CSCI 2470: Deep Learning
Fall 2023
Jessica Patricoski, Ananya Pavuluri, Tuan Pham (Team JAT)

Cellular Age Classification in the Hypothalamus

Introduction

The Webb Lab at Brown recently published a groundbreaking single-cell RNA-seq dataset of the female mouse hypothalamus.[1] Neurons of the hypothalamus regulate homeostatic processes and hormone production, which are dysregulated with age. Sex-specific differences in lifespan, including different hallmarks of aging, are well documented in humans and many other species. Despite findings that females are twice as likely to develop dementia, the influence of biological sex and sex hormones on brain aging is historically understudied.[1-4]

Subsequently, members of the Webb Lab have generated multi-omic data of the hypothalamus at single-cell resolution in male and female mice (currently unpublished). This new data will allow us to analyze the chromatin accessibility component of aging hallmarks of neurons in the hypothalamus alongside previously published gene expression data. Moreover, including male mice in this dataset allows us to uncover sex-specific differences in the aging hypothalamus.

Single-cell resolution ATAC-seq data consists of a very large peak by cell matrix, in which each cell is a sample and each peak is a feature. There are upwards of 250,000 features in this dataset. For this reason, we will focus on the fragments of the X chromosome (ChrX), as it is a sex chromosome that is of utmost importance in aging and the brain.[5,6] This lowers the feature number to a manageable size. The dataset will require extensive preprocessing using the Seurat and Signac libraries in R. Some bottlenecks we foresee with this dataset are the significant sparsity that comes with the

territory of single-cell data. We plan to address this issue by using the cisTopic package in Python.

In this project, we seek to implement a Multi-Layer Perceptron (MLP) pipeline in which the input is the peak vector of a cell, and the output is the likelihood of the cell being from a young or aged mouse brain (binary classification task). Due to the size of the data, the complexity, and the sparsity, Deep Learning is an ideal approach. We will split our data into training and testing groups via cross-validation. We define *model success* in terms of accuracy- as in, the % of times our model correctly classified a cell as young or aged.

A base goal for our project within our given timeframe is to have our model perform better than random classification (correct more than half the time). A target goal is to reach 70% accuracy. A stretch goal is to reach an 80% accuracy and perform hyperparameter analysis to gain biological insight based on what the model uses for classification and how hyperparameters differ in males and females.

This project will hopefully contribute to the larger topic of sex-specific differences in the aging brain, which is crucial to understand. Females have been historically underrepresented in biomedical studies; the practice of conducting studies on only male subjects and generalizing results to all sexes has had major repercussions, and we must better address sex differences in all aspects of biomedicine.

Methodology

Graduate students in the Ashley Webb Lab at Brown performed single-cell ATAC-seq of the hypothalamus in male and female mice at 3 different age points: Young (6 months), Mid-Age (12 months), and Aged (18+ months) (Figure 1). We preprocessed these data by removing Mid-Age mice to begin with a simpler binary classification task, excluding all brain cell types other than neurons, culling the peaks from the X chromosome, and filtering out low-resolution cells. We then analyzed the differentially accessible peaks between Young and Aged female mice (Figure 2). We found 39 differentially accessible peaks on ChrX. The resulting dataset had 11420 female neurons and 12639 male neurons, with 39 DA peaks as features. All preprocessing was done in R using the Seurat and Signac libraries.[7]

