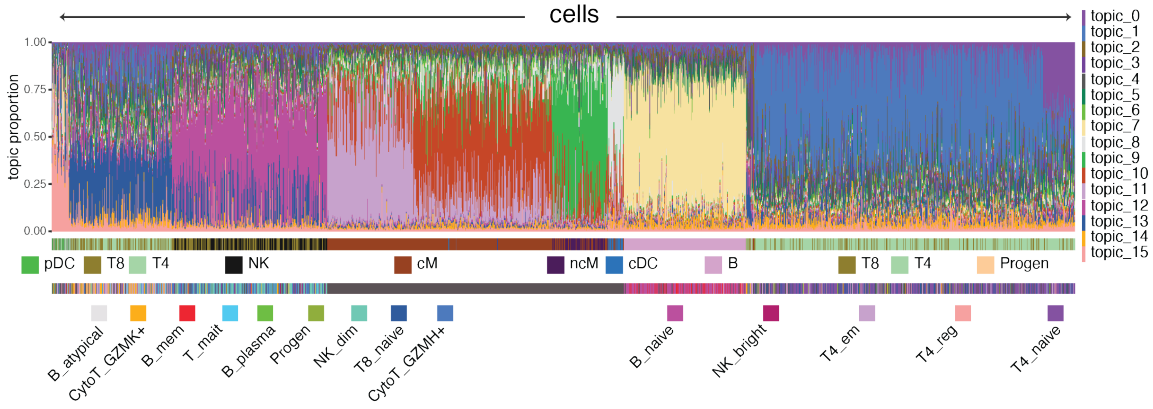

Figure 3.1: **Distinct topic proportions of PBMCs in a scRNA-seq dataset of SLE and healthy individuals correspond to cell type labels** Structure plot of estimated θ values for a subsample of 10,000 cells in the data set. Cells are annotated below by labelled cell types obtained from Perez *et al.* [1]

3.2 Deconvolution of immune cell subsets

Estimated topic proportion profiles show distinct groups of cells within the latent space that align closely with the cell types determined in the original Perez *et al.* paper [1] (Fig.3.1). The topic representation of cells is able to pick up more granularity that potentially describes further distinct cell types, namely in the cells labelled as 'classical monocytes' where there are two distinct groups in the topic space primarily represented by either topic 10 or topic 11, suggesting classical monocyte-labeled subsets could be composed of subsets with differing cell function.

As with the simulated data, a tree structure outlining the hierarchical relationships between cell types is qualitatively assembled using the average latent profiles of immune cell subsets. Similar to the simulated data setting, the constructed tree structure from the real data set aligns with the canonical understanding of immune cell differentiation (Fig.3.2). B cells lie within their own distinct branch of the tree. Myeloid cells, including monocytes and dendritic cells are contained within the tree branch with topics 8 through 11. Topic 15 captures a number of intermediate immune cell-states, including progenitor and proliferating lymphoid cells, proliferating dendritic cells, and plasmablasts. Naive CD4+ T cells are primarily represented by topic 0 and 1, while the remaining CD4+ subtypes are

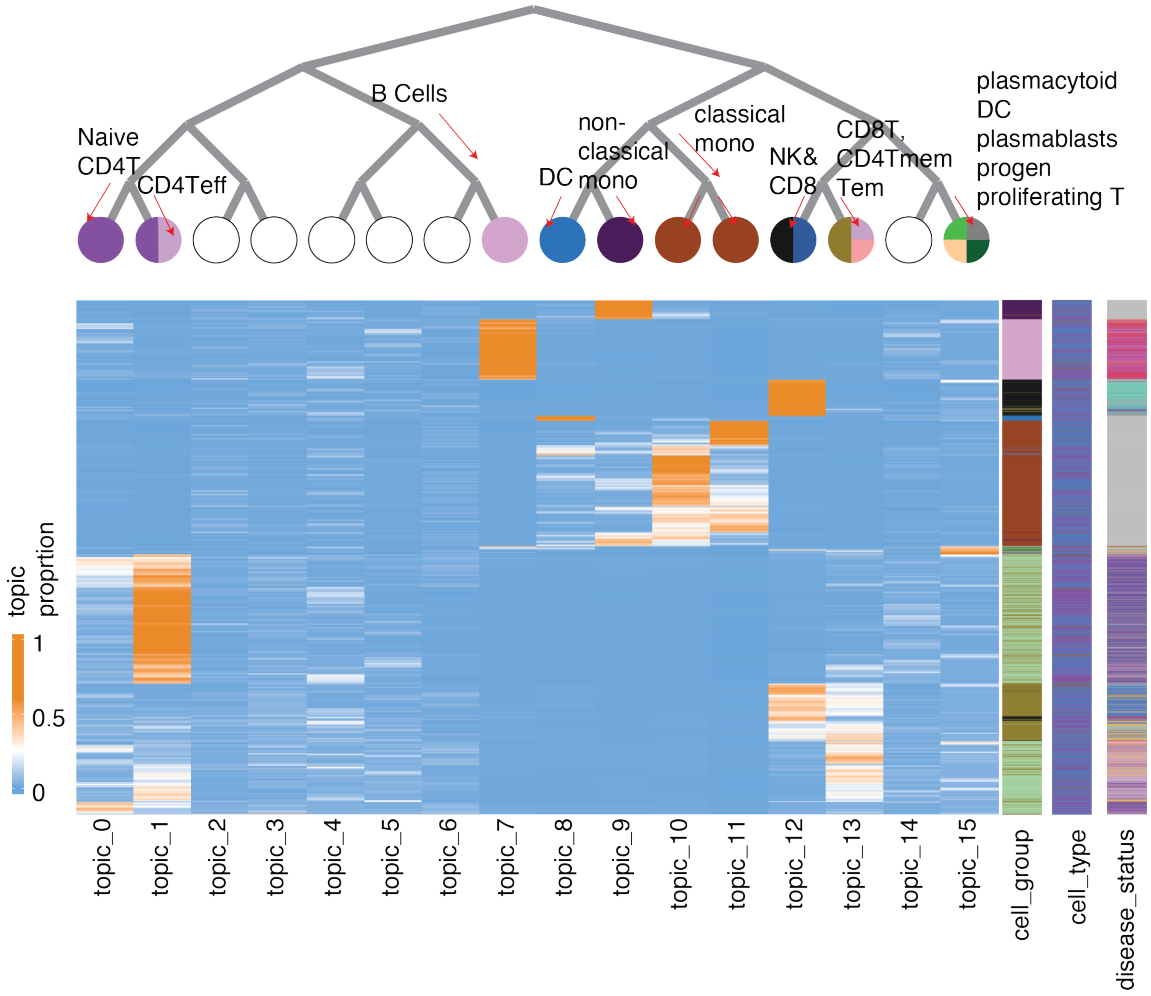

Figure 3.2: **Tree-structure relationships between cell types are reconstructed using latent topic profiles of a real scRNA-seq dataset** The top panel shows the reconstructed tree structure based on average latent topic profile representations of cell types. Below is a heatmap of the latent topic proportion profiles across a subset of 10,000 cells from the dataset.

represented by topic 13 along with non-naive CD8+ T cells. NK cells and naive CD8+ T cells are found in topic 12.

3.2.1 Cell clustering

Louvain clustering [62] using $k = 30$ nearest neighbours in the latent topic space results in clustering that aligns with cell groups described by Perez *et al.* [1] (Fig.3.3, 3.4a), but quickly loses the ability to clearly distinguish cell the more specific cell types (Fig.3.4b). For this purpose, similarity is measured as the proportion of cells with a specified label contained within a cluster, mathematically this is:

$$\text{Similarity}(\text{cell\_label}_j, \text{cluster}_k) = \frac{\sum_{i=0}^{N} \mathbb{I}(\text{cell\_label}(\text{cell}_i) = \text{cell\_label}_j) * \mathbb{I}(\text{cluster}(\text{cell}_i) = \text{cluster}_k)}{\sum_{i=0}^{N} \mathbb{I}(\text{cell\_label}(\text{cell}_i) = \text{cell\_label}_j)}$$

Where N is total number of cells in the dataset and $\text{cell\_label}(\text{cell}_i)$ and $\text{cluster}(\text{cell}_i)$ return the assigned cell label and cluster for cell_i, respectively.

Since the original cell types are assigned from Louvain clustering on UMAP reduced dimensions and canonical marker gene expression, it is unclear as to which clustering method produces higher-quality immune function-relevant cluster assignments.

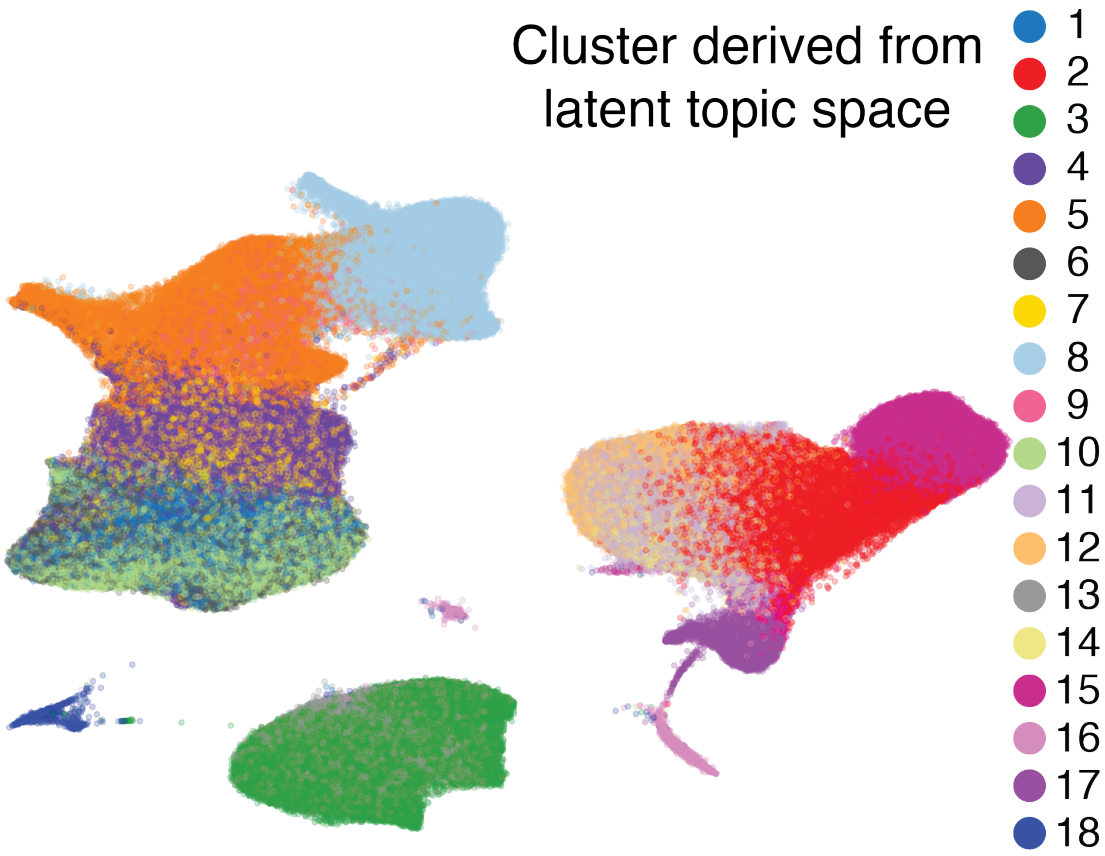

Figure 3.3: **Clustering on latent topic space recovers groups of immune cells belonging to related immune subsets** UMAP projection of scRNA-seq data are coloured by Louvain clusters derived from the latent topic space. UMAP projection values obtained from the original Perez *et al.* study [1]

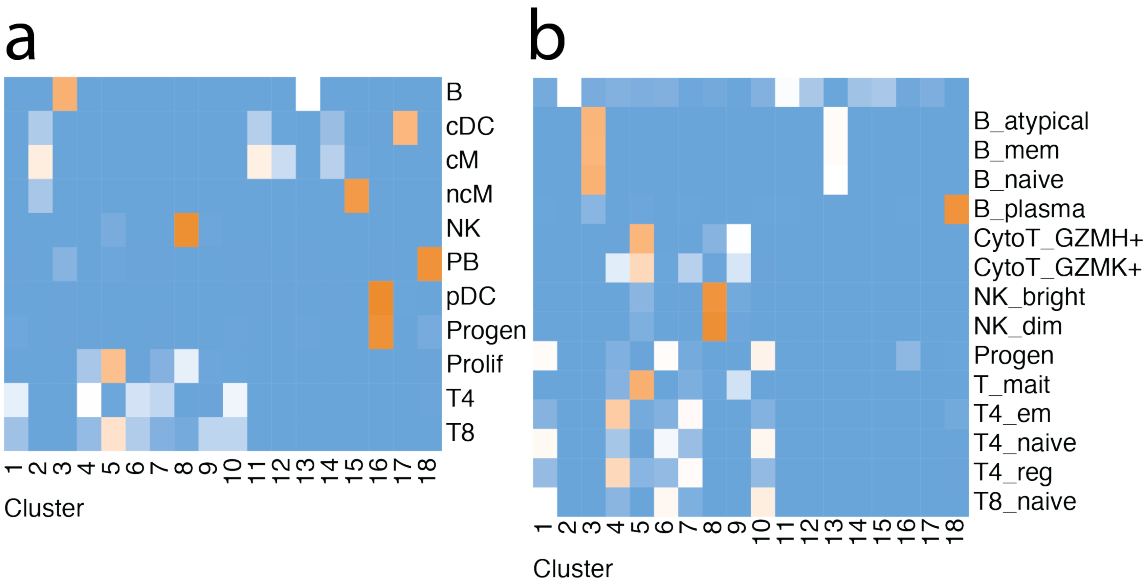

Figure 3.4: **Overlap exists between latent topic clusters and original cell labels** a) Heatmap depicting the similarity matrix between louvain clusters obtained from latent topic space and cell group labels. b) Heatmap depicting the similarity matrix between louvain clusters obtained from latent topic space and cell type labels. Cell clusters are represented along the columns of the heatmap while cell labels are represented along the rows. Cell group and cell types are described in [1]. Cell groups describe more general immune cell subsets while cell types describe specific immune cell subsets.

