Supplementary meterials for Semi-supervised contrastive learning variational autoencoder Integrating single-cell multimodal mosaic datasets

Zihao Wang, Zeyu Wu and Minghua Deng

This supplementary file includes:

1 Dataset

In order to validate our scGCM, we used multiple batches of data obtained from various sequencing technologies. We are currently considering the integration of three modalities: transcriptomics (RNA), chromatin accessibility (ATAC), and surface proteomics (ADT). Detailed information is as follows:

DOGMA-seq Dataset(**?**): This dataset is a human peripheral blood mononuclear cell (PBMC) dataset that simultaneously obtained RNA, ATAC, and ADT data through DOGMA-seq, comprising 4 batches. It is from the Gene Expression Omnibus (GEO), with ID GSE166188.

TEA-seq Dataset(**?**): This dataset is a human peripheral blood mononuclear cell (PBMC) dataset that simultaneously obtained RNA, ATAC, and ADT data through TAE-seq, comprising 5 batches. It is from GEO, with ID GSE158013.

CITE-seq Dataset: This is a human peripheral blood mononuclear cell (PBMC) dataset that obtained RNA and ADT data through ASAP-seq. Here, we used data from two different experiments: one group from the ASAP-CITE sequencing experiment, with two batches from GEO, ID 156473, and another group from a separate CITE-seq experiment, which contains 8 batches and has accurately annotated labels. The data source is https://atlas.fredhutch.org/nygc/multimodal-pbmc.

10X Dataset: This is a human peripheral blood mononuclear cell (PBMC) dataset from the 10X Genomics platform, where RNA and ATAC modalities were measured simultaneously. Here, we considered several batches of data, with data scales of 10,000 and 3,000 cells, respectively. The data source is the official 10X Genomics website: https://www.10xgenomics.com/resources/datasets.

SHARE-seq Dataset: This dataset includes sequencing data from two different tissues, one from the mouse cerebral cortex and one from mouse epidermal cells, where RNA and ATAC modalities were sequenced simultaneously. They are derived from the datasets $Chen\_2019$(**?**) and Ma_2020(**?**).

Xie_2023 Dataset: This dataset is from studies of the mouse frontal cortex and serves as an additional benchmark comparison dataset. In this experiment, the dataset simultaneously measured trimethyl histone (H3K27me3) and gene expression. The data source is Xie_2023(**?**).

2 Preprocessing

Suppose we have cell $n \in \mathcal{N} = \{1, 2, \ldots, N\}$, where N represents the total number of cells. We denote the data matrix for the m modality and cell n as $x_n^m \in N^{D^m}$, where D^m is the number of features in modality m. Here, we generally denote the collection of all sequencing modality data for cell n as $x_n = \{x_n^m\}_{m \in \mathcal{M}_n}$, where $\mathcal{M}_n \in \{RNA, ATAC, ADT\}$ represents the corresponding sequencing modalities. Due to the specific characteristics of each modality, we apply different preprocessing methods to them.

We first process our RNA, ATAC, and ADT data separately using Seurat. For RNA and ADT data, we filter out cells with excessively low or high total expression values, as well as those with a low percentage of mitochondrial genome, for each batch. For ATAC data, we first calculate the total fragment count, transcription start site (TSS) score, and nucleosome signal to filter out low-quality ATAC data in each batch. We then consider the intersection of these batches to obtain our high-quality data. Next, we use Python to process the subsequent data.

For RNA data, we normalize the cells in each batch and then apply log transformation. Finally, we use scanpy(**?**) to identify the 4,000 most variable genes for further downstream analysis. For ATAC data, we apply TF-IDF for normalization. TF-IDF considers the impact of the frequency of feature occurrences, where features with lower frequencies are considered more specific and important. First, we calculate the frequency of each peak in the count matrix, denoted as $IDF_j = log\frac{n}{1+\left|i:x_{n_{i,j}}^{atac} \neq 0\right|}$, where $x_{n_{i,j}}^{atac}$ represents the value of the j peak in the ATAC modality for the i cell. Then, we multiply each frequency value into each peak to obtain the processed count matrix. Finally, we use scanpy to identify the top 20,000 most variable features for analysis. As the dimensionality of the ADT data is relatively low, we use centered log-ratio (CLR) normalization.

For integrating mosaic data, it is more important to align the data features obtained from different sequencing technologies. For RNA and ADT data, their features have clearly labeled names, so we directly take the shared feature set. For ATAC data, we use the Reduce function in Signac to re-merge the features across multiple batches, and subsequently, we perform integration through this peak matrix.

3 Method Comparison

MOFA+. Dataset integration is done using the mofa function in muon. Following the author's tutorial, we preprocess RNA, ATAC, and ADT modalities, then use the model framework muon.tl.mofa to integrate multimodal data, and finally obtain the latent representations. Since the muon package can only handle paired trimodal data, we compare it with our model only in paired modality datasets. https://muon-tutorials.readthedocs.io/en/latest/trimodal/tea-seq/1-TEA-seq-PBMC.html

Multigrate. We use the multigrate package to integrate datasets. Following the author's tutorial, RNA, ATAC, and ADT data are preprocessed, and the mosaic data is integrated using the model framework $multigrate.model.MultiVAE$, yielding the required latent representations. This method can integrate mosaic data, so it serves as a baseline for all datasets in comparison. https://github.com/theislab/multigrate/tree/main

TotalVI. Dataset integration is done using the totalVI function in scvi-tools. Following the author's tutorial, RNA and ADT raw count data are preprocessed, and the multimodal data is integrated using the $scvi.model.TOTALVI$ model framework, yielding the latent representations for RNA and ADT modalities. Since this method only handles RNA-ADT paired data, we only integrate the paired RNA and ADT in datasets that also contain ATAC and compare them with our model. https://docs.scvi-tools.org/en/latest/tutorials/notebooks/multimodal/totalVI.html

MultiVI. Dataset integration is performed using the MultiVI function in scvi-tools. Following the author's tutorial, RNA and ATAC data are preprocessed, and multimodal data is integrated using the $scvi.model.MULTIVI$ model, yielding the latent representations for RNA and ATAC modalities. This method handles RNA-ATAC integration but does not require paired data, so we integrate the RNA and ATAC modalities in our datasets. https://docs.scvi-tools.org/en/latest/tutorials/notebooks/multimodal/MultiVI_tutorial.html

Mowgli. We use the Mowgli package to integrate datasets. Following the author's tutorial, RNA, ATAC, and ADT data are preprocessed, and multimodal data is integrated using the $mowgli.models.MowgliModel$ framework, learning latent representations in the embedded space. This method only handles paired multimodal data, so we compare it with our model only in paired datasets. https://mowgli.readthedocs.io/en/latest/

SnapATAC2. We use the SnapATAC2 package for dataset integration. Following the author's tutorial, RNA, ATAC, and ADT data are preprocessed, and multimodal data is integrated using the $snapatac2.tl.multi\_spectral$ model framework, yielding latent representations in the embedded space. This method only integrates paired multimodal data, so it is compared with our model in paired datasets. https://kzhang.org/SnapATAC2/index.html

4 Evaluation Metrics

4.1 Batch correction Metrics

Graph iLISI: An improved metric based on iLISI, which calculates the inverse Simpson's index to measure the effective batch number in the kNN neighborhood. The value ranges between 1 and the total number of batches N. Graph iLISI uses graph-based distance metrics to introduce graph information. The score is normalized between 0 and 1, where 0 indicates perfect separation and 1 indicates perfect mixing.

Graph connectivity: Measures whether cells with the same label are connected in the kNN graph, with a value between 0 and 1. The higher the value, the greater the connectivity between cells with the same label. Given a label c, assume there are n_c cells with this label. The number of connected components is m_c. The connectivity rate is $p_c = m_c/n_c$, and the average of all labels' p_c is the Graph connectivity. A value of 1 indicates perfect mixing across batches.

4.2 Biological Conservation Metrics

NMI (Normalized Mutual Information): Measures the similarity between two clusterings. The true cell labels are compared with the clusters obtained by the model. NMI ranges between 0 and 1, where 0 indicates no shared information, and 1 indicates complete correlation. Louvain clustering is often used.

ARI (Adjusted Rand Index): Measures the overlap between two clustering results. The Rand Index (RI) is calculated based on random pairs of cells, and ARI is a corrected version of RI. ARI ranges between $[-1, 1]$, where 1 indicates perfect positive correlation, 0 indicates no correlation, and values less than 0 indicate negative correlation.

