JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

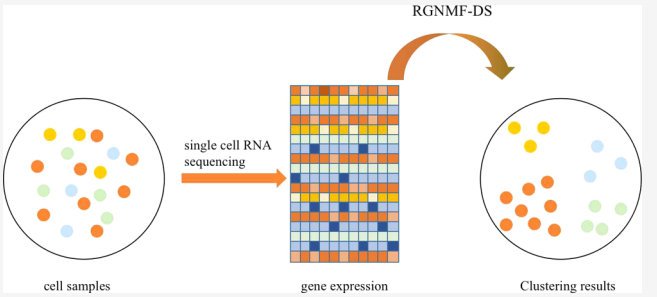
Robust Graph Regularized NMF with Dissimilarity and Similarity Constraints for ScRNA-seq Data Clustering

Zhenqiu Shu,* Qinghan Long, Luping Zhang,* Zhengtao Yu, and Xiao-Jun Wu

Cite This: J. Chem. Inf. Model. 2022, 62, 6271−6286 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The notable progress in single-cell RNA sequencing (ScRNA-seq) technology is beneficial to accurately discover the heterogeneity and diversity of cells. Clustering is an extremely important step during the ScRNA-seq data analysis. However, it cannot achieve satisfactory performances by directly clustering ScRNA-seq data due to its high dimensionality and noise. To address these issues, we propose a novel ScRNA-seq data representation model, termed Robust Graph regularized Non-Negative Matrix Factorization with Dissimilarity and Similarity constraints (RGNMF-DS), for ScRNA-seq data clustering. To accurately characterize the structure information of the labeled samples and the unlabeled samples, respectively, the proposed RGNMF-DS model adopts a couple of complementary regularizers (i.e., similarity and dissimilar regularizers) to guide matrix decomposition. In addition, we construct a graph regularizer to discover the local geometric structure hidden in ScRNA-seq data. Moreover, we adopt the $l_{2,1}$-norm to measure the reconstruction error and thereby effectively improve the robustness of the proposed RGNMF-DS model to the noises. Experimental results on several ScRNA-seq datasets have demonstrated that our proposed RGNMF-DS model outperforms other state-of-the-art competitors in clustering.

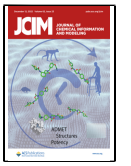


INTRODUCTION

ScRNA-Seq is the high-throughput sequencing of the genome, transcriptome, and epigenomics at the level of a single cell and can analyze the biological process of heterogeneous immune cell subsets and cell-to-cell variation. Compared with traditional RNA sequencing, ScRNA-seq not only shows the advantages of high-throughput and high-depth sequencing but also accurately measures the state of the single cell. Consequently, this can effectively reduce the correlation between cells. Therefore, it can clearly show the heterogeneity between cells in an organism and accurately reflect the molecular biological processes within a specific cell population. Recently, the ScRNA-seq technology has attracted extensive attention in the field of bioinformatics.[1] Clustering plays an important role in ScRNA-seq data analysis, which is closely related to cell heterogeneity analysis, cell differentiation, and other related studies. In the past few years, it is still a hot issue how to accurately cluster the ScRNA-seq data. Single-cell clustering analysis aims at assigning the cells to different clusters one by one according to certain similarities between cells and cells in the gene expression matrix. Thus, the relationship between cells can reveal cell subtypes and infer cell lineage.[2]

Over the past years, several data representation methods, such as initial dimension reduction (e.g., principal component analysis (PCA)),[3] visualization (e.g., T-distributed stochastic neighbor embedding (T-SNE), uniform manifold approximation and projection (UMAP)),[4] similarity quantification (e.g., single-cell interpretation via multikernel learning),[5] spectral clustering (SC),[6] and sparse spectral clustering (SSC),[7] have been widely applied to the representation of ScRNA-seq data. PCA is a classic dimensionality reduction method in pattern recognition. It uses the orthogonal transformation to reduce linear dependencies among samples, resulting in a new set of nonlinear correlation representations, which can capture the characteristics of the cells. T-SNE is the commonly used nonlinear dimensionality reduction method and can reveal the relationship between the cells. It converts the similarity of cells into probabilities and then minimizes the Kullebeck−Leibler divergence by gradient descent algorithm.[8] The SC method[6] finds low-dimensional representation embedded in the data by computing the eigenvectors of the constructed Laplacian matrix. Lu et al.[7] further proposed the SSC algorithm by imposing the sparse regularization constraint on the affinity matrix. Wang et al.[5] developed a novel SIMLR method, which constructs similarity matrices by fusing multiple Gaussian

Received: October 18, 2022
Published: December 2, 2022



kernel functions and then clusters individual cells by performing the spectral clustering algorithm on the similarity matrices. According to the low-rank representation (LRR) theory,[9] Zheng et al.[10] proposed a ScRNA-seq data detection algorithm based on similarity learning to identify cell types.

NMF[11] is a well-known data representation owing to its effectiveness and efficiency. It aims at decomposing a high-dimensional data matrix into the product of two or more low-dimensional non-negative matrices. In the past few decades, a series of variants of NMF have been proposed based on different goals.[12−14] Kong et al.[13] developed a robust NMF approach using the $l_{2,1}$-norm to measure the reconstruction error. Cai et al.[12] put forward to preserve the manifold structure of data in low-dimensional representation by constructing a graph regularizer. In practical applications, to take full advantage of limited label information among data, Wang et al.[14] employed label information to construct a pair of similar and dissimilar constraints, in which the label information can effectively guide matrix decomposition. In the field of single-cell clustering, the NMF methods have also been widely used in recent years. Chen et al.[15] proposed a regularized NMF method based on a deep double random graph for ScRNA-seq data clustering. Xiao et al.[16] proposed a graph regularized NMF method (GRNMF) to discover potential associations between miRNAs and diseases in heterogeneous omics data. To alleviate the influence of noises, Yu et al.[17] developed a robust hypergraph regularized NMF for feature selection and sample clustering in gene expression data.

In practice, the cell type labeling methods (manual labeling plus annotation library, immunofluorescence imaging, and cross-species comparison) can be used to label a small amount of data. In addition, there are some types of cells in clinical data, which can be used as labeled samples. To utilize the known label information, Liu et al. proposed a novel data representation method, called joint-$L_{2,1}$-norm-constraint-based semisupervised feature extraction (L21SFE), to analyze RNA-seq data.[18]

In recent years, the data representation algorithms of ScRNA-seq have received extensive attention. However, there still exist some limitations in analyzing ScRNA-seq data:[19] (1) High dimensionality of the ScRNA-seq data.[20] A single cell usually contains thousands of expressed genes, and thus the genetic information contained in the cell grows exponentially with the growth of the number of cells. Moreover, the gene sequence contains a large amount of redundant information, which seriously affects the performance of the ScRNA-seq data clustering. (2) Reverse transcription and amplification are required in ScRNA-Seq technology.[21] During this process, RNA transcripts may be lost, resulting in them being set to zero if they cannot be recognized in sequencing. This phenomenon is known as the dropout event. However, only a subset of genes are expressed in each cell, and other genes that cannot be expressed are also set to zero. Therefore, it is necessary to accurately identify the true zero values in the ScRNA-seq data and the zero values generated by the dropout event.

To alleviate the aforementioned issues, this work put forward a novel method, called RGNMF-DS, to represent the ScRNA-seq data. To discover the manifold structure embedded into the ScRNA-seq data, we construct a graph regularizer using the p-nearest-neighbor graph and then integrate it into the model. In addition, we employ the $l_{2,1}$-norm to measure the reconstruction error. Therefore, it effectively improves the robustness of the model in dealing with the ScRNA-seq data with noises. Furthermore, we construct a pair of complementary regularizers to guide matrix decomposition by fully utilizing the label information. In addition, we develop an efficient alternating updating algorithm to optimize our proposed model. Experimental results on different ScRNA-seq datasets show the effectiveness of our proposed RGNMF-DS approach.

RELATED WORK

In this section, some ScRNA-seq data representation approaches are introduced in detail.

NMF. NMF[11] is a popular analysis approach to gene expression data. Given a non-negative matrix $X \in \mathbb{R}^{m\times n}$ consisting of n cell samples with m-dimensional gene features, NMF aims at finding two non-negative matrices U and V to approximate the ScRNA-seq data matrix X

$$\min_{U,V} \|X - UV\|_F^2, \text{ s. t., } U \geq 0, V \geq 0 \tag{1}$$

where $\|\cdot\|_F$ represents the Frobenius norm of a matrix, $U \in \mathbb{R}^{m\times k}$ denotes the basis matrix, and $V \in \mathbb{R}^{k\times n}$ stands for the new gene expression matrix. Let $V_i^T = [v_{i1},\cdots, v_{ik}]^T$ denote the i-th line of V, where V_i is considered as the new representation of the i-th data point relative to the new basis matrix U. The nonnegativity constraints of U and V ensure that the factorization allows only additive combinations. Therefore, NMF is considered a part-based learning representation and shows strong interpretability in the psychological and physiological.

Using the multiplicative updating algorithm proposed by Lee[11] to optimize eq 1, we can derive its updating rules as follows

$$U_{ij}^{t+1} \leftarrow U_{ij}^t \frac{(XV^T)_{ij}}{(UVV^T)_{ij}} \tag{2}$$

$$V_{ij}^{t+1} \leftarrow V_{ij}^t \frac{(U^TX)_{ij}}{(U^TUV)_{ij}} \tag{3}$$

In bioinformatics, NMF is widely used to extract useful features belonging to different cell types from microarrays and scRNA-seq.[22,23] Recently, some variants of NMF, such as discriminant NMF (DNMF)[24] and sparse NMF(SNMF),[25] are developed in the field of biology. Specifically, SNMF introduces a regularization term on U or V, which can control the sparseness and generate a more sparse representation. DNMF applies the Fisher criterion to the coefficient matrix. This can maximize the distance between samples in different categories while minimizing the distance between pairs of samples in the same category. The above two methods are all optimized for data dimensionality reduction. However, due to the dropout problem, the ScRNA-seq data usually contains a large amount of noise, so the robustness of the model needs to be further improved.

Robust NMF. The original NMF and its variants usually use the Euclidean distance to measure the error of the approximation.[26] Many studies have shown that it is optimal for zero-mean Gaussian noise. However, it cannot effectively deal with the gene data with sparse noises or outliers. Therefore, Kong[13] proposed a robust NMF (rNMF) approach to handle the outliers and noises, which employs the $l_{2,1}$-norm

https://doi.org/10.1021/acs.jcim.2c01305
J. Chem. Inf. Model. 2022, 62, 6271−6286

