TYPE Methods
PUBLISHED 24 May 2023
DOI 10.3389/fmolb.2023.1184748

Unsupervised neural network for single cell Multi-omics INTegration (UMINT): an application to health and disease

Chayan Maitra[1†], Dibyendu B. Seal[2†], Vivek Das[3*] and Rajat K. De[1*]

[1]Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India, [2]Tatras Data Services Pvt. Ltd., New Delhi, India, [3]Novo Nordisk A/S, Maløv, Denmark

EDITED BY
Frank Emmert-Streib,
Tampere University, Finland

REVIEWED BY
Mai Chan Lau,
Institute of Molecular and Cell Biology (A*STAR), Singapore
Claudia Angelini,
National Research Council (CNR), Italy

*CORRESPONDENCE
Vivek Das,
✉ vivekdas.0687@gmail.com
Rajat K. De,
✉ rajat@isical.ac.in

†These authors have contributed equally to this work

RECEIVED 12 March 2023
ACCEPTED 04 May 2023
PUBLISHED 24 May 2023

CITATION
Maitra C, Seal DB, Das V and De RK (2023), Unsupervised neural network for single cell Multi-omics INTegration (UMINT): an application to health and disease. *Front. Mol. Biosci.* 10:1184748.
doi: 10.3389/fmolb.2023.1184748

Multi-omics studies have enabled us to understand the mechanistic drivers behind complex disease states and progressions, thereby providing novel and actionable biological insights into health status. However, integrating data from multiple modalities is challenging due to high dimensionality and diverse nature of data, and noise associated with each platform. Sparsity in data, non-overlapping features and technical batch effects make the task of learning more complicated. Conventional machine learning (ML) tools are not quite effective against such data integration hazards due to their simplistic nature with less capacity. In addition, existing methods for single cell multi-omics integration are computationally expensive. Therefore, in this work, we have introduced a novel Unsupervised neural network for single cell Multi-omics INTegration (UMINT). UMINT serves as a promising model for integrating variable number of single cell omics layers with high dimensions. It has a light-weight architecture with substantially reduced number of parameters. The proposed model is capable of learning a latent low-dimensional embedding that can extract useful features from the data facilitating further downstream analyses. UMINT has been applied to integrate healthy and disease CITE-seq (paired RNA and surface proteins) datasets including a rare disease Mucosa-Associated Lymphoid Tissue (MALT) tumor. It has been benchmarked against existing state-of-the-art methods for single cell multi-omics integration. Furthermore, UMINT is capable of integrating paired single cell gene expression and ATAC-seq (Transposase-Accessible Chromatin) assays as well.

KEYWORDS

single cell analysis, deep neural networks, unsupervised learning, CITE-seq, ATAC-seq, rare disease, multi-omics integration

1 Introduction

Recent advancements in single cell technologies have provided unprecedented opportunities in analysis of omics data. This allows researchers to probe biological functions at the cellular level while studying embryonic development, immune system or cancer (Griffiths et al., 2018; Papalexi and Satija, 2018; Wills and Mead, 2015). Existing technologies include DROP-seq (Macosko et al., 2015), SMART-seq2 Picelli et al. (2013) and 10x Genomics, which allow measuring mRNA expressions at single cell resolution (scRNA-seq). Most recently, technologies have further scaled up to produce data assays from multiple modalities. This has provided several views of the same cell of interest, thereby refining our definitions of the cellular identity. Multi-omics studies provide better understanding of the underlying biological mechanisms active during disease growth and progression, which were otherwise hidden due to the inclusion of a single omics. They also help us understand the effect of one omics layer on the other (Seal et al., 2020). A few such methods include CITE-seq (Stoeckius et al., 2017) and REAP-seq Peterson et al. (2017) which enable paired measurement of RNA and cell surface proteins. ATAC-seq (Buenrostro et al., 2015) measures chromatin accessibility while other methods like, SNARE-seq (Chen et al., 2019), sci-CAR (Cao et al., 2018) and SHARE-seq Clyde (2021) measure paired gene expression and chromatin accessibility. ScNMT (Clark et al., 2018), on the other hand, integrates single cell chromatin accessibility, DNA methylation and transcriptomics data. However, integration of various modalities of data does not come without its challenges (Lähnemann et al., 2020). High data dimension, sensitivity associated with each platform, zero-inflation due to dropouts (Jiang et al., 2022), and technical batch effects (Luecken et al., 2022) account for the stochasticity and noise in the data. An appropriate organization and analysis of these multi-modal datasets involve clever way of their integration and demand efficient computing paradigm.

At present, several methods exist that can perform the task of integration of single cell omics modalities. Seuratv3 (Stuart et al., 2019) can integrate various single cell omics datasets including RNA-seq, protein expression, chromatin and spatial data, and transfer information between them. Seuratv4 (Hao et al., 2021) provides multi-modal single cell analysis using “weighted nearest neighbor” method. MOFA+ (Argelaguet et al., 2020), based on Bayesian Group Factor Analysis, is another method that generates a low-dimensional representation of the data by integrating two or more omics among gene expression, DNA methylation and chromatin data. In recent years, several neural network-based methods have been developed for the task of single cell multi-omics integration. One such method, GLUE (Cao and Gao, 2022), integrates unpaired samples from single cell multi-omics data and also predicts regulatory interactions. Other methods, like scJoint (Lin et al., 2022) and scMVP (Li et al., 2022), can integrate scRNA-seq and scATAC-seq data. The former is a semi-supervised framework which allows label transfer and joint visualization, while the latter extracts a latent representation from the integrated data using a modified variational autoencoder model. TotalVI (Gayoso et al., 2021), on the other hand, uses an encoder function to learn a joint representation of the data and Bayesian inference to build a latent embedding from single cell RNA and protein expressions. Multigrate (Lotfollahi et al., 2022) develops an alternative pipeline for integrating CITE-seq and single cell ATAC-RNA data for both paired and unpaired samples. It has been used to map multimodal queries to reference atlases and impute missing values. Other standard omics integration methods include UINMF (Kriebel and Welch, 2022), MUON (Bredikhin et al., 2022), scMOC (Eltager et al., 2021) and SIMBA (Chen et al., 2021). A comprehensive review of major single cell multi-omics integration methods can be found in (Stanojevic et al., 2022).

With single cell multi-omics analysis, we can now comprehend the mechanisms underlying complex disease states and progressions at a cellular resolution. It has provided us multiple views of the same patient and cognizance into the individual’s health status. Some diseases, though being rare (often referred to as a rare disease (RD)), cumulatively affect quite a substantial percentage of patients. Overall, there are more than 7,000 variants of RDs. RDs affect patients’ and their families’ quality of life, and have significant societal impact. Due to the rarity of each RD, it is extremely difficult to properly diagnose and treat these individuals, as well as engage them into research to upgrade therapies. With the advancements in omics technologies, molecular understanding of RDs has improved over time leading to their rapid diagnosis. To combine multi-omics data from various technologies, Artificial Intelligence (AI)-based integration techniques are, nevertheless, becoming more and more necessary. Deep learning (DL) methodologies to integrate and query data from several heterogeneous sources may also be utilised to dramatically accelerate the discovery of efficient RD therapies (Bottini et al., 2021). A detailed review (Lee et al., 2022) exploring 332 articles on the application of DL on RDs indicate the rising demand for the use of DL for advancements in diagnosis and therapeutics of RDs.

There are, however, challenges to be addressed while using DL for multi-omics analysis for health and disease. Although different omics measurements are recorded against the same set of cells, they encode different features related to the underlying transcriptional states and activities. Existing methods to handle these datasets are not always capable of extracting features relevant to a biologically significant problem, including cell-type classification/clustering, biomarker identification, disease prediction and drug discovery. Furthermore, sparsity and noise in the data along with differences in platforms producing such high-dimensional datasets and the presence of batch effects add to the complexity of analysis (Lance et al., 2022). An inherent feature of multi-omics data in concern is its complex, non-linear, layered structure. The architecture of a deep neural network also resembles such layered non-linearity. The output from each layer is multiplied by its weight vector to compute the weighted sum, and a non-linear function is then applied over the weighted sum for each node in the layer. The non-linear output is then passed on to the next layer. Thus, deep learning models facilitate learning complex features in an unsupervised manner. However, existing neural network-based methods for single cell multi-omics integration are computationally expensive since they involve substantial amount of parameter training. Further, even though pre-processing of single cell data involves steps that may include scaling/normalization using a specific data distribution, methods for integration of such pre-processed data should be free from making assumptions about data distribution, which is not the case with most of the existing integration models.

All these problems discussed above have encouraged us to develop a robust integration method for single cell multi-omics integration that can be applied to health and disease analysis. Hence, in this work, we have introduced a novel Unsupervised neural network for single cell Multi-omics INTegration (UMINT). UMINT is competent enough to integrate different single cell omics layers of high dimensions with ease. It produces a latent low-dimensional embedding that can extract relevant features from the data, which facilitate further analyses. It can also reconstruct the data with high accuracy. Further, UMINT does not make assumptions about the distribution of data, and can integrate a variable number of omics modalities. In addition, UMINT owns a light-weight architecture and is thus computationally far less expensive than some of the existing unsupervised neural network

