Atlas-level single-cell modeling using a hierarchical mixture of experts embedded topic model

Michael Huang, School of Computer Science

McGill University, Montreal

August, 2023

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of

Master of Computer Science

Abstract

Cells are the basic, fundamental building blocks of life. The differences between individual cells are reflected in the genes they express, and better understanding the differences in their transcriptomic landscapes can grant us novel biological insights. Single-cell RNA sequencing (scRNA-seq) is a technique which allows for quantitative analysis of gene expression at the single-cell resolution. As single-cell RNA sequencing technologies have matured, larger single-cell atlases have become readily available. However, analysis of these large-scale atlases remains challenging for existing methods due to computational costs, model scalability, as well as a lack of model interpretability. Additionally, another limitation of existing single-cell modeling methods is that they don't exploit the biological cell-type hierarchy in populations of cells. In order to address these challenges, we propose a Hierarchical Mixture of Experts Embedded Topic Model (HMoE-scETM) as a mixture of experts extension of the single-cell Embedded Topic Model (scETM). Our extension uses a hierarchical gating network and pre-trained experts to explicitly model known biological hierarchies and increase model scalability, while maintaining interpretability. We achieved superior performance in single-cell clustering experiments compared to state-of-the-art methods, and also show promising extended applications to Bulk RNA-seq data.

Abrégé

Les cellules sont les éléments de base fondamentaux de la vie. Les différences entre les cellules individuelles sont reflétées dans les gènes qu'elles expriment, et une meilleure compréhension des différences de leurs paysages transcriptomiques peuvent nous apporter de nouvelles connaissances biologiques. Le séquençage de l'ARN à l'échelle des cellules uniques (scRNA-seq) est une technique qui permet de l'analyse quantitative de l'expression génique aux résolution des cellules. Avec les avancements dans les technologies séquençages, les grands atlas sont devenus facilement disponibles. Cependant, l'analyse de ces atlas à grande échelle restent difficiles pour les méthodes existantes, en raison des coûts de calculs et l'évolutivité des modèles. Une autre limitation avec les modèles existants est qu'elles n'exploitent pas la hiérarchie biologique qui existe dans les populations des cellules. Afin de relever ces défis, nous proposons un combinaison hiérarchique des experts modèles des sujets (HMoE-scETM), une extension du modèle des sujets sur les cellules individuelles (scETM). Notre extension utilise un réseau de portes hiérarchiques et les experts préformés pour intégrer l'information des hiérarchies biologiques, améliorer l'évolutivité de notre modèle, tandis qu'aussi maintenir l'interprétabilité. Nous avons atteint une performance supérieure par rapport aux méthodes état de l'art dans les expériences de regroupement des cellules individuelles et montre les extensions prometteuses sur le séquençage de l'ARN en gros.

Acknowledgements

I sincerely thank my supervisor Prof. Yue Li for his endless patience, advice, and encouragement, without which this endeavour would not have been possible. Special thanks to Dr. Swapna Seshadri, for her guidance and help during our time working together, your perserverence and dedication are an inspiration to me.

I would also like to thank my friends and family, for supporting me throughout my journey. Lastly I would like to give my most heartfelt thanks to my partner Suri, for your encouragement, through lockdowns and through deadlines, you've been there every step of the way.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AD . . . . . Alzheimer's Disease

ARI . . . . . Adjusted Rand Index

ASW . . . . . Average Silhouette Width

Bseq-sc . . . . Bulk Sequence Single-Cell Deconvolution

CCA . . . . . Canonical Components Analysis

CIBERSORT . Cell-type Identificiation By Estimating Relative Subsets of RNA Transcripts

DNA . . . . . Deoxyribonucleic acid

ELBO . . . . . Evidence Lower Bound

ETM . . . . . Embedded Topic Model

GEO . . . . . Gene Expression Omnibus

GSEA . . . . . Gene Set Enrichment Analysis

GTM-decon . Guided Topic Model for Deconvolution

HBC . . . . . Human Blood Cells

HICA . . . . . Hematopoeic Immune Cell Atlas

HMoE . . . . . Hierarchical Mixture of Experts

LDA . . . . . Latent Dirichlet Allocation

LDVAE . . . . Linear Decoder Variational Autoencoder

MDD . . . . . Major Depressive Disorder

MELE . . . . . Mixture of Explicitly Localized Experts

MILE . . . . . Mixture of Implicitly Localized Experts

MoE . . . . . Mixture of Experts

MoE-Sim-VAE Mixture of Experts Similarity Variational Autoencoder

MuSiC . . . . Multi-Subject Single-Cell Deconvolution

NLP . . . . . Natural Language Processing

NMI . . . . . Normalized Mutal Information

PAM . . . . . Pachinko Allocation Model

PBMC . . . . . Peripheral Blood Mononuclear Cells

PCA . . . . . Principal Components Analysis

RMSE . . . . . Root Mean Squared Error

RNA . . . . . Ribonucleic Acid

SHLDA . . . . Supervised Hierarchical Latent Dirichlet Allocation

SVD . . . . . Singular Value Decomposition

Scaden . . . . Single-Cell Assisted Deconvolutional Deep Neural Network

UMAP . . . . Uniform Manifold Projection

VAE . . . . . Variational Autoencoder

hLDA . . . . . Hierarchical Latent Dirichlet Allocation

hLLDA . . . . Hierarchical Labeled Latent Dirichlet Allocation

scANVI . . . . Single-Cell Annotation using Variational Inference

scETM . . . . Single-Cell Embedded Topic Model

scMM . . . . . Single-Cell Multi-Modal

scRNA-seq . . single-cell Ribonucleic Acid sequencing

scVI . . . . . Single-Cell Variational Inference

Chapter 1

Introduction

Understanding complex diseases is one of the key goals of bioinformatics and its applications in modern medicine. As a result of recent advances in molecular biology techniques, researchers have developed novel computational methods in order to leverage a newfound wealth of data in order to better study complex diseases. In particular, single-cell transcriptomics is a field of molecular biology that has grown significantly in past years, and has yielded discoveries in Cancer [30], Rheumatoid Arthritis [35], and COVID-19 [62]. As single-cell transcriptomics technologies mature, larger atlas level datasets spanning hundreds of thousands to millions of cells have begun to emerge, resulting in the need for new computational methods that are both scalable and interpretable.

The advent of single-cell sequencing techniques has also resulted in an improved understanding of cell-lineages, and how different cell-types relate to one another [9]. Existing single-cell modeling technologies do not exploit these biologically meaningful cell-type lineages. The goal of this thesis is to explore a scalable and interpretable model that incorporates prior knowledge in the form of cell-lineage information. To that end, we propose a Hierarchical Mixture of Experts single-cell Embedded Topic Model, as an extension to the published single-cell Embedded Topic Model [78], with extended applications to cell-type annotation and bulk RNA-seq data.

1.1 Statement of Contribution

Chapter III: Prof. Yue Li conceived the idea and supervised the project, providing feedback on model design, and had weekly meetings to discuss project progress. Michael Huang implemented the HMoE-scETM model, adapting it from scETM [78]. Michael Huang additionally ran all experiments and performed the analyses discussed in the thesis.

Chapter 2

Background and Related Works

2.1 Transcriptomics and RNA Sequencing

Transcriptomics is the study of the expression of Ribonucleic Acid (RNA) transcripts in organisms, allowing researchers to capture the expression of genes in order to understand the their function, identify novel transcripts, and study diseases [31, 39]. Bulk RNA-sequencing is a technique that pools cells in a tissue sample together and measures the gene expression levels within that sample. Bulk RNA-seq has been used in recent decades to study gene expression differences between tissues and in different disease states. In cancer studies, bulk RNA-seq has been used to identify immune signatures associated with specific immune-checkpoint blocker treatments [30]. With advancements in sequencing technologies, bulk RNA-seq data has become cheaper to perform, and large public repositories of data have become widely available for researchers, such as the Genome Expression Omnibus (GEO) [3].

A key limitation to bulk RNA-seq studies, is in the pooling of cells, obscuring the heterogeneity that exists on a cell to cell basis. In order to address this limitation, RNA sequencing on the single-cell scale was first introduced by Tang et al in 2009 [67]. These techniques allowed researchers to identify differences between populations of cells, and observe rare cell-types that had been previously obscured by pooling during bulk RNA-

seq analyses [20]. These single-cell transcriptomics studies have lead to advancements in our understanding of conditions such as Cancer [47], Alzheimer's disease [42], and the novel coronavirus COVID-19 [58]. Although bulk RNA-seq remains the cheaper alternative, recent advancements in scRNA-seq has improved both the throughput and cost effectiveness of single-cell experiments. These advancements have lead to the creation of large scale single-cell atlases, containing hundreds of thousands to millions of cells, such as the Human Cell Atlas project [57], the Tabula Sapiens consortium [10], and the Human Cell Landscape [19]. The advent of these large single-cell atlases necessitates the development of computationally efficient, scalable methods for single-cell modeling.

2.2 Topic Models

Topic models are a statistical tool frequently used in the Natural Language Processing (NLP) field, as a method to discover underlying latent semantic patterns in bodies of text [69]. In an NLP context, the goal is often to categorize different documents and group them based on abstract concepts or "topics", based on the words contained within said documents. Mathematically, topics are underlying hidden variables that correspond to probability distributions of words, the observed variables. While topic models are well suited to modeling discrete bodies of text, they have also been successfully applied in other contexts, such as computer vision, social networks [37], and single-cell modeling [78].

2.2.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is an early and frequently used topic modeling approach developed by Blei, Ng, and Jordan [5]. It was proposed as a generative probabilistic model of a set of documents, where documents are random mixtures over k latent topics, which are each characterized by a distribution over words. This model utilizes

a bag-of-words assumption, whereby the order of words within a document is ignored. The generative process for a document d of length N is described as such:

1. Draw a topic proportion $\theta_d \sim \text{Dirichlet}(\alpha)$
2. For each word w_{dn} in the document d:
   (a) Draw a topic $z_{dn} \sim \text{Multinomial}(\theta_d)$
   (b) Choose a word w_{dn} from $p(w_{dn}|z_{dn}, \beta)$, a multinomial probability distribution conditioned on the topic z_{dn}, and β, a topic by words matrix $(k \times V)$ to be estimated, representing the word probabilities.

Expanding upon the success of LDA in unsupervised topic modeling, variants were developed for unsupervised hierarchical topic modeling, such as Hierarchical latent Dirichlet allocation (hLDA) [18] and Pachinko allocation model (PAM) [34]. Models for supervised hierarchical topic modeling were also proposed, such as supervised hierarchical latent Dirichlet allocation (SHLDA) [50] and hierarchical labeled LDA (hLLDA) [54]. LDA was also successfully applied in bioinformatics settings, including genomic microarray data, DNA K-mer sequences, and fluorescent cell imagery [37].

2.2.2 Embedded Topic Model

Classically, words are represented in a one-hot-encoded representation, whereby a word is represented by a vector of length V, where all entries, except the index corresponding to the specific word w, are zero. However a key limitation to this representation is that it does not capture the relationship between words, where similar words should be closer together in a lower dimensional space. **Word Embeddings** are a continuous, dense representation for words in a fixed vector space, often built using word co-occurrence statistics. In this manner, words that are intuitively similar are represented with similar embeddings, that is to say, vectors that are closer together in the embedding space.

In order to improve upon the interpretability of topic models such as LDA, the embedded topic model (ETM) was proposed by Dieng, Ruiz, and Blei [12]. ETM combines the unsupervised topic modeling capabilities of LDA, and is thus able to uncover interpretable latent semantic structures, with the ability to learn low-dimensional word embedding representations.

ETM simultaneously learns embedding representations of both V words and K topics in an L dimensional space. Unlike traditional topic modeling, such as LDA, where each topic is a complete distribution over the vocabulary V, ETM models a topic k with a topic embedding $\alpha_k \in \mathbb{R}^L$, a dense representation of the topic k in the same space as the word embeddings. The generative process for ETM is similar to that of LDA, where for a document d:

1. Draw the topic proportion $\theta_d \sim \mathcal{LN}(0, I)$, a logistic normal distribution, where: $\delta_d \sim \mathcal{N}(0, I);\ \theta_d = \text{softmax}(\delta_d)$

2. For each word w_n in the document:

   (a) Draw a topic $z_n \sim \text{Categorical}(\theta_d)$

   (b) Draw the word $w_n \sim \text{softmax}(\rho^\top \alpha_{z_n})$, where ρ is the $(L \times V)$ word embedding matrix, and α_{z_n} is the assigned topic embedding

For a corpus $W = \{w_1, \ldots, w_D\}$, the marginal likelihood given the parameters for the word embeddings ρ and the topic embeddings α is:

$$\mathcal{L} = \sum_{d=1}^{D} \log p(w_d | \alpha, \rho) \tag{2.1}$$

However this likelihood is intractable, in order to work around this issue, ETM uses variational inference, using an inference network parameterized by Θ_e following the Variational Autoencoder (VAE) described by Kingma and Welling [27]. The inference network here is an encoder neural network, which ingests a document w_d and outputs a mean and

variance of μ_d, Σ_d, which utilizes a reparameterization trick to allow for backpropagation:

$$\epsilon \sim \mathcal{N}(0, I), \quad \delta_d = \mu_d + \epsilon \Sigma_d \tag{2.2}$$

Formally, the goal is to fit α, ρ, and Θ_e, the topic embeddings, word embeddings, and the inference network (also known as the encoder network) parameters respectively, on the corpus $\{w_1, \ldots, w_D\}$. This is done by maximizing the Evidence Lower Bound (ELBO), where the first term is the reconstruction loss, and the second term is the Kullback–Leibler divergence:

$$\mathcal{L}(\alpha, \rho, \Theta_e) = \sum_{d=1}^{D} \sum_{n=1}^{N_d} \mathbb{E}_q[\log p(w_{nd} \mid \delta_d)] - \sum_{d=1}^{D} KL(q(\delta_d; w_d, \Theta_e) \mid\mid p(\delta_d)) \tag{2.3}$$

2.2.3 Guided Topic Models

A limitation to the interpretability of traditional topic models that they require post-hoc interpretation of the topics. In particular they require investigation of the top words associated with a particular topic to understand the meaning behind the learned topics.

