OXFORD

Briefings in Bioinformatics, 2025, **26(1)**, bbae612
https://doi.org/10.1093/bib/bbae612
Case Study

ADM: adaptive graph diffusion for meta-dimension reduction

Junning Feng[1,2], Yong Liang[3,*], Tianwei Yu[1,*]

[1]School of Data Science, the Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen), 518172 Guangdong, China
[2]Faculty of Innovation Engineering, Macau University of Science and Technology, 999078 Macao Special Administrative Region of China
[3]Chinese Medicine Guangdong Laboratory, Hengqin 519031 Guangdong, China

*Corresponding authors. Yong Liang, Chinese Medicine Guangdong Laboratory, Hengqin 519031 Guangdong, China. E-mail: yongliangresearch@gmail.com; Tianwei Yu, School of Data Science, The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen), Shenzhen, Guangdong 518172, P.R. China; Shenzhen Research Institute of Big Data, Shenzhen, Guangdong 518172, P.R. China. E-mail: yutianwei@cuhk.edu.cn

Abstract

Dimension reduction is essential for analyzing high-dimensional data, with various techniques developed to address diverse data characteristics. However, individual methods often struggle to capture all intricate patterns and complex structures simultaneously. To overcome this limitation, we introduce *ADM (Adaptive graph Diffusion for Meta-dimension reduction)*, a novel meta-dimension reduction method grounded in graph diffusion theory. *ADM* integrates results from multiple dimension reduction techniques, leveraging their individual strengths while mitigating their specific weaknesses.*ADM* utilizes dynamic Markov processes to transform Euclidean space results into an information space, revealing intrinsic nonlinear manifold structures that are hard to capture by conventional methods. A critical advancement in *ADM* is its adaptive diffusion mechanism, which dynamically selects optimal diffusion time scales for each sample, enabling effective representation of multi-scale structures. This approach generates robust, high-quality low-dimensional representations that capture both local and global data structures while reducing noise and technique-specific distortions. We demonstrate *ADM*'s efficacy on simulated and real-world datasets, including various omics data types. Results show that *ADM* provides clearer separation between biological groups and reveals more meaningful patterns compared to existing methods, advancing the analysis and visualization of complex biological data.

Keywords: dimension reduction; adaptive graph; information diffusion; meta-dimension reduction

Introduction

Dimension reduction plays a crucial role in visualizing and comprehending complex datasets [1, 2]. These techniques aim to condense high-dimensional data into a lower-dimensional representation while preserving the data's inherent structure. This process is crucial for uncovering inherent similarities and differences between data points, facilitating an intuitive understanding of the underlying patterns and trends [3, 4].

Dimension reduction techniques have become indispensable for exploratory analysis and pattern discovery across various domains, including computer vision [5], molecular biology [6], and diverse omics fields such as genomics, proteomics, and metabolomics [7–11].

Over the past few decades, researchers have developed numerous dimension reduction techniques to reveal the structure of noisy, high-dimensional data. Early methods such as Principal Component Analysis (PCA) [12], Multidimensional Scaling (MDS) [13], Sammon mapping [14], Isomap [15], and kernel PCA (kPCA) [16] efficiently delineate with linear or nonlinear characteristics by preserving dominant structures, often at the expense of local details. Conversely, methods like Locally Linear Embedding [17], t-SNE [18], LargeVis [19], and Laplacian Eigenmaps [20] prioritize the revelation of local structures but may sacrifice the global coherence of data. Subsequent developments, such as UMAP [2], Hessian Locally Linear Embedding (HLLE) [21], and Kernel Eigenfunction Embeddings [22], have aimed at striking a balance between preserving both global and local structures. However, their effectiveness can be compromised by sensitivity to noise and outliers. Diffusion maps [23] utilize a diffusion operator to explore the data's intrinsic geometry, reducing noise impact but facing challenges in preserving multiscale structures [24]. PHATE [25] advances this concept by integrating manifold learning and information geometry, providing a balanced representation of both global and local structures. Recent advancements have led to more specialized techniques tailored for specific tasks and data types. Deep learning-driven frameworks such as scCRT [26], DREAM [27], and SPDR [28] have been designed for targeted tasks including trajectory inference and batch effect correction in single-cell genomics. Additionally, spectral embedding-based techniques like SnapATAC2 have been developed for specific data formats, particularly in the analysis of single-cell ATAC-seq data [29].

Despite these advancements, each technique, developed based on distinct principles and paradigms, is tailored to specific data characteristics and often struggles to provide a comprehensive data representation. Moreover, each method involves hyperparameters requiring fine-tuning for specific datasets, with different parameter configurations potentially leading to vastly different

Received: August 5, 2024. **Revised:** October 18, 2024. **Accepted:** November 12, 2024

outcomes. Furthermore, due to the inherent noisiness and high dimensionality of real-world datasets, low-dimensional representations inevitably contain distortions from the underlying true structures, which can vary across different techniques.

Consequently, there is growing interest in developing meta-analysis techniques capable of integrating results from various individual methods, leveraging their strengths to achieve a more robust and comprehensive representation while suppressing distortions in individual methods. Some approaches, like fuzzy consensus analysis [30] and supervised learning methods [31–35], primarily focus on combining classification results from multiple datasets. However, meta-dimension reduction technologies require establishing a common space to align different results of the same data accurately. This necessitates further exploration and analysis of the intrinsic relationships between results from various methods. Recently, Ma et al. introduced *Meta-Spec* [36], which quantifies the relative performance of individual techniques in preserving the structure around each data point in the Euclidean space and generates consensus dimension reduction results. While *Meta-Spec* improves upon individual techniques in capturing underlying structure, it encounters challenges when applied to highly complex, high-dimensional data. Its primary limitation lies in offering a static measure that reflects direct or physical distances between samples in Euclidean space. This approach may struggle to capture the dynamic geometric interactions and complex mechanisms of information transmission among samples that are often present in high-dimensional biological datasets. As a result, *Meta-Spec* might not completely represent the intricate relationships and latent structures inherent in complex data.

To address these challenges, we introduce *ADM* (Adaptive graph Diffusion for Meta-dimension reduction), a novel meta-dimension reduction and visualization technique grounded in information diffusion theory. *ADM* employs a dynamic Markov process to simulate information propagation between data points, transforming traditional Euclidean space results into an information space. This approach reveals intrinsic nonlinear manifold structures that conventional distance-preserving techniques may not fully capture.

