Computational and Structural Biotechnology Journal 23 (2024) 3929–3937

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research Article

Cell-to-cell distance that combines gene expression and gene embeddings

Fangfang Guo, Dailin Gan, Jun Li *

Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, Notre Dame, IN 46556, USA

ABSTRACT

The application of large-language models (LLMs) to single-cell gene-expression data has introduced a new type of data that includes a gene-embedding matrix, in addition to the experimentally obtained gene-expression matrix. This paper addresses a fundamental problem in analyzing such data: how to effectively combine the information from both matrices to better define cell-to-cell distance. We identify a computationally feasible solution that demonstrates superior ability to cluster cells of the same type across all six real datasets we tested, underscoring its advantage as a measure of cell-to-cell distance.

1. Introduction

The integration of transformer models [36] and large language models (LLMs) [10] into the analysis of gene-expression data has garnered increasing interest. Recently developed models and tools, such as scBERT [38], Geneformer [34], scGPT [16], and scFoundation [18], have demonstrated their ability to improve predictive capabilities across numerous applications. These models typically employ a transfer learning framework [30], wherein transformer-based architectures are initially trained on extensive collections of single-cell RNA-seq data, comprising thousands of datasets and millions of cells. The models are then fine-tuned on much smaller, user-specific datasets for various applications. This approach has proven to be particularly effective, leveraging the rich, albeit indirect, information about gene interactions embedded in the training data to capture complex gene-gene relationships [38,34,16,18]. However, developing these models from scratch requires substantial resources. The extensive data collection and computational demands render the process prohibitively expensive for most research groups.

GenePT [12] offers an innovative solution that eliminates the need for training models from scratch. It utilizes ChatGPT, a robust pretrained general LLM, to elucidate gene-gene interactions. Specifically, GenePT leverages ChatGPT's embedding function to transform standard NCBI gene descriptions [9] into "embeddings," which are numeric vectors that effectively capture the textual information [29]. This process ensures that details about interactions or functional regulations between genes, typically found in NCBI descriptions, are accurately represented in the embeddings. Thus, these embeddings provide a novel method for summarizing and utilizing prior knowledge about genes and their interactions, serving a similar function to that of transformer-based models developed from scratch, but with significantly greater cost-effectiveness, enhancing its potential for widespread adoption. Furthermore, this method of using ChatGPT-generated embeddings could be broadly applied to encapsulate prior knowledge in textual form for various biological entities beyond genes.

This novel strategy, however, introduces a new problem: how to efficiently utilize the information contained in the embeddings. Typically, original gene expression data is stored as a data matrix. If the scRNA-seq experiment measures the expression of p genes in n cells, the resulting gene-expression matrix is of size $n \times p$, with the element in row i and column j representing the expression of gene j in cell i [22]. Similarly, gene embeddings can be stored as a data matrix. If the dimension of each embedding is d, then this matrix is of size $p \times d$, with the i-th row representing the embedding of gene i in the d-dimensional space. We refer to these as the "gene-expression matrix" and the "gene-embedding matrix", respectively, as illustrated in Fig. 1. The gene-expression matrix encapsulates the experimental data, whereas the gene-embedding matrix provides supplementary information about the features of the genes involved in the experiment. While traditional scRNA-seq data analysis has solely considered the gene-expression matrix [21,4], we now face the challenge of integrating and efficiently utilizing both matrices in our analyses.

The solution to this question may vary by application, and a universal approach is unlikely to exist. This paper explores one specific aspect of this challenge: how to effectively integrate information from two matrices to more accurately define cell-to-cell distance. Measuring the distance between two observations (cells) is a fundamental problem in machine learning and holds intrinsic importance [39]. For instance, many commonly used visualization and classification methods depend entirely on the distance matrix between observations [19]. As shown

* Corresponding author.
E-mail address: jun.li@nd.edu (J. Li).

https://doi.org/10.1016/j.csbj.2024.10.044
Received 10 August 2024; Received in revised form 14 October 2024; Accepted 27 October 2024

Available online 4 November 2024

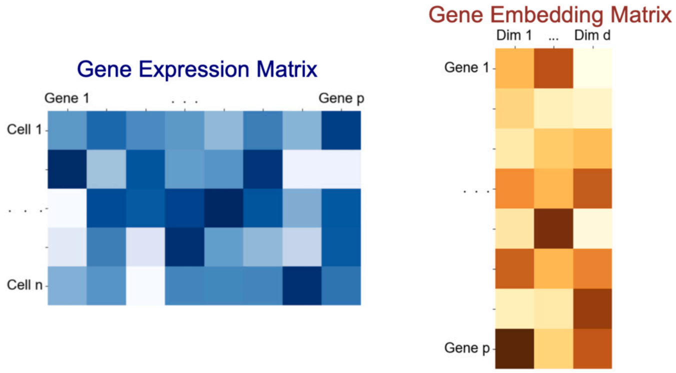

Fig. 1. The new form of data that includes two matrices: a gene-expression matrix and a gene-embedding matrix. Their dimensions are $n \times p$ and $p \times d$, respectively.

in Fig. 2A, traditional methods for assessing cell-to-cell distance focus solely on the gene-expression matrix, typically employing cosine distance between the expression vectors of two cells, which we refer to as "gene-space distance." In contrast, the GenePT paper [12] proposes matrix-multiplying the two matrices to create a cell-embedding matrix of dimensions $n \times d$, and then computing a distance matrix based exclusively on this new matrix, again utilizing cosine distance between cell pairs. This methodology is illustrated in Fig. 2B. Recognizing that the cell-embedding matrix can be viewed an expression matrix in the embedded space, we call this measure "embedded-space distance." However, we suggest that this approach may still be inefficient, a topic that will be further examined in the Results section.

In this paper, we propose a new method to compute cell-to-cell distance that effectively combines the information in the gene-expression and gene-embedding matrices. We will test the efficiency of our method using multiple real scRNA-seq datasets and compare it with both the gene-space distance and the embedded-space distance.

2. Methods

2.1. Limitations of the embedded-space distance

As introduced in the introduction section and illustrated in Figs. 1 and 2, we have two pieces of data: the gene-expression matrix and the gene-embedding matrix. The traditional method for computing cell-to-cell distance, which is called gene-space distance, considers only the gene-expression matrix, calculating the cosine distance between its row vectors to obtain the cell-to-cell distance matrix. The embedded-space distance, proposed by GenePT [12], defines the distance in two steps: first, the gene-expression matrix (dimension: $n \times p$) and the gene-embedding matrix (dimension: $p \times d$) are matrix-multiplied to produce a cell-embedding matrix (dimension: $n \times d$); then, the cell-to-cell distance is determined by calculating the cosine distance between the row vectors of this cell-embedding matrix.

We argue that GenePT's procedure may be problematic in two aspects. First, the initial transformation from the gene space to the embedded space, which represents each cell by d embedded dimensions instead of p genes through matrix multiplication, is linear, although there is no inherent reason why this transformation must be linear. Second, the embedded-space distance is computed solely based on the cell-embedding matrix, but this matrix contains much less information than the gene-expression and gene-embedding matrices combined. Matrix multiplication can be seen as a weighted average of gene embeddings and does not take into account the distances between gene embeddings. In fact, this matrix may even contain less information than the gene-expression matrix alone. This can be easily understood when d is very small: in this case, each cell is represented by far fewer features in the embedded space than the number of genes in the gene space. Consequently, the embedded-space distance may be even less informative than the gene-space distance. Real data analysis will demonstrate such examples, even when d is not small.

