COMP9417 - Machine Learning and Data Mining

AMP Parkinson's Disease Progression Prediction

Group: Import pandas as np

Bill Yang (z5394798)
Hilary Cao (z5308506)
Phoebe Loh (z5361861)
Sean Wibisono (z5360345)
Shakir Ayman Azad (z5342222)

Kaggle Competition: https://www.kaggle.com/competitions/amp-parkinsons-disease-progression-prediction

Kaggle Final Submission: https://www.kaggle.com/code/seanwibi/prot-pep-model

1. Introduction

1.1 What is Parkinson's Disease?

Parkinson's disease (PD) is a chronic and progressive neurodegenerative disorder that affects how the brain is able to process movements, cognition, sleep, and other bodily functions. It is one of the most rapidly growing neurological disorders worldwide, with an expected twelve million cases by 2040. There is currently no known cure for this debilitating condition, but furthering our understanding of various biomarkers using data science will be critical in providing clues for the development of effective treatments.

1.2 What are MDS-UPDRS scores? What features will be used to help predict it?

The Unified Parkinson's Disease Rating Scale (UPDRS) is a quantitative measure of PD severity and is widely considered the gold standard for its evaluation in clinical and research settings (Ivey et al., 2012). It was revised by the Movement Disorder Society (MDS) in 2008 to consist of four parts (Goetz et al., 2008):

- Part I - Non-Motor Aspects of Experiences of Daily Living
- Part II - Motor Aspects of Experiences of Daily Living
- Part III - Motor Examination
- Part IV - Motor Complications

For each part, values are obtained through a separate questionnaire (Goetz et al., 2008a) in which questions are scored on a five-point scale of severity from 0 to 4. For example, Part I has thirteen questions and the question about sleep problems may be labelled at 4 ("I usually do not sleep for most of the night").

In this report, we will be trying to predict all MDS-UPDRS scores in different patients for every 6 months (until 24) the patient has a clinical visit and evaluation[†] . To do this, we will be investigating the features which are the best predictors of these scores. Prior to variable selection, our potential features, as suggested by research, consist of:

- The number of months since the patient's first visit,
- The abundance of proteins and sub-proteins, known as peptides, in the cerebrospinal fluid (CSF) samples collected from patients,
- Whether or not the patient was taking medication such as Levodopa (a dopamine replacement agent to treat PD) during the UPDRS assessment.

All of these features, their relationships, names, and the datasets used will be explained in Section 2. Exploratory Data Analysis.

[†] For conciseness, we will now refer to these four parts (of MDS-UPDRS) as UPDRS_1, UPDRS_2, UPDRS_3 and UPDRS_4 respectively.

1.3 How will this investigation help PD research?

Currently, there are a lot of unknowns about PD that are still yet to be discovered, one of which is that the complete set of proteins involved in PD remains an open research question. That is, not all proteins with a causative effect on PD have been uncovered to this day (Accelerating Medicines Partnership, 2023). As part of our investigation, we are exploring whether any proteins within our feature set exhibit a causal relationship with Parkinson's disease that has not been previously established in past research. In order to delve into this, we will be developing a predictive model involving all protein and peptide abundance levels provided in the datasets from Kaggle (see Section 1.4 Kaggle Competition Context). Narrowing down our feature set through both our exploratory data analysis and our model methodology will allow us to determine if there are any protein/peptide features that significantly affect the UPDRS scores. Any findings will help narrow down the focus area for developing PD cures and treatments by indicating the significance of certain peptides and proteins.

In addition to finding significant proteins that are associated with PD, we may be able to foresee the severity of PD in individual patients in a certain number of months by devising a more accurate prediction of how PD may progress through different patients (denoted by our target variable, UPDRS per month). By creating the most accurate predictive model using the provided features, we could potentially assist healthcare professionals in anticipating disease severity and devising proactive treatment strategies for PD patients.

1.4 Kaggle Competition Context

The project written in this report was submitted to the Featured Kaggle Code competition ‘AMP®-Parkinson's Disease Progression Prediction’, with a $60,000 prize pool. The best model developed within this report was able to achieve a position of tied 20/1100 (top 1.8%) at the time of writing.

2. Exploratory Data Analysis (EDA)

Prior to analysis, conducting exploratory data analysis (EDA) is crucial towards understanding the key elements of the dataset by helping provide insights into the descriptive statistics of the data, the data types and structure of the dataset, and most importantly, for our approach, the correlation between the features and the target variables.

We are provided with a total of four datasets: ‘*train_peptides*’, *train_proteins*’, ‘*train_clinical_data’* and ‘*supplemental_clinical_data’*. A model view developed within PowerBI, as shown below in Figure 1, reveals the relationship between the datasets:

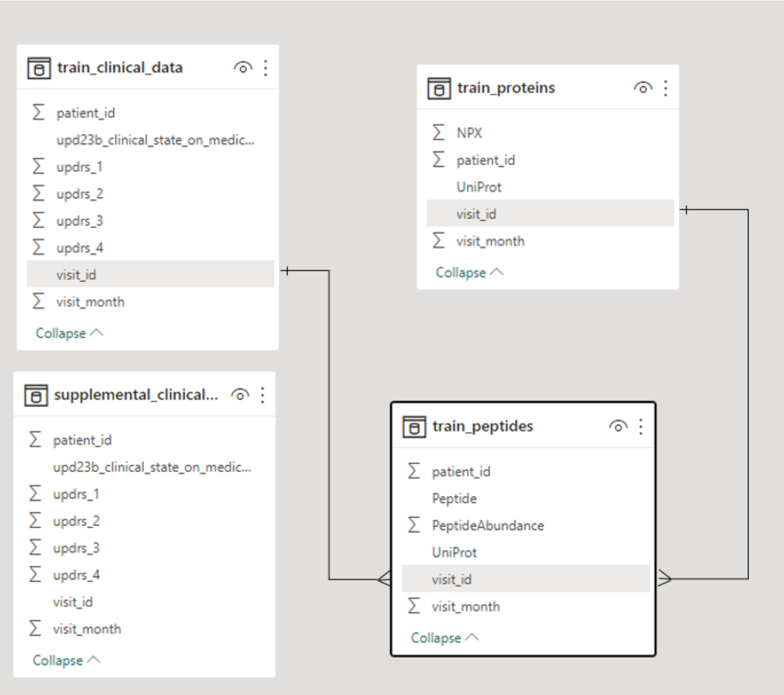


Figure 1: Visual representation of how all datasets provided are linked together

Further terminology explaining what each column of each dataset refers to is detailed in Appendix A.

