# Single-cell multi-omic topic embedding reveals cell-type-specific and COVID-19 severity-related immune signatures

Manqi Zhou[1,†], Hao Zhang[2,†], Zilong Bai[2], Dylan Mann-Krzisnik[3], Fei Wang[2,*], and Yue Li[3,4,5,*]

[1] *Department of Computational Biology, Cornell University*
[2] *Division of Health Informatics, Department of Population Health Sciences, Weill Cornell Medicine*
[3] *Quantitative Life Science, McGill University*
[4] *School of Computer Science, McGill University*
[5] *Mila - Quebec AI Institute*
[*] *Correspondence to few2001@med.cornell.edu, yueli@cs.mcgill.ca*
[†] *Equal contribution*

## Abstract

The advent of single-cell multi-omics sequencing technology makes it possible for researchers to leverage multiple modalities for individual cells and explore cell heterogeneity. However, the high dimensional, discrete, and sparse nature of the data make the downstream analysis particularly challenging. Most of the existing computational methods for single-cell data analysis are either limited to single modality or lack flexibility and interpretability. In this study, we propose an interpretable deep learning method called multi-omic embedded topic model (moETM) to effectively perform integrative analysis of high-dimensional single-cell multimodal data. moETM integrates multiple omics data via a product-of-experts in the encoder for efficient variational inference and then employs multiple linear decoders to learn the multi-omic signatures of the gene regulatory programs. Through comprehensive experiments on public single-cell transcriptome and chromatin accessibility data (i.e., scRNA+scATAC), as well as scRNA and proteomic data (i.e., CITE-seq), moETM demonstrates superior performance compared with six state-of-the-art single-cell data analysis methods on seven publicly available datasets. By applying moETM to the scRNA+scATAC data in human peripheral blood mononuclear cells (PBMCs), we identified sequence motifs corresponding to the transcription factors that regulate immune gene signatures. Applying moETM analysis to CITE-seq data from the COVID-19 patients revealed not only known immune cell-type-specific signatures but also composite multi-omic biomarkers of critical conditions due to COVID-19, thus providing insights from both biological and clinical perspectives.

# 1 Introduction

Multi-omic single-cell high-throughput sequencing technologies opens up new opportunities to interrogate cell-type-specific gene regulatory programs. Single-cell RNA sequencing combined with Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) [1] simultaneously measure the transcriptome and chromatin accessibility in the same cell. CITE-seq [2] measures surface protein and transcriptome data using oligonucleotide-labeled antibodies. By integrating the information from these multiple omics, we can expand our understanding of the genome regulation from multiple perspectives.

However, extracting meaningful biological patterns from the fast-growing multi-omic single-cell data remains a challenge due to several factors [3, 4]. Firstly, multi-omic single-cell technologies are still in the early stages. The cell yield is lower compared to the single-omic technologies such as scRNA-seq. On the other hand, the combined feature dimension is much higher (e.g., genes and peaks). This requires a more deliberate model design that can flexibly distill meaningful cell-type signatures from the multi-modal data while not overfitting the data. Secondly, multi-omic single-cell data are noisier compared with bulk-level or single-cell single-omic data. This calls for a probabilistic model that can infer latent cell types while properly accounting for the statistical uncertainty. Thirdly, the batch effects make it challenging to distinguish biological signals from study-specific confounders. Lastly, multi-omic single-cell are more costly compared to scRNA-seq or scATAC-seq alone. It is therefore highly cost-effective if we can profile single-omic data (e.g., transcriptome) and then predict the unobserved omic (e.g., chromatin accessibility or proteome). Nonetheless, the prediction from one modality to another is a challenging task, particularly from low dimension to high dimension.

Recently, several computational methods were developed to tackle the above multi-modality data integration challenges encountered in multi-omic single-cell data analysis. For instance, SMILE [5] integrates multi-omic data by minimizing the mutual information of the latent representations among the modalities and batches. The totalVI [6] and multiVI [7] integrate CITE-seq data and scRNA+scATAC data via variational autoencoder (VAE) frameworks, respectively. Cobolt [8] is a hierarchical Bayesian generative model to integrate cell modalities. scMM [9] is a mixture-of-experts (MoE) model developed to impute one missing modality conditioned on the other. Multigrate [9] adopted a product-of-experts (PoE) framework to integrate multi-omic data. MOFA+ [10] uses mean-field variational Bayes and coordinate ascent to fit a Bayesian Group Factor Analysis model to integrate the multi-omic data. Seurat V4 [11] integrated multimodal single-cell data through the weighted nearest neighbor algorithm. While many of these methods conferred promising performances in some of the tasks such as cell clustering or modality imputation, they often need to compromise scalability, interpretability, and/or flexibility. In particular, when a neural network is used to encode the high-dimensional multi-omic data, interpretability is traded for flexibility; when a linear model or independent feature assumption is made, flexibility is traded for interpretability and scalability. However, all three are important to reveal cell-type-specific multi-omic signatures that are indicative of gene regulatory programs from large-scale data. Furthermore, most of these methods are entirely data-driven and therefore incapable of fully utilizing the existing biological information such as gene annotations or pathway information.

In this study, we present Multi-Omics Embedded Topic Model (moETM) to integrate multiple molecular modalities at the single-cell level. As one of the main technical contributions, moETM uses product-of-experts to infer latent topics underlying the single-cell multi-omic data and a set of linear decoders to learn shared embedding of topics and multi-omic features (e.g., genes, chromatin accessibility, and/or protein) that can accurately reconstruct the high-dimensional multi-omic data from their low-dimensional latent topic space (**Fig.** 1a). Using stochastic amortized variational inference, moETM is highly scalable to large multi-omic datasets containing over 40,000 cells from scRNA+scATAC-seq and over 200,000 cells from CITE-seq data. Through effectively integrating multiple modalities from multi-omic single-cell sequencing data, moETM seeks to achieve 3 tasks: (1) clustering cells into biologically meaningful clusters to identify sub-celltype indicative of phenotype of interests (**Fig.** 1b); (2) imputing one omic (e.g., single-cell transcriptome or surface proteome) using the other omic (e.g., chromatin accessibility or single-cell transcriptome) (**Fig.** 1c); (3) identifying cell-type signatures, which serve as biomarkers for a target phenotype (**Fig.** 1d). Through comprehensive experiments on seven single-cell multi-omic datasets, we demonstrate moETM's ability comparatively with six state-of-the-art (SOTA) computational methods. We further showcase how moETM facilitates the analysis of the COVID-19 single-cell CITE-seq dataset. Quantitatively, we observe that moETM learns the joint embeddings from multiple modalities with better or comparable bio-conservation, batch-effect correction, and cross-modality imputation compared with the existing methods [5–11]. Furthermore, the topic embedding learned by moETM enables gaining biological insights into the cell-type-specific mutli-omic regulatory elements.

# 2 Results

## 2.1 moETM model overview

As an overview, moETM integrates multiomics data across different experiments or studies with interpretable latent embeddings (**Fig.** 1). It is built upon the widely used variational autoencoder (VAE) [12] to model multi-modal data (**Fig.** 1a). However, to tailor the VAE framework for the single-cell multi-omic data, we made two main contributions on both the encoder and the decoder of the VAE.

The encoder in moETM is a two-layer fully-connected neural network, which infers topic proportion from multi-omic normalized count vectors for a cell. We assume the latent representation of each omic follows a $K$-dimensional independent logistic Normal distribution. Our goal is to effectively combine these distributions into a joint distribution of the multi-omic data. To this end, we take the product of the $K$-dimensional Gaussians (Product-of-Gaussians). Because the Product-of-Gaussians (PoG) is also a Gaussian density function, we can represent the joint latent distribution in closed-form. In principle, this results in a tighter ELBO and therefore more efficient variational inference compared to the mixture of experts (MoE) approaches [13] as adopted in MultiVI/TotalVI [6, 7] and scMM [9]. In particular, these MoE approaches sample $K$-dimensional Gaussian variables for each omic and then take their average. In contrast, our PoG formalism requires sampling only once from the joint Gaussian. Therefore, moETM may confer more robust estimates thanks to the decreased variance in the stochastic variational inference. To obtain interpretable cell embedding, we perform a softmax transformation on the joint Gaussian density. The resulting logistic Normal distribution can be considered as a topic mixture membership for the cell. These topics can be directly mapped to known cell types based on their top gene signatures detected from our linear-decoder (as described below). Because the topic distribution must sum to 1 over the $K$ topics, the inferred topic mixture membership of a cell express statistical uncertainty in the cell embedding.

On the decoder side, inherited from our earlier work [14], moETM employs a linear matrix factorization to reconstruct the normalized count vectors from the cell embedding. Specifically, moETM factorized cell-by-feature matrices into a shared cell-by-topic matrix $\mathbf{\Theta}$, a shared topic-embedding matrix $\boldsymbol{\alpha}$, and $M$ separate feature-embedding matrices $\boldsymbol{\rho}^{(m)}$, where $m \in \{1, \ldots, M\}$ indexes the omics. Since different omics share the same cell-by-topics matrix but had their own feature-embedding matrices, we can explore the relations among cells, topics, and features in a highly interpretable way. This departs from the existing VAE models such as scMM [9], BABEL [15], and Multigrate [16] that used a neural network as the decoder. Another confound in single-cell data are batch effects, which are sources of technical variation. To account for those, we introduced the omic-specific batch-removal factors $\boldsymbol{\lambda}^{(m)} \in \mathbb{R}^{V^{(m)} \times S}$ for each omic $m$ (**Fig.** 1a), which acts as a linear-additive batch-specific bias in reconstructing each modality. By regressing out the batch effects via $\boldsymbol{\lambda}^{(\cdot)}$, moETM can learn biologically meaningful representation in terms of the cell topic mixture or embedding. As detailed in **Methods**, all the parameters in moETM are learned end-to-end by maximizing a common objective function defined as the evidence lower bound (ELBO) of the marginal data likelihood under the framework of amortized variational inference.

## 2.2 Multiomics integration

We performed quantitative evaluations of moETM on the integrated low-dimensional representation comparing with six state-of-the-art multiomics integration methods (SMILE [5], scMM [9], Cobolt [8], MultiVI/TotalVI [6, 7], MOFA+ [10], Seurat V4 [11]) on seven published datasets. Four out of the seven datasets are single-cell transcriptome and chromatin accessibility (gene+peak) datasets and the other 3 are single-cell transcriptome and surface protein expression (gene+protein) datasets measured by Cellular Indexing of Transcriptomes and Epitopes by sequencing (CITE-seq).

The performance of the multiomics integrative task were based on both biological conservation metrics and batch removal metrics (**Methods**). For the biological conservation score, we adopted the common metrics Adjusted Rand Index (ARI) [17] and Normalized Mutual Information (NMI [18]). For evaluating batch-effect removal, we used k-nearest-neighbor batch-effect test (kBET) [19] with graph connectivity (GC).

To make a comprehensive comparison, we used three experimental settings: *i*) 60/40 random split for training and testing with 500 repeats (**Table** 1 and 2, **Supplementary Table** S1 and S2); *ii*) training and testing both on the whole dataset (**Supplementary Table** S3); *iii*) training and testing across different batches (**Supplementary Table** S4 and S5). The number of topics was set to 100 during training based on the robust performance (**Supplementary Fig.** S1). Overall, we obtained consistent results across all 3 settings and therefore chose to focus on describing the results based on the first setting.

We observed that moETM achieved the best overall performance when averaging over all datasets' performance scores among 3 out of 4 evaluation metrics. Similarly, when averaging across gene+protein datasets only, moETM achieved the best overall performance among 3 out of 4 evaluation metrics (**Fig.** 2 middle panel). In particular, moETM conferred the highest averaged ARI, NMI, and GC when either averaging over all datasets or averaging over gene+protein datasets specifically, and the second highest averaged kBET only marginally behind multiVI/totalVI, which might have over-corrected the batch effect at the expense of biological conservation. When averaging over gene+peak datasets, moETM can still achieve the best among 2 out of 4 evaluation metrics. Specifically, moETM ranked the second highest on ARI and NMI and slightly behind Seurat V4, which has a larger standard deviation compared with moETM.

For individual datasets, moETM is either the best or the second best method on 6 out of 7 datasets (except MBC) for different experimental settings in terms of the ARI (**Fig.** 2a,**Table** 1 and 2). One possible reason could be that the sample size of MBC (3293 cells) from which moETM learns high-dimensional peak embeddings is small compared with the other 6 datasets. To assess the benefits of the added features in moETM, we compared moETM with its ablated versions: moETM_rna, moETM_atac, and moETM_protein, where moETM was trained on a single modality. As expected, the performance of moETM on single modality decreased, indicating that moETM could improve its performance by leveraging multiple modalities (**Fig.** 2 right panel).

Similar qualitative conclusions can be drawn based on NMI (**Fig.** 2b, **Supplementary Table** S1, and **Supplementary Table** S2). For kBET (**Fig.** 2c), moETM is the best for the BMMC1 dataset and the second best on BMMC2, HWBC, and HBIC datasets – slightly behind MultiVI/TotalVI. Therefore, while moETM conferred higher biological conservation scores in terms of ARI and NMI, it still maintains a comparable kBET scores on all four datasets compared to MultiVI/TotalVI. Indeed, we observed an excellent balance between the biological conservation and batch effect removal because moETM achieved notably higher GC compared to all methods (**Fig.** 2d). This is because GC is the only metric that is based on both the cell types and batch labels by measuring the similarity among cells of the same type from different batches based on the embedding learned by each method [20].

We postulated that the main reason for the moETM's superior integration performance is it's the Product of Gaussians (PoG) formulation. To that end, we constructed moETM_avg, which replaced PoG with averaging of sampled variables from individual Gaussian distributions similar to the existing VAE models like scMM [9]. As expected, the performance of moETM_avg was worse than moETM in all datasets in terms of both bio-conservation and batch removal evaluation metrics (**Fig.** 2 right panel). Furthermore, since scMM also adopted the average of Gaussian in the encoder, the fact that moETM_avg outperformed scMM indicates the benefits of using the linear decoder, which further improves the multi-omic integration while correcting batch effects across all cells.

We further verified the clustering performance by visualizing cell embeddings using Uniform Manifold Approximation and Projection (UMAP) [21] (**Fig.** 3). Indeed, not only did moETM remove batch effects but also revealed a better representation of cell type clusters. For example, “Plasmablast IGKC-" cells were grouped closely by moETM but were clustered into multiple small parts by SMILE (**Fig.** 3a). Moreover, plasmablast cells from different batches were also mixed better by moETM compared with SMILE, which indicated a better batch-effects correction. “CD4+ T activated" and “CD4+ naive" cells were closer within the same cluster but clearly distinguishable between themselves. In contrast, these two cell types were mixed together by SMILE and scMM. In modeling the BMMC1 dataset (gene+peak), “B1 B" cells and “naive CD20+ B" cells (**Fig.** 3b) were mixed by other methods while better separated by moETM. Taken together, these results show that moETM is able to distinguish similar cell types by capturing biological information in its encoding space while removing batch effects.

## 2.3 Cross-omic imputation

In the case of gene+protein, moETM accurately imputes surface protein expression from gene expression, achieving average Pearson (Spearman) correlation of 0.95, 0.92, and 0.88 (0.94, 0.90, and 0.85) on random split, leave-one-batch, and leave-one-cell-type imputation experiments, respectively (**Supplementary Table** S6). We visualized the reconstructed protein expression against the observed values using the BMMC2 (gene+protein) dataset (**Fig.** 4a). The imputed protein expression is highly linearly correlated with the observed one (**Fig.** 4b), which is what we expected given the high Pearson correlation of 0.95. The runner up methods - namely, scMM and BABEL - also performed well on this task, both achieving a correlation score of 0.94.

Compared to the surface protein imputation task, imputing gene expression from the open chromatin regions is a more challenging task because of the sparser input scATAC-seq signals and the dynamic and often asynchronous interplay between the chromatin states and the transcriptome [22–24]. Nonetheless, moETM achieved a relatively high Pearson (and Spearman) correlation scores of 0.69, 0.65, and 0.58 (and 0.37, 0.35, and 0.32) on random split, leave-one-batch, and leave-one-cell-type experiments. These are notably higher than the corresponding correlation obtained by BABEL (Pearson: 0.65, 0.60, 0.55; Spearman: 0.34, 0.33, 0.30) and scMM (Pearson: 0.63, 0.61, 0.54; Spearman: 0.33, 0.33, 0.28) (**Supplementary Table** S7). Qualitatively, the imputed and the observed gene expression profiles also exhibit similar pattern and linear relationship (**Fig.** 4c, d).

In the previous two imputation applications, modalities were generated from high dimension to low dimension. The imputation from the low dimension to the high dimension is more difficult but nonetheless feasible. Specifically, on the 3 same experimental designs, the Pearson (and Spearman) correlations between the observed and the imputed open chromatin regions from gene expression are 0.58, 0.55, and 0.51 (and 0.33, 0.30, and 0.28) (**Supplementary Table** S8); the Pearson (and Spearman) correlation are between the observed and imputed gene expression from protein expression are 0.65, 0.63, and 0.60 (and 0.41, 0.39, and 0.37) (**Supplementary Table** S9). In contrast, the runner-up method scMM achieved Pearson (and Spearman) correlations of 0.40, 0.29, and 0.37 (and 0.29, 0.25, and 0.21). for imputing chromatin accessibility from gene expression. For imputing gene expression from surface protein, scMM and BABEL also fell behind moETM in terms of both Pearson and Spearman correlations (**Supplementary Table** S9). Qualitatively, the imputed and the observed peaks and gene expression exhibit consistent patterns (**Supplementary Fig.** S2a, c) and strong linear trends and similar patterns (**Supplementary Fig.** S2b, d).

## 2.4 Correlating topic scores of the RNA transcripts and surface proteins for the same genes

As a proof-of-concept, we sought to assess whether the top surface proteins can be mapped to the top genes under the same topic (i.e., following the central dogma). To this end, we trained a 100-topic moETM on the BMMC2 (gene+protein) dataset, which was taken from CITE-seq and consists of over 90,000 cells. For each topic, we calculated the Spearman correlation of topic scores between the 134 pairs of gene and the corresponding translated surface protein (**Fig.** 5a). The correlations ranged from -0.096 to 0.751 with an average of 0.29. Moreover, 96 of the 100 topics have positive correlations, and among those 13 topics have correlations larger than 0.5. In particular, the correlation in topic 40 was 0.576, and the correlation in topic 44 was 0.628.

