# Evaluating the Robustness of Dimensionality Reduction Algorithms Using Sparse Single-Cell ATAC-seq Data

Justin Currie and Morgan Lo

cs2952Q Final Project, Fall 2024

## 1 Abstract

Dimensionality reduction is a critical step in the analysis of single-cell ATAC-seq (scATAC-seq) data due to its inherently sparse and high-dimensional nature. This study evaluates the robustness of three popular dimensionality reduction methods — **Latent Semantic Indexing (LSI)**, **cisTopic**, and **scBasset** — in handling increasing levels of sparsity in scATAC-seq data. Using a well-annotated single-cell multiomic atlas of human bone marrow comprising 69,248 cells from 22 distinct cell types, we create progressively sparser datasets by randomly setting 10%, 20%, 40%, 60%, 80%, and 95% of the non-zero entries in the cell-by-region matrix to zero and introduce synthetic batch effects to assess the performance of each method.

To evaluate the quality of low-dimensional embeddings, we apply **K-means clustering (K=22)** to identify cell-type clusters and use metrics like **Adjusted Rand Index (ARI)**, **Adjusted Mutual Information (AMI)**, and **Homogeneity Score (HS)** to quantify clustering quality. We visualize embeddings using **UMAP** to assess separation of cell types at each sparsity level. Additionally, we infer developmental trajectories using the **VIA trajectory inference algorithm** to evaluate the preservation of biological trajectories in low-dimensional space.

Results reveal that **cisTopic consistently produces the highest-quality embeddings** across all sparsity levels, maintaining better clustering performance compared to LSI and scBasset. While scBasset maintains stable performance at low sparsity, its robustness declines as sparsity increases. LSI performs the worst under high sparsity and struggles to cluster cells according to cell type. Finally, we find that cisTopic embeddings are the most effective at preserving developmental trajectories.

These findings suggest that **cisTopic is the most robust and interpretable method** for reducing dimensionality in sparse scATAC-seq datasets, making it a preferred choice for analyzing datasets with significant sparsity. This study provides a framework for systematically benchmarking dimensionality reduction techniques in scATAC-seq, offering valuable guidance for researchers in the field.

## 2 Introduction

A vast array of proteins bind to the genomes of multicellular organisms. DNA coils around protein complexes called nucleosomes, forming a medium called chromatin. Chromatin can fold into densely packed structures that allow the genome to be compressed within the nucleus of a cell. However, this densely packed chromatin is not accessible to other proteins such as Transcription Factors (TFs), which must be able to bind to DNA to control gene expression. Consequently, short regions of the genome known as *cis*-regulatory elements (CREs) are nucleosome free, making them accessible to regulatory proteins. CREs include regions of DNA just before the start of a gene, called promoters, as well as

regions that interact with promoters through 3D chromatin folding, known as enhancers and repressors. Mechanisms that control chromatin accessibility and the proteins that bind to accessible regions work together to orchestrate precise and cell-type specific regulation of gene expression (Figure 1).

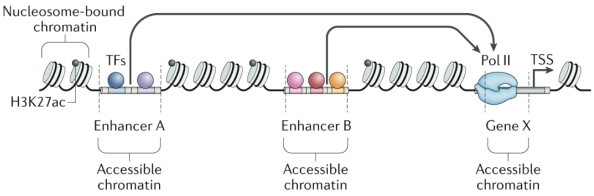


**Figure 1:** TFs bind to accessible regions such as enhacers (left and middle regions) and gene promoters (right region). Interactions between enhancer-bound TFs and gene promoters drive expression changes. Figure from Minnoye, *et-al.*, 2021.

Each cell-type has a unique chromatin accessibility profile that corresponds to its gene expression profile. Allowing accessibility at a specific pattern of promoters, enhancers, and repressors allows for the expression of cell-type specific genes. During development, cells undergo a series of changes as they grow and divide, shifting from progenitor cell-types to differentiated cells with specific functions. Throughout this process, chromatin accessibility is extremely dynamic, allowing the expression of stem-cell specific genes to be repressed while increasing the expression of cell-type specific genes. Therefore, understanding how chromatin accessibility varies over changes in cell-state is fundamental to our understanding of development. Furthermore, chromatin accessibility is frequently misregulated in disease states. For instance, cancer cells frequently have aberrant chromatin accessibility profiles, leading to dangerous gene expression programs that drive uncontrolled cell growth.

Our ability to precisely measure regions of accessible chromatin expanded considerably with the development of Assay for Transposase Accessible Chromatin with sequencing (ATAC-seq) in 2013. The ATAC-seq protocol uses an enzyme called Transposase, which simultaneously cuts DNA at accessible sites and concatenates sequencing adaptors to the ends of the resulting fragments. These fragments can then be sequenced and aligned back to a reference genome. The density of alignment fragments across the genome creates a chromatin accessibility profile that can be used to identify CREs (Figure 2).