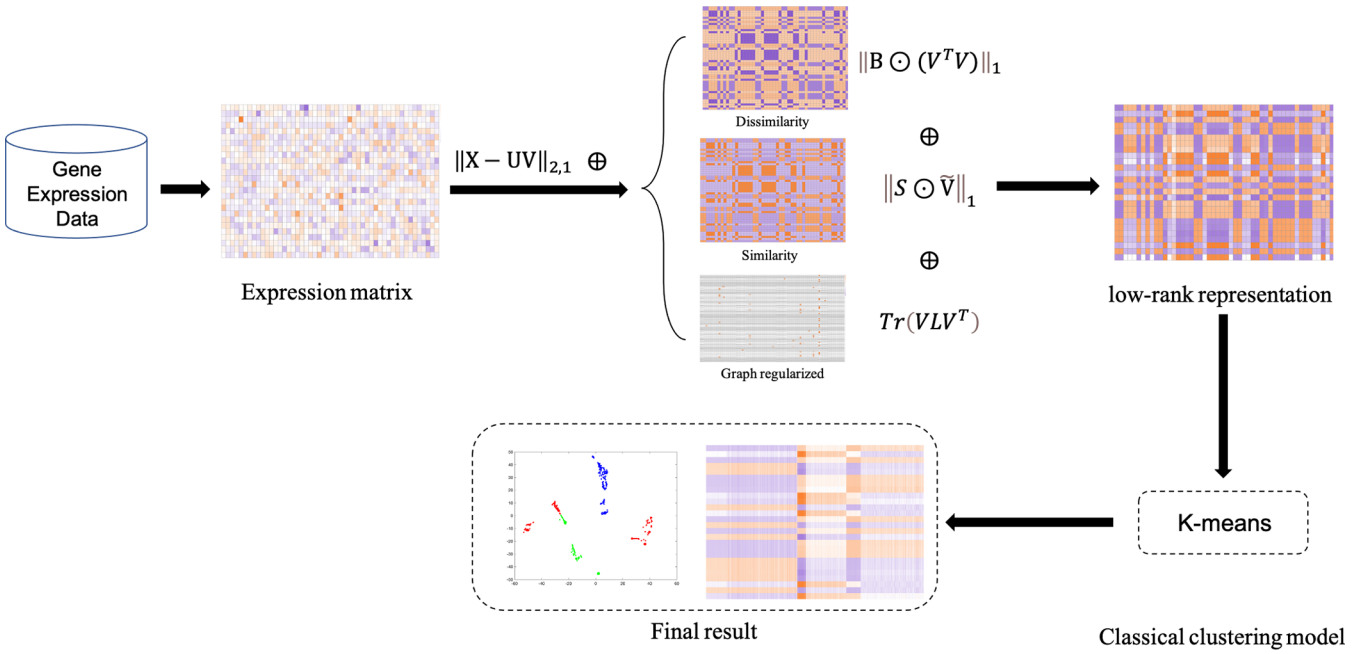


Figure 1. Framework of our RGNMF-DS algorithm.

instead of Frobenius norm to measure the approximation error. Therefore, the objective function of $l_{2,1}$-NMF can be given as follows

$$\min_{U,V} \|X - UV\|_{2,1}, \text{ s. t. } U \geq 0, V \geq 0 = \sum_{j=1}^{n} \left\|x_j - Uv_j\right\| \tag{4}$$

where $\|\cdot\|_{2,1}$ represents the $l_{2,1}$-norm of a matrix. Similarly, using the multiplicative updating algorithm proposed by Kong,[13] we can get the updating rules of eq 4 as follows

$$U_{ij}^{t+1} \leftarrow U_{ij}^{t} \frac{(XHV^T)_{ij}}{(UVHV^T)_{ij}} \tag{5}$$

$$V_{ij}^{t+1} \leftarrow V_{ij}^{t} \frac{(U^T XH)_{ij}}{(U^T UVH)_{ij}} \tag{6}$$

where H is a diagonal matrix and its diagonal elements are represented by

$$H_{jj} = \frac{1}{\sqrt{\sum_{i=1}^{m} (X - UV)_{ij}^2}} = \frac{1}{\|x_j - Uv_j\|} \tag{7}$$

Wu et al.[27] proposed a robust semisupervised NMF (rssNMF) model for single-cell clustering. Its new variable is introduced to absorb the noise of the data, and the marker gene is incorporated into the graph regularization term as prior knowledge. However, in single-cell clustering analysis, the label information of some samples can also be easily obtained. To fully consider the known information, a semisupervised method is proposed in this work.

PROPOSED MODEL

Motivation. One significant characteristic of the ScRNA-seq data is the high-dimensional property. Therefore, we need to reduce its dimensionality before clustering analysis. However, traditional ScRNA-seq data representation methods cannot take full advantage of the prior knowledge and simultaneously ignore the influence of noise or outliers during this procedure. To alleviate these issues, we propose a novel semisupervised learning method, called RGNMF-DS, to represent the high-dimensional ScRNA-seq data from real cells. Figure 1 plots the framework of our proposed RGNMF-DS approach. Specifically, to improve the robustness of the model, we replace the Frobenius norm with the $l_{2,1}$-norm to calculate the loss function. To fully utilize the limited label information of labeled ScRNA-seq samples, we impose dissimilarity and similarity constraints on the coefficient matrix. Furthermore, our proposed model preserves the manifold structure of the ScRNA-seq data by constructing a graph regularizer. Therefore, we can get a low-dimensional representation of the ScRNA-seq data using our RGNMF-DS model, and then adopt the K-means algorithm to get the final clustering results.

Dissimilarity Constraint. Assume that the input matrix $X \in \mathbb{R}^{m \times n}$ denotes the ScRNA-seq data with m genes measured in n single cells, which are composed of two subsets, i.e., $X = [X_l, X_u]$, where $X_l \in \mathbb{R}^{m \times l}$ denotes the l labeled ScRNA-seq samples and $X_u \in \mathbb{R}^{m \times (n-l)}$ denotes the $n - l$ unlabeled ScRNA-seq samples. To utilize the limited label information, we construct the dissimilarity matrix B as follows

$$B_{i,j} = \begin{cases} 1 & \text{if } x_i, x_j \in X_l \text{ and } x_i, x_j \in \text{different classes} \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

We employ the inner product $V^T V \in \mathbb{R}^{n \times n}$ of the low-dimensional representation to describe the similarity between the labeled samples. Therefore, we can construct the dissimilarity constraint as follows

$$\|B \odot (V^T V)\|_1 \tag{9}$$

where $\|\cdot\|_1$ represents the l_1 norm of a matrix. Clearly, the inner product of the new representations of the ScRNA-seq data from different categories becomes smaller when we

https://doi.org/10.1021/acs.jcim.2c01305
J. Chem. Inf. Model. 2022, 62, 6271−6286

minimize eq 9. Therefore, this constraint makes the low-dimensional representations of ScRNA-seq data from different categories more separate from each other.

Similarity Constraint. From eq 8, we can see that the elements of B are assigned to zero when the ScRNA-seq data samples are from the same category. Therefore, there is no concern about the similarities between such samples. To solve this issue, the following similar constraint is constructed from label information

$$\|\overline{S} \odot \tilde{V}\|_1 \tag{10}$$

where $\tilde{V}$ is the pairwise Euclidean distance matrix, and its elements are expressed as follows

$$\tilde{V}_{i,j} = \|v_i - v_j\|^2 \quad \forall\, i, j \in \{1, 2, 3, \cdots, n\} \tag{11}$$

And the similarity matrix $\overline{S}$ is given based on the known label information as follows

$$\overline{S}_{i,j} = \begin{cases} 1 & \text{if } x_i, x_j \in X_l \text{ and } x_i, x_j \in \text{same classes} \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

We can see that the Euclidean distance of the low-dimensional representation of the labeled samples in the same class will shrink as eq 10 decreases, making the heterogeneity of the low-dimensional representation of the same class decrease.

For unlabeled samples, if two samples are highly similar, it can be reasonably assumed that their low-dimensional representations are similar. Therefore, we construct a similarity matrix to preserve the structure information of unlabeled samples, whose elements can be defined as follows

$$\hat{S}_{i,j} = \begin{cases} \exp\left(-\dfrac{\|x_i - x_j\|^2}{\delta^2}\right) & \text{if } (x_i) \in \mathcal{N}_p(x_j) \text{ and } (x_i \text{ and } x_j) \in X_u \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

where $x_i \in \mathcal{N}_p(x_j)$ represents that the i-th sample belongs to the p-nearest neighbors of the j-th sample and $\delta \in \mathbb{R}$ denotes the parameter. Then, we can obtain the total similarity matrix as follows

$$S = \overline{S} + \hat{S} \tag{14}$$

Consequently, we employ the similarity matrix to model the structure information of the ScRNA-seq data and then can give the similarity constraint as follows

$$\|S \odot \tilde{V}\|_1 \tag{15}$$

Graph Regularizer. A natural assumption is that if two points x_i and x_j are close to each other in high-dimensional space, then their low representation v_i and v_j are also close to each other. To consider the local geometric structure of data in matrix decomposition, GNMF[12] is proposed to model the structure of data by constructing a nearest neighbor graph.

We employ the Heat kernel weighting method to construct the weight matrix W of the graph. If nodes i and j are connected, the elements of the matrix are given as follows

$$W_{ij} = e^{-\|x_i - x_j\|^2/\sigma} \tag{16}$$

Using the Euclidean distance as the measure metric, we have

$$d(v_i, v_j) = \|v_i - v_j\|^2 \tag{17}$$

Then, we can measure the smoothness of the low-dimensional representation as follows

$$\begin{aligned} \mathcal{R} &= \frac{1}{2}\sum_{i,j=1}^{n} \|v_i - v_j\|^2 W_{ij} \\ &= \sum_{i=1}^{n} v_i^T v_i D_{ii} - \sum_{i,j=1}^{n} v_i^T v_j D_{ij} \\ &= \mathrm{Tr}(VDV^T) - \mathrm{Tr}(VWV^T) = \mathrm{Tr}(VLV^T) \end{aligned} \tag{18}$$

where $\mathrm{Tr}(\cdot)$ denotes the trace of a matrix and the regularization parameter $\lambda \geq 0$ controls the smoothness of the new representation. D is a diagonal matrix whose entries are column sums of W, $D_{jj} = \sum_j W_{ij}$. $L = D - W$ is the graph Laplacian matrix.

Objective Function. To fully consider the prior knowledge of the ScRNA-seq data, our proposed RGNMF-DS method integrates the aforementioned constraints into matrix factorization using the regularization technology. To effectively handle the ScRNA-seq data with noises or outliers, $l_{2,1}$-norm is used to measure the reconstruction error in our proposed model. Therefore, the objective function of our proposed RGNMF-DS method is given as follows

$$\begin{aligned} \min_{U,V} \|X - UV\|_{2,1} + \lambda \|B \odot (V^T V)\|_1 + \mu \|S \odot \tilde{V}\|_1 \\ + \alpha \mathrm{Tr}(VLV^T) \quad \text{s.t.,}\ U \geq 0,\ V \geq 0 \end{aligned} \tag{19}$$

where λ, μ, and α are three regularization parameters that balance the contributions of different constraints.

OPTIMIZATION ALGORITHM

Since eq 19 is a nonconvex problem for both U and V together, its global optimal solution cannot be achieved theoretically. However, an alternating updating algorithm can be used to solve eq 19. Therefore, we can achieve a local minimization of eq 19 by updating one variable while fixing other variables. The Lagrange function of eq 19 is expressed as follows

$$\begin{aligned} \mathcal{L} = \|X - UV\|_{2,1} + \lambda \|B \odot V^T V\|_1 + \mu \|S \odot \tilde{V}\|_1 \\ + \alpha \mathrm{Tr}(VLV^T) - \mathrm{Tr}(\Phi U) - \mathrm{Tr}(\Psi V) \end{aligned} \tag{20}$$

where Φ and Ψ are the Lagrange multiplier matrices associated with $U \geq 0$ and $V \geq 0$, respectively.

Updating Rule of *U*. By calculating the partial derivatives of eq 20 with respect to U, we have

$$\frac{\partial \mathcal{L}}{\partial U} = -(XQV^T) + UVQV^T - \Phi \tag{21}$$

where Q is a diagonal matrix with the diagonal elements given by

$$Q_{jj} = \frac{1}{\sqrt{\sum_{i=1}^{m} (X - UV)_{ij}^2}} = \frac{1}{\|x_j - Uv_j\|} \tag{22}$$

Using the KKT conditions $\Phi_{ij} u_{ij} = 0$, we get the following equations for u_{ij}

$$-(XQV^T)_{ij} u_{ij} + (UVQV^T)_{ij} u_{ij} = 0 \tag{23}$$

Therefore, the updating rule of the basis matrix U can be obtained as follows

https://doi.org/10.1021/acs.jcim.2c01305
J. Chem. Inf. Model. 2022, 62, 6271−6286

$$u_{ij}^{t+1} \leftarrow u_{ij}^{t} \frac{(XQV^T)_{ij}}{(UVQV^T)_{ij}} \tag{24}$$

Updating Rule of *V*. By taking the partial derivatives of $\mathcal{L}$ w.r.t. V, we can get

$$\frac{\partial \mathcal{L}}{\partial V} = -U^T XQ + U^T UVQ + 2\lambda VB + 4\mu V(A - S) + 2\alpha VL + \Psi \tag{25}$$

where $A \in \mathbb{R}^{n\times n}$ is a diagonal matrix with its diagonal element $A_{ii} = \sum_{j=1}^{n} S_{ij}$. $L = D - W$ is the graph Laplacian matrix. Using the KKT conditions $\psi_{ij} v_{ij} = 0$, we get the following equation for v_{ij}

$$-(U^T XQ + 4\mu VS + 2\alpha VW)_{ij} v_{ij} + (U^T UVQ + 2\lambda VB + 4\mu VA + 2\alpha VD)_{ij} v_{ij} = 0 \tag{26}$$

Therefore, the updating rule of V in elementwise can be written as follows

$$v_{ij}^{t+1} \leftarrow v_{ij}^{t} \frac{(U^T XQ + 4\mu VS + 2\alpha VW)_{ij}}{(U^T UVQ + 2\lambda VB + 4\mu VA + 2\alpha VD)_{ij}} \tag{27}$$

In summary, the proposed RGNMF-DS algorithm is given in Algorithm 1.

