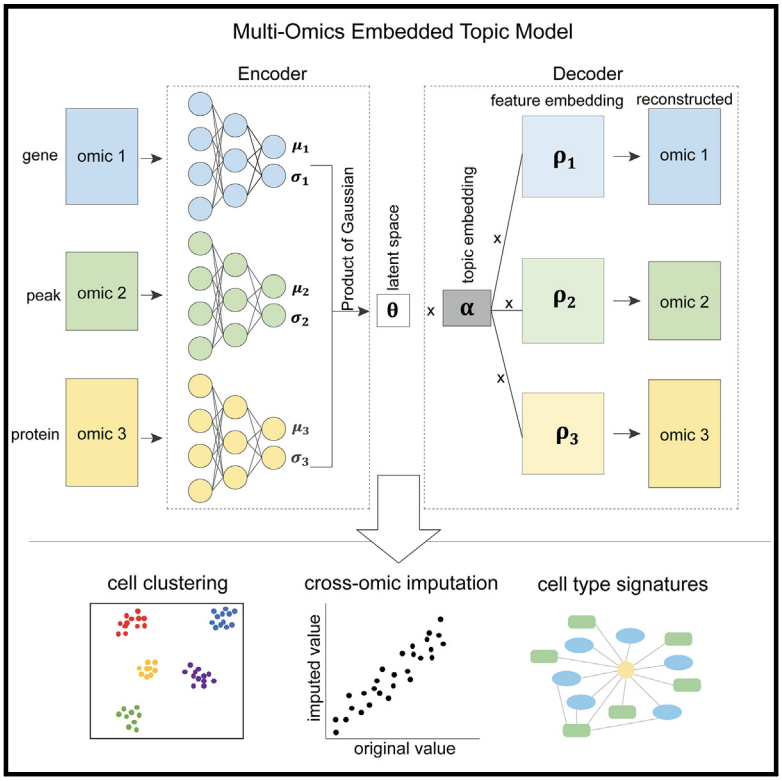
Cell Reports
Methods

Article

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

Graphical abstract



Highlights

- moETM integrates multiple modalities to a low-dimensional representation
- moETM accurately achieves cross-omics imputation
- moETM-inferred topics can elucidate cell-type signatures and gene regulatory programs
- moETM reveals the molecular basis of disease severity on a COVID-19 CITE-seq

Authors

Manqi Zhou, Hao Zhang, Zilong Bai, Dylan Mann-Krzisnik, Fei Wang, Yue Li

Correspondence

few2001@med.cornell.edu (F.W.), yueli@cs.mcgill.ca (Y.L.)

In brief

Zhou et al. develop moETM to integrate single-cell multi-omics data by leveraging product-of-experts framework to infer latent topics and linear decoders to learn multi-omics features. moETM aims to accomplish several objectives, including clustering and identifying sub-cell types, cross-omics imputation, and identifying cell-type signatures.

Zhou et al., 2023, Cell Reports Methods 3, 100563
August 28, 2023
https://doi.org/10.1016/j.crmeth.2023.100563

CellPress

Cell Reports Methods

CellPress
OPEN ACCESS

Article

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

Manqi Zhou,[1,2,7] Hao Zhang,[3,7] Zilong Bai,[2,3] Dylan Mann-Krzisnik,[4] Fei Wang,[2,3,*] and Yue Li[4,5,6,8,*]
[1]Department of Computational Biology, Cornell University, Ithaca, NY 14853, USA
[2]Institute of Artificial Intelligence for Digital Health, Weill Cornell Medicine, New York, NY 10021, USA
[3]Division of Health Informatics, Department of Population Health Sciences, Weill Cornell Medicine, New York, NY 10021, USA
[4]Quantitative Life Science, McGill University, Montréal, QC H3A 0G4, Canada
[5]School of Computer Science, McGill University, Montréal, QC H3A 0G4, Canada
[6]Mila – Quebec AI Institute, Montréal, QC H2S 3H1, Canada
[7]These authors contributed equally
[8]Lead contact
*Correspondence: few2001@med.cornell.edu (F.W.), yueli@cs.mcgill.ca (Y.L.)
https://doi.org/10.1016/j.crmeth.2023.100563

MOTIVATION Single-cell technologies are advancing to enable profiling of multiple modalities within individual cells. However, the sparsity and noise in the data poses challenges for analysis. Most existing computational methods for analyzing single-cell data are either limited to single modality or lack flexibility and interpretability. In this study, we introduce moETM, a unified deep learning model that integrates single-cell multi-omics data by projecting them onto a common topic mixture representation. Furthermore, moETM employs a linear decoder design, which facilitates the interpretability and the discovery of biologically significant patterns.

SUMMARY

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. Here, we propose an interpretable deep learning method called moETM to perform integrative analysis of high-dimensional single-cell multimodal data. moETM integrates multiple omics data via a product-of-experts in the encoder and employs multiple linear decoders to learn the multi-omics signatures. moETM demonstrates superior performance compared with six state-of-the-art methods on seven publicly available datasets. By applying moETM to the scRNA + scATAC data, we identified sequence motifs corresponding to the transcription factors regulating immune gene signatures. Applying moETM to CITE-seq data from the COVID-19 patients revealed not only known immune cell-type-specific signatures but also composite multi-omics biomarkers of critical conditions due to COVID-19, thus providing insights from both biological and clinical perspectives.

INTRODUCTION

Multi-omics single-cell high-throughput sequencing technologies open up new opportunities to interrogate cell-type-specific gene regulatory programs. Single-cell RNA sequencing (scRNA-seq) combined with assay for transposase-accessible chromatin using sequencing (ATAC-seq)[1] simultaneously measure the transcriptome and chromatin accessibility in the same cell. Cellular indexing of transcriptomes and epitopes by sequencing (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-omics single-cell data remains a challenge due to several factors.[3,4] Firstly, the cell yield of multi-omics single-cell technologies is lower compared with the single-omics 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

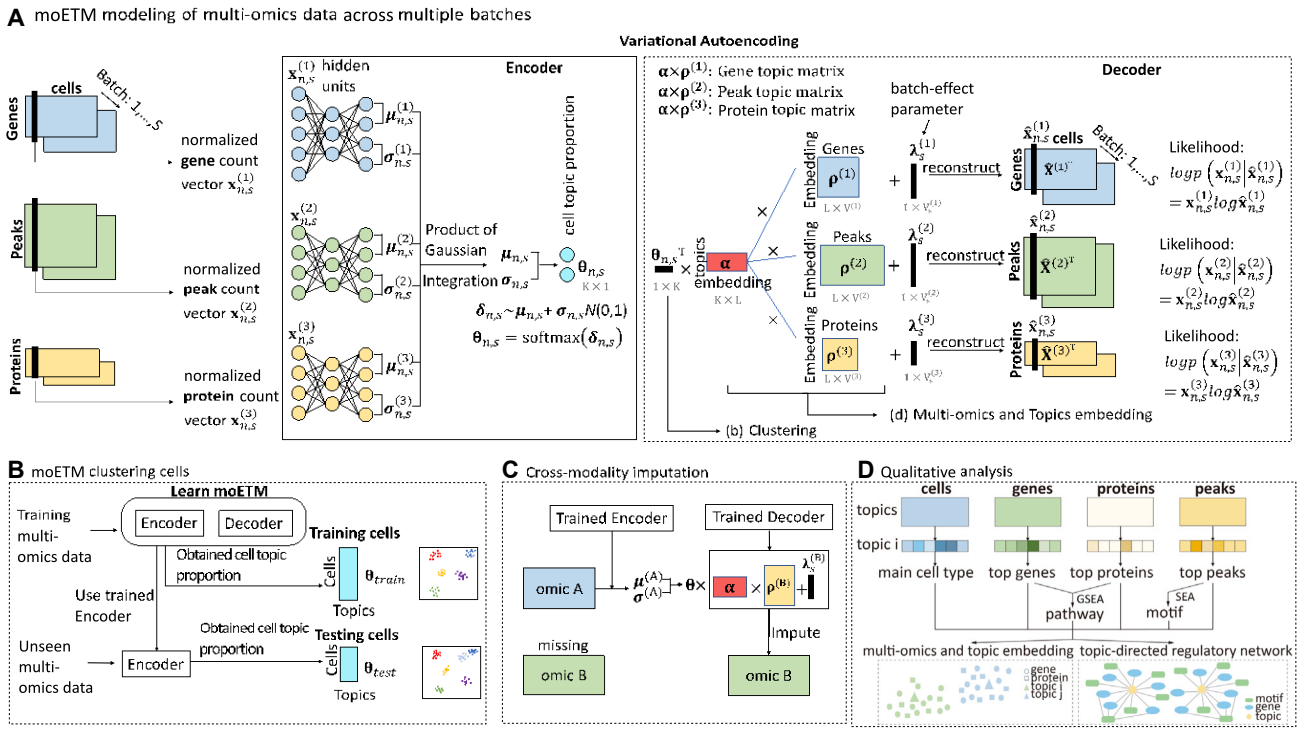


Figure 1. moETM model overview

(A) Modeling single-cell multi-omics data across batches. For details, see STAR Methods.

(B) Evaluating moETM through cell clustering. The integration performance of moETM is evaluated by clustering cells based on their topic proportion and qualitatively evaluated by UMAP visualization.

(C) Cross-modality imputation.

(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.

distill meaningful cell-type signatures from the multi-modal data while not overfitting the data. Secondly, multi-omics single-cell data are noisier compared with bulk-level or single-cell single-omics 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. Finally, multi-omics single-cell data are more costly compared with scRNA-seq or scATAC-seq alone. It is therefore highly cost-effective if we can profile single-omics data and then predict the unobserved omics data.

Recently, several computational methods were developed to tackle the above multi-modality data-integration challenges encountered in multi-omics single-cell data analysis. For instance, SMILE[5] integrates multi-omics 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[10] adopted a product-of-experts framework to integrate multi-omics data. MOFA+[11] uses mean-field variational Bayes and coordinate ascent to fit a Bayesian group factor analysis model to integrate the multi-omics data. Seurat V4[12] 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-omics 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-omics 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 a 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-omics data and a set of linear decoders to learn shared embedding of topics and multi-omics features that can accurately reconstruct the high-dimensional multi-omics data from their low-dimensional latent topic space (Figure 1A).

Through effectively integrating multiple modalities from multi-omics single-cell sequencing data, moETM seeks to achieve three tasks: (1) clustering cells into biologically meaningful clusters to identify the sub-cell type indicative of phenotype of interests (Figure 1B), (2) imputing one omics using the other omics (Figure 1C), and (3) identifying cell-type signatures, which serve as biomarkers for a target phenotype (Figure 1D). Through comprehensive experiments on seven single-cell multi-omics datasets, we demonstrate moETM's ability comparatively with six state-of-the-art 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 crossmodality imputation compared with the existing methods.[5–9,11,12] Furthermore, the topic embedding learned by moETM can be used to gain biological insights into the cell-type-specific multi-omics regulatory elements.

RESULTS

moETM model overview

As an overview, moETM integrates multi-omics data across different experiments or studies with interpretable latent embeddings (Figure 1). It is built upon the widely used VAE[13] to model multi-modal data (Figure 1A). However, to tailor the VAE framework for the single-cell multi-omics 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-omics normalized count vectors for a cell. We assume the latent representation of each omics follows a *K*-dimensional independent logistic normal distribution. Our goal is to effectively combine these distributions into a joint distribution of the multi-omics data. To this end, we take the product of the *K*-dimensional Gaussians, i.e., product-of-Gaussians (PoG). Because the PoG is also a Gaussian density function, we can represent the joint latent distribution in closed form. In principle, this results in a tighter evidence lower bound (ELBO) and therefore more efficient variational inference compared with the MoE approaches[14] as adopted in MultiVI/TotalVI[6,7] and scMM.[9] In particular, these MoE approaches sample *K*-dimensional Gaussian variables for each omics 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 reduced sampling noise from the Monte Carlo approximation procedure. 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 the topic distribution must sum to 1 over the *K* topics; the inferred topic mixture membership of a cell expresses statistical uncertainty in the cell embedding.

