Development of a machine learning pipeline to analyze biological multiple particle tracking datasets

Nels Schimek

A thesis

Submitted in partial fulfillment of the

Requirements for the degree of

Master of Science

University of Washington

2022

Committee:

Elizabeth Nance

Jesse Zalatan

Program authorized to Offer Degree:

Chemistry

University of Washington

Abstract

Development of a machine learning pipeline to analyze biological multiple particle tracking datasets

Nels Schimek

Chair of the Supervisory Committee:

Elizabeth Nance

Department of Chemical Engineering

Multiple Particle Tracking (MPT) has been demonstrated as an important tool for understanding changes to biological environments. MPT studies are capable of generating gigabytes of data across hundreds to thousands of trajectories, making MPT datasets an interesting candidate for machine learning applications. To begin understanding the scope of biological questions that can be answered by coupling MPT datasets with machine learning techniques, an end-to-end data science pipeline is developed building off of recent work in the Nance Lab and applied to three unique datasets. To begin, Principal Components Analysis is applied in order to visualize the spread and distribution of the high dimensional MPT data. Next, a boosted decision tree model, XGBoost, is applied to determine the predictable capability of each dataset, and SHAP values are used to understand model predictions and find the statistical feature driving accurate predictions. Finally, XGBoost models are trained on trajectories from specific diffusion modes to determine any increase in accuracy. Overall, the pipeline presented demonstrates the capability to provide information across multiple biological questions.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank Dr. Elizabeth Nance for the opportunity to conduct research in the Nance lab, and for her support during the past two years. I applied to graduate school develop skills in combining computational and data science tools with neuroscience research, and I am incredibly grateful for the guidance Dr. Nance has giving me while pursuing my graduate education as it has allowed to grow tremendously as a researcher. I would like to thank Dr. David Beck for meeting with me regularly to discuss data science and computational methods and results, as this meeting have helped me strengthen my technical skills and understanding. Finally, I would like to thank Dr. Jesse Zalatan, for reading my thesis and giving insights on how to push future research aims in the Nance lab.

I would like to thank the Nance Lab community for welcoming me into the lab. Even though I worked remotely for the first year of graduate school I still felt just as much as part of the lab as I did when I began working in person. I would like to the Mike McKenna, Hawley Helmbrecht, and Brendan Butler for helping me transition into the research done in the Nance Lab, and for their collaboration and expertise.

I would also like to thank my parents and my sister for their support through my education. I appreciate their willingness to listen to me talk about the ups and downs of graduate school and research, and to give me thoughtful advice.

1. Introduction

The extracellular matrix (ECM) of the brain is a collection of tenascin, proteoglycan, and hyaluronan that exists in the extracellular space (ECS) of brain tissue[1]. Proper structure of the ECM is crucial for normal physiology and functioning of the brain, for example with tissue structural integrity and signaling pathways[2]. However, the structure of the ECM can be impacted by neurological diseases like Alzheimer's and multiple sclerosis[3], making it crucial to develop techniques able to detect structural changes in the ECM.

While probing real-time changes of the ECM microstructure as a result of these diseases is an unsolved challenge, multiple particle tracking (MPT) has gained popularity as a method capable of probing the ECM in real time with nanoscale resolution[4]. Multiple particle tracking is a microscopy technique able to track the motion of thousands of individual nanoparticles in sample while having the resolution to capture the individual trajectories of each nanoparticle[5, 6]. One of the major benefits of MPT is its ability to probe changes to diffusion, rheology, and composition of a wide variety of biological environments at the nanoscale including the vitreous of the eye[7], intestinal mucus[8], viral vectors in the lung[9], and tumor cell migration[10]. MPT has seen extensive use in studies aimed at understanding intracellular trafficking[11] including transport of gene nanocarriers[12, 13] and polymer nanoparticles[14, 15].

One of the most significant applications of MPT has been to probe the brain microenvironment. MPT was used to show that nanoparticles up to a diameter of 114nm could diffuse through brain tissue[16], revealing crucial information about the maximum pore size and pore size distribution and doubling the size of nanoparticles that could be used to penetrate the brain. MPT has also been used to characterize the change in nanoparticle diffusion in the brain

extracellular space (ECS) as a result of oxygen-glucose deprivation [17], as well as changes to diffusion and pore size in the brain ECS due to neural development[18].

Because of the size of datasets generates through MPT studies, the utilization of machine learning tools presents a way to maximize the information extracted from MPT datasets. Two of the most commonly used types of machine learning are supervised learning and unsupervised learning. Supervised learning is used to make an accurate prediction by learning the relationship between one or many inputs to a single output, where the predicted output can be a quantitative value (referred to as regression) or a qualitative value (referred to as classification)[19]. Unsupervised learning is used find structure or relationships between inputs without having access to an output[19]. Unsupervised learning algorithms are commonly used for tasks such as clustering analysis and dimensionality reduction.

Machine learning applications for nanoparticle tracking data have primarily focused on supervised classification tasks. Machine learning has been used to predict the motion type of nanoparticles[20-22], primarily using neural networks and random forests. Neural networks have also been used to predict biological variables from MPT data, including agarose gel stiffness and *in vitro* cell uptake. Most recently XGBoost, a popular boosted decision tree model, was used to predict the biological age of rodents from MPT data[18].

2. Methods

2.1. MPT Data Collection

Nanoparticle diffusion data was previously obtained using multiple particle tracking (MPT) in rat brain slices from three independent experiments: (1) age-dependent diffusion data from brain slices collected from male pups at postnatal (P) day 14, P21, P28, P35 and P70[18] defined as the age dataset; (2) brain-region dependent diffusion data from brain slices from P10 male pups

that included the basal ganglia, cortex, thalamus, striatum, and hippocampus, defined as the region dataset; and (3) diffusion data from P35 male pup brain slices exposed to enzymatic degradation of ECM structures, including chondroitinase (ChABC)-treated, hyaluronidase (HYase)-treated, and non-treated (NT) control slices, defined as the treatment dataset[18].

2.2. Feature generation via diff_classifier

In order to generate features for machine learning, an open-source package called *diff_classifier*[23] was used. *Diff_classifier* is a package developed in the Nance lab to generate statistical features for each trajectory. *Diff_classifier* also calculates locally averaged statistical features for each trajectory. In total, 33 statistical features were generated for each trajectory.

2.3. Principal Components analysis

Principal components analysis (PCA) was applied to each dataset to reduce the number of features from 33 to two to visualize the spread of data and overlap of classes. PCA is an unsupervised learning technique where linear combinations (principal components) of features are calculated, and ranked based on the amount of variance captured[19, 24]. The dataset is then transformed from a high-dimensional feature space into a low-dimensional features space, and the principal components capturing the most variance can be plotted in order to visualize high dimensional data. Trajectories with missing or infinite values were removed and the remaining data was scaled prior to application of the PCA algorithm by subtracting the mean value of each feature and dividing by the standard deviation, using scikit-learn's StandardScaler class[25].

2.4. XGBoost pipeline

The XGBoost software package, which uses a boosted decision tree algorithm, generated classification predictions on each of the three datasets. A boosted decision tree model is built by sequentially adding weak prediction models fit to a subset of the full data set, continuing until a

specific number of weak models is reached or the loss function of the model converges. Final predictions of the model are made through a weighted average of the predictions of each of the weak learners[24]. Data cleaning was done by removing trajectories that had missing or infinite values for any of the features. To prevent overfitting, each dataset was rebalanced through under sampling to ensure each class had the same number of data points. Training, testing, and validation datasets were randomly chosen at a split percentage of 80%/10%/10%. Feature data consisted of the statistical features generated with *diff_classifier*, while the target variable was either age, region, or treatment type. A cross-validation hyperparameter search was used to determine the optimal hyperparameters of an XGBoost model for each of the three different datasets. In order to ensure correct predictions by an XGBoost model were not due to chance, a Y-scrambling method was incorporated where the target variables are randomly reassigned for each feature row, following the procedure published by Li et al[26].

2.5. SHAP values

SHAP (SHapley Additive exPlanations) values were used to determine the importance of each feature to the model's predictive ability. SHAP values are based off the game theory idea of Shapley values[27], and can be used to understand the predictions made by a machine learning model. Specifically, the SHAP software package computing the importance of each feature in the dataset by calculating how much each feature contributes to each prediction instance and averaging the values for each feature[28].

