learning_capstone_final_report

Identifying and predicting potential therapeutic targets

To better understand how to treat genetic disorders or diseased states, it is important to determine which genes or proteins should be targeted. These targets can be identified by uncovering changes in their expression levels in affected individuals compared with unaffected individuals.

Down Syndrome, or Trisomy 21, which results from an extra copy of chromosome 21, affects approximately 1 in every 700 babies born each year in the United States. The genetic aberrations associated with Down Syndrome can delay the individual's development, cause deficits in learning and memory, and lead to a higher prevalence of early-onset neurodegenerative disorders, such as Alzheimer's disease. The proteins and molecular mechanisms that underlie these nervous system defects could serve as potential therapeutic targets. This project set out to identify these proteins and to predict whether the measured protein levels are a result of specific experimental manipulations (or class of mice).

The experimental approach and resulting data

Protein expression data (http://archive.ics.uci.edu/ml/datasets/Mice+Protein+Expression) was collected from wild-type mice and a mouse model for Trisomy 21 (Ts65Dn). These data were donated by Dr. Clara Higuera (University Complutense, Madrid, Spain), Dr. Katheleen Gardiner (University of Colorado, Colorado), and Dr. Krzysztof J. Cios (Virginia Commonwealth University, Virginia), and are available through the UCI Machine Learning Repository.

The mice in the study underwent one of two learning treatments and were treated with or without memantine, an antagonist for the N-methyl-D-aspartate (NMDA) receptor. The NMDA receptor is important for regulating the activity of neurons, and memantine has previously been shown to rescue the performance of Ts65Dn mice in learning and memory tasks.

The dataset includes 8 groups of mice (105-150 mice per group). Wild-type and Ts65n mutant mice were given injections of either saline or memantine. Fifteen minutes after injection, the mice were placed in one of two behavioral assays. One of the assays was the context-shock (CS) assay, where the mice were placed in a novel cage, allowed to explore for 3 minutes, and then given an electric shock. These mice learned to associate the new environment with the aversive stimulus. The other assay was the shock-context (SC) assay, where the mice were placed in the novel cage, immediately given a shock, and then allowed to explore for 3 minutes. These mice do not develop a conditioned fear like the CS groups. From these assays, the groups of mice can be divided into those that learn and those that do not learn:

Mice that learn	**Mice that do not learn**
CS WT saline	CS Ts65Dn saline
CS WT memantine	SC WT saline

Mice that learn	Mice that do not learn
CS Ts65Dn memantine	SC WT memantine
	SC Ts65Dn saline
	SC Ts65Dn memantine

From the 8 groups of mice (1,080 in total), the levels of 77 proteins or protein modifications (attributes) were determined to give 83,160 observations, with 1,346 (approximately 1.6%) of the values missing.

The variables can be defined as the following:

Variable	Type	Category details
Mouse genotype	Categorical	control or Ts65Dn mutant
Drug treatment	Categorical	saline or memantine
Behavioral assay	Categorical	CS or SC
Learning	Categorical	Learn or do not learn
Protein expression levels	Continuous	

The experimental approach and resulting data

Data Wrangling

To begin the data wrangling process, a “class_number” variable was created. Each class was assigned a number between 1 and 8 based on its class, which incorporated the mouse’s genotype, behavioral test given, and whether or not the mouse was given the drug.

' *Ehh gpewvcryq f i vgspyq r*

```
learning1 <- learning %>% mutate(class_number = class)
```

' *Evwnkr Gpewvi wxs '51<*

```
learning1$class_number[learning1$class_number == "c-CS-m"] <- "1"
learning1$class_number[learning1$class_number == "c-SC-m"] <- "2"
learning1$class_number[learning1$class_number == "c-CS-s"] <- "3"
learning1$class_number[learning1$class_number == "c-SC-s"] <- "4"
learning1$class_number[learning1$class_number == "t-CS-m"] <- "5"
learning1$class_number[learning1$class_number == "t-SC-m"] <- "6"
learning1$class_number[learning1$class_number == "t-CS-s"] <- "7"
learning1$class_number[learning1$class_number == "t-SC-s"] <- "8"
```

Using the class number, another new variable was created to indicate whether or not the mouse learned, which is the read out of the experimenters’ learning and memory assay.

' *Ehh pievrmk gspyq r*

```
learning1 <- learning1 %>% mutate (learning = class_number)
```

' *Ewwnkr 5 Apievr 4 Ahs rsxpievr*

```
learning1$learning[learning1$learning == "1"] <- "1"
learning1$learning[learning1$learning == "3"] <- "1"
learning1$learning[learning1$learning == "5"] <- "1"
learning1$learning[learning1$learning == "2"] <- "0"
learning1$learning[learning1$learning == "4"] <- "0"
learning1$learning[learning1$learning == "6"] <- "0"
learning1$learning[learning1$learning == "7"] <- "0"
learning1$learning[learning1$learning == "8"] <- "0"
```

As mentioned, the data set included 1,346 (approximately 1.6%) NA values. To avoid removing these observations and attributes completely, the mean of each class was imputed for the columns that ended with '_N', which are the individual proteins whose levels were measured. This would provide a representative value for the protein being measured dependent upon each class of mouse examined.

' *vitpegi RE zepyiw{ nkl qier sjxli gswiwtsrhmrk gpewn2gpewncryqfiv*

```
learning2 <- learning1 %>% group_by(class) %>%
  mutate_each(funs(replace(., which(is.na(.)), mean(., na.rm=TRUE))), ends_with('_N'))
```

From these wrangling steps, the data set has been cleaned up to begin statistical analyses with the potential for more data wrangling based on the output of the statistics.

Statistical analysis and initial findings

To begin addressing whether there are significant differences in the levels of specific proteins between the different classes examined, the mean protein level was determined for each protein.

The resulting mean values for each protein in one particular class were plotted against those of another class in a scatter plot matrix (Figure 1) to investigate whether or not there is a positive linear relationship between the different classes.

If there is a positive linear relationship, this suggests that the protein levels are not different between the classes, which vary by genotype (wild-type vs. mutant), behavioral test given, drug treatment (drug vs. no drug, or any combination of these. If there are mean values that lie outside the positive linear relationship, this may suggest that the potentially significant difference between the protein levels is a result of the variation(s) between the classes examined.