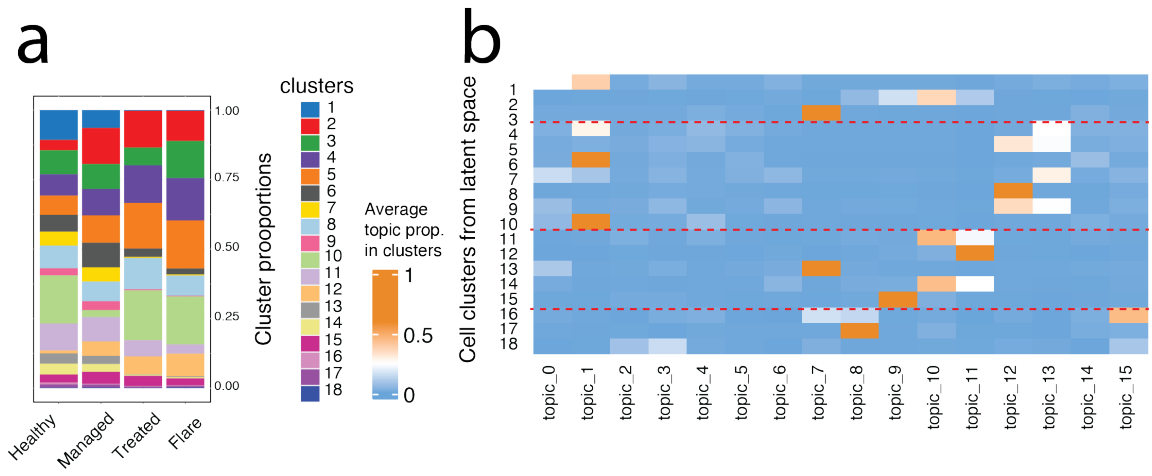

Figure 3.5: **Cluster composition within samples is disease status dependent** a) Cell cluster proportion breakdown by SLE disease status shows cell composition within samples. b) Heatmap showing the median values of latent topic proportions across cells within each Louvain cluster. Clusters are shown along the rows of the heatmap.

3.3 Disease dependent differences in gene expression

From the provided metadata of the scRNA-seq dataset in Perez *et al.* [1], the trained model was analyzed in a disease-stratified fashion. Using the same Louvain clusters determined in the previous section (Fig.3.3), differentially represented clusters were found depending on SLE disease status (Fig.3.5a). Clusters 1, 6, 7, 9 and 11 capture primarily cells from healthy individuals or individuals with managed SLE. Cells in clusters 2 and 12 make up a significantly greater portion of cells from individuals with any level of SLE compared to healthy individuals. Some clusters make up a greater portion of cells from individuals with active disease phenotypes of SLE (treated vs. flare), such as clusters 4 and 5. Finally, cluster 10 appears in each condition except for individuals with managed SLE. The differences in cluster representation suggest that the cellular makeup of PBMCs is associated with disease status.

To understand how each cluster relates to the estimated latent topic representation of cells, the average topic profile for cells in each cluster is constructed (Fig.3.5b). From this, it is shown that many clusters that are differentially represented across disease conditions actually share many features, suggesting that cells in different clusters may correspond

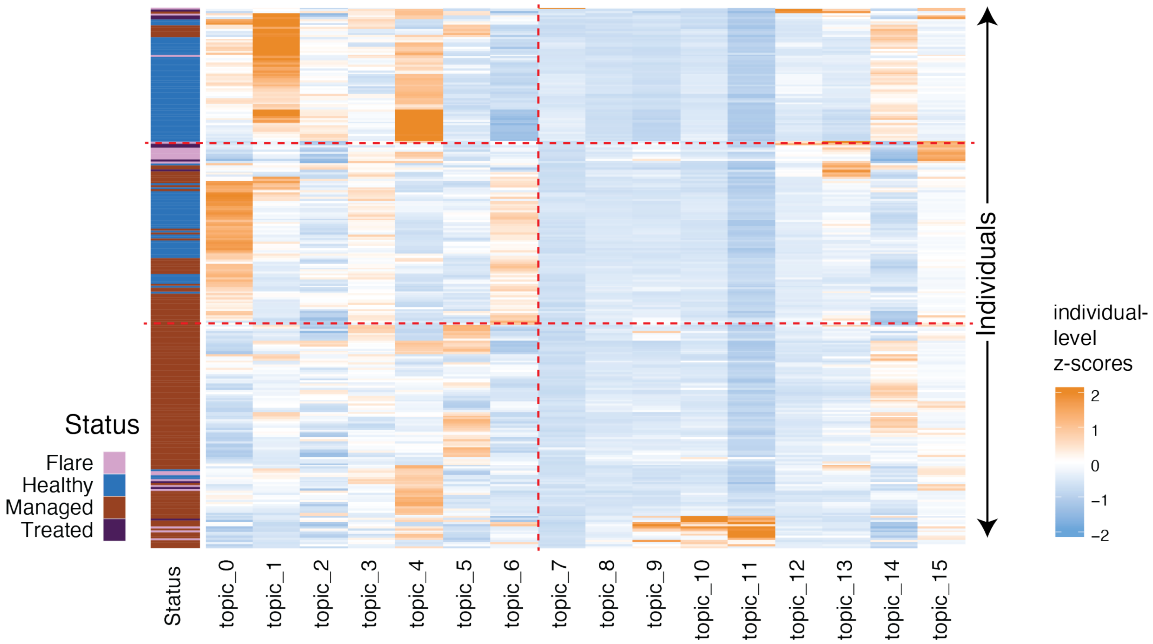

Figure 3.6: **Individuals with identical disease status generally share latent topic patterns** Heatmap shows median latent topic profiles calculated for each individual included in the study. Rows represent each individual and are annotated by disease status.

to the same general cell type, but have some minute differences in gene expression. For example, clusters 4 and 7 both have strong representations by topics 1 and 13 but cells in cluster 7 have a higher representation in topic 0. Cells in these clusters are shown to be most closely aligned with effector memory and regulatory CD4+ T cells (Fig.3.4), yet are distinctly different in their latent topic profiles.

To better understand differences in topic representation as an effect of disease phenotype, the median topic profile for each individual included in the study is constructed (Fig.3.6). From the latent profiles, a large variance in topic profiles across individuals is seen, yet individuals with similar profiles tend to share the same disease status. Wilcoxon rank-sum tests were performed on median latent topic values of individuals between disease status groups (Fig.3.7). Of note, topics 0 and 1 are significantly differentially represented in healthy individuals compared to individuals with SLE of any status. Topic 0 separates cells found in cluster 7 from those in cluster 4, while topic 1 is seen in all clusters containing

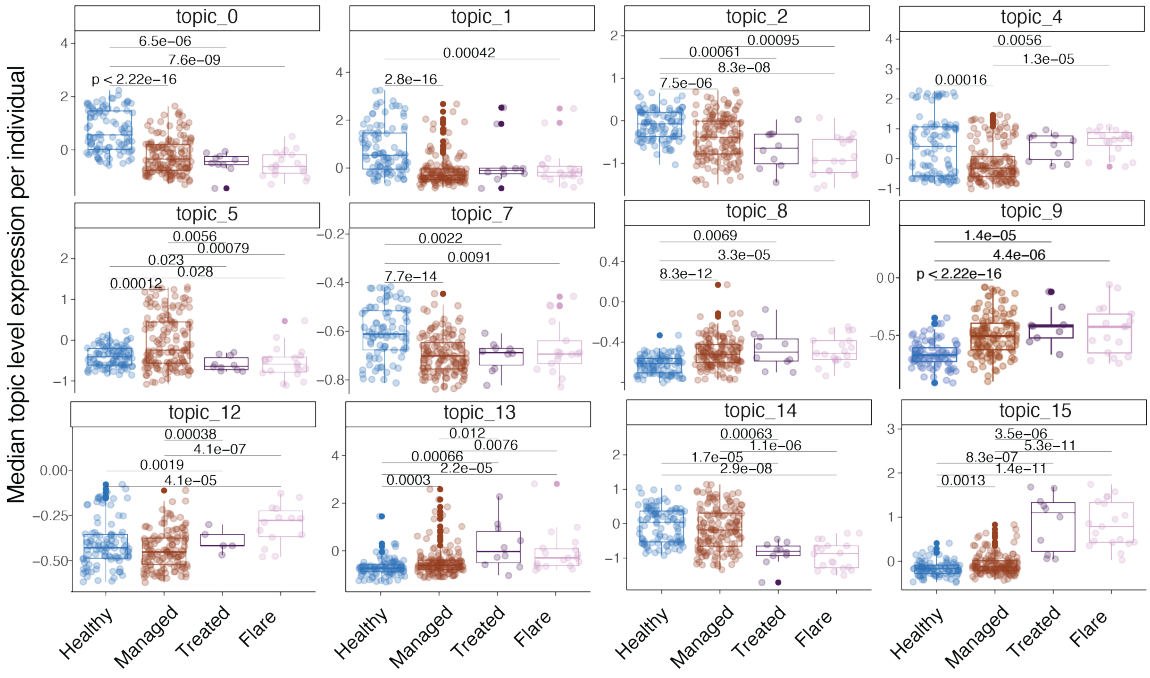

Figure 3.7: **Disease status dependent differences exist in topic representation** Boxplots of median topic expressions per individual grouped by disease status. Topics with significant differences in median topic values between healthy and SLE individuals are shown. Significance between median expression values are calculated using Wilcoxon rank-sum tests between disease status individuals.

CD4+ T cells. Topic 12, which is represented in clusters corresponding to NK cells, CD8+ T cells and proliferating lymphocytes(Fig.3.4, Fig.3.5b), is significantly higher in individuals with treated and flare SLE conditions compared to healthy and managed SLE. Healthy individuals and those with managed SLE have a significantly higher level of topic 14 than those with an active SLE status, while the inverse is true for topic 15, suggesting a higher prevalence of intermediate immune cell states in individuals with active SLE while healthy and managed individuals have a higher prevalence of B cells and certain naive T lymphocytes (Fig.3.4, Fig.3.5b). Showing the differences in topic levels across disease statuses informs our investigation into the differentially regulated biological gene sets suggested by the latent-node level gene embeddings of the model.

3.4 Functional analysis of latent nodes

3.4.1 Node-level gene clustering

Further investigation was conducted on the number of genes that effectively define cell-type specific disease mechanisms using learned node-level gene vectors from LaRCH, namely $\hat{\beta}$ parameters before the tree-based topic-level aggregation. It is assumed that Bayesian variable selection prior can put zero values to unnecessary (or statistically weak) associations for irrelevant genes. Counting only the number of genes significantly deviating from zero, the number of genes needed for each tree node and level was counted (Fig. 3.8). For brevity, genes were clustered into 25 distinctive gene modules by applying the Louvain clustering method on the 10-nearest neighbour gene-gene interaction graph (Fig.3.8). The Bayesian spike-and-slap prior ensured that the node-level gene programs are determined by a subset of genes between 50 to 1800 genes, with a median number of 429. Albeit, more genes were found on nodes 5 (cluster 24) and 7 (cluster 19), which correspond to T- and B-cell groups, respectively.

