# Tumor Cell Prediction at the Single-Cell Level in Pediatric High-Grade Gliomas

**PURM, SUMMER 2022**

**Saraswati V Sridhar (ENGINEERING 2025), Dr. Wenbao Yu (Assistant Professor of Pediatrics, Penn Medicine and CHOP), Dr. Kai Tan (Assistant Professor of Pediatrics, Penn Medicine and CHOP)**

## Background

Pediatric high-grade gliomas, which begin in the glial cells of the Central Nervous System, account for 8-12% of all childhood brain tumors and can be extremely difficult to treat, given their propensity to metastasize quickly, invade brain tissue, place pressure on nearby tissue, and cause intracranial pressure and hydrocephalus. These tumors often cause symptoms such as headaches, seizures, changes in mental function, mood swings, sensory and speech changes, loss of balance and coordination, and changes in pulse and breathing rates. Causes of pediatric high-grade gliomas are not clear, but potential causes include genetic predisposition and exposure to chemicals and radiation. Two major types of high-grade gliomas are diffuse midline gliomas (typically appear as a mass in the middle of the brain, and most commonly appear in the pons, thalamus, spinal cord, and cerebellum) and glioblastomas (typically originate in astrocytes, which provide structural support and energy storage). One major challenge to understanding the complex molecular interactions that underpin these tumors is the vast heterogeneity of the cells. Single-cell RNA sequencing (scRNA-seq) enables this challenge to be overcome by allowing massively parallel profiling of thousands of cells at the single-cell level, resulting in greater resolution of gene expression across a tumor. This can then be combined with machine learning algorithms for a wide range of analytical purposes, including tumor classification.

## Objectives

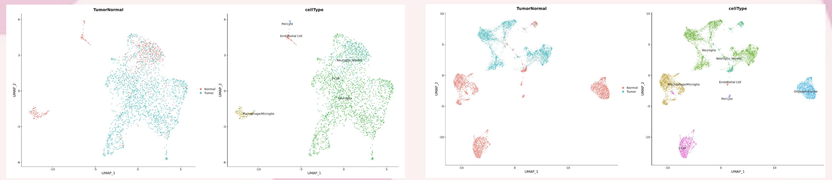
This project aims to develop, implement and evaluate various classification algorithms to classify glioma cells using a computational pipeline based on single-cell RNA data.

- Linear Regression
- Linear Discriminant Analysis
- Support Vector Machine
- Random Foresting
- k-Nearest Neighbor

## Materials and Methods

- Seurat utilized for all data storage and processing
- Algorithms: Linear Regression (Garnett), Linear Discriminant Analysis (scID), Support Vector Machine (scPred), Random Foresting (singleCellNet), and k-Nearest Neighbor (scmapcell)
- Dimensionality reduction conducted with Principal Components Analysis (PCA) and noise reduction conducted with SCT
- Training occurred with both public domain (4 datasets from Gene Expression Omnibus) and lab-generated DMG and GB datasets (10-fold cross validation) and public domain lung and colorectal cancer datasets (5-fold cross validation)
- Predictions visualized using UMAP (Uniform Manifold Clustering and Approximation

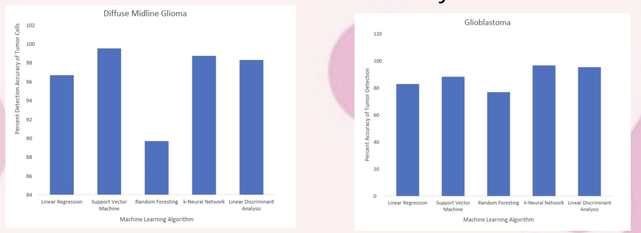
## Results



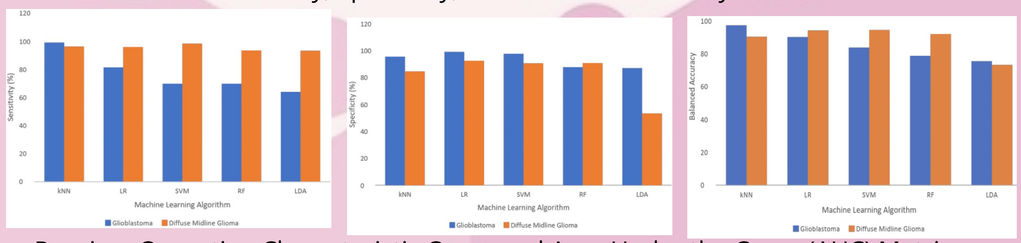

Ground-Truth UMAP Plots for Diffuse Midline Glioma (left) and Glioblastoma (right)

**Metrics for Algorithms trained with Glioma scRNA Data**

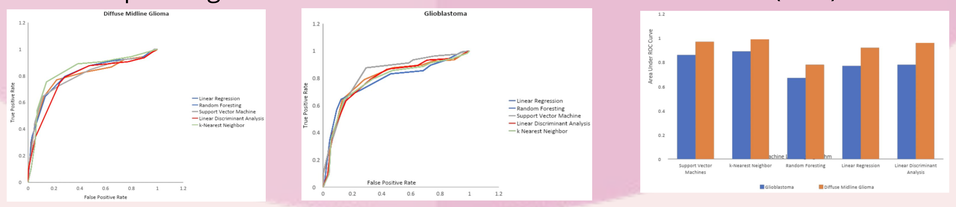
Tumor Detection Accuracy




Sensitivity, Specificity, and Balanced Accuracy




Receiver Operating Characteristic Curve and Area Under the Curve (AUC) Metrics




## Conclusions

- All algorithms correctly assigned the majority of tumor cells to the neuroglia cell type and were generally most accurate when trained with open-source datasets while least accurate when trained with non-glioma datasets, demonstrating low generalizability to non-glioma tumors
- Most common distinguishing gene signatures (features) between tumor and non-tumor cells were AC1, MES2, OPC4, and NPC3
- kNN and SVM algorithms typically had the greatest accuracy with DMG and GB classification, respectively, although the LR and RF algorithms showed the greatest accuracy when trained with non-glioma datasets
- Algorithms trained with lab-generated glioma datasets showed the greatest sensitivity, specificity, and balanced accuracy
- Algorithms trained with open-source glioma datasets showed the greatest AUC in their ROC curves
- Algorithms that achieved the best balanced accuracies, AUC values, sensitivities and specificities varied depending on the tumor type and training datasets utilized
- Evidence demonstrated that a combination of multiple algorithms (i.e. a “mixed approach”) is optimal for classification of high-grade glioma cells

## Future Work

- Exploration of additional classification algorithms implemented on a wider range of glioma datasets
- A deeper analysis of salient mathematical features found in these algorithms
- Optimal combinations of these features to achieve maximum classification accuracy of tumor cells

## Acknowledgements

I would like to acknowledge the following individuals for their guidance, assistance and support throughout this project, as it would not have been possible without them:

- Dr. Wenbao Yu
- Dr. Kai Tan and Dr. Kristina Cole
- Jackie Peng and Samuel Kim
- All members of the Tan Lab for Gene Regulation and all staff at CHOP Center for Childhood Cancer Research