2.1 Investigating Protein and Peptide Data

The *train_peptides* dataset totals six attributes: *‘visit_id’,’visit_month’, ‘patient_id’, ‘UniProt’, ‘Peptide’* and *‘PeptideAbundance’*. Upon initial exploration of this dataset, there are no null values nor duplicates.

The *train_proteins* dataset contains protein expression frequencies aggregated from the peptide data, and totals five attributes: *‘visit_id’,’visit_month’, ‘patient_id’, ‘UniProt’,* and *‘NPX’.* Although there are no null values or duplicates, Appendix B shows the frequency of protein appearances in patient visits, revealing that even the most common proteins only appear in about 40% of visits. There is thus limited data to extract trends between protein/peptide abundances and UPDRS scores.

2.2 Investigating Clinical Data

The *train_clinical_data* dataset contains patients' clinical records on their progression of PD, with their associated CSF samples. This dataset has a total of eight attributes: *'visit_id', 'visit_month', 'patient_id', 'updrs_[1-4]'* and *'upd23b_clinical_state_on_medication'*. Each sample provided has a unique *visit_id* that has its associated protein and peptide data, and the corresponding updrs[1-4] scores, which will be our target variables.

An initial exploration into the data shows that there are no duplicates, however, it should be noted that there are 1038 (40%) null values in '*updrs_4'*. A visual representation of the null values is shown below in Figure 2. It is likely that there are many null UPDRS_4 values as it is dependent on specific medications. The medication state field was not available during testing, so it is disregarded for this investigation.

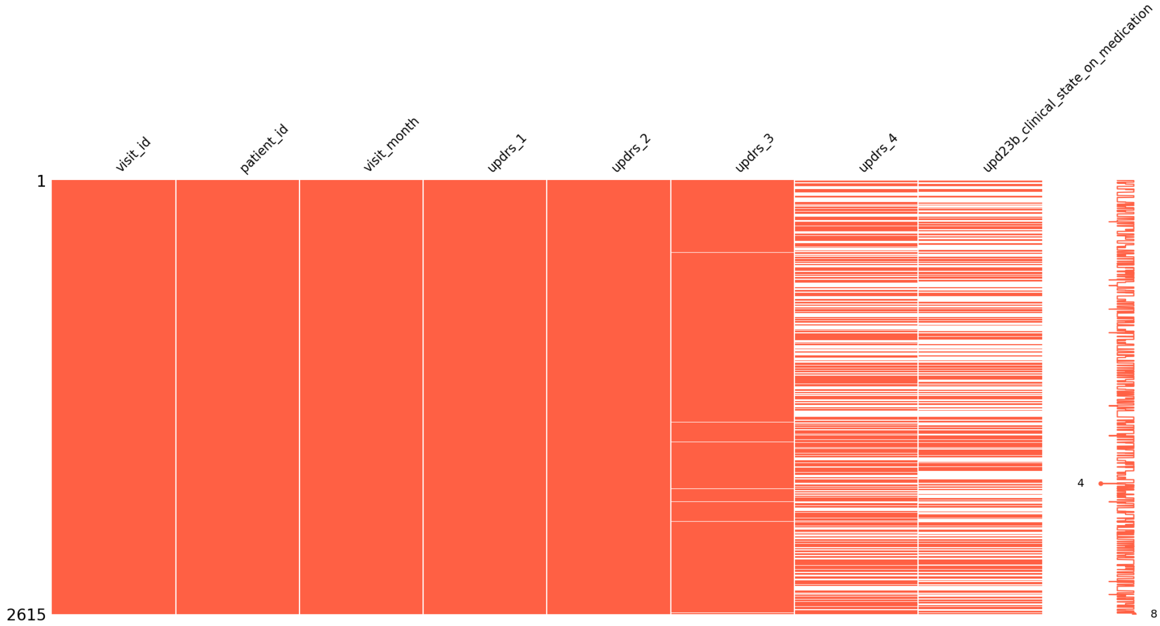


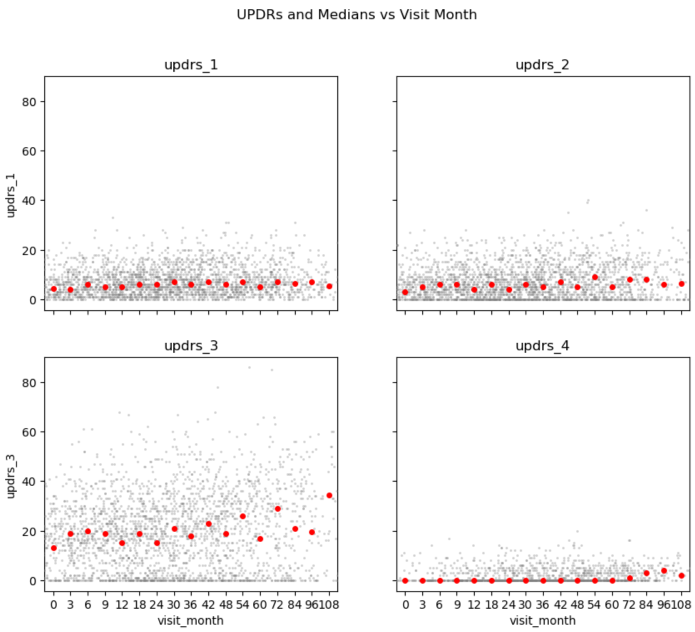
Figure 2: A missingno matrix of train_clinical_data

2.3 Investigating Supplemental Clinical Data

The *supplemental_clinical_data* dataset contains clinical records without associated CSF samples. That is, it can only be appended to *train_clinical_data* without joins to the other two datasets. There are no duplicates however, there are 928 (42%) null values in '*updrs_4'*.