2.6. Diffusion Mode splitting

Each dataset was split into three subsets based on the diffusion mode of each trajectory. The diffusion mode was determined by the alpha feature value (α) from the anomalous diffusion equation. Superdiffusive motion was classified as an $\alpha > 1.1$, Brownian motion at an α between

0.9 and 1.1 inclusively, and subdiffusive was an $\alpha < 0.9$[20, 29]. For each dataset, an XGBoost model using generic hyperparameters was trained 50 times on each data subset and the accuracy was recorded to determine the predictive capability of each diffusion mode compared to training an XGBoost model on each full dataset 50 times. To determine the diffusion mode with the highest predictive power, the distribution of the accuracies of these 50 models was plotted using an empirical cumulative distribution function (ECDF), where the *x*-axis is the accuracy of a given model and the *y*-axis shows the cumulative percentage of data points that are less than that data point. SHAP values were then calculated for the best performing diffusion mode.

3. Thesis Project: Assessment of data science and machine learning pipeline across multiple MPT datasets

3.1. Aims

The aim of this project was to develop a pipeline for applying machine learning tools to MPT datasets and determine its efficacy on three different MPT datasets. Each dataset has a unique target biological variable to predict, which is referred to as a class. The datasets used include the dataset originally used by McKenna et al where the classes are age groups [18], a dataset where the classes are five different brain regions, and a dataset from enzyme induced extracellular matrix (ECM) breakdown experiments from McKenna et al[18]. We first use Principal Component Analysis (PCA) for exploratory data analysis (EDA). PCA can be used to visualize datasets with many features, like MPT datasets, in two dimensions, to see the spread of datapoints from different classes. Datasets with significant overlap between datapoints when plotted in these first two dimensions may be harder to generate accurate predictions for when compared to datasets where classes are separated, so this EDA approach helps quickly determine how successful a

machine learning model might be when making predictions. If EDA shows any separation between the data from distinct classes, as opposed to significant overlap in data points between many or all classes, we apply XGBoost (eXtreme Gradient Boosting)[30], a boosted decision tree model, to train on the data and generate the predictions. Once an accurate model is trained, the Shapley Additive exPlanations [27, 31] package finds the features that most significantly drive model accuracy, which can provide insight into the biological changes causing differences between the classes. To ensure that predictions and SHAP values reflect the model learning real underlying patterns in the data, we apply a Y-scrambling approach to determine how the model would perform if the data was random. Lastly, domain-specific knowledge of the data can be leveraged to further improve model accuracy and gain increased insight into the data. We demonstrate how knowledge of the diffusion mode of a trajectory can be utilized to improve the accuracy of XGBoost model predictions on these datasets. We separate the data into subsets based on diffusion mode and report the effect on results of the XGBoost and SHAP techniques.

3.2. Results

PCA analysis of the age dataset showed significant overlap of datapoints in the P21, P28, and P35 age groups (**Figure 3-1A)**. PCA visualization of only P14, P35, and P70 data showed less overlap between the classes (**Figure 3-1B**), indicating more feasibility of supervised learning for the three classes as opposed to all five classes. Similar trends were observed with the region dataset, where separation occurred between data points from the striatum and cortex, but there was significant overlap between trajectories from the hippocampus, thalamus, and basal ganglia (**Figure 3-1C**). PCA on trajectory data from the striatum, cortex, and hippocampus showed separation between all three classes (**Figure 3-1D**). PCA on the treatment dataset showed significant overlap of NT, HYase, and ChABC groups (**Figure 3-1E**). The overlap was

somewhat reduced by applying PCA to only the NT and ChABC groups (**Figure 3-1F**), however the amount of remaining overlap indicates the dataset will have less predictive power compared to the age and region datasets.

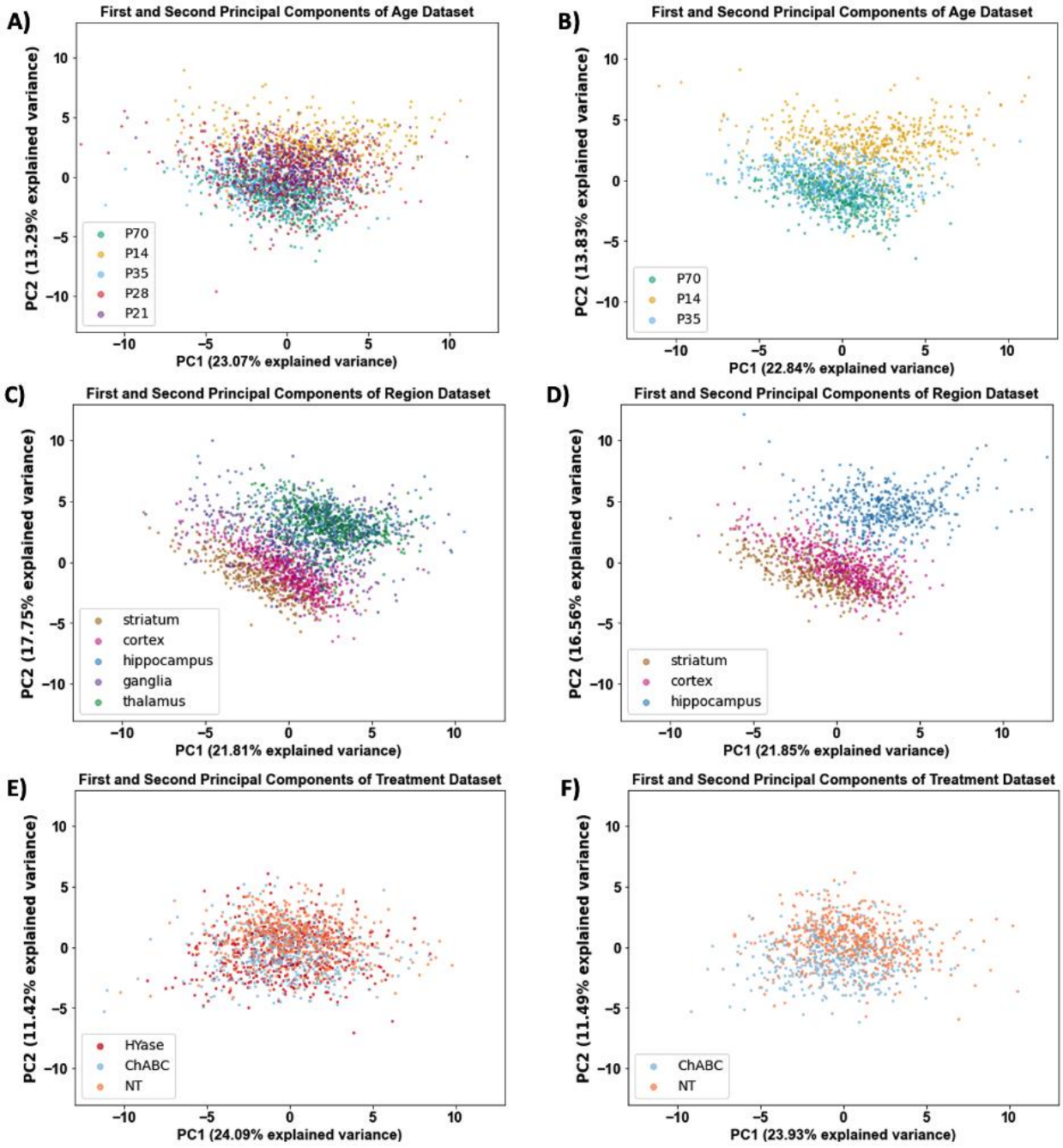

Figure 3-1. PCA results for age, region, and treatment datasets. (A) PCA analysis of P14, P35, and P70 classes from the age dataset showed overlap between P35 and P70 data points and

separation of P14 datapoints. (B) PCA analysis of all five classes from the age dataset showed minimal distinction between classes. (C) PCA analysis of striatum, cortex, and hippocampus classes from the region data set showed separation between classes. (D) PCA analysis of all five classes from the region dataset showed separation of data points from the region and striatum classes, and overlap of datapoints from the hippocampus, ganglia, and thalamus. (E) PCA analysis of the ChABC treated and non-treated classes from the treatment dataset showed overlap between the classes. (F) PCA analysis of all three classes from the treatment dataset showed overlap between all three classes. For all datasets, 13 principal components were needed to reach 85% of the explained variance.

To establish the likelihood of correct predictions occurring by chance, XGBoost models were trained on Y-scrambled data for each dataset and had accuracy scores close to that of random guessing (**Tables 3-1,3-2,3-3**). Analysis of precision, recall, and F1 score metrics of the Y-scrambled models further show the lack of any predictive capacity for these models. Predictions on the age dataset and region dataset with the normal XGBoost model are higher than predictions with the Y-scrambled XGBoost model (+0.52 and +0.56 respectively), and predictions using normal XGBoost of the treatment dataset were higher by 0.19 compared to the Y-scrambled XGBoost model. These results align with the PCA results, where overlap between classes was much higher for the treatment data compared to the age or region data. Predictions using other classes that were removed after PCA analysis, either as replacements or in addition to the classes used in Tables 1, 2, and 3, generally led to decreases in predictive capacity except for the thalamus, which performed similarly to the hippocampus. A confusion matrix of a model trained on all five regions showed that trajectories from the thalamus were almost equally likely to be predicted as hippocampus as they were thalamus, which is one possible reason that interchanging hippocampus for thalamus lead to similar model accuracy (**Figure 3-2**). Hippocampus was chosen over thalamus for the final model because it was predicted correctly at a higher rate when a model was trained on all five regions.

