JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

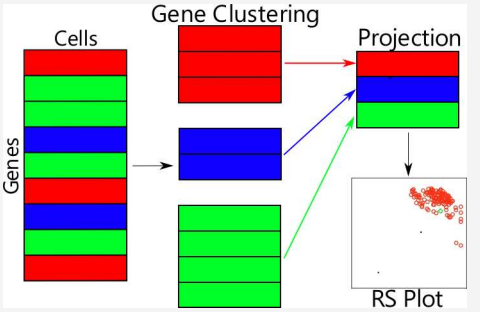
Preprocessing of Single Cell RNA Sequencing Data Using Correlated Clustering and Projection

Yuta Hozumi, Kiyoto Aramis Tanemura, and Guo-Wei Wei*

Cite This: J. Chem. Inf. Model. 2024, 64, 2829−2838 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Single-cell RNA sequencing (scRNA-seq) is widely used to reveal heterogeneity in cells, which has given us insights into cell−cell communication, cell differentiation, and differential gene expression. However, analyzing scRNA-seq data is a challenge due to sparsity and the large number of genes involved. Therefore, dimensionality reduction and feature selection are important for removing spurious signals and enhancing the downstream analysis. We present Correlated Clustering and Projection (CCP), a new data-domain dimensionality reduction method, for the first time. CCP projects each cluster of similar genes into a supergene defined as the accumulated pairwise nonlinear gene−gene correlations among all cells. Using 14 benchmark data sets, we demonstrate that CCP has significant advantages over classical principal component analysis (PCA) for clustering and/or classification problems with intrinsically high dimensionality. In addition, we introduce the Residue-Similarity index (RSI) as a novel metric for clustering and classification and the R-S plot as a new visualization tool. We show that the RSI correlates with accuracy without requiring the knowledge of the true labels. The R-S plot provides a unique alternative to the uniform manifold approximation and projection (UMAP) and t-distributed stochastic neighbor embedding (t-SNE) for data with a large number of cell types.



1. INTRODUCTION

Single cell RNA sequencing (scRNA-seq) reveals heterogeneity within cell types, leading to an understanding of cell−cell communication, cell differentiation, and differential gene expression. With current technology and protocols, more than 20,000 genes can be identified. Numerous data analysis pipelines have been developed to help analyze such complex data.[1−6] Despite improvements in technology that allow for a more accurate reading of genes, the analysis of gene readings remains challenging. Causes of this challenge include dropout event-induced zero expression counts, low sequencing depth leading to low reading counts, general noise, and the high dimensionality of the original data.[7] As a result, dimensionality reduction and feature selection are important for downstream analysis such as removing spurious signals.

Numerous dimensionality reduction and feature selection methods have been proposed for the scRNA-seq data. One such method is ScRNA by non-negative and low-rank representation (SinLRR), which assumes that scRNA-seq has an inherently low rank and attempts to find the smallest rank matrix that captures the original data.[8] Numerous non-negative matrix factorization (NMF) methods with different constraints have also been developed, where the low-dimensional representation of scRNA-seq is a linear combination of the original data and acts as meta-genes.[9−14] Single-cell interpretation via multikernel learning (SIMLR) utilizes multiple kernels to learn a cell−cell similarity metric that generalizes to different biological experiments and experimental procedures.[15] In addition, more traditional approaches, such as principal component analysis (PCA)[16] and its derivatives,[17,18] and visualization techniques, such as uniform manifold approximation and projection (UMAP)[19] and t-distributed stochastic neighbor embedding (t-SNE),[20] have been heavily utilized for scRNA-seq data. Furthermore, deep learning has also been used for dimensionality reduction.[21−26]

Although numerous techniques have been developed, PCA is the most commonly used method for downstream analysis of scRNA-seq data.[27] PCA is a linear dimensionality reduction method, where its goal is to compute the principal components as new features that maximize the variance. The first principal component is a feature that maximizes the variance of the projected data, and each ith principal component is orthogonal to the $i - 1$ principal component that maximizes the variance of the projected data.[28] Single-cell consensus clustering (SC3)[29] utilizes PCA and the eigenvectors of the graph Laplacian induced by Euclidean, Pearson, and Spearman distances and performs a consensus on k-means results obtained from different dimensions using the CSPA algorithm

Special Issue: Machine Learning in Bio-cheminformatics

Received: May 5, 2023
Published: July 4, 2023

https://doi.org/10.1021/acs.jcim.3c00674
J. Chem. Inf. Model. 2024, 64, 2829−2838

to obtain the final cell clustering result. CellChat[30] utilizes the low-dimensional representation of scRNA-seq alongside known interactions between ligands, receptors, and cofactors to predict cell−cell communication, and a user can perform dimensionality reduction prior to utilizing CellChat. DEEPsc[31] is a deep learning method that predicts the probability of a cell belonging to a reference atlas by projecting scRNA-seq to the PCA space of the reference atlas, which can then be used to predict cell types. The popular package Seurat[32] utilizes supervised PCA (SPCA) which finds the projection that captures the weighted nearest neighbor graph of the reference data set for its downstream analysis. In addition to cell clustering, semisupervised and supervised learning methods have been used to classify cell types according to their reference cells by projecting unknown cells to the PCA space of the reference cells.[33,34]

PCA has many advantages, such as computational efficiency and ease in projecting new data into the principal components. However, PCA lacks concrete interpretability and loses the non-negativity of the read-count data. In contrast, the components of NMF are all positive and can be considered metagenes, where metagenes are linear combinations of the original genes. Nonlinear dimensionality reduction methods, such as UMAP, t-SNE, and Isomap, have great performance for low dimensionality that can capture the local structure of the data, but they also lack interpretability due to matrix diagonalization. Moreover, both PCA and traditional nonlinear reduction methods are unstable when the data are reduced to higher dimensions, which is unfavorable for machine learning and deep learning tasks that typically require a large number of features.

We propose a computationally efficient and interpretable dimensionality reduction algorithm for scRNA-seq data called correlated clustering and projection (CCP).[35] CCP begins by clustering genes based on their similarity and then uses the flexibility rigidity index (FRI)[36] to nonlinearly project each gene cluster into a supergene, which is a measure of accumulated gene−gene correlations among cells. Unlike traditional nonlinear reduction methods, CCP bypasses matrix diagonalization, allowing users to select the number of supergenes, which is beneficial for machine learning and deep learning tasks. Furthermore, similar to NMF's metagenes, supergenes are all nonnegative and highly interpretable. We validated CCP's performance on 14 scRNA-seq data sets by varying the number of supergenes and conducting support vector machine classification and *k*-means clustering.

Additionally, we have validated the performance of a novel evaluation metric for dimensionality reduction, called the Residue-Similarity index (RSI).[35] The RSI evaluates the intracluster similarity of cell types or clusters and compares it to their intercluster residual score. As the RSI only requires one set of labels, which can be computed from *k*-means, it can measure the performance of dimensionality reduction for both clustering and classification tasks, without requiring knowledge of the true labels. Furthermore, by analysis of the relationship between samples, the RSI allows for a deeper understanding of the quality of the dimensionality reduction algorithm. We have verified the effectiveness of the RSI alongside CCP on both clustering and classification tasks and introduced the R-S plot as a novel visualization technique for data containing multiple cell types.

2. METHOD

2.1. Correlated Clustering and Projection (CCP). The CCP procedure consists of two steps: gene partitioning and gene projection. Let $\mathcal{Z} \in \mathbb{R}^{M\times I}$ be the log-transformed scRNA-seq data, where M is the number of samples (cells), and I is the number of genes.

2.1.1. Feature Partitioning. The original CCP method used a modified *k*-medoids algorithm for gene clustering; however, we replaced it with a modified *k*-means algorithm for a more stable clustering result. The details of the modified *k*-means clustering method can be found in Section S1.1 of the Supporting materials.

