Downregulation of KRT proteins in the Neuroendocrine System of Amyotrophic Lateral Sclerosis with PT150 as a Multi-Target Neuroprotectant

Samarth Dunakhe[1], Arnav Sharma[1], Aryav Das[1], and Pranav Kulkarni[2]

[1] CortexPD Labs, Troy, MI, USA 48085
[2] University of California, Davis Davis, CA 95616

Abstract. Amyotrophic Lateral Sclerosis (ALS) is a neurodegenerative disease that deteriorates motor neurons. Researchers have discovered proteins and genes such as C9orf72, Nfl, TDP43, SOD1, and FUS that exhibit unique traits in ALS patients. ALS still has a misdiagnosis rate of 55% and an expected lifespan of 3-5 years. This is partially due to the lack of specificity of existing biomarkers. Additionally, ALS mimic diseases, or diseases that mimic ALS very closely in pathology, often increase the diagnostic period as they have to be ruled out before a proper diagnosis can be made. This paper proposes the use of graph-based machine learning to model ALS heterogeneity alongside biological and neurospatial databases to discover and validate an ALS specific biomarker group along side molecular docking to discover a small-molecule therapeutic to modulate this biomarker. Two open source datasets were sourced. First, a gene expression profile with 7,038 gene expressions and second, a neuroproteomic network analysis with 2,105 proteins. To interpret the gene expression profile, a Graph Neural Network (GNN) was designed for classifiying each patient into either ALS, control, or ALS mimic that achieved an accuracy of 98% between ALS and controls and 85% between ALS and mimics. Consequently, Integrated Gradients identified the 20 most differentially expressed genes based on their affect on the GNN's diagnosis of the ALS class. To interpret the proteomic profile, two-sample t-tests and Log2 Fold Change were applied with thresholds at 0.05 and 0.58 respectively. Consequently, this identified the 20 most differentially expressed proteins. After validating the biomarkers considering their neurospatial and pathological contexts, KRT6A, KRT6B, and KRT4 were identified as biomarkers in the pituitary gland with tau specificity of 86% showing potential to reduce the ALS misdiagnosis rate by up to 20% in clinical settings. They were shown to be correlated with cellular differentiation and human developmental pathways. These keratin proteins were down-regulated due to Glucocorticoid receptor a (GRa) inhibiting the activity of AP-1 transcription factor, which is supposed to regulate keratin proteins. Through molecular docking, it was shown that the PT150 ligand could inactivate Glucocorticoid Receptor Alpha and Cyclooxygenase 2 and regulate neuroinflammation and neuroendocrine dysregulation in ALS.

Keywords: Graph Neural Network, Amyotrophic Lateral Sclerosis Biomarkers, Pituitary, Keratin proteins, PT150 Ligand, Multi-target

1 Introduction

ALS is a neurodegenerative disease characterized by the progressive deterioration of upper and lower motor neurons that facilitate all voluntary muscle movement. Upper motor neurons (UMNs) in the motor cortex and brain stem along with lower motor neurons (LMNs) in the spinal cord and brain stem. Degeneration of these motor neurons leads to functional impairments such as muscle atrophy, paralysis, and respiratory failure [29].

ALS is one of the most heterogeneous neurodegenerative diseases due to its genetic variability, as only 15% of ALS cases are known to have genetic origin [16], ALS cases can have varying forms of onset in the limbs or the head/neck which accounts for different progression rates, and the numerous disease mechanisms involved in ALS. This accounts for the 55% misdiagnosis rate and complicated treatment for ALS. Coupled with the 3-5 year average lifespan for 80% of ALS patients after diagnosis, this issue becomes critical to address [28].

Recent research has attempted to improve this misdiagnosis rate through the discovery of several biomarkers.

- NfL: A protein located in the cerebrospinal fluid classified as a biomarker when it is too abundant [27].
- TDP-43 & FUS: A protein and encoding gene respectively, involved in DNA binding processes classified as biomarkers when they are found in the incorrect cellular location (cytoplasm rather than nucleus) [24] [15].
- C9orf72: A gene classified as a biomarker when it has a longer DNA sequence than it should [22].
- SOD1: An encoding gene classified as a biomarker when it encodes enzymes with toxic abilities [14].

These biomarkers have been validated to help researchers understand ALS pathology and improve diagnosis [43] [44] [45]. However, these established biomarkers are unspecific between other diseases along with ALS such as:

- NfL is observed in several Alzheimers Disease (AD) cases, Frontotemporal Dementia (FTD) cases, Multiple Sclerosis (MS) cases, and Guillain-Barré Syndrome (GBS).
- TDP-43 is observed in 90% of Hippocampal Sclerosis (HS) cases, 57% of Alzheimers Disease (AD) cases, $\tilde{5}$0% of Frontotemporal Dementia (FTD) cases [11] [13] [12].
- FUS aggregates is observed in only 3-4% of familial ALS cases, only 1% of sporadic ALS cases, and 6-20% of Frontotemporal Lobar Degeneration (FTLD) cases [15] [9] [10].
- C9orf72 is observed in $\tilde{1}$0-15% of FTD cases and 10% of ALS cases [21] [23].
- SOD1 is observed in only 10-20% of ALS cases [6] [7] [8].

Beyond this lack of specificity, these biomarkers also have very delayed and unreliable onsets as

- NfL abundance rises proportional to neuronal degeneration.
- Cytoplasmic TDP-43 & FUS aggregates are only observable once ALS has progressed significantly due to their slow formation and minature size.
- $\tilde{5}$0% of C9orf72 cases are asymptomatic [5].
- Some SOD1 cases are asymptomatic and SOD1 toxic enzymes are only detectable during mid-stage ALS because of the gradual accumulation.

Because of these fallacies, ALS still has a late 55% misdiagnosis rate, low lifespan from diagnosis, and a limited application of interventional therapeutics as significant neuronal damage has already occurred.

Researchers have applied computational approaches to model ALS heterogeneity using Random Forest (RF) models on MRI data and Support Vector Machines (SVM) on gene expression data [35] [36]. However, these approaches plateau at low accuracies due to their limited data and suboptimal feature aggregation.

Observing these deficiencies, this paper proposes the use of graph-based machine learning to model ALS heterogeneity alongside the use of biological and neurospatial databases to discover and validate a novel, specific, and early-onset ALS biomarker group alongside molecular docking to discover a small-molecule therapeutic to modulate this biomarker. We started with a gene expression profile with 2,943 transcripts [25] and a proteomic network analysis with 2,105 proteins [26]. To interpret the transcriptomic data and correspondingly model ALS heterogeneity, a Graph Neural Network (GNN) was developed on the Graph Sample and Aggregate (GraphSAGE) architecture [19] [17]. Integrated Gradients was then applied to discover which genes were the most important in ALS based on the models decisions [18]. To interpret proteomic data, statistical tests were applied to assess differential expression. To validate potential biomarkers at this stage, neurospatial analysis and referencing with biological databases was conducted which provided insight to biological pathways and previous disease correlation for our identified biomarker group. Based on this identified biomarker group, we utilized molecular docking to analyze which ligand could most effectively bind with our select receptor to regulate the identified biomarker group and also another pathway in ALS through multi-target binding with another receptor. We constructed the State-Of-The-Art method of modeling ALS heterogeneity, a novel biomarker group that improves the misdiagnosis rate of ALS by up to 20% in clinical settings, and a therapeutic that regulates neuroinflammation and neuroendocrine dysregulation in ALS.

2 Methodology

2.1 Data Profile

The first dataset was a batch-controlled, gene expression transcriptome profile with 1117 subjects (397 ALS, 645 control, 75 mimics). The ALS patients were

split between familial and sporadic and the control patients were age and gender matched. These subjects were limited to the population of the Netherlands in Western Europe. The location onset of ALS among these patients were either Bulbar or Spinal. Collected through PAXgene Tubes, 7,038 transcriptomes were expressed and 2,943 of those transcriptomes had a False Discovery Rate (FDR) of less than 0.05. These transcriptomes are associated mainly with RNA binding and intracellular transport events [25].

The second dataset was a cerebrospinal fluid proteomic analysis with 101 subjects (35 sporadic ALS, 10 C9orf72 ALS, 6 SOD1 ALS, 6 asymptomatic C9orf72, 44 age-matched controls). The population for the patients is unknown, however, the population of the USA can be inferred due to the publishing labs location (Emory University). Collected through Tandem Mass Tag-Mass Spectrometry (TMT-MS), 2,105 proteins were expressed [26].

The validation dataset was a gene expression transcriptome profile for Alzheimer's Disease (AD) patients. It consists of postmortem prefrontal cortex (PFC) samples from 467 subjects(310 AD and 157 non-demented control). The age in the dataset ranges from AD subects 53–100 years and controls : 22–106 years. The dataset consistent of 209 females and 258 males. 19,206 differentially expressed transcripts were extracted, 608 of which, were differentially expressed [4].

2.2 Graph Construction and Feature Selection

The dataset was transposed so that each row corresponded with a sample. Each diagnosis in the data was then then labeled with values of 0-2. After the industry standard train-test-split of 80% training and 20% test, feature selection applied. Feature standardization was then applied with *StandardScaler* as some of the gene expression values had different scales.

As the transcriptome profile consisted of 2,943 features resulting in model overfitting, Principal Component Analysis simplified the dataset to 300 of it's most significant features. This amount of 300 was selected after testing with different feature amounts.

To convert the resulting 300 feature dataset into a graph structure suitable for the Graph Neural Network (GNN), we constructed a k-Nearest Neighbors (k-NN) graph to connect patients with similar gene expression profiles. Each patient sample was formed into a node, with edges formed by connecting each node to its 5 nearest neighbors using Euclidean distance. To improve feature aggregation, self-loops were added to preserve each nodes information when being passed through the layers in the GNN [34].