## 2.5 Immune cell-type signatures revealed by multi-omic topics learned from CITE-seq data

To identify cell-type signatures, we associated each topic with the specific cell type that exhibit the highest average topic score across cells. Notably, not all topics were uniquely associated with one single cell type and some topics might be enriched for a combination of multiple cell types. Therefore, only the most distinct topics that were enriched for one cell type were chosen as examples for the downstream cell-type markers analysis. For instances, topic 44 was associated with monocytes, which consists of CD14+ and CD16+ Mono; topic 40 was associated with B cells, which consists of primarily Naive CD20+ B IGKC+ and Naive CD20+ B IGKC- cells; topic 83 was associated with natural killer cells. These are visually easy to detect from the topic mixture probabilities among the individual cells (**Fig.** 5b).

Under each topic, many top genes and top proteins are the known cell-type markers (**Fig.** 5c). For example, under topic 40 (i.e., a B-cell topic), the top genes *CR2*, *SSPN*, and *ADAM28* are known marker genes for B cell; the top proteins CD21, CD20, and CD40 are also marker proteins for B cells according to the CellMarker database [25]. For topic 7, one of top proteins CD11c is a marker protein for dendritic-cell [25]. For topic 83, protein CD16, marker for natural killer cells, is among its top proteins [25].

For topic 44, the top gene *S100A9*'s coding protein is a chemotactic factor for monocytes [26] and is highly expressed during inflammatory processes [27]; among the top proteins for topic 44, CD36 [28], CD33, and CD11c [29] are also markers for monocyte sub-cell-types.

Similarly, monocyte is also enriched in topic 23, which shares the top marker protein CD16 with topic 44 but also contains unique top genes such as *CDKN1C* and *FCGR3A*. While *CDKN1C* is a known marker gene for monocyte [30], *FCGR3A* is up-regulated in CD16+ monocytes as supported by the existing literature [31].

Instead of choosing the top genes or protein from each topic, we performed Gene Set Enrichment Analysis (GSEA) [32, 33] using the topic scores for all of the genes and proteins. Since the BMMC2 dataset is about bone marrow mononuclear cells, we queried the C7 ImmuneSigDB from MSigDb, which is a collection of 5219 gene sets related to immune pathways [34–36]. Across all 100 topics, we identified 2569 enriched gene sets with q-value $< 0.05$ using gene topic scores and 22 enriched gene sets using protein topic scores (**Fig.** 5a). For example, in topic 40, using the gene topic scores, we found a gene set that consists of down-regulated genes in healthy CD4 T cells with respect to healthy B cells [37] (**Fig.** 5d left panel); using the protein topic scores, we found a gene set that consists of up-regulated genes in B cells compared to plasmacytoid dendritic cells (PDC) (**Fig.** 5d right panel) [38].

Furthermore, we projected the topic embeddings and feature embeddings onto a common 2D space using UMAP (**Fig.** 5e). We observed that the top marker genes and the top marker proteins for the cell type clustered together around the corresponding topics. Together, the topics inferred by moETM from the CITE-seq data help link biological relations between genes and proteins via the cell-type-specific topics.

## 2.6 Multi-omic topics of joint genes and chromatin accessibility identified cell-type-specific pathways and regulatory motifs

The topic embedding learned from the scRNA+scATAC data enables us to investigate the relationship between top genes and top peaks (i.e., top open chromatin regions) in the cell-type-specific topics. Given that many top genes are cell-type known markers (**Fig.** 6a), we postulated that the top peaks could be associated with the top genes via *in-cis* or *in-trans* regulatory elements. Different from the previous gene+protein case, one challenge in interpreting the gene+peak multiomic topics is that peaks cannot be matched directly with genes. We proposed two approaches to solve this issue. One is to link peaks to their nearby genes to obtain the peak-neighboring-genes (**Methods**). The other approach is to identify enriched motifs among the top peaks and explore the relationship between genes and motifs via the corresponding transcription factors (TFs) and their target genes.

For the first approach, the top genes and top peak-neighboring-genes in the select topics served as markers for the cell-type-specific gene regulatory programs (**Fig.** 6a).

For example, topic 32 is associated with CD8+ T cell (**Fig.** 6a, b). We zoomed in the topic by examining its top genes and top peaks. Three of the top 5 genes (*TNFRSF9*, *ASTL*, *GZMK*, *DUSP2*, *DGKH*) were related to T cell. In particular, *GZMK* is a marker gene for T cells based on the CellMarker; *TNFRSF9* codes for a signaling protein that promotes expression of cytokines in CD8+ T cells [39]; *DUSP2* encodes an inducible nuclear protein and is highly expressed in T cells [40]. Among the top 5 peak-neighboring-genes (*APBA2*, *PRDX2*, *KLRC4*, *OBSCN*, *XCL2*), *APBA2* is a marker genes for cytotoxic CD8+ T lymphocyte [41]; *XCL2* expression levels substantially increased in CD8+ T cells during T cell activation [42].

As another example, topic 3 is associated with CD4+ T naive cells. Three out of the top 5 genes (*CCR4*, *ADAM12*, *PTPN13*, *MB21D2*, *IL4I1*) and two out of the top 5 peak-neighboring-genes (*INPP4B*, *CCR4*, *PRDX2*, *RORA*, *HIST1H2BD*) are related to T cells. Indeed, *CCR4* is shown to be specifically expressed among naive CD4+ T cells [43]; *ADAM12* is expressed in T cells in the inflamed brain and is a potential target for the treatment of Th1-mediated diseases [44]; *IL4I1* increases the threshold of T-cell activation and partially modulates CD4 T-cell differentiation [45]. For top peak-neighboring-genes, *RORA* is up-regulated among the activated CD4+ T cells [46].

To gain further mechanistic understanding of the inferred topics, we performed GSEA on the topic scores for the genes from the transcriptome modality and the topic scores for the peak-neighboring-genes from the chromatin-accessibility modality. Many enriched gene sets are related to the topic-associated cell types. For topic 3, for instance, one of the enriched gene sets based on the gene topic scores is up-regulated in healthy CD4 T cells with respect to healthy myeloid cells [37] (**Fig.** 6c). This is consistent to an enriched gene set from the peak-neighboring-gene analysis of topic 3, where the gene set consists of a set of genes that were down-regulated in peripheral mononuclear blood cells (PBMC) relative to the memory T cells [47]. Indeed, during T cell immune response, naiv̈e T cells might differentiate into memory cells [48]. Therefore, GSEA further confirmed the cell-type-specific functions of the top genes and peak-neighboring-genes identified via moETM's topics. Interestingly, the top genes and the top peak-neighbouring-genes are often not the same genes. This implies that the peaks and genes provide complementary information to (sometimes the same) cell-type-specific regulatory programs. Therefore, by effectively integrating the scRNA-seq and scATAC-seq data, the inferred multi-omic topics can reveal functional convergence at the pathway level.

Besides using peak-neighboring-genes, as the second approach, we also performed motif enrichment analysis on the top 100 peaks per topic (**Methods**). We then constructed a putative regulatory network by linking the top genes and the enriched motifs via their associated topics (**Fig.** 6d). Interestingly, some of the top genes harbor those enriched motifs, implying that these genes are the putative target genes of the cognate TF. In topic3, for example, one of the enriched motifs corresponds to a TF named FLI1 (p-value = 0.00117), and the top genes *IL4I1* and *PTPN13* are target genes of FLI1 based on the ENCODE Transcription Factor Targets [49, 50]. As another example, one of the enriched motifs for topic 32 correspond to TF MEF2A (p-value = 5.21e-5), whose target genes include the top genes *RGS1*, *EGR1*, *GZMK*, *ASTL*, and *DUSP2* [49, 50]. Therefore, our multi-omic topic analysis suggests that some of the cell-type-specific regulatory programs are implicated with the sequence motifs. Further investigation is needed to establish the hierarchical relation between the TF and the cell lineage.

## 2.7 Prior pathway-informed enrichment

The single-cell multi-omic data are high-dimensional, sparse, and noisy. This is especially the case for the scRNA+scATAC-seq data because of the large number of genes and open chromatin regions. One way to further improve the interpretability of the topics derived from these data is by incorporating prior knowledge such as gene sets or pathway information. In the context of our moETM, this was done by fixing the embeddings-by-genes parameters to the observed pathways-by-genes matrix (**Methods**). Using the 7000 Gene Ontology Biological Process (GO-BP) terms as the pathways-by-genes matrix, we trained the pathway-informed moETM (p-moETM) on the BMMC1 gene+peak dataset.

Quantitatively, p-moETM can achieve comparable cell-clustering performance with ARI 0.72, which is only slightly lower than the default moETM that learned the gene embedding directly from the data (Table 1). We also identified several cell-types-specific topics along with their top genes and peaks (**Supplementary Fig.** S3a-c). Notably, the learned topics-by-embeddings matrix $\alpha$ from p-moETM are essentially the topics-by-pathways matrix. This allows us to directly identify the top pathways for each topic without performing post-hoc GSEA. For instance, topic 25 is associated with B1 B cell (**Supplementary Fig.** S3a). One of its top pathways is related to B cell activation (**Supplementary Fig.** S3d). As another example, topic 8 was enriched for the CD4+ T activated cell, and one of its top pathways was connected to the T cell apoptotic process.

For some topics, their top genes are both the members of the pathway and the cell-type biomarkers. For instance, topic 27 is enriched in the CD4+ T naive cell. One of its top gene *CCR7* is involved in the elimination process of immature T cells. Additionally, topic 41 is enriched for the Transitional B cell. Its top pathways include B cell activation and adaptive immune process. Among its top genes, *TNFAIP3* is in the B cell activation-related pathway. One of its top peaks in chr14: 100207793 - 100208735 is upstream of the promoter of *YY1* (chr14: 100238298 - 100282788), which is a member gene in the B cell activation-related pathway [51].

Furthermore, we experimented a more specific gene set namely the immune signature gene set collections from MSigDB to investigate immune-related pathways implicated in the BMMC1 dataset (**Supplementary Fig.** S4). We identified several cell-type-specific topics that exhibit high scores for meaningful immune pathways. For instance, topic 23 is enriched in naive CD20+ B cells. Two of its top 10 pathways are associated with naive B cells. One of its top genes namely *HLA-DPB1* is up-regulated in naive B cells relative to the plasma cells [52]. One of the top peaks (chr12: 8886393 - 8887019) is upstream from *PHC1* (chr12:8913896 - 8941467), which is also involved in the pathway that genes are up-regulated in naive B cells relative to the plasma cells [52].

## 2.8 Multi-omic topics reveal the molecular basis of COVID-19 severity

As the CITE-seq technology interrogates the expression of surface proteins along with the full transcriptome, it is a promising platform to investigate the immune responses among patients infected by the SARS-CoV-2 virus (COVID-19). Using moETM, we sought to identify clinically relevant molecular signatures from a COVID-19 CITE-seq dataset (HBIC) [53]. The data consist of 781,123 cells from 130 COVID-19 patients with varying degrees of severity due to the viral infection. To establish model confidence, we first performed a quantitative analysis as above. The results showed that moETM could achieve either the highest or the second highest evaluation metrics both in bio-conservation and batch-removal cases (**Table** 2, **Supplementary Table** S2). In particular, moETM ranked first with an ARI value of 0.752 and TotalVI scored second with an ARI value of 0.733. Similarly, moETM and TotalVI attained the highest NMI scores of 0.779 and 0.762, respectively. Both methods also maintained their top performance in terms of batch-correction with TotalVI achieving the highest kBET of 0.197 while moETM coming in second with 0.153. Consistent to the above evaluation (**Table** S2), moETM obtained the best GC score of 0.950 whereas TotalVI achieved the second best of 0.934. Therefore, these quantitative results on the COVID-19 data further suggest that moETM strikes a good balance between biological conservation and batch effect correction in delivering competitive performance among all the SOTA methods.

Qualitatively, we investigated the top features and identified enriched cell types under each topic (**Supplementary Fig.** S5a, b). In particular, topic 42 is enriched for B cells. Among its top 5 genes (*SLC38A11*, *TCL1B*, *IL6*, *TCL1A*, *SYN3*), *IL6* and *TCL1A* are the known maker genes. Also, 3 out of its top 5 proteins (CD19, CR1, CD22, FCGR2A, BAFFR) are marker proteins for B cells. Topic 31 is associated with platelet. Two out of its top 5 genes (*LYVE1*, *RADIL*, *VWF* [54], *TRHDE*, *PPBP*) are marker genes, and one of its top 5 proteins (ITGA2B, KIR3DL1, ITGAX, SELP, FCGR2A) is a marker protein for platelet. Additionally, a previous study has suggested that SELP redistributes to the plasma membrane during platelet activation [55]. The enriched pathways based on GSEA are consistent to the cell-type specificity of those topics (**Supplementary Fig.** S5c). Taking topic 42 as an example, the enriched pathway is the gene set that is down-regulated in CD4 T cells compared with B cells [37]. Because of the shared embedding space, we also observed localization of the top genes and the top proteins for the selected topics via UMAP (**Supplementary Fig.** S5d).

We then leveraged the phenotype severity information among the patients to explore gene and protein signatures related to the COVID-19 phenotypes. Specifically, we utilized COVID metadata information to test whether a topic is significantly over-represented for the severity conditions. Here we considered each topic as a “meta-gene" and associated their up-regulation or down-regulation with the disease phenotypes (**Fig.** 7a, b). We observed that topic 42 is not only enriched for B-cell but also up-regulated among patients with critical COVID status whereas topic 80 is significantly associated with the severe status. Moreover, topic 42 is associated with other demographic features such as age and mainly enriched in the senior group between 70 and 79 years of age (**Fig.** 7a).

Given its disease relevance, we further investigated topic 42 to see whether it elucidates more granular cell types and to some extent whether their top gene/protein signatures can serve as putative biomarkers for COVID critical conditions. First of all, the moETM-inferred cell topic embeddings did not only cluster cells into their primary cell types but also sub-divided B cells into six sub-clusters of known sub-cell-types (**Fig.** 7c and zoom-in view). Intriguingly, aligning the COVID phenotypes with the B cell sub-types revealed that the critical COVID condition corresponded to B malignant cells (**Fig.** 7c, d). B-cell lymphomas start to develop when B lymphocytes, which are in charge of humoral immunity, start to proliferate beyond control. This proliferation turns B cells into malignant cells [56]. The previous study [57] suggested that individuals with certain cancers, such as lymphoma, may be more susceptible to getting severe illness from COVID-19. Furthermore, the top gene *IL6* in topic 42 was consistently expressed at a high level among B cells, including B malignant cells (**Fig.** 7e). Indeed, *IL-6* levels were commonly reported in severely ill patients due to COVID-19 [58, 59]. As another example, topic 21 is also enriched in B malignant cells (**Fig.** 7b). One of its top proteins CD5 (**Supplementary Fig.** S5b) was shown to be highly expressed on malignant cells [60]. Moreover, the previous study [61] suggested that the proportion of CD5+ B cells was significantly reduced in COVID patients. Taken together, our results suggest that *IL-6* or CD5 may be a potential therapeutic target.

# 3 Discussion

Gene regulatory programs involve multi-faceted regulation and can not be understood via only any of the facets alone. Single-cell multi-omic technologies open up venues to interrogate several omics simultaneously in the same cells. As these technologies continue to evolve, computational methods are needed to account for the challenges in modeling the sparse, noisy, and heterogeneous nature of data that are being generated at a rapid pace [3]. In this study, we developed a unified interpretable deep learning model called moETM to integrate single-cell multi-omic data including transcriptome and chromatin accessibility or surface protein, which are the most common types of single-cell multi-omic data to date [4].

Our technical contributions are three-folds. First, via the product-of-experts, moETM effectively integrates multiple omics by projecting them onto a common topic mixture representation. Second, the linear decoder enables the extraction of multi-omic signatures as the top features under each latent topic, which directly reveal marker genes and phenotype markers under topics that are aligned with cell types or phenotype conditions. Third, by efficiently correcting batch effects via a dedicated linear intercept matrix in the decoder, we can integrate multi-omic data from multiple studies, subjects, or technologies, which allows us to exploit the vast amount of multi-omic data in order to obtain biologically diverse and coherent multi-omic topics. Notably, while the last two contributions are inherited from our earlier single-cell transcriptome model namely scETM [14], we consider the success of incorporating them into the multi-omic modeling problem as a substantial departure from the existing method.

To demonstrate the utility of moETM, we benchmarked it with 6 existing state-of-the-art computational methods on 7 published datasets including 4 gene+peak datasets and 3 gene+protein datasets (**Table** 1, 2). Across all datasets, moETM achieved competitive performance in terms of 4 common metrics including the bio-conservation evaluation metrics (i.e., ARI and NMI) and batch-removal evaluation metrics (i.e., kBET and GC). We also confirmed the advantage of using multiple modalities compared with single modality in terms of cell clustering (**Table** 1, 2, **Supplementary Table** S1, S2).

As the vast majority of the single-cell data are still single-omic (e.g., scRNA-seq, scATAC-seq, etc), there are tremendous benefits of imputing one omic from another omic. Because of its joint modeling capabilities, the trained moETM can accomplish this cross-omic imputation task. The imputation can go from a high-dimensional omic to a low-dimensional omic and vice versa. The latter imputation direction is more demanding on the decoder because it needs to “remember" the high-dimensional manifold in its parameter space when decoding the lower-dimensional feature space. In our applications, this involves imputing gene expression (20K) from ATAC peaks (∼100K) and imputing surface protein abundance (∼140) from gene expression (∼20K) and vice versa. In both imputation directions, moETM achieved a higher correlation than scMM and BABEL. Although more challenging, moETM also achieved a reasonable performance when imputing high-dimension from low-dimension.

We also explored the moETM-learned cell-type-specific topics in terms of their top omic features and enriched pathways in light of the supporting evidence from the literature.

For example, in the BMMC2 (gene+protein) dataset, *CR2* is a top gene signature identified by high topic scores in a B-cell-specific-topic topic 40, a marker gene for B cell in the CellMarker database (**Fig.** 5b, c), and a member of an enriched pathway (genes that are down-regulated in CD4 T cells compared with B cells [37]) for the topic. By binding to C3d, *CR2* can lower the threshold for B cell activation in an adaptive immune response [62].

Similarly, in the BMMC2 (gene+protein) dataset, protein CD19 is included in the topic enriched pathway (genes that are down-regulated in CD4 T cells compared with B cells [37]). It is ranked sixth in the B-cell-specific topic 40 and is crucial in determining intrinsic B cell signaling thresholds. Along with other molecules, CD19 functions as the dominant signaling element of a multimolecular complex on the surface of mature B cells [63].

In the BMMC1 (gene+peak) dataset, a top gene feature *IL4I1* in a T-cell-specific topic 3 is a target gene of the enriched motif FLI1 (**Fig.** 6d), which is determined by the top 100 peaks under that topic. *IL4I1* could inhibit human CD4+ and CD8+ T lymphocyte proliferation *in-vitro* [64]. Furthermore, T cells have a high-level expression of FLI1 and the expression decreases after T cell activation [65].

