**Exploring Deep Learning in Single-Cell Studies: PCA and Autoencoder Approaches for Feature Extraction and Clustering of Muscle Cells**

**Abstract**

This project compares the performance of traditional PCA and a deep learning autoencoder for feature extraction in analyzing high-dimensional single-cell RNA sequencing (scRNA-seq) data from the hindlimb muscle of the Green Anole lizard. The primary goal is to assess whether a fully connected neural network can uncover distinct clustering patterns beyond those identified by conventional scRNA-seq tools like Seurat and Scanpy, which rely on PCA for dimensionality reduction. Using UMAP visualization and GMM clustering, the autoencoder model revealed distinct subclusters within the data, potentially revealing heterogeneity in the muscle stem cell (MuSC) population. These findings offer deeper insights into the biological complexity of muscle regeneration, highlighting the potential for incorporating deep learning into single-cell data analysis.

**Introduction**

The Green Anole lizard’s remarkable ability to regenerate its tail after injury provides an excellent model for studying muscle stem cell (MuSC) function. Muscle regeneration involves the activation, proliferation, and differentiation of MuSCs, alongside interactions with other cell types such as macrophages and fibroblasts (De Micheli et al., 2020). To explore these processes, our lab performed single-cell RNA sequencing (scRNA-seq) on lizard hindlimb muscle cells, which offer insights into cell composition and gene expression profile at single-cell resolution.

Computational analysis of scRNA-seq data is crucial to uncover cellular heterogeneity and dynamics. Traditionally, scRNA-seq data analysis is performed using well-established packages like Seurat and Scanpy, which rely on traditional machine learning (ML) algorithms such as Principal Component Analysis (PCA) for dimensionality reduction and the Louvain algorithm for cell clustering (Stuart et al., 2019). However, deep learning (DL), with its ability to model complex relationships using neural networks, remains underutilized in biological research (Ma and Xu, 2022). Unlike PCA, DL can capture intricate patterns, potentially retaining critical information that traditional ML methods may overlook (Brendel et al., 2022).

This project compares the performance of PCA with an autoencoder, a feed-forward fully connected neural network, for learning low-dimensional representations of scRNA-seq data. Unsupervised clustering method was applied to identify distinct cell populations. The predicted cell labels were compared with those generated by Seurat, which serves as a reliable benchmark due to its high reliability. Model performance was evaluated using metrics such as training and validation loss, mean squared error (MSE), silhouette score, and Adjusted Rand Index (ARI), providing a comprehensive evaluation of both dimensionality reduction techniques.

**Methods**

The scRNA-seq data used in this project were obtained from Almada Lab at USC Stem Cell. The dataset consists of cells isolated from the hindlimb of the Green Anole lizard, which is believed to predominantly contain muscle stem cells. The original dataset includes 20,790 gene features across 9,295 cell samples. To reduce computational complexity, the dataset was downsampled to 2,000 randomly selected cells, focusing on the top 2,000 genes exhibiting the highest cell-to-cell variation.

All preprocessing steps were conducted using Seurat packages in R. The downsampled dataset was normalized using Seurat's “LogNormalize” function, which normalizes feature expression for each cell by the total expression, scales by a factor of 10,000, and log-transforms the result (Satijalab). The normalized data was then scaled using Seurat's “ScaleData” function, centering the data to a mean of 0 and scaling it to a variance of 1 (Satijalab). After preprocessing, the gene-by-cell matrix was extracted and loaded in Python for subsequent analysis.

2.1 Dimensional Reduction Model Training: PCA

PCA dimensional reduction was applied to the 2000x2000 scRNA-seq data. The proportion of variance explained by each component was calculated, and the transformed data were visualized by plotting the first two principal components (PCs). The transformed data were visualized in 2D using Uniform Manifold Approximation and Projection (UMAP). Optimal values for n\_neighbors and min\_dist were determined through hyperparameter tuning. True labels, previously generated from Seurat, were overlaid on the plots for interpretation. To assess data separation quality, silhouette scores were calculated for both PCA and UMAP visualizations.