Algorithm 1 RGNMF-DS

Input: The ScRNA-seq data matrix $X \in \mathbb{R}^{m\times n}$;
The parameters λ, μ, α and k.
Initialize: $U = rand(m, k)$ and $V = rand(k, n)$
1 Construct the dissimilarity matrix B by Eq.(8);
2 Construct the similarity matrix $\bar{S}$ by Eq.(12);
3 Calculate the similarity matrix S by Eq.(14);
4 Calculate the weight matrix W of the graph by Eq.(16);
5 Repeat
(1) Update Q by Eq.(22);
(2) Update U by Eq.(24);
(3) Update V by Eq.(27);
Until convergence
Output: U and V

Computational Complexity Analysis. In this subsection, the computational cost of our RGNMF-DS approach is briefly discussed. First, we need to construct four symmetric matrices B, $\bar{S}$, S, and W. Obviously, the values of matrix B (or $\bar{S}$) is not zero only when its row and column are less than l. Therefore, the computational complexity of constructing the matrix B (or $\bar{S}$) is $O(l^2)$. To derive the matrices S and W, we need to construct two p-nearest neighbor graphs. Each row of S or W contains p nonzero elements on average. Therefore, the computational complexity of both S and W is $O(Np)$.

Since the proposed model is optimized by the multiplicative updating algorithm, the main computational cost is mainly caused by updating the matrices Q, U, and V. The computational complexity of updating Q in eq 22 is $O(nmk)$. Similarly, the computational complexities of the updating rules 24 and 27 are $O(n^2m)$ and $O(n^2k)$, respectively. Consequently, the overall complexity of the RGNMF-DS approach is $O(n^2m)$.

CONVERGENCE ANALYSIS

For the updating rules for U and V, we have the following theorem:

Theorem 1: The model in eq 19 is nonincreasing under the updating rules 24 and 27.

The lower bound of our proposed model in eq 19 is greater than zero. Since the last three terms in eq 19 are only related to V, the updating rule of U in RGNMF-DS is the same as that of the original $l_{2,1}$-NMF. Therefore, it is clear to know that eq 19 is nonincreasing under the updating rule 24 according to ref 28.

Therefore, we only need to prove that eq 19 is nonincreasing using rule 27. Therefore, we need to make use of an auxiliary function similar to that used in the Expectation−Maximization algorithm.[29]

Definition: $G(v, v')$ is an auxiliary function for $F(v)$ if the conditions

$$G(v, v') \geq F(v), \quad G(v, v) = F(v) \tag{28}$$

are satisfied.

The auxiliary function is very important due to the following *lemma*.

Lemma1: If G is an auxiliary function of F, then F is nonincreasing under the updating rule

$$v^{t+1} = \arg\min_{v} G(v, v^t) \tag{29}$$

Proof:

$$F(v^{t+1}) \leq G(v^{t+1}, v^t) \leq G(v^t, v^t) = F(v^t) \tag{30}$$

The objective function eq 19 w.r.t. V is written as

$$O_V = \|X - UV\|_{2,1} + \lambda \|B \odot (V^T V)\|_1 + \mu \|S \odot \tilde{V}\|_1 + \alpha \mathrm{Tr}(VLV^T) \tag{31}$$

The first-order derivate of eq 31 w.r.t V_{ab} is given as follows

$$F'_{ab} = \frac{\partial O_V}{\partial V_{ab}} = (-UTXQ + U^T UVQ + 2\lambda VB + 4\mu V(A - S) + 2\alpha VL)_{ab} \tag{32}$$

The second-order derivate of eq 31 w.r.t V_{ab} is expressed as follows

$$F''_{ab} = \frac{\partial^2 O_V}{\partial V_{ab} \partial V_{ab}} = (U^T U)_{aa} Q_{bb} + 2\lambda B_{bb} + 4\mu (A - S)_{bb} + 2\alpha (D - W)_{bb} \tag{33}$$

And the high-order derivate of O_V regarding V_{ab} is given as follows

$$F'''_{ab} = F''''_{ab} = \cdots = 0 \tag{34}$$

Based on the Taylor expansion, we can rewrite the Taylor expansion formula of eq 31 at point V_{ab}^t as follows

https://doi.org/10.1021/acs.jcim.2c01305
J. Chem. Inf. Model. 2022, 62, 6271−6286

$$O_V = F_{ab}(v)$$
$$= F_{ab}(V_{ab}^t) + F_{ab}^{'}(V_{ab}^t)(v - V_{ab}^t) + \frac{F_{ab}^{''}(V_{ab}^t)}{2}(v - V_{ab}^t)^2$$
$$= F_{ab}(V_{ab}^t) + F_{ab}^{'}(V_{ab}^t)(v - V_{ab}^t) + \frac{1}{2}((U^TU)_{aa}Q_{bb} + 2\lambda B_{bb} + 4\mu(A - S)_{bb} + 2\alpha(D - W)_{bb})(v - V_{ab}^t)^2 \quad (35)$$

Then, we give the upper bound auxiliary function for O_V as $G(v, V_{ab}^t)$.

$$G(v, V_{ab}^t) = F_{ab}(V_{ab}^t) + F_{ab}^{'}(V_{ab}^t)(v - V_{ab}^t) + \frac{\frac{1}{2}(U^TUVQ)_{ab} + \lambda(VB)_{ab} + 2\mu(VA)_{ab} + \alpha(VD)_{ab}}{V_{ab}^t} \times (v - V_{ab}^t)^2 \quad (36)$$

It is obvious that $F_{ab}(v) = G(v, V_{ab}^t)$ when $V_{ab}^t = v$. Therefore, we only need to prove $F_{ab}(v) \le G(v, V_{ab}^t)$, which is equivalent to the following form

$$\frac{(U^TUVQ)_{ab} + 2\lambda(VB)_{ab} + 4\mu(VA)_{ab} + 2\alpha(VD)_{ab}}{V_{ab}^t} \ge (U^TU)_{aa}Q_{bb} + 2\lambda B_{bb} + 4\mu(A - S)_{bb} + 2\alpha(D - W)_{bb} \quad (37)$$

It is easy to obtain

$$\frac{(U^TUVQ)_{ab}}{V_{ab}^t} = \frac{\sum_{k=1}^K (U^TU)_{ak} \times V_{kb} \times Q_{bb}}{V_{ab}^t} \ge \frac{(U^TU)_{aa} \times V_{ab} \times Q_{bb}}{V_{ab}^t} = (U^TU)_{aa}Q_{bb} \quad (38)$$

Therefore, the inequality can hold since $\forall(k, b)V_{kb}$ is nonnegative. Similarly, we can obtain three inequalities as follows

$$\frac{VD_{ab}}{V_{ab}^t} = \frac{\sum_{k=1}^K V_{ak} \times B_{kb}}{V_{ab}^t} \ge \frac{V_{ab} \times B_{bb}}{V_{ab}^t} = B_{bb} \quad (39)$$

and

$$\frac{(VA)_{ab}}{V_{ab}^t} = \frac{\sum_{k=1}^K V_{ak} \times A_{kb}}{V_{ab}^t} \ge \frac{V_{ab} \times A_{bb}}{V_{ab}^t} \ge \frac{V_{ab} \times (A - S)_{bb}}{V_{ab}^t} = (A - S)_{bb} \quad (40)$$

and

$$\frac{(VD)_{ab}}{V_{ab}^t} = \frac{\sum_{k=1}^K V_{ak} \times D_{kb}}{V_{ab}^t} \ge \frac{V_{ab} \times D_{bb}}{V_{ab}^t} \ge \frac{V_{ab} \times (D - W)_{bb}}{V_{ab}^t} = (D - W)_{bb} \quad (41)$$

Therefore, eq 37 can hold and $F_{ab}(v) \le G(v, V_{ab}^t)$. The convergence proof of Theorem 1 can be given as follows:

Proof:

Replacing $G(v, V_{ab}^t)$ in eqs 29 by 36 results in the updating rule

$$V_{ab}^{t+1} = V_{ab}^t \frac{F_{ab}^{'}(V_{ab}^t)}{(U^TUVQ + 2\lambda VB + 4\mu VA + 2\alpha VD)_{ab}} = V_{ab}^t \frac{(U^TXQ + 4\mu VS + 2\alpha WV)_{ab}}{(U^TUVQ + 2\lambda VB + 4\mu VA + 2\alpha DV)_{ab}} \quad (42)$$

Since eq 36 is an auxiliary function, F_{ab} is nonincreasing under this updating rule.

EXPERIMENTAL RESULTS

In this section, we conducted extensive experiments to verify the effectiveness of our RGNMF-DS algorithm. First, the RGNMF-DS algorithm is used to reduce the dimension of cell gene expression matrix, and then the low-dimensional expression of the gene expression matrix is used to perform *K*-means clustering. To make a fair comparison, our RGNMF-DS algorithm is compared with several classical single-cell data representation methods on different ScRNA-seq datasets.

Table 1. Details of the Real Single-Cell Datasets

dataset	cells	genes	cell types	species	GSE/ID
Ramskold	33	3575	7	mouse	GSE38495
Human	124	3840	8	human	GSE36552
Pomeroy	42	1379	5	human	
Ning	460	19084	4	human	GSE64016
Gray	890	19020	8	human	GSE81252
Chung	559	57915	4	mouse	GSE75688
Yeo	214	34608	4	human	GSE85908
Limb	3909	1248	6	mouse	GSE109774

Datasets. We conducted clustering experiments on eight constructed ScRNA-seq datasets collected from NCBI and other public databases. The main statistics of these datasets are shown in Table 1. Here, "cells" and "genes", respectively, indicate the number of cells and genes. "cell type" is the number of cell types. All datasets adopt normalized expression levels (FPKM, RPM, or CPM).

Ramskold.[30] The Ramskold dataset includes 33 cell samples with 7 different types. They were human embryonic stem cells, LNCap cells, PC 3 cells, melanoma cell line CTC, melanoma cell line SKMEL 5, UACC 257 cells, and T24 cells from the bladder cancer cell line. Each cell in this dataset sequenced 3575 genes, which were obtained by sequencing a variety of mice using smart-seq techniques.

Human.[31] It consists of 8 kinds of human cells, with a total of 124 cells. The ScRNA-seq data were obtained from human preimplantation embryos and human embryonic stem cells using the ScRNA sequencing technology proposed by Tang et al.[32]

Pomeroy.[33] This dataset collected the gene expression data of DNA microarray from 42 patient samples with 5 different types. It includes malignant gliomas (Mglio), atypical teratoid/rhabdoid tumors (Rhab), primitive neuroectodermal tumors (PNETs), medulloblastoma (MD), and cancerous normal tissues (Ncer).