A key innovation of *ADM* is its adaptive diffusion mechanism, which dynamically selects the optimal diffusion time scale for each sample. This feature enables *ADM* to capture both local and global structures within the data, representing intricate multi-scale structures more effectively. By integrating outputs from multiple dimension reduction methods, *ADM* mitigates noise and reduces technique-specific distortions, leading to more reliable outcomes. These capabilities allow *ADM* to generate more robust and higher-quality low-dimensional representations of high-dimensional datasets, advancing the analysis and visualization of complex biological data across various omics data types, including genomics, proteomics, metabolomics, and epigenomics.

To validate *ADM*'s effectiveness, we conducted comprehensive experiments on simulated and real-world datasets across various omics types. Quantitatively, *ADM* consistently outperformed existing dimension reduction techniques across multiple metrics, including Adjusted Rand Index (*ARI*), Category Consistency Index (CCI), and Average Silhouette Width (*ASW*). Qualitatively, *ADM* demonstrated superior performance in visualizations, enabling clearer separation between biological groups and revealing subtle differences between closely related types. These results highlight *ADM*'s potential for uncovering intricate patterns in complex biological data across various omics modalities.

Methods

Overview of *ADM*

For each output from an individual candidate dimension reduction technique, the *ADM* approach initiates by transforming the Euclidean distance of features into a diffusion operator $\mathcal{P}$. This operator quantifies the likelihood of information propagation between samples through random walks. Subsequently, a sample-specific diffusion process is implemented for each sample to simulate multi-step diffusion. We employ the Breadth-First Search algorithm to adaptively select the appropriate diffusion time scale for each sample. This adaptive strategy considers the inherent heterogeneity within datasets, effectively filters out noise, and prevents over-smoothing. The resulting sample-specific propagation probabilities are then utilized to calculate the diffusion distances. These diffusion distances leverage the dynamic Markov process to link the Euclidean distance with geometric densities, highlighting the intrinsic similarities and differences among samples in the information manifold, serving as a robust metric for quantifying the relative positions within the information space [37].

Next, we combine the diffusion distance matrices from all candidate methods using harmonic averaging and gamma distribution-based normalization to construct a comprehensive meta-diffusion-distance matrix. This distance leverages the advantages of individual candidate techniques, providing a robust representation of the dataset and enabling in-depth exploration of complex relationships among samples. For visualization and further interpretation, the meta-distance matrix can be reduced to a low-dimensional embedding via UMAP, producing plots that reveal the dataset's underlying structure. The overall framework of *ADM* is shown in Fig. 1.

Data preprocessing

In our method, the input data comprises a set of feature matrices generated by various dimension reduction techniques ($N \times \mathcal{X}, \mathcal{X} \in \mathbb{R}^{m \times n}$), known as candidates, where N is the number of candidates, m denotes the total number of samples, and n indicates the feature dimension. Initially, we preprocess each matrix $\mathcal{X}$ to remove outliers potentially skewing the dataset's true structure, thereby preserving the data integrity and consistency. Specifically, outliers are first discovered using the method proposed by Knorr and Ng [38].

For each feature $\mathbf{x}_{.k}$ (i.e., the k^{th} column) in $\mathcal{X}$, we calculate the interquartile range $IQR = Q_3 - Q_1$, where Q_3 is the 75^{th} percentile and Q_1 is the 25^{th} percentile of $\mathbf{x}_{.k}$. We set the neighborhood radius threshold to $d = IQR/3$ for the method by Knorr and Ng [38]. We remove all data points detected as outliers from $\mathbf{x}_{.k}$ to generate a new vector $\tilde{\mathbf{x}}_{.k}$. Our objective is to adjust the original $\mathbf{x}_{.k}$ to resemble the distribution of $\tilde{\mathbf{x}}_{.k}$. To achieve this goal, we employ quantile normalization of $\mathbf{x}_{.k}$ using $\tilde{\mathbf{x}}_{.k}$ as the reference distribution. The effect of this quantile normalization is the outliers in the feature dimension shrink towards the data cloud, while non-outlier data points are changed very little.

Affinity matrix

After preprocessing, we first calculate the Euclidean distance matrix D, in which d_{ij} is the Euclidean distance between sample i and j. To eliminate self-comparisons, we set d_{ii} as infinity. Subsequently, we convert the distance matrix $\mathbf{D}$ into a affinity measure $\mathbf{S}$:

$$\mathbf{S} = 1/\mathbf{D}^{\gamma}, \tag{1}$$

Figure 1. The framework of *ADM*. It takes inputs from various dimension reduction techniques and generates a meta-dimension reduction and visualization output that integrates the strengths of these individual results.

where γ is the parameter for the power transformation, and $\mathbf{S}$ signifies the sample affinities in the Euclidean space. This non-linear scaling mitigates the overwhelming impact of larger distances, thus emphasizing local neighborhood structures.

Adaptive connectivity

In information diffusion theory, connectivity between nodes is crucial for mapping out the routes and patterns of information propagation. To accommodate the complexity of the relationships between samples and capture multi-scale structure, we introduce a two-step procedure to construct a sample connectivity matrix $\mathbf{A}$ based on the distance matrix. This method considers both the overall trends and patterns across the entire dataset, as well as the local context surrounding individual samples. Globally, we set a global percentile threshold α to identify crucial sample interactions across the dataset. For example, setting $\alpha = 5\%$ designates the shortest 5% of distances in $\mathbf{D}$ as globally significant, with $thres = percentile(\mathbf{D}, \alpha)$. If $d_{ij} \leq thres$, the connection between samples i and j is considered substantial on a global scale, indicated by $\mathbf{A}_{ij} = 1$; otherwise, $\mathbf{A}_{ij} = 0$.

Locally, the percentile threshold β is used to adjust the connectivity of each sample according to its immediate neighbors, with $thres_i = percentile(\boldsymbol{d}_{i\cdot}, \beta)$ determining local connections, where $\boldsymbol{d}_{i\cdot}$ represents distances from sample i to all others. If $d_{ik} \leq thres_i$ (where $0 < k \leq n, k \neq i$), we set $\mathbf{A}_{ik} = 1$. Using the two-step procedure to construct the connectivity matrix helps to strike a balance between global exploration and local structure sensitivity.

To establish sample-specific affinity relations, we further generate a sparse adaptive affinity matrix ($\mathcal{U}$) by applying an element-wise multiplication between the affinity matrix $\mathbf{S}$ and the connectivity matrix $\mathbf{A}$:

$$\mathcal{U} = \mathbf{A} \odot \mathbf{S}. \tag{2}$$

This operation effectively strengthens the connections between samples of high similarities while diminishing the influence of those between dissimilar samples. In addition, this affinity measure only accounts for the local distance between samples. Ultimately, we transform these weighted connectivities to a weighted undirected graph $\mathcal{G} = \{V, E\}$, where V represents individual samples, and E denotes the connections defined by matrix $\mathbf{A}$, with edge weights given by $\mathcal{U}$.