Table 3-1. XGBoost metrics for each class in the age dataset, and for the overall model, reported as median and interquartile range from 50 models trained on random, down sampled subsets of data

Evaluation	P14	P35	P70	Total
Normal model				
Accuracy	0.88 ± 0.02	0.80 ± 0.03	0.90 ± 0.04	0.86 ± 0.02
Precision	0.91 ± 0.03	0.79 ± 0.03	0.89 ± 0.02	0.86 ± 0.01
Recall	0.88 ± 0.02	0.80 ± 0.03	0.90 ± 0.04	0.86 ± 0.01
F1 Score	0.89 ± 0.02	0.80 ± 0.03	0.89 ± 0.02	0.86 ± 0.02
Y-Scrambled model				
Accuracy	0.34 ± 0.02	0.34 ± 0.04	0.33 ± 0.02	0.34 ± 0.01
Precision	0.33 ± 0.02	0.34 ± 0.02	0.34 ± 0.02	0.34 ± 0.01
Recall	0.34 ± 0.02	0.34 ± 0.02	0.33 ± 0.02	0.34 ± 0.01
F1 Score	0.34 ± 0.02	0.34 ± 0.02	0.34 ± 0.02	0.34 ± 0.01

Table 3-2. XGBoost metrics for each class in the region dataset, and for the overall model, reported as median and interquartile range from 50 models trained on random, down sampled subsets of data

Evaluation	Cortex	Hippocampus	Striatum	Total
Normal model				
Accuracy	0.88 ± 0.08	0.96 ± 0.01	0.86 ± 0.02	0.90 ± 0.03
Precision	0.85 ± 0.04	0.98 ± 0.01	0.86 ± 0.04	0.90 ± 0.03
Recall	0.88 ± 0.08	0.96 ± 0.01	0.86 ± 0.02	0.90 ± 0.03
F1 Score	0.86 ± 0.06	0.97 ± 0.01	0.86 ± 0.02	0.90 ± 0.03
Y-Scrambled model				
Accuracy	0.34 ± 0.02	0.33 ± 0.02	0.33 ± 0.02	0.34 ± 0.01
Precision	0.33 ± 0.02	0.34 ± 0.02	0.33 ± 0.02	0.34 ± 0.01
Recall	0.33 ± 0.02	0.34 ± 0.02	0.34 ± 0.02	0.34 ± 0.01
F1 Score	0.34 ± 0.02	0.33 ± 0.02	0.34 ± 0.02	0.34 ± 0.01

Table 3-3. XGBoost metrics for each class in the treatment dataset, and for the overall model, reported as median and interquartile range from 50 models trained on random, down sampled subsets of data

Evaluation	ChABC Treated	Non-Treated	Total
Normal Model			
Accuracy	0.66 ± 0.02	0.72 ± 0.03	0.69 ± 0.02
Precision	0.70 ± 0.03	0.68 ± 0.02	0.69 ± 0.02
Recall	0.67 ± 0.02	0.72 ± 0.03	0.69 ± 0.02
F1 Score	0.69 ± 0.02	0.70 ± 0.02	0.69 ± 0.02
Y-scrambled Model			
Accuracy	0.48 ± 0.01	0.52 ± 0.03	0.50 ± 0.01
Precision	0.50 ± 0.01	0.50 ± 0.01	0.50 ± 0.01
Recall	0.48 ± 0.01	0.52 ± 0.03	0.50 ± 0.01
F1 Score	0.49 ± 0.01	0.51 ± 0.02	0.50 ± 0.01

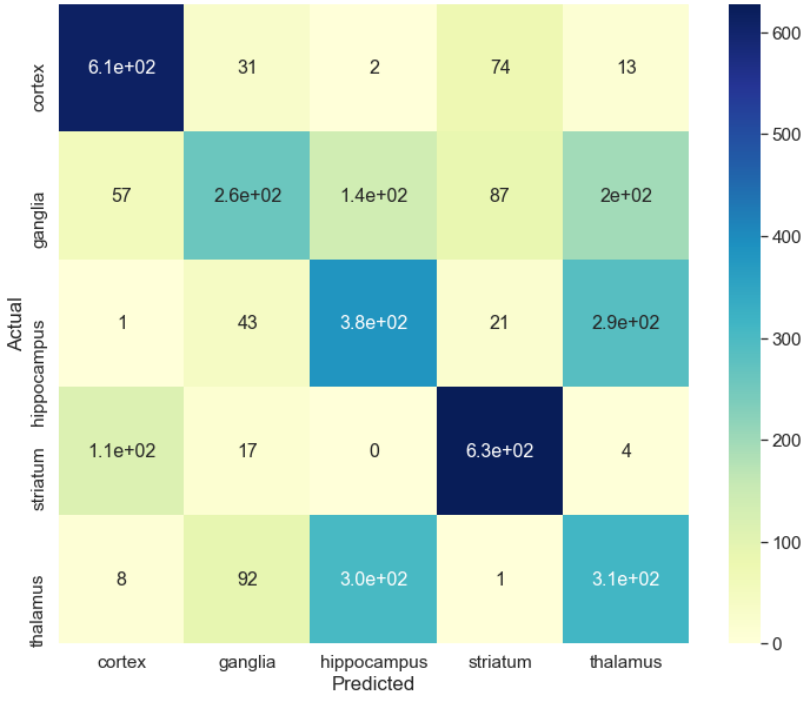

Figure 3-2. Confusion matrix of XGBoost predictions on all five classes from region dataset. Trajectories from the striatum and cortex are predicted with the highest accuracies when all datasets are used. The XGBoost model trained on all five regions predicts trajectories from the thalamus as trajectories from the hippocampus at very similar numbers and predicts a large number of hippocampus trajectories as from the thalamus. Trajectories from the ganglia are most likely to be misclassified as from the hippocampus or the thalamus.

A significant benefit of using a boosted decision tree model on MPT data is that in addition to seeing whether a model can predict complex biological variables, feature importance methods can be used to determine which features the model used to generate predictions and learn patterns in the data. By finding the most important features in an MPT dataset, it is possible to gain insight into how the trajectories differ between independent classes, and therefore begin to understand the biological and chemical differences that led to the model being able to generate predictions much higher than random guessing.

SHAP feature importance from an XGBoost model trained on the age dataset showed that the mean diffusion coefficient at 0.33s (mean_Deff1) was the most important feature, with a magnitude of 2.4 (**Figure 3-3A**). That magnitude was more than double that of the second most important feature, the mean fitted anomalous diffusion coefficient (mean_Dfit), which had a magnitude of 1.1. Four other features had a magnitude above 0.5: mean_boundedness, mean_trappedness, mean_straightness, and mean_fractaldim. The top 12 features were all locally average mean features, demonstrating the value that local averaging of the trajectory data adds to predictive capacity. These magnitude values were all much higher than the SHAP values determined for the Y-scrambled XGBoost model (**Figure 3-3B**). The top feature, kurtosis, had a magnitude of 0.086, only greater than the 15^{th} ranked feature by 0.026. The low magnitude SHAP values and similarity in magnitude across the top 15 features indicates that the high

magnitudes found in the normal XGBoost model are representative of key components of the data that differentiate trajectories from different age groups.

Similar to the results from the age dataset, the top feature for the region dataset was mean_Deff1with a magnitude of 2.3 (**Figure 3-3C**). Mean_fractaldim was the only other feature above 1.0 with a magnitude of 1.3 and mean_Deff2 (mean diffusion coefficient at 3.3s), mean_Dfit, and mean_kurtosis were all above 0.5, indicating the key features driving differentiation between trajectory motion in the different regions. SHAP values of an XGBoost model trained on Y-scrambled data showed features with magnitudes all below 0.04 (**Figure 3-3D**), demonstrating the lack of predictive power from the Y-scrambled features.

SHAP feature importance for an XGBoost model trained on the treatment data showed much lower magnitudes compared to the age and region datasets (**Figure 3-3E**), which was unsurprising due to the comparatively lower prediction accuracies on the treatment dataset. Mean_Deff1 was again the top feature but had a magnitude of only 0.55. Mean_Dfit, with a magnitude of 0.12, was the only other feature with a magnitude above 0.1. SHAP values from an XGBoost model trained on Y-scrambled data from the treatment dataset all fell below 0.015 (**Figure 3-3F**), again showing no predictive power from the Y-scrambled data. The low magnitudes of the SHAP values from the normal XGBoost model on the treatment dataset indicate that mean_Deff1 was the only feature the model could consistently use to different between trajectories from the two classes. Despite the low magnitude, the even lower SHAP values of the Y-scrambled model indicate there are true differences in the diffusion coefficients of trajectories from non-treated and ChABC treated slices that the XGBoost model can identify.

