Single-Cell Analysis via Manifold Fitting: A New Framework for RNA Clustering and Beyond

Zhigang Yao[a,1], **Bingjie Li**[a,1], **Yukun Lu**[a,1], **and Shing-Tung Yau**[b,c,1]

This manuscript was compiled on January 1, 2024

Single-cell RNA sequencing (scRNA-seq) data, susceptible to noise arising from biological variability and technical errors, can distort gene expression analysis and impact cell similarity assessments, particularly in heterogeneous populations. Current methods, including deep learning approaches, often struggle to accurately characterize cell relationships due to this inherent noise. To address these challenges, we introduce scAMF (Single-cell Analysis via Manifold Fitting), a framework designed to enhance clustering accuracy and data visualization in scRNA-seq studies. At the heart of scAMF lies the manifold fitting module, which effectively denoises scRNA-seq data by unfolding their distribution in the ambient space. This unfolding aligns the gene expression vector of each cell more closely with its underlying structure, bringing it spatially closer to other cells of the same cell type. To comprehensively assess the impact of scAMF, we compile a collection of 25 publicly available scRNA-seq data sets spanning various sequencing platforms, species, and organ types, forming an extensive RNA data bank. In our comparative studies, benchmarking scAMF against existing scRNA-seq analysis algorithms in this data bank, we consistently observe that scAMF outperforms in terms of clustering efficiency and data visualization clarity. Further experimental analysis reveals that this enhanced performance stems from scAMF's ability to improve the spatial distribution of the data and capture class-consistent neighborhoods. These findings underscore the promising application potential of manifold fitting as a tool in scRNA-seq analysis, signaling a significant enhancement in the precision and reliability of data interpretation in this critical field of study.

Manifold Fitting | Single-Cell RNA Sequencing Analysis| Unsupervised Clustering | Visualization

Significance Statement

scRNA-seq analysis, crucial for uncovering cellular diversity and disease mechanisms, faces challenges due to technical variability, data complexity, and biological noise. Current scRNA analysis approaches often fail to deliver accurate and stable scRNA-seq clustering. In response, we present a pioneering framework centered on manifold fitting. Diverging from traditional methods that aim for a low-dimensional data representation, our approach fits a low-dimensional manifold within the ambient space and unfolds the data accordingly. This strategy effectively reduces distances between similar cell types while preserving expression information, significantly improving scRNA-seq clustering and visualization compared to existing state-of-the-art techniques. This groundbreaking advancement establishes a new standard in scRNA-seq analysis, opening exciting avenues for future research and potential clinical applications.

Single-cell RNA sequencing (scRNA-seq) (1–3) has become a crucial tool in genomic research, offering unparalleled resolution in dissecting the genomic, transcriptomic, and epigenomic profiles of individual cells. This detailed view is instrumental in deciphering the complex interplay within tissues and the inherent diversity of cellular populations. Advanced analytical methodologies (4–6), encompassing dimensionality reduction, cell clustering, and visualization techniques, enable a deeper understanding of cellular growth tracks and the variations in gene expression. scRNA-seq has provided novel insights into the pathogenesis of diseases such as diabetes (7), Alzheimer's disease (8), and cancer (9). Additionally, the advancement of scRNA-seq technology has set the basis for the growth of multi-omics analysis (10), spatial transcriptomics (11), and the Human Cell Atlas project (12). These advancements not only enhance the depth of cellular analysis but also place these findings in the broader context of tissue structure and function.

Despite the significant contributions of scRNA-seq in genomic research, it faces notable challenges, particularly in dealing with two types of noise: biological noise and measurement error. Biological noise is intrinsic to the cells and can stem from various sources. This includes the over-expression of certain genes, nutrient fluctuations, the specific location of a cell within its tissue or organ environment, and the current state or condition of the cell. In contrast, measurement errors are tied to the technical aspects of the sequencing process. These errors can vary based on the sequencing technology and platform used, as well as the specific methods applied during sequencing. The combination of biological noise and measurement errors adds a layer of complexity and variability to scRNA-seq data, making it challenging to extract accurate and meaningful biological insights.

Researchers have developed various strategies to mitigate the effects of noise and variability in scRNA-seq data. These strategies can be broadly classified into three categories: genomic imputation, graph-based methods, and deep learning networks. Genomic imputation primarily addresses dropout events in scRNA-seq data, where a gene is expressed but fails to be detected. A notable example is CIDR (13), which

Author affiliations: [a]National University of Singapore; [b]Tsinghua University; [c]Harvard University

Z.Y. and S.Y. designed research; Z.Y., B.L., and Y.L. performed research; Z.Y., B.L., Y.L., and S.Y. wrote the paper.

The authors declare no competing interest.

[1]All authors contributed equally to this work.

[2]To whom correspondence should be addressed. E-mail: styau@tsinghua.edu.cn.

calculates the probability of such dropout events and uses this information for imputation. Other genomic imputation methods, including MAGIC (14), SAVER (15), and scImpute (16), have been developed to address dropout events and other forms of noise, each employing unique strategies like Markov process, Bayesian prediction, and machine learning techniques. Graph-based methods, such as Seurat (17), utilize principal component analysis followed by graph clustering to categorize cells. Similarly, SC3 (4) employs a graph-based approach but focuses on consensus clustering. SCANPY (6) constructs a neighborhood graph of cells based on their similarity and then applies Louvain clustering (18) to this graph. Deep learning algorithms, like scDHA (19) and DESC (20), leverage computational models to represent and analyze scRNA-seq data. scDHA combines two autoencoders—a non-negative kernel autoencoder and a Bayesian autoencoder—for clustering. DESC uses a stacked autoencoder for data representation and integrates it with an iterative clustering neural network.

While the strategies mentioned above provide some solutions to the noise issue in scRNA-seq analysis, each has limitations that can impact their effectiveness. Genomic imputation methods rely on assumptions about data distribution and dropout events, introducing the risk of biases or inaccuracies. Graph-based clustering methods, which involve dimensionality reduction, may lead to information loss, potentially obscuring key cellular differences crucial for understanding biological processes. Although deep learning algorithms are often more accurate, their decision-making processes may lack clarity. Furthermore, the specificity of deep networks could limit their adaptability to varied data sets.

Manifold fitting (21–24) is an advanced technique with the potential for processing the scRNA-seq data. This method aims to reconstruct a smooth manifold within the original space where the data is measured, capturing the low-dimensional structure of the data in a manner that minimizes information loss and effectively eliminates noise. The latest innovations in manifold fitting (23) offer solutions to the limitations of existing methods, characterized by three key features. Firstly, it operates in the ambient space without the information loss typically associated with dimensionality reduction in graph clustering. Secondly, manifold fitting is adaptable to diverse data distributions, employing flexible neighborhood definitions, which is a distinct advantage over interpolation methods that rely heavily on specific data distribution assumptions. Finally, this technique is highly interpretable and supported by comprehensive theoretical analysis.