In a more focused study, we analyzed the COVID-19 CITE-seq dataset (gene+protein) and linked moETM-learned immune-specific topics with patient severity conditions due to the infection. Our topic analysis revealed not only immune marker genes but also cell types that are associated with COVID phenotype conditions. In particular, we found that the patients with critical status exhibited high topic probabilities for the B malignant cells. Furthermore, one of the B malignant cell marker genes *IL6* is differentially expressed among these patients compared to patients with mild and no symptoms ( [58,59]).

There are several challenges that are not addressed in moETM [4]. For instance, moETM has the capacity to integrate across multiple batches and modalities but it requires the training data to have all omics measured in the same cells. Given that the transcriptome is shared between the gene+peak and gene+protein data, it is possible to integrate all 3 omics in a mosaic data integration regime while taking into account the data heterogeneity. A more challenging task is to integrate multimodal data without anchored features or cells, which is commonly known as the diagonal integration [4]. Some recent approaches made use of graph representation learning to integrate multi-omic single-cell data at the expense of computational complexity and interpretability [66–68]. Furthermore, given the motif enrichments in our analysis, another natural extension of moETM is to model the sequence information at the upstream of the model training using language models such as the Bidirectional encoder representations from transformer (BERT) model [69]. Indeed, at the decreasing computational cost, we started to see interesting applications of BERT in the related fields including single-cell data modeling [70], genome language understanding [71], and sequence-based gene expression prediction [72].

# 4 Methods

## 4.1 moETM data generative process

The molecular activities in each cell $n$ can be measured with $M$ omics, such as gene expression from transcriptome, surface protein expression, and the open chromatin regions manifested as peaks. For the ease of the following descriptions, we define the entities of genes, proteins and peaks as “features". Profiling those omics in the cell leads to $M$ count vectors $\{\mathbf{x}_n^{(m)}\}_{m=1}^M$, each of which has a dimension $V^{(m)}$ as the number of unique features in omic $m$. Adapting the text-mining analogy, we consider each cell as a “document" written in $M$ languages or modalities (i.e., transcriptome, proteome, chromatin accessibility); each feature from the $m^{th}$ omic is considered as a “word" from the $m^{th}$ vocabulary; each sequencing read is a “token" in the document; the abundance of the reads mapped to the same feature is the “word count" in the document.

The multi-modal document of a cell $n$ can be summarized into a mixture of $K$ latent topics $\boldsymbol{\theta}_n$, which are presumably implicated in each modality (**Fig.** 1a). Inference of these topic mixtures for each cell is accomplished by modeling the distribution of the multi-omic count data $\{\mathbf{x}_n^{(m)}\}_{m=1}^M$ from the topic mixture for the cell and the global topic embedding over the $M$ modalities. The latter are shared among all cells and expressed as $M$ matrices $\{\mathbf{\Phi}^{(m)} \in \mathbb{R}^{K\times V^{(m)}}\}_{m=1}^M$, where a column vector $\boldsymbol{\phi}_k^{(m)} \in \mathbb{R}^{V^{(m)}}$ denotes the $k$-th topic from the $m$-th modality.

To increase information sharing across the omics and the model expressiveness, we further decompose each omic-specific topic embedding matrix $\mathbf{\Phi}^{(m)}$ into the topic embedding $\boldsymbol{\alpha} \in \mathbb{R}^{K\times L}$ and feature embedding $\boldsymbol{\rho}^{(m)} \in \mathbb{R}^{L\times V^{(m)}}$, where $L$ denotes the size of the embedding space. The expected values for the count data for each omic is proportional to the dot product of the cell embedding, topic embedding matrix, and feature embedding matrix: $\mathbf{x}_n^{(m)} \propto \boldsymbol{\theta}_n\boldsymbol{\alpha}\boldsymbol{\rho}^{(m)}$.

Formally, we formulate the data generative process as follows. For each cell indexed by $n \in \{1, \ldots, N\}$, draw a $1 \times K$ topic proportion $\boldsymbol{\theta}_n$ from logistic normal distribution $\boldsymbol{\theta}_n \sim \mathcal{LN}(\mathbf{0}, \mathbf{I})$:

$$\boldsymbol{\delta}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \boldsymbol{\theta}_n = \text{softmax}(\boldsymbol{\delta}_n) = \frac{\exp(\delta_{n,k})}{\sum_{k=1}^{K} \exp(\delta_{n,k})}. \tag{1}$$

For each read $i^{(m)} \in \{1, \ldots, D_n^{(m)}\}$ from the $m^{th}$ modality $w_{n,i^{(m)}}^{(m)}$, draw a feature index $v^{(m)}$ (e.g., the particular transcript or open chromatin region the read was sequenced) from a categorical distribution Cat($\mathbf{r}_{n,.}^{(m)}$):

$$w_{n,i^{(m)}}^{(m)} \sim \prod_{v^{(m)}=1}^{V^{(m)}} [r_{n,v}^{(m)}]^{[w_{n,i^{(m)}}^{(m)} = v^{(m)}]}, \quad x_{n,v}^{(m)} = \sum_{i=1}^{D_n^{(m)}} [w_{n,i}^{(m)} = v]. \tag{2}$$

where $D_n^{(m)}$ and $x_{n,v}^{(m)}$ denote the total number of reads and the read count for feature $v^{(m)}$ for cell $n$ in the $m^{th}$ modality, respectively. The expected rate $r_{n,v^{(m)}}^{(m)}$ of observing feature $v^{(m)}$ in cell $n$ is parameterized as:

$$r_{n,v^{(m)}}^{(m)} = \frac{\exp(\hat{r}_{n,v^{(m)}}^{(m)})}{\sum_{v^{(m)}=1}^{V^{(m)}} \exp(\hat{r}_{n,v^{(m)}}^{(m)})}, \quad \hat{r}_{n,v^{(m)}}^{(m)} = \boldsymbol{\theta}_n \boldsymbol{\alpha} \boldsymbol{\rho}_{.,v^{(m)}}^{(m)} + \lambda_{s(n),v^{(m)}}^{(m)}. \tag{3}$$

where $\boldsymbol{\rho}_{.,v^{(m)}}^{(m)} \in \mathbb{R}^{L \times 1}$ denotes embedding of feature $v^{(m)}$, $\lambda_{s(n),v^{(m)}}^{(m)}$ is the batch-dependent and feature-specific scalar effect, where $s(n)$ indicates the batch index for the $n^{th}$ cell. Notably, the softmax function normalizes the expected observation rates over all features separately with-in each modality to account for different modality size (e.g., there are more peaks than genes, and more transcripts than surface proteins). Another reason for the normalization is to capture feature sparsity (i.e., only a small fraction of features from each modality is non-zero). This is analogous to text mining, where a small fraction of the vocabulary from any language is observed in any given document.

## 4.2 moETM model inference

For the ease of inference, we consider the cell topic embedding $\boldsymbol{\delta}_n$ (before softmax normalization) for cells $n \in \{1, \ldots N\}$ as the latent variables and all the cells are independent. The rest of the parameters including topic embedding $\boldsymbol{\alpha}$, feature embedding $\{\boldsymbol{\rho}^{(m)}\}_{m=1}^{M}$, and batch-effect parameter $\{\boldsymbol{\lambda}^{(m)}\}_{m=1}^{M}$ are treated as point estimates and learned by the model. Let's denote $\hat{\Theta} = \{\boldsymbol{\delta}_n, \boldsymbol{\alpha}, \{\boldsymbol{\rho}^{(m)}\}_{m=1}^{M}, \{\boldsymbol{\lambda}^{(m)}\}_{m=1}^{M})\}$. A principled way to learn those parameters is to maximize the marginal log likelihood:

$$\Theta^* \leftarrow \arg\max_{\Theta} \sum_n \int \log p(\{\mathbf{x}_n^{(m)}\}_{m=1}^{M} \mid \Theta^*) d\boldsymbol{\delta}_n \equiv \arg\max_{\Theta} \sum_n \mathcal{L}_n$$

However, this integral is not tractable. Instead, we took a variational inference approach to optimize the model parameters by maximizing an evidence lower bound (ELBO) of the marginal log likelihood with a proposed variational posterior $q(\boldsymbol{\delta}_n)$ as a surrogate to the true posterior of the cell topic embedding $p(\boldsymbol{\delta}_n \mid \{\mathbf{x}_n^{(m)}\}_{m=1}^M)$:

$$\begin{aligned}\mathcal{L}_n &\geq \mathbb{E}_{q(\boldsymbol{\delta}_n)}\left[\log p(\{\mathbf{x}_n^{(m)}\}_{m=1}^M \mid \boldsymbol{\delta}_n, \Theta^*) + \log p(\boldsymbol{\delta}_n) - \log q(\boldsymbol{\delta}_n)\right] \\ &\equiv \mathbb{E}_{q(\boldsymbol{\delta}_n)}\left[\log p(\{\mathbf{x}_n^{(m)}\}_{m=1}^M \mid \boldsymbol{\delta}_n, \Theta^*)\right] - KL[q(\boldsymbol{\delta}_n)||p(\boldsymbol{\delta}_n)] \equiv ELBO_n \end{aligned} \tag{4}$$

where KL denotes the Kullback-Leibler (KL) divergence between the proposed distribution and the prior (i.e., standard Gaussian with zero mean and identity variance), acting as a regularization when maximizing the data likelihood.

We defined the proposed distribution $q(\boldsymbol{\delta}_n)$ as a product of Gaussians (PoG):

$$q(\boldsymbol{\delta}_n) = \mathcal{N}(\boldsymbol{\delta}_n; \boldsymbol{\mu}^*, \boldsymbol{\sigma}^{*2}), \tag{5}$$

The mean $\boldsymbol{\mu}$ and standard deviation $\boldsymbol{\sigma}$ of the joint Gaussian is computed as:

$$\boldsymbol{\mu}^* = \frac{\sum_{m=1}^M \boldsymbol{\mu}_m \boldsymbol{\sigma}_m^2}{1 + \sum_{m=1}^M \boldsymbol{\sigma}_m^2}, \quad \boldsymbol{\sigma}^{*2} = \frac{\prod_{m=1}^M \boldsymbol{\sigma}_m^2}{1 + \sum_{m=1}^M \boldsymbol{\sigma}_m^2} \tag{6}$$

where $\boldsymbol{\mu}_m$ and $\boldsymbol{\sigma}_m^2$ are the mean and variance of the Gaussian latent embedding for the individual modalities, respectively. Those are output from the encoder neural network (NNET):

$$[\boldsymbol{\mu}_n^{(m)}; \log \boldsymbol{\sigma}_n^{(m)}] = \text{NNET}(\tilde{\mathbf{x}}_n^{(m)}; \mathbf{W}) \tag{7}$$

where $\tilde{\mathbf{x}}_n^{(m)}$ is the normalized counts for each feature as the raw count of the feature divided by the total counts of $m^{th}$ modality in cell $n$, and $\mathbf{W}$ is the parameters for a two-layer feed-forward neural network.

Following stochastic variational inference (SVI) approach, we approximate the above ELBO in Eq (4) by sampling from the proposed joint Gaussian distribution using the reparameterization trick [12]:

$$\begin{gathered}\tilde{\boldsymbol{\delta}}_n \sim \mathcal{N}(\boldsymbol{\mu}^*, \text{diag}(\boldsymbol{\sigma}^*)) = \boldsymbol{\mu}^* + \text{diag}(\boldsymbol{\sigma}^*)\mathcal{N}(\mathbf{0}, \mathbf{I}) \\ ELBO_n \approx \log p(\{\mathbf{x}_n^{(m)}\}_{m=1}^M \mid \tilde{\boldsymbol{\delta}}_n, \Theta) - KL[q(\tilde{\boldsymbol{\delta}}_n)||p(\tilde{\boldsymbol{\delta}}_n)]\end{gathered}$$

Together, the model parameters including the encoder weights are optimized by maximizing the following ELBO via backpropagation:

$$\Theta^*, \mathbf{W}^* \leftarrow \underset{\Theta, \mathbf{W}}{\arg\max} \sum_{n=1}^N \log p(\{\mathbf{x}_n^{(m)}\}_{m=1}^M \mid \tilde{\boldsymbol{\delta}}_n, \Theta^*) - KL[q(\tilde{\boldsymbol{\delta}}_n)||p(\tilde{\boldsymbol{\delta}}_n)] \tag{8}$$

## 4.3 Single-cell multi-omic datasets and preprocessing

There were 7 public datasets included in this study for performance evaluation and model comparison. All 7 datasets are from publicly available repositories. Among them, 4 datasets provide joint profiling of gene expression and open chromatin regions (denoted as “gene+peak" data): the Multiome bone marrow mononuclear cells (BMMC1) dataset from the 2021 NeurIPS challenge consisting of 42,492 cells with 22 cell types from 10 donors across 4 sites [73], the SHARE-seq mouse skin late anagen (MSLAC) dataset containing 34,774 cells with 1 batch and 23 cell types [24], the sci-CAR mouse kidney cells (MKC) dataset from cell samples with 1 batch and 14 cell types [74], and the SHARE-seq mouse brain cells (MBC) dataset containing 3,293 cells with 1 batch and 19 cell types [24]. For the BMMC1 dataset, we take into account two different batch types: one treats a subject (eg. site1 + donor1 as a subject s1d1, site1 + donor2 as a subject s1d2, etc) as a batch (s1d1, s1d2, s1d3, s2d1, s2d4, s2d5, s3d3, s3d6, s3d7, s3d10, s4d1, s4d8, s4d9, 13 batches in total), while the other treats a site (site1 as batch1, site2 as batch2) as a batch (4 batches in total).

For the CITE-seq data measuring transcriptome and surface protein in the same cell, 3 datasets were used in this study: the bone marrow mononuclear cells (BMMC2) dataset from the 2021 NeurIPS challenge from 9 donors and 4 sites [73], the Human White Blood Cell (HWBC) dataset containing 211,000 human peripheral blood mononuclear cells [11], and the Human Blood Immune Cell (HBIC) dataset [53] measuring 647366 peripheral blood mononuclear cells from both COVID patients and healthy patients. Similarly, for the BMMC2 dataset, we consider two different batch types: one treats a subject containing one donor and one site as a batch (12 batches in total), while the other treats a site as a batch (4 batches in total).

All datasets were processed into the format of samples-by-features matrices. For gene+peak datasets, the read count for each gene and peak were first normalized per cell by total counts within the same omic using *scanpy.pp.normalize_total* function in the *scanpy* [75], then log1p transformation was applied. After that, *scanpy.pp.highly_variable_genes* was used to select highly variable genes or peaks.

For the joint profiling of transcriptome and surface protein data (denoted as gene+protein), we used all surface proteins measured by the scADT-seq assay since the number of proteins is much smaller compared with the number of genes or peaks and all of them are highly informative of immune cell functions. The same normalization as in the gene+peak data was performed on the gene+protein data.

## 4.4 Cross-omic imputation

The trained moETM can impute one omic from another omic. Suppose we have two omics namely omic A and omic B. For the training data where both omics are observed, moETM learns a shared topic embedding $\boldsymbol{\alpha}$ and omic-specific feature embedding $\boldsymbol{\rho}^{(A)}$ and $\boldsymbol{\rho}^{(B)}$. For the testing data, suppose without loss of generality that only omic B is observed. To impute omic A, moETM uses the encoder for modality B to generate the topic mixture, which is then input to the decoder for omic A to complete the imputation (**Fig.** 1c).

We evaluated the imputation accuracy using the BMMC1 (gene+peak) and BMMC2 (gene+protein) datasets based on (1) 60/40 random split of training and testing data with 500 repeats to get standard deviation estimate; (2) training on all batches except for one batch and testing on the held-out batch (leave-one-batch); (3) training on all cell types except for one cell type and testing on the held-out cells of that cell type (leave-one-cell-type).

## 4.5 Evaluation metrics

The batch effects correction and biological variance conservation categories were used to assess the efficacy of the integration across multiple modalities. To quantify bio-conservation, we used the Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI), and to measure batch effect removal, we used k-nearest-neighbor batch-effect test (kBET) and Graph Connectivity (GC). Specifically, ARI calculates the degree of similarity between two clusterings and adjusts for the possibility that objects can randomly form the same clusters. NMI normalizes the mutual information to a scale of 0 to 1. While NMI excels in unbalanced clustering or small clusters, ARI is better suited to clusters of similar size [76]. kBET performs hypothesis testing on whether batch labels are distributed differently across cells based on Pearson's $\chi^2$ test [19]. GC measures whether cells of the same type from different batches are close to one another by computing a K nearest-neighbour graph based on the distance between cells in the embedding space [20].

## 4.6 Linking genes to open chromatin regions

We sought to investigate the relation between the top peaks and top genes under the same moETM topic (i.e., $\boldsymbol{\phi}_k^{(m)} = \boldsymbol{\alpha}_k \boldsymbol{\rho}^{(m)}$ for topic $k$ and $m \in \{gene, peak\}$). To assess the *in-cis* relation, we measured the genomic distances between genes and peaks and designated genes that were near peaks as peak-neighboring-genes if they are within 150K base pairs (bp) distance.

Specifically, we first obtained a genes-by-topics matrix $\boldsymbol{\phi}^{\text{(gene)}} = \boldsymbol{\alpha}\boldsymbol{\rho}^{\text{(gene)}}$ and a peaks-by-topics matrix $\boldsymbol{\phi}^{\text{(peak)}} = \boldsymbol{\alpha}\boldsymbol{\rho}^{\text{(peak)}}$.

To transform $\boldsymbol{\phi}^{\text{(peak)}}$ into a peak_to_genes-by-topics matrix $\boldsymbol{\phi}^{\text{(peaks\_to\_genes)}}$, we first derived a binary peaks-to-genes mapping matrix $\mathbf{H}$ with the entries $h_{p,g} = 1$ if the corresponding pair of peak $p$ and gene $g$ are within 150K bp genomic distance and are positively correlated and 0 otherwise.

In detail, we computed the Pearson correlation between gene $g$ and peak $p$ in terms of their topic scores:

$$r_{p,g} = \frac{(\boldsymbol{\phi}_g^{\text{(gene)}} - \bar{\phi}_g^{\text{(gene)}})^\top (\boldsymbol{\phi}_p^{\text{(peak)}} - \bar{\phi}_p^{\text{(peak)}})}{||\boldsymbol{\phi}_g^{\text{(gene)}} - \bar{\phi}_g^{\text{(gene)}}||_2 ||\boldsymbol{\phi}_p^{\text{(peak)}} - \bar{\phi}_p^{\text{(peak)}})||_2}$$

The genome distance between peaks and genes was based on the latest genome build (i.e., hg38 for human) and obtained via the *GenomicRanges* [77] package in R.