2.2 Feature Extraction Model Training: Autoencoder

The data were split into 80% training and 20% testing sets for model training. A fully connected feedforward autoencoder was constructed with an encoder consisting of three dense layers (2048, 512, and 256 nodes) and a latent layer with 50 nodes, followed by a decoder to reconstruct the original input. Dropout layers were added after each encoding and decoding step to prevent model overfitting (**Table 1**), and the optimal dropout rate was determined through hyperparameter tuning. The model was trained with a batch size of 64 for 100 epochs. After training, the performance was evaluated using reconstruction loss, measured by MSE. The latent representation of the data was extracted and visualized in 2D using both PCA and UMAP plots.

**Table 1: Autoencoder Architecture**

|  |  |  |  |
| --- | --- | --- | --- |
| **Layer (type)** | **Output Shape** | **Activation Function** | **Dropout Rate** |
| input\_layer (InputLayer) | (None, 2000) | N/A |  |
| encoder1 (Dense) | (None, 2048) | ReLu |  |
| dropout (Dropout) | (None, 2048) | ReLu | 0.25 |
| encoder2 (Dense) | (None, 512) | ReLu |  |
| dropout\_1 (Dropout) | (None, 512) | ReLu | 0.25 |
| encoder3 (Dense) | (None, 256) | ReLu |  |
| dropout\_2 (Dropout) | (None, 256) | ReLu | 0.25 |
| latent (Dense) | (None, 50) | ReLu |  |
| decoder1 (Dense) | (None, 256) | ReLu |  |
| dropout\_3 (Dropout) | (None, 256) | ReLu | 0.25 |
| decoder2 (Dense) | (None, 512) | ReLu |  |
| dropout\_4 (Dropout) | (None, 512) | ReLu | 0.25 |
| decoder3 (Dense) | (None, 2048) | ReLu |  |
| dropout\_5 (Dropout) | (None, 2048) | ReLu | 0.25 |
| out\_layer (Dense) | (None, 2000) | Linear |  |

2.3 Unsupervised Clustering Analysis on Reduced Dimensions

Gaussian mixture model (GMM) clustering algorithm was applied to both PCA-reduced data and the autoencoder low-dimensional representation to identify cell labels. The optimal number of clusters (K) was determined using the Bayesian Information Criterion (BIC). Clustering was then performed, and the predicted labels were overlaid on UMAP visualizations and compared with the true labels generated from Seurat. Clustering performance was evaluated using the Adjusted Rand Index (ARI).

**Results**

3.1 Data Exploration to Check Sparsity

A key challenge in scRNA-seq analysis is dealing with data sparsity, where many entries are zero. These zeros can result from either the failure to detect a transcript during the experiment or the true biological absence of a transcript (Bouland et al., 2023). Upon examining the normalized and scaled data, I found the dataset had 0.00% sparsity (i.e., no zero entries), but most expression values were close to 0 (**Figure 1**). These near-zero values make it difficult for the model to distinguish meaningful biological patterns from noise.

A blue square with a blue background

Description automatically generated

**Figure 1: Gene Expression Level Heatmap.** A portion of the cells in scRNA-seq data was visualized. The majority of the cells displayed close to 0 expression level.

3.2 Performance of Conventional PCA Dimension Reduction in scRNA-seq Data

Due to the near-zero values in the datasets, the first few principal components (PCs) did not capture a significant portion of the variance (**Figure 2a**)**.** Based on typical scRNA-seq practices, I reduced the data to 50 PCs to ensure sufficient biologically meaningful information for downstream cell clustering analysis. The first two PCs were visualized and overlaid with the true cell labels, yielding a silhouette score of 0.006641759. This score, close to zero, indicated that PCA was unable to capture the true structure of the data, with some overlaps observed between clusters (**Figure 2b**). This aligns with the limitations of PCA, which struggles to capture non-linear relationships, such as biological confounding factors like cell cycle effects (Heumos et al., 2023).