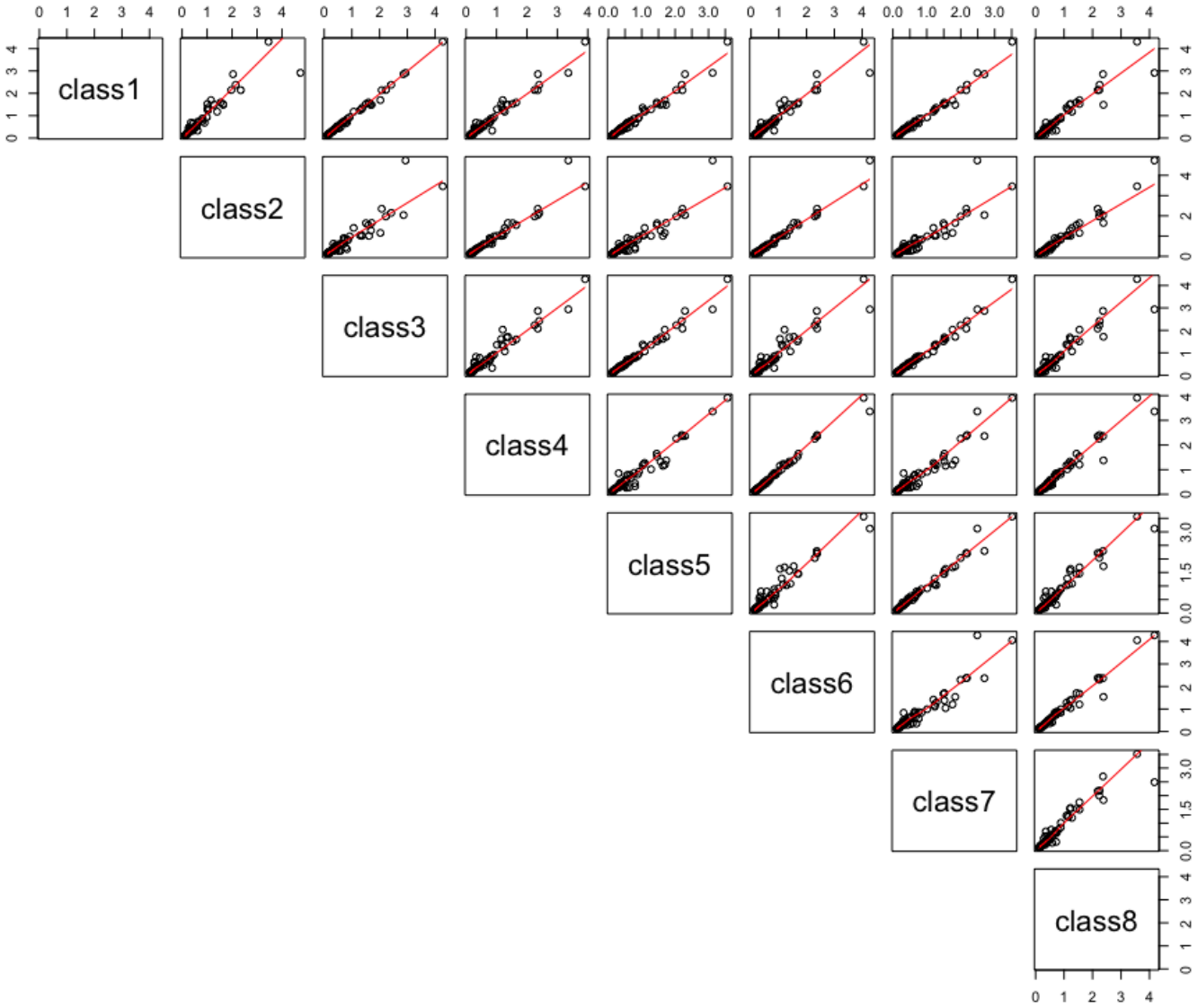

Figure 1. Scatter Plot Matrix by Class

From the resulting plot, there appears to be several outliers, or proteins that vary in their mean levels between the classes.

This leads to the following questions:
1. Which protein levels are different between the classes?
2. Are these differences statistically significant?

To address these questions, more data wrangling was performed. The proteins that display a high correlation to one another would theoretically give the same output in comparison with other proteins. Because of this, collinearity was examined, and a correlation matrix between the mean protein levels of each class was generated.

A correlation plot (Figure 2) reveals the correlation values for each of the proteins, where values close to 1 (dark blue) represent a high positive correlation, values close to 0 (white) represent little to no correlation, and values close to -1 (red) represent a high negative correlation.