2.4 Investigating the Visit_Month Feature

Due to the progressive nature of PD, we first investigated the progression of UPDRS scores versus the visit month as shown in Figure 3 below. Surprisingly, Figure 3 and Appendix C only support a minimal correlation between visit month and UPDRS scores. We would generally expect all scores to increase as PD worsens, supported by Figure 4 showing that UPDRS scores are strongly correlated with each other.



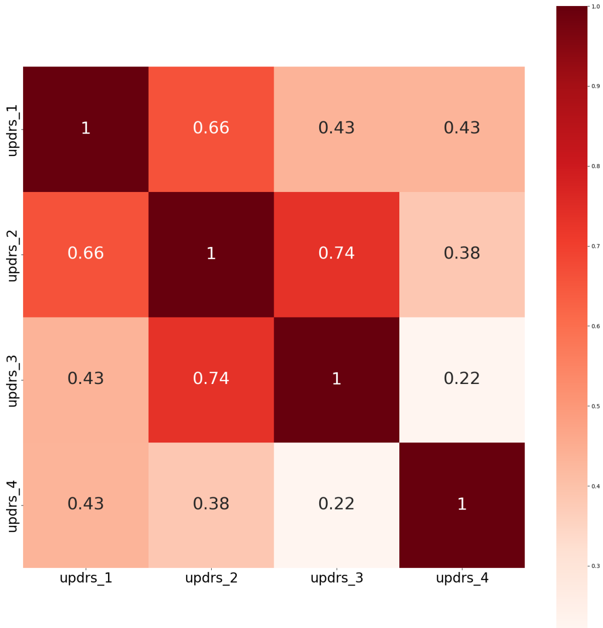


Figure 3: Strip plot of UPDRS Scores and Medians Over Time

Figure 4: Pearson's Correlation Matrix Between Each UPDRS score

2.5 Investigating All Protein and Peptide Features

Upon merging the datasets (through the two joins visualised in Figure 5 and leaving out *supplemental_clinical_data*), trends between all protein and peptide abundances and UPDRS scores were explored through a Spearman correlation matrix, revealing no strong correlation between any individual features (including *NPX* and *PeptideAbundance*) and the targets. The most highly-correlated proteins are shown in Appendix D. However, contrary to our observations, research suggests that protein combinations and peptide-protein relations likely affect the UPDRS scores (Murakami et al., 2019) (Thomas and Beal, 2007). Thus in our models, we expect feature selection and feature engineering will be required to draw predictions from interactions between proteins and peptides.

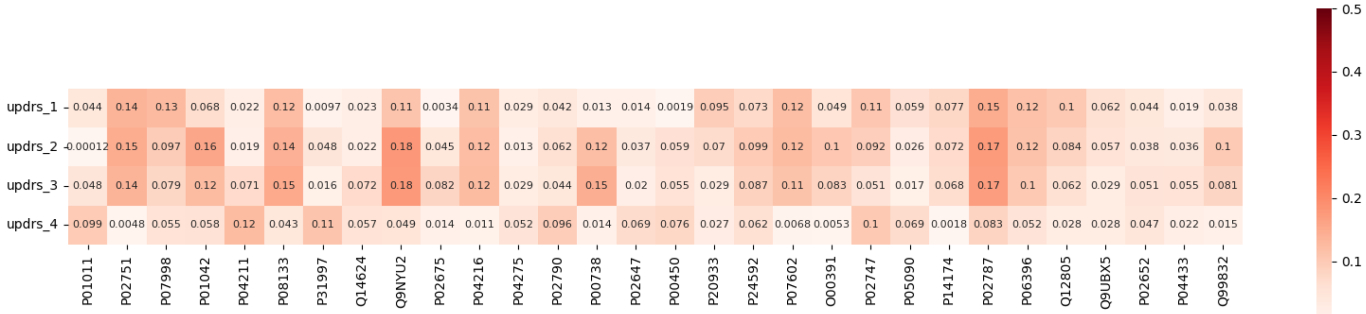


Figure 5: Spearman Correlation matrix between some protein abundance and UPDRS scores.

3. Methodology

3.1 Model Evaluation

Models were evaluated based on the primary metric of the Kaggle competition - Symmetric Mean Absolute Percentage Error (SMAPE). The use of SMAPE as the primary metric means that UPDRS scores close to 0 will have to be predicted with very high accuracy. For example, if the true value is 0, any prediction greater than 0 will incur a 200% penalty to the SMAPE score. On the basis that healthy individuals will typically get a UPDRS score of 0, models with a lower SMAPE may be suitable to act as a proxy for the classification task of predicting whether a patient will develop PD.

Note that the competition organisers changed the metric to "SMAPE + 1", where 1 is added to all truth and prediction values, thus imposing less severe penalties for 0 truth values. We adopted this change for our own evaluation, with the formula:

$$SMAPE + 1 = \frac{1}{n}\sum \frac{2 * |(actual\ \ value + 1) - (predicted\ \ value\ + 1)|}{|(actual\ \ value + 1)| + |(predicted\ \ value\ + 1)|}$$

KFold Cross Validation (CV) was used to evaluate models, splitting the dataset into 5 folds, grouped by *patient_id*, which were further split into train/test data. Grouping by patients was chosen instead of using consecutive folds as it allowed for a better evaluation of how the model would perform on a per-patient basis, which is more reflective of its real-world application. Inconsistent CV scores between folds may indicate, for example, that a model performs better on patients with higher UPDRS scores.

3.2 Baseline Models

Before modelling using protein and peptide data, we first created baseline models that were trained on only clinical data with one feature: the number of months since the patient's first visit. ('*visit_month*') Since PD degenerates over time, using a time-based feature provides a reasonable baseline to later evaluate how using protein and peptide data increases predictive power.

Studies have supported UPDRS scores increasing in a linear fashion over time (Holden et al., 2017), and hence a Linear Regression model was chosen as our baseline model. However, these linear trends were not very apparent in our own dataset, implying that there may be a more complex non-linear relationship between month and UPDRS scores. Therefore, gradient-boosting ensemble methods were also used, namely LightGBM (LGBM) and XGBoost (XGB), which are both based on decision trees. Due to the relatively small amount of data, more data-intensive methods like neural networks were not considered.