3.4.2 Gene set enrichment analysis of node-level gene embeddings

Enrichment analysis of cell-type specific differentially expressed gene sets based on gene embedding values allows for further annotation of cell type relationships to latent topics. Ranked gene set enrichment analysis was performed using the `fgsea` R package [63] with the estimated embedding values $\hat{\beta}$ as the fgsea score. Gene sets used for enrichment analysis were obtained from the DICE database [60], the NHGRI-EBI Catalog of human genome-wide association studies [64], ImmuneSigDB [65], and the hallmark gene sets within MSigDB [66].

Gene sets related to immune function were found to be enriched in a number of relevant latnet nodes, suggesting functional phenotypic features represented through latent topic representation (Fig. A.3). B cell-related genes are enriched at the early stages of the node tree, suggesting a distinct model lineage that these cells follow (Fig.3.9a). an enrichment of B cell-related genes in node 23 is also seen, along with classical monocyte and memory regulatory T cell-related genes. In Fig.3.2, topic 8 (made up of nodes 0, 2, 5, 11, and 23) was determined to correspond to dendritic cells, which canonically share lineage paths with both monocytes and lymphoid cells [2]. Topic 1 (made up of nodes 0, 1, 3, 7, and 16), which is significantly greater in healthy individuals (Fig.3.7), is enriched in genes relating

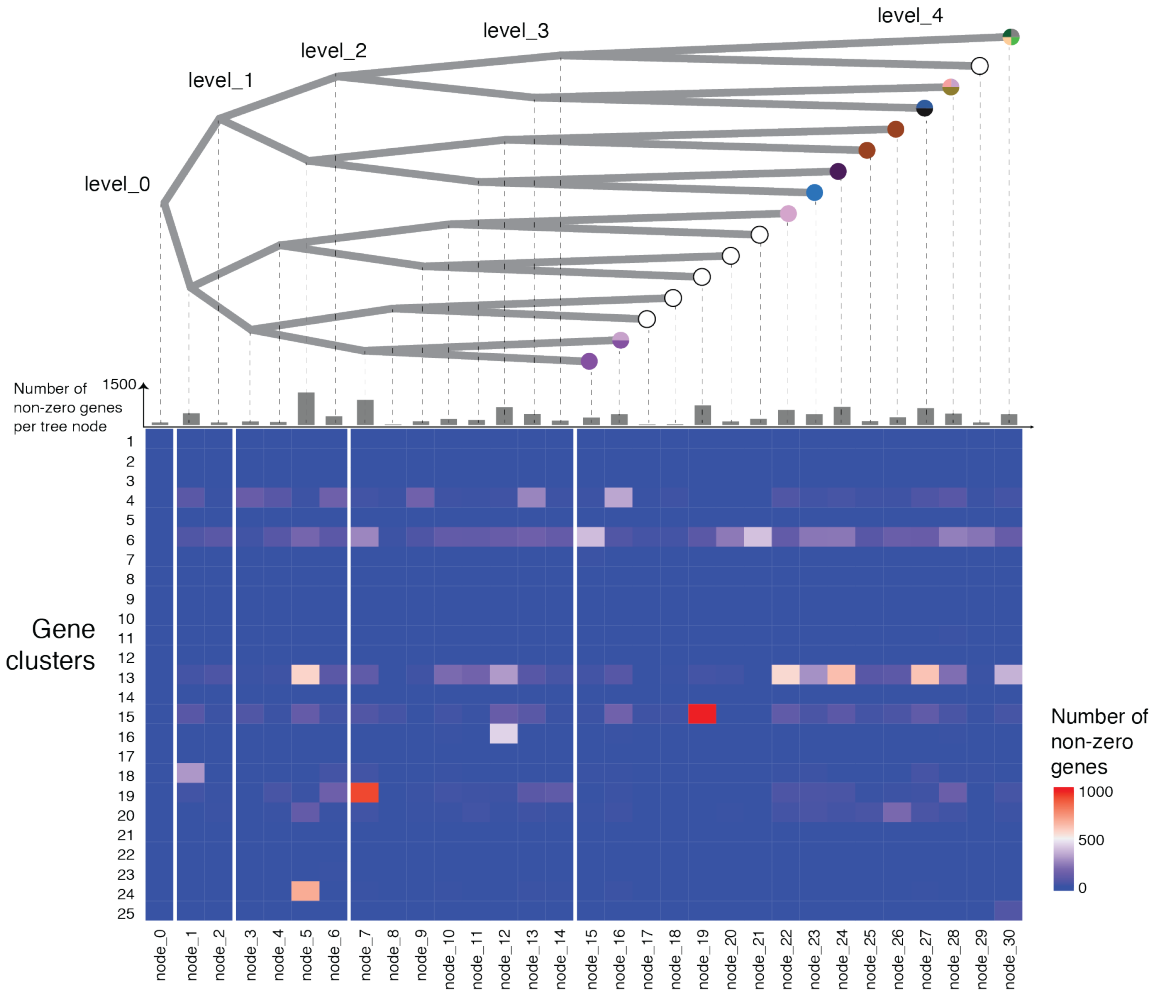

Figure 3.8: **Clustering of genes based on node-level embeddings show overarching gene programs** Heatmap shows number of significantly non-zero genes within each node that belong to gene clusters. Total numbers of non-zero genes per tree node are annotated above.

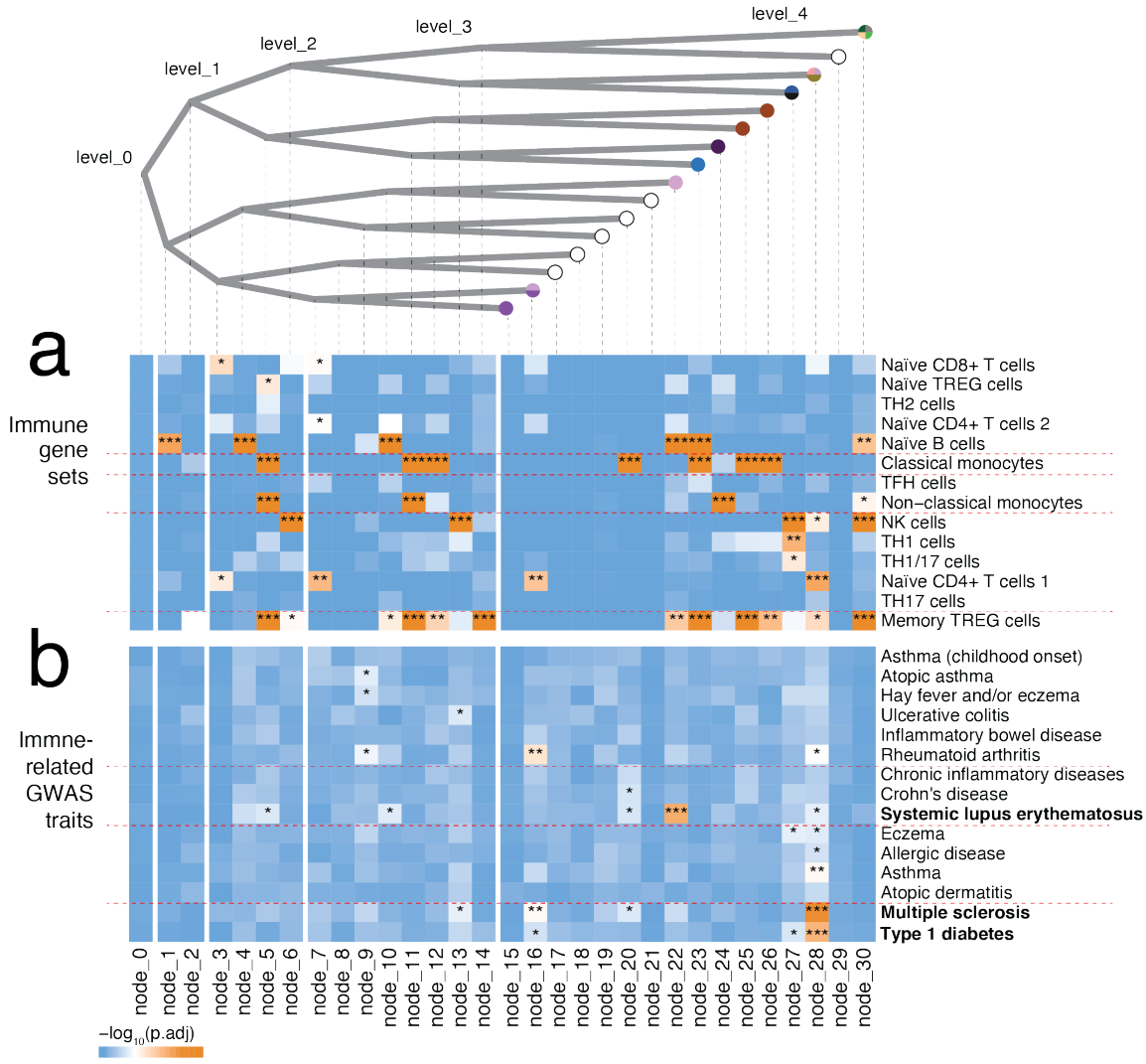

Figure 3.9: **Gene set enrichment analysis shows immune subset and disease relevant features described by latent node features** Heatmap shows the $-\log(p)$ enrichment significance of each gene set per node. Significance annotations are as follows: $***-p<0.0001, **-p<0.001, *-p<0.01$.

to naive T cells, suggesting a loss of naive cells and concomitant increase in activated T cells in individuals with SLE. Gene sets of less prevalent sub-types, such as Th17 and Th2 helper T cell subsets, are not found to be significantly enriched in any node as they are difficult to distinguish amidst a heterogeneous sample.

Enrichment analysis of GWAS (GWAS) gene sets pertaining to immunological disease reveals node-level differences in disease manifestation (Fig.3.9b). Disease-relevant gene sets are mainly found to be enriched in nodes at deeper levels of the tree, suggesting that disease manifestations are cell-type specific, rather than common features across many cells. GWAS genes relating to SLE are enriched primarily in node 22, which corresponds to topic 7. This node also is related to B cell function (Fig.3.8b), which is in line with the fact that SLE is often characterized by abnormal autoreactive B cells [67]. Node 28, corresponding to topic 13, is enriched in gene sets relating to other autoimmune diseases, such as multiple sclerosis and type-1 diabetes, as well as memory Treg cells, naive CD4+ T cells, and NK cells. Autoreactive T cell function is the primary driver of these diseases [68, 69] and the corresponding topic 13 is enriched in individuals with SLE (Fig.3.7), suggesting an additional role of autoreactive T cells in SLE pathogenesis.

3.5 Node-level marker genes

Node-level gene embedding values allow for the identification of significant marker genes that contribute to the gene expression signature of cells represented by specific latent topics. From the node-level marker genes, biological meaning can be extracted to provide insight on cell-type specific features associated with SLE.

3.5.1 Marker gene significance testing

Significant marker genes for each latent node were obtained using parameters learned in the decoder component of LaRCH. $\hat{\beta}$ and $\mathbb{V}[\hat{\beta}]$, calculated using (2.13) and (2.14), were fed directly into the `ashr` R package described in the Stephens *et al.* [70]. This package uses an emprical Bayes approach for large-scale hypothesis testing and false discovery rate estimation. Significant genes are selected as those with `qvalue` < 0.05. The number of significant genes per node is detailed in table A.2. Unique significant genes are used to identify gene features that separate a specific latent tree node from its parent and sibling nodes, therefore playing a role in driving the differentiation from one topic path to another. A significant gene is considered unique to a node if it does not also appear as a significant gene it its parent a direct sibling node. These gene counts are summarized in table A.3.

3.5.2 Marker genes of interest

From the list of significant genes for each tree node, the top ten genes with the greatest absolute $\hat{\beta}$ value from each node are selected for further investigation. The top 80 genes from this list are shown in figure A.2.

Significant node-level immune marker genes

Within this list are several known immune cell type markers (Fig. 3.10).

MAL is a protein coding gene that is selectively expressed during T cell differentiation [71]. It is seen in nodes 7 and 28, which correspond to topics 0, 1, and 13. This is in accordance with the labeling of these topics as representing T cell subsets in figure 3.2. CD69 is a membrane-bound type II C-lectin receptor and is a marker of lymphocyte activation [72]. It is identified as a marker gene in a number of nodes, all which correspond to topics containing lymphocyte subsets (0, 1, 7, 12, 13). Interestingly, this gene is also down regulated in nodes associated with monocyte subsets.