Isolated label F1: Measures the classification accuracy of rare cell types. Cells that appear the least across batches are labeled as 'isolated', and their F1 score is calculated. An Isolated label F1 score of 1 indicates perfect clustering performance on rare cell types across batches.

Graph cLISI: Measures the separation between different cell types. Similar to Graph iLISI, but focuses on cell type labels rather than batch labels. The value ranges between 0 and 1, with 0 indicating low separation of cell types.

5 Train detail

During the entire training process, key hyperparameters can affect the model's performance. We use $k = 30$ to construct the nearest neighbor graph, set the contrastive learning temperature parameter to 0.5, the triplet loss margin to 0.2, and the total loss weight coefficients are 1, 10, 100, 100, and 10, respectively. The model's encoders consist of two layers, with the first layer for connection and the second for the latent representation of the modality, with dimensions of 128 and 16, respectively. The decoder layers also have two layers, with dimensions of 128 and the corresponding dimension for each modality. The learning rate of the neural network autoencoder is 0.001, with a weight decay of $5e^{-4}$, and the model is optimized using the Adam optimizer. The model is implemented based on the PyTorch framework.

6 Result

6.1 scCGM Integration of Transcriptomics and Chromatin Accessibility Paired Data

In the *Chen*_2019 dataset(Figure S1), scGCM and Snap had similar NMI and ARI performance, significantly outperforming other methods, reaching 0.75-0.8. The Mowgli method, based on non-negative matrix factorization, performed the worst, with values of only 0.2 and 0.4. Graph Connectivity and cLISI were mostly even across methods, with Snap leading in il_score_f1, while scGCM ranked third, achieving over 0.5. From the UMAP plot(Figure S3), only scGCM and Snap clearly separated cell types, while the results of other methods were more blurred. Finally, in the Xie_2023 dataset, scGCM achieved the highest NMI and ARI, both exceeding 0.9. The performance of Multigrate, Snap, MultiVI, and Mowgli was relatively close, while the MOFA+ method, based on principal component analysis, performed the worst, with an ARI of only 0.4. Graph Connectivity and cLISI showed consistent results across all methods, but in terms of il_score_f1, scGCM maintained a high level alongside other methods. Reviewing the UMAP plots again, other methods still had some blurred regions in cell type recognition, while scGCM distinctly separated different cell types, with more compact clustering for each type. These results demonstrate that scGCM excels in integrating RNA and ATAC paired data, with precise cell type identification and stable performance across various datasets.

For the Ma_2020 dataset(Figure S2), scGCM's NMI and ARI were significantly better than all other methods, reaching 0.8. The Snap method, based on spectral clustering, performed slightly better than the other methods by about 0.1. Graph Connectivity and cLISI results were generally consistent across all methods, with Snap achieving the highest il_score_f1 of 0.8. As seen in the UMAP plot(Figure S4), other methods had blurred cell type identification, while scGCM successfully clustered different cell types distinctly. Based on the UMAP visualization of marker genes and existing knowledge, we know that the Krt1 gene serves as an important marker in the process of skin cell differentiation. As keratinocytes differentiate from the basal layer to the spinous layer, the expression of Krt1 gene gradually increases. Although there is limited research on the Cmah gene in ORS cells, existing knowledge suggests that, since the Cmah gene is inactive in humans, it may contribute to the aging or degeneration of hair follicles, providing a direction for future research.

6.2 scCGM Integration of Transcriptomics and Protein Data

In another set of mosaic data integration, we compared scGCM with Multigrate(Figure S7-9). The results showed that scGCM's ARI exceeded 0.8, while Multigrate only reached 0.4. Additionally, scGCM outperformed Multigrate in NMI by about 0.1. Although the two methods had similar performance in Graph Connectivity and cLISI, scGCM significantly outperformed Multigrate in iLISI. In summary, scGCM more clearly identifies different cell types when handling this type of data and excels in integrating data from different batches. These results indicate that scGCM is not only well-suited for integrating RNA and ADT data but also excels in cell type identification and demonstrates stable performance, showing outstanding integration results across various datasets. In the UMAP plots of marker genes, we observe that CD14 is highly expressed primarily in CD14 monocytes, while TCF7L2 plays an important role in CD16 monocytes. CLEC10A is highly expressed in cDC2 cells, participating in antigen presentation and playing a key role in the initiation and regulation of the adaptive immune response. The function of this gene indicates its indispensable role in promoting the adaptive immune system's effective response to pathogens. Meanwhile, MS4A1 is a specific marker for B cells, and its expression precisely distinguishes B cells. In the UMAP plots of marker proteins, we see that CD19 and CD22 are highly expressed molecules on the surface of B cells, playing important roles in the development, activation, and function of B cells. CD244 is highly expressed in NK cells, and when it provides a positive stimulatory signal, it enhances NK cell activity, promoting recognition and lysis of target cells and boosting the immune system's clearance capacity.

6.3 scCGM Integration of trimodal mosaic Data

The experimental results for the DOGMA-seq and TEA-seq mosaic datasets are presented in Figures S13-16. The results show that scGCM's ARI was approximately 0.2 higher than the other two methods, and its NMI was at least 0.1 higher. Although scGCM was slightly lower in the Graph Connectivity metric, for other metrics, the methods performed similarly, with scGCM having a slight advantage. From the UMAP plots, it is evident that in the DOGMA-seq dataset, scGCM and Multigrate had clearer clustering

results and performed well in batch integration, while MultiVI performed poorly. In the TEA-seq dataset, scGCM and MultiVI performed well in clustering, while Multigrate performed poorly. All three methods were relatively stable in handling batch effects.

Finally, the experimental results for the 10X-ASAP-DOGMA mosaic dataset are presented in Figures S22. In this dataset, scGCM's ARI and NMI both reached a level of 0.8, while the ARI of the other two methods was only 0.4 and 0.6. In the iLISI metric, scGCM performed the best, demonstrating its excellent integration performance. From the UMAP plots, it is clear that scGCM achieved clearer cell type integration, aiding in a better understanding of biological insights, and scGCM was the only method to successfully achieve full integration of all six batches of data. Based on the UMAP visualization of marker genes and existing knowledge, we know that the FCGR3A gene is primarily expressed in CD16 monocytes and is closely associated with antibody-dependent cell-mediated cytotoxicity reactions in the immune system. The BLK gene enhances B cell signaling by phosphorylating downstream signaling molecules, thereby promoting B cell proliferation and differentiation and supporting antibody production. The PAX5 gene is the master regulator of B cell development, ensuring proper gene expression at different stages of B cell development, while the BCL11A gene acts as a key transcription factor, maintaining B cell differentiation and function, allowing it to play a crucial role in adaptive immunity.These results further demonstrate scGCM's strong capability in integrating RNA, ATAC, and ADT mosaic data. Not only did it excel in cell type identification, but it also showed stable performance across various datasets, making it an effective tool for multimodal single-cell data integration.