## 4.7 Pathway enrichment analysis

For each moETM topic, we performed Gene Set Enrichment Analysis (GSEA) [32] to associate the topic with known pathways or gene sets. In particular, we used each topic to query two gene sets from Molecular signatures database (MSigDB), which are the 5219 Immunologic signature gene sets (C7) and the 7763 Gene Ontology Biological Processes (BP) (C5-BP) terms. For each topic, we ran *GSEAPreranked* on a ranked list of genes based on their corresponding topic scores against every gene set from C7 or C5-BP, and calculated the enrichment score (ES) for over- or under-representation. The statistical significance of the ES was computed based on 1000 permutation test. The gene sets with Benjamini–Hochberg (BH) corrected p-values lower than 0.05 were deemed significant. Similarly, for the scATAC-seq data, the peaks-by-topics matrix was first converted into a peaks_to_genes-by-topics matrix and then provide as input to GSEA pipeline.

## 4.8 Motif enrichment analysis of top peaks from moETM-learned topics

To detect sequence-based regulatory elements for the cell-type-specific topics, we performed motif enrichment analysis using the top 100 peaks that exhibit the highest topic scores under each topic. The 100 sequences corresponding to those top 100 peaks under each topic were extracted from Ensembl database and provided as input to the Simple Enrichment Analysis (SEA) pipeline [78] from the MEME suite [79]. SEA utilizes the STREME motif discovery algorithm [80] to identify known motifs that are enriched in input sequences. For our purpose, we used the HOmo sapiens COmprehensive MOdel COllection (HOCOMOCO) Human (v11) and HOCOMOCO Mouse (v11) motif database [81]. Motifs with Fisher's exact test p-values lower than 0.05 were selected as the enriched motifs.

## 4.9 Differential analysis to detect condition-specific topics

We sought to detect moETM-topics that exhibit significantly higher scores for the conditions of interest such as cell types or phenotypes. Notably, while the cell types were at the single-cell level, the phenotypes were at the subject level (e.g., COVID-19 severity state). The latter means that the cells from the same subject were assigned the same phenotype label. For each dataset, we first split the cells into positive and negative groups, corresponding to the presence and absence of the target condition, respectively. For each topic, we assessed the statistical significance of the topic score increase for the positive group relative to the negative group based on one-sided student t-test. The topics with a Bonferroni-adjusted p-value smaller than 0.001 were considered significant with the label.

## 4.10 Incorporating pathway-informed gene embeddings

In the linear decoder, we reconstruct the cells-by-features matrix by the dot product of the 3 matrices, namely cells-by-topics, topics-by-embedding, and embedding-by-features. By default, the last feature embedding matrix consist of learnable parameters. However, we can instill prior pathway information during the training of moETM by fixing the features embedding to a known gene set. As a result, the topics-by-embedding and embedding-by-features matrices change to topics-by-gene_sets and gene_sets-by-features with only the topics-by-gene_sets as the learnable parameters. This allows us to directly map each topic to each gene set, which may further improve the model interpretability especially if the chosen gene sets were highly relevant to the data. Given that several single-cell multi-omic datasets used in this study were derived from the blood, we utilized the Immunologic signature gene sets collection (C7) from the MSigDB database. Gene sets with fewer than five or more than 1000 genes were filtered out. We then converted the gene set information into a binary gene_sets-by-genes matrix with 0 and 1 indicating the absence and presence of the genes (columns) in the corresponding gene set (rows), respectively. We focused on the gene+peak case by fixing the gene embedding to the gene set while learning the peak embedding as in the default setting. We did not experiment this approach on the gene+protein case, for which the topics learned by the default moETM are sufficiently easy to interpret.

# Code Availability

The moETM code is available at `https://github.com/manqizhou/moETM`.

# Acknowledgement

F.W. would like to acknowledge the support from NIH R01AG076448, R01AG076234, RF1AG072449, NSF 1750326 and 2212175. Y.L. is supported by SERC Alliance Catalyst ALLRP 576153-22, NSERC Discovery Grant DGECR-2019-00253 and CIHR Canada Research Chair (Tier 2) in Machine Learning for Genomics and Healthcare.

# 5 Figures

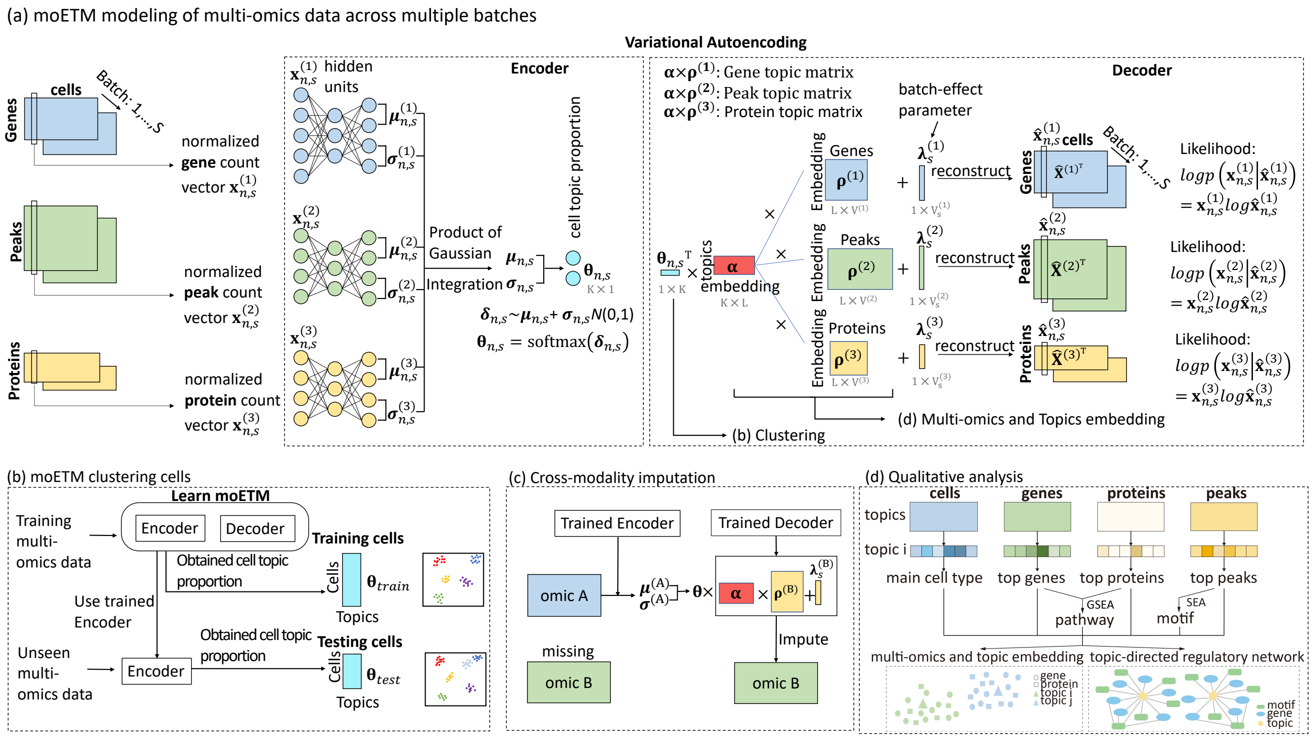

Figure 1: **moETM model overview. a**. Modeling single-cell multi-omics data across batches. In a nutshell, moETM integrates $M$ omics via the product-of-experts (PoE), each of which is a pair of encoder and decoder. For a given cell $n$ from batch $s$, each expert encoder takes one omic $m$ as input $\mathbf{x}_{n,s}^{(m)}$ and produces the mean $\boldsymbol{\mu}_{n,s}^{(m)}$ and log variance $\log((\boldsymbol{\sigma}_{n,s}^{(m)})^2)$ for the omic-specific Gaussian distributed latent embedding variable. The product of these Gaussian densities over the $M$ omics is also a Gaussian, from which we sample a joint logistic Gaussian latent embedding $\boldsymbol{\theta}_{n,s} \sim$ softmax$(\boldsymbol{\mu}_{n,s} + \boldsymbol{\sigma}_{n,s}\mathcal{N}(0, \mathbf{I}))$ to represent the cell. Each linear $m^{th}$ decoder expert then takes the same topic proportion $\boldsymbol{\theta}_{n,s}$ as input and reconstruct the original omic $m$ for the cell with the aid of the global topic embedding $\boldsymbol{\alpha}$ and the omic-specific feature embedding $\boldsymbol{\rho}^{(m)}$. The end-to-end learning of the encoder network parameters and the decoder topic and feature embeddings is accomplished by maximizing the evidence lower bound of the categorical likelihood for the multi-omic count data via backpropagation. **b**. Evaluating moETM through cell clustering. The trained PoE encoders is used to infer the topic proportion of either training $\boldsymbol{\theta}_{train}$ or test data $\boldsymbol{\theta}_{test}$ from their multi-omic data. The integration performance of moETM is evaluated by clustering cells based on their topic proportion and qualitatively evaluated by UMAP visualization. **c**. Cross-omic imputation. To impute the missing omic $B$ (e.g., protein) for a test cell, the trained moETM feeds the observed omic input vector $\mathbf{x}^{(A)}$ to the corresponding encoder expert $A$. The joint Gaussian embedding is then fed to the expert decoder $B$, which takes the inner product of the cell embeddings with its learned topic embedding and feature embedding for omic $B$. **d**. Downstream topic analysis. The learned topics-by-{cells, genes, proteins, peaks} matrices enable identifying cell-type-specific topics, gene signatures, surface protein signatures, and regulatory network motifs, respectively.

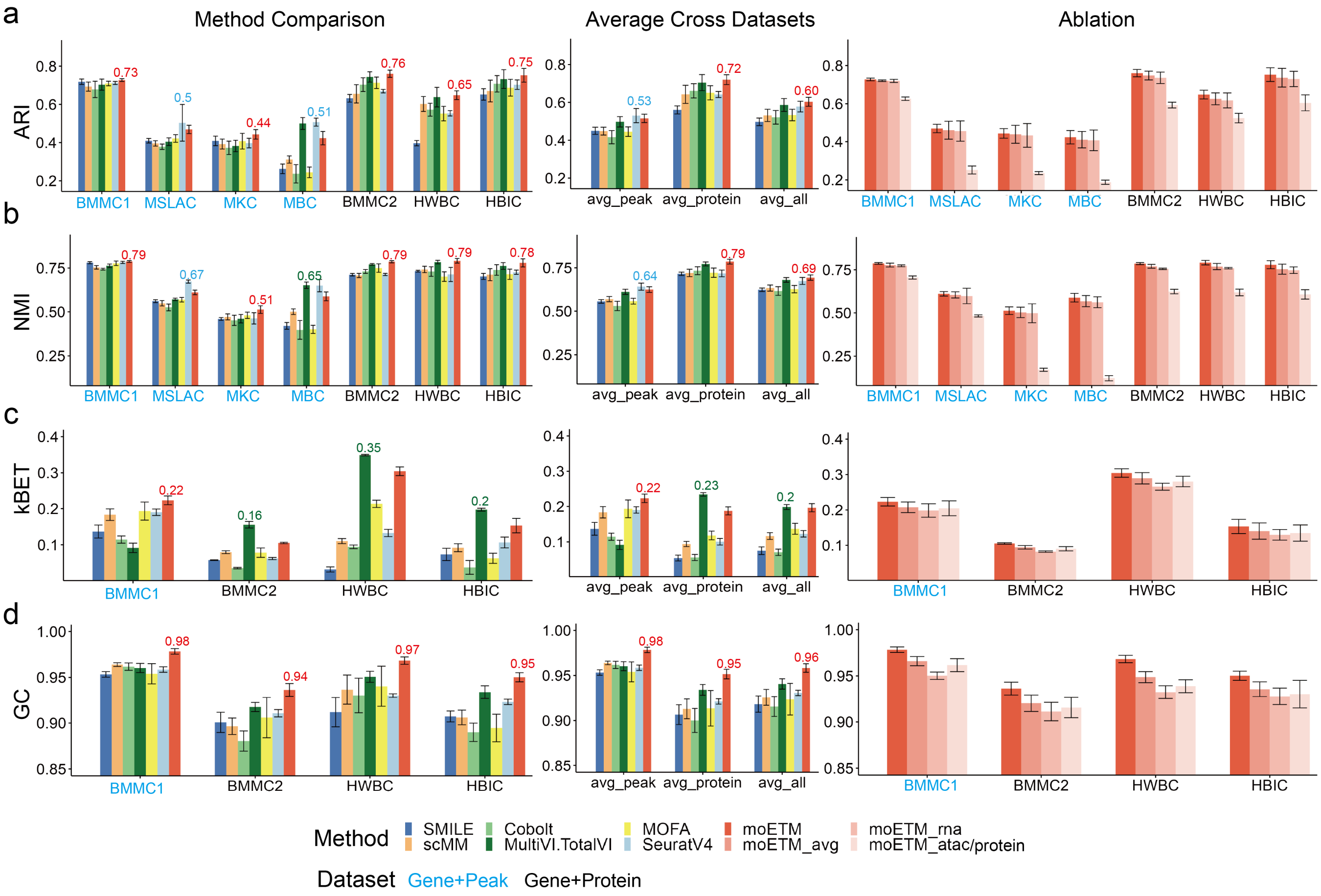

Figure 2: **Methods comparison based on cell clustering.** The left column illustrates the individual performance of each method on each dataset. The 7 datasets are indicated on the x-axis with gene+peak datasets colored in blue and gene+protein datasets colored in black. The evaluation scores for each are shown on y-axis. Ten colors were used to represent 10 different methods including six existing state-of-the-art methods, the proposed moETM model, and 3 of its ablated versions. Within each dataset, the highest value was labeled on the top of the corresponding bar. The middle column is the comparison of averaging values across datasets for each method. The right column is the comparison between moETM and its three ablated versions. Each row represents an evaluation metric. **a**. Adjusted Rand Index (ARI). **b**. Normalized Mutual Information (NMI). **c**. k-nearest neighbour batch effect test (kBET). **d**. Graph connectivity (GC).

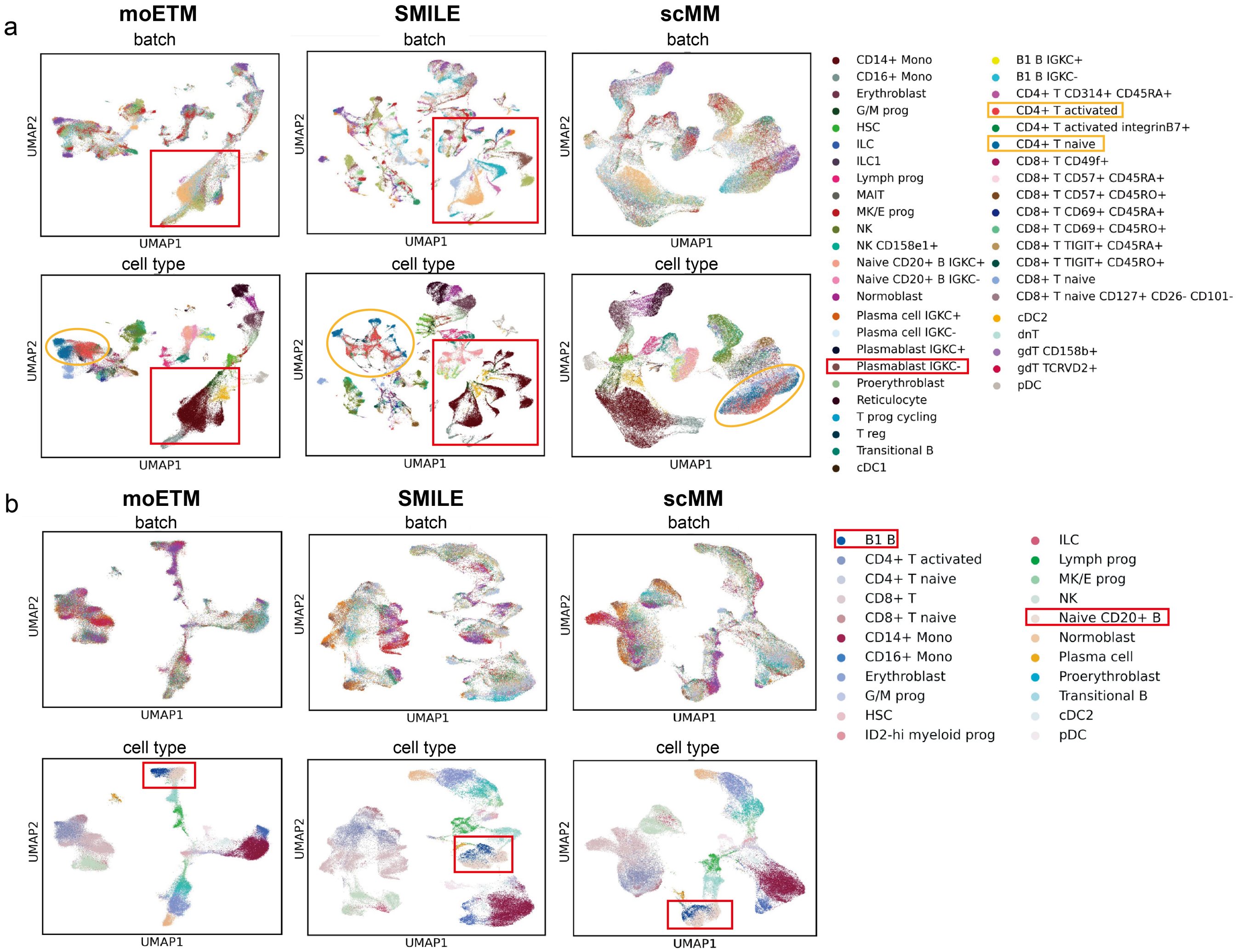

Figure 3: **UMAP visualization of cell clustering. a**. UMAP visualization of moETM, SMILE, and scMM on single-cell CITE-seq from BMMC2 dataset. Each point on the two-dimensional UMAP plots represents a cell. In the upper panel, different colors indicate different batches. In the lower panel, different colors indicate different cell types. **b**. UMAP visualization of moETM, SMILE, and scMM on the gene+peak multiome data from the BMMC1 dataset. Similarly to panel a, the upper and lower panel labelled with batch indices and cell types, respectively. The highlighted clusters and cell types in the legend were described in the main text.