Sample-specific diffusion

We employ the Markov random walk process to explore the nonlinear manifold structure of the dataset. The transition probabilities in this random walk are calculated through row-normalization of matrix $\mathcal{U}$, which renders the adaptive affinity into a Markov-based diffusion operator $\mathcal{P}$:

$$p_{ij} = \frac{\mathcal{U}_{ij}}{\sum_k (\mathcal{U}_{ik})}, \quad for \quad 1 \leq j, k \leq n, \tag{3}$$

where $p_{ij} \in \mathcal{P}$ denotes the probability of information propagating from sample point i to point j. We refer to $\mathcal{P}$ as the diffusion operator, which quantifies the likelihood and intensity of information propagation across the weighted graph.

The time scale τ, depicting the depth of information propagation in a random walk, plays a crucial role in balancing the trade-off between capturing local details and global information. A smaller value of τ emphasizes local structures, often associated with noisy information, while a larger τ captures global structures but may smooth out important details. Hammond et al. [39] pointed out that nodes within a densely connected graph may exchange information more efficiently than their sparsely connected counterparts. Therefore, our approach adaptively selects the optimal τ for each sample by evaluating its spatial position and connectivity to ensure a balanced representation of local and global characteristics.

To achieve this goal, we initially compute the shortest path lengths across all sample pairs within graph $\mathcal{G}$. We record the results in a matrix $\mathbf{L} \in \mathbb{R}^{m \times m}$, where m is the total number of samples, and each element $l_{ij} \in \mathbf{L}$ denotes the shortest path length from node i to node j.

We then calculate the average path length for each sample as a metric of its connectivity within the overall graph:

$$\bar{l}_i = \frac{1}{m-1} \sum_{j=1, j \neq i}^{m} l_{ij}, \tag{4}$$

where $\bar{l}_i$ denotes the average shortest path length for sample i relative to others.

The diffusion steps of sample i are then set to a value that is proportional to its average steps,

$$\tau_i = \bar{l}_i \times \zeta, \tag{5}$$

where ζ is a pre-determined scaling constant. By employing this approach, we ensure the diffusion process is finely tuned to reflect the unique connectivity and structural properties of each sample. Samples on the periphery of the point cloud are allowed more diffusion steps, while samples in the center use fewer diffusion steps. This adaptive process enhances the model's sensitivity to local and global connectivity patterns, facilitating a more precise depiction

study, we use three variants of this indicator, CCI_{raw}, CCI_{umap}, and CCI_{pca}, which measure the preservation of category similarity within the raw distance matrix, as well as the feature matrices after dimension reduction by UMAP and PCA, respectively.

Structural consistency index (*SCI*): the *SCI* metric evaluates the structural fidelity of data after dimension reduction relative to the original, noise-free structure:

$$SCI(\mathcal{D}_o, \mathcal{D}_r) = \frac{1}{m}\sum_{i=1}^{m} \frac{\langle \boldsymbol{d}_o, \boldsymbol{d}_r \rangle}{|\boldsymbol{d}_o| \cdot |\boldsymbol{d}_r|}, \tag{11}$$

where $\mathcal{D}_o$ and

Figure 2. Comparative Performance of the *ADM* and *Meta-Spec* Methods on Simulated Data. (A–F) CCI values of *ADM* and *Meta-Spec* on simulation data of different dimension-reduction outputs. (A–C) The CCI_{raw}, CCI_{umap}, and CCI_{pca} under different noise conditions, given fixed ranking parameters *r*. (D–F) Giving different groups of noise, the CCI_{raw}, CCI_{umap}, and CCI_{pca} under varying ranking parameters. (G–I) The 15 comparative results of the Structural Consistency Index (SCI) across varying signal strengths for the Smiley Dataset, Mammoth Dataset, and Gaussian Dataset, respectively. HLLE is omitted due to very low consistency.

Meta-Spec method. Fig. 2(B) and 2(C) demonstrates the results after applying dimension reduction to the distance matrices using UMAP and PCA, respectively. The results indicate that while both *ADM* and *Meta-Spec* experience a decrease in *CCI* scores with an increase in low-quality group sets, *ADM* consistently achieves higher *CCI* scores than *Meta-Spec* across various levels of noise.

Figure 2(D–F) displays the *CCI* values obtained by the two meta-dimension reduction methods at different r values, ranging from 1 to 20. Simultaneously, we randomly selected ten numbers from the range of 1 to 100 as the number of low-quality sets. It is observed that the *CCI* values for both *ADM* and *Meta-Spec* decrease as r increases, which is expected as the value of *r* determines how many neighbors are considered. However, it is evident that *ADM* outperforms *Meta-Spec* in terms of CCI_{raw}, CCI_{umap}, and CCI_{pca} across all tested r settings.

These experimental findings directly demonstrate *ADM*'s superior capability in mitigating the impact of noise from low-quality candidate methods while preserving signal from high-quality methods. More fundamentally, since *CCI* measures the consistency between sample distances and category labels, the results essentially indicate that the integration of diffusion distance matrices in *ADM* better captures the distribution of the original data compared to the integration of Euclidean distance matrices in *Meta-Spec*.

Simulation to test structure restoration

Dimension reduction plays a crucial roal in retaining essential and representative data features while mitigating complexity. In some specific fields, such as single-cell data analysis, high-noise environments pose significant challenges to dimension reduction techniques in ext. distance the betweenend ext.i7(eTju/F

Figure 3. The Quantitive analysis results of real data. (A) The ARI and NMI results on real dataset. (B) The ASW score across real datasets. The silhouette width score of each dataset on different clusters can be found in Fig. S3 in the supplementary materials. (C) The CCI results across various datasets.

In the experiment, we introduce varying noise levels by adjusting the signal-to-noise ratio φ to simulate realistic high-dimensional, high-noise data. Specifically, we randomly select 20 values of φ from [0.0001, 1] to generate a series of noise $\mathbf{Z}_i, i = 1, \ldots, 20$ for each dataset. For the **Smiley** dataset $\{\mathbf{Y}^1_{\mathbf{Z}_i}\}_{i=1,\ldots,20}$, we set the total number of observation points n as 550 and the feature dimension p as 300. Similarly, for the **Mammoth** $\{\mathbf{Y}^2_{\mathbf{Z}_i}\}_{i=1,\ldots,20}$ and **Gaussian Mixture** dataset $\{\mathbf{Y}^3_{\mathbf{Z}_i}\}_{i=1,\ldots,20}$, we set p to be 300, and 500, respectively, and n at 1000, and 900, respectively.