To better capture the dataset's underlying structure, the PCA-reduced data was visualized using UMAP, which is expected to better represent complex relationships. Several configurations of n\_neighbors (5, 15, 30) and min\_dist (0.1, 0.5, 0.9) were tested, with n\_neighbors = 15 and min\_dist = 0.1 selected due to its higher silhouette score of 0.679026067. Compared to PCA visualization, UMAP provided better cluster separation, confirming its ability to preserve both local and global structures and reveal the complex relationships within data (**Figure 2c**).

While UMAP is often used to capture complex patterns in single-cell RNA sequencing (scRNA-seq) data, it is typically applied to PCA-reduced data for visualization rather than directly performing dimensionality reduction (Heumos et al., 2023). According to some studies, UMAP's performance can decrease as the cell number and sparsity of the data increase, as it relies on constructing a neighborhood graph (Xiang et al., 2021). The similar result was observed in this analysis. When UMAP was directly applied to the whole dataset (2000x2000), the resulting clusters were not well separated and had a very low silhouette score of 0.118827730 compared to UMAP on PCA-reduced data **(Figure 2d).**

a

a

b

A graph with blue dots

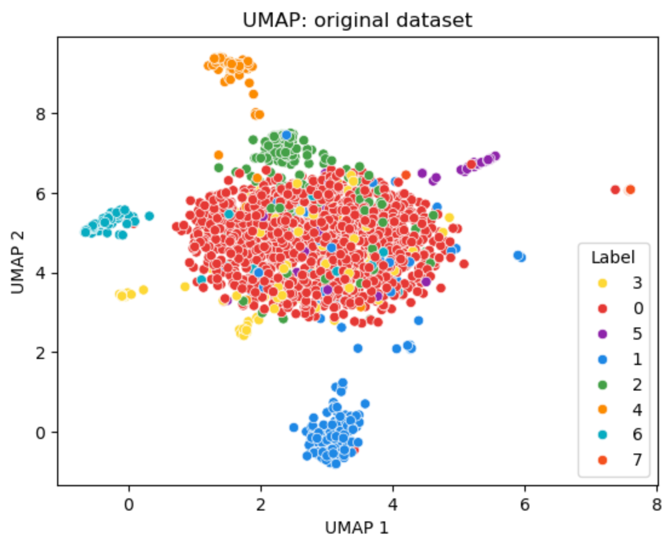
Description automatically generatedA diagram of a graph

Description automatically generated with medium confidence

d

c

A diagram of a map

Description automatically generated with medium confidence

**Figure 2: Performance of Conventional PCA Dimension Reduction.** **a.** Explained variance by each PC. Less than 5% of the variance is captured by the first few PCs. **b.** PCA plot of first two PCs. Despite low variance, distinguishable clusters are observed, indicating that the model captures meaningful biological variation. **c.** UMAP visualization on the PCA-reduced data. Improved cluster separation is seen in the UMAP plot compared to PCA, highlighting its ability to better preserve complex relationships. **d.** UMAP dimension reduction. Poor cluster separation when directly applied to the whole dataset

3.3 Performance of Autoencoder on Feature Extraction in scRNA-seq Data

To better capture complex, intrinsic patterns in scRNA-seq data, I built a basic autoencoder model for feature extraction (**Table 1**). The mean squared error (MSE) loss curves show a significant decrease in loss for both training and testing sets, but the testing set plateaued earlier around 40 epochs out of 100 epochs (**Figure 3a**), indicating potential overfitting. The overall reconstruction loss on the entire dataset was 0.57204. Figure 3b visualizes the loss for each of the 2000 cells, with the majority of cells having a loss value below 1.0.

a

b

A graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of

Description automatically generatedA graph of a cell phone loss

Description automatically generated

**Figure 3: Autoencoder Reconstruction Loss.** a. Reconstruction loss for each cell. b. training and validating loss curves. Dropout rates of 0.5, 0.3, 0.25, and 0.2 were experimented with, and the optimal 0.25 rate was found as it balanced between model underfitting and overfitting and gave the lowest MSE value.

