Part 1:

I built an autoencoder to train on the training dataset and evaluate the loss on the testing dataset. The training dataset's loss is approximately 0.95, while the testing dataset's loss is slightly lower at 0.945 (Figure 1). The similar decreasing trend in losses for both datasets suggests that the model is not overfitting. Next, I applied PCA and t-SNE on the original dataset and the latent space representation obtained from the trained encoder. PCA performed slightly better on the original data, effectively reducing dimensions (Figure 2): it clearly clustered CD19+B cells and part of the dendritic cells. However, it also mixed some dendritic cells and CD14+ monocytes, indicating similar gene expression for these cell types. Other cell types also demonstrated some mixing. In contrast, t-SNE showed improved clustering in the latent space, evidenced by the large distances between different clusters (Figure 3). CD19+B, dendritic, and CD14+ monocytes were clearly separated from other clusters. However, the clustering effect on other cell types remained unclear, similar to PCA. I also examined two random features in the latent space representation and two genes in the original data (Figure 4). The random features in the latent space did not reveal clear clustering effects, performing worse than PCA and t-SNE. In conclusion, t-SNE demonstrated the best performance on the latent space.

Figure 1

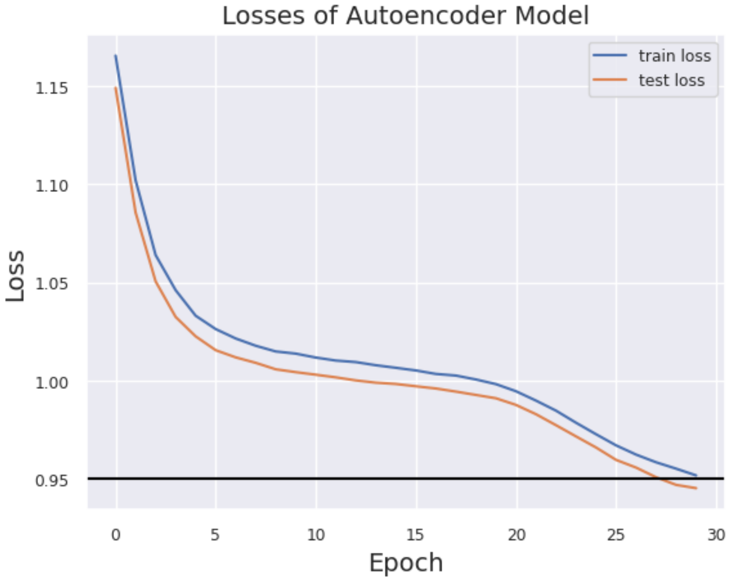


Figure 2

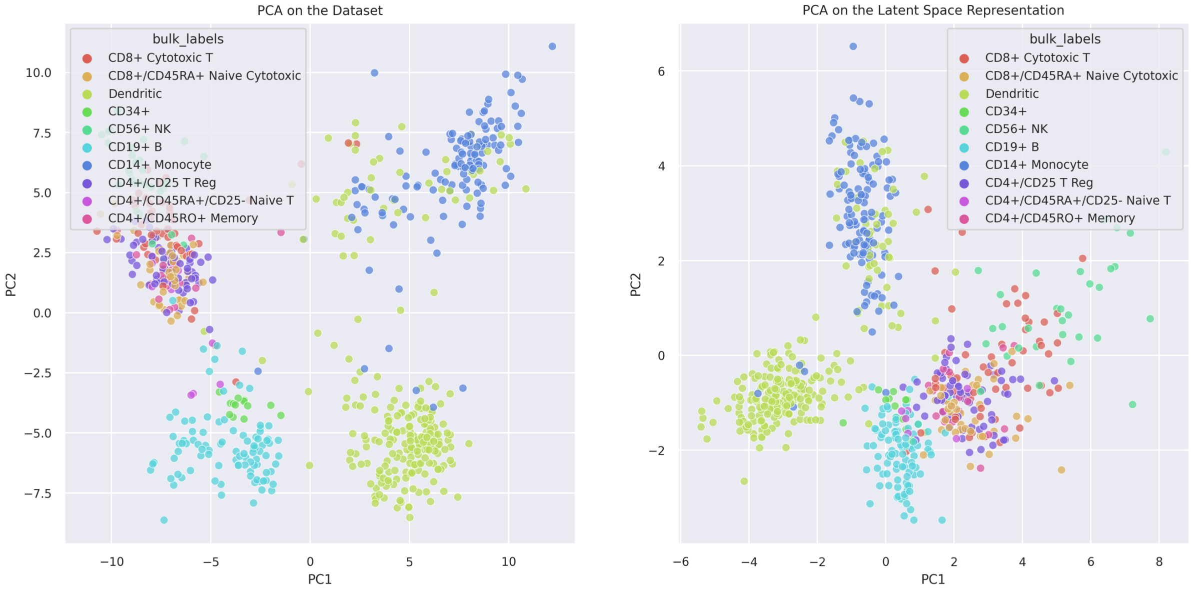
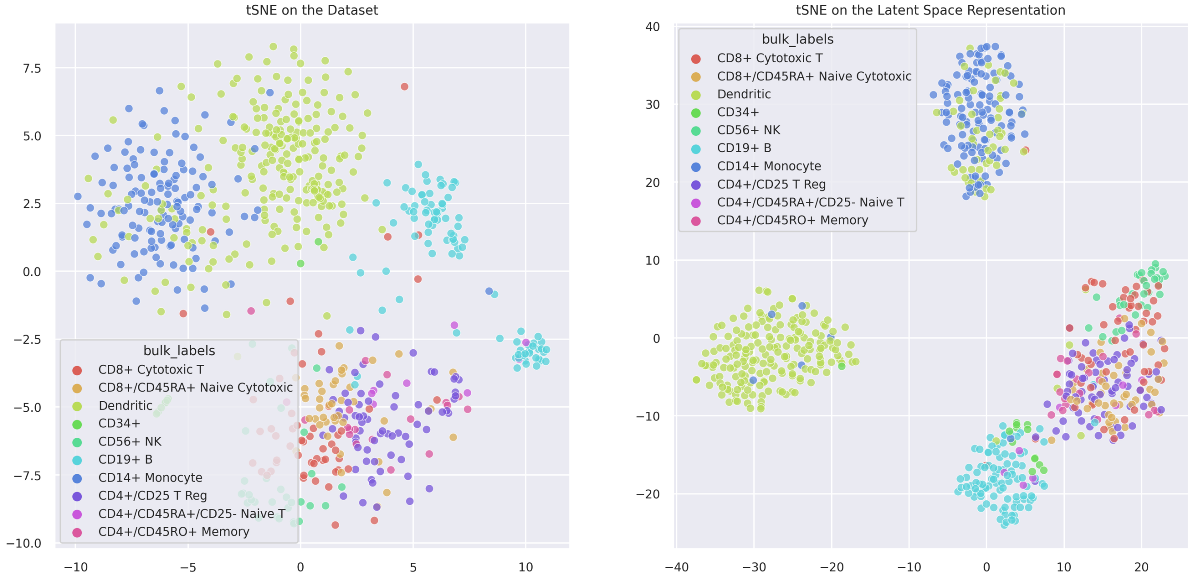


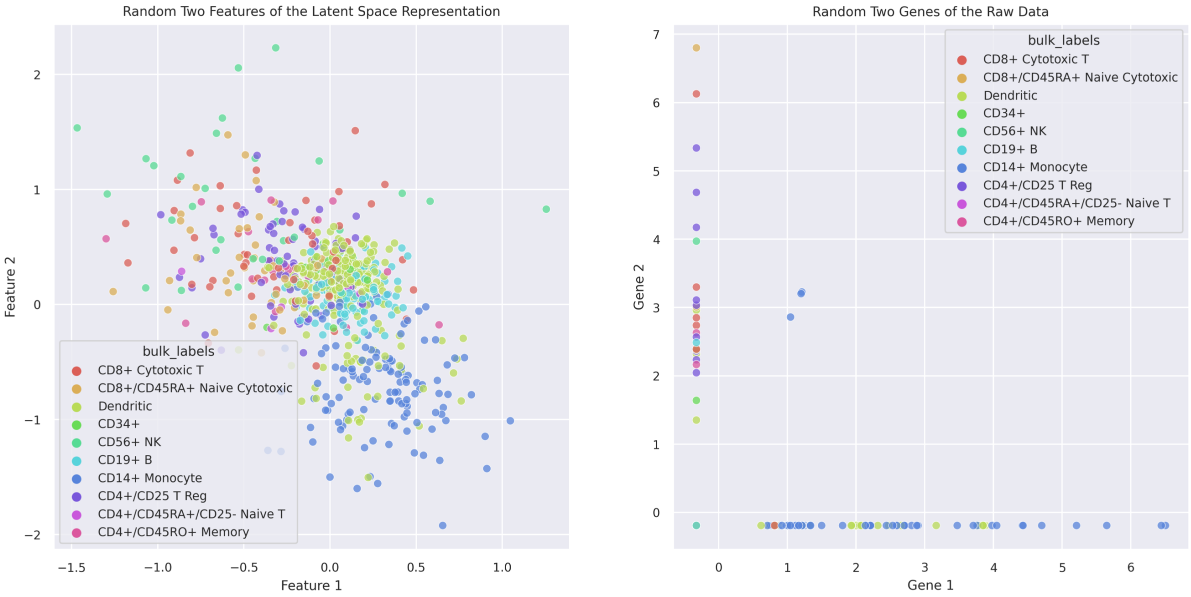
Figure 3



tSNE on the Dataset

tSNE on the Latent Space Representation

Figure 4



Random Two Features of the Latent Space Representation

Random Two Genes of the Raw Data