2.2. A plausible solution: word mover's distance

In seeking a definition of distance that efficiently combines information from both the gene-expression and gene-embedding matrices, we found that our problem is analogous to the problem of document retrieval (see, e.g., [11] for an introduction to document retrieval). In document retrieval, the goal is to find the document most similar to a given document, which involves efficiently defining the distance between two documents.

In this context, each document is a collection of words, each with its appearance frequency in the document. Each word can be represented by an embedding vector, obtained using techniques such as "word2vec" [27]. The challenge here is to define the document-to-document distance based on the embeddings of individual words and their frequency vectors.

We realized that by replacing words with genes, documents with cells, and the frequency vectors of words with gene expression profiles, the two problems become identical, provided that we disregard the order of words as they appear in documents in the document retrieval problem. Therefore, we conducted a comprehensive literature review, focusing primarily on relevant works in computer science and machine learning journals. We identified a metric called "word mover's distance" (WMD) [23], also known as "Earth mover's distance" or "Wasserstein metric," depending on the context and application area. This metric is highly popular and has proven its efficiency across many real datasets (e.g., [37,24]). Below, we briefly introduce WMD as it is applied to our problem.

Consider measuring the distance between cell 1 and cell 2. Let $\tilde{x}_{1j} = x_{1j} / \sum_{k=1}^{p} x_{1k}$ and $\tilde{x}_{2j} = x_{2j} / \sum_{k=1}^{p} x_{2k}$ be the frequency (i.e., standardized expression) of gene j in the two cells. The WMD between these two cells is defined as

$$\mathrm{WMD}(\tilde{x}_1, \tilde{x}_2) = \min_{\Gamma \in \mathbb{R}_+^{p \times p}} \sum_{i=1}^{p} \sum_{j=1}^{p} \Gamma_{ij} d_{i,j} \tag{1}$$

subject to the constraints

$$\sum_{j=1}^{p} \Gamma_{ij} = \tilde{x}_{1i}, \quad \forall i \in \{1, \ldots, p\}, \tag{2}$$

and

$$\sum_{i=1}^{p} \Gamma_{ij} = \tilde{x}_{2j}, \quad \forall j \in \{1, \ldots, p\}. \tag{3}$$

In the above, Γ is the transport plan, Γ_{ij} denotes the amount of "mass" transported from gene i in cell 1 to gene j in cell 2, and $d_{i,j}$ is the Euclidean distance between the embeddings of gene i and gene j. The goal is to find the transport plan Γ that minimizes the overall transportation cost, which is the sum of transported mass (reflected by the gene expression) weighted by the distances between the source and destination of the transportation (reflected by the gene embeddings). Therefore, WMD efficiently combines information from both gene-expression and gene-embedding matrices, making it a promising definition of cell-to-cell distance.

Unfortunately, the computational cost of WMD is very high. The time complexity of computing the WMD distance between a pair of cells containing p genes is $O(p^3 \log p)$ [23]. Considering that the number of genes p is usually larger than 10^3, $p^3 \log p$ becomes quite large. The situation is exacerbated when calculating not just a single distance, but $\frac{1}{2}n(n-1)$ pairwise distances between n cells, where n is often not less than 10^3. Therefore, we were compelled to find a more computationally efficient definition of distance.

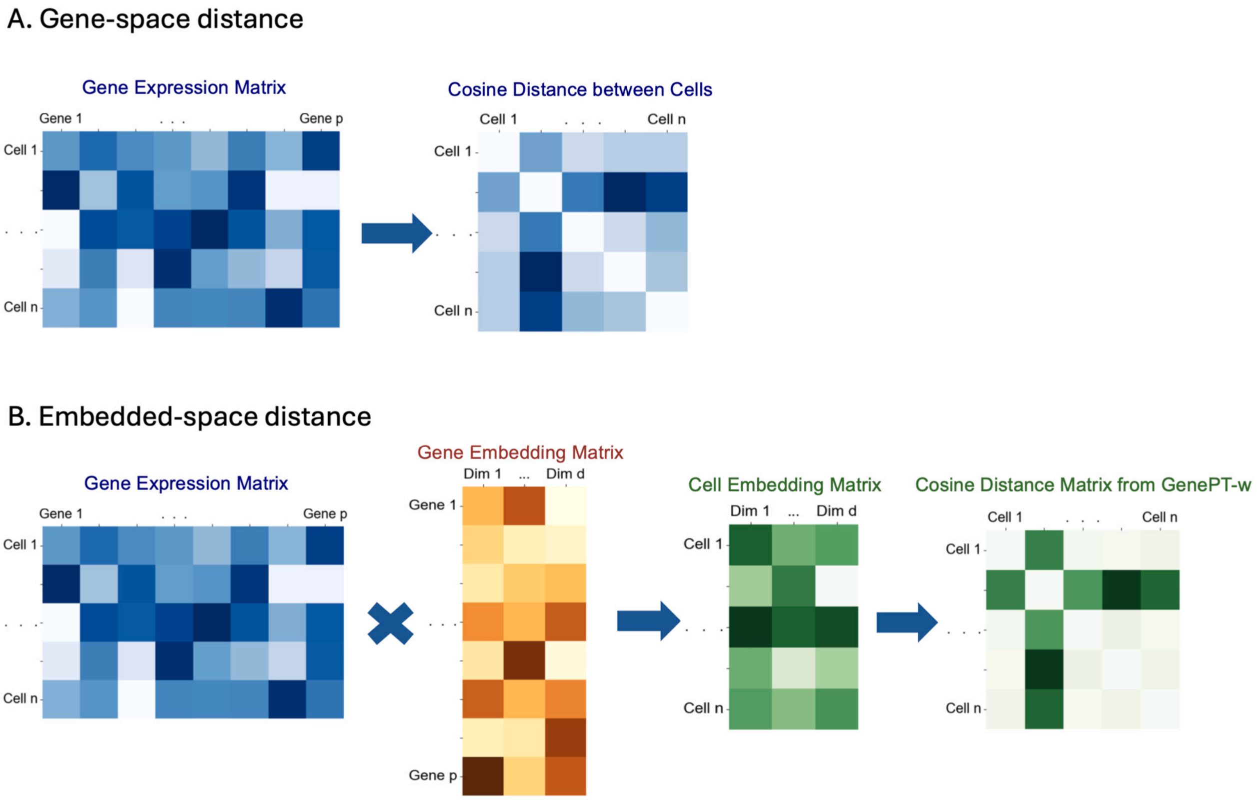

Fig. 2. Current methods for defining cell-to-cell distances. **(A)** gene-space distance, defined as the cosine distance between rows of the gene-expression matrix. **(B)** embedded-space distance, defined as the cosine distance between rows of the cell embedding matrix, which is obtained by the matrix multiplication of the gene-expression and gene-embedding matrices. The dimensions of the cell embedding matrix are $n \times d$.