For each of the above methods, a separate model was trained for each of the UPDRS_[1-4] targets. Rows in which the target variable was missing were dropped, and then each model was trained and evaluated on the test/train data split provided by the KFold method. Default learning parameters were used for all models at this stage.

3.3 Tuned LGBM

Moving forward, LGBM was selected for the following models over XGBM due to their differences in tree growth. LGBM grows leaf-wise while XGBM grows depth-wise, resulting in different trees if full trees are not grown due to early stopping criteria. Leaf-wise growth, which evaluates a split's contribution to the global loss, may outperform depth-wise growth, which only evaluates the loss along a particular branch.

The Kaggle submission SMAPE (using hidden data) for the baseline LGBM model was worse than the local CV SMAPE, indicating that the model was likely overfitted. First, hyperparameters were manually changed, which included reducing the learning rate and the number of estimators within the ensemble model. After seeing improvements, hyperparameter optimization was conducted through a grid search, creating a search space of possible hyperparameter values and evaluating the model at every possible position:

{'learning_rate': 0.002, 'max_depth': 5, 'n_estimators': 20, 'num_leaves': 10}

Motivated by the sparsity of UPDRS_4 measurements, the next section explores a method that was used as a supplementary model to best capture trends without the need for much data.

3.4 Medians Method

A method to predict using the medians of UPDRS scores was developed in response to the large amounts of variance and outliers for the target variables. A matrix with the median of each UPDRS category at each visit month was created, also incorporating the supplementary clinical data to get more representative medians. However, due to the small amount of data, some medians decreased as *'visit_month'* increased**,** clearly not capturing the progressive nature of PD. Therefore, expanding window maximums (the highest median seen until the current visit month) were used to preserve the theoretically monotone relationship between the visit month and UPDRS scores.

Note that this was not used as an individual model but as a supplementary predictor to the following protein and peptide model in the case of missing molecule readings.

3.5 Top Proteins Method

While most protein concentrations showed low Spearman correlation to the target UPDRS scores, the first method of incorporating protein data involved choosing those with the highest correlation to specific scores. These included *'P04180'* for UPRDS_1 and UPDRS_2 and *'O00533'* for UDPRS_3, which also had >40%

appearance rate in readings. It was decided to forgo training a protein model for UPDRS_4 due to the lack of data and any significantly correlated proteins.

Separate LGBM models for UPDRS[1-3] were trained using the above protein NPX values in addition to visit month as features, before hyperparameters were similarly tuned via GridSearch (parameters specified at section 3.7). For predicting UPDRS_4 values in addition to test data in which the readings of the chosen proteins were missing, two options were explored as a supplementary model:

- Previous LGBM model trained on visit month only
- Median model with the prediction given as the expanding window max for that visit month.

In order to explore models with multiple proteins as features, it was first observed that proteins that have higher correlations to target variables often have high correlations with each other (e.g. proteins *'Q92823'* and *'O00533'* have a correlation coefficient of 0.89). Considering the small size of the dataset, including such redundant proteins may not improve the model, and may also unnecessarily increase its dimension and even introduce harmful bias. As we do not have the domain knowledge to see which protein concentrations are biologically expected to increase independently of others, the next section explores the use of Principal Component Analysis to remove redundant proteins and narrow down the features.

3.6 PCA + RandomForest

3.6.1 Data Preparation

Only 1113 out of 2615 visit_ids have associated protein and peptide data and since the protein-peptide data is the main attribute of our model, we removed visits without this data. Imputation with means or medians was not considered as there were too many missing values.

3.6.2 Feature engineering

Due to the many-to-one relationship between proteins and peptides, each protein NPX was joined onto the rows of their corresponding peptides. Therefore, peptides represented their associated proteins while capturing the effect of the protein on UPDRS progression. Interaction effects between NPX and Abundance were explored by multiplying their values. The three features for each peptide now included: i) Its Abundance, ii) its protein NPX, and iii) NPX*Abundance. So in total, we had 968 unique peptides and 2904 attributes. Since not every peptide was present in each CSF sample, those missing values were imputed with 0.

For each UPDRS_[1-3], a separate model was created to model the progression of scores at visit_month+0, visit_month+6, visit_month+12, visit_month+18, visit_month+24 and visit_month+30, totalling 18 different models. UPDRS_4 was not modelled due to insufficient data.

However, having only 1113 samples but 2904 attributes will result in a highly complex and overfitted model. Therefore Principal Component Analysis (PCA) was used to decrease dimensionality while minimising information loss. PCA works by identifying the directions (principal components) in which the data varies the most and projecting the data onto these directions to create a new set of orthogonal features. The first principal component captures the maximum variance in the data, followed by the second principal component, and so on.

First, the data was normalised and fitted to the PCA model. The number of final features for the PCA model was manually trialled in addition to selecting an arbitrary number of features to capture 95-99% of the variance. Although using Maximum Likelihood Estimator to solve for the optimal is usually ideal, it was not applicable as the number of samples was less than the number of features. PCA found that the best performance was at 300 principal components.

3.6.3 Model

Random Forest Regressor was selected as the machine learning model due to the accuracy and efficiency (Winchester et al., 2022) it has shown in previous research of the same nature such as predicting outcomes of chronic kidney disease from EMR data (Zhao, Gu and McDermaid, 2019).

To find the optimal hyperparameters, Randomised Grid Search was used as a normal Grid Search was unfeasible given the computational requirements of PCA. The optimal hyperparameters were:

{‘n_estimators’=100, ‘min_samples_split’=36, ‘min_samples_leaf’=15, ‘max_depth’=None, ‘bootstrap’=True}.

Eighteen different models were trained on data using the PCA-selected features, and KFold cross-validation was similarly used to evaluate them. Note that despite the model’s dependency on the availability of protein-peptide data in CSF, the competition hosts have stated that predictions are not required before a patient’s first CSF reading.