Guided Topic Models were introduced in order to address this weakness, by explicitly providing guidance to topics during learning, so that topics have pre-determined meaning [33]. Seed guided topics have been proposed for language models [61, 63], whereby topics are "seeded" with sets of related words to predetermine a topic of interest. An alternative method used to guide topics is to include explicit topic labels into the inference process. This approach has already been successfully used to directly guide topics in a single-cell setting, whereby the topics directly correspond to annotated cell-types (GTM-decon [66], see Section 2.4.3).

2.3 Mixture of Experts Models

In classification settings, it has been shown that models trained on specific subsets of data outperform models that are trained on the entire problem domain [41]. A model designed to exploit localized experts was proposed by Jacobs et al [23], that introduced what is now known as a Mixture of Experts (MoE) approach, whereby a problem domain is broken down into localized experts, modulated by a gating network. The output of a Mixture of Experts model with I experts is described by Equation 2.4.

$$\hat{y} = \sum_{i}^{I} g_i y_i \tag{2.4}$$

Where y_i is the predicted output of expert i, and g_i is the weight assigned to expert i by the gating network, which is a neural network parameterized by Θ_g and a softmax function, such that the total probability distributed by the gating network sums to 1:

$$\begin{aligned} g &= \text{NN}_{\Theta_g}(x) \\ g_i &= \frac{\exp(g_i)}{\sum_{j=1}^{I} \exp(g_j)} \end{aligned} \tag{2.5}$$

Where $g \in \mathbb{R}^I$ is an I dimensional vector that sums to 1, representing the probabilities associated with each expert. The probability of observing y given an input x for the entire Mixture of Experts is as such:

$$\begin{aligned} P(y|x, \Theta) &= \sum_{i=1}^{I} P(y, i|x, \Theta_e, \Theta_g) \\ &= \sum_{i=1} g_i P(y|i, x, \Theta_e) \end{aligned} \tag{2.6}$$

MoE models are broken down into two main taxonomies. Mixtures of Implicitly Localized Experts (MILE), describes MoE models whereby the problem domain is stochastically partitioned, and includes the original MoE model proposed by Jacobs et al [23].

Mixtures of Explicitly Localized Experts (MELE) are MoE models that have the problem domain partitioned in a non-stochastic manner, and have been found to generally perform better than MILE methods [41]. MELE models partition the problem space such that the local experts can specialize in pre-determined subspaces. These explicitly determined subspaces can be a result of clustering algorithms, or the integration of domain expertise.

2.3.1 Hierarchical Mixture of Experts

Mixture of Experts models can be further expanded to include hierarchical information. Proposed by Jordan et al [25] as an extension of the conventional MoE model, the Hierarchical Mixture of Experts (HMoE) model further divides the problem space under a divide and conquer strategy, whereby the output is weighted by multiple layers of gating networks. The probability model for an HMoE is as such:

$$P(y|x,\Theta) = \sum_{i=1}^{I} g_i(x,\Theta_{g_i}) \sum_{j=1}^{J_i} g_{j|i}(x,\Theta_{g_{j|i}}) P_{ij}(y,\Theta_e) \tag{2.7}$$

Where g_i is the weight assigned by the top level gating network, and $g_{j|i}$ is the weight assigned by the j-th gating network associated with the i-th weight. The gating networks are parameterized by Θ_{g_i} and $\Theta_{g_{i|j}}$ respectively, and can be defined in the same way as a traditional MoE gating network as seen in Equation 2.5, with the caveat that the dimensionality of each gating network is determined by the number of experts or nodes underneath that gate, and may be different for each gating network.

2.4 Related Works

2.4.1 Mixture of Experts in Single-Cell Applications

Mixture-of-Experts Similarity Variational Autoencoder (MoE-Sim-VAE) is an MoE model that explicitly localizes its experts using similarity-based representation learning [28].

MoE-Sim-VAE utilizes a flexible cell-cell ($N \times N$) similarity matrix S, that can be either unsupervised, by performing K-nearest-neighbors clustering to determine an adjacency matrix, or weakly-supervised, incorporating some a priori knowledge as the cell-cell adjacency matrix. MoE-Sim-VAE features a single encoder network, a single gating network, and multiple decoder networks. The gating network is additionally trained to reconstruct the similarity matrix S from the gating network weights. The gating network thus is trained in a manner that explicitly partitions the cells to different experts. A drawback of the MoE-Sim-VAE model is in its lack of interpretability, as the feed forward network decoders obscure biological information contained in the latent embeddings.

scMM was proposed by Minoura et al [44] in order to model single-cell multiomics data, where for the same cells, multiple modalities of data were captured simultaneously. An example application of multi-modal single-cell modeling is CITE-seq data [64], which simultaneously captures transcriptomic and surface protein quantities at the single-cell resolution. scMM utilizes a VAE based approach, with a unique encoder and decoder for each modality. The mixing of experts occurs in the latent embedding, where the experts share the same latent embedding space. The latent embedding is computed by averaging the variational posterior $q_{\phi_m}(z|x_m)$ across modalities. Each modality is reconstructed from the same latent embedding using a modality-specific decoder.

2.4.2 Single-Cell Integration Methods

As larger scRNA-seq datasets become available, the necessity of models for integrative analysis becomes clear. These methods can be designed to cluster cells for the identification of novel cell-types, correct for experimental batch effects, and transfer knowledge to new datasets.

Harmony [29] utilizes a low dimensional Principal Components Analysis (PCA) embedding of single-cell data in order to perform soft clustering, then learns a series of linear correction factors on those clusters. Harmony iterates between clustering on the low dimensional embedding $\hat{Z}$, and learning a series of linear correction factors ϕ until

convergence, whereby batch effects are regressed out in the low-dimension embeddings. **Seurat** [8] utilizes Canonical Correlation Analysis (CCA) of multiple datasets in order to identify linear combinations of genes that are strongly correlated across datasets. This conserved gene correlation structure is then transformed using dynamic time warping to maximally align cells, resulting in a single, integrated, low dimensional projection of the scRNA-seq data. **Scanorama** [22] adapts a mutual nearest neighbors approach initially conceived for pattern matching in images, applied to single-cell data. In order to perform integration, Scanorama compresses the combined scRNA-seq data to a lower dimension using a randomized Singular Value Decomposition (SVD) algorithm. Scanorama then performs all-to-all dataset matching, by taking a cell from a dataset, and querying the remaining cells only from the other datasets for nearest neighbors. A drawback of Harmony, Seurat and Scanorama is that these three models rely on dimensionality reduction for computational scalability, and are not transferrable to new datasets.

Single-Cell Variational Inference (scVI) [38] is a deep learning, variational inference approach that utilizing a VAE architecture. scVI utilizes a zero-inflated negative binomial (ZINB) distribution for its underlying generative process, conditioned on batch labels, a gaussian variable modeling library size differences, and the latent embedding of interest. The decoder for scVI is a typical non-linear deep neural network. In a recent state-of-the-art benchmarking of multiple integration techniques, scVI places third for scRNA-seq integration [40], the two methods that outperformed it are scANVI, which utilizes cell-type annotations, and Scanorama, which is a non-transferrable method. A key drawback of scVI is in its lack of model interpretability, due to the black-box nature of the neural network decoder.

Linear Decoder Variational Autoencoder (LDVAE) [65] is an interpretable extension to scVI, keeping the neural network encoder features, while replacing the neural network decoder with a linear function W, an embedding by genes matrix. This linear decoder provides interpretability, providing a direct link between the latent dimensions and gene

expression. This extension explicitly trades model flexibility in favour of interpretability, at the cost of increasing reconstruction error.

Single-Cell Annotation Using Variational Inference (scANVI) [75] is a semi-supervised extension of scVI, incorporating cell-type annotation information into the generative model. In particular, the generative model for the latent embedding z is conditioned on the cell-type label c, drawn from a multinomial distribution, and a gaussian variable u, representing the within-cell-type variation. Benchmarking for scRNA-seq integration showed that scANVI was the top performer for integration [40]. In addition to integration, because scANVI is a semi-supervised model, it is also able to perform cell-type prediction. A drawback of the scANVI model is that, like scVI, it lacks interpretability in the decoder.

Single-Cell Embedded Topic Model (scETM) [78] is a topic modeling approach to single-cell integration. It is the application of the ETM model described in Section 2.2 applied to single-cell modeling, and is described in detail in Section 3.1.

2.4.3 Single-Cell Reference Based Bulk Deconvolution Methods

Cell-type proportions within a tissue have been shown to be indicators of disease state. Type 2 Diabetes has been characterized as a reduction in Beta cell proportion [60]. Alzheimer's disease is characterized by a reduction in neuronal cell proportion [1]. Although costs of scRNA-seq have been greatly reduced with recent advancements, bulk RNA-seq experiments remain far more economical. It is for this reason that several methods have been created that attempt to predict cell-type mixing proportions in a bulk RNA-seq sample, using scRNA-seq data as a reference, in a task commonly called Deconvolution [66].

Cell-type Identification By Estimating Relative Subsets Of RNA Transcripts (CIBERSORT) [48] is an early marker-based approach that utilizes a cell-type specific gene expression feature matrix and uses linear kernel support vector regression (SVR) to predict relative single-cell abundances in leukocyte populations. Although it is possible to use single-cell data to generate a gene expression feature matrix, CIBERSORT did not utilize single-cell data as a reference for bulk deconvolution. **Bulk Sequence Single-Cell Decon-**

volution (BSeq-SC) [2] and **CIBERSORTx** [49] both extended CIBERSORT to incorporate scRNA-seq reference data in order to generate feature matrices for bulk deconvolution. **Multi-Subject Single-Cell Deconvolution** (MuSiC) [71] is a weighted non-negative least squares regression approach that is marker free, that simultaneously incorporates cross-subject and cross-cell-type variation of gene expression into its deconvolution prediction.

Single-Cell Assisted Deconvolutional Deep Neural Network (Scaden) [43] is a deep learning approach to deconvolution that simulates bulk samples by sampling reference single-cells and artificially pooling them into simulated bulk data with known mixing proportions for training. Deconvolution is thus treated as a supervised regression task, whereby the Scaden neural network learns to predict simulated bulk proportions, and is then later applied to real bulk data. A limitation of this approach, as is often the case with deep learning approaches, is in the black-box like nature of the learned model parameters [4] and the lake of interpretability.

Guided Topic Model for Deconvolution (GTM-decon) [66] is a Bayesian topic modeling approach for deconvolution. GTM-decon learns interpretable cell-type specific gene signatures during inference on a reference scRNA-seq dataset. Each topic is explicitly mapped to a specific cell-type using Bayesian topic prior values based on known cell-type labels of the scRNA-seq data. This differs from traditional topic modeling approaches discussed in Section 2.2, where topics are interpretable but assigned meaning after inference via manual inspection. Due to the one-to-one mapping of topics to specific cell-types, this model can be applied to predict cell-type specific topic mixtures of bulk RNA-seq data.

Chapter 3

Methods

We propose two extensions to the scETM model proposed by Zhao et al [78]. The first is an extension to scETM using a traditional Mixture of Experts approach proposed by Jacobs et al. [23], whereby pre-trained cell-type specific experts are modulated by a gating neural network. The second model is a further extension of the previous Mixture of Experts scETM that exploits the knowledge of hierarchical cell lineages, and the domain knowledge of cell-type lineage tree structures. In both proposed models, we pre-train localized experts on specific cell-types or cell-lineages, and in doing so, explicitly localize the expert models. An alternative interpretation of the model is that by explicitly localizing cell-type experts, we are guiding the topics for each expert to correspond to a specific cell-type.

3.1 Single-Cell Embedded Topic Model

Single-cell Embedded Topic Model (scETM) is an unsupervised generative topic model that is derived from ETM, as described in Section 2.2.2, applied to single-cell RNA-seq data [78]. In the single-cell context, instead of documents, scETM represents a collection of cells $\{x_1, \ldots, x_N\}$, making up the scRNA-seq data X, a cells by genes matrix ($N \times G$). In place of a vocabulary of words, the vocabulary consists of G genes. The scETM generative

process is largely similar to that of ETM, with an adaptation being made in step 2b of ETM's generative process (Section 2.2.2), where a biologically relevant batch correction vector λ is introduced:

$$r_{d,g} = \frac{\exp(\hat{r}_{d,g})}{\sum_{g'} \exp(\hat{r}_{d,g'})}, \ \hat{r}_{d,g} = \theta_d \alpha \rho_g + \lambda_{s(d),g} \tag{3.1}$$

The transcription rate $r_{d,g}$ for gene g in cell d is parameterized by the topic mixture θ_d, the gene embedding ρ_g, and a batch correction term $\lambda_{s(d),g}$ that accounts for biological or experimental batch effects.

Figure 3.1 is model diagram illustrating the design of scETM without the λ batch correction bias term, and also illustrates the probabilistic non-negative matrix factorization interpretation of topic modeling in scETM, where α and ρ represent the topic embedding and gene embedding matrices respectively, seen in the first term in Equation 3.1.

