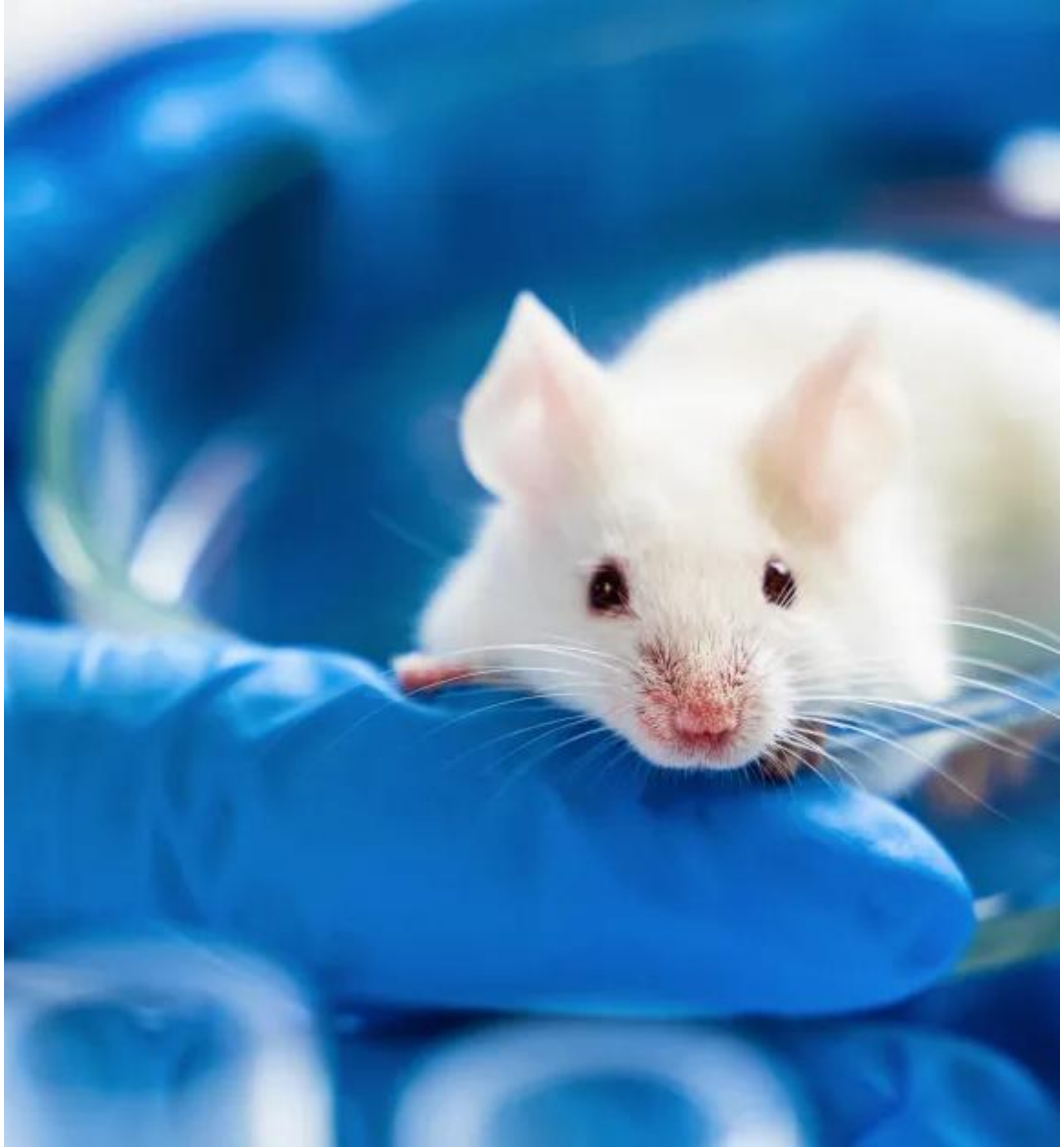


# Classification of Mice Based on Protein Expression Levels

FIRST MAJOR PROJECT

Presentation By Mansi Naik

# Overview

- Problem Statement
- Goal
- Objective
- Dataset Information
- Steps Involved
- Statistics
- Challenges Faced
- Learning
- Conclusion

# Problem Statement

The objective is to identify of proteins that are discriminant between different classes of mice. The dataset include protein expression levels in the cerebral cortex of both control and Down syndrome mice subjected to context fear conditioning. The aim is to classify the mice into eight distinct classes based on genotype, behavior,  and treatment.

