Title

Multi-task learning from multimodal single-cell omics with Matilda

Permalink
https://escholarship.org/uc/item/99k7v7sw

Journal
Nucleic Acids Research, 51(8)

ISSN
0305-1048

Authors
Liu, Chunlei
Huang, Hao
Yang, Pengyi

Publication Date
2023-05-08

DOI
10.1093/nar/gkad157

Peer reviewed

Published online 13 March 2023 Nucleic Acids Research, 2023, Vol. 51, No. 8 **e45**
https://doi.org/10.1093/nar/gkad157

Multi-task learning from multimodal single-cell omics with Matilda

Chunlei Liu[1], Hao Huang[1,2] and Pengyi Yang[1,2,3,*]

[1]Computational Systems Biology Group, Children's Medical Research Institute, Faculty of Medicine and Health, The University of Sydney, Westmead, NSW 2145, Australia, [2]School of Mathematics and Statistics, The University of Sydney, Sydney, NSW 2006, Australia and [3]Charles Perkins Centre, The University of Sydney, Sydney, NSW 2006, Australia

Received October 20, 2022; Revised January 28, 2023; Editorial Decision February 19, 2023; Accepted February 21, 2023

ABSTRACT

Multimodal single-cell omics technologies enable multiple molecular programs to be simultaneously profiled at a global scale in individual cells, creating opportunities to study biological systems at a resolution that was previously inaccessible. However, the analysis of multimodal single-cell omics data is challenging due to the lack of methods that can integrate across multiple data modalities generated from such technologies. Here, we present Matilda, a multi-task learning method for integrative analysis of multimodal single-cell omics data. By leveraging the interrelationship among tasks, Matilda learns to perform data simulation, dimension reduction, cell type classification, and feature selection in a single unified framework. We compare Matilda with other state-of-the-art methods on datasets generated from some of the most popular multimodal single-cell omics technologies. Our results demonstrate the utility of Matilda for addressing multiple key tasks on integrative multimodal single-cell omics data analysis. Matilda is implemented in Pytorch and is freely available from https://github.com/PYangLab/Matilda.

INTRODUCTION

Recent development of multimodal single-cell omics technologies enables multiple modalities of cellular regulatory circuitry to be simultaneously profiled in individual cells (1). Data generated from these technologies create new opportunities for integrative analysis of cellular programs that are inaccessible from analysing each data modality alone and hence promise to provide a more holistic characterization of cellular systems at single-cell resolution (2). A large number of computational methods have been developed for single-cell RNA-sequencing (scRNA-seq) data to perform tasks such as data simulation (3), dimension reduction (4) and classification of cell types (5,6), and feature selection (7,8). While methods designed for scRNA-seq data analysis can be applied to analyse RNA modality in multimodal single-cell omics data, most of them cannot take advantage of other available data modalities and therefore could not fully utilize the information embedded in such data. The lack of computational methods that can integrate across data modalities is a key issue in multimodal single-cell omics data analysis and greatly hinder biological discovery from such data (9,10).

Here we present Matilda, a neural network-based multi-task learning method for integrative analysis of multimodal single-cell omics data (Figure 1A). Although previously methods developed for scRNA-seq data analysis typically address different tasks (e.g. data simulation, cell type classification) independently, a key observation in Matilda is that many common tasks in multimodal single-cell omics data analysis are closely related to each other. The modularity nature of neural networks employed in Matilda makes it well-suited for integrating multiple data modalities and performing multiple tasks in a single unified framework. For example, the data simulated by the variational autoencoder (VAE) (11), a key component of Matilda, can be augmented to the original data to improve cell type classification. By leveraging such relationships, Matilda simultaneously performs data simulation, dimension reduction, cell type classification, and feature selection across data modalities (Figure 1A), therefore, achieving multiple key tasks in integrative analysis of multimodal single-cell omics data.

Matilda performs data simulation, cell type classification, and feature selection for single-cell multimodal omics data in a single multi-task learning framework. To evaluate the performance of Matilda on multiple tasks in multimodal single-cell omics data analysis, we applied Matilda to a collection of datasets generated from popular multimodal single-cell omics technologies including those profiling three modalities using TEA-seq (gene expression [RNA], cell surface proteins [ADT], and chromatin

[*]To whom correspondence should be addressed. Tel: +61 2 9351 3039; Fax: +61 2 9351 4534; Email: pengyi.yang@sydney.edu.au

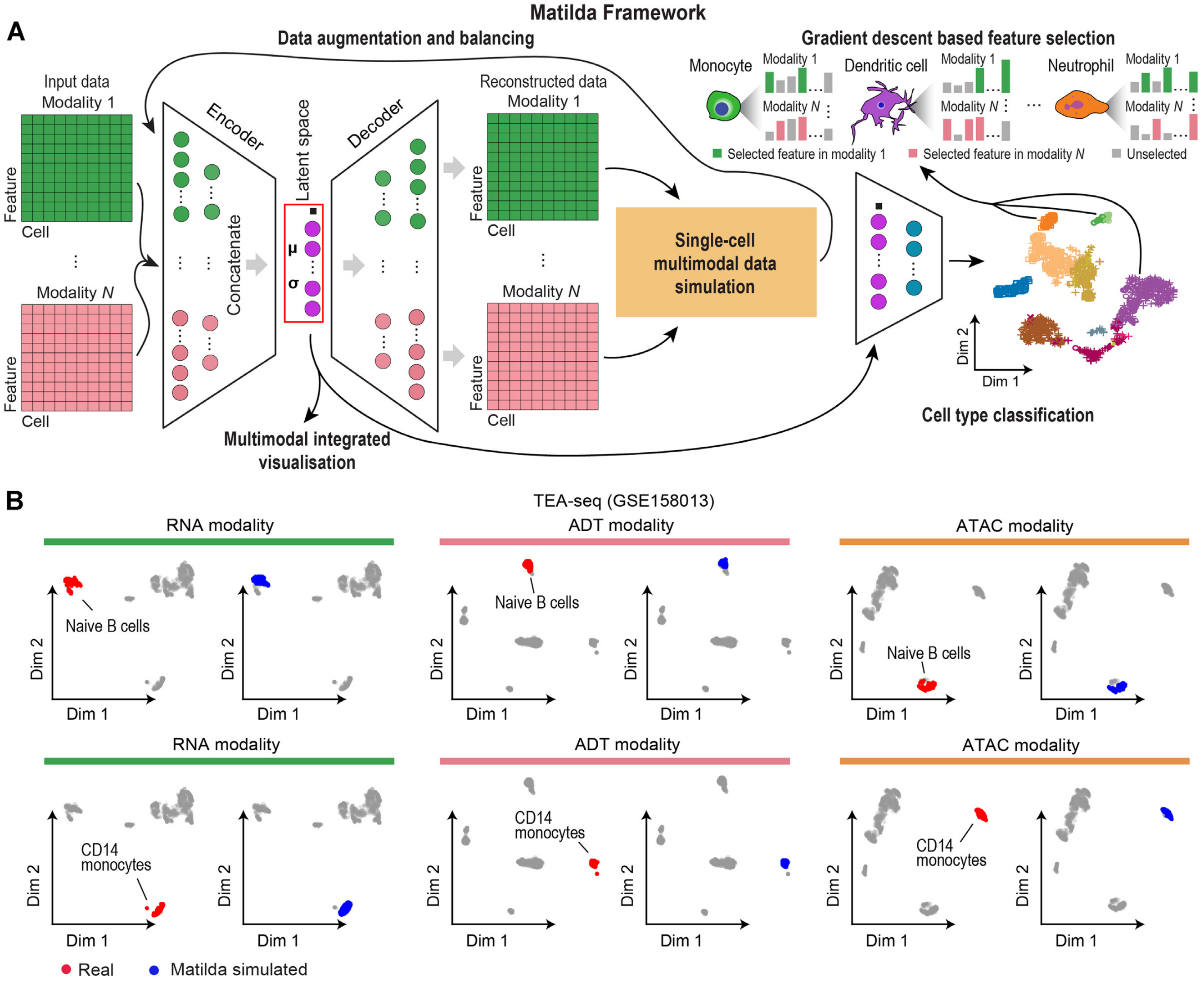

Figure 1. Matilda framework and multimodal single-cell data simulation. (**A**) Schematic summary of the main components in Matilda framework, including multimodal single-cell data simulation, data augmentation, multimodal integrated visualization, cell type classification, and gradient descent-based feature selection. (**B**) UMAP visualization of cell-type-specific simulations of RNA, ADT, and ATAC modalities in the TEA-seq dataset (GSE158013) using Matilda. The Upper and lower panels show real (red), and Matilda simulated (blue) naïve B cells and CD14 monocytes, respectively.

accessibility [ATAC]) (12), and those profiling two modalities using CITE-seq (RNA and ADT) (13–15) and SHARE-seq (RNA and ATAC) (16). While there are currently few methods specifically designed for data simulation, cell type classification, and feature selection using multiple modalities in these datasets, various methods (e.g. Sparsim (17) for data simulation, scClassify (5) for cell type classification, MAST (8) for feature selection) have been developed for single-cell RNA-sequencing (scRNA-seq) data and therefore can be applied using the RNA modality in these datasets. Using a range of evaluation criteria, we show that Matilda outperforms other state-of-the-art methodologies designed for various tasks using single or multiple data modalities. Our results demonstrate the utility of Matilda as the first comprehensive method for addressing multiple key tasks in multimodal single-cell omics data analysis.