Ning.[34] This dataset includes 247 H1-Fucci cells and 213 H1 cells, and there are 19084 genes in each cell. This dataset is used to verify the ability of the Oscope method to distinguish oscillating genes in ScRNA-seq data of unsynchronized cell population.

https://doi.org/10.1021/acs.jcim.2c01305
J. Chem. Inf. Model. 2022, 62, 6271−6286

Table 2. Clustering ACC of Different Algorithms

Algorithm	Ramskold	Human	Pomeroy	Ning	Gray	Chung	Yeo	Limb
K-means	0.7020	0.5618	0.4609	0.4935	0.0848	0.1216	0.0569	0.1129
NMF	0.3939	0.2419	0.5476	0.2826	0.1820	0.3506	0.3925	0.3914
SC	0.7727	0.6035	0.3609	0.4370	0.0429	0.1202	0.0569	0.1241
GNMF	0.8485	0.7581	0.7381	0.4717	0.5944	0.5206	0.7523	0.8596
$l_{2,1}$-NMF	0.8788	0.7016	0.5714	0.3935	0.1764	0.5367	0.5140	0.6475
SSC	0.8182	0.6022	0.3109	0.4543	0.0483	0.0561	0.1215	0.0614
SIMLR	0.6162	0.5645	0.3565	0.4739	0.1041	0.0787	0.1433	0.1624
t-SNE	0.6566	0.6586	0.4087	0.5250	0.1519	0.0465	0.1020	0.1613
SinNLRR	0.5253	0.5161	0.4652	0.4348	0.1309	0.0650	0.1332	0.7991
rssNMF	0.9091	0.7500	0.7381	0.4717	0.6360	0.3685	0.7523	0.6933
SC3	0.7879	**0.7984**	0.6667	0.5391	0.8483	0.3953	0.7617	0.8808
S3NMF	0.6061	0.5887	0.5952	0.2870	0.4899	0.2934	0.7243	0.4612
DENMF	0.8182	0.6129	0.8095	0.3848	0.3712	0.6530	0.7654	0.7586
DSINMF	0.7879	0.6855	0.8571	0.5522	0.5843	0.4812	0.8692	0.7994
CASSL	0.4815	0.4194	0.4865	0.4703	0.3288	0.6877	0.4065	0.3836
RGNMF-DS	**0.9697**	0.7500	**0.8810**	**0.7891**	**0.8202**	**0.7245**	**0.9252**	**0.9854**

Table 3. Clustering NMI of Different Algorithms

Algorithm	Ramskold	Human	Pomeroy	Ning	Gray	Chung	Yeo	Limb
K-means	0.8228	0.6674	0.0298	0.1541	0.0978	0.0058	0.0088	0.0991
NMF	0.3337	0.1074	0.3728	0.0052	0.0177	0.0641	0.0496	0.2288
SC	0.8613	0.6804	0.0594	0.1204	0.0085	0.0016	0.0113	0.1129
GNMF	0.8858	0.7798	0.6572	0.0692	0.5878	0.1148	0.6616	0.8292
$l_{2,1}$-NMF	0.892	0.7016	0.4105	0.0718	0.0203	0.1694	0.1755	0.7057
SSC	0.9282	0.7019	0.0272	0.1018	0.0271	0.0048	0.0836	0.0209
SIMLR	0.7325	0.6891	0.0567	0.2449	0.1119	0.0241	0.1136	0.1542
t-SNE	0.7250	0.7342	0.1525	0.2508	0.1467	0.0014	0.0968	0.1521
SinNLRR	0.5548	0.6745	0.2008	0.1108	0.1422	0.0239	0.1096	0.6575
rssNMF	0.8898	0.7773	0.6654	0.0692	0.6610	0.0958	0.6477	0.7493
SC3	0.8484	0.8325	0.6701	0.4838	0.8177	0.1515	0.7203	0.8802
S3NMF	0.6582	0.6875	0.5796	0.0092	0.5858	0.0190	0.6717	0.6371
DENMF	0.8373	0.6534	0.6339	0.1367	0.3428	0.1254	0.5342	0.6381
DSINMF	0.9100	0.7325	0.7519	0.2926	0.6994	0.0986	0.7082	0.7888
CASSL	0.4188	0.2420	0.2749	0.0616	0.0787	0.0781	0.0547	0.0479
RGNMF-DS	**0.9619**	**0.8340**	**0.8232**	**0.5278**	**0.8337**	**0.3009**	**0.7914**	**0.9480**

Table 4. Clustering ARI of Different Algorithms

Algorithm	Ramskold	Human	Pomeroy	Ning	Gray	Chung	Yeo	Limb
K-means	0.6125	0.4489	-0.0020	0.1349	0.0498	0.0002	-0.0011	0.0998
NMF	0.0365	-0.0019	0.2231	-0.0020	0.0050	0.0166	0.0451	0.1873
SC	0.6986	0.5274	0.0309	0.1012	-0.0003	-0.0023	-0.0015	0.1125
GNMF	0.8302	0.6569	0.5497	0.0292	0.4163	0.0413	0.6519	0.8350
$l_{2,1}$-NMF	0.7816	0.5564	0.2479	0.0413	0.0045	0.0948	0.1912	0.5565
SSC	0.8302	0.4949	0.0105	0.1074	0.0147	−0.0008	0.0806	0.0178
SIMLR	0.5238	0.4455	0.0177	0.1716	0.0694	0.0127	0.1109	0.1605
t-SNE	0.4688	0.5658	0.0683	0.2004	0.1382	−0.0003	0.0733	0.1588
SinNLRR	0.2064	0.4381	0.1367	0.1070	0.1257	0.0048	0.1019	0.6116
rssNMF	0.8445	0.6496	0.5571	0.0292	0.5201	0.0015	0.6401	0.6639
SC3	0.6706	**0.7745**	0.5225	0.2751	0.6858	−0.0377	0.6636	0.8858
S3NMF	0.4279	0.4791	0.4077	-0.0020	0.3630	0.0019	0.6477	0.3971
DENMF	0.6470	0.5723	0.5644	0.0896	0.2050	0.2334	0.5549	0.6555
DSINMF	0.8028	0.5668	0.7038	0.2069	0.4915	0.0759	0.7231	0.7707
CASSL	0.1270	0.0819	0.0953	0.1016	0.0486	0.1403	0.0418	0.056
RGNMF-DS	**0.9427**	0.7392	**0.7624**	**0.5952**	**0.7745**	**0.4147**	**0.8195**	**0.9722**

Gray.[35] This dataset contains 890 cells from 8 types, and each cell is composed of 1920 genes. It was obtained by performing ScRNA-seq on cells during hepatocyte-like differentiation in two-dimensional (2D) culture and three-dimensional (3D) LBs.

Chung.[36] In the dataset, the samples were collected from several patients with primary breast cancer and several patients

https://doi.org/10.1021/acs.jcim.2c01305
J. Chem. Inf. Model. 2022, 62, 6271−6286

Table 5. Clustering Purity of Different Algorithms

Algorithm	Ramskold	Human	Pomeroy	Ning	Gray	Chung	Yeo	Limb
K-means	0.7727	0.7124	0.4652	0.4957	0.0856	0.1225	0.0592	0.1256
NMF	0.4242	0.3468	0.5476	0.4630	0.2742	0.7335	0.4159	0.5692
SC	0.8283	0.7379	0.4630	0.4830	0.0446	0.1222	0.0600	0.1364
GNMF	0.8485	0.8306	0.7619	0.4870	0.6045	0.7335	0.7804	0.8972
$l_{2,1}$-NMF	0.8788	0.7903	0.5714	0.4848	0.2764	0.7746	0.5607	0.8043
SSC	0.9091	0.7890	0.4630	0.5127	0.0627	0.1222	0.1215	0.0786
SIMLR	0.7020	0.7339	0.4630	0.5804	0.1227	0.1222	0.1433	0.1624
t-SNE	0.7121	0.7567	0.5217	0.5804	0.1519	0.1222	0.1207	0.1613
SinNLRR	0.5758	0.6935	0.5674	0.4848	0.1408	0.1222	0.1332	0.7991
rssNMF	0.9091	0.8306	0.7619	0.4870	0.6427	0.7335	0.7710	0.8273
SC3	0.8788	**0.8790**	0.6667	0.6913	0.8483	0.7335	0.7757	0.8961
S3NMF	0.6364	0.7177	0.6905	0.4630	0.6449	0.7335	0.7804	0.7409
DENMF	0.8182	0.7500	0.8095	0.5152	0.4685	0.7496	0.7654	0.7899
DSINMF	0.6667	0.7339	0.8571	0.6065	0.7506	0.4848	0.8692	0.8798
CASSL	0.5556	0.4274	0.5135	0.4813	0.3356	0.7292	0.4252	0.3905
RGNMF-DS	**0.9697**	**0.8790**	**0.8810**	**0.7891**	**0.8629**	**0.8157**	**0.9252**	**0.9854**

with lymph node metastasis. A total of 559 single cells are involved, such as four breast cancer subtypes: Lumina A, Lumina B, HER 2, and triple-negative breast cancer (TNBC). 57915 genes in each cell have been sequenced.

Yeo.[37] This dataset includes 214 cells from 4 subtypes, in which each cell contains 222 genes. It is used to analyze the changes in alternative splicing during the differentiation of motor neurons.

Limb.[38] The Limb dataset contains 3909 cells with 6 subtypes, and each cell includes 1248 genes. It was obtained by single-cell RNA sequencing of cells in 20 tissues from 3-month-old mice. These data reveal gene expression in cell populations with unclear characteristics. It is widely used to compare gene expression of common cell types among tissues.

Evaluation Metrics. Four widely used metrics were used to evaluate the performance of our RGNMF-DS algorithm. The first metric is the accuracy (ACC),[39,40] which calculates the percentage of correctly predicted class labels among the obtained P from the algorithm and the true class labels T. Therefore, the ACC can be calculated by the following formulation

$$\mathrm{ACC} = \frac{\sum_{i=1}^{n} \delta(t_i, \mathrm{map}(p_i))}{n} \tag{43}$$

where t_i and p_i denote the true and predicted class labels of i-th cell, respectively; $\delta(m, n)$ represents the indicator function; and n is the number of single cells. If m is different from n, this function is set to 0. Otherwise, it is set to 1. $\mathrm{map}(p_i)$ is an optimal permutation function that maps clustering labels to true labels, which is obtained by applying the Hungarian algorithm.[41]

The second metric is normalized mutual information (NMI),[42] which is often used to evaluate clustering results in information retrieval and feature selection. X is the prediction set obtained by each algorithm, and T is the truth set of the dataset. NMI is calculated as follows

$$\mathrm{NMI}(T, X) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{MP}(t_i)\ln \mathrm{MP}(t_i)}{\sqrt{\sum_{j=1}^{n} MP(t_i)\ln \mathrm{MP}(t_i)}} \tag{44}$$

where $\mathrm{MP}(t_i, x_j)$ is the joint probability that an arbitrarily chosen sample belongs to both sample types t_i and x_j. $\mathrm{MP}(t_i)$ and $\mathrm{MP}(x_j)$ refer to the probability that randomly selected samples belong to sample types t_i and x_j, respectively.

The third metric is the adjusted rand index (ARI),[43] which is a common index to measure the similarity between the clustering results and the labels provided by the dataset. Rand index (RI) is calculated as follows:

$$\mathrm{RI} = \frac{a + b}{C_n^2} \tag{45}$$

where a is defined as the number of instance pairs that are divided into the same category in the actual category and the same cluster in the clustering result, b is defined as the number of instance pairs that are divided into different categories in actual categories and different clusters in clustering results, and n represents the total number of instances. C_n^2 represents how many combinations of any two samples belong to one class. The value range of the RI is $[0, 1]$, and RI cannot guarantee that its values randomly divided by clustering results are close to 0. Therefore, the adjusted rand index (ARI) can be calculated as

$$\mathrm{ARI} = \frac{\mathrm{RI} - E[\mathrm{RI}]}{\max(\mathrm{RI}) - E[\mathrm{RI}]} \tag{46}$$

It is clear to see that the value range of ARI is $[-1, 1]$, and the larger the value, the better the clustering effect.