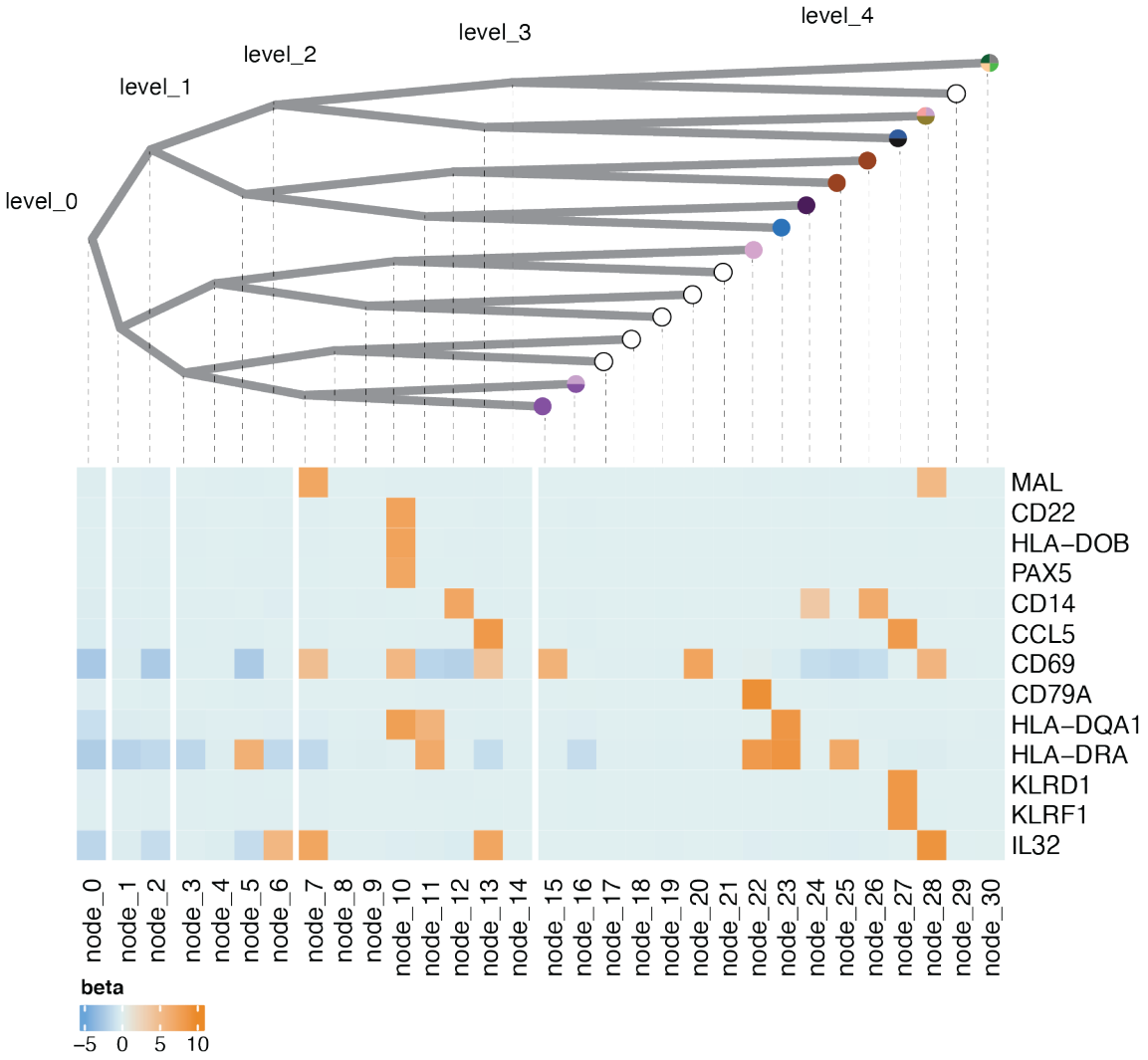

Figure 3.10: **Canonical immune subset markers are captured in latent tree nodes of the topic model** Heatmap shows estimated node-wise gene embedding values $\hat{\beta}$ of genes previously identified as marker gene for immune cell subsets.

NK cells are identified by KLRD1 and KLRF1, two killer cell lectin like receptors which are markers of node 27, the leaf node of topic 12 [73].

In nodes corresponding to topic 7, which is almost exclusively present in B cells, a number of B cell markers are present. In node 10 and 22, the two final nodes in the path of topic 7, CD22, HLA-DOB, PAX5, and CD79a are uniquely expressed in B cells. [74–77] HLA-DRA is identified as a marker gene in node 22 along with a number of other tree nodes that correspond with topics 8-11 which are associated with dendritic cells and monocytes. HLA-DRA is a HLA class II molecule that is a component of the major histocompatibility complex (MHC) II and is expressed on the surface of various antigen presenting cells (APCs) including B lymphocytes, dendritic cells, and monocytes [78]. CD14, a well known marker of monocyte [79], is a top gene in nodes 12, 24, and 26 which correspond to topics 9, 10, and 11, topics associated with monocytes.

From the analysis in this chapter, node-level immune cell markers are congruent with cell type assignments presented in Perez *et al.* [1]. The presentation of immune cell type marker genes as significant genes in tree nodes demonstrates the ability for the LaRCH model results to inform cell type assignments to scRNA-seq datasets using their latent topic representations.

Significant node-level genes with potential SLE disease association

Analysis of significant genes outside of canonical immune subset markers paired with the SLE disease status dependent differential topic representation provides insight into potential cell-type specific mechanisms associated with disease phenotypes.

Patients experiencing SLE flares show significantly elevated representation of topic 12 3.7. This topic is associated with CD8+ T lymphocytes and nodes in this topic contain GZMH, GZMB, and PRF1 as marker genes. It has been shown that patients with SLE show elevated levels of GZMH+ cytotoxic CD8+ T cells [1]. Similarly, PRF1 and GZMB expressing CD8+ T cells have been correlated with SLE disease activity [80].

Topic 7, which contains node 22, has a significantly lower representation in patients with SLE. Node 22 has a number of significant genes associated with B cell function, including BANK1. BANK1 encodes a scaffolding protein expressed predominantly in B cells. Variants in this gene have been shown to elevate risk of SLE by increasing susceptibility to autoantibodies [81].

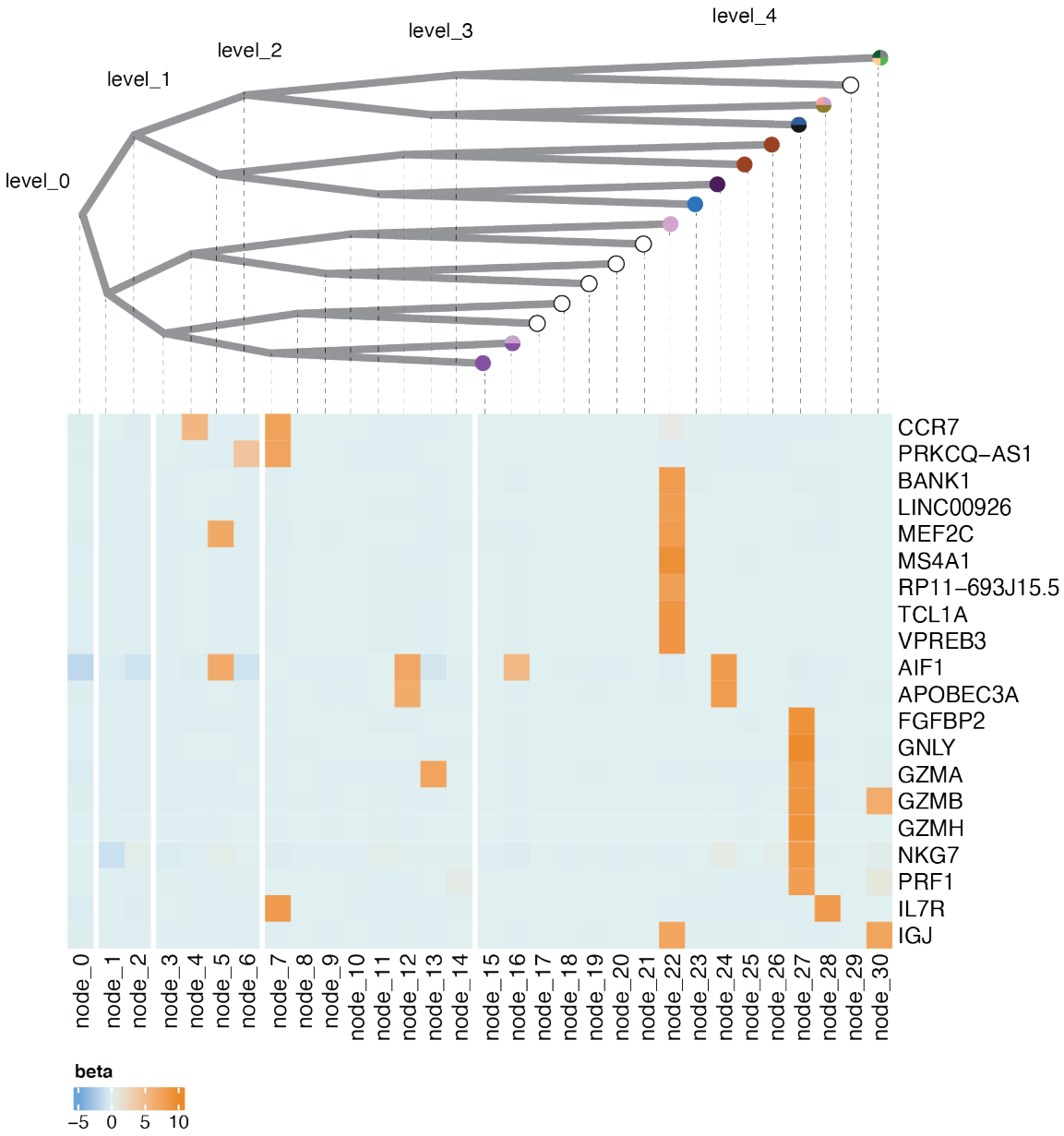

Figure 3.11: **Node-level gene embeddings reveal potential genes of interest with regards to SLE pathogenesis** Heatmap shows estimated node-wise gene embedding values $\hat{\beta}$ of genes with relevant immune function to SLE disease mechanisms.

Patients with SLE show elevated representation in topic 9. The corresponding nodes of this topic, nodes 12 and 24 contain AIF1 and APOBEC3A as significant genes. AIF1 has been shown to be linked with the activation of macrophages and is implicated in several inflammatory diseases [82], making it a viable candidate for exploration of macrophage related mechanisms driving SLE. APOBEC3A is a part of the APOBEC3 family of cytidine deaminase. APOBEC3A gene expression is induced by type I interferon (IFN-I) response during viral infections and plays an important role in defence against viral infection [83]. Patients with SLE have been shown to have elevated expression of APOBEC3A compared to healthy controls, especially in those with flares [84] and there is justification for further exploration into the role it plays in autoimmune disease [85].

Topics 0 and 1 latent values are significantly higher in healthy individuals than patients with SLE 3.7. A shared node in the paths of these topics is node 7 which contains the marker genes CCR7 and PRKCQ-AS1. CCR7 is a chemokine receptor that is critical to the function of naïve T cells [86]. This suggests an expected reduction of naïve CD4+ T cells in patients with SLE. PRKCQ-AS1 is a long non-coding RNA (lncRNA) that has recently been associated with the regulation of the expression of IL-1β, IL-6, and IL-8 in fibroblasts with TNF-α stimulation [87], as well as the regulation of glycolysis and mitochondrial functions in thyroid carcinoma [88]. SNPs in this gene have also been associated with asthma, eczema, and allergic rhinitis [89], making it a viable candidate for exploration in the context of CD4+ T cell driven SLE manifestation.

Chapter 4

Conclusion

4.1 Summary of findings

In this thesis, we present LaRCH, a neural topic model with a built-in node tree structure that enforces hierarchical relationships between latent topics through shared additive embedding features. A spike-and-slab prior informed gene embedding model parameters and provides a natural method of marker gene selection and the ability to perform interpretable latent feature detection. From multiple tests on realistic simulated scRNA-seq datasets, we show that LaRCH is effective at cell type discrimination in heterogeneous samples of immune cell gene expression profiles. By enforcing an underlying topic structure, we allow for an intuitive representation of nested relationships between cell types with shared lineage paths.

Application of LaRCH on a single-cell gene expression dataset across multiple disease phenotypes of systemic lupus erythematosus shows disease dependent hierarchical genomic features. Our findings show cell type specific differences in gene expression in individuals with SLE in addition to overarching cellular features present across all cell types, suggesting disease mediated altered immune function. Investigation into specific node-level marker genes shows the detection of canonical immune subset gene markers and genes with significant roles in immune function. Specifically, we detect differences in the expression of APOBEC3A in monocyte cell populations and PRKQ-AS1 in CD4+ T lymphocyte populations. APOBEC3A is an enzyme encoding gene that has been shown to be upregulated in patients with SLE, while PRKQ-AS1 is a lncRNA that has recently been associated with aberrant immune function and may play a key role in immunological disorders. Both genes show potential impact on the disease mechanisms of SLE and other autoimmunological disorders and are promising avenues for further investigation.