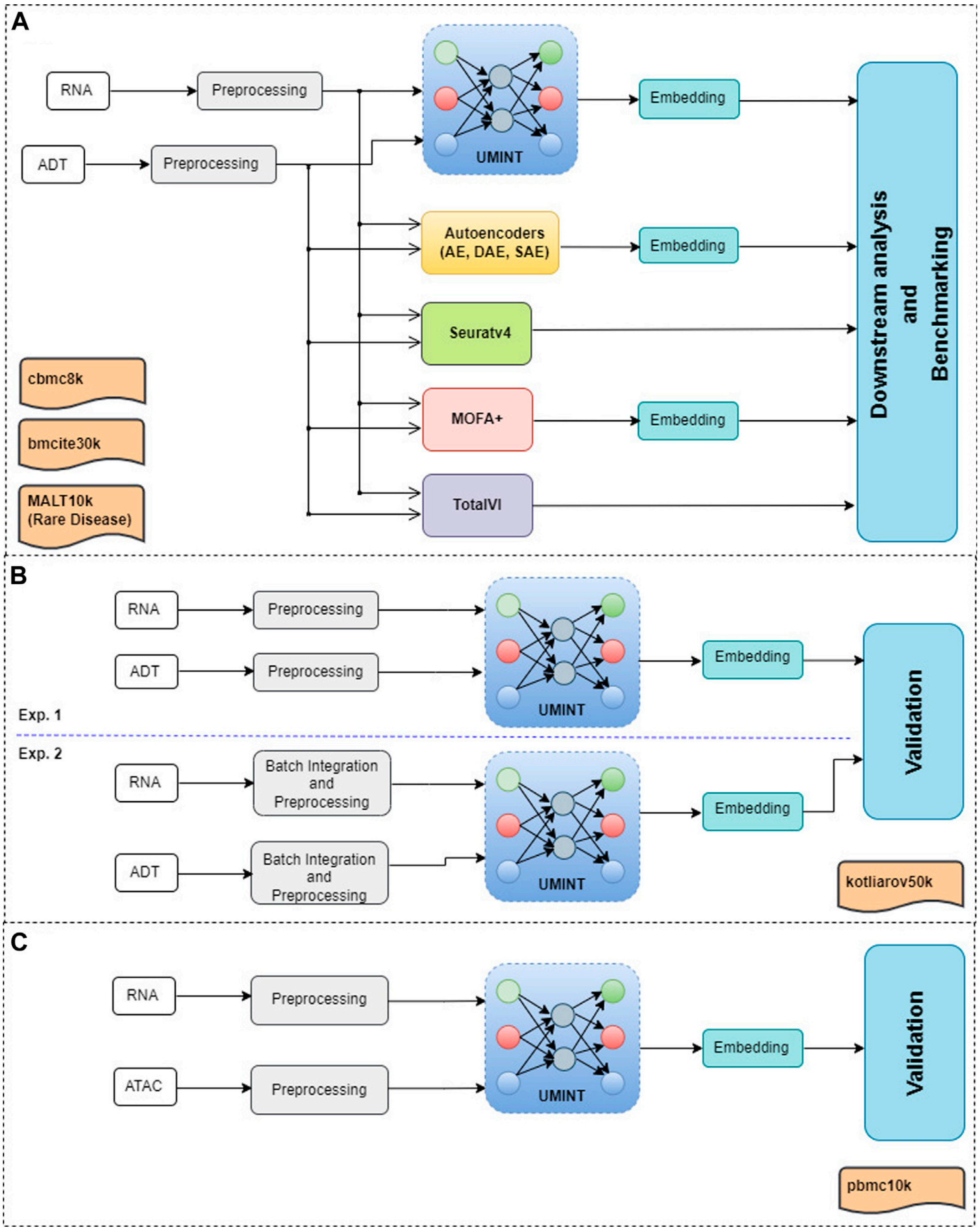


FIGURE 1
A graphical abstract showing the overall workflow of the methods used for evaluation and benchmarking of the proposed method for single cell multi-omics integration, called UMINT. Panel **(A)** shows the primary experiments conducted in this work, as described in Section 3.1 and Section 3.2. Each modalities in *cbmc*8*k*, *bmcite*30*k* and *MALT*10*k* have first been preprocessed using Seuratv4. The preprocessed datasets have been fed as input to UMINT, Autoencoder-based methods (AE, SAE and DAE), Seuratv4, MOFA+ and TotalVI. Seuratv4 and TotalVI are capable of producing a latent low-dimensional embedding and subsequently find cell clusters. The embedding produced by UMINT, AE-based methods and MOFA + have been subjected to k-means and hierarchical clustering. The clustering performance of all the methods have then been compared. Panel **(B)** shows the experiments conducted for *kotliarov*50*k*, where preprocessed data from each modality without batch integration (Exp. 1) and with batch-integration (Exp. 2) done using Seuratv4 SCTransform () have been separately fed as input to UMINT, and the clustering performance on the embeddings generated by UMINT in both these cases have been compared, as explained in Section 3.3. Panel **(C)** depicts the experiments carried out on *pbmc*10*k*, where the two modalities (RNA and ATAC) have been preprocessed using MUON and the integrated embedding produced by UMINT has been assessed for its clustering performance, as explained in Section 3.4.

based methods (like those based on autoencoder networks), used for single cell multi-omics integration. The performance of UMINT has been demonstrated on multiple publicly available healthy and disease datasets. These comprise four CITE-seq datasets, one of which contains cells from MALT tumor, a rare variant of malignant lymphoma. We have benchmarked the results against several existing state-of-the-art algorithms used for single cell multi-omics integration, which fall under different categories. In-silico experimental results compare favourably for UMINT against these state-of-the-art methods. Additionally, the performance of UMINT has been validated through integration of an auxiliary multimodal single cell paired gene expression and ATAC-seq (chromatin accessibility) data. Finally, as a further extension to this work, UMINT has been used to integrate bulk multi-omics data with more than two omics layers, which it has been able to execute with ease. Thus, UMINT's ability to integrate widely heterogenous omics data (CITE-seq, paired RNA-seq and ATAC-seq) with varying number of omics layers boosts its utility as a powerful integration model and makes it completely fit in with the *status quo*.

The remaining part of this article has been organized as follows. Section 2 explains the methodology behind the proposed single cell multi-omics integration technique, called UMINT. Section 3 describes *in silico* experimental results obtained on different datasets used in this work and provides a detailed comparison against other existing methods for single cell multi-omics integration. In Section 4, the strengths and limitations of the proposed method are discussed along with concluding remarks.

2 Methodology

This section describes the datasets used in the experiments, the methodology used for data pre-processing and the proposed neural network model, called UMINT, for single cell multi-omics integration. Figure 1 shows a graphical abstract illustrating the overall workflow of the methods used for single cell multi-omics integration, and the procedures conducted for preprocessing, embedding, validation and benchmarking performed in this work.

2.1 Data acquisition and pre-processing

Initially, four publicly available CITE-seq datasets, viz., *cbmc*8*k* (Stoeckius et al., 2017), *MALT*10*k* (Li et al., 2021), *bmcite*30*k* (Stuart et al., 2019) and *kotliarov*50*k* (Kotliarov et al., 2020), have been used in this work. *MALT*10*k* dataset consists of cells from a MALT tumor, a rare kind of malignant lymphoma (Oh et al., 2006). The datasets have been downloaded as count matrices and pre-processed via Seuratv4 (Hao et al., 2021). For scRNA-seq part in these datasets, we have normalized them by library size to sum up to 10,000, applied a logarithmic transformation, extracted highly variable genes, and finally scaled them linearly (with default parameters). The protein expressions/antibody-derived tag (ADT) datasets have been normalized using the centered log-ratio transformation (Stoeckius et al., 2017). Three proteins, viz., *CCR*5, *CCR*7 and *CD*10, have been removed from *cbmc*8*k* dataset due to poor abundance. The first three datasets, viz., *cbmc*8*k*, *MALT*10*k* and *bmcite*30*k*, have been used to evaluate the performance of UMINT and compare the results against other state-of-the-art algorithms. The fourth dataset, viz., *kotliarov*50*k*, contains filtered cells with highly variable genes only. It has been preprocessed via Seuratv4 and used to assess other performance criteria of the proposed methodology. Another auxilliary single cell multimodal dataset, downloaded from 10x Genomics, has been used at a later stage of the work. It contains paired ATAC and gene expression data from human PBMCs with granulocytes removed through cell sorting (processed with ARC 1.0.0 pipeline). It has been preprocessed via MUON (Bredikhin et al., 2022), and used to evaluate the performance of UMINT on paired RNA-seq and ATAC-seq data. The summary of the single cell multi-omics datasets used in this work have been listed in Table 1.

2.2 Unsupervised neural network for single cell multi-omics INTegration (UMINT)

In this work, we have developed a deep Unsupervised neural network for single cell Multi-omics INTegration (UMINT). UMINT is a non-recurrent feed-forward neural network that is efficient enough to integrate variable number of omics layers and extract a latent embedding at a reduced dimension. The network structure of UMINT represents a novel neural network architecture as shown in Figure 2.

Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_m$ be m datasets corresponding to m different omics modalities having n samples (cells) each with $d_1, d_2, \ldots, d_m$ features (RNAs in case of gene expression data, ADTs in case of protein expression data or Peaks in case of transposase-accessible chromatin data) respectively. The UMINT architecture consists of two sub-architectures - an encoder and a decoder. The encoder accepts data from multi-omics datasets presented to the *Input layer*, transports them through one or more *Modality encoding layer*(s) and integrates them in the final layer, known as the *Bottleneck layer*. The decoder accepts the embedded output from the *Bottleneck layer*, transports them through one or more *Modality decoding layer*s and finally tries to reconstruct the original data at the *Reconstruction layer*. In this work, we have used only one layer each for modality encoding and decoding. However, UMINT may contain multiple such layers based on the requirement. In order to improve generalization capability and reduce dimension of the latent embedding, the number of neurons at the *Bottleneck layer* has been kept smaller than the number of neurons in the *Input layer*.

2.2.1 Forward propagation

The *Input layer* of UMINT consists of m different modules, each of which accepts input from a data modality. The number of neurons in each of the modules in the *Input layer* is equal to the dimensions of the individual data modalities $d_1, d_2, \ldots, d_m$ respectively. In this layer, UMINT tries to find a suitable projection for each of the data modalities that may be good enough to get integrated in subsequent layers. Each module in the first *Modality encoding layer* shares a dense connection with the corresponding modules of the *Input layer*. The first *Modality encoding layer* containing m modules thus accepts data from m modules in the *Input layer* as input, and obtains m different projections. Let $\mathbf{a}_{ji}$ and $\mathbf{h}_{ji}$ be the input to and the output from the *ith* module in the *jth* layer respectively. Then, for the *Input* and *Modality encoding* layers, we have

$$\begin{aligned} \mathbf{a}_{1i} &= \mathbf{x}_i & \mathbf{h}_{1i} &= \mathbf{a}_{1i} \\ \mathbf{a}_{2i} &= \mathbf{W}_{1i}\mathbf{h}_{1i} + \mathbf{b}_{1i} & \mathbf{h}_{2i} &= ReLU\left(\mathbf{a}_{2i}\right) \end{aligned} \tag{1}$$

where $\mathbf{x}_i$ is a sample in *ith* modality, and $ReLU(\mathbf{y}) = max(\mathbf{0}, \mathbf{y})$. The term $\mathbf{W}_{1i}$ denotes the weights between *ith* module of the *Input layer* and *ith* module of the first *Modality encoding layer*, and $\mathbf{b}_{1i}$ denotes

TABLE 1 Summary of datasets used for evaluation of UMINT.