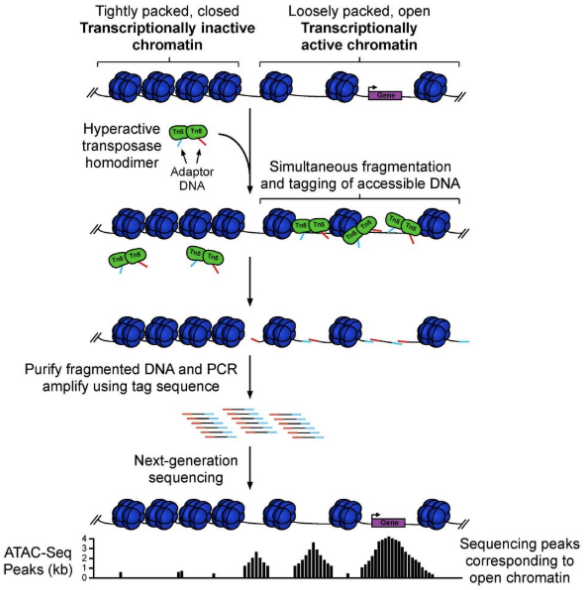


**Figure 2:** Schematic of the ATAC-seq workflow. Figure from AtacSeqWorkshop.

The recent advancement of single-cell sequencing technology has allowed us to begin profiling chromatin accessibility at single-cell resolution, a technique known as single-cell ATAC-seq (scATAC-seq). Bulk ATAC-seq data (which aggregates accessible DNA fragments across an entire tissue sample) is diffi-

cult to analyze due to high levels of cell-type heterogeneity within tissue samples. Generating scATAC-seq data allows us to identify the variety of cell-types existing in a given sample and analyze cell-type specific accessibility patterns. However, this dramatic increase in resolution comes at the cost of data quality. During single-cell sequencing preparation, many accessible DNA-fragments are lost, making ATAC-seq data extremely sparse. It is estimated that just 1-10% of accessible regions are detected per cell, resulting in a region $\times$ cell matrix that may have $\sim$3% nonzero entries (1). The data is also very high-dimensional, as over 100,000 accessible regions are often identified over all cell-types. Therefore, it is of great interest to obtain information-rich, low-dimensional embeddings of profiled cells from scATAC-seq experiments. Ideally, these latent embeddings should reflect cell-type identity, allowing us to identify separate cell-types in tissue samples and infer developmental trajectories.

Here, we assess the quality of the latent embeddings produced by three dimensionality-reduction methods for scATAC-seq, data: Latent Semantic Indexing (LSI), CisTopic, and scBasset. Using a well-annotated scATAC-seq atlas of the human bone marrow (2), we evaluate each method in terms of its ability to create embeddings that result in accurate cell-type clusterings and inferred developmental trajectories. We consider the robustness of each method to dropout by progressively increasing sparsity in the data. We additionally evaluate the robustness of each method to sequencing-based batch effects.

# 3 Data and Methods

## 3.1 Dataset

A benchmarking study requires a dataset with a well-defined ground truth. Therefore, we chose to use a 69,248 cell multiomic atlas of the human bone marrow (2). This means that both chromatin accessibility (ATAC-seq) and gene expression (RNA-seq) were profiled for all cells in the dataset. While multiomic profiling exacerbates sparsity, it also allows for accurate cell-type annotation through expertly curated analysis of cell-type specific marker genes. The dataset contains cells from 22 annotated cell-types, from hematopoietic stem cells (HSCs) to several kinds differentiated red and white blood cells. Since the bone marrow generates new blood cells at a remarkably high rate, the cell-types and differentiation trajectories represented in this dataset have been extensively studied.

### 3.1.1 Data Preparation

Since all three of our dimensionality-reduction algorithms use a regions $\times$ cells counts matrix as input, we simply downloaded this matrix from the NCBI Gene Expression Omnibus under accession GSE194122. The resulting 116,468 region $\times$ 69,248 cell matrix had already undergone basic quality control, including removal of low-quality cells (cells with very few accessible regions), and suspected doublets (cells that likely include data from two separate cells). As a technical consequence of library preparation, scATAC-seq dropout is an inherently random phenomenon. While some regions may be more accessible than others, this is not reflected in scATAC-seq data, as a maximum of two DNA fragments can be produced from a given region in a given cell (one for each allele). Therefore, the counts matrix consists entirely of 0s, 1s, and 2s. Therefore, we mimicked increasing levels of dropout by randomly selecting nonzero elements in our counts matrix and setting them to 0. We created a series of counts matrices with 10%, 20%, 40%, 60%, 80%, and 95% of the nonzero entries in the original matrix removed. Another common issue with single-cell sequencing data are batch effects. A large single-cell sample is typically sequenced in several batches. Due to slight differences in the execution of the library preparation protocols for each batch, separate batches may have different sequencing depths, or different degrees of dropout. While the original counts matrix had been batch-corrected, we mimicked a simple, yet dramatic batch effect by introducing a 40% dropout rate in a random subset of 50% of the cells, creating two separate batches with different sequencing depths.

## 3.2 Dimensionality Reduction Methods

Dimensionality reduction is a crucial step in the analysis of single-cell ATAC-seq data due to its inherently high dimensionality and sparsity. Reducing the dimensionality of the region-by-cell matrix allows for more efficient storage, visualization, and clustering of cells into biologically meaningful groups. In this study, we evaluate the performance of three widely used dimensionality reduction methods: **Latent Semantic Indexing (LSI)**, **cisTopic**, and **scBasset**.

### 3.2.1 Latent Semantic Indexing (LSI)

Latent Semantic Indexing (LSI) is a matrix factorization technique that identifies patterns in sparse, high-dimensional data by capturing latent structure through **Singular Value Decomposition (SVD)**. Originally developed for text analysis, LSI has been adapted for use in single-cell ATAC-seq analysis, where the goal is to reduce the high-dimensional region-by-cell matrix into a lower-dimensional space.

The LSI process begins by applying Term Frequency-Inverse Document Frequency (TF-IDF) normalization to the counts matrix to account for the variability in region accessibility across cells. Then, SVD is applied to decompose the matrix into three components:

$$X = U\Sigma V^T$$

where $X$ is the original matrix, $U$ and $V$ are orthogonal matrices, and $\Sigma$ is a diagonal matrix containing singular values. The first $k$ components are retained, effectively reducing the dimensionality to a $k$-dimensional space. The resulting low-dimensional embeddings are used for downstream clustering, trajectory inference, and visualization.

