Machine Learning for scRNA-Seq Cell Type Classification

BioE 245 Final Project – Spring 2025

Nathan Lanclos

Initial comparisons for dimensionality reduction involved PCA on the original data (1A), t-SNE on the original data (1B), and PCA on an early autoencoder's 32D latent space (1C). In that initial comparison, t-SNE provided the clearest separation of the annotated PBMC cell types, particularly excelling where PCA on original data showed significant overlap. However, subsequent refinements to the autoencoder, specifically a configuration with a 64-dimensional latent space, no L1 regularization, L2 regularization, and dropout yielded significantly improved visualization results when PCA was applied to its latent space (Table 1). This revised autoencoder approach produced well-defined clusters for major cell types like Dendritic cells, CD14+ Monocytes, and CD19+ B cells, and also offered clearer separation among several T-cell and NK cells. Visually, the PCA of this improved autoencoder's latent space now performs comparably to, and for some cell type distinctions, arguably better than t-SNE on the original data.

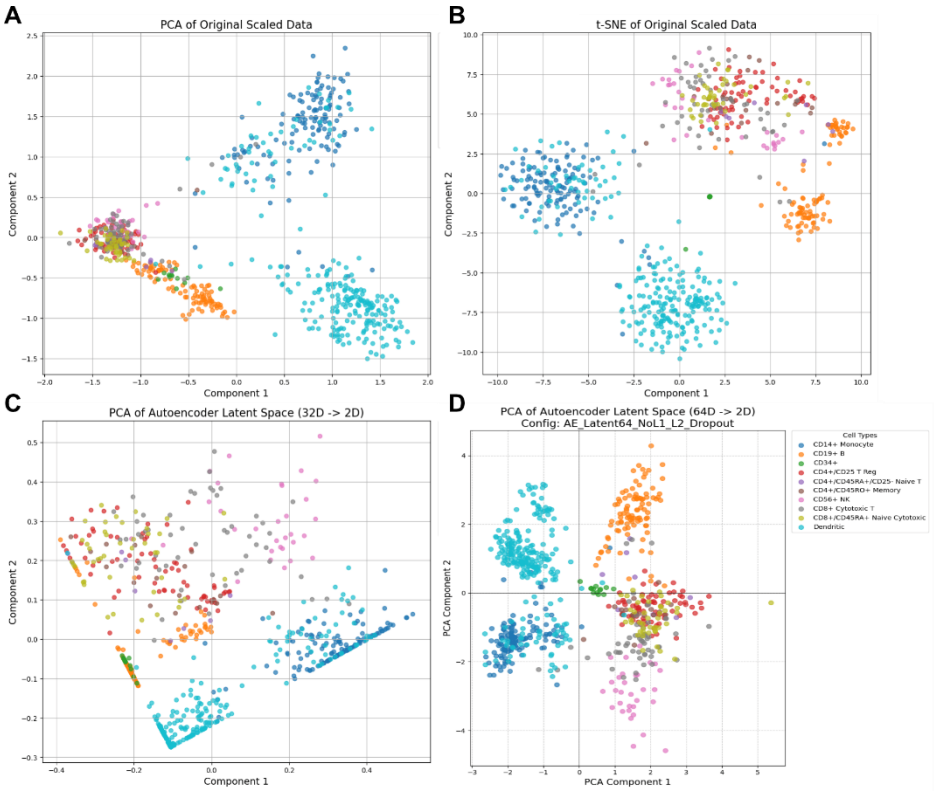

Figure 1. Dimensionality Reduction Plots. (A) PCA/original data, (B) t-SNE/original data, (C) PCA/AE latent (baseline), (D) PCA/AE latent (optimized)

Table 1. Baseline and Optimized Autoencoder Performance

Autoencoder Architecture	MSE
AE (L1 + 32dim/relu + adam) - 217,501 param.	0.0171
AE (L2 + 64dim/linear + adam + dropout) - 475,197 param.	0.0161

For the cell type classification task, we explored two primary model architectures: Random Forest, chosen for its robustness and strong baseline performance on tabular data, and a Multi-Layer Perceptron (MLP), selected for its ability to capture complex non-linear relationships and to align with the autoencoder. Both classifiers were initially trained on the original scaled gene expression data and subsequently on the latent space representations. The Random Forest on original data achieved an accuracy of 0.81, and on the autoencoder latent space, an accuracy of 0.76. The MLP, however, demonstrated better performance on the original data, achieving an accuracy of 0.84, which also remained unchanged after hyperparameter tuning. The MLP performed better on the latent space than the Random Forest, reaching 0.81 accuracy but this was still lower than its performance on the original, high-dimensional gene expression data. To further optimize, hyperparameter tuning (HP) was applied using GridSearchCV for the Random Forest and Keras Tuner for the MLP. Interestingly, hyperparameter tuning did not yield an improvement over the baseline for either model architecture or data representation. Overall, the MLP directly applied to the original scaled gene data provided the highest