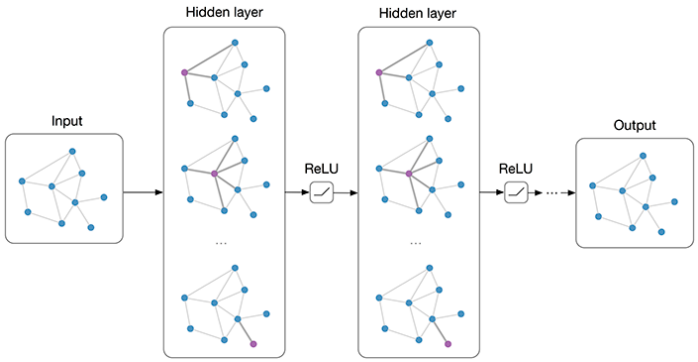


Fig. 1: Self loops in an example Graph Neural Network [46]

The resulting adjacency matrix was computed using kneighbors_graph() from scikit-learn, and the edges were extracted and formatted into a PyTorch Geometric edge_index tensor. These two parameters were the inputs for the GraphSAGE-based GNN.

2.3 GraphSAGE Architecture

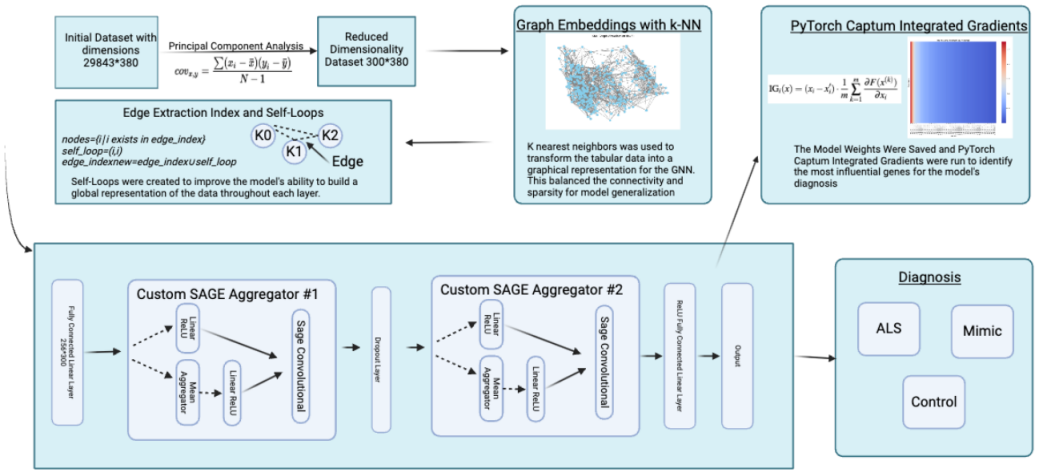


Fig. 2: The Graph Neural Networks Architecture [40]

GraphSAGE (Graph Sample and Aggregate) is a GNN architecture designed for inductive learning. This allows the GNN to generate dynamic node embeddings with information from their neighbors in the feature space, described

as "Neighborhood Aggregation", rather than relying on fixed feature vectors. GraphSAGE works by sampling a fixed number of neighbors and applying functions (such as mean and pooling) to combine their features allwoing each node to capture local and global patterns [19].

After the fully connected input layer described in "Graph Construction and Feature Selection", the GNN consists of two GraphSAGE layers and one hidden layer, and a fully-connected output layer described below.

The **first SAGEConv layer** aggregated the information for each node using neighborhood aggregation computed as:

$$\text{AGGREGATE}^{(l)}(i) = \frac{1}{|\mathcal{N}(i)| + 1} \left(h_i^{(l)} + \sum_{j \in \mathcal{N}(i)} h_j^{(l)} \right),$$

$$\begin{aligned}
\text{AGGREGATE}^{(l)}(i) &= \text{Aggregated feature representation for node } i \text{ at layer } l. \\
h_i^{(l)} &= \text{Feature representation of node } i \text{ at layer } l. \\
\mathcal{N}(i) &= \text{Set of neighboring nodes of node } i \text{ in the graph.} \\
|\mathcal{N}(i)| &= \text{Number of neighbors of node } i. \\
h_j^{(l)} &= \text{Feature representation of a neighboring node } j \text{ at layer } l. \\
\sum_{j \in \mathcal{N}(i)} h_j^{(l)} &= \text{Summation of feature representations of all neighbors of node } i \text{ at layer } l. \\
\frac{1}{|\mathcal{N}(i)| + 1} &= \text{Bessels correction that corects bias}
\end{aligned}$$

The **hidden layer** refined the node representations by applying Dropout regularization where 50% of the neurons were randomly set to zero druing training to prevent overfitting. It also employed ReLU activation enabling non-linearity for the graphs operations

The **second SAGEConv layer** further aggregated each nodes information allowing the model to incorporate the secondary neighbors for each node in the feature space, described as "Heirarchical Aggregation". The final node information was then simplified into embeddings which were put into the output layer.

Finally, the **fully-connected output layer** mapped the node embeddings to the disease labels (ALS, MIM, or control) through logits which were converted to probabilities with a softmax function.

The GNN was also tested on

2.4 Integrated Gradients

Integrated Gradients (IG) from PyTorch Captum library quantified the attribution of each gene to the GNN's mapping of the ALS class.

IG method works through a path integral approach. IG interpolates between a baseline input x' which represents the mean of the feature values across the

data samples and the actual input. This generates a series of intermediate inputs. This traces a straight line in the feature space from the baseline input to the actual input. At each step, the gradient of the model's output for each feature is computed, capturing how each inputs contribution to the models output [18]. The resulting equation, that approximates the integral with a summation, is given by:

$$\mathrm{IG}_i(x) = (x_i - x_i') \cdot \frac{1}{m} \sum_{k=1}^{m} \frac{\partial F(x^{(k)})}{\partial x_i}$$

$$\begin{aligned}
\mathrm{IG}_i(x) &= \text{Integrated Gradient attribution for feature } i \text{ at input } x. \\
x_i &= \text{Actual value of feature } i \\
x_i' &= \text{The mean of the feature values } i, \\
m &= \text{Number of interpolation steps} \\
x^{(k)} &= \text{Intermediate input obtained by linearly interpolating between } x' \text{ and } x. \\
F(x) &= \text{Model's prediction function, mapping input } x \text{ to an output score.} \\
\frac{\partial F(x^{(k)})}{\partial x_i} &= \text{Gradient of the model's output}
\end{aligned}$$

The 20 most attributed transcriptomes in that dataset for its diagnosis of the ALS label were considered significant.

2.5 Statistical Analysis

The differential expression between each protein was analyzed using Log2FC and Two-Sample t-tests to measure relative protein abundance and expression respectively. Positive Log2FC indicated upregulation and negative Log2FC indicated downregulation, with the threshold at Log2FC > 0.58 . The p-values from the t-tests had a threshold at $p < 0.05$. The 20 proteins with the lowest p-values and the most extreme Log2FC values were considered significant.

2.6 Neurospatial and database validation

Analysis with UniProt and Gene Ontology (GO) was then performed to clarify which biological pathways and processes the 20 most differentially expressed proteins and genes were involved in [20] [32]. We identified major pathways with cutoffs at $p < 0.05$ and fold enrichment > 4.

the 20 most differentially expressed genes and proteins were then analyzed using the Allen Brain Atlas (ABA) and Human Protein Atlas (HPA) [31] [33]. The ABA and HPA provided the expression levels for each protein and gene across the brain alongside tissue-level cellular quantification for each protein and gene.

2.7 Molecular Docking

To perform in-silico drug discovery, a computational technique used to predict how two molecules will interact with each other called "Molecular Docking" was used.

First 10,000 molecular sequences were selected from PubChemPy in the form of SMILE sequences which are computationally understandable molecular sequences. Filtering was then conducted based on Lipinski's rule of 5, which is a set of rules that state an orally active drug has no more than one violation of the following criteria:

- There are no more than 5 hydrogen bond donors
- There are no more than than 10 hydrogen bond acceptors
- The molecular mass is less than 500 daltons
- The octanol-water partition coefficient does not exceed 5

The resulting molecules were then put into the Protein Ligand Binding Affinity Using Pretrained Transformers (PLAPT) model, which utilizes pre-trained Transformers and a Branching Neural Network (BNN) to predict protein-ligand binding affinities with State-Of-The-Art accuracies [41]. The resulting molecule (ligand) was put into Autodock Vina which provided detail into the ligands Blood-Brain-Barrier permeability and its relevance in clinical settings [42].

3 Results

3.1 Model Metrics

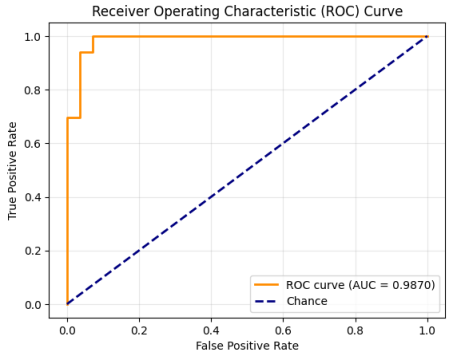


Fig. 3: The AUC-ROC graph for ALS vs. Controls showing the trade-off between the true positives and the false positives.

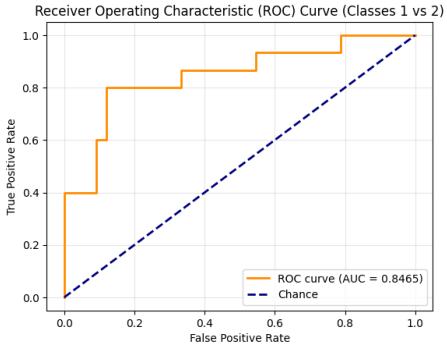


Fig. 4: The AUC-ROC graph for ALS vs. Mimics showing the trade-off between the true positives and the false positives.