4.2 Limitations

The use cases of the LaRCH model are limited to samples of heterogeneous cell populations with highly variable gene features. Granularity of cell type detection proves to be a challenge and cell subsets with large overlapping gene expression profiles are difficult to discern from one another. Training a model with a deeper latent tree did not yield better results in this aspect. This means that the application of this model would not be suitable for several uses cases such as constructing cancer phylogenies or the detection of rare cell subsets, where detection of minute differences is necessary.

Using a fixed perfect binary tree structure may lead to over parameterization of the model where many components of the tree are unused, resulting in an ill-fitting tree structure to represent the data. Unfortunately, including a tree-fitting component to the model would greatly increase computational burden as the search space would be super-exponential. *Post-hoc* tree pruning could help solve this problem, but would result in the loss of information in the model. In LaRCH, we combat over-fitting at the embedding parameter level by implementing Bayesian priors that encourage model sparsity.

Training of the LaRCH model requires intensive computational resources. LaRCH is implemented in Python using a machine-learning library (`PyTorch` [58]), which often has specialized requirements for hardware, such as graphics processing units (GPUs), with sufficient processing capabilities and memory capacity. All model training experiments presented in this thesis were executed on a Nvidia Linux X64 Display Driver GPU. Smaller scRNA-seq datasets of 2,000 cells across 50,000 genes requires on average 12 minutes to complete model training with a tree depth of 5 in 1,000 epochs. However, on a large gene expression dataset of 1.2 million cells across 30,000 genes, training required over 60 hours to complete.

4.3 Future directions

The LaRCH model may be readily expanded to include other HTS data types. Integration of ChIP-seq data of transcription factor binding affinity or ATAC-seq data of DNA accessibility profiles would be natural additions to this model. Multi-omic integration could provide valuable insights into the gene mechanistic drivers of cell type differentiation.

To understand overall trends in up and down-regulated gene sets, one could relax element-

wise sparsity in facour of set-wise sparsity and fine-mapping of genes, building upon the Sum of Single Effects (SuSiE) method of variable selection outlined in Wang *et al.* [90]. Such a formulation would likely result in a more scalable iterative coordinate-wise algorithm, while assigning high probability mass on causal/anchor features for each independent effect. In fact, implementing a tree-structured latent feature space is also a promising extension to less computationally intensive factorization-based learning algorithms such as NMF. Such algorithms are able to yield low-dimensional embeddings of genomic data, maintaining many use cases of a neural topic model, in an more computationally efficient manner.

Finally, using the results of the LaRCH model, a number of validation experiments can be performed. From the genes described in Chapter 3.5.2, genetic perturbation experiments on healthy and disease models to uncover the downstream effects of differential expression of these genes. From these experiments, there is a potential for bettering our understanding of mechanisms of immune function involved in driving the pathogenesis of autoimmune disorders including SLE.

References

[1] R. K. Perez, M. G. Gordon, M. Subramaniam, M. C. Kim, G. C. Hartoularos, S. Targ, Y. Sun, A. Ogorodnikov, R. Bueno, A. Lu, M. Thompson, N. Rappoport, A. Dahl, C. M. Lanata, M. Matloubian, L. Maliskova, S. S. Kwek, T. Li, M. Slyper, J. Waldman, D. Dionne, O. Rozenblatt-Rosen, L. Fong, M. Dall'Era, B. Balliu, A. Regev, J. Yazdany, L. A. Criswell, N. Zaitlen, and C. J. Ye, "Single-cell RNA-seq reveals cell type-specific molecular and genetic associations to lupus," *Science*, vol. 376, p. eabf1970, Apr. 2022.

[2] D. D. Chaplin, "Overview of the immune response," *J. Allergy Clin. Immunol.*, vol. 125, pp. S3–23, Feb. 2010.

[3] C. A. Janeway, P. Travers, M. Walport, and M. Shlomchik, "Principles of innate and adaptive immunity," in *Immunobiology: The Immune System in Health and Disease. 5th edition* (P. Austin, E. Lawrence, S. Gibbs, E. Hunt, and M. Morales, eds.), New York: Garland Science, 2001.

[4] A. Torang, P. Gupta, and D. Klinke, "An elastic-net logistic regression approach to generate classifiers and gene signatures for types of immune cells and t helper cell subsets," *BMC Bioinformatics*, vol. 20, Aug. 2019.

[5] P. Fang, X. Li, J. Dai, L. Cole, J. A. Camacho, Y. Zhang, J. Yong, J. Wang, X.-F. Yang, and H. Wang, "Immune cell subset differentiation and tissue inflammation," *Journal of Hematology & Oncology*, vol. 11, July 2018.

[6] E. Schönherr and H.-J. Hausser, "Extracellular matrix and cytokines: A functional unit," *Developmental Immunology*, vol. 7, no. 2–4, pp. 89–101, 2000.

[7] S. J. Turner, J. Li, and B. E. Russ, "Chapter 5 - epigenetics mechanisms driving immune memory cell differentiation and function," in *Epigenetics of the Immune System*

(D. Kabelitz and J. Bhat, eds.), vol. 16 of *Translational Epigenetics*, pp. 117–137, Academic Press, 2020.

[8] E. L. Pearce, "Metabolism as a driver of immunity," *Nature Reviews Immunology*, vol. 21, pp. 618–619, Sept. 2021.

[9] N. M. Chapman, S. Shrestha, and H. Chi, *Metabolism in Immune Cell Differentiation and Function*, pp. 1–85. Springer Netherlands, 2017.

[10] C. Liu, K. Omilusik, C. Toma, N. S. Kurd, J. T. Chang, A. W. Goldrath, and W. Wang, "Systems-level identification of key transcription factors in immune cell specification," *PLOS Computational Biology*, vol. 18, p. e1010116, Sept. 2022.

[11] R. ter Horst, M. Jaeger, S. P. Smeekens, M. Oosting, M. A. Swertz, Y. Li, V. Kumar, D. A. Diavatopoulos, A. F. Jansen, H. Lemmers, H. Toenhake-Dijkstra, A. E. van Herwaarden, M. Janssen, R. G. van der Molen, I. Joosten, F. C. Sweep, J. W. Smit, R. T. Netea-Maier, M. M. Koenders, R. J. Xavier, J. W. van der Meer, C. A. Dinarello, N. Pavelka, C. Wijmenga, R. A. Notebaart, L. A. Joosten, and M. G. Netea, "Host and environmental factors influencing individual human cytokine responses," *Cell*, vol. 167, pp. 1111–1124.e13, Nov. 2016.

[12] W. Zhang, J. Ferguson, S. M. Ng, K. Hui, G. Goh, A. Lin, E. Esplugues, R. A. Flavell, C. Abraham, H. Zhao, and J. H. Cho, "Effector cd4+ t cell expression signatures and immune-mediated disease associated genes," *PLoS ONE*, vol. 7, p. e38510, June 2012.

[13] A. Y. Rudensky, "Regulatory t cells and foxp3," *Immunological Reviews*, vol. 241, pp. 260–268, Apr. 2011.

[14] I. I. Ivanov, B. S. McKenzie, L. Zhou, C. E. Tadokoro, A. Lepelley, J. J. Lafaille, D. J. Cua, and D. R. Littman, "The orphan nuclear receptor rorγt directs the differentiation program of proinflammatory il-17+ t helper cells," *Cell*, vol. 126, pp. 1121–1133, Sept. 2006.

[15] S. F. Ziegler and J. H. Buckner, "Foxp3 and the regulation of treg/th17 differentiation," *Microbes and Infection*, vol. 11, pp. 594–598, Apr. 2009.

[16] J. Geginat, M. Paroni, S. Maglie, J. S. Alfen, I. Kastirr, P. Gruarin, M. De Simone, M. Pagani, and S. Abrignani, "Plasticity of human cd4 t cell subsets," *Frontiers in Immunology*, vol. 5, Dec. 2014.

[17] C. S. Field, F. Baixauli, R. L. Kyle, D. J. Puleston, A. M. Cameron, D. E. Sanin, K. L. Hippen, M. Loschi, G. Thangavelu, M. Corrado, J. Edwards-Hicks, K. M. Grzes, E. J. Pearce, B. R. Blazar, and E. L. Pearce, "Mitochondrial integrity regulated by lipid metabolism is a Cell-Intrinsic checkpoint for treg suppressive function," *Cell Metab.*, vol. 31, pp. 422–437.e5, Feb. 2020.

[18] S. Yang, C. Xie, Y. Chen, J. Wang, X. Chen, Z. Lu, R. R. June, and S. G. Zheng, "Differential roles of TNFα-TNFR1 and TNFα-TNFR2 in the differentiation and function of CD4+Foxp3+ induced treg cells in vitro and in vivo periphery in autoimmune diseases," *Cell Death Dis.*, vol. 10, p. 27, Jan. 2019.

[19] H. Xiao, J. Jiao, L. Wang, S. O'Brien, K. Newick, L.-C. S. Wang, E. Falkensammer, Y. Liu, R. Han, V. Kapoor, F. K. Hansen, T. Kurz, W. W. Hancock, and U. H. Beier, "HDAC5 controls the functions of foxp3(+) t-regulatory and CD8(+) T cells," *Int. J. Cancer*, vol. 138, pp. 2477–2486, May 2016.

[20] R. M. Thomas, H. Sai, and A. D. Wells, "Conserved intergenic elements and DNA methylation cooperate to regulate transcription at the il17 locus," *J. Biol. Chem.*, vol. 287, pp. 25049–25059, July 2012.

[21] Y. Jiang, Y. Liu, H. Lu, S.-C. Sun, W. Jin, X. Wang, and C. Dong, "Epigenetic activation during T helper 17 cell differentiation is mediated by tripartite motif containing 28," *Nat. Commun.*, vol. 9, p. 1424, Apr. 2018.

[22] D. Pham, S. Sehra, X. Sun, and M. H. Kaplan, "The transcription factor etv5 controls T_H17 cell development and allergic airway inflammation," *J. Allergy Clin. Immunol.*, vol. 134, pp. 204–214, July 2014.

[23] A. Wagner, C. Wang, J. Fessler, D. DeTomaso, J. Avila-Pacheco, J. Kaminski, S. Zaghouani, E. Christian, P. Thakore, B. Schellhaass, E. Akama-Garren, K. Pierce, V. Singh, N. Ron-Harel, V. P. Douglas, L. Bod, A. Schnell, D. Puleston, R. A. Sobel, M. Haigis, E. L. Pearce, M. Soleimani, C. Clish, A. Regev, V. K. Kuchroo, and N. Yosef, "Metabolic modeling of single th17 cells reveals regulators of autoimmunity," *Cell*, vol. 184, pp. 4168–4185.e21, Aug. 2021.

[24] D. J. Puleston, F. Baixauli, D. E. Sanin, J. Edwards-Hicks, M. Villa, A. M. Kabat, M. M. Kamiński, M. Stanckzak, H. J. Weiss, K. M. Grzes, K. Piletic, C. S. Field,

M. Corrado, F. Haessler, C. Wang, Y. Musa, L. Schimmelpfennig, L. Flachsmann, G. Mittler, N. Yosef, V. K. Kuchroo, J. M. Buescher, S. Balabanov, E. J. Pearce, D. R. Green, and E. L. Pearce, "Polyamine metabolism is a central determinant of helper T cell lineage fidelity," *Cell*, vol. 184, pp. 4186–4202.e20, Aug. 2021.

[25] X. Chen, W. Su, T. Wan, J. Yu, W. Zhu, F. Tang, G. Liu, N. Olsen, D. Liang, and S. G. Zheng, "Sodium butyrate regulates Th17/Treg cell balance to ameliorate uveitis via the Nrf2/HO-1 pathway," *Biochem. Pharmacol.*, vol. 142, pp. 111–119, Oct. 2017.

[26] J. T. Gaublomme, N. Yosef, Y. Lee, R. S. Gertner, L. V. Yang, C. Wu, P. P. Pandolfi, T. Mak, R. Satija, A. K. Shalek, V. K. Kuchroo, H. Park, and A. Regev, "Single-Cell genomics unveils critical regulators of th17 cell pathogenicity," *Cell*, vol. 163, pp. 1400–1412, Dec. 2015.

[27] G. C. Tsokos, "Systemic lupus erythematosus," *N. Engl. J. Med.*, vol. 365, pp. 2110–2121, Dec. 2011.

[28] J. Tian, D. Zhang, X. Yao, Y. Huang, and Q. Lu, "Global epidemiology of systemic lupus erythematosus: a comprehensive systematic analysis and modelling study," *Annals of the Rheumatic Diseases*, vol. 82, pp. 351–356, Oct. 2022.