MATERIALS AND METHODS

Datasets and preprocessing

TEA-seq dataset. TEA-seq enables simultaneous single-cell profiling of transcripts, epitopes, and chromatin accessibility (12). The processed matrices of TEA-seq data from measuring PBMC were downloaded from the NCBI Gene Expression Omnibus (GEO) under the accession number GSE158013, with raw RNA expression, ADT expression, and peak accessibility (ATAC) measured for the same cells in four data batches. We summarized the matrix of ATAC from peak level to gene activity scores using the 'CreateGeneActivityMatrix' function in Seurat package (14). Genes with fewer than 1% quantifications across cells in each of the three modalities were removed, respectively. This resulted in a dataset with 6310 (9855 RNA, 46 ADT, 17141 ATAC); 6545 (9852 RNA, 46 ADT, 17081 ATAC); 6534

(9911 RNA, 46 ADT, 16552 ATAC); and 6748 (9859 RNA, 46 ADT, 16620 ATAC) numbers of cells in each of the four data batches. The cell type information was obtained from the original study and for each of the four data batches the number of cell types are 11, 11, 10 and 10.

CITE-seq dataset by Stephenson et al. This CITE-seq dataset measures PBMC from healthy individuals and from COVID-19 patients (15). Only the data from healthy individuals were used in this study. The raw matrices of RNA and ADT and the annotation of cells to their respective cell types from the original study were downloaded from the EMBL-EBI ArrayExpress database under the accession number E-MTAB-10026, with 30313 healthy cells from Cambridge medical centre (batch 1) and 64262 healthy cells from NCL medical centre (batch 2). RNA and ADT in this dataset were filtered by removing those that expressed in less than 1% of the cells and cell types were filtered by removing those that have less than 10 cells. After filtering, there are 30313 cells from 17 cell types (10668 RNA, 192 ADT) in batch 1 and 64257 cells from 16 cell types (10618 RNA, 192 ADT) in batch 2 of the dataset for downstream analysis.

CITE-seq dataset by Hao et al. The raw RNA and ADT matrices from this CITE-seq dataset generated by Hao *et al.* (14) from PBMC 2 were downloaded from NCBI GEO under the accession number GSE164378. The dataset contains two batches and the cells in both batches were annotated to 31 cell types. As the above, RNA and ADT in this dataset were filtered by removing those that expressed in <1% of the cells and cell types were filtered by removing those that have less than 10 cells. This resulted in 67090 cells (11 451 RNA, 228 ADT) in batch 1 and 94674 cells (12 347 RNA, 228 ADT) in batch 2 of the dataset.

CITE-seq dataset by Ramaswamy et al. The raw RNA and ADT matrices of PBMC from three healthy donors in this CITE-seq dataset generated by Ramaswamy *et al.* (13) were downloaded from NCBI GEO under the accession number GSE166489. Each patient sample corresponds to one data batch. After filtering RNA and ADT expressed in less than 1% of the cells and discarding cell types that have fewer than 10 cells, we obtained 8641 cells and 26 cell types in batch 1 (11062 RNA, 189 ADT), 9523 cells and 26 cell types in batch 2 (10 801 RNA, 189 ADT), and 10 410 cells and 28 cell types (11 039 RNA, 189 ADT) in batch 3 of this dataset.

SHARE-seq dataset. The SHARE-seq data that measures RNA and ATAC from matched cells in mouse skin samples were downloaded from NCBI GEO under the accession number GSE140203 (16). The dataset contains raw count of RNA and ATAC of cells annotated to 22 cell types. Similar to the above, we first removed peaks with no expression across cells, and then summarized the ATAC data from peak level into gene activity scores using the 'CreateGeneActivityMatrix' function in Seurat. We filtered out RNA and ATAC quantified in fewer than 1% of the cells and cell types that have less than 10 cells, resulting in a dataset with 32231 cells (8926 RNA, 14 034 ATAC) for the subsequent analyses.

Matilda design

Multi-task learning architecture. The multi-task neural networks in Matilda consist of multimodality-specific encoders and decoders in a variational autoencoder (VAE) component for data simulation and a fully-connected classification network for cell type classification. The encoders in the VAE component are shareable for both data simulation and classification tasks, and consist of one learnable point-wise parameter layer and one fully-connected layer to the input layer. Because ADT modality has significantly fewer features than RNA and ATAC modalities, we set empirically, based on model selection, the numbers of neurons for encoders of RNA, ADT, and ATAC modalities to be 185, 30, and 185, respectively. To learn a latent space that integrates the information from across modalities, we concatenated the output from the encoder trained from each data modality to perform joint learning using a fully-connected layer with 100 neurons, followed by a VAE reparameterization process (11). Next, the fully-connected layer of the latent space is split into two branches with one branch fed into the decoders and the other branch fed into the fully-connected classification network. For the decoder branch, it consists of multiple decoders each corresponds to an input data modality. Each decoder consists of one fully-connected layer to the output layer that has the same number of neurons as the features in the corresponding data modality. For each fully-connected layer in the VAE component, batch normalization (18), shortcut (19) were utilized in the model. ReLU activation was used in all fully-connected layers except in the reparameterization process. Dropout ($r = 0.2$) was utilized only for fully-connected layers in encoders. For the classification branch, it consists of the latent space as input to a fully-connected layer with a dimension equal to the number of cell types in the training data. The fully-connected layer outputs a probability vector for cell type prediction through a SoftMax function.

Loss function. Let X be the single-cell multimodal omic data from N modalities, the VAE component of Matilda contains two procedures: (i) the encoders encode each modality in the data X individually, and concatenate them for joint learning. This process projected the high-dimensional X into a low-dimensional latent space. We denote the posterior distribution of this process as $q_\theta(z|X)$, where θ is the learnable parameter of the neural network in this procedure; (ii) the decoders reconstruct the low-dimensional latent space to the high-dimensional original data space. We denote the posterior distribution of this process as $p_\varphi(X|z)$, where φ is the learnable parameter of the neural network in this procedure. The loss function of the data simulation component can be represented as the negative log-likelihood with a regularizer:

$$\begin{aligned} L_{sim}\ (\theta, \varphi) = & -E_{z\sim q_\theta(z|X)}\left[log p_\varphi\ (X|z)\right] \\ & + KL(q_\theta\ (z|X)\ ||p\ (z)) \qquad (1) \end{aligned}$$

The first term is the reconstruction loss using the expectation of negative log-likelihood. This term encourages the decoder to learn to reconstruct the original data X using the low-dimensional representation z. The second term is the Kullback-Leibler (KL) divergence between the encoder's distribution $q_\theta(z|X)$ and $p(z)$, where $p(z)$ is specified as a standard Normal distribution as $p(z) \sim N(0, 1)$. This divergence measures the information loss when using $q_\theta(z|X)$ to represent $p(z)$. The encoder network parameters are in turn optimized using stochastic gradient descent via backpropagation, which is made possible by the reparameterization trick (11).

For the loss function of the classification component, we use cross-entropy loss with label smoothing (20). Label smoothing is a regularizer technique, which replaces one-hot real label vector y_{real} with a mixture of y_{real} and the uniform distribution:

$$y_{ls} = (1 - \alpha) \times y_{real} + \alpha/K \tag{2}$$

where K is the number of label classes, and α is a hyperparameter that determines the amount of smoothing. Then, the classification loss can be represented as:

$$L_{cla} = -\Sigma_{i=1}^{i=K} y_{ls}^{i} log y_{output}^{i} \tag{3}$$

where y_{output}^{i} is the predicted label for the i^{th} cell.

To learn Matilda, we combined the simulation loss and classification loss to give the following overall loss function:

$$L_{sum} = L_{sim} + \lambda \times L_{cla} \tag{4}$$

where λ is a weighting coefficient that determines the importance of the classification term against the data simulation term from Matilda.

Data augmentation and balancing strategy. During the model training process, Matilda performs data augmentation and balancing using simulated data from the VAE component. Specifically, Matilda first ranks the cell types in the training dataset by the number of cells in each type. The cell type corresponding to the median number is used as the reference and those that have smaller numbers of cells are augmented to have the same number of cells as the median using VAE simulated single-cell multimodal data for each cell type. Cell types that have larger numbers of cells than the median number are randomly down-sampled to match the median number of cells as well. This strategy helps Matilda to mitigate imbalanced cell type distribution in the data (21) and better learn the molecular features of under-represented and rare cell types.

Joint feature selection from multiple modalities. Leveraging its neural network architecture, Matilda implements two approaches, i.e. integrated gradient (IG) (22) descent and saliency (23) based procedures, to detect the most informative features simultaneously from each of all data modalities. Specifically, for the IG method, to assess the importance of each feature, the trained model was used for backpropagation of the partial derivatives from the output units of the classification network to the input units of the encoders, where each input unit represents an individual feature from a given modality in the input data X. The importance score of each input feature of each cell is determined by approximating the integral gradients of the model's output to its input:

$$S_j = \int_{\tau=0}^{1} X_j \times \frac{\partial F(\tau \times X)}{\partial X_j} d\tau \tag{5}$$

where F F represents the classification branch of the multi-task neural networks, and $\frac{\partial F(\tau \times X)}{\partial X_j}$ is the gradient of $F(X)$ along with the j^{th} feature. We aggregated these derivatives across cells within each cell type. These aggregated gradients indicate the importance of each feature from each data modality in predicting each cell type. The top-ranked features from each cell type can be selected based on their aggregated derivatives for subsequent analyses. For the saliency method, a cell-type-specific importance score of a feature j is computed using the derivative:

$$S_j = \left.\frac{\partial F(X)}{\partial X}\right|_{X_j} \tag{6}$$

The magnitude of the derivative S_j indicates the effect of feature j on the classification score.