3.7 Final Model: Combination of Median & Protein/Peptide Models

The final-selected model, after evaluation of cross-validation scores, was the Top Proteins model using the medians method as a supplementary for missing protein readings.
Parameters selected from GridSearch were:

{'learning_rate': 0.1, 'max_depth': -1, 'n_estimators': 100, 'num_leaves': 31}

4. Results

4.1 Local Cross-Validation Results

The following graph (Figure 6) displays the average SMAPE scores across all folds using grouped KFold cross-validation described above. All models were trained as previously described, with the “Top Proteins" model supplemented with the Medians method giving the best performance, with a **53.76 SMAPE score**.

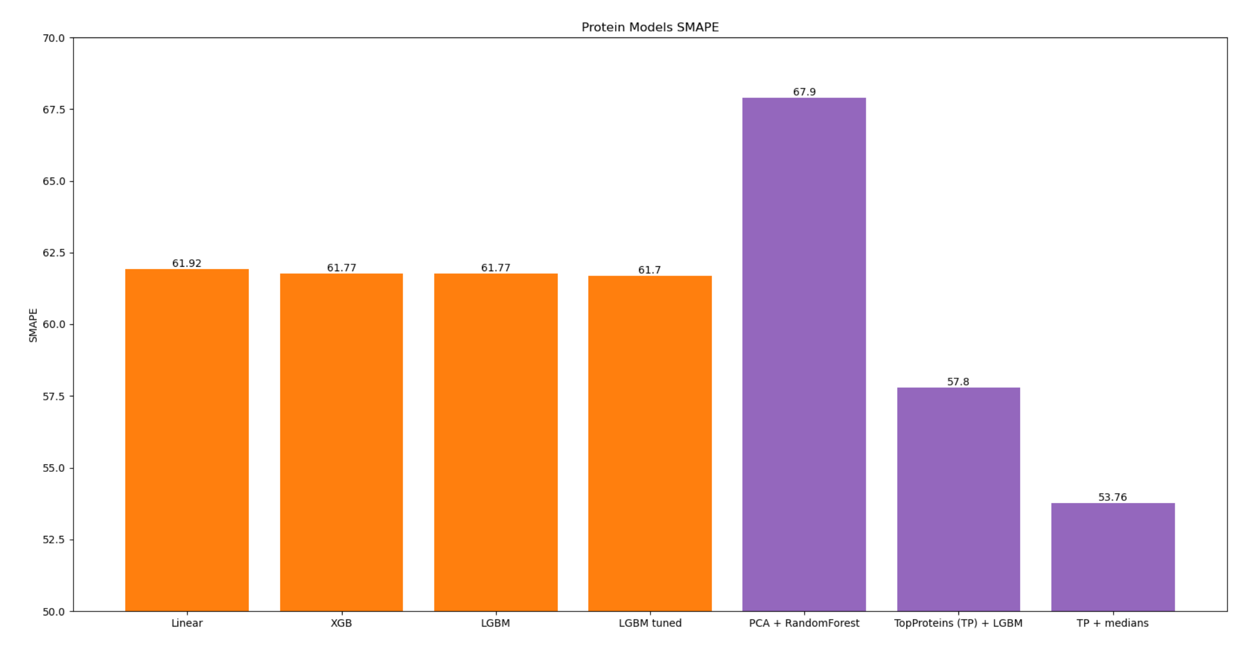


Figure 6: Graph of average SMAPE scores of each model

4.2 Kaggle Results

Our best score on the Kaggle leaderboard, which was evaluated using the same “SMAPE + 1” metric on a hidden dataset, was **56.3** with the Top Proteins model using medians as the supplementary predictor. At the time of writing, this was tied for 20th out of 1100 teams (top 1.8%). For reference, the current best score is 54.1.

5. Discussion

5.1 Baseline Models

All baseline models had similar performances, with XGBoost and LightGBM gradient boosting algorithms slightly outperforming Linear Regression. If the underlying relationship between visit month and UPDRS scores was truly linear, Linear Regression should be the most accurate model as it is able to better extrapolate beyond the observed values, especially with this limited data set. Therefore, we deduced that UPDRS scores did not increase linearly with visit month, at least not in this dataset. Tuning LGBM parameters also gave a slight boost to the cross-validation score, successfully reducing overfitting**,** to act as a baseline model for comparison against future protein and peptide models.

5.2 Protein/Peptide Models

The first method of incorporating protein data, the 'Top Proteins' model, saw a notably improved performance with a **53.7 SMAPE**. Furthermore, using medians to predict when protein readings were missing proved to be a more successful approach than using the previous LGBM model trained on '*visit_month*'. This was perhaps due to the lack of training data for UPDRS_4 proving difficult for the LGBM model to predict such scores accurately, whereas the expanding window maximum medians method was able to capture the progressive trends of PD without much data. Despite the selected proteins having a relatively low Spearman correlation to UPDRS (the highest being 0.23), this performance boost using just two proteins was promising and encouraged our successive exploration of PCA to incorporate the rest of the data.

However, the second method involving PCA to narrow down protein and peptide features performed significantly worse in cross-validation, despite trying multiple parameters for the selection of principal components as described in Section 3 - Methodology. While the purpose of applying PCA was to isolate the informative proteins and remove noise, it is probable that too many initial features were provided, especially with a limited dataset where protein measurements were spread 'too thin'. However, without the domain knowledge to know which proteins are expected to be biologically correlated with each other, it is not feasible to remove proteins based on their individual Spearman correlation. Individual correlations do not provide enough information to determine non-significant protein features, especially in this multivariate model where combinations of proteins are likely to be stronger biomarkers (Murakami et al., 2019).

5.3 Kaggle Results

It was noted that the Kaggle leaderboard results (**56.3)** were slightly worse than local cross-validation scores, despite extensive hyperparameter tuning to reduce overfitting. Since patients were different in the hidden test data, this may be indicative that the predictors used in our protein models were not as strong on these other patients. This fact emphasises PD being a highly individualised disease and the challenge of finding protein signals that apply to a range of patients.

Overall Insights

While it was difficult to model the complex and highly individual relationship of protein abundance to UPDRS scores with sparse readings, performance improvements for the Top Proteins model indicate the existence of some, albeit inconsistent, relationship between protein abundance and PD progression. We also saw the strength of very simple methods, including the use of medians, to model complex trends.