The fourth metric is purity, which measures the degree of existence of data points of major true type in each type. It is obtained by the weighted sum of purity values of individual clusters. Purity is given as follows

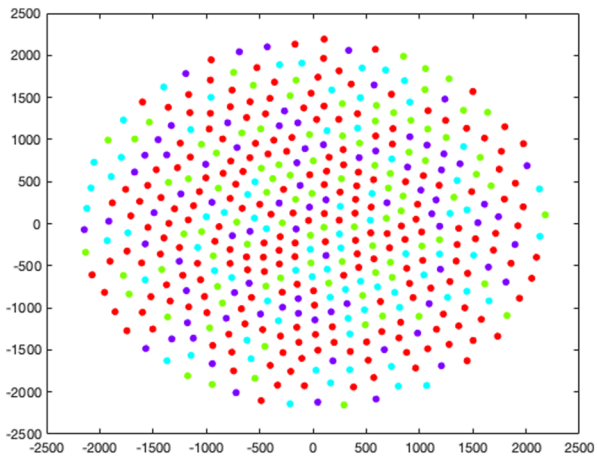
$$\mathrm{Purity} = \sum_{i=1}^{c} \frac{n_i}{n} P(K_i)$$

$$P(K_i) = \frac{1}{n_i} \max_j (n_i^j) \tag{47}$$

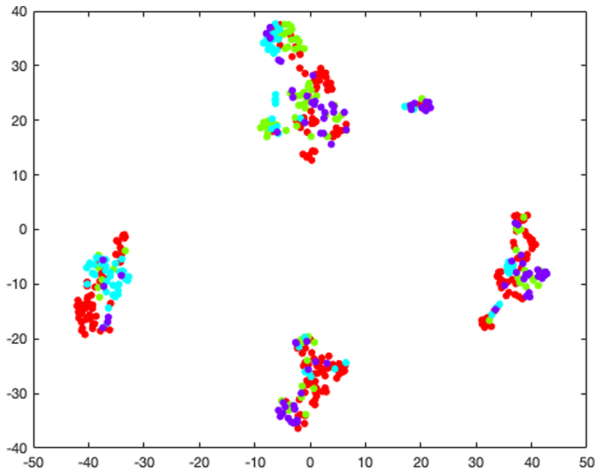
where K_i is the specific cluster size of n_i, n_i^j is the number of the i-th input class that are assigned to the j-th cluster, c is the number of the clusters, and n is the total number of the data points. Generally speaking, higher purity value means better clustering performance.

In the experiment, we used the above metrics to evaluate the performances of all methods from different aspects. The larger

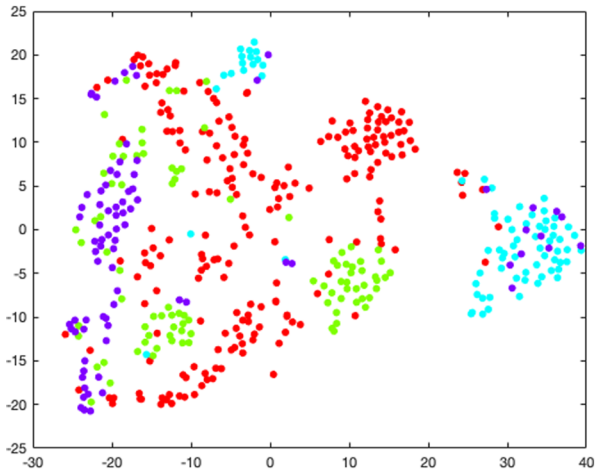
https://doi.org/10.1021/acs.jcim.2c01305
J. Chem. Inf. Model. 2022, 62, 6271−6286



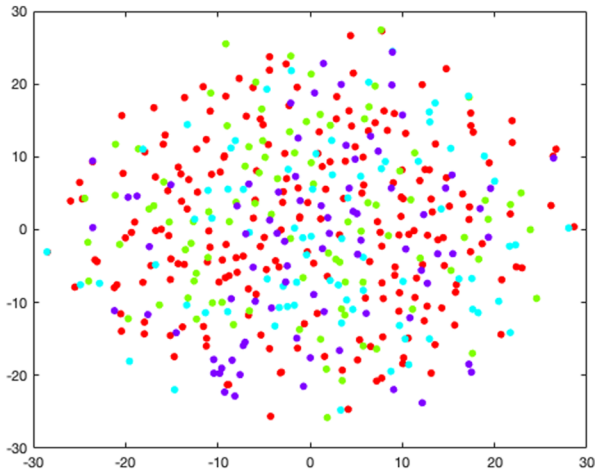
(a) Visualization results with SSC



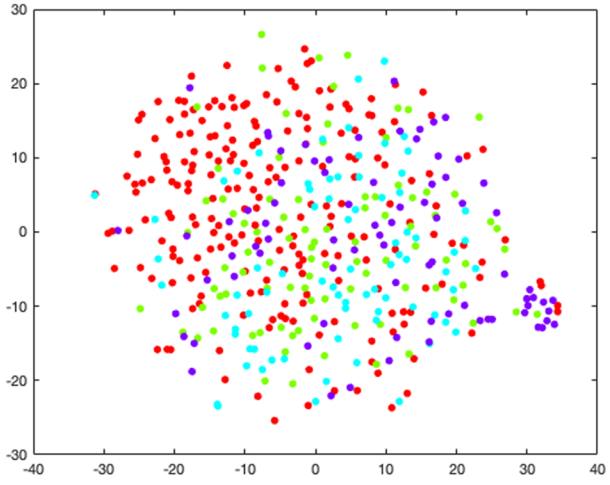
(b) Visualization results with SimLR



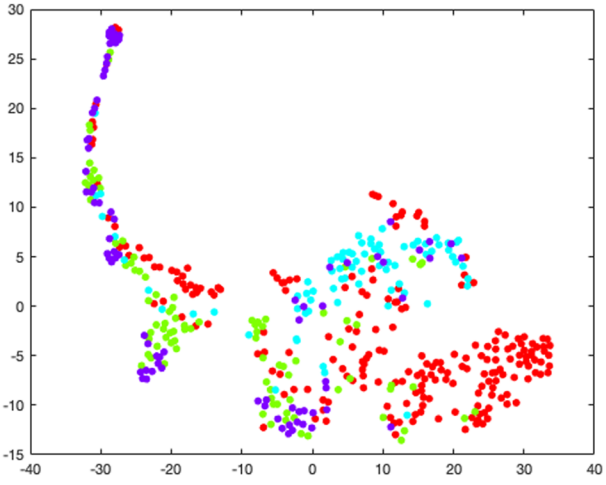
(c) Visualization results with SinNLRR



(d) Visualization results with NMF



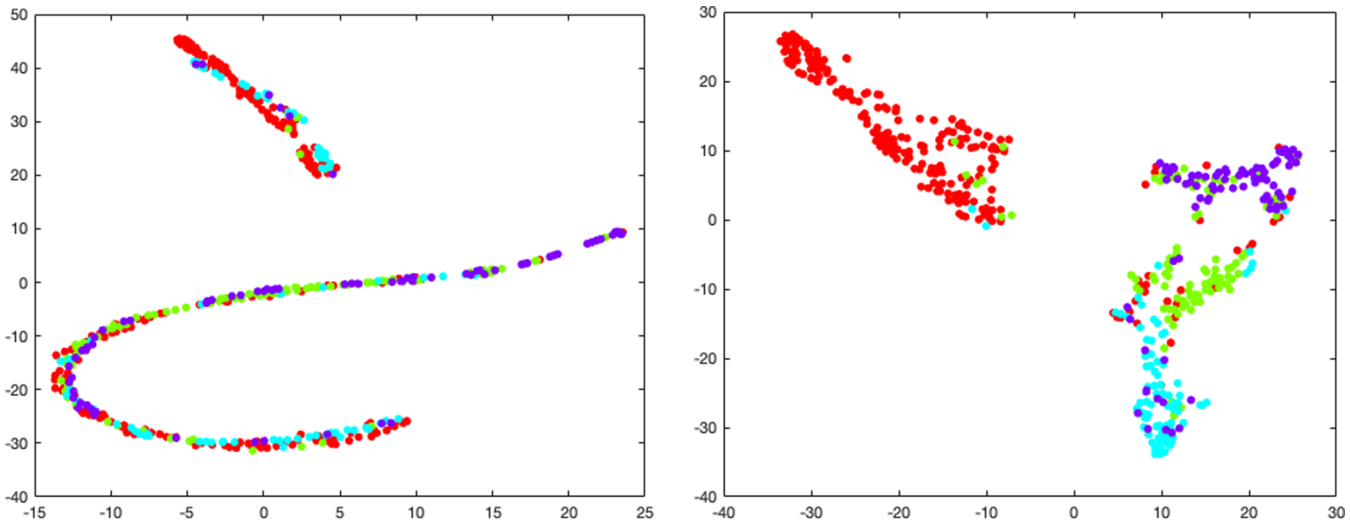
(e) Visualization results with $l_{2,1}$-NMF



(f) Visualization results with GNMF

Figure 2. continued

https://doi.org/10.1021/acs.jcim.2c01305
J. Chem. Inf. Model. 2022, 62, 6271−6286



(g) Visualization results with rssNMF (h) Visualization results with RGNMF-DS

Figure 2. Visualization results of different algorithms on the Ning dataset.

their values, the better the performance of the corresponding method. It indicates that a larger proportion of cells are assigned to the correct type.

Comparison Algorithms. We compared the proposed method with some classical ScRNA-seq clustering methods and summarized them as follows:

(1) *K*-means[44] is the most basic and commonly used clustering algorithm.
(2) NMF[11] is the original NMF model.
(3) SC[6] is a clustering method based on graph theory.
(4) GNMF[12] incorporates the manifold structure into the matrix decomposition.
(5) $l_{2,1}$-NMF[13] is a robust NMF approach by adopting $l_{2,1}$-norm loss function.
(6) SSC[7] clusters the ScRNA-seq data samples derived from the union of subspaces by employing sparse representation techniques.
(7) SIMLR[5] is a spectral clustering method. It learns that reliability distance metrics through multiple kernels are most suitable for the data structure.
(8) t-SNE[45] is a nonlinear dimensionality reduction algorithm, allowing us to display clustering results in a two-dimensional or three-dimensional space.
(9) SinNLRR[10] is a ScRNA-seq data representation method based on non-negative low-rank representations.
(10) rssNMF[27] is a robust and semisupervised NMF model.
(11) SC3[46] is a highly parallel and quantitative method for measuring gene expression in single cells.
(12) S3NMF[47] is a semisupervised symmetric non-negative matrix factorization clustering model.
(13) DENMF[48] is a semisupervised dual embedding model that unifies dimensionality reduction and clustering processes.
(14) DSINMF[15] is a new deep matrix factorization method for clustering single cell.
(15) CASSL[49] is a semisupervised learning ScRNA-seq clustering method.

