Data Mining in Proteomic Database from Depressed Brain Tissue

Michelle Chen, Mert Ketenci, Jung Suk Lee,
Aishwarya Vedantaramanujam Srinivas, Yimin Wang

Dr. Maura Boldrini, Dr. Hanga Galfalvy, Dr. Lewis Brown

Abstract

One person dies due to suicide every 40 seconds globally. Major depressive disorder (MDD) is projected to become the leading disease burden globally by 2030. The pathogenesis of depression and suicide is unclear and there is a high non-response rate to the most used selective serotonin reuptake inhibitors antidepressants (SSRI). We leveraged statistical analysis and machine learning to investigate 36 postmortem human hippocampi and reveal proteins differentially expressed, as measured by shot gun proteomics mass spectrometry, in non-psychiatric controls, suicide and non-suicide MDD, and between SSRI-treated and untreated MDD. All subjects underwent psychological autopsy, brain neuropathology and toxicology exams. We used the non-parametric Benjamini-Hochberg procedure with a maximum false positive rate of 5%. In this pilot study, we found significant differences in protein levels between control, treated, and untreated MDD, and among suicide and non-suicide MDD. Significant differences in expression of: RB6I2, DIRA1, CN166, MA2A1, PRS8, HPT, HS71L, and VGF were found between control vs. MDD; while, DIRA1, PRS8, RB6I2, ATD3A, MA2A1, and ENLP differentiated suicide vs. non-suicide MDD (all $p<0.05$). Higher concentrations of RB6I2 in non-psychiatric controls versus MDD ($p=0.000070$) suggests alterations in neurotransmitter release at nerve terminals in MDD. Lower ATD3A in suicide versus non-suicide MDD ($p=0.000853$) reveals mitochondrial protein synthesis gap which may adversely affects ATP production downstream and worsen lethargy symptoms. This proteomics and data mining approach helps uncover molecules involved in the pathogenesis of MDD and suicide and identify new molecular treatment targets to develop novel drugs to treat these severe conditions.

Introduction

According to the World Health Organization, one person dies due to suicide every 40 seconds globally. As suicide rate climbs with over 16 per 100,000 individuals dying by suicide annually, Major depressive disorder (MDD) is projected to become the leading disease burden globally by 2030. The pathogenesis of depression and suicide is not clear and there is a high non-response rate for current solutions involving antidepressants. Thus, under the mentorship of Dr. Maura Boldrini and Dr. Hanga Galfalvy, our team leveraged brain proteomics data to understand at a molecular level how proteins are differentially expressed in individuals with MDD, MDD with treatment, and control, as well as depressed individuals who died by suicide versus those who did not.

One of the major challenges for us during this project was the small sample size of our dataset, which precluded us from performing large scale machine learning for predictive analysis. However, the benefit was that we were able to take a more exploratory approach and examine factors such as ‘Age’, ‘Gender’, and ‘GAS scores’, which may have serious implications regarding the development and worsening of MDD diagnoses within patients. In addition to quantifying the relationship between depression and certain protein levels, we are also interested in uncovering the biological roles of these proteins and how they may affect processes downstream. Using our analysis, we hoped to reveal meaningful insights into how exactly depression manifests itself from a proteomics perspective and how other demographics-related features may play a significant role as well.

Related Work & Your Contributions

Due to the biological and medical-heavy emphasis of this Capstone project, our team familiarized ourselves with biological procedures and terms by reading several research papers below. The first paper we referenced was ***“Neuropsychological function and suicidal behavior: attention control, memory and executive dysfunction in suicide attempt”*** **by Keulp et al., 2012**.[1] Subjects who had attempted suicide in the past were observed having deficits in specific components of attention control, memory and working memory when asked to perform motor and language fluency tasks, compared to healthy subjects. This helped us realize that there may be some biological and chemical changes which occur in those who attempt suicide that drastically alter their normal behavior.

The hippocampus is the part of the brain which is highly involved with the formation of memories and mood-regulation, which were found to be deficit in subjects that attempted suicide. One of our mentors, Dr. Boldrini was interested in seeing differences in protein expression in the Dentate Gyrus (an area within the hippocampus) of suicide versus non-suicide as well as group-wise MDD, MDD*SSRI, and control groups. We referenced many of her prior papers which involved the DG and were related to our specific Capstone topic including: **“*Resilience Is Associated With Larger Dentate Gyrus, While Suicide Decedents With Major Depressive Disorder Have Fewer Granule Neurons”*** **by Boldrini et al., 2019 [2]**, which details how early life adversity increases MDD and suicide risk and potentially affects DG plasticity. The paper reported smaller DG and fewer granular neurons in MDD samples.

The last paper we referenced was called ***“Molecular Aging in Human Prefrontal Cortex Is Selective and Continuous Throughout Adult Life”*** **by Erraji-Benchekroun et al., 2005**. We were recommended by Dr. Boldrini to read this paper because it explains the huge role aging plays on the neuro-plasticity of the brain over time. Drawing inspiration from the paper, we were able to investigate whether some proteins undergo age-related transcriptional changes that can influence suicidal behavior in MDD, MDD*SSRI patients (See 'Observing effects of Aging on MDD and suicidal behavior' section above). We also referenced this paper to learn what was the optimal way to visualize protein expression and interactions amongst our variables.

Our contribution to this brain proteomics topic includes leveraging various statistical analysis and natural language processing techniques to uncover 11 key proteins which play an influential role in ATP production and neurotransmitter regulation. Using this hybrid technical and biological approach, we can form more robust conclusions about our data and uncover novel avenues in which the research community can leverage to learn more about MDD.

Discussion of dataset

The dataset we worked with consists of two parts: 1) post-mortem samples measuring various protein levels collected from the Dentate Gyrus (DG) of 36 individuals, 12 clinically-diagnosed with MDD, 12 diagnosed with MDD and also administered selective serotonin reuptake inhibitor (SSRIs) / antidepressant treatment, and lastly 12 control individuals and 2) demographic data for each of the 36 individuals. For the first dataset, the type of protein (1131 total) measured was listed as each row while columns were designated as the individual samples. Meanwhile, the demographics dataset provides information related to the Group (Control, MDD, MDD*SSRI), sex, age, ethnicity, and GAS (Global Assessment of Functioning).

Our data was presented in a long format, where we have a row for each patient and protein combination. In order to leverage the standard data analysis tools, we converted the data to a wide format where we have a patient per row and proteins on the columns. Other data preparations such as deleting spaces, lowercasing the strings, and finding erroneous data were done. In particular, there was one data point, a protein of a patient with a value of 0. This was replaced by null as opposed to an average. Numeric data was also log adjusted as advised. For each of the 36 post-mortem brains, since the lab obtained two technical replicates from each of the initial samples and already adjusted for the differing volumes of the replicates, we averaged the two measures to create just one observation per patient brain.