On the decoder side, inherited from our earlier work,[15] moETM employs a linear matrix factorization to reconstruct the normalized count vectors from the cell embedding. Our working hypothesis is that the encoder creates a linearly separable space for the decoder to achieve a good reconstruction when the two networks are trained end-to-end. Specifically, the decoder factorizes the cell-by-feature matrices into a shared cell-by-topic matrix Q, a shared topic-embedding matrix a, and *M* separate feature-embedding matrices $r^{\delta m \flat}$, where $m \in f1; .; Mg$ indexes the omics. Since different omics share the same cells-by-topics matrix but have 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,[16] and Multigrate[10] that used a neural network as the decoder. Another main challenge in single-cell data analysis is the batch effects, which are sources of technical variation. To account for those, we introduced the omics-specific batch removal factors $l^{\delta m \flat} \in R^{V^{\delta m \flat} 3S}$ for each omics *m* (Figure 1A), which act as linear-additive batch-specific biases in reconstructing each modality. By regressing out the batch effects via $l^{\delta \$ \flat}$, moETM can learn biologically meaningful representations in terms of the cell topic mixture and the topic/feature embedding. As detailed in STAR Methods, all the parameters in moETM are learned end-to-end by maximizing a common objective function defined as the ELBO of the marginal data likelihood under the framework of amortized variational inference.

Multi-omics integration

We performed quantitative evaluations of moETM on the integrated low-dimensional representation compared with six state-of-the-art multi-omics integration methods (SMILE,[5] scMM,[9] Cobolt,[8] MultiVI/TotalVI,[6,7] MOFA+,[11] and Seurat V4[12]) on seven published datasets. Four out of the seven datasets are single-cell transcriptome and chromatin accessibility (gene + peak) datasets and the other three are single-cell transcriptome and surface protein expression (gene/transcript + protein) datasets measured by CITE-seq.

The performance of the multi-omics integrative task was based on both biological conservation metrics and batch removal metrics (STAR Methods). For the biological conservation score, we adopted the common metrics including 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: (1) 60/40 random split for training and testing with 500 repeats (Tables S1 and S2), (2) training and testing both on the whole dataset (Table S3), and (3) training and testing across different batches (Tables S4 and S5). The number of topics was set to 100 during the training based on the robust performance (Figure S1). Overall, we obtained consistent results across all three 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 or gene + protein datasets' performance scores among three out of four evaluation metrics (Figure 2). It conferred the second-highest averaged kBET, which is only marginally behind multiVI/totalVI. The latter two methods 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 performance among two out of four evaluation metrics. Specifically,

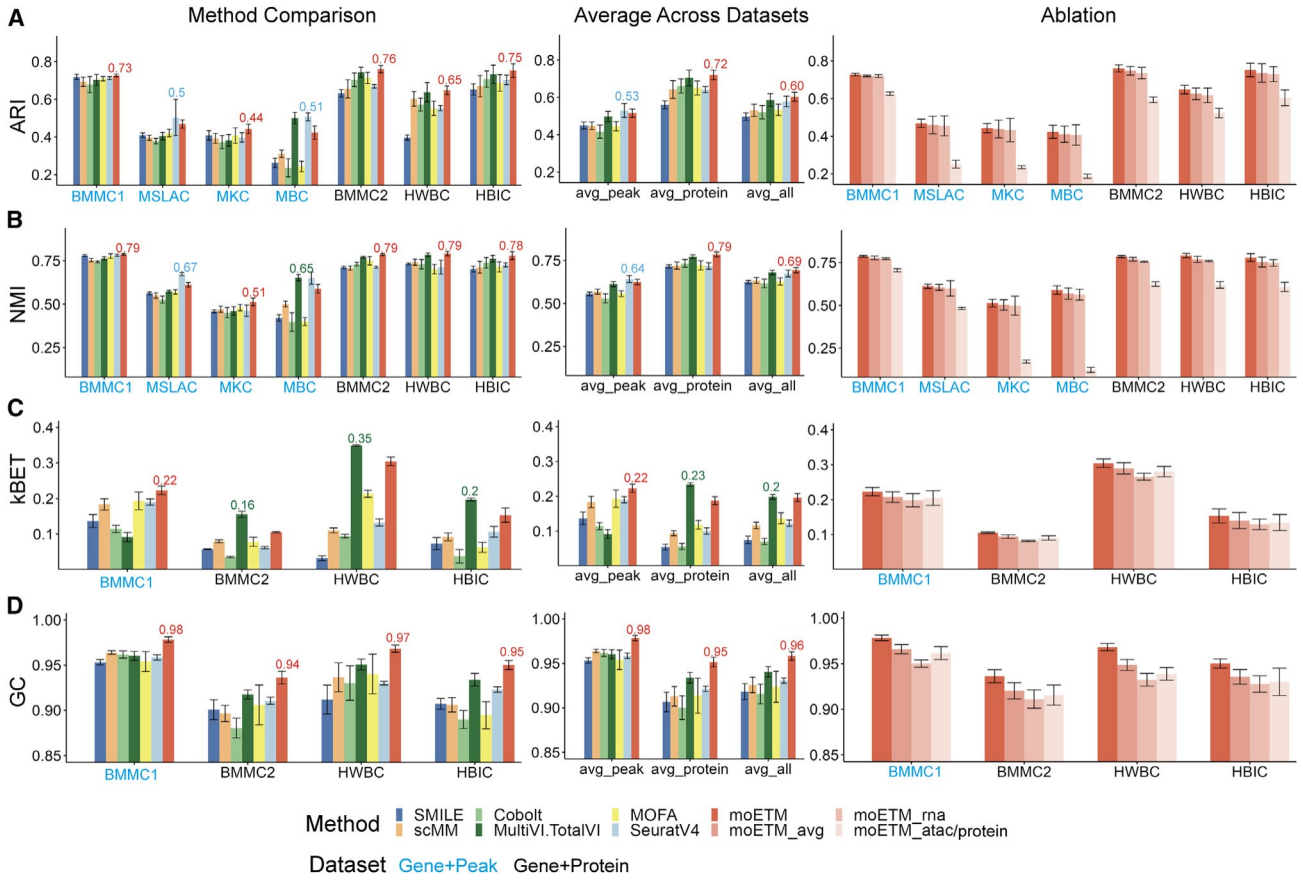


Figure 2. Methods comparison based on cell clustering

The left column illustrates the individual performance of each method on each dataset. The seven datasets are indicated on the x axis with gene + peak datasets colored in blue and gene + protein datasets colored in black. 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 neighbor batch effect test (kBET).

(D) Graph connectivity (GC).

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 six out of seven datasets (except MBC) for different experimental settings in terms of the ARI (Figure 2A; Tables S1 and S2). One possible reason could be that the sample size of MBC (3,293 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 omic. As expected, the performance of moETM on single modality decreased, indicating that moETM could improve its performance by leveraging multiple modalities (Figure 2, right panel).

Similar quantitative conclusions can be drawn based on NMI (Figure 2B; Tables S1 and S2). For kBET (Figure 2C), moETM is the best for the BMMC1 dataset and the second best on the other datasets—slightly behind MultiVI/TotalVI. Therefore, while moETM conferred higher biological conservation scores in terms of ARI and NMI, it still maintains comparable kBET scores on all four datasets compared with MultiVI/TotalVI. Indeed, we observed an excellent balance between the biological conservation and batch effect removal because moETM achieved notably higher GC compared with all other methods (Figure 2D). This is because GC is the only metric that is based on both 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 moETM's superior integration performance is its 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 such as scMM.[9] As expected, the performance of moETM_avg was worse than moETM in all

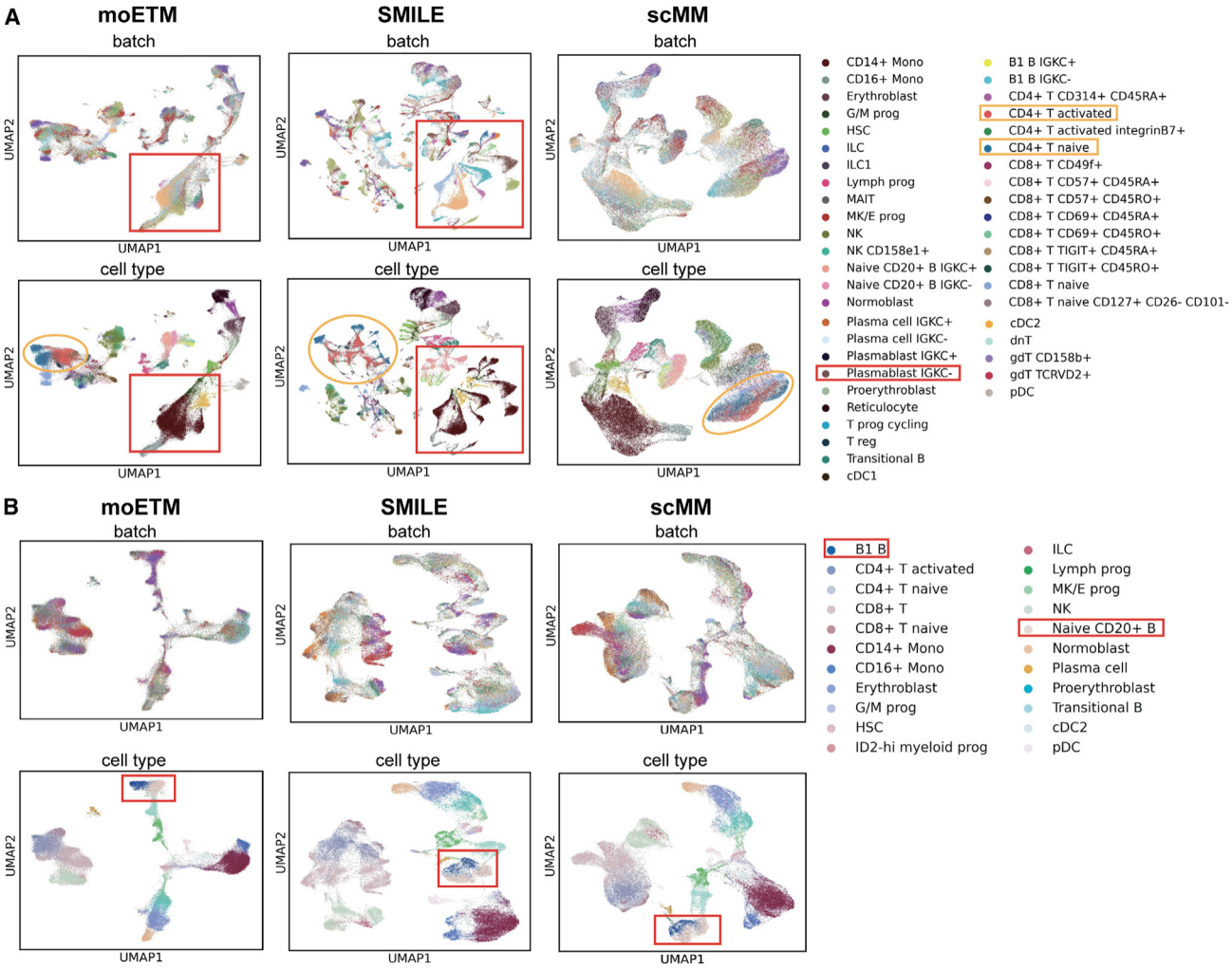


Figure 3. UMAP visualization of cell clustering

(A) UMAP visualization of moETM, SMILE, and scMM on single-cell CITE-seq from BMMC2 dataset colored by batches (upper panel) and cell types (lower panel).