6. Conclusion

The validation of Parkinson’s Disease biomarkers are critical clues for the improvement of treatment efficacy. This project achieved some degree of success in identifying proteins that exhibited minor statistical significance in predicting the progression of UPDRS scores, specifically 'P04180' for UPRDS_1 and UPDRS_2, and 'O00533' for UPDRS_3. However, their inferior performance on hidden Kaggle data indicated that these were not strong signals and are not suitable as universal PD biomarkers. In addition, the effectiveness of using simple medians of scores has also been demonstrated, potentially able to estimate the severity of patients’ future conditions in an easily accessible way.

The biggest challenge we encountered was developing a strategy to identify and narrow down the informative proteins, especially given our limited familiarity with the biological subject matter. This emphasised the importance of truly understanding the data and discouraged the blind use of complicated machine learning techniques, as we saw limited success using PCA to narrow down the model features. We believe that researching proteins that are intercorrelated or have already been ruled out in PD research, will potentially lead to more successful feature selection.

Whilst there are many protein interactions we have not explored, our investigation provides evidence that there are **no proteins in our feature set that can be used as effective biomarkers on their own**, substantiated by our Kaggle leaderboard position. Our methods on datasets containing new proteins can quickly determine if any single protein is a useful biomarker of Parkinson’s progression, and may indicate that priority should be given to studies that analyse combinations of proteins with other proteins or even additional genetic and laboratory data.

7. References

Accelerating Medicines Partnership (2023). *AMP®-Parkinson's Disease Progression Prediction*. [online] kaggle.com. Available at: https://www.kaggle.com/competitions/amp-parkinsons-disease-progression-prediction/data [Accessed 11 Apr. 2023].

Ebrahimi-Fakhari, D., Wahlster, L. and McLean, P.J. (2012). Protein degradation pathways in Parkinson's disease: curse or blessing. *Acta Neuropathologica*, 124(2), pp.153–172. doi:https://doi.org/10.1007/s00401-012-1004-6.

Emamzadeh, F.N. and Surguchov, A. (2018). Parkinson's Disease: Biomarkers, Treatment, and Risk Factors. *Frontiers in Neuroscience*, [online] 12(612). doi:https://doi.org/10.3389/fnins.2018.00612.

Goetz, C., Fahn, S., Martinez-Martin, P., Poewe, W., Sampaio, C., Stebbins, G., Stern, M., Tilley, B., Dodel, R., Dubois, B., Holloway, R., Jankovic, J., Kulisevsky, J., Lang, A., Lees, A., Leurgans, S., Lewitt, P., Nyenhuis, D., Olanow, W. and Rascol, O. (2008a). *MDS-UPDRS*. [online] Available at: https://www.movementdisorders.org/MDS-Files1/PDFs/Rating-Scales/MDS-UPDRS_English_FINAL.pdf [Accessed 14 Apr. 2023].

Goetz, C.G., Tilley, B.C., Shaftman, S.R., Stebbins, G.T., Fahn, S., Martinez-Martin, P., Poewe, W., Sampaio, C., Stern, M.B., Dodel, R., Dubois, B., Holloway, R., Jankovic, J., Kulisevsky, J., Lang, A.E., Lees, A., Leurgans, S., LeWitt, P.A., Nyenhuis, D. and Olanow, C.W. (2008). Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results. *Movement Disorders*, [online] 23(15), pp.2129–2170. doi:https://doi.org/10.1002/mds.22340.

Holden, S.K., Finseth, T., Sillau, S.H. and Berman, B.D. (2017). 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), pp.47–53. doi:https://doi.org/10.1002/mdc3.12553.

Ivey, F.M., Katzel, L.I., Sorkin, J.D., Macko, R.F. and Shulman, L.M. (2012). The Unified Parkinson's Disease Rating Scale as a predictor of peak aerobic capacity and ambulatory function. *Journal of rehabilitation research and development*, [online] 49(8), pp.1269–76. doi:https://doi.org/10.1682/jrrd.2011.06.0103.

Murakami, H., Tokuda, T., El-Agnaf, O.M.A., Ohmichi, T., Miki, A., Ohashi, H., Owan, Y., Saito, Y., Yano, S., Tsukie, T., Ikeuchi, T. and Ono, K. (2019). Correlated levels of cerebrospinal fluid pathogenic proteins in drug-naïve Parkinson's disease. *BMC Neurology*, [online] 19(1). doi:https://doi.org/10.1186/s12883-019-1346-y.

Siderowf, A., Concha-Marambio, L., Lafontant, D.-E., Farris, C.M., Ma, Y., Urenia, P.A., Nguyen, H., Alcalay, R.N., Chahine, L.M., Foroud, T., Galasko, D., Kieburtz, K., Merchant, K., Mollenhauer, B., Poston, K.L., Seibyl, J., Simuni, T., Tanner, C.M., Weintraub, D. and Videnovic, A. (2023). Assessment of heterogeneity among participants in the Parkinson’s Progression Markers Initiative cohort using α-synuclein seed amplification: a cross-sectional study. *The Lancet Neurology*, [online] 22(5), pp.407–417. doi:https://doi.org/10.1016/S1474-4422(23)00109-6.

Thomas, B. and Beal, M.F. (2007). Parkinson’s disease. *Human Molecular Genetics*, [online] 16(R2), pp.R183–R194. doi:https://doi.org/10.1093/hmg/ddm159.

Winchester, L., Barber, I., Lawton, M., Ash, J., Liu, B., Evetts, S., Hopkins-Jones, L., Lewis, S., Bresner, C., Malpartida, A.B., Williams, N., Gentlemen, S., Wade-Martins, R., Ryan, B., Holgado-Nevado, A., Hu, M., Ben-Shlomo, Y., Grosset, D. and Lovestone, S. (2022). Identification of a possible proteomic biomarker in Parkinson’s disease: discovery and replication in blood, brain and cerebrospinal fluid. *Brain Communications*, [online] 5(1). doi:https://doi.org/10.1093/braincomms/fcac343.