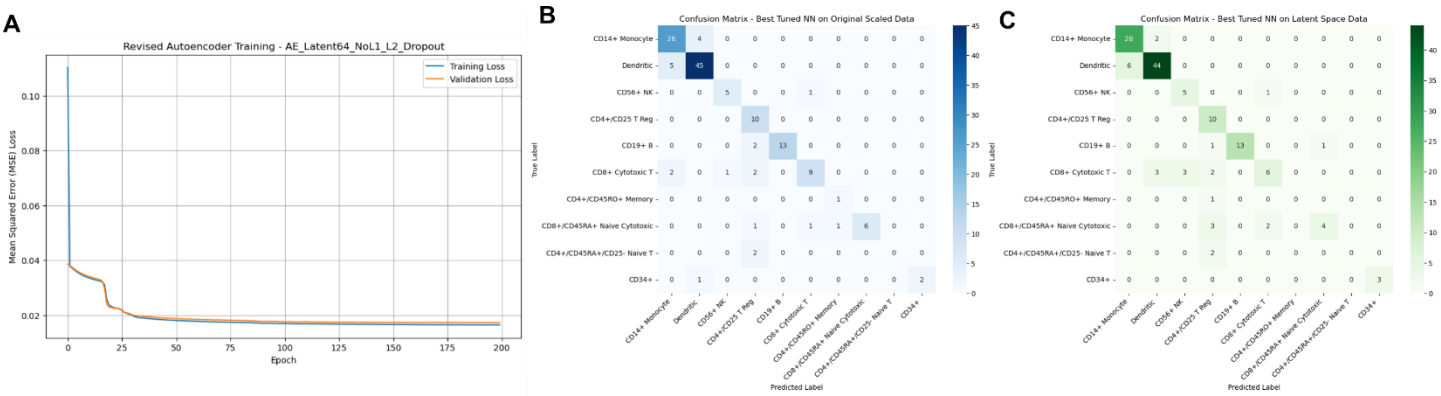

Figure 2. Autoencoder training and Classifier Confusion Matrices. (A) Training curve of optimized autoencoder (blue = training, orange = validation), (B) Confusion matrix of tuned MLP on original data, (C) Confusion matrix of tuned MLP on autoencoder latent space

classification accuracy at 0.84, suggesting that for this dataset and the chosen classifier architectures, the learned autoencoder latent spaces did not enhance, and in some cases hindered, the classification performance compared to using the original feature set with a capable non-linear classifier. The model struggles the most with "CD4+/CD25 T Reg" and "CD8 Cytotoxic T" cell types, but otherwise has very low error rates (2B and 2C). To make these tools accessible and applicable for further development, they were packaged into a streamlit web app that can be run on any machine with the project virtual environment. (Figure 3).

Table 2. Classifier Accuracy by Data Representation and Model Configuration

Data Representation	Classifier Accuracy			
	Random Forest	Random Forest (HP)	MLP	MLP (HP)
Original Data	0.81	0.81	0.84	0.84
Autoencoder Latent	0.76	0.76	0.81	0.81

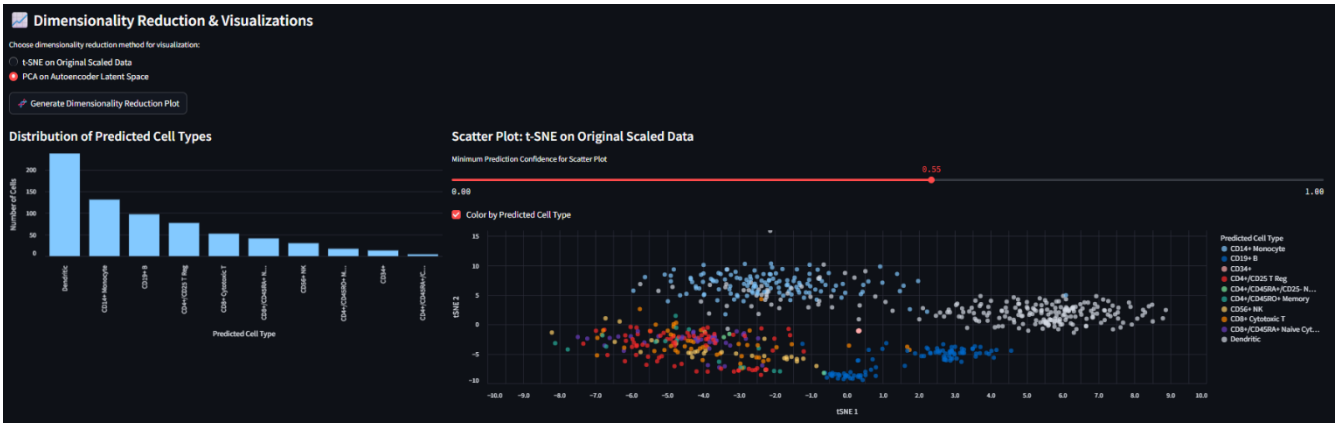

Figure 3. Screenshot of Streamlit UI for visualizing cell type predictions and dimension reduction plots