# Goal

In this project, we’re using machine learning to help classify different types of mice based on the proteins found in their brains. We're not just focused on getting accurate predictions, We also want to understand what these proteins can tell us about the mice. By doing this, we hope to learn how factors like genetics, treatment, and behavior affect brain activity, and build a model that could support future research in biology and medicine.

# Objectives

*Understand the Data:*

Start by exploring the dataset to see what kind of information it contains and identify any missing or inconsistent values.

*Prepare the Data:*

Clean and preprocess the data so it’s ready for machine learning. This includes handling null values, normalizing values, and encoding labels.

## *Analyze Patterns:*

Use data visualization and statistical tools to find patterns or trends in the protein expression data that may relate to mouse type.

## *Build and Compare Models:*

Train different machine learning models to classify the mice and compare their performance.

## *Evaluate and Interpret:*

Use metrics like accuracy and confusion matrix to evaluate how well the models work, and try to understand which features (proteins) are most important in the classification.

# Dataset Information

- *Instances*- 1080
(Each mouse has 15 measurements for each protein)

- *Features*- 80
(77 protein expression values + 3 metadata columns: Mouse ID, Genotype, Treatment)

- *Classes*- 8
(Based on combinations of Genotype, Behavior, and Treatment)

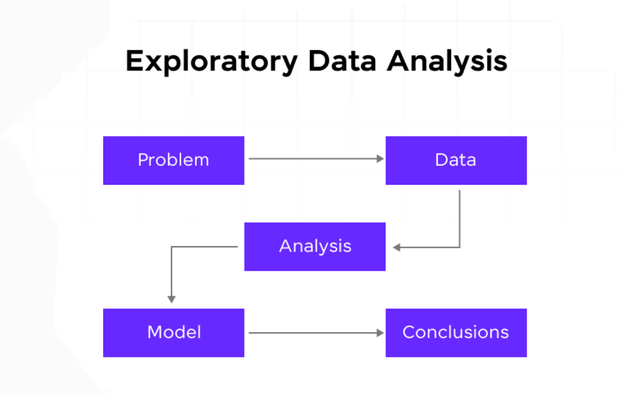
# Steps Involved

## *Data Loading & Cleaning*

Import the dataset, check for missing values, and handle any inconsistencies.

## *Exploratory Data Analysis (EDA)*

Use visualizations and summary statistics to understand patterns in the data.



Exploratory Data Analysis

Problem

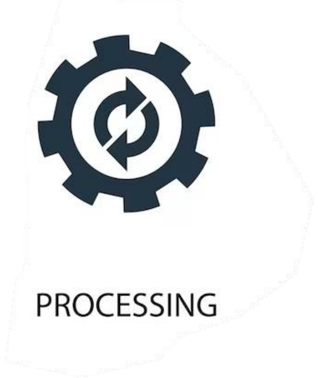
Data

Analysis

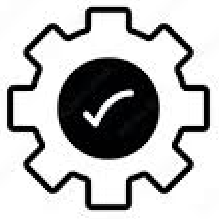
Model

Conclusions

## *Preprocessing*



Normalize the data, encode class labels, and prepare it for machine learning models.



## *Feature Selection*

Reduce dimensionality by identifying the most informative protein features.

## *Model Building*

Train multiple models (e.g., Random Forest, SVM) to classify the mice.