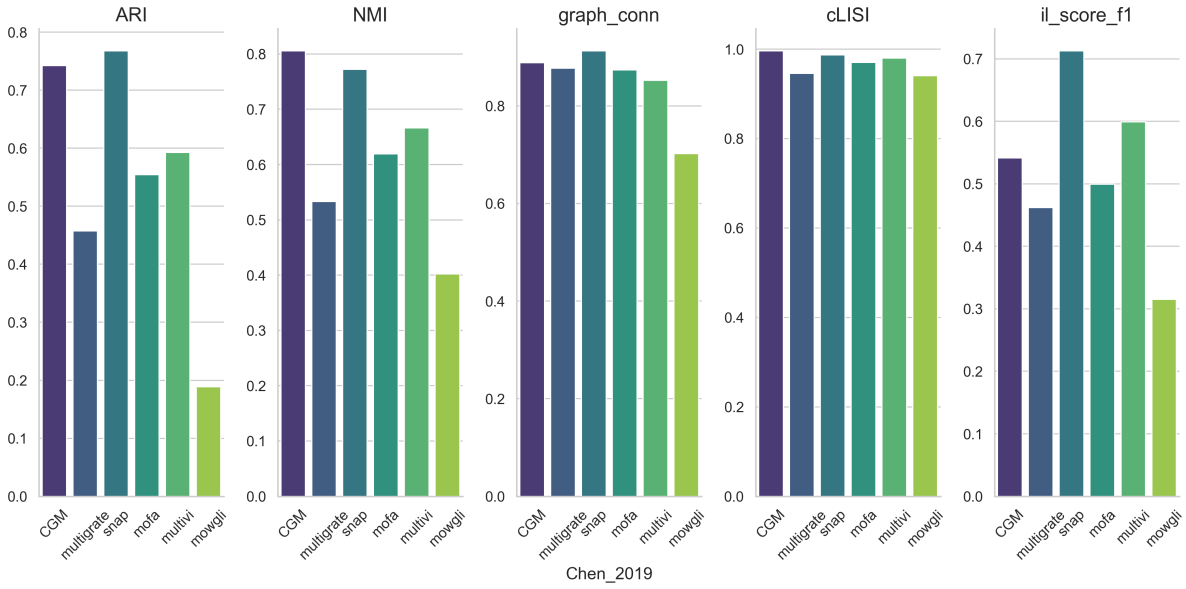

Figure S1: **benchmarking of performance in** *Chen_2019* **dataset.**

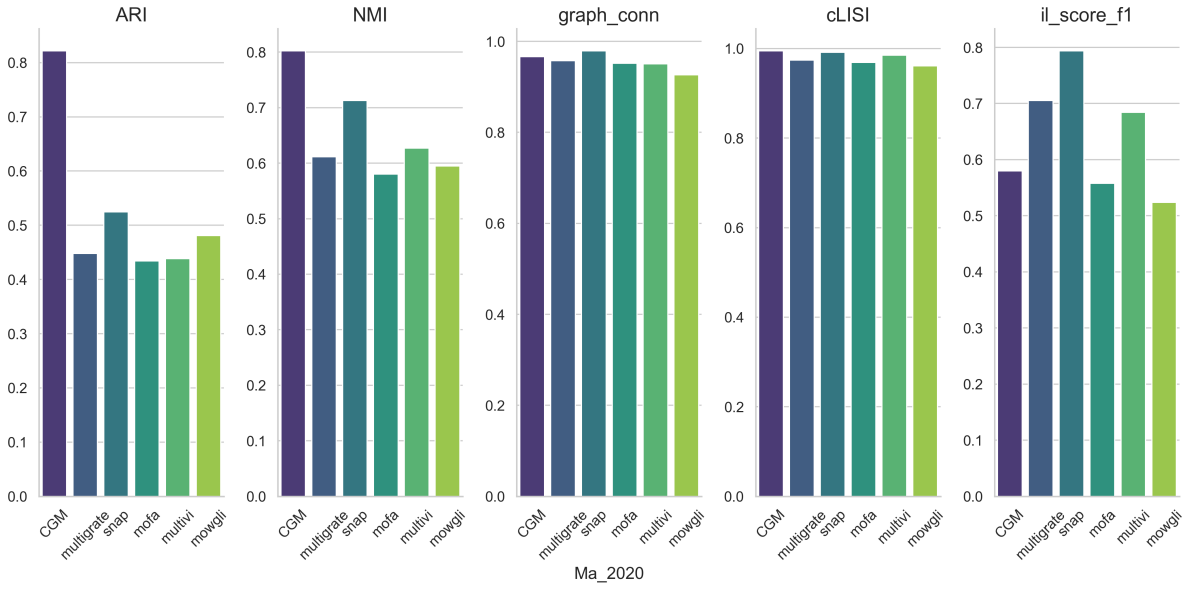

Figure S2: **benchmarking of performance in** Ma_2020 **dataset.**

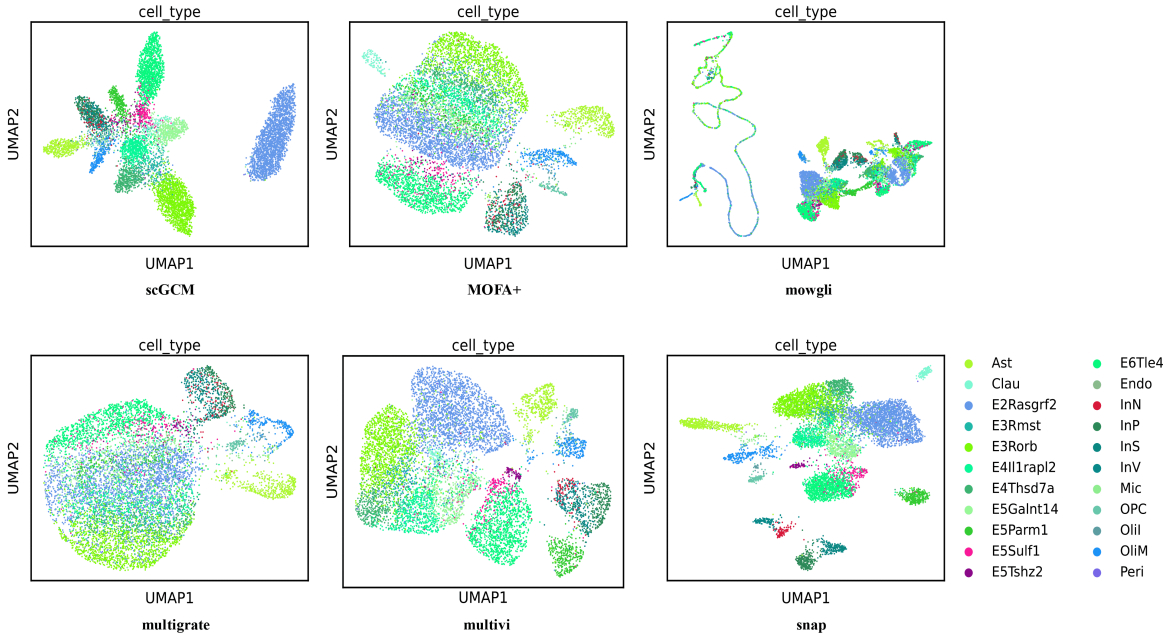

Figure S3: **UMAP visualization of cell embeddings obtained by scGCM and five other strategies in** *Chen*_2019 **dataset.**

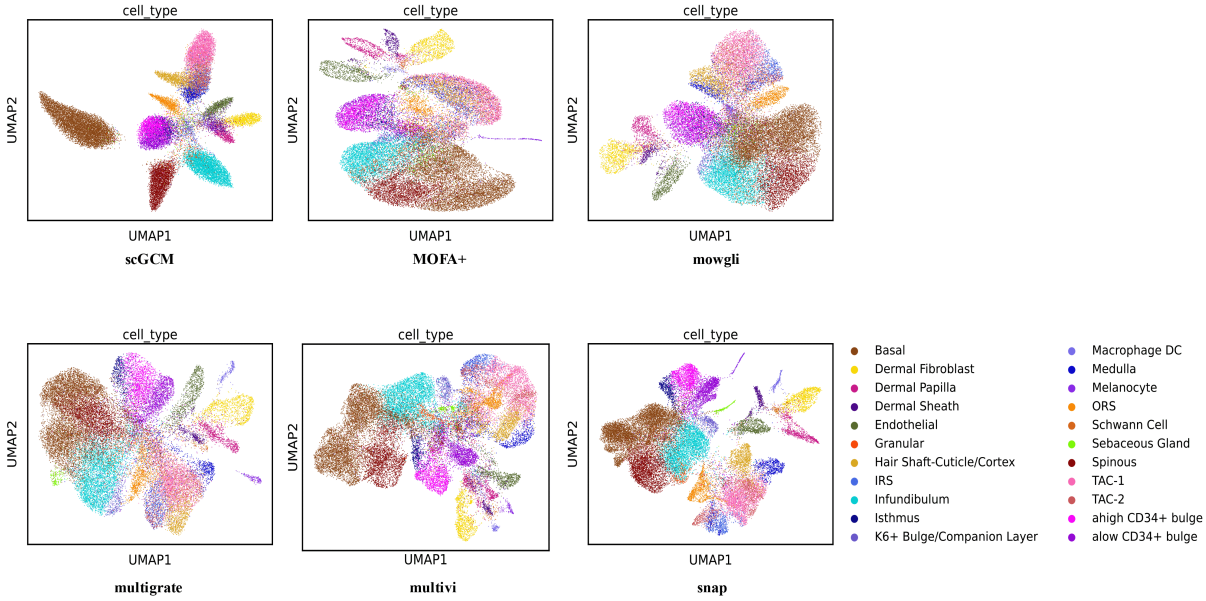

Figure S4: **UMAP visualization of cell embeddings obtained by scGCM and five other strategies in** Ma_2020 **dataset.**