Let $\mathcal{Z} = \{\mathbf{z}^1, ..., \mathbf{z}^i, ..., \mathbf{z}^I\}$ be the rows of $\mathcal{Z}$ or the gene vector, and $\mathbf{z}^i \in \mathbb{R}^M$. CCP implements *k*-means clustering described in S1.1, but the clustering is done on the genes. Hence, we get clusters $Z^1, ..., Z^N$, $\mathcal{Z} = \uplus_{n=1}^{N} Z^n$, $N \ll I$.

Let $S = \{1, ..., I\}$ be the enumeration of the original genes. Then, we can partition $S = \{S^1, ..., S^N\}$, using the *k*-means clustering results, by setting $S^n = \{i | \mathbf{z}^i \in Z^n\}$, i.e., S^n is the number of genes in the nth cluster.

2.1.2. Feature Projection. With the gene partitioning, we define $\mathbf{z}_m^{S^n} \in \mathbb{R}^{S^n}$ as the S^n genes in the mth cell. These genes are projected into a supergene x_m^n using the flexibility rigidity index (FRI). Denote $\left\|\mathbf{z}_i^{S^n} - \mathbf{z}_j^{S^n}\right\|$ as some metric between cell i and cell j for the cluster of S^n genes, and the gene−gene correlation between the two cells are defined by $C_{ij}^{S^n} = \Phi(\left\|\mathbf{z}_i^{S^n} - \mathbf{z}_j^{S^n}\right\|; \eta^{S^n}, \tau, \kappa)$, where Φ is the correlation kernel, and η^{S^n}, τ and $\kappa > 0$ are parameters. Commonly used metrics include the Euclidean, Manhattan, and Wasserstein distances. In addition, the correlation kernels satisfy the following conditions

$$\Phi(\left\|\mathbf{z}_i^{S^n} - \mathbf{z}_j^{S^n}\right\|; \eta^{S^n}, \tau, \kappa) \to 0, \quad \text{as} \left\|\mathbf{z}_i^{S^n} - \mathbf{z}_j^{S^n}\right\| \to \infty \tag{1}$$

$$\Phi(\left\|\mathbf{z}_i^{S^n} - \mathbf{z}_j^{S^n}\right\|; \eta^{S^n}, \tau, \kappa) \to 1, \quad \text{as} \left\|\mathbf{z}_i^{S^n} - \mathbf{z}_j^{S^n}\right\| \to 0 \tag{2}$$

Commonly used kernel functions are the radial basis functions. In particular, we use the generalized exponential function

$$\Phi(\left\|\mathbf{z}_i^{S^n} - \mathbf{z}_j^{S^n}\right\|; \eta^{S^n}, \tau, \kappa) = \begin{cases} e^{-\left(\frac{\left\|\mathbf{z}_i^{S^n} - \mathbf{z}_j^{S^n}\right\|}{\eta^{S^n}\tau}\right)^{\kappa}} & \left\|\mathbf{z}_i^{S^n} - \mathbf{z}_j^{S^n}\right\| < r_c^{S^n} \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

where $r_c^{S^n}$ is the cutoff distance, and η^{S^n} is the scale, which are defined by the data. κ is the power, and τ is a scale parameter.

Pairwise gene−gene correlation matrix $C^{S^n} = \{C_{ij}^{S^n}\}$ reveals cell−cell interactions and can also be viewed mathematically as the weight of the edges in a weighted graph, given the cutoff $r_c^{S^n}$. The cutoff $r_c^{S^n}$ is taken as the 2-standard deviations of the pairwise distances. η^{S^n} can then be viewed as the algebraic connectivity, which is defined as the average minimal distance between the cluster of genes

https://doi.org/10.1021/acs.jcim.3c00674
J. Chem. Inf. Model. 2024, 64, 2829−2838

Table 1. Accession ID, Source Organism, and the Counts for Samples, Genes, Cell Types, and Normalization for 14 Data Sets

Accession ID	Reference	Organism	Samples	Genes	Cell types	Normalization
GSE45719	Deng[40]	Mouse	300	22431	8	RPKM
GSE59114	Kowalczyk[41]	Mouse	1428	8422	6	TPM
GSE67835	Darmanis[42]	Human	420	22084	8	CPM
GSE75748 cell	Chu[43]	Human	1018	19097	7	TPM
GSE75748 time	Chu[43]	Human	758	19189	6	TPM
GSE82187	Gokce[44]	Mouse	705	18840	10	TPM
GSE84133 h1	Baron[45]	Human	1937	20125	14	TPM
GSE84133 h2	Baron[45]	Human	1724	20125	14	TPM
GSE84133 h3	Baron[45]	Human	3605	20125	14	TPM
GSE84133 h4	Baron[45]	Human	1308	20125	14	TPM
GSE84133 m1	Baron[45]	Mouse	822	14878	13	TPM
GSE84133 m2	Baron[45]	Mouse	1064	14878	13	TPM
GSE89232	Breton[46]	Human	957	20689	4	TPM
GSE94820	Villani[47]	Human	1140	26593	5	TPM

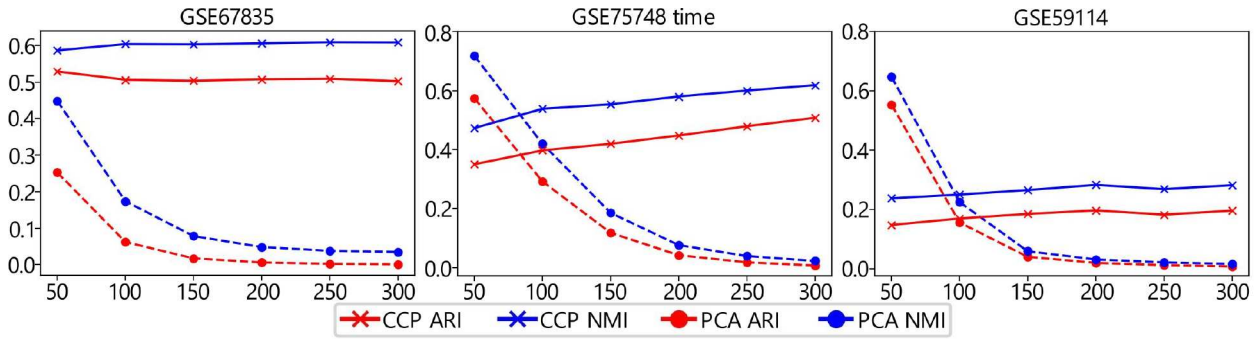


Figure 1. ARI and NMI of the clustering results of CCP and PCA on GSE67835 and GSE75748 time and GSE59114 data. The red and blue lines correspond to CCP and PCA, respectively. A total of 20 random initializations were used to test the reduction, and for each reduction, a total of 30 random initializations were used to obtain the clustering results from k-means clustering. The averages of the ARI and NMI were obtained. For CCP, all the tests utilize $\tau = 6$ and $\kappa = 2$ for the exponential kernel.

$$\eta^{S^n} = \frac{\sum_{m=1}^{M} \min_{\mathbf{z}_j^{S^n}} \left\| \mathbf{z}_m^{S^n} - \mathbf{z}_j^{S^n} \right\|}{M} \tag{4}$$

Using the correlation function, we can project S^n genes into a supergene using the FRI for the ith sample

$$x_i^n = \sum_{m=1}^{M} \Phi\left(\left\| \mathbf{z}_i^{S^n} - \mathbf{z}_m^{S^n} \right\|; \eta^{S^n}, \tau, \kappa \right) \tag{5}$$

By performing the projection of all gene clusters, we get the lower dimensional supergene representation for the ith sample (cell) $\mathbf{x}_i = (x_i^1, ..., x_i^N)^T$.

2.2. Evaluation Metric. In this section, we introduce the Residue Similarity Index (RSI) and its scores. Details on the Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), Balanced Accuracy (BA), and Silhouette score can be found in Section S1.2 of the Supporting materials.