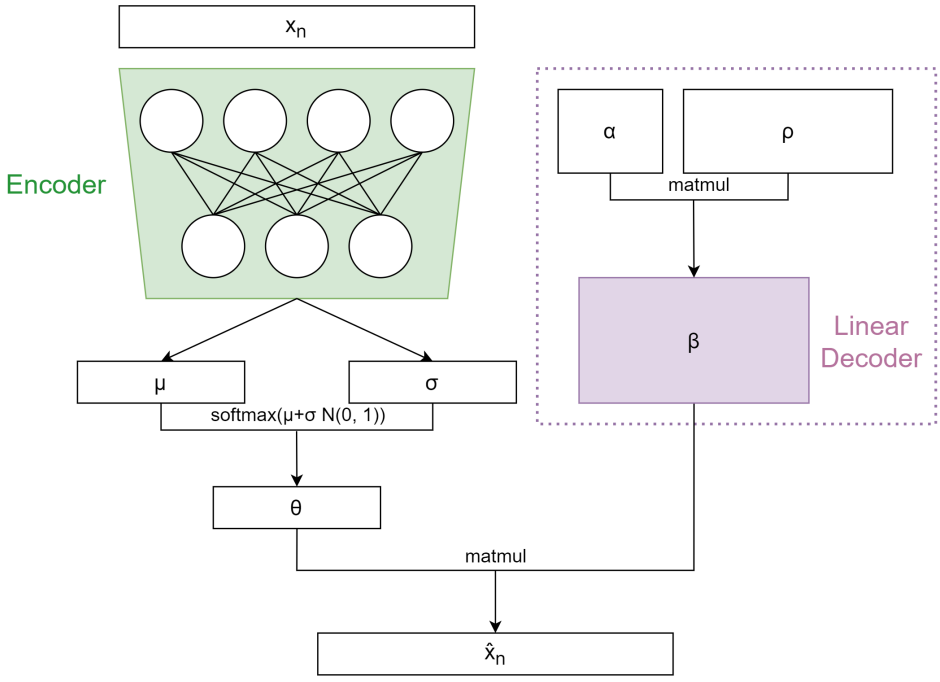

Figure 3.1: scETM model diagram

The inference process for training an scETM model is as follows:

Algorithm 1: scETM Inference

Initialize Model Parameters, Θ_e, α, ρ
for *Epoch* $\leftarrow 1, 2, \ldots,$ *MaxEpochs* **do**
  for $x_i \leftarrow$ *Sample Minibatch* **do**
    Compute $\mu_i, \sigma_i \leftarrow \text{NN}(x_i; \Theta_e)$
    Sample $\epsilon \sim \mathcal{N}(0, I)$
    Perform reparameterization trick $\delta_i = \mu + \sigma_i \epsilon$
    Perform softmax $\theta_i \leftarrow \text{softmax}(\delta_i)$
    Compute NLL $\leftarrow x_i \log \theta_i \alpha \rho$
    Compute KL given μ_i, σ_i
  end
  Compute ELBO given NLL and KL, and compute gradient for backprop
  Update model parameters Θ_e, α, ρ
end

Figure 3.2: scETM training algorithm

3.2 Mixture of Experts scETM

The simplest form of MoE architecture is to simply have multiple experts governed by a gating network. We utilize a series of pre-trained scETM units, each trained on a unique cell-type, to function as the experts. Formally, consider a dataset of single cells $X = \{x_1, \ldots, x_N\}$, which we then partition into C unique cell-types. We then train C different scETM experts, one for each cell-type, each with a latent dimension L, and then combine the scETM units together, using a gating network $\text{GN}(x) = g \in \mathbb{R}^C$. We can formalize the reconstruction process to be as such:

$$\hat{x}_n = \sum_{k=1}^{C} g_k \theta_{n,k} \beta_k \tag{3.2}$$

Where θ_k is the latent representation generated by the k-th encoder, and g_k is the k-th index output of the gating network. We can also consider this as a formulation to be equivalent to a concatenation of our model components, as seen in Figure 3.3. Where

we compute the weighted topic embedding $g_k\theta_k$, and concatenate the gating network weighted embeddings:

$$\theta_{\text{MoE}} = g_1\theta_1 \oplus \ldots \oplus g_C\theta_C \tag{3.3}$$

Yielding $\theta_{\text{MoE}} \in \mathbb{R}^{CL}$ a combined latent embedding, where for a single cell x_n, its MoE latent embedding θ_n is $(1 \times CL)$. At the same time, we can consider the pre-trained decoders for our C experts, parameterized by α_k and ρ_k, and vertically concatenate them together to form a unified linear decoder that has dimensionality $(CL \times G)$:

$$\beta_{\text{MoE}} = \alpha_1\rho_1 \oplus \ldots \oplus \alpha_C\rho_C \tag{3.4}$$

The utility of conceptualizing this model in a matrix concatenation format is in maintaining the interpretability of the model outputs, in particular, we are still able to use the combined latent embedding for visualization, and can still perform differential gene expression analysis on the combined linear decoder. Additionally, each expert can be trained on a subset of the data in parallel, allowing for faster model training.

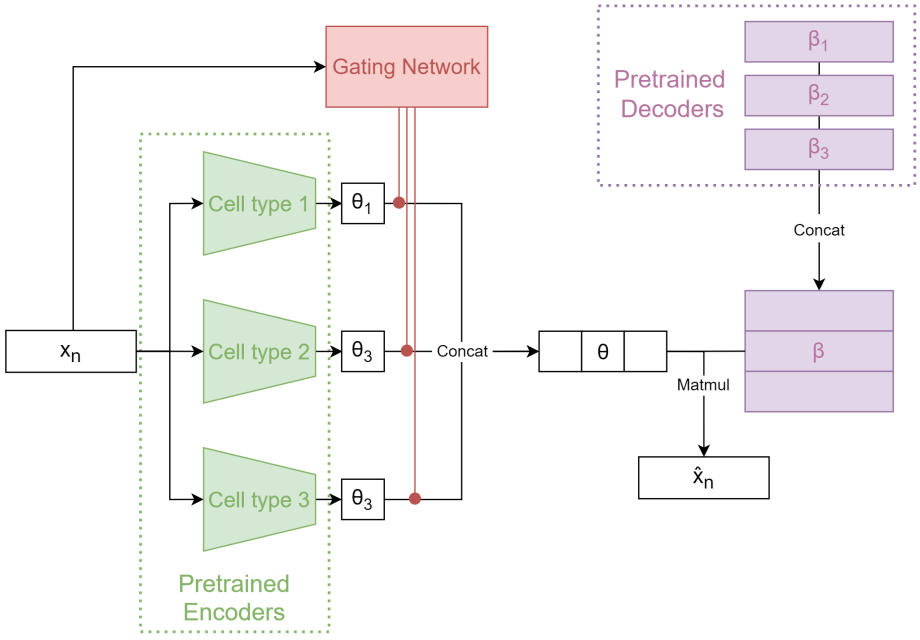

Figure 3.3: Mixture of Experts scETM Model

If we are to incorporate a reconstruction bias term λ as mentioned in scETM for each expert, we can extend Equation 3.2:

$$\hat{x}_n = \sum_{k=1}^{C} g_k \theta_{n,k} \beta_k + \sum_{k=1}^{C} g_k \lambda_k \tag{3.5}$$

The training process for the MoE-scETM model is largely the same as the training process for scETM, whereby cell-type specific experts are trained following Algorithm 1:

Algorithm 2: HMoE/MoE Model Inference

for *Cell-type c in reference scRNA-seq data into cell-type specific subsets* **do**
  Train cell-type specific scETM$_c$ expert using Algorithm1
  $\Theta_c, \alpha_c, \rho_c \leftarrow$ scETM$_c$
end
Initialize model gating network parameters Θ_g, (in the case of HMoE additional: $\Theta_{i|g}$)
for *Epoch* $\leftarrow 1, 2, \ldots,$ *MaxEpochs* **do**
  for $x_i \leftarrow$ *Sample Minibatch* **do**
    Compute $\theta_1, \theta_2, \ldots, \theta_C \leftarrow$ Encoder$_c(x; \Theta_c)$
    Compute $g \leftarrow \text{GN}(x_i; \Theta_g)$
    Compute $\hat{x} \leftarrow \sum_{c=1}^{C} g_c \theta_c \beta_c$
    Compute KL $\leftarrow \sum_{c=1}^{C} g_c \text{KL}_c$
  end
  Compute ELBO given NLL and KL, and compute gradient
  Finetune model parameters Θ_g
end

Figure 3.4: MoE Training Algorithm

3.3 Hierarchical Mixture of Experts scETM

In a biological setting, single-cells often have well studied developmental hierarchies [9]. We can consider the previous MoE-scETM to be a special case of a hierarchical model where there is a single hierarchical level. We can extend this to a well defined tree structure as seen in Figure 3.5. We can consider our Hierarchical Mixture of Experts to be a tree, where each leaf is a localized scETM expert, and each internal node determines some cell-type grouping, eventually all connecting back to a root node. An internal node N contains a gating network, a feed-forward neural network with a fixed output size $|g| = |\text{Children}(N)|$, the number of children, and a softmax layer, such that the total weight at each node is distributed among its children.

In our setting, we can consider the weight at each node to be the probability assigned by the node to each of its leaves, and thus the final weight assigned to a leaf is the cumulative product of all weights tracing the path from the root to the leaf. The reconstructed gene expression for cell n is as follows:

$$\hat{x}_n = \sum_{k=1}^{K} \theta_k \beta_k \prod_j g_{j|k} \tag{3.6}$$

Where θ_k and β_k are the softmax of the latent embedding δ_k obtained from the k-th leaf encoder after sampling via the reparameterization trick as described in equation 2.2 and the linear decoder for the k-th leaf respectively, and $g_{j|k}$ represents the gating weights associated for the path from the root to leaf node k. The training algorithm for fine-tuning the HMoE model follows Algorithm 2 for MoE, with additional parameters for internal node gating networks.

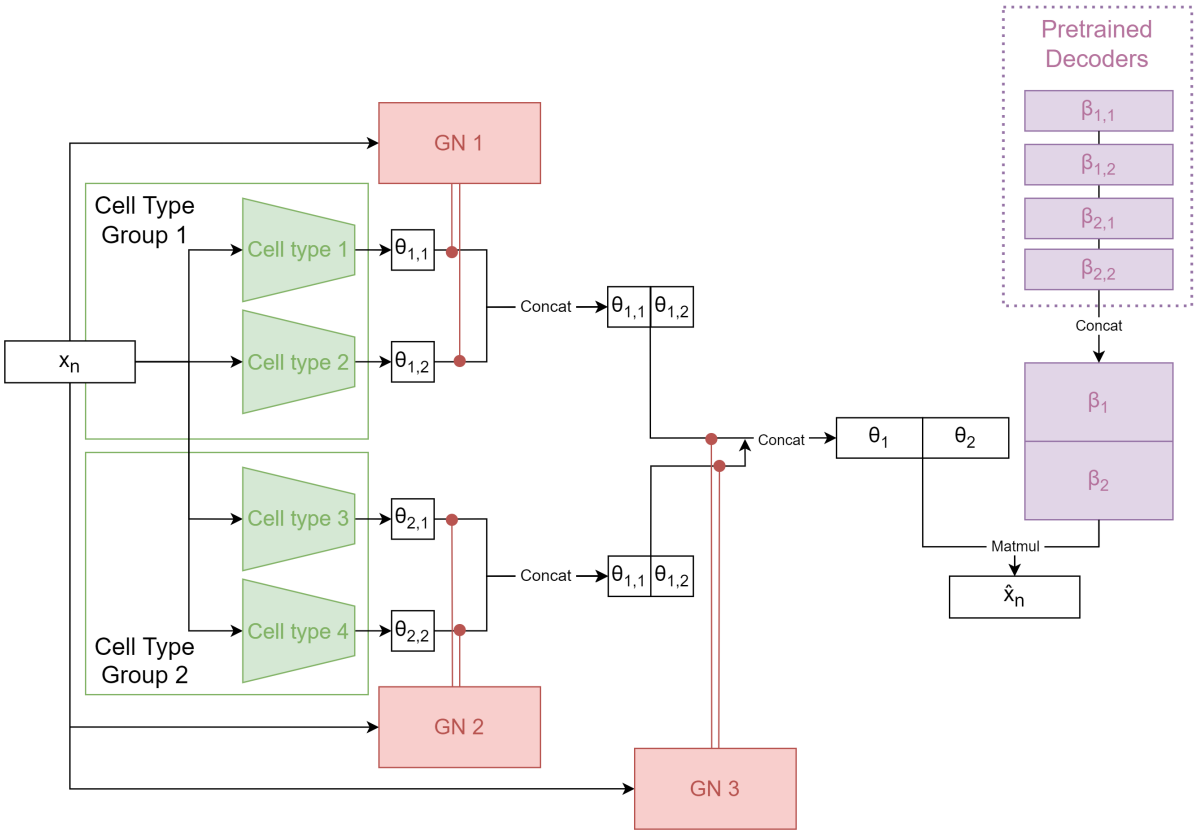

Figure 3.5: Hierarchical Mixture of Experts scETM Model

3.3.1 Hierarchical Tree Structure

The HMoE implementation requires a priori domain knowledge to first build the hierarchical tree. For immune cells, the cell lineage tree is well studied, branching from Hematopoietic Stem and Progenitor cells into Lymphoid and Myeloid branches [14, 17, 52]. An example immune cell lineage tree adapted from Janeway et al [24] and Fang et al [14] is illustrated in Figure 3.6. Additional hierarchical information can be included in an HMoE model, such as sub-cell-types, for example CD8+ and CD4+ T cells falling under a T cell node. For brain cells, we use a lineage tree that separates brain cells into three major cell groups: endothelial/stromal cells, glial cells, and neuronal cells, used by Zhang et al [77].

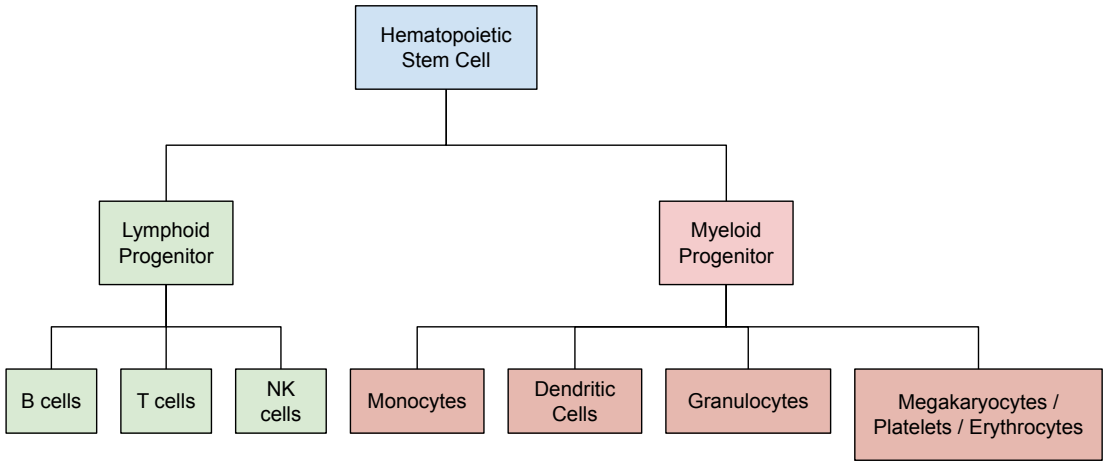

Figure 3.6: Immune cell developmental lineage tree.