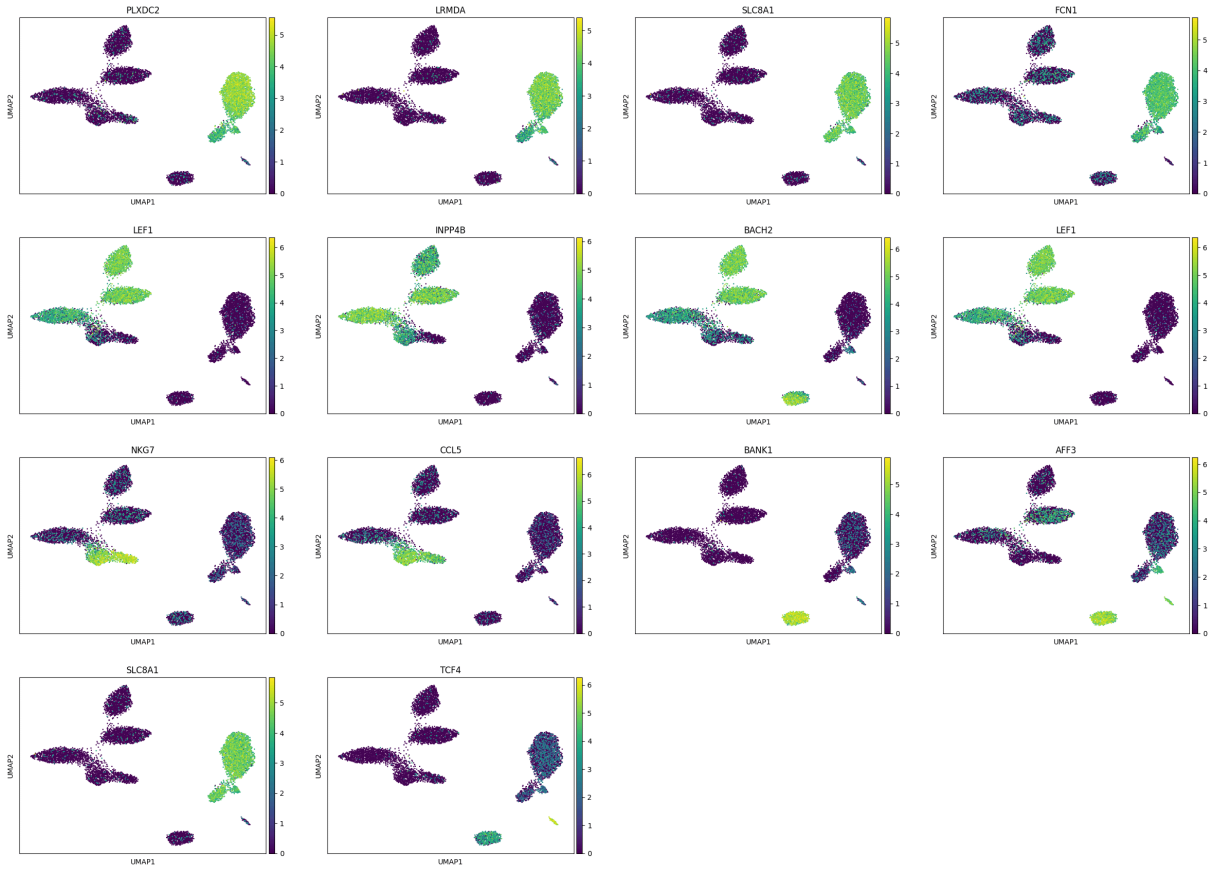

Figure S5: **UMAP visualization of Gene expression over all single-cell samples in** $10X\_PBMC$ **dataset.**

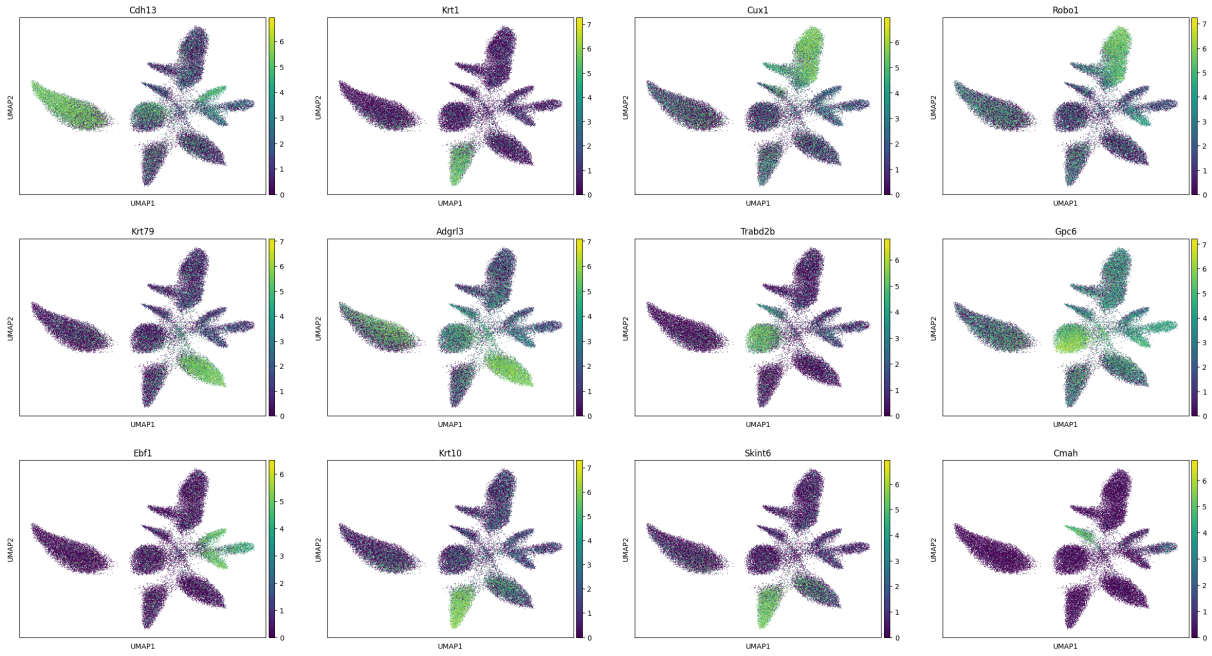

Figure S6: **UMAP visualization of Gene expression over all single-cell samples in** Ma_2020 **dataset.**