Experimental Results and Analysis. *Comparison of Cell Type Identification.* In this section, we compared RGNMF-DS with other state-of-the-art methods for ScRNA-seq data clustering. Tables 2−5 recorded the clustering performances of all methods in different real datasets. According to these clustering results, we can give the following observations:

(a) It is clear to see that the $l_{2,1}$-NMF method outperforms NMF in single-cell clustering on all datasets. This is because $l_{2,1}$-NMF fully considers the ScRNA-seq mixed with noises by adopting the $l_{2,1}$-norm to measure the reconstruction error. Therefore, the experimental results show that $l_{2,1}$-NMF is more robust than NMF in ScRNA-seq clustering. In addition, we can see that GNMF can achieve superior performances than NMF on all ScRNA-seq datasets. The main reason is that the former takes full advantage of the manifold structure of the ScRNA-seq data by constructing a graph regularizer. Therefore, GNMF can effectively capture the semantic information of the ScRNA-seq data compared with NMF.
(b) It can be found that the RGNMF-DS algorithm achieves the best clustering performance on all datasets in comparison with NMF, $l_{2,1}$-NMF, and GNMF. The main reason is that our proposed RGNMF-DS algorithm fully considers more prior knowledge than other competitors. Especially, our proposed RGNMF-DS method not only explores the manifold structure of the ScRNA-seq data using the regularization technology but also adopts $l_{2,1}$-norm to improve its robustness in handling the noise problems.
(c) It is worth noting that our proposed RGNMF-DS algorithm outperforms the rssNMF on all ScRNA-seq datasets. Note that both RGNMF-DS and rssNMF belong to semisupervised learning methods. Specifically, rssNMF incorporates marker genes as prior information into the graph regularization term. Our proposed RGNMF-DS algorithm fully utilizes the limited label information by constructing a couple of complementary regularizers (i.e., similarity and dissimilar regularizers) to guide the matrix decomposition. Experimental results demonstrate the superiority of the proposed RGNMF-DS method in ScRNA-seq data clustering.

https://doi.org/10.1021/acs.jcim.2c01305
J. Chem. Inf. Model. 2022, 62, 6271−6286

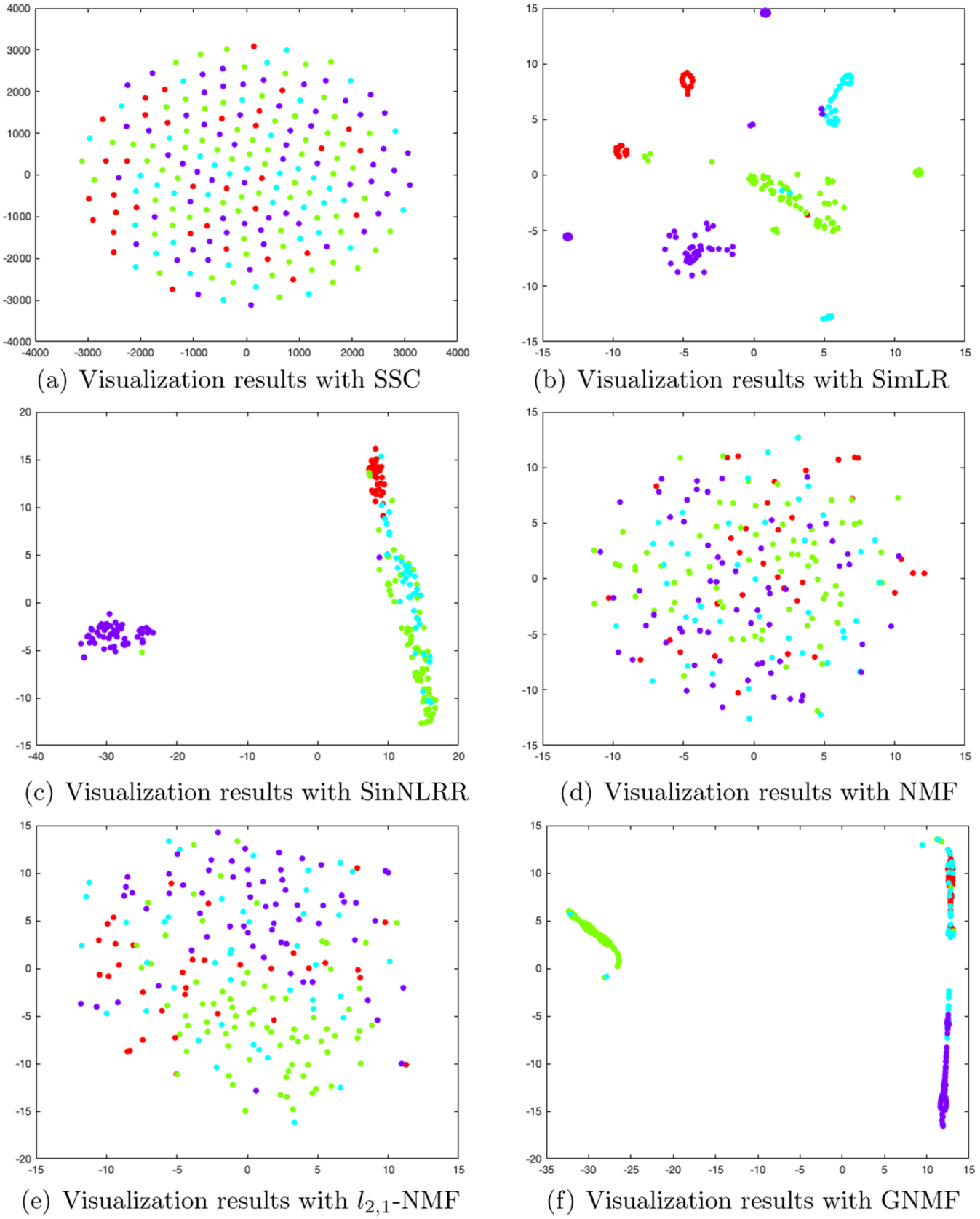
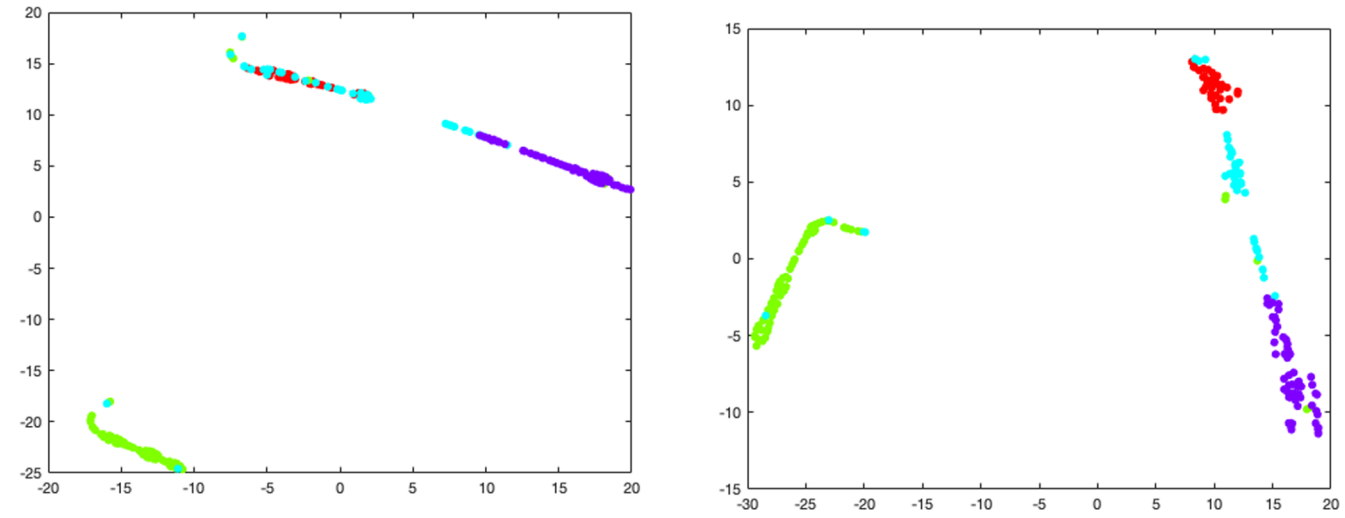


Figure 3. continued

https://doi.org/10.1021/acs.jcim.2c01305
J. Chem. Inf. Model. 2022, 62, 6271−6286



(g) Visualization results with rssNMF (h) Visualization results with RGNMF-DS

Figure 3. Visualization results of different algorithms on the Yeo dataset.

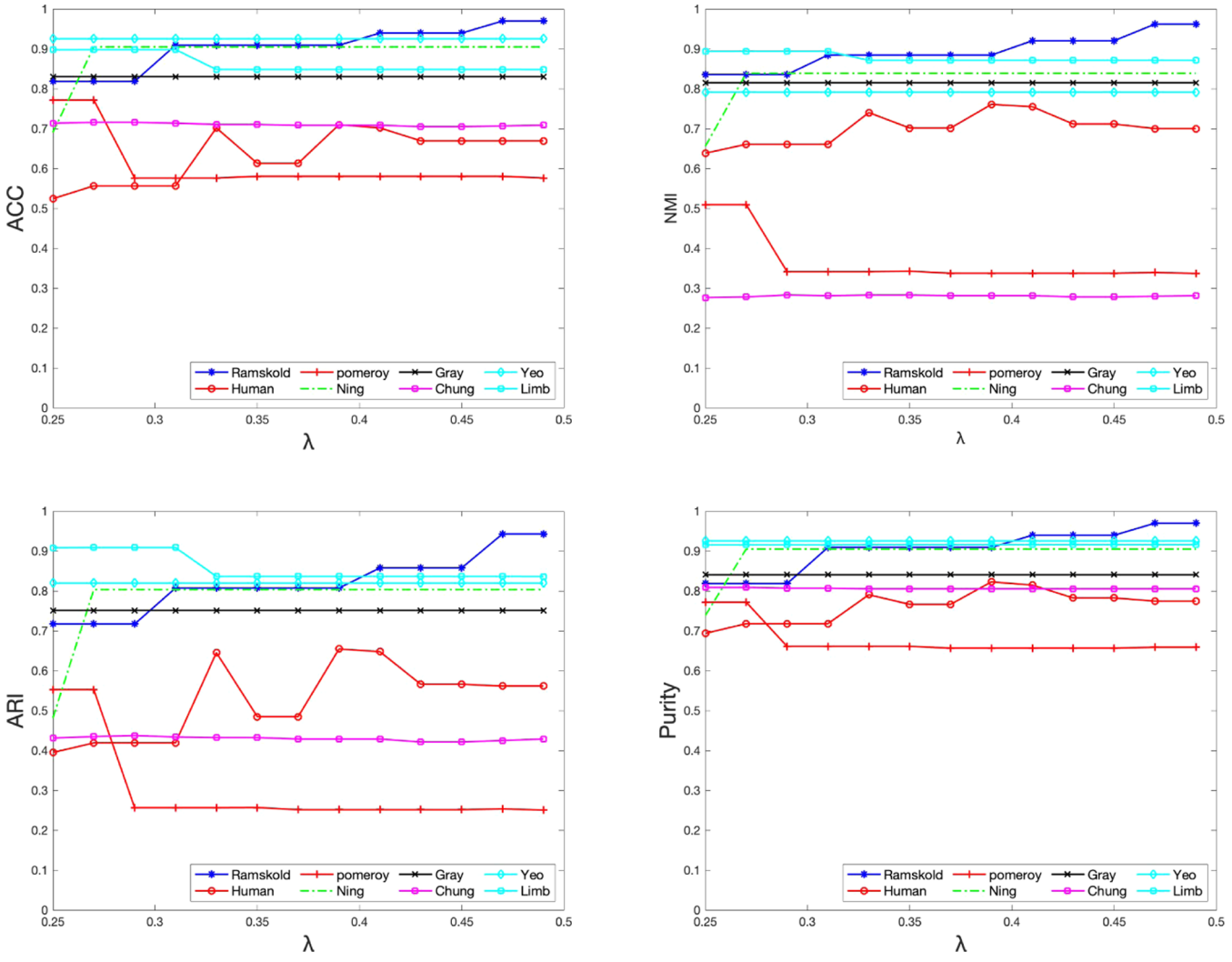


Figure 4. Clustering performance of all algorithms varied the parameter λ on all ScRNA-seq datasets.