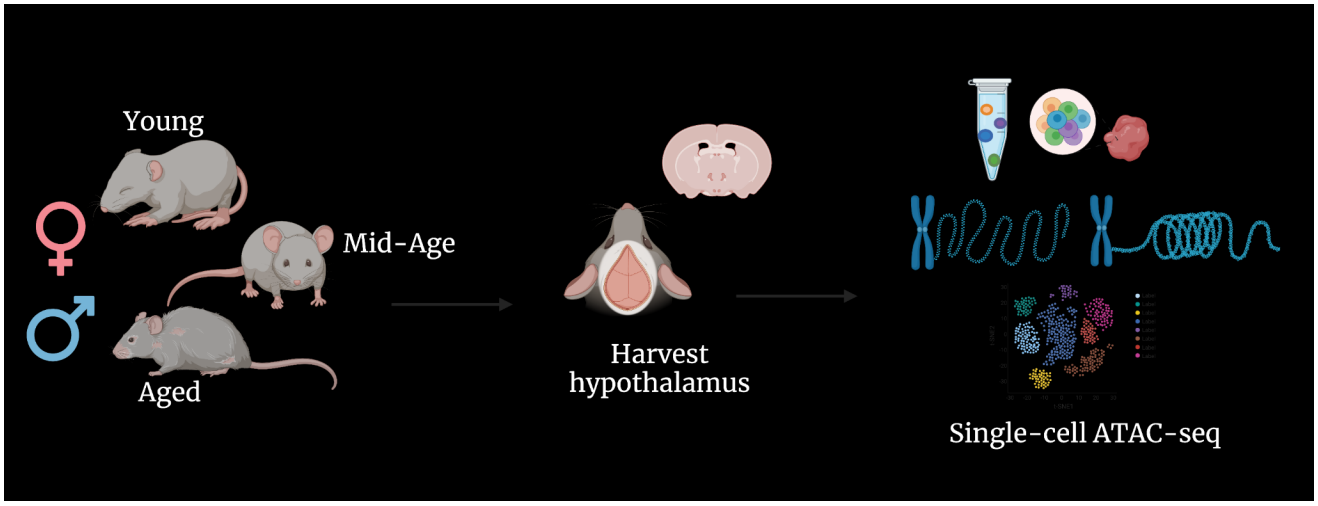

Figure 1. Overview of whole hypothalamus single-cell ATAC-seq experiment.

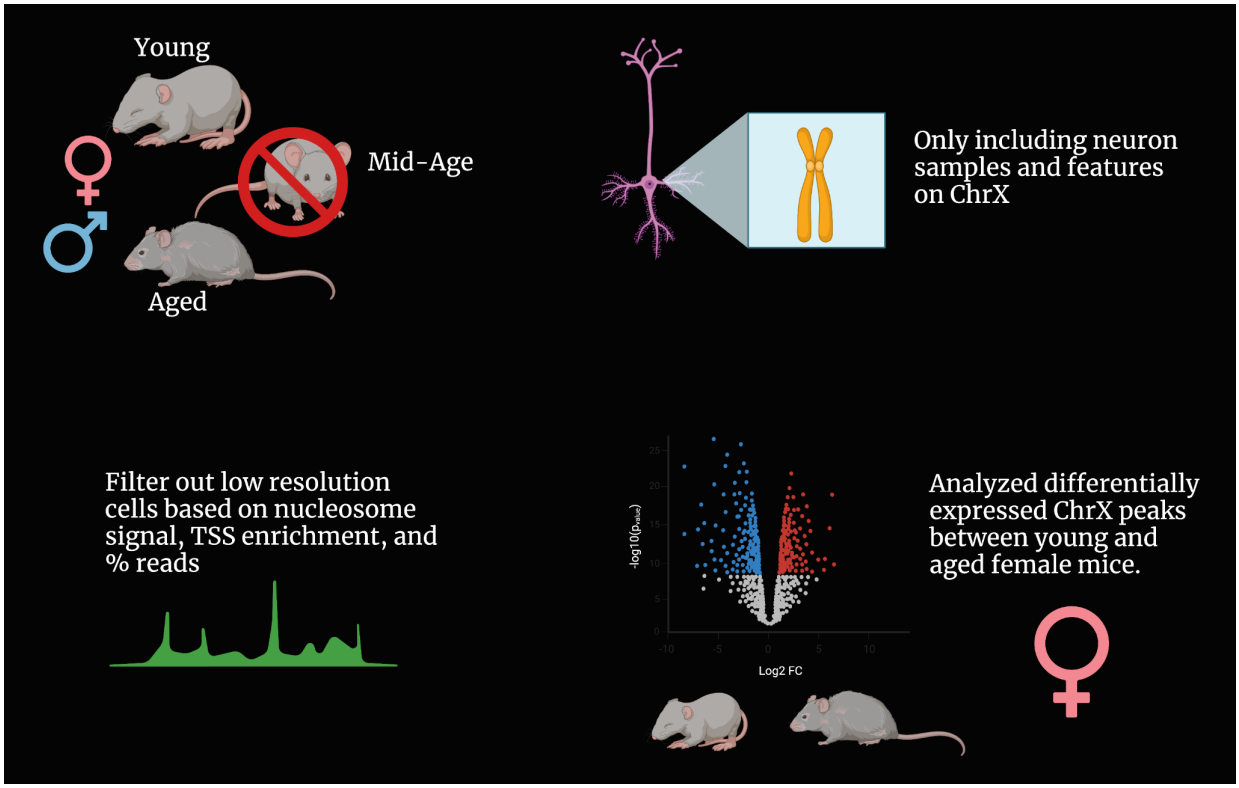

Figure 2. Preprocessing and feature engineering overview.