While LSI is computationally efficient and fast, it tends to perform poorly on sparse scATAC-seq data due to its sensitivity to noise. As sparsity increases, the first few components of SVD may capture technical artifacts rather than true biological signals, leading to poor clustering performance.

### 3.2.2 Topic Modeling

cisTopic is a topic modeling approach designed to identify regulatory elements in chromatin accessibility data. The method models region-by-cell accessibility as a mixture of **latent topics**, where each topic represents a group of co-accessible genomic regions that are likely to share a regulatory function. cisTopic is based on **Latent Dirichlet Allocation (LDA)**, a probabilistic graphical model widely used in natural language processing.

The cisTopic model assumes that each region belongs to one or more "topics" and that each cell can be represented as a mixture of topics. Specifically, for each region $r$ in cell $c$, the probability of accessibility is given by:

$$P(r \mid c) = \sum_{k=1}^{K} P(r \mid z_k) P(z_k \mid c)$$

where $z_k$ represents a latent topic, and $K$ is the total number of topics. The algorithm infers the probabilities $P(r \mid z_k)$ and $P(z_k \mid c)$ using variational inference.

By clustering cells in the space of these latent topics, cisTopic enables the identification of biologically meaningful groupings of cells and regions. Unlike LSI, cisTopic is more robust to sparsity since it leverages the assumption that chromatin accessibility is driven by a limited number of underlying biological processes. This approach allows for better clustering and trajectory inference under high sparsity conditions, as shown in our results.

#### 3.2.3 scBasset

scBasset is a **deep learning-based approach** that leverages **Convolutional Neural Networks (CNNs)** to model chromatin accessibility patterns at the single-cell level. Unlike traditional matrix factorization or topic modeling approaches, scBasset directly learns a nonlinear mapping from the raw accessibility data to a low-dimensional embedding space. This approach enables scBasset to capture complex, non-linear relationships between genomic regions and cell-type-specific chromatin accessibility.

The scBasset model takes the binary region-by-cell accessibility matrix as input and encodes each region as a **one-hot-encoded DNA sequence**. This sequence is passed through multiple convolutional layers that detect patterns of accessibility, similar to how convolutional layers in image classification models detect features like edges and shapes. The output of the convolutional layers is passed through fully connected layers to produce a **low-dimensional embedding** for each cell. This embedding can then be used for clustering, trajectory inference, and visualization.

Unlike LSI and cisTopic, scBasset explicitly incorporates DNA sequence information into the embedding process, allowing it to capture region-level regulatory information that is missed by purely matrix-based methods. However, scBasset requires substantially more computational resources and longer training times compared to LSI and cisTopic. Its performance is also more sensitive to hyperparameter tuning, such as learning rate and batch size. Despite these challenges, scBasset has been shown to produce high-quality embeddings, particularly for low-sparsity datasets.

### 3.3 Evaluation Workflow

After obtaining 30-dimensional embeddings for each condition (all levels of sparsity and with batch effects) using each dimensionality-reduction technique, we evaluated the quality of the embeddings using the following steps:

1. To find cell-type clusters in each embedding, we ran $k$-means clustering with $k = 22$ to reflect the 22 separate cell-types in the dataset. For evaluating the persistence of batch effects, we set $k = 2$ to reflect the presence of the 2 artificially introduced batches.
2. To evaluate the quality of each clustering, we used three separate metrics: Adjusted Random Index (ARI), Adjusted Mutual Information (AMI) and Homogeneity Score (HS). These metrics take in a proposed labeling and ground-truth labeling and evaluate the degree of similarity between the two. Each method is label-agnositc (the cell-type index need not be the same in the predicted and true labeling) and has a range of [0,1], with higher values indicating higher-quality clusterings.
3. We further applied the Uniform Manifold Approximation and Projection (UMAP) algorithm to reduce each embedding to just two dimensions, allowing us to visualize the degree to which different cell-types and batches cluster in each embedding.
4. We then inferred developmental trajectories from each embedding using VIA, a graph-based trajectory inference algorithm that reconstructs cell lineages based on lazy teleporting random walks integrated with Markov Chain Monte Carlo refinement (6). This allowed us to qualitatively analyze whether well-defined cell-type trajectories were present in each embedding.

## 4 Results

### 4.1 Latent embeddings produced by CisTopic best reflect cell-type identity

To assess the quality of cell-type separation in the low-dimensional embeddings produced by LSI, cisTopic, and scBasset, we applied the clustering and evaluation methods described in Section 3. The dataset includes 69,248 cells from 22 distinct cell types, which serve as ground-truth labels for clustering evaluation.

For each embedding, we ran K-means clustering with $K = 22$ and measured clustering performance using Adjusted Rand Index (ARI), Adjusted Mutual Information (AMI), and Homogeneity Score (HS). UMAP was used to visualize the separation of cells at different sparsity levels, allowing for qualitative assessment of each method's ability to maintain distinct cell-type clusters.

**Clustering Results** We evaluated the clustering performance of LSI, cisTopic, and scBasset embeddings across datasets with increasing sparsity (10%, 20%, 40%, 60%, 80%, and 95%). The clustering metrics (ARI, AMI, and Homogeneity Score) for each sparsity level are shown in Figure 3.