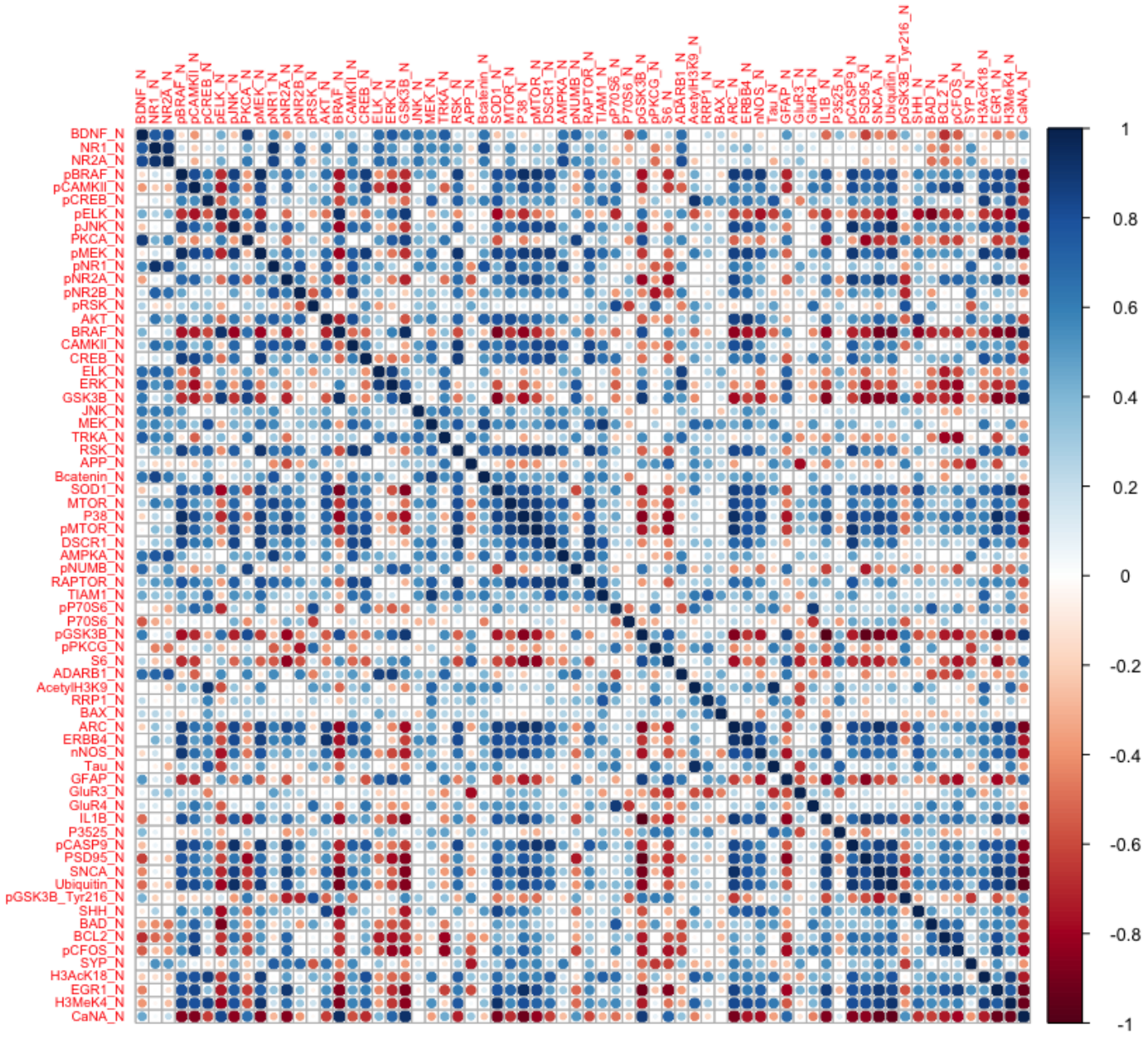

Figure 2. Correlation Plot

The combinations of proteins that have a correlation of 0.95 or greater were identified.

In total, 14 combinations were extracted, and from these combinations, 9 proteins were removed.

After removing all columns except class-number and each protein tested, a repeated measures ANOVA was performed to examine whether the protein levels between classes 1-8 are significantly different. The results were also plotted as a box plot for each protein. From the plots, some ***potential*** general trends were observed:

- Little to no differences between classes

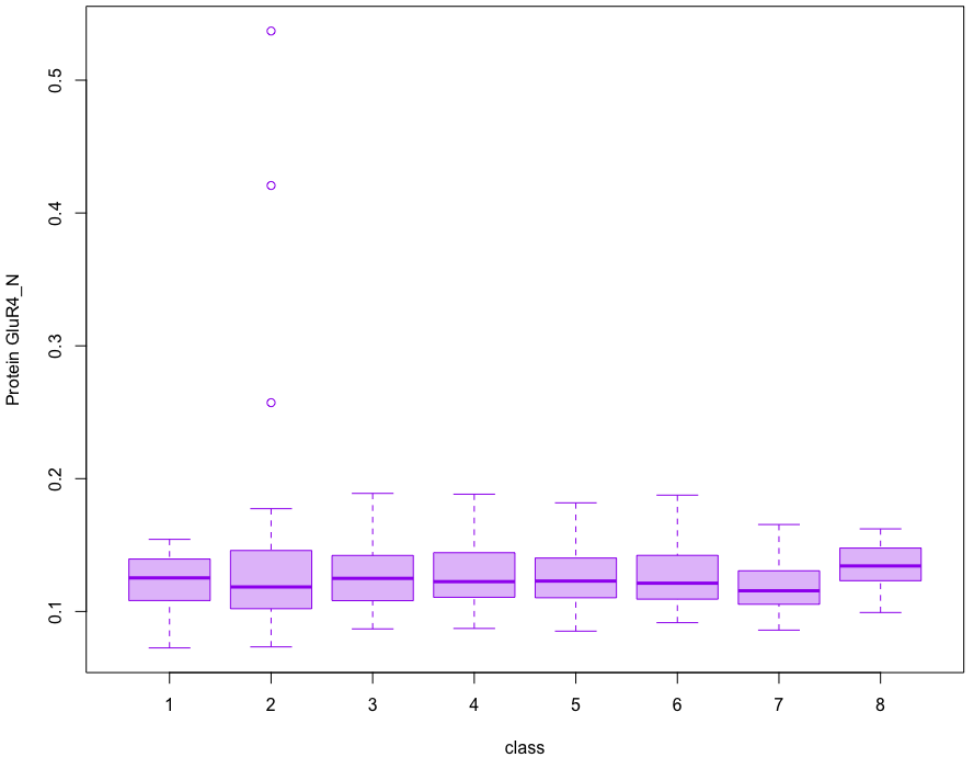

Figure 3. Little to no difference between classes

- Differences between genotype

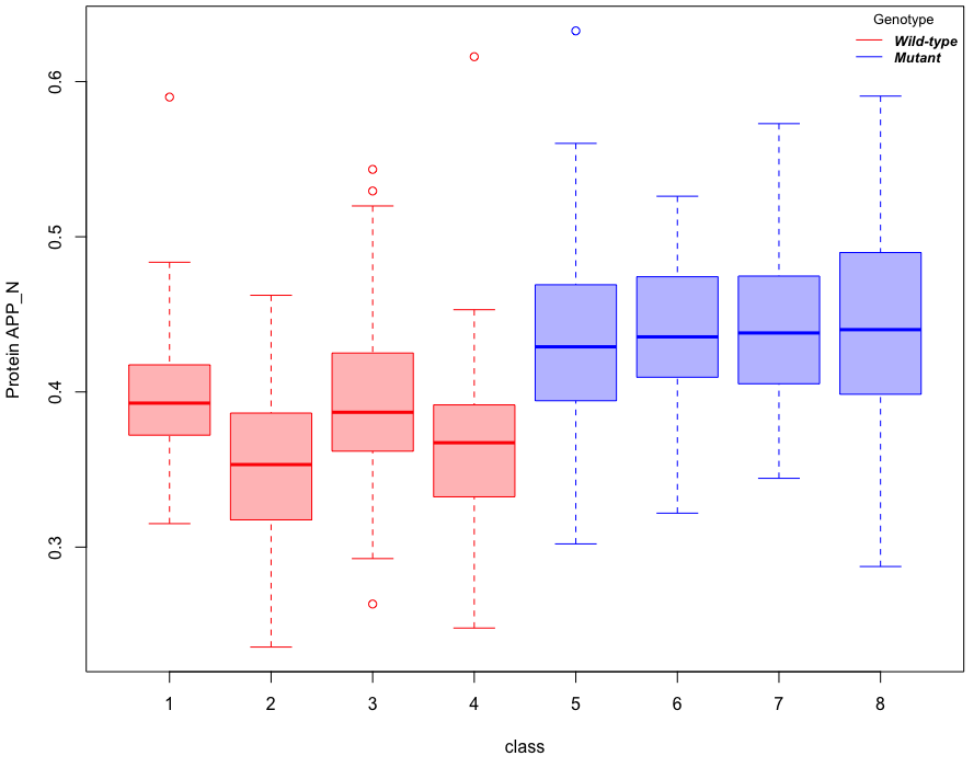

Figure 4. Differences between genotypes

- Differences between behavior test given

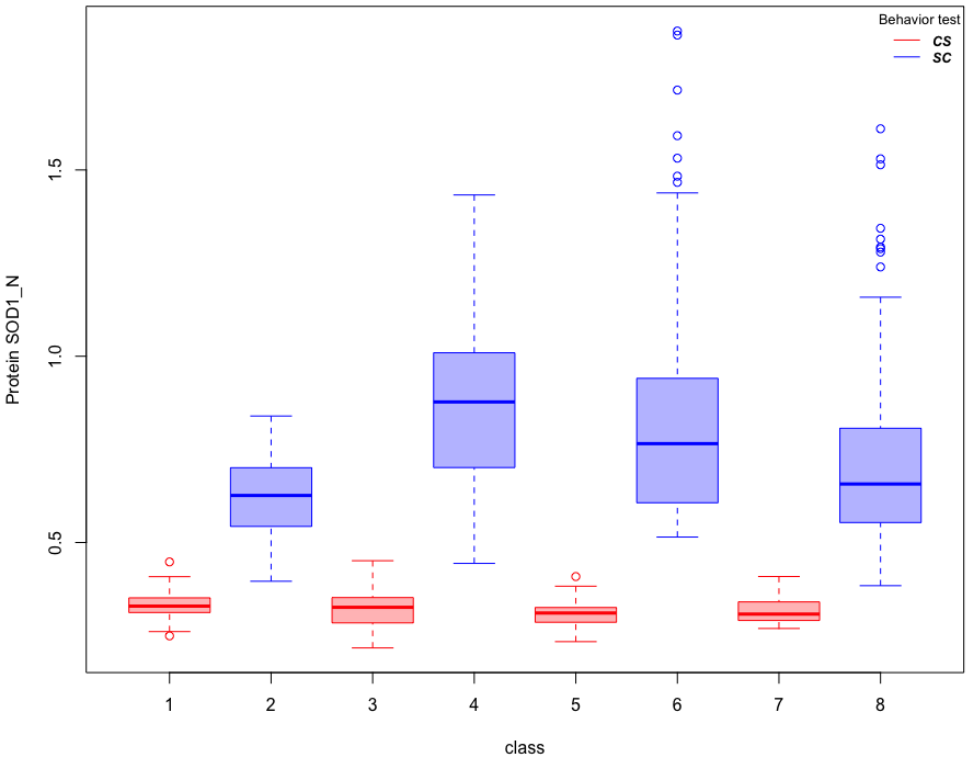

Figure 5. Differences between behavior test given

- Differences between drug given and not given

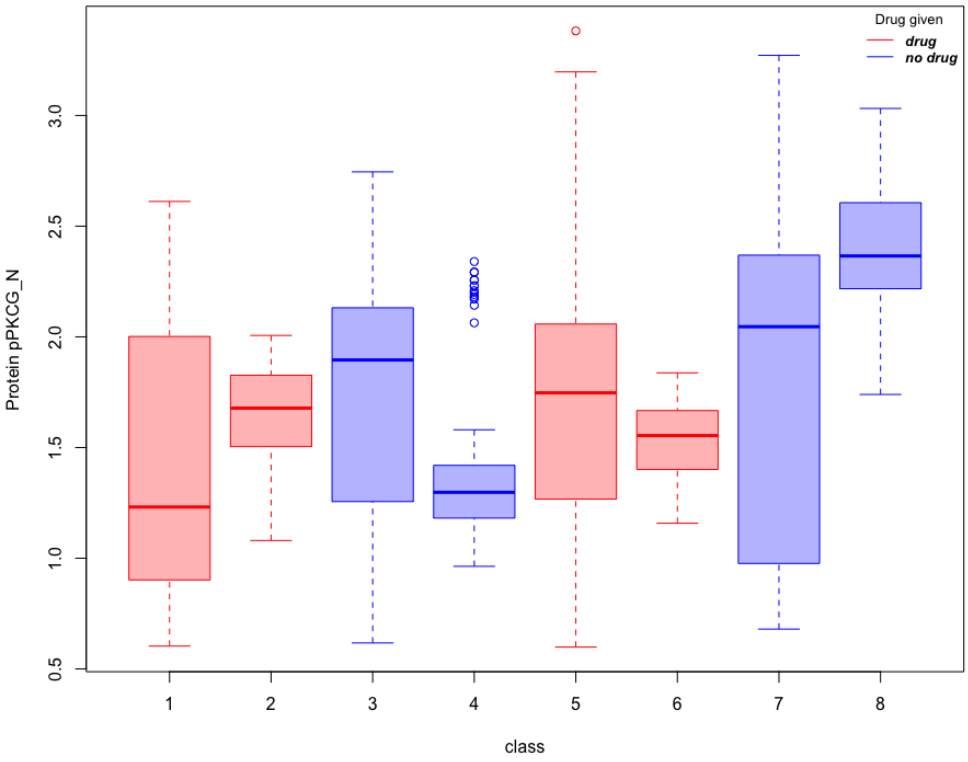

Figure 6. Differences between drug and no drug. For example, Class 6 and Class 8 differ only by whether or not the drug was provided.