Then, we apply 12 individual dimension reduction techniques (including PCA, UMAP, PHATE, etc.) and a meta-dimension reduction technique *Meta-Spec* as comparative methods. We test some of these individual techniques with 2 different parameter settings, and finally obtained 15 individual candidates. *ADM* and *Meta-spec* use these 15 candidates as inputs. We evaluate the performance of *ADM*, against these 15 comparative methods using the SCI metric.

Figure 2(G–I) displays the performance comparison of 16 different methods applied to three distinct datasets $\{\mathbf{Y}^1_{\mathbf{Z}_i}\}_{i=1,\ldots,20}$, $\{\mathbf{Y}^2_{\mathbf{Z}_i}\}_{i=1,\ldots,20}$, and $\{\mathbf{Y}^3_{\mathbf{Z}_i}\}_{i=1,\ldots,20}$, respectively. Each boxplot shows the distribution of the achieved *SCI* values under 20 different levels of noise. The results show that meta-visualization techniques, including *ADM* and *Meta-Spec*, outperform individual dimension reduction methods in maintaining the core structure of the data. Across the three data sets, the *SCI* values of *ADM* are higher than those of *Meta-Spec*. Notably, *ADM* is characterized by narrower boxplots, which is more pronounced in the Mammoth and Gaussian Mixture datasets, indicating its superior consistency across datasets with varying levels of noise. This highlights *ADM*'s strong capacity for noise reduction, signal detection, and accurate reconstruction of the data's original structure through its dynamic information diffusion mechanism.

Results on real data

To comprehensively evaluate the performance of our proposed *ADM* method in dimension reduction and visualization, we conducted an extensive analysis across eight diverse publicly available datasets. These datasets span a wide range of single-cell and spatial omics modalities, including single-cell RNA sequencing data from various tissues and species, multi-omics data combining scRNA-seq and miRNA-seq, single-cell ATAC-seq data, metabolomics and spatial proteomics data. This diverse selection enables us to assess the versatility and effectiveness of *ADM* in handling various types of high-dimensional biological data.

Our analytical approach involves a systematic comparison of *ADM* with Meta-spec and 12 individual dimension reduction methods. For each dataset, we initially applied 12 individual techniques with varying parameter settings, generating 16 candidate outputs that we save for further analysis. Subsequently, we utilize

Figure 4. Visualization of the Oihane dataset. The abbreviations in the figure correspond to their full names: Astro: astrocytes, Endo: endothelial, Epend: ependymal, Hyb: hybrid, Micro: microglia, Neur: neurons, Oligo: oligodendrocytes. Due to space limitations, we present the outcomes of 16 randomly selected individual techniques. The complete comparative results can be found in Fig. S7 in the supplementary materials.

meta-dimension reduction methods, including *ADM* and the comparison method *Meta-Spec*, to integrate these results. The overall accuracy in terms of preserving between-cell type differences based on true cell labels is summarized by *ARI* and *NMI*, which are calculated after UMAP dimension-reduction of the respective distance matrices (Fig. 3A). Additionally, we employed the *ASW* score to evaluate clustering quality, measuring both intra-cluster cohesion and inter-cluster separation, shown as Fig. 3(B). To avoid artifacts caused by clustering techniques, we also analyzed the preservation of true cell class information in the distance matrices before and after further dimension reduction (Fig. 3C). We discuss the detailed results for each of these real datasets in the following sub-sections.

Oihane dataset: mouse midbrain and striatum scRNA-seq data.

The first dataset is the single-cell RNA sequencing data of the mouse midbrain and striatum, named Oihane [40]. This dataset comprises expression profiles from 1337 single cells and 13 446 genes, encompassing various cell types including oligodendrocytes, microglia, neurons, astrocytes, endothelial cells, ependymal cells, and hybrid cells. After filtering for genes expressed in at least 20% of cells, we retained 217 genes for downstream analysis.

Figure 4 illustrates the visualization outcomes of various individual dimension reduction techniques and two meta-reduction techniques applied to the Oihane dataset. Among the individual techniques, UMAP and t-SNE excel in separating different categories, with t-SNE exhibiting poor intra-class compactness. PHATE, Isomap, and LEIM capture the global manifold structure of the data but are less effective in differentiating between cell types. In contrast, meta-dimension reduction techniques can leverage the strengths of individual reduction techniques to maintain proximity relationships among data points while revealing the overall manifold structure of the dataset.

In comparison to *Meta-Spec*, *ADM* demonstrates superior performance by further expanding the distance between cell types while simultaneously maintaining intra-class tightness. In the visualization, the distances produced by *Meta-Spec* tend to identify oligodendrocytes as two distinct categories, one of them overlapping with microglia, whereas the diffusion distance generated by *ADM* closely arranges oligodendrocyte cells and creates clear separation from other categories. For Endothelial and neurons, *ADM* exhibits a clear boundary, whereas in the *Meta-Spec* visualization, these two cell types overlap with blurred boundaries. Additionally, in *Meta-Spec*, some astrocytes intertwine with neurons in the visualization, while the *ADM* results form two distinct clusters with clear boundaries between them.

Additionally, Fig. 3(A) demonstrates that *ADM* is superior in clustering accuracy. It offers significantly higher clustering precision with an *ARI* of 0.7314 and *NMI* of 0.7331 compared to

Figure 5. Visualization of Gutierrez dataset. The complete comparative results can be found in Fig. S8 in the supplementary materials.

the *ARI* (0.4070) and *NMI* (0.5517) of *Meta-Spec*. *ASW* score of *ADM* is nearly 0.65, substantially higher than *Meta-Spec*'s score of approximately 0.35, indicating better-defined and more cohesive clusters. Furthermore, Fig. 3C illustrates that *ADM* consistently outperforms *Meta-Spec* in preserving category similarities. This suggests that the adaptive diffusion distance produced by the proposed *ADM* can serve as a robust representation of cells for subsequent analysis.

Gutierrez dataset: human lymphocyte population.

The dataset is single-cell RNA sequencing (scRNA-seq) data from human lymphocyte populations [41]. The original data contains 2,036 cells and 33,694 genes. After filtering for genes expressed in at least 30% of cells, we retained 631 genes for downstream analysis. This dataset includes various cell types such as iNKT, MAIT, $\gamma\delta$T cells, NK cells, and CD4+ and CD8+ T cells, which demonstrated a high degree of functional and phenotypic similarity.