During the Natural Language Processing portion of our research, we also worked with protein function text data which had been scraped from uniprot.org.

Exploratory Data Analysis

Feature Selection Methods

A major portion of this research project was focused on deriving a set of key proteins that were useful in differentiating suicide versus non-suicide samples, as well as depressed vs. non-depressed samples. In order to identify such proteins, the team employed an array of feature selection techniques: T-tests, Random Forest Classifier, PCA, and Nonparametric tests such as Kruskal-Wallis and Benjamini-Hochberg. The T-test enabled us to come up with 80 important proteins in 95% confidence interval, not adjusted for multi-testing. SYFA_HUMAN, ATD3A_HUMAN and MBOA7_HUMAN were among the highest 3 (Refer to Figure 1.)

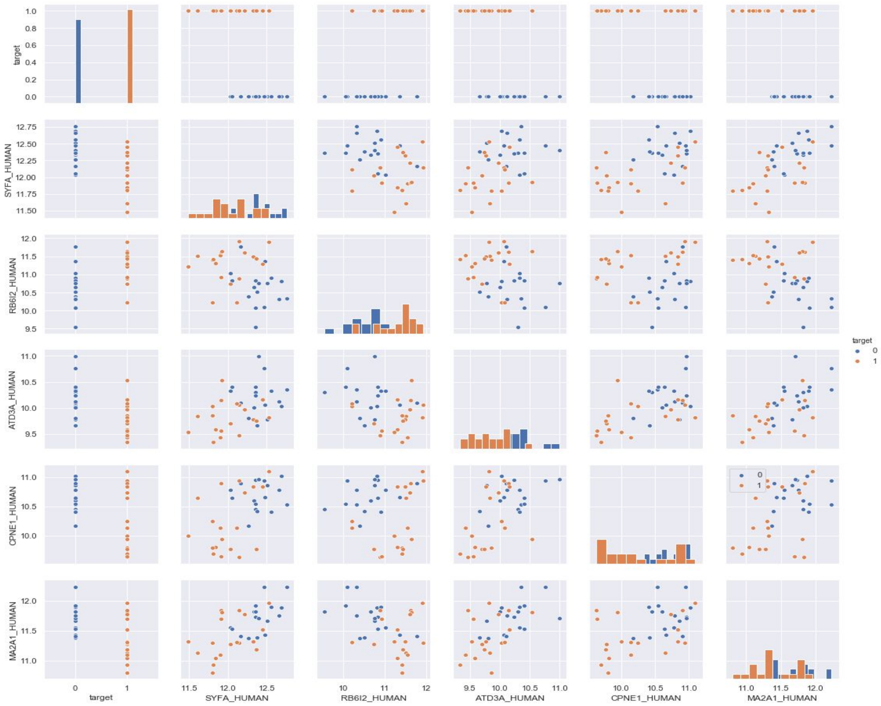


Figure 1. T-Test Visualization

Using the Random Forest Classifier, over 175 proteins were identified to have feature importance greater than 0. The plot shows top most 10 important proteins. CPNE1_HUMAN was identified to have the highest importance among them all, and we have visualized its distribution as follows (See Figure 2.)

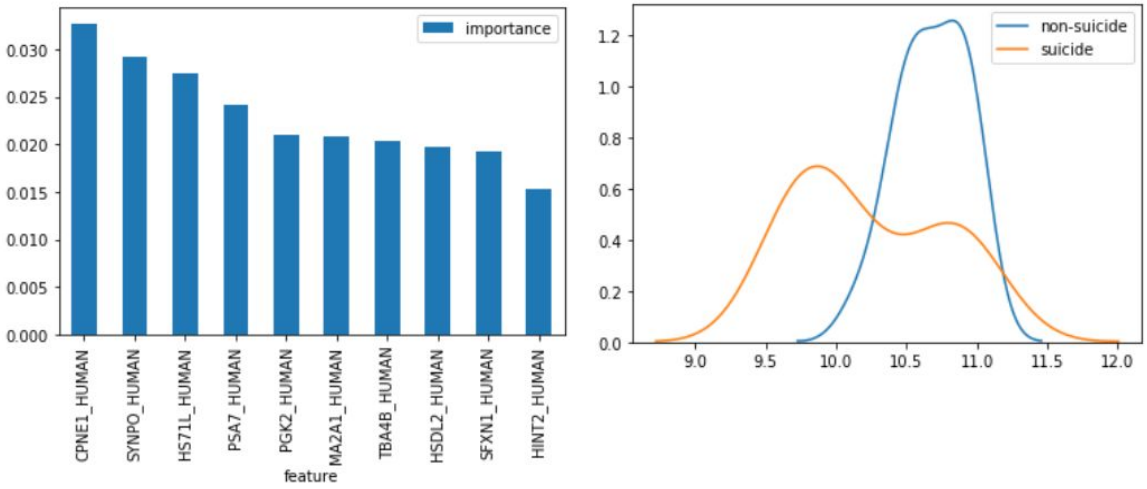


Figure 2. Top 10 proteins using Random Forest Feature Selection

The PCA study that we performed on the data sought to find the lowest number of linear combinations of proteins that would be able to explain most of the variance. We have found that 90% of the variance can be explained by using 20 super proteins that are formed by combining the existing 1134 proteins.

Lastly, we also leveraged non-parametric tests such as Kruskal-Wallis and Benjamini-Hochberg for feature selection. The advantage of nonparametric tests is that they do not assume anything about the distribution of the features. Likewise, since we have less than 30 samples in each of the 3 groups individually, parametric tests to check altered expression of proteins among groups is not applicable. Thus, we used a nonparametric Kruskal-Wallace test to test if there are proteins significantly differing among the 3 groups - MDD, C and MDD*SSRI. Interestingly, the protein VGF_HUMAN is shown to be statistically significant among all groups. Its distribution is shown on the right in the boxplot.

