FGMSF：Multi-Stage Deep Fusion Clustering with Manifold Preservation and Adaptive Gating for Robust Single-Cell Multi-Omics Integration

Junhao Zou[1,2] • Shuang Quan Li[1] • Liu Liu[3] • Jiaqi Liu[2, 4, *]

1 School of Automation, Harbin University of Science and Technology, Harbin, Heilongjiang, 150080, China.

2 The First Affiliated Hospital, Cardiovascular Lab of Big Data and Imaging Artificial Intelligence, Hengyang Medical School, University of South China, Hengyang, Hunan, 421001, China.

3 School of Computer, University of South China, Hengyang, Hunan, 421001, China.

4 Hunan Provincial Key Laboratory of Multi-omics and Artificial Intelligence of Cardiovascular Diseases, University of South China, Hengyang, Hunan, 421001, China.

***Corresponding author. E-mail: liujqeureka@163.com**

This is a preprint. The paper has not yet been peer-reviewed or published.

Abstract

With the development of single-cell multimodal sequencing technology, compared with traditional single-modal analysis, multimodal data can provide more comprehensive and rich information from multiple biological levels, thereby more accurately characterizing cell state and functional characteristics. However, multimodal data still faces many challenges in the fusion and analysis process, such as misalignment between modals, data redundancy, severe noise interference, and inconsistent scales. Therefore, how to fully explore the complementary information between multimodal data, improve the accuracy of cell type recognition, and explore the potential relationship between cells has become a key scientific issue in current single-cell multimodal analysis. To this end, this paper proposes a multi-stage fusion cluster analysis model (FGMSF) based on the forgetting gate mechanism. The model first preserves the original manifold structure of the data through manifold fitting in the preprocessing stage, providing structural priors for subsequent fusion. In the first stage, the model achieves global feature integration through data alignment and modal fusion, while maintaining the original modal information to a certain extent, thereby retaining more local details for subsequent reconstruction and clustering tasks. In the second stage, the model constructs independent manifold spaces suitable for reconstruction and clustering tasks, and designs an EM heuristic optimization strategy based on the forgetting gate structure to maximize the effective transfer of information at the bottleneck layer, thereby improving clustering performance and reconstruction quality.Ultimately, the model can output a low-dimensional embedded manifold space while achieving high-quality reconstruction results and efficient and accurate clustering effects. In the evaluation of multiple downstream tasks, FGMSF not only significantly improved clustering performance, but also showed excellent performance in feature dimensionality reduction, batch effect removal, multimodal data integration, cell trajectory inference, etc., and also had good biological interpretability. This method provides new theoretical support and practical reference for in-depth analysis of multimodal single-cell data.

Keywords Multi-stage fusion • Manifold learning • Forgetting gate • EM heuristic optimization strategy • Multitask

1 | Introduction

With the rapid development of single-cell sequencing technology and R&D platforms, single-cell multiomics cluster analysis has brought unprecedented opportunities to identify cell types, elucidate cell-to-cell heterogeneity, and explore cell functions. Previously, tasks such as data analysis and cell clustering from single-cell data on RNA gene expression were not sufficient to gain insight into the complex structure of cells [1]. Therefore, most of the current methods are based on multi-view joint analysis of multiple modal characteristics of single cells, which can better understand the heterogeneity and intrinsic connections between cells [2]. Even so, single-cell data presents great challenges due to its high-dimensional characteristics, sparsity, noise, and large number of zero-value expression problems [3]. However, manifold learning, as a nonlinear dimensionality reduction method, can effectively alleviate these problems. The method is based on the manifold assumption, which states that high-dimensional data is usually nested on top of a low-dimensional nonlinear manifold in a higher-dimensional space. By mining the manifold structure of the data, the manifold learning method maps the high-dimensional data to the low-dimensional space while retaining its original geometric features and internal associations as much as possible [4, 5]. In view of the fact that single-cell data often contain complex cell differentiation trajectories and continuous lineage relationships, the introduction of manifold learning can help to reveal cell differentiation pathways and their internal relationships more accurately [6], thereby improving the accuracy and biological explanatory power of cluster analysis.

In recent years, many single-cell multi-omics cluster analysis models have emerged, and all of them have been improved in terms of their respective advantages, providing various new ideas and methods for single-cell data analysis. BREM-SC [7] is a single-cell multiomics data integration model based on the Bayesian framework, which is one of the earliest proposed models for cluster analysis of CITE-seq data. By combining random-effects models and Bayesian inference, it overcomes the limitations of unimodal analysis and provides a more complete understanding of cell function and regulatory mechanisms. It can also handle batch effects due to differences in technology or experimental conditions. This model not only accurately infers the correlation between cell state and omics, but also models the unique characteristics of each cell. TotalVI [8] is a deep variational autoencoder model whose core idea is to embed data from both modalities into a shared latent space to capture the commonalities and differences between multiple omics. Leveraging the expressive power of deep learning and the flexibility of Bayesian inference, it excels in denoising, integration, and joint analysis of single-cell transcriptome and proteome data, but models may have high requirements for input data quality. CiteFuse [9] is a multi-omics integration method for single-cell transcriptome data specifically used for antibody tagging, by independently calculating the similarity matrix between two different omics datasets, then applying an effective fusion algorithm to merge the matrix based on the similarity, and finally clustering the merged similarity matrix based on the existing graph fusion algorithm. It not only enables the analysis of RNA and ADT data separately, but also enables joint analysis in a shared potential space, including data integration, modal difference alignment, dimensionality reduction, and visualization, providing a one-stop solution from integration to visualization. Seurat [10] is a powerful and widely used single-cell data analysis toolkit for complete workflows from data preprocessing to advanced analysis. It supports the preprocessing, clustering and integration analysis of single-cell data, which is suitable for a variety of scenarios of multi-omics data integration, and can also select and combine different functional modules according to research needs. Specter [11] uses spectral embedding and graph theory methods to efficiently and accurately integrate data of different modalities, and integrate data of different modalities through efficient neighboring structures. The core idea of the model is to uncover the biology of cell types and states by constructing and manipulating adjacency maps between cells to generate a unified latent representation that eliminates modal differences while retaining unique information about each modality. Cobolt [12] fuses multiple omics data mainly through co-clustering and shared similarity learning, and uses matrix factorization and low-dimensional embedding techniques to process high-dimensional data, so as to obtain consistent cell clustering results. scMM [13] uses a probabilistic graph model to jointly model the latent factors of multiple omics data based on Bayesian inference, so as to realize the integration and clustering of cross-omics data. By sharing the potential space, scMM can effectively fuse data from different omics, and combine dimensionality reduction and adaptive clustering strategies to improve the clustering effect and robustness of analysis.

Because various objective problems during the data extraction process, the quality of single-cell data may be greatly reduced, one of the most important issues is the 'Dropout' event in single-cell data [14, 15], which refers to the phenomenon that in single-cell RNA sequencing (scRNA-seq) or other single-cell technologies, the true expression value of some genes is not detected due to technical limitations or biological factors, so that the expression value is zero in the data. And it has strong sparsity and non-randomness, so that the zero value is not uniformly distributed, but biased towards low-expression genes. The distribution of the data usually appears as a discrete negative binomial distribution or a zero-bloated distribution. In order to solve the problem of zero bloat and over-dispersion for single-cell counting data, Davide Risso et al. proposed the ZINB model to effectively capture the characteristics of dropout data [16], thereby improving the analytical performance of the model. Subsequently, Gökcen Eraslan et al. also used this ZINB model to further develop an innovative single-cell cluster analysis method [17], which effectively improved the downstream analysis performance. Based on the above analysis, we find that many models only focus on solving the clustering problem, but the improvement of data quality can also provide effective information support for subsequent research. If we can further optimize the data quality while improving the clustering effect, this undoubtedly reflects the comprehensive advantages of the model at two levels. This comprehensive improvement not only enhances the ability to structure the data, but also lays a more solid foundation for downstream analysis tasks, so as to discover the heterogeneity and intrinsic connections between cells, and improve the reliability and performance of the model in practical applications.

In order to solve the above challenges, this paper proposes a multi-stage fusion clustering model fgmsf based on forgetting gate by processing the manifold structure of data, aiming to achieve feature dimension reduction, batch effect removal, data integration, trajectory inference and improve the quality of data, while still improving the clustering effect.Fgmsf is an end-to-end single-cell multi omics clustering analysis method. The main process is shown in Figure 1a. After the single-cell multi omics data is modeled, high-quality reconstructed data, low dimensional manifold space and efficient clustering results are obtained, which are then used for multi downstream task analysis and the corresponding results and analysis are given.The specific model architecture is shown in Figure 1b. Fgmsf is based on the variational autoencoder. First, three kinds of fitting data with manifold structure are obtained by manifold fitting after three kinds of conversion of the original data, and then the best fitting data are selected through discrimination, and then the random noise is added to the two encoders of the model. The encoder adopts a partially shared coding mode to achieve the alignment of data feature space, which can also better remove batch effects, and then splice the alignment features to complete the first stage of fusion.Then Bayesian reasoning is carried out, and feature space fusion is carried out on the bottleneck layer to obtain the popular space, and then the second stage fusion is completed. In order to further improve the clustering effect and optimize the data quality, in the second stage, we used two different fusion strategies to build an independent manifold space for the clustering task and the data reconstruction process, so as to maximize the effective transmission of information flow in the bottleneck layer. Specifically, in the reconstruction process, the feature space weighted form is used for fusion, and then the reconstructed data are optimized by independent decoder and ZINB model. In the clustering task, we designed an EM heuristic optimization strategy based on forgetting gate (EM-FG) to fuse the feature space to improve the clustering effect.

The article is structured as follows, and in Section 2 we give a detailed description of the FGMSF method and the specific analysis steps. In Section 3, we present the results and the experimental proof for the model. In Section 4, we perform a multi-task downstream analysis and meanwhile also giving the relevant biological validation. Finally, in Section 5, we made a correlation conclusion.