As shown in Fig. 5, current individual dimension-reduction techniques such as UMAP, t-SNE, iMDS, and PHATE face significant challenges in distinguishing these subtypes. They struggle to segregate any one subtype distinctly or allocate these cells to appropriate categories.

ADM can reveal subtle differences between cell subtypes. In the visualization analysis, we identified four clear clusters representing a continuum of lineages from adaptive to innate characteristics. Cluster 1 primarily consists of adaptive T cells (CD4+ and CD8+) expressing molecules related to antigen presentation and specificity recognition. Cluster 2 encompasses iNKT, MAIT, and Vδ2 T cells, subtypes that express genes associated with rapid immune responses, reflecting their innate-like T cell features. Cluster 3 includes $\lambda\delta$1 and $\lambda\delta$2 T cells, which are characterized by gene expressions related to innate immune surveillance and cytotoxic functions. Cluster 4 is composed of NK cells, whose expression profiles are directly linked to innate immune responses.

Compared to the *Meta-Spec* method, which could partially distinguish some cell subtypes, *ADM* is notably clearer in terms of category boundary definition and inter-class distance. The *ARI* and *NMI* values of *ADM* are slightly higher than those of *Meta-Spec* (Fig. 3A). By leveraging the mechanism of information diffusion, *ADM* not only enhanced the accuracy of cell type identification but also provided new perspectives on the functional balance of immune cells in immunological defense, offering important insights for a deeper understanding of the complexity of the immune system and the development of targeted immune therapeutic strategies.

Quake dataset: mixed cell types of human immune and respiratory systems.

The dataset from Isakova et al. [42] initially comprised 1,676 single cells profiled across 23,341 genes. We applied a filtering criterion to retain genes expressed in at least 25% of the cells, resulting in a final set of 3,431 genes for subsequent analyses.

Figure 6. Visualization of the Brain5k dataset. The complete comparative results can be found in Fig. S9 in the supplementary materials.

It includes a diverse range of cell types, including stromal cells, lung epithelial cells, lung endothelial cells, B cells, leukocytes, monocytes, T cells, classical monocytes, ciliated columnar cells of the tracheobronchial tree, myeloid cells, and natural killer cells. These cell types represent various components of the human immune and respiratory systems.

Figure 7 shows the visualization of *ADM* and representative comparison methods. TSNE, LEIM, and UMAP demonstrate their ability to distinguish T cells and lung endothelial cells from other cell types with noticeable clarity. However, the intra-class aggregation of these individual methods is not as good as meta-dimension reduction methods *Meta-Spec* and *ADM*.

Compared to *Meta-Spec*, *ADM* exhibits more apparent separation between cell types such as lung epithelial cells, epithelial cells and stromal cells while maintaining higher inter-class separation and intra-class cohesion. Particularly, the distinction between classical monocyte and monocyte is more pronounced in *ADM*'s visualization.

Quantitative metrics, including *ARI* and *NMI*, further support the performance differences between *ADM* and *Meta-Spec*. *ADM* demonstrates an *ARI* of 0.32 and an *NMI* of 0.526, whereas *Meta-Spec* exhibits an *ARI* of 0.30 and an *NMI* of 0.510. Additionally, *ADM* also achieves higher *CCI* scores than *Meta-Spec* on this dataset.

CCLE dataset: cancer cell lines bulk sequencing data.

The CCLE dataset [43] is a comprehensive cancer cell line resource, encompassing over 20 different types of cancer cell lines derived from a variety of tissues, including skin, central nervous system, soft tissue, ovary, blood, and lymphoid tissues. Integrating multimodal data such as gene expression, micro-RNA expression, and protein expression, CCLE offers a rich information resource for the study of cancer mechanisms, identification of biomarkers, and discovery of new therapeutic targets. In this work, we focused particularly on the gene and micro-RNA expression data provided by CCLE, analyzed as CCLE_mRNA and CCLE_miRNA, respectively. The CCLE_mRNA dataset details the expression levels of genes within different cancer cell lines, while the CCLE_miRNA dataset provides the expression profiles of micro-RNAs, using bulk RNA sequencing technique. We analyzed mRNA and micro-RNA expression sequencing results from 901 cell lines within CCLE, which include 14 997 mRNA and 700 micro-RNAs, respectively.

As shown in supplementary Fig. S5 and Fig. 8, dimension-reduction and visualization based on mRNA and miRNA expression data indicate that most clustering methods effectively differentiated cancer cell lines derived from blood and lymphoid tissues from others. Meta-reduction techniques, especially *ADM*, exhibits excellent intra-group compactness, reflecting cellular similarity at the molecular level. Single dimension-reduction techniques, although capable of identifying blood and lymphoid cells, result in more blurred cluster boundaries and lower overall density separation.

In the visualization of central nervous system cells, the output of *ADM* is clustered into a tight group, whereas the *Meta-Spec*

Figure 7. Visualization of the Quake dataset. The abbreviations in the figure correspond to their full names: BC: B cell, CCCT: ciliated columnar cell of tracheobronchial tree, ClMono: classical monocyte, EpLung: epithelial cell of lung, Leuk: leukocyte, LungEnd: lung endothelial cell, Mono: monocyte, Myel: myeloid cell, NK: natural killer cell, Strom: stromal cell, TC: T cell. The complete comparative results can be found in Fig. S10 in the supplementary materials.

technique displays more dispersed data points. *ADM* also demonstrates superior performance in clustering colon cells, whereas the *Meta-Spec* technique failed to differentiate this cell category distinctly. These observations suggest that *ADM* is more effective in maintaining intrinsic cellular similarity. Quantitatively, the *ADM* technique outperformed *Meta-Spec* for mRNA, with an *ARI* of 0.3123 and a *NMI* of 0.5452 (Fig. S4 of supplementary materials), compared to the *ARI* of 0.2664 and *NMI* of 0.5019 for *Meta-Spec* (Fig 3). For CCLE_miRNA, *ADM*'s *ARI* is 0.2233, and *NMI* is 0.4015, surpassing *Meta-Spec*'s *ARI* of 0.2081 and *NMI* of 0.3652.

Performance of ADM on Other types of omics data

To further evaluate the broad applicability of *ADM* across various omics datset, we analyzed its performance on three distinct datasets: ATAC-seq, metabolomics, and spatial proteomics data.