3.4 Datasets and Preprocessing

Single-cell reference datasets were preprocessed using the *scanpy* python library [72], following the preprocessing tutorial guidelines. For reference datasets, cells with less than 200 genes expressed were removed, and genes present in less than 3 cells were discarded. During transfer training, only overlapping genes between reference and target dataset

Dataset	Accession #	# Cells	# Genes	# Cell-types
AD [42]	SYN18485175	70634	17775	8
COVID19 Atlas [62]	EMTAB10026	647360	24929	18
HBC [73]	GSE149938	7643	19813	7
Hematopoietic Immune Cell Atlas [56]	ERP122984	829081	22240	18
Lake Frontal Cortex [32]	GSE97930	10319	34305	8
MDD [46]	GSE144136	78886	28839	8
PBMC [13]	GSE132044	17980	23863	9
Tabula Sapiens [55]	GSE201333	483152	58870	177

Table 3.1: Single-Cell Datasets

Dataset	Accession #	# Samples	# Genes
Purified Immune Bulk [45]	GSE107011	127	58183
ROSMAP Prefrontal Cortex [51]	SYN3505724	41	11675

Table 3.2: Bulk RNA-seq Datasets

were used. Single-cell datasets after preprocessing are summarized in Table 3.1, and Bulk RNA-seq datasets are summarized in Table 3.2. During training, the reference single-cell data was split into 81% training, 9% validation, and 10% withheld testing. The validation data was used for model early-stopping to reduce overfitting.

3.5 Model Implementation

The Adam optimizer [26] was used for pre-training cell-type specific scETM experts, as well as for fine-tuning the gating network parameters in both MoE-scETM and HMoE-scETM. The learning rate for the cell-type experts was set to 1e-3, and during fine-tuning the expert parameters were frozen and the gating networks learning rates were set to 1e-4. A weight decay term of 1.2e-6 was included for the L2 regularization of neural network parameters. Each cell-type specific expert was assigned a latent embedding dimension of 5, and the topic and gene embedding dimensions were set to 400. Each scETM encoder network and each MoE/HMoE gating network had two internal layers of dimensional-

ity (128, 128). Each model was trained until the NELBO of the withheld validation set stopped decreasing. The hierarchical tree structure used for the models are illustrated in Figure 3.7, left to right the trees are for the HMoE models trained on HICA, PBMC, MDD, and AD datasets as reference respectively.

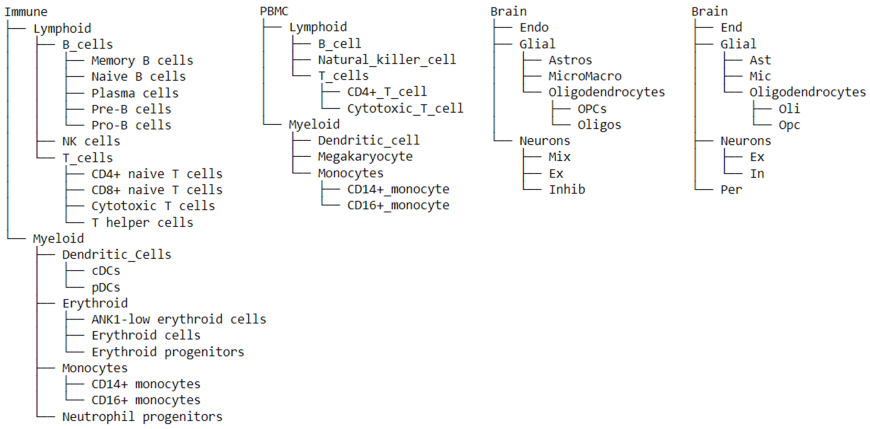

Figure 3.7: Hierarchical cell-type lineage tree used for HMoE models.

3.6 Evaluation Metrics

Adjusted Rand Index (ARI) is a clustering metric that measures the overlap between two data clusterings. ARI compares the pair-wise labelling agreement; in a classification perspective, it considers the number of true positives - the number of pairs that are in the same cluster in both clusterings, and the number of true negatives; the number of pairs that are in different clusters for both clusterings. An ARI of 1 indicates perfect clustering overlap between two sets of labels, whereas an ARI of 0 indicates random labeling. **Normalized Mutual Information** (NMI) is another clustering metric that evaluates the quality of a clustering. NMI is the mutual information score normalized by the entropy of the known class and predicted clustering labels. The NMI score will range between 0 for uncorrelated labels, and 1 for perfectly matching labels. **Average Silhouette Width** (ASW) is a clustering metric that compares the distance between clusters and the tightness of clusters. Silhouette Width (SW) measures the difference between the inter-cluster dis-

tance and the intra-cluster distance [59], whereby a high SW would indicate that a cluster is tight and well separated from other clusters. ASW is the average of Silhouette Widths over all clusters, ranging from -1 to 1, where for a very good clustering the inter-cluster distance is much greater than the intra-cluster distance for all clusters.

$$ASW = \frac{1}{|C|} \sum_{c \in |C|} SW_c \tag{3.7}$$

Where the Silhouette Width of a single cell i is obtained by:

$$SW_i = \frac{b_i - a_i}{\max a_i, b_i} \tag{3.8}$$

The inter-cluster distance b is defined as the mean distance of a data point $i \in C_I$ to the nearest other cluster k, where $d(.)$ is a distance function, in this case the euclidean distance on the embedding dimension:

$$b_i = \min_{J \neq I} \frac{1}{|C_J|} \sum_{j \in C_J} d(i, j) \tag{3.9}$$

The intra-cluster distance a is defined as the mean distance between all pairs of points from the same cluster:

$$a_i = \frac{1}{|C_I| - 1} \sum j \in C_I d(i, j) \tag{3.10}$$

ARI, NMI, and ASW were computed using their respective python modules from the *scikit-learn* package [53]. For ARI and NMI, one set of clusterings is the true cell-type labels from the dataset annotations. The second data clustering used is the predicted labels, generated by computing the Louvain [6] clusterings on the model's latent embeddings. The Louvain clusterings were computed using the *scanpy* python library [72]. For ASW, the distance is computed on the latent embeddings, and the labels used are the cell-type annotations, independent of any clustering algorithm.

For cell-type annotation prediction, multi-class classification accuracy was evaluated. In the case of zero-shot transferring, cell-types were manually aligned between datasets. The true labels used were the original authors annotations, and the predicted labels are the HMoE-scETM model gating network weights. For imputation we consider the root mean squared error (RMSE) of the imputed genes, and the pearson correlation coefficient between imputed and originally masked genes, implemented using the python library *scipy* [70].

3.7 Cell-type Annotation

In order to perform cell-type annotation, we first train an scETM expert for each cell-type, to the granularity of the labels available. After fine-tuning the Mixture of Experts or Hierarchical Mixture of Experts model, we can consider the gating weights associated with each leaf to be the probability weights attributed to each specific cell-type expert. In other words, the weights at the leaves can be considered to be a cell-type prediction, where the weights $g = \{g_1, \ldots, g_C\}$ is a vector that sums up to 1, and each weight g_i is obtained from the following equation:

$$g_i = \prod_j g_{j|i} \tag{3.11}$$

Where the weight associated with a cell-type expert leaf i is the product of the gating weights from the root to leaf i, and the gating weight is the i-th index of associated gating network output. The cell-type label predicted for a given sample x_n is given by $\text{argmax}_i(g_i)$, the index corresponding to the cell-type with the highest gating weight probability.

3.8 Applications to Bulk RNA-seq Data

As an exploratory analysis of potential extensions to this method, we applied the HMoE-scETM model via zero-shot transfer to two bulk RNA-seq datasets, ROSMAP [51], a set of heterogenous brain prefrontal cortex samples, and GSE107011 [45], a set of purified bulk immune cells. We trained two HMoE-scETM models using AD scRNA-seq data [42] and Hematopoietic Immune cells [56] respectively. Both models were trained on only genes overlapping between reference and target bulk. The gene-expression counts of the bulk samples were normalized per bulk sample in the same manner as the single-cell data, such that each sample would sum to 1.

3.9 Marker Gene and Gene Set Enrichment Analysis

In the original scETM model, the learned topics by genes β matrix can be analyzed for both marker gene enrichment and gene set enrichment. In our HMoE extension, we combine the separate linear decoders together via concatenation as described equation 3.4, where for C experts, each with T topics, we yield a $\beta_{\text{HMoE}} \in \mathbb{R}^{CT,G}$. For topic t, we can sort $\beta_{t,:}$ and compare the top genes against CellmarkerDB [76], a unified database of cell-type specific marker genes.

Gene set enrichment analysis (GSEA) can be performed to identify functional pathways that are overrepresented in a set of genes, and can elucidate phenotypic differences. GSEA to identify topic specific enriched pathways was performed using the *gseapy* python package [15]. In order to determine which topics to investigate, we performed permutation tests using the *scipy* implementation [70], conditioned on COVID-19 disease status (healthy vs severe/critical). The test statistic used was difference of means, and sampling was performed 100000 times. We applied Bonferonni correction for multiple hypothesis testing for an adjusted p-value ¡ 0.05.

For topic t, the un-normalized $\beta_{t,:}$ was used as the gene ranking metric. The Enrichr [74] gene set *COVID-19_Related_Gene_Sets_2021* was provided to *gseapy*.

3.10 Ablation Study

In order to evaluate the improvements yielded by incorporating cell-type information and cell lineage information, we compared the performance of baseline scETM model against the Mixture of Experts and Hierarchical Mixture of Experts implementations on the MDD, PBMC and Tabula Sapiens datasets. All models were trained on the same 90% training data split and evaluated on the same 10% withheld test set. For the PBMC and MDD datasets, each cell-type specific expert was trained with 5 topics. For the Tabula Sapiens dataset, in the absence of a complete cell-type lineage, we only compared MoE-scETM and scETM, where the mixture of experts aggregated cell-types together under their 4 major lineages (immune, epithelial, endothelial, stromal), with 50 topics per cell lineage. The scETM baseline used for all experiments had the same total embedding dimension as the respective HMoE and MoE model.

3.11 Model Comparisons

We compared our Hierarchical Mixture of Experts model against three existing published methods deep-learning based single-cell modelling methods: scVI [38], LDVAE [65], and scANVI [75]. All three models were implemented using the *scvi-tools* python library [16], and were run following their respective tutorials. All three models were used to evaluate clustering metrics on withheld test cells, while scVI and LDVAE were compared against for zero-shot transfer clustering evaluation. All models were parameterized such that they had the same size latent embedding, and had their encoders set to two layers.

Chapter 4

Results

4.1 Clustering Performance

For clustering performance on withheld test cells (Figure 4.1), the HMoE model both with leaf-bias (HMoE-LB) terms and without leaf-bias (HMoE-XB) terms perform comparably to LDVAE, scVI and scANVI in terms of ARI and NMI on the two brain cell datasets (AD, MDD). On the two immune cell datasets (PBMC, HICA) the HMoE models tend to outperform existing methods in ARI and NMI. When evaluating zero-shot transfer performance (Figure 4.2), HMoE models performed better when using either MDD or AD as reference and zero-shot transferred embeddings to either the Lake Frontal Cortex dataset, AD, and MDD datasets. HMoE models however, performed weaker than the LDVAE and scVI models transferring from the PBMC dataset to the HBC dataset. In all tasks, the HMoE models tend to outperform existing methods in terms of ASW, with the HMoE model without leaf bias generally performing slightly better.

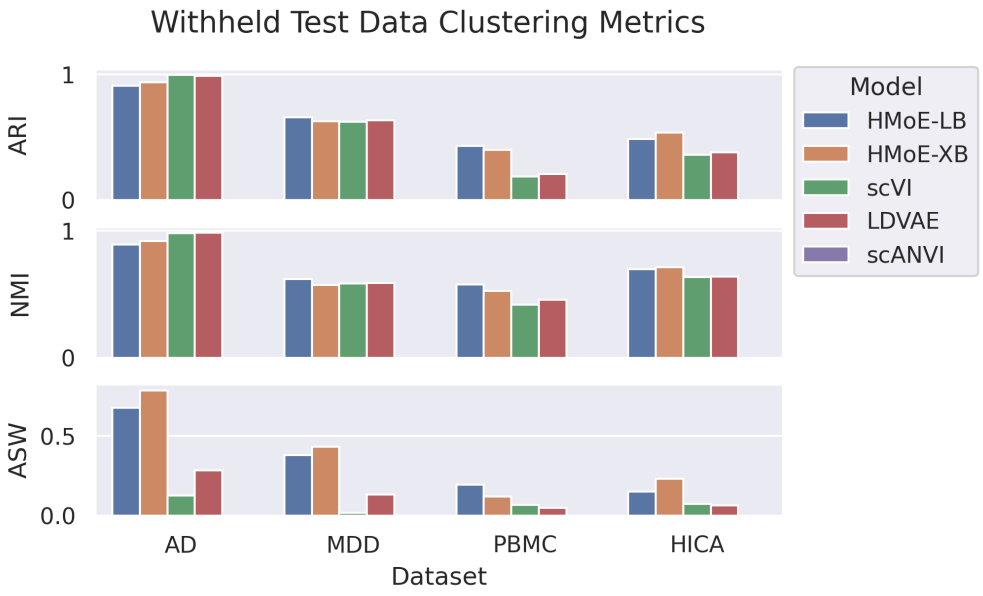

Figure 4.1: Model clustering performance on withheld test cells.

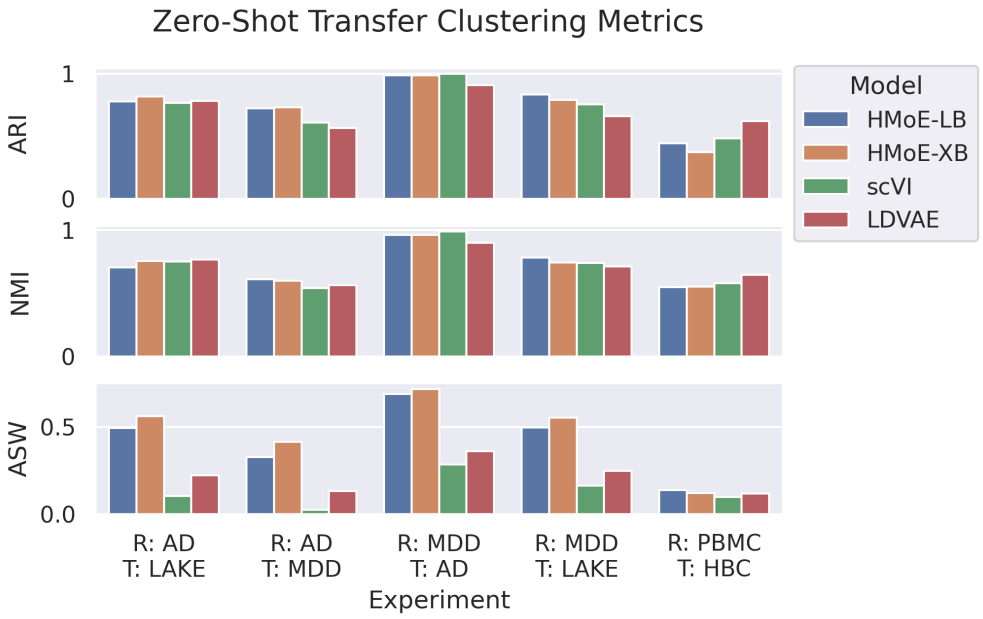

Figure 4.2: Model clustering zero-shot transfer performance.