To further explore the patterns captured by the autoencoder, the latent layer was extracted and visualized using both PCA and UMAP. Comparing the PCA plots between the autoencoder latent representation (**Figure 4a**) and PCA-reduced data (**Figure 2b**), the autoencoder showed improved cluster separation, especially in clusters 6, 3, 7, and 5. However, the silhouette score for the autoencoder was lower (-0.158318207). Similarly, UMAP visualization of the autoencoder data (**Figure 4b**) revealed a different clustering result from the PCA-reduced data (**Figure 2c**), with a silhouette score of -0.044356871. Despite the lower silhouette scores, the autoencoder separated the largest cluster (Cluster 0) into multiple subclustering patterns. Differential gene expression analysis in my lab identified Cluster 0 as muscle stem cells (MuSCs), which are known to exhibit heterogeneity across different cellular states such as quiescence, activation, and differentiation. While the low silhouette scores might traditionally suggest poor clustering, the visualization suggests that the autoencoder could have captured complex biological phenomena, such as transitions between different cell states and communication between distinct cell types.

b

a

A diagram of a graph

Description automatically generated with medium confidenceA diagram of a number of colored dots

Description automatically generated with medium confidence

**Figure 4: Latent Space Visualizations Using PCA and UMAP.** a. PCA based on latent layer. b. UMAP based on the latent layer. The same parameters were used as in PCA-reduced data.

3.4 Performance of Unsupervised Clustering Models on PCA-reduced Data and Autoencoder Latent Representation

Clustering analysis is an important step in scRNA-seq to determine cell identities and enable biological interpretations. To further evaluate the differences between two dimensionality reduction methods, GMM clustering was applied to both PCA-reduced data and autoencoder latent representations. GMM, a probabilistic model-based clustering technique, performs soft clustering, which is particularly beneficial for clusters of diverse shapes and orientations. ARI scores were calculated for each transformation and are presented in Table 2.

The ARI scores indicate that the GMM clustering method performed better on PCA-reduced data compared to the autoencoder latent layer (**Table 2**). However, visualization of the clusters revealed notable differences between the two methods. For the autoencoder latent space, GMM-generated labels appeared more aligned with the distribution of data, as the largest original cluster was separated into multiple cell types (**Figure 5b vs. Figure 4b).** In contrast, for PCA-reduced data, the true labels better captured distinct cell types, with GMM-generated labels providing less separation (**Figure 5a vs. Figure 2c**).

To quantify these observations, silhouette scores were compared between true and GMM-predicted labels. In the autoencoder latent space, the silhouette score increased with GMM-generated labels, suggesting that the clustering aligned better with local data structures. However, for PCA-reduced data, the silhouette score decreased when using GMM-generated labels, indicating that the true labels were more representative of the data’s structure.

The clustering analysis suggested that the autoencoder potentially captured more subtle, complex patterns in the data, revealing biologically meaningful subclusters.

**Table 2: Model comparison based on GMM clustering results.** ARI, silhouette scores based on true and predicted labels were calculated.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Clustering algorithm** | **ARI** | **silhouette score (true labels)** | **silhouette score (predicted labels)** |
| PCA Dimension Reduction | GMM | 0.691358498 | 0.679026067 | 0.45482964 |
| Autoencoder latent layer | GMM | 0.152460044 | -0.044356871 | 0.06067727 |

b

a

A chart with different colored dots

Description automatically generatedA diagram of a cell

Description automatically generated with medium confidence

**Figure 5: GMM Clustering Analysis Results Visualization. a.** GMM clustering applied to PCA-reduced data. Compared to true labels (Figure 2c), GMM was unable to separate cluster 1 into two distinct cell types. **b**. GMM clustering applied to autoencoder latent representation. Compared to true labels (Figure 4b), GMM separated cluster 0 into multiple subclusters, indicating its potential in identifying subcluster relationships within data.