Dataset	Description	#Cells	#RNAs	#ADTs/ #Peaks	Batches present	Healthy/ Disease	Source
cbmc8k (CITE-seq)	scRNAseq and antibody sequencing of CBMCs	8,617	20,501	13	No	Healthy	Stoeckius et al. (2017)
MALT10k (CITE-seq)	Cells from a dissociated Extranodal Marginal Zone B-Cell Tumour (MALT) stained with TotalSeq-B antibodies	8,412	33,538	17	No	Rare disease	Li et al. (2021)
bmcite30k (CITE-seq)	scRNA-seq profiles measured alongside a panel of antibodies from bone marrow	30,672	17,009	25	Yes	Healthy	Stuart et al. (2019)
kotliarov50k (CITE-seq)	CITE-seq profiling of 82 surface proteins and transcriptomes of 53,201 single cells from healthy high and low influenza-vaccination responders	58,654	32,738	87	Yes	Healthy	Kotliarov et al. (2020), Lotfollahi et al. (2022)
pbmc10k (paired RNA-seq and ATAC-seq)	Single cell multiome ATAC and gene expression data from cryopreserved human peripheral blood mononuclear cells (PBMCs) of a healthy female donor	11,909	36,601	108,377	No	Healthy	Bredikhin et al. (2022)

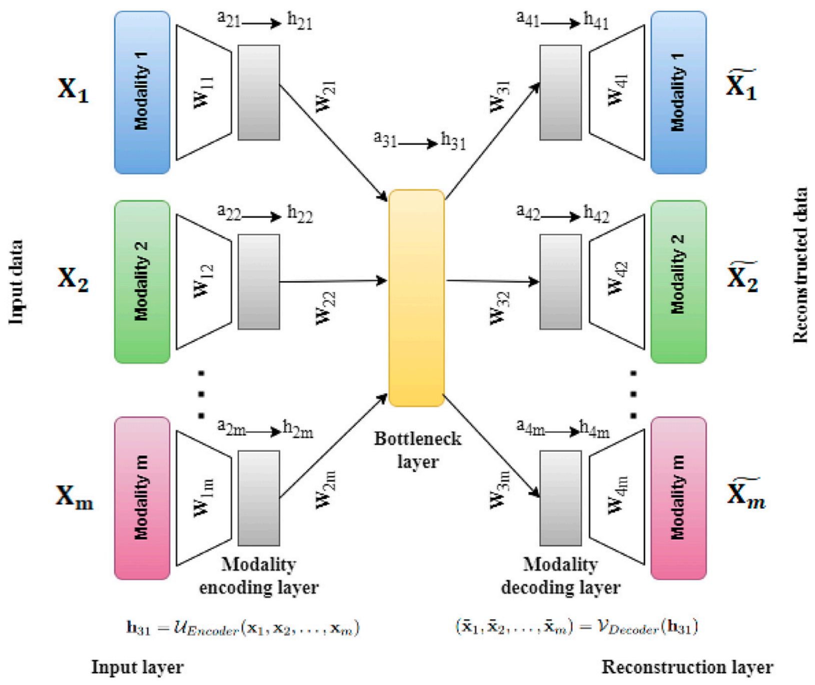


FIGURE 2
Architecture of UMINT showing propagation of input data through the network. Eq. 1 shows how the *Modality encoding layer* encodes each modality fed as input to UMINT. At the *Bottleneck layer*, integration of these modalities is performed using Eq. 2. The encoding process that combines the above mentioned steps is represented in Eq. 3. Eq. 4 shows how the *Modality decoding layer* tries to decode individual modalities and produce reconstructions. The decoding process is represented in Eq. 5. Once a reconstruction is produced, the loss is calculated using Eq. 6. The error is then propagated backwards through the network and the trainable parameters are updated accordingly.

the bias terms of the nodes in *ith Modality encoding layer*. Finally, the outputs from the *Modality encoding layer*(s) are projected onto a lower dimensional space in the *Bottleneck layer*. The final *Modality encoding layer* and the *Bottleneck layer* are fully connected. If $\mathbf{W}_{2i}$ represents the weights between the *ith* module of the final *Modality encoding layer* and the *Bottleneck layer*, then we have

$$\mathbf{a}_{31} = \sum_{i=1}^{m} (\mathbf{W}_{2i}\mathbf{h}_{2i}) + \mathbf{b}_{21} \qquad \mathbf{h}_{31} = ReLU(\mathbf{a}_{31}) \qquad (2)$$

where $\mathbf{b}_{21}$ denotes the bias terms of the nodes in *Bottleneck layer*. This concludes the process of encoding. The overall function of the encoder network can thus be represented as

$$\mathbf{h}_{31} = \mathcal{U}_{Encoder}\left(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\right) \tag{3}$$

Reconstruction of the original data is done by the decoder network in exactly the opposite manner to that of encoding. The integrated embedding coming out of the *Bottleneck layer* is projected onto the *Modality decoding layer*(s) which consists of the same number of modules as that in the *Modality encoding layer*(s). The number of neurons in each module of the *Modality decoding layer* is identical to that used in the modules in the *Modality encoding layer*. The last layer of the decoder is the *Reconstruction layer* which tries to reconstruct the original data from the respective modules in the final *Modality decoding layer*. The process of decoding can be expressed as

$$\begin{array}{ll} \mathbf{a}_{4i} = \mathbf{W}_{3i}\mathbf{h}_{31} + \mathbf{b}_{3i} & \mathbf{h}_{4i} = ReLU\left(\mathbf{a}_{4i}\right) \\ \mathbf{a}_{5i} = \mathbf{W}_{4i}\mathbf{h}_{4i} + \mathbf{b}_{4i} & \mathbf{h}_{5i} = \mathbf{a}_{5i} = \tilde{\mathbf{x}}_i \end{array} \tag{4}$$

where $\mathbf{W}_{ji}$ and $\mathbf{b}_{ji}$ represent weights and biases for the *ith* module in the *jth* layer respectively. Thus, the decoder function is given by

$$\left(\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots, \tilde{\mathbf{x}}_m\right) = \mathcal{V}_{Decoder}\left(\mathbf{h}_{31}\right) \tag{5}$$

where $\tilde{\mathbf{x}}_i$ denotes the reconstruction for $\mathbf{x}_i$.

2.2.2 Objective function

In this scope of work, UMINT has been initially used to integrate scRNA-seq and single cell protein expression data. Subsequently, it has been used to integrate scRNA-seq and ATAC-seq data. Thus, for each dataset, paired RNA and ADT assays, or paired RNA and ATAC assays form the inputs to different modules of the *Input Layer* of UMINT. For each cell $\mathbf{x}$, UMINT tries to find an optimal reconstruction $\tilde{\mathbf{x}}$ of the input data retaining as much information as possible, thereby minimizing the reconstruction error $\|\mathbf{x} - \tilde{\mathbf{x}}\|$. The reconstruction error is contributed by reconstruction loss from each modality. In order to avoid biasness arising out of number of dimensions in the input modalities, we have introduced a balancing parameter λ_i. Additionally, in order to limit over-fitting, we have used an $L1$ regularization on the nodes' activities to allow sparsity of nodes' outputs and an $L2$ regularization on the weight values since $L2$ regularization tries to shift weight values towards zero. Both $L1$ and $L2$ regularizations minimize the model complexity. In this work, $L1$ and $L2$ regularizations have been controlled using regularization parameters α and β respectively. The objective function thus becomes

$$\mathcal{L}_{UMINT} = \frac{1}{n}\sum_{i=1}^{m}\lambda_i\|\mathbf{X}_i - \tilde{\mathbf{X}}_i\|^2 + \sum_{i=1}^{m}\sum_{j=2}^{4}\left(\alpha\|\mathbf{h}_{ji}\|_1 + \beta\|\mathbf{W}_{ji}\|_2\right) \tag{6}$$

Values of the regularization parameters α and β have been set to 0.0001 and 0.001 respectively, as recommended in literature (Chaudhary et al., 2018). UMINT has been trained for 25 epochs using Adam Optimizer (Kingma and Ba, 2017) with a batch size of 16 and Eq. 6 as the loss function. During the forward pass, the data is fed as input to the encoder. A lossy reconstruction of the input data is produced by the decoder at the *Reconstruction layer*. The error value is then propagated backwards, and the weights and biases are updated for a better reconstruction in the next forward pass.

2.3 Latent low-dimensional embedding and clustering

At the outset of this work, the proposed integration model, called UMINT, has been used to integrate RNA and protein expression data. Once trained to reconstruct the input data, UMINT is capable of learning a latent low-dimensional embedding that extracts relevant features from the integrated data. Here, we have used UMINT to extract a latent embedding of 64 dimensions. This latent embedding has been used in the subsequent step for downstream analysis in order to explore its effectiveness. We have used agglomerative hierarchical clustering and k-means clustering algorithm on this latent embedding to cluster the cell-types for each of the datasets used in the study. The performance of UMINT has then been compared against existing benchmark methods used for multi-omics integration. We have used two measures, viz., Adjusted Rand Index (ARI) and Fowlkes Mallows Index (FMI) scores, to measure the degree of agreement between the actual and predicted cell-types, for all the methods used in comparison including UMINT. The actual cell types corresponding to the ground truth data have been obtained from the corresponding source datasets mentioned in Table 1.

For two sets of cluster labels, the overlap between them is represented by a contingency table $\mathbf{C} = [c_{ij}]$, where c_{ij} indicates the total number of points belonging to both *ith* cluster of the first set and *jth* cluster of the second set. ARI is an external cluster validity index, and is thus defined as

$$ARI = \frac{\sum_{ij}\binom{c_{ij}}{2} - \left[\sum_i\binom{p_i}{2}\sum_j\binom{q_j}{2}\right]\Big/\binom{N}{2}}{\frac{1}{2}\left[\sum_i\binom{p_i}{2} + \sum_j\binom{q_j}{2}\right] - \left[\sum_i\binom{p_i}{2}\sum_j\binom{q_j}{2}\right]\Big/\binom{N}{2}} \tag{7}$$

where $p_i = \sum_j c_{ij}$, $q_j = \sum_i c_{ij}$ and $N = \sum_{ij} c_{ij}$ respectively. An ARI value close to 1 indicates good resemblance between two clusters. Similarly, FMI, another external evaluation index used to measure the similarity between two sets of cluster labels, is defined as

$$FMI = \frac{TP}{\sqrt{(TP + FP)(TP + FN)}} \tag{8}$$

where TP, FP and FN denotes the count of True Positives, False Positives and False Negatives respectively. The FMI score lies between 0 and 1, and a high value implies a good similarity between two clusters.

At a subsequent stage of the work, UMINT has been evaluated on another multiome dataset containing paired gene expresssion and ATAC-seq data. After preprocessing each modality and reducing them to highly variable features, UMINT has been used to extract a latent 64-dimensional embedding by integrating the RNA and ATAC assays. This latent embedding has been further subjected to agglomerative hierarchical clustering and k-means clustering. The embedding quality has been assessed using external evaluation criteria like, ARI and FMI, as explained above.

3 Results

UMINT has been applied on a variety of datasets containing cells from both healthy donors as well as donors with a disease, and its performance has been evaluated over multiple steps as depicted in the graphical abstract shown in Figure 1. Initially, it has been benchmarked on three CITE-seq datasets, viz., *cbmc*8*k*, *MALT*10*k* and *bmcite*30*k*, where *cbmc*8*k* and *bmcite*30*k* contain cells from healthy samples, and *MALT*10*k* contains cells from rare lymphoma (MALT). The latent embedding produced by UMINT has been first compared with that produced by Autoencoder (AE)-based architectures. Subsequently, the UMINT-generated embedding has been compared with three other state-of-the-art single cell multi-omics integration methods. Section 3.1 and Section 3.2 describe results of comparison of UMINT against AE and other state-of-the-art methods on these three datasets. Thereafter, UMINT has been further tested for its performance on another CITE-seq dataset *kotliarov*50*k* having multiple batches. For evaluating the batch correction performance of UMINT, we have performed two different experiments on *kotliarov*50*k* dataset, with and without batch integration. Finally, as an extended utility, UMINT has been used to integrate paired gene expression and ATAC-seq data. The integrated embedding generated by UMINT has then been validated through clustering techniques. Additionally, UMINT has been validated on a bulk multi-omics cancer dataset for integration and classification.