4.2 Ablation Study

Ablation studies were conducted to evaluate the effects of extending the original scETM model to the Mixture of Experts and Hierarchical Mixture of Experts on three different datasets (Figure 4.3). For the MDD brain dataset, there is mostly parity between HMoE and MoE clustering performance, whereas for the PBMC ablation study there is a significant increase in clustering performance from scETM to MoE for ARI and NMI, and a slight further increase in performance when incorporating the Hierarchical Mixture of Experts (Figure 4.3, middle column). For the Tabula Sapiens dataset, the dataset was split into 4 experts representing general cell-lineage compartments (Immune, Endothelial, Epithelial, and Stromal). Explicitly training separate components for the major cell-lines conferred slight improvements to clustering performance compared to a comparable scETM model. All models performed poorly in terms of ASW on that task, notably there are many more unique cell-type labels in the Tabula Sapiens dataset relative to other datasets (Table 3.1).

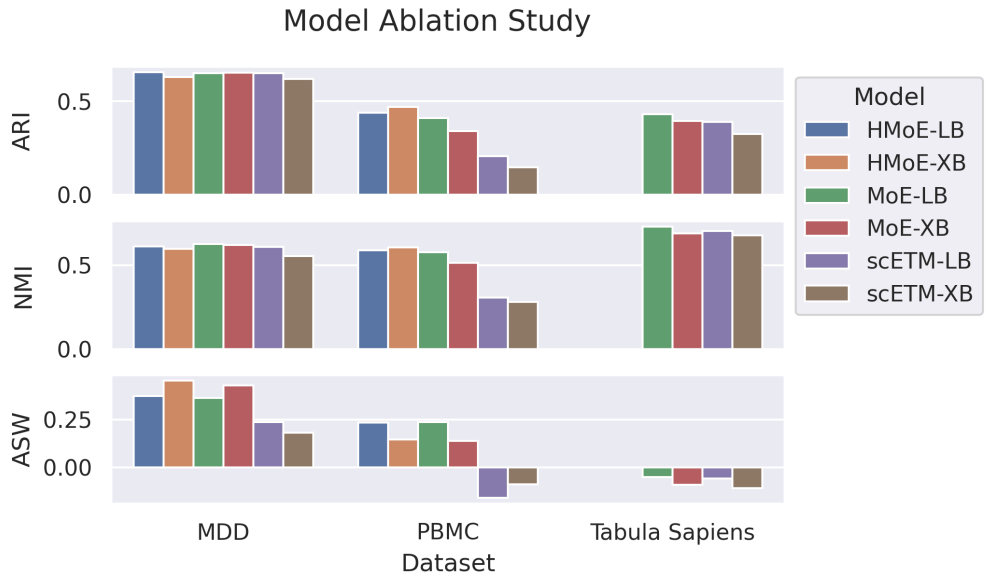

Figure 4.3: Model Clustering Performance on withheld test cells.

4.3 Imputation Performance

We evaluated the performance of the different models in terms of imputation accuracy when 10% of genes were randomly masked, and repeated this experiment for five repetitions. Figure 4.4 shows that the imputation performance of the HMoE models for withheld test cells in the PBMC and HICA immune cell datasets performed modestly better on average compared to other models in terms of RMSE. However imputation RMSE was slightly inferior compared to scVI, LDVAE, and scANVI on the test cells for the brain datasets (AD and MDD). In terms of Pearson R, the HMoE models generally perform slightly worse compared to the benchmark methods.

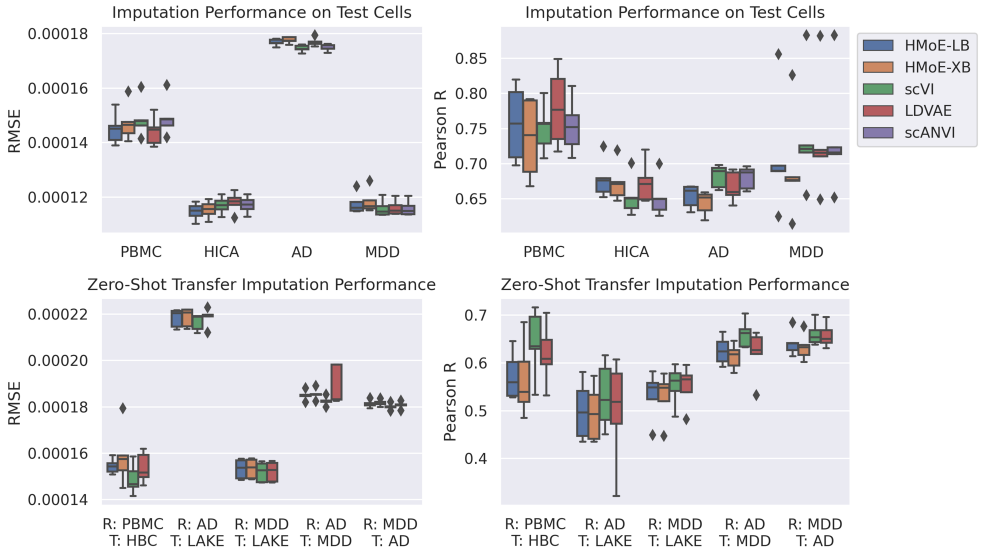

Figure 4.4: Imputation RMSE and Pearson correlation coefficient.

4.4 Cell-Type Annotation

As shown in Figures 4.5 and 4.6, HMoE models perform well on in terms of predicting cell-type annotations on both test data, achieving 98.3% and 91.0% prediction accuracy on withheld test cells for HMoE models trained on AD and MDD datasets as reference. Even when there are many more cell-types, increasing cell-type label specificity with sub-cell-type labels, as seen in Figure 4.6, the HMoE model gating network weights can still predict the cell-type annotations with a high degree of accuracy (76.5% accuracy).

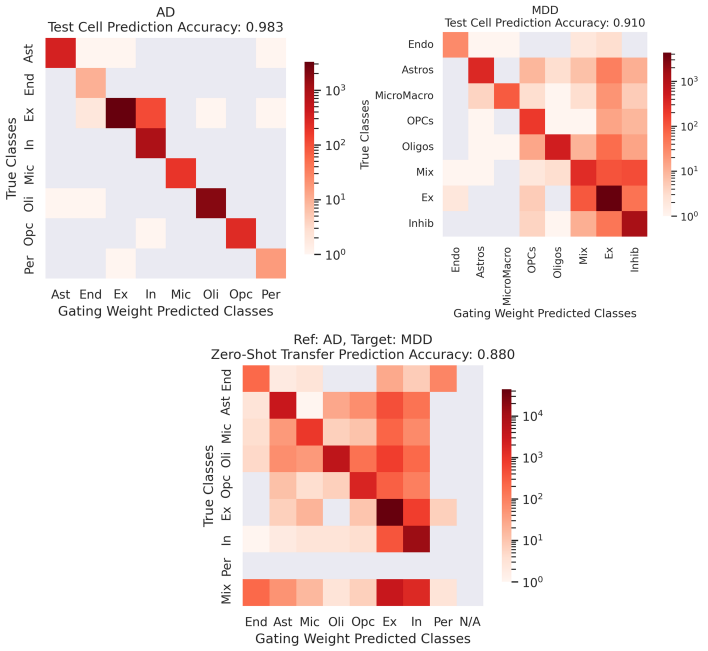

Figure 4.5: Cell-type annotation confusion matrix for Brain cells.

The HMoE model is also able to predict cell-type annotations when zero-shot transferred between datasets, as seen in the bottom panel of Figure 4.5. For an HMoE model trained on AD as reference, when transferred to MDD data, the model maintains a high accuracy of 88.0%, with the majority of prediction errors occurring due to a cell-type misalignment between reference and target. The prediction accuracy when removing unaligned cell-type during transfer is 94.5%, comparable to the original cell-type prediction accuracy seen in the top two panels of Figure 4.5.

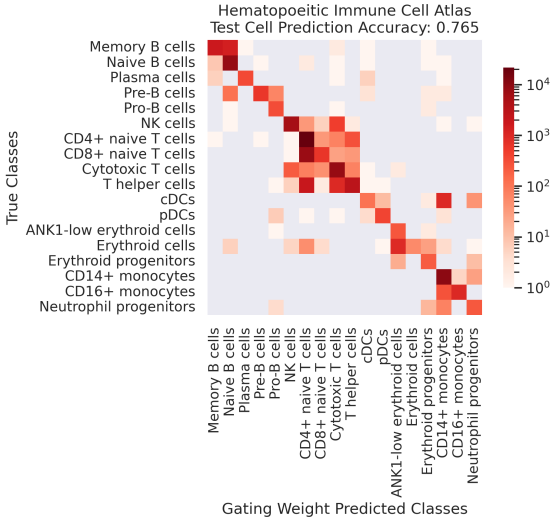

Figure 4.6: Cell-type annotation confusion matrix for Hematopoietic Immune withheld test cells.

4.5 Applications to Bulk RNA-seq Data

Figure 4.7 illustrates the HICA-HMoE model's performance when zero-shot transferred to a purified immune bulk RNA-seq dataset (GSE107011 [45]). The model is able to generally predict the correct corresponding cell-types. The model is able to accurately predict NK cells, B cells, Monocytes, and Dendritic cells in the purified bulk. The model struggles the most with differentiating the CD4+ T cell from the CD8+ Naive T cell and Cytotoxic T cells, but still maintains a distinct gating-weight distribution for CD4+ T cells from other T cell subtypes.

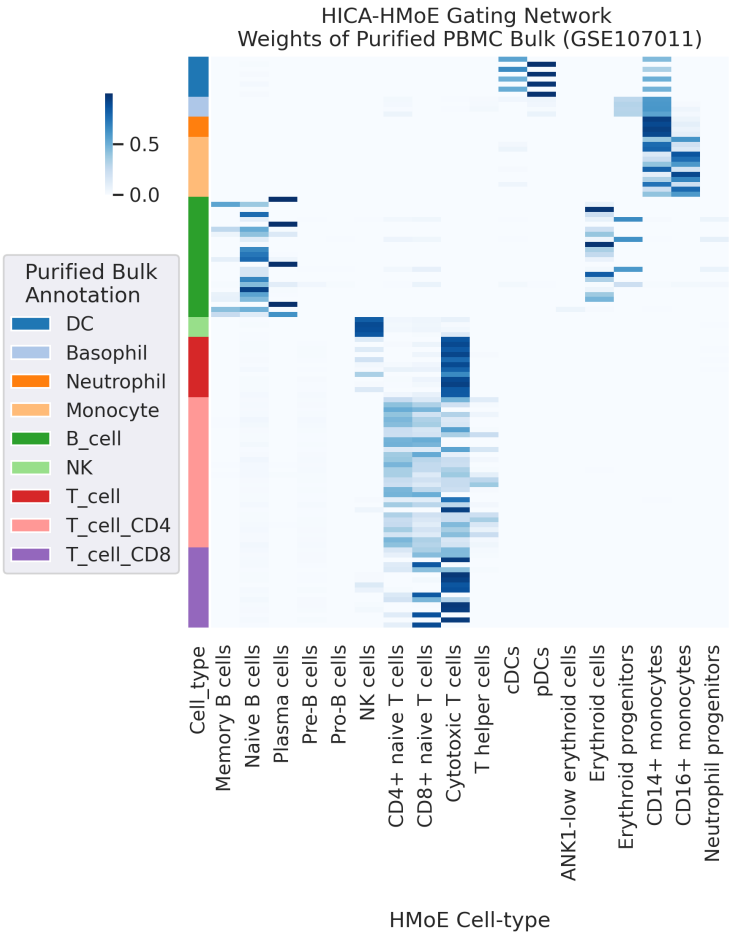

Figure 4.7: Purified bulk immune prediction using HMoE gating weights.

When applied to heterogeneous bulk data, the gating network weights can be interpreted as deconvolution predictions, inferring cell-type mixing proportions. Figure 4.8 shows the overlapping cell-types between annotated known mixing proportions of 41 brain prefrontal cortex bulk RNA-seq samples (left heatmap) and the HMoE inferred cell-type proportions (right heatmap). The model achieves a modest mean Spearman correlation R of 0.617.

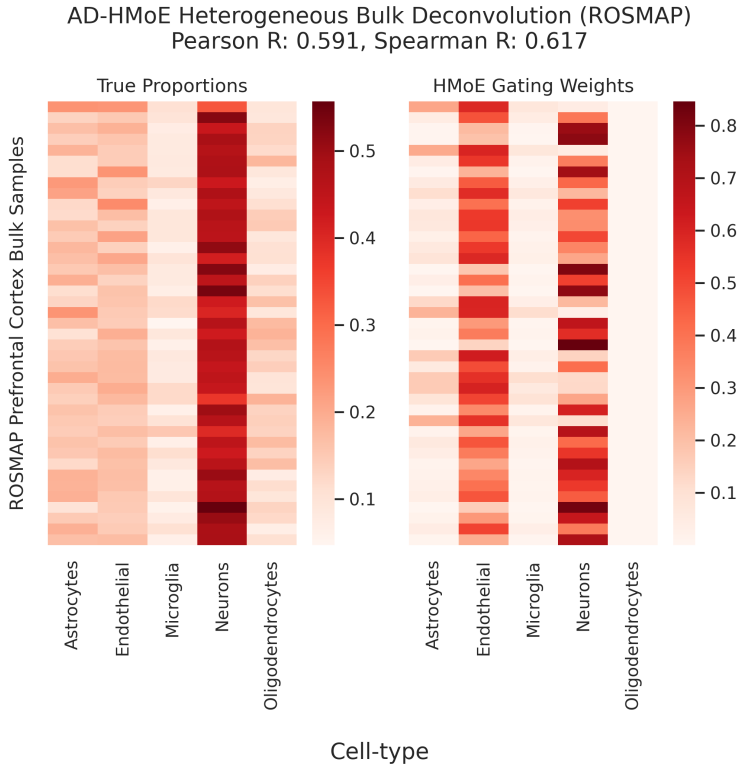

Figure 4.8: Heterogeneous bulk prefrontal cortex deconvolution using HMoE gating weights.