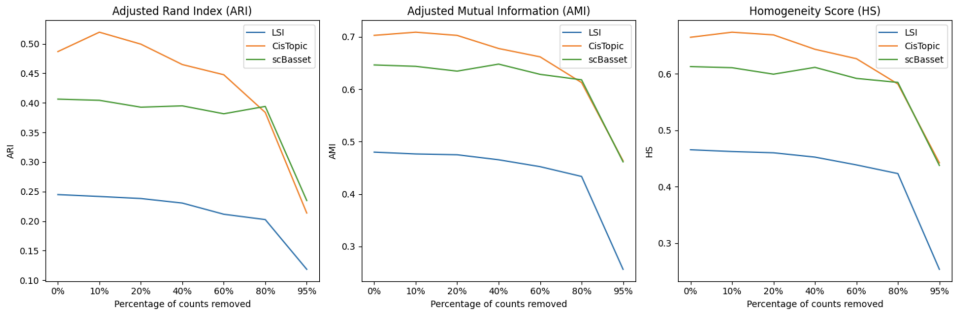


**Figure 3:** Clustering metrics (ARI, AMI, and Homogeneity Score) for LSI, cisTopic, and scBasset embeddings across increasing levels of sparsity. Higher scores indicate better clustering performance.

**Quantitative Analysis** As sparsity increases, clustering performance declines for all methods, but the degree of decline varies significantly across LSI, cisTopic, and scBasset.

- **LSI**: LSI experiences a rapid loss in clustering performance as sparsity increases. At 10% and 20% sparsity, LSI achieves moderate ARI, AMI, and Homogeneity scores, but its performance drops sharply at 80% sparsity and beyond. By 95% sparsity, clustering metrics are near zero, indicating that LSI embeddings fail to separate cell types.
- **cisTopic**: cisTopic exhibits strong robustness to sparsity. ARI, AMI, and Homogeneity scores remain high, and while some decline is observed as sparsity increases, cisTopic still outperforms LSI and performs the same as scBasset. This robustness can be attributed to the topic modeling framework, which captures meaningful latent features that persist despite sparsity.
- **scBasset**: scBasset maintains strong performance up to 80% sparsity. However, its performance drops at 80% and 95% sparsity, where the clustering metrics decline substantially. This suggests that scBasset struggles to extract meaningful features from extremely sparse input data.

**Qualitative Analysis** To provide a qualitative perspective, we visualized UMAP embeddings of the low-dimensional representations at various sparsity levels (0% to 95%) for LSI, cisTopic, and scBasset. The UMAP plots are shown in Figure 4. Each point represents a cell, and colors correspond to ground-truth cell-type annotations. Clusters corresponding to distinct cell types should be well-separated, with minimal overlap.

At low sparsity (10%), all methods produce distinct clusters corresponding to each of the 22 cell types. However, as sparsity increases, qualitative differences between the methods become more pronounced.

- **LSI**: Clusters begin to merge at 40% sparsity, and by 95% sparsity, the embedding is nearly homogeneous, with no clear cluster separation.
- **cisTopic**: cisTopic maintains clear separation of clusters even at 60% sparsity, but merges at 95% sparsity.

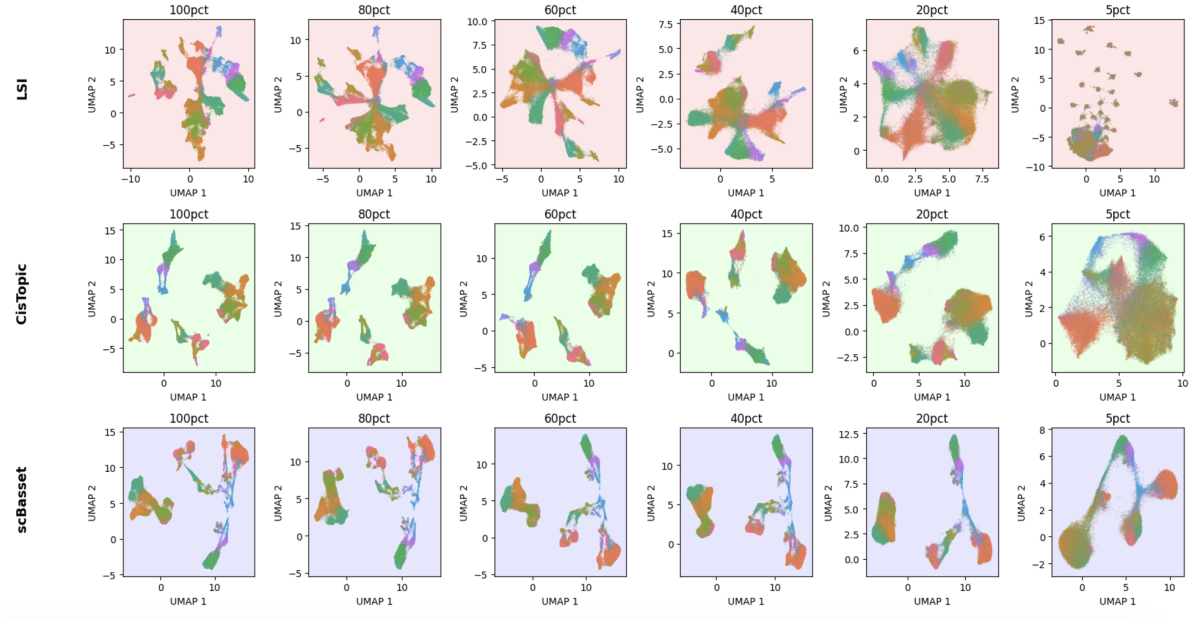


**Figure 4:** UMAP visualizations of embeddings produced by LSI, cisTopic, and scBasset at sparsity levels of 10%, 40%, and 95%. Each point represents a cell, and colors correspond to the annotated cell type. cisTopic maintains clear clusters at high sparsity, while LSI and scBasset embeddings lose structure at higher sparsity levels.