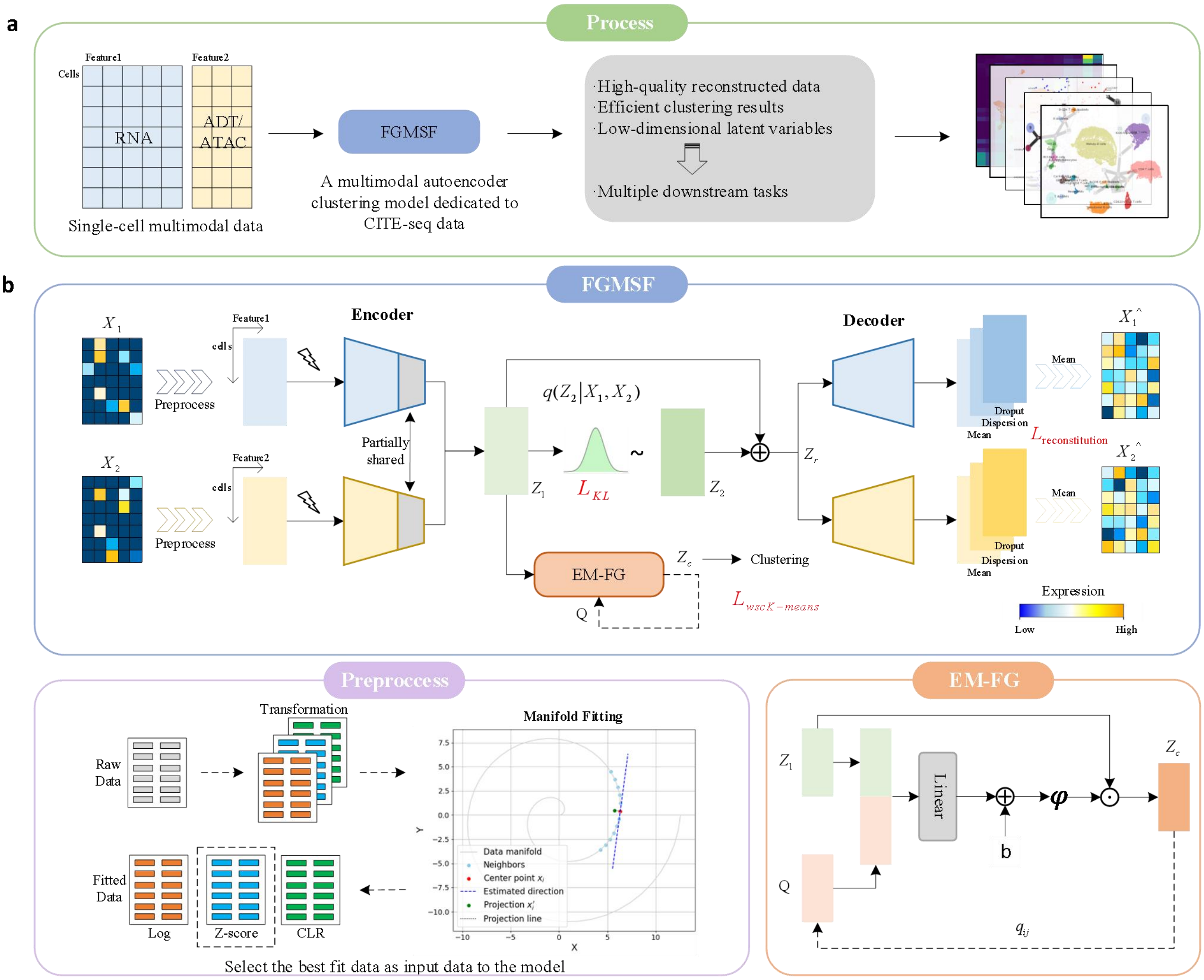


Fig.1 First, the single-cell multimodal data are input into the fgmsf model, and then the reconstructed data, clustering results and low dimensional manifold space are obtained. Finally, the multi downstream task analysis is carried out (a). Fgmsf model structure block diagram (b), first the original data through manifold fitting to get the best fitting data, then add the random

noise input model, after two parts share the encoder, carry out feature stitching, and then carry out Bayesian reasoning and feature space fusion to get two independent manifold spaces. In the reconstruction and clustering, use weighted fusion strategy and em-fg optimization strategy respectively, and finally get high-quality reconstructed data and effective clustering results.

2 | Materials and Methods

2.1 | Datasets

7 sets of RNA and ADT datasets were used in this paper, among which the PBMC dataset was available on the 10X website (https://www.10xgenomics.com/datasets), and the cell type labels is available from GitHub of Specter (https://github.com/canzarlab/Specter). There were also cells from the spleen and lymph nodes of two wild-type mice (biological replicates), which were stained with 111 antibodies or 208 antibodies for each mouse over two days, yielding 4 sets of datasets (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi,GSE150599), and the two sets of 111 antibody staining data were recombined into a new dataset SLN111, and the two groups of 208 antibody staining data were recombined into a new dataset SLN206. Its cell type labels is available on TotalVI (https://github.com/YosefLab/totalVI_reproducibility) on GitHub. 2 sets of RNA and ATAC datasets were also used. PBMCs of human peripheral blood mononuclear cells were expressed as pbmc3k and pbmc10k, respectively, which can be obtained from the 10X genomics website (https://www.10xgenomics.com/datasets), and the cell abels are transferred by Signac (v1.4.0) from the annotated datasets.

2.2 | Methods

We used a total of 9 datasets for validation, including 7 datasets containing RNA and ADT counts and 2 datasets containing RNA and ATAC counts, as well as their cell type labels. The whole process of the method is divided into three parts: pretreatment, model training and downstream tasks.

1. Preproccess stage

We use three data transformations on the original data, and then perform manifold fitting [18] to obtain the fitting data of the three transformations. Finally, we screen the best fitting data as the final input data of the model through clustering discrimination. The three transformations are Logarithmic Transformation (log), Z-score Normalization (Z-score) and Centered Log-Ratio Transform (CLR).

Log transformation is a commonly used method in single-cell data preprocessing. It is mainly used to narrow the expression span, reduce the dominance of highly expressed genes, and stabilize the variance of the data, making the data closer to the normal distribution, which is conducive to subsequent dimension reduction, clustering and other analysis.

Z-score normalization is used to convert gene expression data into a distribution with a mean of 0 and a standard deviation of 1, so as to eliminate the influence of different gene expression levels and make each gene comparable in dimension reduction and clustering.

Log ratio transform is mainly used for the preprocessing of proportional data, which is commonly used in single-cell multimodal data analysis, such as protein data (ADT) of CITE-seq. It can effectively deal with compositional data, avoid the pseudo correlation problem caused by the sum of proportional values constraint, and make the data more suitable for dimension reduction, clustering and other operations .

$$y = \log(\frac{x_i}{g(x)}), g(x) = \left(\prod_{j=1}^{D} x_j\right)^{1/D} \tag{1}$$

where x_i is the ith component in the sample, $g(x)$ is the geometric mean of all components of the sample, and *D* is the number of genes.

The datasets transformed are labeled as $y_1, \cdots, y_n$, respectively. For a transformed data $y \in \{y_1, \cdots, y_n\}$, the manifold hypothesis assumes that the *y* can be viewed as a noisy representation of the underlying low-dimensional manifold *M* of dimension *d* ($d \ll D$), with *D* as the number of genes. The fitting process includes two parts, direction estimation and projection estimation.

Direction estimation. The purpose of finding the projection direction is to determine the direction where the point should move closer to the underlying manifold. For any point *y*, its projection direction can be expressed as $d_v = F(y) - y$, where *F(y)* is local weighted average of y. *F(y)* is defined by the following formula [19]:

$$F(y) = \sum_i \alpha_i(y) y^{(i)} \tag{2}$$

where defined as:

$$\alpha_i(y) = \frac{\tilde{\alpha}_i(y)}{\tilde{\alpha}(y)'}, \tilde{\alpha}(y) = \sum_{i \in l_y} \tilde{\alpha}_i(y) \tag{3}$$

$$\tilde{\alpha}_i(y) = \begin{cases} (1 - \frac{\left\| y - y^{(i)} \right\|_2^2}{r_0^2})^k, \left\| y - y^{(i)} \right\|_2 \le r_0 \\ 0 \end{cases} \tag{4}$$

and *k* > 2 (*k* is a fixed integer, usually set to 3) ensures smoothness.

Computing the shared nearest neighborhood. For each $y^{(i)}$, we denote $N_p(i)$ as *p* (default by 15) nearest neighborhood of $y^{(i)}$, as determined by a given metric. $N_p(i)$ contains the key sample points where $y^{(i)}$ should pay attention to. Hence, the shared nearest neighborhood (SNN) of $y^{(i)}$ and $y^{(j)}$ is defined as:

$$SNN(i,j) = \left| N_p(i) \cap N_p(j) \right| \tag{5}$$

The final projection direction $F(y^{(i)})$ is defined as:

$$\mathbb{R}(i) = \arg_{S\subset y,|S|=p} \max \sum_{y^{(i)}\in\mathbb{R}(i)} SNN(i,j) \tag{6}$$

$$F(y^{(i)}) = \frac{1}{|\mathbb{R}(i)|} \sum_{y^{(i)}\in\mathbb{R}(i)} y^{(i)} \tag{7}$$

Projection estimation. Considering that single-cell data has high-dimensional features and irregular noise, this method simplifies projection estimation $\mathrm{G(y}^{(i)})$ by identifying the maximum density point on the line connecting $y^{(j)}$ and $F(y^{(i)})$, and replacing the original $\mathrm{G(y}^{(i)})$ with that point. The calculation process is as follows:

$$G(y^{(i)}) = \arg\max_{y_t} \rho(y_t) \tag{8}$$

$$y_t = y^{(i)} + t(F(y^{(i)}) - y^{(i)}) \tag{9}$$

$$\rho(y^{(i)}) = \frac{1}{\sum_{y^{(i)}\in\mathbb{R}(i)} \left\| y^{(i)} - y^{(j)} \right\|_2^2} \tag{10}$$

The fitting data is represented as :

$$X = \left\{ G(y^{(i)}) \right\}_{i=i}^{N} \tag{11}$$

In the end, we directly used simple k-means clustering for discrimination and selected the best fitting data as the input for the model.

2. Training stage

FGMSF reduces dimensionality while preserving some inherent connections in the data through the first stage of fusion, then imports its connections during the second stage of fusion, and finally maximizes the information flow of the bottleneck layer through two independent manifold spaces.