The GNN performed at a 98% test accuracy + 99% AUC between ALS and control as shown in Figure 3 and a 85% test accuracy + 85% between ALS and Mimics as shown in Figure 4. This reveals the GNN's ability to model ALS heterogeneity as these metrics are State-Of-The-Art for ALS modeling approaches that use machine learning detailed in Figure 5.

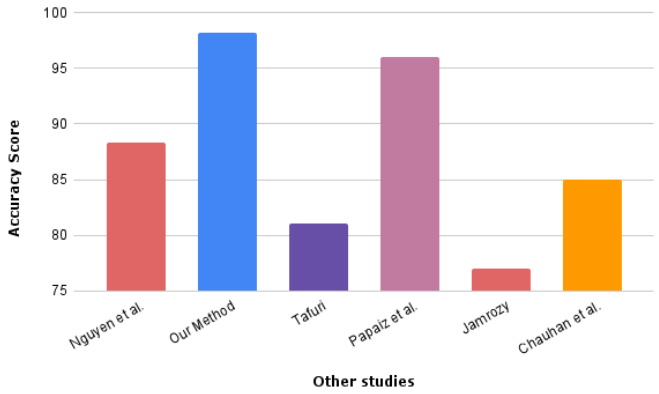


Fig. 5: The accuracy scores of [35] on gene expression data; [36] on MRI scans; [37] from the PRO-ACT Database; [38] on MRI data; and [39] on vowel speaking data in comparison to the GNN

The confusion matrix in Figure 6 shows that control subjects (Class 0) were predicted with perfect precision, ALS subjects (Class 1) were predicted with only 1 false negative and 4 false positives and Mimic subjects (Class 2) were predicted with 5 false positives and 10 true positives. This shows that the GNNs generalizations over classes was also State-Of-The-Art.

The model was also tested on an Alzheimer's Disease gene expression profile to ensure the GNN did not overfit on the ALS gene expression profile. The GNN achieved an accuracy and AUC of 97% and 98% respectively on this dataset with changes only in the data structuring with column transposing, class labeling, etc. This showed that the GNN is applicable in contexts beyond just ALS due to its generalizability and lack of over/under-fitting.

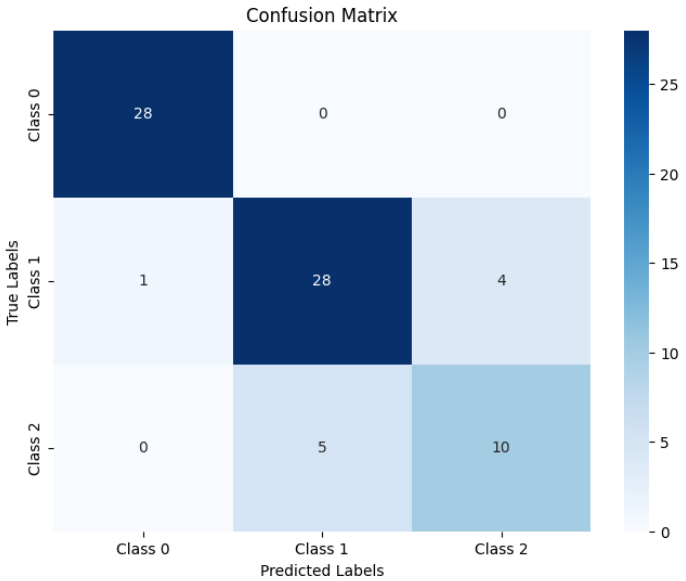


Fig. 6: The Confusion Matrix showing the GNN's performance across each class

3.2 Contextualized Biomarker Group

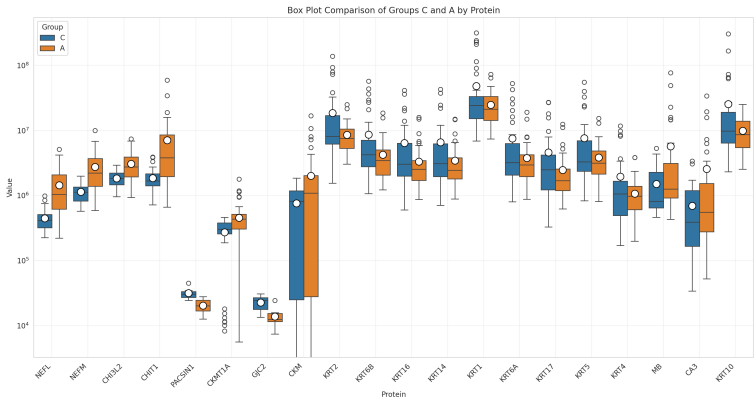


Fig. 7: The box plot shows each protein's significance in comparison to each other.

Significant differential expression was identified in the keratin proteins KRT6A, KRT6B, and KRT4 as shown in Figure 7 and the statistical test results indicated significant differential downregulation in their expressions as shown in

Variant	**Log2FC**	**p-value**
KRT6B	-0.99896	0.02440
KRT6A	-0.97817	0.03420
KRT4	-0.83941	0.03903

Table 1: Table of Variants with Log2FC and p-values

Paper	**Accuracy**	**AUC**	**Method Used**
Our Method	97	98	GraphSAGE GNN
Alshamlan et al.	84	_	SVM
Deng et al.	_	96	MLR

Table 2: Table of Variants with Log2FC and p-values

Table 2. Through further analysis with Gene Ontology and Uniprot [20] [32], it was discovered that the keratin proteins were prevalent in developmental and cellular pathways such as "tissue development" and "cellular differentiation". After cross-referencing with the highest attributed genes on Captum IG, it was shown that the keratin proteins and genes were expressed in similar pathways.

Neurospatial analysis revealed that the keratin proteins were expressed in the Pituitary Gland with tau specificity at 0.86 which is critical when considering how the Pituitary Gland has been linked to growth hormone and cortisol dysregulation in ALS. This discovery shows that the keratins are the reason for the hormone dysregulation. This proccess is further detailed in Figure 8.

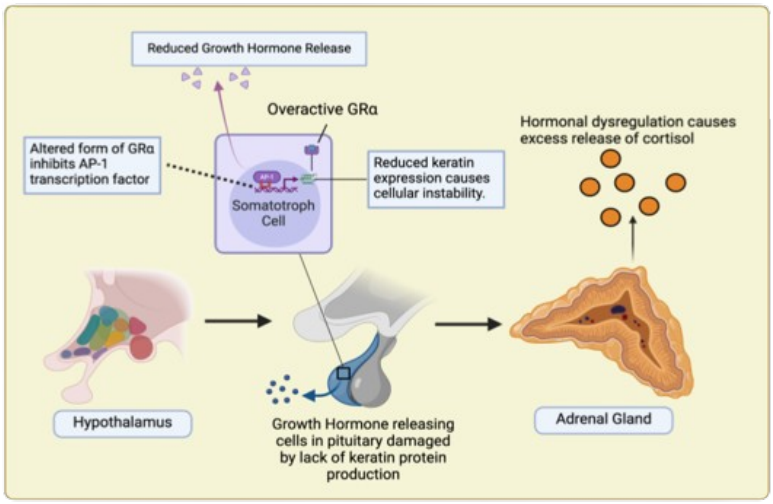
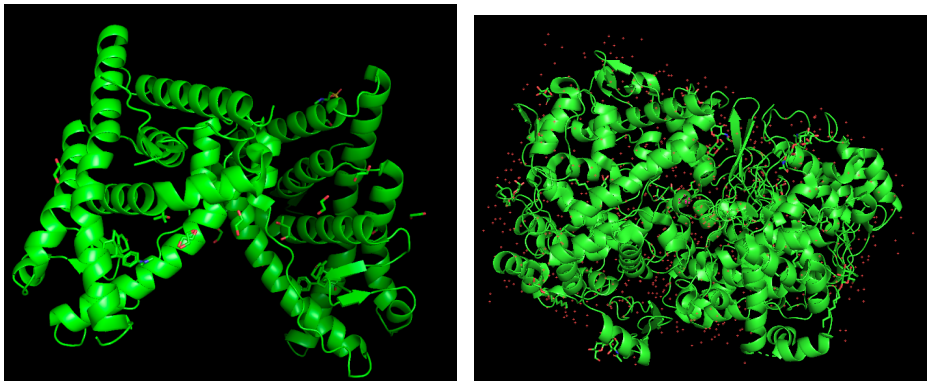


Fig. 8: The pathway of the newly found involvement of keratin proteins in hormone dysregulation.