Our MLP network consisted of 2 hidden layers, each with Leaky ReLU activation, 2 Dropout layers with a 50% Dropout rate to prevent overfitting, and an output layer with sigmoid activation. The model was constructed using the Tensorflow library in Python.[8]

Female data was shuffled randomly to ensure broad training samples and split into 50% training and 50% testing data. Our model was trained only on female samples but tested on female and male samples separately to assess whether a model trained on female data could also be generalized to male data. We hypothesized that our model would perform poorly on male data, given that aging is a sex-specific process (Figure 3).

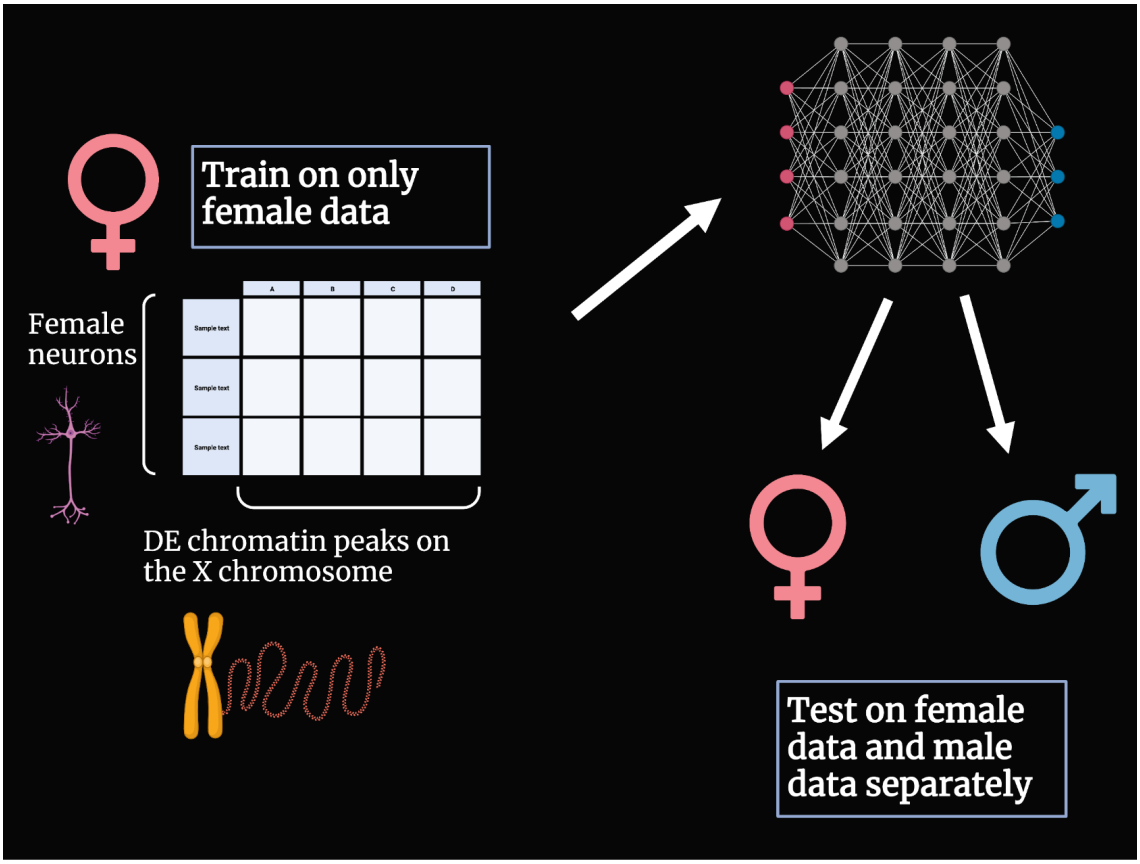

Figure 3. Overview of training and testing pipeline of the MLP model. Training data are represented by a cell-by-peak matrix of female neurons. The female dataset is split into 50% training and 50% testing subsets, and the model is tested on both male and female data.

Results

The model performed above baseline (50% accuracy, indicating random performance) for the Young/Aged binary classification on female testing data. However, precision, recall, and F1 score metrics indicate a significant skewness toward Young predictions (Table 1). This skewness is also displayed in the confusion matrix (Figure 4).