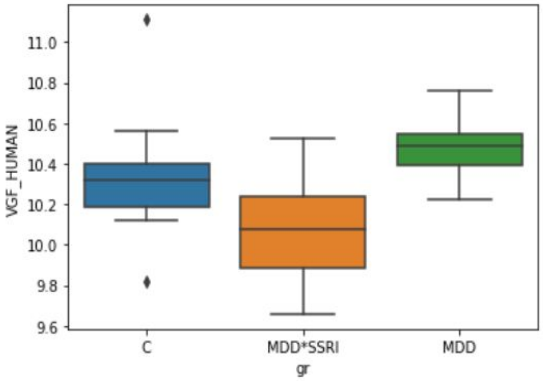
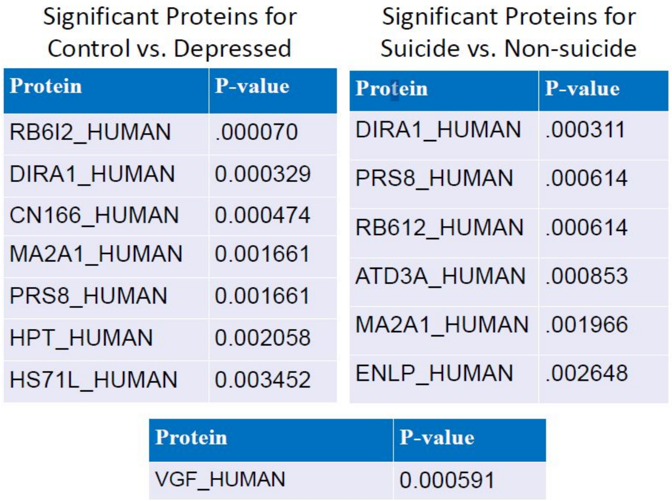


Figure 3. Kruskal-Wallace Results

With thousands of proteins that we are working with, it would be easy to run into significance in our tests by sheer chance. Therefore, we applied two corrective adjustments to our p-values to account for multiple testing. This is to avoid "p-value hunting." Using Bonferroni Adjustment[11], Only one protein, DIRA1_HUMAN, was significantly different between the suicide groups and non-suicide groups.

Bonferroni Correction can be a very harsh adjustment. Therefore, we also leveraged Benjamini-Hochberg Procedure[12]. Procedure focuses on controlling for discoveries that are false i.e. false discovery rate. We uncovered 11 key proteins collectively using the Kruskal-Wallace test and Benjamini-Hochberg procedure.

Table 1. 11 Key Proteins from Feature Selection



Significant Proteins for Control vs. Depressed

Protein	P-value
RB6I2_HUMAN	.000070
DIRA1_HUMAN	0.000329
CN166_HUMAN	0.000474
MA2A1_HUMAN	0.001661
PRS8_HUMAN	0.001661
HPT_HUMAN	0.002058
HS71L_HUMAN	0.003452

Significant Proteins for Suicide vs. Non-suicide

Protein	P-value
DIRA1_HUMAN	.000311
PRS8_HUMAN	.000614
RB612_HUMAN	.000614
ATD3A_HUMAN	.000853
MA2A1_HUMAN	.001966
ENLP_HUMAN	.002648

Protein	P-value
VGF_HUMAN	0.000591

Examining the relationship between GAS levels and Protein Levels

For each individual sample, we were also provided GAS scores which range from 1-100 and used by medical professionals to evaluate the social, occupational, and psychological functioning of individuals. Since this study works with post-mortem data, these scores were derived using information from interviews with close family and friends of the individuals. We are interested in observing how the expression of some of the key proteins we discovered during feature selection changes with fluctuations in GAS score by treatment group and suicide status. Our team was keen on analyzing these GAS scores because prior research and our initial analysis confirms that GAS scores are correlated with MDD diagnosis, in that lower GAS scores are linked with MDD (and MDD*SSRI) and vice versa. We see higher overall GAS scores amongst control samples.

Analysis shows that our subjects gas scores generally fall into the lower and higher end of the spectrum. From a group perspective, Control has the highest mean GAS score, followed by MDD, and MDD candidates who were treated. This is a bit surprising because we expect MDD patients to feel worse due to their condition, while MDD patients who are treated should be improving and overall feel happier. However, MDD candidates also experience greater variance in scores, with one sample having a GAS score of <20, which is even less than the mean MDD*SSRI score. Given this, we wanted to see how key protein expression varies based on GAS. All of the key proteins, with the exception of RB6I2, express an increasing trend. However, when we colored the points by group, we can see a much different story. For instance, with PRS8, we were able to come up with an interesting hypothesis that can be broken down into two parts: treated groups experience increasing protein levels along with increases in GAS while controls and MDD groups experience the opposite. For the majority of the proteins, treatment groups tend to congregate with the lowest GAS scores, followed by MDD, and controls with the highest GAS. This tells us that perhaps the treatment is not as effective at boosting mood and overall life satisfaction. See plots below as an illustration of our points.

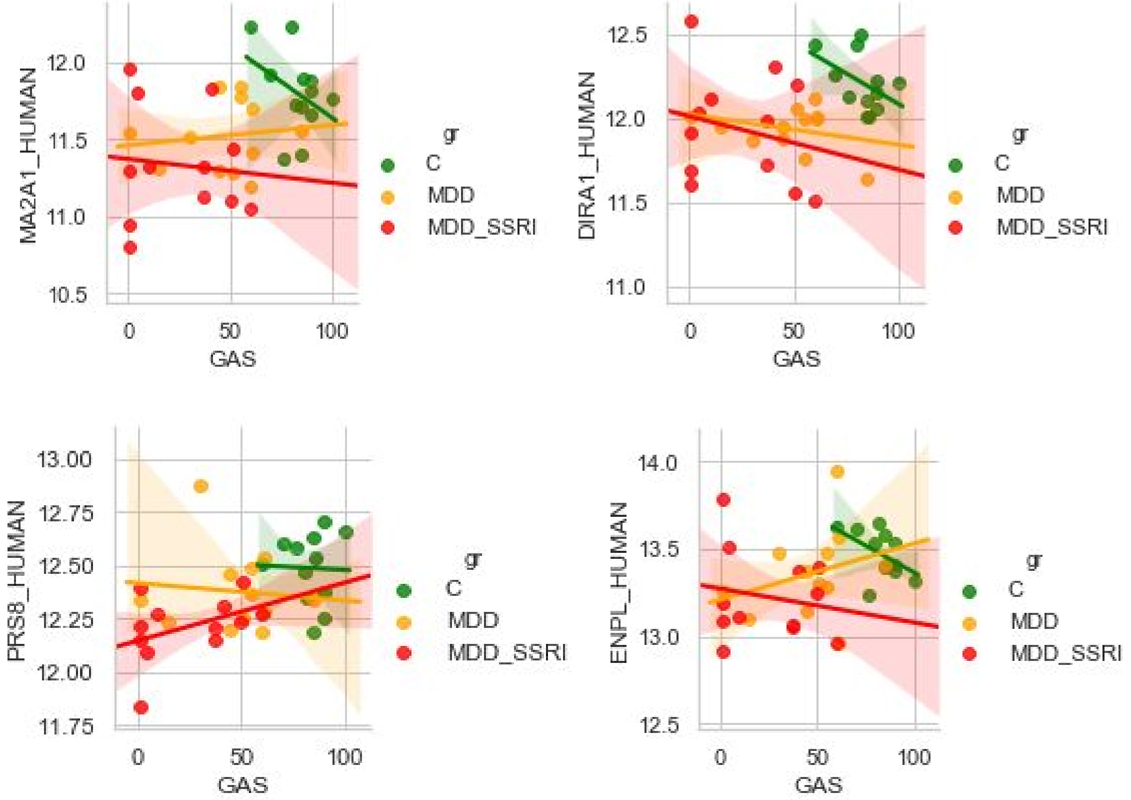
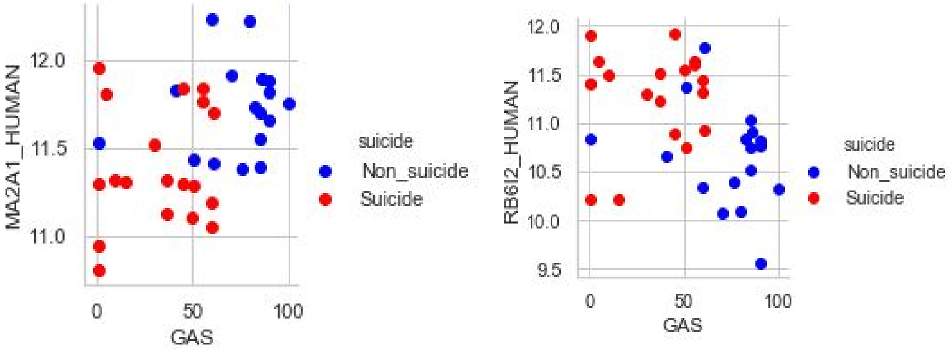