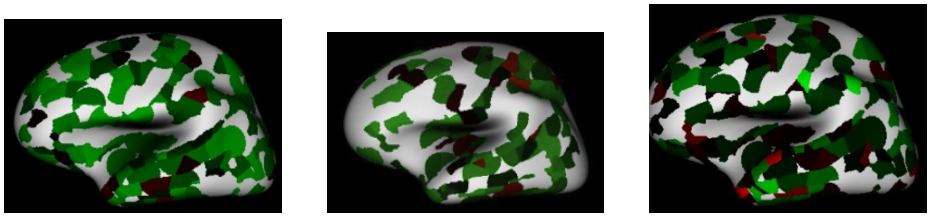
3.3 Drug Candidate



(a) Glucocorticoid Receptor α (b) Cyclooxygenase 2

Fig. 10: The molecular structures of Glucocorticoid Receptor α and 3D structure of Cyclooxygenase 2 [48]

Glucocorticoid Receptor α (GRα) was identified a regulatory receptor that was shown to inhibit the activity of the AP-1 transcription factor which is supposed



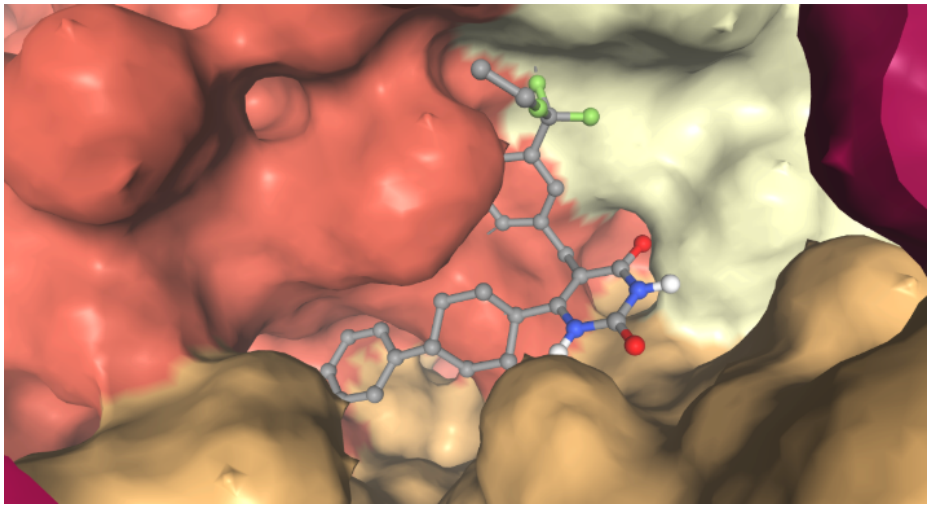
(a) Spatial distribution of KRT4.

(b) Spatial distribution of KRT6B.

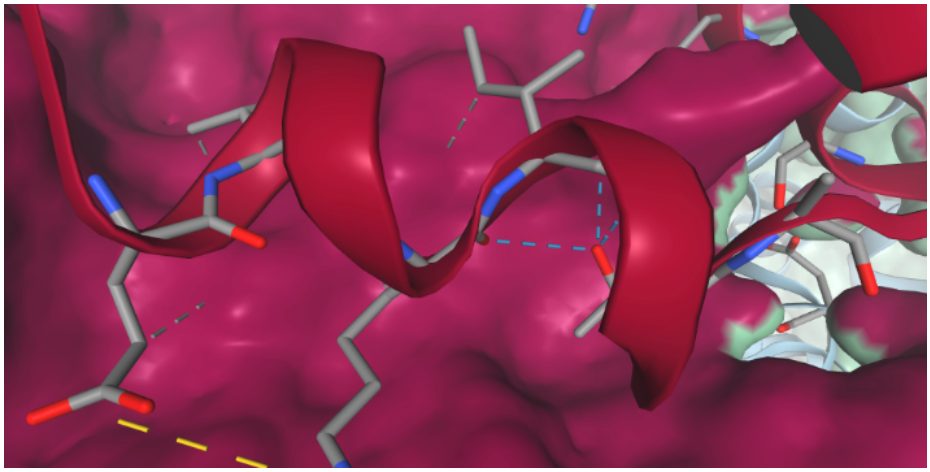
(c) Spatial distribution of KRT6A.

Fig. 9: Brain maps showing the expression (z-score) distribution of KRT4, KRT6B, and KRT6A across the brain.

to regulate keratin expression in response to stimuli such as neuronal activity and growth factors. Cyclooxygenase 2 (COX2) was identified as an important enzyme involved in the inflammatory response of ALS motor neuron degeneration. Specifically, COX2 produces prostaglandins, which mediate inflammation and pain. AP-1 is built from a heterodimeric complex of proteins from the JUN, FOS, ATF, and MAF families which regulate genes involved in cell differentiation, apoptosis, proliferation, and immune responses.



(a) Glucocorticoid Receptor α



(b) Cyclooxygenase 2

Fig. 11: Simulated binding environments between PT150 to Glucocorticoid Receptor α and Cyclooxygenase 2 [42]

The ligand PT150 was identified as small-molecule modulator for GRα activity. As shown in Figure 11, PT150 was able to bind with GRα at -7.5 kcal/mol inhibiting GRαs ability to inhibit targets such as AP-1. This is turn, allows for AP-1 to regulate KRT6A, KRT6B, and KRT4. Furthermore, PT150 indicated potential as a multi-target therapeutic agent as it was able to bind with Cyclooxygenase 2 (COX 2) at-7.5kcal/mol which in turn, partially inhibited

neuroinflammation, which is important in ALS as neuroinflammation directly contributes to motor neuron degeneration through it's application of oxidative stress, release of cytokines, activation of Glial cells, and disruption of the Blood-Brain-Barrier (BBB).

PT150 also demonstrated BBB permeability through it's lipophilicity which is how well a compound dissolves in fats, oils, and non-polar solvents. In this case, the lipophilicity of PT150 indicated that it could dissolve into the BBB which led to it crossing the BBB.

PT150 has recently been practiced in COVID-19 as it has been shown to modulate glucocorticoid and androgen receptors, which inhibits SARS-CoV-2 in human bronchial epithelial cells in turn, demonstrating PT150's antiviral properties against SARS-CoV-2 [1]. PT150 is also currently under clinical development by Palisades Therapeutics in Phase I trials for both prostate and ovarian cancers. PT150 acts in these diseases by targeting glucocorticoid and androgen receptors, which play a role in the progression of prostate cancer and translocate in overaiance cancer [2] [3]. This shows that beyond our in-silico tests, PT150 has real potential to be clinically applied as a therapeutic for ALS that targets the neuroendocrine system, regulated keratin proteins and in turn, regulates hormone release and production.

4 Discussion

4.1 Limitations and Future Directions

From a **data standpoint**, longitudinal sampling with a more diverse population from 6 continents would help observe if keratin proteins could act as population-generalized ALS biomarkers.

From a **clinical standpoint**, validation of the keratin proteins cellular-level roles in ALS can be carried out through genetic engineering (e.g, CRISPR, Cas9), in-vivo assays using tissue-level expression assays such as western blotting, and behavioral analysis of keratin dysregulation in animal models.

From a **drug development standpoint**, PT150 has already been clinically applied as a drug but its ability to regulate keratin proteins can be researched using in-vivo ALS mouse models to assess hormone regulation and techniques such as immunofluorescence or qPCR to quantify and confirm keratin protein regulation.

4.2 Research Concerns

The threat of **keratin contamination** was addressed in this research. Although the proteomic dataset did not provide information regarding keratin filtering, KRT4 is found on the face and genitals, KRT6B is found on hair and hands/feet, KRT6A is found on hands/feet and nails. The CSF analysis could not have been contaminated by these keratin proteins, however, statistical analysis with UniProt, log2FC, and Two-Sample t-tests counteract this claim and the prevalence of keratin proteins being revealed in associated neurodegenerative diseases

such as alzheimers [30] further proves that the identified keratin proteins were not contaminants.

5 Conclusion

In conclusion, this research made the SOTA method for modeling ALS heterogeneity with applications to other neurodegenerative diseases as the method also recieved SOTA accuracy on modeling Alzheimer's data which is important as ALS heterogeneity is the #1 factor for all ALS misdiagnosis. This research also made the first-ever correlation of KRT6A, KRT6B, and KRT4 to ALS as specific and early-onset biomarkers which has potential to reduce the misdiagnosis rate through the biomarkers specificity and potential to improve the efficacy of interventional therapeutics through early diagnosis. Lastly, this researcher also made the first-ever correlation of PT150 to ALS as a BBB-permeable and multi-target therapeutic which is clinically applicable as PT150's efficacy has been validated through its practicion in COVID-19 and Prostate/Ovarian cancer

References

1. Barnes, B.J., Adrover, J.M., Baxter-Stoltzfus, A., Borczuk, A., Cools-Lartigue, J., Crawford, J.M., et al. "Clinical stage molecule PT150 is a modulator of glucocorticoid and androgen receptors with antiviral activity against SARS-CoV-2." *mBio*, 2021. https://doi.org/10.1128/mBio.00665-21 https://pmc.ncbi.nlm.nih.gov/articles/PMC7738205/
2. Palisades Therapeutics. "PT150 in Prostate Cancer: Phase I Clinical Trial Data and Mechanistic Insights." *Pharmaceutical Technology*, 2023. https://www.pharmaceutical-technology.com/data-insights/pt-150-palisades-therapeutics-prostate-cancer-likelihood-of-approval/
3. Palisades Therapeutics. "PT150 in Ovarian Cancer: Potential Applications in Chemosensitization." *Pharmaceutical Technology*, 2023. https://www.pharmaceutical-technology.com/data-insights/pt-150-palisades-therapeutics-ovarian-cancer-likelihood-of-approval/
4. Yu, Wuhan, Yu, Weihua, Yang, Yan, and Lü, Yang. "Exploring the Key Genes and Identification of Potential Diagnosis Biomarkers in Alzheimer's Disease Using Bioinformatics Analysis." *Frontiers in Aging Neuroscience* https://doi.org/10.3389/fnagi.2021.602781
5. Murphy, E., Fischer, M.W., Pottier, C., Brown, R., Hardy, N.J., Pickering-Brown, C.G., Rademakers, R.: Age-related penetrance of the C9orf72 repeat expansion. *Scientific Reports* https://doi.org/10.1038/srep44843
6. Nguyen, H.P., Van Broeckhoven, C., van der Zee, J.: ALS Genes in the Genomic Era and Their Implications for FTD. *Journal of Neurology, Neurosurgery & Psychiatry* **93**(2), 201-214 (2018). https://doi.org/10.1136/jnnp-2020-326447 https://jnnp.bmj.com/content/93/2/201
7. Leblond, C.S., Kaneb, H.M., Dion, P.A., Rouleau, G.A.: Dissection of Genetic Factors Associated with Amyotrophic Lateral Sclerosis. *Expert Review of Neurotherapeutics* https://doi.org/10.1586/14737175.2016.1182480