4.6 Marker Gene Enrichment

We investigated the marker gene enrichment for the learned beta-matrices for the HMoE-scETM model trained on the HICA dataset. Figure 4.9 shows the top 10 genes for the top 5 topics enriched for their corresponding cell-type specific marker genes obtained from CellmarkerDB [76]. We can observe that there is a significant amount of marker genes captured in the top genes for cell-type specific topics.

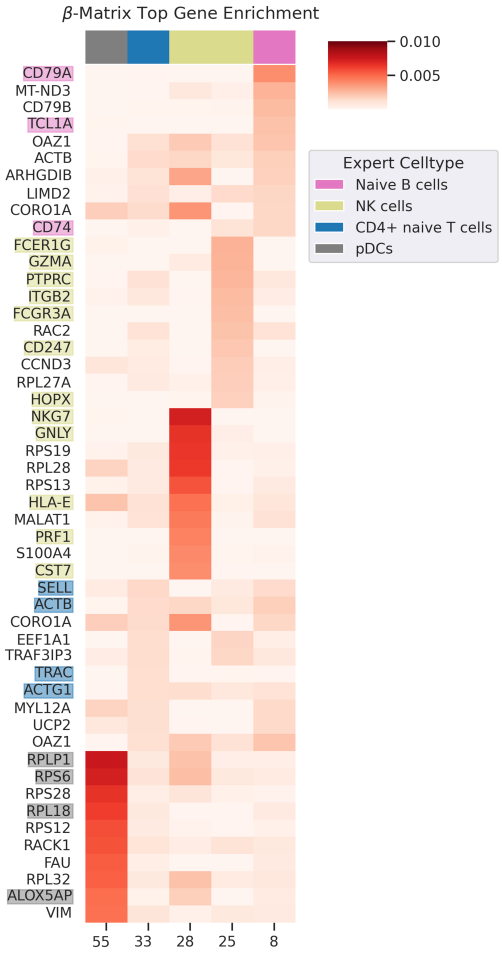

Figure 4.9: HMoE-scETM β-matrix top marker genes.

4.7 Model Interpretability

Figure 4.10 shows the UMAP projections of the latent embeddings generated by models trained on HICA cells applied to withheld test cells. The top row corresponds to HMoE models, and clearly show distinct clusters for lymphoid cells, with specific B cell subtypes distinctly separate from the T cell subtypes and the NK cells. Separate from the lymphoid clusters, there are additionally distinct myeloid clusters.

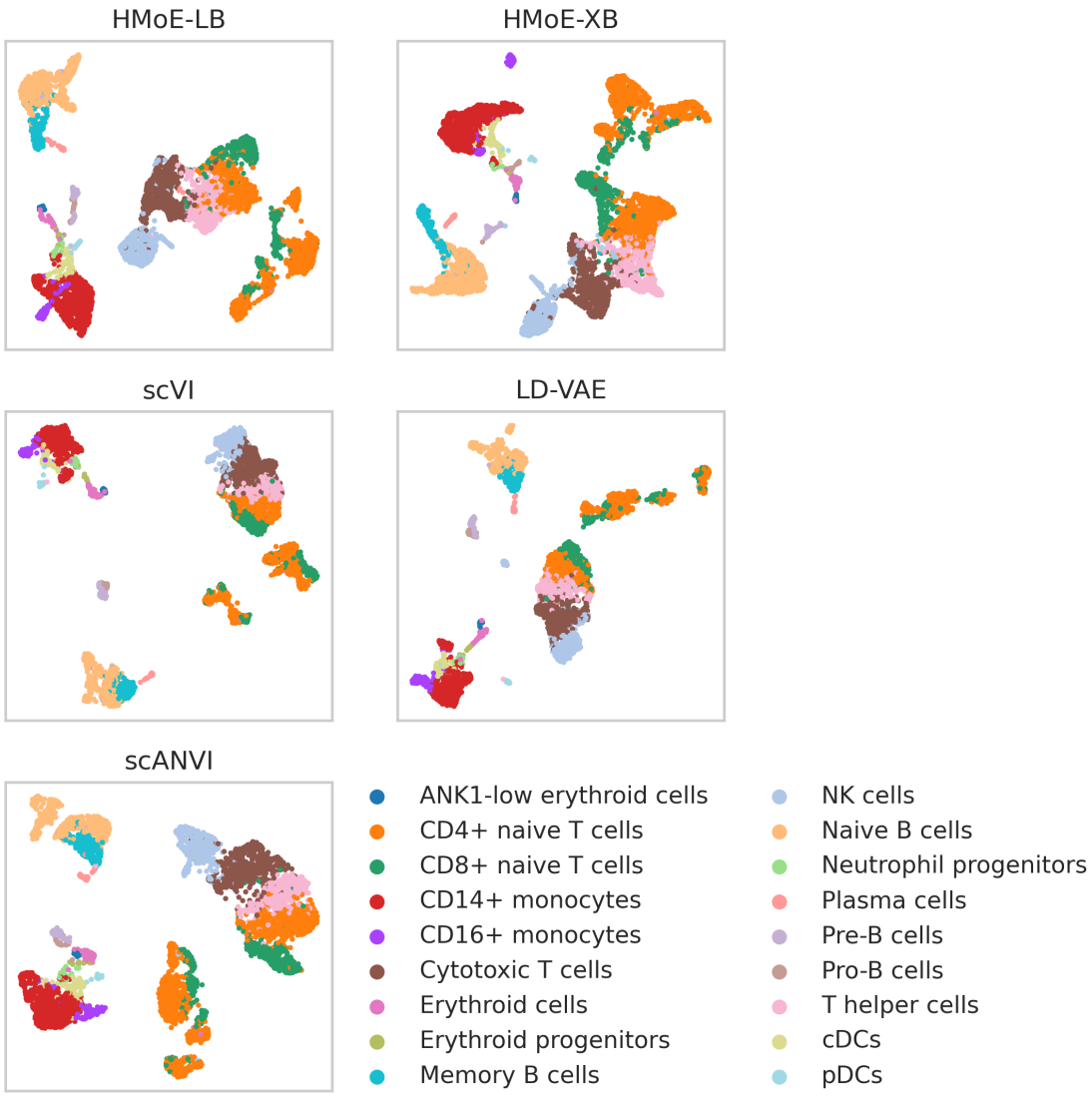

Figure 4.10: Hematopoietic Immune Cell Atlas UMAP projection of latent embeddings.

Figure 4.11 shows the latent embeddings of a 5000 randomly sampled COVID-19 immune cells when we applied zero-shot transfer of an HMoE-scETM model trained on HICA as reference. We can observe that the COVID19 cell-type labels (left legend, left row colors) strongly correspond to corresponding HICA cell-type labels, where column colors correspond to cell-type specific experts (right legend). For example B cells in the COVID19 dataset (blue) is predominantly enriched for topics corresponding to Memory B Cells, Naive B cells, and Pre-B cells. Additionally we can observe that the COVID-19 cells cluster together largely by cell-type when clustering via rows.

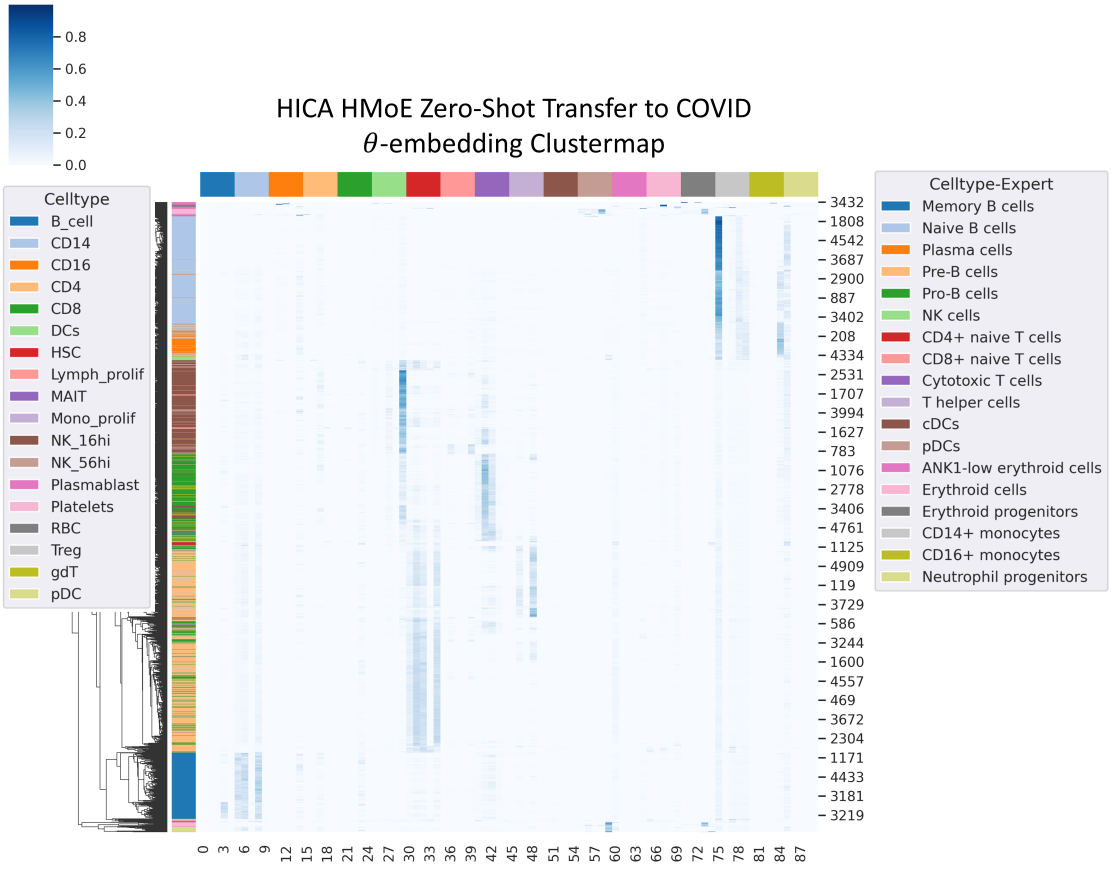

Figure 4.11: Zero-shot transfer θ embedding of COVID19 data, trained on HICA as reference.

4.8 Zero-Shot Transfer Gene Set Enrichment

We applied the HICA reference HMoE-scETM model to COVID-19 data, and observed the theta embedding (A subsampling of 5000 cells is illustrated in Figure 4.11). We then performed permutation tests using the difference in mean theta values over all topics, and sorted for the top differential topics between COVID-19 healthy and COVID-19 severe/critical patients. Topic 31 and Topic 75 were the two most differentially expressed topics. Topic 31 had a higher theta value for healthy patients compared to COVID-19 severe/critical patients, whereas Topic 75 was the inverse. The leading edge plots for the top gene sets for those two topics computed using the *gseapy* package, with the gene set *COVID-19_Related_Gene_Sets_2021* are shown in Figure 4.12. The test-statistic magnitude distribution is shown in the plot on the left. The top right leading edge curve is the top COVID-19 related gene set that is increased in healthy patients for Topic 31, a CD4+ T cell topic. The bottom right leading edge curve is the top COVID-19 related gene set that is increased in COVID-19 severe/critical patients for Topic 75, a CD14+ Monocyte topic.

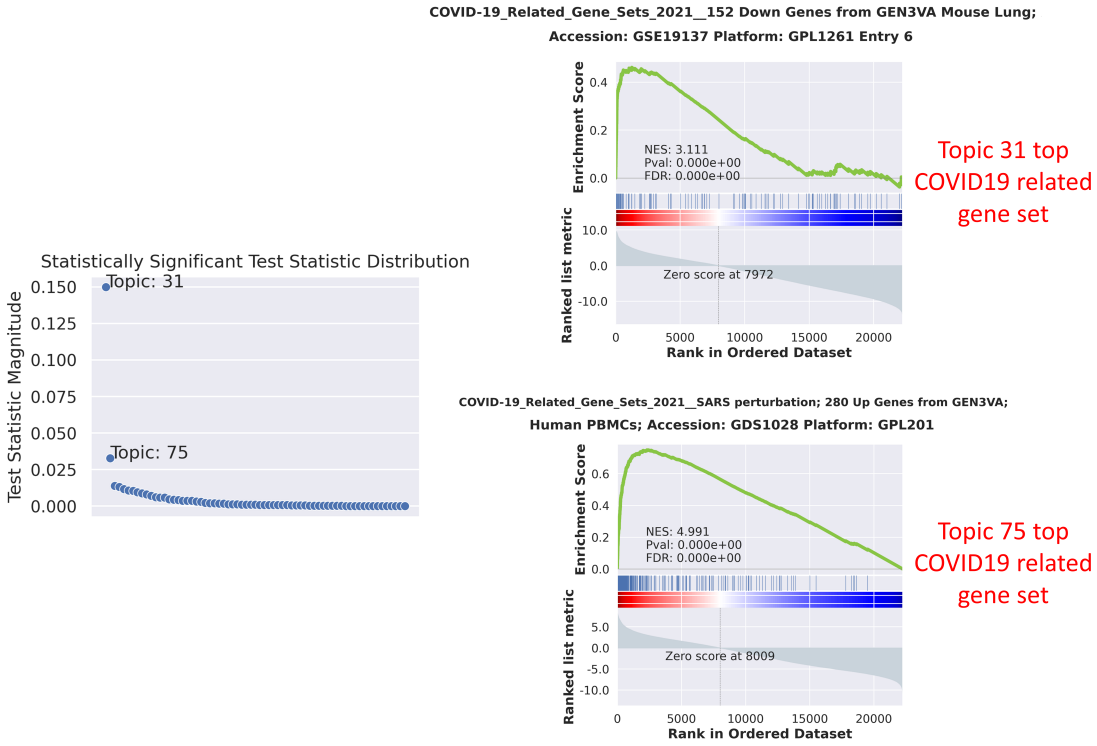

Figure 4.12: Top COVID-19 related gene-sets obtained from top differential topics.