Matilda model training. Matilda adopts a two-step training strategy. In the first step, i.e. before augmentation and balancing, we train a network from scratch. In the second step, i.e. after augmentation and balancing, we inherit the weights from the first step as the initial value and fine-tune the networks using augmented and balanced data. Several key hyper-parameters may impact the performance of Matilda. These include the number of layers in the neural networks, the number of neurons in each layer, the parameter λ that balances the VAE data reconstruction and cell type classification in the multi-tasking learning, and other parameters such as learning rate, number of epochs, batch size, and dropout rate. To optimize these hyper-parameters, we used the training datasets of CITE-seq, SHARE-seq, and TEA-seq to evaluate the model performance with different parameter combinations based on measurements including (a) the distance between the umap of simulated data and real data and (b) the classification accuracy before and after data augmentation. These allowed us to determine the following Matilda settings that were used in subsequent experiments. Specifically, for both steps in the training process, batch size was set to 64 cells in learning from all datasets. The epoch was set to 30 for all datasets except the CITE-seq dataset generated by Hao *et al.* (GSE164378) which contains the largest number of cells. Since large datasets do not need many training epochs for the neural networks to converge, we set this to 10 for this CITE-seq dataset (GSE164378) for improving training efficiency. The parameter λ for balancing loss function in multi-tasking learning was empirically set to 0.1 for all datasets and the parameter α in label smoothing was set to 0.1 according to (24). In the first stage, we empirically determined the learning rate of 0.02 in the training process. In the second stage, we fine-tuned the networks with an initial learning rate of 0.02 for the first half of epochs and 0.002 for the second half of epochs. In Matilda, all input data modalities were normalized by the 'NormalizeData' function in Seurat (14) and then scaled using a z-score transformation to a similar range.

Settings for other classification methods

CHETAH. Raw count matrices of RNA modality from each dataset were used as input for CHETAH (v1.8.0) (6) and the function 'CHETAHclassifier' was used to perform cell type classification, following the author's tutorial (https://github.com/jdekanter/CHETAH).

scmapCell. Raw count matrices of RNA modality from each dataset were first normalized using 'NormalizeData' function in Seurat and then used as input for scmap (v1.14.0) (25) as suggested (https://github.com/hemberg-lab/scmap). By default, the top 500 most informative genes were used and the function 'scmapCell2Cluster' annotates cells in the query dataset to their respective cell types based on the reference data.

scClassify. Raw count matrices of RNA modality from each dataset were first normalized using the 'NormalizeData' function in Seurat and then used as input for scClassify (v1.4.0) (5). The default parameters, e.g. tree = 'HOPACH', algorithm = 'WKNN', selectFeatures = 'limma', similarity = 'pearson', were used as suggested in the pipeline (https://github.com/SydneyBioX/scClassify).

singleCellNet. Raw count matrices of RNA modality from each dataset were first normalized using the 'NormalizeData' function in Seurat and then used as input for singleCellNet (v0.1.0) (26). 'scn_train' function with the default parameters of nTopGenes = 10, nRand = 70, nTrees = 1000, nTopGenePairs = 25, dLevel = 'newAnn', colName_samp = 'cell' was used for training the model. The trained models were subsequently used for predicting the cell types for cells in the query data using 'scn_predict' and 'assess_comm' with default parameters (https://github.com/pcahan1/singleCellNet).

CelliD. The raw count matrices of RNA modality from each dataset were used as input for CelliD (v1.0.0) (27). Following the author's pipeline (https://github.com/RausellLab/CelliD), we use the function 'RunMCA' to perform Multiple Correspondence Analysis (MCA) dimension reduction for both reference and query data. Then extract gene signatures in each cell type using the function 'GetGroupGeneSet' with default parameters dims = 1:50, n.features = 200, group.by = 'cell.type'. The cell-to-cell matching and label transferring across data were generated using the function 'RunCellHGT'.

scID. Raw count matrices of RNA modality from each dataset were first normalized using the 'NormalizeData' function in Seurat and then used as input for the R package scID (v2.2) (28). Following the author's tutorial (https://github.com/BatadaLab/scID), we used the function 'scid_multiclass' with default parameters for identifying cell types in the query datasets.

UMINT. UMINT package version (c084930) (29) was used in this study. Following the author's tutorial (https://github.com/deeplearner87/UMINT), raw count matrices of RNA and/or ATAC modalities from each dataset were first normalized using the 'NormalizeData' function in Seurat, followed by 'FindVatiableFeatures', 'ScaleData' and 'RunPCA'. Raw count matrices of ADT modality were normalized using the 'NormalizeData' function with the parameter normalization.method = 'CLR', margin = 2, followed by 'ScaleData' and 'RunPCA'. The multimodal embeddings from UMINT were obtained and used for cell type classification.

Settings for other simulation methods

SPARSim. Following the author's pipeline (https://gitlab.com/sysbiobig/sparsim), raw count matrices of RNA modality from each dataset were used as input for SPARSim (v0.9.5) (17). Data were first normalized using the 'scran_normalization' function in SPARSim package and data parameters were estimated by 'SPARSim_estimate_parameter_from_data' function. The function 'SPARSim_simulation' was then used for generating simulated data.

cscGAN. Following the author's pipeline (https://github.com/imsb-uke/scGAN), raw count matrices of RNA modality from each dataset were first normalized using the 'process_files' function in cscGAN (Github version 988ad95) (30). Default parameters and training iteration of 6000 was used for model training and the 'run_exp' function was used for generating simulated data from the trained model.

ACTIVA. Following the author's pipeline (https://github.com/SindiLab/ACTIVA), raw count matrices of RNA modality from each dataset were used as input for ACTIVA (31). Data was first pre-processed using the 'Scanpy_IO' function in ACTIVA package (v0.0.3). Then, the model was trained using the 'ACTIVA' function with the default parameters. The function 'generate_subpopulation' was then used for generating simulated data.

Settings for other dimension reduction methods

Seurat. Seurat package (v4.1.0) (14) was used for dimension reduction of all CITE-seq datasets. The raw count matrices of RNA and ADT were used as input, which were then normalized by the 'NormalizeData' function in Seurat. By default, the top 2000 most variable genes were selected from RNA modality by 'FindVariableFeatures' function and data are subsequently scaled by 'ScaleData' function. Data from ADT modality were processed similarly as those of RNA modality, except using parameters of normalization.method = 'CLR' and margin = 2 in the in 'NormalizeData' function, as suggested in the author's pipeline (https://satijalab.org/seurat/). PCA was performed using the 'runPCA' function and the function 'FindMultiModalNeighbors' integrates RNA and ADT modalities using the PCA results. The joint visualization of RNA and ADT were generated using 'wnn.umap' function.

totalVI. The totalVI procedure implemented in the scvi-tools package (v0.15.0) (4) was used for dimension reduction of all CITE-seq datasets. Following the author's tutorial (https://github.com/scverse/scvi-tools), the raw count matrices of RNA and ADT were first normalized using the 'normalize_total' and 'log1p' functions and then the top 4000 most variable genes were selected using the 'highly_variable_genes' function.

The data were subsequently used as input for model training using 'scvi.model.TOTALVI.setup_anndata', 'scvi.model.TOTALVI', and 'train' functions in scvi-tools. The latent space of RNA and ADT modalities was generated using the 'get_latent_representation' function.

Conos. Conos package (v1.4.5) (32) was used for dimension reduction of the SHARE-seq dataset. Following the author's pipeline (https://github.com/kharchenkolab/conos), the raw count matrices of RNA and ATAC were normalized by the 'basicP2proc' function in pagoda2 package (v1.0.8), where recommended parameters of n.odgenes = 3e3, nPcs = 30, min.cells.per.gene = -1, make.geneknn = FALSE, and n.cores = 1 were used. Next, the joint graph was built using buildGraph with k = 15, k.self = 5, k.self.weigh = 0.01, ncomps = 30, n.odgenes = 5e3, and space = 'PCA' in Conos. The joint visualization of RNA and ATAC were generated using 'largeVis' in function 'embedGraph' with default parameter alpha = 1/2.

MultiVI. The MultiVI procedure implemented in the scvi-tools (v0.15.0) (33) was used for dimension reduction of the SHARE-seq dataset. Following the author's pipeline (https://github.com/scverse/scvi-tools), the raw count matrices of RNA and gene activity score matrices from ATAC and the paired matrix of RNA and ATAC were used as input. These data were first concatenated using the 'organize_multiome_anndatas' function in scvi-tools and then used for model training using 'scvi.model.MULTIVI.setup_anndata', 'scvi.model.MULTIVI' and 'train' functions in scvi-tools. The latent space of RNA and ATAC modalities was generated using the 'get_latent_representation' function.