2.2.1. Residue-Similarity Index and Scores. We present the residue score (R score), similarity score (S score), and R-S index (RSI).[35] Let the data be represented as $\{(\mathbf{x}_m, y_m) | \mathbf{x}_m \in \mathbb{R}^N, y_m \in \mathbb{Z}_L, 1 \le m \le M\}$, where $\mathbf{x}_m$ is the mth gene vector, y_m is the label or the cluster assignment, and L is the number of classes or clusters. Assume that there is a partition of data $\mathcal{X}$ according to the labels or cluster assignments. That is, $C_l = \{\mathbf{x}_m \in \mathcal{X} | y_m = l\}$ and $\uplus_0^{L-1} \mathcal{C}_l = \mathcal{X}$.

The R score is defined as the interclass sum of distance. For a given data $\mathcal{X}$ with assignment $y_m = l$, the R-score is defined as

$$R_m = R(\mathbf{x}_m) = \frac{1}{R_{\max}} \sum_{\mathbf{x}_j \notin C_l} \left\| \mathbf{x}_m - \mathbf{x}_j \right\| \tag{6}$$

where $R_{\max} = \max_{\mathbf{x}_m, \mathbf{x}_m \in \mathcal{X}} R_m$. The similarity (S) score is defined as the intraclass average of distance, defined as

$$S_m = S(\mathbf{x}_m) = \frac{1}{|C_l|} \sum_{\mathbf{x}_j \in C_l} \left(1 - \frac{\left\| \mathbf{x}_m - \mathbf{x}_j \right\|}{d_{\max}} \right) \tag{7}$$

where $d_{\max} = \max_{\mathbf{x}_i, \mathbf{x}_j \in \mathcal{X}} \left\| \mathbf{x}_i - \mathbf{x}_j \right\|$, and $|C_l|$ is the number of data in class C_l. Both R_m and S_m are bounded by 0 and 1, and the larger the better for a given data set.

The class residue index (CRI) and the class similarity index (CSI) can then be defined as the average of the R-score and S-score of each of the classes. That is $\mathrm{CRI}_l = \frac{1}{|C_l|} \sum_m R_m$, and $\mathrm{CSI}_l = \frac{1}{|C_l|} \sum_m S_m$. Then, the residue index (RI) and the similarity index (SI) can be defined as $\mathrm{RI} = \frac{1}{L} \mathrm{CRI}_l$ and $\mathrm{SI} = \frac{1}{L} \mathrm{CSI}_l$, respectively.

Using the RI and SI, the residue similarity disparity can be computed by taking RSD = RI − SI, and the residue-similarity index (RSI) can be computed as RSI = 1 − |RI − SI|.

https://doi.org/10.1021/acs.jcim.3c00674
J. Chem. Inf. Model. 2024, 64, 2829−2838

3. RESULTS

CCP was benchmarked against PCA on 14 data sets, and the data set details can be found in Table 1. The data was normalized using either reads per kilobase of transcript per million (RPKM), transcript per million (TPM), or counts per million (CPM). For each data set, CCP was used to obtain the number of supergenes as N = 50, 100, 150, 200, 250, and 300. The parameters κ and τ of the exponential kernel were searched over κ = 1, 2 and τ = 1, 2, ..., 6 and set to τ = 6 and κ = 2 for the exponential kernel. To test the reduction, 20 random seeds were used for CCP and PCA, and for each reduction, 30 random initializations of k-means were used to obtain cluster labels. After the cluster labels were obtained, the ARI and NMI were computed by comparing the results to the labeled cell types, and the averages were visualized. For each figure, the red and blue lines represent CCP and PCA, respectively, and the star and dotted markers indicate the ARI and NMI, respectively.

3.1. CCP Benchmark. Figure 1 shows the performance of CCP and PCA on three data sets, GSE67835, GSE75748 time, and GSE59114 data. For GSE67835, CCP outperforms PCA in all of the dimensions we have tested. For GSE75748 time, CCP outperforms PCA for 50 supergenes and above. GSE75748 time shows an increase in performance as the number of gene dimensions increased. PCA exhibits instability as N increases, which is noticeable from their decrease in performance from N = 50 to 150 for both data sets. CCP does not perform well on GSE59114 because both ARI and NMI are less than 0.3 for all the dimensions we have tested. CCP's performance may be poor due to the low intrinsic dimensionality of GSE59114. In other words, the number of gene clusters is inherently small, leading to redundant clusters. GSE59114, in particular, only has 8,422 genes, whereas other data have over 15,000 genes.

In order to verify CCP's performance, the residue similarity index (RSI) was calculated for the k-means clustering result of the gene partitioning in CCP. Figure 2 shows the RSI of the k-means clustering on the genes at various numbers of cell clusters (k). The top row shows the clustering result for GSE59114, which had poor CCP performance, and the bottom row shows the clustering result for GSE67825, which had good CCP performance. For each number of clusters, 10 random initializations were used for the k-means clustering, and the averages of the RI, SI, and RSI were obtained. The red, blue, and green lines correspond to the RI, SI, and RSI, respectively. The RSI can be used to check the quality of the clustering, where the peak in the RSI suggests the optimal number of clusters and, in the case of CCP, the intrinsic dimensionality of the data. The right column shows the 2D visualization of the genes using t-SNE. The samples were colored according to their cluster labels. The t-SNE visualization of GSE59114 shows the k-means clustering result when k = 8 was selected. The t-SNE visualization of GSE67835 shows the k-means clustering result when k = 64 was selected. Seven of the 64 clusters were colored, and the points colored in green are the rest of the genes.

Notice that in GSE59114, there is a noticeable peak in the RSI score at k = 8 clusters, whereas in GSE67835, the peak is flat and occurs at about k = 32−64 clusters. This suggests that the intrinsic dimensionality is about 8 for GSE59114, which is unfavorable for CCP. On the other hand, the intrinsic dimension of GSE67835 is much higher, which is more

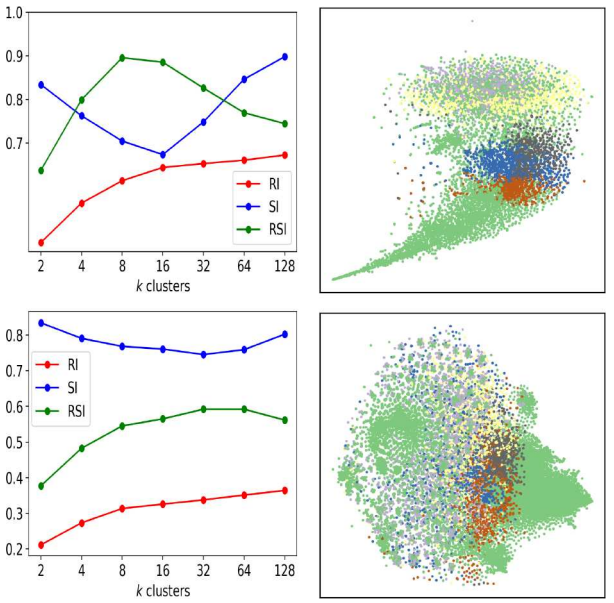


Figure 2. RI, SI, and RSI of the gene clustering of GSE59114 and GSE67835. k-means clustering was performed with k = 2, 4, 8, 16, 32, 64, and 128 gene clusters. For each number of clusters, 10 random initializations were utilized, and the averages of the RI, SI, and RSI were obtained. The red, blue, and green lines correspond to the RI, SI, and RSI, respectively. We use t-SNE to visualize the genes in 2D. For GSE59114, k = 8 clusters were obtained, and the genes were colored according to their cluster assignment. For GSE67835, k = 64 cell types were obtained. Seven random gene clusters were colored, and the rest of the clusters were colored in green.

suitable for CCP. Notice that the clusters have distinct boundaries, supporting the relatively low dimensionality of the data. On the other hand, the GSE67835 data are not well-clustered even at k = 64. Notice that the orange and blue genes have some outliers, and the purple genes are not well-clustered. This suggests that the number of optimal gene clusters is larger, which suggests high gene dimensionality and favors CCP.

