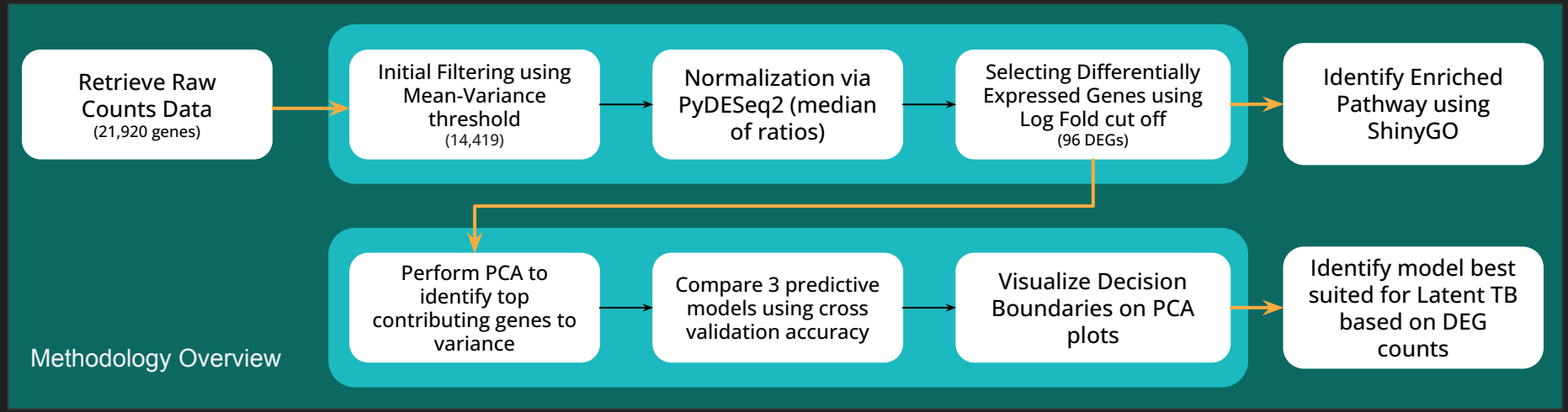
# Latent Tuberculosis Modelling

Walker Azam

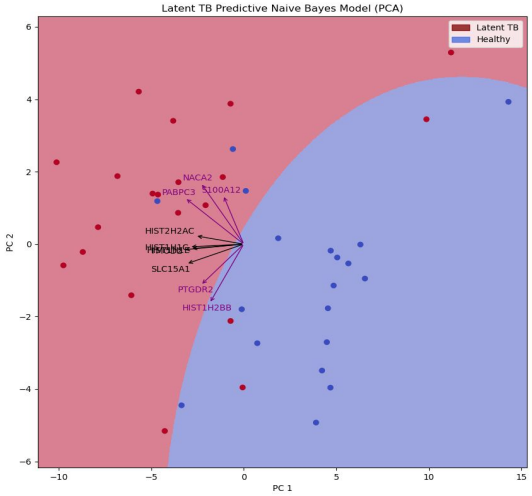
PABIO 536 - Lightning Talk

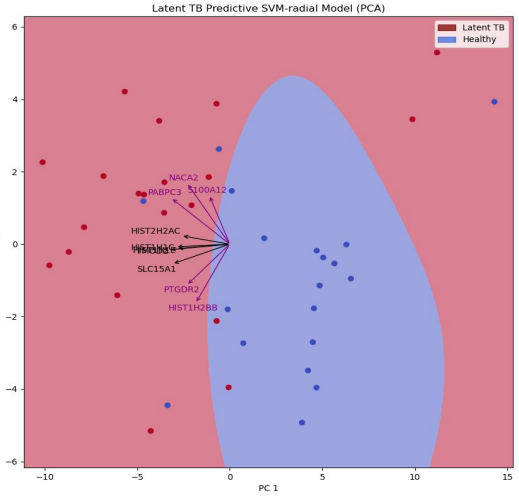
# Background

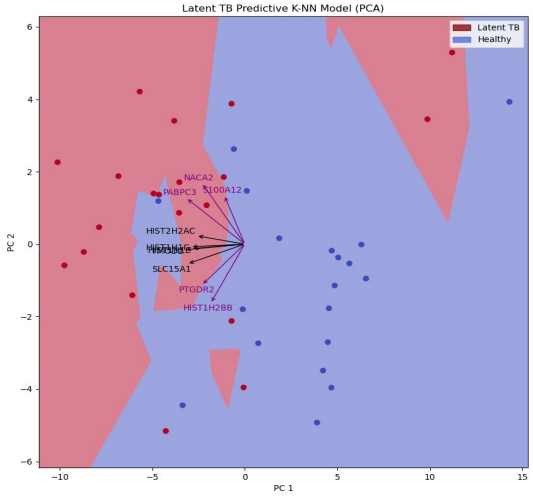
- Latent Tuberculosis remains a challenge due to difficulties in detection, screening, and effective treatment development and adherence
- Project Goals
  - Identify Differentially Expressed Genes associated with Latent TB in Human CD4 T-Cells
  - Compare different predictive models to better understand Latent TB modelling within samples