## *Model Evaluation*

Assess model performance using accuracy, confusion matrix, and cross-validation

## *Result Interpretation*

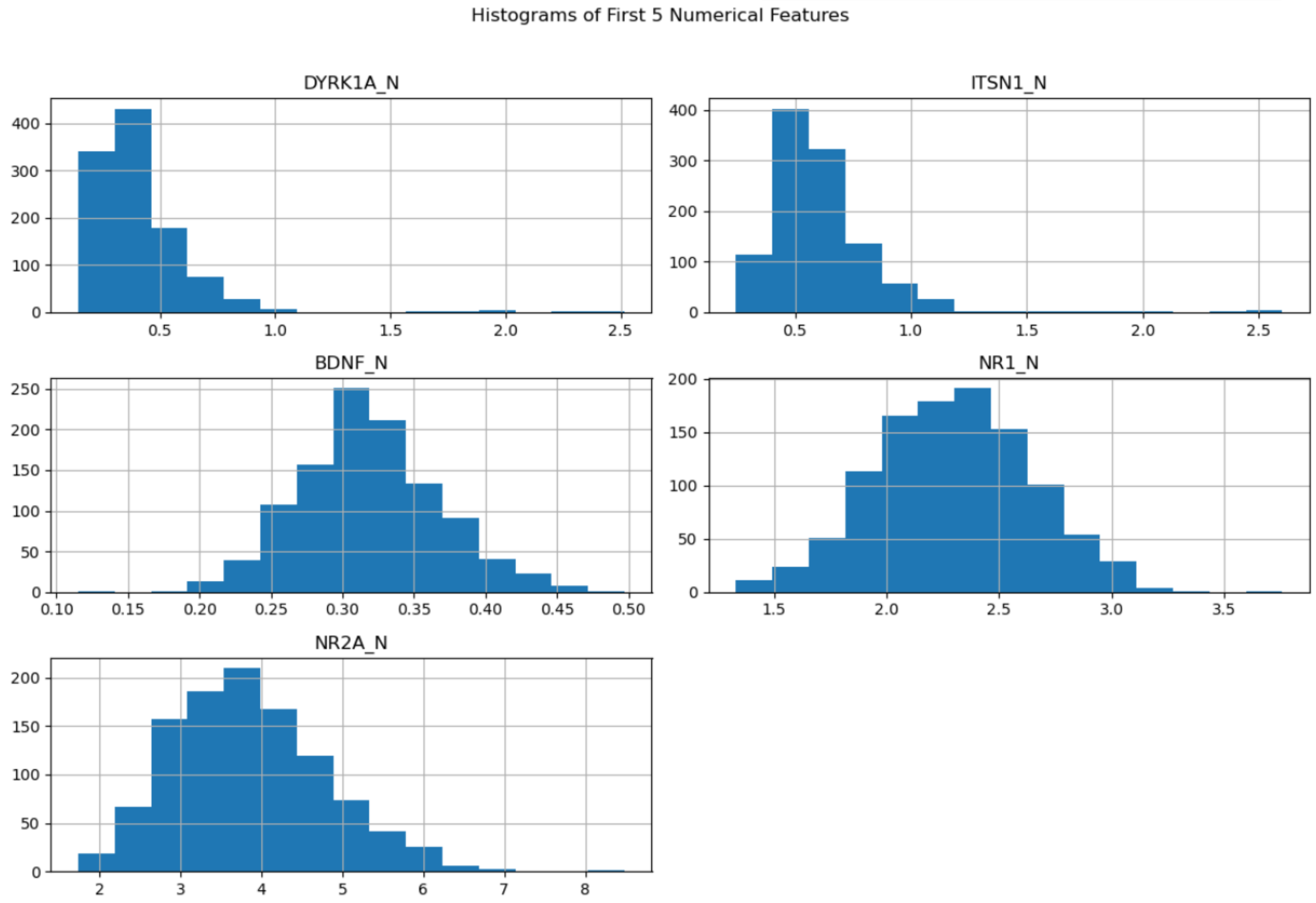
Analyze which proteins and features contributed most to the classification.

## *Conclusion & Insights*

Summarize findings and biological implications of the classification results.