Brain5k Dataset: the scATAC-seq of mouse cortical neurons The Brain5k dataset consists of 10x ATAC-seq data from adult mouse cortical neurons [44], comprises 2317 cells with 155 093 features, categorized into 10 distinct cell types. As shown in Fig. 6, traditional dimension reduction methods (e.g., Isomap, Sammon mapping, UMAP) struggle to effectively differentiate cell types, while methods like kPCA, MDS, t-SNE, and PHATE only broadly distinguish major cell groups. In contrast, meta-dimension reduction methods like *Meta-Spec* and *ADM* successfully capture the hierarchical relationships among cell types. These methods clearly separate major neuronal subtypes (such as L2/3 IT, L4, L6 CT) while maintaining continuity between closely related cell types. *ADM* further refines these results, precisely distinguishing annotated cell types and revealing potential developmental and functional associations,such as the gradual relationships among L2/3 IT, L4, L5 IT, and L6 CT neurons, aligning with their spatial arrangement and functional connections within the cortex.

Quantitative metrics further support these visual observations. *ADM*'s *ARI* of 0.5152 is significantly higher than *Meta-Spec*'s 0.3443. Similarly, *ADM*'s *NMI* of 0.5343, compared to *Meta-Spec*'s 0.5018, better reflects its efficacy in preserving the information within the dataset.

Hbrc dataset: the metabolomics data of Human Breast Cancer To demonstrate *ADM*'s versatility across different omics modalities, we analyzed a human breast cancer metabolomics dataset [45], comprising 1327 samples with 489 features, categorized into 7 different cancer subtypes or stages. The visualization (Supplementary Fig. S6) shows that *ADM* technique tends to cluster the basal and claudin-low subtypes closely, which is consistent with their shared triple-negative status. Similarly, Luminal A (LumA) and Luminal B (LumB) subtypes show proximity in the reduced space, reflecting their underlying

Figure 8. Visualization of the CCLE_miRNA dataset. The abbreviations in the figure correspond to their full names: AG: autonomic ganglia, Bo: bone, Br: breast, CNS: central nervous system, En: endometrium, Fb: fibroblast, HL: haematopoietic and lymphoid tissue, Ki: kidney, LIN: large intestine, LIV: liver, LUN: lung, Oe: oesophagus, Ov: ovary, Pa: pancreas, Sk: skin, STI: soft tissue, STO: stomach, Th: thyroid, UAT: upper aerodigestive tract, UT: urinary tract. The complete comparative results can be found in Fig. S11 in the supplementary materials.

hormonal similarities (both typically ER+ and/or PR+). These observations demonstrate *ADM*'s ability to capture and represent biologically relevant relationships in the reduced dimensional space, potentially aiding in the identification of metabolic signatures associated with different breast cancer subtypes.

Spleen Dataset: the spatial proteomics data of mouse spleen The spatial proteomics data from mouse spleen [46] comprises 2568 spots and 21 proteins. Visualization results show that traditional techniques struggle to differentiate between various categories or represent their relationships effectively (Fig. 9). In contrast, meta-dimensionality reduction methods, *Meta-Spec* and *ADM* not only distinguish different immune cell types but also capture their spatial developmental relationships. Marginal Zone Macrophages are visually positioned as a bridge between Red Pulp Macrophages in the red pulp and Marginal Metallophilic Macrophages near the white pulp, reflecting their intermediate location in the spleen. Furthermore, *ADM* clearly separates T cells and B cells in the visualization, further demonstrating its robustness in capturing underlying data structures through information diffusion.

Across all three datasets, visualization results and quantitative analysis using *ARI*, *NMI*, *ASW*, and *CCI* metrics consistently demonstrate *ADM* 's superior performance over *Meta-Spec* (Fig. 3 and Figs S2, S4 of supplementary materials). *ADM* excels in revealing intra-class compactness, inter-class differentiation, and preserving category consistency. In summary, while *Meta-Spec* provides valuable data insights, *ADM* demonstrates superior performance in clustering purity and fidelity of data representation.

Efficacy of *ADM* in single-cell annotation

To assess the practical utility of the *ADM* method, we analyzed several single-cell transcriptomics datasets, which comprised various cell types with initially concealed labels. The dimension reduction and visualization were executed as outlined in Section 3.3. Subsequently, we applied the DBSCAN algorithm [47] for density-based spatial clustering to delineate distinct clusters within the two-dimensional projection. We hypothesized these clusters correspond to either homogeneous cell types or similar cellular states. Differential expression analysis across these clusters was conducted using Seurat [48], providing the foundation for each cluster's annotation.

Figure 10(A) and 10(B) shows that *ADM* highlighted the upregulated expression patterns of the MAG and APOD genes in Cluster 1, both identified as markers for oligodendrocyte cells [49–54]. This suggests that this cluster predominantly consists of oligodendrocyte cells, aligning with experimental findings from the dataset [40]. However, alternative dimension reduction methods depicted more dispersed expressions of MAG and APOD across clusters. Additionally, in Clusters 5 and 6, differential expression patterns of PLPP3 (Fig. 10C) and PLTP (Fig. 10D) were observed,

Figure 9. Visualization of the Spleen dataset. The complete comparative results can be found in Fig. S12 in the supplementary materials.

which are markers for astrocytes [51] and endothelial cells [52], respectively. These patterns are consistent with the cell annotation labels [40].

ADM's ability is further corroborated by the precise alignment of key gene expression with established cellular markers, as demonstrated in the visualized clusters. This effectively validates *ADM*'s robust capability to perform stable dimension reduction and highlights its effectiveness in preserving the intricate structures of biological data.

Discussion and conclusion

In this study, we presented ***ADM***, a novel meta-dimension reduction method designed to tackle the challenges associated with comprehending and visualizing complex, high-dimensional datasets.

ADM leverages graph diffusion theory and integrates outputs from multiple dimension reduction methods, addressing three major limitations of existing approaches while providing a more robust framework for data analysis and visualization.

- **First**, *ADM* excels at capturing both **local and global structures** within the data. Its adaptive diffusion mechanism dynamically selects the optimal diffusion time scale for each sample, enabling it to represent intricate multi-scale structures more effectively.
- **Second**, *ADM* is proficient in **mitigating noise and distortions**. By integrating information from diverse dimension reduction techniques and applying a dynamic Markov process, it reduces the influence of noise and method-specific distortions, leading to more reliable outcomes.
- **Third**, *ADM* reveals **dynamic nonlinear structures** by transforming results from Euclidean space into an information space. This transformation allows *ADM* to uncover intrinsic nonlinear manifolds and capture dynamic geometric relationships between samples—interactions that traditional static methods often fail to identify.