3.1 Comparison with autoencoder-based unsupervised neural network models

An AE network (Hinton and Salakhutdinov, 2006), which is an unsupervised neural network used for dimension reduction, being the closest resemblance to UMINT, we have first compared it with a regular AE network and its variations, viz., Denoising AE (DAE) and Sparse AE (SAE), both in terms of architectural difference and performance.

3.1.1 Comparison with autoencoder-based models with respect to number of trainable parameters and execution time

Similar to an AE network, UMINT also tries to reconstruct the original input as explained in Section 2.2. However, there is a difference between the two. The input layer in AE shares a dense connection with the first hidden layer, whereas, the connections between the *Input layer* and the first *Modality encoding layer* in UMINT is not dense. This reduces the number of parameters to be trained, drastically. Although, in this work, UMINT has been used to integrate single cell RNA and protein expression data, it is quite capable of integrating any number of omics layers. Let us consider that the input to UMINT consists of data from m modalities having n samples each with $d_1, d_2, \ldots, d_m$ dimensions respectively. As shown in Figure 2, UMINT consists of the same number of modules in the *Modality encoding layer* as that in the *Input layer*. If the number of neurons in each of the module of this *Modality encoding layer* are $n_1, n_2, \ldots, n_m$ respectively, then the total number of trainable parameters (TP_{UMINT}) between the *Input layer* and the *Modality encoding layer* in the encoder network becomes

$$TP_{UMINT} = \sum_{i=1}^{m} d_i n_i \tag{9}$$

Considering an input of similar dimensions, if an AE network is employed to achieve this same task of integration, the number of trainable parameters (TP_{AE}) between the input layer and the first hidden layer becomes

$$TP_{AE} = \left(\sum_{i=1}^{m} d_i\right)\left(\sum_{i=1}^{m} n_i\right) \tag{10}$$

Thus, the reduction in the number of trainable parameters ($TP_{Reduction}$) in the encoder network is given by

$$TP_{Reduction} = TP_{AE} - TP_{UMINT} = \sum_{i=1}^{m}\left(d_i \sum_{j=1, j\neq i}^{m} n_j\right) \tag{11}$$

A reduction in the number of trainable parameters by the same amount is also available at the *Reconstruction layer* of the decoder network. Hence, the total reduction ($TP_{TotalReduction}$) in the number of trainable parameters in UMINT is given by

$$TP_{TotalReduction} = 2 \times TP_{Reduction} = 2 \times \sum_{i=1}^{m}\left(d_i \sum_{j=1, j\neq i}^{m} n_j\right) \tag{12}$$

This is a massive improvement over AEs considering more than one modality of data to be integrated. UMINT network reduces to a regular AE network if a single modality is used which, however, does not serve the purpose of integration.

We have further recorded the execution time taken by both UMINT and the different variations of AE for integration of single cell multi-omics data. This experiment has been repeated multiple times with different training and test datasets to ensure stability of results. As shown in Figure 3, we have observed that integration using UMINT has been much faster as compared to that obtained using different AE-based networks, like a regular AE, DAE and SAE. Thus, we can say that UMINT not only has a light-weight architecture than AE-based networks, but it is also computationally less expensive than them.

3.1.2 Comparison with autoencoder-based models with respect to performance

Initially, the performance of UMINT has been compared with that of a standard AE and its variations (DAE, SAE) with respect to their reconstruction capability and the strength of the latent low-dimensional embeddings produced at the bottleneck layer by each network. As mentioned earlier, RNA and protein expression data form the input to different modules of the *Input Layer* of UMINT which has then been used to reconstruct the input data. On a similar note, RNA and protein expression data have been stacked together to form the input to the AE-based networks. The AE-based networks reconstruct the combined input by passing it through a series of layers. In the process, they learn to extract useful features at the bottleneck layer, where a latent embedding of 64 dimensions is produced, similar to UMINT. The models UMINT, AE, DAE and SAE have been trained keeping all hyper-parameter values identical. For each modality, the amount of correlation between the original data and its reconstruction, has then been computed for UMINT and all the AE-based models using Pearson correlation coefficient. We have then defined an Overall Reconstruction Score (ORS) to assess the reconstruction performance of UMINT against that of the

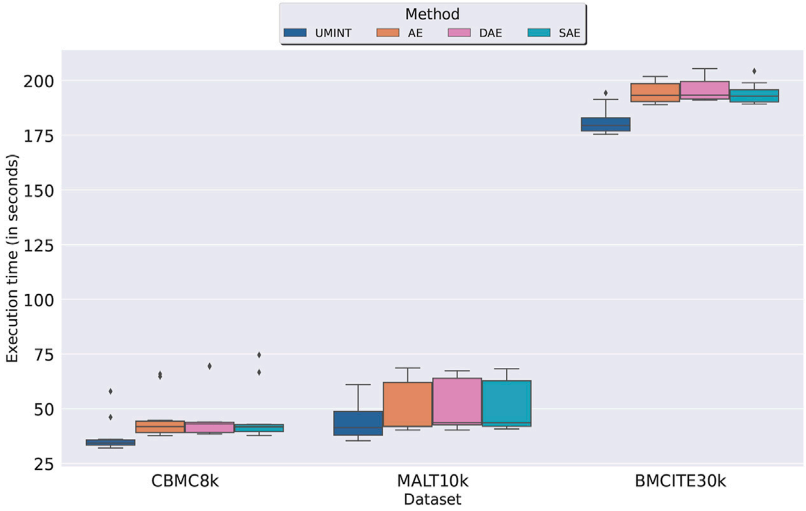


FIGURE 3
Comparison of execution time taken for single cell multi-omics data integration by UMINT with that taken for integration by different variations of AE.

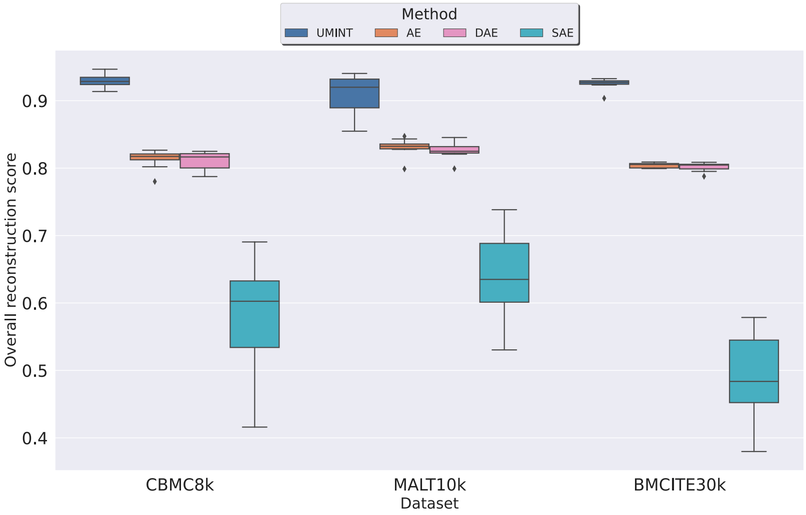


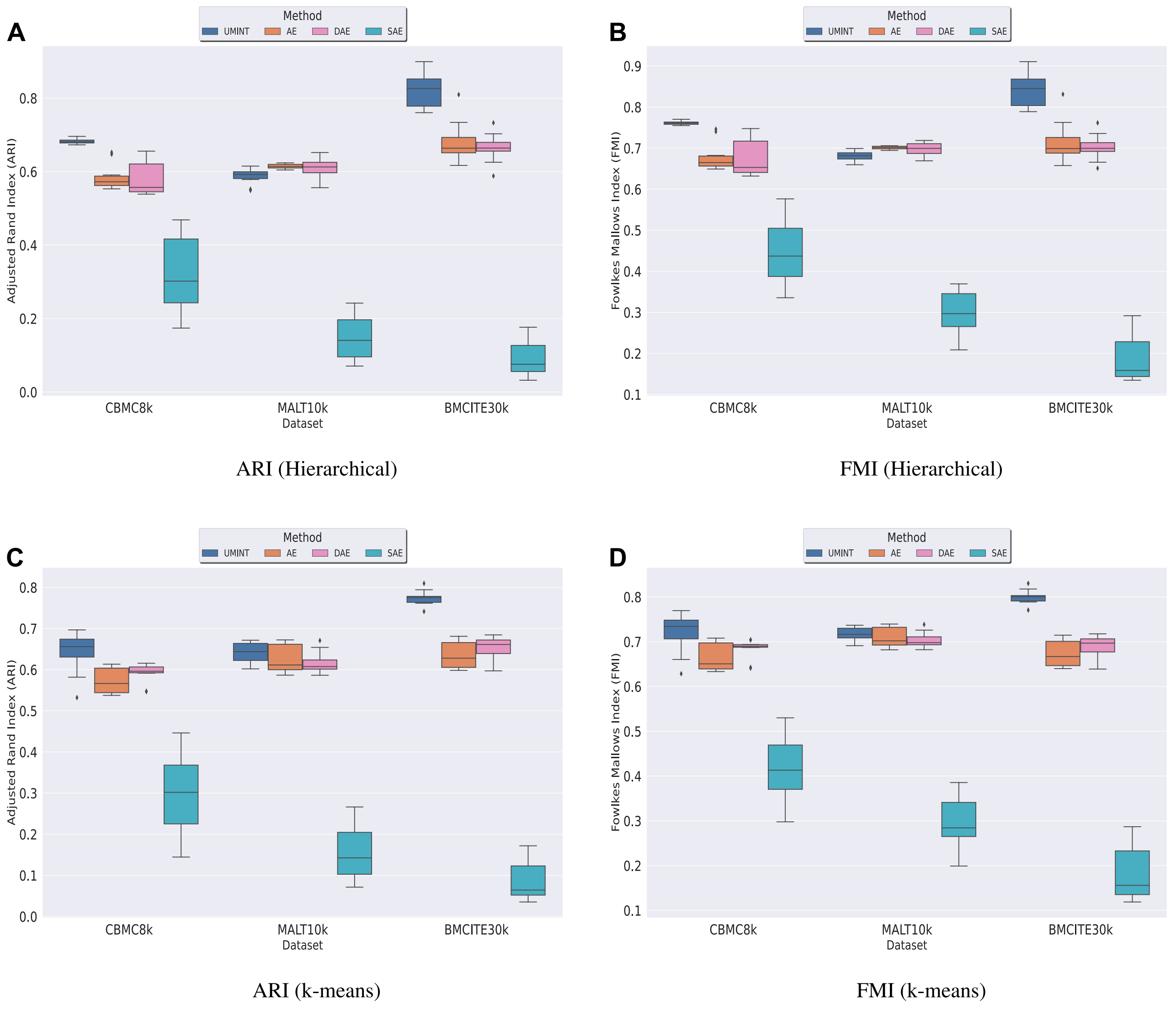
FIGURE 4
Comparison of performance of UMINT with that of a regular AE, and its variations DAE and SAE with respect to overall reconstruction of RNA and ADT modalities in *cbmc*8*k*, *MALT*10*k* and *bmcite*30*k* datasets.

