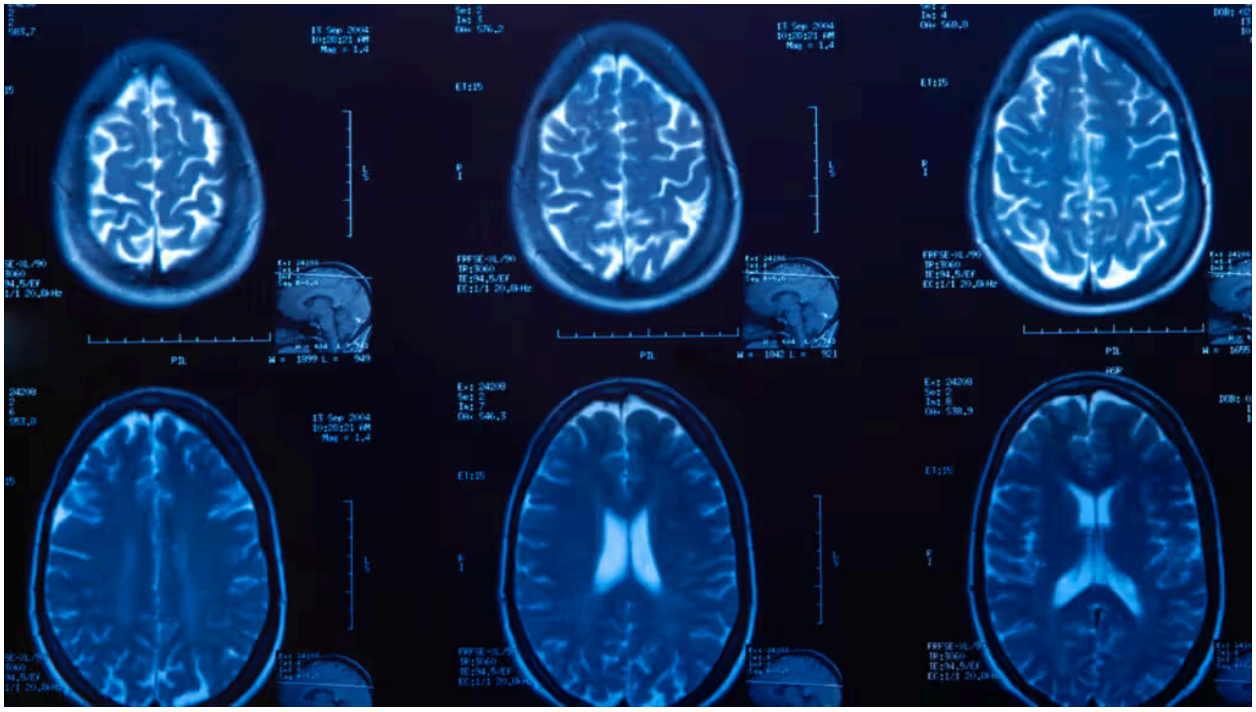


Predicting Parkinson's disease progression by protein abundance

Nicole Shum

Image: MRI scans of a Parkinson's disease patient depicting changes in the structure of the brain. (source: c_bell/Getty Images)

0. Abstract

Parkinson's disease (PD) is a progressive neurodegenerative disorder characterized by worsening motor and non-motor symptoms, driven by complex pathophysiological processes including abnormal protein aggregation. Current treatments primarily manage symptoms, highlighting a critical need for therapies that can slow disease progression. This study investigates the potential of protein abundance as a predictive biomarker for PD progression, leveraging machine learning techniques on clinical and proteomic data from the Accelerating Medicines Partnership® Parkinson's Disease (AMP-PD) initiative. We utilized a dataset comprising Unified Parkinson's Disease Rating Scale (UPDRS) scores and normalized protein expression (NPX) values for 248 PD patients. Exploratory data analysis revealed varying patterns in UPDRS scores across disease stages and identified several proteins significantly correlated with different symptom domains, implicating roles for vascular function, synaptic regulation, energy metabolism, and immune response. A predictive modeling pipeline was established, using supervised regression models (linear regression and CatBoost) to forecast UPDRS scores at 6, 12, and 18 months. Through rigorous model comparison and hyperparameter tuning, a CatBoost model with optimized parameters consistently demonstrated superior predictive performance, achieving the lowest Root Mean Squared Error (RMSE) and highest R^2 across most UPDRS subscales and timepoints. While the models showed promise in capturing overall disease progression trends, challenges remain, particularly in predicting scores at the extremes of the UPDRS scale. This study underscores the feasibility of using protein abundance with machine learning to predict PD progression, offering a robust starting point for future research aimed at developing targeted therapies and enhancing our understanding of this complex disease.

1. Introduction

1.1 Parkinson's disease

Parkinson's disease (PD) is a complex neurodegenerative disorder that affects nearly 1 million people in the US. It is most prevalent in older adults, and in the US, the overall prevalence in persons ages 45 and older is 572 people per 100,000 (Marras et al., 2018; Willis et al., 2022). It is generally characterized by a worsening ability to move over time and may include tremors, stiffness, slowing of movement, and trouble balancing. In advanced cases of the disease, dyskinesia (involuntary, uncontrolled, and repetitive movements) and psychosis (auditory and visual hallucinations or delusions) are also common (Mayo Clinic, 2024). On the molecular level, this is generally associated with Lewy bodies, or abnormal protein deposits in the brain, and dopamine-deficient neurons in the substantia nigra, a brain region crucial for movement control. However, its full pathology involves multiple areas of the nervous system, neurotransmitters, and protein aggregates. Because of the complex pathophysiology of this disease, there are currently no treatments that slow the neurodegenerative process. Furthermore, the exact cause of this disease is still unknown. Still, it is generally assumed that PD results from the interaction of genetic and environmental factors that affect fundamental biological processes on the cellular level and lead to the formation of abnormal protein aggregates (Kalia & Lang, 2015).

In patients with PD, abnormal protein aggregates form as a result of mutations in the genes that encode certain proteins. In healthy patients, protein synthesis starts with the transcription of DNA into mRNA by RNA polymerase. Then, amino acids are synthesized from the mRNA transcript by ribosomes,

with one amino acid corresponding to each codon, or three-nucleotide sequence. Specific sequences of amino acids correspond to a specific peptide, with the lengths of these sequences varying. Because of the hydrophobicity and charge of each specific amino acid, the peptides take on unique shapes in response to the environment and neighboring amino acids in a process called protein folding. A sequence of peptides makes up a full protein, with the lengths of these sequences varying. Each protein has a unique structure for its function (Sanvictores & Farci, 2020). However, in PD patients, mutations in the genetic code, even just the replacement of one nucleotide, result in different amino acids being synthesized, which results in the formation of irregularly folded proteins. These proteins cannot function properly, and signaling cascades in the body may result in the buildup of these proteins in cells (Griffiths et al., 2000).

Because of the strong correlation between PD and abnormal protein aggregates in the body, many studies have used the abundance of certain proteins as a detector or predictor of how PD progresses (McNaught et al., 2006). An increased risk of developing PD is associated with a genetic history, specifically with the mutation of the gene GBA, which encodes the protein β-glucocerebrosidase. B-glucocerebrosidase is an enzyme active in lysosomes, the structures inside cells that contain degradative enzymes (Noyce et al., 2012). Lysosomes often act as a recycling center in the cell and are used to break down non-essential cell products or components and reuse their components for other cellular processes. Therefore, it can be extrapolated that a mutation in the gene that encodes this protein would result in non-functional lysosomes and may cause a buildup of non-essential proteins, resulting in the protein aggregates characteristic of PD. Sidransky et al. (2009) showed that the odds ratio for any GBA mutation was greater than 5 in PD patients compared to healthy controls. Another relevant protein is α-synuclein, which Lewy bodies are known to be composed of consistently. α-synuclein positive aggregates follow a non-random progression in the central nervous system, progressing from the back of the brain to the front of the brain, correlating with more and more severe symptoms of PD (Licker et al., 2009).

1.2 Machine learning in disease progression

Recently, the use of artificial intelligence has been widespread in solving real-world problems, including helping clinicians identify the correct prognosis for a patient. For instance, Xiao et al. (2019) applied nine prediction models, spanning statistical, machine learning, and neural network approaches, to predict the severity of chronic kidney disease based on blood biochemical features and demographic information. They were able to identify multiple features of blood tests that showed predictive ability for chronic kidney disease severity. Another study, Pinto et al. (2020), predicted the disease progression and outcomes of multiple sclerosis based on clinical information, using a machine learning exploratory framework. Many more of these studies have been conducted on other diseases, with great impact on their fields, leading to a greater understanding of the pathologies of these diseases. Based on the success of these models in predicting the progression of other diseases, using machine learning models in predicting PD shows great promise.

1.3 Study overview

While early PD symptoms are controlled by dopaminergic therapy, treatment-related motor complications, such as motor fluctuations and dyskinesia, limit treatment in the later stages of PD

(Schapira & Olanow, 2004), therefore, there is a prevalent need to develop therapies that can slow the progression of PD to decrease the number of late-stage PD patients. This study aims to understand the most relevant proteins that predict PD progression, to further the efforts of developing these therapies. First, a preliminary analysis will be done of a dataset made up of clinical data of real PD patients, including their PD symptom progression and recorded protein abundance. Then, supervised and unsupervised machine learning methods will be applied to these datasets to build a predictive model of how protein abundance predicts PD progression. The predictive model will predict each category of the UPDRS score of a patient in 6 months, 12 months, and 18 months based on protein and peptide abundance at month 0. This illumination of the pathology of PD will hopefully lead to the development of late-stage PD therapies and show the role of machine learning in predicting the progression of PD and other neurodegenerative disorders.

2. Data overview and exploratory data analysis (EDA)

2.1 Data overview

This study will primarily use four datasets: `train_peptides.csv`, `train_proteins.csv`, `train_clinical_data.csv`, and `supplemental_clinical_data.csv`. Each dataset is representative of real PD patients and sourced from the Accelerating Medicines Partnership® Parkinson's Disease (AMP®PD), which is managed by the Foundation of the National Institutes of Health (FNIH). AMP®PD is a public-private partnership between government, industry, and nonprofit organizations that aims to identify and validate diagnostic, prognostic, and/or disease progression biomarkers for PD through clinical profiling of PD patients. The full dataset comprises over 1000 subjects with complex clinical data, and the training set used in this study includes 248 PD patients with accompanying protein and peptide data (Kaggle, 2023).

2.2 Clinical data structure

The clinical data is provided by `train_clinical_data.csv` and `supplemental_clinical_data.csv`, and gives an overview of the patients in this study and their PD progression over multiple clinic visits. `train_clinical_data.csv` comprises 248 PD patients that have accompanying protein and peptide data for training predictive models, and `supplemental_clinical_data.csv` has data on an additional 771 PD patients without accompanying protein and peptide data, providing supplemental context about the typical progression of PD. The two datasets have the same 8 columns and were merged to form an overview of PD progression for preliminary data analysis. For the preliminary, exploratory data analysis, outliers (790 rows) were removed via IQR detection. The resulting dataframe had 4048 rows and 8 columns, with 996 unique patients. The columns of this dataset are listed in Table 1.