ADM's performance has been extensively evaluated on both simulated and real-world datasets, spanning various domains, include single-cell RNA sequencing (scRNA-seq) and other omics data such as proteomics, metabolomics, and epigenomics. On the one hand, findings from simulated datasets demonstrated *ADM*'s efficacy in mitigating interference from low-quality outputs and restoring the underlying true data structure. On the other hand, extensive results from real-world datasets confirmed that *ADM* generates **higher-quality low-dimensional representations**, providing clearer separation between biological groups and revealing more meaningful patterns across various omics data types compared to existing dimension reduction methods.

However, *ADM* faces two main challenges that warrant future research. First, due to the necessity to process multiple dimension

Figure 10. Visualization of marker genes on the dimensionality reduction results of kPCA, iMDS, t-SNE, UMAP, meta-spec, and ADM. (A) Visualization of the distribution of MAG in the Oihane dataset. (B) Visualization of the distribution of APOD. (C) Visualization of the distribution of PLPP3. (D) Visualization of the distribution of PLTP.

reduction results and perform graph diffusion, the computational cost of *ADM* may be higher than that of single dimension reduction methods. This trade-off between computational complexity and representation quality should be considered when applying *ADM* to very large datasets. Second, while *ADM* has the ability to discern and prioritize high-quality outputs, it can still be influenced by extreme inputs, such as the HLLE results. Thus, efforts to improve the computational efficiency of *ADM*, particularly for large-scale datasets, through algorithm optimization or parallel computing techniques would be beneficial. Additionally, enhancing *ADM*'s robustness against extreme inputs by refining the algorithm to more effectively identify and mitigate the influence of outlier results would further improve its performance and reliability.

In conclusion, *ADM* represents a promising approach in meta-dimension reduction, offering a robust, adaptive, and comprehensive solution for researchers analyzing complex high-dimensional datasets. By addressing key challenges in capturing multi-scale structures, mitigating noise, and revealing dynamic nonlinear relationships, *ADM* enhances our ability to visualize intricate data patterns across various scientific domains. With continued development and optimization, *ADM* has the potential to become a powerful, general-purpose tool for high-dimensional data analysis.

Key Points

- We develop a novel meta-dimension reduction method called *ADM*, which utilizes graph diffusion theory and dynamic Markov processes to integrate multiple dimension reduction techniques, enhancing visualization and data analysis.
- *ADM* adapts the time scale of information diffusion according to sample-specific characteristics, transforming spatial metrics into dynamic diffusion distances to explore the dataset's intricate patterns and structures.
- *ADM* demonstrates superior robustness in analyzing high-dimensional and noisy datasets compared to current techniques.
- Extensive evaluations on simulations and diverse public omics datasets demonstrate *ADM*'s capability to achieve clearer separation among different biological conditions and groups, thus improving biological interpretations and insights.

Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

Conflict of interest: The authors declare no competing interests.

Funding

This work was partially supported by the National Key R&D Program of China (2022ZD0116004), the TianYuan funds for Mathematics of the National Science Foundation of China (Grant No. 12326604), Guangdong Talent Program (2021CX02Y145), Guangdong Provincial Key Laboratory of Big Data Computing, and Shenzhen Key Laboratory of Cross-Modal Cognitive Computing (ZDSYS20230-626091302006).

Availability of data and materials

The ADM package source code is available at https://github.com/Seven595/ADM. For full result reproduction, including code and datasets, please refer to https://github.com/Seven595/ADMReproduce. The datasets presented in this study can be also acquired from the following websites or accession numbers: (1) the mouse midbrain and striatum scRNAseq data (GSE148393); (2) the human lymphocyte population dataset (GSE81772); (3) the mixed cell types of human immune and respiratory systems (GSE151334); (4) the spatial proteomics data from mouse spleen data (GSE198353). (5) the cancer cell lines bulking sequencing data (https://depmap.org/portal/data_page/?tab=allData); (6) the scATAC-seq data of mouse cortical neurons is available at http://download.gao-lab.org/GLUE/dataset/10x-ATAC-Brain5k.h5ad ; (7) the metabolomics data of human breast cancer is available at https://github.com/EddieFua/medNet.

References

1. Donoho D. 50 years of data science. *J Comput Graph Stat* 2017;**26**:745–66. https://doi.org/10.1080/10618600.2017.1384734.
2. McInnes L, Healy J, Melville J. UMAP: uniform manifold approximation and projection for dimension reduction. *Journal of Open Source Software (JOSS)*. 2018;**3**:861.
3. Wattenberg M, Viégas F, Johnson I. How to use t-SNE effectively. *Distill* 2016;**1**:e2. https://doi.org/10.23915/distill.00002.
4. Wang Y, Huang H, Rudin C. *et al*. Understanding how dimension reduction tools work: An empirical approach to deciphering t-SNE, UMAP, TriMAP, and PaCMAP for data visualization. *J. Mach. Learn. Res.* 2021;**22**:1–73.
5. Cheng J, Liu H, Wang F. *et al*. Silhouette analysis for human action recognition based on supervised temporal t-SNE and incremental learning. *IEEE Trans Image Process* 2015;**24**:3203–17. https://doi.org/10.1109/TIP.2015.2441634.
6. Olivon F, Elie N, Grelier G. *et al*. Metgem software for the generation of molecular networks based on the t-SNE algorithm. *Anal Chem* 2018;**90**:13900–8. https://doi.org/10.1021/acs.analchem.8b03099.
7. Dorrity MW, Saunders LM, Queitsch C. *et al*. Dimensionality reduction by umap to visualize physical and genetic interactions. *Nat Commun* 2020;**11**:1537. https://doi.org/10.1038/s41467-020-15351-4.
8. Meng C, Zeleznik OA, Thallinger GG. *et al*. Dimension reduction techniques for the integrative analysis of multi-omics data. *Brief Bioinform* 2016;**17**:628–41. https://doi.org/10.1093/bib/bbv108.
9. Hie B, Peters J, Nyquist SK. *et al*. Computational methods for single-cell RNA sequencing. *Annu Rev Biomed Data Sci* 2020;**3**:339–64. https://doi.org/10.1146/annurev-biodatasci-012220-100601.
10. Chen G, Ning B, Shi T. Single-cell RNA-seq technologies and related computational data analysis. *Front Genet* 2019;**10**:441123. https://doi.org/10.3389/fgene.2019.00317.
11. Narayan A, Berger B, Cho H. Assessing single-cell transcriptomic variability through density-preserving data visualization. *Nat Biotechnol* 2021;**39**:765–74. https://doi.org/10.1038/s41587-020-00801-7.
12. Moon TK, Stirling WC. *Mathematical Methods and Algorithms for Signal Processing*. 1st ed. Pearson; 1999.
13. Carroll JD, Arabie P. Multidimensional scaling. In: Birnbaum MH (ed.), *Handbook of Perception and Cognition (Second Edition), Measurement, Judgment and Decision Making*, pp. 179–250. Academic Press; 1998.
14. Sammon JW. A nonlinear mapping for data structure analysis. *IEEE Trans Comput* 1969;**C-18**:401–9. https://doi.org/10.1109/T-C.1969.222678.
15. Tenenbaum JB, de Silva V, Langford JC. A global geometric framework for nonlinear dimensionality reduction. *Science* 2000;**290**:2319–23. https://doi.org/10.1126/science.290.5500.2319.
16. Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Kernel principal component analysis. In: *International Conference on Artificial Neural Networks*, pp. 583–8. Berlin, Heidelberg: Springer, 1997. https://doi.org/10.1007/BFb0020217.
17. Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. *Science* 2000;**290**:2323–6. https://doi.org/10.1126/science.290.5500.2323.
18. Vactions.