AE-based models, based on the omics modalities used for integration as follows

$$ORS = \frac{1}{m}\sum_{i=1}^{m}\rho_i \tag{13}$$

where ρ_i is the Pearson correlation coefficient value between the pairwise distances in the original *ith* data modality and the pairwise distances in its reconstructed counterpart. As shown in Figure 4, we have observed that UMINT has outperformed all the AE-based networks with respect to overall reconstruction of the omics modalities, for all the three datasets used for evaluation, with Median Correlation Coefficient (MCC) values and corresponding *p*-values as recorded in Supplementary Table S1. All the experiments have been repeated multiple times with different training and test datasets divided in a 80 : 20 ratio. The results of the test for statistical significance thus obtained, have made us infer that UMINT is capable of producing better overall reconstructions than AE-based methods.

Thereafter, we have compared the latent low-dimensional embedding produced by UMINT with that produced by a standard AE and its variations. Once trained to create a lossy reconstruction of the input, the latent representation has been extracted from the bottleneck layer of UMINT and all AE-based



ARI (Hierarchical)

FMI (Hierarchical)

ARI (k-means)

FMI (k-means)

FIGURE 5
Comparison of clustering performance of UMINT against that of AE-based methods when agglomerative hierarchical clustering is used, as measured by **(A)** ARI and **(B)** FMI; **(C)** and **(D)** show clustering performance of UMINT compared to AE-based methods, as measured by ARI and FMI respectively, when k-means clustering algorithm is used.

TABLE 2 A theoretical comparison between UMINT and other methods used for comparison.

Method	Methodology	Produces latent embedding (Yes/No)	Support cell-type clustering (Yes/No)	Omics integration supported for	Makes assumption about data distribution (Yes/No)	Can reconstruct original data (Yes/No)
UMINT	Neural network based	Yes	No	Both single cell and bulk	No	Yes
Autoencoder	Neural network based	Yes	No	Both single cell and bulk	No	Yes
Seuratv4	Graph based	Yes	Yes, via Louvain, Leiden and SLM	Single cell only	No	No
MOFA+	Matrix factorization based	Yes	No	Both single cell and bulk	Yes	Yes
TotalVI	Neural network based	Yes	No (recommends using Scanpy)	Single cell only	Yes	Yes

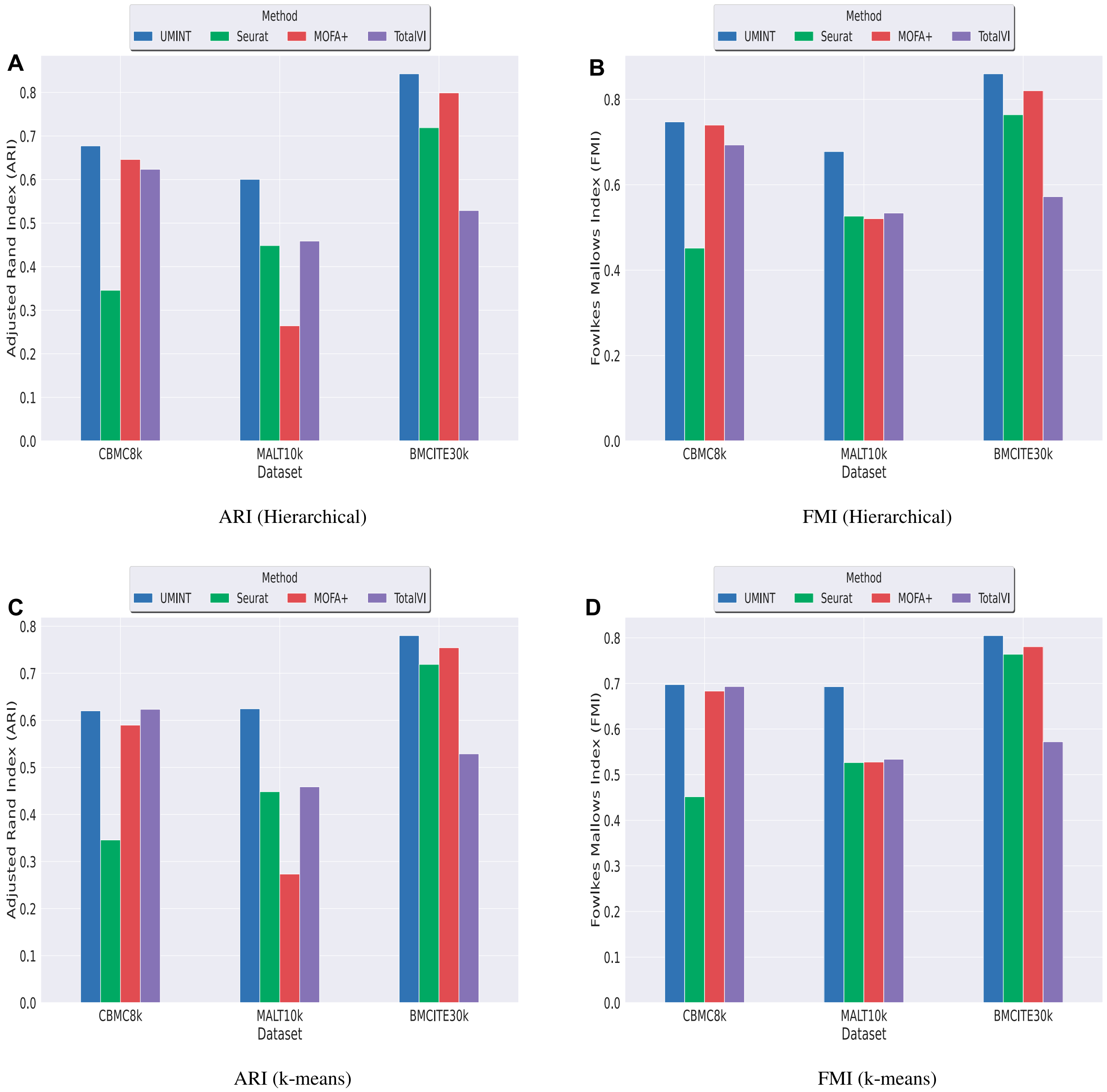
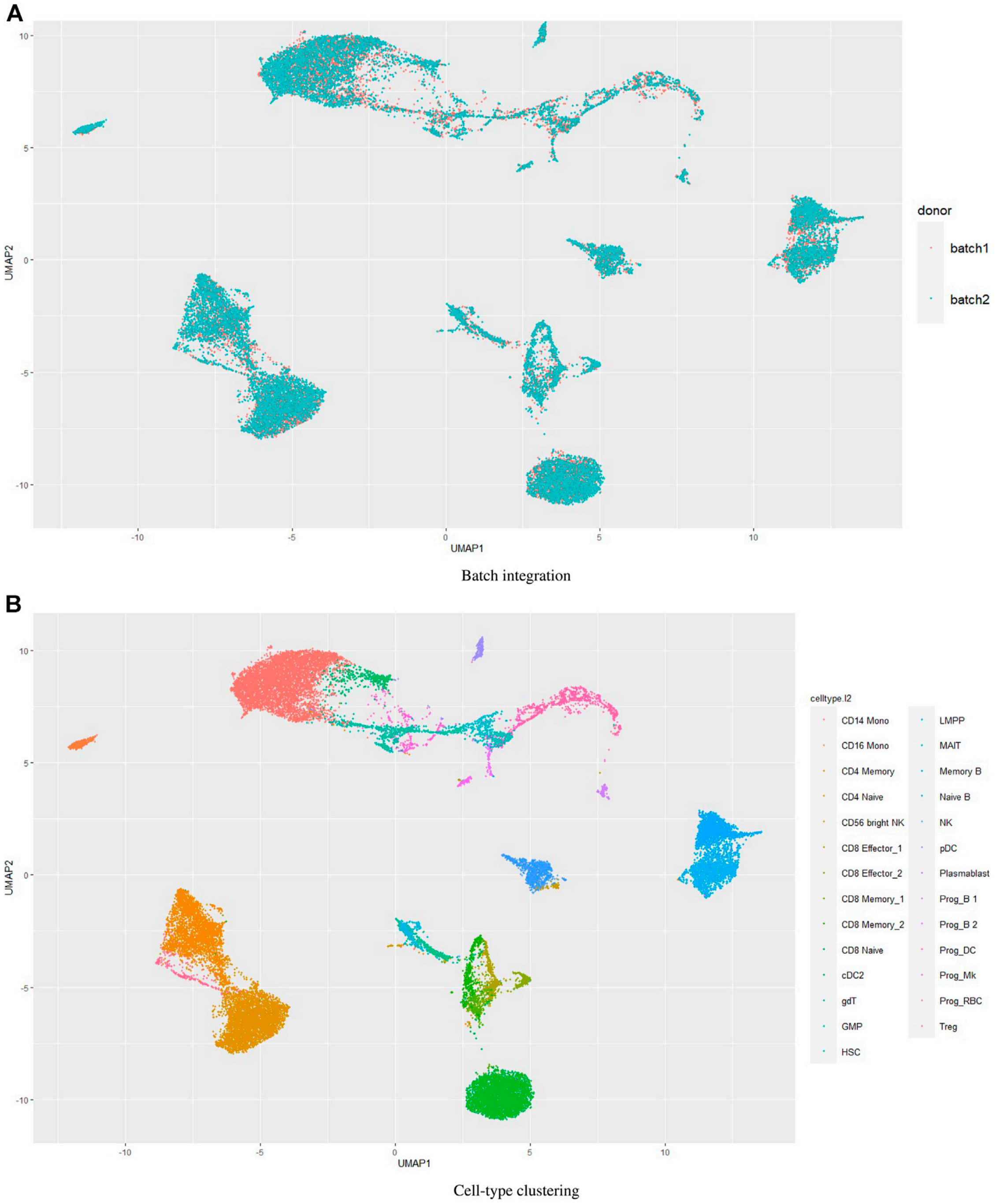


FIGURE 6
Comparison of clustering performance of UMINT against Seuratv4, MOFA+ and TotalVI when agglomerative hierarchical clustering is used, as measured by **(A)** ARI and **(B)** FMI; **(C)** and **(D)** show performance of each method, as measured by ARI and FMI respectively, when k-means algorithm is used for clustering.

models. Cell-type clustering on this latent embedding using k-means and agglomerative hierarchical clustering has been performed to validate the effectiveness of UMINT and compare it with AE-based models using ARI and FMI scores, as discussed earlier.

We have observed that for *cbmc8k* and *bmcite30k* datasets, the latent representation produced by UMINT has been more representative of the cell clusters when compared with that of AE-based methods. For *MALT10k* dataset, when hierarchical clustering algorithm is used, UMINT embedding has produced similar ARI and FMI scores to that obtained on embedding produced by AE-based methods. This is indicated by both ARI and FMI scores as shown in Figures 5A–D. Median ARI (MARI), Median FMI (MFMI) scores along with the corresponding *p*-values obtained using UMINT and the AE-based models on these three datasets for both hierarchical and k-means clustering algorithms have been shown in Supplementary Table S2. All the experiments have been repeated multiple times with different training and test datasets to ensure stability of the results.