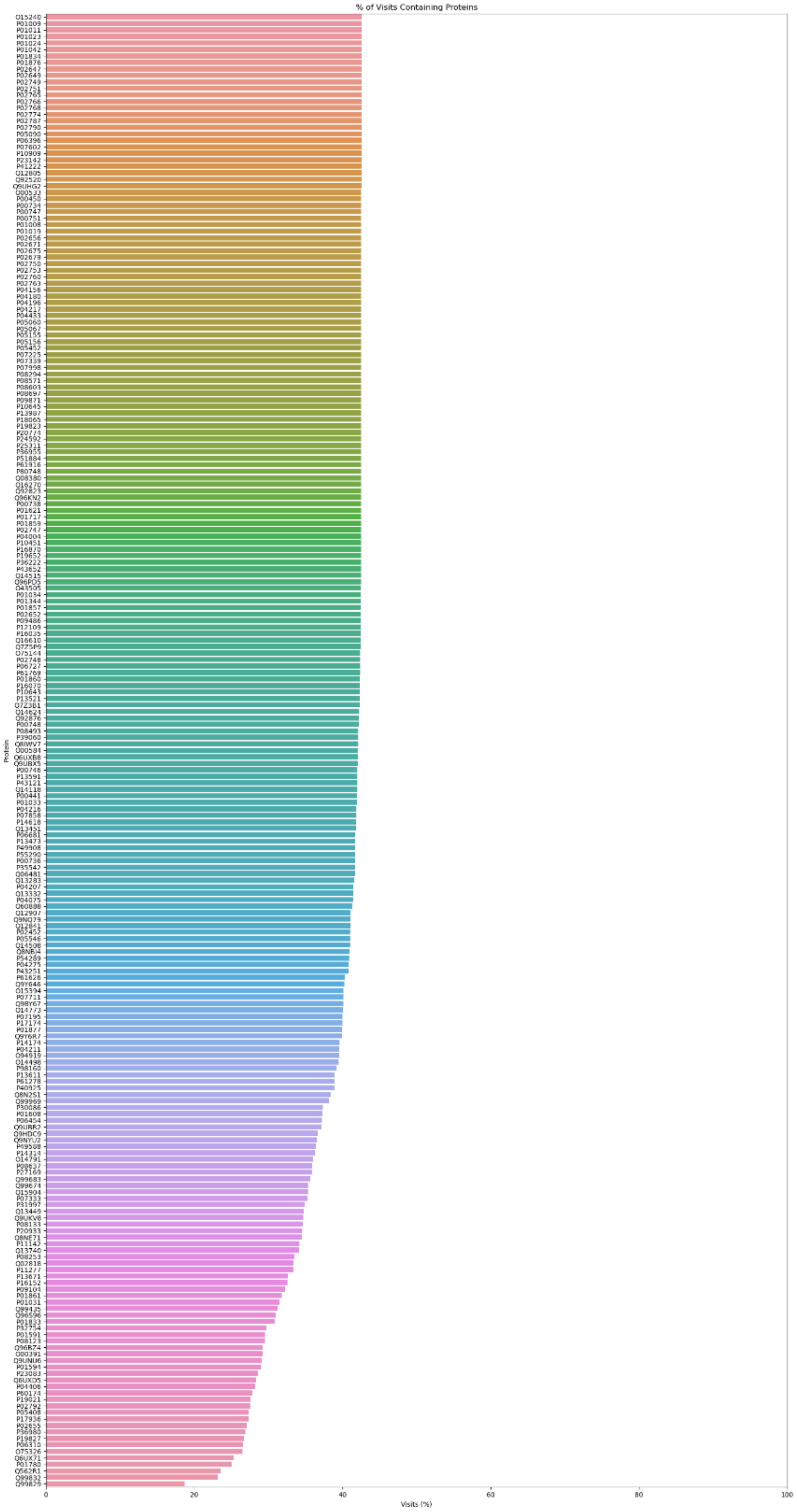
Zhao, J., Gu, S. and McDermaid, A. (2019). Predicting outcomes of chronic kidney disease from EMR data based on Random Forest Regression. *Mathematical Biosciences*, 310(0025-5564), pp.24–30. doi:https://doi.org/10.1016/j.mbs.2019.02.001.

8. Appendix

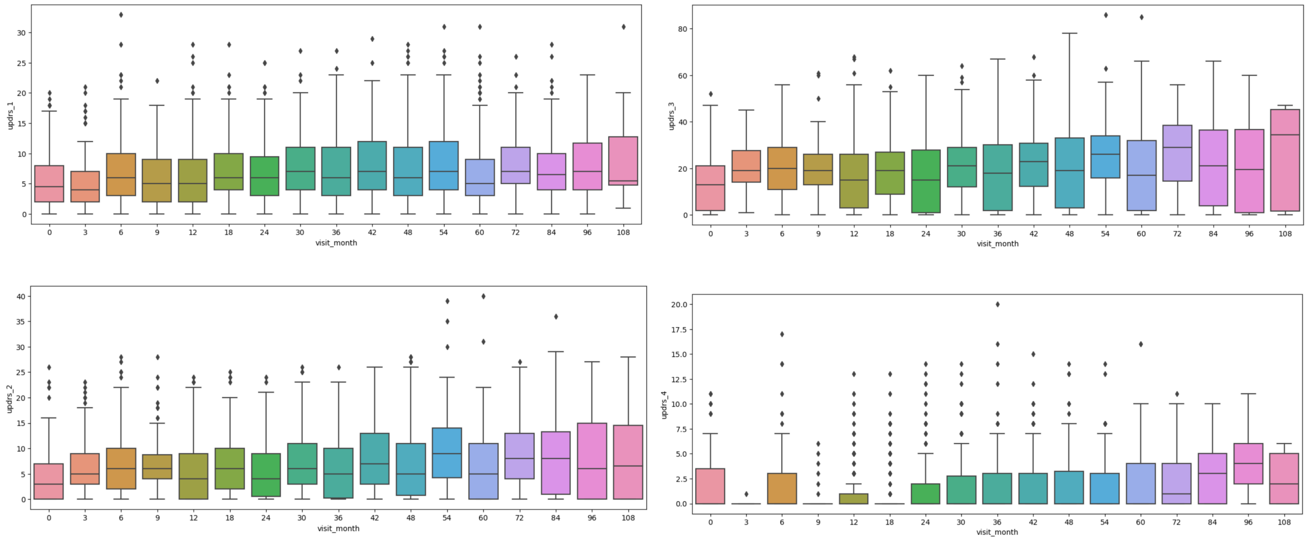
Appendix A: A table for terminology for each column of each dataset used