2.3. Attempts with other WMD variants

Variants of WMD with higher computational efficiency have been proposed in the literature of document retrieval, and we have considered many of them.

For example, Relaxed WMD (RWMD) [23] reduces the computational complexity to $O(p^2 \log p)$ by dropping one of the two sets of constraints in WMD. However, in our application of defining cell-to-cell distance, RWMD always equals 0, rendering it useless. The proof of this property is given in Section 1 of the Supplementary Material.

Another computationally efficient method is the Word Centroid Distance (WCD) [23], which approximates documents by averaging the embedding vectors of all words to form a centroid. However, studies have demonstrated that this method performs significantly worse than the original WMD [8,23].

To improve WCD, it was proposed to combine WCD with Iterative Constrained Transfers (ICT) for pruning [3]. ICT reduces the computation of WMD by iteratively narrowing the search space after pre-sorting with WCD. However, the additional effectiveness brought by the pruning strategy is limited in high-dimensional spaces, where the distance distribution between points tends to be uniform, affecting the effectiveness of pruning [6].

2.4. The scHOTT algorithm

After extensive exploration, we have settled on the Hierarchical Optimal Topic Transport (HOTT) algorithm [41], which was also originally developed in the context of natural language processing. We have adapted it to single-cell RNA-seq data, and we refer to it as scHOTT. Below, we provide a brief introduction to this methodology. A more detailed mathematical description is available in Sections 2 and 3 of the Supplementary Material.

The computation of WMD between two cells is notably intensive due to the large number of genes, p, in a cell's expression profile. scHOTT addresses this issue by introducing a hierarchical structure by adding a layer termed "cell functions" (it is termed "topics" in the original publication) between the expression profile of a cell and the expression of individual genes. The method posits that the expression profile can be decomposed into a set of K cell functions, each with a specific weight, referred to as function weights (originally "distribution of topics"). Each cell function is subsequently composed of a set of r genes, each also weighted, referred to as gene weights (originally "distribution of features"). It is important to note that these K cell functions are shared among all cells, though each cell has its own set of weights for these functions. Moreover, each cell function comprises a fixed set of r genes with a fixed set of gene weights, which are independent of its function weight within a cell. Additionally, the same gene may contribute to multiple functions with varying probabilities, which reflects the biological complexity where genes often participate in multiple biological processes and pathways. In the scHOTT algorithm, a probabilistic model known as Latent Dirichlet Allocation (LDA) (developed by [7] and applied to genomic data in previous studies such as [40,17,1]) is employed to identify the K cell functions and to determine the function and gene weights.

As depicted in Fig. 3, WMD between a pair of cells is computed in a bottom-up approach. Initially, the WMD distance between each pair of cell functions is determined by considering the Euclidean distance between a pair of gene embeddings as $d_{i,j}$ in Equation (1) and treating the gene weights as the frequencies $\tilde{x}_{1i}$ and $\tilde{x}_{2j}$ in Equations (2) and (3). This step involves a computational complexity of $O(r^3 \log r)$. Subsequently, the WMD distance between each pair of cells is computed by utilizing the WMD distance between a pair of cell functions, calculated in the first step, as $d_{i,j}$ in Equation (1) and treating the function weights

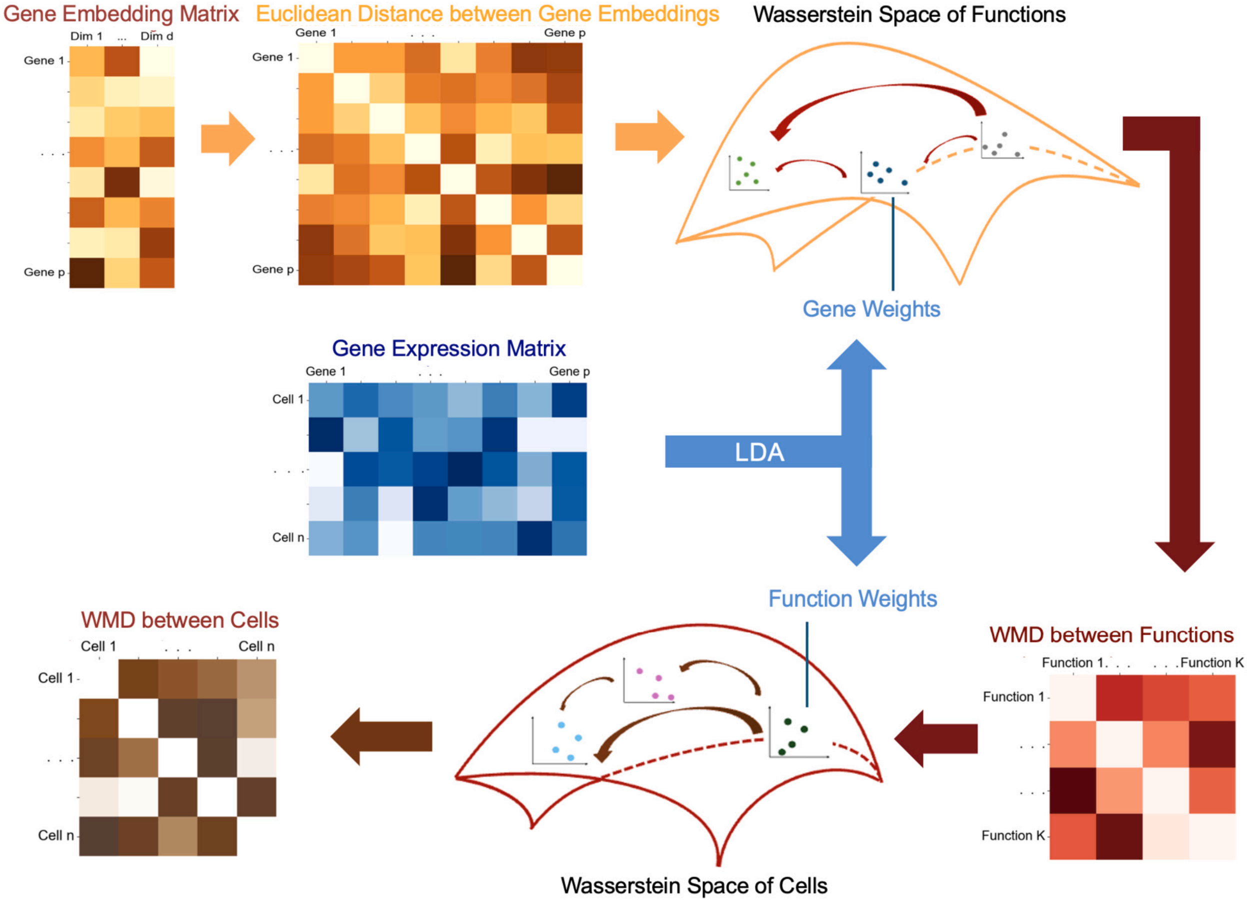

Fig. 3. The Workflow of the scHOTT Algorithm. In the middle row, the LDA algorithm provides the function weights and gene weights. In the top row, from left to right, the Euclidean distances between gene embeddings are first computed and then combined with the gene weights to calculate the WMDs between functions. In the bottom row, from right to left, the WMDs between functions are combined with the function weights to determine the WMDs between cells.