Figure 4. Protein levels vs. GAS Regression Results

In the plot below, we label using different colors, suicide versus non suicide individuals, to see how the protein expressions fluctuate with respect to GAS. We observe that suicide candidates experience lower expression together with lower GAS, with the exception of RB6I2. Please reference the below scatterplots of MA2A1_HUMAN and RB6I2_HUMAN for comparison.

Figure 5. Protein levels vs. GAS Scatterplots



Examining Aging Effects related to Protein Levels

Aging has been shown to lead to morphologic and functional changes in the brain and is associated with an increased risk for psychiatric and neurological disorders that may impair daily functions such as mood-regulation, memory, and cognition (Erraji-Benchekroun et al., 2005). To track these age-related changes in the past, researchers have attempted to profile protein expression in prefrontal cortex samples and successfully identified roughly 500 proteins which underwent transcriptional age-related changes (Erraji-Benchekroun et al., 2005). Our team sought to investigate similar age-related patterns using our proteomics data from the DG, which is located in the hippocampus, an area highly involved with learning, memory-creation, and mood-regulation.

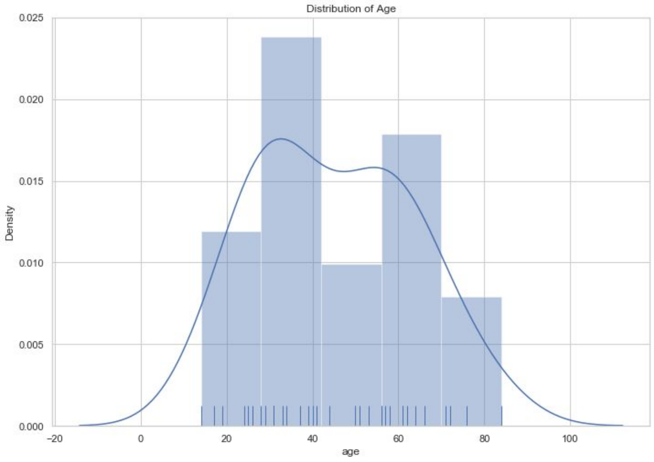


Figure 6. Distribution of Age of samples

The distribution of age amongst post-mortem samples is a bimodal distribution with the greatest density of ages occurring between 27-42 years old and 57-70 years old (Figure 6.). Using the protein features derived during the feature-selection phase, we plotted the protein expression along the y-axis and age along the x-axis. Below, we plotted a key protein discovered from the nonparametric analysis and colored the data points based on group (Figure 7.)

Some key observations are that, it protein expression levels of RB6l2, all increase with age as illustrated with three separate lines by groups. Using Ordinary Least Squares regression, we are left with significant coefficients (p value = 0 < 0.05) for both MDD and MDD*SSRI variables which are 0.8141 and 1.0138. However, when we look at the interaction term between age and group, we see that the interaction is not significant for any of the groups (p > 0.05). We can also further facet the axes by ‘sex’. Looking at RB6I2 again, we see that there are wider confidence intervals meaning higher variance of expression for females versus males. The confidence intervals for males is more tight perhaps due to the larger sample size for males and how standard error is typically inversely proportional to group size. (See figures below.).