Building upon these advancements, we invent a novel framework named scAMF to address the persistent noise issue for scRNA-seq data. scAMF, inspired by recent advances in (23), is elegantly designed to accommodate the analysis of scRNA-seq data. Compared with existing scRNA-seq analysis frameworks, scAMF carries several notable advantages. Firstly, scAMF markedly enhances the spatial distribution of data, achieving a more pronounced agglomeration within classes while ensuring clearer separation between them. This refinement is pivotal for enabling more precise and accurate clustering in subsequent analyses. Secondly, scAMF comprehensively integrates various data transformation methods and clustering algorithms. This synergy allows for the efficient processing of clustering and visualization tasks across different scRNA-seq data platforms, demonstrating superior performance compared to existing algorithms. Most notably, scAMF introduces an innovative self-supervised approach to autonomously determine optimal clustering outcomes, marking a significant stride towards auto-machine learning in scRNA-seq analysis. Considering these advancements, scAMF emerges as a promising and potentially transformative tool in scRNA-seq analysis, setting a new benchmark for future research.

Before diving into the details, we showcase the performance of scAMF in clustering on the Kolodziejczyk data. This mouse embryo stems cell set (25) contains 704 cells in 3 classes (lif, 2i, a2i) with 38658 genes. This study focuses on the potential molecular mechanisms governing stem cell differentiation and maintenance. Fig. 1 shows the T-SNE (26) visualization between raw and scAMF-processed data. While the data points belonging to different classes in the raw data are entangled together, scAMF manages to separate classes and form a distinct aggregated pattern within each class. This evidently suggests a much higher clustering accuracy, highlighting the great potential of scAMF for scRNA-seq analysis.

In the following sections, we will explore scAMF further in several aspects on 25 scRNA-seq data sets (Results), with detailed procedure and methodology (Materials and Methods).

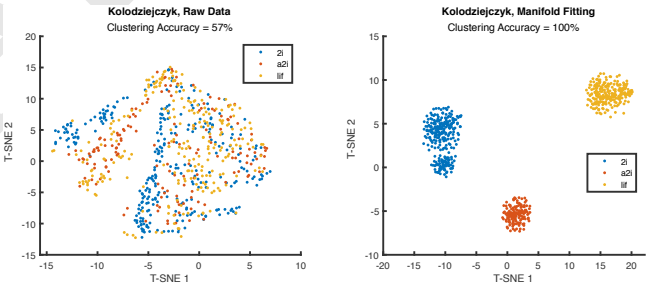

Fig. 1. An intuitive illustration of scAMF's outstanding performance. The Kolodziejczyk data contains 704 cells with 38658 genes. The clustering accuracy of k-means on raw data is 57%, whereas the clustering accuracy of k-means rises to 100% after scAMF. This evidently highlights the great potential of scAMF for scRNA-seq analysis.

Results

An overview of scAMF. scAMF consists of three modules for the analysis of scRNA-seq data: raw data transformation, manifold fitting, unsupervised clustering, and validation (Fig. 2**a**). Integrating these modules makes scAMF a unified framework for automatic machine learning of scRNA-seq data with high accuracy.

scAMF first transforms the data using three transformation methods (Fig. 2**b** and Methods A). The role of this module is to improve the signal-to-noise ratio in the data, including reducing excessive variance in highly expressed genes and partially correcting for batch effects, ensuring a more accurate and reliable data representation. Due to the diverse origins and sequencing platforms of scRNA-seq, different transformations are required to process the data