Table 1
Summary of the six datasets used in this study.

Dataset	# Cells	# Cell Types	Tissue	References
Aorta	9,625	12	Human ascending aorta	Li et al. [25]
Artery	14,352	9	Human carotid artery	Alsaigh et al. [2]
Bones	9,380	7	Human knee	Chou et al. [14]
Myeloid	13,178	21	Human cancers	Cheng et al. [13]
Pancreas	14,767	15	Human pancreas	Tran et al. [35]
Multiple Sclerosis	21,312	18	Human Brain	Schirmer et al. [33]

as the frequencies $\tilde{x}_{1i}$ and $\tilde{x}_{2j}$ in Equations (2) and (3). This phase has a computational complexity of $O(K^3 \log K)$.

Consequently, the total computational complexity of computing WMD between all cell pairs is $O\left(\frac{n(n-1)}{2}K^3 \log K\right)+O\left(\frac{K(K-1)}{2}r^3 \log r\right)\approx O\left(\frac{n(n-1)}{2}K^3 \log K\right)$, since $n \gg K$ while K is on the same scale as r. This complexity is significantly lower than the $O\left(\frac{n(n-1)}{2}p^3 \log p\right)$ complexity of the traditional method. For our computations, we set $K = 90$ and $r = 50$. In our real-dataset analysis involving 3000 highly variable genes (i.e., $p = 3000$), the computational load is approximately $(p/K)^3 \log p/\log K > 6 \times 10^4$ times faster. Before computing the pairwise WMD between cells, the computation of LDA takes some time, which we have found to be similar in scale to the computation of pairwise WMD between cells. Thus, the overall computational time using this strategy is markedly shorter than the traditional method of WMD computation.

3. Results

3.1. Data collection and pre-processing

The GenePT paper [12] evaluated its performance using six real datasets, all of which are utilized in this study. A brief summary of these datasets is provided in Table 1, with more detailed information available in Section 4 of the Supplementary Material.

The normalized and logarithmic-transformed data for the Aorta [25], Myeloid [13], and Multiple Sclerosis [33] datasets were directly downloaded from the GenePT study. The Artery [2], Bones [14], and Pancreas [35] datasets were obtained from the original publications and subsequently normalized and logarithmic-transformed by us. For all datasets, 3000 highly variable genes were used for the analysis.

Gene embeddings were obtained by embedding the gene names using the "text-embedding-3-large" model offered by OpenAI [29]. The length of each embedding is 3,072.

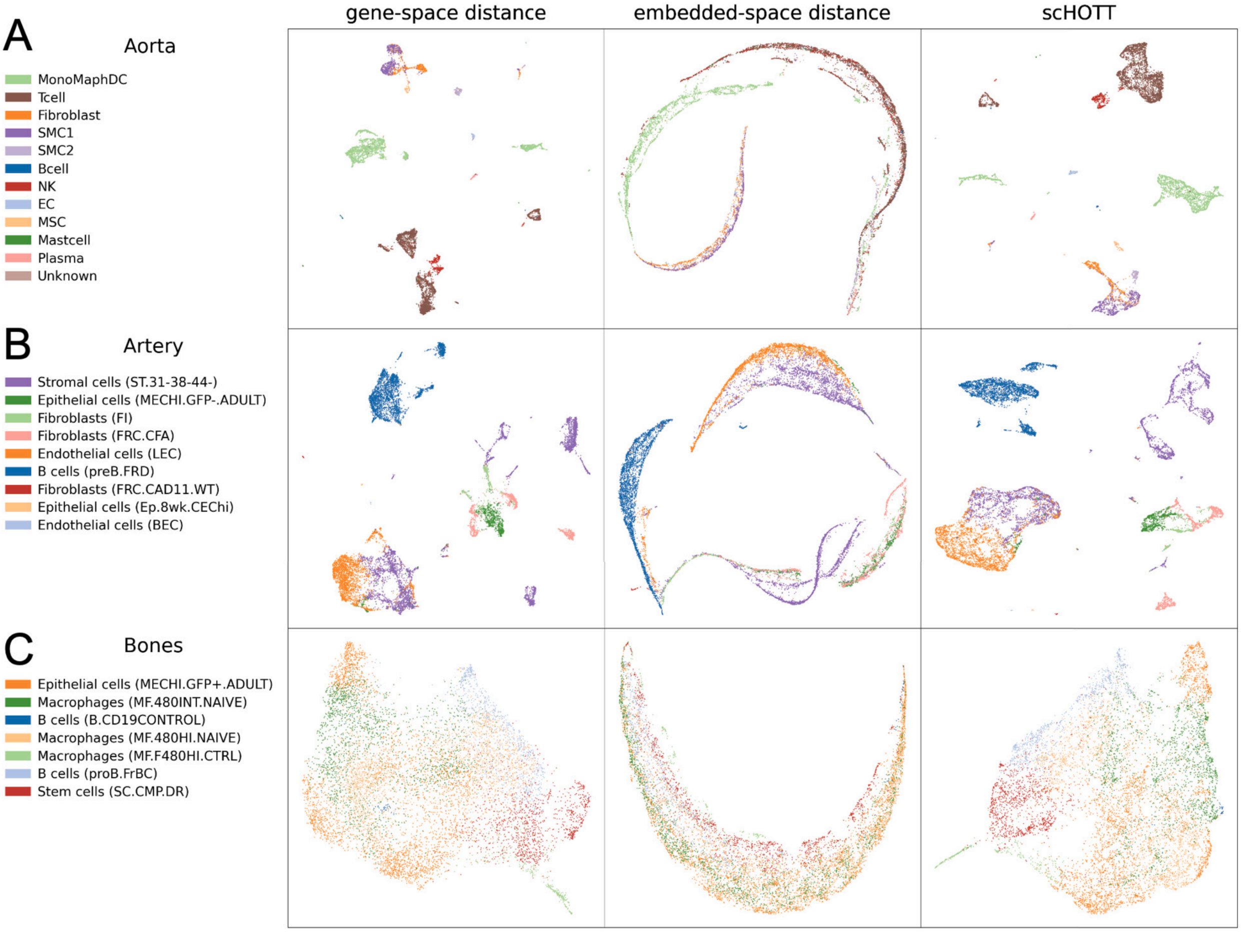

Fig. 4. 2D UMAP visualizations using three different distance metrics (from left to right: gene-space distance, embedded-space distance, and scHOTT) in three different datasets (from top to bottom: Aorta, Artery, and Bones). In each plot, a cell is represented by a point colored according to cell type. Ideally, cells of the same color should cluster together.

We computed our scHOTT distance using the default settings: $K = 90$ functions and $r = 50$ genes per function. The computation time varied from 50 minutes to 3.5 hours, depending on the number of cells in each dataset, on an Apple MacBook Pro 16-inch with an M2 Max chip featuring a 12-core CPU.