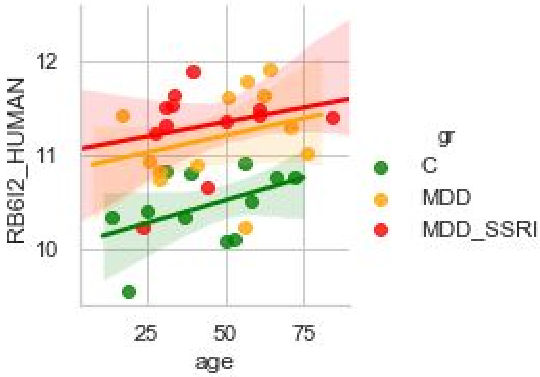


Figure 7. OLS for RB6I2 levels vs. Age

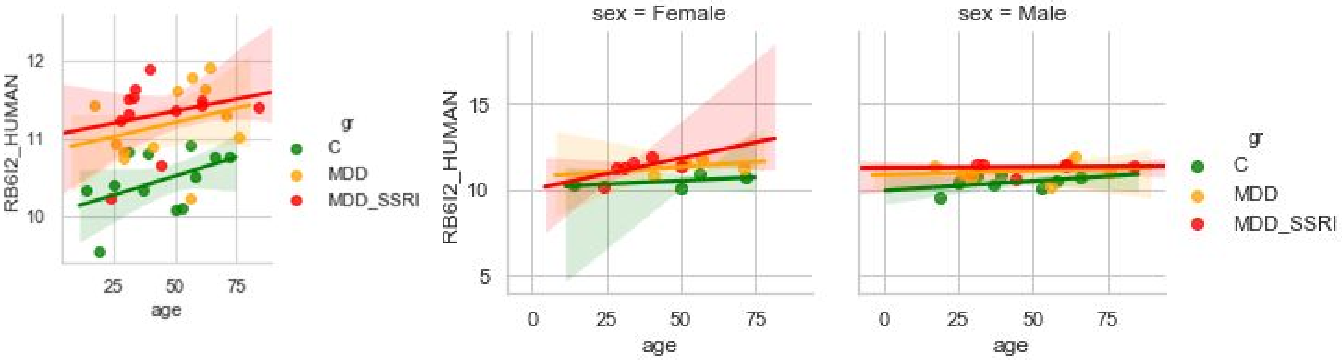


Figure 8. Female versus Male RB6I2 levels vs. Age

For key proteins that were found to have a significant impact on suicide versus non suicide status, we went ahead and plotted protein expression as a function of age and visualized a regression line as well as a 95% confidence interval for the two subject categories (suicide and non_suicide). For proteins MA2A1, DIRA1, PRS8, ENPL, and ATD3A, the regression line for non-suicide subjects (blue) across all ages is above the suicide subjects' regression line (red). Whereas, for RB6l2, the red suicide regression line is higher than the non-suicide line for all ages meaning that our suicide subjects seem to experience higher protein expression compared to that of non-suicide subjects. For some of the proteins (RB6I2 and DIRA1), there does seem to be a prominent trend amongst the regression lines with respect to age (DIRA1 coef = -0.2539 and p = 0.001 < 0.05, RB6I2 coef = -0.0042 and p = 0.045 < 0.05). You can better visualize these patterns by looking at the following plots which depict information for all samples including those from the control group. If we were to remove the control group, since no one in the control group commited suicide, we see that the same patterns hold but not as prominently. For instance, looking at RB6I2 expression, the suicide group has a lower expression, and becomes non-significant (p = 0.422 > 0.05) perhaps due to our small sample size. The same occurs for DIRA1 (-0.1521, p = 0.228 >0.05). Another interesting aspect to note is regarding DIRA1 protein: the blue regression line actually crosses the red suicide regression line at age = 75, which may indicate an aging effect on the decreasing concentration of DIRA1 in the Dentate Gyrus over time (See below figure.). However, when we examine the interaction between suicide status and aging, it is not significant (ie. RB6I2 -> p = 0.863 > 0.05, coef = -0.0026 and DIRA1 -> p = 0.398 > 0.05, coef = 0.0066)

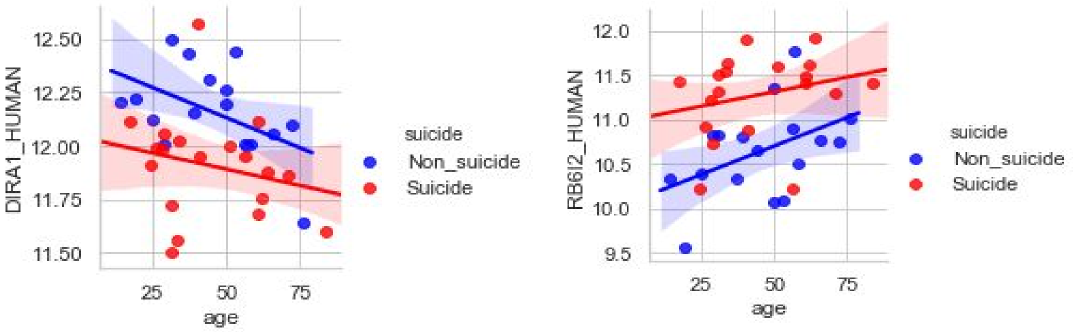


Figure 9. Age regression faceted by suicide status

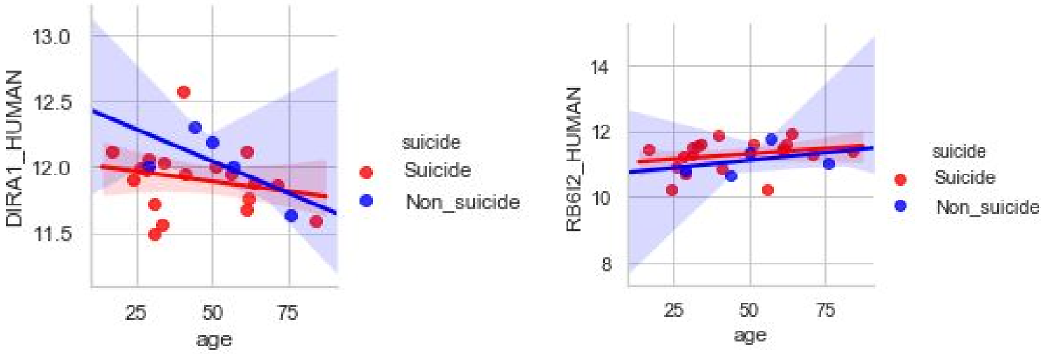


Figure 9.2. After removing outliers

Testing Aging and Protein level Interaction

When we rerun the interaction analysis from a group-wise perspective and remove the outliers and controls, we see that PRS8_HUMAN was the only protein with a significant Age and MDD*SSRI interaction term of coef = -0.0102 and p = 0.008 < 0.05. If we add back the controls, Age and MDD becomes significant (p = 0.0064), but Age and MDD*SSRI is not (p = 0.366). It is important to be aware of these interactions since aging can affect the natural process in which protein levels fluctuate and are produced.

Defining Network of Clusters for Suicide vs. Non-suicide and Control, MDD, MDD and Treated

In this section we are going to talk about the similarity graph between clusters and observe whether our key, selected protein features can effectively discriminate samples. The two groups of clusters we looked at were suicide vs non-suicide and control, MDD, MDD*SSRI.

To form the network one must define the notion of similarity or how well two random variables vary together. A standard measure for such value would be correlation between two random variables. Using the features we have selected, we can look at the correlation between patients without knowing the actual labels. In other words, without having any information if the patient in in control group, treatment or did suicide.

This approach is going to give us a correlation matrix where rows and columns are patients. The diagonals are going to have highest correlation value 1 as a patient is fully correlated with itself. Using the correlation matrix, now one can form a graph between the patients by forming an edge between patients that have higher correlation than a selected threshold value.