Index	Column name	Column description
1	`visit_id`	The unique ID code for each visit, in the format "[patient_id]_[visit_month]". *(character)*

2	`patient_id`	The unique ID code for each patient. *(character)*
3	`visit_month`	Number of months since the patient's first visit, ranging from 0 (first visit) to 108 months. *(integer)*
4	`updrs_1`	The patient's score for part I of the UPDRS: non-motor aspects of daily experiences (e.g., mood and behavior). Self-reported by the patient, guided by a provider. *(double)*
5	`updrs_2`	The patient's score for part II of the UPDRS: motor aspects of daily experiences. Self-reported by the patient, guided by a provider. *(double)*
6	`updrs_3`	The patient's score for part III of the UPDRS: motor examination by a provider. *(double)*
7	`updrs_4`	The patient's score for part IV of the UPDRS: motor complications (e.g., dyskinesias, dystonia, or motor fluctuation), examined by a provider. *(double)*
8	`upd23b_clinical_state_on_medication`	Whether the patient was on medication ("On") or not ("Off"). *(factor)*

Table 1. Descriptions of each column in the clinical dataset. Variable class is indicated in italics in the column description.

A key component of understanding PD progression is understanding the Unified Parkinson's Disease Rating Scale (UPDRS) from the Movement Disorder Society (MDS). There are four subcategories, each covering a distinct category of symptoms, each requiring comprehensive evaluation or surveying by a healthcare provider. Part I considers non-motor experiences of daily living, Part II considers motor experiences of daily living, Part III comprises a motor examination, and Part IV considers motor complications resulting from treatment. Each category has a certain number of PD signs or symptoms, each rated on a 5-point Likert scale (ranging from 0 to 4) (Holden et al., 2018). The examination is composed of both self-administered surveys and direct examinations by a PD specialist. Generally, higher scores in each category indicate more severe symptoms. However, the maximum scores in each category differ: 52 points for Part I, 52 points for Part II, 132 points for Part III, and 24 points for Part IV (Movement Disorder Society, 2008).

The clinical dataset was generally of high quality, with relatively low amounts of missing data. The most missing data was found in `updrs_4` and `upd23b_clinical_state_on_medication`. Missing data in `updrs_4` reflected the importance of understanding the UPDRS, as UPDRS IV considers medication-related motor complications like dyskinesias, dystonia, or motor fluctuation (Movement Disorder Society, 2008). Therefore, for patients who did not take medication, this data is missing and reflects a purposeful exclusion of data. For `upd23b_clinical_state_on_medication`, since Parkinson's medication is quick-acting and removed from the body rapidly (Kaggle, 2023), it was

assumed that if no data was recorded, the patient was not on medication at that time. This reduced the number of NA values in the dataset from 2428 to 0. In addition, most columns of the dataframe were categorized as the appropriate class in R, excluding `patient_id` and `upd23b_clinical_state_on_medication`. `patient_id` was reclassified as character from double, and `upd23b_clinical_state_on_medication` was reclassified as factor from character. Finally, as mentioned above, outliers were removed from the dataset through IQR detection. Overall, the low amount of missing values and minimal data processing showed the clinical dataset to be a high-quality dataset and ready for preliminary data analysis.

2.3 Protein and peptide data structure

The protein and peptide data are provided by `train_proteins.csv` and `train_peptides.csv`, respectively. `train_proteins.csv` comprises the protein abundance of 248 patients, with 227 unique proteins. In total, the protein dataset has 233,741 rows and 5 columns. The columns of the protein dataset are described in Table 2. `train_peptides.csv` comprises the peptide abundance of the same 248 patients, with 968 unique peptides. Each peptide corresponds to a certain protein; however, it is important to note that the frequencies of component peptides and their respective proteins are not one-to-one, as many proteins contain repeated copies of a given peptide. The peptide dataset has 981,834 rows and 6 columns, which are described in Table 3. The 248 patients in these datasets are the same as the 248 patients in `train_clinical_data.csv` and will be used to train the predictive models later in this study. The protein and peptides data are of high quality, with no missing data. No outliers were removed from this dataset for the preliminary analysis.

Index	Column name	Column description
1	`visit_id`	The unique ID code for each visit, in the format "[patient_id]_[visit_month]". *(character)*
2	`visit_month`	Number of months since the patient's first visit, ranging from 0 (first visit) to 108 months. *(integer)*
3	`patient_id`	The unique ID code for each patient. *(character)*
4	`UniProt`	The UniProt ID code for the associated protein. *(character)*
5	`NPX`	Normalized Protein eXpression (NPX), or the frequency of the protein's occurrence in the sample. *(double)*

Table 2. Descriptions of each column in the protein dataset. Variable class is indicated in italics in the column description.

Index	Column name	Column description
1	`visit_id`	The unique ID code for each visit, in the format

		"[patient_id]_[visit_month]". *(character)*
2	`visit_month`	Number of months since the patient's first visit, ranging from 0 (first visit) to 108 months. *(integer)*
3	`patient_id`	The unique ID code for each patient. *(character)*
4	`UniProt`	The UniProt ID code for the associated protein. Usually, there are multiple peptides per protein. *(character)*
5	`Peptide`	The sequence of amino acids included in the peptide, in single-letter abbreviations for each amino acid. *(character)*
6	`PeptideAbundance`	The frequency of the peptide in the sample. *(double)*

Table 3. Descriptions of each column in the peptide dataset. Variable class is indicated in italics in the column description.

To fully understand the protein and peptide dataset, it is imperative to understand the UniProt system of protein classification. The Universal Protein Resource (UniProt) assigns a unique ID code to every protein (e.g., "O00391" for sulfhydryl oxidase 1), and provides a comprehensive summary of each protein on their website, including categories like the protein's name, sequence, organism of origin, function, active site, structure, and many more (Coudert et al., 2023). Through the use of their documented API, a dataframe of requested UniProt IDs may be accessed, along with specified columns providing summaries of the aforementioned categories. All unique UniProt IDs from the protein dataset were input into this API, resulting in `uniprot_data`, a dataframe with 227 rows and 4 columns. The columns of this dataset are listed in Table 4. The goal of this dataset is to easily access data about relevant proteins when analyzing the protein and peptide dataset. Additionally, it should be noted that the sequence of each peptide in `Peptide` is given using the single-letter abbreviations of amino acids (e.g., R for arginine, G for glycine, W for tryptophan, etc.).

Index	Column name	Column description
1	`Entry`	The UniProt ID code for the associated protein. *(character)*
2	`Protein.names`	Commonly used names for the protein. *(character)*
3	`Function..CC.`	Summary of the known functions of the protein (including citations). *(character)*
4	`Involvement.in.disease`	Summary of the known associations with diseases of the protein (including citations). *(character)*

Table 4. Descriptions of each column in the UniProt dataset. Variable class is indicated in italics in the column description.

2.4 Typical disease progression

To characterize the clinical progression of PD, clinical data from 996 patients were analyzed. These results were stratified by disease duration, which was approximated by the number of months since the patient's first visit. Disease duration was classified into early-stage (less than 2 years), mid-stage (3-5 years), and late-stage (6 or more years). To show the progression of PD symptoms, the scores from the UPDRS Parts I-IV were analyzed.

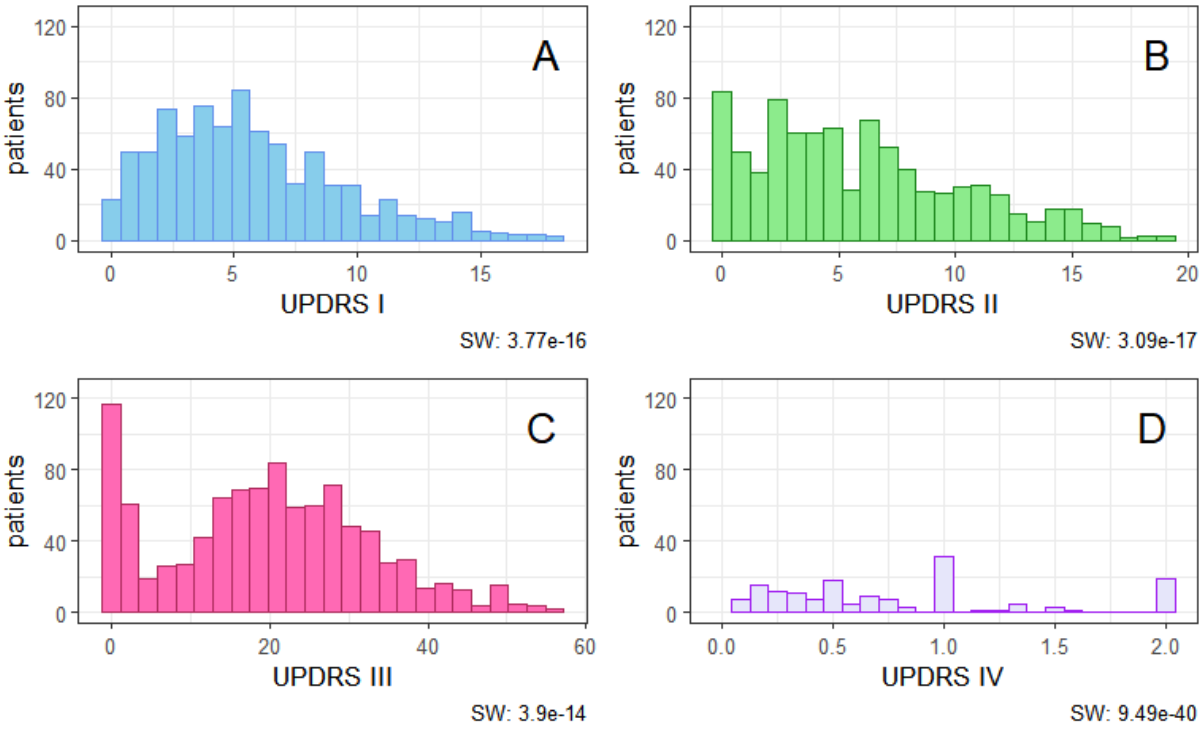


Figure 1. Distribution of mean UPDRS scores per patient across all stages. The Shapiro-Wilk's p-value is indicated at the bottom right of each graph; none showed normal distributions. **A.** UPDRS I score, or non-motor (mental) experiences; maximum score possible: 52. **B.** UPDRS II score, or motor experiences; maximum score possible: 52. **C.** UPDRS III score, or motor examination; maximum score possible: 132. **D.** UPDRS IV score, or treatment-related motor complications; maximum score possible: 24; only counting for patients on medication.