(B) UMAP visualization of moETM, SMILE, and scMM on the gene + peak multi-omics data from the BMMC1 dataset.

datasets in terms of all evaluation metrics (Figure 2, right panel). Furthermore, since scMM also adopted the average of the Gaussian samples from the encoder, the fact that moETM_avg outperformed scMM indicates the benefits of using the linear decoder, which further improves the multi-omics integration while correcting batch effects across all cells.

We further verified the clustering performance by visualizing the cell topic mixture embeddings using Uniform Manifold Approximation and Projection (UMAP)[21] (Figure 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 (Figure 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+ T 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 (Figure 3B) were mixed by other methods while better separated by moETM.

In addition, we visually compared the cell clustering by the individual modalities with that by the integrated modalities via UMAP (Figure S2). Similarly, we also compared the cell clustering based on the individual omics of single-cell transcriptome and surface proteins with the clustering based on the integrated RNA + protein topic mixture using the BMMC2 dataset. We observed that the integrated topic representation led to a more coherent cell clustering compared with the topic mixture of individual omics (Figure S2). For example, the ''CD14$^+$ Mono'' cells were grouped more closely by the integrated topic mixture compared with their unimodal counterparts (Figure S2A; ARI: 0.735 by RNA + peak in contrast to 0.690 by RNA only and 0.648 by peak only). Similarly, the plasmablast IGKC– cells also formed tighter clusters in the integrated RNA + protein

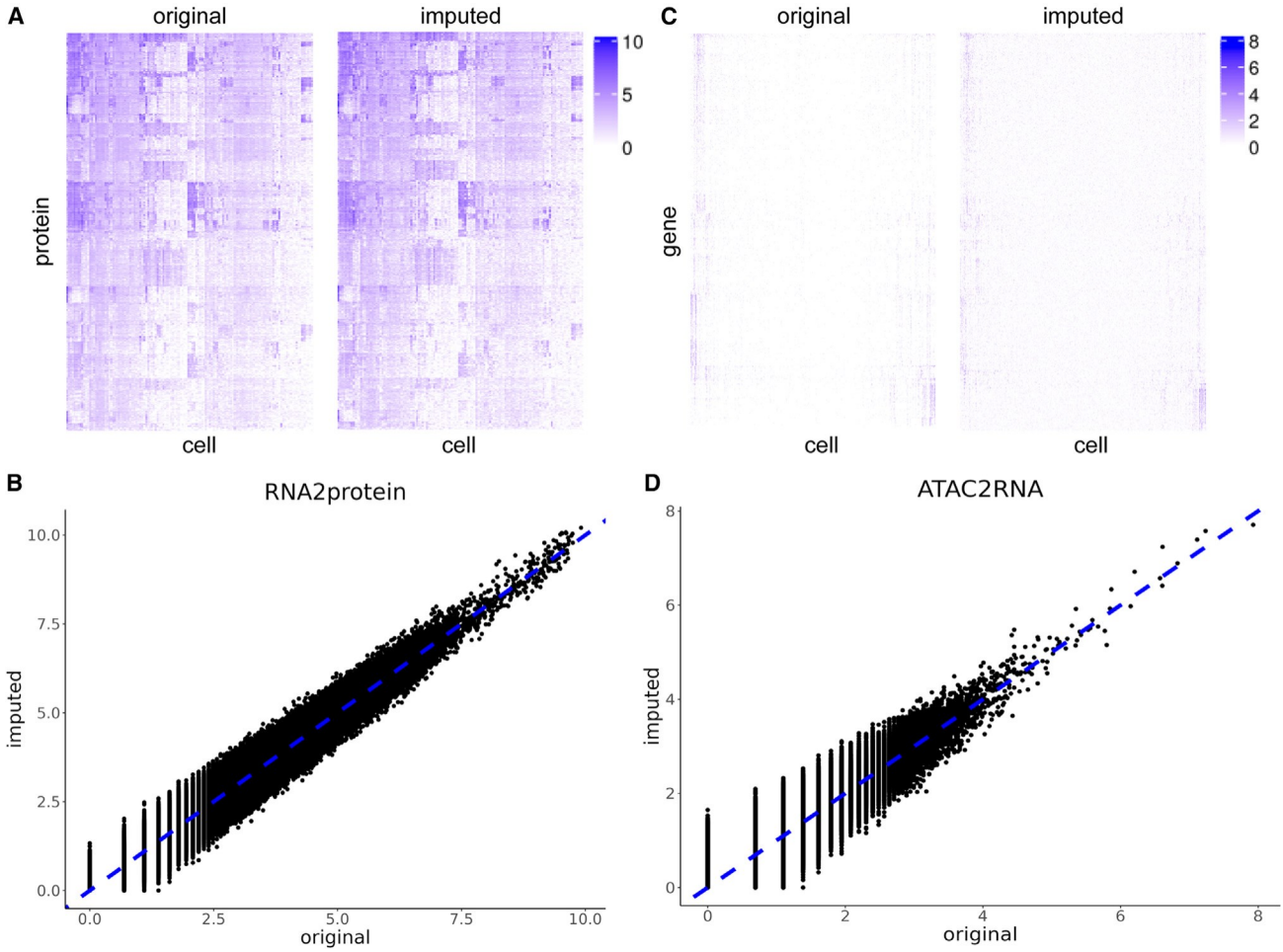


Figure 4. Cross-omics imputation

(A) Heatmap of original protein and imputed protein values from gene expression using the BMMC2 CITE-seq dataset. The column and row orders are the same for the two heatmaps.

(B) Scatterplot of original and imputed surface protein expression. The diagonal line is in blue color.

(C and D) Heatmap and scatterplot of the original and imputed gene expression from chromatin accessibility on the BMMC1 dataset. For more results, see Figure S3.

embedding space (Figure S2B; ARI: 0.734 by RNA + protein in contrast to 0.688 by transcriptome only and 0.590 by surface proteins only).

Therefore, moETM was able to improve cell clustering by integrating multiple modalities. 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.

Cross-omics imputation

In the case of gene + protein, moETM accurately imputes surface protein expression from gene expression for the BMMC2 dataset, 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 (Table S6). We visualized the reconstructed protein expression against the observed values using the BMMC2 (gene + protein) dataset (Figure 4A). The imputed protein expression is highly linearly correlated with the observed one (Figure 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 with 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 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) (Table S6). Qualitatively, the imputed and the observed gene expression profiles also exhibit similar pattern and linear relationship (Figures 4C and 4D).

In the previous two imputation applications, low-dimensional modalities were generated from high-dimensional modalities. The imputation from the low dimension to the high dimension is more difficult but nonetheless feasible. Specifically, on the three 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); the Pearson (and Spearman) correlation 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) (Table S6). 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 (Table S6). Qualitatively, the imputed and the observed peaks and gene expression exhibit consistent patterns (Figures S3A and S3C) and strong linear trends (Figure S3B and S3D).

Correlating RNA transcripts with surface proteins and in cis chromatin accessibility regions

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 generated by CITE-seq over 90,000 cells. For each topic, we calculated the Spearman correlation of topic scores between the 134 pairs of the gene transcripts and the corresponding translated surface proteins (Figure 5A). The correlations ranged from 0.096 to 0.751 with an average of 0.29. In particular, 96 of the 100 topics have positive correlations. Among them, 13 topics have correlations larger than 0.5.

To further quantify the known transcript-protein as well as gene-peak regulatory relations, we computed their Spearman correlations across topics inferred from the BMMC datasets. We paired a peak with a gene if it was within 150k bp distance from the transcription start site of the gene. The overall distribution of the correlations for transcript-protein pairs and gene-peak pairs were both significantly higher than 0 ($p < 2.2e16$; one-sample t test) and comparable with the correlations calculated directly from the observed data (Figures S4A and S4C).

Notably, 90% of transcript-protein pairs exhibited positive correlations and nine pairs displayed correlations exceeding 0.5 (Figure S4E). Nonetheless, several transcript-protein pairs exhibited low or negative correlations. Several factors could contribute to these low correlations. Firstly, random noise may hinder correlations between genes and proteins. Secondly, dynamic cellular processes at the single-cell level can give rise to variations between cells, leading to a decrease in correlations.[25] For example, transcriptional bursting or delays between transcription and translation will lead to asynchronous behavior of gene and protein during the cell cycle, thereby reducing the correlations between gene and protein expression levels.[25] Particularly, a number of transcript-protein pairs displayed negative correlations (Figure S4E). This phenomenon has also been observed in previous studies.[26–28] For instance, Li et al. reported a mismatch between mRNA and protein expression levels, including the *CD69*-CD69 pair.[27] One possible cause might be due to the impact of other biological processes overriding the effects of transcription.[28] Taking *CD69*-CD69 as an example, the *CD69* gene may undergo post-translational modifications such as differential glycosylation.[29] The transcribed *CD69* mRNA molecules can be translated to a 22.5 kDa polypeptide, which can further be differentially glycosylated to two different subunits. These subunits can be randomly combined to form different receptors, leading to a reduction in the abundance of the CD69 protein.[29] If the influence of post-translational modifications surpasses the impact of protein synthesis, it can give rise to a negative correlation. Furthermore, the *CD69* mRNA transcripts are unstable. The level of *CD69* mRNA could decline rapidly with a half-life of less than 60 min.[29] While mRNA molecules degrade over time, protein levels may maintain relatively stable or continue to accumulate. If the rate of mRNA degradation surpasses that of protein synthesis, a negative correlation could emerge.

In addition to investigating correlations across topics, we calculated correlations across cells by computing Spearman correlations in terms of observed values and reconstructed values based on the BMMC datasets (Figures S4A–S4D). The correlations based on both the observed and reconstructed data across cells were significantly greater than 0, indicating consistent relations among transcript-protein and gene-peak pairs captured at the cell level. However, the correlations from the reconstructed values are higher than those from the observed values (Figures S4B and S4D). This is because the observed values may contain random noise or batch effects compared with reconstructed values by moETM, which can be considered as denoised and confounder-corrected values of the gene/protein/peak signals.

Immune cell-type signatures revealed by multi-omics 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, we chose to describe a selected subset of the topics based on their distinctly enriched cell types and heatmap visualization (Figure 5B).

For instance, topic 44 was associated with monocytes and consists of $CD14^+$ and $CD16^+$ Mono; topic 40 was associated with B cells and 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 (Figure 5B).