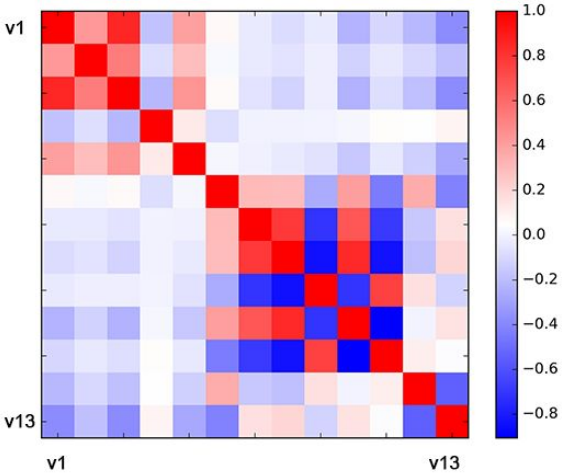


Figure 10. An example covariance matrix that will later lead to adjacency matrix

It is important to note that we are going to use only positive correlation values while forming the edges as we only want to form edges within the groups, not between. The result can be seen below. The green nodes are patients that did suicide and greens are those that did not. We see that the selected features are very good at identifying the groups.

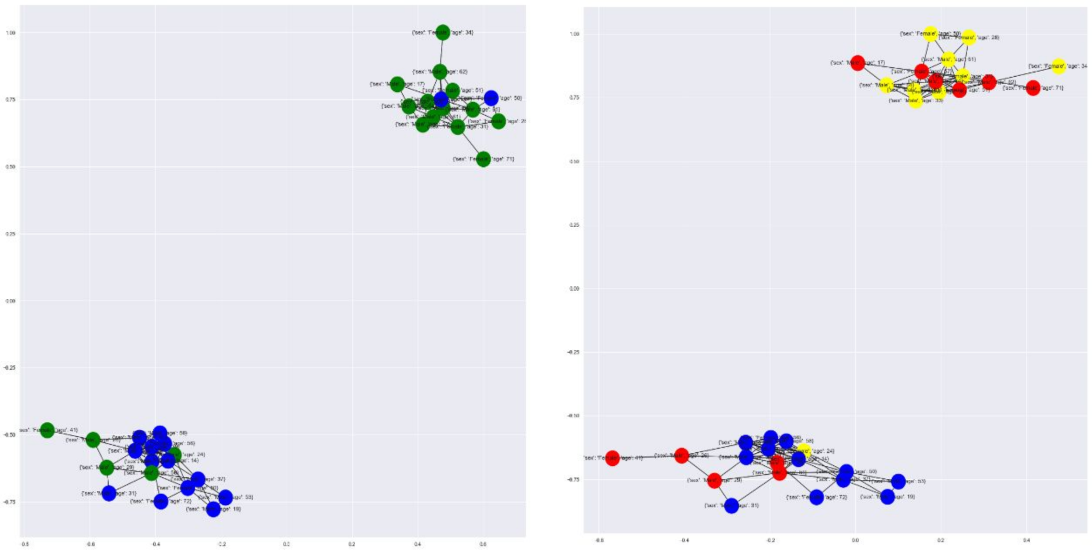


Figure 11. Distinct clustering for non-suicide vs. suicide samples and control vs. depressed

Another thing one can ask is how well those features identify the control, treatment and non-treatment groups. Coloring the nodes with respect to such labels yield the figure above. Here Blue nodes represent control, Yellow nodes represent MDD*SSRI and Red nodes represent MDD.

The next logical follow up question to ask at that point is whether we are able to cluster red (MDD) and Yellow(MDD*SSRI) separately? The answer is unsurprisingly no. This is because, If that was possible we could cluster them on the previous slide. One can see the results regarding this below (Figure 12.).

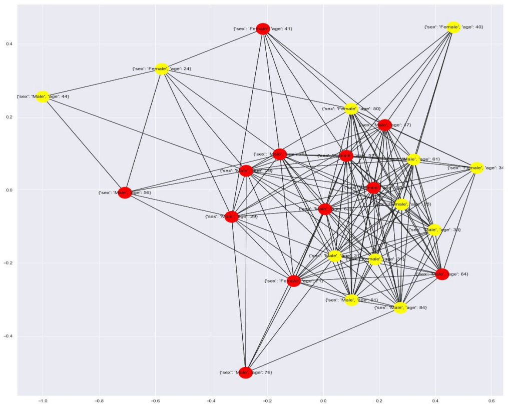


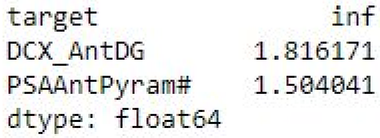
Figure 12. Little discrimination between depressed and treated versus depressed

Analysis of Cell Counts and Proteins levels correlation

Two other interesting question to ask are “How does depression affect the reproduction of the cells?” and “Are there cells that enable differentiating between individuals that commited suicide or not?”. In this section we are going to give some examples regarding correlation between proteins and cell types. The highest positive correlation occurs between Doublecortin Dentate Gyrus cells and ANXA 2 proteins.

- **Doublecortin Dentate Gyrus:** The dentate gyrus is thought to contribute to the formation of memories, and to play a role in depression.
- **ANXA 2**: ANXA 2 plays a role in the regulation of cell growth and signal transmission.

We have also found that Doublecortin Dentate Gyrus cells play an important role in differentiating between suicidal and non-suicidal patients. We observed that on average non-suicidal patients have approximately 58% more cells in Doublecortin Dentate Gyrus compared to suicidal patients. The results of the test are shown below (Figure 13.).



```
target              inf
DCX_AntDG      1.816171
PSAAntPyram#   1.504041
dtype: float64
```

Figure 13. Results of cell count and protein level correlation

Cell Counts - Kruskal Test

We explored the significance of cell counts among groups and also between suicide/non-suicide patients. The Mann Whitney U test between suicide and non suicide groups resulted in no significant cell counts. Kruskal test among the three groups showed that NestinAnt cell count is significantly different among the groups with a p-value of 0.002758. Further investigation is needed to discover why this might be the case.

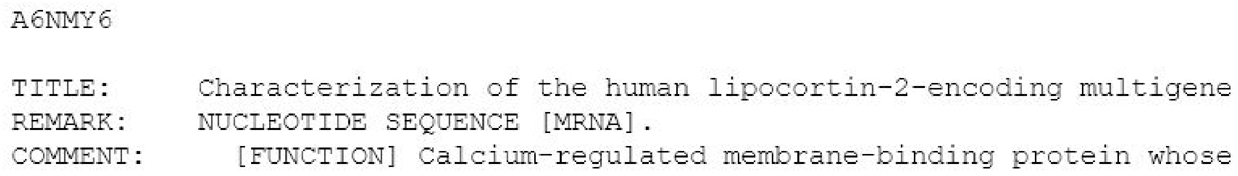
Demographics and Smoking

We additionally conducted testing by separating the cell counts as well as the protein expressions into two groups based on Caucaisian vs Non-Caucasian and Smoking vs Non-Smoking. Interesting questions such as whether certain ethnicities are more exposed to higher levels of proteins that are correlated with higher depression levels. Additionally, it can shed light on whether a smoking habits can lead to neurodegenerative diseases. Unfortunately, under Bonferroni Adjustments and Benjamini-Hochberg Procedures, none of the proteins and cell counts were significantly different among these groups.