UPDRS Part I scores, aggregated by patient, exhibited a mean of 5.707, a median of 5, and a standard deviation of 3.725. The interquartile range (IQR) was 5.143, with scores spanning from a minimum of 0 to a maximum of 18. UPDRS Part II scores showed a mean of 5.828, a median of 5, and a standard deviation of 4.400. The IQR was 6.333, with values ranging from 0 to 19. UPDRS Part III was the most variable among the parts, with a mean of 19.69, a median of 20, and a standard deviation of

12.910. The IQR was 17.98, and scores ranged from 0 to 56. UPDRS Part IV had a mean score of 0.01849, a median of 0, and a standard deviation of 0.434. The IQR was 0, and scores ranged from 0 to 2.

In Parkinson's patients, the percentage of sessions on medication varied greatly, exhibiting a trimodal distribution at 0%, 40%, and 100%. The number of sessions also varied, ranging from 2 to 16. The distributions of these variables can be found in Appendix 8.1.

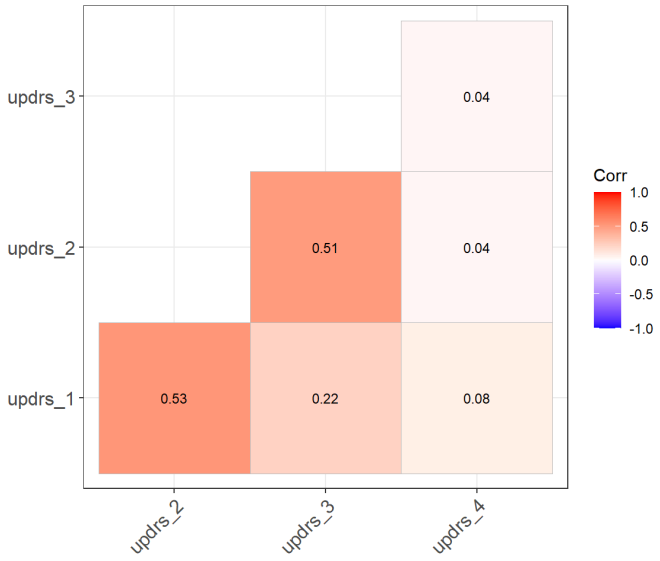


Figure 2. Correlation between different UPDRS scores.

Additionally, Pearson correlation scores were calculated between UPDRS scores of patients, shown in Figure 2. The highest correlations were between UPDRS II and UPDRS I (0.53) and between UPDRS II and UPDRS III (0.51). This indicates that when feature engineering for predictive modeling, it may be important to include other UPDRS scores as predictors.

2.5 Relevant proteins & peptides

To characterize overall protein expression profiles across all stages of disease progression, we computed summary statistics for the aggregate protein abundance per protein. The mean aggregate abundance was 2.54×10^6, with a median of 1.02×10^5. The standard deviation was 2.11×10^7, indicating a high degree of variability in protein levels across proteins. The interquartile range (IQR) was 5.19×10^5, while the range spanned from 2.23×10^3 to 3.10×10^8. The distribution of mean protein abundance is shown in Figure 3, log transformed because of the high variability in protein abundance.

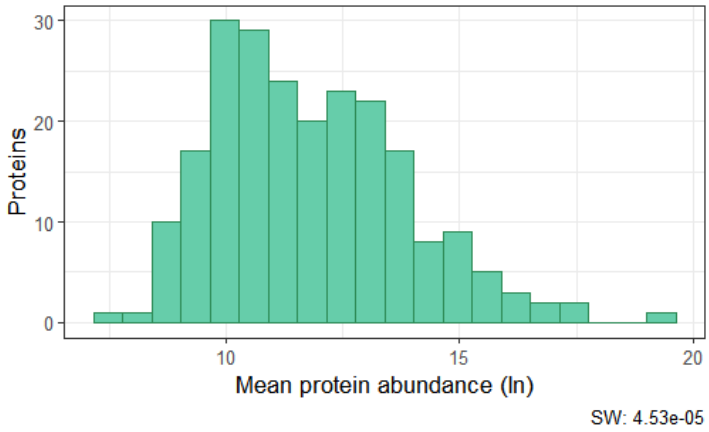


Figure 3. Distribution of mean protein abundance per protein across all stages, log transformed. The Shapiro-Wilk's p-value is indicated at the bottom right of the graph. This did not show a normal distribution.

We chose to exclude peptides from further analysis due to their one-to-one correlation with proteins, which reflects their direct derivation from the same parent molecules. Including both proteins and their corresponding peptides would introduce redundancy, as each peptide's abundance is inherently tied to that of its protein. Additionally, multiple peptides can originate from a single protein, and these repeated measures may artificially inflate the influence of certain proteins in predictive modeling. To avoid this convolution and maintain model interpretability, we focused solely on protein-level data.

2.6 Exploration of key factors affecting disease progression

To better understand the molecular determinants of Parkinson's disease progression, we conducted exploratory analyses examining the relationship between protein expression and clinical severity, as measured by the Unified Parkinson's Disease Rating Scale (UPDRS). We first calculated pairwise correlations between normalized protein expression (NPX) values and each UPDRS subscore. Although most individual correlations were modest in magnitude, several proteins consistently showed strong negative associations with multiple UPDRS parts, suggesting their potential as biomarkers of disease burden. The full list of the top 50 proteins most correlated with each UPDRS subscore is provided in Appendix 8.2; however, the top five proteins for each UPDRS subscore will be discussed here in more detail. Every protein has a function, and looking at the relationship between this function and symptoms presented in Parkinson's disease may enhance our interpretation of later predictive models.

For UPDRS Part I, the top five most strongly correlated proteins were Q06481 (APPH; -0.203), P05060 (Secretogranin-1; -0.182), P04180 (LCAT; -0.182), P17174 (Aspartate aminotransferase; -0.176), and P14618 (Pyruvate kinase (PKM); -0.165). APPH (Q06481) regulates hemostasis and G-protein signaling, suggesting a role in maintaining vascular and synaptic homeostasis (Petersen et al., 1994); a reduced concentration could impact neuroinflammatory or cognitive symptoms. Secretogranin-1 (P05060) is a neuroendocrine secretory protein, and its peptides have been implicated in modulating neurotransmission (Coudert et al., 2023); dysregulation could therefore influence mood or affective

symptoms. LCAT (P04180) plays a critical role in lipid metabolism and is expressed in the brain, where it modulates cholesterol in lipoproteins, essential for synaptic membrane fluidity and function (Tessier, 1976). Because of the observed negative correlation between UPDRS I and LCAT, it may be assumed that decreased LCAT indicates a lack of metabolic function in breaking apart protein aggregates in the brain. Aspartate aminotransferase (P17174) regulates glutamate levels, a major excitatory neurotransmitter, and is involved in neuroprotection (Shen et al., 2011); decreased activity could contribute to glutamate toxicity, relevant to mood disturbances and cognitive decline. Lastly, Pyruvate kinase PKM (P14618) governs cellular energy metabolism and also acts as a transcriptional regulator (Dombrauckas et al., 2005). Impairments in its activity may impact neuronal energy supply and gene expression, which could underlie fatigue, apathy, and other non-motor symptoms prevalent in PD.

Similar trends were observed for UPDRS Part II, where APPH (Q06481) again emerged as the most strongly correlated protein (-0.229), followed by P04180 (LCAT; -0.218), P05060 (Secretogranin-1; -0.215), O15240 (VGF; -0.214), and P43121 (MUC18; -0.201). Like with UPDRS I, lack of APPH (Q06481), LCAT (P04180), and Secretogranin-1 (P05060) may have possible effects on microvascular function and synaptic activity that could influence motor planning and execution. VGF (O15240), a neuropeptide precursor, plays essential roles in synaptic plasticity, neurogenesis, and energy balance. Specific VGF-derived peptides such as TLQP-21 and TLQP-62 have been shown to regulate metabolism, stress responses, and memory (Coudert et al., 2023), processes that indirectly support motor functioning by maintaining neuronal health and resilience. Finally, MUC18 (P43121) modulates endothelial integrity and neural crest cell adhesion (Anfosso et al., 2001). While its role in the brain is less established, its involvement in signaling and calcium dynamics may influence the vascular and neuronal microenvironment relevant to motor symptom progression.

For UPDRS Part III, O15240 (VGF; -0.229), P13521 (Secretogranin-2; -0.223), O00533 (L1CAM-like; -0.222), P0560 (Secretogranin-1; -0.206), and Q06481 (APPH; -0.201) were the top-ranked. These proteins suggest a strong involvement of synaptic plasticity, neurosecretory pathways, and cell adhesion in Parkinson's motor dysfunction. As above, lack of VGF (O15240), Secretogranin-1 (P0560), and APPH (Q06481) all had effects on motor symptom progression, similar to UPDRS II. Secretogranin-2 (P13521) is a granin-family neuroendocrine protein critical for secretory granule formation and peptide release, notably producing secretoneurin, which has neurotrophic and neuroprotective roles (Hotta et al., 2009). The neural cell adhesion molecule L1-like protein (O00533) supports neurite outgrowth, regulates GABAergic synapse efficiency, and maintains neuronal positioning and survival, mechanisms vital for motor coordination and recovery from neurodegeneration (Coudert et al., 2023). Collectively, these proteins point to impaired neurosecretory granule function, disrupted neuroplasticity, and weakened neuronal adhesion as contributing factors to the motor deficits assessed in UPDRS Part III.

Finally, for Part IV, P04217 (Alpha-1B-glycoprotein; -0.182), P05155 (C1 Inhibitor; -0.149), P04211 (Immunoglobulin lambda variable 7-43; -0.145), P02774 (Vitamin D-binding protein; -0.136), and P02747 (Complement C1q subunit C; -0.126) showed the strongest correlations. These proteins highlight a potential link between immune modulation, inflammation, and treatment-related complications in Parkinson's Disease (PD). While the precise function of Alpha-1B-glycoprotein (P04217) remains unknown, it has been associated with immune regulation and cancer biology and may reflect systemic responses to chronic neurodegeneration (Bachtiar et al., 2010). C1 inhibitor (P05155), a serpin that tightly regulates the complement cascade and coagulation pathways, may protect against

neuroinflammatory damage and blood-brain barrier disruption (Aulak et al., 1993), both relevant in the context of levodopa-induced dyskinesias and other motor complications. Immunoglobulin lambda variable region protein (P04211) represents adaptive immune activity (Teng & Papavasiliou, 2007) and could point to peripheral immune activation in PD progression or as a response to long-term pharmacologic treatment. Vitamin D-binding protein (P02774) supports transport of vitamin D, which has neuroprotective and anti-inflammatory roles, while also participating in actin scavenging and macrophage activation, suggesting a role in immune cell regulation and tissue homeostasis (Nagasawa et al., 2005). Lastly, C1q subunit C (P02747) initiates the classical complement cascade, mediating immune responses to neuronal debris and synaptic remodeling, processes that may be exacerbated by treatment-induced fluctuations (Coudert et al., 2023). Together, these proteins implicate dysregulation of immune and inflammatory pathways as contributors to the motor complications captured by UPDRS Part IV, possibly reflecting both disease pathology and responses to dopaminergic therapies.