We add the preprocessed RNA and ADT/ATAC data to random noise to obtain two partially shared encoder inputs, which are expressed as:

$$X_1 = X_R + k_R \times N_R, X_2 = X_A + k_A \times N_A \tag{12}$$

Where $X_R \in \mathbb{R}^{N\times D_R}$ and $X_A \in \mathbb{R}^{N\times D_A}$ are the pretreated RNA and ADT/ATAC data, and N is the number of cells, D_R and D_A are the characteristic dimensions of RNA and ADT/ATAC data. N_R and N_A are the random noises that are added to both (the mean is 0 and the variance is 1). k_R and k_A are the coefficients of the random noise, set to 2 and 1.5, respectively.

We use the variational autoencoder as the overall architecture of the model [20] to meet the needs of manifold learning for nonlinear mapping and dimensionality reduction. The model extracts features through two partially shared encoders E_1 and E_2 [21], through feature splicing and Bayesian inference, two feature spaces Z_1 and Z_2 are obtained, which are expressed as:

$$X_1^c = E_1(X_1), X_2^c = E_2(X_2) \tag{13}$$

$$Z_1 = X_1^c \oplus X_2^c, Z_2 = \mu_1 + \sigma_1 \times \varepsilon \tag{14}$$

In the process of reconstruction, the weighted fusion of the two feature spaces is used to obtain the reconstructed manifold space Z_r , which can be expressed as:

$$Z_{\mathrm{r}} = Z_2 + \eta Z_1 \tag{15}$$

where X_1^c and X_2^c are the output of the two parts that share the encoder. μ_1 and σ_1 are the mean and standard deviation of Z_1 , and ε is the noise sampled from the standard normal distribution. η is a weight factor of Z_1 and is set to 0.3.

In the clustering process, we use the fusion method of EM amnesia optimization strategy to obtain the clustering manifold space $Z_c = \alpha \times Z_1$ for clustering analysis.

$$\alpha = \varphi(W_\alpha \cdot [Z_c, Q] + b) \tag{16}$$

where α is the coefficient matrix of Z_1 , φ is the sigmoid activation function, and W_α and b are the learnable parameters. Q is the soft allocation weight matrix, which is calculated as follows:

$$Q = [q_{ij}]_{m\times n}, q_{ij} \geq 0, \sum_{j=1}^{n} q_{ij} = 1 \tag{17}$$

where the calculation of q_{ii} can be found in the clustering loss.

Finally, the reconstructed manifold space Z_r passes through two completely independent decoders D_1 and D_2 , and three independent fully connected layers are added after the output of each decoder to obtain the three parameters M , θ and $\prod$ of the ZINB model [17], which are defined as:

$$\hat{X}_1 = D_1(Z_r) \; ; \; \hat{X}_2 = D_2(Z_r) \tag{18}$$

$$M_R = X_1^{\wedge} = diag(s_i^r) \times \exp(w_\mu^r \hat{X}_1) \tag{19}$$

$$\theta_R = \exp(w_\theta^r \hat{X}_1) \; ; \; \Pi_R = sigmoid(w_\pi^r \hat{X}_1) \tag{20}$$

$$M_A = X_2^{\wedge} = diag(s_i^a) \times \exp(w_\mu^a \hat{X}_2) \tag{21}$$

$$\theta_A = \exp(w_\theta^a \hat{X}_2) \; ; \; \Pi_A = sigmoid(w_\pi^a \hat{X}_2) \tag{22}$$

where M_R , θ_R and Π_R are the mean, dispersion and exit probability of RNA data in ZINB loss, respectively, and M_A , θ_A and Π_A are the average, dispersion and exit probability of ADT/ATAC data in ZINB loss, respectively, M_R and M_A are also used as reconstruction data of RNA and ADT/ATAC. w_u^r , w_θ^r , w_π^r , w_u^a , w_θ^a an w_π^a is the learnable parameter for the six fully ligated layers, and s_i^r and s_i^r are the factor sizes for the RNA and ADT/ATAC data.

2.3 | Parameter configuration

The model performed 400 iterations in pre-training and

100 iterations in training, and also set an iterative convergence threshold to avoid meaningless multiple trainings. Due to the large difference in the feature dimension between ADT and ATAC, and the difference in dimensionality between different datasets, we will set different encoding and decoding structures according to different datasets. For the RNA and ADT datasets, the two partially shared encoders are set to {256,64,32,16} and {64,32,16}, respectively, and the two decoders are set to {32,64,256} and {32,64}, respectively. For the RNA and ATAC datasets, the two partially shared encoders are set to {256,64,32,16} and {256,64,32,16}, respectively, and the two decoders are set to {32,64,256} and {32,64,256}, respectively.

2.4 | Loss functio

1. ZINB loss

We use the zero-bloated negative binomial (ZINB) distribution for both RNA and ADT/ATAC data as the reconstructed loss function of the model [17]. First, define the representation of the negative binomial (NB) distribution as:

$$\begin{aligned} & NB(X_R \mid \mu_R, \theta_R) \\ & = \frac{\Gamma(X_R+\theta_R)}{X_R\Gamma(\theta_R)}\left(\frac{\theta_R}{\theta_R+\mu_R}\right)^{\theta_R}\left(\frac{\theta_R}{\theta_R+\mu_R}\right)^{X_R} \end{aligned} \tag{23}$$

$$\begin{aligned} & NB(X_A \mid \mu_A, \theta_A) \\ & = \frac{\Gamma(X_A+\theta_A)}{X_A\Gamma(\theta_A)}\left(\frac{\theta_A}{\theta_A+\mu_A}\right)^{\theta_A}\left(\frac{\theta_A}{\theta_A+\mu_A}\right)^{X_A} \end{aligned} \tag{24}$$

where μ_R , θ_R , μ_A and θ_A are the mean and dispersion of RNA and ADT/ATAC data in the *NB* distribution, respectively.

The *ZINB* distribution can be parameterized as an *NB* distribution with an additional coefficient of exit probability, which represents the weight of the mass of the zero-probability point, which can be expressed as:

$$\begin{aligned} & ZINB(X_R \mid \pi_R, \mu_R, \theta_R) \\ & = \pi_R\delta_0(X_R) + (1-\pi_R)NB(X_R \mid \mu_R, \theta_R) \end{aligned} \tag{25}$$

$$\begin{aligned} & ZINB(X_A \mid \pi_A, \mu_A, \theta_A) \\ & = \pi_A\delta_0(X_A) + (1-\pi_A)NB(X_A \mid \mu_A, \theta_A) \end{aligned} \tag{26}$$

Where π_R and π_A are exit probabilities for RNA and ADT/ATAC data, respectively.

Eventually, the two *ZINB* loss functions of this model are the refactoring loss functions, which are defined as:

$$L_R = -\log(ZINB(X_R \mid \pi_R, \mu_R, \theta_R)) \tag{27}$$

$$L_A = -\log(ZINB(X_A \mid \pi_A, \mu_A, \theta_A)) \tag{28}$$

2. Kullback-Leibler(KL) Divergence Loss

The KL divergence term plays a regularization role in the total loss function of the VAE base model, which makes the potential representation of the encoder output close to the prior distribution and improves the generalization ability [22]. Its KL loss function is expressed as:

$$\begin{aligned} L_{KL} & = D_{KL}(q(z \mid x) \| p(z)) \\ & = \frac{1}{2}(1+\log(\sigma^2)-\mu^2-\sigma^2) \end{aligned} \tag{29}$$

where μ_1 and σ_1 are the mean and standard deviation of feature space Z_1 .

3. Clustering loss

We use a weighted soft-allocated K-means clustering method [23] to cluster clustered manifold space Z_c , which can generate smooth cluster assignment weights instead of hard assignments to a single cluster center. This can improve the robustness and clustering effect of the model to noise, and most importantly, it can provide clustering guidance for EM heuristic optimization strategies, iterate on the basis of forgetting gates to optimize the clustered manifold space, improve the representation ability, and retain internal information. In this process, the scaled squared Euclidean distance between the eigenvector z_i of Z_c and each of the centricities v_i can be expressed as :

$$d_{ij} = \|z_i - v_j\|^2 \tag{30}$$

The mean of each row is then subtracted from that row to calculate its normalized distance matrix:

$$td_{ij} = d_{ij} - \frac{1}{K}\sum_{j=1}^{K} d_{ij} \tag{31}$$

The normalized soft-allocation weight q_{ij} is calculated using exponential operation and double normalization, and from this weight the soft-assigned weights matrix Q mentioned above is formed:

$$q_{ij} = \frac{\exp(-td_{ij})}{\sum_{j=1}^{K}\exp(-td_{ij})} \tag{32}$$

$$q_{ij} = \frac{{q_{ij}}^2}{\sum_{j=1}^{K}{q_{ij}}^2} \tag{33}$$

Finally, the weighted soft classification K-means clustering loss is calculated using the weighted squared Euclidean distance of d_{ij} , as follows:

$$L_{\text{wsc}K-means} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} d_{ij}q_{ij} \tag{34}$$

4. Total loss

By calculating the sum of two reconstructed losses for RNA and ADT/ATAC data, one KL loss and one clustering loss, the final overall loss is expressed as:

$$L_{total} = L_R + L_A + \alpha L_{KL} + \beta L_{wscK-means} \tag{35}$$

where the weight α of KL loss and the weight of cluster loss β are set at 0.01 and 0.001, respectively, and the clustering loss ($\beta = 0$) is not used in the pre-training stage, and the KL loss is not used in the first half of the pre-training ($\alpha = 0$).

2.5 | Evaluate metrics

1. Adjusted Rand Index (ARI)

ARI is a measure of the consistency between clustering results and true labels [24], ranging from -1 to 1, with closer to 1 indicating better clustering. We use the ARI value calculated by the adjusted_rand_score function in the sklearn library, which is defined as:

$$ARI = \frac{(RI - \mathrm{E}[RI])}{\max(RI) - \mathrm{E}[RI]} \tag{36}$$

where RI is the Rand Index, which indicates the proportion of sample pairs whose clustering results are consistent with the real label. $\mathrm{E}[RI]$ is the expected value of the Rand index in the case of random clustering, and $\max(RI)$ is the maximum possible Rand index.