Batch integration

Cell-type clustering

FIGURE 7
Performance of UMINT on *bmcite30k* dataset with respect to **(A)** batch integration, **(B)** cell-type clustering.

3.2 Comparison with other state-of-the-art methods for single cell multi-omics integration

Subsequently, we have compared the performance of UMINT with three other state-of-the-art methods - Seuratv4 (Hao et al., 2021), MOFA+ (Argelaguet et al., 2020) and TotalVI (Gayoso et al., 2021). We have chosen these three methods since they represent three different categories of algorithms - Graph based, Matrix factorization based and Neural network based, developed for single cell multi-omics integration (Stanojevic et al., 2022). To ensure a fair comparison between all these methods, we have followed the same preprocessing pipeline for all the datasets used for comparison. The effectiveness of UMINT has once again been demonstrated by clustering the cell-types on the latent low-dimensional embedding produced by it. It may be mentioned here that Seuratv4 is capable of producing an integrated low-dimensional representation through weighted-nearest neighbor

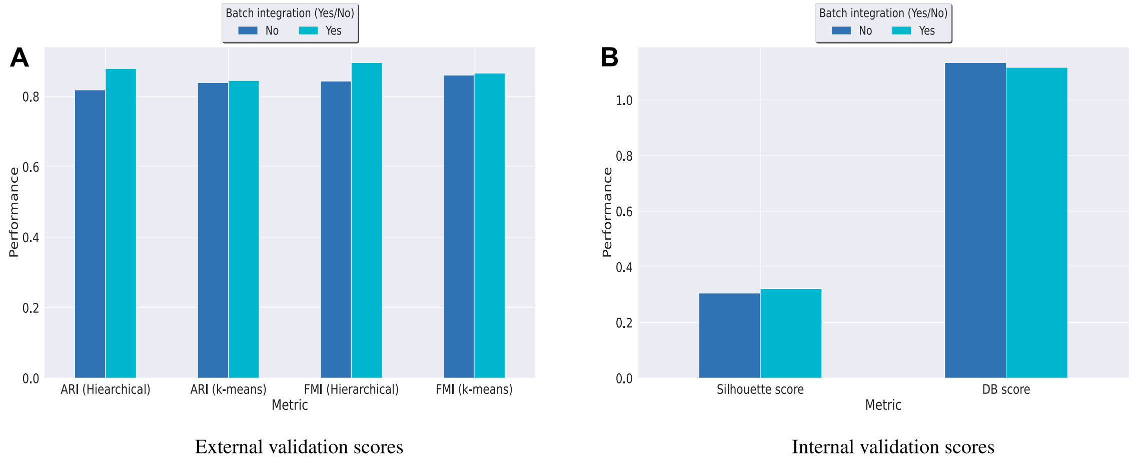


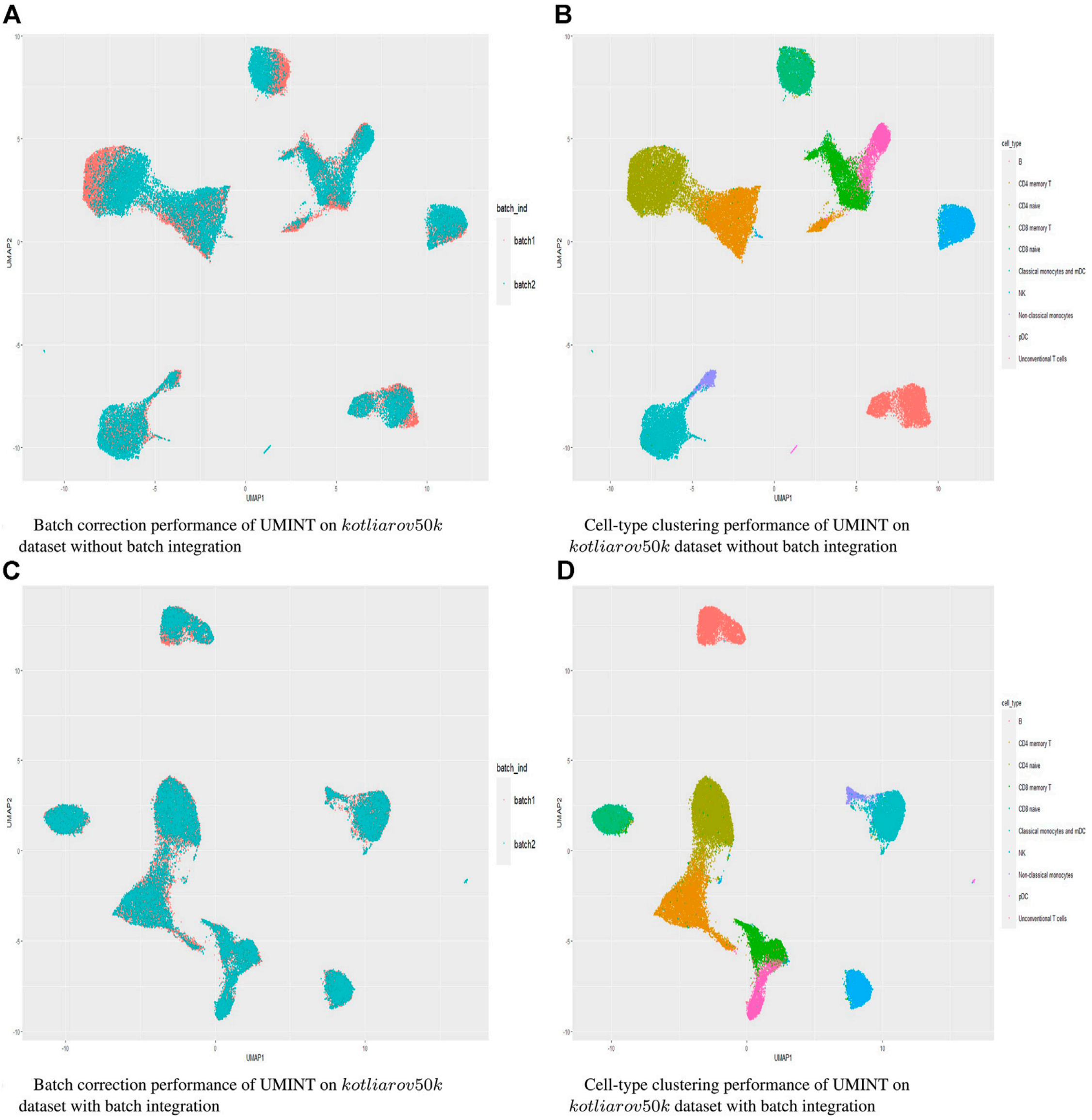
FIGURE 8
Clustering performance of UMINT-generated embeddings obtained from RNA and ADT data with and without batch integration on the *kotliarov*50*k* dataset as measured by **(A)** external validity indices and **(B)** internal validity indices.

analysis, and also find cell clusters from the integrated embedding using Louvain (Blondel et al., 2008), Leiden (Traag et al., 2019) or SLM (Waltman and Van Eck, 2013) community-detection algorithms. TotalVI, on the other hand, integrates the data through variational inferencing and autoencoding, and uses the standard Scanpy (Wolf et al., 2018) pipeline for clustering on the latent embedding. MOFA+, however, only produces a low-dimensional representation of the integrated data, as in the case of UMINT. The methods used in this work for comparison have been compared theoretically in Table 2. Thus, for Seuratv4 and TotalVI, we have extracted the cluster labels, while for MOFA+, we have extracted the factors representing the low-dimensional embedding and used hierarchical and k-means clustering on the same, similar to UMINT, as illustrated in the graphical abstract shown in Figure 1. Interestingly, UMINT has outperformed all three methods in terms of ARI and FMI scores, validated both by hierarchical and k-means clustering. Figure 6 shows the average ARI and FMI scores obtained using UMINT plotted against the scores obtained by the three benchmark methods.

3.3 Performance of UMINT on multi-batch datasets

Batch effects in single cell datasets pose great challenges in data integration and compromises the results (Haghverdi et al., 2018; Tran et al., 2020). We wondered how UMINT would perform when there are batches in the data. The dataset *bmcite*30*k* used in this work contains two batches. However, we have not performed batch integration on this dataset. Figures 4–6 show the performance of UMINT on *bmcite*30*k* dataset when no batch integration has been performed. We have further observed that batches present in the *bmcite*30*k* projection by UMINT have been well integrated and are thus inseparable. Additionally, cell clusters obtained on UMINT projection are cohesive and well separated too. Thus, we can say that besides cell-type clustering, UMINT may have the potential to integrate batches in data efficiently, as shown in Figure 7. However, validation on a single dataset might not establish the strength of UMINT in terms of batch correction since the *bmcite*30*k* dataset itself may not have strong batch effects.

Hence, in order to reinforce our findings, we have performed a few more experiments on another dataset *kotliarov*50*k*. This dataset, collected from (Lotfollahi et al., 2022), contains filtered data for 52,117 cells with highly variable genes (3,999), and two batches of RNA and protein expressions each. Moreover, it contains expression values for 87 proteins, a lot more than the three other datasets used in this work. We have first integrated batches using Seuratv4 SCTransform () (Hao et al., 2021) module with default parameters and fed the batch integrated RNA and ADT datasets to UMINT. The low-dimensional embedding produced by UMINT has then been evaluated for clustering and batch integration performance. In another experiment, we have fed the preprocessed RNA and ADT datasets (without batch integration) into UMINT, and evaluated the low-dimensional embedding produced by it for clustering and batch integration performance too. Apart from two external validity indices, we have used two internal validity indices - silhouette coefficient (Rousseeuw, 1987) and Davies Bouldin (DB) index (Davies and Bouldin, 1979) to measure the clustering performance of the UMINT-generated embeddings on the omics data with and without batch integration. Interestingly, we have observed that when the UMINT embedding has been generated from the RNA and ADT data without batch integration, the ARI, FMI, Silhouette and DB scores achieved have been quite close to those achieved when UMINT embedding has been generated from batch integrated RNA and ADT data, as shown in Figure 8. Thus, it is clear that even without batch integration, UMINT can extract most relevant features from the data that can act as input to further downstream investigations. However, the UMINT-generated embedding obtained on batch integrated data has shown better batch correction performance than that obtained on data without batch integration. From Figures 9A, C, it can be observed that batches in *kotliarov*50*k* data remain separable if batch integration is not performed on the dataset explicitly. This explains why



Batch correction performance of UMINT on *kotliarov*50*k* dataset without batch integration

Cell-type clustering performance of UMINT on *kotliarov*50*k* dataset without batch integration

Batch correction performance of UMINT on *kotliarov*50*k* dataset with batch integration

Cell-type clustering performance of UMINT on *kotliarov*50*k* dataset with batch integration

FIGURE 9
(A) and **(B)** show batch correction and clustering performance of UMINT on the *kotliarov*50*k* dataset without batch integration; **(C)** and **(D)** show similar results on the *kotliarov*50*k* dataset with batch integration.