3.2. Residue-Similarity Index Comparison. The residue-similarity index (RSI) has been shown to correlate with classification accuracy in ref 35. In this section, we use the RSI for classification and clustering on the 14 data sets from Table 1. We use CCP to process each data set with the same parameters as the previous section with 20 random initializations. For classification, we use 5-fold cross-validation with 10 random seeds and the support vector machine to predict cell types. We used balanced accuracy (BA) to measure the performance of the classification. Then, using the same 5-fold cross-validation, we calculate the RSI, where we obtain the RI, SI, and RSI from the test set, similar to ref 35. For clustering, we compute the RSI for PCA and CCP using the k-means clustering labels and the true labels. Additionally, using the k-means clustering labels, we compute the Silhouette score to compare the results with the RSI. Full details of the benchmark procedure can be found in S2.1 of the Supporting Materials.

In general, we have found no correlation between the Silhouette scores and RSI for clustering results. Additionally, we have found that BA and the RSI correlate in classification results.

We found that the RSI correlates with the classification accuracy in many of our tests. Figure 3 shows the RSI for

https://doi.org/10.1021/acs.jcim.3c00674
J. Chem. Inf. Model. 2024, 64, 2829−2838

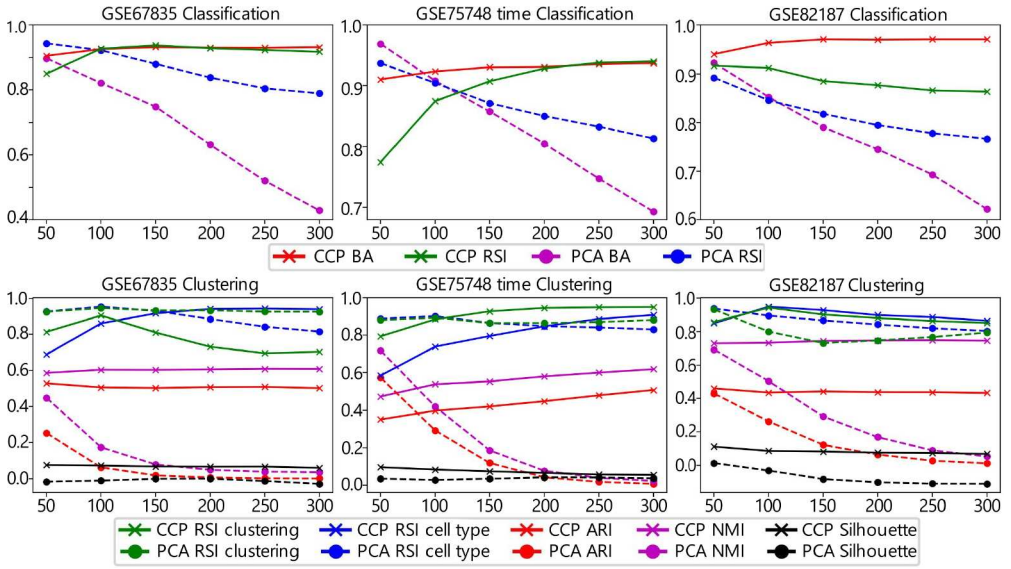


Figure 3. Comparison of the RSI in classification and clustering problems for GSE67835, GSE75748 time, and GSE82187 data at reduced dimensions N = 50, 100, 150, 200, 250, and 300. CCP was used to reduce the original data dimension using τ = 6 and κ = 2 for the exponential kernel. The top and bottom rows correspond to the classification and clustering results, respectively. For classification, a support vector machine was used. True labels were used to compute the RSI for the 5-fold cross-validation. For clustering, the RSI were computed using the cluster labels from k-means clustering and the true labels.

classification and clustering problems for GSE67835, GSE75748 time, and GSE82187 data. CCP was used to reduce the original data using τ = 6 and κ = 2 for the exponential kernel. The top row corresponds to classification results, and the bottom row corresponds to clustering results. Notice that for classification results, all three data sets show a correlation between BA and the RSI. The RSI on classification results for GSE67835 shows a plateau at about 150 supergenes, which corresponds to the plateau of the BA. This suggests that the optimal dimension is about 150. The RSI on classification results for GSE75748 shows a plateau at about 200 supergenes, even though BA plateaus at about 150 supergenes. Even though the accuracy plateaued earlier, this suggests that the optimal dimension is 200 gene clusters. In addition, since GSE75748 time observes cell differentiation at different times, it is possible that some cells are at different stages in their cell cycles, as suggested in the literature.[43] This suggests that there are many intermediate stages in the cell differentiation. The RSI on classification results on GSE82187 shows a small decrease as the number of supergenes increases. This suggests that the optimal dimension is smaller than those of the GSE67835 and GSE75748 time. Lastly, the RSI decreases for all three data sets when PCA is utilized, which corresponds to the decrease in BA.

For the clustering results, the RSI using the k-means labels and the true cell types are similar. Even though the ARI and NMI of PCA decrease as the number of gene clusters increases, the RSI remains consistent. This suggests that PCA cannot differentiate clusters at higher dimensions. CCP, on the other hand, shows a correlation with both of the RSI scores.

Additional examples of utilizing the RSI on classification and clustering problems can be found in Section S2.2 of the Supporting materials.

Figure 4 shows the overall clustering performance of CCP and PCA. The bars show the mean ARI and NMI values across the different numbers of components. Notice that for both ARI and NMI, CCP significantly outperforms PCA.

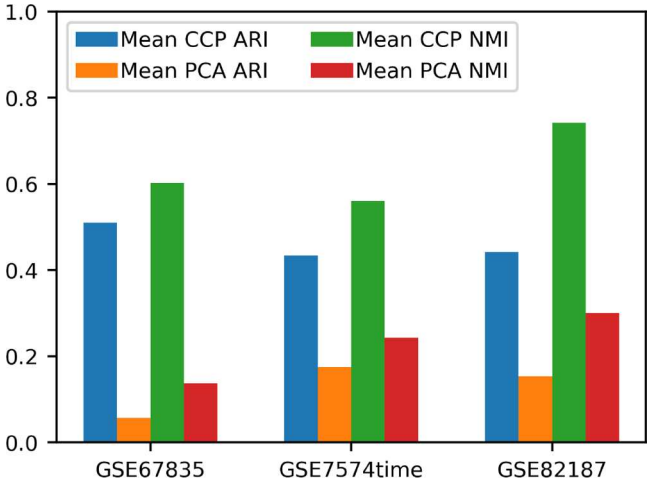


Figure 4. Comparison of CCP and PCA clustering on GSE67835, GSE75748 time, and GSE82187 data. CCP was used to reduce the original data dimension using τ = 6 and κ = 2 for the exponential kernel. The blue, orange, green, and red bars correspond to mean CCP ARI, mean PCA ARI, mean CCP NMI, and mean PCA NMI, respectively. Here, the average was taken over different dimensions.

Figure 5 shows the overall classification performance of CCP and PCA. The bars show the mean BAs across different numbers of dimensions. Notice that for the mean BA, CCP significantly outperforms PCA.