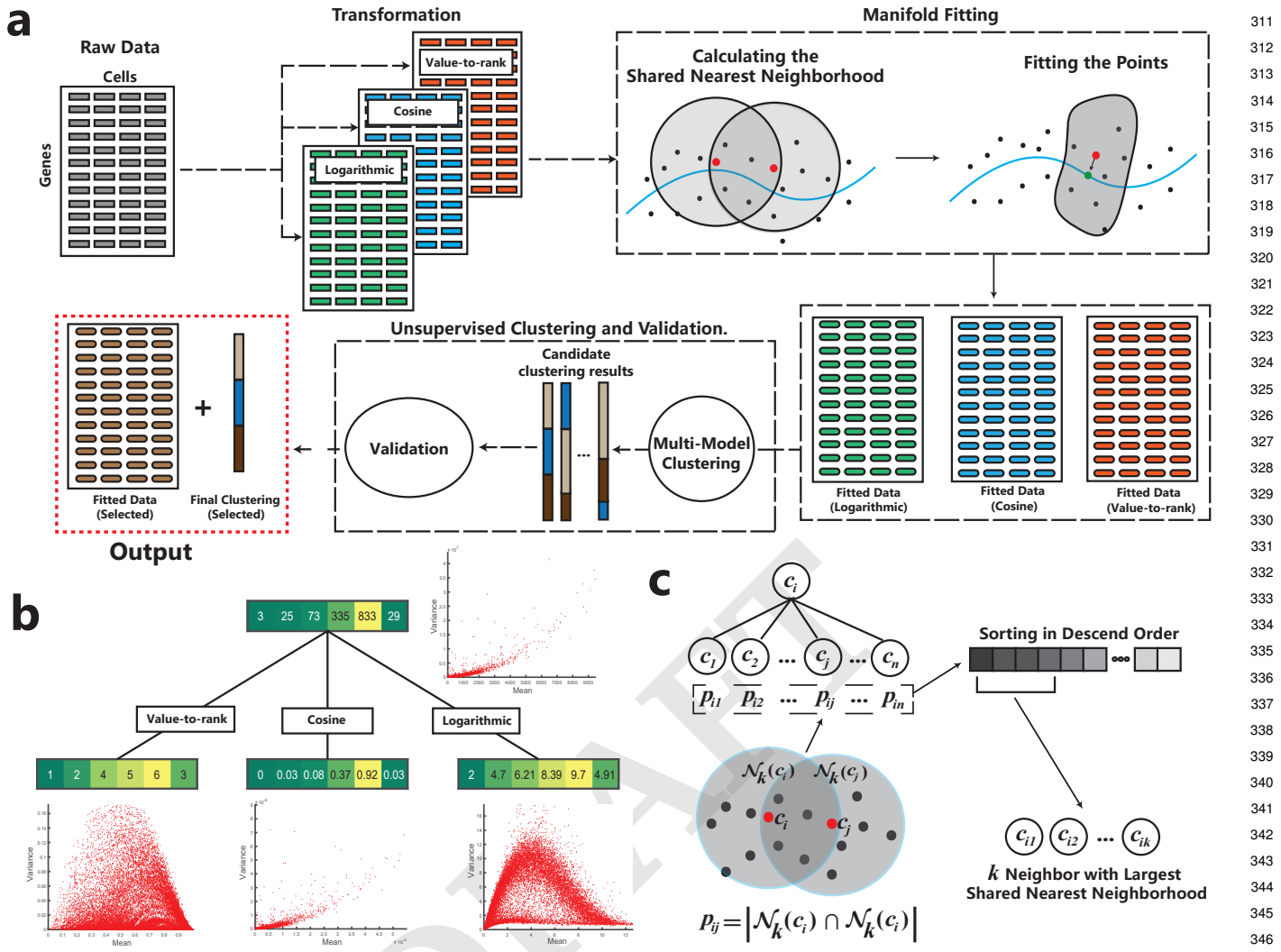

Fig. 2. The scAMF framework for the analysis of scRNA-seq data. (a) A schematic overview of the scAMF pipeline: data transformation, manifold fitting, unsupervised clustering and validation, and final outputs. (b) An illustration of the mechanism of value-to-rank, cosine, and logarithmic transformations. The colored bars demonstrate randomly generated example sequences and the corresponding transformed results. The scatter plots depict the mean-variance distribution of the raw and transformed Kolodziejczyk data (25). (c). Illustration of neighborhood selection based on the shared nearest neighborhood: Each cell is denoted as c_i, with its k-nearest neighborhood represented by $\mathcal{N}_k(c_i)$. The shared nearest neighborhood p_{ij} quantifies the overlap between $\mathcal{N}_k(c_i)$ and $\mathcal{N}_k(c_j)$, indicating common points. This method refines the neighborhood of c_i by selecting cells with the highest shared nearest neighborhood count. This process results in a more accurate neighborhood definition for c_i.

effectively. Note that scAMF employs an unsupervised approach to determine the most suitable transformation method in its pipeline. This strategy is innovative among existing scRNA-seq analysis methods.

For each transformed scRNA-seq data, scAMF performs the manifold fitting algorithm (Methods B, (23)), with suitable modifications tailoring to scRNA data. To highlight, scAMF uses neighborhoods determined by the shared nearest neighbor metric (27), which is more efficient in the measurement of high-dimensional data than the classical Euclidean or correlation-based metrics. A demonstration of the principle that uses the shared nearest neighbor algorithm to determine neighborhoods is shown (Fig. 2**c**). Moreover, scAMF simplifies the two steps outlined in the original approach by Yao et al (23), rendering it a more straightforward and expeditious method. Following the fitting process, the data is unfolded in the ambient space. This results in a reduction of intra-class spacing and an augmentation of inter-class spacing, creating conditions more conducive to effective clustering in subsequent analyses.

For the fitted data, scAMF employs four very basic fast clustering algorithms, which include but are not limited to spectral clustering and three agglomerative hierarchical clustering algorithms with different linkage settings to accommodate various structural complexities within scRNA-seq data sets (Methods C). The utilization of multi-model clustering in scAMF is connected to SC3 (4). By combining these clustering techniques, scAMF ensures a comprehensive analysis, enabling the identification of subtle and distinct clustering patterns that a single-method approach might overlook. However, a key distinction from SC3 is that we do not rely on the consensus of multiple clustering results. Instead, we select the optimal clustering result, as introduced subsequently.

The unsupervised clustering validation module (Methods C) aims to identify the most accurate clustering result among

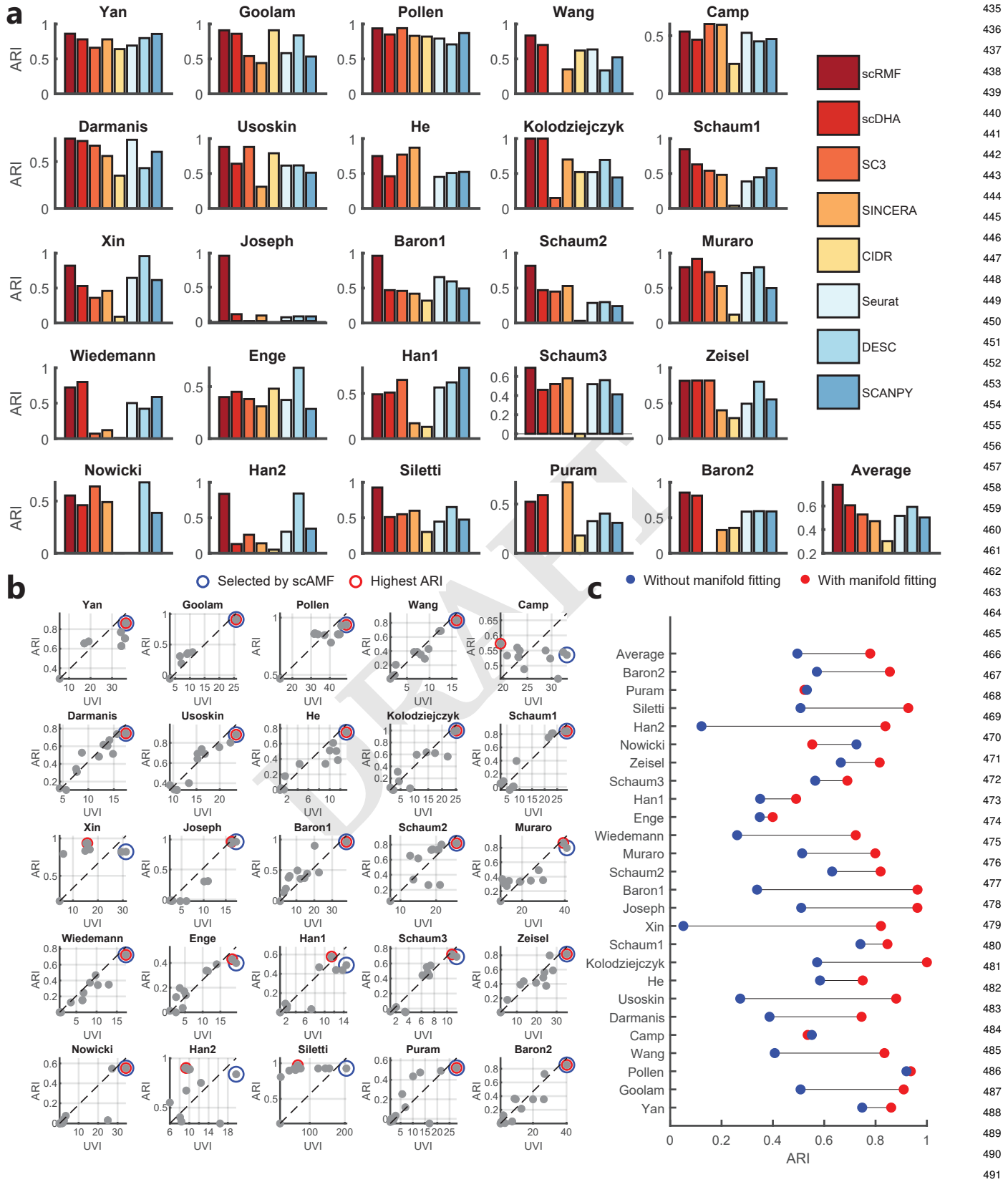

Fig. 3. Performance of scAMF on 25 scRNA-seq data sets. (a) Clustering performance comparison of scAMF and other methods, measured by ARI. The 25 plots display ARI values for each data set, while the lower right plot shows the average ARI, with scAMF surpassing other clustering methods. (b) Scatter plots of ARI against UVI for all combinations of transformation methods and clustering algorithms in each data set with manifold fitting. In most cases, combinations with the highest UVI values align with those exhibiting the highest ARI, underscoring the rationality of scAMF's selection procedure. (c) ARI of scAMF with and without manifold fitting. The inclusion of manifold fitting consistently enhances scAMF's clustering performance.