**Conclusions**

This project compared conventional PCA dimensionality reduction with an autoencoder-based approach for feature extraction in scRNA-seq data. The results suggested that the autoencoder demonstrated the potential to uncover more nuanced data structures, such as subclusters within the muscle stem cell population. However, the autoencoder model showed signs of overfitting, as indicated by the early stabilization of the testing loss and high MSE. To further enhance model performance, future efforts should focus on adjusting neural network structures and training processes, such as the number of neural layers, batch size, and epochs. GMM clustering analysis revealed that autoencoder latent space facilitated better identification of subclusters compared to PCA-reduced data. However, the low silhouette score suggested that the clustering performance requires optimization. More advanced clustering algorithms commonly utilized in scRNA-seq datasets, such as Louvain, should be considered in future work to better handle the sparsity issue (Wickramasinghe & Muthukumarana, 2022). Additionally, downstream scRNA-seq analyses, such as differential gene expression analysis, are necessary to confirm whether the autoencoder truly captured biologically meaningful signals.

Overall, these findings highlight the potential of deep learning approaches to improve single-cell studies. These findings are critical for understanding cellular heterogeneity, especially in muscle tissues where distinct cell populations, such as MuSCs, undergo dynamic changes. Deep learning model’s potential to capture transitions between cellular states provide insights into cellular processes such as activation, differentiation, and response to injury for muscle regeneration. Incorporating AI methods in biological studies facilitates deeper understanding of cellular behavior and development in both health and disease.

**References**

1. Ashani Wickramasinghe, & Saman Muthukumarana. (2022). Assessing the impact of the density and sparsity of the network on community detection using a Gaussian mixture random partition graph generator. *International Journal of Information Technology*, *14*(2), 607–618. <https://doi.org/10.1007/s41870-022-00873-5>
2. ‌Brendel, M., Su, C., Bai, Z., Zhang, H., Elemento, O., & Wang, F. (2022). Application of Deep Learning on Single-cell RNA Sequencing Data Analysis: A Review. *Genomics, Proteomics & Bioinformatics*, *20*(5), 814–835. <https://doi.org/10.1016/j.gpb.2022.11.011>
3. Bouland, G. A., Mahfouz, A., & Marcel. (2023). Consequences and opportunities arising due to sparser single-cell RNA-seq datasets. *Genome Biology*, *24*(1). <https://doi.org/10.1186/s13059-023-02933-w>
4. De Micheli, A. J., Laurilliard, E. J., Heinke, C. L., Ravichandran, H., Fraczek, P., Soueid-Baumgarten, S., De Vlaminck, I., Elemento, O., & Cosgrove, B. D. (2020). Single-Cell Analysis of the Muscle Stem Cell Hierarchy Identifies Heterotypic Communication Signals Involved in Skeletal Muscle Regeneration. *Cell Reports*, *30*(10), 3583-3595.e5. <https://doi.org/10.1016/j.celrep.2020.02.067>
5. Heumos, L., Schaar, A. C., Lance, C., Litinetskaya, A., Drost, F., Zappia, L., Lücken, M. D., Strobl, D. C., Henao, J., Curion, F., Schiller, H. B., & Theis, F. J. (2023). Best practices for single-cell analysis across modalities. *Nature Reviews Genetics*, 1–23. <https://doi.org/10.1038/s41576-023-00586-w>
6. ‌Ma, Q., & Xu, D. (2022). Deep learning shapes single-cell data analysis. *Nature Reviews Molecular Cell Biology*, *23*(5), 303–304. <https://doi.org/10.1038/s41580-022-00466-x>
7. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, III WMM, Hao Y, Stoeckius M, Smibert P, Satija R (2019). “Comprehensive Integration of Single-Cell Data.” *Cell*, 177, 1888-1902. [doi:10.1016/j.cell.2019.05.031](https://doi.org/10.1016/j.cell.2019.05.031), <https://doi.org/10.1016/j.cell.2019.05.031>.
8. ‌Xiang, R., Wang, W., Yang, L., Wang, S., Xu, C., & Chen, X. (2021). A Comparison for Dimensionality Reduction Methods of Single-Cell RNA-seq Data. *Frontiers in Genetics*, *12*. <https://doi.org/10.3389/fgene.2021.646936>

‌