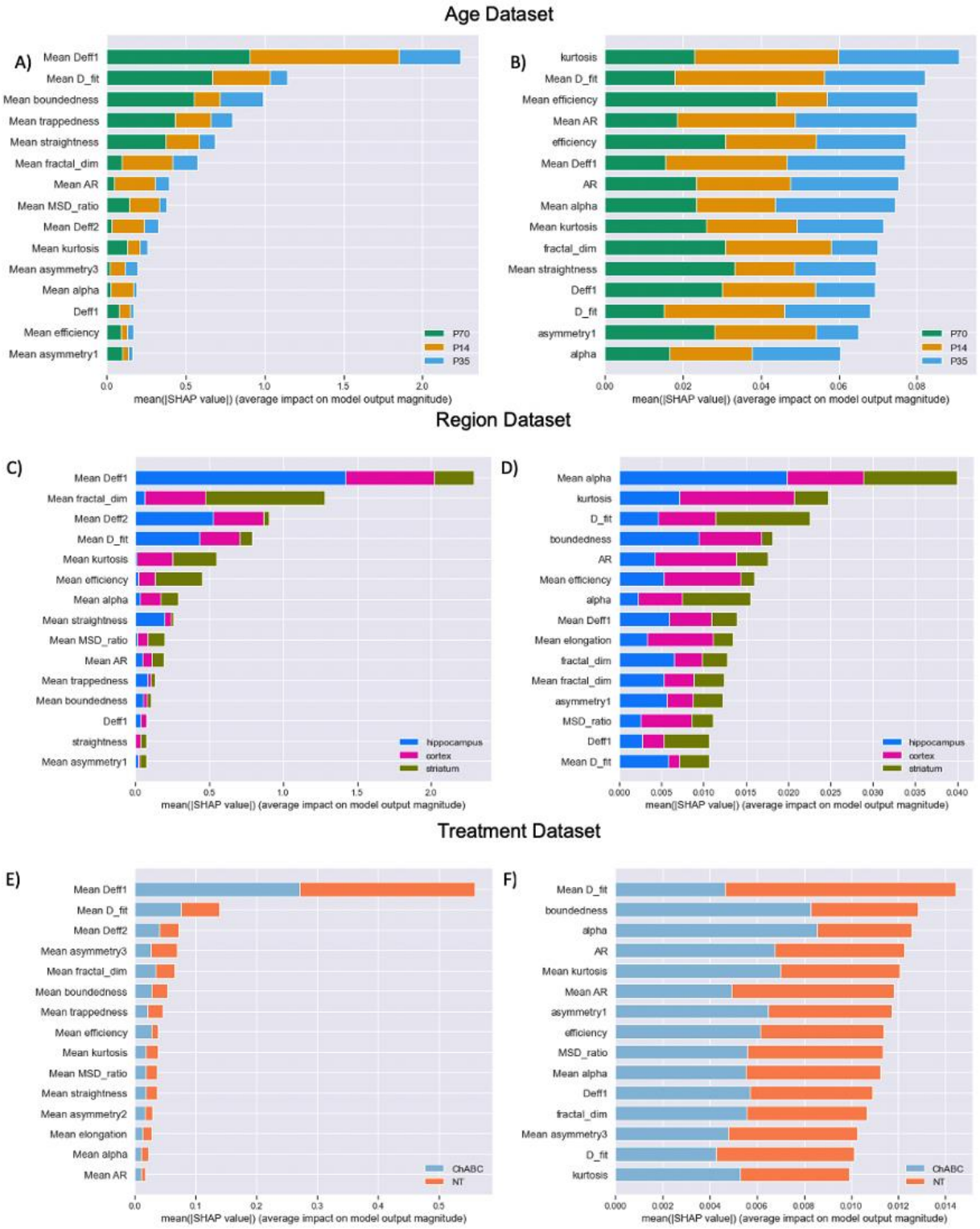
Age Dataset
A)
Mean Deff1
Mean D_fit
Mean boundedness
Mean trappedness
Mean straightness
Mean fractal_dim
Mean AR
Mean MSD_ratio
Mean Deff2
Mean kurtosis
Mean asymmetry3
Mean alpha
Deff1
Mean efficiency
Mean asymmetry1
P70
P14
P35
0.0
0.5
1.0
1.5
2.0
mean(|SHAP value|) (average impact on model output magnitude)
B)
kurtosis
Mean D_fit
Mean efficiency
Mean AR
efficiency
Mean Deff1
AR
Mean alpha
Mean kurtosis
fractal_dim
Mean straightness
Deff1
D_fit
asymmetry1
alpha
P70
P14
P35
0.00
0.02
0.04
0.06
0.08
mean(|SHAP value|) (average impact on model output magnitude)
Region Dataset
C)
Mean Deff1
Mean fractal_dim
Mean Deff2
Mean D_fit
Mean kurtosis
Mean efficiency
Mean alpha
Mean straightness
Mean MSD_ratio
Mean AR
Mean trappedness
Mean boundedness
Deff1
straightness
Mean asymmetry1
hippocampus
cortex
striatum
0.0
0.5
1.0
1.5
2.0
mean(|SHAP value|) (average impact on model output magnitude)
D)
Mean alpha
kurtosis
D_fit
boundedness
AR
Mean efficiency
alpha
Mean Deff1
Mean elongation
fractal_dim
Mean fractal_dim
asymmetry1
MSD_ratio
Deff1
Mean D_fit
hippocampus
cortex
striatum
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
0.040
mean(|SHAP value|) (average impact on model output magnitude)
Treatment Dataset
E)
Mean Deff1
Mean D_fit
Mean Deff2
Mean asymmetry3
Mean fractal_dim
Mean boundedness
Mean trappedness
Mean efficiency
Mean kurtosis
Mean MSD_ratio
Mean straightness
Mean asymmetry2
Mean elongation
Mean alpha
Mean AR
ChABC
NT
0.0
0.1
0.2
0.3
0.4
0.5
mean(|SHAP value|) (average impact on model output magnitude)
F)
Mean D_fit
boundedness
alpha
AR
Mean kurtosis
Mean AR
asymmetry1
efficiency
MSD_ratio
Mean alpha
Deff1
fractal_dim
Mean asymmetry3
D_fit
kurtosis
ChABC
NT
0.000
0.002
0.004
0.006
0.008
0.010
0.012
0.014
mean(|SHAP value|) (average impact on model output magnitude)

Figure 3-3. **SHAP values of normal and Y-scrambled models for each dataset.** The top six SHAP features for (A) the normal model of on the age dataset had values greater than 0.5 with the mean diffusion coefficient at 0.33s being the top feature, (B) while the top feature for the Y-scrambled model had a value of only 0.09. Five features for (C) the normal model of the region dataset were greater than 0.5 with the mean diffusion coefficient at 0.33s at the top, (D) while the highest value feature for the Y-scrambled dataset was 0.04. For the treatment dataset, (E) only the mean diffusion coefficient at 0.33s was above 0.5, and (F) the highest Y-scrambled feature was 0.014.

While XGBoost models outperformed random guessing for each dataset, increasing model accuracy by providing data with the most predictive capability could reveal more insights about biological and chemical differences between classes that may by obscured by noise. Equations fit to the nanoparticle trajectories determine the type of motion they experience, making it possible to separate the data into subsets based on particle motions classified as superdiffusive, subdiffusive, or classical regular Brownian motion. To ensure that any differences were not caused by the presence of more super-diffusive datapoints compared to the other modes or by having a wider distribution of data points, the percentage of each diffusion mode occurring in each dataset was visualized, shown in **Figure 3-4**. The diffusion mode with the highest percentage of data points in the age dataset (**Figure 3-4A**) was sub-diffusive motion, consisting of 49% of P14 data points, 68% of P35 data points, and 68% of P70 data points. Superdiffusive data points accounted for 42% of P14 data points, 24% of P35 datapoints, and 25% of P70 datapoints. The lowest percent was Brownian motion, accounting for 9% of P14 datapoints, 8% of P35 datapoints, and 7% of P70 datapoints. For the region dataset (**Figure 3-4B),** subdiffusive was the most frequent diffusion mode, making up 61% of cortex datapoints, 53% of hippocampus datapoints, and 72% of striatum datapoints. Superdiffusive datapoints were 33% of cortex datapoints, 38% of hippocampus datapoints, and 23% of striatum datapoints. Brownian datapoints were the least frequent, being 6% of cortex datapoints, 9% of hippocampus

datapoints, and 6% of striatum datapoints. For the treatment dataset (**Figure 3-4C)** subdiffusive datapoints were again the most common and accounted for 61% of ChABC data points and 68% of NT datapoints. Superdiffusive datapoints made up 31% of ChABC datapoints and 25% of NT datapoints, while Brownian datapoints were 9% of ChABC datapoints and 8% of NT datapoints.