[29] C. Adamichou and G. Bertsias, "Flares in systemic lupus erythematosus: diagnosis, risk factors and preventive strategies," *Mediterranean Journal of Rheumatology*, vol. 28, pp. 4–12, Jan. 2017.

[30] D. Furst, A. Clarke, A. Fernandes, T. Bancroft, W. Greth, and S. Iorga, "Incidence and prevalence of adult systemic lupus erythematosus in a large us managed-care population," *Lupus*, vol. 22, pp. 99–105, Oct. 2012.

[31] J. S. Nusbaum, I. Mirza, J. Shum, R. W. Freilich, R. E. Cohen, M. H. Pillinger, P. M. Izmirly, and J. P. Buyon, "Sex differences in systemic lupus erythematosus," *Mayo Clinic Proceedings*, vol. 95, pp. 384–394, Feb. 2020.

[32] J. M. P. Woo, C. G. Parks, S. Jacobsen, K. H. Costenbader, and S. Bernatsky, "The role of environmental exposures and gene–environment interactions in the etiology of systemic lupus erythematous," *Journal of Internal Medicine*, vol. 291, pp. 755–778, Feb. 2022.

[33] C.-F. Kuo, M. J. Grainge, A. M. Valdes, L.-C. See, S.-F. Luo, K.-H. Yu, W. Zhang, and M. Doherty, "Familial aggregation of systemic lupus erythematosus and coaggregation of autoimmune diseases in affected families," *JAMA Internal Medicine*, vol. 175, p. 1518, Sept. 2015.

[34] F. Ceccarelli, C. Perricone, P. Borgiani, C. Ciccacci, S. Rufini, E. Cipriano, C. Alessandri, F. R. Spinelli, A. Sili Scavalli, G. Novelli, G. Valesini, and F. Conti, "Genetic factors in systemic lupus erythematosus: Contribution to disease phenotype," *Journal of Immunology Research*, vol. 2015, pp. 1–11, 2015.

[35] M. Teruel and M. E. Alarcón-Riquelme, "The genetic basis of systemic lupus erythematosus: What are the risk factors and what have we learned," *Journal of Autoimmunity*, vol. 74, pp. 161–175, 2016.

[36] V. R. Moulton and G. C. Tsokos, "T cell signaling abnormalities contribute to aberrant immune cell function and autoimmunity," *Journal of Clinical Investigation*, vol. 125, pp. 2220–2227, May 2015.

[37] R. M. Talaat, S. F. Mohamed, I. H. Bassyouni, and A. A. Raouf, "Th1/th2/th17/treg cytokine imbalance in systemic lupus erythematosus (sle) patients: Correlation with disease activity," *Cytokine*, vol. 72, pp. 146–153, Apr. 2015.

[38] A. Suárez-Fueyo, D. F. Barber, J. Martínez-Ara, A. C. Zea-Mendoza, and A. C. Carrera, "Enhanced phosphoinositide 3-kinase δ activity is a frequent event in systemic lupus erythematosus that confers resistance to activation-induced t cell death," *The Journal of Immunology*, vol. 187, pp. 2376–2385, Sept. 2011.

[39] D. M. Gravano and K. K. Hoyer, "Promotion and prevention of autoimmune disease by cd8+ t cells," *Journal of Autoimmunity*, vol. 45, pp. 68–79, Sept. 2013.

[40] M. D. Luecken and F. J. Theis, "Current best practices in single-cell rna-seq analysis: a tutorial," *Molecular Systems Biology*, vol. 15, June 2019.

[41] V. Y. Kiselev, T. S. Andrews, and M. Hemberg, "Challenges in unsupervised clustering of single-cell rna-seq data," *Nature Reviews Genetics*, vol. 20, pp. 273–282, Jan. 2019.

[42] I. Korsunsky, N. Millard, J. Fan, K. Slowikowski, F. Zhang, K. Wei, Y. Baglaenko, M. Brenner, P.-r. Loh, and S. Raychaudhuri, "Fast, sensitive and accurate integration

of single-cell data with harmony," *Nature methods*, vol. 16, no. 12, pp. 1289–1296, 2019.

[43] J. Liu, C. Gao, J. Sodicoff, V. Kozareva, E. Z. Macosko, and J. D. Welch, "Jointly defining cell types from multiple single-cell datasets using liger," *Nature protocols*, vol. 15, no. 11, pp. 3632–3662, 2020.

[44] M. Tschannen, O. Bachem, and M. Lucic, "Recent advances in autoencoder-based representation learning," *arXiv preprint arXiv:1812.05069*, 2018.

[45] D. P. Kingma and M. Welling, "Auto-Encoding variational bayes," Dec. 2013.

[46] R. Lopez, J. Regier, M. B. Cole, M. I. Jordan, and N. Yosef, "Deep generative modeling for single-cell transcriptomics," *Nat. Methods*, vol. 15, pp. 1053–1058, Dec. 2018.

[47] C. H. Grønbech, M. F. Vording, P. N. Timshel, C. K. Sønderby, T. H. Pers, and O. Winther, "scvae: variational auto-encoders for single-cell gene expression data," *Bioinformatics*, vol. 36, no. 16, pp. 4415–4422, 2020.

[48] Y. Zhao, H. Cai, Z. Zhang, J. Tang, and Y. Li, "Learning interpretable cellular and gene signature embeddings from single-cell transcriptomic data," *Nat. Commun.*, vol. 12, pp. 1–15, Sept. 2021.

[49] Y. Zhang, M. S. Khalilitousi, and Y. P. Park, "Unraveling dynamically encoded latent transcriptomic patterns in pancreatic cancer cells by topic modeling," *Cell Genomics*, vol. 3, no. 9, p. 100388, 2023.

[50] A. B. Dieng, F. J. R. Ruiz, and D. M. Blei, "Topic modeling in embedding spaces," *Trans. Assoc. Comput. Linguist.*, vol. 8, pp. 439–453, Dec. 2020.

[51] S. Domcke and J. Shendure, "A reference cell tree will serve science better than a reference cell atlas," *Cell*, vol. 186, pp. 1103–1114, Mar. 2023.

[52] T. Griffiths, M. Jordan, J. Tenenbaum, and D. Blei, "Hierarchical topic models and the nested chinese restaurant process," *Adv. Neural Inf. Process. Syst.*, vol. 16, 2003.

[53] M. Isonuma, J. Mori, D. Bollegala, and I. Sakata, "Tree-Structured Neural Topic Model," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, eds.), (Online), pp. 800–806, Association for Computational Linguistics, July 2020.

[54] L. Michielsen, M. Lotfollahi, D. Strobl, L. Sikkema, M. J. T. Reinders, F. J. Theis, and A. Mahfouz, "Single-cell reference mapping to construct and extend cell-type hierarchies," *NAR Genom Bioinform*, vol. 5, p. lqad070, July 2023.

[55] D. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv. org*, Dec. 2014.

[56] D. P. Kingma, T. Salimans, and M. Welling, "Variational dropout and the local reparameterization trick," in *Advances in Neural Information Processing Systems 28* (C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, eds.), pp. 2575–2583, Curran Associates, Inc., 2015.

[57] T. J. Mitchell and J. J. Beauchamp, "Bayesian variable selection in linear regression," *J. Am. Stat. Assoc.*, vol. 83, pp. 1023–1032, Dec. 1988.

[58] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, and Others, "Pytorch: An imperative style, high-performance deep learning library," *Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[59] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch." Oct. 2017.

[60] B. J. Schmiedel, D. Singh, A. Madrigal, A. G. Valdovino-Gonzalez, B. M. White, J. Zapardiel-Gonzalo, B. Ha, G. Altay, J. A. Greenbaum, G. McVicker, G. Seumois, A. Rao, M. Kronenberg, B. Peters, and P. Vijayanand, "Impact of genetic polymorphisms on human immune cell gene expression," *Cell*, vol. 175, pp. 1701–1715.e16, Nov. 2018.

[61] C. Hafemeister and R. Satija, "Normalization and variance stabilization of single-cell rna-seq data using regularized negative binomial regression," *Genome Biology*, vol. 20, Dec. 2019.

[62] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Stat. Mech.*, vol. 2008, p. P10008, Oct. 2008.

[63] G. Korotkevich, V. Sukhov, N. Budin, B. Shpak, M. N. Artyomov, and A. Sergushichev, "Fast gene set enrichment analysis." Feb. 2021.

[64] E. Sollis, A. Mosaku, A. Abid, A. Buniello, M. Cerezo, L. Gil, T. Groza, O. Güneş, P. Hall, J. Hayhurst, A. Ibrahim, Y. Ji, S. John, E. Lewis, J. A. L. MacArthur, A. McMahon, D. Osumi-Sutherland, K. Panoutsopoulou, Z. Pendlington, S. Ramachandran, R. Stefancsik, J. Stewart, P. Whetzel, R. Wilson, L. Hindorff, F. Cunningham, S. A. Lambert, M. Inouye, H. Parkinson, and L. W. Harris, "The NHGRI-EBI GWAS catalog: knowledgebase and deposition resource," *Nucleic Acids Res.*, vol. 51, pp. D977–D985, Jan. 2023.

[65] J. Godec, Y. Tan, A. Liberzon, P. Tamayo, S. Bhattacharya, A. J. Butte, J. P. Mesirov, and W. N. Haining, "Compendium of immune signatures identifies conserved and species-specific biology in response to inflammation," *Immunity*, vol. 44, no. 1, pp. 194–206, 2016.

[66] A. Liberzon, C. Birger, H. Thorvaldsdóttir, M. Ghandi, J. P. Mesirov, and P. Tamayo, "The molecular signatures database hallmark gene set collection," *Cell systems*, vol. 1, no. 6, pp. 417–425, 2015.

[67] S. Karrar and D. S. Cunningham Graham, "Abnormal b cell development in systemic lupus erythematosus," *Arthritis Rheumatol.*, vol. 70, pp. 496–507, Apr. 2018.

[68] J. van Langelaar, L. Rijvers, J. Smolders, and M. M. van Luijn, "B and t cells driving multiple sclerosis: Identity, mechanisms and potential triggers," *Front. Immunol.*, vol. 11, p. 760, May 2020.

[69] S. V. Gearty, F. Dündar, P. Zumbo, G. Espinosa-Carrasco, M. Shakiba, F. J. Sanchez-Rivera, N. D. Socci, P. Trivedi, S. W. Lowe, P. Lauer, N. Mohibullah, A. Viale, T. P. DiLorenzo, D. Betel, and A. Schietinger, "An autoimmune stem-like cd8 t cell population drives type 1 diabetes," *Nature*, vol. 602, pp. 156–161, Feb. 2022.

[70] M. Stephens, "False discovery rates: a new deal," *Biostatistics*, vol. 18, pp. 275–294, Apr. 2017.

[71] M. A. Alonso and S. M. Weissman, "cdna cloning and sequence of mal, a hydrophobic protein associated with human t-cell differentiation.," *Proceedings of the National Academy of Sciences*, vol. 84, no. 7, pp. 1997–2001, 1987.

[72] D. Cibrián and F. Sánchez-Madrid, "Cd69: from activation marker to metabolic gatekeeper," *European journal of immunology*, vol. 47, no. 6, pp. 946–953, 2017.

[73] B. Cózar, M. Greppi, S. Carpentier, E. Narni-Mancinelli, L. Chiossone, and E. Vivier, "Tumor-infiltrating natural killer cells," *Cancer discovery*, vol. 11, no. 1, pp. 34–44, 2021.

[74] J. Moyron-Quiroz, S. Partida-Sánchez, R. Donís-Hernández, C. Sandoval-Montes, and L. Santos-Argumedo, "Expression and function of cd22, a b-cell restricted molecule," *Scandinavian journal of immunology*, vol. 55, no. 4, pp. 343–351, 2002.

[75] U. M. Nagarajan, J. Lochamy, X. Chen, G. W. Beresford, R. Nilsen, P. E. Jensen, and J. M. Boss, "Class ii transactivator is required for maximal expression of hla-dob in b cells," *The Journal of Immunology*, vol. 168, no. 4, pp. 1780–1786, 2002.

[76] P. G. Chu and D. A. Arber, "Cd79: a review," *Applied Immunohistochemistry & Molecular Morphology*, vol. 9, no. 2, pp. 97–106, 2001.

[77] M. M. Desouki, G. R. Post, D. Cherry, and J. Lazarchick, "Pax-5: a valuable immunohistochemical marker in the differential diagnosis of lymphoid neoplasms," *Clinical medicine & research*, vol. 8, no. 2, pp. 84–88, 2010.

[78] P. J. van den Elsen, T. M. Holling, H. F. Kuipers, and N. van der Stoep, "Transcriptional regulation of antigen presentation," *Current Opinion in Immunology*, vol. 16, no. 1, pp. 67–75, 2004.