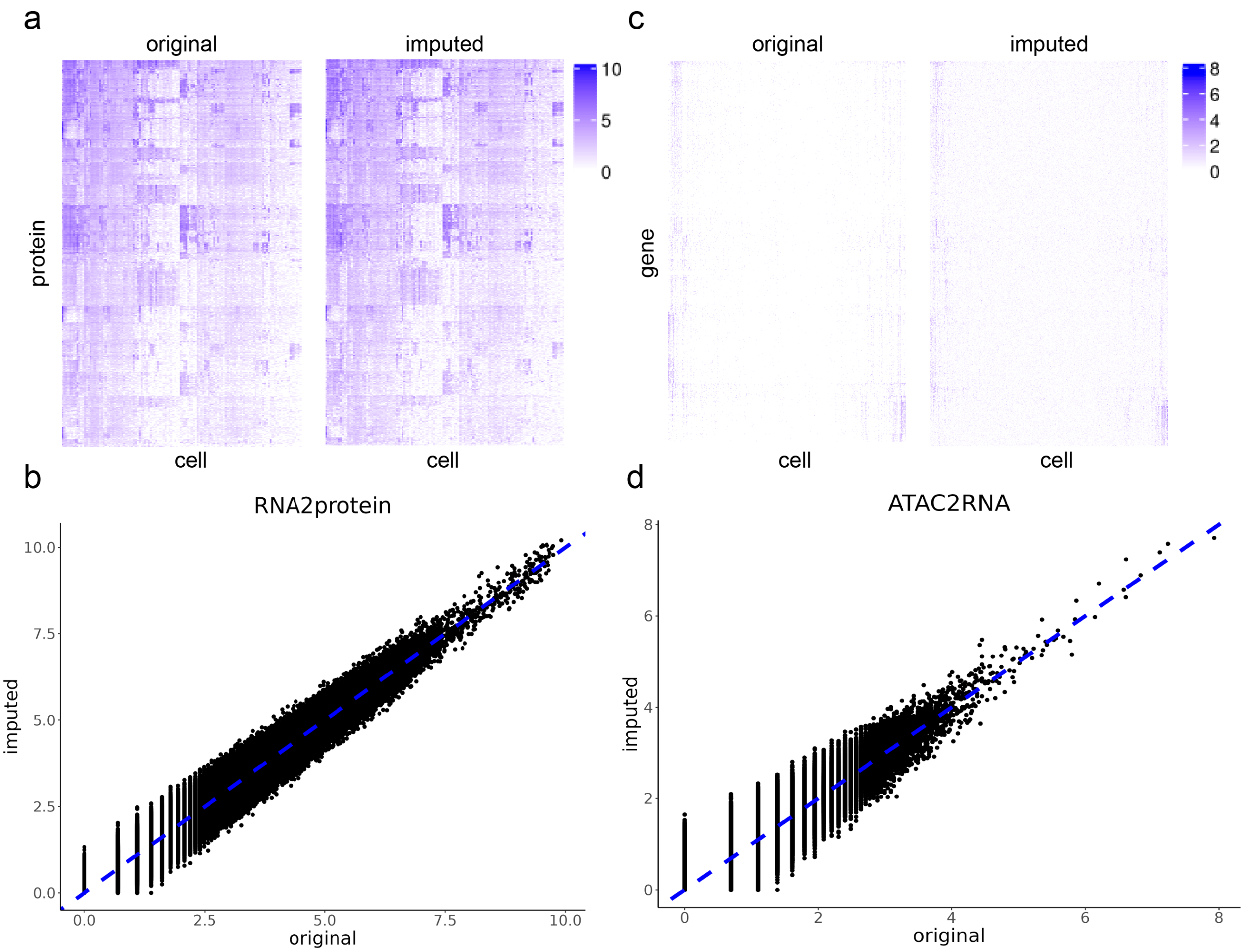

Figure 4: **Cross-omic imputation. a**. Heatmap of original protein and imputed protein values from gene expression using the BMMC2 CITE-seq dataset. We trained moETM on 60% of the cells with observed protein+gene omics and used the trained moETM to impute the protein expression based on the gene expression for the remaining 40% of the test cells. The two heatmaps correspond to the original and imputed protein expression, respectively. The columns are the randomly sampled 5000 test cells, and the rows are the surface proteins. For visual comparison, the column and row orders are the same for the two heatmaps. The color intensities are proportional to original or imputed protein expression over the cells. **b**. Scatter plot of original and imputed surface protein expression. The same values shown in panel a were displayed as scatter plot in this panel. The x-axis and y-axis represent the original and imputed protein expression values of the test cells, respectively. The diagonal line is in blue color. The more similar the reconstructed value is with the original value, the closer it is with the blue line, **c** & **d**. Heatmap and scatterplot of the original and imputed gene expression from chromatin accessibility on the BMMC1 dataset. The imputation results were shown in the same way as in panel a and c. We trained moETM on 60% of the cells with observed gene+peak omics. We then applied the trained moETM to the 40% test cells by imputing their gene expression based on their open chromatin regions (i.e., peaks). The original and imputed gene expression of the test cells were compared qualitatively in the heatmap and scatterplot. We also illustrated the imputation results from the low dimensional omic to the high dimensional omic in **Supplementary Fig.** S2.

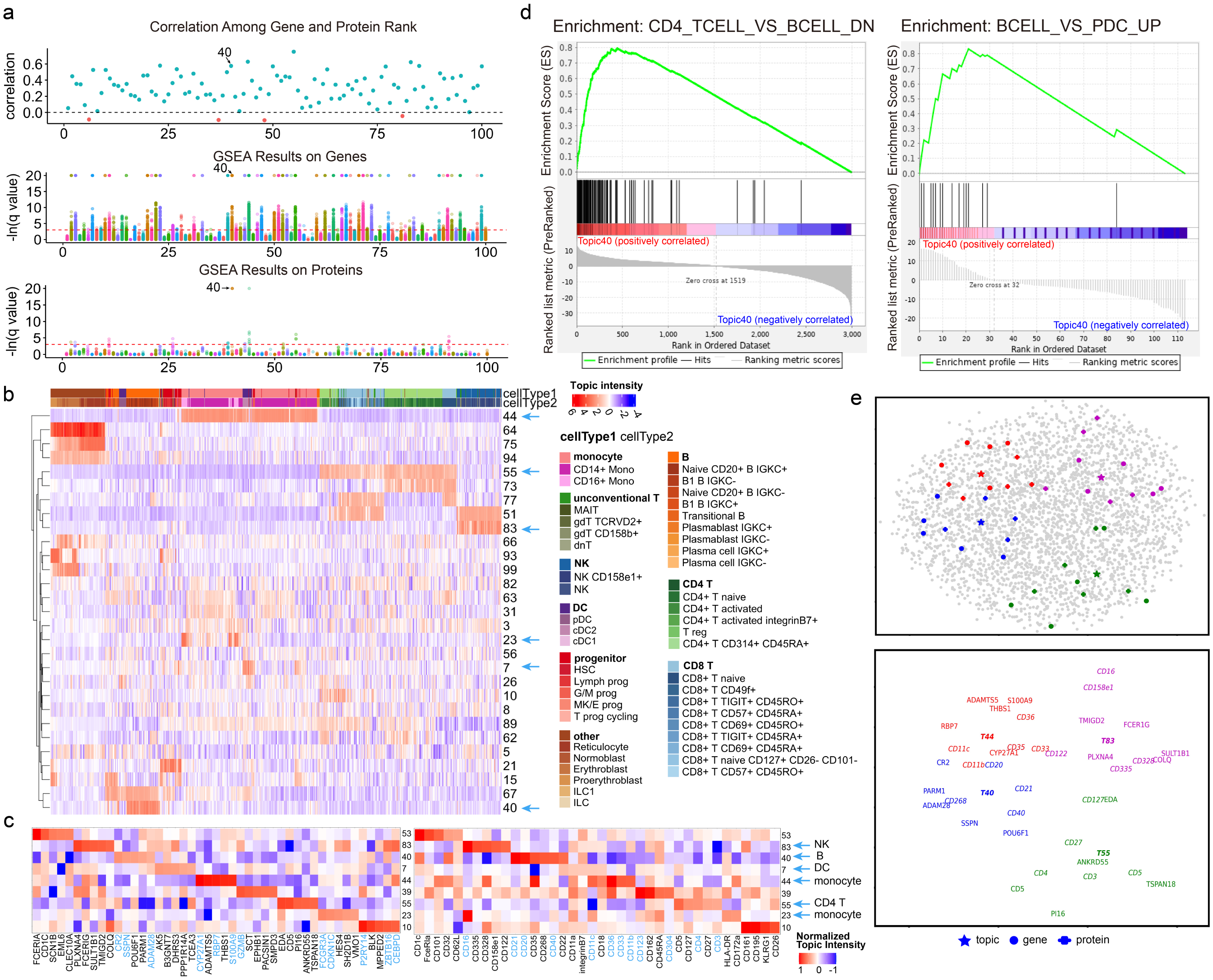

Figure 5: **Topic analysis of gene+protein CITE-seq data. a**. Protein-RNA correlations and pathway enrichments for the 100 topics learned from the CITE-seq BMMC2 data. In each plot, the x-axis is the 100 topics and the y-axis is either the protein-RNA correlation or the pathway enrichment scores in terms of -ln q-value. The top panel is the Spearman correlation between the RNA and protein expression for the same genes under each topic. Correlations above 0 are labeled blue and correlations below 0 are labeled red. The middle and the bottom panels are the corresponding GSEA enrichments of gene and protein topic scores, respectively. The dots correspond to the tested immunologic signature gene sets from MSigDB. Different colors represent different gene sets. **b**. Topics embedding of 10,000 sub-sampled cells from the BMMC2 dataset. Only the topics (rows) with the sum of absolute values greater than the third quartile across all sampled cells (columns) were shown. The two color bars display two tiers of annotations for the 9 broad cell types (cellType1) and 45 fine-grained cell types (cellType2). The topics that were labelled with arrows were described in the main text. **c**. Genes and proteins signatures of the select topics. The left and right panel display the topics-by-genes and topics-by-proteins heatmap, respectively. The top genes and proteins that are known cell-type markers based on CellMarker or literature search are highlighted in blue. For visualization purposes, we divided the topic values by the maximum absolute value within the same topic such that the topic scores range between -1 and 1. **d**. GSEA leading-edge analysis of Topic 40. The left panel is the GSEA result of gene topic scores on a significantly enriched gene set (q-value $<$ 0.001), which contains down-regulated genes in CD4 T cells relative to the CD19 B cells. Similarly, the right panel displays an enriched gene set (q-value $<$ 0.001), based on the protein topic scores for the same topic. The gene set contains up-regulated genes in B cells relative to plasmacytoid dendritic cells (pDC). **e**. UMAP visualization of the genes, proteins, and topics via their shared embedding space. Genes, proteins, and topics were labeled by star, circle and cross shapes on the top panel, respectively. Topics 40, 44, 55, 83 were colored in blue, red, green, and purple, respectively. The bottom panel displays the corresponding topic indices and gene symbols highlighted on the top panel.

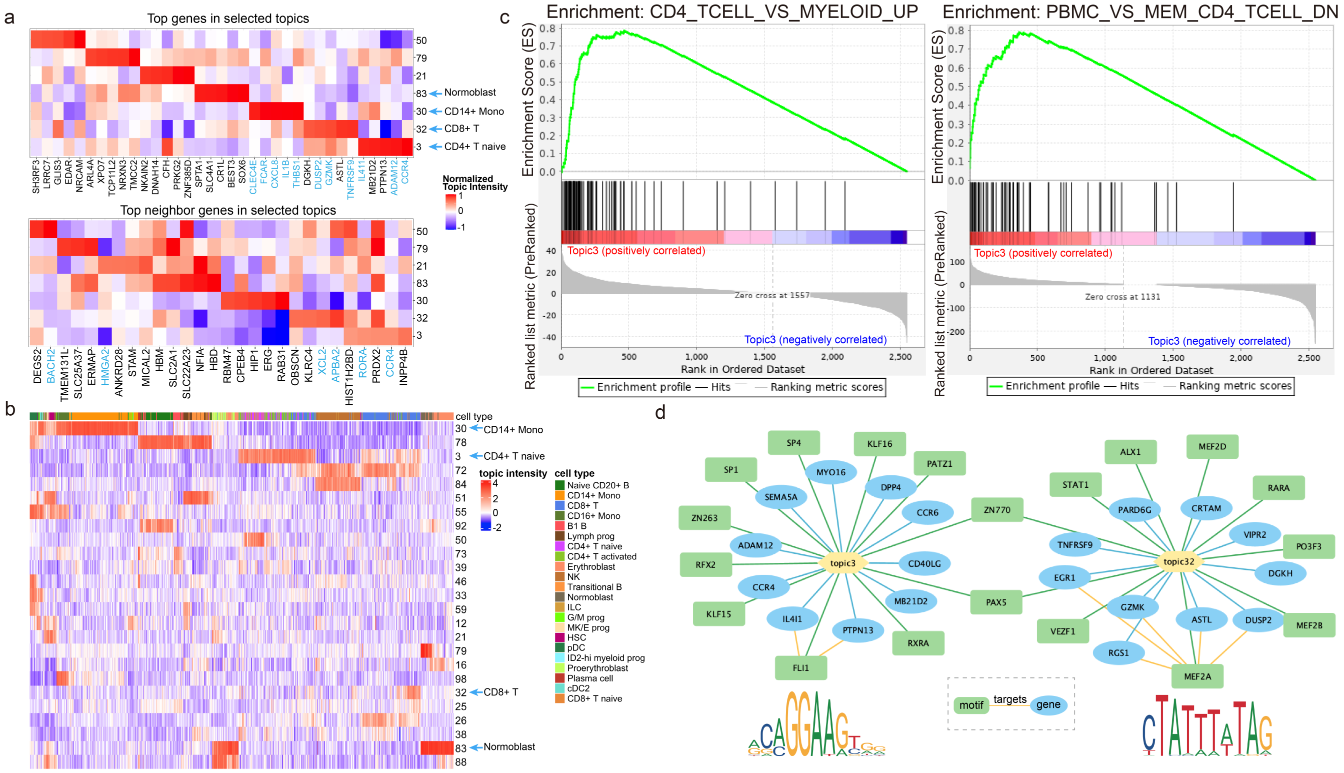

Figure 6: **Topic analysis of single-cell gene+peak data from the BMMC1 dataset. a.** Top genes and top peak-neighbour-genes of the select topics. The heatmap displays the top features (columns) for 7 out of 100 topics, which were selected based on their cell-type enrichments. The top signatures that are related to the enriched cell types based on CellMarker or literature search are highlighted in blue. For visualization purposes, we divided the topic values by the maximum absolute value within the same topic such that the topic scores range between -1 and 1. **b**. Topic embedding of cells from the BMMC1 dataset. The heatmap displays the embedding profiles of topics (rows) for 10,000 randomly sampled cells (columns) from the BMMC1 dataset. Only the topics with the sum of absolute values larger than the third quartile over the 10K cells are shown. The color bar on the top of the heatmap indicate the cell types with the text annotations shown in the legend. The columns and rows were ordered based on agglomerative hierarchical clustering with Euclidean distance and complete linkage **c**. GSEA leading edge analysis of Topic 3. The left panel is the GSEA result using gene topic scores and the right panel is the GSEA result using peak-neighboring-gene topic scores. The barcode in the middle are the genes that belong to the corresponding gene sets, namely the up-regulated genes in CD4 T cell relative to the Myeloid cells and the down-regulated genes in PBMC relative to the memory CD4 T cell for the gene and peak modalities of the same topic, respectively. **d**. Topic-directed regulatory networks based on motif enrichment analysis. The blue ellipses represent genes and the green rectangles represent enriched motifs. The bottom left and right motif logos correspond to the transcription factors (TFs) FLI1 and MEF2A, respectively. The yellow edges between motifs and genes indicate known TF-target associations based on ENCODE TF Targets dataset.

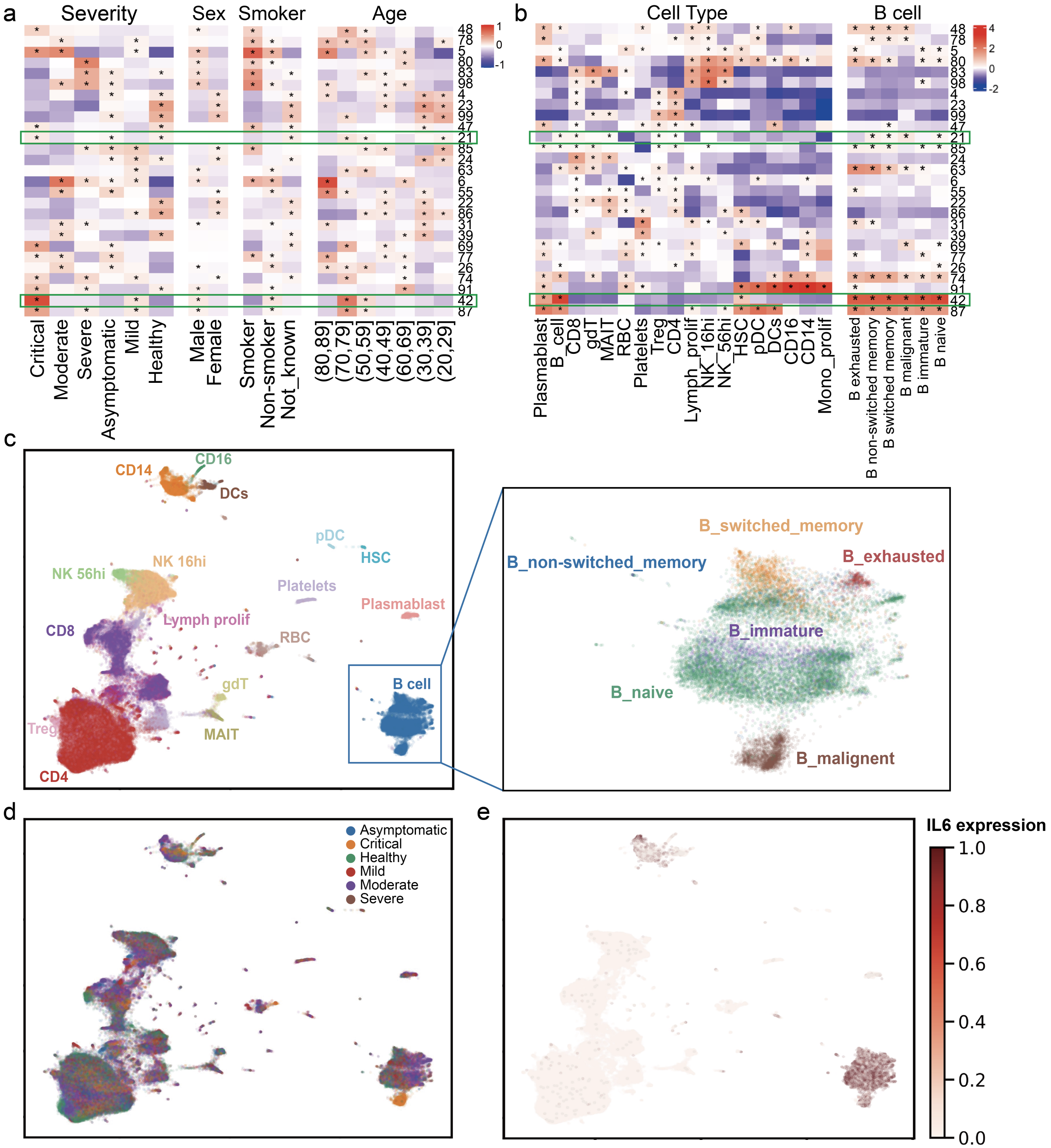

Figure 7: **Topic association with the COVID-19 severity status. a**. Differential analysis of severity states, sex, smoking history, and age. The color intensity values correspond to the differences of average topic scores between the positive cells and negative cells for each attribute (i.e., columns) and each topic (i.e., rows). Asterisks indicate Bonferroni-adjusted p-value $<$ 0.001 based on one-sided t-test of up-regulated topics for each label. The results on the highlighted topic 21 and 42 were described in the main text. **b**. Differential analysis of topics across cell types. The heatmap on the left displays the topic associations with each of the 18 cell types, and the one on the right associates the same topics with 6 fine-grained B-cell subtypes. Similarly, asterisks indicate adjusted p-value < 0.001 for the t-test of up-regulated topics in each label. **c**. UMAP visualization of cell clustering. Colors indicate 18 cell types. The right panel shows a zoom-in version of the B-cell clustering with color indicating the 6 B-cell subypes. **d**. UMAP visualization with cells colored by source subjects' severity states due to COVID-19 infection. **e**. Normalized gene expression of *IL6* among the cells on the same UMAP.