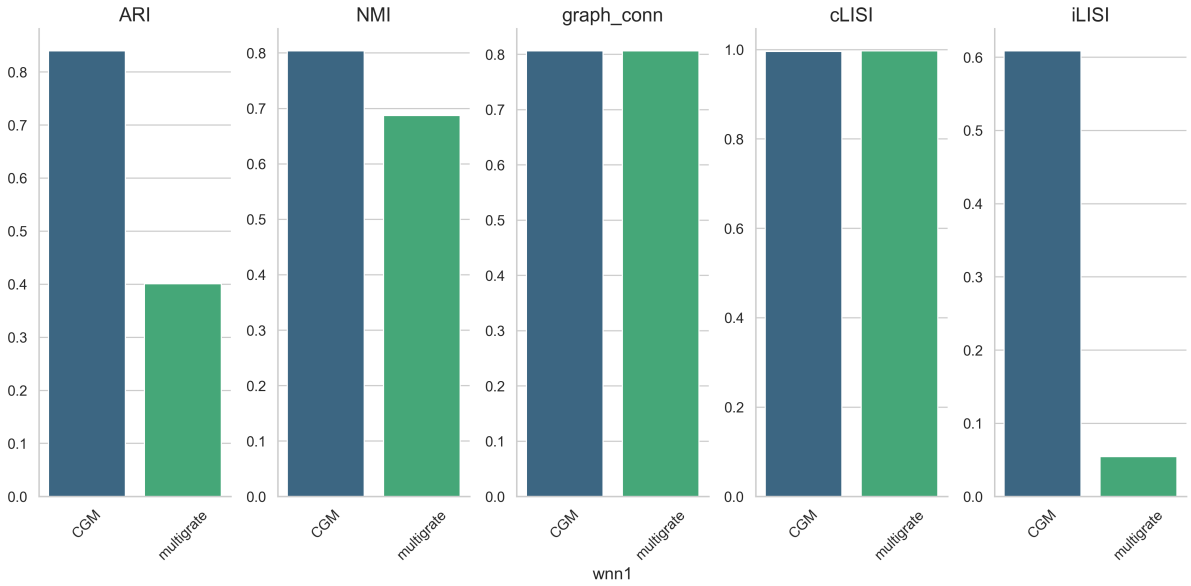

Figure S7: **benchmarking of performance in CITE-seq mosaic dataset.**

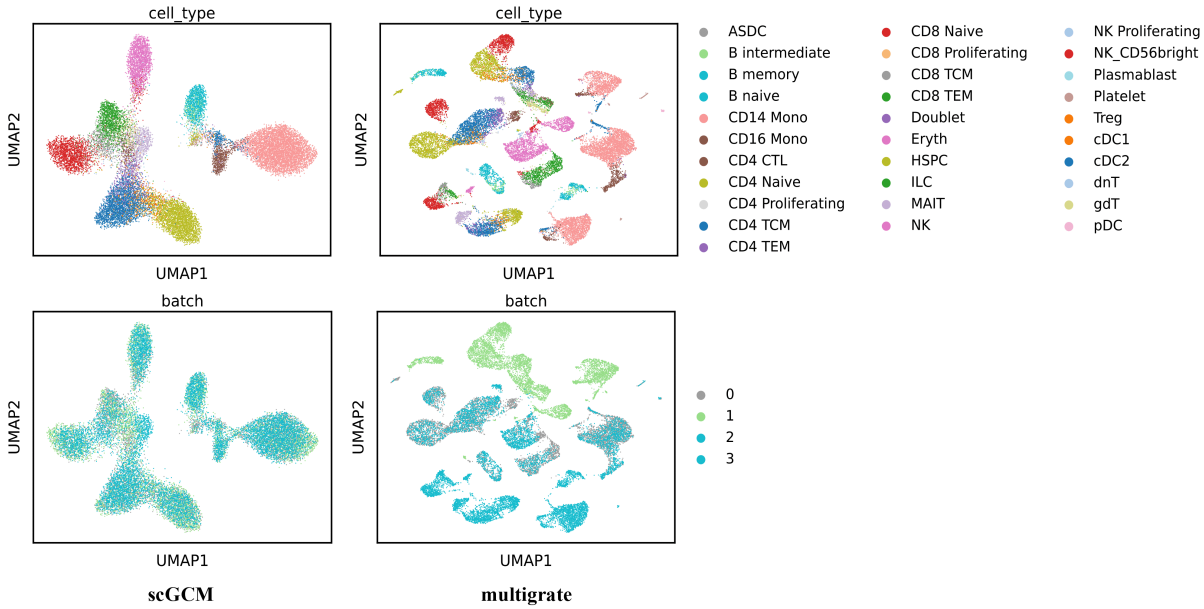

Figure S8: **UMAP visualization of cell embeddings obtained by scGCM and multigrate in CITE-seq mosaic dataset.**

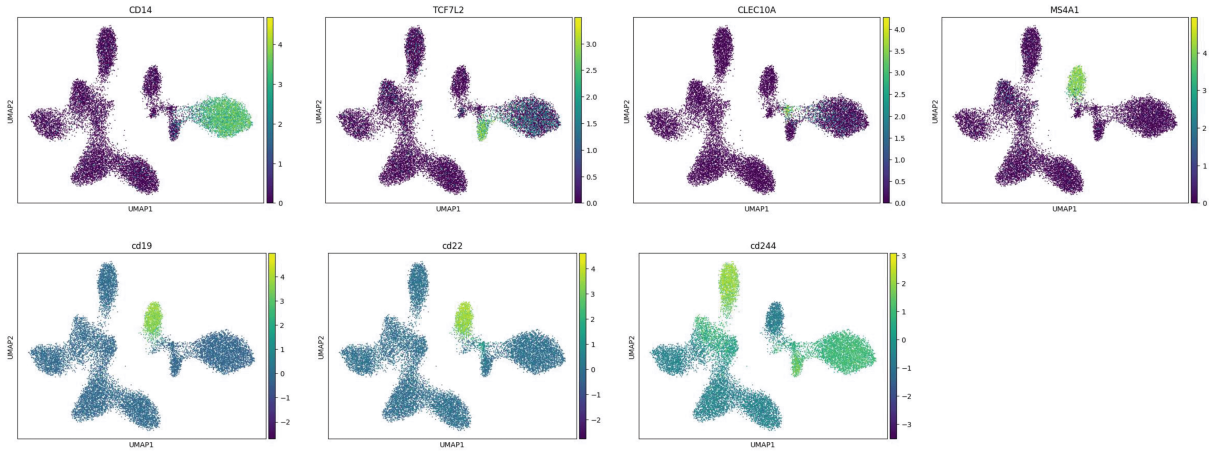

Figure S9: **UMAP visualization of marker gene and protein expression in CITE-seq mosaic dataset.**

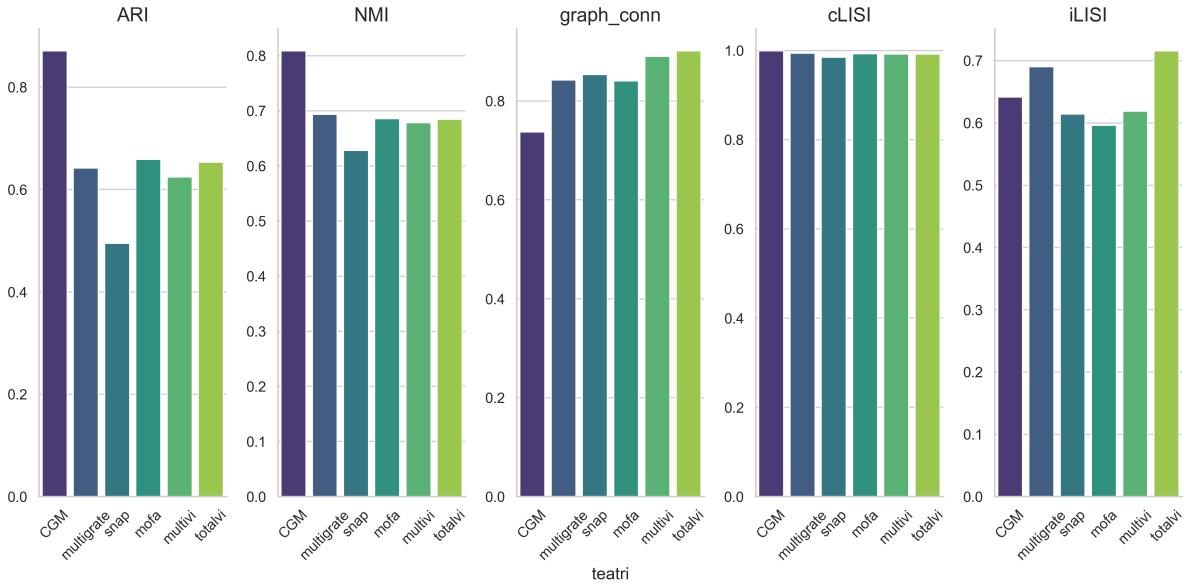

Figure S10: **benchmarking of performance in Tea-seq dataset .**

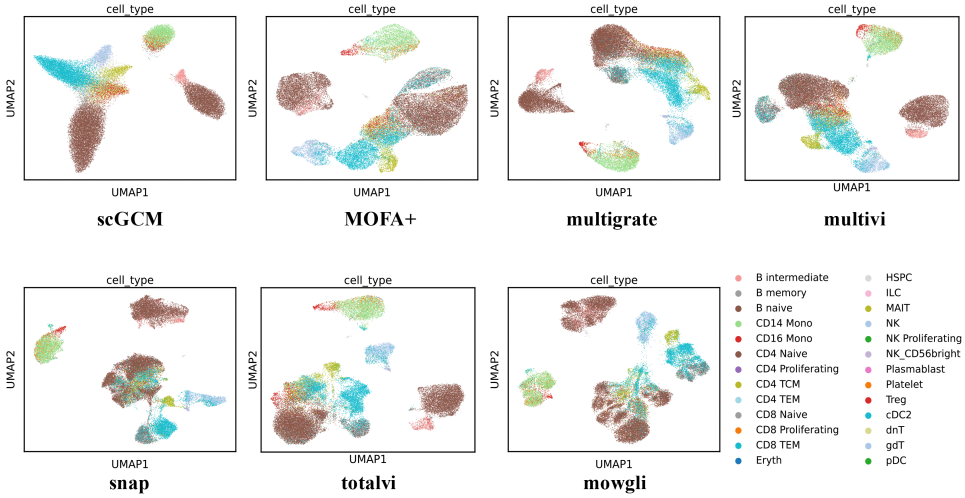

Figure S11: **UMAP visualization of cell embeddings obtained by scGCM and six other strategies in Tea-seq dataset.**