8. Kaur, S.J., McKeown, S.R., Rashid, S.: Mutant SOD1 Mediated Pathogenesis of Amyotrophic Lateral Sclerosis. *Gene* https://doi.org/10.1016/j.gene.2015.11.049 https://www.nature.com/articles/srep44606
9. The ALS Association: Genetics of ALS. https://www.als.org/research/als-research-topics/genetics
10. Neumann, M., Rademakers, R., Roeber, S., Baker, M., Kretzschmar, H.A., Mackenzie, I.R.: A new subtype of frontotemporal lobar degeneration with FUS pathology. *Brain* (2009). https://doi.org/10.1093/brain/awp214 https://link.springer.com/article/10.1007/s00401-010-0698-6
11. Jarmuż-Szymczak, M., Saran, T., Bandurski, T., Kuśmirek, S., Małodobra-Mazur, M., Stępień, A.: Incidence and morphology of secondary TDP-43 proteinopathies – Part 1. *Folia Neuropathologica* (2023).
12. University of Kentucky: LATE: Changes in Neurodegenerative Diseases.
13. Josephs, K.A., Murray, M.E., Whitwell, J.L., Parisi, J.E., Petrucelli, L., Jack, C.R., Dickson, D.W.: Updated TDP-43 in Alzheimer's Disease Staging Scheme. *Acta Neuropathologica* (2022). https://doi.org/10.1007/s00401-021-02391-w
14. Rosen, D.R., Siddique, T., Patterson, D., et al.: Mutations in Cu/Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis. *Nature* https://doi.org/10.1038/362059a0
15. Kwiatkowski, T.J. Jr., Bosco, D.A., Leclerc, A.L., et al.: Mutations in the FUS/TLS gene on chromosome 16 cause familial amyotrophic lateral sclerosis. *Science* https://doi.org/10.1126/science.1166066
16. Gregory, J.M., Fagegaltier, D., Phatnani, H., Harms, M.B.: Genetics of Amyotrophic Lateral Sclerosis. *Current Genetic Medicine Reports* (2020). https://doi.org/10.1007/s40142-020-00194-8
17. Scarselli, F., Gori, M., Tsoi, A.C., Hagenbuchner, M., Monfardini, G.: The graph neural network model. *IEEE Transactions on Neural Networks* (2009). https://doi.org/10.1109/TNN.2008.2005605
18. Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S., Reblitz-Richardson, O.: Captum: A unified and generic model interpretability library for PyTorch. *arXiv preprint* (2020). https://doi.org/10.48550/arXiv.2009.07896
19. Hamilton, W.L., Ying, R., Leskovec, J.: Inductive representation learning on large graphs. *Advances in Neural Information Processing Systems* (2017). https://doi.org/10.48550/arXiv.1706.02216
20. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G.: Gene ontology: Tool for the unification of biology. *Nature Genetics* (2000). https://doi.org/10.1038/75556
21. Renton, A.E., Majounie, E., Waite, A., Simon-Sanchez, J., Rollinson, S., Gibbs, J.R., Schymick, J.C., Laaksovirta, H., van Swieten, J.C., Myllykangas, L., Kalimo, H., Paetau, A., Abramzon, Y., Remes, A.M., Kaganovich, A., Scholz, S.W., Duckworth, J., Ding, J., Harmer, D.W., Hernandez, D.G., Johnson, J.O., Mok, K., Ryten, M., Trabzuni, D., Guerreiro, R.J., Orrell, R.W., Neal, J., Murray, A., Pearson, J., Jansen, I.E., Sondervan, D., Seelaar, H., Blake, D., Young, K., Halliwell, N., Callister, J.B., Toulson, G., Richardson, A., Gerhard, A., Snowden, J., Mann, D., Neary, D., Nalls, M.A., Peuralinna, T., Jansson, L., Isoviita, V.M., Kaivorinne, A.L., Hölttä-Vuori, M., Ikonen, E., Sulkava, R., Benatar, M., Wuu, J., Chiò, A., Restagno, G., Borghero, G., Sabatelli, M., Con-

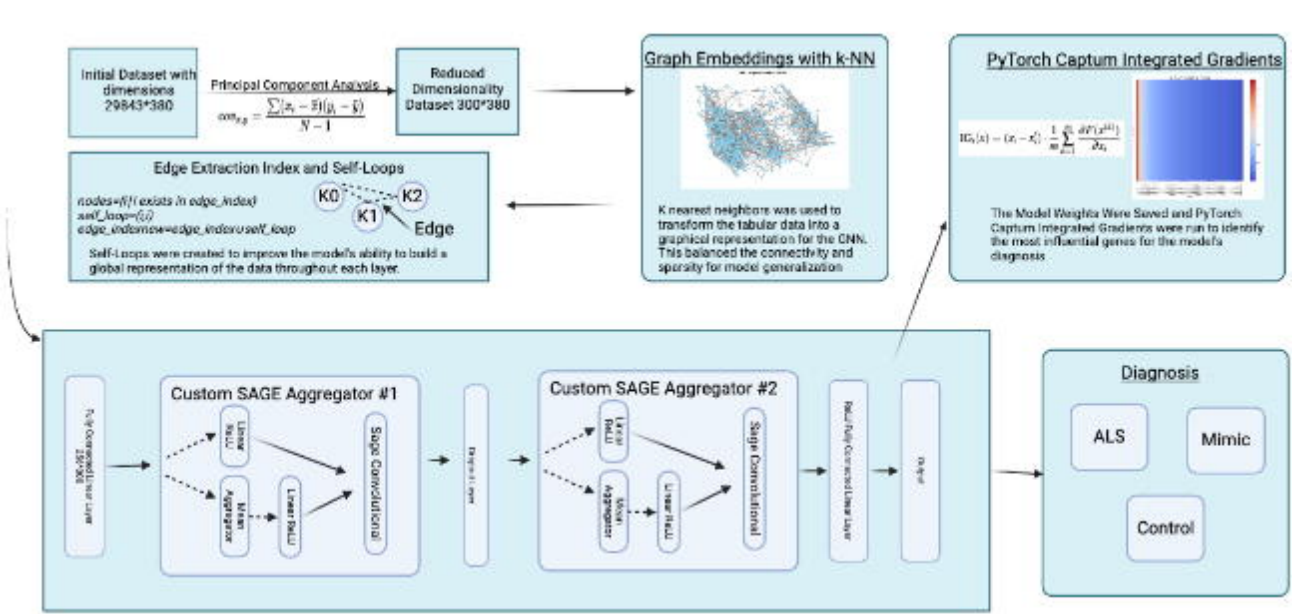
sortium, I., Heckerman, D., Rogaeva, E., Zinman, L., Rothstein, J.D., Sendtner, M., Drepper, C., Eichler, E.E., Alkan, C., Abdulla, M.H., Pack, S.D., Dutra, A., Pak, E., Hardy, J., Singleton, A.B., Williams, N.M., Heutink, P., Pickering-Brown, S., Morris, H.R., Tienari, P.J., Traynor, B.J.: A hexanucleotide repeat expansion in C9orf72 is the cause of chromosome 9p21-linked ALS-FTD. *Neuron* (2011). https://doi.org/10.1016/j.neuron.2011.09.010 https://pmc.ncbi.nlm.nih.gov/articles/PMC3322422/