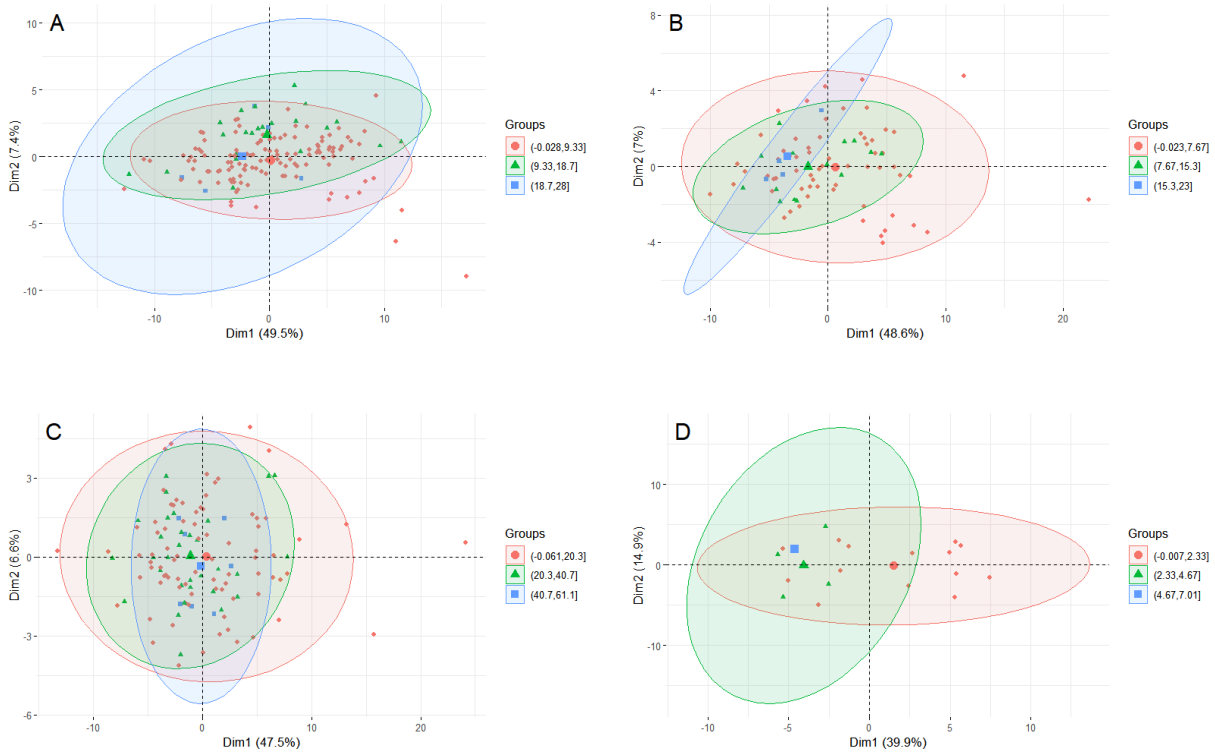


Figure 4. PCA plots showing variation in protein expression across visits, with samples grouped by tertiles of UPDRS scores. Ellipses represent 95% confidence intervals for each UPDRS tertile group. **A.** UPDRS I. **B.** UPDRS II. **C.** UPDRS III. **D.** UPDRS IV.

To further examine patterns in protein expression, we performed principal component analysis (PCA) on the full protein dataset and visualized the first two dimensions, which together captured a substantial proportion of the total variance. Samples were grouped into tertiles based on their UPDRS scores to assess whether protein expression patterns differed across disease severity levels. As shown in Figure 4, PCA plots for each UPDRS part (A-D) revealed partial separation of score tertiles, with

overlapping but distinct clusters and 95% confidence ellipses. This separation was most pronounced in UPDRS I and II, suggesting that certain protein expression signatures are more predictive of non-motor and daily living impairments. While overlap remains, these preliminary visualizations support the feasibility of using protein-based features to stratify patients by disease severity.

2.7 Discussion of EDA results

This section summarizes a comprehensive analysis of Parkinson's disease (PD) progression using clinical and protein data from 996 patients. Disease severity was assessed using the Unified Parkinson's Disease Rating Scale (UPDRS) Parts I-IV and stratified by disease stage. UPDRS Part III (motor symptoms) showed the greatest variability among patients. Medication usage patterns varied widely across visits, and UPDRS subscores were moderately correlated with each other, particularly between Parts I and II, and II and III, highlighting their potential interdependence in modeling disease burden. Protein abundance data were also analyzed, and peptides were excluded to avoid redundancy, given their derivation from proteins. The protein dataset showed highly variable expression levels, prompting log transformation for downstream analysis.

Further exploration linked protein expression to clinical severity. Several statistical questions emerge: Does protein abundance affect symptom severity? Are there patient subgroups with distinct protein expression profiles and UPDRS patterns? Which proteins are most predictive of specific symptom domains? Several proteins were consistently and negatively correlated with UPDRS scores, suggesting they may serve as biomarkers for disease burden. For example, APPH, LCAT, and Secretogranin-1 were negatively associated with non-motor symptoms (UPDRS I), daily living impairments (UPDRS II), and motor dysfunction (UPDRS III), implicating roles in vascular function, synaptic regulation, and energy metabolism. Motor complications (UPDRS IV) were associated with proteins involved in immune response and inflammation, such as C1 inhibitor and complement proteins. Principal component analysis revealed partial clustering of patients by disease severity, particularly for non-motor symptoms, supporting the idea that protein expression profiles may help stratify patients by disease stage and symptom profile. These findings generate testable hypotheses for future modeling and biomarker validation studies.

3. Methodology

3.1 Data manipulation

The primary goal of this study was to build predictive models capable of forecasting future values of Unified Parkinson's Disease Rating Scale (UPDRS) scores at multiple timepoints. Specifically, separate regression models were trained for each combination of UPDRS subscore (Parts I-IV) and future offset (6, 12, and 18 months). The modeling approach followed a standard train-test paradigm, where 80% of the data was randomly assigned to a training set and the remaining 20% was reserved for testing. This partitioning ensured that model performance was evaluated on unseen data, providing a realistic estimate of predictive generalization.

To prepare the data, protein abundance measurements were first pivoted from long to wide format such that each row represented a unique visit and each column represented the normalized protein expression (NPX) value for a distinct UniProt identifier. Clinical data, including patient identifiers,

visit month, and UPDRS scores, were then merged into this dataset using `visit_id` as the linking key. This integration allowed each modeling instance to incorporate both molecular and clinical features from a single time point as predictors. Missing data were handled in a two-pronged fashion. Rows with missing future UPDRS scores (i.e., the target variable) were dropped from each model-specific dataset to ensure that only complete cases were included in training and evaluation. For the protein features, missing values were assumed to reflect low or undetectable abundance. Accordingly, these NA values were imputed as zeros. This assumption is reflective of the data collection, which employed a broad-spectrum protein sensor, where non-detection typically implies negligible concentration rather than measurement error or data loss (Leca-Bouvier & Blum, 2005).

To evaluate model performance, root mean squared error (RMSE) was used as the primary metric. RMSE quantifies the average magnitude of the prediction error and is particularly appropriate for continuous-valued outcomes such as UPDRS scores. It is sensitive to large deviations between predicted and observed values, thus favoring models that produce accurate and consistent estimates.

3.2 Feature engineering

Feature engineering was conducted to assess the impact of dimensionality reduction on model performance. Specifically, the top 50 proteins most correlated with each UPDRS subscore were selected as input features. Correlation was computed using the training data only, and selection was repeated separately for each model configuration. Models using these reduced sets of features were compared with baseline models trained on all available protein features. This comparison helped evaluate whether restricting input to the most relevant proteins improved predictive accuracy or model robustness.

3.3 Model 0: Linear regression

The first model, referred to as Model 0, served as a baseline or “naive” comparator for the more complex machine learning approaches developed later in the analysis. This model used standard multiple linear regression to predict future UPDRS scores from protein abundance features. Linear regression was selected for its simplicity, interpretability, and widespread use in biomedical research, particularly for initial exploratory modeling.

By fitting a linear relationship between input features (protein NPX values) and the target UPDRS score at a future time point, this model established a benchmark level of performance. It assumed additive and linear effects from each protein, without capturing potential interactions or nonlinear dependencies. Given the high dimensionality and possible multicollinearity among protein features, this model is likely limited in capturing the full complexity of the underlying biology. However, its role was not to maximize predictive accuracy, but rather to act as a foundational reference against which the performance gains of more sophisticated models (e.g., CatBoost) could be objectively compared.

3.4 Model A: CatBoost

Model A employed CatBoost, a gradient boosting decision tree algorithm developed for high performance on tabular data. CatBoost was selected due to its ability to handle categorical variables natively, robustness to overfitting, and strong performance in many real-world machine learning tasks. Additionally, CatBoost efficiently manages missing values and automatically accounts for feature

interactions and nonlinearity, making it especially well-suited for complex biological datasets where relationships between protein expression and clinical scores may not be purely additive or linear.

For this model, all numeric protein abundance features were used without filtering, and missing protein values were imputed as zero, under the assumption that undetected proteins had negligible abundance. CatBoost was trained with default hyperparameters, and performance was evaluated on a held-out 20% test set using RMSE.

3.5 Model B: CatBoost with feature selection

Model B followed the same CatBoost framework as Model A but incorporated feature engineering to restrict the input to only the top 50 proteins most strongly correlated with each UPDRS score. This step aimed to reduce noise and overfitting by excluding weak or uninformative features. The correlation-based filtering was applied independently for each UPDRS score, allowing the model to focus on the most relevant protein signals per outcome. By reducing dimensionality, Model B aimed to improve generalizability and interpretability without sacrificing predictive performance. Like Model A, Model B used CatBoost's default training settings, and model accuracy was again assessed using RMSE on the test set.

3.6 Model C: CatBoost with hyperparameter tuning

Model C extended the CatBoost modeling approach by incorporating a grid search to optimize hyperparameters. Grid search is a brute-force method for tuning model parameters by exhaustively evaluating a predefined set of parameter combinations. In this case, key CatBoost parameters such as learning rate, depth, and number of iterations were varied across reasonable ranges (For this model: learning_rate = {0.01, 0.05, 0.1}, depth = {4, 6, 8}, iterations = {300, 500}).

Each combination was evaluated using cross-validation on the training set, and the parameter set yielding the lowest average validation root mean square error (RMSE) was selected. The final model was retrained on the full training data using the best-found parameters and evaluated on the test set. This process aimed to maximize predictive accuracy by tailoring the model to the specific characteristics of each UPDRS prediction task.