outcomes produced by various transformations and clustering methods. This module operates on the principle that cells of the same type exhibit greater similarities while heterogeneous samples do not. For each clustering result, scAMF calculates the similarity ratio of intra-class cells to inter-class cells. The clustering result with the highest similarity ratio is then selected as the most accurate one. Consequently, this process also determines the optimal transformation and manifold-fitting representation of the data.

By implementing scAMF, we can classify high-variance genes into two distinct groups: cluster-irrelevant genes and cluster-related genes. Cluster-irrelevant genes showcase variations not linked to specific cell types, essentially representing "housekeeping" genes essential for basic cellular functions but not indicative of particular cell types. In contrast, cluster-related genes exhibit significant variance across different classes, reflecting genes more indicative of specialized cellular functions and characteristics. scAMF identifies cluster-related genes based on clustering results, and these genes can be visualized. This facilitates a more focused and informative visualization, highlighting genes most relevant to the specific cell types and functions identified in the scRNA-seq data.

scAMF enhances the performance of scRNA-seq clustering. To evaluate the performance of scAMF in clustering, we compare it to seven cutting-edge scRNA analysis methods, including scDHA (19), SC3 (4), SINCERA (28), CIDR (13), Seurat (29), DESC (20), and SCANPY (6), on 25 scRNA-seq data sets with known cell types (Materials). Note that the true cell type is solely used for performance evaluation.

In evaluating clustering performance, scAMF demonstrates a notable superiority over other methods. We primarily employ the adjusted rand index (ARI) as our chief evaluation metric, complemented by the normalized mutual information (NMI) and accuracy (ACC) metrics (SI Appendix, B). scAMF consistently registers the highest ARI scores across a majority of the data sets, averaging 0.78. This performance significantly eclipses that of its closest competitor, scDHA, which achieves an average score of 0.61 (Fig. 3**a** and SI Appendix Tab. S2). The statistical significance of scAMF's superior ARI is further validated by a one-sided Wilcoxon signed-rank test (p-value 6.665×10^{-4}). Additionally, scAMF's exemplary performance in clustering is similarly evident in the NMI and ACC metrics (SI Appendix Tab. S3 and S4).

The strong performance of scAMF is thanks in part to its unsupervised clustering validation module. It selects the clustering result that maximizes the UVI (Methods C) from combined transformation methods and clustering algorithms. Notably, the chosen result aligns with the highest clustering accuracy. This relationship is evident by comparing UVI and ARI across all clustering results for each data set (Fig. 3**b**). In 16 out of 25 data sets, the highest UVI coincides with the highest ARI. Additionally, in 6 data sets, the clustering indicated by the peak UVI achieves the second-highest ARI. This highlights scAMF's ability to effectively identify the optimal clustering result, contributing to automated machine learning with high accuracy.

We conduct ablation studies to highlight the critical role of manifold fitting in scAMF. The impact of this step is illustrated by comparing ARI for scAMF with and without manifold fitting, as depicted in Fig. 3**c**. The results reveal a significant decrease in ARI; specifically, the ARI of scAMF without manifold fitting model drops by an average of 0.28 across nearly all data sets tested. This decline highlights the pivotal role of manifold fitting in enhancing scAMF's performance. Additionally, the effectiveness of the unsupervised clustering validation module is considerably reduced without manifold fitting (SI Appendix, Fig. S1). In scenarios where manifold fitting is excluded while retaining other steps, only 7 data sets show a consistent correlation between UVI and ARI. This evidence collectively underscores manifold fitting as a cornerstone module in the scAMF framework, integral to its effectiveness.

scAMF offers a spatial layout that better supports class separation. To intrinsically evaluate the improvement of scAMF brought to scRNA-seq data sets, we employ specific metrics to evaluate spatial distributions post-scAMF application, focusing on intra/inter-class distances and neighborhood purity.

Intra-class distance calculates the average distance from a sample to others within its class, indicating the compactness of each class. Conversely, inter-class distance measures the average distance to a sample from different classes, illustrating class separation. Scatter plots across six typical data sets (Fig. 4**a**) show a trend towards increased class distinction under scAMF, indicated by the upper-left movement of data points. Further, we introduce a ratio of intra-class to inter-class distance for each point. A lower ratio signifies enhanced class separation and tighter clustering within the same class. Fig. 4**a** visually confirms scAMF's role in improving spatial class separation.

The neighborhood purity of a sample is defined as the proportion of neighboring samples of the same type as the sample itself. It can be assessed before and after the scAMF framework. In our analysis, we choose a neighborhood size of ten and compute the neighborhood purity for all data sets. By comparing the results, we observed a consistent increase in neighborhood purity in the data sets processed through scAMF (Fig. 4**b**). This enhancement in neighborhood purity provides strong evidence of scAMF's effectiveness in improving class distinction within the scRNA-seq data.

scAMF provides better visualization. In this part, we show that scAMF's visualization performance surpasses those of widely-used methods like T-SNE (26), uniform manifold approximation and projection (UMAP)(30), and the classical principle component analysis (PCA)(31). We visualize two typical data, Kolodziejczyk and Usoskin, where different classes are indicated by various colors (Fig. 5**a**).

In analyzing the Kolodziejczyk data (top row of Fig. 5**a**), PCA demonstrates limited class differentiation, with greater variation within classes than between them. T-SNE, akin to PCA, manages some separation but indicates potential within-class divisions, likely arising from batch effects in the data set. UMAP, however, inaccurately splits the "2i" class into two distinct clusters. In contrast, scAMF effectively segregates the three classes, achieving clear distinction and creating densely packed clusters for each class, thus demonstrating its superior clustering capability.