- **scBasset**: At 40% sparsity, scBasset maintains clear clusters, but at 80% and 95% sparsity, clusters merge, leading to a loss of distinct cell-type separation.

**Summary of Clustering Performance** The clustering analysis reveals differences in the robustness of LSI, cisTopic, and scBasset. The key findings are as follows:

- **cisTopic demonstrates the highest robustness to sparsity**, maintaining the clearest cluster separation and higher clustering metrics (ARI, AMI, and Homogeneity) across all sparsity levels.
- **LSI fails to maintain clusters** as sparsity increases. Clustering metrics drop to near zero at 80% and 95% sparsity, indicating that LSI embeddings do not preserve biological structure in highly sparse data.
- **scBasset achieves moderate robustness**, with strong performance at 10% and 20% sparsity. However, its performance declines at 80% and 95% sparsity, indicating that convolutional networks may struggle to generalize when large proportions of accessibility information are lost.

These findings demonstrate that **cisTopic is the most robust and reliable method** for generating embeddings that accurately reflect cell-type identity, even under extreme sparsity conditions. cisTopic's topic modeling framework provides a strong advantage in capturing biologically meaningful latent features, which are less affected by sparsity compared to linear approaches like LSI or convolutional models like scBasset.

## 4.2 LSI embeddings are susceptible to batched data

We next asked whether each of our three methods could create informative embeddings in the presence of a strong batch effect. We separated our data into two batches, one with a 40% reduction in overall counts, simulating differential read depth between sequencing batches (see section 3.1.1). Following the same workflow performed for the dropout analysis, we evaluated the degree of similarity between $k$-means clusterings produced from each embedding with ground-truth batch labels ($k = 2$) and ground-truth cell-type labels ($k = 22$).

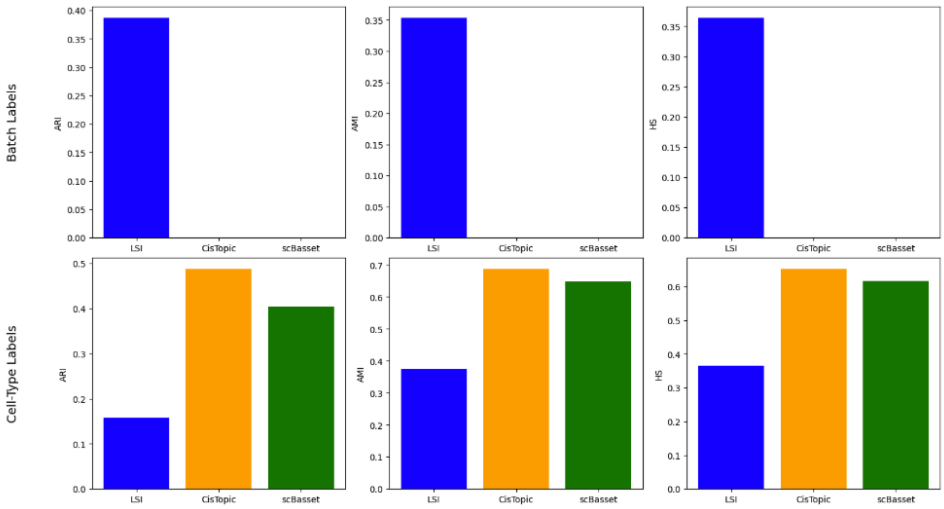


**Figure 5:** Evaluation of clusterings of cell embeddings on ground-truth batch labels and ground-truth cell-type labels. High metric scores on cell-type labels and low scores on batch labels indicate robustness to batch effects.

Figure 5 demonstrates that the embeddings produced by LSI are heavily impacted by the batch effect. The clusterings produced from the LSI embedding align much better with the batch labels than they do with the cell-type labels, indicating that the separate batches are well-defined in the embedding. Conversely, the embeddings produced by CisTopic and scBasset do not reflect a large division between batches and maintain high separation by cell-type. The performance of the clusterings generated from CisTopic and scBasset on the cell-type labels are comparable to these methods' performance on the unbatched data, while LSI performance decreases on the batched data (see Figure 3). These results are visually illustrated with UMAP embeddings in Figure 6. The LSI embedding illustrates a high degree of overall separation between batches with non-distinct cell-type clusters, while cells originating in each batch are interspersed across the CisTopic and scBasset embeddings. It is worth noting that some cell-type clusters within the scBasset embedding appear to separate by batch, while this effect is less evident in the CisTopic embedding.

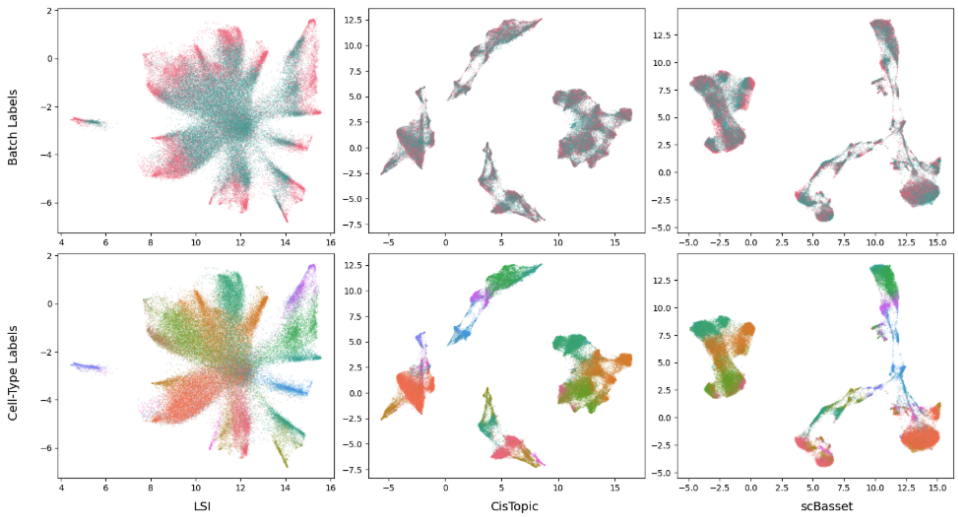


**Figure 6:** UMAP visualizations of embeddings produced by LSI, CisTopic, and scBasset on batched data. Plots in the top panel are colored by batch label, and plots in the bottom panel are colored by cell-type label.