4. Results

4.1 Model performance

To evaluate predictive performance across different modeling strategies, we compared the four models in forecasting UPDRS scores at +6, +12, and +18 month intervals. Performance was assessed using root mean square error (RMSE) and R^2, with lower RMSE and higher R^2 indicating superior performance. Model C consistently outperformed the other approaches across nearly all UPDRS subscales and timepoints. This model achieved the lowest RMSE and highest R^2 in 10 out of the 12 prediction tasks, demonstrating substantial improvements in both error reduction and explained variance compared to the baseline. The results for each model can be found in Table 5.

For UPDRS I, Model C yielded the best performance across all three timepoints, with RMSEs decreasing from 6.797 (Model 0) to 4.669 (Model C) and R^2 improving from -0.766 to 0.166. Similar trends were observed for UPDRS II, where Model C again dominated, especially at the 12-month mark,

achieving the highest R^2 (0.393) and the lowest RMSE (3.831). Prediction of UPDRS III scores, typically the most challenging due to Parkinson's disease motor complications having higher variance and symptom complexity (Harrison et al., 2008), showed the largest RMSE values across models. Nonetheless, Model C significantly improved accuracy compared to Model 0, reducing RMSE from 17.421 to 10.796 at 18 months and increasing R^2 from -0.73 to 0.218. For UPDRS IV, results were more mixed. Although Model B had the lowest RMSE at 18 months (2.292), Model C remained competitive and achieved the best or near-best R^2 across all three timepoints.

	Model 0	*Model A*	*Model B*	*Model C*
UPDRS I, +6 months	6.797 / -0.766	5.470 / 0.070	5.548 / 0.043	**4.669** / **0.166**
UPDRS I, +12 months	5.773 / -0.266	5.506 / 0.109	5.690 / 0.049	**4.549** / **0.214**
UPDRS I, +18 months	6.606 / -0.304	5.921 / 0.090	6.077 / 0.042	**5.010** / **0.232**
UPDRS II, +6 months	6.206 / -0.402	6.074 / 0.115	6.321 / 0.041	**4.670** / **0.206**
UPDRS II, +12 months	5.984 / 0.069	6.217 / 0.160	6.558 / 0.065	**3.831** / **0.393**
UPDRS II, +18 months	7.521 / -0.354	5.983 / 0.169	6.233 / 0.098	**5.105** / **0.318**
UPDRS III, +6 months	13.744 / -0.215	12.530 / 0.193	12.988 / 0.133	**11.525** / **0.256**
UPDRS III, +12 months	14.82 / -0.001	14.433 / 0.221	15.255 / 0.129	**11.548** / **0.371**
UPDRS III, +18 months	17.421 / -0.73	13.434 / 0.047	13.378 / 0.055	**10.796** / **0.218**
UPDRS IV, +6 months	3.012 / -0.123	3.123 / 0.137	3.094 / 0.153	**3.000** / **0.182**
UPDRS IV, +12 months	**2.578** / -0.538	3.104 / -0.043	3.077 / -0.024	2.592 / **0.187**
UPDRS IV, +18 months	3.268 / -0.66	2.370 / 0.098	**2.292** / 0.156	2.628 / **0.196**

Table 5. RMSE and R^2 values for each predictive model. Each cell is formatted in the form "[RMSE] / [R^2]". The best model is highlighted in bold.

Figure 5 depicts the actual vs. predicted values of the two best-performing models, both from Model C. In this plot, the true observed values (actuals) are plotted on the x-axis, while the model's predicted values are plotted on the y-axis. This plot helps identify patterns of bias (e.g., consistently over- or under-predicting at certain value ranges), the presence of outliers, and heteroscedasticity (non-constant variance in errors), all of which are important for assessing model reliability and guiding improvements. From these plots, we can see that while our final predictive model does a good job of capturing the overall relationship between protein abundance and disease progression, it tends to overestimate and struggles when the true UPDRS scores are 0.

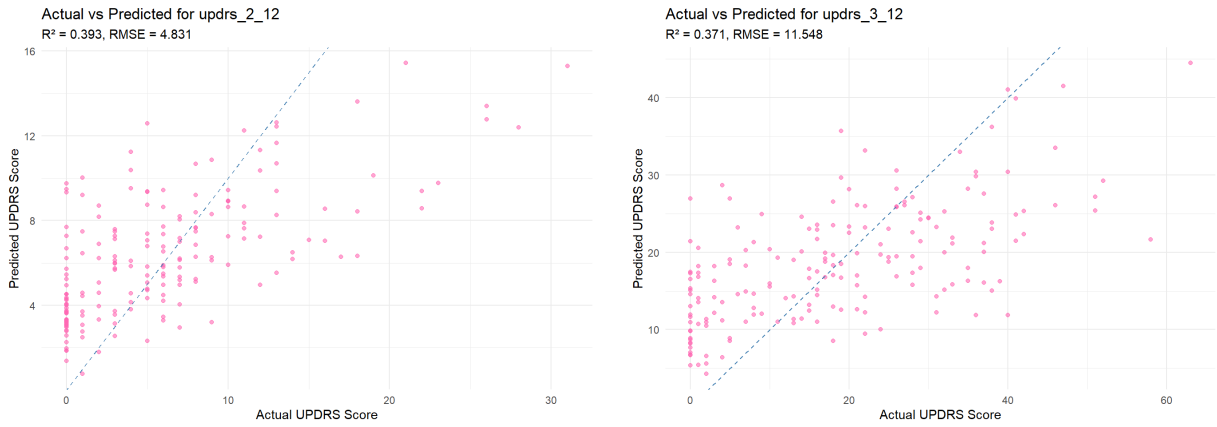


Figure 5. Actual vs. predicted values scatter plots for the best-performing models overall (Model C). See Appendix 8.3 for plots of the best-performing models for each UPDRS and time.

4.2 Most relevant features

Taking a look at the most relevant features of the best models may illuminate the functional relationship between certain proteins and Parkinson's disease progression. A full list of the 20 most important features and their importance scores for each model can be found in Appendix 8.4, but the overall most important features per model will be discussed here. As P04180, P04211, P43121, P17174, P04217, Q06481, and P13521 were already discussed in the exploratory data analysis, we will look at the ones remaining. The proteins remaining suggest a multifaceted relationship with PD, primarily revolving around key biological processes including immune response and inflammation, metabolic regulation, and cellular protein homeostasis.

Several proteins, such as ICOS ligand (O75144), Complement C3 (P01024), and various immunoglobulins (e.g., P01591, P01857, P01860, P01877, P23083, P01594, P80748, P04211, Q9Y6R7), are integral to the immune system (Teng & Papavasiliou, 2007). Their involvement points to the significant role of neuroinflammation in PD pathology, where chronic activation of immune cells and complement pathways can contribute to the degeneration of dopaminergic neurons. Similarly, proteins like Haptoglobin (P00738) and Coagulation factor XII (P00748), which are implicated in oxidative stress

response and coagulation/inflammation, respectively (Fasano, 2011), further underscore the inflammatory and oxidative stress landscape characteristic of PD.

Beyond inflammation, a substantial group of proteins highlights the importance of metabolic integrity, particularly iron and lipid metabolism, in PD. Ceruloplasmin (P00450), Hemopexin (P02790), and Ferritin light chain (P02792) are crucial for iron homeostasis (Park et al., 1999), and their dysregulation could contribute to the iron accumulation and oxidative stress seen in PD brains. Similarly, Apolipoprotein C-III (P02656), Retinol-binding protein 4 (P02753), and Phosphatidylcholine-sterol acyltransferase (P04180) suggest that aberrant lipid and vitamin transport/metabolism may impact neuronal health and function (Chan et al., 2008).

Finally, proteins involved in cellular processes and protein quality control, such as Ras GTPase-activating protein-binding protein 1 (Q13283) involved in stress granule formation, Peptidyl-prolyl cis-trans isomerase FKBP5 (Q13451) in protein folding and stress responses, and E3 ubiquitin-protein ligase UBR1 (Q8IWV7) in protein degradation (Coudert et al., 2023), are highly relevant to the protein aggregation and misfolding that defines PD. Other proteins like Neural cell adhesion molecule 2 (O15394), SPARC (P09486), and Cell surface glycoprotein MUC18 (P43121) indicate that disruptions in cellular adhesion and extracellular matrix interactions could also play a role in neuronal dysfunction (Coudert et al., 2023).

5. Discussion and reflection

5.1 Study limitations

While this study provides promising insights into the potential of protein abundance to predict Parkinson's disease (PD) progression, several limitations must be acknowledged. First, computational constraints limited the scale and complexity of the modeling pipeline. More powerful hardware would allow for more extensive hyperparameter tuning, larger ensemble models, and the exploration of deeper architectures, all of which could enhance model performance. Additionally, there was a lack of detailed metadata about the dataset, including specifics about how and when protein measurements were collected, the protocols used, and potential batch effects. This lack of transparency hinders the interpretability and reproducibility of results.

A significant limitation is the lack of demographic information about the patients, including age, sex, disease subtype, and medication history. These factors are known to influence both protein expression and disease progression and could act as confounders or modifiers in the predictive models. Including such information in future analyses would likely improve the models' accuracy and generalizability. Despite these limitations, the study lays a strong foundation for further research using machine learning to understand and ultimately slow the progression of PD.

5.2 Future considerations

Overall, this study presents exciting prospects for predictive modeling in the world of healthcare. These models, while not having the best accuracy, provide a strong starting point for building the models. From a modeling standpoint, the presence of numerous zero values in the target variables posed a challenge, potentially indicating ceiling or floor effects in UPDRS scoring or irregular follow-up data. One proposed solution is to implement a hybrid modeling approach, first using a binary classification

model to predict whether a patient's UPDRS score will increase, followed by a regression model (e.g., CatBoost) to estimate the magnitude of that change. This two-stage approach may more accurately capture the underlying clinical dynamics.

6. Conclusion

This study successfully demonstrated the utility of protein abundance, coupled with machine learning methodologies, in predicting the progression of Parkinson's disease. By leveraging a comprehensive dataset from the AMP-PD initiative, we were able to build and evaluate predictive models that offer promising insights into how proteomic signatures correlate with future changes in UPDRS scores. The consistent outperformance of the hyperparameter-tuned CatBoost model across various UPDRS subscales and timepoints highlights the power of advanced machine learning algorithms in uncovering complex biological relationships that may not be apparent through traditional statistical methods.