The main challenge with the Usoskin data (bottom row of Fig. 5**a**) is the overlapping nature of different classes, making complete separation challenging. Techniques like PCA, T-

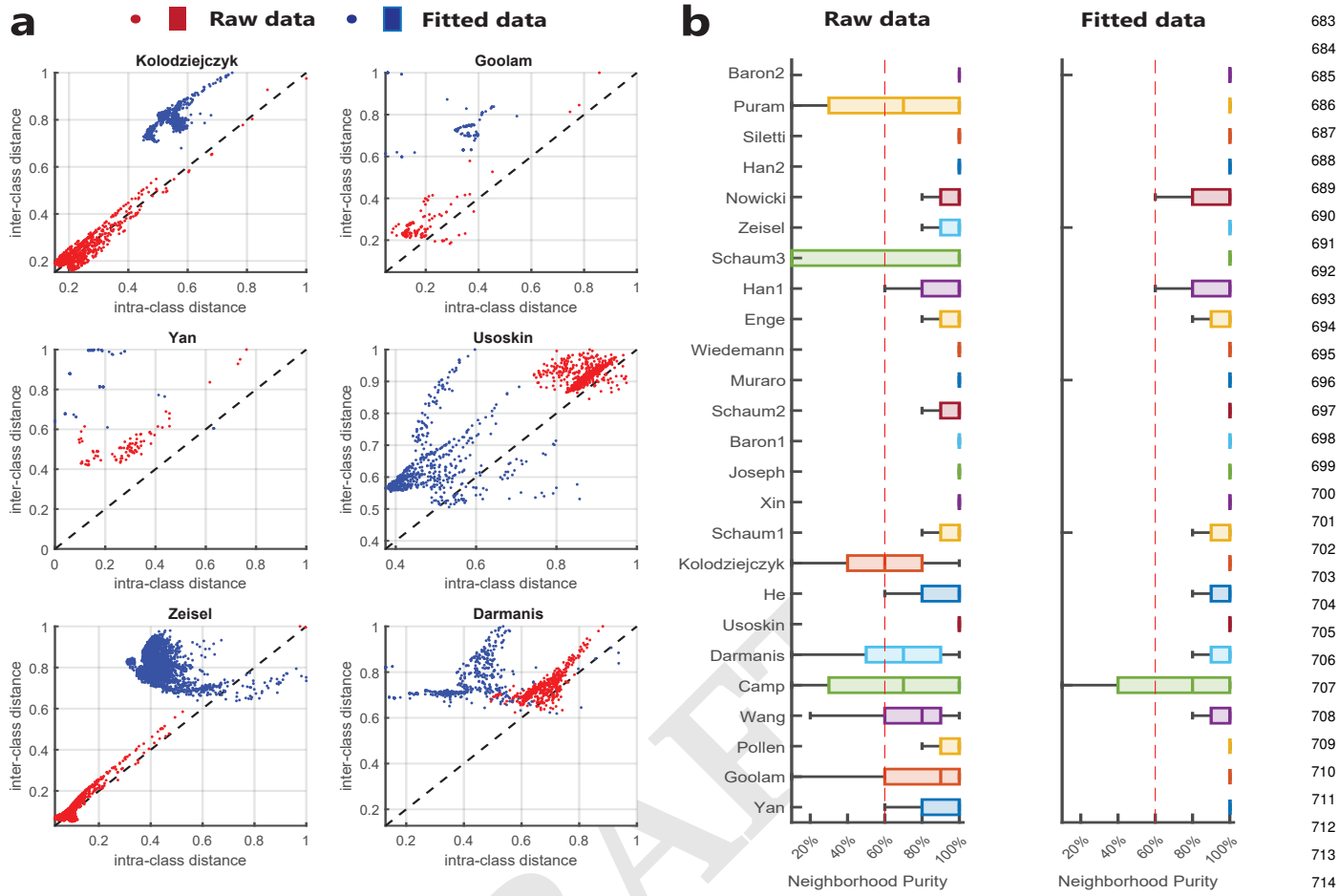

Fig. 4. Space distribution comparison between raw data and manifold-fitted data. (a) Representations of intra-class distance against inter-class distance plots in 6 data sets: Kolodziejczyk, Goolam, Yan, Usoskin, Zeisel, and Darmanis. Fitted data sets exhibit higher intra-inter-class distance ratios, indicating improved underlying structures for clustering post-scAMF. The box plots further confirm that scAMF achieves higher intra-inter-class distance ratios. (b) Box charts of raw and fitted data neighborhood purity for all 25 data sets. The majority of fitted data sets show higher neighborhood purity values than raw data sets, supporting the notion that scAMF contributes to a more accurate representation of scRNA-seq data structure.

SNE, and UMAP struggle to distinguish between these classes adequately. T-SNE and UMAP show some improvement over PCA, yet this enhancement is still limited. In contrast, scAMF effectively differentiates all classes while minimizing the misplacement of cells. Notably, scAMF highlights the bridging role of the NP group. scAMF's visualization reveals a connection between NP neurons and the PEP, TH, and NF groups, while these groups maintain distinct separation. This observation aligns with biological insights into NP neurons (32), known for their diverse neurochemical expressions, including TRPV1 and TRPA1 channels, reflecting their roles in various sensory functions such as heat sensation and chemical response.

To assess the effectiveness of scAMF's visualization capability, we have quantitatively compared it with other scRNA-seq analysis methods using the silhouette index. The silhouette index (33), ranging from -1 to 1, serves as a measure of visualization clarity and accuracy. Higher values indicate better visualization quality. This comparison is visually represented in a box plot (Fig. 5**b**), showing the silhouette index distribution across all data sets. The plot illustrates a notable shift of the scAMF box towards 1, indicating superior visualization performance compared to other methods. Detailed silhouette index scores for each method are available (SI Appendix, Tab. S5). Although scDHA may achieve a higher silhouette index in a few data sets, scAMF stands out for its consistent performance and lower variability, making it a more reliable choice for scRNA-seq data visualization.

Algorithm 1 scAMF Framework

Input: $\mathcal{X}$: raw expression matrix $(N \times D)$
Output: $\mathcal{Z}$: fitted data $(N \times D)$, $\mathcal{C}$: final clustering result $(N \times 1)$
1: Transform $\mathcal{X}$ to $\{\mathcal{Y}_i : i = 1, ..., n_T\}$ using transformation methods $\{T_i : i = 1, ..., n_T\}$
2: Obtain fitted data $\mathcal{Z}_i$ from each $\mathcal{Y}_i$ through manifold fitting
3: Cluster each $\mathcal{Z}_i$ using a set of basic clustering methods $\{C_j : j = 1, ..., n_C\}$
4: Obtain a set of candidate clustering results $\{\mathcal{C}_{ij}\}$
5: Calculate UVI for each $\mathcal{C}_{ij}$
6: Choose the $\mathcal{C}_{ij}$ with the largest UVI as the final clustering result $\mathcal{C}$
7: Choose the corresponding $\mathcal{Z}_i$ with respect to $\mathcal{C}$ as $\mathcal{Z}$, the final fitted data
return $\mathcal{Z}$, $\mathcal{C}$