XGBoost models were then fit to subsets of data consisting to each of the different diffusion modes, and a dataset with all diffusion modes, for each of the three datasets. For each data subset, 50 models were trained, and the final model accuracy was recorded. For the age dataset, the cumulative distribution of accuracies for models trained only on super-diffusive datapoints had a 50% point of 0.87, compared to 0.85 for models trained on Brownian data points or all diffusion modes, and 0.83 for models trained on subdiffusive data points (**Figure 3-4D**). The cumulative distribution of accuracies for the region dataset showed the 50% point for subdiffusive data points at 0.89, followed by all diffusion modes at 0.87, subdiffusive datapoints at 0.86, and Brownian datapoints at 0.84 (**Figure 3-4E**). For the treatment dataset, the 50% point of the cumulative distributions were nearly identical for models trained on superdiffusive datapoints or Brownian motion data points at 0.70, with models trained on all diffusion modes having a 50% point of 0.69 and the distribution for subdiffusive data points being 0.67 (**Figure 3-4F**).

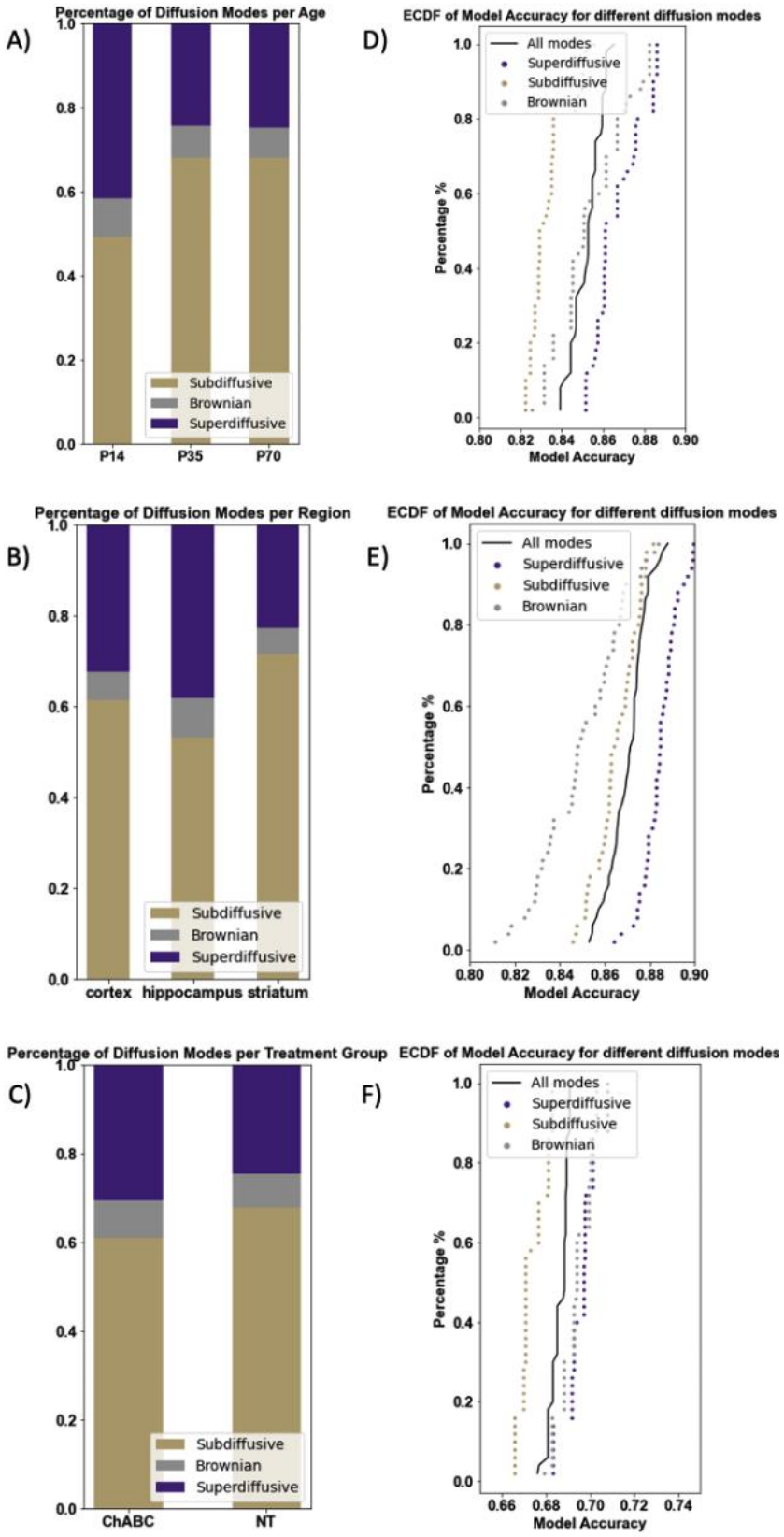
A)
Percentage of Diffusion Modes per Age
Subdiffusive
Brownian
Superdiffusive
P14
P35
P70
D)
ECDF of Model Accuracy for different diffusion modes
All modes
Superdiffusive
Subdiffusive
Brownian
Percentage %
Model Accuracy
B)
Percentage of Diffusion Modes per Region
cortex
hippocampus
striatum
E)
C)
Percentage of Diffusion Modes per Treatment Group
ChABC
NT
F)

Figure 3-4. Percentage and model accuracy of diffusion modes across data sets. (A) Subdiffusive data points were the highest percentage for the age dataset (gold), followed by superdiffusive (purple) and then Brownian (gray). (B) For the region dataset, subdiffusive data points consisted of the highest percentage of data points, with superdiffusive being second and Brownian data points being third. (C) For the treatment dataset subdiffusive datapoints were again the highest percentage, followed by superdiffusive and lastly Brownian data points. (D) XGBoost models trained on only superdiffusive data from the age dataset performed best, with models trained only on subdiffusive datapoints performing worst. (E) XGBoost models trained on superdiffusive data from the region dataset only outperformed models trained on all diffusion modes, only subdiffusive datapoints, and only Brownian datapoints. (F) Models trained only on superdiffusive or Brownian datapoints from the treatment dataset performed similarly, followed by models trained on all modes, with models trained on only subdiffusive data performing worst.

3.3. Discussion and conclusion

The consistency of the machine learning pipeline described in this paper across multiple datasets demonstrates the feasibility of applying a machine learning approach to MPT studies. Our PCA visualizations coupled with the results from the XGBoost models show that models will struggle with data from classes that overlap in these plots and will have the highest accuracy when these plots show distinct separation between classes. By first applying an unsupervised dimensionality reduction algorithm to MPT data, it is possible to quickly determine whether time and resources should be spent training a supervised machine learning on the dataset, or whether more data cleaning or data collection may be needed. The application of the Y-scrambling method shows that the technique can properly ensure that the accuracy of the supervised learning models is indeed caused by the model uncovering patterns in the data.

For the age dataset, the overlap in data between the P21, P28, and P35 classes indicates that any microenvironmental changes between ages are not strongly discernable with the current method of data acquisition and processing, and therefore all three classes should not be included in the same XGBoost model. The random-guess accuracy and low SHAP values of the XGBoost model trained on Y-scrambled age data further indicate that there are clear differences in

diffusion between the P14, P35, and P70 microenvironments that the model can detect. The overlap in the PCA plots of the hippocampus, basal ganglia, and thalamus in the region dataset indicates these three classes should not be added to the same XGBoost model. The difficulty in distinguishing between these regions may be due to similarities in the extracellular microenvironment of the hippocampus, basal ganglia, and thalamus, or it could be a result of variable sampling between regions. In this study, 10 videos were recorded and quantified from both the striatum and cortex, but only 6 were collected from each of the hippocampus, basal ganglia, and thalamus. As a result, highlighted by the output of the Y-scrambled model, the data collected from those three regions may have provided enough of a difference in feature values for a model to separate one of the classes from the striatum and cortex, but not enough to differentiate between the three regions with 6 videos themselves. Future experimental work could increase video collection in the regions that contain the least number of particles tracking for a more even sample size across brain regions, or account for subregions within the specified brain regions. Depending on the available samples, some brain regions are smaller and provide less accessible tissue area per slice or are not present in all slices. Lastly, there is heterogeneity within a given brain region that can vary slice to slice. If notable variability in feature values exist between slices or videos within a given class, it could prevent the identification of differences between classes and in turn reduce predictive capability.