Visualization. To more intuitively observe the clustering results of different algorithms, we projected their low representation into two-dimensional space using t-SNE technology. In this paper, we only drew a scatter diagram of

https://doi.org/10.1021/acs.jcim.2c01305
J. Chem. Inf. Model. 2022, 62, 6271−6286

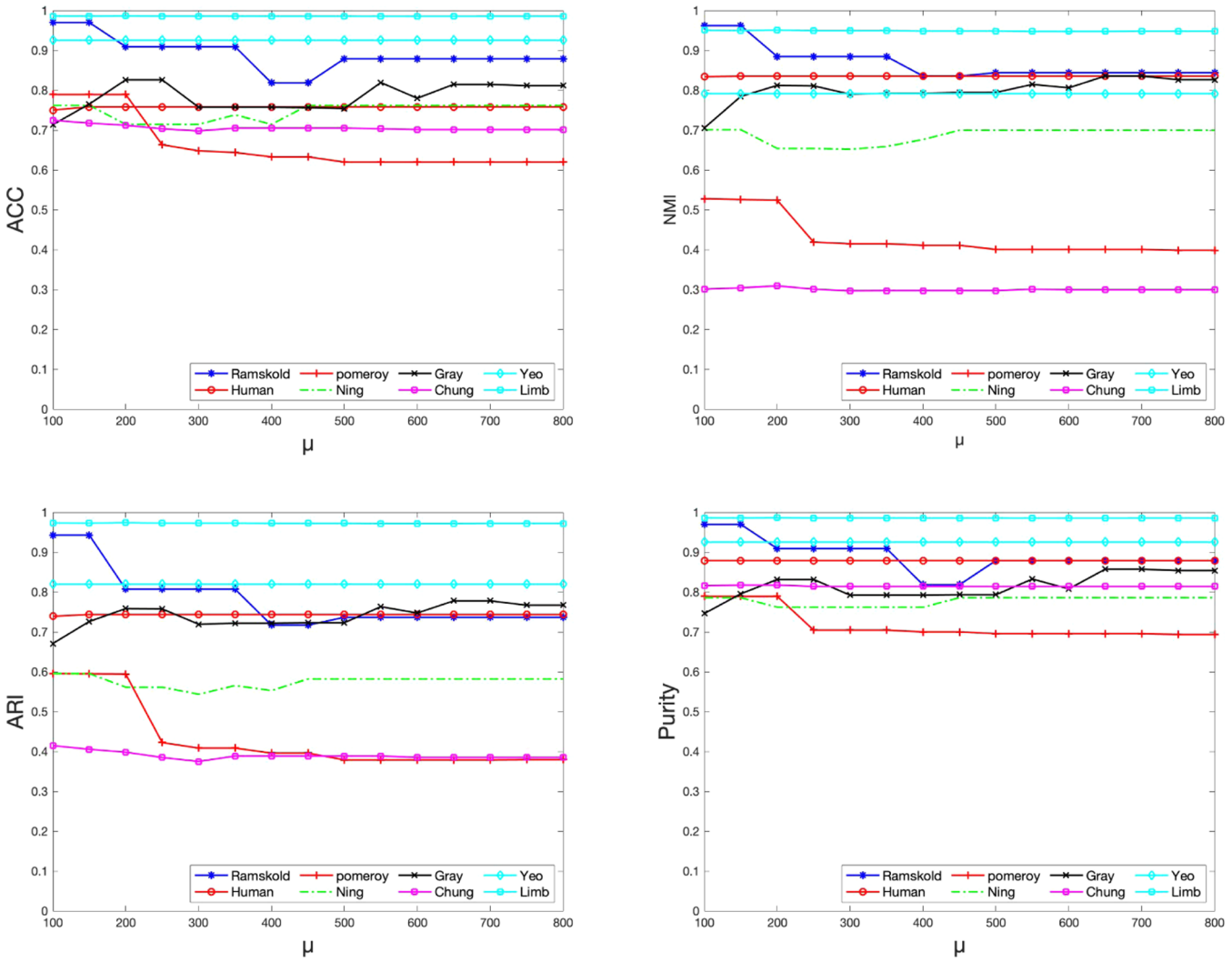


Figure 5. Clustering performances of all algorithms varied the parameter μ on all ScRNA-seq datasets.

the clustering results on Ning and Yeo datasets due to the space limited. Figures 2 and 3 show the clustering performances of different algorithms on the Ning and Yeo datasets. From Figures 2 and 3, we can intuitively observe that the proposed RGNMF-DS algorithm achieves the best clustering performances among these algorithms.

Parameter Setting. In this subsection, we carried out some experiments to assess the sensitivity of these parameters. Our proposed RGNMF-DS method includes three parameters λ, μ, and α. In each ScRNA-seq dataset, we randomly selected 20% of ScRNA-seq data samples as labeled datasets and the remaining as unlabeled samples. The values of these parameters in the proposed algorithm and other comparison algorithms were set to $[0, 0.001, 0.01, 0.1, 1, 10, 100, 1000]$, respectively. For a fair comparison, the number of nearest neighbors for all relevant models was empirically set to $p = 5$.

Specially, we varied one parameter by fixing other parameters on the ScRNA-seq datasets. Figure 4 shows the performance of the proposed model with different values of the parameter λ on all eight datasets. We can find that the clustering performances of our proposed RGNMF-DS method are relatively stable within a certain range of the parameter λ. Figure 5 shows the performances of the proposed model under different settings of parameter μ on all eight datasets. It can be observed that our proposed model achieves better performance when the parameter value range is between $[200, 800]$. Figure 6 plots the performances of the proposed RGNMF-DS model varied with the parameter α. It is clear to see that the proposed model maintains relatively stable performance within a large range of parameter α.

CONCLUSIONS

In this work, we propose a novel ScRNA-seq data representation algorithm, named RGNMF-DS, for clustering. In RGNMF-DS, both the limited label information and the intrinsic manifold structure of the ScRNA-seq data are fully utilized using the regularization technology. In addition, the reconstruction error is measured by adopting the $l_{2,1}$ norm in RGNMF-DS. The proposed model is optimized by developing an iterative updating algorithm. Meanwhile, we provide the convergence proof of the proposed RGNMF-DS algorithm. Experimental results on several ScRNA-seq datasets show that our RGNMF-DS algorithm outperforms other competitors in clustering.

https://doi.org/10.1021/acs.jcim.2c01305
J. Chem. Inf. Model. 2022, 62, 6271−6286

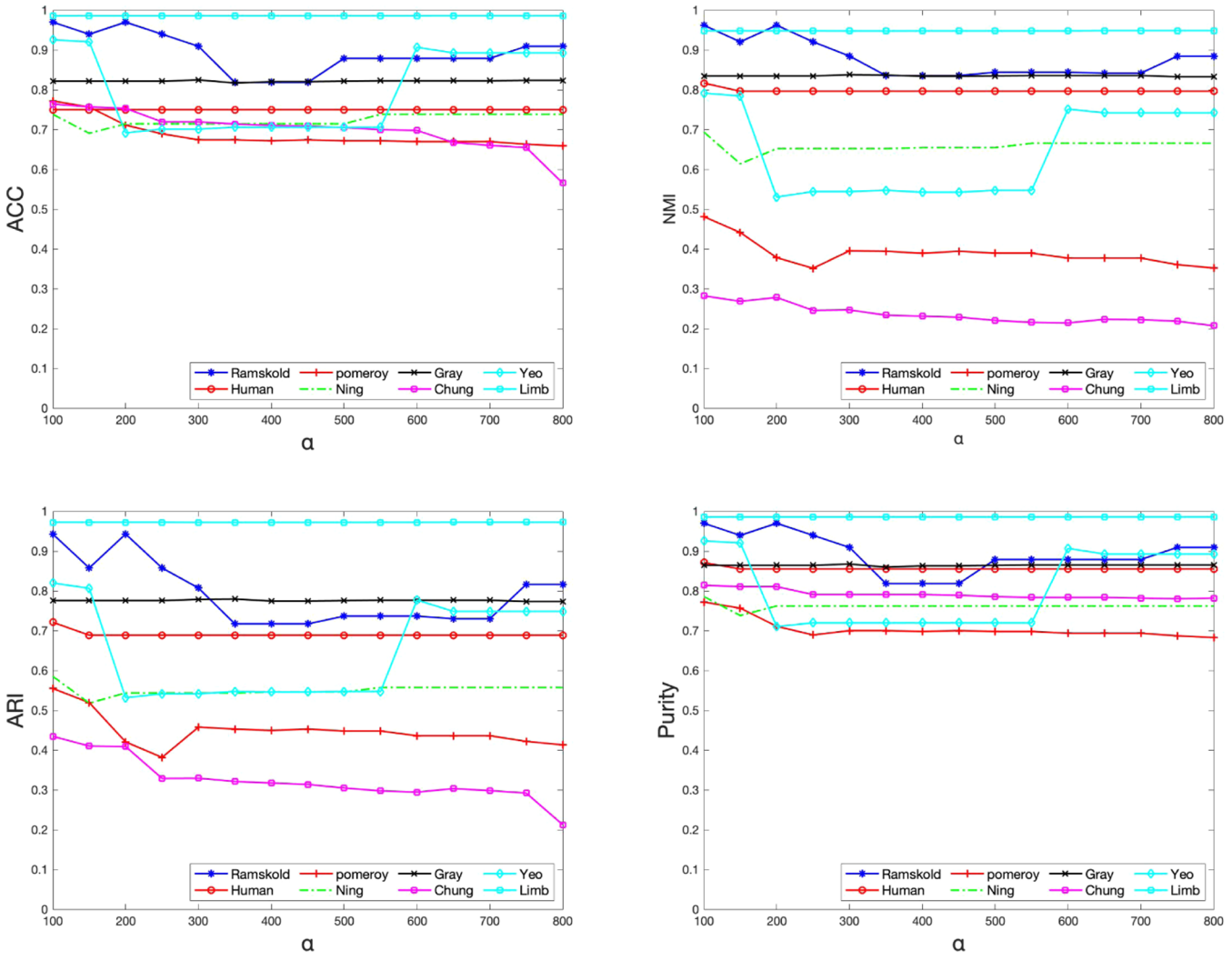


Figure 6. Clustering performances of all algorithms varied the parameter α on all ScRNA-seq datasets.

■ ASSOCIATED CONTENT

Data Availability Statement

The ScRNA-seq datasets used in this work are available for download from GEO under the following accession numbers: Ramskold (GSE38495), Human (GSE36552), Ning (GSE64016), Gray (GSE81252), Chung (GSE75688), Yeo (GSE85908), Limb (GSE109774). The source code and datasets of this work have been available on GitHub: https://github.com/szq0816/RGNMF-DS.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.2c01305.

Visualizations of our proposed RGNMF-DS on other scRNA-seq datasets (PDF).

■ AUTHOR INFORMATION

Corresponding Authors

Zhenqiu Shu − *Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650093, China;* orcid.org/0000-0001-5737-3383; Email: shuzhenqiu@163.com

Luping Zhang − *Library of Kunming Medical University, Kunming 650031, China*; Email: lupingzhangkm@126.com

Authors

Qinghan Long − *Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650093, China*

Zhengtao Yu − *Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650093, China*

Xiao-Jun Wu − *Jiangsu Provincial Engineering Laboratory of Pattern Recognition and Computational Intelligence, Jiangnan University, Wuxi 214122, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.2c01305

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Z.S. received funding from the National Natural Science Foundation of China [grant nos. 61603159, 62162033] and Yunnan Foundation Research Projects [grant nos.

202201AT070154, 202101BE070001-056]. Z.Y. received funding for equipment from Yunnan Provincial Major Science and Technology Special Plan Projects [grant nos. 202002AD080001, 202103AA080015].