Column	Definition
[multiple_datasets] visit_id	ID code to distinguish visits, key for all datasets.
[multiple_datasets] visit_month	The relative month from the first visit taken by the patient (subtracted from the first month).
[multiple_datasets] patient_id	ID code to distinguish patients.
[train_clinical & supplementary] UPDRS_[1-4]	The patient's score for part N of the Movement Disorder Society revised Unified Parkinson's Disease Rating Scale (MDS-UPDRS).
[train_clinical & supplementary] upd23b_clinical_state_on_medication	Whether or not the patient was taking medication during the UPDRS assessment. These medications such as Levodopa (a dopamine replacement agent to treat Parkinson's) wear off quickly so patients commonly take the UPDRS_3 questionnaire (which medications mainly affect) twice a month, one with and one without medication.
[train_proteins] UniProt	The UniProt ID code for the associated protein - think of it as a name for a particular protein. This column can be pivoted so that each protein can be treated as a feature.
[train_proteins] NPX	Stands for the Normalised Protein Expression and measures the frequency of the protein's occurrence in the sample. protein abundance values derived from mass spectrometry readings of cerebrospinal fluid (CSF) samples taken from patients.
[train_peptides] Peptide	The sequence of amino acids included in the peptide. There can be multiple peptides per protein.
[train_peptides] PeptideAbundance	The frequency of the amino acid in the sample.

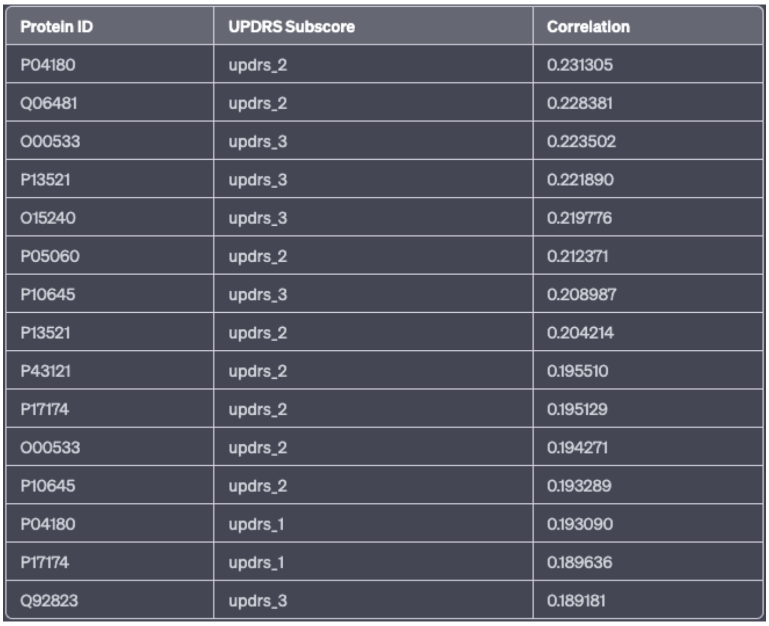
Appendix B: Barplot of Protein Appearance Frequency in Visits



Appendix C: Box plots of visit_month with updrs[1-4] scores.

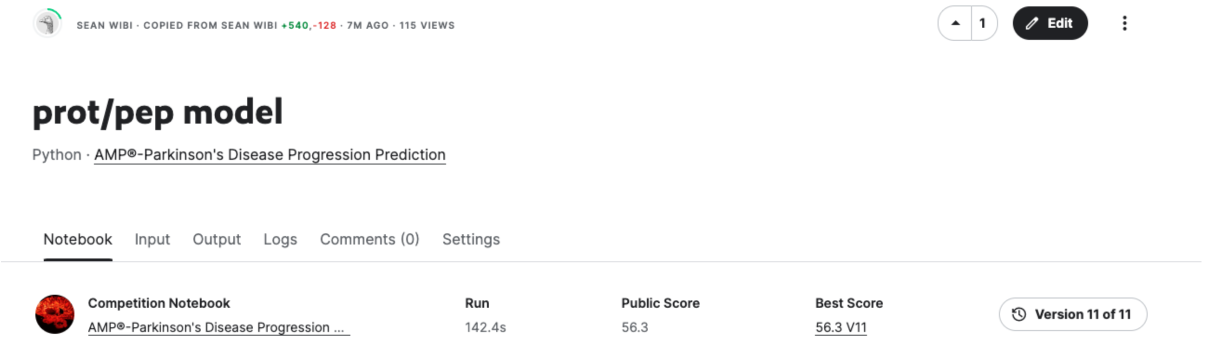


Appendix D: Proteins with the highest correlation to UPDRS scores.



Protein ID	UPDRS Subscore	Correlation
P04180	updrs_2	0.231305
Q06481	updrs_2	0.228381
O00533	updrs_3	0.223502
P13521	updrs_3	0.221890
O15240	updrs_3	0.219776
P05060	updrs_2	0.212371
P10645	updrs_3	0.208987
P13521	updrs_2	0.204214
P43121	updrs_2	0.195510
P17174	updrs_2	0.195129
O00533	updrs_2	0.194271
P10645	updrs_2	0.193289
P04180	updrs_1	0.193090
P17174	updrs_1	0.189636
Q92823	updrs_3	0.189181

Appendix E: Kaggle leaderboard score



SEAN WIBI · COPIED FROM SEAN WIBI +540,-128 · 7M AGO · 115 VIEWS

prot/pep model

Python · AMP®-Parkinson's Disease Progression Prediction

Competition Notebook	Run	Public Score	Best Score	
AMP®-Parkinson's Disease Progression ...	142.4s	56.3	56.3 V11	Version 11 of 11