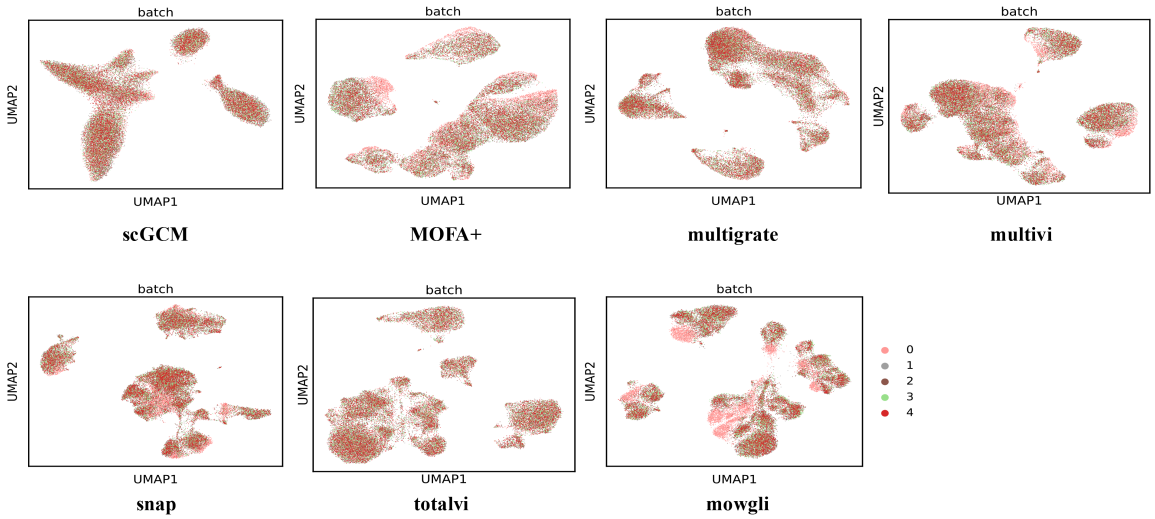

Figure S12: **UMAP visualization of batch in Tea-seq dataset.**

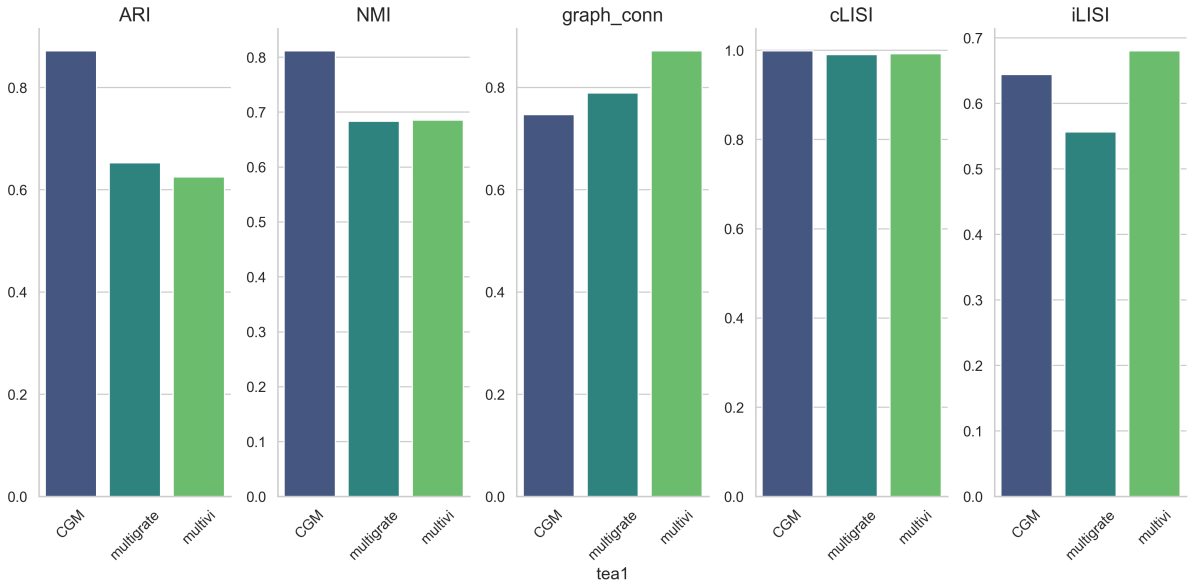

Figure S13: **benchmarking of performance in Tea-seq mosaic dataset.**

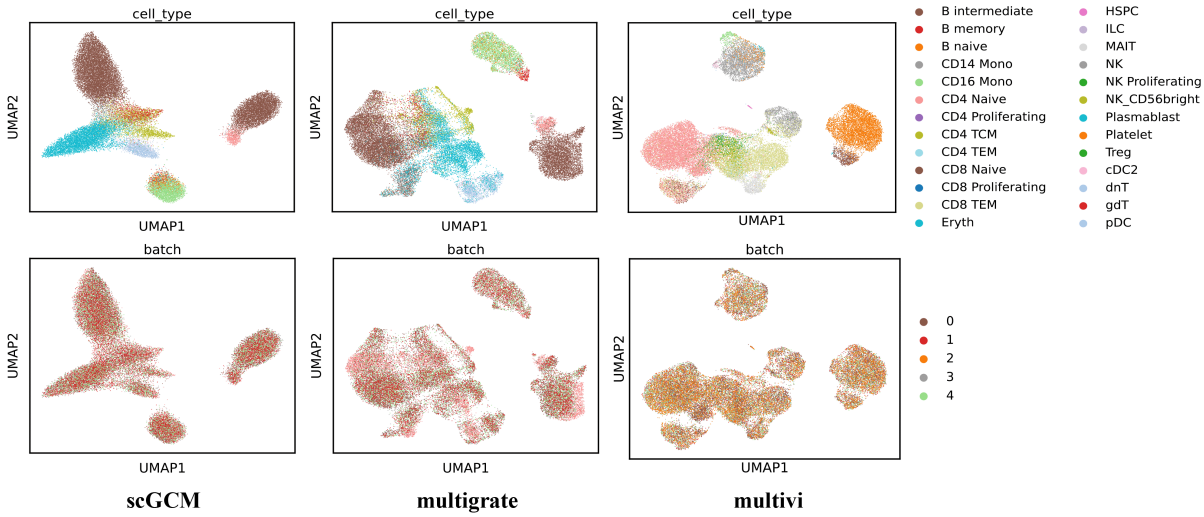

Figure S14: **UMAP visualization of cell embeddings obtained by scGCM,multigrate and multivi in Tea-seq mosaic dataset.**