TABLE 3 ARI and FMI scores obtained on applying k-means and hierarchical clustering on the UMINT-generated latent embedding of *pbmc*10*k* multiome dataset.

k-means		Hierarchical	
ARI	**FMI**	**ARI**	**FMI**
0.69	0.74	0.73	0.77

performance on *kotliarov*50*k* dataset without explicit batch correction is not as good as that on the same dataset with batch correction. Thus, there is further scope of improvement for UMINT in terms of batch correction performance. Figures 9B, D show cell-type clustering performance of UMINT on the *kotliarov*50*k* dataset, without and with batch integration respectively.

3.4 Performance of UMINT on paired RNA-seq and ATAC-seq data

Finally, the performance of UMINT has been assessed on another multiome dataset containing paired gene expression and ATAC-seq assays. This *pbmc*10*k* dataset has been first preprocessed via MUON (Bredikhin et al., 2022) and reduced to highly variable features only. These reduced datasets have been further processed to match the cells in the two modalities. The two paired assays, RNA and ATAC, have then been fed as input to UMINT, which has successfully extracted a latent low-dimensional embedding out of the integrated data. In order to validate the embedding quality, we have used both hierarchical and k-means clustering techniques on the UMINT-generated embedding and measured the clustering performance using ARI and FMI scores.

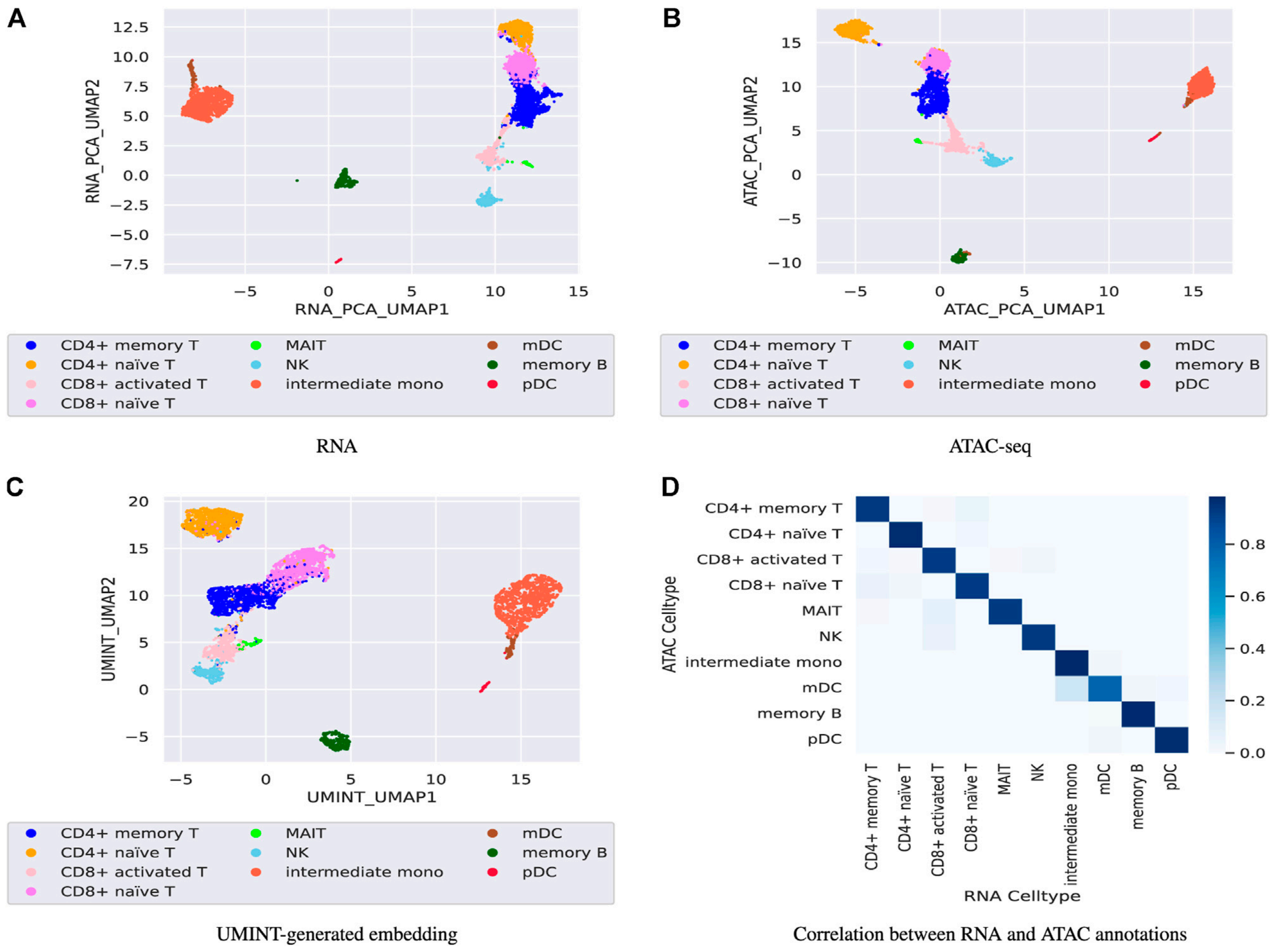


FIGURE 10
Figures **(A)**, **(B)** show UMAP projections on the individual RNA-seq and ATAC-seq data respectively after PCA-based dimension reduction while, Figure **(C)** show UMAP projection on the UMINT-generated latent embedding of *pbmc10k* multiome dataset. Figure **(D)** shows the correlation between the RNA and ATAC annotations.

The average scores over multiple runs of this experiment has been reported in Table 3, which shows that the UMINT-generated embedding has been pretty efficient in clustering the cell-types. The UMAP-projections on the individual modalities (after PCA-based dimension reduction) and on the embedding produced by UMINT have been illustrated in Figures 10A–C respectively. Cell-type annotation for the UMAP plot on the UMINT-generated embedding has been performed using the RNA-annotations since the ATAC annotations highly correlate to the RNA annotations as shown in Figure 10D. Thus, we can say that besides CITE-seq data, UMINT is competent enough to integrate paired RNA-seq and ATAC-seq assays too.

3.5 Performance of UMINT on bulk multi-omics data

As an extension to this work, in order to support the claim that UMINT can integrate a variable number of omics layers, we have further assessed UMINT for its integration performance on bulk expression datasets with more than two modalities. TCGA multi-omics data for Liver Hepatocellular Carcinoma (LIHC) from TCGA portal (now relocated to Genomic Data Commons https://gdc.cancer.gov/), have been used for this purpose. Pre-processed datasets have been collected from our earlier work (Seal et al., 2020) in which we tried to estimate gene expression surrogates from genetic and epigenetic features through a DL pipeline. This dataset contains three omics layers - DNA methylation (DNAm), Copy Number Variation (CNV) and RNA-seq. It contains 404 paired samples out of which 359 are cancer and 45 are normal. The procedures conducted in this separate experiment and their corresponding results have been described in Supplementary Section S2.

4 Discussion and conclusion

In this work, we have introduced a novel deep Unsupervised neural network for single cell Multi-omics INTegration (UMINT). We have used UMINT to integrate heterogenous single cell omics modalities and extract meaningful projections at reduced dimensions. These features have been further used for clustering.

The effectiveness of UMINT has been first demonstrated on four publicly available CITE-seq datasets, and compared on three of them with some other state-of-the-art algorithms used for single cell multi-omics integration. One of the three datasets used for benchmarking corresponds to MALT rare disease, which establishes the applicability of UMINT on rare diseases as well. Thereafter, the performance of UMINT has been assessed on an auxiliary dataset containing paired gene expression and ATAC-seq assays.

The strengths or advantages of UMINT are many-fold. UMINT-generated latent embedding has been proved to produce better clustering than that generated using AE-based methods. UMINT has a light-weight architecture in terms of the number of trainable parameters. Even then, UMINT-generated reconstruction has been better than that produced by AE-based methods. When evaluated against other state-of-the-art algorithms, UMINT has displayed superior performance over every other method used for comparison, across most of the evaluation criteria on all the three CITE-seq datasets. The fact that UMINT can integrate both CITE-seq, and paired RNA-seq and ATAC-seq data fortifies its strength over other existing single cell multi-omics integration methods. Moreover, UMINT does not make assumptions about the underlying data distribution, thus making it more robust. Finally, UMINT has been found to be competent enough to integrate bulk multi-omics datasets too. It has been able to produce better reconstructions for bulk omics data than that obtained using a standard AE. UMINT-extracted features from bulk multi-omics data, have also shown superior classification of tumour and normal samples. Integration of both single cell and bulk multi-omics datasets imply that UMINT supports integration of widely heterogenous and varying number of omics modalities. Very few such integration methods exist that can efficiently integrate features from both single cell and bulk multi-omics, and can handle variable number of omics layers. UMINT's capacity to embed features from healthy and disease omics (including a rare disease) also demonstrates its applicability across varying health conditions.

UMINT, however, is susceptible to batch effects to some extent. It has been able to correct batches for *bmcite*30*k* dataset well (though the inherent batch effect in this dataset is subject to further investigations), while for *kotliarov*50*k* data, integration has been compromised by a tolerable amount due to batch effect. Further, there is a huge imbalance of features in CITE-seq, and the paired RNA-seq and ATAC-seq data. Despite this imbalance, the overall embedding produced by UMINT remains unaffected. In the current scope of work, we have not explored the integration of other high throughput omics modalities like spatial transcriptomics. Sequencing-based spatial transcriptomics data like 10x Visium are still not available at a single cell resolution. They are at spot level which may contain around 10–30 cells per spot on an average which hinders pairing of input samples (cells and spots). Presently, deconvolution methods are still a better choice for interrogating spatial trancriptomics with single cell transcriptomics. We have also not optimized the model at this stage for image-based data, hence multi-omics spatial data, like NanoString GeoMx, MERSCOPE using MERFISH technology, cannot be still used for integration. Inclusion of such omics layer(s) will need inclusion of Convolutional Neural Network (CNN)-based DL models into the existing UMINT architecture. This would further allow us to better understand the overall contribution of each omics layer at a single cell and spatial level to decipher regulatory systems biology on top of scRNA-seq, scATAC-seq and protein expression data with a spatial location. The potentiality of UMINT to select features from each of the input modalities has also not been explored in the current scope of work. Instead, UMINT has been used to extract relevant features from the integrated data at a low dimension. All these remain as a future extension and a scope for improvement for UMINT to identify key molecular anchors in development and disease biology.

Nevertheless, UMINT can capture better variability among high-dimensional datasets and produce robust low-dimensional embedding which can significantly assist further downstream analyses. A reduction in the number of trainable parameters also makes UMINT far less computationally expensive than existing neural network models based on AEs. Thus, we are able to provide a robust and efficient unsupervised deep learning model for single cell multi-omics integration.