Multigrate. The Multigrate method (34) was used for the dimension reduction of all datasets. Following the author's pipeline (https://github.com/theislab/multigrate), the raw count matrices of RNA, ADT or ATAC were used as input. RNA and ATAC data were first normalized using the 'normalize_total' and 'log1p' functions, and then the top 4000 most variable genes were selected using the 'highly_variable_genes' function. For ADT data, we perform CLR transformation using 'clr' function. Next, we combine the multimodality data using 'organize_multiome_anndatas' function and train the Multigrate models using 'MultiVAE.setup_anndata', 'MultiVAE', and 'train' functions in the Multigrate package (v0.0.2). The latent space was generated using the 'get_latent_representation' function.

Settings for other feature selection methods

We performed feature selection from the RNA modality of each dataset using a collection of methods: (i) simple one-sided t-test and Wilcoxon rank sum test, (ii) popular methods based on differential expression analysis including Limma (v3.48.3) (7) and MAST (v.1.2.1) (8), (iii) methods based on maximizing classification performance including logistic regression (LR) and receiver operating curve (ROC) implemented in the 'FindMarkers' function in Seurat 2 and (iv) deep learning based feature selection methods, including PROPOSE and scCapsNet with the following settings:

PROPOSE. The PROPOSE procedure (35) was used for feature selection of RNA modality from all datasets. Following the author's pipeline (https://github.com/iancovert/propose), the raw count matrices were first binarized to {0,1} according to the sign of the values and then used for model training using the 'PROPOSE' function in propose package (Github version 41fd568) with the number of marker genes as 100 and other parameters as default.

scCapsNet. The scCapsNet (Github version b21ca07) procedure was used for feature selection of RNA modality from all datasets. The raw count matrices were normalized using the 'log2' function in the numpy package. Following the author's pipeline (https://github.com/wanglf19/scCaps), the models were trained using the default network and parameters.

Performance evaluation

Cell type classification evaluation. We evaluated the accuracy of a cell type classification model by calculating their average accuracy as the sum of the accuracy in all cell types divided by the number of cell types in a dataset. The average accuracy of all cell types accounts for the performance of a classification model in both the major and minor cell types. We used two pipelines, referred to as 'intra-dataset' and 'inter-dataset' classification, for cell type classification model evaluation. While intra-dataset classification splits training and test data from one batch of a dataset, inter-dataset classification splits training and test data from different batches in a dataset. For intra-dataset classification, we performed five-fold cross-validation repeated five times with different seeding on each batch of each dataset. For inter-dataset classification, we select the common features and cell types from different batches in the same dataset with different data batches and train on one batch and test on another batch.

Simulation evaluation. We used the correlation heatmaps to visualize the correlation structure of select features in each data modality of each dataset. Specifically, we first applied the functions 'modelGeneVar' and 'getTopHVGs' from the scran R package (v1.20.1) (36) to select the top 100 high variable genes (HVGs) based on their variability calculated from each data modality in each dataset (except the ADT modality of the TEA-seq dataset) and then calculated pairwise Pearson's correlation coefficients from these HVGs across all cells in each dataset. Since the ADT modality of the TEA-seq dataset only contains 46 ADTs, we used all of them in the correlation analysis and heatmap visualization. For comparison to other simulation methods in RNA modality, we used the same visualization methods as above for each simulation method and also quantified the performance of each simulation method by calculating the overall Pearson's correlation of real and simulated data and represented these as boxplots.

Dimension reduction evaluation. We used the performance of a simple k-means clustering algorithm to assess cell type clustering on dimension reduced dataset generated from each modality integration and dimension reduction method. Similar to cell type classification, we used intra-dataset and inter-dataset for assessing cell type clustering. In particular, we used the latent space of the test dataset obtained either from five-fold cross-validation or a data batch for cell type clustering and compared the concordance between the clustering output and the cell type labels from their original studies. The five-fold cross-validation procedure in the intra-dataset clustering was repeated five times with different seeding. We assessed the clustering concordance using four evaluation metrics, including Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), Fowlkes-Mallows index (FM), and Jaccard index (Jaccard). Briefly, let N be the number of cells in the dataset, $\boldsymbol{U} = \{U_1, U_2, \ldots, U_R\}$ be the cell type annotation from the original study, and $\boldsymbol{V} = \{V_1, V_2, \ldots, V_c\}$ be the partition generated by clustering, the pairs between $\boldsymbol{U}$ and $\boldsymbol{V}$ can be classified into one of the four types: (i) N_{11}: the number of pairs that are in the same partition in both $\boldsymbol{U}$ and $\boldsymbol{V}$; (ii) N_{00} : the number of pairs that are in different partitions in $\boldsymbol{U}$ and $\boldsymbol{V}$; (iii) N_{01}: the number of pairs that are in the same partition in $\boldsymbol{U}$ but in different partitions in $\boldsymbol{V}$; (iv) N_{10}: the number of pairs that are in different partitions in $\boldsymbol{U}$ but in the same partition in $\boldsymbol{V}$. Given the above notation, we defined the ARI, NMI, FM, Jaccard metrics as follows:

$$ARI\,(\boldsymbol{U}, \boldsymbol{V}) = \frac{2(N_{00}N_{11} - N_{01}N_{10})}{(N_{00} + N_{01})(N_{01} + N_{11}) + (N_{00} + N_{10})(N_{10} + N_{11})} \quad (7)$$

$$Jaccard\ (\boldsymbol{U}, \boldsymbol{V}) = \frac{N_{11}}{N_{11} + N_{10} + N_{01}} \quad (8)$$

$$FM\,(\boldsymbol{U}, \boldsymbol{V}) = \sqrt{\left(\frac{N_{11}}{N_{11} + N_{01}}\right)\left(\frac{N_{11}}{N_{11} + N_{10}}\right)} \quad (9)$$

$$NMI\ (\boldsymbol{U}, \boldsymbol{V}) = \frac{I\,(\boldsymbol{U}; \boldsymbol{V})}{H(\boldsymbol{U}) + H(\boldsymbol{V})} \quad (10)$$

where $I(\boldsymbol{U}; \boldsymbol{V})$ is the mutual information between $\boldsymbol{U}$ and $\boldsymbol{V}$, defined as

$$I\ (\boldsymbol{U}; \boldsymbol{V}) = \Sigma_{i=1}^{R}\ \Sigma_{j=1}^{C} \frac{|U_i \cap U_j|}{N} log_2 \frac{N\,|U_i \cap V_j|}{|U_i|\,|V_j|} \quad (11)$$

and $H(\cdot)$ is the entropy of partitions, in which $H(\boldsymbol{U})$ and $H(\boldsymbol{V})$ are calculated

$$H\ (\boldsymbol{V}) = -\Sigma_{j=1}^{C} \frac{|V_j|}{N} log \frac{|V_j|}{N} \quad (12)$$

$$H\ (\boldsymbol{U}) = -\Sigma_{i=1}^{R} \frac{|U_i|}{N} log \frac{|U_i|}{N} \quad (13)$$

Feature selection evaluation. We used the classification of each cell type to evaluate the performance of features selected for that cell type. Specifically, we used a 'one-vs-all' procedure in that we classified each cell type against all remaining cell types using the top 100 features selected for that cell type from different feature selection methods. Note that only Matilda selected features from all data modalities whereas the other feature selection methods were designed for analysing gene expression data and thus used only to select features from RNA modality of each dataset. The classification accuracy for each cell type was calculated using the 'intra-dataset' procedure in that feature selection was conducted on training datasets and their utility/effectiveness in cell type classification were verified on test datasets generated from five-fold cross-validation repeated five times.

Running time evaluation. We evaluated running time on a server with AMD(R) Ryzen processor CPU (16 cores and 64 Gb total memory) and one RTX3090 graphics processing unit. We used the CITE-seq datasets generated by Hao *et al.* (GSE164378) and Ramaswamy *et al.* (GSE166489) to benchmark the running time, given the large numbers of cells in these two datasets. To evaluate the impact of the number of cells from the training datasets, we kept the number of cells to 2k in the test dataset and varied the number of cells in the training datasets from 1k, 2k, 3k, 5k, 10k, 20k, to 30k. Similarly, to evaluate the impact of the number of cells from test dataset, we kept the number of cells in the training dataset to 2k and varied the those in the test datasets from 1k, 2k, 3k, 5k, 10k, 20k, to 30k as above. The elapsed run time was evaluated by the R function 'system.time()' and Python function 'time.time()' for methods implemented using R and Python, respectively.

RESULTS

Multimodal single-cell data simulation

We applied Matilda to five recent multimodal single-cell omics datasets including a TEA-seq dataset that profiles RNA, ADT and ATAC modalities in human PBMC samples; three CITE-seq datasets that profile RNA and ADT modalities in human PBMC samples; and a SHARE-seq dataset that profiles RNA and ATAC modalities in mouse skin samples (Supplementary Figure S1). To test if Matilda is able to simulate multimodal omics data in a cell-type-specific manner, we first visualized cells using each modality on UMAPs (Figure 1B and Supplementary Figure S2) and highlighted cells from representative cell types using real and Matilda simulated data. We found that Matilda not only precisely simulates each data modality in a cell-type-specific manner but also denoizes the outliers in the real data, (e.g. ADT modality of B cells and CD14 cells in CITE-seq data; Supplementary Figure S2a).