Our exploratory analysis further illuminated potential molecular determinants of PD progression, identifying specific proteins with strong correlations to distinct motor and non-motor symptom domains. These findings reinforce the growing understanding of PD as a multifactorial disorder influenced by neuroinflammation, metabolic dysregulation, and impaired protein homeostasis. While the predictive accuracy, particularly for extreme UPDRS scores, indicates room for refinement, the models developed here provide a strong foundation for future advancements. This research contributes to the broader effort to decode the intricate pathology of PD, paving the way for the identification of novel biomarkers, the development of more personalized treatment strategies, and ultimately, a path toward therapies that can effectively slow or halt the relentless progression of this devastating neurodegenerative condition.

7. References

Anfosso, F., Bardin, N., Vivier, E., Sabatier, F., Sampol, J., & Dignat-George, F. (2001). Outside-in signaling pathway linked to CD146 engagement in human endothelial cells. *The Journal of biological chemistry, 276*(2), 1564-1569. https://doi.org/10.1074/jbc.M007065200

Aulak, K. S., Davis, A. E., 3rd, Donaldson, V. H., & Harrison, R. A. (1993). Chymotrypsin inhibitory activity of normal C1-inhibitor and a P1 Arg to His mutant: evidence for the presence of overlapping reactive centers. *Protein science : a publication of the Protein Society, 2*(5), 727-732. https://doi.org/10.1002/pro.5560020504

Bachtiar, I., Kheng, V., Wibowo, G. A., Gani, R. A., Hasan, I., Sanityoso, A., ... & Tai, S. (2010). Alpha-1-acid glycoprotein as potential biomarker for alpha-fetoprotein-low hepatocellular carcinoma. *BMC research notes, 3*, 1-8.

Chan, D. C., Chen, M. M., Ooi, E. M., & Watts, G. F. (2008). An ABC of apolipoprotein C-III: a clinically useful new cardiovascular risk factor?. International journal of clinical practice, 62(5), 799–809. https://doi.org/10.1111/j.1742-1241.2007.01678.x

Coudert, E., Gehant, S., De Castro, E., Pozzato, M., Baratin, D., Neto, T., ... & Bridge, A. (2023). Annotation of biologically relevant ligands in UniProtKB using ChEBI. *Bioinformatics, 39*(1), btac793.

Dombrauckas, J. D., Santarsiero, B. D., & Mesecar, A. D. (2005). Structural basis for tumor pyruvate kinase M2 allosteric regulation and catalysis. *Biochemistry, 44*(27), 9417-9429. https://doi.org/10.1021/bi0474923

Fasano A. (2011). Zonulin and its regulation of intestinal barrier function: the biological door to inflammation, autoimmunity, and cancer. *Physiological reviews, 91*(1), 151–175. https://doi.org/10.1152/physrev.00003.2008

Goetz, C. G., Tilley, B. C., Shaftman, S. R., Stebbins, G. T., Fahn, S., Martinez-Martin, P., ... & LaPelle, N. (2008). Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results. Movement disorders: official journal of the Movement Disorder Society, 23(15), 2129-2170.

Griffiths, W. M., Miller, A. J. F., Suzuki, J. H., Lewontin, D. T., & Gelbart, R. (2000). Chapter 14—Mutation, repair, and recombination. *An introduction to Genetic Analysis; WH Freemand and Company: New York, NY, USA*.

Harrison, M. B., Wylie, S. A., Frysinger, R. C., Patrie, J. T., Huss, D. S., Currie, L. J., & Wooten, G. F. (2009). UPDRS activity of daily living score as a marker of Parkinson's disease progression. *Movement disorders: official journal of the Movement Disorder Society, 24*(2), 224-230.

Holden, S. K., Finseth, T., Sillau, S. H., & Berman, B. D. (2018). Progression of MDS-UPDRS scores over five years in de novo Parkinson disease from the Parkinson's progression markers initiative cohort. *Movement disorders clinical practice, 5*(1), 47-53.

Hotta, K., Hosaka, M., Tanabe, A., & Takeuchi, T. (2009). Secretogranin II binds to secretogranin III and forms secretory granules with orexin, neuropeptide Y, and POMC. *The Journal of endocrinology, 202*(1), 111-121. https://doi.org/10.1677/JOE-08-0531

Kaggle, Kirsch, L., Dane S., Adam S., & Dardov, V. AMP®-Parkinson's Disease Progression Prediction. https://kaggle.com/competitions/amp-parkinsons-disease-progression-prediction, 2023. Kaggle.

Kalia, L. V., & Lang, A. E. (2015). Parkinson's disease. *The Lancet, 386*(9996), 896-912.

Leca-Bouvier, B., & Blum, L. J. (2005). Biosensors for protein detection: a review. *Analytical Letters, 38*(10), 1491-1517.

Licker, V., Kövari, E., Hochstrasser, D. F., & Burkhard, P. R. (2009). Proteomics in human Parkinson's disease research. Journal of proteomics, 73(1), 10-29.

Marras, C., Beck, J. C., Bower, J. H., Roberts, E., Ritz, B., Ross, G. W., ... & Parkinson’s Foundation P4 Group. (2018). Prevalence of Parkinson’s disease across North America. *NPJ Parkinson's disease, 4*(1), 21.

Mayo Clinic Staff. (2024, September 27). *Parkinson’s disease*. Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/parkinsons-disease/symptoms-causes/syc-2036055

McNaught, K. S. P., & Olanow, C. W. (2006). Protein aggregation in the pathogenesis of familial and sporadic Parkinson's disease. *Neurobiology of aging, 27*(4), 530-545.

Nagasawa, H., Uto, Y., Sasaki, H., Okamura, N., Murakami, A., Kubo, S., Kirk, K. L., & Hori, H. (2005). Gc protein (vitamin D-binding protein): Gc genotyping and GcMAF precursor activity. *Anticancer research, 25*(6A), 3689-3695.

Noyce, A. J., Bestwick, J. P., Silveira-Moriyama, L., Hawkes, C. H., Giovannoni, G., Lees, A. J., & Schrag, A. (2012). Meta-analysis of early nonmotor features and risk factors for Parkinson disease. *Annals*

of neurology, 72(6), 893-901.

Park, Y. S., Suzuki, K., Taniguchi, N., & Gutteridge, J. M. (1999). Glutathione peroxidase-like activity of caeruloplasmin as an important lung antioxidant. *FEBS letters, 458*(2), 133–136. https://doi.org/10.1016/s0014-5793(99)01142-4

Petersen, L. C., Bjørn, S. E., Norris, F., Norris, K., Sprecher, C., & Foster, D. C. (1994). Expression, purification and characterization of a Kunitz-type protease inhibitor domain from human amyloid precursor protein homolog. *FEBS letters, 338*(1), 53-57.

Pinto, M. F., Oliveira, H., Batista, S., Cruz, L., Pinto, M., Correia, I., ... & Teixeira, C. (2020). Prediction of disease progression and outcomes in multiple sclerosis with machine learning. *Scientific reports, 10*(1), 21038.

Sanvictores, T., & Farci, F. (2020). Biochemistry, primary protein structure.

Schapira, A. H., & Olanow, C. W. (2004). Neuroprotection in Parkinson disease: mysteries, myths, and misconceptions. *Jama, 291*(3), 358-364.

Shen, H., Damcott, C., Shuldiner, S. R., Chai, S., Yang, R., Hu, H., Gibson, Q., Ryan, K. A., Mitchell, B. D., & Gong, D. W. (2011). Genome-wide association study identifies genetic variants in GOT1 determining serum aspartate aminotransferase levels. *Journal of human genetics, 56*(11), 801-805. https://doi.org/10.1038/jhg.2011.105

Sidransky, E., Nalls, M. A., Aasly, J. O., Aharon-Peretz, J., Annesi, G., Barbosa, E. R., ... & Ziegler, S. G. (2009). Multicenter analysis of glucocerebrosidase mutations in Parkinson's disease. *New England Journal of Medicine, 361*(17), 1651-1661.

Teng, G., & Papavasiliou, F. N. (2007). Immunoglobulin somatic hypermutation. *Annual review of genetics, 41*, 107-120. https://doi.org/10.1146/annurev.genet.41.110306.130340

Tessier P. (1976). L'exophtalmie dans les maladies de Crouzon et d'Apert [Exophthalmos in Crouzon's disease and in Apert's disease]. *Bulletins et memoires de la Societe francaise d'ophtalmologie, 88*, 357-361.

Willis, A. W., Roberts, E., Beck, J. C., Fiske, B., Ross, W., Savica, R., ... & Parkinson’s Foundation P4 Group. (2022). Incidence of parkinson disease in North America. *NPJ Parkinson's Disease, 8*(1), 170.

Xiao, J., Ding, R., Xu, X., Guan, H., Feng, X., Sun, T., ... & Ye, Z. (2019). Comparison and development of machine learning tools in the prediction of chronic kidney disease progression. Journal of translational medicine, 17, 1-13.

8. Appendix

8.1 Typical disease progression

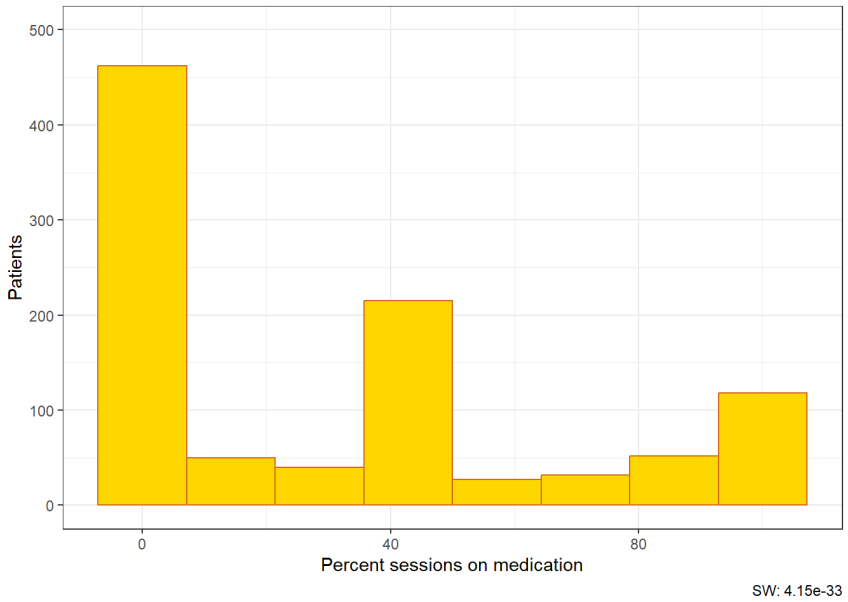


Figure 6. Distribution of the percentage of sessions on medication per patient. Shapiro-Wilks score is indicated in the bottom right corner; not normal.

