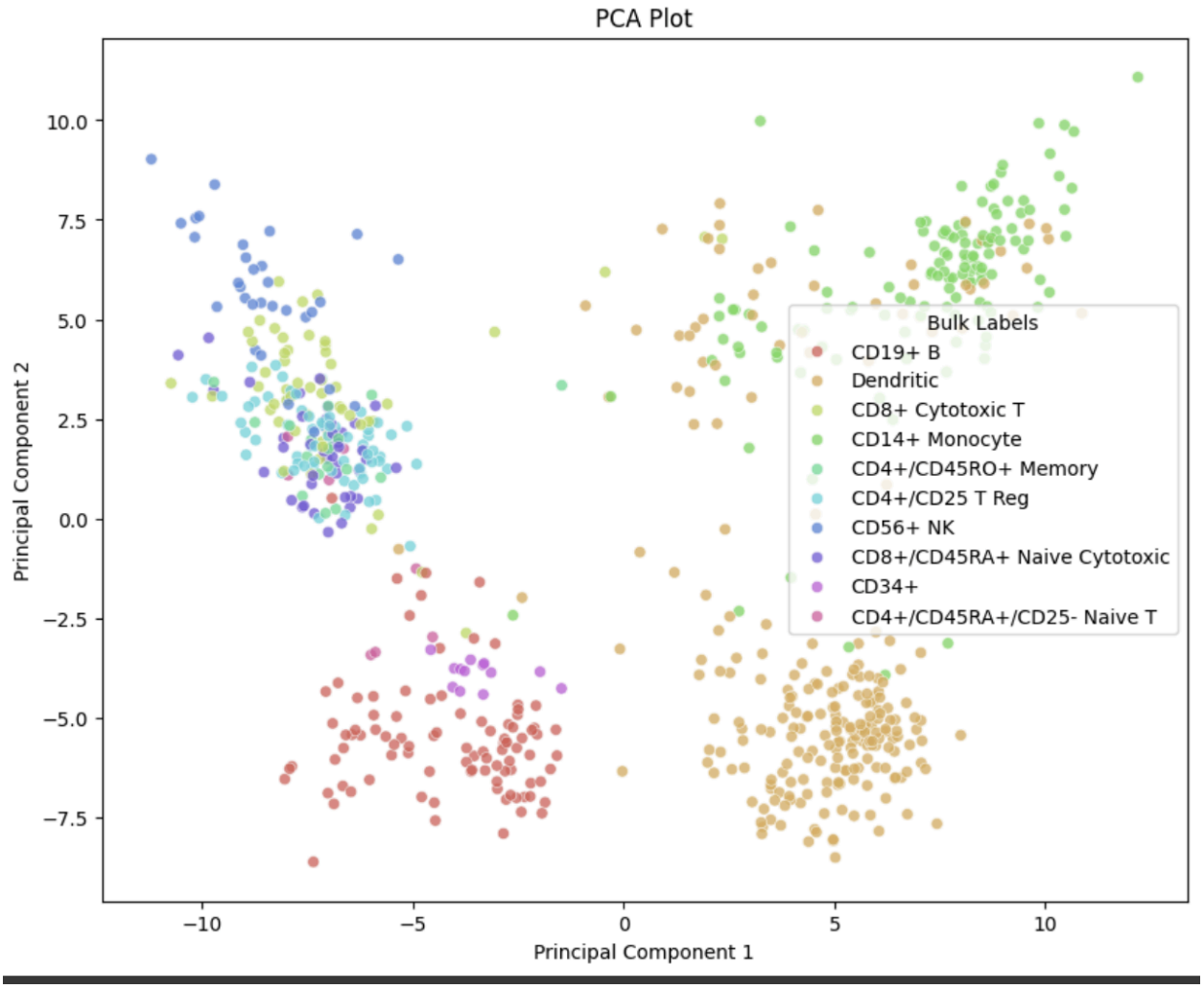
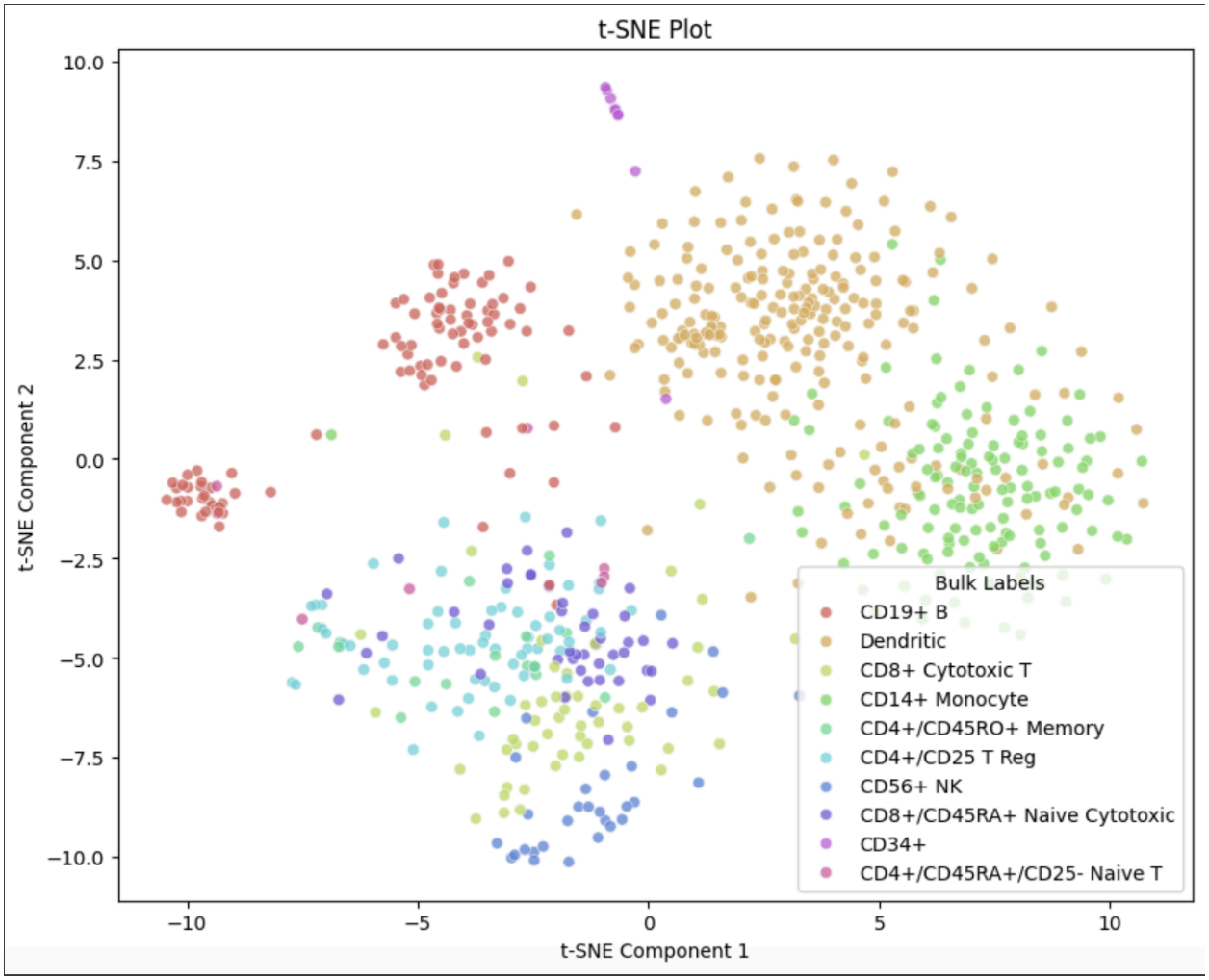
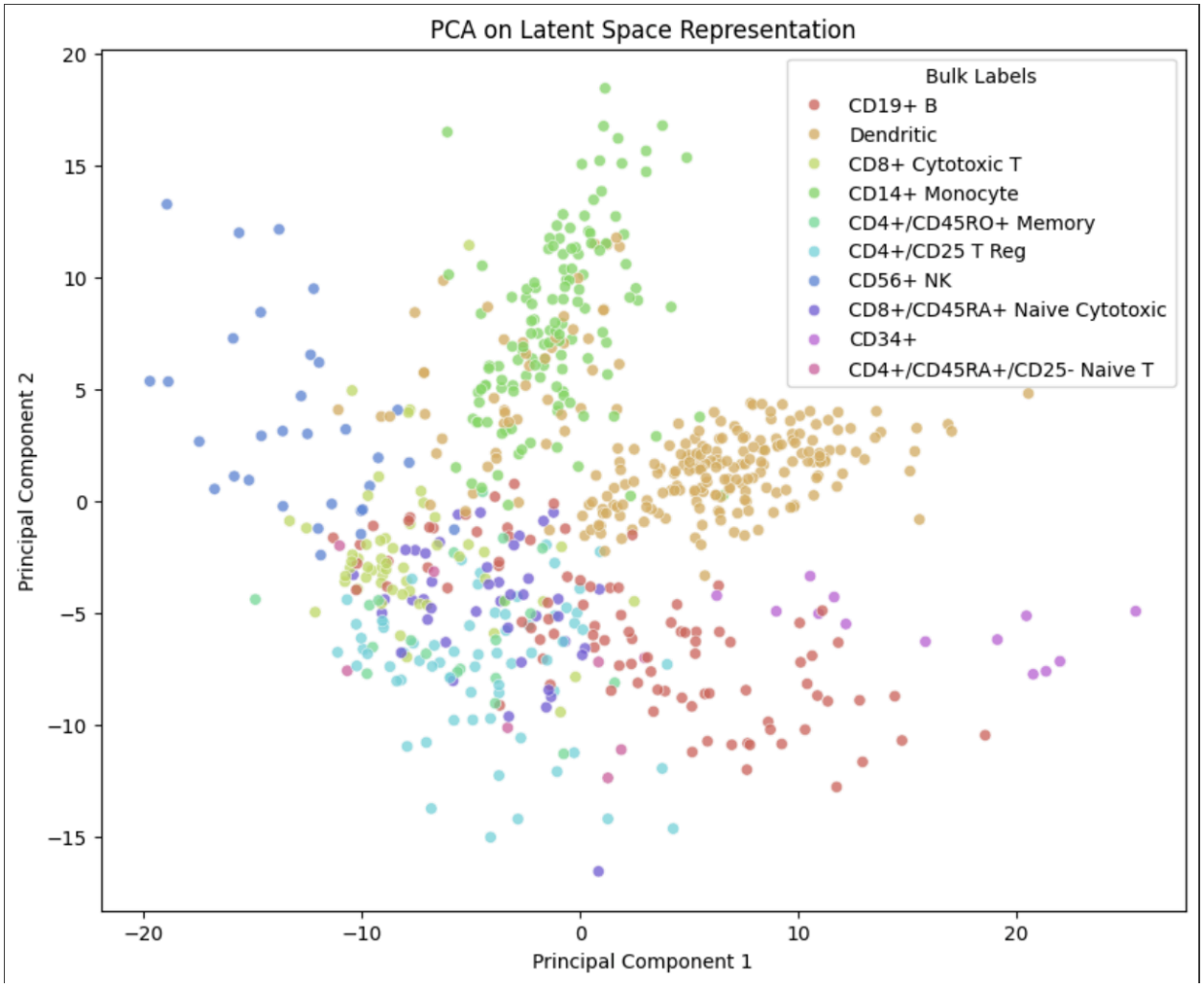
Part 1:
Among the three dimensionality reduction approaches—PCA, t-SNE, and PCA Latent Representation, I think t-SNE performed the best. Although the clusters are not as well defined as in PCA, the cells seem more separated according to their cell type labels compared to PCA and autoencoder. The t-SNE algorithm is good at capturing local structure and preserving the local distances between points, making it particularly effective for visualizing high-dimensional data in lower dimensions especially in this case with lots of features. There is more clarity of a separation between the cell types, especially when looking at “CD4+/CD25 T Reg”, “CD14+ Monocyte”, and the “CD8+/CD45RA+ Naive Cytotoxic ”, which has more separation compared to PCA and the autoencoder. PCA doesn’t capture the intricate local relationships present in the data and the autoencoder doesn't have as clear clusters as PCA, therefore making it worse at visualization than t-SNE.