Data availability statement

UMINT has been implemented in Python 3. The codes to reproduce the results are freely available at https://github.com/deeplearner87/UMINT. GitHub repository has been organized into three main directories—Preprocessing, Proposed and Benchmarking. The Preprocessing directory (https://github.com/deeplearner87/UMINT/tree/main/Preprocessing) contains codes (R scripts and IPython notebooks) for preprocessing the datasets used in this study. The Proposed directory (https://github.com/deeplearner87/UMINT/tree/main/Proposed) contains the script for the proposed method umint.py and notebooks for the pipeline executed on various datasets. Notebooks corresponding to the comparative analysis made in this work are contained in the directory Benchmarking (https://github.com/deeplearner87/UMINT/tree/main/Benchmarking). The preprocessed datasets used in this work can be downloaded from https://doi.org/10.5281/zenodo.7723340. UMINT can be executed on any standard computing platform with at least 8 GB RAM on a Windows/Linux/CentOS platform with python 3.7+ installed in it.

Author contributions

Conceptualization of methodology and framework: CM, DS, VD, and RD. Data curation, data analysis, formal analysis, visualization, implementation: CM and DS. Investigation, code review, validation, original draft preparation: DS, CM, and VD. Reviewing, editing, overall supervision: VD and RD.

Acknowledgments

DS thanks the Data Science Laboratory, A. K. Choudhury School of Information Technology, University of Calcutta, Kolkata, India for providing the HPC infrastructure to execute and test the models. RD acknowledges the grant by DST-NSF provided to him through IDEAS-TIH, Indian Statistical Institute, Kolkata, India.

Conflict of interest

VD works as a Senior Research Scientist at Novo Nordisk A/S, Måløv. DS works as an Associate Data Scientist at Tatras Data Services Pvt. Ltd. None of them has received any funds for this work.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2023.1184748/full#supplementary-material

References

Argelaguet, R., Arnol, D., Bredikhin, D., Deloro, Y., Velten, B., Marioni, J. C., et al. (2020). Mofa+: A statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol.* 21, 111–117. doi:10.1186/s13059-020-02015-1

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* 2008, P10008. doi:10.1088/1742-5468/2008/10/p10008

Bottini, S., Emmert-Streib, F., and Franco, L. (2021). Editorial: AI and multi-omics for rare diseases: Challenges, advances and perspectives. *Front. Mol. Biosci.* 8, 719978. doi:10.3389/fmolb.2021.719978

Bredikhin, D., Kats, I., and Stegle, O. (2022). Muon: Multimodal omics analysis framework. *Genome Biol.* 23, 42–12. doi:10.1186/s13059-021-02577-8

Buenrostro, J. D., Wu, B., Chang, H. Y., and Greenleaf, W. J. (2015). Atac-seq: A method for assaying chromatin accessibility genome-wide. *Curr. Protoc. Mol. Biol.* 109, 21.29.1–21.29.9. doi:10.1002/0471142727.mb2129s109

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. doi:10.1038/s41587-022-01284-4

Cao, J., Cusanovich, D. A., Ramani, V., Aghamirzaie, D., Pliner, H. A., Hill, A. J., et al. (2018). Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* 361, 1380–1385. doi:10.1126/science.aau0730

Chaudhary, K., Poirion, O. B., Lu, L., and Garmire, L. X. (2018). Deep learning–based multi-omics integration robustly predicts survival in liver cancer. *Clin. Cancer Res.* 24, 1248–1259. doi:10.1158/1078-0432.CCR-17-0853

Chen, S., Lake, B. B., and Zhang, K. (2019). High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat. Biotechnol.* 37, 1452–1457. doi:10.1038/s41587-019-0290-0

Chen, H., Ryu, J., Vinyard, M. E., Lerer, A., and Pinello, L. (2021). Simba: Single-cell embedding along with features. *bioRxiv*. doi:10.1101/2021.10.17.464750

Clark, S. J., Argelaguet, R., Kapourani, C.-A., Stubbs, T. M., Lee, H. J., Alda-Catalinas, C., et al. (2018). scnmt-seq enables joint profiling of chromatin accessibility dna methylation and transcription in single cells. *Nat. Commun.* 9, 781–789. doi:10.1038/s41467-018-03149-4

Clyde, D. (2021). Share-seq reveals chromatin potential. *Nat. Rev. Genet.* 22, 2. doi:10.1038/s41576-020-00308-6

Davies, D. L., and Bouldin, D. W. (1979). "A cluster separation measure," in IEEE Transactions on Pattern Analysis and Machine Intelligence, 224–227. doi:10.1109/TPAMI.1979.4766909

Eltager, M., Abdelaal, T., Mahfouz, A., and Reinders, M. J. (2021). scmoc: Single-cell multi-omics clustering. *bioRxiv*. doi:10.1101/2021.02.24.432644

Gayoso, A., Steier, Z., Lopez, R., Regier, J., Nazor, K. L., Streets, A., et al. (2021). Joint probabilistic modeling of single-cell multi-omic data with totalvi. *Nat. Methods* 18, 272–282. doi:10.1038/s41592-020-01050-x

Griffiths, J. A., Scialdone, A., and Marioni, J. C. (2018). Using single-cell genomics to understand developmental processes and cell fate decisions. *Mol. Syst. Biol.* 14, e8046. doi:10.15252/msb.20178046

Haghverdi, L., Lun, A. T., Morgan, M. D., and Marioni, J. C. (2018). Batch effects in single-cell rna-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* 36, 421–427. doi:10.1038/nbt.4091

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., Zheng, S., Butler, A., et al. (2021). Integrated analysis of multimodal single-cell data. *Cell* 184, 3573–3587.e29. doi:10.1016/j.cell.2021.04.048

Hinton, G. E., and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science* 313, 504–507. doi:10.1126/science.1127647

Jiang, R., Sun, T., Song, D., Li, J. J., Zhong, X., Ye, Z., et al. (2022). Statistics or biology: The zero-inflation controversy about scrna-seq data. *Genome Biol.* 23, 1–8. doi:10.1080/10903127.2022.2126912

Kingma, D. P., and Ba, J. (2017). Adam: A method for stochastic optimization.

Kotliarov, Y., Sparks, R., Martins, A. J., Mulè, M. P., Lu, Y., Goswami, M., et al. (2020). Broad immune activation underlies shared set point signatures for vaccine responsiveness in healthy individuals and disease activity in patients with lupus. *Nat. Med.* 26, 618–629. doi:10.1038/s41591-020-0769-8

Kriebel, A. R., and Welch, J. D. (2022). Uinmf performs mosaic integration of single-cell multi-omic datasets using nonnegative matrix factorization. *Nat. Commun.* 13, 780–817. doi:10.1038/s41467-022-28431-4

Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D. J., Hicks, S. C., Robinson, M. D., et al. (2020). Eleven grand challenges in single-cell data science. *Genome Biol.* 21, 31–35. doi:10.1186/s13059-020-1926-6

Lance, C., Luecken, M. D., Burkhardt, D. B., Cannoodt, R., Rautenstrauch, P., Laddach, A. C., et al. (2022). Multimodal single cell data integration challenge: Results and lessons learned. *bioRxiv*. doi:10.1101/2022.04.11.487796

Lee, J., Liu, C., Kim, J., Chen, Z., Sun, Y., Rogers, J. R., et al. (2022). Deep learning for rare disease: A scoping review. *J. Biomed. Inf.* 135, 104227. doi:10.1016/j.jbi.2022.104227

Li, L., Dugan, H. L., Stamper, C. T., Lan, L. Y.-L., Asby, N. W., Knight, M., et al. (2021). Improved integration of single-cell transcriptome and surface protein expression by linq-view. *Cell Rep. Methods* 1, 100056. doi:10.1016/j.crmeth.2021.100056

Li, G., Fu, S., Wang, S., Zhu, C., Duan, B., Tang, C., et al. (2022). A deep generative model for multi-view profiling of single-cell rna-seq and atac-seq data. *Genome Biol.* 23, 20–23. doi:10.1186/s13059-021-02595-6

Lin, Y., Wu, T.-Y., Wan, S., Yang, J. Y., Wong, W. H., and Wang, Y. (2022). Scjoint integrates atlas-scale single-cell rna-seq and atac-seq data with transfer learning. *Nat. Biotechnol.* 40, 703–710. doi:10.1038/s41587-021-01161-6

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

Luecken, M. D., Büttner, M., Chaichoompu, K., Danese, A., Interlandi, M., Müller, M. F., et al. (2022). Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* 19, 41–50. doi:10.1038/s41592-021-01336-8

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* 161, 1202–1214. doi:10.1016/j.cell.2015.05.002

Oh, S. Y., Kim, W. S., Ryoo, B.-Y., Kang, H. J., Park, Y. H., Kim, K., et al. (2006). Intestinal marginal zone b-cell lymphoma of malt type: Clinical manifestation and outcome of a rare disease. *Blood* 108, 4742. doi:10.1182/blood.v108.11.4742.4742

Papalexi, E., and Satija, R. (2018). Single-cell rna sequencing to explore immune cell heterogeneity. *Nat. Rev. Immunol.* 18, 35–45. doi:10.1038/nri.2017.76

Peterson, V. M., Zhang, K. X., Kumar, N., Wong, J., Li, L., Wilson, D. C., et al. (2017). Multiplexed quantification of proteins and transcripts in single cells. *Nat. Biotechnol.* 35, 936–939. doi:10.1038/nbt.3973

Picelli, S., Björklund, Å. K., Faridani, O. R., Sagasser, S., Winberg, G., and Sandberg, R. (2013). Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* 10, 1096–1098. doi:10.1038/nmeth.2639

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20, 53–65. doi:10.1016/0377-0427(87)90125-7

Seal, D. B., Das, V., Goswami, S., and De, R. K. (2020). Estimating gene expression from dna methylation and copy number variation: A deep learning regression model for multi-omics integration. *Genomics* 112, 2833–2841. doi:10.1016/j.ygeno.2020.03.021

Stanojevic, S., Li, Y., and Garmire, L. X. (2022). Computational methods for single-cell multi-omics integration and alignment. arXiv preprint arXiv:2201.06725

Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P. K., Swerdlow, H., et al. (2017). Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* 14, 865–868. doi:10.1038/nmeth.4380

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., III, et al. (2019). Comprehensive integration of single-cell data. *Cell* 177, 1888–1902. doi:10.1016/j.cell.2019.05.031

Traag, V. A., Waltman, L., and Van Eck, N. J. (2019). From louvain to leiden: Guaranteeing well-connected communities. *Sci. Rep.* 9, 5233–5312. doi:10.1038/s41598-019-41695-z

Tran, H. T. N., Ang, K. S., Chevrier, M., Zhang, X., Lee, N. Y. S., Goh, M., et al. (2020). A benchmark of batch-effect correction methods for single-cell rna sequencing data. *Genome Biol.* 21, 12–32. doi:10.1186/s13059-019-1850-9

Waltman, L., and Van Eck, N. J. (2013). A smart local moving algorithm for large-scale modularity-based community detection. *Eur. Phys. J. B* 86, 471–514. doi:10.1140/epjb/e2013-40829-0

Wills, Q. F., and Mead, A. J. (2015). Application of single-cell genomics in cancer: Promise and challenges. *Hum. Mol. Genet.* 24, R74–R84. doi:10.1093/hmg/ddv235

Wolf, F. A., Angerer, P., and Theis, F. J. (2018). Scanpy: Large-scale single-cell gene expression data analysis. *Genome Biol.* 19, 15–5. doi:10.1186/s13059-017-1382-0