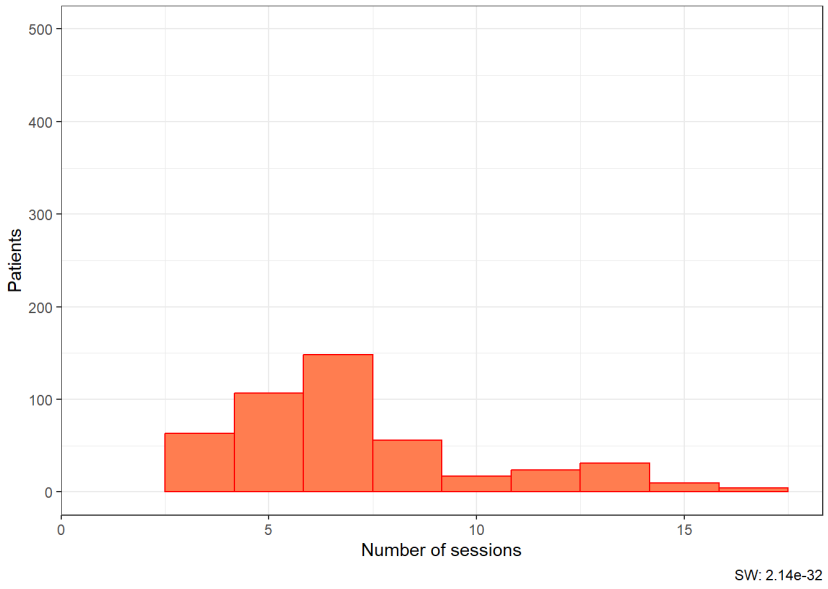


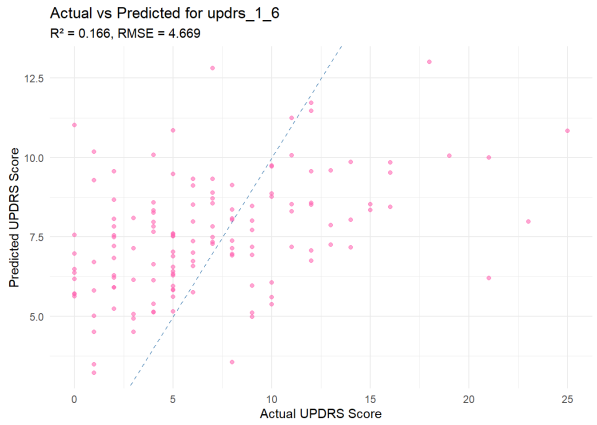
Figure 7. Distribution of the number of sessions per patient. Shapiro-Wilks score is indicated in the bottom right corner; not normal.

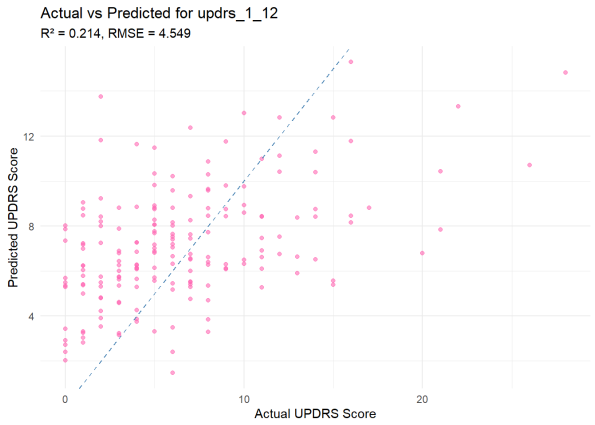
8.2 Exploration of key factors affecting disease progression

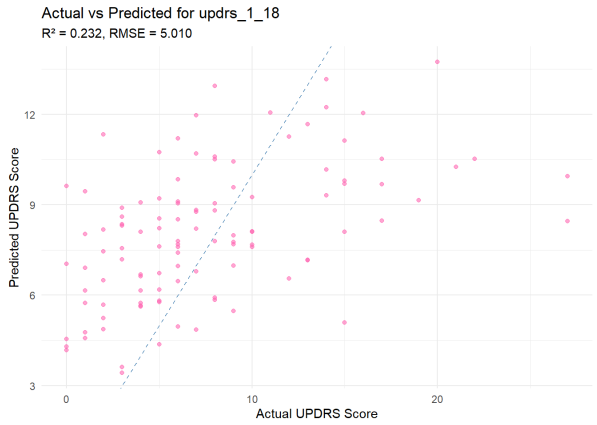
Subscore	Protein (Correlation)
UPDRS I	Q06481 (-.2), P05060 (-.18), P04180 (-.18), P17174 (-.18), P14618 (-.17), Q9BY67 (-.16), P43121 (-.16), O15240 (-.16), P02649 (-.16), Q15904 (-.15), P10645 (-.15), P13591 (-.15), O00533 (-.15), P13521 (-.14), P02787 (-.14), P23142 (-.14), P08571 (-.14), P02751 (-.14), P61916 (-.14), P04156 (-.14), P05067 (-.14), Q13332 (-.14), P13611 (-.14), P14314 (-.14), P19021 (-.14), P05452 (-.13), P09871 (-.13), Q92520 (-.13), P13987 (-.13), Q9UHG2 (-.13), Q99829 (-.13), P61278 (-.13), Q92823 (-.13), P02747 (-.13), Q14118 (-.12), Q99674 (-.12), Q6UXB8 (-.12), Q13449 (-.12), Q16610 (-.12), P09486 (-.12), Q12907 (-.12), Q02818 (-.12), P61769 (-.12), P07602 (-.12), P39060 (-.12), Q8NBJ4 (-.12), Q7Z3B1 (-.12), Q9NYU2 (-.12), P98160 (-.12), P05155 (-.11)
UPDRS II	Q06481 (-.23), P04180 (-.22), P05060 (-.22), O15240 (-.21), P43121 (-.2), P17174 (-.19), Q9BY67 (-.19), P13521 (-.19), Q9NYU2 (-.19), P02787 (-.19), O00533 (-.18), Q15904 (-.18), P10645 (-.18), P14618 (-.18), Q13332 (-.18), P02649 (-.18), P19021 (-.18), Q92823 (-.17), P02753 (-.17), O75326 (-.17), P05067 (-.17), P09104 (-.17), O14498 (-.17), P04156 (-.17), Q16610 (-.16), Q92520 (-.16), O15394 (-.16), P40925 (-.16), Q8NBJ4 (-.16), Q9NQ79 (-.16), Q99829 (-.16), P13591 (-.16), P11142 (-.15), P39060 (-.15), Q14118 (-.15), Q99674 (-.15), P55290 (-.15), Q7Z3B1 (-.15), P08133 (-.14), P04075 (-.14), P04216 (-.14), Q96KN2 (-.14), P61278 (-.14), Q9UHG2 (-.14), P07602 (-.14), P01861 (.14), P07195 (-.13), P49908 (-.13), Q12907 (-.13), Q02818 (-.13)
UPDRS III	O15240 (-.23), P13521 (-.22), O00533 (-.22), P05060 (-.21), Q06481 (-.2), P05067 (-.19), Q9BY67 (-.19), Q13332 (-.19), P17174 (-.19), P43121 (-.19), P10645 (-.19), Q92823 (-.18), Q9NYU2 (-.18), P55290 (-.18), Q92520 (-.18), P09104 (-.18), O15394 (-.18), P40925 (-.17), P02787 (-.17), Q7Z3B1 (-.17), P11142 (-.17), Q16610 (-.17), P14618 (-.17), Q9NQ79 (-.17), P04180 (-.16), P02649 (-.16), P19021 (-.16), Q14118 (-.16), Q99674 (-.16), Q14508 (-.15), P61278 (-.15), Q6UXD5 (-.15), P04075 (-.15), P13591 (-.15), P04156 (-.15), Q99435 (-.15), P08133 (-.15), O14498 (-.15), O60888 (-.15), Q8NBJ4 (-.15), Q14515 (-.14), Q15904 (-.14), Q96KN2 (-.14), P07195 (-.14), P02753 (-.14), Q9UHG2 (-.14), O43505 (-.13), P04216 (-.13), P01861 (.13), P01877 (.13)
UPDRS IV	P04217 (-.18), P05155 (-.15), P04211 (-.15), P02774 (-.14), P02747 (-.13), P39060 (-.12), Q06481 (-.12), P16152 (.12), P07225 (-.11), P20774 (-.11), P00746 (-.11), P23142 (-.11), P17174 (-.11), P02749 (-.11), P02790 (-.11), P31997 (.1), Q9NQ79 (-.1), P09486 (-.1), P06681 (-.1), P24592 (-.1), P05408 (-.1), P98160 (-.1), P19652 (-.1), P01857 (-.1), P10643 (-.1), Q99683 (-.1), P61916 (-.1), P01011 (-.1), P19021 (-.1), O75326 (-.1), Q9UKV8 (-.09), O00584 (-.09), P61626 (-.09), P05090 (-.09), O15394 (-.09), Q96BZ4 (-.09), P02753 (-.09), P00450 (-.09), P02649 (-.09), P00736 (-.08), O14498 (-.08), P02656 (-.08), P23083 (.08), Q14508 (-.08), P02787 (-.08), P01023 (-.08), P01780 (-.08), Q6UXB8 (-.08), P05060 (-.08), Q14515 (-.08)

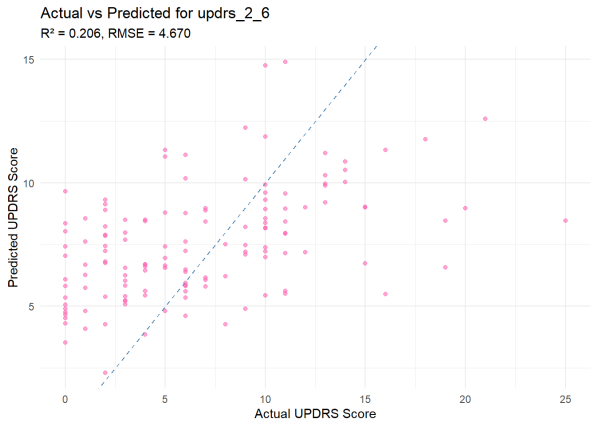
Table 6. Top 50 most correlated proteins per model (by UniProt code) and correlation (in parentheses).