Text mining

Much of our prior work so far has been quantitative analysis related to protein expression. We are now interested in learning more about the biological function and context surrounding our key proteins. Websites such as The National Center of Biotechnology Information (ncbi.nlm.nih.gov) contain an abundance of information on the research of these proteins. Other websites of interest include uniprot.org, genecards.org, and the Columbia University Library Web of Science database. For this second project report, we focused on the National Center for Biotechnology Information website and uniprot.org. While uniprot.org has a simple built in interface which allows users to download text straight from site, ncbi.nlm.nhi.gov does not, which makes extracting data a bit more challenging. However, by incorporating Python package, Selenium, a web-based automation tool, we were able to extract key pieces of information (ie. summaries, article titles, etc.) pertaining to certain proteins from the ncbi site. By creating this automated pipeline for text mining, we can effectively extract large amounts of text data that may lead to meaningful insights about the protein's role in various diseases and normal body functions. We found that many of our key proteins has not been studied extensively in academia or in medicine, which implicates the significance of our Capstone findings in furthering depression proteomics studies.

We first started with NCBI.org website, and iterated through all the proteins of our interest. Example web page of NCBI.org protein can be found here, https://www.ncbi.nlm.nih.gov/protein/A0JP26. Below is a capture of one of the 1126 proteins' and the titles, remarks, and the comments of the literatures that have been written about them.



```
A6NMY6

TITLE:      Characterization of the human lipocortin-2-encoding multigene
REMARK:     NUCLEOTIDE SEQUENCE [MRNA].
COMMENT:      [FUNCTION] Calcium-regulated membrane-binding protein whose
```

This will be a great value add to the team who are currently accessing these databases in a semi-automated way that requires significant level of human effort. Next step is to get more detailed data by actually going into these papers using the web scraping tools and extract out the abstracts of all the papers. This will empower the group to do more natural language processing on those scraped data

Natural Language Processing Techniques to Understand Biological Role of Proteins

Our goal for this section is to create an automated script which accepts input text data and outputs relevant insights pertaining to protein-related language. For the data retrieved from uniprot.org, columns include 'entry name' (protein name), 'Function', 'Involvement with Disease', 'Interactions' (with other proteins), etc. During the cleaning and preprocessing phase, we removed strange characters, punctuation, lower cased, lemmatized, and removed the stop words. These steps are essential when it comes to working with text data. Next, we had to examine the data a bit more to see if there were any syntactic nuances that had to be addressed. For instance, there are often references to publications tacked on at the end of a sentence (i.e., '{ECO:0000269|PubMed:12194967') that may influence the distribution of words and context of the sentence. Thus, using regex, we removed such instances.

We used both CountVectorizer and TfidfVectorizer to create word frequency distributions of the text data. CountVectorizer counts the numbers instances of each word, while TfidfVectorizer will do that and also downweight words which occur overly frequently. Below is the Tf-idf frequency distribution plot for VGF_HUMAN. We can see that 'depression' is among the top 10 most frequent words (see Figure 14. below). This protein is also involved with 'memory', 'neurogenesis', and 'neuroplasticity', which all confirm its importance to the dentate hippocampus (memory-formation) and cellular brain processes.

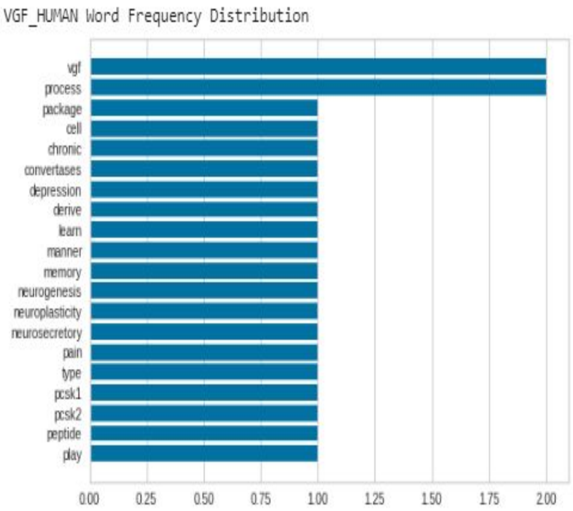


Figure 14. VGF_HUMAN word frequency distribution results

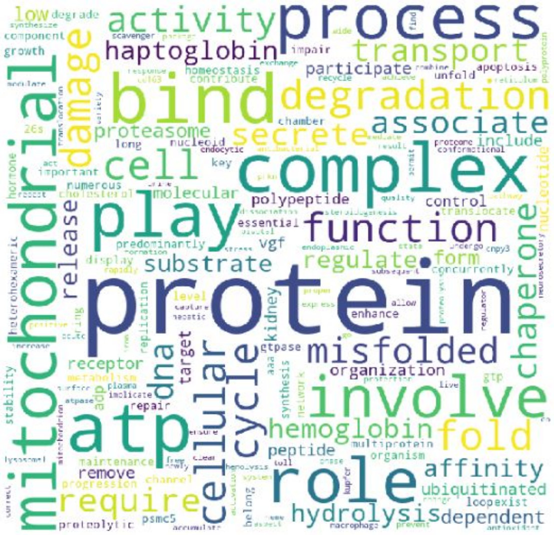


Figure 15. Word cloud visualization

We can also get a better idea of the most prominent words that occur in sentences for all the proteins by using a word cloud visualization (above in Figure 15.). Aside from 'protein', we see words which reference the functions of these proteins on a cellular level (i.e. 'atp', 'mitochondrial', 'hemoglobin'). However, we are also interested in seeing how our key proteins relate to one another. From a semantic perspective we can accomplish this using Universal Sentence Encoder and encoding each sentence as a 512 dimensional vector. From there, we created a similarity metric and applied it to the text for each pairs of proteins.