The p values for the comparisons in protein levels are between either 0 and 0.001 or 0.001 and 0.01.

Together, these analyses suggest that there are significant differences between at least 2 of the classes for each protein examined. This leads to further questions of whether or not a specific protein and its expression levels can be predicted within one of the 8 classes of proteins. Moreover, it would be interesting to determine and predict the proteins that are affected by particular drugs within individuals of varying genetic and behavioral backgrounds.

Applying machine learning to predict experimental class

Are there specific relationships between the proteins examined and the various classes of mice? Given the findings above, a model was generated that investigates the relationship between the expression levels of a particular protein and each class of mice examined. Supervised learning was employed using logistic regression analyses with the goal being to come up with a low-complexity model that can determine whether the expression level of a particular protein is from a specific class of mice. This was done by fitting a regression curve for the categorical variable class/class_number.

The analyses focused on the protein levels as being the independent variables and the class number as the dependent variable, as the aim is to be able to predict the categorical variable (class number) based on the expression levels of the proteins.

A two-parameter logistic regression model was utilized. This model assumes that the protein expression levels will account for the class of mice that display those levels. The model also uses combinatorial modeling to determine the optimal goodness-of-fit, where all combinations of proteins are examined for each class number, except when the proteins are in combination with themselves. In addition, this model utilizes one hot encoding by creating dummy variables to replace the values 1-8 in class_number with either 0 or 1, where a “1” would indicate that it is in that particular class and “0” would indicate that it is not within that class.

To test the machine learning technique, the models were run with the proteins that give the lowest AIC for each class, where the AIC measures the goodness-of-fit of the model. The lower the AIC value, the better the model is. If there is a close to perfect fit, then it would be suggested that the measured protein levels would indicate a particular mouse to be categorized in a specific class. From testing the models, the observed and predicted values were calculated for each class, where any value greater or equal to 0.5 would be considered a “1”, and anything less than 0.5 would be “0”.

The values were plotted for each class to visualize the comparison between the observed and predicted values

- Class 1

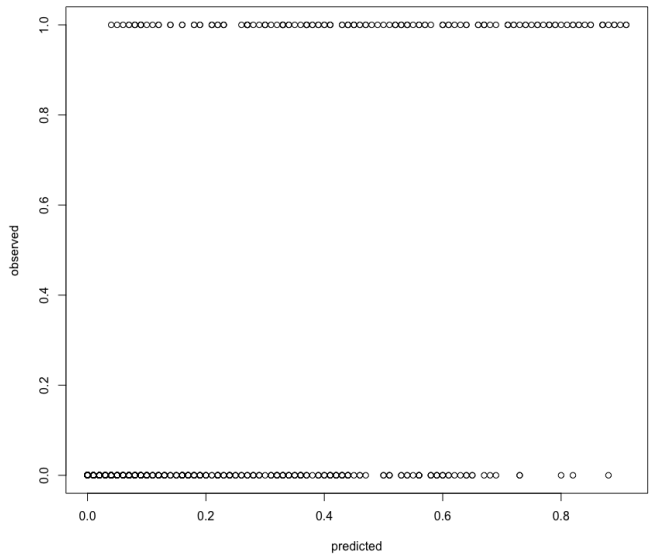

Figure 7. Class 1.

- Class 2

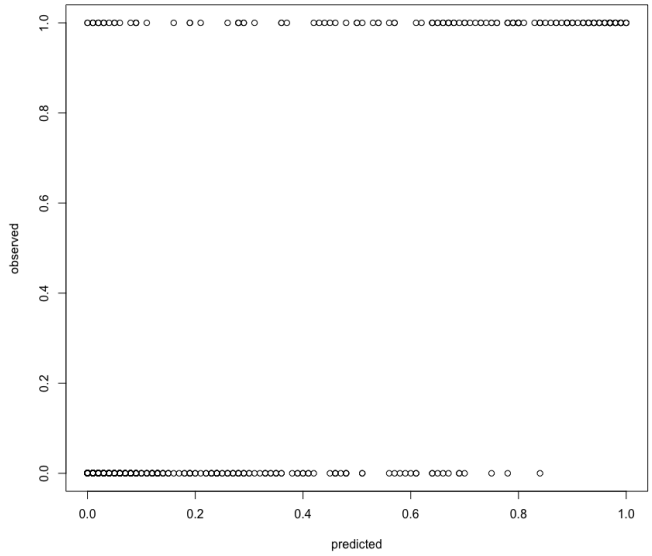

Figure 8. Class 2.

- Class 3

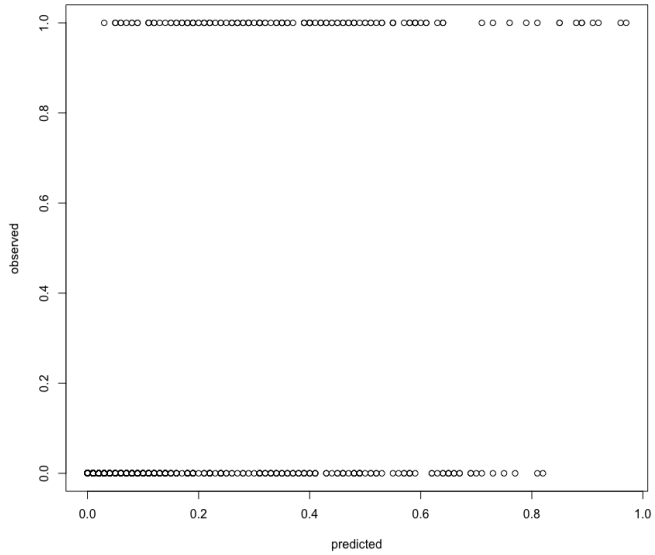

Figure 9. Class 3.