To further assess the performance of Matilda on data simulation, we compared the correlation structure of highly variable genes (HVGs) by each data modality using real data and those simulated by Matilda (Figure 2A–C and Supplementary Figure S3). We found that data simulated by Matilda closely resemble the correlation structure of real data across all modalities. While no other methods are currently available for simulating multimodal single-cell omics data besides Matilda, various methods have been developed for simulating from scRNA-seq data (3). We, therefore, compared the simulation results of Matilda on RNA modality with those generated from scGAN (30), a simulation method based on deep generative adversarial networks, ACTIVA (31), a deep learning method based on adversarial

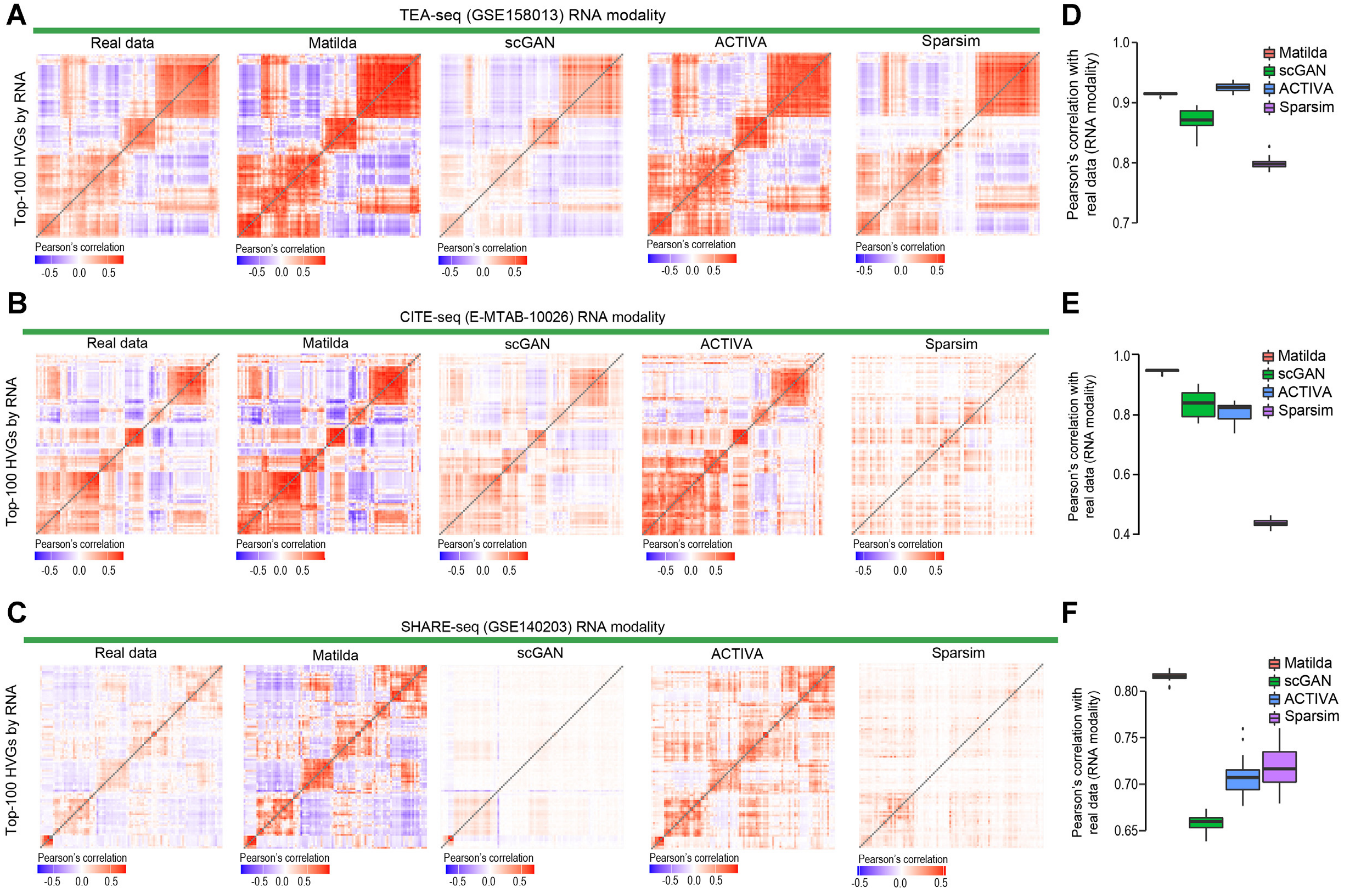

Figure 2. (**A**–**C**) Heatmap visualization of the correlation structure of RNA modality of real and simulated TEA-seq dataset (GSE158013), CITE-seq (E-MTAB-10026) and SHARE-seq dataset (GSE140203) using Matilda, scGAN, ACTIVA and Sparsim. The top-100 highly variable genes (HVGs) selected from the RNA modality of the real data were used for the heatmap. (**D**–**F**) Pearson's correlation of simulated data from each simulation method with real data using RNA modality for TEA-seq dataset (GSE158013), CITE-seq (E-MTAB-10026) and SHARE-seq dataset (GSE140203). Centre line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range; points, outliers.

VAE, and Sparsim (17), one of the best performing simulation methods based on mixture modelling (3). We found that in most cases data simulated from Matilda for the RNA modality better preserve the correlation structure in the real data compared to alternative methods as quantified in Figure 2D–F. These results demonstrate the ability of Matilda on simulating multiple data modalities in a cell-type-specific manner in multimodal single-cell omics datasets.

Multimodal data integration and dimension reduction

During model training, Matilda learns to combine and reduce the feature dimensions of multimodal single-cell omics data to a latent space using its VAE component in the framework (Figure 1A). The trained VAE of Matilda thus can be used for multimodal feature integration and dimension reduction of both the training and new data. Several alternative methods are available for such tasks. These include Seurat (14) and totalVI (4), which are designed for integrating RNA and ADT modalities in CITE-seq data; Conos (32) and multiVI (33), which are designed for integrating RNA and ATAC modalities such as these in SHARE-seq data; and Multigrate (34), which is not limited to specific paired assays and can be applied to both bi- and tri-modality data. Comparing to these methods, we found that the dimension reduced data from Matilda shows significantly better cell type separation under UMAP projection (Figure 3A, B).

To further quantify these visual observations, we clustered the dimension reduced data generated from each method using a simple *k*-means clustering algorithm and analysed the concordance of the clustering output with the cell type labels provided from their original studies using a panel of concordance metrics including ARI, NMI, FM, and Jaccard index (see Materials and Methods). We found that in most cases Matilda generated dimension reduced datasets led to higher clustering concordance with respect to the original cell type labels across all datasets irrespective of the metrics (Figure 3C–E and Supplementary Figure S4). These results demonstrate the superior performance of Matilda for integrating and reducing feature dimensions in multimodal single-cell omics data and its utility for subsequent applications such as data visualization and clustering of cell types.

Cell type classification using multiple data modalities

To evaluate Matilda on cell type classification using multimodal single-cell omics data, we performed both five-fold