4. DISCUSSION

4.1. CCP. Like other dimensionality reduction algorithms, CCP has its advantages and disadvantages. CCP nonlinearly projects each cluster of similar genes into a supergene. Supergenes are highly interpretable: each supergene represents a measure of a cluster of genes' accumulated pairwise nonlinear correlations with the same cluster of genes in all other cells for a given cell. Similar to NMF, supergenes are non-negative,

https://doi.org/10.1021/acs.jcim.3c00674
J. Chem. Inf. Model. 2024, 64, 2829−2838

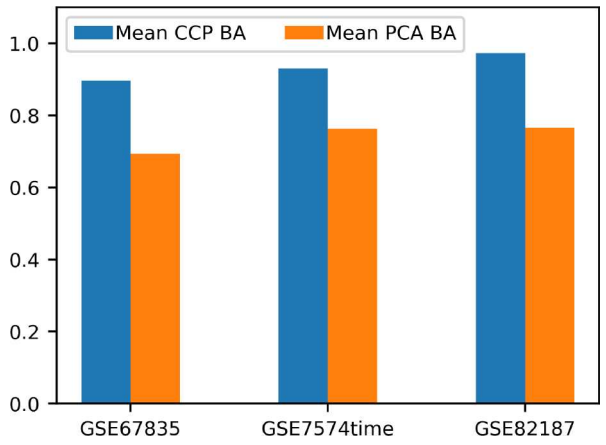


Figure 5. Comparison of CCP and PCA classification on GSE67835, GSE75748 time, and GSE82187 data. CCP was used to reduce the original data dimension using $\tau = 6$ and $\kappa = 2$ for the exponential kernel. The blue and orange bars correspond to the mean BAs of CCP and PCA, respectively. Here, the average was taken over different dimensions.

which is important for downstream analysis such as differential gene expression analysis.

Since CCP is a data-domain method, it bypasses matrix diagonalization. One limitation of many dimensionality reduction algorithms is their dependence on matrix diagonalization. In scRNA-seq data, the number of genes is typically larger than 5,000, which gives rise to the "curse of dimensionality". When the number of features is large, every sample may appear to be equidistant from one another, which makes many machine learning algorithms unable to find meaningful clusters in the data. CCP, on the other hand, partitions the genes into clusters and computes the pairwise gene−gene correlations across all cells, which avoids the curse of dimensionality.

Even though CCP has shown success in many scRNA-seq data sets, it does have limitations. CCP does not perform well for data sets with a low intrinsic dimension. As shown in Figure 2, GSE59114 and GSE94228 have a low intrinsic dimension, and as a result, their clustering results also suffered.

In addition, many scRNA-seq data sets are sparse due to low signal-to-noise ratio and dropout events. Therefore, CCP will most likely benefit from data imputation.

4.2. RSI. The RSI is a useful tool for assessing the performance of dimensionality reduction for both clustering and classification problems. In the following section, we compare the RSI to the traditional clustering metrics, ARI and NMI, and also to the Silhouette score. Then, we discuss the RSI and its connection with classification accuracy.

4.2.1. RSI for Clustering. Compared with the ARI and NMI, which measure the similarity between two sets of labels, the RSI evaluates the performance using only one set of labels. In this study, the ARI and NMI were used to compare the true labels with the clustering labels. However, in practice, such true labels are not available. The RSI, on the other hand, can evaluate the effectiveness of clustering without the need for original labels. This is similar to the Silhouette score, which measures the separations between clusters. However, when there are multiple clusters, the Silhouette score becomes difficult to interpret because it measures whether a sample belongs to its current cluster assignment or to the nearest neighboring cluster. Therefore, it is often used to evaluate the optimal number of clusters rather than evaluating different parameters while fixing the number of clusters. The RSI can evaluate the effectiveness of different parameters while fixing the number of clusters.

4.2.2. RSI for Classification. Using the RSI for cell types, we have shown that the RSI correlates with classification accuracy. Additionally, the RI and SI indicate how well the clusters separate from each other. The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) is a metric commonly used to evaluate classification effectiveness. However, the AUC-ROC is a better metric for binary classification problems, and its interpretation is more challenging for multiclass problems. The RSI, on the other hand, can handle problems with more than two cell types. Lastly, the RSI uses the features and labels to compute the scores; therefore, it can also demonstrate the effectiveness of dimensionality reduction algorithms in conjunction with classification problems.

The RSI can also be utilized for visualizing each class or cluster, which we have called a residue-similarity (R-S) plot. In order to showcase the R-S plot, we compare it with traditional visualization techniques used in scRNA-seq data, namely, t-SNE and UMAP. CCP was used to reduce the dimensionality. The 5-fold cross-validation was used to divide the data into 5 parts, where 4 parts were used to train the support vector machine classifier and 1 part was used to test the classifier. Then, residue and similarity scores were computed for each sample and plotted according to their true cell type. Samples were then colored according to their predicted labels from the support vector machine classifier. The x-axis and y-axis correspond to residue and similarity scores, respectively. Both residue and similarity scores range from 0 to 1, where 1 is the most optimal, and the top-right corner indicates well-separated and clustered reduction. However, it is important to note that having a balance of both scores is important, as shown in Hozumi et al. (2022).[35] For t-SNE and UMAP, the original data were log-transformed, and genes with variance less than 10^{-6} were removed prior to the reduction. Samples were then plotted and colored according to their cell types.

Figure 6 shows a comparison between the R-S and 2D plots of UMAP and t-SNE for the GSE75748 time data. CCP was used to generate 200 supergenes with $\tau = 6$ and $\kappa = 2$. For the UMAP and t-SNE plots, the reduction was directly applied to the log-transformed original data. In ref 43, Chu obtained snapshots at different times of embryonic stem (ES) differentiation from pluripotency to definitive endoderm (ED) over 4 days at 0, 12, 24, 36, 72, and 96 h. Noticeably, cells recorded at 72 and 96 h are mixed in UMAP and t-SNE plots and misclassified in the R-S plot. This finding is consistent with ref 43, where cells from 72 and 96 h were relatively homogeneous. In a biological sense, this may indicate that cell differentiation had mostly completed by 72 h, such that not much of the further process of cell differentiation was observed at 96 h. In the t-SNE and UMAP plots, we can see a pattern similar to that of the R-S plot. There are 2 subclusters of the 12 h samples. Additionally, the 72 and 96 h samples form one large cluster, which is consistent with the R-S plot's findings. Most notably, there is a large difference between the ES cell at 0 h and ES cells at different times in all visualizations, and there is no misclassification of the 0 h state with cells from 72 and 96 h states, indicating that the cells have indeed differentiated from the original pluripotent state.

https://doi.org/10.1021/acs.jcim.3c00674
J. Chem. Inf. Model. 2024, 64, 2829−2838