2. Normalized Mutual Information (NMI)

NMI is normalized by dividing the mutual information by the average information entropy (or maximum information entropy) [25], so that the range is [0,1]. NMI is expressed as:

$$NMI(U,V) = \frac{2MI(U,V)}{H(U)+H(V)} \tag{37}$$

where $H(U)$ and $H(V)$ are denoted as the entropy of the predicted label U and the true label V, respectively, representing their respective uncertainties. $MI(U,V)$ is mutual information.

3. Adjusted Mutual Information (AMI)

AMI is the result of adjusting for Mutual Information (MI), which takes into account the impact of randomness on mutual information [26]. The value range of AMI is $[-1,1]$, and the specific formula is expressed as:

$$AMI(U,V) = \frac{MI(U,V) - \mathrm{E}[MI(U,V)]}{\max(H(U),H(V)) - \mathrm{E}[MI(U,V)]} \tag{38}$$

whereis the mathematical expectation of mutual information, $\mathrm{E}[MI(U,V)]$ which is used to remove the influence of random factors on mutual information.

3 | Results and Discussion

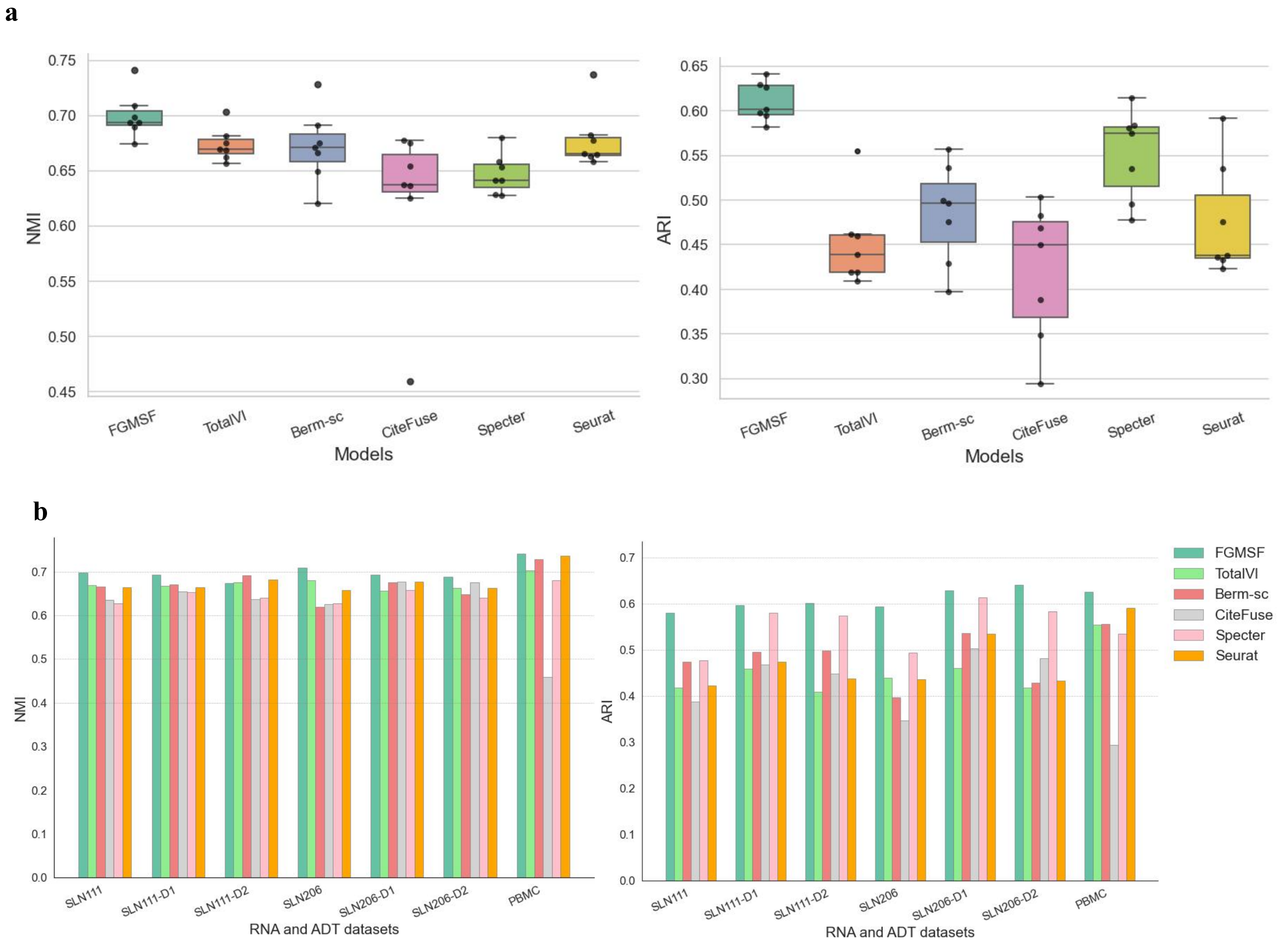


Fig.2 FGMSF was compared with the other 5 methods, boxplot (a), histogram (b). This reflects that FGMSF has good stability and excellent clustering ability, and the ARI index has been greatly improved, indicating that the intra-class clustering ability has been further improved.

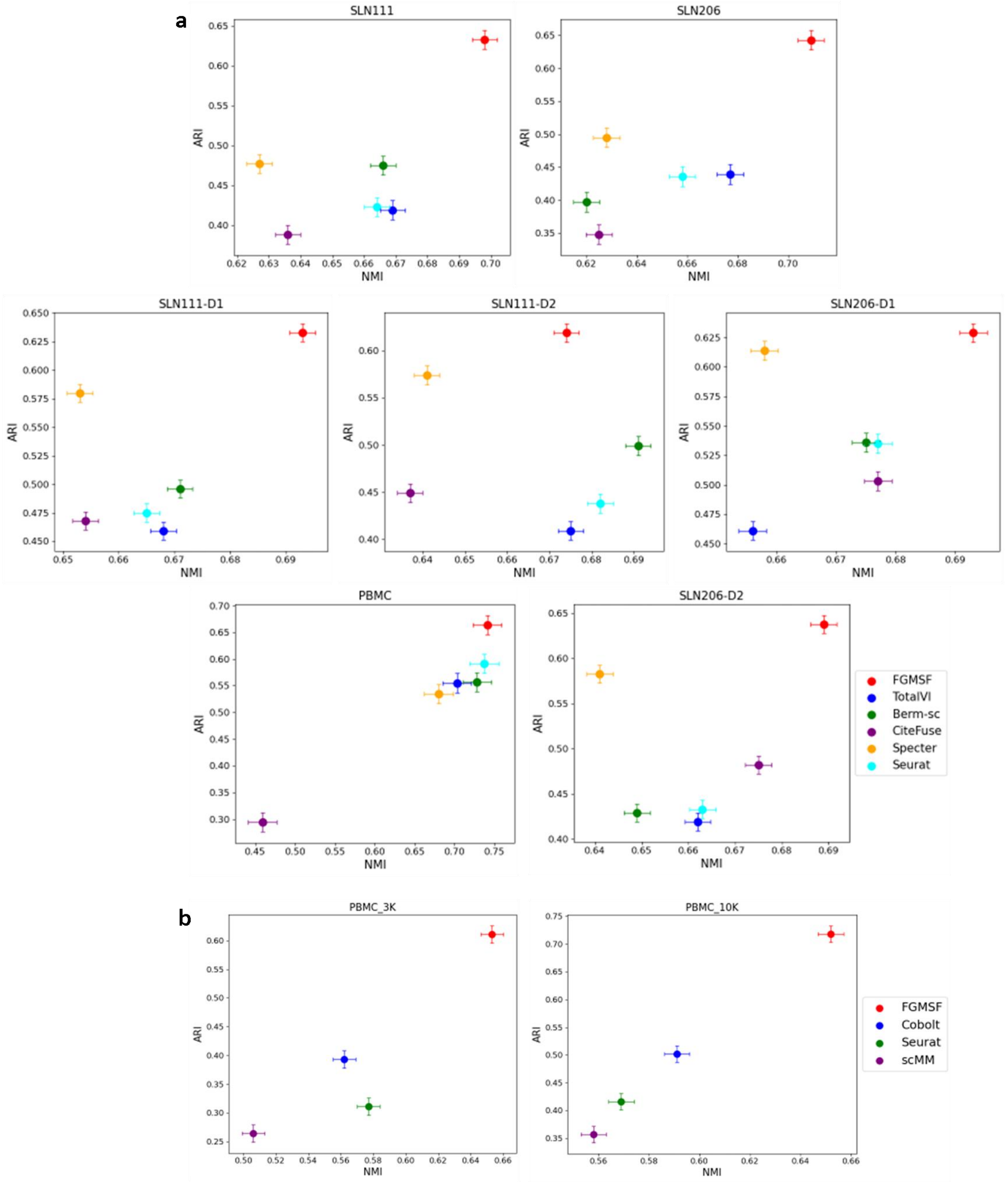


Fig.3 Model comparison dot plot on RNA and ADT datasets (a), model comparison dot plot on RNA and ATAC datasets(b).

3.1 | Comparison with different models

We evaluated the clustering performance of FGMSF on 9 datasets using ARI and NMI clustering indicators, and for RNA and ADT datasets, there were 5 single-batch datasets (SLN111-D1, SLN111-D2, SLN206-D1, SLN206-2, PBMC) and 2 two-batch datasets (SLN111, SLN206), and 5 single-cell multiomics models (BREM-SC, TotalVI, CiteFuse, Seurat, Sputter). For both RNA and ATAC datasets, they were single-batch datasets and were compared to three single-cell multiomics models (Seurat, Cobolt, scMM). In Figure 2 and 3, we use boxplots, histograms, and dot plots to demonstrate the performance of FGMSF on clustering. Specifically, FGMSF has significant advantages in both stability and clustering performance, particularly in terms of ARI metrics.

As shown in the box plot in Figure 3a, the height of the FGMSF box is the highest in both indicators, indicating that there is a high level of accuracy on different datasets.Moreover, the box is also the narrowest, indicating that the clustering results on different datasets have good stability.