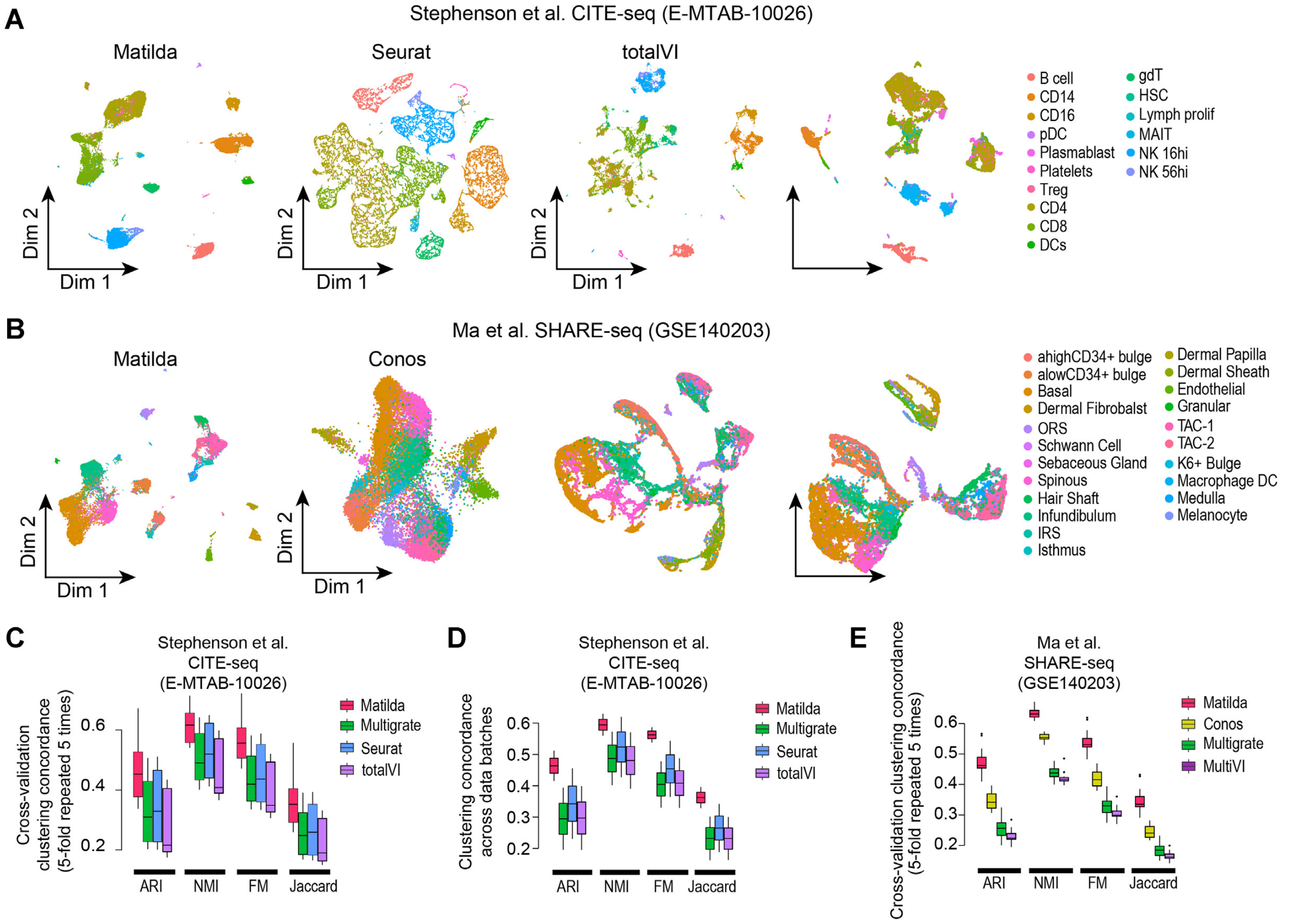

Figure 3. Assessment of Matilda for multimodal integrated dimension reduction and visualization. (A, B) Visualization and (C–E) quantification of joint multimodal dimension reduction results. (**A**) Visualizations of CITE-seq data (E-MTAB-10026) using Matilda, Seurat, totalVI, and Multigrate. (**B**) Visualizations of SHARE-seq data (GSE140203) using Matilda, Conos, MultiVI, and Multigrate. Cells are colour-coded by their types on the UMAPs. Quantifications were based on *k*-means clustering concordance using dimension reduced data from each method and the cell-type annotation from the original publication by ARI, NMI, FM, and Jaccard index. Either (**D**) 5-fold cross-validation repeated five times with different random seedings or (**D**) data from different batches from CITE-seq data (E-MTAB-10026) were used for capturing the variability in quantifications. (**E**) Quantification of *k*-means clustering concordance using dimension reduced data from Matilda, Conos, Multigrate, and MultiVI with cell-type annotations from the original study by ARI, NMI, FM, and Jaccard index on SHARE-seq data (GSE140203). Centre line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range; points, outliers.

cross-validation (repeated 5 times) and training and test using different batches within each dataset (Supplementary Figure S1b). While several methods have been developed recently for transferring cell type labels across different data modalities for multimodal single-cell omics data (37–39), there are currently few methods specifically designed for cell type classification by using all data modalities from such data. To this end, we resorted to comparing methods that are developed for cell type classification from scRNA-seq data by using RNA modality only (40) and UMINT (31), a method designed for integrating multiple data modalities to low-dimensional embeddings which can be used for cell type classification. We found that Matilda classifies cells significantly more accurately across all datasets under both the cross-validation settings (Figure 4A) and those from training and test using different batches within each dataset (Figure 4B) than other state-of-the-art cell type classification methods that use only RNA modality or those from using integrated embeddings generated by UMINT. The breakdown of the classification results from training and test using each pair of data batches reveals that Matilda led to higher cell type classification accuracy across all pairs in all four datasets that contain multiple data batches (Figure 4C).

To test if the performance of Matilda is impacted by the reduced size of the training data, we performed a stratified sampling of each cell type from CITE-seq and TEA-seq datasets generated by Ramaswamy *et al.* (13) and Swanson *et al.* (12), respectively, 80%, 50% and 20% of cells and trained each classification model using these subsampled datasets. We found that the performance of Matilda is largely maintained even when the model was trained on a small proportion of cells from the original datasets (Supplementary Figure S5). It is worth noting that the improved cell type classification accuracy of Matilda is not a sacrifice in speed on model training or classification of test

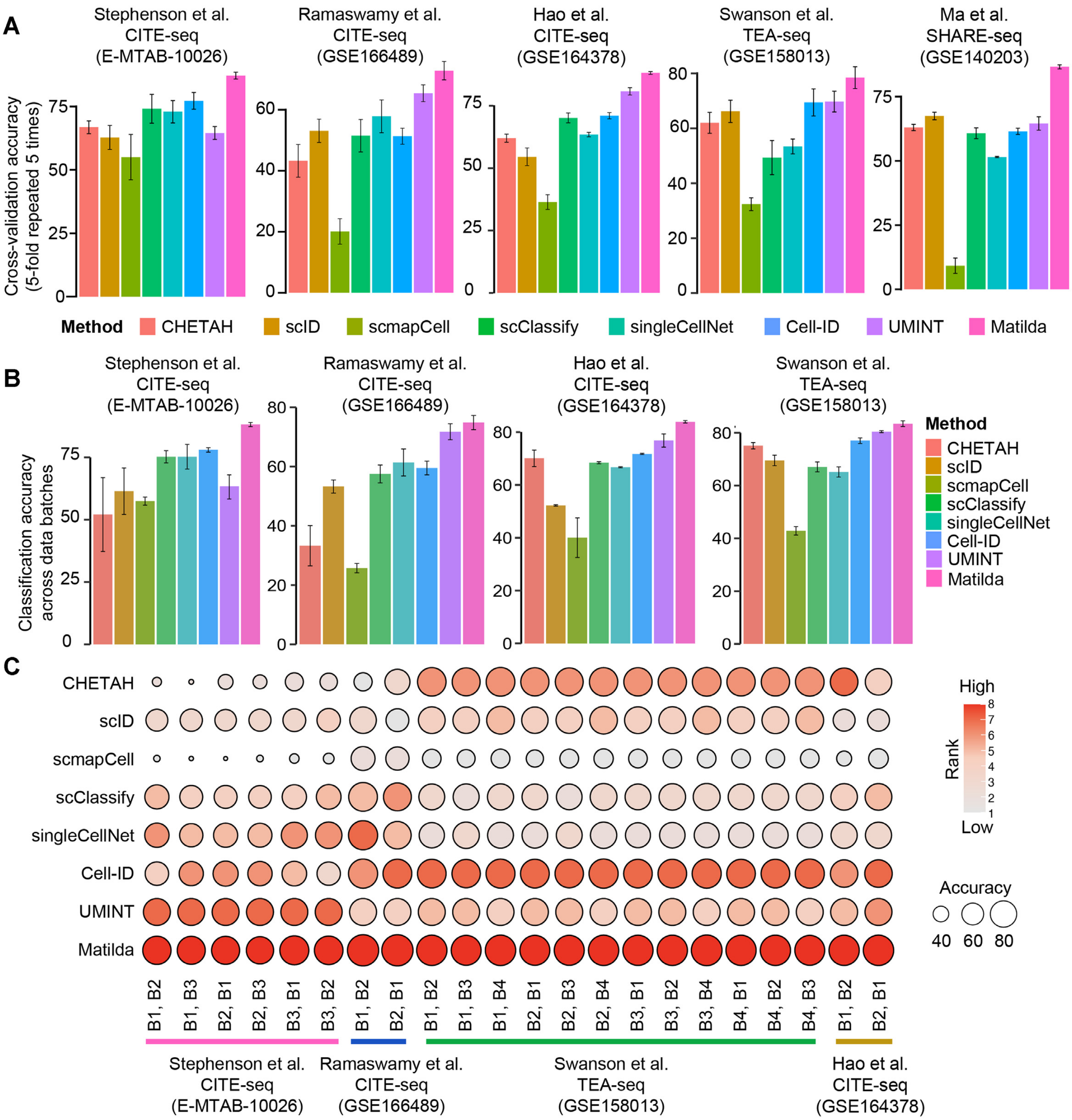

Figure 4. (A, B) Cell type classification of each multimodal single-cell omics data. Either (**A**) 5-fold cross-validation repeated five times with different random seedings or (**B**) data from different batches were used for benchmarking the performance of each method. Error bar, SD. (**C**) Ranking summary of cell type classification accuracy across data batches for each method.

data (Supplementary Figure S6). Since Matilda uses multi-task learning and the simulated data from the VAE component for data augmentation, we also evaluated the impact of these procedures on cell type classification accuracy. We found that, across all five datasets, multi-task learning indeed improved cell type classification than learning each task independently (Supplementary Figure S7a), and data augmentation resulted in better performance than those without (Supplementary Figure S7b). Together, these results demonstrate the utility of multi-task learning and data augmentation from simulation for improving cell type classification and highlight Matilda's increased cell type classification accuracy using multimodalities compared to alternative methods that use only RNA modality.