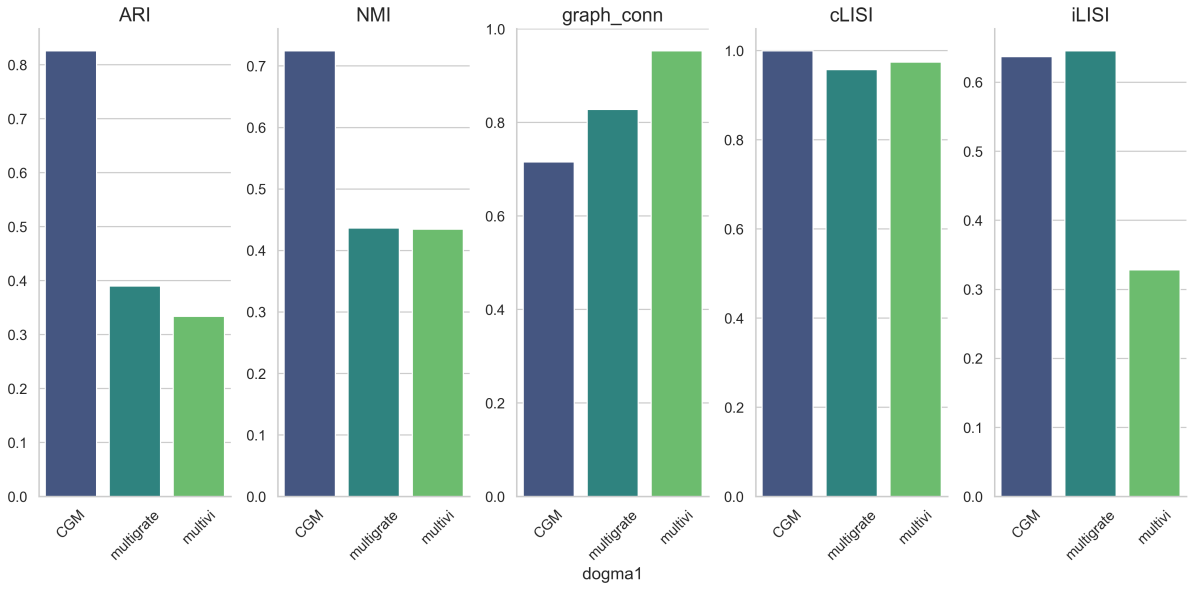

Figure S15: **benchmarking of performance in Dogma-seq mosaic dataset.**

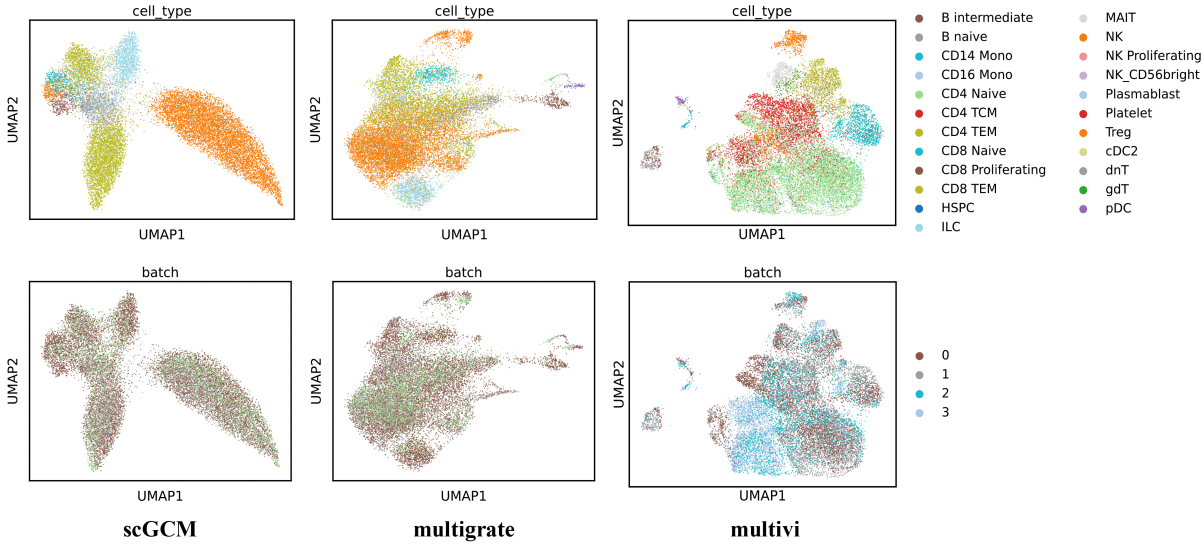

Figure S16: **UMAP visualization of cell embeddings obtained by scGCM,multigrate and multivi in Dogma-seq mosaic dataset.**

Figure S17: **UMAP visualization of marker gene expression in TEA-seq dataset.**

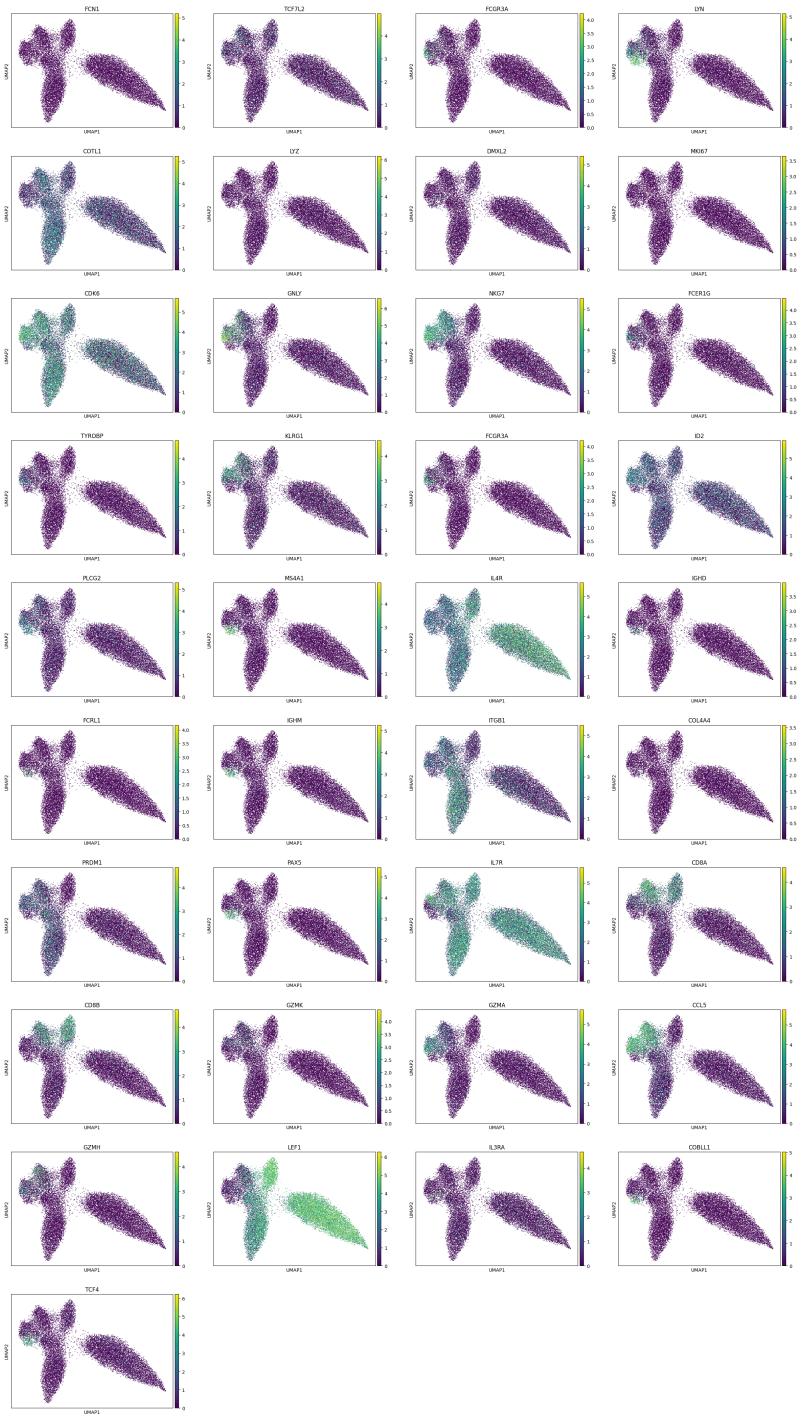

Figure S18: **UMAP visualization of marker gene expression in Dogma-seq dataset.**