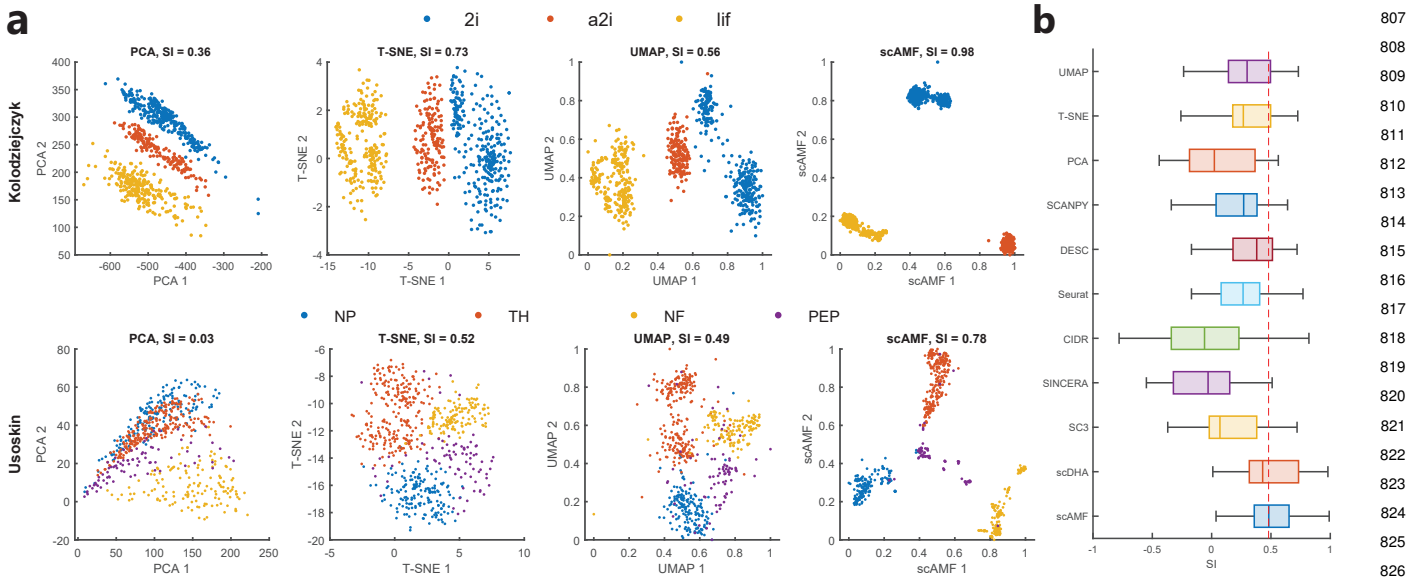

Fig. 5. 2D Visualizations and silhouette indexes of scAMF. (a) Color-coded representations of Kolodziejczyk and Usoskin data using T-SNE, UMAP, PCA, and scAMF. We also report the silhouette index for all plots, measuring the compactness of clusters within the same class and the separation between different classes. (b) A box plot of silhouette indexes for 8 clustering methods along with PCA, T-SNE and UMAP across all 25 data sets. scAMF significantly outperforms other methods in terms of silhouette index.

Materials and Methods

Materials. Our study collects 25 diverse scRNA data sets centered on cellular and molecular biology, encompassing embryonic, fetal, and adult tissues from humans and mice. These data sets offer valuable insights into various stages of development and disease, each with unique characteristics and sourced from different studies, as detailed by their respective GEO accession codes and research citations. They provide a rich resource for understanding biological processes and heterogeneity at a single-cell level, employing a range of protocols, including Smart-Seq2, STRT-Seq, and 10X Genomics. All data sets are freely available for download, along with detailed information for each data set (SI Appendix, F and Tab. S1).

Methods. Suppose we have a data set $\mathcal{X} = \{x^{(i)}\}_{i=1}^{N}$, where $\mathcal{X}$ represents a collection of scRNA-seq raw data. Each vector $x^{(i)}$ corresponds to the expression values $[x_1^{(i)}, \ldots, x_D^{(i)}]$ of the i-th cell across D genes. The transformed data is denoted as $\mathcal{Y} = \{y^{(i)}\}_{i=1}^{N}$, where each transformed cell expression $y^{(i)}$ consists of $[y_1^{(i)}, \ldots, y_D^{(i)}]$, still encompassing D genes. Finally, the fitted data is represented by $\mathcal{Z} = \{z^{(i)}\}_{i=1}^{N}$.

A. Raw Data Transformation. Raw data are transformed in parallel, allowing us to select the most suitable transformation. As we will show, it would achieve the best possible condition for downstream analysis. While multiple transformation methods can be chosen, we apply three of them in our experiments.

- **Value-to-rank Transformation (T_1).** Instead of looking at the actual levels of gene expression, the value-to-rank transformation focuses on where each gene stands in the order within a cell. This method transforms the expression levels to their respective ranks within each cell expression vector. Specifically, the transformation is defined component-wisely with each $y_k^{(i)}$ being the rank of $x_k^{(i)}$ in the ascending order of $x^{(i)}$.
- **Cosine Transformation (T_2).** To minimize the impact of varying expression magnitudes across cells, this transformation normalizes the expression vectors of each cell to have a unit Euclidean norm. For an expression vector $x^{(i)}$, the normalized vector is given by $y^{(i)} = x^{(i)}/\|x^{(i)}\|_2$ where $\|x^{(i)}\|_2$ represents the Euclidean norm of $x^{(i)}$.
- **Logarithmic transformation (T_3).** Logarithmic transformation is also applied to rectify the skewness commonly present in scRNA-seq data. Each expression vector $x_k^{(i)}$ in vector $x^{(i)}$ is transformed as follows
$$y_k^{(i)} = \log_2(1 + x_k^{(i)}).$$
This approach stabilizes variance across varying expression levels, making the data more suitable for subsequent analysis.

B. Manifold Fitting. The data sets transformed are labelled as $\mathcal{Y}_1, \cdots, \mathcal{Y}_{n_T}$, respectively. For a transformed data $\mathcal{Y} \in \{\mathcal{Y}_i : i = 1, ..., n_T\}$, with D as the dimension of the ambient space (i.e., the number of genes), the manifold hypothesis assumes that the $\mathcal{Y}$ can be viewed as a noisy representation of the underlying low-dimensional manifold $\mathcal{M}$ of dimension d ($d \ll D$). This implies that the true (noiseless) samples lie on a simpler, more basic structure but are obscured or altered due to noise. Hence, the fitting aims to construct an estimated manifold $\widehat{\mathcal{M}}$ from the noisy data $\mathcal{Y}$, or equivalently, approximately project $\mathcal{Y}$ onto $\widehat{\mathcal{M}}$. Specifically, the main steps of the manifold fitting algorithm (23) include direction estimation and projection estimation.