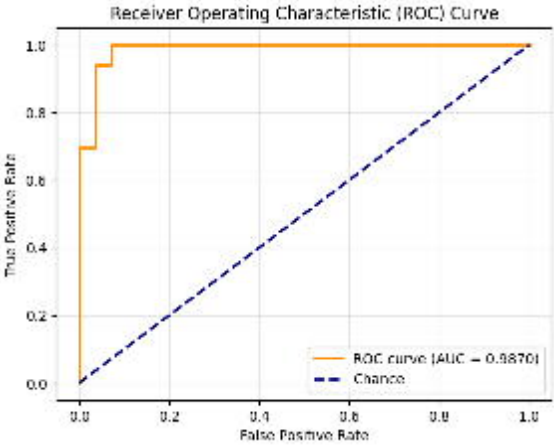
22. DeJesus-Hernandez, M., Mackenzie, I.R., Boeve, B.F., Boxer, A.L., Baker, M., Rutherford, N.J., Nicholson, A.M., Finch, N.A., Flynn, H., Adamson, J., Kouri, N., Wojtas, A., Sengdy, P., Hsiung, G.Y., Karydas, A., Seeley, W.W., Josephs, K.A., Coppola, G., Geschwind, D.H., Wszolek, Z.K., Feldman, H., Knopman, D.S., Petersen, R.C., Miller, B.L., Dickson, D.W., Boylan, K.B., Graff-Radford, N.R., Rademakers, R.: Expanded GGGGCC hexanucleotide repeat in noncoding region of C9orf72 causes chromosome 9p-linked FTD and ALS. *Neuron* (2011). https://doi.org/10.1016/j.neuron.2011.09.011 https://pubmed.ncbi.nlm.nih.gov/21944778/
23. Majounie, E., Renton, A.E., Mok, K., Dopper, E.G., Waite, A., Rollinson, S., Chiò, A., Restagno, G., Nicolaou, N., Simon-Sanchez, J., van Swieten, J.C., Abramzon, Y., Johnson, J.O., Sendtner, M., Pamphlett, R., Orrell, R.W., Mead, S., Sidle, K.C., Houlden, H., Rohrer, J.D., Morrison, K.E., Pall, H., Talbot, K., Ansorge, O., Hernandez, D.G., Arepalli, S., Sabatelli, M., Mora, G., Corbo, M., Giannini, F., Calvo, A., Englund, E., Borghero, G., Floris, G.L., Remes, A.M., Laaksovirta, H., McCluskey, L., Trojanowski, J.Q., Van Deerlin, V.M., Schellenberg, G.D., Nalls, M.A., Drory, V.E., Lu, C.S., Yeh, T.H., Ishiura, H., Takahashi, Y., Tsuji, S., Le Ber, I., Brice, A., Drepper, C., Williams, N., Kirby, J., Shaw, P.J., Hardy, J., Tienari, P.J., Heutink, P., Morris, H.R., Pickering-Brown, S., Traynor, B.J.: Frequency of the C9orf72 hexanucleotide repeat expansion in patients with amyotrophic lateral sclerosis and frontotemporal dementia: a cross-sectional study. *The Lancet Neurology* (2012). https://doi.org/10.1016/S1474-4422(12)70043-1 https://pubmed.ncbi.nlm.nih.gov/22692064/
24. Neumann, M., Sampathu, D.M., Kwong, L.K., Truax, A.C., Micsenyi, M.C., Chou, T.T., Bruce, J., Schuck, T., Grossman, M., Clark, C.M., McCluskey, L.F., Miller, B.L., Masliah, E., Mackenzie, I.R., Feldman, H., Feiden, W., Kretzschmar, H.A., Trojanowski, J.Q., Lee, V.M.: Ubiquitinated TDP-43 in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. *Science* (2006). https://doi.org/10.1126/science.1134108
25. Van Rheenen, W., Diekstra, F.P., Harschnitz, O., Westeneng, H.J., van Eijk, K.R., Saris, C.G.J., Groen, E.J.N., van Es, M.A., Blauw, H.M., van Vught, P.W.J., Veldink, J.H., van den Berg, L.H.: Whole blood transcriptome analysis in amyotrophic lateral sclerosis: A biomarker study. *PLOS One* (2018). https://doi.org/10.1371/journal.pone.0198874
26. Trautwig, A.N., Fox, E.J., Dammer, E.B., Shantaraman, A., Ping, L., Duong, D.M., Levey, A., Lah, J.J., Fournier, C.N., McEachin, Z.T., Glass, J.D., Seyfried, N.T.: Network analysis of the cerebrospinal fluid proteome reveals shared and unique differences between sporadic and familial forms of amyotrophic lateral sclerosis. *bioRxiv* (2024). https://doi.org/10.1101/2024.02.29.582840
27. Benatar, M., Zhang, L., Wang, L., Granit, V., Statland, J., Barohn, R., Swenson, A., Ravits, J., Jackson, C., Burns, T.M., Trivedi, J., Pioro, E.P., Caress, J., Katz, J., McCauley, J.L., Rademakers, R., Malaspina, A., Ostrow, L.W., Wuu, J., CReATe Consortium: Validation of serum neurofilaments as prog-

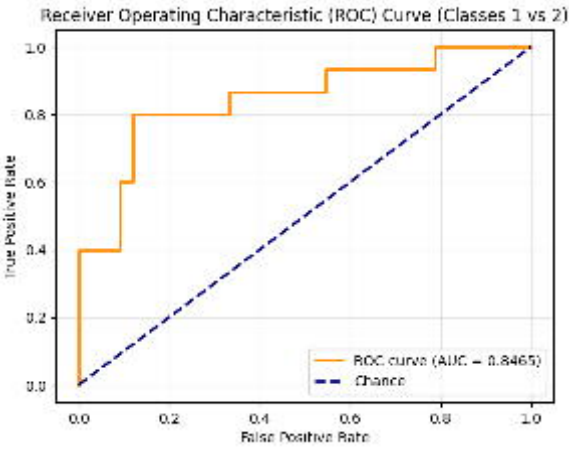
nostic and potential pharmacodynamic biomarkers for ALS. *Neurology* (2020). https://doi.org/10.1212/WNL.0000000000009559
28. Target ALS: False Positives and False Negatives: How ALS Can Be Misdiagnosed. (2022). https://www.targetals.org/2022/02/02/false-positives-and-false-negatives-how-als-can-be-misdiagnosed/
29. The ALS Association: ALS Facts (2011). http://web.alsa.org/site/PageServer?pagename=ALSA_Facts
30. Richens, J.L., Spencer, H.L., Butler, M., Cantlay, F., Vere, K.-A., Bajaj, N., Morgan, K., O'Shea, P.: Rationalising the role of Keratin 9 as a biomarker for Alzheimer's disease. *Scientific Reports* **6**, 22962 (2016). https://doi.org/10.1038/srep22962
31. Hawrylycz, M., Lein, E., Guillozet-Bongaarts, A., et al.: An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature* (2012). https://doi.org/10.1038/nature11405
32. The UniProt Consortium: UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Research* (2025). https://doi.org/10.1093/nar/gkae1010
33. Thul, P.J., Lindskog, C.: The human protein atlas: A spatial map of the human proteome. *Protein Science* (2018). https://doi.org/10.1002/pro.3307
34. Lampert, M., Scholtes, I.: The Self-Loop Paradox: Investigating the Impact of Self-Loops on Graph Neural Networks. *arXiv preprint arXiv:2312.01721* (2023). https://doi.org/10.48550/arXiv.2312.01721
35. Nguyen, H.-L., Vu, D.-L., Le, H.-C.: Exploiting Machine Learning And Gene Expression Analysis in Amyotrophic Lateral Sclerosis Diagnosis. In: *Proceedings of ICCE 2024* (2024). https://doi.org/10.1109/icce62051.2024.10634725
36. Tafuri, B., Milella, G., Filardi, M., Giugno, A., Zoccolella, S., Tamburrino, L., Gnoni, V., Urso, D., De Blasi, R., Nigro, S.S., Logroscino, G.: Machine learning-based radiomics for Amyotrophic Lateral Sclerosis diagnosis. *Expert Systems With Applications* (2023). https://doi.org/10.1016/j.eswa.2023.122585
37. Papaiz, F., Dourado, M.E.T., de Medeiros Valentim, R.A., et al.: Ensemble-imbalance-based classification for amyotrophic lateral sclerosis prognostic prediction: identifying short-survival patients at diagnosis. *BMC Medical Informatics and Decision Making* (2024). https://doi.org/10.1186/s12911-024-02484-5
38. Jamrozy, M., Maj, E., Bielecki, M., Bartoszek, M., Gołębiowski, M., Kuzma-Kozakiewicz, M.: Machine learning classificatory as a tool in the diagnosis of amyotrophic lateral sclerosis using diffusion tensor imaging parameters collected with 1.5T MRI scanner: A case study. *Electronic Journal of General Medicine* (2023). https://doi.org/10.29333/ejgm/13536
39. Chauhan, R., Sharma, U.: Exploiting speech tremors: machine learning for early diagnosis of amyotrophic lateral sclerosis. *Engineering Research Express* **6**(4), 045224 (2024). https://doi.org/10.1088/2631-8695/ad7d62
40. Lenail, B.: NN-SVG: Publication-Ready Neural Network Architecture Schematics. *Journal of Open Source Software* (2019). https://doi.org/10.21105/joss.00747
41. Rose, T., Monti, N., Anand, N., Shen, T.: PLAPT: Protein-Ligand Binding Affinity Prediction Using Pretrained Transformers. *bioRxiv* (2023). https://doi.org/10.1101/2024.02.08.575577v3
42. Bugnon, M., Röhrig, U.F., Goullieux, M., Perez, M.A.S., Daina, A., Michielin, O., Zoete, V.: SwissDock 2024: major enhancements for small-molecule docking with Attracting Cavities and AutoDock Vina. *Nucleic Acids Research* **52**(W1), W324–W332 (2024). https://doi.org/10.1093/nar/gkae300
43. Majumder, V., Gregory, J. M., Barria, M. A., Green, A., & Pal, S.: TDP-43 as a potential biomarker for amyotrophic lateral sclerosis: a systematic review and