3.2. Evaluation based on UMAP visualization

We first assessed the performance of scHOTT on the UMAP [26] plot and compared it with gene-space distance and embedded-space distance. The UMAP plot is a highly popular visualization tool in scRNA-seq data, known for its ability to clearly display cell clusters [5]. Ideally, different cell types should form their own clusters. UMAP allows users to input their own definitions of distance. Therefore, by examining how well different cell types are clustered together and separated from other cell types in a UMAP plot, we can evaluate the quality of the distance definition.

We ran UMAP using three different distances: scHOTT, gene-space distance, and embedded-space distance. Fig. 4 shows the UMAP plots for the Aorta [25], Artery [2], and Bones [14] datasets, and Fig. 5 shows the UMAP plots for the Myeloid [13], Pancreas [35], and Multiple Sclerosis [33] datasets. In each UMAP, the points represent the cells and are colored by their cell types.

These plots reveal interesting patterns. First, cells of the same type typically form partial ring shapes in all UMAP plots based on embedded-space distance (i.e., all plots in the middle column). This is undesired, as cell types do not form tight clusters, indicating difficulty in using a clustering algorithm to separate different cell types.

Second, both the scHOTT distance (i.e., plots in the right column) and gene-space distance (i.e., plots in the left column) yield relatively tight clusters, each roughly consisting of cells from the same cell type. However, there are instances in every dataset where cell clusters based on scHOTT distance are better separated than those based on gene-space distance. Below we provide several examples.

In the Aorta dataset, the UMAP generated using gene-space distance shows MSCs (light orange) and fibroblast cells (dark orange) intermixing with indistinct boundaries. Similarly, the boundaries between T cells (dark brown) and NK cells (red) are unclear. In contrast, the separations between these cell types are clearer in the results obtained using the scHOTT distance.

In the Artery dataset, the UMAP reveals partial intermixing between endothelial cells (LEC) (dark orange) and stromal cells (purple), as well as between epithelial cells (MECHi.GFP-.ADULT) (dark green) and fibroblasts (FRC.CFA) (pink) when using the gene-space distance. The separations are more distinct when using the scHOTT distance.

The Bones dataset presents a considerable challenge; however, the scHOTT distance yields better results. The UMAP generated using gene-space distance shows stem cells (SC.CMP.DR) (red) intermixed with B cells (proB.FrBC) (light blue), and three types of macrophages (MF.480HI.NAIVE, MF.480INT.NAIVE, MF.F480HI.CTRL) also intermingle. The scHOTT distance significantly reduces these overlaps.

In the Myeloid dataset, the scHOTT distance significantly reduces the overlap between two types of macrophages: Macro_LYVE1 (red) and Macro_C1QC (dark blue).

In the Pancreas dataset, both distance measures effectively separate most cell types, but the scHOTT distance results in slightly tighter clus-

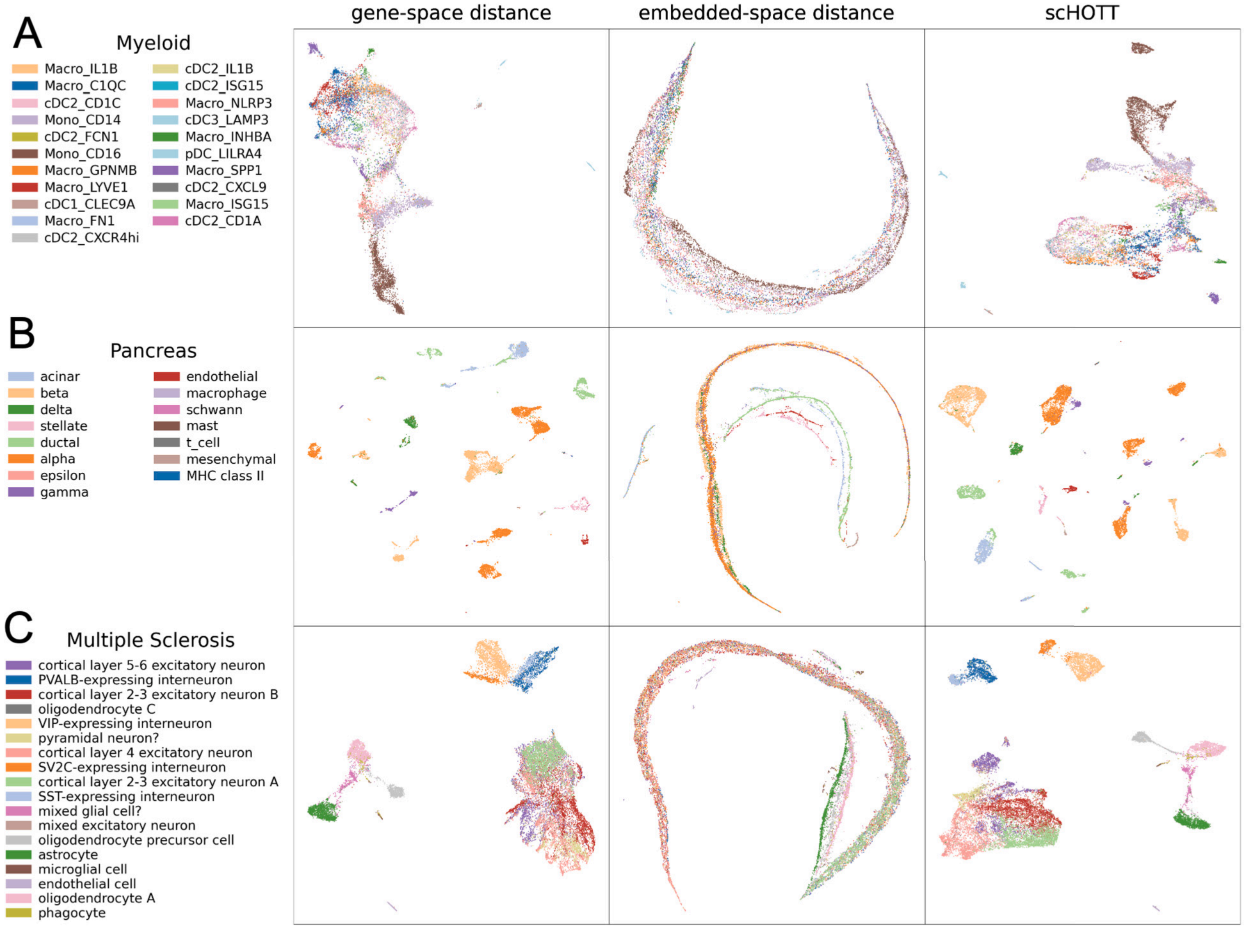

Fig. 5. 2D UMAP visualizations using three different distance metrics (from left to right: gene-space distance, embedded-space distance, and scHOTT) in three different datasets (from top to bottom: Myeloid, Pancreas, and Multiple Sclerosis). In each plot, a cell is represented by a point colored according to cell type. Ideally, cells of the same color should cluster together.

ters for alpha cells (dark orange) and beta cells (light orange) compared to the gene-space distance.

In the Multiple Sclerosis dataset, the scHOTT distance significantly enhances the separation of cell types within the large cluster at the bottom of the plot, particularly between cortical layer 2-3 excitatory neuron A (light green) and cortical layer 4 excitatory neuron (light red). Additionally, it markedly improves the distinction among the four cell types at the top of the plot.