- **B1. Direction estimation.** The purpose of finding the projection direction is to determine the direction where the point should move closer to the underlying manifold. For any point y, its projection direction can be expressed as $d_y = F(y) - y$, where $F(y)$ is obtained through the local contraction of y. In (23), $F(y)$ is defined by the following formula:
$$F(y) = \sum_i \alpha_i(y) y^{(i)},$$
with $\alpha_i(y)$ defined as:
$$\tilde{\alpha}_i(y) = \begin{cases} (1 - \frac{\|y - y^{(i)}\|_2^2}{r_0^2})^k, & \|y - y^{(i)}\|_2 \le r_0; \\ 0, & \text{otherwise}; \end{cases} \quad [1]$$
$$\tilde{\alpha}(y) = \sum_{i \in I_y} \tilde{\alpha}_i(y), \quad \alpha_i(y) = \frac{\tilde{\alpha}_i(y)}{\tilde{\alpha}(y)},$$
and $k > 2$ (3 by default) ensures smoothness, and I_y contains indices of members in $\mathcal{B}_D(y, r_0)$. In our implementation, we refine $F(y)$ by addressing two key aspects. 1) recognizing the limitations of Euclidean distance for scRNA-seq data, we have

adopted shared nearest neighbor metrics—a method better suited for high-dimensional data—to accurately determine the neighboring points involved in computation $F(y)$. 2) we have set the parameter to $k = 0$ in [1] to simplify the calculating process. This simplification is practical, as our objective is solely to project the points without the need for the smoothness of the output manifold. The rest is the detailed calculation of $F(y)$.

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

$$\mathrm{SNN}\,(i,j) = \left|\mathcal{N}_p(i) \cap \mathcal{N}_p(j)\right|.$$

- **Determine the neighborhood.** Using the SNN of $y^{(i)}$ and $y^{(j)}$, the refined p-nearest neighborhood of $y^{(i)}$ is defined by

$$\mathbb{B}(i) = \arg \max_{S\subset\mathcal{Y},\ |S|=p} \sum_{y^{(j)}\in S} \mathrm{SNN}(i,j).$$

- **Calculate the projection direction.** The projection direction $F(y^{(i)})$ of $y^{(i)}$ is defined by

$$F(y^{(i)}) = \frac{1}{|\mathbb{B}(i)|} \sum_{y^{(j)}\in\mathbb{B}(i)} y^{(j)}.$$

- **B2. Projection estimation.** Upon determining $F(y^{(i)})$, we construct a cylinder centered at $y^{(i)}$ with $F(y^{(i)}) - y^{(i)}$ as its major axis, following (23). The weighted average of the samples, denoted as $G(y^{(i)})$, within this cylinder, serves as an estimate for the projection. To accommodate the ultra-high nature and irregular noise patterns in scRNA-seq data, we simplify $G(y^{(i)})$ by identifying the point of maximum density along the line connecting $y^{(i)}$ and $F(y^{(i)})$ and using this point as a substitute for the original $G(y^{(i)})$. Precisely, the simplification is

$$G(y^{(i)}) = \arg\max_t \rho(y^{(i)} + t(F(y^{(i)}) - y^{(i)})).$$

Here,

$$\rho(y^{(i)}) = \frac{1}{\sum_{y^{(j)}\in\mathbb{B}(i)} \|y^{(i)} - y^{(j)}\|_2^2}$$

quantifies the density of $\rho(y^{(i)})$, with higher values indicating denser regions. This modification simplifies the analysis while accommodating the complex nature of scRNA-seq. The fitted data is then represented by

$$\mathcal{Z} = \{z^{(i)}\}_{i=1}^N, \qquad z^{(i)} = G(y^{(i)}).$$

C. Unsupervised Clustering and Validation. Upon outputting $\mathcal{Z}$, one may simply utilize fairly basic clustering methods. These could include but are not limited to

- *Spectral Clustering (C_1)*: group the data by analyzing the connectivity (adjacent matrix) among data points. Subsequently, it uses the k-means algorithm to perform clustering within this reduced-dimensional space.
- *Agglomerative Hierarchical Clustering (C_2-C_4)*: creates a hierarchy of clusters, with each being characterized by its approach to linkage. Single Linkage calculates the distance between clusters based on the closest pair of points, one from each cluster. Average Linkage takes the average distance between all pairs of points across two clusters. Centroid Linkage defines the distance between clusters as the distance between their respective centroids.

We remark that the clustering module in scAMF necessitates a predetermined number of cell types, denoted as K. Researchers can provide this number based on biological insights. To ensure a fair comparison, we employ the true number of cell types for clustering in all compared methods that require K as input (SI Appendix, D).

By implementing various transformations T_i and clustering methods C_j, where $1 \le i \le n_T$ and $1 \le j \le n_C$, multiple distinct clustering outcomes $\mathcal{C}_{ij}$ can be obtained. Each outcome is derived from the fitted data $\mathcal{Z}_i$ and categorizes cells into K groups. To report the optimal clustering result, we further introduce an unsupervised validation index (UVI) based on the similarity ratio to measure the justifiability of each $\mathcal{C}_{ij}$. We denote the label of the h-th cell in $\mathcal{C}_{ij}$ by $\mathcal{C}_{ij}^{(h)}$. The similarity ratio estimates the clustering quality based on two primary factors: average intra-class and inter-class similarity. Specifically, the similarity between $z^{(h)}$ and $z^{(\ell)}$, denoted as $s(h,\ell)$, is defined as:

$$s(h,\ell) = \begin{cases} 1 & \text{if } z^{(h)} \in \mathcal{N}_q(\ell) \text{ or } z^{(\ell)} \in \mathcal{N}_q(h); \\ 0 & \text{otherwise,} \end{cases}$$

where q is adaptively determined as $\min\{N/10, 40\}$. The average intra-class similarity, represented as $a(z^{(h)}, \mathcal{C}_{ij})$, evaluates the cohesion within a cluster as

$$a(z^{(h)}, \mathcal{C}_{ij}) = \frac{1}{|\{(h,\ell) : \mathcal{C}_{ij}^{(h)} = \mathcal{C}_{ij}^{(\ell)}\}|} \sum_{\mathcal{C}_{ij}^{(h)} = \mathcal{C}_{ij}^{(\ell)}} s(h,\ell).$$

Furthermore, the inter-class similarity, measuring the separation between different clusters, is denoted by

$$b(z^{(h)}, \mathcal{C}_{ij}) = \frac{1}{|\{(h,\ell) : \mathcal{C}_{ij}^{(h)} \neq \mathcal{C}_{ij}^{(\ell)}\}|} \sum_{\mathcal{C}_{ij}^{(h)} \neq \mathcal{C}_{ij}^{(\ell)}} s(h,\ell).$$

Hence, the UVI through the similarity ratio is defined by

$$\mathrm{UVI}(\mathcal{Z}, \mathcal{C}_{ij}) = \frac{\sum_h a(z^{(h)}, \mathcal{C}_{ij})}{\sum_h b(z^{(h)}, \mathcal{C}_{ij})}.$$

A higher UVI indicates superior clustering performance, marked by greater similarity within clusters than between them. We select the clustering result with the highest UVI from all $\mathcal{C}_{ij}$, and denote it by $\mathcal{C}$, as the final output (refer to **Algorithm** 1).