- Dataset: GEO dataset for Latent Tuberculosis CD4 T-Cell response within Human (GSE99373)
  - Contained raw counts for 39 samples (19 Healthy, 20 with Latent TB) and 21920 genes



# PCA Plots with Model Decision Boundaries



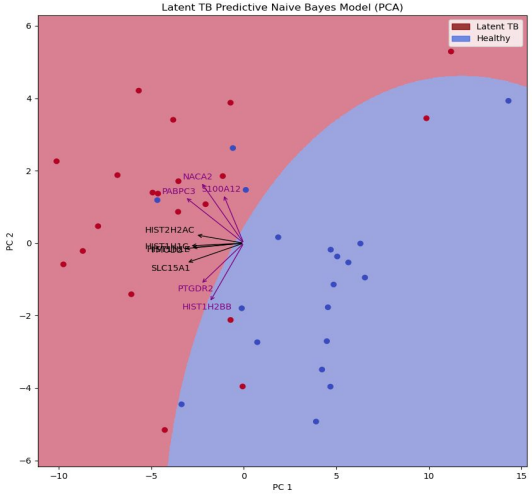



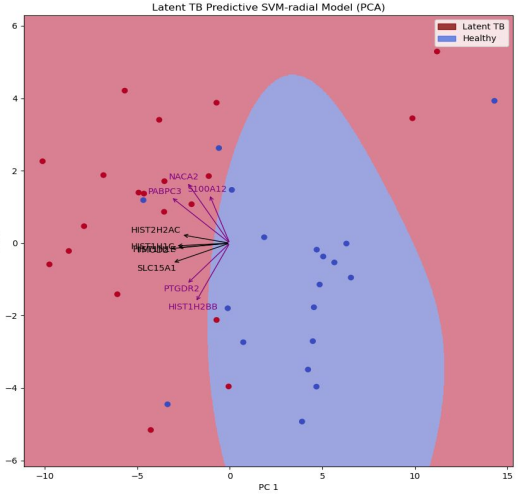



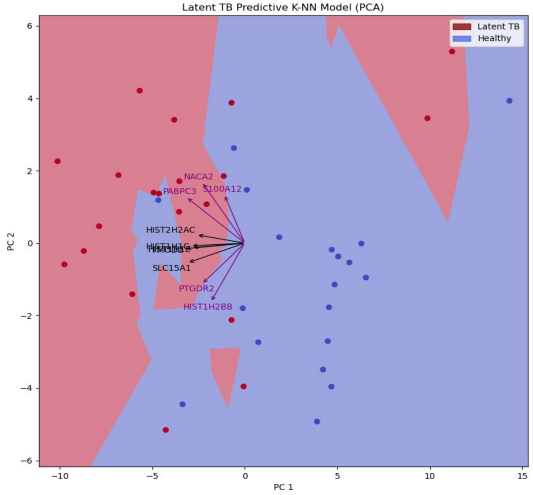

- Plots of PC1 vs PC2 explain ~42% total variance, with top 5 genes driving variance per PC overlaid in arrows (PC1: black, PC2: purple)

What Model Performed Best?