# 6 Tables

Table 1: Adjusted Rand Index (ARI) scores of cell clustering based on the embedding learned by 10 different models from four gene+peak datasets (i.e., columns). For each dataset, we split the cells into 60% training and 40% test cells. The experiments were evaluated based on 500 random splits to record the mean and standard deviation of the performances of each method. When comparing between moETM and six SOTA methods for each dataset, the highest ARI scores is in bold and the second highest is in blue. The 10 models (i.e., rows) include six SOTA methods, our proposed moETM using PoE, moETM using MoE, and two single-omic ETM models trained on only RNA and ATAC, respectively.

| Metrics | Methods | Genes + Peaks | | | |
|---|---|---|---|---|---|
| | | BMMC | MSLAC | MKC | MBC |
| ARI | SMILE | 0.719 (0.014) | 0.410 (0.012) | 0.409 (0.025) | 0.263 (0.025) |
| | scMM | 0.693 (0.024) | 0.396 (0.014) | 0.392 (0.026) | 0.311 (0.019) |
| | Cobolt | 0.678 (0.043) | 0.378 (0.014) | 0.373 (0.035) | 0.237 (0.048) |
| | MultiVI | 0.703 (0.029) | 0.405 (0.020) | 0.382 (0.030) | 0.500 (0.031) |
| | MOFA+ | 0.710 (0.012) | 0.421 (0.020) | 0.407 (0.041) | 0.244 (0.028) |
| | Seurat V4 | 0.712 (0.008) | **0.504** (0.096) | 0.398 (0.025) | **0.507** (0.021) |
| | moETM | **0.727** (0.007) | 0.469 (0.022) | **0.443** (0.025) | 0.423 (0.035) |
| | moETM-average | 0.720 (0.004) | 0.460 (0.047) | 0.439 (0.047) | 0.410 (0.043) |
| | moETM-rna | 0.720 (0.008) | 0.455 (0.053) | 0.433 (0.062) | 0.407 (0.054) |
| | moETM-atac | 0.626 (0.009) | 0.252 (0.021) | 0.135 (0.008) | 0.188 (0.011) |

Table 2: Adjusted Rand Index (ARI) scores of cell clustering of 3 CITE-seq gene+protein datasets. Same as in **Table** 1, we split the cells into 60% training and 40% testing and evaluated each method based on how well their learned embedding cluster cells into known cell types.

| Metrics | Methods | Genes + Proteins | | |
|---|---|---|---|---|
| | | BMMC | HWBC | HBIC |
| ARI | SMILE | 0.632 (0.020) | 0.397 (0.014) | 0.652 (0.030) |
| | scMM | 0.655 (0.048) | 0.602 (0.039) | 0.670 (0.057) |
| | Cobolt | 0.703 (0.036) | 0.572 (0.034) | 0.708 (0.043) |
| | TotalVI | 0.744 (0.027) | 0.638 (0.051) | 0.733 (0.049) |
| | MOFA+ | 0.713 (0.030) | 0.552 (0.038) | 0.687 (0.044) |
| | Seurat V4 | 0.670 (0.009) | 0.553 (0.014) | 0.703 (0.025) |
| | moETM | **0.760** (0.019) | **0.648** (0.023) | **0.752** (0.036) |
| | moETM-average | 0.748 (0.023) | 0.626 (0.031) | 0.737 (0.049) |
| | moETM-rna | 0.736 (0.030) | 0.617 (0.039) | 0.730 (0.040) |
| | moETM-protein | 0.593 (0.015) | 0.523 (0.025) | 0.604 (0.042) |

# References

1. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).

2. Stoeckius, M. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nature methods* **14**, 865–868 (2017).

3. Lähnemann, D. *et al.* Eleven grand challenges in single-cell data science. *Genome biology* **21**, 1–35 (2020).

4. Argelaguet, R., Cuomo, A. S., Stegle, O. & Marioni, J. C. Computational principles and challenges in single-cell data integration. *Nature biotechnology* **39**, 1202–1215 (2021).

5. Xu, Y., Das, P. & McCord, R. P. Smile: mutual information learning for integration of single-cell omics data. *Bioinformatics* **38**, 476–486 (2022).

6. Gayoso, A. *et al.* Joint probabilistic modeling of single-cell multi-omic data with totalvi. *Nature methods* **18**, 272–282 (2021).

7. Ashuach, T., Gabitto, M. I., Jordan, M. I. & Yosef, N. Multivi: deep generative model for the integration of multi-modal data. *bioRxiv* (2021).

8. Gong, B., Zhou, Y. & Purdom, E. Cobolt: integrative analysis of multimodal single-cell sequencing data. *Genome biology* **22**, 1–21 (2021).

9. Minoura, K., Abe, K., Nam, H., Nishikawa, H. & Shimamura, T. scmm: Mixture-of-experts multimodal deep generative model for single-cell multiomics data analysis. *bioRxiv* (2021).

10. Argelaguet, R. *et al.* Mofa+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome biology* **21**, 1–17 (2020).

11. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587 (2021).

12. Kingma, D. P. & Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

13. Wu, M. & Goodman, N. Multimodal generative models for scalable weakly-supervised learning. *Advances in Neural Information Processing Systems* **31** (2018).

14. Zhao, Y., Cai, H., Zhang, Z., Tang, J. & Li, Y. Learning interpretable cellular and gene signature embeddings from single-cell transcriptomic data. *Nature communications* **12**, 1–15 (2021).

15. Wu, K. E., Yost, K. E., Chang, H. Y. & Zou, J. Babel enables cross-modality translation between multiomic profiles at single-cell resolution. *Proceedings of the National Academy of Sciences* **118**, e2023070118 (2021).

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

17. Hubert, L. & Arabie, P. Comparing partitions. *Journal of classification* **2**, 193–218 (1985).

18. Danon, L., Diaz-Guilera, A., Duch, J. & Arenas, A. Comparing community structure identification. *Journal of statistical mechanics: Theory and experiment* **2005**, P09008 (2005).

19. Büttner, M., Miao, Z., Wolf, F. A., Teichmann, S. A. & Theis, F. J. A test metric for assessing single-cell rna-seq batch correction. *Nature methods* **16**, 43–49 (2019).

20. Luecken, M. D. *et al.* Benchmarking atlas-level data integration in single-cell genomics. *Nature methods* **19**, 41–50 (2022).

21. McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

22. Shema, E., Bernstein, B. E. & Buenrostro, J. D. Single-cell and single-molecule epigenomics to uncover genome regulation at unprecedented resolution. *Nature Genetics* **51**, 19–25 (2019).

23. Lynch, A. W. *et al.* Mira: Joint regulatory modeling of multimodal expression and chromatin accessibility in single cells. *Nature Methods* **19**, 1097–1108 (2022).

24. Ma, S. *et al.* Chromatin potential identified by shared single-cell profiling of rna and chromatin. *Cell* **183**, 1103–1116 (2020).

25. Zhang, X. *et al.* Cellmarker: a manually curated resource of cell markers in human and mouse. *Nucleic acids research* **47**, D721–D728 (2019).

26. Crowe, L. A. *et al.* S100a8 & s100a9: Alarmin mediated inflammation in tendinopathy. *Scientific reports* **9**, 1–12 (2019).

27. Wang, S. *et al.* S100a8/a9 in inflammation. *Frontiers in immunology* **9**, 1298 (2018).

28. Woo, M.-S., Yang, J., Beltran, C. & Cho, S. Cell surface cd36 protein in monocyte/macrophage contributes to phagocytosis during the resolution phase of ischemic stroke in mice. *Journal of Biological Chemistry* **291**, 23654–23661 (2016).

29. Ong, S.-M. *et al.* A novel, five-marker alternative to cd16–cd14 gating to identify the three human monocyte subsets. *Frontiers in immunology* **10**, 1761 (2019).

30. Hu, Y. *et al.* Genetic landscape and autoimmunity of monocytes in developing vogt–koyanagi–harada disease. *Proceedings of the National Academy of Sciences* **117**, 25712–25721 (2020).

31. Metcalf, T. U. *et al.* Human monocyte subsets are transcriptionally and functionally altered in aging in response to pattern recognition receptor agonists. *The Journal of Immunology* **199**, 1405–1417 (2017).

32. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102**, 15545–15550 (2005).

33. Mootha, V. K. *et al.* Pgc-1$\alpha$-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature genetics* **34**, 267–273 (2003).

34. Godec, J. *et al.* Compendium of immune signatures identifies conserved and species-specific biology in response to inflammation. *Immunity* **44**, 194–206 (2016).

35. Liberzon, A. *et al.* The molecular signatures database hallmark gene set collection. *Cell systems* **1**, 417–425 (2015).

36. Liberzon, A. *et al.* Molecular signatures database (msigdb) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).

37. Hutcheson, J. *et al.* Combined deficiency of proapoptotic regulators bim and fas results in the early onset of systemic autoimmunity. *Immunity* **28**, 206–217 (2008).

38. Nakaya, H. I. *et al.* Systems biology of vaccination for seasonal influenza in humans. *Nature immunology* **12**, 786–795 (2011).

39. Fröhlich, A. *et al.* Comprehensive analysis of tumor necrosis factor receptor tnfrsf9 (4-1bb) dna methylation with regard to molecular and clinicopathological features, immune infiltrates, and response prediction to immunotherapy in melanoma. *EBioMedicine* **52**, 102647 (2020).

40. Lang, R. & Raffi, F. A. Dual-specificity phosphatases in immunity and infection: an update. *International journal of molecular sciences* **20**, 2710 (2019).

41. Cari, L., Nocentini, G., Migliorati, G. & Riccardi, C. Potential effect of tumor-specific treg-targeted antibodies in the treatment of human cancers: A bioinformatics analysis. *Oncoimmunology* **7**, e1387705 (2018).

42. Fox, J. C. *et al.* Structural and agonist properties of xcl2, the other member of the c-chemokine subfamily. *Cytokine* **71**, 302–311 (2015).

43. Song, K. *et al.* Characterization of subsets of cd4+ memory t cells reveals early branched pathways of t cell differentiation in humans. *Proceedings of the National Academy of Sciences* **102**, 7916–7921 (2005).

44. Liu, Y. *et al.* Adam12 is a costimulatory molecule that determines th1 cell fate and mediates tissue inflammation. *Cellular & molecular immunology* **18**, 1904–1919 (2021).

45. Puiffe, M.-L. *et al.* Il4i1 accelerates the expansion of effector cd8+ t cells at the expense of memory precursors by increasing the threshold of t-cell activation. *Frontiers in immunology* **11**, 600012 (2020).

46. Haim-Vilmovsky, L. *et al.* Mapping rora expression in resting and activated cd4+ t cells. *PloS one* **16**, e0251233 (2021).

47. Abbas, A. R., Wolslegel, K., Seshasayee, D., Modrusan, Z. & Clark, H. F. Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. *PloS one* **4**, e6098 (2009).

48. Kaech, S. M., Wherry, E. J. & Ahmed, R. Effector and memory t-cell differentiation: implications for vaccine development. *Nature Reviews Immunology* **2**, 251–262 (2002).

49. Feingold, E. & Pachter, L. The encode (encyclopedia of dna elements) project. *Science* **306**, 636–640 (2004).

50. Consortium, E. P. A user's guide to the encyclopedia of dna elements (encode). *PLoS biology* **9**, e1001046 (2011).

51. Köhler, S. *et al.* The human phenotype ontology in 2021. *Nucleic acids research* **49**, D1207–D1217 (2021).

52. Good, K. L., Avery, D. T. & Tangye, S. G. Resting human memory b cells are intrinsically programmed for enhanced survival and responsiveness to diverse stimuli compared to naive b cells. *The Journal of Immunology* **182**, 890–901 (2009).

53. Stephenson, E. *et al.* Single-cell multi-omics analysis of the immune response in covid-19. *Nature medicine* **27**, 904–916 (2021).

54. Kanaji, S., Fahs, S. A., Shi, Q., Haberichter, S. L. & Montgomery, R. Contribution of platelet vs. endothelial vwf to platelet adhesion and hemostasis. *Journal of Thrombosis and Haemostasis* **10**, 1646–1652 (2012).

55. O'Leary, N. A. *et al.* Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation. *Nucleic acids research* **44**, D733–D745 (2016).

56. Hodson, D. J. *et al.* Regulation of normal b-cell differentiation and malignant b-cell survival by oct2. *Proceedings of the National Academy of Sciences* **113**, E2039–E2046 (2016).

57. Bonuomo, V. *et al.* Covid-19 (sars-cov-2 infection) in lymphoma patients: A review. *World Journal of Virology* **10**, 312 (2021).

58. Jones, S. A. & Hunter, C. A. Is il-6 a key cytokine target for therapy in covid-19? *Nature Reviews Immunology* **21**, 337–339 (2021).

59. Sabaka, P. *et al.* Role of interleukin 6 as a predictive factor for a severe course of covid-19: retrospective data analysis of patients from a long-term care facility during covid-19 outbreak. *BMC infectious diseases* **21**, 1–8 (2021).

60. Boyd, S. D., Natkunam, Y., Allen, J. R. & Warnke, R. A. Selective immunophenotyping for diagnosis of b-cell neoplasms: immunohistochemistry and flow cytometry strategies and results. *Applied immunohistochemistry & molecular morphology: AIMM/official publication of the Society for Applied Immunohistochemistry* **21**, 116 (2013).

61. Laing, A. G. *et al.* A dynamic covid-19 immune signature includes associations with poor prognosis. *Nature medicine* **26**, 1623–1635 (2020).

62. Paul Hannan, J. The structure-function relationships of complement receptor type 2 (cr2; cd21). *Current Protein and Peptide Science* **17**, 463–487 (2016).

63. Wang, K., Wei, G. & Liu, D. Cd19: a biomarker for b cell development, lymphoma diagnosis and therapy. *Experimental hematology & oncology* **1**, 1–7 (2012).

64. Lasoudris, F. *et al.* Il4i1: an inhibitor of the cd8+ antitumor t-cell response in vivo. *European journal of immunology* **41**, 1629–1638 (2011).

65. Zhang, X. K. *et al.* The transcription factor fli-1 modulates marginal zone and follicular b cell development in mice. *The Journal of Immunology* **181**, 1644–1654 (2008).

66. Wen, H. *et al.* Graph neural networks for multimodal single-cell data integration. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, 4153–4163 (Association for Computing Machinery, New York, NY, USA, 2022). URL `https://doi.org/10.1145/3534678.3539213`.

67. Cao, Z.-J. & Gao, G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nature Biotechnology* 1–9 (2022).

68. Wang, J. *et al.* scgnn is a novel graph neural network framework for single-cell rna-seq analyses. *Nature communications* **12**, 1–11 (2021).

69. Devlin, J., Chang, M., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C. & Solorio, T. (eds.) *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 4171–4186 (Association for Computational Linguistics, 2019). URL `https://doi.org/10.18653/v1/n19-1423`.

70. Yang, F. *et al.* scbert as a large-scale pretrained deep language model for cell type annotation of single-cell rna-seq data. *Nature Machine Intelligence* **4**, 852–866 (2022).

71. Ji, Y., Zhou, Z., Liu, H. & Davuluri, R. V. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics* **37**, 2112–2120 (2021).

72. Avsec, Ž. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods* **18**, 1196–1203 (2021).

73. Luecken, M. D. *et al.* A sandbox for prediction and integration of dna, rna, and proteins in single cells. In *NeurIPS 2021 Datasets and Benchmarks Track (Round 2)* (2021).

74. Cao, J. *et al.* Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* **361**, 1380–1385 (2018).

75. Wolf, F. A., Angerer, P. & Theis, F. J. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology* **19**, 1–5 (2018).

76. Romano, S., Vinh, N. X., Bailey, J. & Verspoor, K. Adjusting for chance clustering comparison measures. *The Journal of Machine Learning Research* **17**, 4635–4666 (2016).

77. Lawrence, M. *et al.* Software for computing and annotating genomic ranges. *PLoS computational biology* **9**, e1003118 (2013).

78. Bailey, T. L. & Grant, C. E. Sea: Simple enrichment analysis of motifs. *bioRxiv* (2021).

79. Bailey, T. L., Johnson, J., Grant, C. E. & Noble, W. S. The meme suite. *Nucleic acids research* **43**, W39–W49 (2015).

80. Bailey, T. L. Streme: accurate and versatile sequence motif discovery. *Bioinformatics* **37**, 2834–2840 (2021).

81. Kulakovskiy, I. V. *et al.* Hocomoco: towards a complete collection of transcription factor binding models for human and mouse via large-scale chip-seq analysis. *Nucleic acids research* **46**, D252–D259 (2018).