Under each cell-type-enriched topic, many top genes and top proteins are the known cell-type markers (Figure 5D). For example, under topic 40 (i.e., a B cell topic), the top genes *CR2*, *SSPN*, and *ADAM28* are known marker genes for B cells; the top proteins CD21, CD20, and CD40 are also marker proteins for B cells according to the CellMarker database.[30] For topic 7, one of top proteins CD11c is a marker protein for dendritic cells (DCs).[30] For topic 83, protein CD16, marker for natural killer cells,

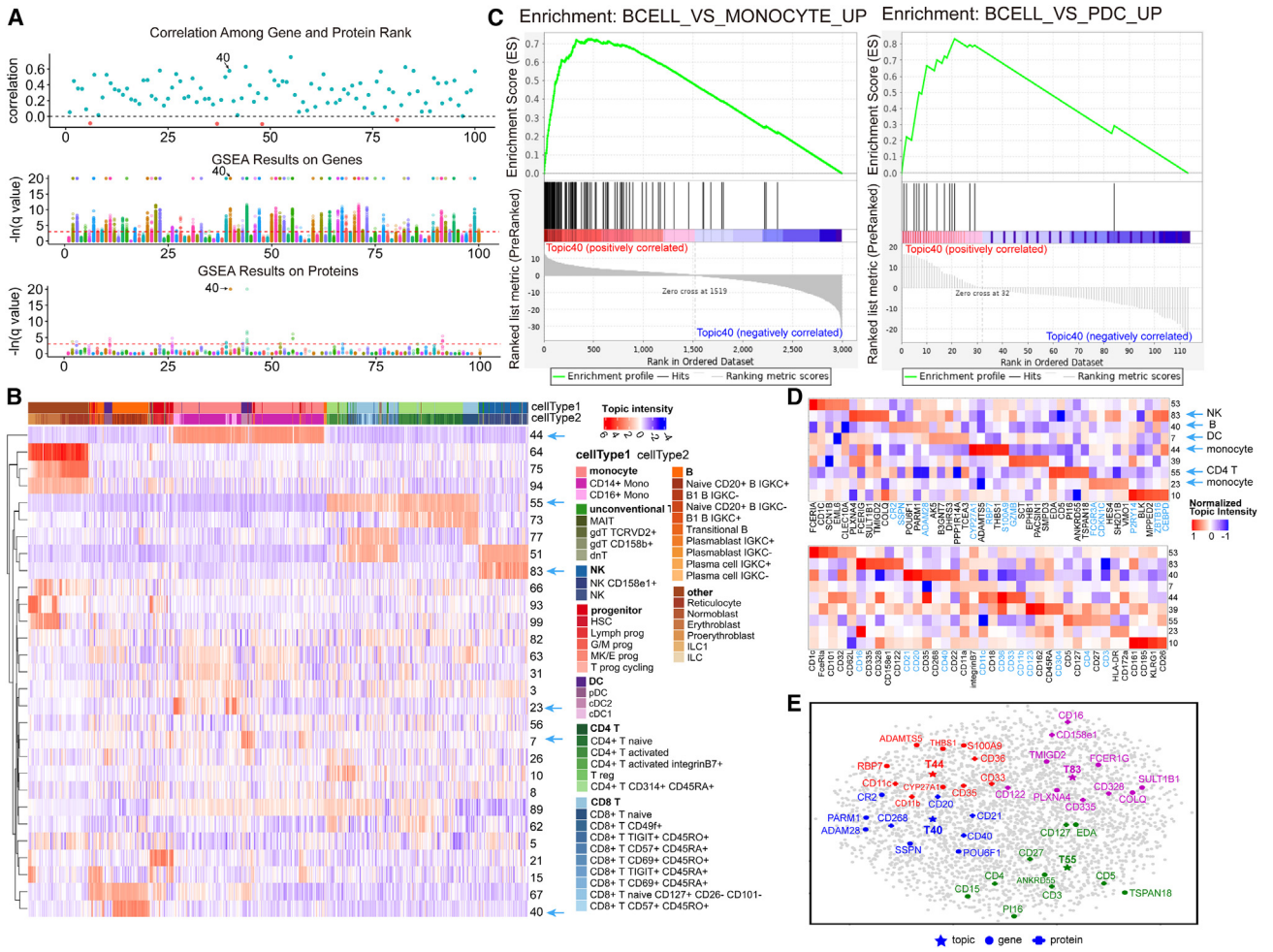


Figure 5. Topic analysis of gene + protein CITE-seq data

(A) Protein-RNA correlations and pathway enrichments for the 100 topics learned from the CITE-seq BMMC2 data. In the middle and the bottom panels, 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 with the sum of absolute values greater than the third quartile across all sampled cells 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). (C) GSEA leading-edge analysis of topic 40. The left and right panel represent significantly enriched gene sets (q value < 0.001) based on gene topic scores and protein topic scores, respectively.

(D) Genes and proteins signatures of the select topics. The upper and lower panels display the topics-by-genes and topics-by-proteins heatmap. The top genes and proteins that are known cell-type markers based on CellMarker or literature search are highlighted in blue.

(E) UMAP visualization of the genes, proteins, and topics via their shared embedding space. The corresponding topic indices and gene/protein symbols were highlighted by corresponding colors.

is among its top proteins.[30] For topic 44, the top gene *S100A9*'s coding protein is a chemotactic factor for monocytes[31] and is highly expressed during inflammatory processes[32]; among the top proteins for topic 44, CD36,[33] CD33, and CD11c[34] are also markers for monocyte sub-cell types. Similarly, the 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,[35] *FCGR3A* is upregulated in $CD16^{+}$ monocytes as supported by the existing literature.[36]

Moreover, we performed gene set enrichment analysis (GSEA)[37,38] using the topic scores for all of the genes and proteins. Because BMMCs are immune cells, we queried the C7 ImmuneSigDB from MSigDb, which is a collection of 5,219 gene sets related to immune pathways.[39–41] Across all 100 topics, we identified 2,569 enriched gene sets with q values < 0.05 using gene topic scores and 22 enriched gene sets using protein topic scores (Figure 5A). For example, in topic 40, using the gene topic scores, we found a gene set that consists of upregulated genes in B cells with respect to monocytes[42] (Figure 5C, left panel); using the protein topic scores, we found a gene set that consists of upregulated genes in B cells compared with plasmacytoid DCs (Figure 5C, right panel).[42]

Furthermore, we projected the topic embeddings and feature embeddings onto a common 2D space using UMAP (Figure 5E). We observed that the top marker genes and the top marker proteins for the cell type clustered together around the corresponding topics, implying a well-aligned embedding space within and across these modalities. Together, the results suggested that the cell-type-enriched topics inferred by moETM from the CITE-seq data reveal meaningful biological relations between genes and proteins.

Joint multi-omics topic analysis 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 in the cell-type-specific topics. Given that many top genes are known cell-type markers (Figure 6A), we postulated that the top peaks could be associated with the top genes via in *cis* or in *trans* regulatory elements. One challenge in interpreting the gene + peak multi-omics 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 (STAR 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 (Figure 6A). For example, topic 32 is associated with $CD8^+$ T cells (Figures 6A and 6B). We zoomed-in the topic by examining its top genes and top peaks. Three of the top 5 genes (TNFRSF9, *ASTL*, GZMK, DUSP2, and *DGKH*) were related to T cells. In particular, *GZMK* is a marker gene for T cells based on CellMarker[30]; *TNFRSF9* codes for a signaling protein that promotes expression of cytokines in $CD8^+$ T cells[43]; *DUSP2* encodes an inducible nuclear protein and is highly expressed in T cells.[44] Among the top 5 peak-neighboring genes (APBA2, *PRDX2*, *KLRC4*, *OBSCN*, and XCL2), *APBA2* is a marker gene for cytotoxic $CD8^+$ T lymphocytes[45]; *XCL2* expression levels substantially increased in $CD8^+$ T cells during T cell activation.[46] As another example, topic 3 is associated with $CD4^+$ T naive cells. Three out of the top 5 genes (CCR4, ADAM12, *PTPN13*, *MB21D2*, and IL4I1) and two out of the top 5 peak-neighboring genes (*INPP4B*, CCR4, *PRDX2*, RORA, and HIST1H2BD) are related to T cells. Indeed, *CCR4* is shown to be specifically expressed among naive $CD4^+$ T cells[47]; *ADAM12* is expressed in T cells in the inflamed brain and is a potential target for the treatment of Th1-mediated diseases[48]; *IL4I1* increases the threshold of T cell activation and partially modulates CD4 T cell differentiation.[49] For top peak-neighboring genes, *RORA* is upregulated among the activated $CD4^+$ T cells.[50]

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 (Figure 6D). 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 upregulated in healthy CD4 T cells compared with healthy myeloid cells[51] (Figure 6C). This is consistent with 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 upregulated in naive CD4 T cells relative to the DC.[52] 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 transcripts and the top peak-neighboring genes do not often correspond to 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-omics 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 (STAR Methods; Figure 6D). We then constructed a putative regulatory network by linking the top genes and the enriched motifs via their associated topics (Figures 6E and 6F). Interestingly, some of the top genes harbor those enriched motifs, implying that these genes are the putative target genes of the cognate TF. In topic 3, for example, one of the enriched motifs corresponds to a TF named FLI1 ($p = 0.00117$), and the top genes *IL4I1* and *PTPN13* are target genes of FLI1 based on the ENCODE Transcription Factor Targets.[53,54] As another example, one of the enriched motifs for topic 32 corresponds to TF MEF2A ($p = 5.21e\text{-}5$), whose target genes include the top genes *RGS1*, *EGR1*, *GZMK*, *ASTL*, and *DUSP2*.[53,54]

We further expanded our topic-network analysis by including enriched pathways and cell type information (Figures 6E and S5). We defined the *intra-connections* within the same topic as edges between the topic nodes and cell type nodes. We also established *inter-connections* between genes and external nodes including motifs and pathways. Specifically, the top genes under each topic could serve as members of enriched pathways or target genes for enriched motifs. For instance, in topic 32, gene *DUSP2* is a target gene for enriched motifs MEF2A ($p = 5.21e\text{-}5$, permutation test) and PAX5 ($p = 1.53e\text{-}5$, permutation test), while also being a member gene in four enriched pathways and three of them are upregulated gene sets in the enriched cell type of T cells (UNSTIM_VS._ACD3_ACD28_STIM_WT_CD4_TCELL_DN, NKT_CELL_VS._ALPHAALPHA_CD8_TCELL_DN, UNSTIM_VS._ACD3_ACD28_STIM_WT_CD4_TCELL_DN). Similarly, in topic 30, gene *FCAR* is a target gene for one enriched motif CEBPB ($p = 1.27e5$, permutation test) and a member of three enriched pathways, two of which are related to gene sets upregulated in the enriched cell type of monocyte (MONOCYTE_VS._MDC_UP, PBMC_MRKAD5_HIV_1_GAG_POL_NEF_AGE_20_50YO_1DY_UP, MONOCYTE_VS._MDC_DAY7_FLU_VACCINE_UP). Likewise, in topic 3, gene *DPP4* is a target gene for two enriched motifs PAX5 ($p = 5.39e\text{-}4$, permutation test) and SP1 ($p = 1.13e\text{-}4$, permutation test) and a member of 11 enriched pathways where four of them (NAIVE_TCELL_VS._NKCELL_UP, NAIVE_TCELL_VS._MONOCYTE_UP, NAIVE_CD4_TCELL_VS._MONOCYTE_UP, NAIVE_CD4_TCELL_VS._DC_UP) are upregulated gene sets in $CD4^+$ T naive cells. Those connections highlighted a consistent