# PCA Plots with Model Decision Boundaries




- Plots of PC1 vs PC2 explain ~42% total variance, with top 5 genes driving variance per PC overlaid in arrows (PC1: black, PC2: purple)
- SVM model performed the best (82%), whereas K-NN performed weakest (62%). Naive Bayes performed around 73% accuracy. Predicting using only PC 1 and PC 2 saw only slight decrease in predictive accuracy, but no major changes in inter-model performances
- SVM's decision boundary 'trusts' the training data least stringent, allowing for better prediction

# Implications and Limitations

- The identified Differentially Expressed Genes serve as a decently strong feature set for predictive modelling
  - Distance based models sometimes may not pick up true signals
- Could be used to cut down on testing/screening for Latent TB
- Top genes associated with each PC:
  - PC1: SLC15A1, TMOD2, and HIST1H1E, HIST1H1C, HIST2H2AC
  - PC2: PTGDR2, PABPC3, S100A12, NACA2, HIST1H2BB.
- Limitations
  - Small, ‘wide’, dataset → limits model variations, reflects a lot of actual limits in human studies
  - More to compare common algorithms rather than create a robust predictive model

Thank You!

Any Questions?

# Appendix - Model Accuracies

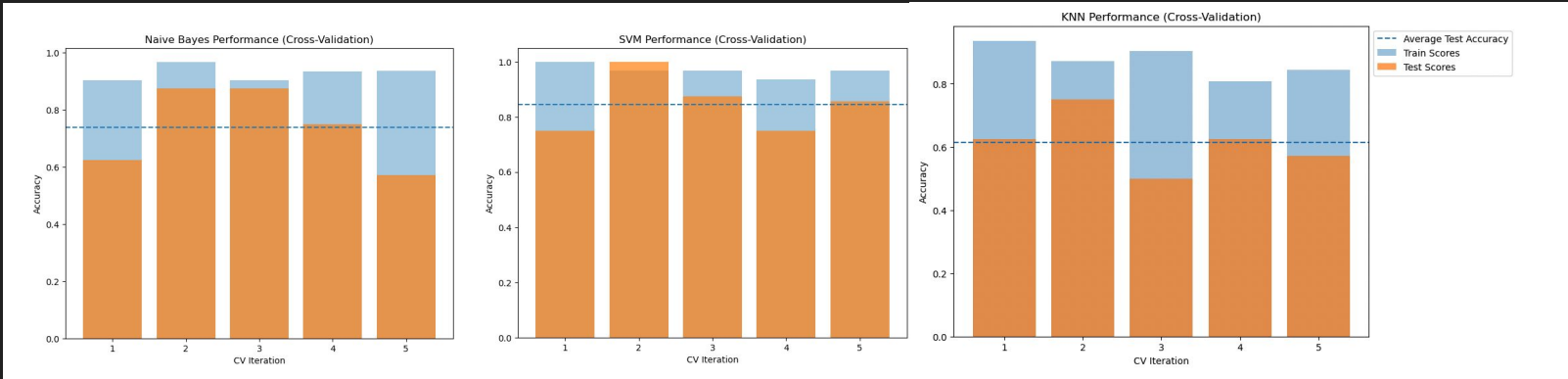



Figure 2: Cross Validation Accuracies Across 3 Models

Each model was compared using a 5-fold cross validation across the dataset of 39 samples and 96 genes. SVM used a radial kernel and K-NN used 2 neighbors. SVM performed the best at an average of 84% accuracy, whereas K-NN performed the worst at 62%. K-NN shows most evidence of overtraining as indicated by large difference in testing and training scores