Chapter 5

Discussion

The purpose of this thesis was to explore the utility of incorporating cell-type information to extend scETM. This process was inspired during contributions to the GTM-decon project, with the general idea being: to capture the cell-type specific topics guided by the cell-type prior as seen in GTM-decon. In a sense, MoE-scETM can be seen as a deep learning analogue to GTM-decon, where-by training specific experts for different cell-types, and then combining the experts into the unified MoE-scETM model, the topics associated with a specific expert have been guided - that is to say, those topics are guided to a known cell-type.

5.1 Clustering

In comparing the clustering results in Figures 4.1, 4.2 we observe that the ARI and NMI correlate strongly together, but don't necessarily increase together with ASW. However a strength of ASW as a clustering metric is that it is independent of clustering algorithm, whereas ARI and NMI both require a clustering algorithm, such as Leiden [68] or Louvain [6]. Additionally, this requires the clustering method itself to be fine-tuned for additional hyperparameters. ASW however, only relies on distances computed on the latent embedding space, and ground truth cell-type annotations that are also already required

by ARI and NMI. In terms of ARI and NMI, HMoE is competitive with scVI, LDVAE, and scANVI (Figure 4.1, 4.2). The HMoE models outperformed benchmark methods in terms of ARI, NMI, for the PBMC and HICA test datasets, and performed comparably on the MDD and AD datasets. In terms of ASW, HMoE models outperformed scVI, LDVAE, and scANVI in all tasks.

In our Ablation study, we find that for the MDD test set, the HMoE and MoE models result in no increases in ARI or NMI but result in a fairly significant increase in ASW. For the PBMC immune cells, HMoE and HMoE models both outperform scETM, and there is some increase in performance going from MoE to HMoE, mostly in terms of ARI and NMI, indicating that for well studied cell-type lineages, hierarchical information can yield improvements in performance. For the Tabula Sapiens dataset, there was no cell-type specific experts, but rather, for each major cell lineage, each consisting of dozens of unique cell-types. In situations where there aren't enough cells for many cell-types to properly train a cell-type specific expert, and for situations where there isn't a well studied cell lineage, there can still be small increases in performance from an scETM model just by subdividing the problem space. However the very poor ASW for both MoE and scETM models for the Tabula Sapiens dataset indicates that there is much overlap between the different cell-type labels. Although we can divide human cell lineages in to five broad categories: epithelial, endothelial, stromal/mesenchymal, immune, and neural [7], it appears necessary to guide topics on a more specific, cell-type level, and that grouping large lineages is not sufficient for more significant performance gains from scETM.

5.2 Cell-Type Annotation

As shown in Figure 4.5, the HMoE-scETM model performs well on cell-type classification for brain cells on the withheld test set for both AD reference and MDD reference. Although there is a slight decrease in performance for the MDD dataset, that may be due to the neuronal cell-type labels not being as distinctive in the MDD dataset, in particular

the ”Mix” cell-type has the highest rate of misclassification. Figure 4.5 also shows that the zero-shot transferred model can yield a high degree of classification accuracy. Even when including cell-type labels that do not overlap between the two datasets, where Pericytes and ”Mix” labeled cells are missing from the MDD and AD datasets respectively, the AD-reference HMoE model is able to achieve 88% accuracy. Furthermore, when removing unaligned cells from the task, the model can achieve an accuracy of 94.5%, comparable to both AD reference and MDD reference models when only evaluating test-set cells. Additionally, we can observe that the unaligned ”Mix” cell-type is a neuronal cell-type label for the MDD dataset, and is mostly predicted as Excitatory and Inhibitory neurons by the AD-reference model.

When evaluating cell-type annotation accuracy for a dataset with a larger number of distinct cell-types, including sub-cell-types as labels (Hematopoietic Immune Cell Atlas, Figure 4.6), classification errors tended to occur in a square grid pattern, mostly making classification errors within sub-cell-types. This is notable as not all misclassifications are equal, where in this biological setting, the distance between classes is not identical, for example B cell subtypes are more similar than they would be to a erythroid cell.

5.3 Bulk Deconvolution

The results in Figure 4.7 show good results when the HMoE-scETM model is applied to purified bulk data using single-cells as reference. In particular, the predicted cell-types for purified bulk appear very accurate. The HMoE model is able to very easily predict between general cell-types, such as the B cells, T cells, Monocytes, and Dendritic Cells. The model does appear to struggle to a degree in separating CD4+ T cells from the other T cell types, but still does distinctly predict higher weights for the CD4+ T cell expert compared to the other labeled T cell purified bulk samples. Also notable is the Basophil and Neutrophil cell-types in the purified bulk data don’t have exact matching cells in the reference. These are largely predicted to be Monocytes. Basophils and Neutrophils are

types of granulocytes, which are closely related to Monocytes [9, 17]. This indicates that the model can, to a degree, accommodate for cell-types that are missing in the reference data.

The zero-shot transfer deconvolution of heterogenous bulk data shows that the model has potential (Figure 4.8), but likely needs to be expanded in order to improve performance. Our model is able to achieve a modest mean Spearman R of 0.617 on the ROSMAP bulk brain data, placing it slightly below the mean performance of models designed specifically for bulk deconvolution we evaluated in GTM-decon [66]. Avenues such as model fine-tuning on artificially simulated bulk data as seen in Scaden [43] or fine-tuning directly on the real bulk data may improve deconvolution results.

5.4 Gene Set and Marker Enrichment

Figure 4.9 shows that the top genes for cell-type specific topics are enriched for known marker genes from those cell-types. For example, for Natural Killer cells, Topic 25 has 6 out of the top 10 genes representing known marker genes, from CellmarkerDB [76]. These genes are also supported by the surrounding literature in relation to NK cells, such as FCER1G [11], GZMA [36], and PTPRC [21]. This is notable as the HMoE-scETM model is not provided prior marker information, and is still able to identify cell-type specific marker genes.

When applying zero-shot transfer of the HMoE-scETM model, trained on HICA as reference, transferred to COVID-19 immune cells, we were able to observe differential topics, and for those topics found biologically relevant gene sets. In particular for Topic 31 (Figure 4.12, top right), a topic that is positively enriched in healthy patients compared to COVID-19 severe/critical patients, the top COVID-19 related gene set is for down genes in COVID. Whereas for Topic 75 (Figure 4.12, bottom right), which is higher in COVID-19 severe/critical patients compared to healthy patients, the top COVID-19 related gene set is for genes that are up perturbed in COVID. This indicates that there is potential utility in

investigating gene-sets corresponding to differential topics even when applying a model that was not trained on the specific phenotype or condition of interest.

5.5 Model Limitations

The current HMoE-scETM model requires prior cell-type lineage information.The developmental lineage of immune cells has been well documented, and is supported by a wealth of literature [9, 52]. However not all tissues have the degree of well studied cell-type differentiation lineages. Additionally, this structure imposes a significant requirement of domain knowledge in order to fully leverage the hierarchical nature of the HMoE model.

The current implementation of the HMoE model emphasizes the zero-shot transfer learning task, and doesn't exploit the potential to improve model performance in a transfer learning setting. Currently the framework is limited to fine-tuning model parameters on the same cells used to train the cell-type experts. Additionally the current implementation of the model only permits the fine-tuning of the gating network parameters, as the topics can drift and lose cell-type specificity, resulting in a loss of the annotation and deconvolution applications.

Lastly the current HMoE implementation requires fully labeled cell-type information, and does not support semi-supervised learning, unlike scANVI [75], which allows for a combination of labeled and unlabeled cells for training. This limitation could be addressed by modifications to allow for fine-tuning the model on unlabeled cells, and could potentially simultaneously address the previously mentioned limitations on fine-tuning.

Chapter 6

Conclusion and Future Work

In this thesis, we developed a cell-type guided topic model extension to scETM using a hierarchical mixture of explicitly localized experts. We were able to perform clustering of withheld test cells and additionally able to perform clustering on an entirely different dataset using zero-shot transfer learning. For clustering, our model is competitive with, or outperforms scVI, LDVAE, and scANVI, three existing single-cell modeling techniques. We also show that the mixture of experts formulation allows for the use of the gating network in order to perform accurate cell-type annotation. Additionally we performed exploratory analysis of potential applications of our model to bulk RNA-seq data as a deconvolution tool. Lastly we showed that our model yielded interpretable results, showing that cell-type specific topics are enriched for known cell-type specific marker genes, and that topic specific gene-sets yield meaningful biological pathways when performing gene set enrichment analysis.

Future work includes exploring model fine-tuning when transferred to new data, in both single-cell and bulk applications. In particular, model improvements with deconvolution specifically in mind, such as fine-tuning on simulated bulk or fine-tuning on real bulk data is of interest. The current expert pre-training, and model fine-tuning paradigm only allows for cell-types that have sufficient training examples, and doesn't allow for unlabeled cells. Future work to include extensions that allow for semi-supervised learning

would enable the model to better utilize training datasets and also allow for more flexible fine-tuning. Lastly the current Hierarchical Mixture of Experts model design requires significant prior domain knowledge, whereby future work exploring potential automated hierarchical lineage tree building using unsupervised methods may result in improvements in ease of model use.

Bibliography

[1] ANDRADE-MORAES, C. H., OLIVEIRA-PINTO, A. V., CASTRO-FONSECA, E., DA SILVA, C. G., GUIMARAES, D. M., SZCZUPAK, D., PARENTE-BRUNO, D. R., CARVALHO, L. R., POLICHISO, L., GOMES, B. V., ET AL. Cell number changes in alzheimer's disease relate to dementia, not to plaques and tangles. *Brain 136*, 12 (2013), 3738–3752.

[2] BARON, M., VERES, A., WOLOCK, S. L., FAUST, A. L., GAUJOUX, R., VETERE, A., RYU, J. H., WAGNER, B. K., SHEN-ORR, S. S., KLEIN, A. M., ET AL. A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure. *Cell systems 3*, 4 (2016), 346–360.

[3] BARRETT, T., WILHITE, S. E., LEDOUX, P., EVANGELISTA, C., KIM, I. F., TOMASHEVSKY, M., MARSHALL, K. A., PHILLIPPY, K. H., SHERMAN, P. M., HOLKO, M., ET AL. Ncbi geo: archive for functional genomics data sets—update. *Nucleic acids research 41*, D1 (2012), D991–D995.

[4] BENÍTEZ, J. M., CASTRO, J. L., AND REQUENA, I. Are artificial neural networks black boxes? *IEEE Transactions on neural networks 8*, 5 (1997), 1156–1164.

[5] BLEI, D. M., NG, A. Y., AND JORDAN, M. I. Latent dirichlet allocation. *Journal of machine Learning research 3*, Jan (2003), 993–1022.

[6] BLONDEL, V. D., GUILLAUME, J.-L., LAMBIOTTE, R., AND LEFEBVRE, E. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment 2008*, 10 (2008), P10008.

[7] BRESCHI, A., MUÑOZ-AGUIRRE, M., WUCHER, V., DAVIS, C. A., GARRIDO-MARTÍN, D., DJEBALI, S., GILLIS, J., PERVOUCHINE, D. D., VLASOVA, A., DOBIN, A., ET AL. A limited set of transcriptional programs define major cell types. *Genome research 30*, 7 (2020), 1047–1059.

[8] BUTLER, A., HOFFMAN, P., SMIBERT, P., PAPALEXI, E., AND SATIJA, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology 36*, 5 (2018), 411–420.

[9] CHEN, J., RÉNIA, L., AND GINHOUX, F. Constructing cell lineages from single-cell transcriptomes. *Molecular aspects of medicine 59* (2018), 95–113.

[10] CONSORTIUM*, T. S., JONES, R. C., KARKANIAS, J., KRASNOW, M. A., PISCO, A. O., QUAKE, S. R., SALZMAN, J., YOSEF, N., BULTHAUP, B., BROWN, P., ET AL. The tabula sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science 376*, 6594 (2022), eabl4896.

[11] CRINIER, A., DUMAS, P.-Y., ESCALIÈRE, B., PIPEROGLOU, C., GIL, L., VILLACRECES, A., VÉLY, F., IVANOVIC, Z., MILPIED, P., NARNI-MANCINELLI, É., ET AL. Single-cell profiling reveals the trajectories of natural killer cell differentiation in bone marrow and a stress signature induced by acute myeloid leukemia. *Cellular & molecular immunology 18*, 5 (2021), 1290–1304.

[12] DIENG, A. B., RUIZ, F. J., AND BLEI, D. M. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics 8* (2020), 439–453.

[13] DING, J., ADICONIS, X., SIMMONS, S. K., KOWALCZYK, M. S., HESSION, C. C., MARJANOVIC, N. D., HUGHES, T. K., WADSWORTH, M. H., BURKS, T., NGUYEN,

L. T., ET AL. Systematic comparison of single-cell and single-nucleus rna-sequencing methods. *Nature biotechnology 38*, 6 (2020), 737–746.

[14] FANG, P., LI, X., DAI, J., COLE, L., CAMACHO, J. A., ZHANG, Y., JI, Y., WANG, J., YANG, X.-F., AND WANG, H. Immune cell subset differentiation and tissue inflammation. *Journal of hematology & oncology 11*, 1 (2018), 1–22.

[15] FANG, Z., LIU, X., AND PELTZ, G. Gseapy: a comprehensive package for performing gene set enrichment analysis in python. *Bioinformatics 39*, 1 (2023), btac757.

[16] GAYOSO, A., LOPEZ, R., XING, G., BOYEAU, P., VALIOLLAH POUR AMIRI, V., HONG, J., WU, K., JAYASURIYA, M., MEHLMAN, E., LANGEVIN, M., ET AL. A python library for probabilistic analysis of single-cell omics data. *Nature biotechnology 40*, 2 (2022), 163–166.

[17] GEERING, B., STOECKLE, C., CONUS, S., AND SIMON, H.-U. Living and dying for inflammation: neutrophils, eosinophils, basophils. *Trends in immunology 34*, 8 (2013), 398–409.

[18] GRIFFITHS, T., JORDAN, M., TENENBAUM, J., AND BLEI, D. Hierarchical topic models and the nested chinese restaurant process. *Advances in neural information processing systems 16* (2003).

[19] HAN, X., ZHOU, Z., FEI, L., SUN, H., WANG, R., CHEN, Y., CHEN, H., WANG, J., TANG, H., GE, W., ET AL. Construction of a human cell landscape at single-cell level. *Nature 581*, 7808 (2020), 303–309.

[20] HAQUE, A., ENGEL, J., TEICHMANN, S. A., AND LÖNNBERG, T. A practical guide to single-cell rna-sequencing for biomedical research and clinical applications. *Genome medicine 9*, 1 (2017), 1–12.