[79] S. Goyert, E. Ferrero, S. Seremetis, R. Winchester, J. Silver, and A. Mattison, "Biochemistry and expression of myelomonocytic antigens.," *Journal of immunology (Baltimore, Md.: 1950)*, vol. 137, no. 12, pp. 3909–3914, 1986.

[80] P. Blanco, V. Pitard, J.-F. Viallard, J.-L. Taupin, J.-L. Pellegrin, and J.-F. Moreau, "Increase in activated cd8+ t lymphocytes expressing perforin and granzyme b correlates with disease activity in patients with systemic lupus erythematosus," *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology*, vol. 52, no. 1, pp. 201–211, 2005.

[81] E. M. Dam, T. Habib, J. Chen, A. Funk, V. Glukhova, M. Davis-Pickett, S. Wei, R. James, J. H. Buckner, and K. Cerosaletti, "The bank1 sle-risk variants are associated with alterations in peripheral b cell signaling and development in humans," *Clinical immunology*, vol. 173, pp. 171–180, 2016.

[82] D. De Leon-Oliva, C. Garcia-Montero, O. Fraile-Martinez, D. L. Boaru, L. García-Puente, A. Rios-Parra, M. J. Garrido-Gil, C. Casanova-Martín, N. García-Honduvilla, J. Bujan, *et al.*, "Aif1: function and connection with inflammatory diseases," *Biology*, vol. 12, no. 5, p. 694, 2023.

[83] M. Taura, J. A. Frank, T. Takahashi, Y. Kong, E. Kudo, E. Song, M. Tokuyama, and A. Iwasaki, "Apobec3a regulates transcription from interferon-stimulated response elements," *Proceedings of the National Academy of Sciences*, vol. 119, no. 20, p. e2011665119, 2022.

[84] D. Perez-Bercoff, H. Laude, M. Lemaire, O. Hunewald, V. Thiers, M. Vignuzzi, H. Blanc, A. Poli, Z. Amoura, V. Caval, *et al.*, "Sustained high expression of multiple apobec3 cytidine deaminases in systemic lupus erythematosus," *Scientific Reports*, vol. 11, no. 1, p. 7893, 2021.

[85] Y.-T. Liu and X.-Y. Meng, "The underexplored role of apobec3 enzymes in autoimmune diseases," *Rheumatology*, p. kead663, 2023.

[86] M. R. Britschgi, A. Link, T. K. A. Lissandrin, and S. A. Luther, "Dynamic modulation of ccr7 expression and function on naive t lymphocytes in vivo," *The Journal of Immunology*, vol. 181, no. 11, pp. 7681–7688, 2008.

[87] Q. Zhao, J. Liu, X. Ouyang, W. Liu, P. Lv, S. Zhang, and J. Zhong, "Role of immune-related lncrnas–prkcq-as1 and egot in the regulation of il-1β, il-6 and il-8 expression in human gingival fibroblasts with tnf-α stimulation," *Journal of Dental Sciences*, vol. 18, no. 1, pp. 184–190, 2023.

[88] X. Zhang, Y. Zhong, L. Liu, C. Jia, H. Cai, J. Yang, B. Wu, and Z. Lv, "Fasting regulates mitochondrial function through lncrna prkcq-as1-mediated igf2bps in papillary thyroid carcinoma," *Cell Death & Disease*, vol. 14, no. 12, p. 827, 2023.

[89] Y. Shi, Y. Niu, P. Zhang, H. Luo, S. Liu, S. Zhang, J. Wang, Y. Li, X. Liu, T. Song, *et al.*, "Characterization of genome-wide str variation in 6487 human genomes," *Nature Communications*, vol. 14, no. 1, p. 2092, 2023.

[90] G. Wang, A. Sarkar, P. Carbonetto, and M. Stephens, "A simple new approach to variable selection in regression, with application to genetic fine mapping," *Journal of*

the Royal Statistical Society Series B: Statistical Methodology, vol. 82, pp. 1273–1300, Dec. 2020.

[91] W. Bailis, J. A. Shyer, J. Zhao, J. C. G. Canaveras, F. J. Al Khazal, R. Qu, H. R. Steach, P. Bielecki, O. Khan, R. Jackson, Y. Kluger, L. J. Maher, 3rd, J. Rabinowitz, J. Craft, and R. A. Flavell, "Distinct modes of mitochondrial metabolism uncouple T cell differentiation and function," *Nature*, vol. 571, pp. 403–407, July 2019.

[92] M. A. Sallin, S. Sakai, K. D. Kauffman, H. A. Young, J. Zhu, and D. L. Barber, "Th1 differentiation drives the accumulation of intravascular, non-protective CD4 T cells during tuberculosis," *Cell Rep.*, vol. 18, pp. 3091–3104, Mar. 2017.

[93] M. Angela, Y. Endo, H. K. Asou, T. Yamamoto, D. J. Tumes, H. Tokuyama, K. Yokote, and T. Nakayama, "Fatty acid metabolic reprogramming via mTOR-mediated inductions of PPARγ directs early activation of T cells," *Nat. Commun.*, vol. 7, p. 13683, Nov. 2016.

Appendix A

Supplementary tables and figures

Th1						
Study	α CD3	α CD28	IL-2	IL-12	TNF-α	α IL-4
Field 2020 [17]	5 μg/mL	0.5 μg/mL	100 U/mL	10 ng/mL	10 μg/mL	—
Bailis 2019 [91]	beads	beads	5 ng/mL	2 ng/mL	10 μg/mL	—
Sallin 2017 [92]	3 μg/mL	1 μg/mL	0.12-30 ng/mL	0.12-30 ng/mL	10 μg/mL	—
Yang 2019 [18]	1 μg/mL	1 μg/mL	—	10 ng/mL	5 μg/mL	100 ng/mL
Thomas 2012 [20]	1 μg/mL	1 μg/mL	—	10 ng/mL	10 μg/mL	—
Jiang 2018 [21]	5 μg/mL	5 μg/mL	—	10 ng/mL	10 μg/mL	—
Pham 2014 [22]	2 μg/mL	0.5 μg/mL	—	5 ng/mL	10 μg/mL	—
Wagner 2021 [23]	1 μg/mL	1 μg/mL	—	20 ng/mL	—	—
Puleston 2021 [24]	5 μg/mL	2 μg/mL	100 U/mL	10 ng/mL	4 μg/mL	—
Th2						
Study	α CD3	α CD28	IL-2	IL-4	α IL-12	α IFNγ
Field 2020 [17]	5 μg/mL	0.5 μg/mL	100 U/mL	10 ng/mL	10 μg/mL	10 μg/mL
Angela 2016 [93]	—	—	25 U/mL	10 U/mL	—	1 μg/mL
Thomas 2012 [20]	1 μg/mL	1 μg/mL	—	40 ng/mL	10 μg/mL	10 μg/mL
Jiang 2018 [21]	5 μg/mL	5 μg/mL	—	10 ng/mL	—	10 μg/mL
Pham 2014 [22]	2 μg/mL	0.5 μg/mL	—	10 ng/mL	—	10 μg/mL
Wagner 2021 [23]	1 μg/mL	1 μg/mL	—	20 ng/mL	—	—
Puleston 2021 [24]	5 μg/mL	2 μg/mL	100 U/mL	10 ng/mL	—	4 μg/mL

Table A.1: **Table summary of CD4+ polarization conditions** A meta-analysis of various studies with *in vitro* CD4+ subset polarization condition protocols. Included are the polarization conditions for Th1s and Th2s across 10 studies. Units are presented as described in the original studies.

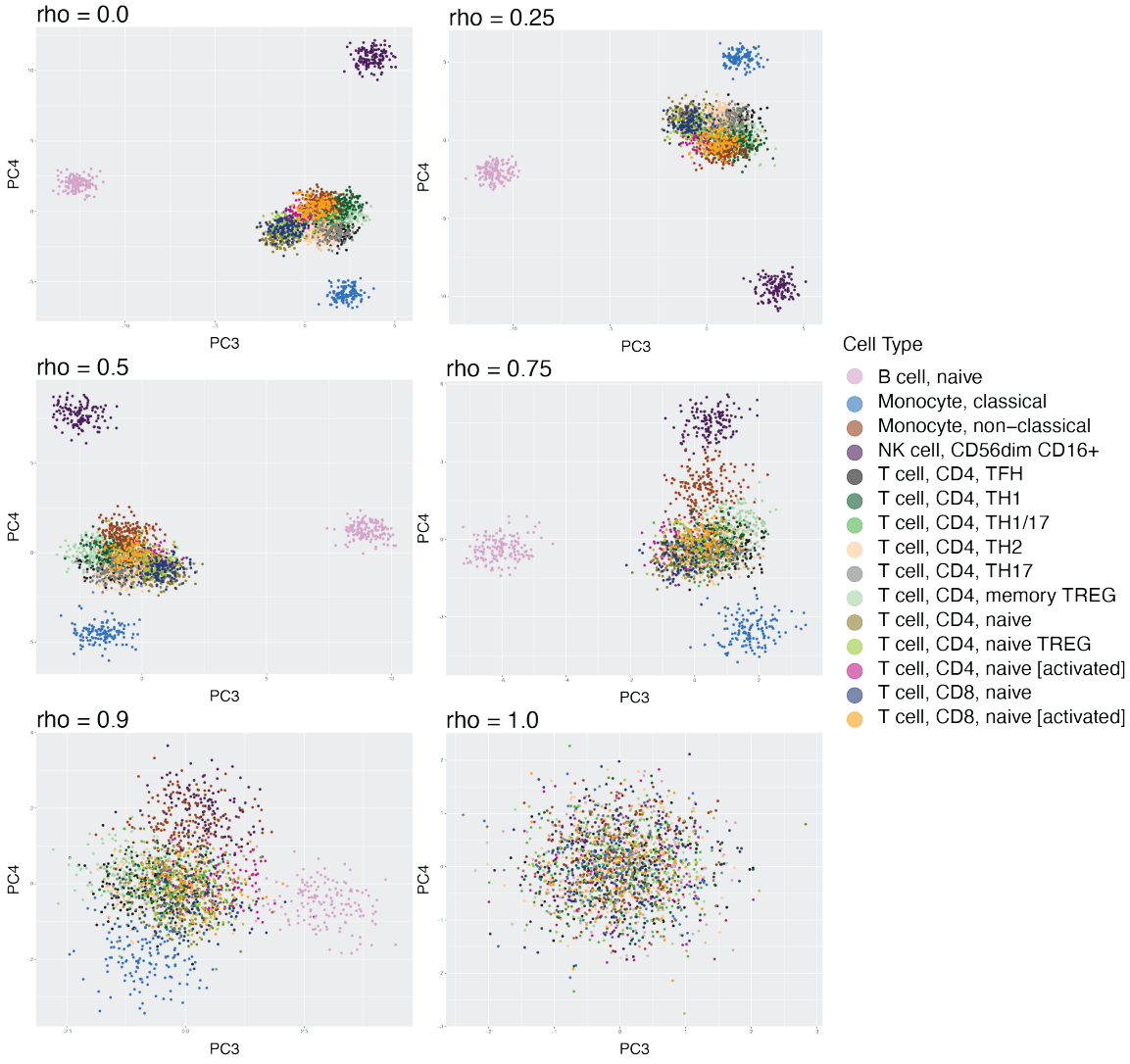

Figure A.1: **Simulated scRNA-seq datasets exhibit realistic distribution of expression patterns across various noise levels** Shown are the plots of third and fourth PCs for generated immune cell scRNA-seq datasets from the simulation scheme. Datasets generated using various noise proportions $\rho = \{0.0, 0.25, 0.5, 0.75, 0.9, 1.0\}$ are shown and coloured by cell type.

Node	No. significantly upregulated genes	No. significantly downregulated genes
Level 0		
node_0	0	157
Level 1		
node_1	525	242
node_2	133	100
Level 2		
node_3	215	71
node_4	198	69
node_5	1705	289
node_6	485	123
Level 3		
node_7	1399	102
node_8	47	6
node_9	227	29
node_10	361	54
node_11	582	54
node_12	1451	160
node_13	677	74
node_14	280	63
Level 4		
node_15	830	16
node_16	823	56
node_17	41	3
node_18	49	2
node_19	1172	93
node_20	334	19
node_21	486	33
node_22	913	81
node_23	808	43
node_24	1185	100
node_25	764	59
node_26	1003	74
node_27	1096	102
node_28	894	44
node_29	286	7
node_30	638	87

Table A.2: **Table outlining the number of significant genes for each latent node** Upregulated genes are identified as genes with an embedding value of $\hat{\beta} > 0$ to a significance level of $\alpha = 0.05$. Inversely, downregulated genes are those with an embedding values of $\hat{\beta} < 0$.