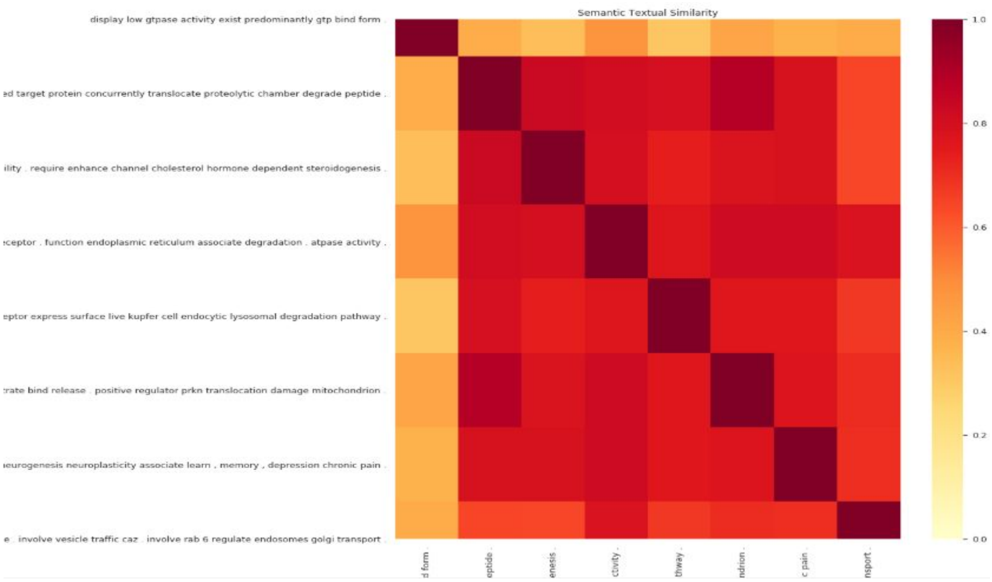


Figure 16. Similarity scores displayed as heatmap

We can see in the above heatmap in Figure 16., that row 2, 6 and row 4,6,7,8 have above 0.8 semantic textual similarity. If we take a look at the table below, we see that for PRS8_HUMAN (row 2 text) and HS71L_HUMAN (row 6 text), there is some similarity within the cleaned text in that both proteins are involved with misfolded protein formation and/or damage, and are both atp-dependent processes.

Table 2. Overlapping roles seen through protein function text

Row 2: PRS8_HUMAN	Row 6: HS71L_HUMAN
'component 26s proteasome , multiprotein complex involve atp dependent degradation ubiquitinated protein . complex play key role maintenance protein homeostasis remove misfolded damage protein , impair cellular function , remove protein function long require . , proteasome participate numerous cellular processes , including cell cycle progression , apoptosis , DNA damage repair . psmc5 belong heterohexameric ring aaa protein unfold ubiquitinated target protein concurrently translocate proteolytic chamber degrade peptide .'	'molecular chaperone implicate wide variety cellular process , include protection proteome stress , fold transport newly synthesize polypeptide , activation proteolysis misfolded protein formation dissociation protein complex . play pivotal role protein quality control system , ensure correct fold protein , fold misfolded protein control target protein subsequent degradation . achieve cycle atp bind , atp hydrolysis adp release , mediate co chaperone . affinity polypeptide regulate nucleotide bind state . atp bind form , low affinity substrate protein . , hydrolysis atp adp , undergo conformational change increase affinity substrate protein . go repeat cycle atp hydrolysis nucleotide exchange , permit cycle substrate bind release . positive regulator prkn translocation damage mitochondrion .'

Critical Discussion of Results

In this pilot study, we found significant group differences in the level of proteins between control, treated, and non-treated as well as among suicide and non-suicide brains. RB6I2, DIRA1, CN166, MA2A1, PRS8, HPT, HS71L, and VGF were significant in differentiating control vs. depressed samples while, DIRA1, PRS8, RB6I2, ATD3A, MA2A1, and ENLP were significant proteins differentiating suicide vs.non-suicide samples. They are discovered using the Benjamini-Hochberg procedure that keeps the false discovery rate controlled at < 0.05. From our network graphs, we can reasonably discriminate and cluster control versus depressed samples and suicide versus non-suicide samples, but were unable to do so for depressed versus depressed and treated samples. The major contextual reason for why we did not observe any clear patterns between the depressed groups is that individual response to antidepressants is highly variable and may confound the effects of antidepressants in treating individuals. Ideally, more samples would be needed to create an even larger network graph and to more accurately examine their clustering patterns.

Using our NLP techniques, we can effectively learn more about our key proteins and see how they affect individuals. RB6I2 and ATD3A are two noteworthy proteins which warrant further discussion. We discovered higher concentrations of RB6I2 amongst depressed, suicidal samples. Using TF-IDF word frequency distributions, we observed RB6I2 to be highly involved in the regulation of neurotransmitter release at nerve terminals, and thus excess levels may result in neurotransmitter system dysfunction and contribute to depression-like symptoms due to neurological molecular imbalances. Meanwhile, word frequency distributions revealed that ATD3A and HS71L were both found to be highly involved with mitochondrial protein synthesis. The differential expression of ATD3A in particular within suicide and non-suicide brains may reveal a mitochondrial protein synthesis gap which adversely affects ATP production downstream and further escalate MDD / depression-like symptoms (ie. Lethargy).

Limitations, conclusions, and future work

We were able to leverage various types of statistical and NLP techniques to thoroughly examine the brain proteomics levels of control, MDD, and MDD + treated samples. This proteomics and data mining approach helps us to uncover molecules involved in the pathogenesis of MDD and suicide. Such information can be useful in identifying new molecular treatment targets to develop novel drugs to treat these severe conditions and provide temporary solutions for symptoms related to ATP production and neurotransmitter regulation in the brain which are experienced by those with deficit or excess protein levels.

One of the major limitations of our work were the low number of samples (total 36 samples) when conducting our analysis. We were able to overcome this by using nonparametric statistical testing which did not assume any particular distribution for our protein features. Another limitation was the amount of available research and literature on specific proteins. This made it difficult for us to use NLP since these techniques are heavily dependent on the textual inputs. However, since many proteins had either little or no text literature, it was difficult to understand these proteins' functionality and thus tie this information back to our quantitative results. On the other hand, since many of the key proteins we discovered to be correlated with MDD and suicide had not been researched in the past, our results have immense potential to pave the way for further research in these areas.

This leads us to our future work and next steps in the area. Our mentors have already expressed excitement in presenting our work at various medical conferences next year. Additionally, this pilot research project will provide the medical research community additional confidence to seek out more brain samples to test our assumptions on. The more samples we have, the more accurately we can measure these patterns and and further empower the research community who are forerunners in advancing our society's well being.