- Class 4

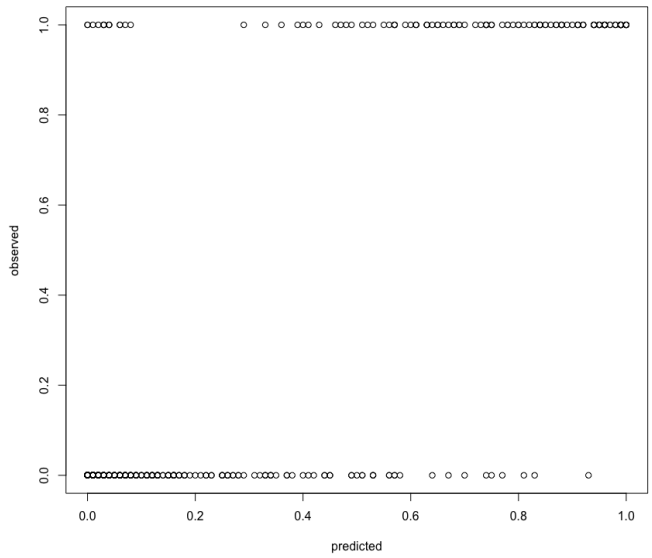

Figure 10. Class 4.

- Class 5

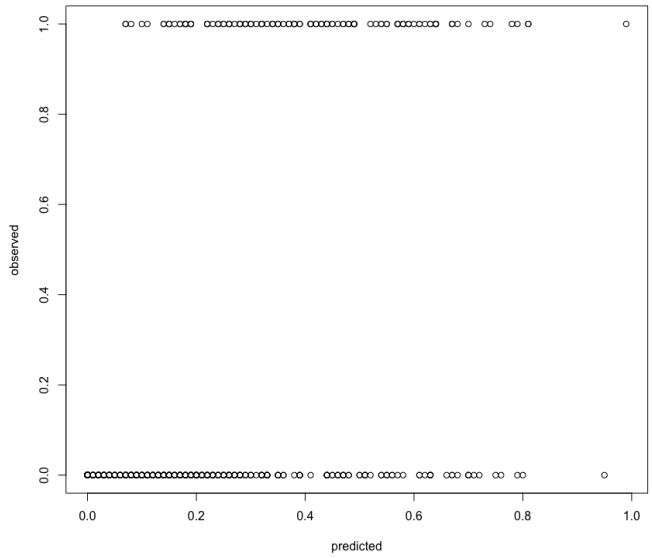

Figure 11. Class 5.

- Class 6

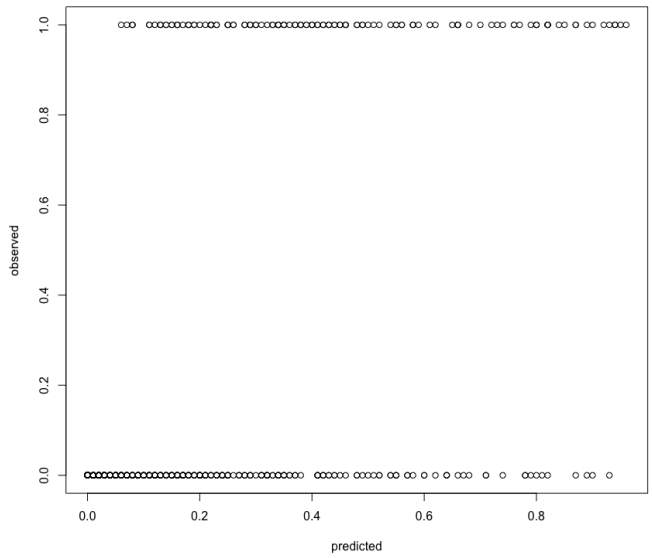

Figure 12. Class 6.

- Class 7

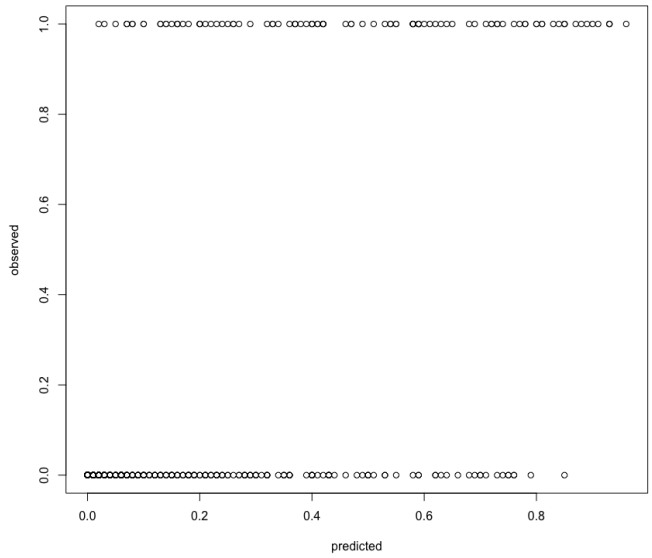

Figure 13. Class 7.

- Class 8

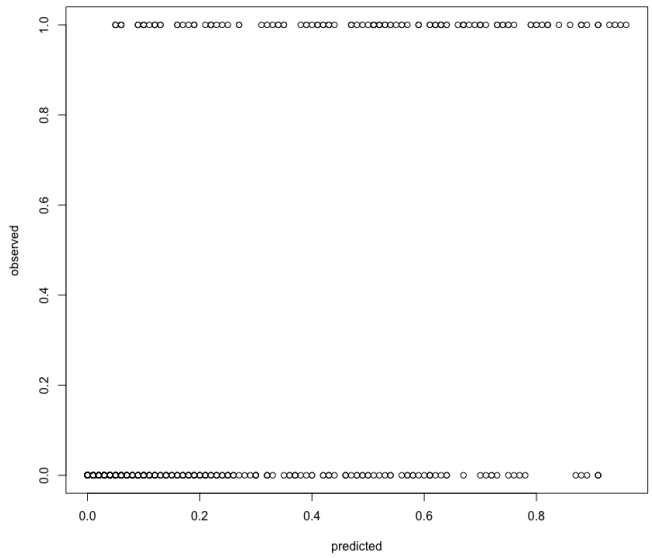

Figure 14. Class 8.