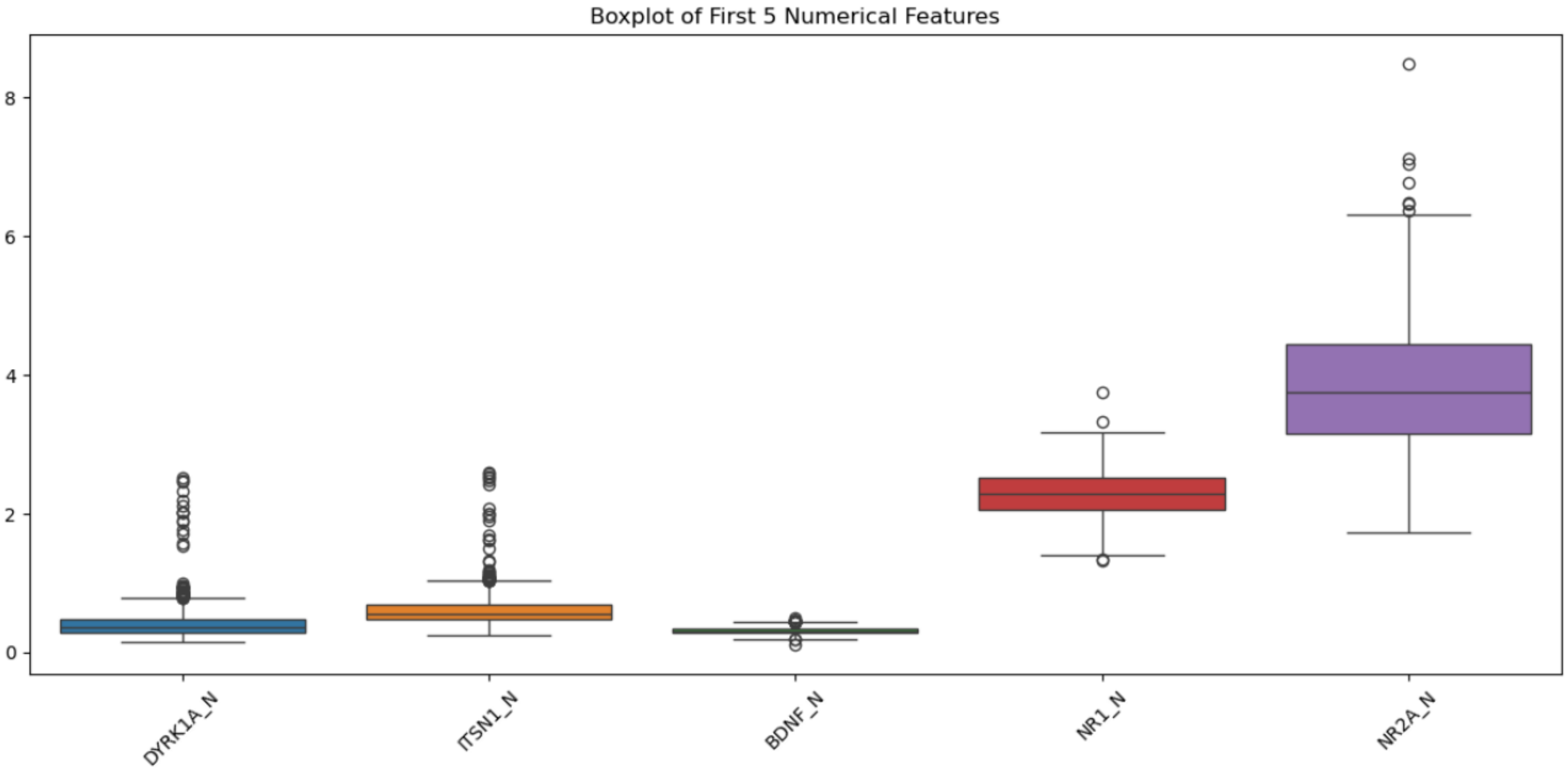
Conclusion

# Histogram



A histogram shows how often different values appear in a dataset.
Each bar represents a range of protein expression values, and the height shows how many mice fall into that range.
It helps us understand the shape of the data—whether it's normally distributed, skewed, or has multiple peaks.
This is useful for spotting imbalances or patterns in the protein data.
Overall, histograms help decide if the data needs normalization or transformation before modeling.