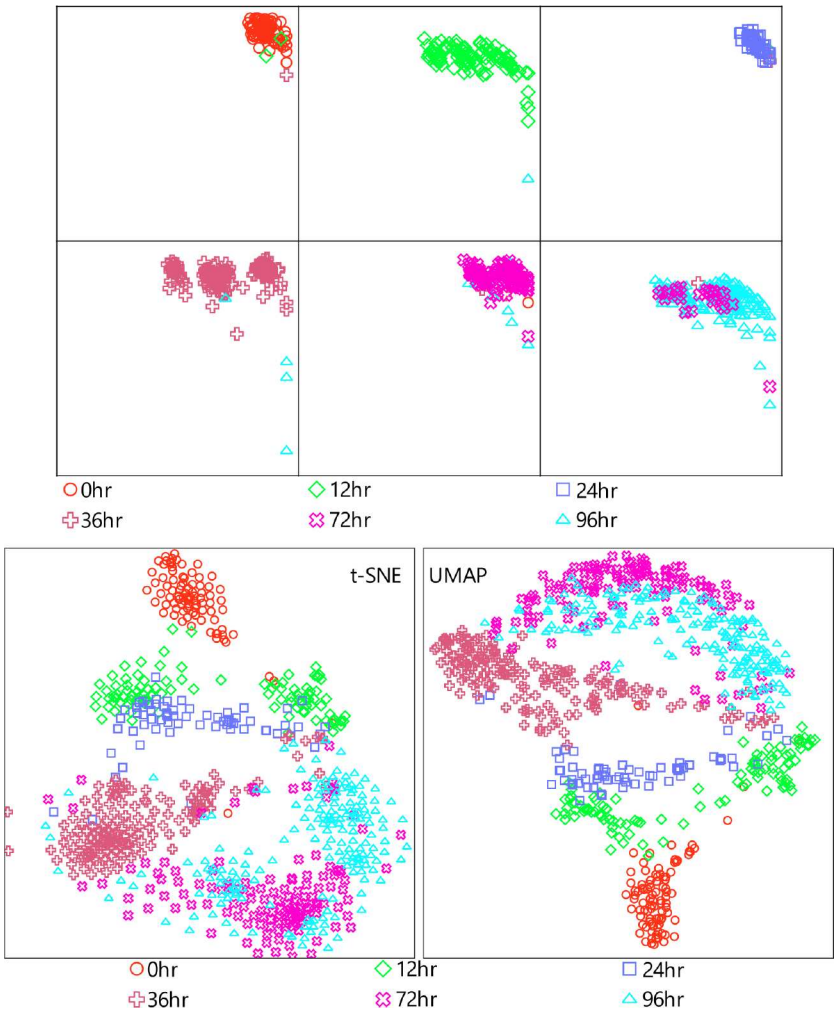


Figure 6. R-S plot, CCP assisted t-SNE plot, and standard t-SNE plots of GSE75748 time data. CCP was used to reduce the scRNA-seq data to 200 supergenes using $\tau = 6$ and $\kappa = 2$. The 5-fold cross-validation was used to split the data into 5 parts, where 4 parts were used for training and 1 part was used for testing the support vector machine classifier. RS scores were computed for the testing set, and all 5 folds were visualized. Each section corresponds to one of the 6 true cell types, and the sample's color and marker correspond to the predicted label from the support vector machine classifier. For t-SNE and UMAP, the data was log-transformed, and any genes with less than 10^{-6} variance were removed before applying the reduction. Samples were colored according to their cell types.

Figure 7 shows a comparison between the R-S plot and 2D plots of UMAP and t-SNE of the GSE75748 cell data. CCP was used to reduce the dimension to 100 supergenes with $\tau = 6$ and $\kappa = 2$. In ref 43, Chu obtained snapshots of lineage-specific progenitor cells that differentiated from H1 human embryonic stem (ES) cells. These differentiated cells include neuronal progenitor cells (NPCs), endoderm derivative cells (DECs), endothelial cells (ECs), trophoblast-like cells (TB), human foreskin fibroblasts (HFFs), and undifferentiated H1 and H9 human ES cells. Not surprisingly, all 3 visualizations show that undifferentiated ES cells H1 and H9 are clustered together, indicating that these two ES cells are relatively homogeneous, which agrees with Chu's findings. In the R-S plot, we see that all but 1 DEC sample are classified incorrectly, whereas in UMAP and t-SNE plots, DEC samples do not form a distinct cluster and have a super cluster forming with the H1, H9, and DEC cluster. In addition, all 3 visualizations show 2 clusters of NPC samples, but CCP is able to classify NPC samples correctly. Notice that in the R-S plot there are a few misclassifications of ECs and DECs, and in UMAP, these two clusters are adjacent to one another. This is consistent with a small number of misclassified EC and DEC groups shown in the RS plot. Since ECs are derivatives of mesoderm, it has been suggested by refs 37−39 that mesoderm and DECs may have developed and differentiated from a common progenitor pool.

5. CONCLUSION

CCP is a novel dimensionality reduction method that projects each cluster of similar genes into a supergene defined as accumulated pairwise nonlinear gene−gene correlations among cells. We have shown that CCP is able to differentiate cell types and also preserve the similarity along the trajectory of cellular differentiation. In addition, since CCP works exclusively in the data-domain, it does not rely on matrix diagonalization, and its results are easily interpretable. It significantly outperforms PCA for problems with intrinsically high dimensionality.

We also show that the RSI is a novel metric for evaluating the effectiveness of dimensionality reduction algorithms. Since it correlates with accuracy but does not rely on knowing the true labels of the data, it can be applied to improve both

https://doi.org/10.1021/acs.jcim.3c00674
J. Chem. Inf. Model. 2024, 64, 2829−2838

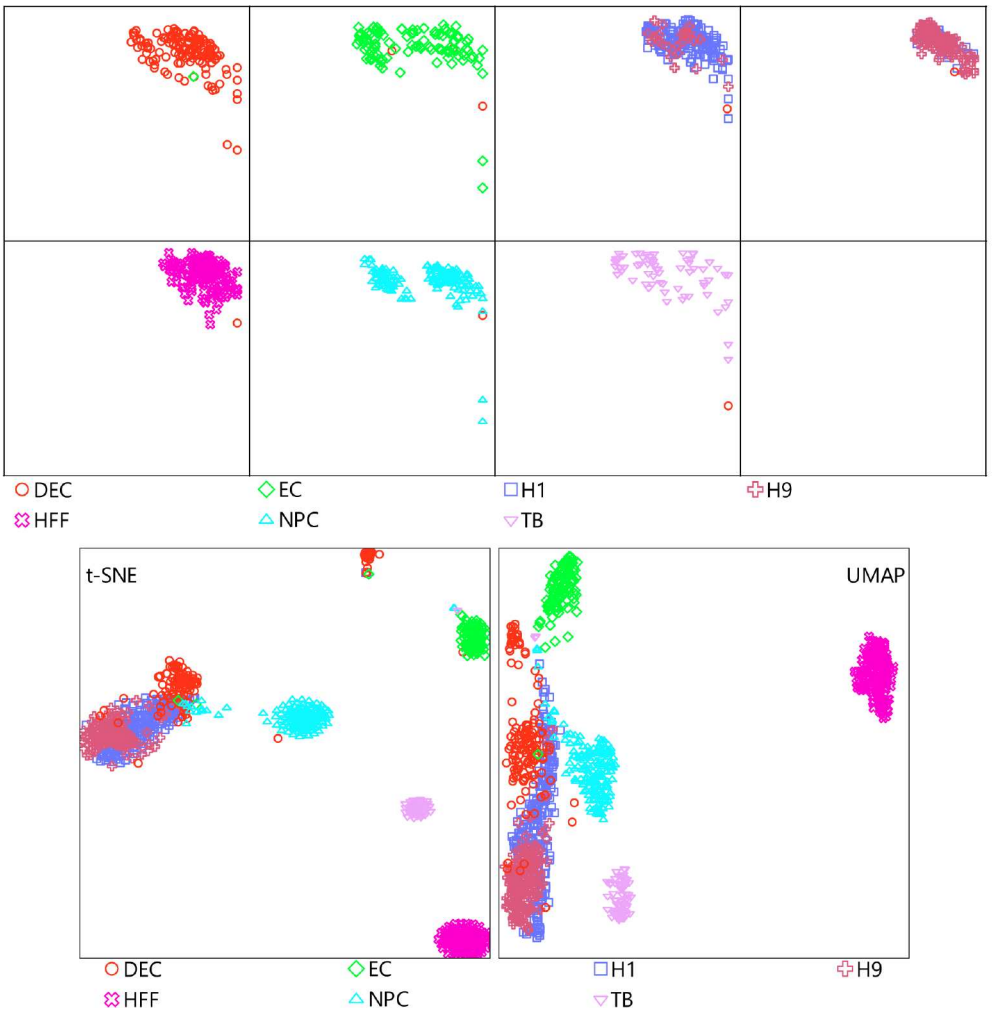


Figure 7. RS plot and CCP assisted UMAP and t-SNE plots of GSE75748 cell. CCP was used to reduce the scRNA-seq data to 100 components using $\tau = 6$ and $\kappa = 2$. 5-Fold cross-validation was used to split the data into 5 parts, where 4 parts were used for training and 1 part was used for testing the k-NN classifier. The RS score was computed for the testing set, and all 5 folds were visualized. Each section corresponds to 1 of the 7 true cell types, and the sample's color and marker correspond to the predicted label from the k-NN classifier. For t-SNE and UMAP, the data was log-transformed, and any genes with less than 10^{-6} variance were removed before applying the reduction. Samples were colored according to their cell types.

clustering and classification. In addition, the RSI can be used to vary the number of clusters and obtain insight into the optimal number of cell types. This information can be used to filter out data where CCP may not perform well, because CCP works best when the intrinsic dimensionality of the data, i.e., the number of gene features, is relatively high. Lastly, the R-S plot is introduced as a new visualization tool that works well for problems with a large number of cell types.