Discussion

In this study, we present scAMF, a novel framework enhancing scRNA-seq analysis through manifold fitting. scAMF insightfully represents the hidden low-dimensional structure, addressing technical variability and biological noise inherent in scRNA-seq data. Rigorous experimentation demonstrates the superior performance of scAMF in recovering distorted RNA expression data and improving clustering compared to existing techniques.

In our experiment, the true number of cell types is used as input to implement and evaluate scAMF and other clustering methods. Actually, determining the number of cell types is an open problem in RNA-seq analysis. There are two strategies for determining the number of cell types. The first strategy utilizes statistical techniques like the silhouette coefficient (33) or Dunn index (34), which may sometimes overestimate the true number. Alternatively, as suggested in (35), biologists should predetermine the number of cell types based on their research needs, opting for more clusters for detailed exploration or fewer for simplified analysis. While biologist-provided counts might slightly deviate from the true number of cell types, scAMF exhibits considerable flexibility, effectively managing both overestimation and underestimation (SI Appendix, Tab S7). For instance, even when the number of cell types is overestimated by three, scAMF still achieves a notable average ARI of 0.66 for all

data sets, outperforming other methods using the true number of cell types.

Future research directions for the manifold fitting approach used in scAMF could focus on expanding its application to other omics data like proteomics and metabolomics to tackle similar high-dimensionality and noise challenges. Additionally, enhancing the computational efficiency of scAMF, particularly for processing increasingly large data sets, could be achieved through more efficient algorithms or parallel computing. Moreover, integrating scAMF with emerging spatial transcriptomics data would offer a more comprehensive understanding of tissue architecture and cell-to-cell interactions, enabling spatial mapping of cell types and states.

Data and Code Archival. The Matlab-based implementation of scAMF, encompassing a demo of the scAMF pipeline, all data used in this article, and evaluation functions, is accessible via https://github.com/zhigang-yao/scAMF.

ACKNOWLEDGMENTS. Z.Y. has been supported by MOE Tier 2 grant (A-0008520-00-00, A-8001562-00-00) and Tier 1 grant (A-0004809-00-00, A-8000987-00-00) at the National University of Singapore; B.L. is supported by MOE Tier 2 grant A-0008520-00-00; Y.L. is supported by Research Scholarship at the National University of Singapore. Z.Y. thanks and appreciates the constructive discussion with Jun Liu at Harvard University.

1. F Tang, et al., mRNA-seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).
2. A Shalek, et al., Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498**, 236–240 (2013).
3. R Grindberg, et al., RNA-sequencing from single nuclei. *Proc. Natl. Acad. Sci.* **110**, 19802–19807 (2013).
4. V Kiselev, et al., SC3: consensus clustering of single-cell RNA-seq data. *Nat. Methods* **14**, 483–486 (2017).
5. Y Hao, et al., Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587 (2021).
6. F Wolf, P Angerer, F Theis, SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 1–5 (2018).
7. S Sachs, et al., Targeted pharmacological therapy restores β-cell function for diabetes remission. *Nat. Metab.* **2**, 192–209 (2020).
8. H Mathys, et al., Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* **570**, 332–337 (2019).
9. Y Katzenelenbogen, et al., Coupled scRNA-seq and intracellular protein activity reveal an immunosuppressive role of TREM2 in cancer. *Cell* **182**, 872–885 (2020).
10. Z Miao, B Humphreys, A McMahon, J Kim, Multi-omics integration in the age of million single-cell data. *Nat. Rev. Nephrol.* **17**, 710–724 (2021).
11. L Moses, L Pachter, Museum of spatial transcriptomics. *Nat. Methods* **19**, 534–546 (2022).
12. A Regev, et al., The human cell atlas. *Elife* **6**, e27041 (2017).
13. P Lin, M Troup, J Ho, CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data. *Genome Biol.* **18**, 1–11 (2017).
14. D Dijk, et al., MAGIC: A diffusion-based imputation method reveals gene-gene interactions in single-cell RNA-sequencing data. *bioRxiv:10.1101/111591* (2017).
15. M Huang, et al., SAVER: gene expression recovery for single-cell RNA sequencing. *Nat. Methods* **15**, 539–542 (2018).
16. W Li, J Li, An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nat. Commun.* **9**, 997 (2018).
17. T Stuart, et al., Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902 (2019).
18. V Blondel, J Guillaume, R Lambiotte, E Lefebvre, Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* p. 10008 (2008).
19. D Tran, et al., Fast and precise single-cell data analysis using a hierarchical autoencoder. *Nat. Commun.* **12**, 1029 (2021).
20. X Li, et al., Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis. *Nat. Commun.* **11**, 2338 (2020).
21. C Fefferman, S Ivanov, Y Kurylev, M Lassas, H Narayanan, Fitting a putative manifold to noisy data in *Conference On Learning Theory*. (PMLR), Vol. 75, pp. 688–720 (2018).
22. C Fefferman, S Ivanov, M Lassas, H Narayanan, Fitting a manifold of large reach to noisy data. *J. Topol. Analysis* (2023).
23. Z Yao, J Su, B Li, Manifold fitting: An invitation to statistics. *arXiv:2304.07680* (2023).
24. Z Yao, J Su, ST Yau, Manifold fitting with CycleGAN. *Proc. Natl. Acad. Sci. (In press)* (2023).
25. A Kolodziejczyk, et al., Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. *Cell Stem Cell* **17**, 471–485 (2015).
26. L Van der Maaten, G Hinton, Visualizing data using T-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
27. RA Jarvis, EA Patrick, Clustering using a similarity measure based on shared near neighbors. *IEEE Transactions on Comput.* **100**, 1025–1034 (1973).
28. M Guo, H Wang, S Potter, J Whitsett, Y Xu, SINCERA: a pipeline for single-cell RNA-seq profiling analysis. *PLoS Comput. Biol.* **11**, e1004575 (2015).
29. Y Hao, et al., Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat. Biotechnol.* (2023).
30. L McInnes, J Healy, J Melville, UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv:1802.03426* (2018).
31. H Hotelling, Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **24**, 417 (1933).
32. D Usoskin, et al., Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. *Nat. Neurosci.* **18**, 145–153 (2015).
33. P Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).
34. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *J. Cybern.* **3**, 32–57 (1973).
35. L Chen, W Wang, Y Zhai, M Deng, Deep soft k-means clustering with self-training for single-cell RNA sequence data. *NAR Genomics Bioinforma.* **2**, lqaa039 (2020).