## 4.3 CisTopic embeddings best reflect ground-truth developmental trajectories

Informative embeddings of scATAC-seq data should not only reflect the diversity of cell-types in the original tissue sample, but also the relationships between them. To understand how chromatin accessibility changes throughout development, we must be able identify the series of stages a cell goes through as it progresses from a progenitor cell-type to a differentiated one. Therefore, it is important to generate cell-type embeddings that reflect developmental trajectories. We evaluated the extent to which each method encodes well-studied developmental trajectories using VIA, a graph-based trajectory inference algorithm that reconstructs cell lineages using a low-dimensional cell embedding as input (6).

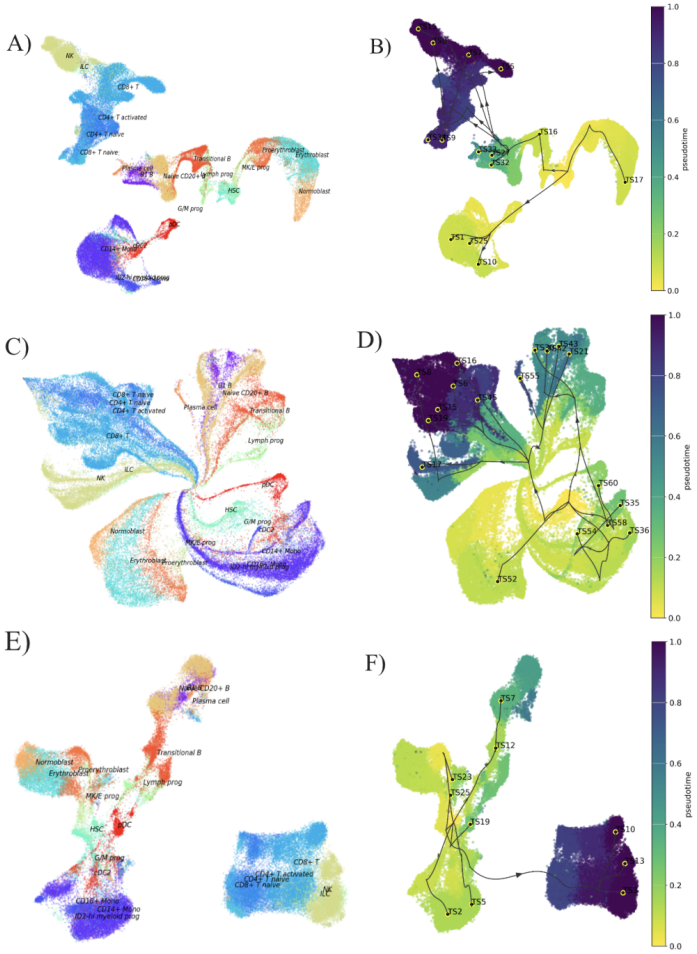


**Figure 7:** UMAP representations of CisTopic (A,B), LSI (C,D) and scBasset (E,F) embeddings generated from uperturbed data. Panels A, C, and E are colored and labeled according to cell-type and panels B, D, and F are colored according to pseudotimes computed by VIA. Black lines with arrows indicate VIA-inferred developmental trajectories.

Figure 7 illustrates UMAP representations of the embeddings generated from each dimensionality-reduction method using the original dataset (without any additional dropout or batch effects introduced). Panels A, C, and E are colored and labeled according to cell-type. Hematopoietic Stem Cells (HSCs) are multipotent stemm cells in the bone marrow that give rise to the incredible diversity of blood cell-types, including oxygen carrying red blood cells and the white blood cells that compose the immune system. HSCs initially give rise myeloid (G/M), erythroid (MK/E), and lymphoid (lymph) progenitor cells. Myeloid progentiors ultimately give rise to different types of monocytes, erythroid progentiors ultimately give rise to normoblasts (red blood cells), and lymphoid progentors give rise to the array of T, B, and Natural Killer (NK) cells in the immune system. Therefore, a well-structured embedding should illustrate HSCs merging into the three main progenitor types, which then produce three separate branches ultimatley leading to normoblasts, monocytes, and T, B, and NK cells. This general structure is highly

apparent in the cisTopic embedding and largely apparent in the scBasset embedding, but appears less defined in the LSI embedding. Panels B, D, and F are colored according to VIA computed pseudotime. Pseudotime reflects the stemness of a cell, or its capacity to give rise to new cell types. Therefore, HSCs and the surrounding progenitor cell types should have low pseudotimes, and differentiated cells at the end of each branch have higher pseudotimes. This trend is generally reflected in all three embeddings.