3.3. *Evaluation based on kNN classification*

Although UMAP provides rich details for comparison, it does not provide a statistic for an overall impression or conclusion of the comparison. Therefore, in this section, we use the performance of a k-nearest neighbor (kNN) classifier, a supervised classification method, to evaluate the performance of different distances in a more concise and objective manner.

We chose kNN as the classifier because its performance directly relies on the quality of the distance metric used to identify neighbors [31]. The kNN classifier assigns a test sample to the majority class of its nearest neighbors in the training data. A good distance metric assigns small values to those “true” neighbors, which are expected to be of the same class label (i.e., cell type in our problem), making the kNN classifier more likely to give the correct classification. Thus, the performance of kNN provides a good measure of the quality of the distance.

Following the GenePT paper [12], we summarize the performance of the kNN classifier using four summary statistics: accuracy, precision, recall, and F1-score. Fig. 6 presents these metrics for the six real datasets. We find that overall, the four summary statistics are highly consistent with each other for each dataset. Therefore, we select one of them, the F1-score, for further discussion.

scHOTT outperforms the other two methods in all six datasets. Compared to the gene-space distance and the embedded-space distance, the average improvement of scHOTT across the six datasets is 35.35% and 58.37%, respectively. Its advantage is especially substantial in two datasets: Bones (F1-score 0.82 vs. 0.52 vs. 0.48) and Multiple Sclerosis (0.66 vs. 0.21 vs. 0.31). These statistics underscore the effectiveness and robustness of scHOTT, highlighting its substantial advantage over the other two methods in measuring cell-to-cell distances in scRNA-seq data.

4. Discussion

The application of LLMs and ChatGPT to gene-expression data has introduced new opportunities and challenges. The gene-embedding matrix adds valuable information to the traditionally obtained gene-expression matrix, yet how to effectively utilize this additional information remains unclear. This paper addresses a fundamental aspect of this problem: how to combine information from both matrices to better define cell-to-cell distance in single-cell RNA-seq data. We have discovered the intrinsic similarity of this problem to the document retrieval problem and identified an algorithm called scHOTT that can be computed in a reasonable amount of time. Across all six real datasets we examined, scHOTT demonstrated superior performance in terms of separating different cell types in the UMAP plot and achieving better classification accuracy in kNN classification.

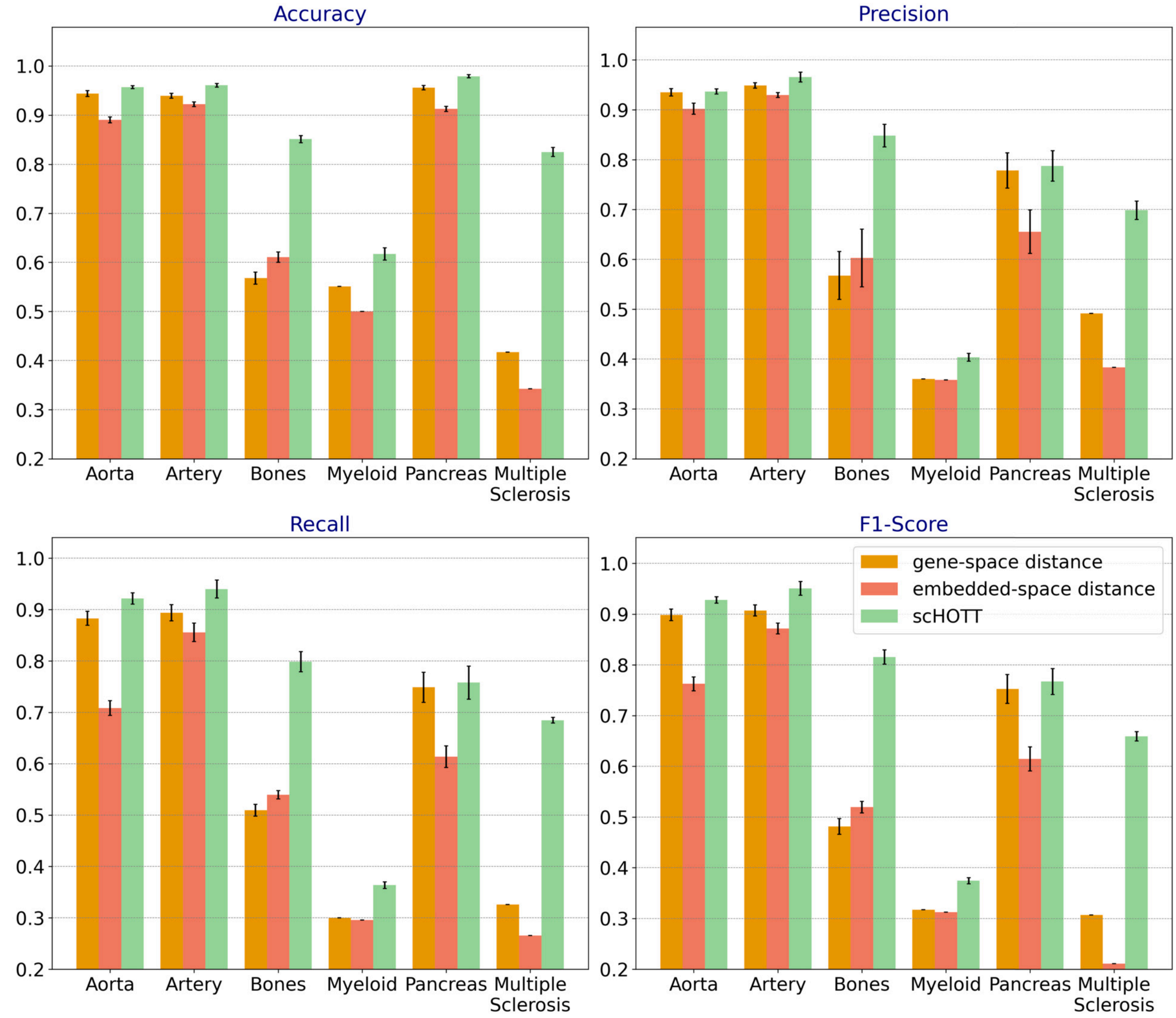

Fig. 6. Performance of three different distances (gene-space distance, embedded-space distance, and scHOTT) across six real datasets. The performance is measured using Accuracy, Precision, Recall, and F1-Score, as shown in different sub-figures. Error bars represent the standard deviation based on results from 10 random seeds.

The scHOTT distance has two parameters: the number of functions K and the number of genes per function r. For all real datasets we considered, we used the default values: $K = 90$ and $r = 50$. Fig. 7 shows the performance of scHOTT on the Multiple Sclerosis dataset under different values of K and r. It appears that scHOTT's performance is quite stable across a wide range of parameter values.