Interestingly, the accuracy was significantly lower when testing on male data. However, the model assigned the Young classification to all samples, with zero samples classified as Aged, resulting in a precision, recall, and F1 Score of zero (Table 1).

	Training Accuracy	Testing Accuracy	Precision	Recall	F1 Score
♀ **Females**	65.43%	64.72%	0.0056	4.655e-04	8.595e-04
♂ **Males**	N/A	35.80%	0.000**	0.000**	0.000**

Table 1. Comparison of metrics across both sexes. The model performs above the baseline of 50% testing accuracy (random behavior) on female data. Analysis revealed that a model trained on only female data results in very low accuracy when tested on male data. **These metrics are due to the model classifying all male samples as Young and no samples as Aged.

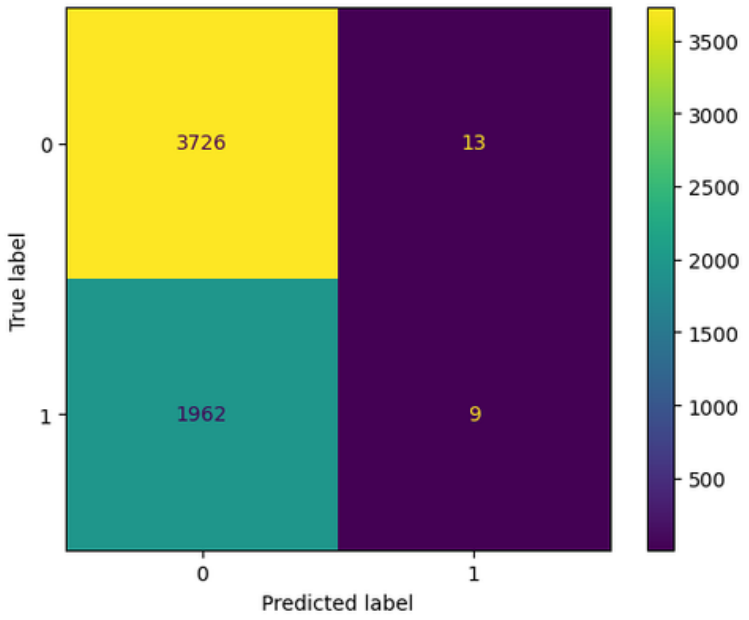

Figure 4. Confusion matrix for female testing data. Zero represents Young, and one represents Aged.

Discussion

Challenges

Dataset-related bottlenecks

Working with a large dataset is challenging. The raw data for this project was in a 10 GB file, which is a 64000+ by 271000+ sparse matrix. Even upon subsetting and performing feature selection, our data was still very spare. Moreover, our data had significantly more Young mice than Aged mice, contributing to the skewness of the classifications. Although one cannot always expect equal representation in training data, especially in datasets that take significant labor to generate, this is a hurdle we have yet to overcome. One possible solution is to downsample the number of Young mice, so there is an equal representation of Young and Aged mice during training. Moreover, dimensionality reduction analysis via UMAP showed no clear boundaries between Aged and Young female neurons, as we expected separable clusters representing each class but did not observe this (Figure 5).