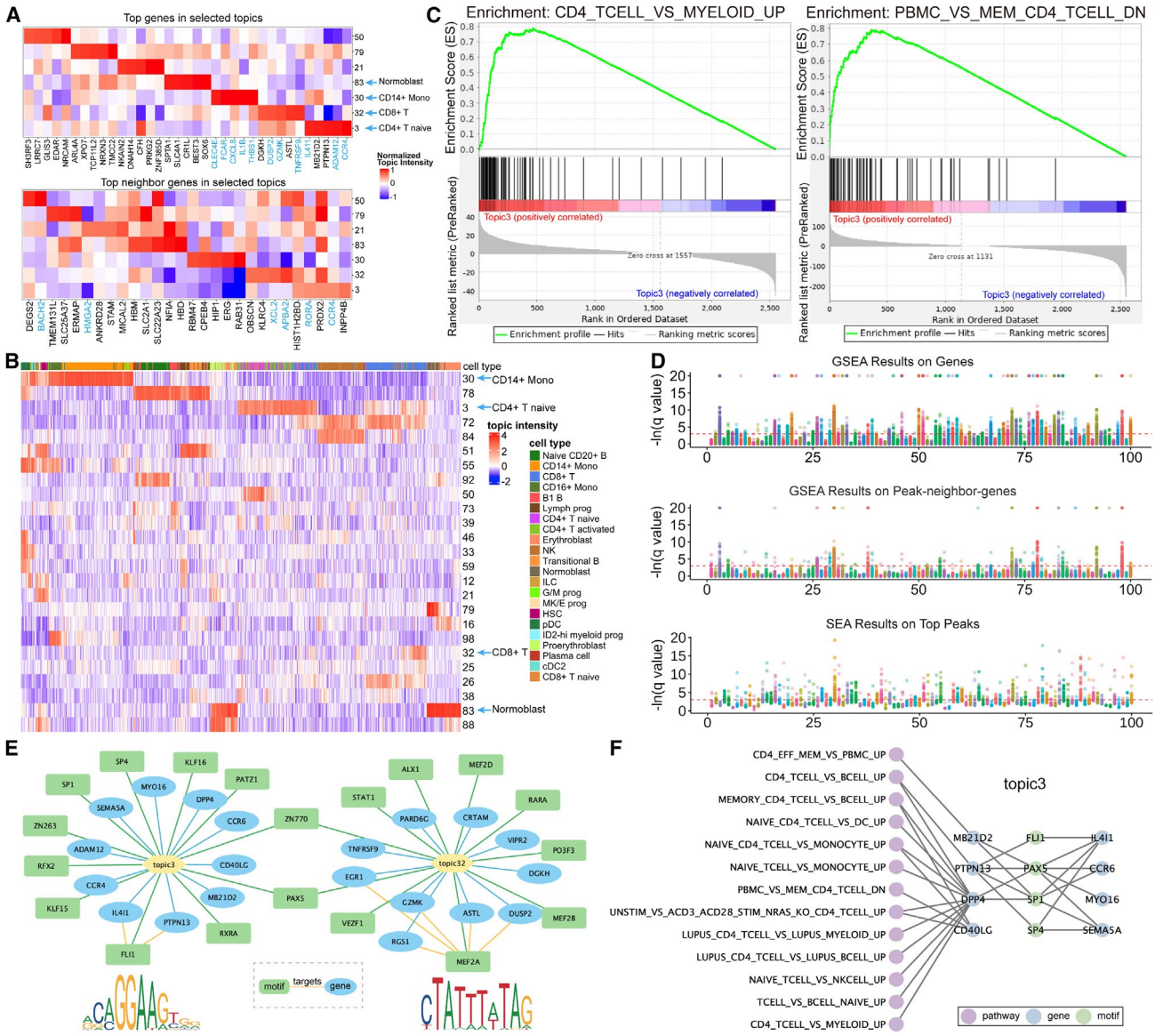


Figure 6. Topic analysis of single-cell gene + peak data from the BMMC1 dataset

(A) Top genes and top peak-neighbor genes of the select topics.

(B) Topic embedding of cells from the BMMC1 dataset.

(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.

(D) Pathway enrichment and motif enrichment for the 100 topics.

(E) Topic-directed regulatory networks based on motif enrichment analysis. The yellow edges indicate known TF-target associations based on ENCODE TF Targets dataset.

(F) Topic-directed regulatory networks incorporating enriched pathways, motifs, and top genes. For more results, see Figure S5.

regulatory relationship across motifs and pathways under inferred topics.

Therefore, our multi-omics topic analysis suggests that some of the cell-type-specific regulatory programs are implicated with the sequence motifs and pathways. Further investigation is needed to establish the hierarchical relation between the TF and the cell lineage.

Prior pathway-informed enrichment

The single-cell multi-omics 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 (STAR Methods). Using the 7,000 gene ontology biological process 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 an ARI of 0.72, which is only slightly lower than the default moETM that learned the gene embedding directly from the data (Table S1). We also identified several cell-type-specific topics along with their top genes and peaks (Figure S6C–S6H). Notably, the learned topics-by-embeddings matrix a from p-moETM is 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 cells (Figure S6C). One of its top pathways is related to B cell activation (Figure S6A). As another example, topic 8 was enriched for the $CD4^+$ T activated cells, 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 (Figure S6, left panel). For instance, topic 27 is enriched in the $CD4^+$ T naive cells. One of its top genes *CCR7* is involved in the elimination process of immature T cells. In addition, 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 gene member in the B cell activation-related pathway.[55]

Furthermore, we experimented using a more specific gene set namely the immune signature gene set collections from MSigDB to investigate immune-related pathways implicated in the BMMC1 dataset (Figure S6, right panel). 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 upregulated in naive B cells relative to the plasma cells.[56] One of the top peaks (chr12: 8886393–8887019) is upstream of *PHC1* (chr12:8913896–8941467), which is also involved in the pathway where genes are upregulated in naive B cells relative to the plasma cells.[56]

Multi-omics 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).[57] 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 S2). In particular, moETM ranked first with an ARI value of 0.752. 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 came 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.

Qualitatively, we investigated the top features and identified enriched cell types under each topic (Figures S7A and S7B). 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 marker genes. Also, three out of its top 5 proteins (CD19, CR1, CD22, FCGR2A, and BAFFR) are marker proteins for B cells. Topic 31 is associated with platelets. Two out of its top 5 genes (*LYVE1*, *RADIL*, VWF,[58] *TRHDE*, and *PPBP*) are marker genes, namely *VWF* and *PPBP*, and one of its top 5 proteins (ITGA2B, KIR3DL1, ITGAX, SELP, and FCGR2A) is a marker protein (i.e., ITGA2B) for platelets. In addition, a previous study has suggested that SELP redistributes to the plasma membrane during platelet activation.[59] The enriched pathways based on GSEA are consistent to the cell-type specificity of those topics (Figure S7C). Taking topic 42 as an example, the enriched pathway is the gene set that is downregulated in CD4 T cells compared with B cells.[51] Because of the shared embedding space, we also observed localization of the top genes and the top proteins for the selected topics via UMAP (Figure S7D).

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 upregulation or downregulation with the disease phenotypes (Figures 7A and 7B). We delved into topics based on their distinctly enriched cell types, heatmap visualization (Figure S7A), and differential analysis by the phenotypes (e.g., COVID severity) (STAR Methods). For example, we observed that topic 42 is not only enriched for B cell but also upregulated 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 (Figure 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 (Figure 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 (Figures 7C and 7D). B cell lymphomas start to develop when B lymphocytes, which

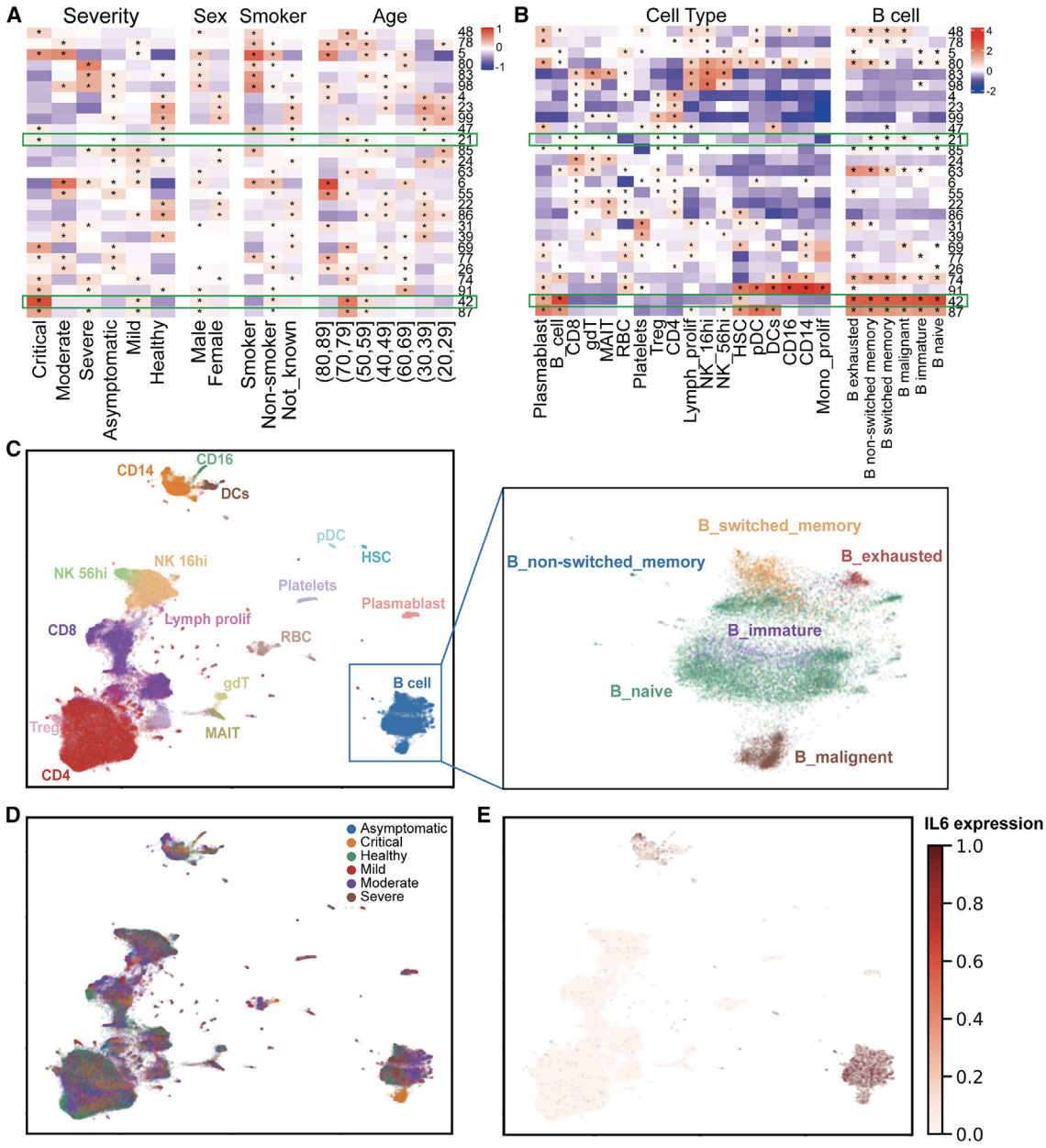


Figure 7. Topic association with the COVID-19 severity status

(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 and each topic. Asterisks indicate Bonferroni-adjusted p value < 0.001 based on one-sided t test of up-regulated topics for each label.

(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.

(C) UMAP visualization of cell clustering. The right panel shows a zoom-in version of the B cell clustering with color indicating the six B cell subtypes.

(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.

are in charge of humoral immunity, start to proliferate beyond control. This proliferation turns B cells into malignant cells.[60] The previous study[61] suggested that individuals with certain cancers, such as lymphoma, may be more susceptible to getting severe illness from COVID-19. The top gene *IL6* in topic 42 was consistently expressed at a high level among B cells, including although not specifically in B malignant cells (Figure 7E). *IL6* levels were commonly reported in severely ill patients due to COVID-19.[62,63] As another example, topic 21 is also enriched in B malignant cells (Figure 7B). One of its top proteins CD5

(Figure S7B) was shown to be highly expressed on malignant cells.[64] Moreover, the previous study[65] suggested that the proportion of $CD5^+$ B cells was significantly reduced in COVID patients. Taken together, our results suggest that *IL6* or CD5 may be a potential therapeutic target.