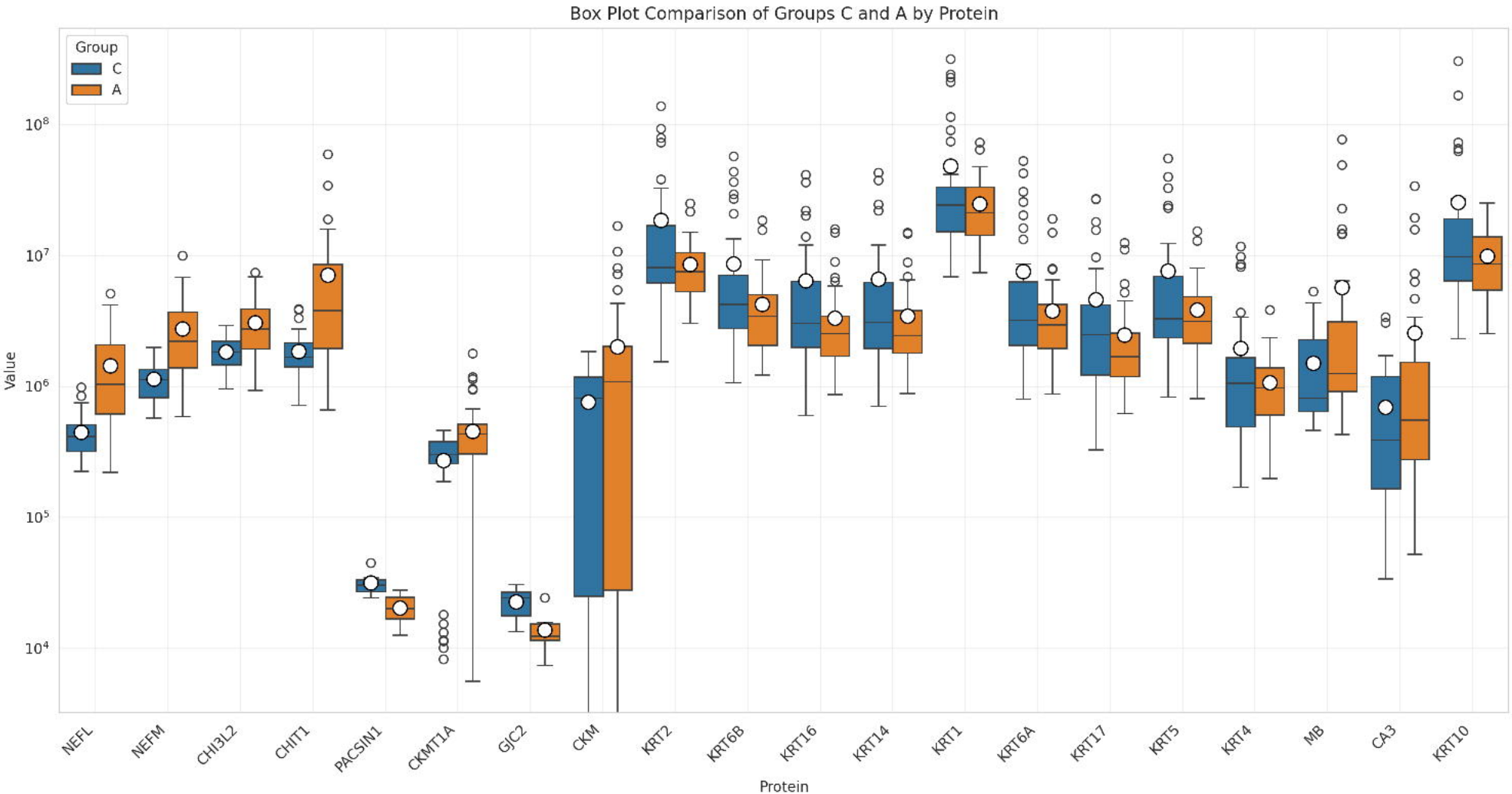
meta-analysis. *BMC Neurology* (2018). https://doi.org/10.1186/s12883-018-1091-7

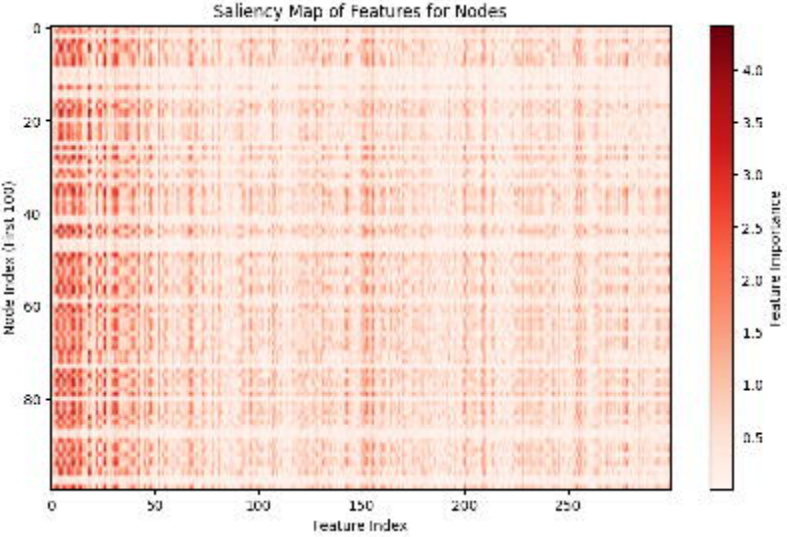
44. May, S., Hornburg, D., Schludi, M. H., Arzberger, T., Rentzsch, K., Schwenk, B. M., Grässer, F. A., Mori, K., Kremmer, E., Banzhaf-Strathmann, J., Mann, M., Meissner, F., & Edbauer, D.: C9orf72 FTLD/ALS-associated Gly-Ala dipeptide repeat proteins cause neuronal toxicity and Unc119 sequestration. *Acta Neuropathologica* (2014). https://doi.org/10.1007/s00401-014-1329-4
45. Gendron, T. F., C9ORF72 Neurofilament Study Group, Daughrity, L. M., Heckman, M. G., Diehl, N. N., Wuu, J., Miller, T. M., Pastor, P., Trojanowski, J. Q., Grossman, M., Berry, J. D., Hu, W. T., Ratti, A., Benatar, M., Silani, V., Glass, J. D., Floeter, M. K., Jeromin, A., Boylan, K. B., & Petrucelli, L.: Phosphorylated neurofilament heavy chain: A biomarker of survival for C9ORF72-associated amyotrophic lateral sclerosis. *ArX* (2017). https://doi.org/10.1002/ana.24980
46. Osanlou, K., Guettier, C., Cazenave, T., & Jacopin, E.: Planning and Learning: Path-Planning for Autonomous Vehicles, a Review of the Literature. *arXiv* (2023). https://doi.org/2207.13181 https://arxiv.org/abs/2207.13181
47. BioRender: BioRender - Science Illustration Tool. *BioRender.com* (2025). https://biorender.com *Accessed: February 10, 2025*
48. Schrödinger, LLC.: PyMOL – The PyMOL Molecular Graphics System (Version 2.5). *PyMOL* (2015). https://pymol.org/2/

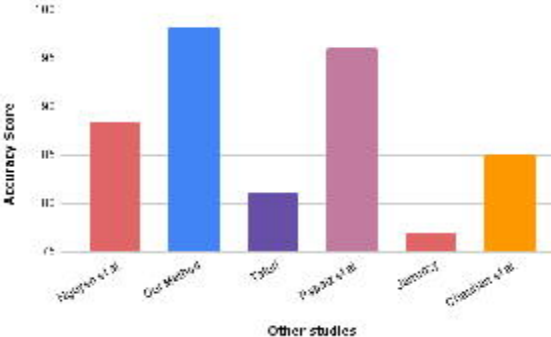

Initial Dataset with dimensions 29843*380
Principal Component Analysis
$cov_{x,y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N-1}$
Reduced Dimensionality Dataset 300*380
Graph Embeddings with k-NN
K nearest neighbors was used to transform the tabular data into a graphical representation for the CNN. This balanced the connectivity and sparsity for model generalization
PyTorch Captum Integrated Gradients
$IG_i(x) = (x_i - x_i') \cdot \frac{1}{m}\sum_{k=1}^{m} \frac{\partial F(x'^{(k)})}{\partial x_i}$
The Model Weights Were Saved and PyTorch Captum Integrated Gradients were run to identify the most influential genes for the model's diagnosis
Edge Extraction Index and Self-Loops
nodes=f(i) if exists in edge_index
self_loop=(i,i)
edge_indexnew=edge_index+self_loop
K0
K1
K2
Edge
Self-Loops were created to improve the model's ability to build a global representation of the data throughout each layer.
Custom SAGE Aggregator #1
Linear ReLU
Mean Aggregator
Linear ReLU
Sage Convolutional
Custom SAGE Aggregator #2
Linear ReLU
Mean Aggregator
Linear ReLU
Sage Convolutional
Relu Fully Connected Linear Layer
Output
Diagnosis
ALS
Mimic
Control

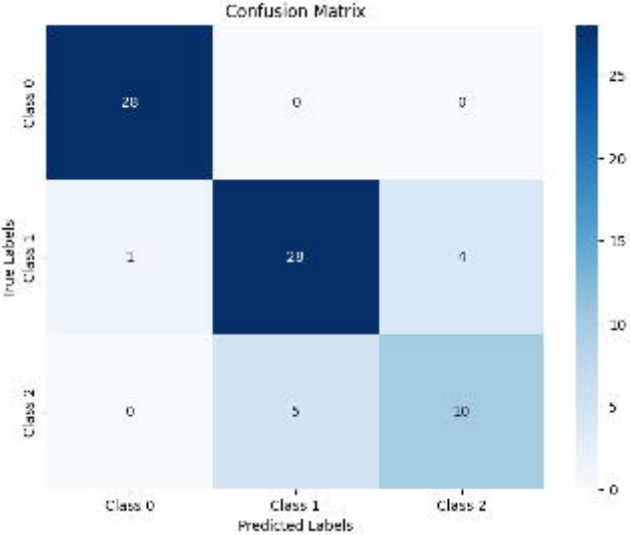

Receiver Operating Characteristic (ROC) Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (AUC = 0.9870)
Chance

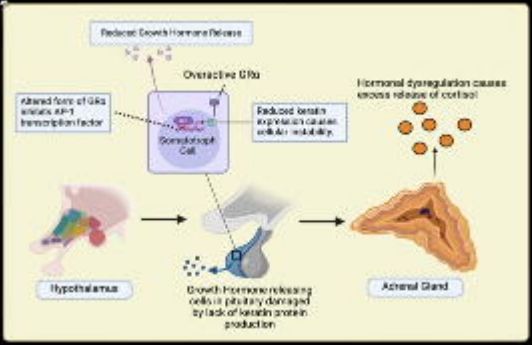

Receiver Operating Characteristic (ROC) Curve (Classes 1 vs 2)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (AUC = 0.8465)
Chance