# Box plot

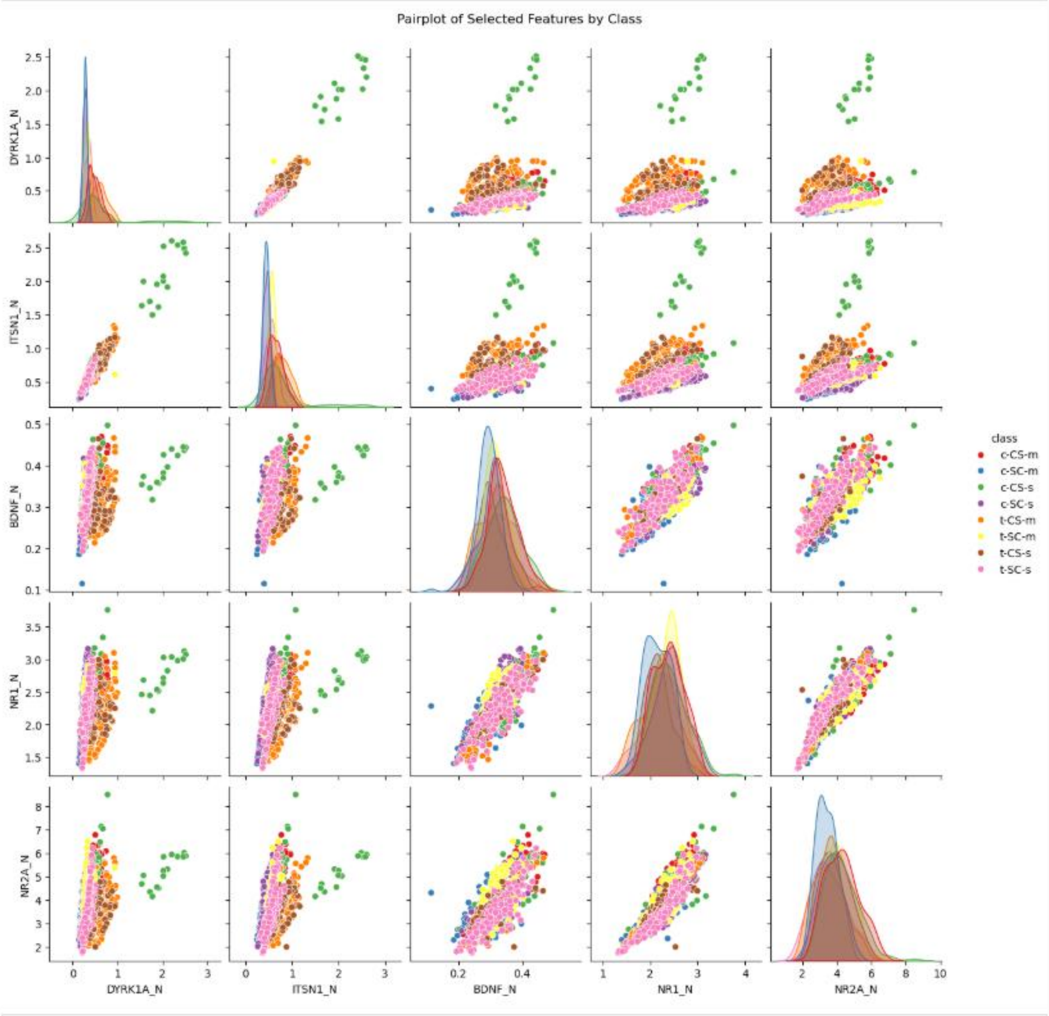


A box plot helps visualize the distribution of protein expression data. The line inside the box shows the median, representing the central value. The box itself captures the middle 50% of the data, and its length indicates the spread or variability.

The whiskers extend to the typical range of values, while any points outside them are considered outliers.

These outliers may reflect unusual protein activity or potential biological differences in the mice.

# Pair Plot



A pair plot shows the relationship between multiple protein features at once.
It creates scatter plots for every pair of proteins, helping us see how they vary together.
Patterns like clusters, trends, or linear relationships become more visible in this format.
It also reveals whether certain groups of mice are naturally separating based on protein expression.
This makes it a great tool for early feature exploration before applying machine learning.

# Challenges Faced

*High Dimensionality:*
With over 70 protein features, it was hard to decide which ones were truly important without overfitting the model.

*Class Imbalance:*
Some mouse groups had fewer samples, which made training balanced models difficult.

*Missing & Noisy Data:*
A few values were missing or looked inconsistent, which required careful cleaning and imputation.

*Overlapping Classes:*

Some groups had very similar protein patterns, which made it tough for the model to separate them clearly.

*Model Tuning:*

Finding the right algorithms and parameters took time, since each performed differently on the complex data.

# Experience & Learnings

This project helped us understand how machine learning can be used to study protein data in mice. We learned how to clean and analyze data using visual tools like box plots and histograms. Trying different models showed us how each has its own strengths and weaknesses. Most importantly, we saw how data science can help uncover patterns in biology in a fun and meaningful way.

# Conclusion

In this project, we used machine learning to classify different groups of mice based on their brain protein levels. By analyzing the data and testing different models, we were able to predict how factors like treatment, behavior, and genetics affect protein expression. Our results show that machine learning is a powerful tool for understanding complex biological patterns. This approach can support future research in neuroscience and drug development.

# Thank you