The true positive, true negative, false positive, and false negative values were determined, and the accuracy of the model was calculated ((true positive + true negative)/total). The accuracy values for each class are:

Class	Accuracy
1	0.887
2	0.944
3	0.877
4	0.957
5	0.887
6	0.888
7	0.924
8	0.908

From this, it can be suggested that the measured expression levels of two proteins can differentiate whether or not the expression level values belong within a class or not, as the model for each class displays high accuracy.

Identifying specific proteins that are affected by therapeutic drug in Ts65Dn mice

Being able to identify specific proteins that are affected by drug treatment may provide an avenue of exploration for future therapeutic targets of diseases and disorders. The data was subset to identify proteins in Ts65Dn mutant mice given the CS behavior, which allows for learning to occur, that vary in expression levels with (Class 5) and without (Class 7) the memantine drug. This would provide possible protein targets that promote learning in Ts65Dn mutant mice and potentially those affected by Down Syndrome.

Proteins that exhibit a statistically significant difference in expression levels between these mutant mice were identified using t-test analyses. From this analysis, 36 proteins were extracted.

It is also plausible that the drug would affect the control mice similarly to that seen in the mutant mice. To determine which of these 36 proteins are specific to mutant mice, proteins that display a statistically significant difference in expression levels between wild-type mice given the CS behavior with (Class 1) or without (Class 3) the memantine drug were identified. From this analysis, 30 proteins were identified.

From these 2 lists of proteins, 19 of the proteins are unique to being different in the Ts65Dn mutant mice (Class 5 and Class 7).

These analyses identify potential therapeutic targets for ameliorating learning and memory deficits in individuals affected by Down Syndrome. Some of these potential targets are:

1 - The AMPA receptor GluR3 – Synaptic plasticity, or the ability of the neurons to alter their synaptic strength or synaptic activity contributes to learning and memory. AMPA receptors play an important role in regulating this synaptic plasticity, where increased levels of the receptors enhance synaptic plasticity.

From the analyses here, Ts65Dn mutant mice subjected to the behavior that promotes learning and given no drug display a reduction in GluR3 levels compared with wild-type mice of the same behavior and drug treatment (see Figure 7: Class 7 purple vs. Class 3 orange). When these mutant mice are given the memantine drug, there is a rescue in GluR3 levels, as they increase to be more similar to wild-type mice (see Figure 7: Class 7 purple vs. Class 5 blue). This suggests that GluR3 may enhance synaptic plasticity in Ts65Dn mutant mice.

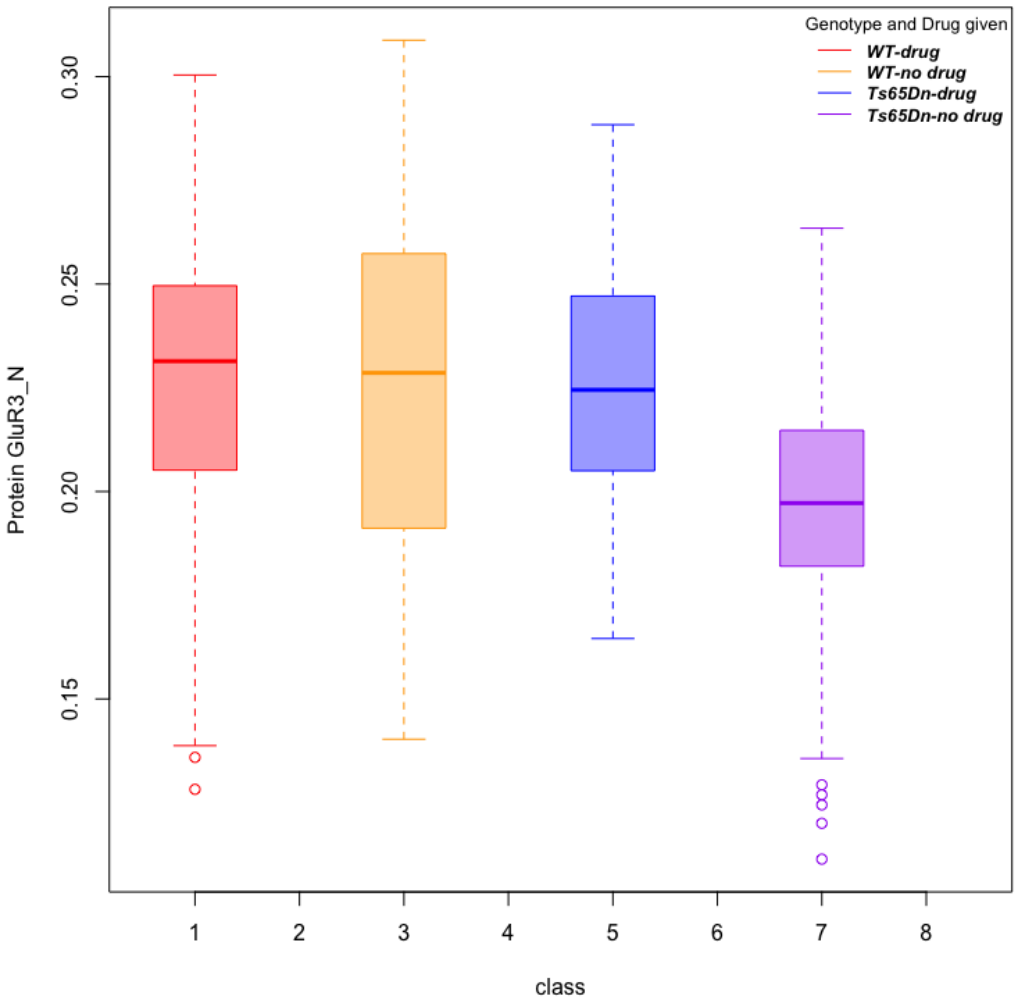

Figure 7. GluR3 protein expression levels are rescued in Ts65Dn mice given memantine

2 - The transcription factor CREB – The cAMP responsive element binding protein (CREB) is a protein found within the cell nucleus and modulates the transcription of target genes. The targets of CREB influence learning and memory. It has been found to have roles in long-term memory and other forms of memory, including spatial and social learning.