DISCUSSION

Gene regulatory programs involve multi-faceted regulation and cannot be understood via a single-omics approach alone. 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-omics data including transcriptome and chromatin accessibility or surface proteins, which are the most common types of single-cell multi-omics data to date.[4]

Our technical contributions are 3-fold. 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-omics 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-omics data from multiple studies, subjects, or technologies, which allows us to exploit the vast amount of multi-omics data to obtain biologically diverse and coherent multi-omics topics.

To demonstrate the utility of moETM, we benchmarked it with six existing state-of-the-art computational methods on seven published datasets including four gene + peak datasets and three gene + protein datasets (Tables S1 and S2). Across all datasets, moETM achieved competitive performance based on four common evaluation metrics. We also confirmed the advantage of using multiple modalities compared with single modality in terms of cell clustering. Moreover, because of its joint modeling capabilities, the trained moETM can accomplish this cross-omics imputation task. 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 dimensions from low dimensions.

We also explored the moETM-learned cell-type-specific topics in terms of their top omics features and enriched pathways in the light of the supporting evidence from the literature. moETM is able to detect immune cell-type signatures and identify cell-type-specific pathways and regulatory motifs. 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.

Limitations of the study

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. 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-omics single-cell data at the expense of computational complexity and interpretability.[66–68]

STAR+METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - moETM data generative process
  - moETM model inference
  - Single-cell multi-omic datasets and preprocessing
  - Cross-omic imputation
  - Evaluation metrics
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Linking genes to open chromatin regions
  - Pathway enrichment analysis
  - Motif enrichment analysis of top peaks from moETM-learned topics
  - Differential analysis to detect condition-specific topics
  - Incorporating pathway-informed gene embeddings

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.crmeth.2023.100563.

ACKNOWLEDGMENTS

F.W. would like to acknowledge the support from NIH R01AG076448, R01AG076234, RF1AG072449, NSF 1750326, and 2212175. Y.L. is supported by NSERC 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.

AUTHOR CONTRIBUTIONS

M.Z. and H.Z. designed and performed the experiments under supervision of F.W. and Y.L. All authors read and approved the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: March 20, 2023
Revised: March 31, 2023
Accepted: July 28, 2023
Published: August 18, 2023

REFERENCES

1. Buenrostro, J.D., Wu, B., Litzenburger, U.M., Ruff, D., Gonzales, M.L., Snyder, M.P., Chang, H.Y., and Greenleaf, W.J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. Nature *523*, 486–490.
2. Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P.K., Swerdlow, H., Satija, R., and Smibert, P. (2017). Simultaneous epitope and transcriptome measurement in single cells. Nat. Methods *14*, 865–868.
3. Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D.J., Hicks, S.C., Robinson, M.D., Vallejos, C.A., Campbell, K.R., Beerenwinkel, N., Mahfouz, A., et al. (2020). Eleven grand challenges in single-cell data science. Genome Biol. *21*, 1–35.
4. Argelaguet, R., Cuomo, A.S.E., Stegle, O., and Marioni, J.C. (2021). Computational principles and challenges in single-cell data integration. Nat. Biotechnol. *39*, 1202–1215.
5. Xu, Y., Das, P., and McCord, R.P. (2022). Smile: mutual information learning for integration of single-cell omics data. Bioinformatics *38*, 476–486.
6. Gayoso, A., Steier, Z., Lopez, R., Regier, J., Nazor, K.L., Streets, A., and Yosef, N. (2021). Joint probabilistic modeling of single-cell multi-omic data with totalvi. Nat. Methods *18*, 272–282.
7. Ashuach, T., Gabitto, M.I., Jordan, M.I., and Yosef, N. (2021). Multivi: Deep Generative Model for the Integration of Multi-Modal Data. Preprint at bioRxiv. https://doi.org/10.1101/2021.08.20.457057.
8. Gong, B., Zhou, Y., and Purdom, E. (2021). Cobolt: integrative analysis of multimodal single-cell sequencing data. Genome Biol. *22*, 351–421.
9. Minoura, K., Abe, K., Nam, H., Nishikawa, H., and Shimamura, T. (2021). Scmm: Mixture-Of-Experts Multimodal Deep Generative Model for Single-Cell Multiomics Data Analysis. Preprint at bioRxiv. https://doi.org/10.1101/2021.02.18.431907.
10. Lotfollahi, M., Litinetskaya, A., and Theis, F.J. (2022). Multigrate: Single-Cell Multi-Omic Data Integration. Preprint at bioRxiv. https://doi.org/10.1101/2022.03.16.484643.
11. Argelaguet, R., Arnol, D., Bredikhin, D., Deloro, Y., Velten, B., Marioni, J.C., and Stegle, O. (2020). Mofa+: a statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biol. *21*, 111–117.
12. Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., 3rd, Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., et al. (2021). Integrated analysis of multimodal single-cell data. Cell *184*, 3573–3587.e29.
13. Kingma, D.P., and Welling, M. (2013). Auto-encoding Variational Bayes. Preprint at arXiv. https://doi.org/10.48550/arXiv.1312.6114.
14. Wu, M., and Goodman, N. (2018). Multimodal generative models for scalable weakly-supervised learning. Adv. Neural Inf. Process. Syst. *31*.
15. Zhao, Y., Cai, H., Zhang, Z., Tang, J., and Li, Y. (2021). Learning interpretable cellular and gene signature embeddings from single-cell transcriptomic data. Nat. Commun. *12*, 5261–5315.
16. Wu, K.E., Yost, K.E., Chang, H.Y., and Zou, J. (2021). Babel enables cross-modality translation between multiomic profiles at single-cell resolution. Proc. Natl. Acad. Sci. USA *118*, e2023070118.
17. Hubert, L., and Arabie, P. (1985). Comparing partitions. J. Classif. *2*, 193–218.
18. Danon, L., Diaz-Guilera, A., Duch, J., and Arenas, A. (2005). Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment* 2005, P09008.
19. Büttner, M., Miao, Z., Wolf, F.A., Teichmann, S.A., and Theis, F.J. (2019). A test metric for assessing single-cell rna-seq batch correction. Nat. Methods *16*, 43–49.
20. Luecken, M.D., Büttner, M., Chaichoompu, K., Danese, A., Interlandi, M., Mueller, M.F., Strobl, D.C., Zappia, L., Dugas, M., Colomé-Tatché, M., and Theis, F.J. (2022). Benchmarking atlas-level data integration in single-cell genomics. Nat. Methods *19*, 41–50.
21. McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform Manifold Approximation and Projection for Dimension Reduction. Preprint at arXiv. https://doi.org/10.48550/arXiv.1802.03426.
22. Shema, E., Bernstein, B.E., and Buenrostro, J.D. (2019). Single-cell and single-molecule epigenomics to uncover genome regulation at unprecedented resolution. Nat. Genet. *51*, 19–25.
23. Lynch, A.W., Theodoris, C.V., Long, H.W., Brown, M., Liu, X.S., and Meyer, C.A. (2022). Mira: Joint regulatory modeling of multimodal expression and chromatin accessibility in single cells. Nat. Methods *19*, 1097–1108.
24. Ma, S., Zhang, B., LaFave, L.M., Earl, A.S., Chiang, Z., Hu, Y., Ding, J., Brack, A., Kartha, V.K., Tay, T., et al. (2020). Chromatin potential identified by shared single-cell profiling of rna and chromatin. Cell *183*, 1103–1116.e20.
25. Liu, Y., Beyer, A., and Aebersold, R. (2016). On the dependency of cellular protein levels on mrna abundance. Cell *165*, 535–550.
26. Jayapal, K.P., Philp, R.J., Kok, Y.J., Yap, M.G.S., Sherman, D.H., Griffin, T.J., and Hu, W.S. (2008). Uncovering genes with divergent mrna-protein dynamics in streptomyces coelicolor. PLoS One *3*, e2097.
27. Li, J., Zhang, Y., Yang, C., and Rong, R. (2020). Discrepant mrna and protein expression in immune cells. Curr. Genomics *21*, 560–563.
28. Koussounadis, A., Langdon, S.P., Um, I.H., Harrison, D.J., and Smith, V.A. (2015). Relationship between differentially expressed mrna and mrna-protein correlations in a xenograft model system. Sci. Rep. *5*, 10775.
29. Radulovic, K., and Niess, J.H. (2015). Cd69 is the crucial regulator of intestinal inflammation: a new target molecule for ibd treatment? J. Immunol. Res. *2015*, 497056.
30. Zhang, X., Lan, Y., Xu, J., Quan, F., Zhao, E., Deng, C., Luo, T., Xu, L., Liao, G., Yan, M., et al. (2019). Cellmarker: a manually curated resource of cell markers in human and mouse. Nucleic Acids Res. *47*, D721–D728.
31. Crowe, L.A.N., McLean, M., Kitson, S.M., Melchor, E.G., Patommel, K., Cao, H.M., Reilly, J.H., Leach, W.J., Rooney, B.P., Spencer, S.J., et al. (2019). S100a8 & s100a9: Alarmin mediated inflammation in tendinopathy. Sci. Rep. *9*, 1463–1512.
32. Wang, S., Song, R., Wang, Z., Jing, Z., Wang, S., and Ma, J. (2018). S100a8/a9 in inflammation. Front. Immunol. *9*, 1298.
33. Woo, M.-S., Yang, J., Beltran, C., and Cho, S. (2016). Cell surface cd36 protein in monocyte/macrophage contributes to phagocytosis during the resolution phase of ischemic stroke in mice. J. Biol. Chem. *291*, 23654–23661.
34. Ong, S.-M., Teng, K., Newell, E., Chen, H., Chen, J., Loy, T., Yeo, T.W., Fink, K., and Wong, S.C. (2019). A novel, five-marker alternative to cd16–cd14 gating to identify the three human monocyte subsets. Front. Immunol. *10*, 1761.
35. Hu, Y., Hu, Y., Xiao, Y., Wen, F., Zhang, S., Liang, D., Su, L., Deng, Y., Luo, J., Ou, J., et al. (2020). Genetic landscape and autoimmunity of monocytes in developing vogt–koyanagi–harada disease. Proc. Natl. Acad. Sci. USA *117*, 25712–25721.
36. Metcalf, T.U., Wilkinson, P.A., Cameron, M.J., Ghneim, K., Chiang, C., Wertheimer, A.M., Hiscott, J.B., Nikolich-Zugich, J., and Haddad, E.K. (2017). Human monocyte subsets are transcriptionally and functionally altered in aging in response to pattern recognition receptor agonists. J. Immunol. *199*, 1405–1417.
37. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA *102*, 15545–15550.
38. Mootha, V.K., Lindgren, C.M., Eriksson, K.F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstråle, M., Laurila, E., et al. (2003). Pgc-1a-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat. Genet. *34*, 267–273.

39. Godec, J., Tan, Y., Liberzon, A., Tamayo, P., Bhattacharya, S., Butte, A.J., Mesirov, J.P., and Haining, W.N. (2016). Compendium of immune signatures identifies conserved and species-specific biology in response to inflammation. Immunity *44*, 194–206.

40. Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P., and Tamayo, P. (2015). The molecular signatures database hallmark gene set collection. Cell Syst. *1*, 417–425.

41. Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., and Mesirov, J.P. (2011). Molecular signatures database (msigdb) 3.0. Bioinformatics *27*, 1739–1740.

42. Nakaya, H.I., Wrammert, J., Lee, E.K., Racioppi, L., Marie-Kunze, S., Haining, W.N., Means, A.R., Kasturi, S.P., Khan, N., Li, G.M., et al. (2011). Systems biology of vaccination for seasonal influenza in humans. Nat. Immunol. *12*, 786–795.

43. Fröhlich, A., Loick, S., Bawden, E.G., Fietz, S., Dietrich, J., Diekmann, E., Saavedra, G., Fröhlich, H., Niebel, D., Sirokay, J., et al. (2020). 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.

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

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

46. Fox, J.C., Nakayama, T., Tyler, R.C., Sander, T.L., Yoshie, O., and Volkman, B.F. (2015). Structural and agonist properties of xcl2, the other member of the c-chemokine subfamily. Cytokine *71*, 302–311.

47. Song, K., Rabin, R.L., Hill, B.J., De Rosa, S.C., Perfetto, S.P., Zhang, H.H., Foley, J.F., Reiner, J.S., Liu, J., Mattapallil, J.J., et al. (2005). Characterization of subsets of cd4+ memory t cells reveals early branched pathways of t cell differentiation in humans. Proc. Natl. Acad. Sci. USA *102*, 7916–7921.

48. Liu, Y., Bockermann, R., Hadi, M., Safari, I., Carrion, B., Kveiborg, M., and Issazadeh-Navikas, S. (2021). Adam12 is a costimulatory molecule that determines th1 cell fate and mediates tissue inflammation. Cell. Mol. Immunol. *18*, 1904–1919.

49. Puiffe, M.-L., Dupont, A., Sako, N., Gatineau, J., Cohen, J.L., Mestivier, D., Lebon, A., Prévost-Blondel, A., Castellano, F., and Molinier-Frenkel, V. (2020). Il4i1 accelerates the expansion of effector cd8+ t cells at the expense of memory precursors by increasing the threshold of t-cell activation. Front. Immunol. *11*, 600012.

50. Haim-Vilmovsky, L., Henriksson, J., Walker, J.A., Miao, Z., Natan, E., Kar, G., Clare, S., Barlow, J.L., Charidemou, E., Mamanova, L., et al. (2021). Mapping rora expression in resting and activated cd4+ t cells. PLoS One *16*, e0251233.

51. Hutcheson, J., Scatizzi, J.C., Siddiqui, A.M., Haines, G.K., 3rd, Wu, T., Li, Q.Z., Davis, L.S., Mohan, C., and Perlman, H. (2008). Combined deficiency of proapoptotic regulators bim and fas results in the early onset of systemic autoimmunity. Immunity *28*, 206–217.

52. Abbas, A.R., Baldwin, D., Ma, Y., Ouyang, W., Gurney, A., Martin, F., Fong, S., van Lookeren Campagne, M., Godowski, P., Williams, P.M., et al. (2005). Immune response in silico (iris): immune-specific genes identified from a compendium of microarray expression data. Gene Immun. *6*, 319–331.

53. ENCODE Project Consortium; and Pachter, L. (2004). The encode (encyclopedia of dna elements) project. Science *306*, 636–640.

54. The ENCODE Project Consortium (2011). A user's guide to the encyclopedia of dna elements (encode). PLoS Biol. *9*, e1001046.

55. Köhler, S., Gargano, M., Matentzoglu, N., Carmody, L.C., Lewis-Smith, D., Vasilevsky, N.A., Danis, D., Balagura, G., Baynam, G., Brower, A.M., et al. (2021). The human phenotype ontology in 2021. Nucleic Acids Res. *49*, D1207–D1217.

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

57. Stephenson, E., Reynolds, G., Botting, R.A., Calero-Nieto, F.J., Morgan, M.D., Tuong, Z.K., Bach, K., Sungnak, W., Worlock, K.B., Yoshida, M., et al. (2021). Single-cell multi-omics analysis of the immune response in covid-19. Nat. Med. *27*, 904–916.

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

59. O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., et al. (2016). Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. *44*, D733–D745.

60. Hodson, D.J., Shaffer, A.L., Xiao, W., Wright, G.W., Schmitz, R., Phelan, J.D., Yang, Y., Webster, D.E., Rui, L., Kohlhammer, H., et al. (2016). Regulation of normal b-cell differentiation and malignant b-cell survival by oct2. Proc. Natl. Acad. Sci. USA *113*, E2039–E2046.

61. Bonuomo, V., Ferrarini, I., Dell'Eva, M., Sbisà, E., Krampera, M., and Visco, C. (2021). Covid-19 (sars-cov-2 infection) in lymphoma patients: A review. World J. Virol. *10*, 312–325.

62. Jones, S.A., and Hunter, C.A. (2021). Is il-6 a key cytokine target for therapy in covid-19? Nat. Rev. Immunol. *21*, 337–339.

63. Sabaka, P., Koscálová, A., Straka, I., Hodosy, J., Lipták, R., Kmotorková, B., Kachlíková, M., and Kusnírová, A. (2021). 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 Infect. Dis. *21*. 308–8.

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

65. Laing, A.G., Lorenc, A., Del Molino Del Barrio, I., Das, A., Fish, M., Monin, L., Muñoz-Ruiz, M., McKenzie, D.R., Hayday, T.S., Francos-Quijorna, I., et al. (2020). A dynamic covid-19 immune signature includes associations with poor prognosis. Nat. Med. *26*, 1623–1635.

66. Wen, H., Ding, J., Jin, W., Wang, Y., Xie, Y., and Tang, J. (2022). 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). https://doi.org/10.1145/3534678.3539213.

67. Cao, Z.-J., and Gao, G. (2022). Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. Nat. Biotechnol. *40*, 1458–1466.

68. Wang, J., Ma, A., Chang, Y., Gong, J., Jiang, Y., Qi, R., Wang, C., Fu, H., Ma, Q., and Xu, D. (2021). scgnn is a novel graph neural network framework for single-cell rna-seq analyses. Nat. Commun. *12*, 1882–1911.

69. Jordan, M.I., Ghahramani, Z., Jaakkola, T.S., and Saul, L.K. (1999). An introduction to variational methods for graphical models. Mach. Learn. *37*, 183–233.

70. Luecken, M.D., Burkhardt, D.B., Cannoodt, R., Lance, C., Agrawal, A., Aliee, H., Chen, A.T., Deconinck, L., Detweiler, A.M., Granados, A.A., et al. (2021). A sandbox for prediction and integration of dna, rna, and proteins in single cells. In NeurIPS 2021 Datasets and Benchmarks Track (Round 2).

71. Cao, J., Cusanovich, D.A., Ramani, V., Aghamirzaie, D., Pliner, H.A., Hill, A.J., Daza, R.M., McFaline-Figueroa, J.L., Packer, J.S., Christiansen, L., et al. (2018). Joint profiling of chromatin accessibility and gene expression in thousands of single cells. Science *361*, 1380–1385.

72. Wolf, F.A., Angerer, P., and Theis, F.J. (2018). Scanpy: large-scale single-cell gene expression data analysis. Genome Biol. *19*. 15–5.

73. Romano, S., Vinh, N.X., Bailey, J., and Verspoor, K. (2016). Adjusting for chance clustering comparison measures. J. Mach. Learn. Res. *17*, 4635–4666.

74. Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T., and Carey, V.J. (2013). Software for computing and annotating genomic ranges. PLoS Comput. Biol. *9*, e1003118.

75. Bailey, T.L., and Grant, C.E. (2021). Sea: Simple Enrichment Analysis of Motifs. Preprint at bioRxiv. https://doi.org/10.1101/2021.08.23.457422.

76. Bailey, T.L., Johnson, J., Grant, C.E., and Noble, W.S. (2015). The meme suite. Nucleic Acids Res. *43*, W39–W49.

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

78. Kulakovskiy, I.V., Vorontsov, I.E., Yevshin, I.S., Sharipov, R.N., Fedorova, A.D., Rumynskiy, E.I., Medvedeva, Y.A., Magana-Mora, A., Bajic, V.B., Papatsenko, D.A., et al. (2018). Hocomoco: towards a complete collection of transcription factor binding models for human and mouse via large-scale chip-seq analysis. Nucleic Acids Res. *46*, D252–D259.

STAR+METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
BMMC CITE-seq	Luecken et al.[69]	GSE194122
BMMC Multiome data	Luecken et al.[69]	GSE194122
Mouse skin cell SHARE-seq	Ma et al.[24]	GSE140203
Mouse brain cell SHARE-seq	Ma et al.[24]	GSE140203
Mouse kidney cell sci-CAR	Cao et al.[70]	GSE117089
PBMC CITE-seq	Hao et al.[12]	GSE164378
COVID-19 PBMC CITE-seq	Stephenson et al.[57]	https://www.covid19cellatlas.org/
Software and algorithms		
moETM	This paper	https://github.com/manqizhou/moETM https://doi.org/10.5281/zenodo.8104798
scanpy	Wolf et al.[71]	https://github.com/scverse/scanpy
anndata	Virshup et al.[72]	https://github.com/scverse/anndata
biomaRt	Durinck et al.[73]	https://rdrr.io/bioc/biomaRt/man/
seurat	Hao et al.[12]	https://satijalab.org/seurat/
totalVI	Gayoso et al.[6]	https://github.com/YosefLab/scvi-tools
SMILE	Xu et al.[5]	https://github.com/rpmccordlab/SMILE
scMM	Minoura et al.[9]	https://github.com/kodaim1115/scMM
Cobolt	Gong et al.[8]	https://github.com/epurdom/cobolt
MultiVI	Ashuach et al.[7]	https://zenodo.org/record/5762077
MOFA+	Argelaguet et al.[11]	https://github.com/bioFAM/MOFA2

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Yue Li (yueli@cs.mcgill.ca).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- All data used in this study is publicly available. The peripheral blood mononuclear cells CITE-seq measuring from both COVID patients and healthy patients is available at the website https://www.covid19cellatlas.org/. The other datasets used are available under the NCBI GEO accession numbers as listed in the key resources table.
- All original code has been deposited at https://doi.org/10.5281/zenodo.8104798 and https://github.com/manqizhou/moETM and is publicly available as of the date of publication.
- Any additional information required to reanalyze the data reported in this work paper is available from the lead contact upon request.

METHOD DETAILS

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 θ_n, which are presumably implicated in each modality (Figure 1A). Inference of these topic mixtures for each cell is accomplished by modeling the distribution of the multi-omic count data $\{x_n^{(m)}\}_{m=1}^{M}$ from the topic mixture for the cell and learning the global topic embedding over the M modalities. The latter are shared among all cells and expressed as M matrices $\{F^{(m)} \in \mathbb{R}^{K \times V^{(m)}}\}_{m=1}^{M}$, where a row vector $\alpha_k^{(m)} \in \mathbb{R}^{V^{(m)}}$ 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 $F^{(m)}$ into the topic embedding $\alpha \in \mathbb{R}^{K \times L}$ and feature embedding $\rho^{(m)} \in \mathbb{R}^{L \times V^{(m)}}$, 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: $x_n^{(m)} \propto \theta_n \alpha \rho^{(m)}$.