Recognit 2000;**33**:1475–85. https://doi.org/10.1016/S0031-3203(99)00138-7.
33. Parisi F, Strino F, Nadler B. *et al*. Ranking and combining multiple predictors without labeled data. *Proc Natl Acad Sci* 2014;**111**:1253–8. https://doi.org/10.1073/pnas.1219097111.
34. Liu Z-G, Pan Q, Dezert J. *et al*. Combination of classifiers with optimal weight based on evidential reasoning. *IEEE Trans Fuzzy Syst* 2017;**26**:1217–30. https://doi.org/10.1109/TFUZZ.2017.2718483.
35. Mohandes M, Deriche M, Aliyu SO. Classifiers combination techniques: A comprehensive review. *IEEE Access* 2018;**6**:19626–39. https://doi.org/10.1109/ACCESS.2018.2813079.
36. Ma R, Sun ED, Zou J. A spectral method for assessing and combining multiple data visualizations. *Nat Commun* 2023;**14**:780. https://doi.org/10.1038/s41467-023-36492-2.
37. Bertagnolli G, De Domenico M. Diffusion geometry of multiplex and interdependent systems. *Phys Rev E* 2021;**103**:042301. https://doi.org/10.1103/PhysRevE.103.042301.
38. Knorr EM, Ng RT. A unified notion of outliers: Properties and computation. In: *Third International Conference on Knowledge Discovery and Data Mining (KDD)*, AAI Press, Newport Beach, CA. 1997, pp. 219–22.
39. Hammond D, K, Gur Y, Johnson CR. Graph diffusion distance: A difference measure for weighted graphs based on the graph Laplacian exponential kernel. In:*In 2013 IEEE Global Conference on Signal and Information Processing*, pp. 419–22. Austin, TX, USA: IEEE, 2013.
40. Huarte OU, Kyriakis D, Heurtaux T. *et al*. Single-cell transcriptomics and in situ morphological analyses reveal microglia heterogeneity across the nigrostriatal pathway. *Front Immunol* 2021;**12**:639613. https://doi.org/10.3389/fimmu.2021.639613.
41. Gutierrez-Arcelus M, Teslovich N, Mola AR. *et al*. Lymphocyte innateness defined by transcriptional states reflects a balance between proliferation and effector functions. *Nat Commun* 2019;**10**:687. https://doi.org/10.1038/s41467-019-08604-4.
42. Isakova A, Neff N, Quake SR. Single-cell quantification of a broad RNA spectrum reveals unique noncoding patterns associated with cell types and states. *Proc Natl Acad Sci* 2021;**118**:e2113568118. https://doi.org/10.1073/pnas.2113568118.
43. Ghandi M, Huang FW, Jané-Valbuena J. *et al*. Next-generation characterization of the cancer cell line encyclopedia. *Nature* 2019;**569**:503–8. https://doi.org/10.1038/s41586-019-1186-3.
44. Cao Z-J, Gao G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nat Biotechnol* 2022;**40**:1458–66. https://doi.org/10.1038/s41587-022-01284-4.
45. Cai Q, Yinghao F, Lyu C. *et al*. A new framework for exploratory network mediator analysis in omics data. *Genome Res* 2024;**34**:642–54. https://doi.org/10.1101/gr.278684.123.
46. Long Y, Ang KS, Sethi R. *et al*. Deciphering spatial domains from spatial multi-omics with spatialglue. *Nat Methods* 2024;**21**:1–10.
47. Hahsler M, Piekenbrock M, Doran D. DBSCAN: fast density-based clustering with R. *J Stat Softw* 2019;**91**:1–30.
48. Hao Y, Stuart T, Kowalski MH. *et al*. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat Biotechnol* 2024;**42**:293–304. https://doi.org/10.1038/s41587-023-01767-y.
49. Zhang J, Velmeshev D, Hashimoto K. *et al*. Neurotoxic microglia promote TDP-43 proteinopathy in progranulin deficiency. *Nature* 2020;**588**:459–65. https://doi.org/10.1038/s41586-020-2709-7.
50. Pool A-H, Wang T, Stafford DA. *et al*. The cellular basis of distinct thirst modalities. *Nature* 2020;**588**:112–7. https://doi.org/10.1038/s41586-020-2821-8.
51. Huang M, Modeste E, Dammer E. *et al*. Network analysis of the progranulin-deficient mouse brain proteome reveals pathogenic mechanisms shared in human frontotemporal dementia caused by GRN mutations. *Acta Neuropathol Commun* 2020;**8**:1–25. https://doi.org/10.1186/s40478-020-01037-x.
52. Ye Emily W, Pan L, Zuo Y. *et al*. Detecting activated cell populations using single-cell RNA-seq. *Neuron* 2017;**96**:313–329.e6. https://doi.org/10.1016/j.neuron.2017.09.026.
53. Skinnider MA, Squair JW, Kathe C. *et al*. Cell type prioritization in single-cell data. *Nat Biotechnol* 2021;**39**:30–4. https://doi.org/10.1038/s41587-020-0605-1.
54. Congxue H, Li T, Yingqi X. *et al*. CellMarker 2.0: An updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data. *Nucleic Acids Res* 2023;**51**:D870–6.

Briefings in Bioinformatics, 2025, **26(1)**, bbae612
https://doi.org/10.1093/bib/bbae612
Case Study