# Supplementary Information

Manqi Zhou[1,†], Hao Zhang[2,†], Zilong Bai[2], Dylan Mann-Krzisnik[3], Fei Wang[2,*], Yue Li[2,3,4,*]

[1]Department of Computational Biology, Cornell University [2]Division of Health Informatics, Department of Population Health Sciences, Weill Cornell Medicine [3]Quantitative Life Science, McGill University [4]School of Computer Science, McGill University [5]Mila - Quebec AI Institute

[†]Equal contribution

[*]Correspondence: few2001@med.cornell.edu, yueli@cs.mcgill.ca

# S1 Supplementary Tables

Table S1: Evaluation of cell clustering by NMI, kBET, and GC on 4 genes+peaks single-cell multi-omic datasets. The experiments were the same as described in **Table** 1. Only the BMMC data have multiple batches (13 batches) and therefore evaluated by the kBET and GC scores. The other 3 datasets were evaluated by NMI and ARI only based on their cell-type labels (**Table** 1).

| Metrics | Methods | Genes + Peaks | | | |
|---|---|---|---|---|---|
| | | BMMC | MSLAC | MKC | MBC |
| NMI | moETM | **0.787** (0.005) | 0.612 (0.013) | **0.513** (0.022) | 0.590 (0.025) |
| | SMILE | 0.780 (0.006) | 0.562 (0.008) | 0.460 (0.008) | 0.420 (0.019) |
| | scMM | 0.754 (0.010) | 0.550 (0.016) | 0.471 (0.017) | 0.501 (0.016) |
| | Cobolt | 0.744 (0.005) | 0.526 (0.020) | 0.452 (0.029) | 0.398 (0.053) |
| | MultiVI | 0.763 (0.010) | 0.572 (0.007) | 0.461 (0.024) | **0.653** (0.018) |
| | MOFA+ | 0.777 (0.013) | 0.570 (0.013) | 0.480 (0.018) | 0.401 (0.023) |
| | Seurat V4 | 0.782 (0.006) | **0.673** (0.009) | 0.463 (0.032) | 0.651 (0.035) |
| | moETM-average | 0.777 (0.011) | 0.605 (0.018) | 0.503 (0.030) | 0.569 (0.033) |
| | moETM-rna | 0.773 (0.005) | 0.599 (0.045) | 0.498 (0.055) | 0.562 (0.032) |
| | moETM-atac | 0.705 (0.009) | 0.482 (0.006) | 0.172 (0.009) | 0.123 (0.015) |
| kBET | moETM | **0.223** (0.012) | - | - | - |
| | SMILE | 0.137 (0.018) | - | - | - |
| | scMM | 0.184 (0.016) | - | - | - |
| | Cobolt | 0.115 (0.010) | - | - | - |
| | MultiVI | 0.092 (0.013) | - | - | - |
| | MOFA+ | 0.193 (0.025) | - | - | - |
| | Seurat V4 | 0.190 (0.009) | - | - | - |
| | moETM-average | 0.207 (0.015) | - | - | - |
| | moETM-rna | 0.198 (0.019) | - | - | - |
| | moETM-atac | 0.205 (0.021) | - | - | - |
| GC | moETM | **0.978** (0.003) | - | - | - |
| | SMILE | 0.953 (0.003) | - | - | - |
| | scMM | 0.964 (0.002) | - | - | - |
| | Cobolt | 0.962 (0.004) | - | - | - |
| | MultiVI | 0.960 (0.005) | - | - | - |
| | MOFA+ | 0.954 (0.011) | - | - | - |
| | Seurat V4 | 0.959 (0.003) | - | - | - |
| | moETM-average | 0.966 (0.005) | - | - | - |
| | moETM-rna | 0.950 (0.004) | - | - | - |
| | moETM-atac | 0.962 (0.007) | - | - | - |

Table S2: Evaluation of cell clustering of 3 CITE-seq gene+protein datasets based on NMI, kBET, and GC. The experiments were the same as described in **Table** 2.

| Metrics | Methods | Genes + Proteins | | |
|---|---|---|---|---|
| | | BMMC | HWBC | HBIC |
| NMI | moETM | **0.786** (0.006) | **0.791** (0.013) | **0.779** (0.023) |
| | SMILE | 0.712 (0.007) | 0.732 (0.005) | 0.702 (0.017) |
| | scMM | 0.708 (0.012) | 0.742 (0.018) | 0.712 (0.034) |
| | Cobolt | 0.731 (0.011) | 0.730 (0.027) | 0.739 (0.030) |
| | TotalVI | 0.770 (0.005) | 0.784 (0.010) | 0.762 (0.019) |
| | MOFA+ | 0.750 (0.024) | 0.701 (0.027) | 0.715 (0.029) |
| | Seurat V4 | 0.714 (0.006) | 0.714 (0.040) | 0.726 (0.012) |
| | moETM-average | 0.770 (0.011) | 0.769 (0.018) | 0.753 (0.029) |
| | moETM-rna | 0.755 (0.004) | 0.759 (0.010) | 0.748 (0.019) |
| | moETM-protein | 0.625 (0.013) | 0.620 (0.019) | 0.608 (0.027) |
| kBET | moETM | 0.105 (0.002) | 0.304 (0.012) | 0.153 (0.020) |
| | SMILE | 0.058 (0.001) | 0.032 (0.007) | 0.073 (0.017) |
| | scMM | 0.080 (0.004) | 0.110 (0.007) | 0.092 (0.011) |
| | Cobolt | 0.036 (0.002) | 0.095 (0.005) | 0.038 (0.019) |
| | TotalVI | **0.156** (0.009) | **0.349** (0.002) | **0.197** (0.004) |
| | MOFA+ | 0.078 (0.013) | 0.213 (0.010) | 0.063 (0.014) |
| | Seurat V4 | 0.062 (0.003) | 0.133 (0.010) | 0.107 (0.015) |
| | moETM-average | 0.094 (0.005) | 0.290 (0.016) | 0.140 (0.023) |
| | moETM-rna | 0.082 (0.002) | 0.266 (0.010) | 0.130 (0.015) |
| | moETM-protein | 0.090 (0.006) | 0.280 (0.015) | 0.135 (0.023) |
| GC | moETM | **0.936** (0.007) | **0.968** (0.004) | **0.950** (0.005) |
| | SMILE | 0.901 (0.011) | 0.912 (0.016) | 0.907 (0.006) |
| | scMM | 0.897 (0.009) | 0.937 (0.016) | 0.906 (0.008) |
| | Cobolt | 0.880 (0.011) | 0.930 (0.019) | 0.890 (0.010) |
| | TotalVI | 0.918 (0.005) | 0.951 (0.006) | 0.934 (0.007) |
| | MOFA+ | 0.906 (0.022) | 0.940 (0.022) | 0.895 (0.015) |
| | Seurat V4 | 0.911 (0.004) | 0.930 (0.002) | 0.923 (0.003) |
| | moETM-average | 0.920 (0.009) | 0.949 (0.006) | 0.936 (0.008) |
| | moETM-rna | 0.911 (0.010) | 0.932 (0.007) | 0.928 (0.009) |
| | moETM-atac | 0.916 (0.011) | 0.939 (0.007) | 0.930 (0.015) |

Table S3: Evaluation of embedding-based cell clustering of the 4 gene+peak datasets. Different from the one listed in **Table** 1 and **Supplementary Table** S1, we trained and tested each model on all of the cells from each dataset. Because the cell type labels were not used in training any of the model, this is still an unbiased evaluation.

| Metrics | Methods | Genes + Peaks | | | |
|---|---|---|---|---|---|
| | | BMMC | MSLAC | MKC | MBC |
| ARI | moETM | **0.735** (0.006) | 0.515 (0.008) | **0.584** (0.019) | 0.468 (0.019) |
| | SMILE | 0.723 (0.009) | 0.477 (0.009) | 0.439 (0.024) | 0.301 (0.018) |
| | scMM | 0.693 (0.007) | 0.412 (0.005) | 0.420 (0.017) | 0.333 (0.026) |
| | Cobolt | 0.664 (0.009) | 0.400 (0.010) | 0.394 (0.024) | 0.303 (0.039) |
| | MultiVI | 0.697 (0.007) | 0.413 (0.013) | 0.403 (0.014) | 0.502 (0.025) |
| | MOFA+ | 0.709 (0.007) | 0.489 (0.010) | 0.424 (0.017) | 0.403 (0.026) |
| | Seurat V4 | 0.706 (0.008) | **0.547** (0.001) | 0.403 (0.001) | **0.538** (0.0001) |
| NMI | moETM | **0.798** (0.005) | 0.662 (0.008) | **0.643** (0.020) | 0.601 (0.022) |
| | SMILE | 0.784 (0.009) | 0.617 (0.011) | 0.47 (0.028) | 0.445 (0.036) |
| | scMM | 0.744 (0.006) | 0.583 (0.014) | 0.488 (0.025) | 0.524 (0.031) |
| | Cobolt | 0.738 (0.009) | 0.568 (0.015) | 0.464 (0.020) | 0.428 (0.066) |
| | MultiVI | 0.755 (0.008) | 0.580 (0.020) | 0.470 (0.026) | 0.669 (0.013) |
| | MOFA+ | 0.769 (0.006) | 0.631 (0.026) | 0.496 (0.021) | 0.574 (0.028) |
| | Seurat V4 | 0.782 (0.008) | **0.702** (0.001) | 0.488 (0.001) | **0.694** (0.0001) |
| kBET | moETM | **0.234** (0.008) | - | - | - |
| | SMILE | 0.115 (0.005) | - | - | - |
| | scMM | 0.133 (0.007) | - | - | - |
| | Cobolt | 0.102 (0.011) | - | - | - |
| | MultiVI | 0.068 (0.011) | - | - | - |
| | MOFA+ | 0.163 (0.011) | - | - | - |
| | Seurat V4 | 0.153 (0.002) | - | - | - |
| GC | moETM | **0.974** (0.003) | - | - | - |
| | SMILE | 0.965 (0.007) | - | - | - |
| | scMM | 0.970 (0.008) | - | - | - |
| | Cobolt | 0.961 (0.010) | - | - | - |
| | MultiVI | 0.957 (0.004) | - | - | - |
| | MOFA+ | 0.964 (0.008) | - | - | - |
| | Seurat V4 | 0.952 (0.004) | - | - | - |

Table S4: Evaluation of embedding-based clustering by leaving-one-**subject**-out (Section 4.3). Each method was trained on $B-1$ subjects and tested on the held-out subject. The highest and second highest score per dataset were highlighted in bold and blue, respectively.

| Metrics | Methods | Genes + Peaks | Genes + Proteins |
|---|---|---|---|
| | | BMMC | BMMC |
| ARI | moETM | **0.779** (0.071) | **0.776** (0.071) |
| | SMILE | 0.766 (0.082) | 0.703 (0.099) |
| | scMM | 0.743 (0.120) | 0.737 (0.084) |
| | Cobolt | 0.732 (0.105) | 0.726 (0.013) |
| | MultiVI/TotalVI | 0.739 (0.095) | 0.740 (0.107) |
| | MOFA+ | 0.759 (0.098) | 0.713 (0.034) |
| | Seurat V4 | 0.730 (0.095) | 0.686 (0.120) |
| NMI | moETM | **0.819** (0.037) | **0.823** (0.029) |
| | SMILE | 0.808 (0.036) | 0.803 (0.030) |
| | scMM | 0.780 (0.056) | 0.804 (0.026) |
| | Cobolt | 0.750 (0.150) | 0.799 (0.034) |
| | MultiVI/TotalVI | 0.774 (0.043) | 0.810 (0.038) |
| | MOFA+ | 0.783 (0.042) | 0.758 (0.027) |
| | Seurat V4 | 0.782 (0.050) | 0.744 (0.023) |

Table S5: Evaluation of embedding-based clustering by leaving-one-**site**-out (Section 4.3). Each method was trained on $B-1$ sites and tested on the held-out site. The highest and second highest score per dataset were highlighted in bold and blue, respectively.

| Metrics | Methods | Genes + Peaks | Genes + Proteins |
|---|---|---|---|
| | | BMMC | BMMC |
| ARI | moETM | **0.735** (0.084) | **0.772** (0.053) |
| | SMILE | 0.714 (0.095) | 0.686 (0.075) |
| | scMM | 0.700 (0.106) | 0.665 (0.066) |
| | Cobolt | 0.693 (0.068) | 0.653 (0.057) |
| | MultiVI/TotalVI | 0.704 (0.086) | 0.694 (0.085) |
| | MOFA+ | 0.716 (0.078) | 0.668 (0.073) |
| | Seurat V4 | 0.693 (0.097) | 0.666 (0.149) |
| NMI | moETM | **0.796** (0.036) | **0.810** (0.033) |
| | SMILE | 0.780 (0.041) | 0.798 (0.040) |
| | scMM | 0.760 (0.033) | 0.771 (0.022) |
| | Cobolt | 0.768 (0.062) | 0.759 (0.043) |
| | MultiVI/TotalVI | 0.762 (0.038) | 0.787 (0.050) |
| | MOFA+ | 0.780 (0.042) | 0.768 (0.041) |
| | Seurat V4 | 0.756 (0.038) | 0.737 (0.010) |

Table S6: Imputing surface protein expression from gene expression. The best score per evaluation metric is in bold.

| Methods | random split | | leave-one-batch | | leave-one-cell-type | |
|---|---|---|---|---|---|---|
| | Pearson | Spearman | Pearson | Spearman | Pearson | Spearman |
| moETM | **0.95** (0.02) | **0.94** (0.02) | **0.92** (0.02) | **0.90** (0.03) | **0.88** (0.04) | **0.85** (0.03) |
| BABEL | 0.94 (0.03) | 0.92 (0.03) | 0.90 (0.02) | 0.87 (0.03) | 0.85 (0.03) | 0.81 (0.03) |
| scMM | 0.94 (0.03) | 0.91 (0.04) | 0.91 (0.03) | 0.89 (0.03) | 0.83 (0.04) | 0.78 (0.05) |

Table S7: Imputing gene expression from chromatin accessibility. The best score per evaluation metric is in bold.

| Methods | random split | | leave-one-batch | | leave-one-cell-type | |
|---|---|---|---|---|---|---|
| | Pearson | Spearman | Pearson | Spearman | Pearson | Spearman |
| moETM | **0.69** (0.02) | **0.37** (0.03) | **0.65** (0.04) | **0.35** (0.04) | **0.58** (0.05) | **0.32** (0.03) |
| BABEL | 0.65 (0.03) | 0.34 (0.03) | 0.6 (0.03) | 0.33 (0.03) | 0.55 (0.05) | 0.30(0.02) |
| scMM | 0.63 (0.03) | 0.33 (0.04) | 0.61 (0.04) | 0.33 (0.03) | 0.54 (0.05) | 0.28 (0.04) |

Table S8: Imputing chromatin accessibility from gene expression. The best score per evaluation metric is in bold.

| Methods | random split | | leave-one-batch | | leave-one-cell-type | |
|---|---|---|---|---|---|---|
| | Pearson | Spearman | Pearson | Spearman | Pearson | Spearman |
| moETM | **0.58** (0.03) | **0.33** (0.02) | **0.55** (0.04) | **0.30** (0.03) | **0.51** (0.03) | **0.28** (0.04) |
| BABEL | 0.38 (0.02) | 0.27 (0.03) | 0.34 (0.03) | 0.23 (0.02) | 0.31 (0.03) | 0.18 (0.03) |
| scMM | 0.40 (0.03) | 0.29 (0.03) | 0.37 (0.04) | 0.25 (0.03) | 0.33 (0.03) | 0.21 (0.03) |

Table S9: Imputing gene expression from surface protein expression. The best score per evaluation metric is in bold.

| Methods | random split | | leave-one-batch | | leave-one-cell-type | |
|---|---|---|---|---|---|---|
| | Pearson | Spearman | Pearson | Spearman | Pearson | Spearman |
| moETM | **0.65** (0.03) | **0.41** (0.03) | **0.63** (0.02) | **0.39** (0.03) | **0.60** (0.03) | **0.37** (0.02) |
| BABEL | 0.62 (0.03) | 0.39 (0.03) | 0.60 (0.04) | 0.37 (0.03) | 0.57 (0.02) | 0.33 (0.03) |
| scMM | 0.61 (0.02) | 0.37 (0.03) | 0.60 (0.03) | 0.36 (0.04) | 0.54 (0.04) | 0.30 (0.03) |

# S2 Supplementary Figures

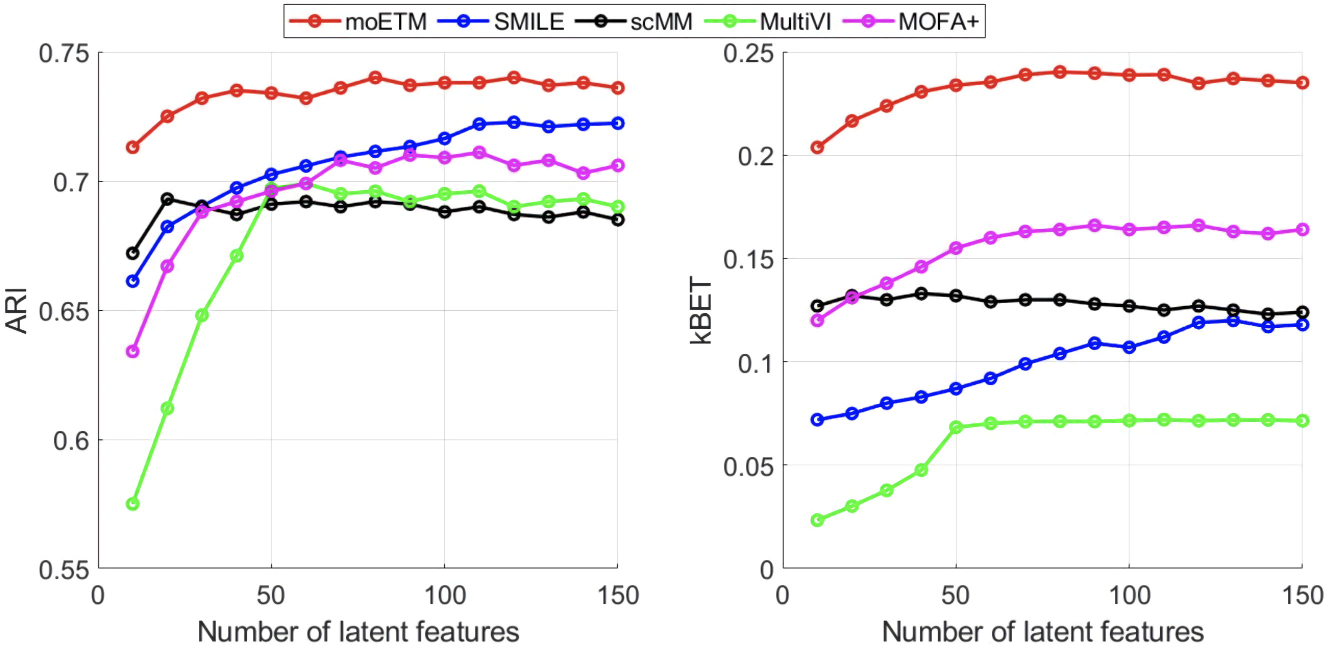

Figure S1: **Evaluation metric values as a function of the numbers of latent dimensions on the BMMC1 dataset.**