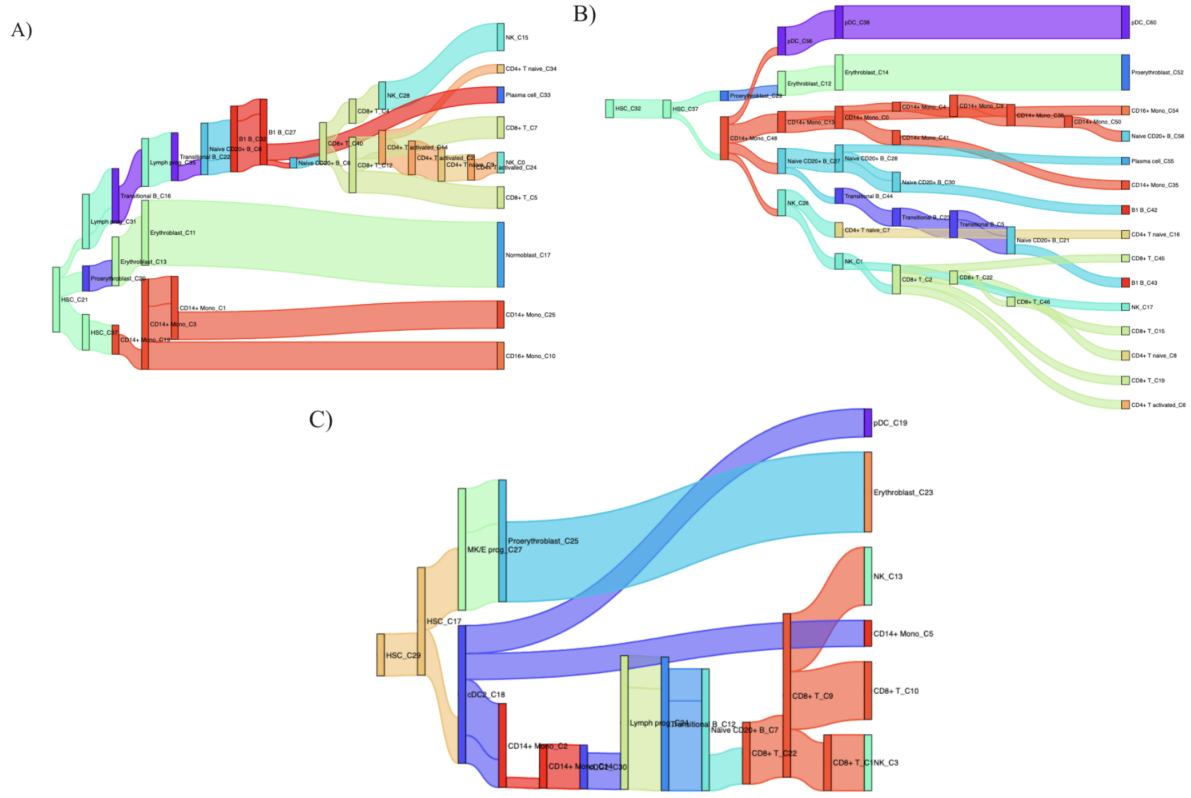


**Figure 8:** Differentiation flow plots illustrating VIA-inferred developmental trajectories generated using embeddings from CisTopic (A), LSI (B), and scBasset (C). Differentiation progresses from left to right. Widths reflect the relative number of cells of a given type at a given differentiation stage.

The developmental trajectories indicated by the black arrows overlaying the pseudotime plots in Figure 7 are represented explicitly in Figure 8. Here, our observations regarding representation of expected trajectories in each embedding are more clearly visualized. CisTopic clearly recovers the three main expected branches of differention: monocyte development, erythroid development, and the development of T, B, and NK cells. While correctly placing HSCs at the root of the differentiation tree and recoverying the erythroid developmental trajectory, VIA incorrectly represents CD14+ Monocytes, a terminally differentiated cell-type, as giving rise to B, T, and NK cells when the LSI embedding is used as input. The trajectories inferred from the scBasset embedding are similarly problematic, with cDC2 (dendritic) cells giving rise to the entire immune system. While LSI and scBasset may produce cleaner results with different trajectory inference algorithms, this analysis suggests that CisTopic embeddings best represent ground-truth cell-type lineages.

# 5 Discussion

The primary objective of this study was to evaluate the robustness of three dimensionality reduction methods — LSI, cisTopic, and scBasset — in generating biologically meaningful embeddings from sparse scATAC-seq data. Our analysis focused on the ability of each method to maintain cell-type cluster integrity, resist batch effects, and preserve developmental trajectories under increasing sparsity. The key findings and their broader implications are discussed below.

## 5.1 Robustness to Sparsity

The robustness of dimensionality reduction methods to sparsity is critical for scATAC-seq analysis due to the inherently sparse nature of the data. Our results demonstrate that cisTopic is the most effective method for preserving biologically relevant embeddings as sparsity increases. While scBasset maintained strong performance up to 80% sparsity, its performance degraded significantly at 95% sparsity. LSI showed the poorest performance, with clustering metrics (ARI, AMI, and Homogeneity Score) approaching zero at high sparsity levels.

The superior performance of cisTopic can be attributed to its topic modeling approach, which identifies meaningful latent features rather than relying on simple linear transformations like LSI. By learning co-accessibility patterns across regions, cisTopic generates embeddings that maintain biological relationships, even when the majority of counts are zero. In contrast, LSI, which relies on matrix decomposition, struggles to distinguish biological variation from noise as sparsity increases. scBasset's use of convolutional neural networks allows for the detection of complex patterns, but the reliance on training data makes it more sensitive to extreme sparsity.

These results suggest that cisTopic should be the preferred method for scATAC-seq datasets with high sparsity.