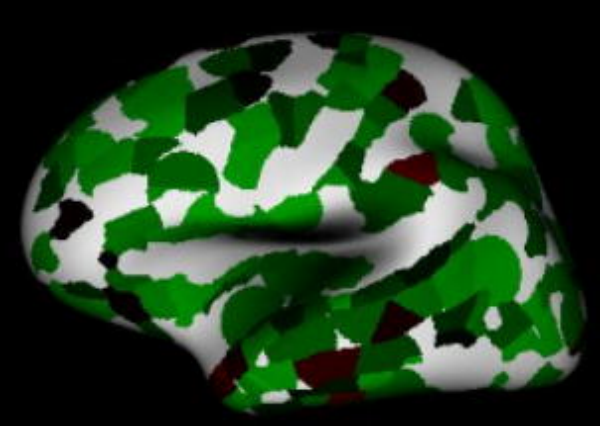

Box Plot Comparison of Groups C and A by Protein
Group
C
A
Value
10^8
10^7
10^6
10^5
10^4
NEFL
NEFM
CHI3L2
CHIT1
PACSIN1
CKMT1A
GJC2
CKM
KRT2
KRT6B
KRT16
KRT14
KRT1
KRT6A
KRT17
KRT5
KRT4
MB
CA3
KRT10
Protein

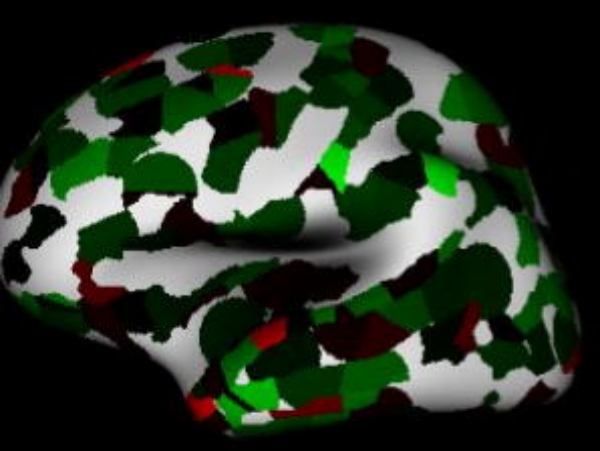

Saliency Map of Features for Nodes
Node Index (First 100)
Feature Index
Feature Importance
0
20
40
60
80
0
50
100
150
200
250
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0

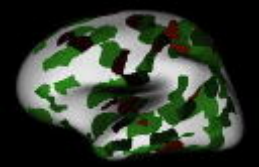

Accuracy Score
Other studies

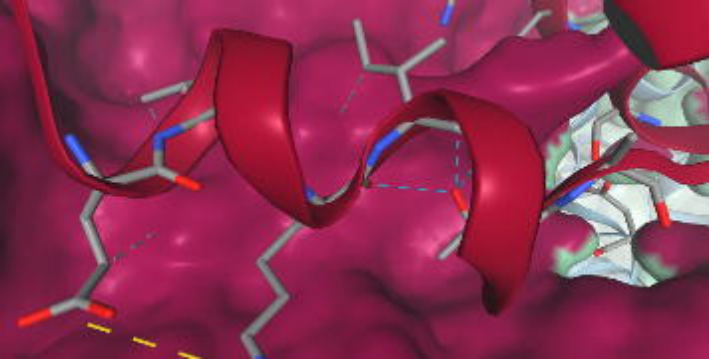

Confusion Matrix
True Labels
Class 0
Class 1
Class 2
20
0
0
1
29
1
0
5
10
Class 0
Class 1
Class 2
Predicted Labels
25
20
15
10
5
0

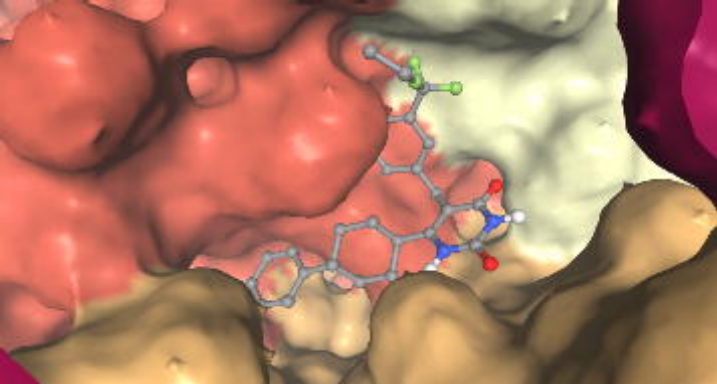

Reduced Growth Hormone Release
Overactive GRs
Altered form of GRs
Reduced keratin
Somatotropic Cell
Hormonal dysregulation causes excess release of cortisol
Hypothalamus
Growth Hormone releasing cells in pituitary damaged by lack of keratin protein production
Adrenal Gland

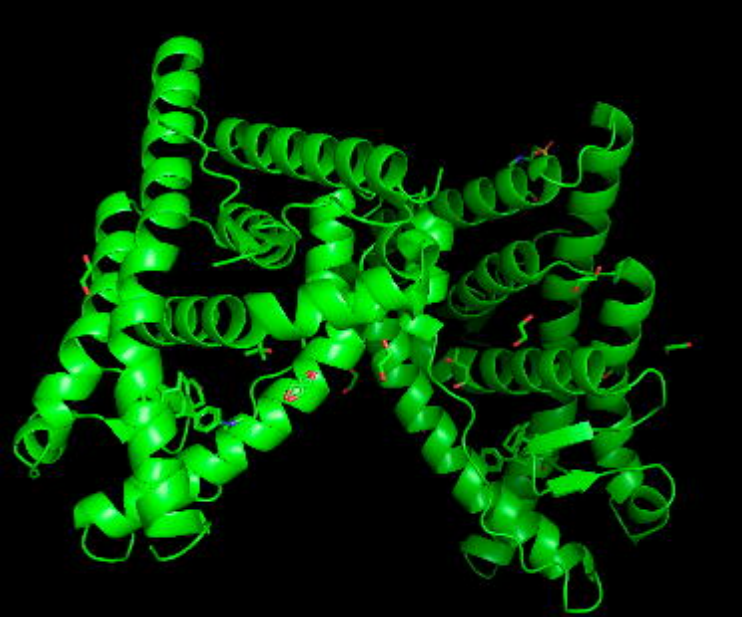


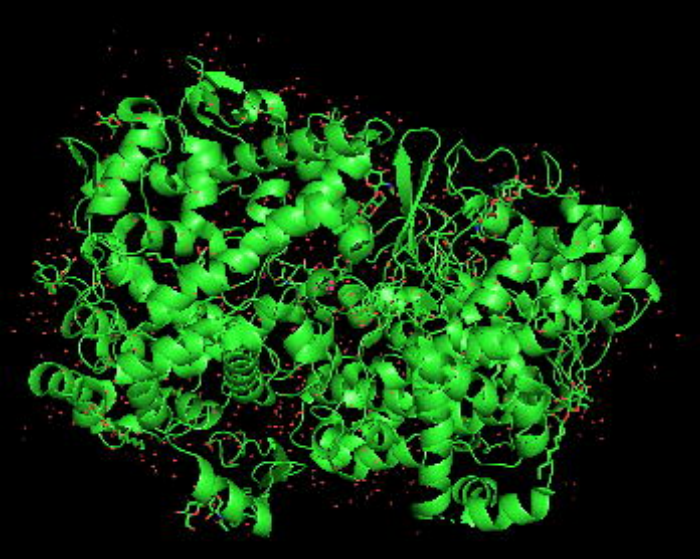


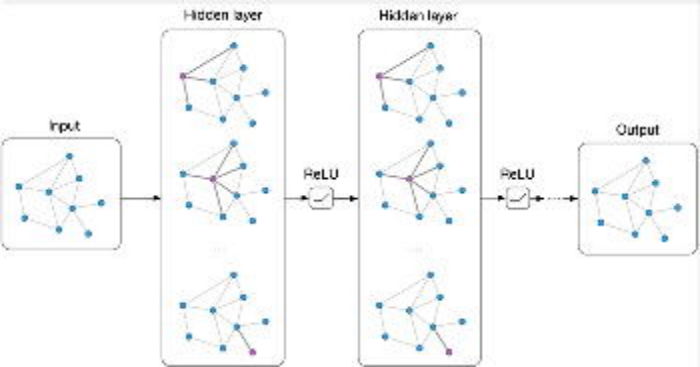


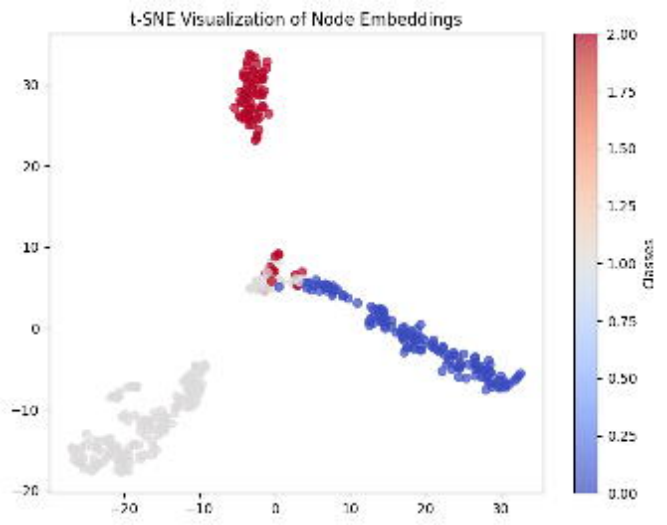









Input
Hidden layer
ReLU
Hidden layer
ReLU
Output


t-SNE Visualization of Node Embeddings
30
20
10
0
−10
−20
−20
−10
0
10
20
30
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Classes