REFERENCES

(1) Cole, M. B.; Risso, D.; Wagner, A.; DeTomaso, D.; Ngai, J.; Purdom, E.; Dudoit, S.; Yosef, N. Performance Assessment and Selection of Normalization Procedures for Single-cell RNA-Seq. *Cell Syst.* **2019**, *8*, 315−328.
(2) Zhang, S.; Li, X.; Lin, Q.; Wong, K.-C. Review of Single-cell RNA-seq Data Clustering for Cell Type Identification and Characterization. 2020, arXiv preprint arXiv:2001.01006 https://arxiv.org/abs/2001.01006.
(3) Ma, S.; Dai, Y. Principal Component Analysis based Methods in Bioinformatics Studies. *Briefings Bioinf.* **2011**, *12*, 714−722.
(4) Haghverdi, L.; Lun, A. T.; Morgan, M. D.; Marioni, J. C. Batch Effects in Single-cell RNA-sequencing Data are Corrected by Matching Mutual Nearest Neighbors. *Nat. Biotechnol.* **2018**, *36*, 421−427.
(5) Wang, B.; Zhu, J.; Pierson, E.; Ramazzotti, D.; Batzoglou, S. Visualization and Analysis of Single-cell RNA-seq Data by Kernel-based Similarity Learning. *Nat. Methods* **2017**, *14*, 414−416.
(6) Von Luxburg, U. A Tutorial on Spectral Clustering. *Stat. Comput.* **2007**, *17*, 395−416.
(7) Lu, C.; Yan, S.; Lin, Z. Convex Sparse Spectral Clustering: Single-view to Multi-view. *IEEE Trans. Image Process.* **2016**, *25*, 2833−2843.
(8) Butler, A.; Hoffman, P.; Smibert, P.; Papalexi, E.; Satija, R. Integrating Single-cell Transcriptomic Data Across Different Conditions, Technologies, and Species. *Nat. Biotechnol.* **2018**, *36*, 411−420.
(9) Liu, G.; Lin, Z.; Yan, S.; Sun, J.; Yu, Y.; Ma, Y. Robust Recovery of Subspace Structures by Low-rank Representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 171−184.
(10) Zheng, R.; Li, M.; Liang, Z.; Wu, F.-X.; Pan, Y.; Wang, J. SinNLRR: a Robust Subspace Clustering Method for Cell Type Detection by Non-negative and Low-rank Representation. *Bioinformatics* **2019**, *35*, 3642−3650.
(11) Lee, D. D.; Seung, H. S. Learning the Parts of Objects by Non-negative Matrix Factorization. *Nature* **1999**, *401*, 788−791.
(12) Cai, D.; He, X.; Han, J.; Huang, T. S. Graph Regularized Nonnegative Matrix Factorization for Data Representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *33*, 1548−1560.
(13) Kong, D.; Ding, C.; Huang, H.Robust Nonnegative Matrix Factorization Using l21-norm. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*; ACM Press, 2011; pp 673−682 DOI: 10.1145/2063576.2063676.
(14) Wang, D.; Gao, X.; Wang, X. Semi-supervised Nonnegative Matrix Factorization via Constraint Propagation. *IEEE Trans. Cybern.* **2016**, *46*, 233−244.
(15) Lan, W.; Chen, J.; Chen, Q. Detecting Cell Type from Single Cell RNA Sequencing based on Deep Bi-stochastic Graph Regularized Matrix Factorization. *bioRxiv* **2022**, DOI: 10.1101/2022.05.16.492212.
(16) Xiao, Q.; Luo, J.; Liang, C.; Cai, J.; Ding, P. A Graph Regularized Non-negative Matrix Factorization Method for Identifying MicroRNA-disease Associations. *Bioinformatics* **2018**, *34*, 239−248.
(17) Yu, N.; Gao, Y.-L.; Liu, J.-X.; Wang, J.; Shang, J. Robust Hypergraph Regularized Non-negative Matrix Factorization for Sample Clustering and Feature Selection in Multi-view Gene Expression Data. *Hum. Genomics* **2019**, *13*, 1−10.
(18) Liu, J.-X.; Wang, D.; Gao, Y.-L.; Zheng, C.-H.; Shang, J.-L.; Liu, F.; Xu, Y. A Joint-L2, 1-norm-constraint-based Semi-supervised Feature Extraction for RNA-Seq Data Analysis. *Neurocomputing* **2017**, *228*, 263−269.
(19) Li, X.; Wong, K.-C. Single-cell Rna Sequencing Data Interpretation by Evolutionary Multiobjective Clustering. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2019**, *17*, 1.
(20) Wang, C.; Gao, Y.-L.; Liu, J.-X.; Kong, X.-Z.; Zheng, C.-H. Single-cell RNA Sequencing Data Clustering by Low-rank Subspace Ensemble Framework. *IEEE/ACM Trans. Comput. Biol. Bioinf* **2020**, 1.
(21) Li, R.; Wang, Z.; Guan, J.; Zhou, S. Effectively Clustering Single Cell RNA Sequencing Data by Sparse Representation. *IEEE/ACM Trans. Comput. Biol. Bioinf* **2021**, 1.
(22) Shao, C.; Höfer, T. Robust Classification of Single-cell Transcriptome Data by Nonnegative Matrix Factorization. *Bioinformatics* **2017**, *33*, 235−242.
(23) Ye, Y.; Li, J. J. NMFP: a Non-negative Matrix Factorization Based Preselection Method to Increase Accuracy of Identifying mRNA Isoforms from RNA-seq Data. *BMC Genomics* **2016**, *17*, No. 11.
(24) Zhilong, J.; Zhang, X.; Guan, N.; Bo, X.; Barnes, M. R.; Luo, Z. Gene Ranking of RNA-Seq Data via Discriminant Non-Negative Matrix Factorization. *PLoS One* **2015**, *10*, No. e0137782.
(25) Kim, H.; Park, H. Sparse Non-negative Matrix Factorizations via Alternating Non-negativity-constrained Least Squares for Microarray Data Analysis. *Bioinformatics* **2007**, *23*, 1495−1502.
(26) Yang, Q.; Yin, X.; Kou, S.; Wang, Y. Robust Structured Convex Nonnegative Matrix Factorization for Data Representation. *IEEE Access* **2021**, *9*, 155087−155102.
(27) Wu, P.; An, M.; Zou, H.-R.; Zhong, C.-Y.; Wang, W.; Wu, C.-P. A Robust Semi-supervised NMF Model for Single Cell RNA-seq Data. *PeerJ* **2020**, *8*, e10091.
(28) Lee, D.; Seung, H. S.Algorithms for Non-negative Matrix Factorization. In *Advances in Neural Information Processing Systems*; MIT Press, 2001; Vol. *13*, pp 556−562.
(29) Dempster, A. P.; Laird, N. M.; Rubin, D. B. Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. R. Stat.Soc. B* **1977**, *39*, 1−22.
(30) Ramsköld, D.; Luo, S.; Wang, Y.-C.; Li, R.; Deng, Q.; Faridani, O. R.; Daniels, G. A.; Khrebtukova, I.; Loring, J. F.; Laurent, L. C.; et al. Full-length mRNA-Seq from Single-cell Levels of RNA and Individual Circulating Tumor Cells. *Nat. Biotechnol.* **2012**, *30*, 777−782.
(31) Yan, L.; Yang, M.; Guo, H.; Yang, L.; Wu, J.; Li, R.; Liu, P.; Lian, Y.; Zheng, X.; Yan, J.; et al. Single-cell RNA-Seq Profiling of Human Preimplantation Embryos and Embryonic Stem Cells. *Nat. Struct. Mol. Biol.* **2013**, *20*, 1131−1139.
(32) Tang, F.; Barbacioru, C.; Bao, S.; Lee, C.; Nordman, E.; Wang, X.; Lao, K.; Surani, M. A. Tracing the Derivation of Embryonic Stem Cells From the Inner Cell Mass by Single-cell RNA-Seq Analysis. *Cell Stem Cell* **2010**, *6*, 468−478.
(33) Pomeroy, S. L.; Tamayo, P.; Gaasenbeek, M.; Sturla, L. M.; Angelo, M.; McLaughlin, M. E.; Kim, J. Y. H.; et al. Prediction of Central Nervous System Embryonal Tumour Outcome Based on Gene Expression. *Nature* **2002**, *415*, 436−442.
(34) Leng, N.; Chu, L.-F.; Barry, C.; Li, Y.; Choi, J.; Li, X.; Jiang, P.; Stewart, R. M.; Thomson, J. A.; Kendziorski, C. Oscope Identifies Oscillatory Genes in Unsynchronized Single-cell RNA-seq Experiments. *Nat. Methods* **2015**, *12*, 947−950.
(35) Camp, J. G.; Sekine, K.; Gerber, T.; Loeffler-Wirth, H.; Binder, H.; Gac, M.; Kanton, S.; Kageyama, J.; Damm, G.; Seehofer, D.; et al. Multilineage Communication Regulates Human Liver Bud Development from Pluripotency. *Nature* **2017**, *546*, 533−538.
(36) Chung, W.; Eum, H. H.; Lee, H.-O.; Lee, K.-M.; Lee, H.-B.; Kim, K.-T.; Ryu, H. S.; Kim, S.; Lee, J. E.; Park, Y. H.; et al. Single-cell RNA-seq Enables Comprehensive Tumour and Immune Cell Profiling in Primary Breast Cancer. *Nat. Commun.* **2017**, *8*, No. 15081.
(37) Song, Y.; Botvinnik, O. B.; Lovci, M. T.; Kakaradov, B.; Liu, P.; Xu, J. L.; Yeo, G. W. Single-cell Alternative Splicing Analysis with Expedition Reveals Splicing Dynamics During Neuron Differentiation. *Mol. Cell* **2017**, *67*, 148−161.

(38) The Tabula Muris Consortium, Overall coordination, Logistical coordination; et al. Single-cell Transcriptomics of 20 Mouse Organs Creates a Tabula Muris. *Nature* **2018**, *562*, 367−372.
(39) Wagner, S.; Wagner, D. *Comparing Clusterings: an Overview*; Universität Karlsruhe Fakultät für Informatik, Karlsruhe, 2007.
(40) Cai, D.; He, X.; Han, J. Document Clustering using Locality Preserving Indexing. *IEEE Trans. Knowl. Data Eng.* **2005**, *17*, 1624−1637.
(41) Kuhn, H. W. The Hungarian Method for The Assignment Problem. *Naval Res Logist. Q.* **1955**, *2*, 83−97.
(42) Strehl, A.; Ghosh, J. Cluster Ensembles−a Knowledge Reuse Framework for Combining Multiple Partitions. *J. Mach. Learn Res.* **2002**, *3*, 583−617.
(43) Hubert, L.; Arabie, P. Comparing Partitions. *J. Classif.* **1985**, *2*, 193−218.
(44) Lloyd, S. Least Squares Quantization in PCM. *IEEE Trans. Inf. Theory* **1982**, *28*, 129−137.
(45) Linderman, G. C.; Rachh, M.; Hoskins, J. G.; Steinerberger, S.; Kluger, Y. Fast Interpolation-based T-SNE for Improved Visualization of Single-cell RNA-seq Data. *Nat. Methods* **2019**, *16*, 243−245.
(46) Nakamura, T.; Yabuta, Y.; Okamoto, I.; Aramaki, S.; Yokobayashi, S.; Kurimoto, K.; Sekiguchi, K.; Nakagawa, M.; Yamamoto, T.; Saitou, M. SC3-seq: a Method for Highly Parallel and Quantitative Measurement of Single-cell Gene Expression. *Nucleic Acids Res.* **2015**, *43*, e60.
(47) Jia, Y.; Liu, H.; Hou, J.; Kwong, S. Semisupervised Adaptive Symmetric Non-negative Matrix Factorization. *IEEE Trans. Cybern.* **2021**, *51*, 2550−2562.
(48) Wu, W.; Kwong, S.; Hou, J.; Jia, Y.; Ip, H. H. S. Simultaneous Dimensionality Reduction and Classification via Dual Embedding Regularized Nonnegative Matrix Factorization. *IEEE Trans. Image Process.* **2019**, *28*, 3836−3847.
(49) Seal, D. B.; Das, V.; De, R. K. CASSL: A Cell-type Annotation Method for Single Cell Transcriptomics Data Using Semi-supervised Learning. *Appl. Intell.* **2022**, 1−19, DOI: 10.1007/s10489-022-03440-4.

https://doi.org/10.1021/acs.jcim.2c01305
J. Chem. Inf. Model. 2022, 62, 6271−6286