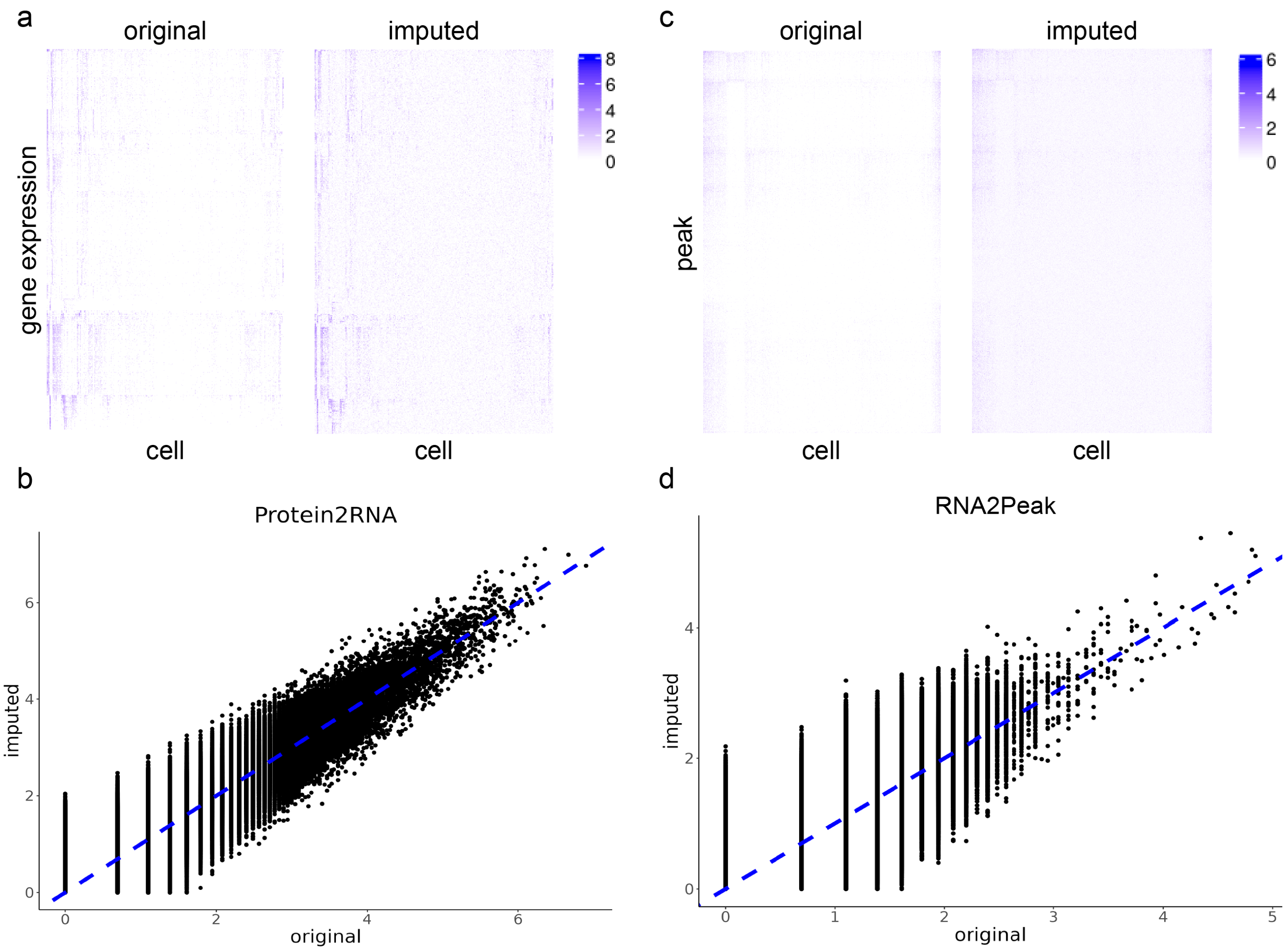

Figure S2: **Cross-omic imputation from low dimension to high dimension. a**. Heatmap of original and imputed gene expression from protein values using the BMMC2 CITE-seq dataset. In each heatmap, the columns are the randomly sampled 5000 cells, and the rows are genes. The order of cells and genes are the same in the two heatmaps. **b**. Scatter plot of original and imputed gene expression values. The x-axis is the original expression and the y-axis represents the imputed values. The diagonal line is in blue. **c**. Heatmap of the original and imputed peak values from gene expression on the BMMC1 dataset. Rows are peaks and columns are randomly sampled 5000 cells. **d**. Scatter plot of the original and imputed peak values.

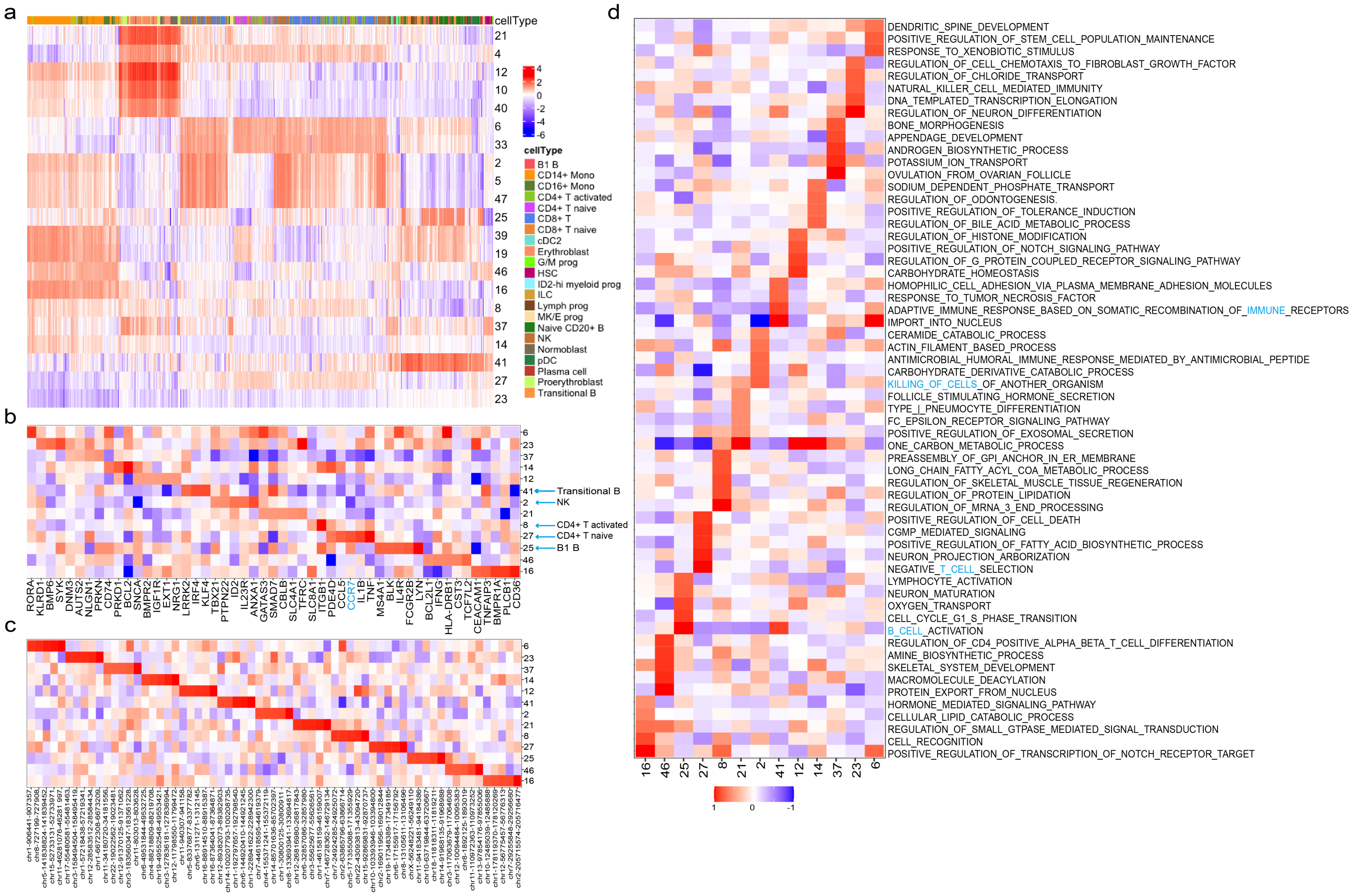

Figure S3: **GO-informed cell and features embedding learned by our moETM on the BMMC1 dataset.** The gene embedding matrix of moETM was fixed to the gene sets of Gene Ontology Biological Processes from MSigDB during the training on the BMMC1 gene+peak data. **a**. Cell topic-embedding. Columns are cells and rows are topics. The top bar indicate the cell types. Color intensities are proportional to the topic embedding of the cells. **b**. Top genes for the select topics. Rows are the select topics and columns are the top 5 genes per topic. Marker genes were colored in blue. Cell-type-specific topics were labeled by the arrows. **c**. Top peaks for the same set of topics as in panel b. **d**. The top 5 pathways for each of the selected topics. Cell-type-associated pathways were colored in blue. The color intensities are the topic embedding values for the pathways.

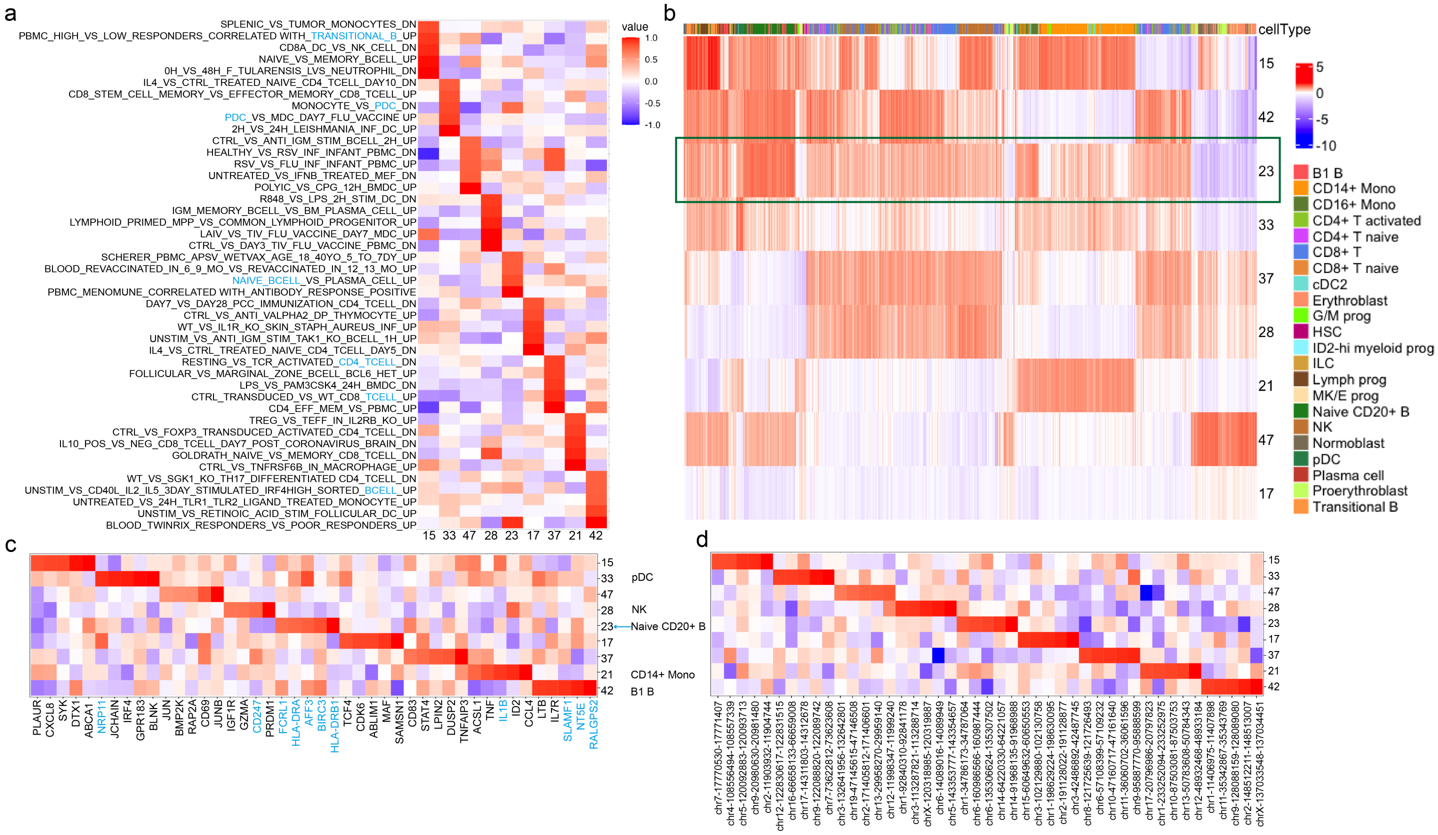

Figure S4: **GO-informed cell and features embedding learned by our moETM on the BMMC1 dataset. a**. The top 5 pathways for each of the selected topics. Cell-type-associated pathways were colored in blue. The color intensities are the topic embedding values for the pathways. **b**. Cell topic-embedding. Columns are cells and rows are topics. The top bar indicate the cell types. Color intensities are proportional to the topic embedding of the cells. The highlighted topic 23 was discussed in the main text. **c**. Top genes for the select topics. Rows are the select topics and columns are the top 5 genes per topic. Marker genes were colored in blue. Cell-type-specific topics were labeled by the arrows. **d**. The top 5 pathways for each of the selected topics. Cell-type-associated pathways were colored in blue. The color intensities are the topic embedding values for the pathways.

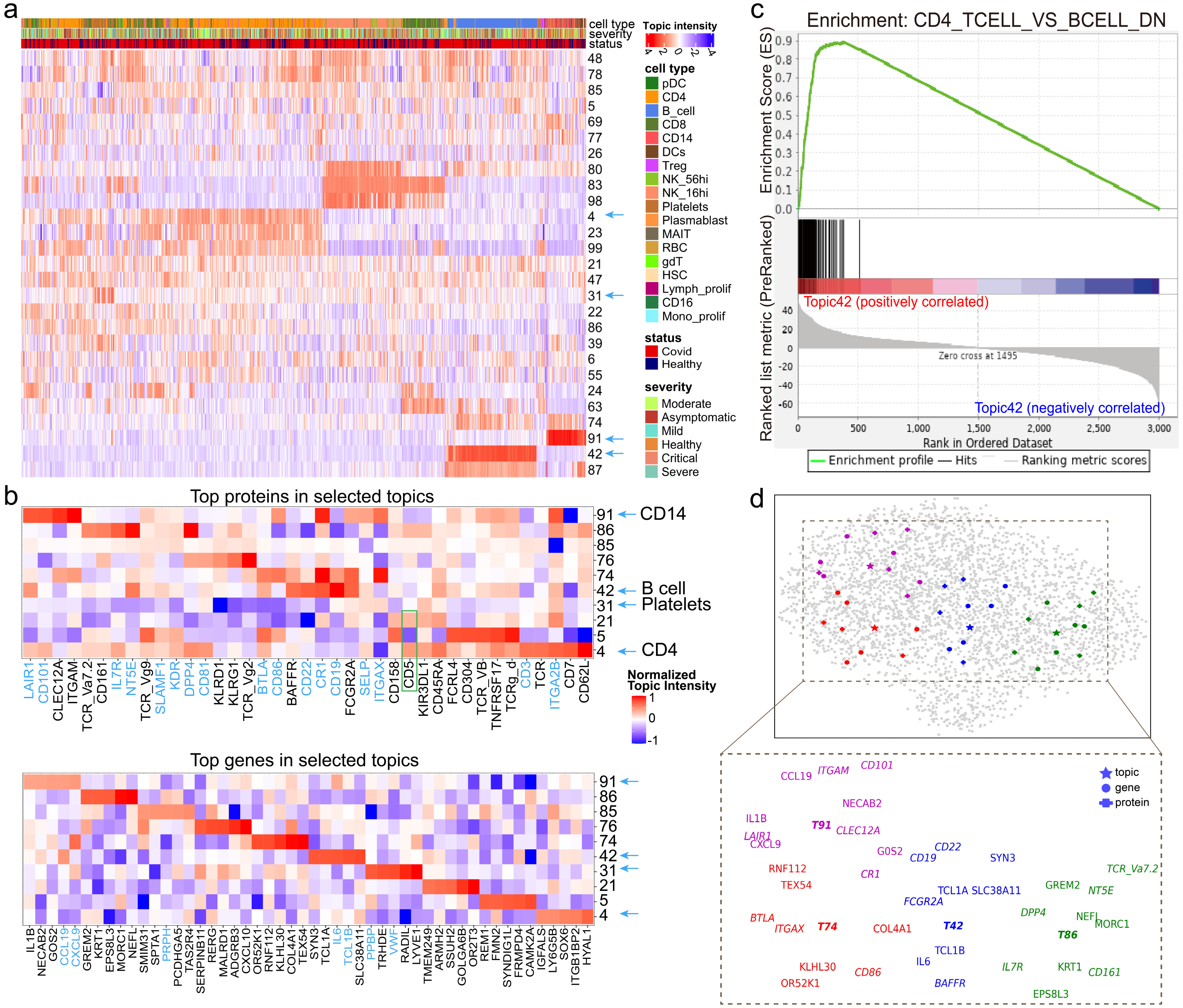

Figure S5: **Topic analysis of single-cell COVID-19 CITE-seq dataset. a**. Topics intensity of cells sub-sampled (n=10000). Only the topics with the sum of absolute values larger than the third quartile across all sampled cells were shown. The three color bars show cell types, disease severity, and disease status. **b**. Top proteins and top genes per select topic. The marker genes and proteins based on CellMarker or literature search are colored in blue. For visualization purposes, we divided the topic values by the maximum absolute value within the same topic. **c**. GSEA lead-edge analysis of topic 42. The enriched gene set contains genes that were down-regulated in CD4 T cells relative to B cells. The barcode in the middle are the genes that belong to the corresponding gene set. **d**. UMAP visualization of the genes, proteins, and topics via their shared embedding space. Genes, proteins, and topics were labeled by star, circle and cross shapes on the top panel, respectively. Topics 42, 74, 86, and 91 were colored in blue, red, green, and purple, respectively. The bottom panel displays the corresponding topic indices and gene symbols highlighted on the top panel.