Combined with Figure 2b, we can see that in the BREM-SC, CiteFuse and Seurat models, there are abnormal clustering results on a certain dataset, and the stability of all the comparison model methods on the ARI index is not good, indicating that their models only pay attention to the global effect and the inter-class clustering effect is stable, but do not pay attention to the local effect, that is, the intra-class clustering performance is not very good. However, FGMSF not only greatly improves the ARI index, but also has a certain stability, which truly improves the ability of intra-class clustering. Although the clustering performance improvement on the NMI index is small, it is basically the best result, except for a little deficiency on the SLN111-D2 datase.

In Figure 3, we give the model comparison dot plots of RNA and ADT datasets and RNA and ATAC datasets, respectively, and we can intuitively see the results of simultaneous comparison of ARI and NMI indicators, and we can see that the clustering performance of FGMSF is still outstanding.

3.2 | Ablation experiments

We believe that relying solely on Bayesian inference has certain limitations in data reconstruction, and it is difficult to completely retain the original cell information and its internal connections, which loses the details and potential information of the original data to a certain extent. Therefore, this paper does not adopt the method of variational processing of the two codes, but chooses to retain the detailed features of the original data through feature stitching. Subsequently, on this basis, feature space fusion is introduced to integrate multiple information and construct a more biologically significant manifold space, which fully combines the ability of Bayesian inference to reconstruct the data and the detailed information after the original data is encoded. In order to maximize the information flow of the bottleneck layer, we construct independent manifold spaces for decoding reconstruction and dimensionality reduction clustering. Specifically, the decoding and reconstruction process has a higher dependence on the details of the original information retained by the encoder and the global view provided by variational inference [27] to ensure the completeness and accuracy of data reconstruction. However, in the process of dimensionality reduction clustering, more emphasis is placed on extracting representative and discriminant features to improve the clustering effect and reveal the potential connections between cells [28].

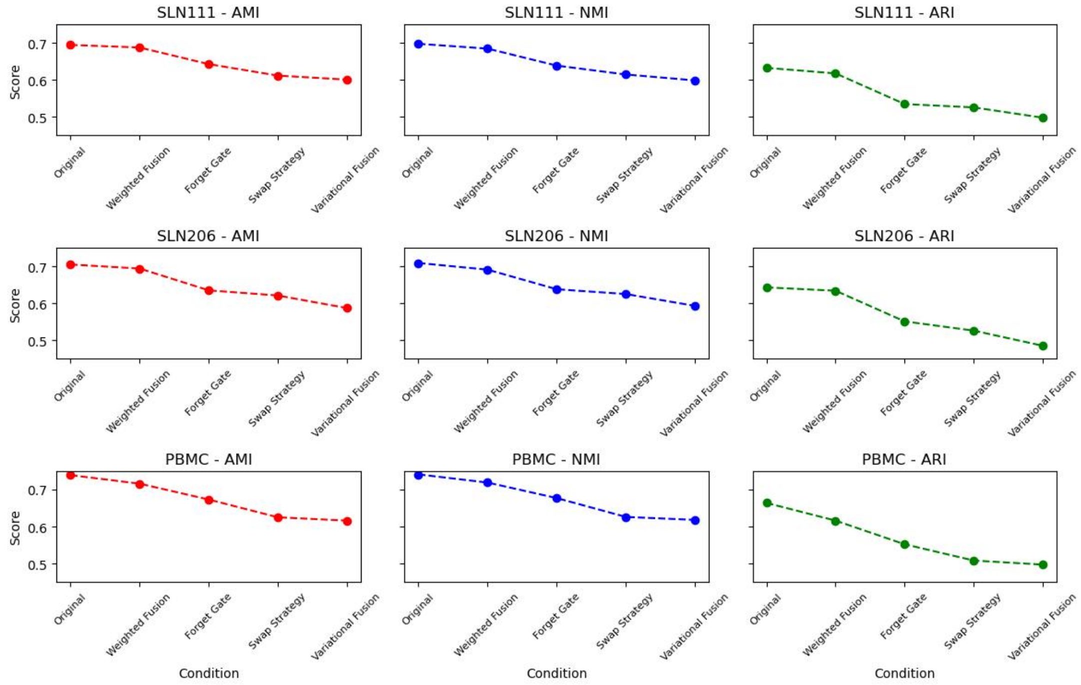


Fig.4 Ablation experiments of 5 modes on the SLN206, SLN111, and PBMC datasets.

In order to verify the two-stage fusion proposed in this paper and the two fusion strategies in different processes, ablation experiments are carried out on the two designs to verify their improvement on the performance of the model. Firstly, in order to verify that it is indeed necessary to design different fusion strategies for reconstruction and clustering, construct an independent manifold space, and design two ablation models and use one of the fusion strategies for verification. Secondly, in order to verify that the two different fusion strategies are specially designed for different processes, and to further prove the correctness of the previous theoretical reasoning, a model with the two fusion strategies interchangeable is designed for verification. Finally, in order to verify the importance of fusing the original detail information into the manifold space, a model with variational processing after coding was designed for verification. In this way, a total of 5 types of comparison were generated, namely the original model, the model with only weighted fusion, the model with only forgotten gate fusion, the fusion strategy swap model, and the model with direct variational fusion.

The results of the ablation experiment are shown in Figure 4, which only shows the ablation results of some of the datasets (the ablation experiments for the remaining datasets can be found in Supplementary

Material Figure 1). It is evident that only the original model's indicator is the highest in each of the different indicators of the dataset, which shows that our inference is correct. The model needs to be fused in two stages, and the feature details are included to make the manifold space have a better representation, and it can also be shown that if you want to achieve the double improvement of reconstruction and clustering, you need independent manifold space to support. Finally, we can also observe a significant performance improvement on the ARI metric, which not only illustrates the effectiveness of the FGMSF model for the intra-class clustering of the dataset in this paper, but also shows the effectiveness of the model method improvement.

3.3 | Clustering & labeling

As shown in Figure 5, the clustering results of the FGMSF model on the SLN206 dataset are shown. Through UMAP dimensionality reduction visualization and cell type annotation, it can be observed that the model can not only achieve effective cell classification, but also retain the intrinsic connection between data to a certain extent. This indicates that the FGMSF model can balance the clustering accuracy and biological correlation well (the UMAP clustering results of the remaining datasets are shown in Supplementary Material Figure 2).

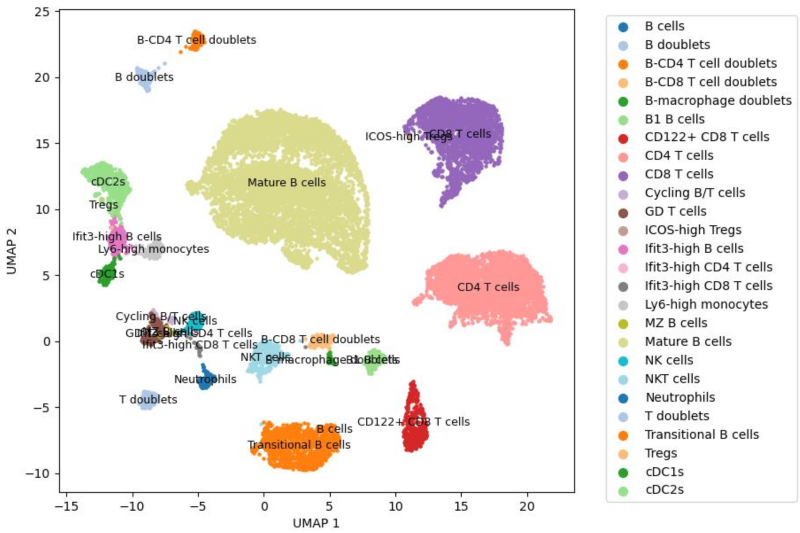


Fig.5 The clustering and labeling on the SLN206 dataset are clear, and there is a certain relationship between the clusters.

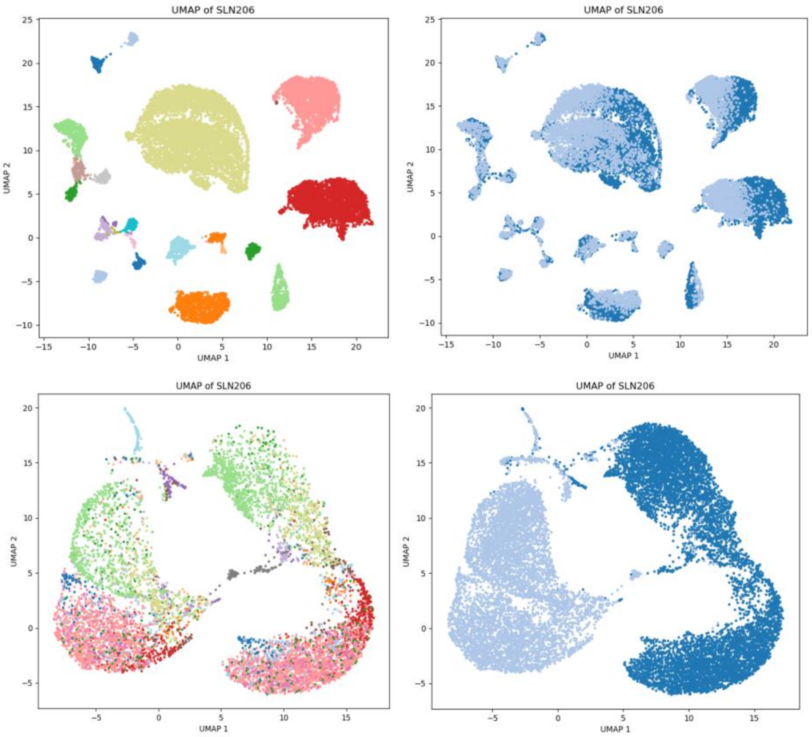


Fig.6 Debatching task on the SLN206 dataset. The first row is the clustered manifold space of the FGMSF model, the second row is the raw data, the first column is the clustering results, and the second column is the batch results.