Node	No. unique significantly upregulated genes	No. unique significantly downregulated genes
Level 0		
node_0	0	157
Level 1		
node_1	525	154
node_2	133	40
Level 2		
node_3	188	19
node_4	183	19
node_5	1662	250
node_6	476	78
Level 3		
node_7	1394	88
node_8	38	1
node_9	213	14
node_10	344	17
node_11	292	31
node_12	995	41
node_13	622	26
node_14	249	26
Level 4		
node_15	429	12
node_16	620	8
node_17	37	1
node_18	45	1
node_19	1098	69
node_20	244	12
node_21	400	19
node_22	870	49
node_23	639	34
node_24	1022	61
node_25	214	22
node_26	445	24
node_27	959	49
node_28	687	30
node_29	229	5
node_30	608	44

Table A.3: **Table outlining the number of unique significant genes for each latent node** Upregulated genes are identified as genes with an embedding value of $\hat{\beta} > 0$ to a significance level of $\alpha = 0.05$. Inversely, downregulated genes are those with an embedding values of $\hat{\beta} < 0$. Genes are considered unique to a node if it does not also appear as a significant gene in its parent and sibling node.

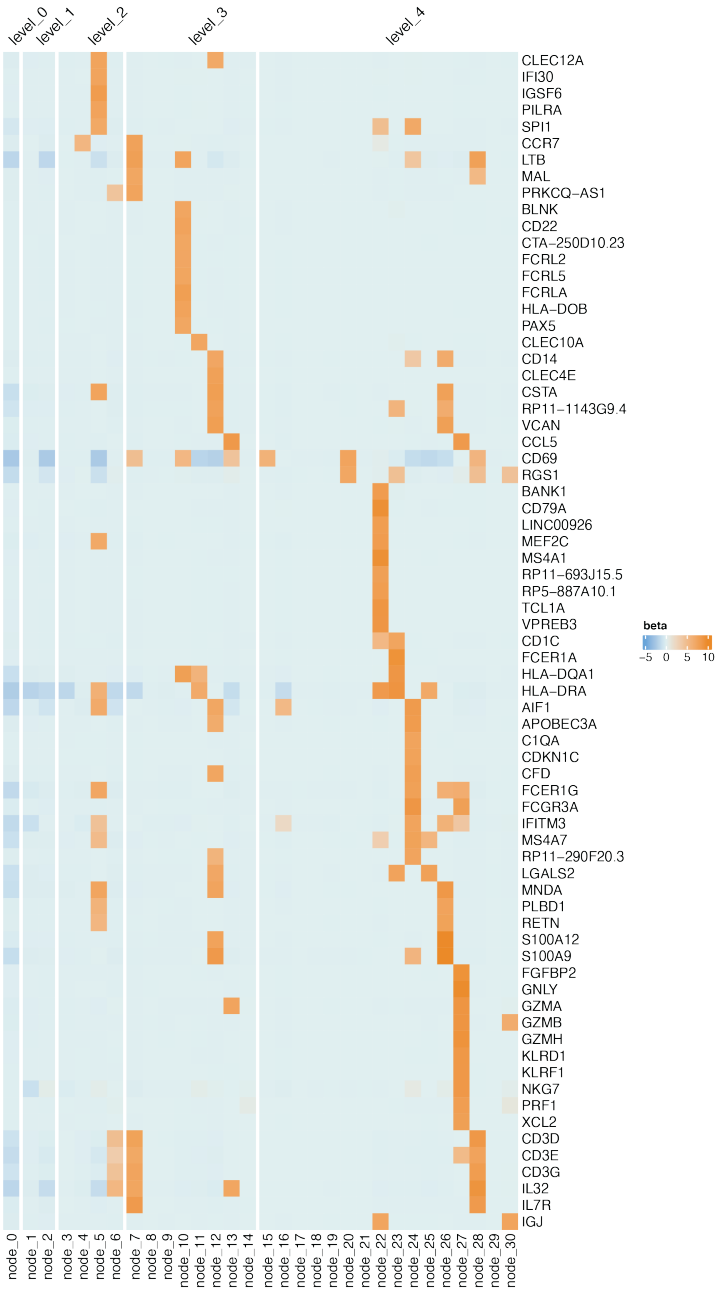

Figure A.2: **Heatmap of top genes shows candidates for further analysis** Heatmap shows estimated gene embedding values $\hat{\beta}$ of top genes across all latent tree nodes. Genes are selected as the 80 gene with the greatest absolute $\hat{\beta}$ values amongst the aggregated pool of the top ten genes per latent tree node.

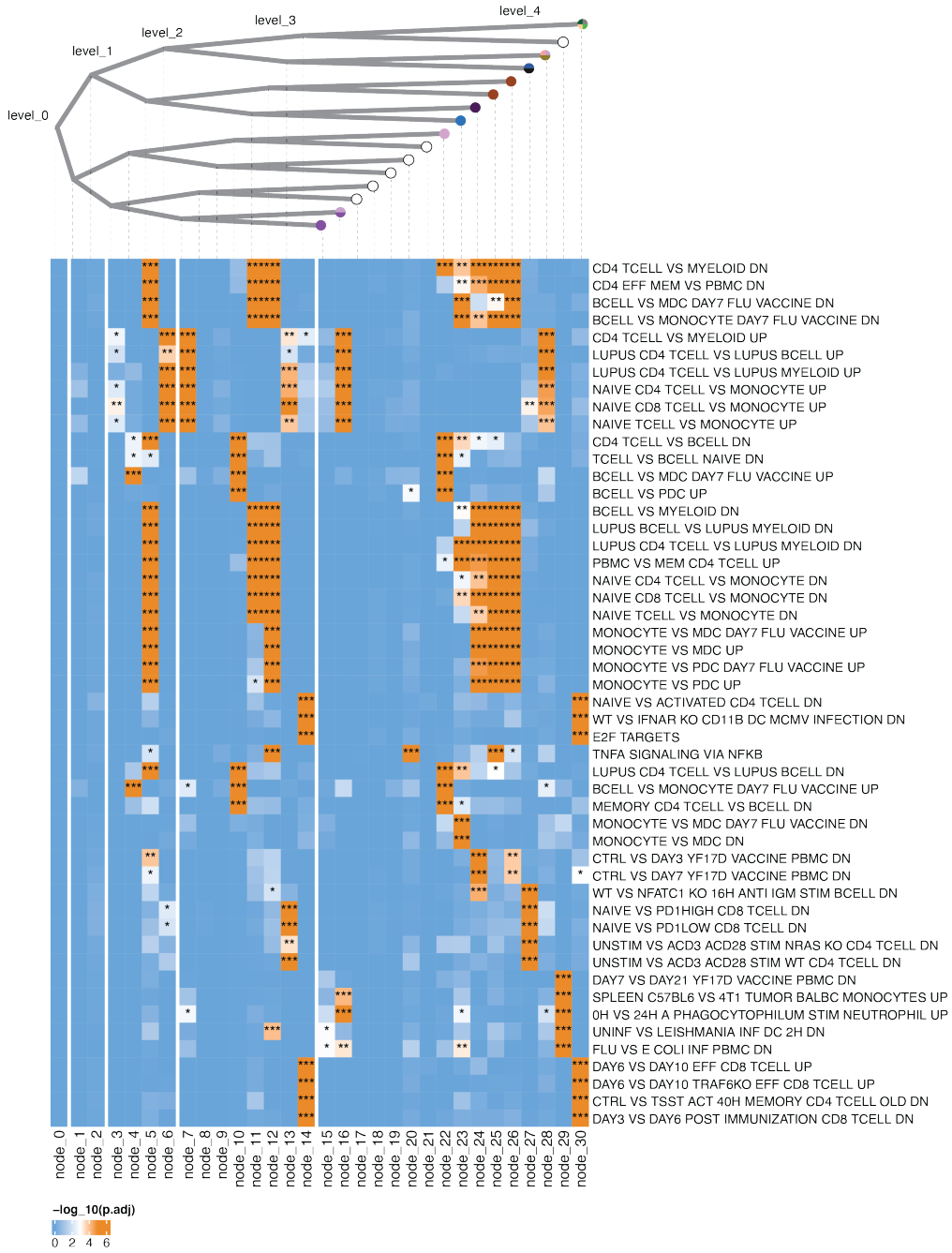

Figure A.3: **Enrichment analysis of gene sets related to immune function show cell function described by latent node embeddings** Heatmap shows the $-\log(p)$ enrichment significance of each gene set per node. Significance annotations are as follows: $***-p<0.0001, **-p<0.001, *-p<0.01$

Appendix B

Materials and methods for *in vitro* experiments

B.1 CD4+ T cell isolation

For *in vitro* cell culture, male wild-type BL6 mice were bred and kept in specific pathogen-free (SPF) conditions at the Centre for Molecular Medicine and Therapeutics (CMMT) mouse facility at BC Children's Hospital Research Institute. All animals were cared for in compliance with the Canadian Council on Animal care and the University of British Columbia Animal Care Committee (Protocol numbers A19-024, A19-0273 and A21-0266).

Spleen and lymph nodes were harvested from mice at 8-14 weeks of age. Following organ harvest, naïve CD4+ T cells were isolated by negative selection using a StemCell Mouse CD4+ T Cell Isolation Kit according to the manufacturer's protocol (StemCell Technologies, Cat: 19852).

B.2 CD4+ T cell subset polarization

All isolated CD4+ T cells were cultured in T cell media (TCM) containing RPMI 1640 media (Corning) supplemented with 10% fetal bovine serum (FBS) (Gibco), 2mM L-Glutamine, 100 U/mL penicillin/streptomycin, 55 μM β-Mercaptoethanol (Life Technologies). Cells were activated using 5μg/mL plate-bound anti-CD3 (BioXCell), 2 μg/mL soluble anti-CD8 (BioXCell) and 10 ng/mL recombinant human (rh) IL-2 (Peprotech) at a concentration of 1.5×10^6 cells/mL.

In addition to the above media, Helper T cell subtypes were polarized as follows:

- T_H0: No additional cytokines or antibodies
- T_H1: 4 μg/mL anti-mouse IL-4 (BioXCell), 10 ng/mL recombinant murine (rm) IL-12
- Regulatory T cell (T_{reg}): 10 ng/mL rh TGF-β1 (Peprotech)
- Non-pathogenic T_H17: 10 ng/mL rmIL-6 (Peprotech), 5 ng/mL TGF-β1

- Pathogenic T_H17 (T_H17p): 10 ng/mL rmIL-6, 5 ng/mL TGF-β1, 10 ng/mL rmIL-1β (Peprotech), 10 ng/mL rmIL-23 (BioLegend)

72-hours after initial activation, cells were split with fresh media containing the same polarization conditions with the exception anti-CD3 and anti-CD28 which were omitted. On day 4, cells were collected for downstream flow cytometry analysis.

B.3 Flow cytometry analysis

Cells collected for flow cytometry analysis are spun, supernatant discarded, and resuspended in TCM supplemented with 10 ng/mL rhIL-2 and 1.5 μg/mL Brefeldin A (eBioscience). A subsample of cells were further restimulated with phorbol 12-myristate (PMA) (25 ng/mL; Sigma) and ionomycin (250 ng/mL; Sigma). Following resuspension, all samples are incubated for 4 hours before staining.

Surface proteins are stained by resuspending cell samples in FACS buffer comprised of Phosphate Buffered Saline (PBS) (Corning), 2% FBS, and surface marker antibody cocktail containing PerCP-Cy5.5 anti-mouse CD69 (1:200, clone: H1.2F3 BioLegend), FITC anti-mouse CD25 (1:200, clone: PC61 BioLegend), eFluor 780 Fixable Viability Dye (1:10000, eBioscience), BV785 anti-mouse CD8α (1:200, clone: 53-6.7, BioLegend), and BV605 anti-mouse CD4 (1:200, clone: RM4-4, BioLegend). Cells are incubated in antibody cocktail for 30 minutes at room temperature in the dark.

To detect intracellular proteins, all samples were fixed and permeabilized using the Foxp3 / Transcription Factor Staining Buffer Set (eBiosciences), followed by intracellular marker antibody cocktail containing Alexa Fluor 700 monoclonal antibody FOXP3 (1:200, clone: FJK-16s, eBiosciences), BV510 anti-mouse RORγt (1:200, clone: Q31-378), BV421 anti-T-bet (1:200, clone: 4B10, BioLegend), APC anti-mouse IL-17A (1:200, clone: TC11-18H10.1, BioLegend), and PE IFN-γ (1:200, clone: XMG1.2, BioLegend). Cells are incubated for 60 minutes at room temperature in the dark. Sample were resuspended in FACS buffer and FACS data was acquired with a BD LSRFortessa X-20 flow cytometer. FlowJo software was used for analysis.