[21] HESSLEIN, D. G., PALACIOS, E. H., SUN, J. C., BEILKE, J. N., WATSON, S. R., WEISS, A., AND LANIER, L. L. Differential requirements for cd45 in nk-cell function

reveal distinct roles for syk-family kinases. *Blood, The Journal of the American Society of Hematology 117*, 11 (2011), 3087–3095.

[22] HIE, B., BRYSON, B., AND BERGER, B. Efficient integration of heterogeneous single-cell transcriptomes using scanorama. *Nature biotechnology 37*, 6 (2019), 685–691.

[23] JACOBS, R. A., JORDAN, M. I., NOWLAN, S. J., AND HINTON, G. E. Adaptive mixtures of local experts. *Neural computation 3*, 1 (1991), 79–87.

[24] JANEWAY JR, C., TRAVERS, P., WALPORT, M., AND SHLOMCHIK, M. *Immunobiology: 5th Edition*. Garland Science, 2001.

[25] JORDAN, M. I., AND JACOBS, R. A. Hierarchical mixtures of experts and the em algorithm. *Neural computation 6*, 2 (1994), 181–214.

[26] KINGMA, D. P., AND BA, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[27] KINGMA, D. P., AND WELLING, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[28] KOPF, A., FORTUIN, V., SOMNATH, V. R., AND CLAASSEN, M. Mixture-of-experts variational autoencoder for clustering and generating from similarity-based representations on single cell data. *PLoS computational biology 17*, 6 (2021), e1009086.

[29] KORSUNSKY, I., MILLARD, N., FAN, J., SLOWIKOWSKI, K., ZHANG, F., WEI, K., BAGLAENKO, Y., BRENNER, M., LOH, P.-R., AND RAYCHAUDHURI, S. Fast, sensitive and accurate integration of single-cell data with harmony. *Nature methods 16*, 12 (2019), 1289–1296.

[30] KUKSIN, M., MOREL, D., AGLAVE, M., DANLOS, F.-X., MARABELLE, A., ZINOVYEV, A., GAUTHERET, D., AND VERLINGUE, L. Applications of single-cell and bulk rna sequencing in onco-immunology. *European journal of cancer 149* (2021), 193–210.

[31] KUKURBA, K. R., AND MONTGOMERY, S. B. Rna sequencing and analysis. *Cold Spring Harbor Protocols 2015*, 11 (2015), pdb–top084970.

[32] LAKE, B. B., CHEN, S., SOS, B. C., FAN, J., KAESER, G. E., YUNG, Y. C., DUONG, T. E., GAO, D., CHUN, J., KHARCHENKO, P. V., ET AL. Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nature biotechnology 36*, 1 (2018), 70–80.

[33] LI, C., CHEN, S., XING, J., SUN, A., AND MA, Z. Seed-guided topic model for document filtering and classification. *ACM Transactions on Information Systems (TOIS) 37*, 1 (2018), 1–37.

[34] LI, W., AND MCCALLUM, A. Pachinko allocation: Dag-structured mixture models of topic correlations. In *Proceedings of the 23rd international conference on Machine learning* (2006), pp. 577–584.

[35] LI, X., SUN, H., LI, H., LI, D., CAI, Z., XU, J., AND MA, R. A single-cell rna-sequencing analysis of distinct subsets of synovial macrophages in rheumatoid arthritis. *DNA and Cell Biology 42*, 4 (2023), 212–222.

[36] LIEBERMAN, J. Granzyme a activates another way to die. *Immunological reviews 235*, 1 (2010), 93–104.

[37] LIU, L., TANG, L., DONG, W., YAO, S., AND ZHOU, W. An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus 5*, 1 (2016), 1–22.

[38] LOPEZ, R., REGIER, J., COLE, M. B., JORDAN, M. I., AND YOSEF, N. Deep generative modeling for single-cell transcriptomics. *Nature methods 15*, 12 (2018), 1053–1058.

[39] LOWE, R., SHIRLEY, N., BLEACKLEY, M., DOLAN, S., AND SHAFEE, T. Transcriptomics technologies. *PLoS computational biology 13*, 5 (2017), e1005457.

[40] LUECKEN, M. D., BÜTTNER, M., CHAICHOOMPU, K., DANESE, A., INTERLANDI, M., MÜLLER, M. F., STROBL, D. C., ZAPPIA, L., DUGAS, M., COLOMÉ-TATCHÉ, M., ET AL. Benchmarking atlas-level data integration in single-cell genomics. *Nature methods 19*, 1 (2022), 41–50.

[41] MASOUDNIA, S., AND EBRAHIMPOUR, R. Mixture of experts: a literature survey. *Artificial Intelligence Review 42* (2014), 275–293.

[42] MATHYS, H., DAVILA-VELDERRAIN, J., PENG, Z., GAO, F., MOHAMMADI, S., YOUNG, J. Z., MENON, M., HE, L., ABDURROB, F., JIANG, X., ET AL. Single-cell transcriptomic analysis of alzheimer's disease. *Nature 570*, 7761 (2019), 332–337.

[43] MENDEN, K., MAROUF, M., OLLER, S., DALMIA, A., MAGRUDER, D. S., KLOIBER, K., HEUTINK, P., AND BONN, S. Deep learning–based cell composition analysis from tissue expression profiles. *Science advances 6*, 30 (2020), eaba2619.

[44] MINOURA, K., ABE, K., NAM, H., NISHIKAWA, H., AND SHIMAMURA, T. A mixture-of-experts deep generative model for integrated analysis of single-cell multiomics data. *Cell reports methods 1*, 5 (2021).

[45] MONACO, G., LEE, B., XU, W., MUSTAFAH, S., HWANG, Y. Y., CARRÉ, C., BURDIN, N., VISAN, L., CECCARELLI, M., POIDINGER, M., ET AL. Rna-seq signatures normalized by mrna abundance allow absolute deconvolution of human immune cell types. *Cell reports 26*, 6 (2019), 1627–1640.

[46] NAGY, C., MAITRA, M., TANTI, A., SUDERMAN, M., THÉROUX, J.-F., DAVOLI, M. A., PERLMAN, K., YERKO, V., WANG, Y. C., TRIPATHY, S. J., ET AL. Single-nucleus transcriptomics of the prefrontal cortex in major depressive disorder implicates oligodendrocyte precursor cells and excitatory neurons. *Nature neuroscience 23*, 6 (2020), 771–781.

[47] NAVIN, N. E. The first five years of single-cell cancer genomics and beyond. *Genome research 25*, 10 (2015), 1499–1507.

[48] NEWMAN, A. M., LIU, C. L., GREEN, M. R., GENTLES, A. J., FENG, W., XU, Y., HOANG, C. D., DIEHN, M., AND ALIZADEH, A. A. Robust enumeration of cell subsets from tissue expression profiles. *Nature methods 12*, 5 (2015), 453–457.

[49] NEWMAN, A. M., STEEN, C. B., LIU, C. L., GENTLES, A. J., CHAUDHURI, A. A., SCHERER, F., KHODADOUST, M. S., ESFAHANI, M. S., LUCA, B. A., STEINER, D., ET AL. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nature biotechnology 37*, 7 (2019), 773–782.

[50] NGUYEN, V.-A., YING, J. L., AND RESNIK, P. Lexical and hierarchical topic regression. *Advances in neural information processing systems 26* (2013).

[51] PATRICK, E., TAGA, M., ERGUN, A., NG, B., CASAZZA, W., CIMPEAN, M., YUNG, C., SCHNEIDER, J. A., BENNETT, D. A., GAITERI, C., ET AL. Deconvolving the contributions of cell-type heterogeneity on cortical gene expression. *PLOS Computational Biology 16*, 8 (2020), e1008120.

[52] PAUL, F., ARKIN, Y., GILADI, A., JAITIN, D. A., KENIGSBERG, E., KEREN-SHAUL, H., WINTER, D., LARA-ASTIASO, D., GURY, M., WEINER, A., ET AL. Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell 163*, 7 (2015), 1663–1677.

[53] PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., MICHEL, V., THIRION, B., GRISEL, O., BLONDEL, M., PRETTENHOFER, P., WEISS, R., DUBOURG, V., VANDERPLAS, J., PASSOS, A., COURNAPEAU, D., BRUCHER, M., PERROT, M., AND DUCHESNAY, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 12* (2011), 2825–2830.

[54] PETINOT, Y., MCKEOWN, K., AND THADANI, K. A hierarchical model of web summaries. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (2011), pp. 670–675.

[55] PISCO, A., AND CONSORTIUM, T. S. Tabula sapiens single-cell dataset, Apr 2021.

[56] REGEV, A., LI, B., KOWALCZYK, M., DIONNE, D., TICKLE, T., LEE, J., ROZENBLATT-ROSEN, O., ASHENBERG, O., TABAKA, M., SHEKAHAR, K., SLYPER, M., AND WALDMAN, J. The census of immune cells, 2020.

[57] REGEV, A., TEICHMANN, S. A., LANDER, E. S., AMIT, I., BENOIST, C., BIRNEY, E., BODENMILLER, B., CAMPBELL, P., CARNINCI, P., CLATWORTHY, M., ET AL. The human cell atlas. *elife 6* (2017), e27041.

[58] REN, X., WEN, W., FAN, X., HOU, W., SU, B., CAI, P., LI, J., LIU, Y., TANG, F., ZHANG, F., ET AL. Covid-19 immune features revealed by a large-scale single-cell transcriptome atlas. *Cell 184*, 7 (2021), 1895–1913.

[59] ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics 20* (1987), 53–65.

[60] SASAKI, H., SAISHO, Y., INAISHI, J., WATANABE, Y., TSUCHIYA, T., MAKIO, M., SATO, M., NISHIKAWA, M., KITAGO, M., YAMADA, T., ET AL. Reduced beta cell number rather than size is a major contributor to beta cell loss in type 2 diabetes. *Diabetologia 64*, 8 (2021), 1816–1821.

[61] SONG, Z., HU, Y., VERMA, A., BUCKERIDGE, D. L., AND LI, Y. Automatic phenotyping by a seed-guided topic model. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (2022), pp. 4713–4723.

[62] STEPHENSON, E., REYNOLDS, G., BOTTING, R. A., CALERO-NIETO, F. J., MORGAN, M., TUONG, Z. K., BACH, K., SUNGNAK, W., WORLOCK, K. B., YOSHIDA, M., ET AL. The cellular immune response to covid-19 deciphered by single cell multiomics across three uk centres. *medRxiv* (2021), 2021–01.

[63] STEUBER, F., SCHNEIDER, S., AND SCHOENFELD, M. Embedding semantic anchors to guide topic models on short text corpora. *Big Data Research 27* (2022), 100293.

[64] STOECKIUS, M., HAFEMEISTER, C., STEPHENSON, W., HOUCK-LOOMIS, B., CHATTOPADHYAY, P. K., SWERDLOW, H., SATIJA, R., AND SMIBERT, P. Simultaneous epitope and transcriptome measurement in single cells. *Nature methods 14*, 9 (2017), 865–868.

[65] SVENSSON, V., GAYOSO, A., YOSEF, N., AND PACHTER, L. Interpretable factor models of single-cell rna-seq via variational autoencoders. *Bioinformatics 36*, 11 (2020), 3418–3421.

[66] SWAPNA, L. S., HUANG, M., AND LI, Y. Guided-topic modelling of single-cell transcriptomes enables sub-cell-type and disease-subtype deconvolution of bulk transcriptomes. *bioRxiv* (2023).

[67] TANG, F., BARBACIORU, C., WANG, Y., NORDMAN, E., LEE, C., XU, N., WANG, X., BODEAU, J., TUCH, B. B., SIDDIQUI, A., ET AL. mrna-seq whole-transcriptome analysis of a single cell. *Nature methods 6*, 5 (2009), 377–382.

[68] TRAAG, V. A., WALTMAN, L., AND VAN ECK, N. J. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports 9*, 1 (2019), 5233.

[69] VAYANSKY, I., AND KUMAR, S. A. A review of topic modeling methods. *Information Systems 94* (2020), 101582.

[70] VIRTANEN, P., GOMMERS, R., OLIPHANT, T. E., HABERLAND, M., REDDY, T., COURNAPEAU, D., BUROVSKI, E., PETERSON, P., WECKESSER, W., BRIGHT, J., ET AL. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods 17*, 3 (2020), 261–272.

[71] WANG, X., PARK, J., SUSZTAK, K., ZHANG, N. R., AND LI, M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nature communications 10*, 1 (2019), 380.

[72] WOLF, F. A., ANGERER, P., AND THEIS, F. J. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology 19* (2018), 1–5.

[73] XIE, X., LIU, M., ZHANG, Y., WANG, B., ZHU, C., WANG, C., LI, Q., HUO, Y., GUO, J., XU, C., ET AL. Single-cell transcriptomic landscape of human blood cells. *National Science Review 8*, 3 (2021), nwaa180.

[74] XIE, Z., BAILEY, A., KULESHOV, M. V., CLARKE, D. J., EVANGELISTA, J. E., JENKINS, S. L., LACHMANN, A., WOJCIECHOWICZ, M. L., KROPIWNICKI, E., JAGODNIK, K. M., ET AL. Gene set knowledge discovery with enrichr. *Current protocols 1*, 3 (2021), e90.

[75] XU, C., LOPEZ, R., MEHLMAN, E., REGIER, J., JORDAN, M. I., AND YOSEF, N. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Molecular systems biology 17*, 1 (2021), e9620.

[76] ZHANG, X., LAN, Y., XU, J., QUAN, F., ZHAO, E., DENG, C., LUO, T., XU, L., LIAO, G., YAN, M., ET AL. Cellmarker: a manually curated resource of cell markers in human and mouse. *Nucleic acids research 47*, D1 (2019), D721–D728.

[77] ZHANG, Z., WIENCKE, J. K., KELSEY, K. T., KOESTLER, D. C., MOLINARO, A. M., PIKE, S. C., KARRA, P., CHRISTENSEN, B. C., AND SALAS, L. A. Hierarchical deconvolution for extensive cell type resolution in the human brain using dna methylation. *Frontiers in Neuroscience 17* (2023), 1198243.

[78] ZHAO, Y., CAI, H., ZHANG, Z., TANG, J., AND LI, Y. Learning interpretable cellular and gene signature embeddings from single-cell transcriptomic data. *Nature communications 12*, 1 (2021), 5261.