For the treatment dataset, the enzymes used in this study degrade the same components of the ECM, therefore it was unsurprising that the datapoints from ChABC and HYase overlap when visualized using PCA. The overlap in the PCA visualization of the non-treated and treated classes is more surprising, as HYase and ChABC are known to degrade ECM components, and in particular perineural nets (PNNs)[18, 32, 33]. One possible reason for the overlap in these groups is

that the changes to nanoparticle diffusion caused by ECM degradation are not severe enough to cause the statistical features to be different for many of the nanoparticle trajectories. We see this highlighted when looking at Y-scrambled model results and the SHAP values, as the best XGBoost model accuracy was only 19% above random guessing, and only one feature (mean_Deff1) had a high impact on the model.

The importance of locally averaged features across all three datasets indicates local averaging is a significant contributor to model performance and can be beneficial to include in future MPT machine learning work. It may also highlight that it is not individual trajectories that provide robust information to a dataset, but the distribution and diversity of nanoparticle motion within a given subspace. Research into the structure of the brain ECS has shown there to be void spaces connected via tortuous channels[34-36] and reservoirs of space next to cells[37]. As noted in McKenna et al, it is likely that neighboring nanoparticles diffusing in these voids will be experiencing nearly identical environments and have similar diffusive behavior, while trajectories moving in an ECS channel of varying width will have different diffusive behavior. The ECS is also known to vary across age, brain region, and disease[34, 35], so the fact that locally averaged features were the top feature across all three datasets supports the idea these features are detecting changes in the motion of neighboring nanoparticles resulting from structural differences in the ECM between the different ages, regions, or treatment conditions.

The variance of the features with the highest feature importance across different datasets demonstrates that SHAP analysis is a robust technique for MPT data analysis. Our results demonstrate that if we can understand how specific features relate to the interactions that nanoparticles undergo with the microenvironment, we can better understand the unique changes occurring in different biological conditions. While mean_Deff1 and mean_Dfit were in the top 5

features for each dataset, the other most important features that differ between datasets can provide additional insight into structural and environmental changes. For example, the other top 5 features for the age dataset are mean boundedness, mean trappedness, and mean straightness. These features primarily relate to how far a nanoparticle travels for each time step and whether it is moving within a confined space[21]. For the region dataset top features are mean_fractaldim, mean_kurtosis, and mean_Deff2, which relate to the space taken up by the trajectory (fractaldim) and asymmetry of point distribution of the trajectory (kurtosis)[22].

Despite having the greatest number of trajectories by percentage in each dataset, models trained on sub-diffusive trajectories consistently underperformed those trained on only super-diffusive trajectories or all trajectories. This may hint that simply having larger datasets and more trajectories would not be an effective approach to improving model accuracy on MPT datasets. In general, the results seem to show that faster moving particles lead to models with higher accuracy. One reason that super-diffusive particles lead to models with higher accuracies could be that increased movement better captures the biological differences within the ECS between classes. Based on our data, it is unclear if having a higher number of super-diffusive trajectories within a class improves accuracy relative to other classes.

For the age data, despite having the lowest percentage of super-diffusive trajectories, the P70 class is predicted at the highest rate, followed by P14 with the highest percentage of super-diffusive trajectories. Studies looking at changes in diffusion during the neurodevelopmental process in rats has shown that the ECS volume fraction decreases with age[38]. Interestingly, the authors saw a large decrease in volume fraction between postnatal days 10-21, but not between postnatal day 21 and adults aged 90-120 days. Since some of the most important features for the age dataset are mean_boundedness, mean_trappedness, and mean_straightness, which relate to

the confinement of the nanoparticle trajectory, super-diffusive trajectories may be better at detecting the differences in ECS volume fraction between postnatal day 14, which would have a relatively large volume fraction, and day 70, which would have a smaller volume fraction.

For the region data, the hippocampus has both the highest percentage of super-diffusive trajectories as well as the highest model accuracy. Research looking at ECM matrix composition in the rat brain showed regionally dependent expression of aggrecan, brevican, and tenascin-R[39], where the hippocampus had the highest signal intensity of all the analyzed regions. It is possible that the data from super-diffusive trajectories provide the predictive model with a high amount of information due to a greater ability to navigate the dense ECM in this region. Super-diffusive trajectories may also be better at capturing the structural differences caused by the regional differences in protein expressions, and specifically how these changes are affecting the fractal dimension and kurtosis of the trajectories across different regions.

The non-treated class for the treatment data has a lower percentage of super-diffusive trajectories compared to the ChABC-treated class but has a higher accuracy. One potential reason could be that having more super-diffusive trajectories improves accuracy up until a certain threshold, when accuracy becomes saturated and adding more data points no longer provides information to the model. When brain slices are treated with ChABC, the effective diffusion coefficient has been shown to increase. While the super-diffusive nanoparticles may be better at reflecting this change in diffusion relative to trajectories of other diffusion modes, as mentioned earlier the degree of degradation may not sufficient enough for a machine learning model to pick up the microstructural changes. The loss of PNNs due to ChABC treatment may have also caused neuronal cell death, an inflammatory response, and activation of cells such as

microglia[40], which could alter the ECS and ECM, mitigating the effects of the loss of ECM structure.

In this project, I develop a methodology to apply data science and machine learning tools to MPT data to probe underlying biological changes using three distinct experimental datasets. While MPT has long been a technique used to understand complex biological environments, our methodology increases the insights extracted from a given MPT dataset. Specifically, we demonstrate how unsupervised learning can be leveraged to understand the predictive capacity of a dataset, and how supervised learning and feature importance techniques can be applied to begin understanding microenvironmental and structural changes in the brain ECS. By applying the methodology to three unique datasets, we show that PCA can feasibly be applied to any other MPT datasets to determine whether there is data separation in the dataset, and XGBoost and SHAP analysis will be most effective on datasets where PCA is able to separate data.

4. Continuing and Future work

4.1. Application of pipeline to other MPT datasets

While the methodology presented in this paper is successful for predicting neurodevelopmental age, brain region, and ECM treatment, more work is needed to determine the extent at which machine learning and MPT can be applied to predict other complex biological variables and probe microenvironmental differences. Within the context of the brain, the methodology could be applied to predict the spectrum of disease progression or severity as opposed to binary classifications of either healthy or not healthy. Planned future work within the lab aims to apply this pipeline to two models of neurodegeneration, namely rotenone treated brain slices and oxygen-glucose deprivation treated brain slices.

4.2. Development of MPT trajectory database

As the number of MPT datasets generated in the Nance lab continues to grow, machine learning will continue to be able to be used to answer interesting biological questions. However due to the size of each dataset, it is crucial to have a standardized data management system in place. Future work will include the development of a database storing all current tracked nanoparticles, with the ability to add future MPT datasets into the database. Developing a database will streamline the process of collecting subsets of data for machine learning problems, for example collecting all super-diffusive trajectories from a single dataset. It would also make it possible to combine data from multiple datasets that have a variable held constant, for example utilizing all trajectories tracked in the cortex from multiple datasets.

4.3. Deep learning applications

Past research has shown that a deep learning approach can be used for nanoparticle trajectory classification tasks by applying a convolutional neural network (CNN) to the raw trajectory data[20, 22]. For each of these datasets, future work can investigate whether or not a deep learning approach on raw trajectory data, as opposed to the statistical features used in this study, can improve the accuracy of predictions on classes which these current methods struggle. Synthetic data can also be generated using a Generative Adversarial Network deep learning approach, which has been shown to improve CNN accuracy in medical classification tasks[41].

5. Summary of published/preparing to be published work

Nano-based probes for the brain extracellular environment

Jeremy R. Filteau, Brandan P. Butler, Nels Schimek, Elizabeth Nance

An extensive and growing number of studies demonstrates the critical role of the brain extracellular space (ECS) in normal development, aging, and in response to injury or disease.

Therefore, developing technologies to probe this dynamic and heterogeneous environment is of increasing interest. Nanomaterials are a promising platform for probing the brain ECS, given their highly tunable properties and compatibility with quantitative imaging. In this chapter, we discuss design considerations for engineering nanomaterial probes for application in the brain parenchyma. We begin with background on the brain ECS microenvironment, followed by discussion on how changes in diffusion, rheology, and composition reflect development, disease, or injury-mediated remodeling. We highlight the spectrum of models used to probe the brain environment, from synthetic engineered systems (i.e. hydrogels) to *ex vivo* slices/organoids, and *in vivo* animal models of injury and disease. Equally important to the choice of model system is the design of the nanomaterial probe. We discuss common physicochemical parameters for nano-based probes, including size, shape, surface charge, surface chemistry. We next analyze the various techniques used to quantify the behavior of nanoparticles in the brain, which include techniques such as integrative optical imaging, real-time iontophoresis, and particle tracking microscopy. Lastly, we examine active research in improving nanomaterial design for quantifying stimuli responsiveness, advancing super-resolution imaging techniques, and employing machine learning and artificial intelligence to understand and predict underlying biological changes to the brain microenvironment.