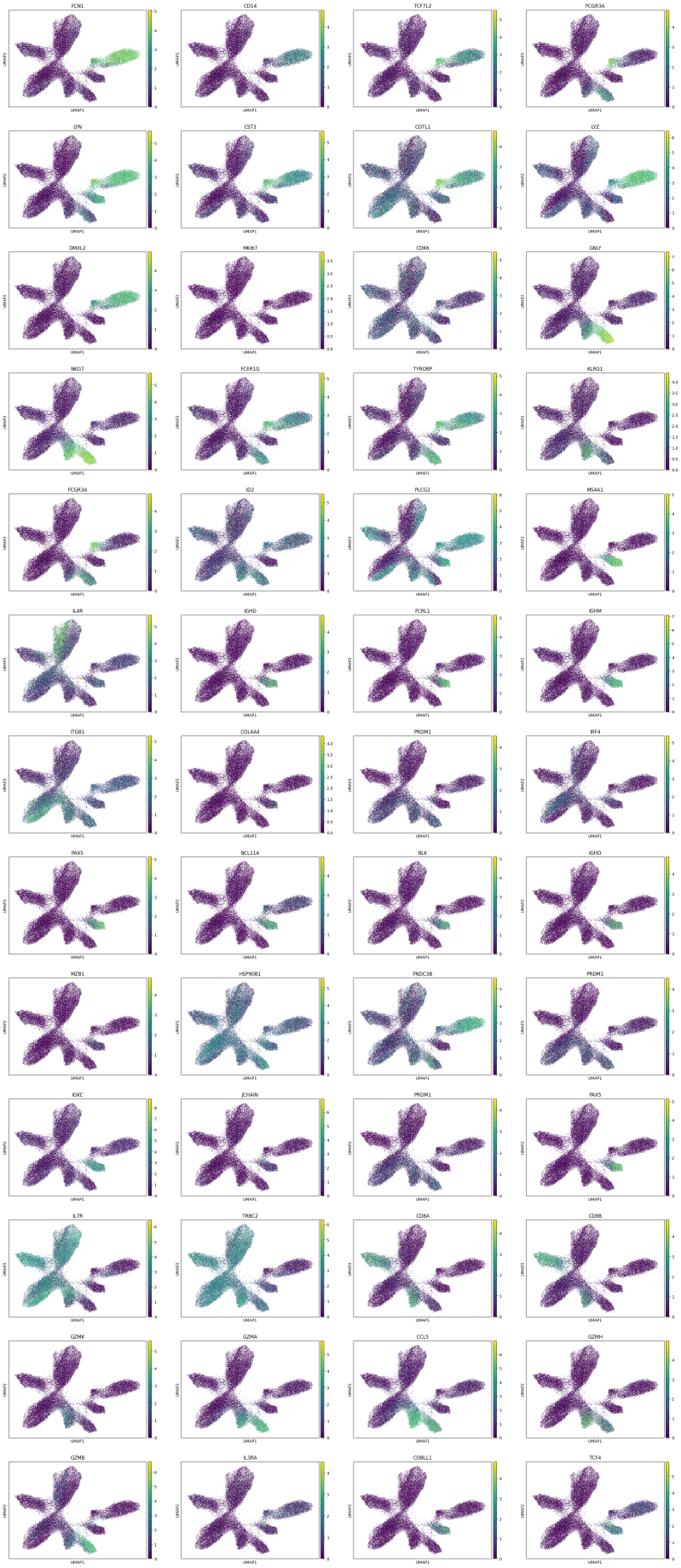

Figure S19: **UMAP visualization of marker gene expression in 10X-ASAP-DOGMA dataset.**

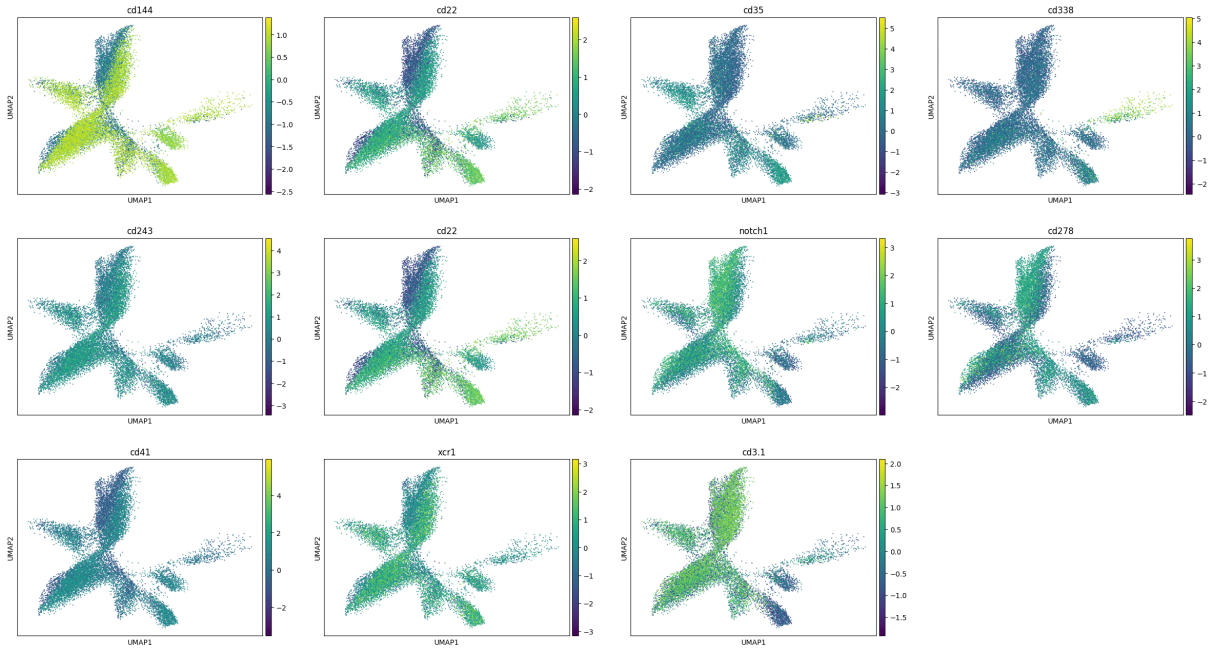

Figure S20: **UMAP visualization of marker protein weight in 10X-ASAP-DOGMA dataset.**

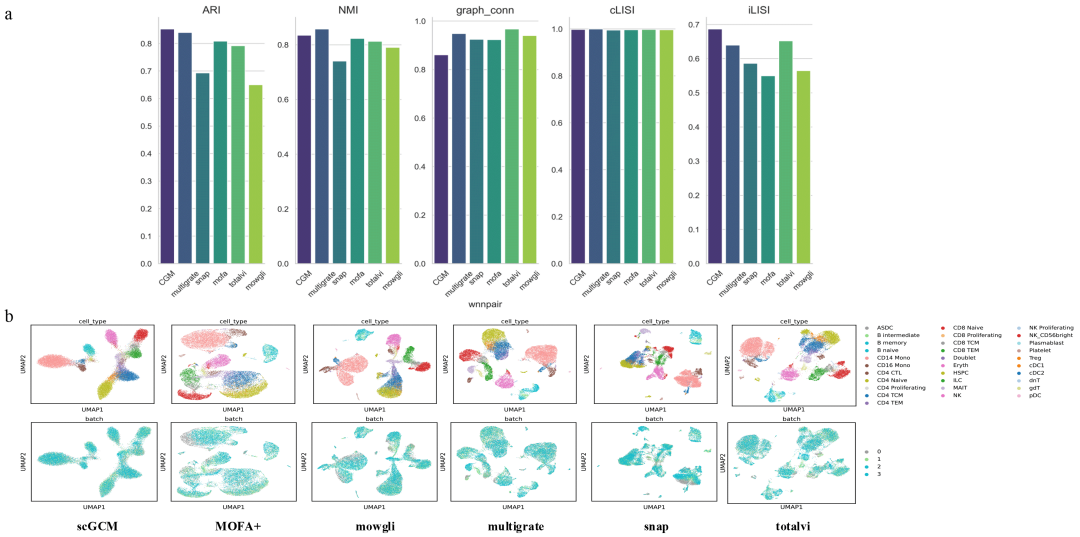

Figure S21: **Evaluation obtained with scGCM on RNA and ADT paired data. a.** benchmarking of performance in CITE-seq dataset . **b.** UMAP visualization of cell embeddings obtained by scGCM and five other strategies in CITE-seq dataset.

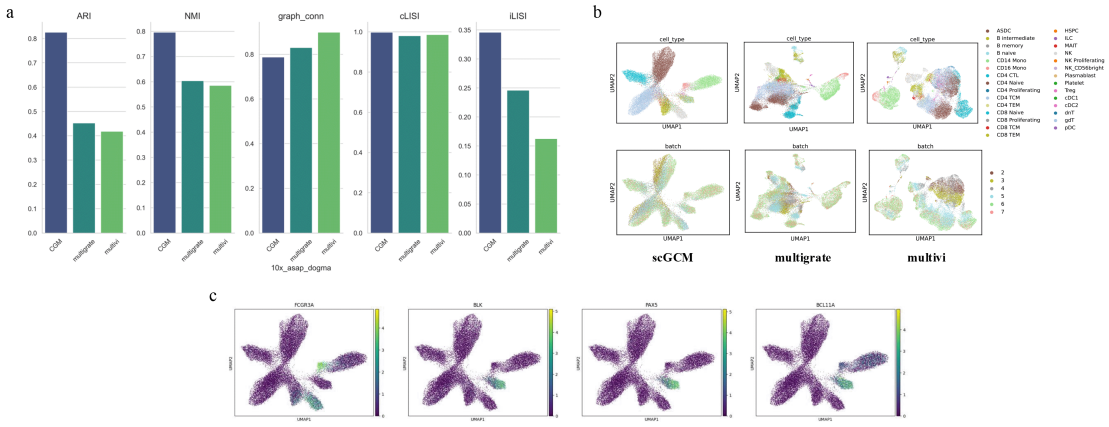

Figure S22: **Evaluation obtained with scGCM on RNA,ATAC and ADT mosaic data. a.** benchmarking of performance in 10X-ASAP-DOGMA dataset . **b.** UMAP visualization of cell embeddings obtained by scGCM and two other strategies in 10X-ASAP-DOGMA dataset. **c.** Gene expression of FCGR3A, BLK, PAX5 and BCl11A over all single-cell samples in 10X-ASAP-DOGMA dataset.