The analyses performed here reveal that Ts65Dn mutant mice subjected to the behavior that promotes learning and given no drug display a reduction in CREB expression levels (see Figure 8: Class 7 purple vs. Class 3 orange). After memantine exposure, there is an increase in CREB expression levels, similar to those observed in wild-type mice (see Figure 8: Class 7 purple vs. Class 5 blue). This suggests that CREB may promote learning and memory in Ts65Dn mice.

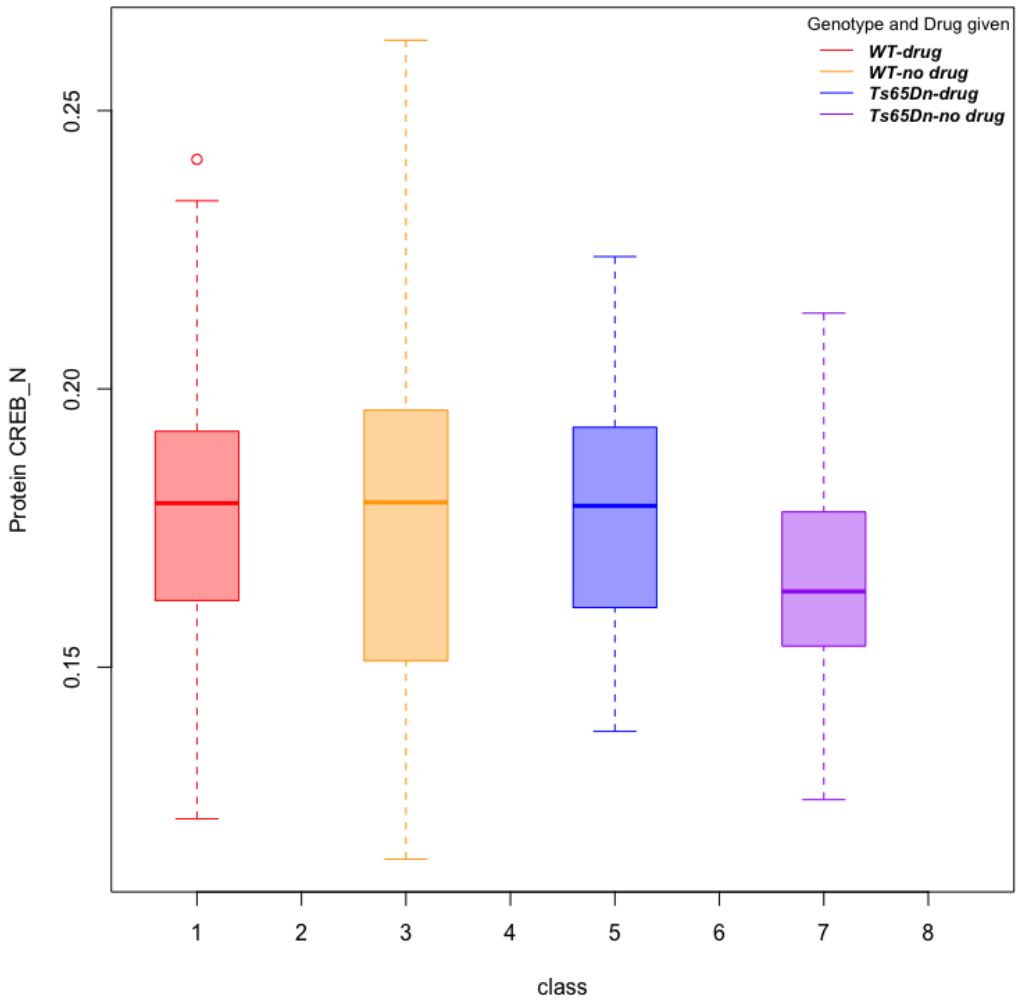

Figure 8. CREB protein expression levels are rescued in Ts65Dn mice given memantine

Future directions and recommendations

The analyses and findings described above provide a model for predicting which experimental group the expression levels of a particular protein belong to. In addition, they identify specific proteins that may serve as therapeutic targets for learning and memory in those individuals affected by Down Syndrome.

While the data set provides measurements of 77 different proteins, it does not include information regarding the biological role of these proteins. By knowing which proteins function within the same or parallel biological pathways with other proteins, it is plausible to find a relationship (positive or negative) between the measured levels of the protein in the different classes of mice tested. This may provide useful to hypothesize whether a protein that was not examined would be affected by genotype, drug treatment, or behavioral assay. However, it is noted that novel proteins are continually discovered to play a role in defined biological processes, so the data set and analyses would need to evolve with these new discoveries.

Further research could expand the analyses and models performed on the data set. Here, logistic regression was used on the data set, but having another data set to use as a “test” in the machine learning analysis, would have proven useful in testing the model derived from the original “training” data set.

In addition, identification of potential therapeutic targets focused on 2 classes in this analysis. Future exploration could focus on the relationships between other classes to determine which proteins may be involved in learning and memory function in individuals with Trisomy 21. For example, how do the protein expression levels differ in Ts65Dn mice given the memantine drug and put through the context-shock (CS) learning behavior (Class 5) compare to the same mutant mice with drug exposure given the shock-context (SC) non-learning behavior (Class 6)?

The findings from the analyses described could prove to be useful for future research, including:

- Pursuing research on novel pharmacotherapies for learning and memory within the lab using Ts65DN mice followed by clinical trials in individuals with Trisomy 21. Here, 2 proteins of interest were described (GluR3 and CREB), but through further analysis, more protein targets could be identified.
- With other data sets, the analyses could be modified to predict and identify proteins that affect other developmental disorders and diseases.