Machine learning applications to multiple particle tracking data to evaluate the brain extracellular space

Nels Schimek, David Beck, Michael McKenna, Elizabeth Nance

Multiple particle tracking (MPT) is a microscopy technique capable of simultaneously tracking hundreds to thousands of nanoparticles in a biological sample and has been used extensively in recent years to characterize the brain extracellular space. Machine learning techniques have been applied to MPT datasets in order to predict the diffusion mode of nanoparticle trajectories as well as more complex biological variables. In this study, we develop a machine learning pipeline and evaluate its effectiveness across three different MPT datasets: varying age, varying region, and a non-treated versus treated condition. We utilize unsupervised learning, supervised classification, and feature importance calculations to determine which datasets can be predicted with the highest accuracy, and to glean biological insights from each dataset. Finally, we determine the effect that the diffusion mode of a trajectory has on training a supervised machine learning model.

Nels Schimek
206.999.1268 | nlsschim@uw.edu | https://www.linkedin.com/in/nlsschim/

Current Masters Thesis Candidate experienced with data science approaches for neurological and disease research. Pursuing a career in biotechnology with interests in disease research, chemical biology, data visualization, and machine learning. Experience with scientific communication, biological data analysis, software package development, machine learning, and cloud and high-performance computing.

Education

2020- 2022 M.Sc. in Applied Chemical Science and Technology, University of Washington, Seattle

Expected thesis defense: Spring 2022

Thesis title: Machine Learning Methods for Analyzing Biological Multiple Particle Tracking Data

2016-2020 B.S. in Biochemistry. University of Washington, Seattle

Skills:

Technical: Python (Numpy, Matplotlib, Pandas, scikit-learn), Matlab, R, Azure, git, GitHub, CLI, Excel

Organizational: Public speaking, scientific writing, team building, mentorship, proposal writing

Research Experience

Graduate research assistant, University of Washington Sept 2020 - Present

Disease Directed Engineering lab in the Department of Chemical Engineering; PI: Dr. Elizabeth Nance

- Development and testing of diff_predictor, an open-source software package for machine learning analysis of multiple particle tracking data
- Application of diff_predictor to generate machine learning predictions and feature importance values to gain insight into age, region, and treatment-based biological changes in the brain extracellular space.
- Taught and developed learning modules for TEXTILE, a lab developed framework for teaching data science to high school, undergraduate, and graduate researchers.

Undergraduate research assistant, University of Washington Sept 2016 - Sept 2020

Department of Neurosurgery; PI: Dr. Pierre Mourad

- Design and execution of neuromodulation experiments in rodents and human using medical ultrasound to probe the visual cortex
- Developed MATLAB scripts to visualize and perform statistical analysis on EEG data

Work Experience

Student assistant, University of Washington Department of Biology Oct 2021 – Feb 2022

- Integration and unit testing of SEIRS+, an open-source python package for dynamic epidemiological modeling.
- Deployment of SEIRS+ package for simulation-based case studies on spread and methods of intervention for COVID19

Publications

Schimek, N., Beck, D., Mckenna, M., Nance, E *. Multiple Particle Tracking for Predictions of Biological Differences. In preparation.

Schimek, N., Shackelford, D., Beck, D., Nance, E *. diff_predictor: Machine learning for visualization, prediction, and feature importance of multiple particle tracking data. *Journal of Open Source Software.* In preparation.

Schimek, N., Filteau J., Butler, B., Nance, E.* Nano-based probes for the brain extracellular environment. In revision

Schimek, N., Burke-Conte, Z., Abernethy, J., Schimek, M., Burke-Conte, C., Bobola, M., Stocco, A., Mourad, P *. (2020). Repeated Application of Transcranial Diagnostic Ultrasound Towards the Visual Cortex Induced Illusory Visual Percepts in Healthy Participants. *Front. Hum. Neurosci.* 14, 66. doi:10.3389/fnhum.2020.00066.

Bobola, M. S., Chen, L., Ezeokeke, C. K., Kuznetsova, K., Lahti, A. C., Lou, W., **Schimek, N.,** Mourad, P *. (2018). A Review of Recent Advances in Ultrasound, Placed in the Context of Pain Diagnosis and Treatment. *Curr. Pain Headache Rep.* 22, 60. doi:10.1007/s11916-018-0711-7.

Mentoring

Undergraduate mentor, Nance Lab, University of Washington June 21 - Present

Continuously mentored two undergraduate students in the Nance Lab in introductory programming, software development, and data science.

Hackathon tutorial instructor, UW Department of Chemical Engineering Jan 3, 22 – Jan 14, 22
Assisted Chemical Engineering undergraduates in learning introductory Python and data science topics.

Funding and Awards

UW Azure Cloud Computing Credits for Research. Dec. 2021 - June 2022
UW eScience Institute

UW Azure Cloud Computing Credits for Research. Dec. 2020 - June 2021
UW eScience Institute

Summer Undergraduate Research Program grant. June 2018 - Aug. 2018
Washington NASA Space Grant Consortium

Bibliography

1. Zimmermann DR, Dours-Zimmermann MT. Extracellular matrix of the central nervous system: from neglect to challenge. Histochemistry and Cell Biology. 2008;130(4):635-53. doi: 10.1007/s00418-008-0485-9.
2. Krishnaswamy VR, Benbenishty A, Blinder P, Sagi I. Demystifying the extracellular matrix and its proteolytic remodeling in the brain: structural and functional insights. Cellular and Molecular Life Sciences. 2019;76(16):3229-48. doi: 10.1007/s00018-019-03182-6.
3. Lau LW, Cua R, Keough MB, Haylock-Jacobs S, Yong VW. Pathophysiology of the brain extracellular matrix: a new target for remyelination. Nature Reviews Neuroscience. 2013;14(10):722-9. doi: 10.1038/nrn3550.
4. Nance EA, Woodworth GF, Sailor KA, Shih T-Y, Xu Q, Swaminathan G, Xiang D, Eberhart C, Hanes J. A Dense Poly(Ethylene Glycol) Coating Improves Penetration of Large Polymeric Nanoparticles Within Brain Tissue. Science Translational Medicine. 2012;4(149):149ra19-ra19. doi: doi:10.1126/scitranslmed.3003594.
5. Selvaggi L, Salemme M, Vaccaro C, Pesce G, Rusciano G, Sasso A, Campanella C, Carotenuto R. Multiple-Particle-Tracking to investigate viscoelastic properties in living cells. Methods. 2010;51(1):20-6. doi: https://doi.org/10.1016/j.ymeth.2009.12.008.
6. Valentine MT, Perlman ZE, Gardel ML, Shin JH, Matsudaira P, Mitchison TJ, Weitz DA. Colloid surface chemistry critically affects multiple particle tracking measurements of biomaterials. Biophys J. 2004;86(6):4004-14. doi: 10.1529/biophysj.103.037812. PubMed PMID: 15189896.
7. Xu Q, Boylan NJ, Suk JS, Wang YY, Nance EA, Yang JC, McDonnell PJ, Cone RA, Duh EJ, Hanes J. Nanoparticle diffusion in, and microrheology of, the bovine vitreous ex vivo. J Control Release. 2013;167(1):76-84. Epub 20130128. doi: 10.1016/j.jconrel.2013.01.018. PubMed PMID: 23369761; PMCID: PMC3693951.
8. Abdulkarim M, Agulló N, Cattoz B, Griffiths P, Bernkop-Schnürch A, Borros SG, Gumbleton M. Nanoparticle diffusion within intestinal mucus: Three-dimensional response analysis dissecting the impact of particle surface charge, size and heterogeneity across polyelectrolyte, pegylated and viral particles. Eur J Pharm Biopharm. 2015;97(Pt A):230-8. Epub 20150204. doi: 10.1016/j.ejpb.2015.01.023. PubMed PMID: 25661585.
9. Duncan GA, Kim N, Colon-Cortes Y, Rodriguez J, Mazur M, Birket SE, Rowe SM, West NE, Livraghi-Butrico A, Boucher RC, Hanes J, Aslanidi G, Suk JS. An Adeno-Associated Viral Vector Capable of Penetrating the Mucus Barrier to Inhaled Gene Therapy. Mol Ther Methods Clin Dev. 2018;9:296-304. Epub 20180322. doi: 10.1016/j.omtm.2018.03.006. PubMed PMID: 30038933; PMCID: PMC6054694.
10. Bloom RJ, George JP, Celedon A, Sun SX, Wirtz D. Mapping Local Matrix Remodeling Induced by a Migrating Tumor Cell Using Three-Dimensional Multiple-Particle Tracking. Biophys J. 2008;95(8):4077-88. doi: https://doi.org/10.1529/biophysj.108.132738.
11. Kim AJ, Hanes J. The emergence of multiple particle tracking in intracellular trafficking of nanomedicines. Biophys Rev. 2012;4(2):83-92. Epub 2012/02/03. doi: 10.1007/s12551-012-0066-y. PubMed PMID: 28510091.
12. Suh J, Wirtz D, Hanes J. Efficient active transport of gene nanocarriers to the cell nucleus. Proc Natl Acad Sci U S A. 2003;100(7):3878-82. Epub 2003/03/18. doi: 10.1073/pnas.0636277100. PubMed PMID: 12644705.