Feature selection from multiple data modalities

Finally, the neural network trained for cell type classification in Matilda can be used for multimodal feature selection using methods such as integrated gradient (IG) descent (22)

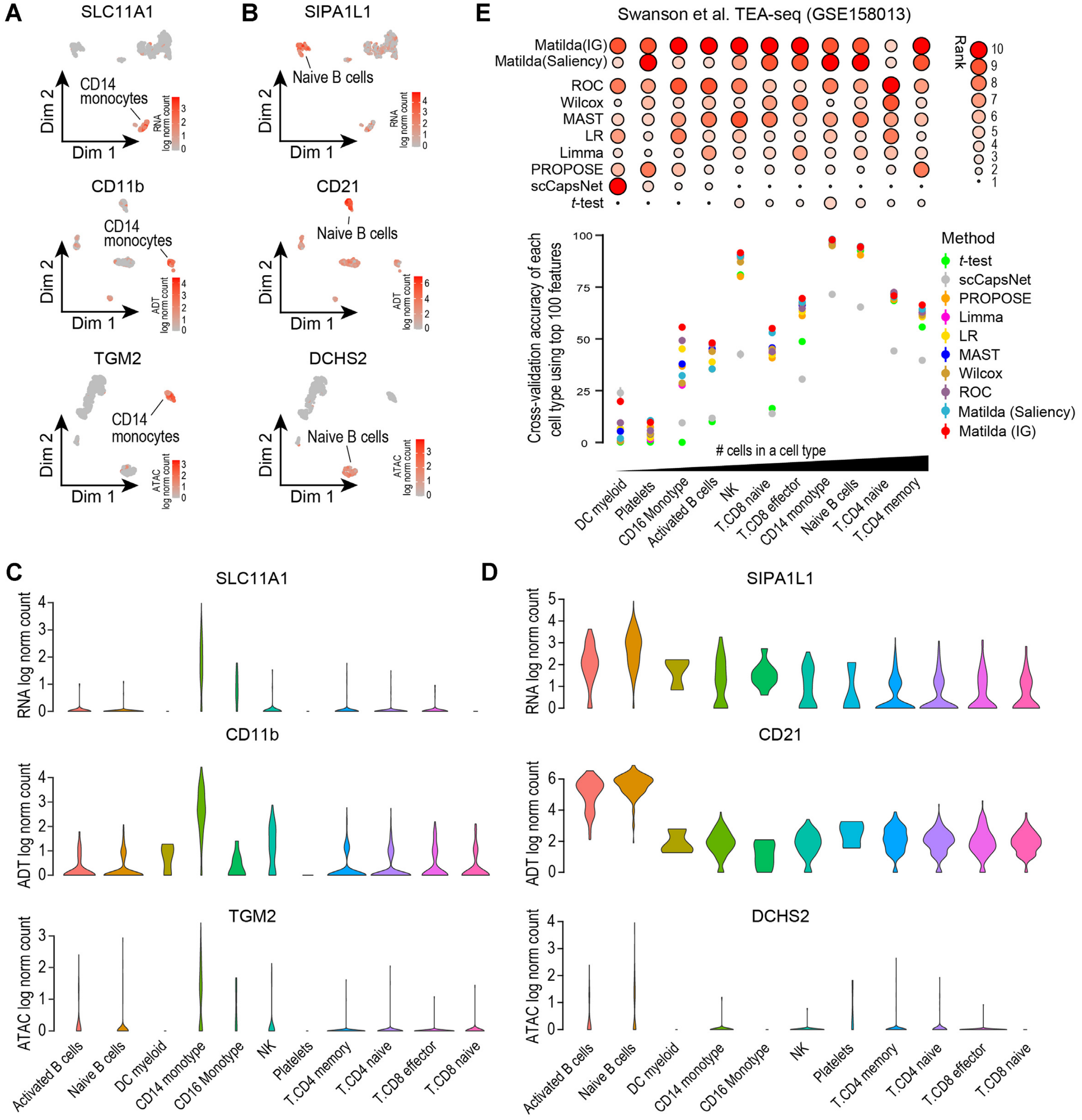

Figure 5. (**A, B**) For TEA-seq data (GSE158013), UMAPs highlight representative markers selected from each of the three modalities for CD14 monocytes and Naïve B cells. (**C, D**) violin plots of levels of selected markers for CD14 monocytes and Naïve B cells in their respective modalities across all cell types. (**E**) Classification of each cell type in TEA-seq data (GSE158013) using features selected by different methods. Cell types are arranged from low to high based on the number of cells in each cell type. Feature selection methods are also ranked based on the performance of their selected features in classifying each cell type (upper panel). Error bar, SE.

and saliency procedures (23), and thus can lead to the selection of cell-type-specific features across all available modalities in the datasets. Figure 5A, B visualize top-ranked features selected by Matilda using IG for CD14 monocytes and Naïve B cells, respectively, in each data modality in the TEA-seq dataset. The RNA and ADT expression levels and the ATAC activity of selected genes across all cell types in the dataset are shown in Figure 5C, D. As expected, these analyses reveal that features selected by Matilda for each data modality show expression specificity towards their respective cell types, demonstrating their potential usage for characterizing cell identity and their underlying molecular programs.

To evaluate the top features selected by Matilda across multiple data modalities and those selected from RNA modality using popular methods such as *t*-test and limma

(7), and those specifically designed for scRNA-seq (e.g. MAST (8), ROC), and recently proposed deep learning feature selection methods, including PROPOSE (35) and scCapsNet (41), we compared their utility in classifying each cell type in each dataset. We found that cell-type-specific features selected by Matilda from multiple data modalities on average resulted in more accurate discrimination of their respective cell types as shown by the scatter plot and the overall rankings of methods in each dataset (Figure 5E and Supplementary Figure S8). Within the two feature selection methods implemented in Matilda, IG appears to perform slightly better than saliency and is hence the recommended approach in Matilda for feature selection from multimodal single-cell omics data. Together, these results demonstrate Matilda as a useful approach for feature selection from multiple data modalities for cell type characterization and other downstream analyses.

DISCUSSION

The key motivation for using multi-task learning in Matilda is that many common tasks in single-cell multimodal omics data analysis are interrelated. Learning these tasks in parallel may therefore improve the performance of the model on each individual task. Furthermore, the rationale for using neural network models in Matilda is due to their modularity which fits well with the multiple data modalities and tasks. This allows the integration of data modalities and information sharing of tasks which together enable complementary information to be extracted and hence lead to more accurate characterizations of cellular programs. With the advance in single-cell multimodal omics technologies, we expect more data modalities to become available in the near future. The modularity and flexibility of Matilda allow integration when additional modality becomes available in such data.

One common criticism of neural network-based learning models is that a large number of examples need to be provided during the training process. We demonstrated in our experiments that Matilda's performance in cell type classification is largely maintained even with a relatively small number of cells in the training datasets. This may be due to the data simulation and augmentation component implemented in Matilda which increases the number of cells in the training datasets, especially for the rare cell types. However, dealing with cell types with an extremely small number of cells is still a challenge and may require alternative approaches.

While the current implementation of Matilda deals with datasets profiling discrete cell types, studies that look at transitional processes such as development and organogenesis create datasets with transient cell types. To analyse such datasets will require reformatting the loss function in the Matilda framework such as changing the classification component to a regression component. The potential mismatch of cell types in the training and query datasets may also have a significant impact on the performance of Matilda. A solution may be to utilize the prediction probability of the neural network for deciding whether a cell in a query dataset should be classified or not. These form the key directions for our future work.

Various methods have been developed for label transfer across modalities using different single-cell omics data (e.g. scRNA-seq, scATA-seq) (37–39,42,43). Such label transfer methods are distinguished from methods such as Matilda and UMINT that integrate multiple data modalities in the same cells (referred to as 'vertical integration') (44) since the embeddings of cells generated from label transfer methods are from individual data modalities. While the embeddings generated from label transfer methods can provide useful alignment of data modalities, they could not be directly used for multimodality cell type classification as performed by Matilda and UMINT.

In sum, Matilda is so far the first method for simultaneous simulation and supervised classification of cells using multiple modalities in single-cell multimodal omics data. It is also the first method for joint feature selection from multiple data modalities. Matilda addresses multiple key tasks in single-cell multimodal omics data analysis in a single unified framework.

DATA AVAILABILITY

All the datasets used in this study are publicly available. The 'TEA-seq dataset' was downloaded from NCBI GEO under the accession number GSE158013. The 'CITE-seq dataset by Stephenson et al' was downloaded from the EMBL-EBI Array Express database under the accession number E-MTAB-10026. The 'CITE-seq dataset by Hao et al' was downloaded from NCBI GEO under the accession number GSE164378. The 'CITE-seq dataset by Ramaswamy et al' was downloaded from NCBI GEO under the accession number GSE166489. The 'SHARE-seq' was downloaded from NCBI GEO under the accession number GSE140203. Matilda was implemented using PyTorch (version 1.9.1) with code available at https://github.com/PYangLab/Matilda.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We think the intellectual feedback from the Computational Systems Biology Group at Children's Medical Research Institute (CMRI) and engagement from the colleagues at the School of Mathematics and Statistics, The University of Sydney and Sydney Precision Bioinformatics Alliance.

Author contributions: Conceptualization, C.L., P.Y.; Methodology, C.L., P.Y.; Investigation, C.L., H.H., P.Y.; Data generation, C.L., H.H.; Writing – Original Draft, C.L., P.Y.; Writing – Review & Editing, all authors; Supervision, P.Y.; Funding Acquisition, P.Y.