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 θ_n from logistic normal distribution $\theta_n \sim LN(0, I)$:

$$\delta_n \sim N(0, I); \theta_n = \mathrm{softmax}(\delta_n) = \frac{\exp(\delta_{n,k})}{\sum_{k=1}^{K} \exp(\delta_{n,k})}. \quad \text{(Equation 1)}$$

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 $\mathrm{Cat}(r_{n,:}^{(m)})$:

$$w_{n,i^{(m)}}^{(m)} \sim \prod_{v^{(m)}=1}^{V^{(m)}} \left[r_{n,v}^{(m)}\right]^{\left[w_{n,i^{(m)}}^{(m)} = v^{(m)}\right]}. \quad \text{(Equation 2)}$$

where $D_n^{(m)}$ denotes the total number of reads. The expected rate $r_{n,v^{(m)}}^{(m)}$ of observing feature $v^{(m)}$ in cell n is parameterized as:

$$r_{n,v^{(m)}}^{(m)} = \frac{\exp\left(b_{n,v^{(m)}}^{(m)}\right)}{\sum_{v^{(m)}=1}^{V^{(m)}} \exp\left(b_{n,v^{(m)}}^{(m)}\right)}; b_{n,v^{(m)}}^{(m)} = \theta_n \rho_{:,v}^{(m)} + l_{s(n),v^{(m)}}^{(m)} \quad \text{(Equation 3)}$$

where $\rho_{:,v^{(m)}}^{(m)} \in \mathbb{R}^{L \times 1}$ denotes embedding of feature $v^{(m)}$, $l_{s(n),v^{(m)}}^{(m)}$ 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 within 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 only a small fraction of the unique words are draw from the entire vocabulary for any given document.

The likelihood for cell n can be expressed as multinomial distribution:

$$p(w_n | r_n) \propto \prod_{i=1}^{D_n^{(m)}} \prod_{v^{(m)}=1}^{V^{(m)}} \left[r_{n,v}^{(m)}\right]^{\left[w_{n,i^{(m)}}^{(m)} = v^{(m)}\right]} = \prod_{v^{(m)}=1}^{V^{(m)}} \left[r_{n,v}^{(m)}\right]^{\sum_{i=1}^{D_n^{(m)}} \left[w_{n,i^{(m)}}^{(m)} = v^{(m)}\right]} = \prod_{v^{(m)}=1}^{V^{(m)}} \left[r_{n,v}^{(m)}\right]^{x_{n,v}^{(m)}}$$

where $x_{n,v}^{(m)} = \sum_{i=1}^{D_n^{(m)}} \mathbb{1}[w_{n,i}^{(m)} = v]$ denotes the read count for feature $v^{(m)}$ for cell n in the m^{th} modality. As a result, we can more conveniently express the data likelihood in terms of *the read count*:

$$p(x_n | r_n) = \prod_{v^{(m)}=1}^{V^{(m)}} \left[r_{n,v}^{(m)}\right]^{x_{n,v}^{(m)}} \quad \text{(Equation 4)}$$

moETM model inference

For the ease of inference, we consider the cell topic embedding δ_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 α, feature embedding $\{\rho^{(m)}\}_{m=1}^{M}$, and batch-effect parameter $\{l^{(m)}\}_{m=1}^{M}$ are treated as point estimates and learned by the model. Let's denote $\Theta = \{\delta_n; \alpha; \{\rho^{(m)}\}_{m=1}^{M}; \{l^{(m)}\}_{m=1}^{M}\}$. A principled way to learn those parameters is to maximize the log marginal likelihood:

$$\Theta) \arg\max_{\Theta} \sum_n \log p\left(x_n^{(m)}{}_{m=1}^{M} \Theta\right) \arg\max_{\Theta} \sum_n L_n$$

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 $q(d_n)$ as a surrogate to the true posterior of the cell topic embedding $p(d_n|\{x_n^{(m)}\}_{m=1}^{M})$:

$$L_n = \log \int p\left(\{x_n^{(m)}\}_{m=1}^{M}; d_n\right) dd_n$$

$$= \log \int q(d_n) \frac{p\left(\{x_n^{(m)}\}_{m=1}^{M}; d_n\right)}{q(d_n)} dd_n$$

$$= \log E_{q(d_n)}\left[\frac{p\left(\{x_n^{(m)}\}_{m=1}^{M}; d_n\right)}{q(d_n)}\right]$$

$$\geq E_{q(d_n)}\left[\log \frac{p\left(\{x_n^{(m)}\}_{m=1}^{M}; d_n\right)}{q(d_n)}\right] \quad \text{(Equation 5)}$$

$$= E_{q(d_n)}\left[\log p\left(\{x_n^{(m)}\}_{m=1}^{M}, d_n\right) - \log q(d_n)\right]$$

$$= E_{q(d_n)}\left[\log p\left(\{x_n^{(m)}\}_{m=1}^{M}, d_n\right)\right] - E_{q(d_n)}\left[\log \frac{q(d_n)}{p(d_n)}\right]$$

$$= E_{q(d_n)}\left[\log p\left(\{x_n^{(m)}\}_{m=1}^{M}, d_n\right)\right] - KL[q(d_n)\|p(d_n)] \equiv ELBO_n \quad \text{(Equation 6)}$$

where Equation 5 follows the Jensen's inequality[69] and 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 $q(d_n)$ as a product of Gaussians (PoGs):

$$q(d_n) = N(d_n; \mu; \sigma^2); \quad \text{(Equation 7)}$$

The mean μ and standard deviation σ of the joint Gaussian is computed as:

$$\mu = \frac{\sum_{m=1}^{M} \mu_m \sigma_m^{-2}}{1 + \sum_{m=1}^{M} \sigma_m^{-2}}; \quad \sigma^2 = \frac{1}{1 + \sum_{m=1}^{M} \sigma_m^{-2}} \quad \text{(Equation 8)}$$

where μ_m and σ_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):

$$\mu_n^{(m)}, \log \sigma_n^{(m)} = \mathrm{NNET}\left(\tilde{x}_n^{(m)}; W\right) \quad \text{(Equation 9)}$$

where $x_n^{(m)}$ 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 W is the parameters for a two-layer feedforward neural network.

We approximate the above ELBO in Equation 6 by sampling from the proposed joint Gaussian distribution using the reparameterization trick[13]:

$$\tilde{d}_n \sim N(\mu, \mathrm{diag}(\sigma)) = \mu + \mathrm{diag}(\sigma) N(0, I)$$

$$ELBO_n^{(m)} = \log p\left(x_n^{(m)}, \{d_n^{(m)}\}_{m=1}^{M}; \Theta\right) - KL(q(d_n^{(m)}) \| p(d_n^{(m)}))$$

where the KL divergence has closed form:

$$\begin{aligned} KL(q(d_n^{(m)}) \| p(d_n^{(m)})) &= E_{q(d_n^{(m)})}[\log q(d_n^{(m)}) - E_{q(d_n^{(m)})}[\log p(d_n^{(m)})] \\ &= -\frac{1}{2}\log(\sigma_n^{(m)2}) + \frac{1}{2}\log(\sigma_0^2) + \frac{1}{2}\frac{\sigma_n^{(m)2}}{\sigma_0^2} + \frac{1}{2}\frac{(\mu_n^{(m)} - \mu_0)^2}{\sigma_0^2} - \frac{1}{2} \\ &= \frac{1}{2}\left[-\log(\sigma_n^{(m)2}) + \mu_n^{(m)2} + \sigma_n^{(m)2} - 1\right] \end{aligned} \quad \text{(Equation 10)}$$

Together, with the Multinomial likelihood defined in 4 and KL divergence in 10, we can express the ELBO in its approximate closed-form using the sampled latent variable:

$$\begin{aligned} ELBO &= \sum_{n=1}^{N}\sum_{m=1}^{M} \log p\left(x_n^{(m)}, \{d_n^{(m)}\}_{m=1}^{M}; \Theta\right) - KL(q(d_n^{(m)}) \| p(d_n^{(m)}))_{n=} \\ &= \sum_{v^{(m)}=1}^{M}\sum_{n;v}^{(m)} x_{n;v}^{(m)} \log r_{n;v}^{(m)} + \frac{1}{2}\left[\log(\sigma_n^{(m)2}) - \mu_n^{(m)2} - \sigma_n^{(m)2} + 1\right]_{m=1} \end{aligned} \quad \text{(Equation 11)}$$

where $r_{n;v}^{(m)}$ is defined in 3. The model parameters including the encoder weight W and the decoder weights $\Theta = \{\alpha; \rho\}$ are optimized by maximizing the above ELBO via backpropagation:

$$\Theta; W) = \arg\max_{\Theta;W} \sum_{n=1}^{N} \log p\left(x_n^{(m)}, \{d_n^{(m)}\}_{m=1}^{M}; \Theta\right) - KL(q(d_n^{(m)}) \| p(d_n^{(m)})) \quad \text{(Equation 12)}$$

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).

1. 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,[70],
2. SHARE-seq mouse skin late anagen (MSLAC) dataset containing 34,774 cells with 1 batch and 23 cell types,[24],
3. sci-CAR mouse kidney cells (MKC) dataset from cell samples with 1 batch and 14 cell types,[71],
4. 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 (e.g., 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.

1. Bone marrow mononuclear cells (BMMC2) dataset from the 2021 NeurIPS challenge from 9 donors and 4 sites,[70],
2. Human White Blood Cell (HWBC) dataset containing 211,000 human peripheral blood mononuclear cells,[12],
3. Human Blood Immune Cell (HBIC) dataset[57] measuring 647,366 peripheral blood mononuclear cells from both COVID patients and healthy patients.

Similarly, for the BMMC2 dataset, we consider two different batch types: one treats donors as batches (12 batches in total), while the other treats sites as batches (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*,[72] 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.

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 α and omic-specific feature embedding $\rho^{(A)}$ and $\rho^{(B)}$. 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 (Figure 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).

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 0–1. While NMI excels in unbalanced clustering or small clusters, ARI is better suited to clusters of similar size.[73] kBET performs hypothesis testing on whether batch labels are distributed differently across cells based on Pearson's χ^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]

QUANTIFICATION AND STATISTICAL ANALYSIS

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., $\beta_k^{(m)} = \alpha_k \rho^{(m)}$ 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 $\beta^{(gene)} = \alpha\rho^{(gene)}$ and a peaks-by-topics matrix $\beta^{(peak)} = \alpha\rho^{(peak)}$.

To transform $\beta^{(peak)}$ into a peak_to_genes-by-topics matrix $\beta^{(peaks\_to\_genes)}$, we first derived a binary peaks-to-genes mapping ma-trix 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:

$$r_{p,g} = \frac{\left(\beta_g^{(gene)} - \bar{\beta}_g^{(gene)}\right)^{\top}\left(\beta_p^{(peak)} - \bar{\beta}_p^{(peak)}\right)}{\left\|\beta_g^{(gene)} - \bar{\beta}_g^{(gene)}\right\|_2 \left\|\beta_p^{(peak)} - \bar{\beta}_p^{(peak)}\right\|_2}$$

The genome distance between peaks and genes was based on the latest genome build (i.e., hg38 for human) and obtained via the *GenomicRanges*[74] package in R.

Pathway enrichment analysis

For each moETM topic, we performed Gene Set Enrichment Analysis (GSEA)[37] 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 sig-nificant. 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.

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[75] from the MEME suite.[76] SEA utilizes the STREME motif discovery algorithm[77] to identify known motifs that are enriched in input se-quences. For our purpose, we used the HOmo sapiens COmprehensive MOdel COllection (HOCOMOCO) Human (v11) and HOCOMOCO Mouse (v11) motif database.[78] Motifs with Fisher's exact test p values lower than 0.05 were selected as the enriched motifs.

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 pheno-types. 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's *t* test. The topics with a Bonferroni-adjusted p value smaller than 0.001 were considered significant with the label.

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.