13. Bausinger R, von Gersdorff K, Braeckmans K, Ogris M, Wagner E, Bräuchle C, Zumbusch A. The Transport of Nanosized Gene Carriers Unraveled by Live-Cell Imaging. Angewandte Chemie International Edition. 2006;45(10):1568-72. doi: https://doi.org/10.1002/anie.200503021.
14. Lai SK, Hida K, Chen C, Hanes J. Characterization of the intracellular dynamics of a non-degradative pathway accessed by polymer nanoparticles. Journal of controlled release : official journal of the Controlled Release Society. 2008;125(2):107-11. Epub 2007/10/25. doi: 10.1016/j.jconrel.2007.10.015. PubMed PMID: 18053606.
15. Suh J, Choy K-L, Lai SK, Suk JS, Tang BC, Prabhu S, Hanes J. PEGylation of nanoparticles improves their cytoplasmic transport. Int J Nanomedicine. 2007;2(4):735-41. PubMed PMID: 18203439.
16. Nance EA, Woodworth GF, Sailor KA, Shih TY, Xu Q, Swaminathan G, Xiang D, Eberhart C, Hanes J. A dense poly(ethylene glycol) coating improves penetration of large polymeric nanoparticles within brain tissue. Sci Transl Med. 2012;4(149):149ra19. doi: 10.1126/scitranslmed.3003594. PubMed PMID: 22932224; PMCID: PMC3718558.
17. Joseph A, Liao R, Zhang M, Helmbrecht H, McKenna M, Filteau JR, Nance E. Nanoparticle-microglial interaction in the ischemic brain is modulated by injury duration and treatment. Bioengineering & Translational Medicine. 2020;5(3):e10175. doi: https://doi.org/10.1002/btm2.10175.
18. McKenna M, Shackelford D, Ferreira Pontes H, Ball B, Nance E. Multiple Particle Tracking Detects Changes in Brain Extracellular Matrix and Predicts Neurodevelopmental Age. ACS Nano. 2021;15(5):8559-73. doi: 10.1021/acsnano.1c00394.
19. Gareth James DWTHRT. An introduction to statistical learning : with applications in R: New York : Springer, [2013] ©2013; 2013.
20. Granik N, Weiss LE, Nehme E, Levin M, Chein M, Perlson E, Roichman Y, Shechtman Y. Single-Particle Diffusion Characterization by Deep Learning. Biophys J. 2019;117(2):185-92. Epub 20190622. doi: 10.1016/j.bpj.2019.06.015. PubMed PMID: 31280841; PMCID: PMC6701009.
21. Wagner T, Kroll A, Haramagatti CR, Lipinski H-G, Wiemann M. Classification and Segmentation of Nanoparticle Diffusion Trajectories in Cellular Micro Environments. PLOS ONE. 2017;12(1):e0170165. doi: 10.1371/journal.pone.0170165.
22. Kowalek P, Loch-Olszewska H, Szwabiński J. Classification of diffusion modes in single-particle tracking data: Feature-based versus deep-learning approach. Physical Review E. 2019;100(3):032410. doi: 10.1103/PhysRevE.100.032410.
23. Curtis C, Rokem A, Nance E. diff_classifier: Parallelization of multi-particle tracking video analyses. J Open Source Softw. 2019;4(36). Epub 20190410. doi: 10.21105/joss.00989. PubMed PMID: 31431940; PMCID: PMC6701843.
24. Hastie T, Tibshirani R, Friedman JH. The Elements of Statistical Learning: Data Mining, Inference, and Prediction: Springer; 2009.
25. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V. Scikit-learn: Machine learning in Python. the Journal of machine Learning research. 2011;12:2825-30.
26. Li H, Tamang T, Nantasenamat C. Toward insights on antimicrobial selectivity of host defense peptides via machine learning model interpretation. Genomics. 2021;113(6):3851-63. doi: https://doi.org/10.1016/j.ygeno.2021.08.023.

27. Shapley LS. A Value for N-Person Games. Santa Monica, CA: RAND Corporation; 1952.
28. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Proceedings of the 31st International Conference on Neural Information Processing Systems; Long Beach, California, USA: Curran Associates Inc.; 2017. p. 4768–77.
29. Burnecki K, Kepten E, Garini Y, Sikora G, Weron A. Estimating the anomalous diffusion exponent for single particle tracking data with measurement errors - An alternative approach. Scientific Reports. 2015;5(1):11306. doi: 10.1038/srep11306.
30. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; San Francisco, California, USA: Association for Computing Machinery; 2016. p. 785–94.
31. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee S-I. From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence. 2020;2(1):56-67. doi: 10.1038/s42256-019-0138-9.
32. Vedunova M, Sakharnova T, Mitroshina E, Perminova M, Zakharov Y, Pimashkin A, Dityatev A, Mukhina I. Seizure-like activity in hyaluronidase-treated dissociated hippocampal cultures. Frontiers in Cellular Neuroscience. 2013;7. doi: 10.3389/fncel.2013.00149.
33. Chu P, Abraham R, Budhu K, Khan U, De Marco Garcia N, Brumberg JC. The Impact of Perineuronal Net Digestion Using Chondroitinase ABC on the Intrinsic Physiology of Cortical Neurons. Neuroscience. 2018;388:23-35. Epub 20180710. doi: 10.1016/j.neuroscience.2018.07.004. PubMed PMID: 30004010; PMCID: PMC6338339.
34. Syková E, Nicholson C. Diffusion in brain extracellular space. Physiol Rev. 2008;88(4):1277-340. doi: 10.1152/physrev.00027.2007. PubMed PMID: 18923183; PMCID: PMC2785730.
35. Thorne RG, Nicholson C. In vivo diffusion analysis with quantum dots and dextrans predicts the width of brain extracellular space. Proceedings of the National Academy of Sciences. 2006;103(14):5567-72. doi: doi:10.1073/pnas.0509425103.
36. Vanharreveld A, Crowell J, Malhotra SK. A STUDY OF EXTRACELLULAR SPACE IN CENTRAL NERVOUS TISSUE BY FREEZE-SUBSTITUTION. J Cell Biol. 1965;25(1):117-37. doi: 10.1083/jcb.25.1.117. PubMed PMID: 14283623.
37. Hrabetova S, Cognet L, Rusakov DA, Nägerl UV. Unveiling the Extracellular Space of the Brain: From Super-resolved Microstructure to In Vivo Function. The Journal of Neuroscience. 2018;38(44):9355-63. doi: 10.1523/jneurosci.1664-18.2018.
38. Lehmenkühler A, Syková E, Svoboda J, Zilles K, Nicholson C. Extracellular space parameters in the rat neocortex and subcortical white matter during postnatal development determined by diffusion analysis. Neuroscience. 1993;55(2):339-51. doi: 10.1016/0306-4522(93)90503-8. PubMed PMID: 8377929.
39. Dauth S, Grevesse T, Pantazopoulos H, Campbell PH, Maoz BM, Berretta S, Parker KK. Extracellular matrix protein expression is brain region dependent. J Comp Neurol. 2016;524(7):1309-36. doi: 10.1002/cne.23965. PubMed PMID: 26780384.
40. Bonneh-Barkay D, Wiley CA. Brain extracellular matrix in neurodegeneration. Brain Pathol. 2009;19(4):573-85. Epub 20080725. doi: 10.1111/j.1750-3639.2008.00195.x. PubMed PMID: 18662234; PMCID: PMC2742568.
41. Chen RJ, Lu MY, Chen TY, Williamson DFK, Mahmood F. Synthetic data in machine learning for medicine and healthcare. Nat Biomed Eng. 2021;5(6):493-7. doi: 10.1038/s41551-021-00751-8. PubMed PMID: 34131324.