Transcriptomics profiling of Parkinson's disease progression subtypes reveals distinctive patterns of gene expression

Journal of Central Nervous System Disease
Volume 17: 1–17

DOI: 10.1177/11795735241286821
Sage

Carlo Fabrizio[1], Andrea Termine[1] and Carlo Caltagirone[2]

[1]Data Science Unit, Santa Lucia Foundation IRCCS, Rome, Italy. [2]Department of Clinical and Behavioral Neurology, Santa Lucia Foundation IRCCS, Rome, Italy.

ABSTRACT

BACKGROUND: Parkinson's Disease (PD) varies widely among individuals, and Artificial Intelligence (AI) has recently helped to identify three disease progression subtypes. While their clinical features are already known, their gene expression profiles remain unexplored.

OBJECTIVES: The objectives of this study were (1) to describe the transcriptomics characteristics of three PD progression subtypes identified by AI, and (2) to evaluate if gene expression data can be used to predict disease subtype at baseline.

DESIGN: This is a retrospective longitudinal cohort study utilizing the Parkinson's Progression Markers Initiative (PPMI) database.

METHODS: Whole blood RNA-Sequencing data underwent differential gene expression analysis, followed by multiple pathway analyses. A Machine Learning (ML) classifier, namely XGBoost, was trained using data from multiple modalities, including gene expression values.

RESULTS: Our study identified differentially expressed genes (DEGs) that were uniquely associated with Parkinson's disease (PD) progression subtypes. Importantly, these DEGs had not been previously linked to PD. Gene-pathway analysis revealed both distinct and shared characteristics between the subtypes. Notably, two subtypes displayed opposite expression patterns for pathways involved in immune response alterations. In contrast, the third subtype exhibited a more unique profile characterized by increased expression of genes related to detoxification processes. All three subtypes showed a significant modulation of pathways related to the regulation of gene expression, metabolism, and cell signaling. ML revealed that the progression subtype with the worst prognosis can be predicted at baseline with 0.877 AUROC, yet the contribution of gene expression was marginal for the prediction of the subtypes.

CONCLUSION: This study provides novel information regarding the transcriptomics profiles of PD progression subtypes, which may foster precision medicine with relevant indications for a finer-grained diagnosis and prognosis.

KEYWORDS: Parkinson's disease, RNA-seq, precision medicine, subtyping

RECEIVED: February 16, 2024. **ACCEPTED:** August 22, 2024.

TYPE: Original Research Article

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Ricerca Corrente grants (Italian Ministry of Health) from the Santa Lucia Foundation IRCCS Linea di Ricerca 1 - Neuroscienze Cliniche e Neuroriabilitazione.

SUPPLEMENTAL MATERIAL Supplemental material for this article is available online.

CORRESPONDING AUTHOR: Data Science Unit, Santa Lucia Foundation IRCCS, Via Ardeatina, 306/354, Rome 00179, Italy.
Email: c.fabrizio@hsantalucia.it

Introduction

Parkinson's Disease (PD), the prevailing neurodegenerative movement disorder, is experiencing a faster rise in prevalence than other neurological disorders over the last years.[1,2] The primary pathological feature is the accumulation of misfolded, aggregated α-synuclein in the substantia nigra and other brain regions, which contributes to movement disorders like bradykinesia in combination with either rest tremor, rigidity, or both.[3,4]

PD is a heterogeneous condition, with variations in clinical features, symptoms, and rate of progression.[5] This variability has prompted a number of studies investigating the existence of PD subtypes. To this extent, PD is a well-suited model for precision medicine which, taking individual variability into account, emphasizes fine-grained diagnostics to enhance treatment effectivenes.[6] One of the challenges in PD research is to assign each affected individual to a specific disease cluster, in order to find phenotypic subgroups that may have a particularly good response to specific treatments.[3]

While the majority of research concerning data-driven clustering in PD has centred on disease subtyping using baseline cross-sectional data, mounting evidence suggests that PD has a highly heterogeneous progression.[7,8] Therefore, any static subtype defined at the baseline may not well account for disease progression patterns. Accordingly, PD subtypes instability is particularly pronounced in the early stages of the disease[9,10] and advanced PD patients exhibit many clinical similarities despite early-stage heterogeneity.[11,12] The

hypothesis of heterogeneous progression in PD found further support in a 2021 study, where a predictive model found that patients show non-sequential, overlapping disease progression trajectories over eight distinct disease states, finally suggesting that static subtype assignment might be ineffective at capturing the full spectrum of PD progression.[8]