FUNDING

National Health and Medical Research Council (NHMRC) Investigator Grant [1173469 to P.Y.]. Funding for open access charge: Sydney Research Accelerator (SOAR) prize.

Conflict of interest statement. None declared.

REFERENCES

1. Stuart,T. and Satija,R. (2019) Integrative single-cell analysis. *Nat. Rev. Genet.*, **20**, 257–272.
2. Zhu,C., Preissl,S. and Ren,B. (2020) Single-cell multimodal omics: the power of many. *Nat. Methods*, **17**, 11–14.
3. Cao,Y., Yang,P. and Yang,J.Y.H. (2021) A benchmark study of simulation methods for single-cell RNA sequencing data. *Nat. Commun.*, **12**, 6911.
4. 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.
5. Lin,Y., Cao,Y., Kim,H.J., Salim,A., Speed,T.P., Lin,D.M., Yang,P. and Yang,J.Y.H. (2020) scClassify: sample size estimation and multiscale classification of cells using single and multiple reference. *Mol. Syst. Biol.*, **16**, e9389.
6. de Kanter,J.K., Lijnzaad,P., Candelli,T., Margaritis,T. and Holstege,F.C.P. (2019) CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing. *Nucleic Acids Res.*, **47**, e95.
7. Ritchie,M.E., Phipson,B., Wu,D., Hu,Y., Law,C.W., Shi,W. and Smyth,G.K. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **43**, e47.
8. Finak,G., McDavid,A., Yajima,M., Deng,J., Gersuk,V., Shalek,A.K., Slichter,C.K., Miller,H.W., McElrath,M.J., Prlic,M. *et al.* (2015) MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.*, **16**, 278.
9. Ma,A., McDermaid,A., Xu,J., Chang,Y. and Ma,Q. (2020) Integrative methods and practical challenges for single-cell multi-omics. *Trends Biotechnol.*, **38**, 1007–1022.
10. Efremova,M. and Teichmann,S.A. (2020) Computational methods for single-cell omics across modalities. *Nat. Methods*, **17**, 14–17.
11. Kingma,D.P. and Welling,M. (2014) Auto-encoding variational bayes. arXiv doi: https://arxiv.org/abs/1312.6114, 10 December 2022, preprint: not peer reviewed.
12. Swanson,E., Lord,C., Reading,J., Heubeck,A.T., Genge,P.C., Thomson,Z., Weiss,M.D., Li,X., Savage,A.K., Green,R.R. *et al.* (2021) Simultaneous trimodal single-cell measurement of transcripts, epitopes, and chromatin accessibility using TEA-seq. *Elife*, **10**, e63632.
13. Ramaswamy,A., Brodsky,N.N., Sumida,T.S., Comi,M., Asashima,H., Hoehn,K.B., Li,N., Liu,Y., Shah,A., Ravindra,N.G. *et al.* (2021) Immune dysregulation and autoreactivity correlate with disease severity in SARS-CoV-2-associated multisystem inflammatory syndrome in children. *Immunity*, **54**, 1083–1095.
14. Hao,Y., Hao,S., Andersen-Nissen,E., Mauck,W.M., 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.
15. 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.
16. 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.
17. Baruzzo,G., Patuzzi,I. and Di Camillo,B. (2020) SPARSim single cell: a count data simulator for scRNA-seq data. *Bioinforma. Oxf. Engl.*, **36**, 1468–1475.
18. Ioffe,S. and Szegedy,C. (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. arXiv doi: https://arxiv.org/abs/1502.03167, 02 March 2015, preprint: not peer reviewed.
19. He,K., Zhang,X., Ren,S. and Sun,J. (2016) Deep residual learning for image recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Las Vegas, NV, USA, pp. 770–778.
20. Müller,R., Kornblith,S. and Hinton,G. (2020) When does label smoothing help? arXiv doi: https://arxiv.org/abs/1906.02629, 10 June 2020, preprint: not peer reviewed.
21. He,H. and Garcia,E.A. (2009) Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.*, **21**, 1263–1284.
22. Sundararajan,M., Taly,A. and Yan,Q. (2017) Axiomatic attribution for deep networks. arXiv doi: https://arxiv.org/abs/1703.01365, 13 June 2017, preprint: not peer reviewed.
23. Simonyan,K., Vedaldi,A. and Zisserman,A. (2014) Deep inside convolutional networks: visualising image classification models and saliency maps. arXiv doi: https://arxiv.org/abs/1312.6034, 19 April 2014, preprint: not peer reviewed.
24. Liu,Z., Luo,W., Wu,B., Yang,X., Liu,W. and Cheng,K.-T. (2020) Bi-real net: binarizing deep network towards real-network performance. *Int. J. Comput. Vis.*, **128**, 202–219.
25. Kiselev,V.Y., Yiu,A. and Hemberg,M. (2018) scmap: projection of single-cell RNA-seq data across data sets. *Nat. Methods*, **15**, 359–362.
26. Tan,Y. and Cahan,P. (2019) SingleCellNet: a computational tool to classify single cell RNA-seq data across platforms and across species. *Cell Syst.*, **9**, 207–213.
27. Cortal,A., Martignetti,L., Six,E. and Rausell,A. (2021) Gene signature extraction and cell identity recognition at the single-cell level with cell-ID. *Nat. Biotechnol.*, **39**, 1095–1102.
28. Boufea,K., Seth,S. and Batada,N.N. (2020) scID uses discriminant analysis to identify transcriptionally equivalent cell types across single-cell RNA-seq data with batch effect. *Iscience*, **23**, 100914.
29. Maitra,C., Seal,D.B., Das,V. and De,R.K. (2022) UMINT: unsupervised neural network for single cell multi-omics integration. bioRxiv doi: https://doi.org/10.1101/2022.04.21.489041, 22 April 2022, preprint: not peer reviewed.
30. Marouf,M., Machart,P., Bansal,V., Kilian,C., Magruder,D.S., Krebs,C.F. and Bonn,S. (2020) Realistic in silico generation and augmentation of single-cell RNA-seq data using generative adversarial networks. *Nat. Commun.*, **11**, 166.
31. Heydari,A.A., Davalos,O.A., Zhao,L., Hoyer,K.K. and Sindi,S.S. (2022) *ACTIVA*: realistic single-cell RNA-seq generation with automatic cell-type identification using introspective variational autoencoders. *Bioinformatics*, **38**, 2194–2201.
32. Barkas,N., Petukhov,V., Nikolaeva,D., Lozinsky,Y., Demharter,S., Khodosevich,K. and Kharchenko,P.V. (2019) Joint analysis of heterogeneous single-cell RNA-seq dataset collections. *Nat. Methods*, **16**, 695–698.
33. Ashuach,T., Gabitto,M.I., Jordan,M.I. and Yosef,N. (2021) MultiVI: deep generative model for the integration of multi-modal data. bioRxiv doi: https://doi.org/10.1101/2021.08.20.457057, 20 August 2021, preprint: not peer reviewed.
34. Lotfollahi,M., Litinetskaya,A. and Theis,F.J. (2022) Multigrate: single-cell multi-omic data integration. bioRxiv doi: https://doi.org/10.1101/2022.03.16.484643, 17 March 2022, preprint: not peer reviewed.
35. Covert,I., Gala,R., Wang,T., Svoboda,K., Sümbül,U. and Lee,S.-I. (2022) Predictive and robust gene selection for spatial transcriptomics. bioRxiv doi: https://doi.org/10.1101/2022.05.13.491738, 26 December 2022, preprint: not peer reviewed.
36. Lun,A.T.L., McCarthy,D.J. and Marioni,J.C. (2016) A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Research*, **5**, 2122.
37. Stuart,T., Butler,A., Hoffman,P., Hafemeister,C., Papalexi,E., Mauck,W.M., Hao,Y., Stoeckius,M., Smibert,P. and Satija,R. (2019) Comprehensive integration of single-cell data. *Cell*, **177**, 1888–1902.
38. Lin,Y., Wu,T.-Y., Wan,S., Yang,J.Y.H., Wong,W.H. and Wang,Y.X.R. (2022) scJoint integrates atlas-scale single-cell RNA-seq and ATAC-seq data with transfer learning. *Nat. Biotechnol.*, **40**, 703–710.
39. 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.
40. Abdelaal,T., Michielsen,L., Cats,D., Hoogduin,D., Mei,H., Reinders,M.J.T. and Mahfouz,A. (2019) A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biol.*, **20**, 194.
41. Wang,L., Nie,R., Yu,Z., Xin,R., Zheng,C., Zhang,Z., Zhang,J. and Cai,J. (2020) An interpretable deep-learning architecture of capsule networks for identifying cell-type gene expression programs from single-cell RNA-sequencing data. *Nat. Mach. Intell.*, **2**, 693–703.
42. Peng,T., Chen,G.M. and Tan,K. (2021) GLUER: integrative analysis of single-cell omics and imaging data by deep neural network. bioRxiv doi: https://doi.org/10.1101/2021.01.25.427845, 26 January 2021, preprint: not peer reviewed.

43. Demetci,P., Santorella,R., Sandstede,B. and Singh,R. (2021) Unsupervised integration of single-cell multi-omics datasets with disparities in cell-type representation. bioRxiv doi: https://doi.org/10.1101/2021.11.09.467903, 11 November 2021, preprint: not peer reviewed.

44. 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.