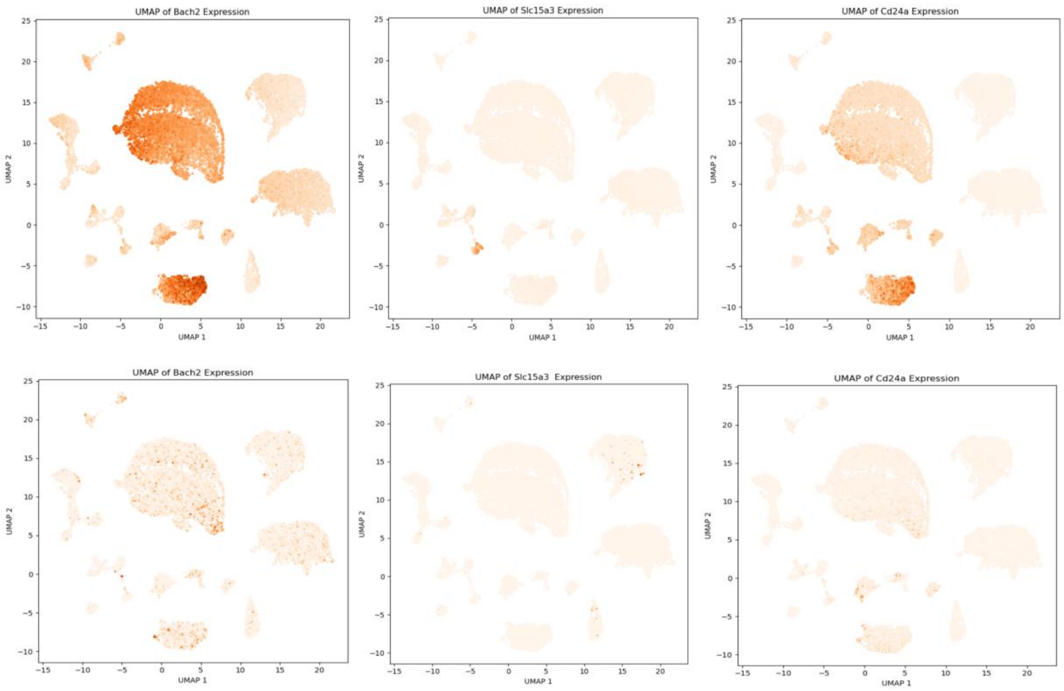


Fig.7 Comparison of marker genes on the SLN206 dataset. The first row reconstructs the data, and the second row is the raw data, with each column being a different marker gene.

4 | Downstream analysis

4.1 | De-batch effect

We assessed the debatching effect capacity of FGMSF, as shown in Figure 6. As can be seen from the comparison chart, the FGMSF model significantly reduces the impact of batch effect while achieving high-quality clustering. This indicates that the model is highly robust and adaptable in terms of cross-batch integration, and the clustered manifold space constructed by it has high expressive ability (the clustering, cell type annotation, and debate results of the FGMSF model on the SLN111 dataset are presented in Supplementary Material Figure 3).

4.2 | Marker gene identification

In Figure 7, by comparing the performance of the reconstructed data with the original data, we can observe that the reconstructed data can more significantly express the marker genes of different cell clusters, while the original data performs poorly in characterizing the marker genes.To a large extent, this phenomenon indicates that the data reconstruction process improves the quality of the data and allows for a more accurate representation of the biological differences between cell clusters. This improvement may be due to the fact that the refactoring method effectively reduces technical noise and enhances the biological signal, thus providing more reliable underlying data support for downstream analysis [29]. This result further validates the importance of data reconstruction in single-cell transcriptome analysis, especially in the potential applications of cell type identification and functional elucidation (the labeled protein map of the SLN206 dataset is shown in Supplementary Material Figure 4).

The following is an explanation and analysis of the mark gene to illustrate its validity in biological significance. The results showed that Bach2 gene was highly expressed in both mature B cells and transition B cells, which was usually related to its important role in B cell differentiation and function regulation. Bach2 is a transcriptional repressor that balances the proliferation and differentiation of B cells by regulating the expression of key genes during antibody production. In Mature B cells, it further regulates antibody production and humoral immune responses, allowing cells to respond appropriately when confronted with antigens [30, 31]. SLC15A3 acts as a proton-coupled oligopeptide transporter and is involved in the regulation of signaling pathways in immune cells, particularly in inflammatory responses. Neutrophils require a high degree of activity and responsiveness in response to infection and inflammation, so more SLC15A3 may be needed to support their function. Studies have shown that SLC15A3 function in endolysosomes and is associated with the conduction of inflammatory signals [32]. And neutrophils show SLC15A3 high expression levels at homeostasis. CD24a is an important marker molecule during B cell development, and its expression changes dynamically at different stages of B cell differentiation. CD24a is typically expressed higher in transitional B cells, a critical transition period for the differentiation of immature B cells into mature B cells. Studies have shown that CD24a expression is higher in the early stages of B cell maturation [33], then gradually decreases as it differentiates into mature antibody-producing cells .

4.3 | Genetic difference analysis

The volcano map simultaneously displays the expression changes of each gene, allowing for rapid identification of important differentially expressed genes. Figure 8 (a, b) shows the analysis of multiple genetic differences between mature B cells and CD4 T cells, CD8 T cells and transitional B cells. And there are many proven findings [34-36], such as Setbp1, Pik3ap1, Slc15a3, and Klhl14 being highly expressed in MatureB cells, and Cd40lg and Cd28 being highly expressed in CD4T cells. Prkcq, Sla2, Cd7, and Nkg7 are highly expressed in CD8 T cells, while Vpreb3, Eaf2, and Iglc1 are highly expressed in transitional B cells.

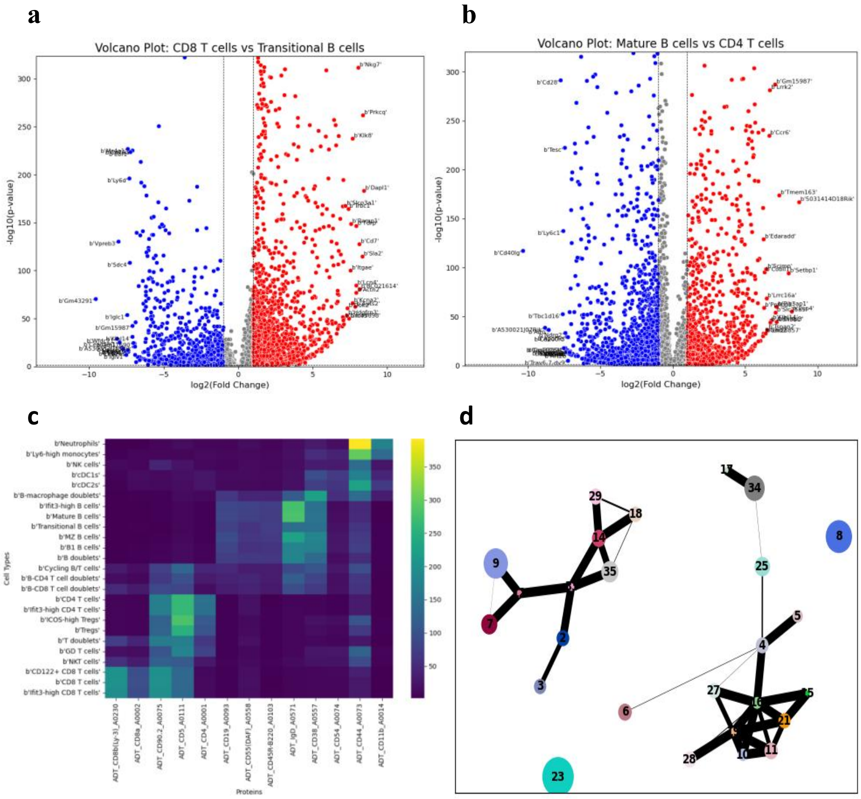


Fig.8 The volcano plot (a, b) shows the gene differences between the different clusters, the protein heat map (c) shows the protein expression differences between the different clusters, and the cell cluster trajectory map (d) shows the potential connections between the different clusters.

4.4 | Protein differences analysis

The protein heat map is shown in Figure 8c, where we found that different B cell types were clustered together and shared multiple significantly enriched proteins. ADT_CD44_A0073 is enriched in neutrophil granules, and this protein plays an important role in neutrophil chemotaxis and aggregation. ADT_lgD_A0571 exhibited the highest enrichment fraction in both Transitional B cells and Mature B cells, reflecting its key function during B cell maturation and differentiation. ADT_CD5_A0111 is significantly enriched in Tregs, ICOS-high Tregs, lfit3-high CD4 T cells, CD4 T cells, and its role in T cell signaling and immune regulation has been extensively studied and confirmed, and the expression level of CD5 in Tregs is related to its regulatory properties [37], especially in highly autoreactive Tregs.