■ ASSOCIATED CONTENT

Data Availability Statement

All data was processed and is available at https://github.com/hozumiyu/SingleCellDataProcess. The code needed to reproduce this paper's result can be found at https://github.com/hozumiyu/CCP-for-Single-Cell-RNA-Sequencing. CCP is made available through our Web server at https://weilab.math.msu.edu/CCP/ or through the source code https://github.com/hozumiyu/CCP. The source code of the RSI and R-S plot can be found at https://github.com/hozumiyu/RSI.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.3c00674.

Details of the metrics, and benchmark on the other data described in Table 1 (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Guo-Wei Wei − *Department of Mathematics, Department of Electrical and Computer Engineering, and Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, Michigan 48824, United States;* orcid.org/0000-0002-5781-2937; Email: weig@msu.edu

Authors

Yuta Hozumi − *Department of Mathematics, Michigan State University, East Lansing, Michigan 48824, United States;* orcid.org/0000-0003-4674-2379

Kiyoto Aramis Tanemura − *Department of Mathematics, Michigan State University, East Lansing, Michigan 48824, United States*

https://doi.org/10.1021/acs.jcim.3c00674
J. Chem. Inf. Model. 2024, 64, 2829−2838

Complete contact information is available at:
https://pubs.acs.org/10.1021/acs.jcim.3c00674

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported in part by NIH grants R01GM126189 and R01AI164266, NSF grants DMS-2052983, DMS-1761320, and IIS-1900473, NASA grant 80NSSC21M0023, MSU Foundation, Bristol-Myers Squibb 65109, and Pfizer.

■ REFERENCES

(1) Hwang, B.; Lee, Ji Hyun; Bang, Duhee Single-Cell RNA Sequencing Technologies and Bioinformatics Pipelines. *Exp. Mol. Med.* **2018**, *50* (8), 1−14.
(2) Andrews, T. S; Yu Kiselev, Vladimir; Davis, McCarthy; Martin, Hemberg Tutorial: Guidelines for the Computational Analysis of Single-Cell RNA Sequencing Data. *Nat. Protoc.* **2021**, *16* (1), 1−9.
(3) Luecken, M. D; Theis, Fabian J Current Best Practices in Single-Cell RNA-Seq Analysis: A Tutorial. *Mol. Syst. Biol.* **2019**, *15* (6), No. e8746.
(4) Chen, G.; Ning, Baitang; Shi, Tieliu Single-Cell RNA-Seq Technologies and Related Computational Data Analysis. *Front. Genet.* **2019**, *10*, 317.
(5) Petegrosso, R.; Li, Zhuliu; Kuang, Rui Machine Learning and Statistical Methods for Clustering Single-Cell RNA-Sequencing Data. *Briefings Bioinf* **2020**, *21* (4), 1209−1223.
(6) Li, W. V.; Li, Jingyi Jessica A Statistical Simulator Scdesign for Rational Scrna-Seq Experimental Design. *Bioinformatics* **2019**, *35* (14), i41−i50.
(7) Lähnemann, D.; Köster, Johannes; Szczurek, Ewa; McCarthy, Davis J; Hicks, Stephanie C; Robinson, Mark D; Vallejos, Catalina A; Campbell, Kieran R; Beerenwinkel, Niko; Ahmed, Mahfouz; et al. Eleven Grand Challenges in Single-Cell Data Science. *Genome Biol.* **2020**, *21* (1), 31.
(8) Zheng, R.; Li, Min; Liang, Zhenlan; Wu, Fang-Xiang; Pan, Yi; Wang, Jianxin SINNLRR: A Robust Subspace Clustering Method for Cell Type Detection by Non-Negative and Low-Rank Representation. *Bioinformatics* **2019**, *35* (19), 3642−3650.
(9) Shu, Z.; Long, Qinghan; Zhang, Luping; Yu, Zhengtao; Wu, Xiao-Jun Robust Graph Regularized NMF With Dissimilarity and Similarity Constraints for SCRNA-Seq Data Clustering. *J. Chem. Inf. Model.* **2022**, *62* (23), 6271−6286.
(10) Wu, P.; Mo, An; Zou, Hai-Ren; Zhong, Cai-Ying; Wang, Wei; Wu, Chang-Peng A Robust Semi-Supervised NMF Model for Single Cell RNA-Seq Data. *PeerJ.* **2020**, *8*, No. e10091.
(11) Lan, W.; Chen, J. 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.
(12) Xiao, Q.; Luo, Jiawei; Liang, Cheng; Cai, Jie; Ding, Pingjian A Graph Regularized Non-Negative Matrix Factorization Method for Identifying MicroRNA-Disease Associations. *Bioinformatics* **2018**, *34* (2), 239−248.
(13) Yu, N.; Gao, Ying-Lian; Liu, Jin-Xing; Wang, Juan; Shang, Junliang Robust Hypergraph Regularized Non-Negative Matrix Factorization for Sample Clustering and Feature Selection in Multi-View Gene Expression Data. *Hum. Genomics* **2019**, *13* (S1), 46.
(14) Liu, J.-X.; Wang, Dong; Gao, Ying-Lian; Zheng, Chun-Hou; Shang, Jun-Liang; Liu, Feng; Xu, Yong A Joint-l2, 1-Norm-Constraint-Based Semi-Supervised Feature Extraction for RNA-Seq Data Analysis. *Neurocomputing* **2017**, *228*, 263−269.
(15) Wang, B.; Zhu, Junjie; Pierson, Emma; Ramazzotti, Daniele; Batzoglou, Serafim Visualization and Analysis of Single-Cell RNA-Seq Data by Kernel-Based Similarity Learning. *Nat. Methods* **2017**, *14* (4), 414−416.
(16) Ma, S.; Dai, Ying Principal Component Analysis Based Methods in Bioinformatics Studies. *Briefings Bioinf* **2011**, *12* (6), 714−722.
(17) Park, S.; Zhao, Hongyu Sparse Principal Component Analysis With Missing Observations. *Ann. Appl. Stat* **2019**, *13* (2), 1016−1042.
(18) Townes, F. W.; Hicks, Stephanie C.; Aryee, Martin J.; Irizarry, Rafael A. Feature Selection and Dimension Reduction for Single-Cell RNA-Seq Based on a Multinomial Model. *Genome Biol.* **2019**, *20* (1), 295.
(19) McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. 2018, arXiv:1802.03426. *arXiv*. https://arxiv.org/abs/1802.03426 (accessed 2023-06-30).
(20) Kobak, D.; Berens, Philipp The Art of Using T-SNE for Single-Cell Transcriptomics. *Nat. Commun.* **2019**, *10* (1), 5416.
(21) Lopez, R.; Regier, Jeffrey; Cole, Michael B; Jordan, Michael I; Yosef, Nir Deep Generative Modeling for Single-Cell Transcriptomics. *Nat. Methods* **2018**, *15* (12), 1053−1058.
(22) Torroja, C.; Sanchez-Cabo, Fatima Digitaldlsorter: Deep-Learning on Scrna-Seq to Deconvolute Gene. expression data. *Front. Genet.* **2019**, *10*, 978.
(23) Yuan, M.; Chen, Liang; Deng, Minghua ScMRA: A Robust Deep Learning Method to Annotate ScRNA-Seq Data With Multiple Reference Datasets. *Bioinformatics* **2022**, *38* (3), 738−745.
(24) Luo, Z.; Xu, Chenyu; Zhang, Zhen; Jin, Wenfei A Topology-Preserving Dimensionality Reduction Method for Single-Cell RNA-Seq Data Using Graph Autoencoder. *Sci. Rep.* **2021**, *11* (1), 20028.
(25) Wang, D.; Gu, Jin VASC: Dimension Reduction and Visualization of Single-Cell RNA-Seq Data by Deep Variational Autoencoder. *Genomics, Proteomics Bioinf.* **2018**, *16* (5), 320−331.
(26) Lin, E.; Mukherjee, Sudipto; Kannan, Sreeram A Deep Adversarial Variational Autoencoder Model for Dimensionality Reduction in Single-Cell RNA Sequencing Analysis. *BMC Bioinf* **2020**, *21* (1), 64.
(27) Zhou, H. J.; Li, Lei; Li, Yumei; Li, Wei; Li, Jingyi Jessica PCA Outperforms Popular Hidden Variable Inference Methods for Molecular QTL Mapping. *Genome Biol.* **2022**, *23* (1), 210.
(28) Jolliffe, I. T.; Cadima, Jorge Principal Component Analysis: A Review and Recent Developments. *Philos. Trans. R. Soc., A* **2016**, *374* (2065), 20150202.
(29) Kiselev, V. U.; Kirschner, Kristina; Schaub, Michael T; Andrews, Tallulah; Yiu, Andrew; Chandra, Tamir; Natarajan, Kedar N; Wolf, Reik; Barahona, Mauricio; Green, Anthony R; et al. SC3: Consensus Clustering of Single-Cell RNA-Seq Data. *Nat. Methods* **2017**, *14* (5), 483−486.
(30) Jin, S.; Guerrero-Juarez, Christian F; Zhang, Lihua; Chang, Ivan; Ramos, Raul; Kuan, Chen-Hsiang; Myung, Peggy; Plikus, Maksim V; Nie, Qing Inference and Analysis of Cell-Cell Communication Using CellChat. *Nat. Commun.* **2021**, *12* (1), 1088.
(31) Maseda, F.; Cang, Zixuan; Nie, Qing DEEPsc: A Deep Learning-Based Map Connecting Single-Cell Transcriptomics and Spatial Imaging Data. *Front. Genet.* **2021**, *12*, 636743.
(32) Hao, Y.; Stephanie, Hao; Erica, Andersen-Nissen; Mauck, William M, III; Zheng, Shiwei; Butler, Andrew; Lee, Maddie J; Wilk, Aaron J; Darby, Charlotte; Zager, Michael; et al. Integrated Analysis of Multimodal Single-Cell Data. *Cell* **2021**, *184* (13), 3573−3587.
(33) Pliner, H. A.; Shendure, Jay; Cole, Trapnell Supervised Classification Enables Rapid Annotation of Cell Atlases. *Nat. Methods* **2019**, *16* (10), 983−986.
(34) Zhang, Z.; Luo, Danni; Zhong, Xue; Choi, Jin Huk; Ma, Yuanqing; Wang, Stacy; Mahrt, Elena; Guo, Wei; Stawiski, Eric W; Modrusan, Zora; et al. SCINA: a semi-supervised subtyping algorithm of single cells and bulk samples. *Genes* **2019**, *10* (7), 531.
(35) Hozumi, Y.; Wang, R.; Wei, G.-W. CCP: Correlated Clustering and Projection for Dimensionality Reduction. 2022, arXiv:2206.04189. *arXiv*. https://arxiv.org/abs/2206.04189 (accessed 2023-06-30).