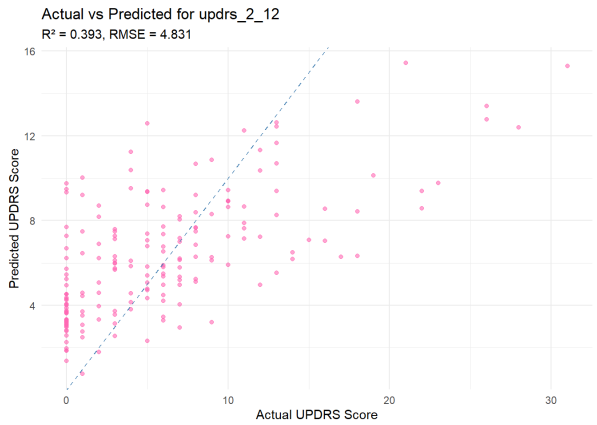
8.3 Model performance

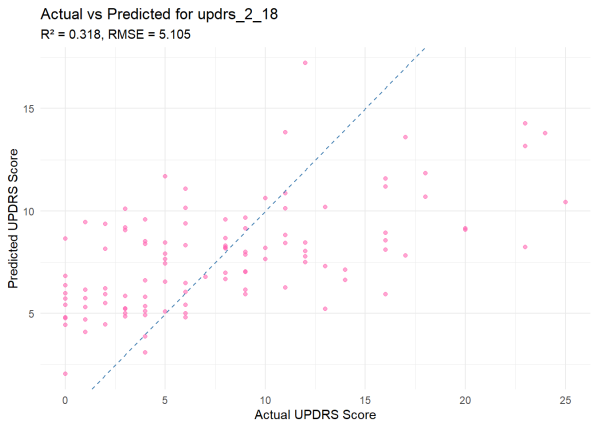












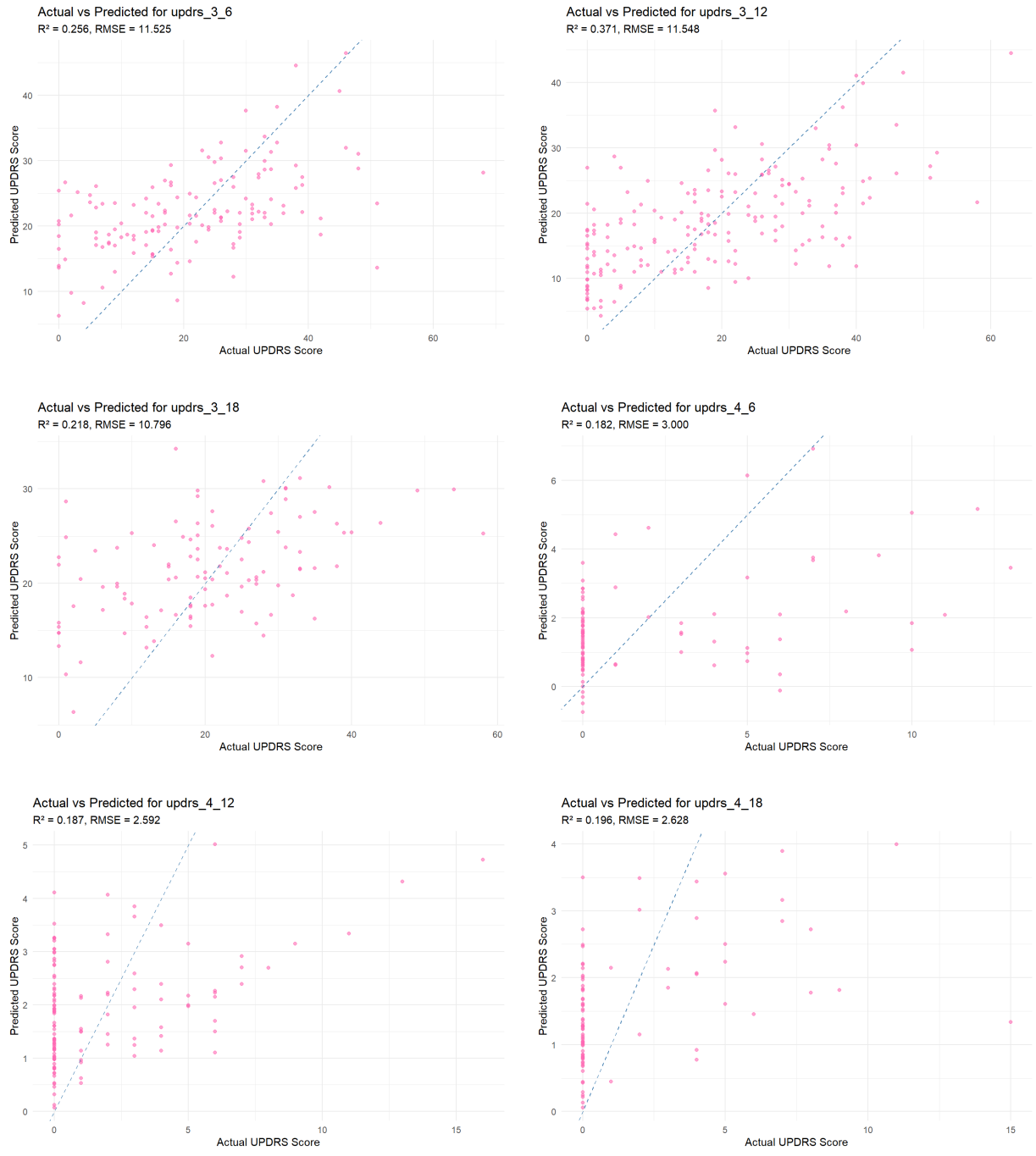


Figure 8. Actual vs. predicted values scatter plots for the best-performing models overall (Model C), for each UPDRS subscore and time.

8.4 Most relevant features

Model	Feature (Importance)

UPDRS I, +6 months	Q9Y6R7 (3.36), P04180 (2.77), P80748 (1.82), P02790 (1.64), P04211 (1.35), P01031 (1.27), Q8IWV7 (1.24), P01876 (1.24), P01024 (1.12), P08253 (1.1), P01009 (1.09), P32754 (1.06), P01877 (.96), P02655 (.95), P05060 (.94), P01717 (.9), P01860 (.88), Q96PD5 (.87), P01594 (.86), P02792 (.86)
UPDRS I, +12 months	P43121 (3.21), P04180 (2.79), Q6UXB8 (2.14), P80748 (1.82), P17174 (1.57), Q06481 (1.52), P36980 (1.51), P02790 (1.44), P02792 (1.3), O75144 (1.28), O00533 (1.27), P01009 (1.27), P02749 (1.22), P02748 (1.15), P16870 (1.15), P01860 (1.12), P00441 (1.07), Q8IWV7 (1.04), P00736 (1.04), P01780 (1.03)
UPDRS I, +18 months	Q8IWV7 (2.58), P02792 (2.27), P04217 (1.74), P04207 (1.74), P14314 (1.64), P01876 (1.59), P61916 (1.59), Q8NBJ4 (1.52), Q562R1 (1.45), P02749 (1.43), P43121 (1.37), O00584 (1.32), P17174 (1.22), P08637 (1.15), P19021 (1.11), P02790 (1.1), P01780 (1.03), P05060 (.97), P01591 (.97), P02675 (.97)
UPDRS II, +6 months	P27169 (3.51), Q9Y6R7 (3.17), P01860 (2.48), P02753 (2.17), P00450 (1.8), P80748 (1.79), P04180 (1.63), P13473 (1.58), P02763 (1.56), P02656 (1.54), P36980 (1.33), P08253 (1.17), P17174 (1.16), P02671 (1.11), P01877 (1.09), P01876 (1.03), P00748 (.95), Q7Z5P9 (.92), P13521 (.91), P61916 (.9)
UPDRS II, +12 months	Q6UXB8 (2.78), P04180 (2.37), P02753 (2.37), P36980 (2.06), P02656 (2.03), Q06481 (1.87), P27169 (1.84), O75144 (1.71), P43121 (1.71), Q9Y6R7 (1.69), P01860 (1.68), P01877 (1.61), P01591 (1.36), P00738 (1.27), P05060 (1.27), Q9Y646 (1.24), P01042 (1.22), P01009 (1.21), P01594 (1.19), Q16270 (1.18)
UPDRS II, +18 months	P27169 (2.65), P08253 (2.47), P23083 (2.05), P02753 (1.95), P04180 (1.8), Q06481 (1.5), P02763 (1.43), P25311 (1.4), P01009 (1.4), P08637 (1.36), Q9Y6R7 (1.33), P61278 (1.23), P02792 (1.2), P01594 (1.17), P01780 (1.17), Q16270 (1.12), P01861 (1.11), P54289 (1.06), Q14624 (1.), P31997 (.97)
UPDRS III, +6 months	P01591 (2.5), P01877 (2.28), Q06481 (2.), P27169 (1.97), P02753 (1.94), P02655 (1.65), Q96PD5 (1.52), P60174 (1.47), P00738 (1.42), P13521 (1.4), P04433 (1.35), O00533 (1.34), P30086 (1.31), Q12907 (1.29), P36980 (1.21), P00748 (1.17), P18065 (1.13), O00391 (1.1), P07998 (1.05), Q99969 (1.05)
UPDRS III, +12 months	P02753 (3.14), P00738 (2.37), P13521 (1.89), P01877 (1.88), P43121 (1.33), P01859 (1.25), P01857 (1.22), P32754 (1.19), P01861 (1.17), P02656 (1.12), P01876 (1.07), P36980 (1.07), P27169 (1.05), P02655 (1.03), Q6UXB8 (1.03), P05060 (.99), P61278 (.97), P01042 (.96), O00533 (.96), P08294 (.94)
UPDRS III, +18 months	P08253 (2.19), P27169 (2.11), P00748 (2.04), Q13451 (1.82), P01877 (1.31), P01857 (1.28), P04004 (1.26), P02655 (1.19), P80748 (1.18), P01608 (1.17), P04156 (1.14), P13591 (1.14), O00533 (1.1), Q99683 (1.08), P05060 (1.04), P02763 (.97), P49908 (.96), P18065 (.94), P01717 (.94), Q6UXB8 (.92)
UPDRS IV, +6 months	P04217 (6.47), P09486 (2.14), O75144 (2.03), P02656 (2.01), P04211 (1.84), P02747 (1.76), Q12907 (1.69), P80748 (1.55), Q8IWV7 (1.5), Q13332 (1.44), P23083 (1.34), P00734 (1.31), P02766 (1.29), P07225 (1.24), P19021 (1.23), P01877 (1.21), P01011

	(1.17), P08294 (1.13), Q9NYU2 (1.13), P06310 (1.13)
UPDRS IV, +12 months	Q12907 (3.81), P04211 (2.69), P04217 (2.62), P01024 (1.55), O15394 (1.5), P00736 (1.27), P02763 (1.18), P23083 (1.14), P01009 (1.12), P00734 (1.11), O14791 (1.11), P16070 (1.09), Q9NYU2 (1.08), P01877 (1.06), O00391 (1.01), Q9UBX5 (.99), P01594 (.98), P31997 (.94), P08493 (.93), P01860 (.91)
UPDRS IV, +18 months	P23083 (3.99), Q13283 (2.45), P01857 (2.31), P02452 (2.2), P01594 (1.68), P02655 (1.66), P00441 (1.51), P06454 (1.48), P43251 (1.46), P01876 (1.4), P31997 (1.4), P04217 (1.37), P07225 (1.34), O14791 (1.27), P10451 (1.24), P11142 (1.22), P54289 (1.22), P49588 (1.16), P04211 (1.1), Q12841 (1.06)

Table 7. Top 20 most important features per model (by UniProt code) and importance (in parentheses). All features are from the best model (Model C)