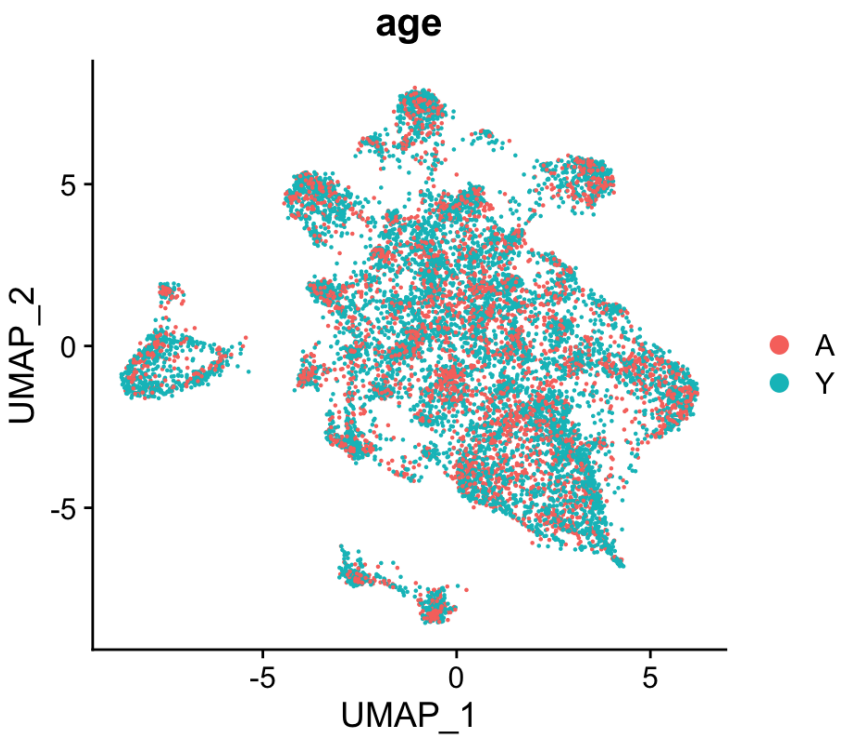

Figure 5. UMAP plot of female neurons clustered based on age reveals no clear boundaries between Aged vs. Young cells.

Overfitting

At times throughout testing our model, we noticed significant overfitting. To combat this, we had to determine the optimal combination of Dropout layers/rates and regularization methods via trial-and-error that would produce a decent accuracy with no overfitting.

Reflection

Base, target, and stretch goals

Our model performed above our base goal of 50% testing accuracy on female testing data, with an accuracy of ~65%. However, we did not meet our target goal of a 70% accuracy. Given that single-cell ATAC-seq is a technique that has recently gained

popularity, working with such data requires thorough domain knowledge for feature engineering (such as which genes are behind certain peaks or which chromosomal regions are involved in natural aging processes) and that the questions we are addressing are novel and underexplored, such bottlenecks are expected.

Future directions

Many steps can be taken to improve the accuracy of our model. First, using differentially accessible peaks across all chromosomes and not just ChrX may lead to a better outcome. Furthermore, we must do more thorough investigations of the differentially accessible peaks. It will be interesting to see whether the majority of DA peaks are upregulated in Young cells, as this would provide a biological basis for the skewness of our classifications. We must also explore the genes associated with these peaks to determine which ones are biologically relevant to age. In the long run, we hope to incorporate single-cell RNA-seq into the pipeline. We expect a multimodal dataset with gene expression and chromatin accessibility profiles to improve model accuracy significantly. We also hope to explore other types of networks and algorithms, such as Support Vector Machine and XGBoost, in solving this problem.

Takeaways

This project taught us that increasing model complexity is not always the correct path. After removing hidden layers and reducing the number of neurons in each layer, our accuracy increased significantly. Moreover, this project taught us to prioritize explainability, especially for tasks requiring a thorough understanding of features with the strongest effects on output. We must create an explainable and interpretable model to contribute valuable biological insight for our questions. Understanding which peaks most strongly influence output will allow us to search for which genes/proteins are associated with those regions and whether they are relevant for aging.

References

1. Hajdarovic KH, Yu D, Hassell LA, et al. Single-cell analysis of the aging female mouse hypothalamus. *Nat Aging*. 2022;2(7):662-678. doi:10.1038/s43587-022-00246-4
2. Hanamsagar R, Bilbo SD. Sex differences in neurodevelopmental and neurodegenerative disorders: Focus on microglial function and neuroinflammation during development. *J Steroid Biochem Mol Biol*. 2016;160:127-133. doi:10.1016/j.jsbmb.2015.09.039
3. Taylor CM, Pritschet L, Yu S, Jacobs EG. Applying a Women's Health Lens to the Study of the Aging Brain. *Front Hum Neurosci*. 2019;13:224. doi:10.3389/fnhum.2019.00224
4. 2021 Alzheimer's disease facts and figures. *Alzheimers Dement*. 2021;17(3):327-406. doi:10.1002/alz.12328
5. Roberts AL, Morea A, Amar A, et al. Age acquired skewed X chromosome inactivation is associated with adverse health outcomes in humans. *Elife*. 2022;11:e78263. Published 2022 Nov 22. doi:10.7554/eLife.78263
6. Zhao C, Gong G. Mapping the effect of the X chromosome on the human brain: Neuroimaging evidence from Turner syndrome. *Neurosci Biobehav Rev*. 2017;80:263-275. doi:10.1016/j.neubiorev.2017.05.023
7. Stuart T, Butler A, Hoffman P, et al. Comprehensive Integration of Single-Cell Data. *Cell*. 2019;177(7):1888-1902.e21. doi:10.1016/j.cell.2019.05.031
8. Abadi M, Agarwal A, Barham P, et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Available from: https://www.tensorflow.org