## 5.2 Impact of Batch Effects

Batch effects are a significant challenge in single-cell sequencing analysis, often arising from differences in sequencing depth, reagent quality, or other experimental factors. We introduced a synthetic batch effect by reducing read depth by 40% in half of the cells. The resulting analysis revealed that LSI embeddings were highly susceptible to batch effects, as clusters aligned more with batch labels than with cell-type labels. In contrast, both cisTopic and scBasset embeddings displayed strong robustness to batch effects, with minimal separation of cells by batch in UMAP visualizations.

The batch susceptibility of LSI can be explained by its reliance on linear transformations, which are unable to disentangle batch-specific variation from biological variation. The robustness of cisTopic is likely due to its probabilistic modeling of topic distributions, which allows for more flexibility in capturing cross-sample similarities. scBasset's robustness may stem from the convolutional network's ability to generalize across varying input distributions.

Given these findings, LSI should be used with caution in studies where batch effects are present. cisTopic, with its demonstrated robustness, is better suited for batch-corrected or multi-batch datasets. While scBasset also showed batch robustness, its greater computational demands and sensitivity to hyperparameters make it a less practical choice for large datasets.

## 5.3 Preservation of Developmental Trajectories

A key goal of scATAC-seq analysis is to capture cellular differentiation processes and reconstruct developmental trajectories. Using VIA trajectory inference, we assessed each method's ability to preserve known hematopoietic differentiation trajectories. The embeddings produced by cisTopic most clearly reflected expected cell lineage trajectories, including well-defined branches for the development of erythroid, myeloid, and lymphoid lineages. While scBasset embeddings captured some trajectory structure, the inferred paths showed erroneous developmental relationships, such as dendritic cells giving rise to lymphocytes. LSI embeddings produced the least accurate trajectories, with VIA incorrectly placing monocytes as progenitors for lymphoid cells.

These results further support the superiority of cisTopic for downstream trajectory analysis. The ability to accurately capture lineage information is crucial for understanding developmental processes in health and disease. The robust trajectory preservation by cisTopic suggests that it can effectively represent continuous biological processes.

## 5.4 Computational Efficiency

While performance is the primary consideration when selecting dimensionality reduction methods, computational efficiency is also an important factor. LSI is the fastest method due to its reliance on linear algebraic techniques, making it suitable for exploratory analysis and large datasets where speed is critical. cisTopic, while slower than LSI, is computationally efficient relative to scBasset, as its topic modeling approach requires iterative optimization but avoids the large-scale training demands of deep learning models. scBasset, on the other hand, requires significant computational resources and time for training, especially on large datasets with many cells and regions.

For studies where computational efficiency is paramount, LSI may be used as an initial exploratory tool, but for high-quality analysis and robust embeddings, cisTopic should be prioritized. scBasset, while effective in low-sparsity scenarios, may not justify the computational cost for most practical applications.

## 5.5 Limitations and Future Directions

While our study provides a comprehensive analysis of dimensionality reduction methods for scATAC-seq, several limitations should be noted. First, we only evaluated three methods, and future studies could incorporate other state-of-the-art techniques, such as Harmony and UMAP, for comparison. Second, the introduction of synthetic sparsity and batch effects may not fully capture the complexity of real-world batch effects observed in experimental data. Applying these methods to experimentally generated multi-batch scATAC-seq datasets would provide more generalizable insights.

Additionally, our analysis focused on one dataset (a human bone marrow multiomic atlas). While this dataset is well-annotated and suitable for benchmarking, future research should validate our findings across diverse datasets, including those with distinct tissue types and experimental protocols.

Finally, future improvements to scBasset's architecture may address its sensitivity to sparsity. Incorporating dropout-aware training strategies or self-supervised learning approaches could increase the model's robustness to missing data. The development of computationally efficient implementations for scBasset could also increase its usability in large-scale scATAC-seq studies.

## 5.6 Conclusions

This study comprehensively evaluates three dimensionality reduction methods for scATAC-seq data, focusing on their robustness to sparsity, resistance to batch effects, and preservation of developmental trajectories. Our key conclusions are as follows:

- **cisTopic is the most robust method**, demonstrating superior performance in cell-type clustering, batch effect resistance, and trajectory preservation, even under high sparsity.
- **LSI is computationally efficient but performs poorly** in the presence of batch effects and high sparsity, making it suitable only for initial exploratory analysis.
- **scBasset offers strong performance** but requires significant computational resources and is less robust to high sparsity than cisTopic.

Given its performance across all dimensions, cisTopic is the recommended method for scATAC-seq dimensionality reduction, especially when datasets are expected to be sparse or contain batch effects. This study highlights the importance of selecting dimensionality reduction methods that are not only computationally efficient but also biologically meaningful, enabling more accurate analysis of cellular heterogeneity and developmental trajectories.

# 6 References


1. **Granja, J.M., Corces, M.R., Pierce, S.E., et al.** (2021). ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nature Genetics*, **53**, 403–411. DOI: 10.1038/s41588-021-00790-6

2. **Luecken, M.D., Theis, F.J.** (2021). A sandbox for prediction and integration of DNA, RNA, and proteins in single cells. *NeurIPS*. DOI: 10.1038/s41588-021-00790-6

3. **Gonzales-Blas, C.B., Dey, K.K., Kalluri, S., et al.** (2019). cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. *Nature Methods*, **16**, 397–400. DOI: 10.1038/s41592-019-0367-1

4. **Yuan, H., Kelley, D.R.** (2022). scBasset: sequence-based modeling of single-cell ATAC-seq using convolutional neural networks. *Nature Methods*, **19**, 1088–1096. DOI: 10.1038/s41592-022-01562-8

5. **Genome Biology Article.** (2019). Accessed from Genome Biology: 10.1186/s13059-019-1854-5

6. **Nature Communications Article.** (2021). Accessed from Nature Communications: 10.1038/s41467-021-25773-3