In the scHOTT algorithm, each gene may contribute to multiple cell functions. An alternative approach could involve dividing genes into non-overlapping groups, though this strategy would fall outside the LDA framework. To explore this, we divided the 3000 genes in the Aorta dataset into 100 random, non-overlapping groups, each containing 30 genes, and calculated the cell-to-cell WMD for each group. The overall cell-to-cell distances were then obtained by summing up the 100 WMDs from the groups. We refer to this measure as the "random-split distance." Fig. S2 displays the accuracy, precision, recall, and F1-score for both the random-split distance and our scHOTT distance. It is evident that the random-split distance significantly underperforms compared to scHOTT. Furthermore, although calculating the random-split distance is considerably faster than computing WMD traditionally, it is still more than ten times slower than scHOTT.

With the intention of focusing on the use of the gene-embedding matrix, we deliberately narrowed the scope of our paper, concentrating our comparisons on distances defined in the original space (which ignores the gene-embedding matrix), the embedded space (the intuitive and previously sole method utilizing the gene-embedding matrix), and our combination of both matrices. Consequently, we did not conduct a comprehensive review of various advanced distance measures in the original space that do not use gene embeddings, such as those described in [28,15,20,32]. In the Supplementary Material, we briefly explored the use of PHATE [28], an advanced distance measure in the original space, for distance computation. We found that although the PHATE distance measure performs significantly better than those in both the embedded-space and gene-space, it is still clearly outperformed by our scHOTT distance.

Our performance comparison is based on UMAP plots and numerical measures derived from kNN classification. One could also consider other measures, such as Silhouette score and Adjusted Rand Index (ARI), although both have clear limitations. The Silhouette score performs poorly in evaluating non-globular-shaped clusters. ARI compares the true cell type labels with the inferred cell type labels, and thus, it depends on the choice of clustering methods. Fig. S5 presents the Silhouette score and ARI for the six datasets. It is evident that our scHOTT distance continues to significantly outperform the other distances when evaluated using these metrics.

Finally, while the scHOTT algorithm is significantly faster than the original WMD algorithm, calculating all pairwise distances between

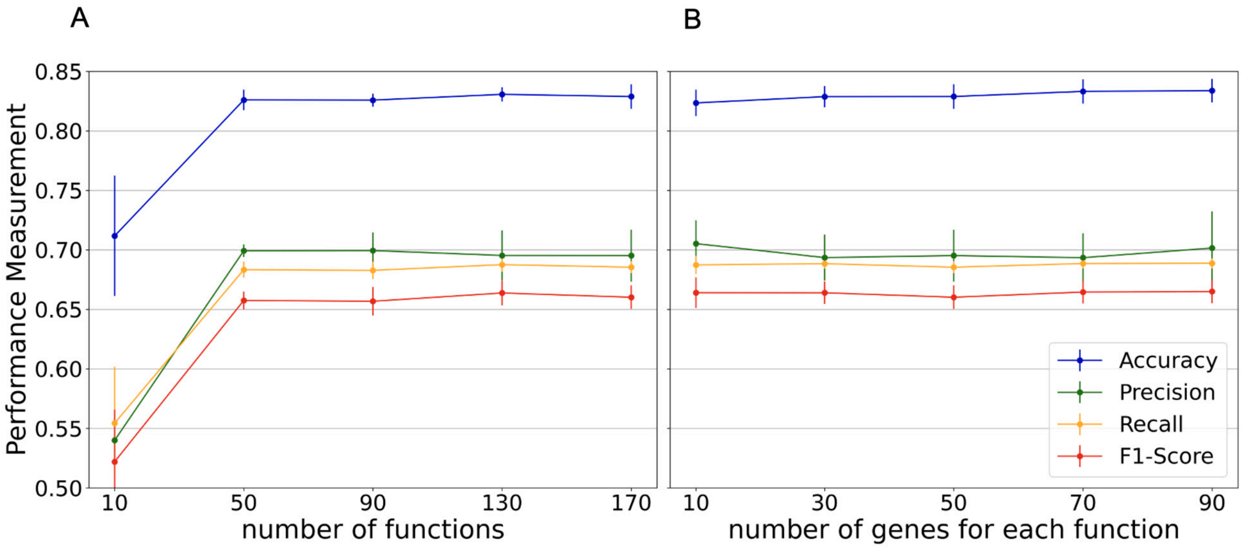

Fig. 7. Changes in scHOTT's performance on the Multiple Sclerosis dataset with respect to **(A)** K, the number of functions, and **(B)** r, the number of genes for each function. Lines of different colors represent Accuracy, Precision, Recall, and F1-Score, respectively. Error bars represent the standard deviation based on results from 10 random seeds.

cells still requires 3.5 hours for datasets containing approximately 20,000 cells. Developing an even faster algorithm remains an area of ongoing interest.

The Python code for the scHOTT algorithm is available at https://github.com/Fangfang-Guo/scHOTT.

CRediT authorship contribution statement

Fangfang Guo: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation. **Dailin Gan:** Writing – review & editing, Writing – original draft, Formal analysis, Data curation. **Jun Li:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors have no conflicts of interest to declare.

Acknowledgement

This work is supported by the National Institutes of Health (R01CA252878 and R01CA280097 to J.L.) and the DOD BCRP Breakthrough Award, Level 2 (W81XWH2110432 to J.L.).

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csbj.2024.10.044.

References

[1] Adossa Nigatu A, Rytkönen Kalle T, Elo Laura L. Dirichlet process mixture models for single-cell rna-seq clustering. Open Biol 2022;11(4):bio059001.

[2] Alsaigh Tom, Evans Doug, Frankel David, Torkamani Ali. Decoding the transcriptome of calcified atherosclerotic plaque at single-cell resolution. Commun Biol 2022;5(1):1084.

[3] Atasu Kubilay, Mittelholzer Thomas. Linear-complexity data-parallel Earth mover's distance approximations. In: International conference on machine learning. PMLR; 2019. p. 364–73.

[4] Bacher Rhonda, Kendziorski Christina. Design and computational analysis of single-cell rna-sequencing experiments. Genome Biol 2016;17:1–14.

[5] Becht Etienne, McInnes Leland, Healy John, Dutertre Charles-Antoine, Kwok Immanuel WH, Ng Lai Guan, et al. Dimensionality reduction for visualizing single-cell data using umap. Nat Biotechnol 2019;37(1):38–44.

[6] Beyer Kevin, Goldstein Jonathan, Ramakrishnan Raghu, Shaft Uri. When is "nearest neighbor" meaningful? In: Database theory—ICDT'99: 7th international conference Jerusalem, Israel, January 10–12, 1999 proceedings 7. Springer; 1999. p. 217–35.

[7] Blei David M, Ng Andrew Y, Jordan Michael I. Latent Dirichlet allocation. J Mach Learn Res Jan. 2003;3:993–1022.

[8] Brokos Georgios-Ioannis, Malakasiotis Prodromos, Androutsopoulos Ion. Using centroids of word embeddings and word mover's distance for biomedical document retrieval in question answering. arXiv preprint. arXiv:1608.03905, 2016.

[9] Brown Garth R, Hem Vichet, Katz Kenneth S, Ovetsky Michael, Wallin Craig, Ermolaeva Olga, et al. Gene: a gene-centered information resource at ncbi. Nucleic Acids Res 2015;43(D1):D36–42.