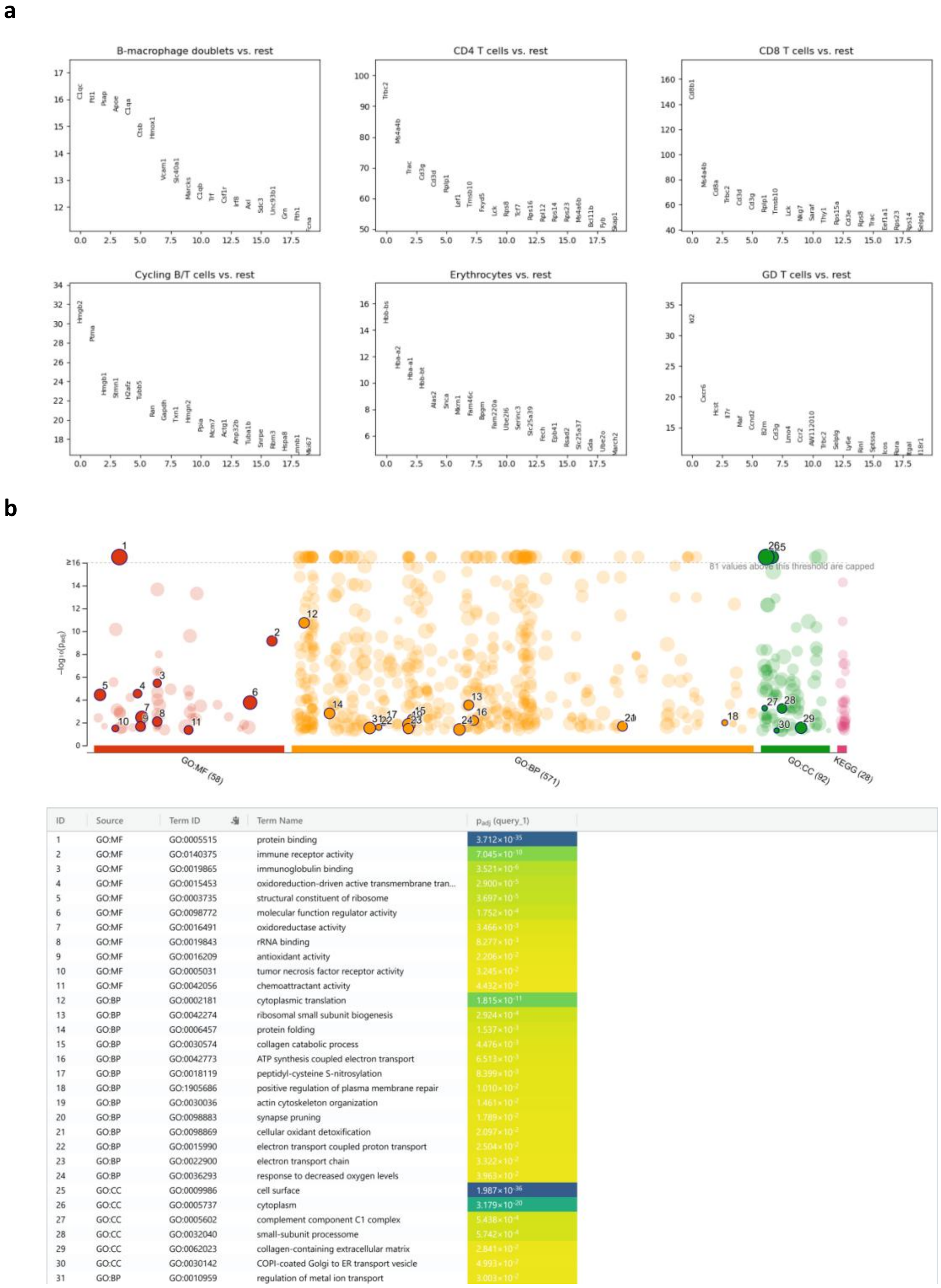
These analysis results further support the protein expression characteristics in each cell cluster and provide an important basis for understanding the function of immune cells. We also found that some subtype cell clusters of the same type had identical protein expression between them (the corresponding protein heatmap is shown in Figure 5 of the material), which made the ADT data not useful in post-fusion clustering and could interfere with model clustering. This is also the reason why the model does not have a significant improvement effect on NMI indicators, and cannot better distinguish different cell subtypes, resulting in an insignificant overall clustering effect.

4.5 | Trajectory analysis

Figure 8d shows the trajectory of the cell clusters, showing the connections between cells including regulatory T cells, natural killer T cells, marginal band (MZ) B cells, and GD T cells. We have found many proven connections for a variety of cell clusters, which also shows that the latent space not only represents the differences between cell clusters, but also retains the intrinsic connections between clusters. Specifically, between NKT cells29 and B-CD8 T3 cells, CD8+ NKT cells express both the T cell markers TCRβ and CD3, as well as the NK cell receptors CD49b and NKG2D. However, there is also a link between cD122+ CD8 T cells7, ICOS-high Tregs cells12, and CD8 T 9 cells, and naturally occurring CD8(+) CD122(+) T cells are also Tregs cells with the ability to suppress T cell responses and suppress autoimmune and allogeneic immunity [38]. The link between Ifit3-high B14 cells and Tregs33 cells is more likely to show that Ifit3-high B cells can regulate the PI3K/AKT signaling pathway by binding to the JAK-PI3K signaling pathway, which in turn affects the expression of IFNγ and the stability of FOXP3 in Treg cells.

4.6 | Enrichment analysis

The differentially expressed gene sequences for some of the cell clusters are shown in Figure 10a (the differentially expressed gene sequences of all cell clusters are shown in Supplementary Material Figure 6). For these differentially expressed genes, we performed a gene ontology (GO) enrichment analysis and demonstrated the enriched biological pathways in Figure 10b.



ID	Source	Term ID	Term Name	p_{adj} (query_1)
1	GO:MF	GO:0005515	protein binding	3.712×10^{-35}
2	GO:MF	GO:0140375	immune receptor activity	7.045×10^{-10}
3	GO:MF	GO:0019865	immunoglobulin binding	3.521×10^{-6}
4	GO:MF	GO:0015453	oxidoreduction-driven active transmembrane tran...	2.900×10^{-5}
5	GO:MF	GO:0003735	structural constituent of ribosome	3.697×10^{-5}
6	GO:MF	GO:0098772	molecular function regulator activity	1.752×10^{-4}
7	GO:MF	GO:0016491	oxidoreductase activity	3.466×10^{-3}
8	GO:MF	GO:0019843	rRNA binding	8.277×10^{-3}
9	GO:MF	GO:0016209	antioxidant activity	2.206×10^{-2}
10	GO:MF	GO:0005031	tumor necrosis factor receptor activity	3.245×10^{-2}
11	GO:MF	GO:0042056	chemoattractant activity	4.432×10^{-2}
12	GO:BP	GO:0002181	cytoplasmic translation	1.815×10^{-11}
13	GO:BP	GO:0042274	ribosomal small subunit biogenesis	2.924×10^{-4}
14	GO:BP	GO:0006457	protein folding	1.537×10^{-3}
15	GO:BP	GO:0030574	collagen catabolic process	4.476×10^{-3}
16	GO:BP	GO:0042773	ATP synthesis coupled electron transport	6.513×10^{-3}
17	GO:BP	GO:0018119	peptidyl-cysteine S-nitrosylation	8.399×10^{-3}
18	GO:BP	GO:1905686	positive regulation of plasma membrane repair	1.010×10^{-2}
19	GO:BP	GO:0030036	actin cytoskeleton organization	1.461×10^{-2}
20	GO:BP	GO:0098883	synapse pruning	1.789×10^{-2}
21	GO:BP	GO:0098869	cellular oxidant detoxification	2.097×10^{-2}
22	GO:BP	GO:0015990	electron transport coupled proton transport	2.504×10^{-2}
23	GO:BP	GO:0022900	electron transport chain	3.322×10^{-2}
24	GO:BP	GO:0036293	response to decreased oxygen levels	3.963×10^{-2}
25	GO:CC	GO:0009986	cell surface	1.987×10^{-36}
26	GO:CC	GO:0005737	cytoplasm	3.179×10^{-20}
27	GO:CC	GO:0005602	complement component C1 complex	5.438×10^{-4}
28	GO:CC	GO:0032040	small-subunit processome	5.742×10^{-4}
29	GO:CC	GO:0062023	collagen-containing extracellular matrix	2.841×10^{-2}
30	GO:CC	GO:0030142	COPI-coated Golgi to ER transport vesicle	4.993×10^{-2}
31	GO:BP	GO:0010959	regulation of metal ion transport	3.003×10^{-2}

Fig.9 Gene sequences of different clusters (a), GO enrichment analysis (b).

The significantly enriched entries were mainly concentrated in GO:MF and GO:BP, and in the GO:MF category, the significantly enriched entries were mainly concentrated in GO:MF and GO:BP, and in the GO:MF category, the significantly enriched items included protein binding (GO:0005515), immune receptor activity (GO:0041075), antioxidant activity (GO:0016209), etc., indicating that there were more active molecular interactions and immune-related processes in the samples. In the GO:BP category, significant enrichment of items such as ribosomal subunit generation (GO:0042254), protein folding (GO:0006457), and oxidative phosphorylation (GO:0006119) suggests that the samples are involved in more active protein synthesis, folding, and energy metabolism. In addition, the spatial localization of these processes is further supported by significant enrichment of items in the GO:CC category, such as cytoplasm (GO:0005737) and cell surface (GO:0009986).

These results together reveal the possible existence of important biological processes such as immune regulation, protein homeostasis, and antioxidant defense in the samples, and provide a basis for further study of their underlying biological mechanisms [39].

5 | Conclusions

This article proposes a multi-stage fusion clustering model FGMSF based on forgetting gates, which adopts an end-to-end form and constructs fitting data with manifold structure using manifold fitting. Deep clustering analysis is conducted on single-cell multi omics datasets. This model proposes a two-stage fusion approach that preserves the detailed features of the original data as much as possible by fusing the feature space of the bottleneck layer, and adopts different fusion strategies for the reconstruction and clustering processes in the second stage fusion. In addition, we have specifically designed an EM heuristic optimization strategy based on the forget gate structure, which removes redundant information, maximizes bottleneck layer information flow, and deeply explores effective representation information in the manifold space, significantly improving clustering ability while ensuring the manifold. The FGMSF model not only improves the quality of reconstructed data while enhancing clustering performance, but also meets the multiple requirements of single-cell clustering models for spatial alignment, de batch effects, and preserving the connections between data.

In the downstream analysis, we verified the significant improvement of the model in terms of clustering effect and reconstructed data quality from multiple dimensions through a variety of tasks. Specifically, in the evaluation of the clustering effect, we found that the three clustering indicators showed significant improvement in both comparison with other models and performance in ablation experiments, especially in ablation experiments, where the advantages of secondary fusion technology and amnesia gate fusion strategy were clearly verified. In the UMAP clustering plot and the debatching effect plot, the clustering ability was further demonstrated. In addition, the cell cluster trajectory plot also validates the model's superior ability to preserve the intrinsic connections of the data. In the assessment of the quality of the reconstructed data, we demonstrated that the model was able to better reconstruct the data by graphically representing the expression of the marker genes, especially highlighting the differences in the expression of the marker genes in different clusters. The results of the volcano map and the protein heat map also provide further evidence of the biological validity and accuracy of the reconstructed data. Finally, enrichment analysis was conducted by obtaining differential gene sequences, and important biological processes were identified, providing more possibilities for biomedical analysis.