t-SNE Plot
t-SNE Component 2
t-SNE Component 1
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
−10
−5
0
5
10
Bulk Labels
CD19+ B
Dendritic
CD8+ Cytotoxic T
CD14+ Monocyte
CD4+/CD45RO+ Memory
CD4+/CD25 T Reg
CD56+ NK
CD8+/CD45RA+ Naive Cytotoxic
CD34+
CD4+/CD45RA+/CD25- Naive T



Part 2:

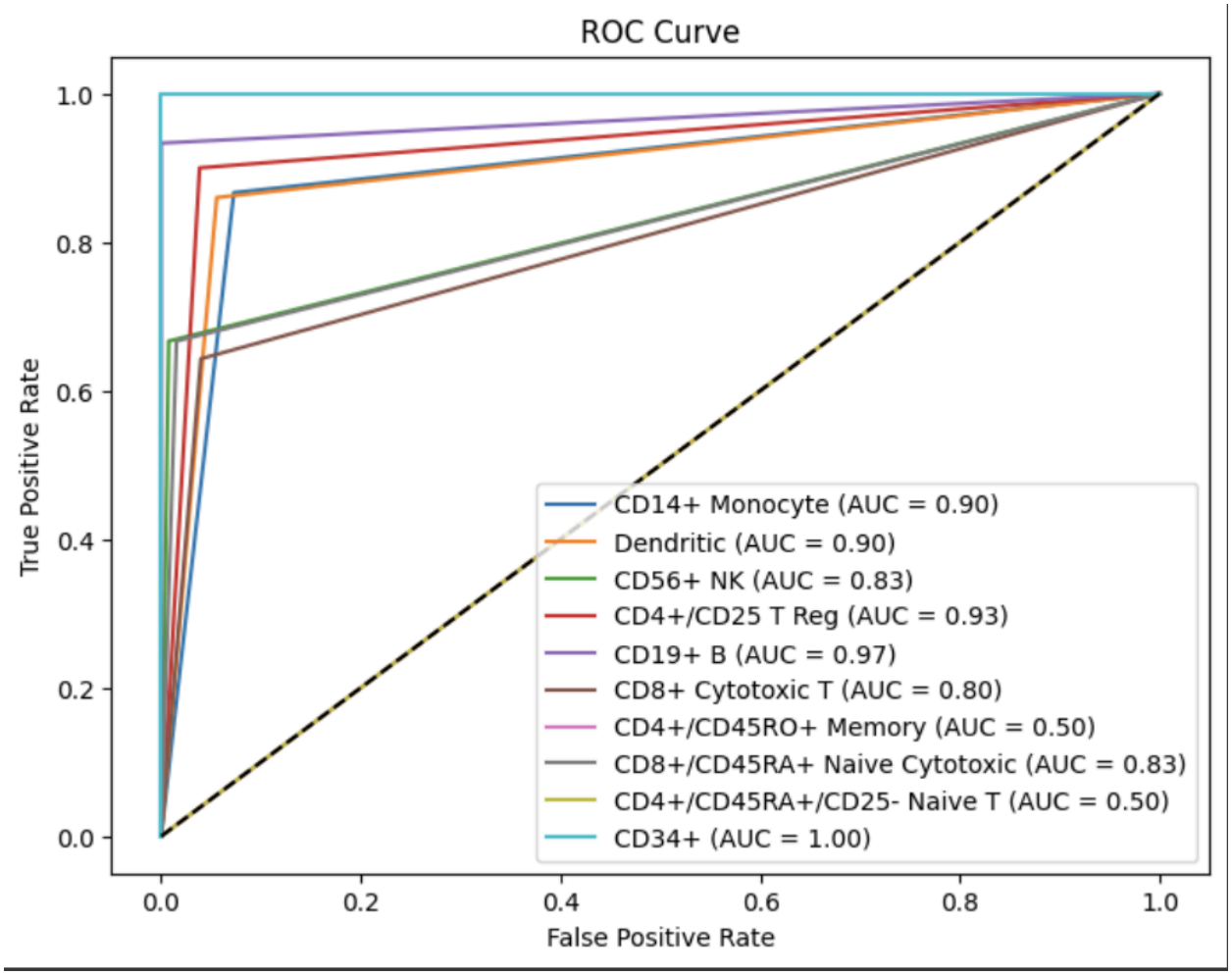
The model I built is a Stacking Classifier, a form of ensemble learning where multiple base classifiers are combined to improve predictive performance. For this project, I used four

XGBoost classifiers as base models, each having different parameters of estimators (10, 50, 90, 130) and max depth (2, 3, 4, 5). These base models are then combined using a Logistic Regression meta-model, which learns to weigh the predictions of each base model to make the final classification decision, as mentioned in discussion. To look at model evaluation, I calculated the accuracy of classification, and also printed a classification report using sklearn.metrics. My model achieves around 81% accuracy. When looking at the classification report which gives information on precision, recall, and F1-scores, the model struggles with classes such as "CD8+/CD45RA+ Naive Cytotoxic," failing to detect any instances, as indicated by a recall score of 0.00, however does better on classes such as "CD4+/CD45RA+/CD25- Naive T" and "CD34+" which have a recall score of 1.0. Below, I have shown the classification report produced from the model as well as ROC curves for each of the cell type classes to evaluate.

Model Classification Report:

	precision	recall	f1-score	support
0	0.76	0.87	0.81	30
1	0.90	0.86	0.88	50
2	0.80	0.67	0.73	6
3	0.64	0.90	0.75	10
4	1.00	0.93	0.97	15
5	0.64	0.64	0.64	14
6	1.00	0.00	0.00	1
7	0.75	0.67	0.71	9
8	1.00	0.00	0.00	2
9	1.00	1.00	1.00	3
accuracy			0.81	140
macro avg	0.85	0.65	0.65	140
weighted avg	0.83	0.81	0.81	140

ROC Curve:


ROC Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
CD14+ Monocyte (AUC = 0.90)
Dendritic (AUC = 0.90)
CD56+ NK (AUC = 0.83)
CD4+/CD25 T Reg (AUC = 0.93)
CD19+ B (AUC = 0.97)
CD8+ Cytotoxic T (AUC = 0.80)
CD4+/CD45RO+ Memory (AUC = 0.50)
CD8+/CD45RA+ Naive Cytotoxic (AUC = 0.83)
CD4+/CD45RA+/CD25- Naive T (AUC = 0.50)
CD34+ (AUC = 1.00)