[10] Brown Tom, Mann Benjamin, Ryder Nick, Subbiah Melanie, Kaplan Jared D, Dhariwal Prafulla, et al. Language models are few-shot learners. Adv Neural Inf Process Syst 2020;33:1877–901.

[11] Castells Pablo, Fernandez Miriam, Vallet David. An adaptation of the vector-space model for ontology-based information retrieval. IEEE Trans Knowl Data Eng 2006;19(2):261–72.

[12] Chen, Yiqun, Zou, James. Genept: a simple but effective foundation model for genes and cells built from chatgpt, bioRxiv, 2023.

[13] Cheng Sijin, Li Ziyi, Gao Ranran, Xing Baocai, Gao Yunong, Yang Yu, et al. A pan-cancer single-cell transcriptional atlas of tumor infiltrating myeloid cells. Cell 2021;184(3):792–809.

[14] Chou Ching-Heng, Jain Vaibhav, Gibson Jason, Attarian David E, Haraden Collin A, Yohn Christopher B, et al. Synovial cell cross-talk with cartilage plays a major role in the pathogenesis of osteoarthritis. Sci Rep 2020;10(1):10868.

[15] Coifman Ronald R, Lafon Stéphane. Diffusion maps. Appl Comput Harmon Anal 2006;21(1):5–30.

[16] Cui Haotian, Wang Chloe, Maan Hassaan, Pang Kuan, Luo Fengning, Duan Nan, et al. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. Nat Methods 2024:1–11.

[17] Deek Rebecca A, Li Hongzhe. A zero-inflated latent Dirichlet allocation model for microbiome studies. Front Genet 2021;11:602594.

[18] Hao Minsheng, Gong Jing, Zeng Xin, Liu Chiming, Guo Yucheng, Cheng Xingyi, et al. Large-scale foundation model on single-cell transcriptomics. Nat Methods 2024:1–11.

[19] Hastie Trevor, Tibshirani Robert, Friedman Jerome H, Friedman Jerome H. The elements of statistical learning: data mining, inference, and prediction, vol. 2. Springer; 2009.

[20] Huguet Guillaume, Tong Alexander, De Brouwer Edward, Zhang Yanlei, Wolf Guy, Adelstein Ian, et al. A heat diffusion perspective on geodesic preserving dimensionality reduction. Adv Neural Inf Process Syst 2024;36.

[21] Jovic Dragomirka, Liang Xue, Zeng Hua, Lin Lin, Xu Fengping, Luo Yonglun. Single-cell rna sequencing technologies and applications: a brief overview. Clin Transl Med 2022;12(3):e694.

[22] Kotliar Dylan, Veres Adrian, Nagy M Aurel, Tabrizi Shervin, Hodis Eran, Melton Douglas A, et al. Identifying gene expression programs of cell-type identity and cellular activity with single-cell rna-seq. eLife 2019;8:e43803.

[23] Kusner Matt, Sun Yu, Kolkin Nicholas, Weinberger Kilian. From word embeddings to document distances. In: International conference on machine learning. PMLR; 2015. p. 957–66.

[24] Li Changchun, Ouyang Jihong, Li Ximing. Classifying extremely short texts by exploiting semantic centroids in word mover's distance space. In: The world wide web conference; 2019. p. 939–49.

[25] Li Yanming, Ren Pingping, Dawson Ashley, Vasquez Hernan G, Ageedi Waleed, Zhang Chen, et al. Single-cell transcriptome analysis reveals dynamic cell populations and differential gene expression patterns in control and aneurysmal human aortic tissue. Circulation 2020;142(14):1374–88.

[26] McInnes Leland, Healy John, Melville James. Umap: uniform manifold approximation and projection for dimension reduction. arXiv preprint. arXiv:1802.03426, 2018.

[27] Mikolov Tomas, Chen Kai, Corrado Greg, Dean Jeffrey. Efficient estimation of word representations in vector space. arXiv preprint. arXiv:1301.3781, 2013.

[28] Moon Kevin R, Van Dijk David, Wang Zheng, Gigante Scott, Burkhardt Daniel B, Chen William S, et al. Visualizing structure and transitions in high-dimensional biological data. Nat Biotechnol 2019;37(12):1482–92.

[29] OpenAI. New embedding models and api updates. https://openai.com/blog/new-embedding-models-and-api-updates, 2024. [Accessed 8 August 2024].

[30] Pan Sinno Jialin, Yang Qiang. A survey on transfer learning. IEEE Trans Knowl Data Eng 2009;22(10):1345–59.

[31] Pele Ofir, Werman Michael. Fast and robust Earth mover's distances. In: 2009 IEEE 12th international conference on computer vision. IEEE; 2009. p. 460–7.

[32] Rosen, Yanay, Roohani, Yusuf, Agrawal, Ayush, Samotorcan, Leon, Tabula Sapiens Consortium, Quake, Stephen R, Leskovec, Jure. Universal cell embeddings: a foundation model for cell biology, bioRxiv, 2023, pp. 2023–11.

[33] Schirmer Lucas, Velmeshev Dmitry, Holmqvist Staffan, Kaufmann Max, Werneburg Sebastian, Jung Diane, et al. Neuronal vulnerability and multilineage diversity in multiple sclerosis. Nature 2019;573(7772):75–82.

[34] Theodoris Christina V, Xiao Ling, Chopra Anant, Chaffin Mark D, Al Sayed Zeina R, Hill Matthew C, et al. Transfer learning enables predictions in network biology. Nature 2023;618(7965):616–24.

[35] Tran Hoa Thi Nhu, Ang Kok Siong, Chevrier Marion, Zhang Xiaomeng, Yee Shin Lee Nicole, Goh Michelle, et al. A benchmark of batch-effect correction methods for single-cell rna sequencing data. Genome Biol 2020;21:1–32.

[36] Vaswani Ashish, Shazeer Noam, Parmar Niki, Uszkoreit Jakob, Jones Llion, Gomez Aidan N, et al. Attention is all you need. Adv Neural Inf Process Syst 2017;30.

[37] Wu Lingfei, Yen Ian EH, Xu Kun, Xu Fangli, Balakrishnan Avinash, Chen Pin-Yu, et al. Word mover's embedding: from word2vec to document embedding. arXiv preprint. arXiv:1811.01713, 2018.

[38] Yang Fan, Wang Wenchuan, Wang Fang, Fang Yuan, Tang Duyu, Huang Junzhou, et al. scbert as a large-scale pretrained deep language model for cell type annotation of single-cell rna-seq data. Nat Mach Intell 2022;4(10):852–66.

[39] Yang Liu, Jin Rong. Distance metric learning: a comprehensive survey. Michigan State Universiy. 2006;2(2):4.

[40] Yang Qi, Xu Zhaochun, Zhou Wenyang, Wang Pingping, Jiang Qinghua, Juan Liran. An interpretable single-cell rna sequencing data clustering method based on latent Dirichlet allocation. Brief Bioinform 2023;24(4):bbad199.

[41] Yurochkin Mikhail, Claici Sebastian, Chien Edward, Mirzazadeh Farzaneh, Solomon Justin M. Hierarchical optimal transport for document representation. Adv Neural Inf Process Syst 2019;32.