6 | Comparison methods

Since the dominant models of different dataset combinations are different, we looked for a total of 7 representative groups of models to be compared with two different datasets, including 5 groups of participants for the comparison of RNA and ADT datasets, and 3 groups for comparison of RNA and ATACdatasets.BREM-SC(https://github.com/tarot0410/BREMSC), CiteFuse(https://github.com/SydneyBioX/CiteFuse), Seurat(v4.0.4,https://github.com/satijalab/seurat), TotalVI(https://github.com/YosefLab/totalVI_reproducibility),Cobolt(https://github.cm/epurdom/cobolt),scMM(https://github.com/kodaim1115/scMM)andSpecter(https://github.com/canzarlab/Specter) are used as competing methods.

References

1. Mereu, E., Lafzi, A., Moutinho, C., Ziegenhain, C., McCarthy, D. J., Álvarez-Varela, A., ... & Heyn, H. (2020). Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. Nature biotechnology, 38(6), 747-755. https://doi.org/10.1101/630087
2. Wani, S. A., & Quadri, S. M. K. (2023). Evaluation of computational methods for single cell multi-omics integration. Procedia Computer Science, 218, 2744-2754. https://doi.org/10.1016/j.procs.2023.01.246
3. Feng, X., Zhang, H., Lin, H., & Long, H. (2023). Single-cell RNA-seq data analysis based on directed graph neural network. Methods, 211, 48-60. https://doi.org/10.1016/j.ymeth.2023.02.008
4. Meilă, M., & Zhang, H. (2024). Manifold learning: What, how, and why. Annual Review of Statistics and Its Application, 11(1), 393-417. https://doi.org/10.1146/annurev-statistics-040522-115238
5. Golchin, E., & Maghooli, K. (2014). Overview of manifold learning and its application in medical data set. International journal of biomedical engineering and science (IJBES), 1(2). https://airccse.com/ijbes/papers/1214ijbes03.pdf%E3%80%82

6. Nguyen, N. D., Blaby, I. K., & Wang, D. (2019). ManiNetCluster: a novel manifold learning approach to reveal the functional links between gene networks. BMC genomics, 20, 1-14. https://doi.org/10.1186/s12864-019-6329-2
7. Wang, X., Sun, Z., Zhang, Y., Xu, Z., Xin, H., Huang, H., ... & Chen, W. (2020). BREM-SC: a bayesian random effects mixture model for joint clustering single cell multi-omics data. Nucleic acids research, 48(11), 5814-5824. https://doi.org/10.1093/nar/gkaa314
8. Gayoso, A., Steier, Z., Lopez, R., Regier, J., Nazor, K. L., Streets, A., & Yosef, N. (2021). Joint probabilistic modeling of single-cell multi-omic data with totalVI. Nature methods, 18(3), 272-282. https://doi.org/10.5281/zenodo.4330368
9. Kim, H. J., Lin, Y., Geddes, T. A., Yang, J. Y. H., & Yang, P. (2020). CiteFuse enables multi-modal analysis of CITE-seq data. Bioinformatics, 36(14), 4137-4143. 10.1093/bioinformatics/btaa282
10. Butler, A., Hoffman, P., Smibert, P., Papalexi, E., & Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nature biotechnology, 36(5), 411-420. https://doi.org/10.1038/nbt.4096
11. Ringeling, F. R., & Canzar, S. (2021). Linear-time cluster ensembles of large-scale single-cell RNA-seq and multimodal data. Genome research, 31(4), 677-688. https://www.genome.org/cgi/doi/10.1101/gr.267906.120
12. Gong, B., Zhou, Y., & Purdom, E. (2021). Cobolt: integrative analysis of multimodal single-cell sequencing data. Genome biology, 22, 1-21. https://doi.org/10.1186/s13059-021-02556-z
13. Minoura, K., Abe, K., Nam, H., Nishikawa, H., & Shimamura, T. (2021). A mixture-of-experts deep generative model for integrated analysis of single-cell multiomics data. Cell reports methods, 1(5). https://doi.org/10.5281/zenodo.5149733
14. Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D. J., Hicks, S. C., Robinson, M. D., ... & Schönhuth, A. (2020). Eleven grand challenges in single-cell data science. Genome biology, 21, 1-35. https://doi.org/10.1186/s13059-020-1926-6
15. Qiu, P. (2020). Embracing the dropouts in single-cell RNA-seq analysis. Nature communications, 11(1), 1169. https://doi.org/10.1038/s41467-020-14976-9
16. Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S., & Vert, J. P. (2018). A general and flexible method for signal extraction from single-cell RNA-seq data. Nature communications, 9(1), 284. https://doi.org/10.1038/s4146-017-02554-5
17. Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S.,& Theis, F. J. (2019). Single-cell RNA-seq denoising using a deep count autoencoder. Nature communications, 10(1), 390. https://doi.org/10.1038/s41467-018-07931-2
18. Yao, Z., Li, B., Lu, Y., & Yau, S. T. (2024). Single-cell analysis via manifold fitting: A framework for RNA clustering and beyond. Proceedings of the National Academy of Sciences, 121(37), e2400002121. https://doi.org/10.1073/pnas.2400002121
19. Yao, Z., Su, J., Li, B., & Yau, S. T. (2023). Manifold Fitting. arXiv preprint arXiv:2304.07680. https://arxiv.org/abs/2304
20. Kingma, D. P., & Welling, M. (2013, December). Auto-encoding variational bayes. https://arxiv.org/abs/1312.6114
21. He, Z., Hu, S., Chen, Y., An, S., Zhou, J., Liu, R., ...& Ying, X. (2024). Mosaic integration and knowledge transfer of single-cell multimodal data with MIDAS. Nature biotechnology, 42(10), 1594-1605. https://doi.org/10.1038/s41587-023-02040-y
22. Pereyra, G., Tucker, G., Chorowski, J., Kaiser, Ł., & Hinton, G. (2017). Regularizing neural networks by penalizing confident output distributions. arXiv preprint arXiv:1701.06548. https://doi.org/10.48550/arXiv.1701.06548
23. Zhu, B., Bedeer, E., Nguyen, H. H., Barton, R., & Henry, J. (2020). Improved soft-k-means clustering algorithm for balancing energy consumption in wireless sensor networks. IEEE Internet of Things Journal, 8(6), 4868-4881. https://doi.org/10.48550/arXiv.1701.06548
24. Steinley, D., Brusco, M. J., & Hubert, L. (2016). Thevariance of the adjusted Rand index. Psychological methods, 21(2), 261. https://doi.org/10.1037/met0000049
25. Mahmoudi, A., & Jemielniak, D. (2024). Proof of biased behavior of Normalized Mutual Information. Scientific Reports, 14(1), 9021. https://doi.org/10.1038/s41598-024-59073-9
26. Schütze, H., Manning, C. D., & Raghavan, P. (2008). Introduction to information retrieval (Vol. 39, pp. 234-265). Cambridge: Cambridge University Press. https://doi.org/10.1038/s41598-024-59073-9
27. Alemi, A. A., Fischer, I., Dillon, J. V., & Murphy, K. (2016). Deep variational information bottleneck. arXiv preprint arXiv:1612.00410. https://arxiv.org/abs/1612.00410
28. Jiang, Z., Zheng, Y., Tan, H., Tang, B., & Zhou, H. (2016). Variational deep embedding: An unsupervised and generative approach to clustering. arXiv preprint arXiv:1611.05148. https://doi.org/10.48550/arXiv.1611.05148
29. Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., ... & Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. science, 286(5439), 531-537. https://doi.org/10.1126/science.286.5439.531
30. Muto, A., Ochiai, K., Kimura, Y., Itoh-Nakadai, A., Calame, K. L., Ikebe, D., ... & Igarashi, K. (2010). Bach2 represses plasma cell gene regulatory network in B cells to promote antibody class switch. The EMBO journal, 29(23), 4048-4061. https://doi.org/10.1038/emboj.2010.257
31. Hu, Q., Xu, T., Zhang, W., & Huang, C. (2022). Bach2 regulates B cell survival to maintain germinal centers and promote B cell memory. Biochemical and biophysical research communications, 618, 86-92. https://doi.org/10.1016/j.bbrc.2022.06.009
32. Yu, S., Yang, J., Zhang, R., Guo, Q., & Wang, L. (2024). SLC15A3 is transcriptionally regulated by HIF1α and p65 to worsen neuroinflammation in experimental ischemic stroke. Molecular Neurobiology, 1-16. https://doi.org/10.1007/s12035-024-04191-8
33. Mensah, F. F., Armstrong, C. W., Reddy, V., Bansal, A. S., Berkovitz, S., Leandro, M. J., & Cambridge, G. (2018). CD24 expression and B cell maturation shows a novel link with energy metabolism: potential implications for patients with myalgic encephalomyelitis/chronic fatigue syndrome. Frontiers in immunology, 9, 2421. https://doi.org/10.3389/fimmu.2018.02421
34. Troutman, T. D. (2014). B-Cell Adapter for Phosphoinositide 3-Kinase Is a Signaling Adapter in the Toll-Like Receptor/Interleukin-1 Receptor Superfamily (Doctoral dissertation). https://hdl.handle.net/2152.5/3328
35. Ham, H., Hirdler, J. B., Bihnam, D. T., Mao, Z., Gicobi, J. K., Macedo, B. G., ... & Billadeau, D. D. (2025). Lysosomal NKG7 restrains mTORC1 activity to promote CD8+ T cell durability and tumor control. Nature Communications, 16(1), 1628. https://doi.org/10.1038/s41467-025-56931-6
36. Lelliott, E. J., Ramsbottom, K. M., Dowling, M. R., Shembrey, C., Noori, T., Kearney, C. J., ... & Oliaro, J. (2022). NKG7 enhances CD8+ T cell synapse efficiency to limit inflammation. Frontiers in immunology, 13, 931630. https://doi.org/10.3389/fimmu.2022.931630
37. Tung, J. W., Kunnavatana, S. S., Herzenberg, L. A., & Herzenberg, L. A. (2001). The regulation of CD5 expression in murine T cells. BMC molecular biology, 2, 1-13. https://doi.org/10.1186/1471-2199-2-5
38. Liu, J., Chen, D., Nie, G. D., & Dai, Z. (2015). CD8+

CD122+ T-cells: a newly emerging regulator with central memory cell phenotypes. Frontiers in immunology, 6, 494. https://doi.org/10.3389/fimmu.2015.00494

39. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., ... & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. Nature genetics, 25(1), 25-29. https://doi.org/10.1038/75556