(36) Xia, K.; Opron, Kristopher; Wei, Guo-Wei Multiscale Multiphysics and Multidomain Models—Flexibility and Rigidity. *J. Chem. Phys.* **2013**, *139* (19), 194109.
(37) Yu, P.; Pan, Guangjin; Yu, Junying; Thomson, James A FGF2 Sustains NAONAG and Switches the Outcome of bmp4-Induced Human Embryonic Stem Cell Differentiation. *Cell Stem Cell* **2011**, *8* (3), 326−334.
(38) Rodaway, A.; Takeda, Hiroyuki; Koshida, Sumito; Broadbent, Joanne; Price, Brenda; Smith, James C; Patient, Roger; Holder, Nigel Induction of the Mesendoderm in the Zebrafish Germ Ring by Yolk Cell-Derived TGF-Beta Family Signals and Discrimination of Mesoderm and Endoderm by FGF. *Development* **1999**, *126* (14), 3067−3078.
(39) Tada, S.; Era, Takumi.; Furusawa, Chikara.; Sakurai, Hidetoshi.; Nishikawa, Satomi.; Kinoshita, Masaki.; Nakao, Kazuki.; Chiba, Tsutomu.; Nishikawa, Shin-Ichi. Characterization of Mesendoderm: A Diverging Point of the Definitive Endoderm and Mesoderm in Embryonic Stem Cell Differentiation Culture. *Development* **2005**, *132*, 4363.
(40) Deng, Q.; Ramsköld, Daniel; Reinius, Björn; Sandberg, Rickard Single-Cell RNA-Seq Reveals Dynamic, Random Monoallelic Gene Expression in Mammalian Cells. *Science* **2014**, *343* (6167), 193−196.
(41) Kowalczyk, M. S.; Tirosh, Itay; Heckl, Dirk; Nageswara Rao, Tata; Dixit, Atray; Haas, Brian J; Schneider, Rebekka K; Wagers, Amy J; Ebert, Benjamin L; Regev, Aviv Single-Cell RNA-Seq Reveals Changes in Cell Cycle and Differentiation Programs Upon Aging of Hematopoietic Stem Cells. *Genome Res.* **2015**, *25* (12), 1860−1872.
(42) Darmanis, S.; Sloan, Steven A; Zhang, Ye; Martin, Enge; Caneda, Christine; Shuer, Lawrence M; Gephart, Melanie G Hayden; Barres, Ben A; Quake, Stephen R A Survey of Human Brain Transcriptome Diversity at the Single Cell Level. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112* (23), 7285−7290.
(43) Chu, L.-F.; Ning, Leng; Zhang, Jue; Hou, Zhonggang; Mamott, Daniel; Vereide, David T; Choi, Jeea; Kendziorski, Christina; Stewart, Ron; Thomson, James A Single-Cell RNA-Seq Reveals Novel Regulators of Human Embryonic Stem Cell Differentiation to Definitive Endoderm. *Genome Biol.* **2016**, *17*, 173.
(44) Gokce, O.; Stanley, Geoffrey M; Treutlein, Barbara; Norma, F Neff; Camp, J. Gray; Malenka, Robert C; Rothwell, Patrick E; Fuccillo, Marc V; Südhof, Thomas C; Quake, Stephen R Cellular Taxonomy of the Mouse Striatum As Revealed by Single-Cell RNA-Seq. *Cell Rep* **2016**, *16* (4), 1126−1137.
(45) Baron, M.; Veres, Adrian; Wolock, Samuel L; Faust, Aubrey L; Gaujoux, Renaud; Vetere, Amedeo; Hyoje Ryu, Jennifer; Wagner, Bridget K; Shen-Orr, Shai S; Klein, Allon M; et al. A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter−And Intra-cell Population Structure. *Cell Syst* **2016**, *3* (4), 346−360.e4.
(46) Breton, G.; Zheng, Shiwei; Valieris, Renan; Tojal da Silva, Israel; Satija, Rahul; Nussenzweig, Michel C Human Dendritic Cells (DCs) Are Derived From Distinct Circulating Precursors That Are Precommitted to Become CD1C+ or CD141+ DCs. *J. Exp. Med.* **2016**, *213* (13), 2861−2870.
(47) Villani, A.-C.; Satija, Rahul; Reynolds, Gary; Sarkizova, Siranush; Shekhar, Karthik; Fletcher, James; Griesbeck, Morgane; Butler, Andrew; Zheng, Shiwei; Lazo, Suzan; et al. Single-Cell RNA-Seq Reveals New Types of Human Blood Dendritic Cells, Monocytes, and Progenitors. *Science* **2017**, *356* (6335), No. eaah4573.

https://doi.org/10.1021/acs.jcim.3c00674
J. Chem. Inf. Model. 2024, 64, 2829−2838