Recently, α-synuclein Seed Amplification Assays (SAA) was identified as a promising biomarker for the biochemical diagnosis of PD,[13] yet this necessitates a cerebrospinal fluid (CSF) sample to be detected, which might not always be readily available. Conversely, peripheral blood is a more accessible sample type and can be subjected to molecular-level analysis, which could provide further details on biomarkers for a finer-grained diagnosis. The identification of disease subtypes in such a complex disease is pivotal to advance therapeutics,[14] and RNA-Seq allows for a broad scope view of the biochemical landscape of a specific phenotype.[15]

Research on PD blood transcriptomics is consistently highlighting the association of inflammatory pathways, oxidative stress, and mitochondrial processes with the disease,[15,16] also demonstrating that immune cell subtypes play a role in its transcriptomic changes.[17] Nonetheless, it was noted that RNA-Seq data is often ignored in Machine Learning (ML) studies of PD,[18] meaning that the potential of this data source remains to be fully exploited.

Efforts in PD progression subtyping research focus on detecting distinct classes of patients based on unique progression patterns, emphasizing the importance of incorporating time as a dimension. Artificial Intelligence (AI) algorithms play a crucial role in managing the complexity of time series data, enhancing result reliability, and enabling hypothesis-free experiments.

A pivotal study for PD subtyping employed clustering analysis at baseline and performed a longitudinal evaluation, but it was based on cross-sectional data analysis, thus overlooking the temporal dimension.[5] The most recent attempt in 2022 introduced an intriguing approach, combining NMF-reduced PD representations with Gaussian Mixture Model clustering; however, it lacked a clear temporal framework, resulting in non-overlapping clusters for patients at the latest time point.[19] Contrastingly, a 2019 study by Zhang et al harnessed a Long Short Term Memory (LSTM) model to identify three PD progression subtypes.[20] LSTM is an AI architecture specifically designed to handle sequential data, such as time series.[21] The analysis of comprehensive clinical and biological data resulted in the identification of three distinct subtypes: in brief, subtype I (S1) starts with mild motor and non-motor symptoms, and motor impairment increases with a moderate rate over time; subtype II (S2) has moderate motor and non-motor symptoms at baseline, with a slow progression rate; subtype III (S3) has significant motor and non-motor symptoms at baseline, and its impairment progresses rapidly over time, thus accounting for a worse prognosis.[20] An improved iteration of this approach, using an LSTM coupled with a Deep Progression Embedding (DPE) model, was shared as a preprint in 2021, aligning with earlier findings but awaiting peer-review.[22] Other authors developed their own algorithm for the identification of progression subtypes,[23] but the heterogeneity in the results dependent on the features selected for analysis, and unavailability of clustered subject IDs, made us prone to focusing on PD progression subtypes identified in[20].

The transcriptional profiles of PD subtypes with distinct progression (rapid vs slow) profiles were compared using classical statistical techniques and microarray technology, and more than 200 differentially expressed genes were found.[24] More recently, multivariate data analysis with AI techniques is allowing for the identification of data-driven subtypes of PD,[20] and although PD subtypes with distinctive progression phenotypes have been identified, their transcriptomics profiles remain unexplored. In fact, to the best of our knowledge, the transcriptomics characteristics of PD progression subtypes have never been taken into account in a multivariate analysis of longitudinal data using AI methods.

The present study has two main objectives: (1) to describe the transcriptomics profile of PD progression subtypes, and (2) to subsequently evaluate the usefulness of gene expression data in predicting disease subtype at baseline. The present paper aims to reveal the biological characteristics of disease progression subtypes. We expect to find partially distinct characteristics of gene expression, which should account for the separate identity of the disease subtypes. The identification of unique transcriptomic traits associated with the subtypes may foster precision medicine in PD, with relevant indications for a finer-grained diagnosis and prognosis. Finally, we make available comprehensive results tables and code scripts, fostering the formulation of hypotheses for further experiments on PD subtypes.

Methods

Workflow overview

Data from the PPMI database were used for both of the objectives of this study: (1) to identify the transcriptomics characteristics of the disease progression subtypes, and (2) to train the ML model aimed at evaluating the usefulness of gene expression data in predicting disease subtypes at baseline. First, data were gathered and the cohort of study was defined, as described in subsection 2.2. RNA-Seq data were preprocessed (subsection 2.3) and then a differential expression analysis was performed as described in subsection 2.4 The resulting DEGs were further analyzed through pathway analyses, as described in subsection 2.5 Following cohort definition, the ML classifier was trained to predict the cluster at baseline, as described in subsection 2.6, then its behaviour was investigated using XAI methods (subsection 2.7). The research workflow is schematized in Figure 1. The R code used to perform data analysis can be found on GitHub (https://github.com/217c/parkinson_subtypes_rnaseq).

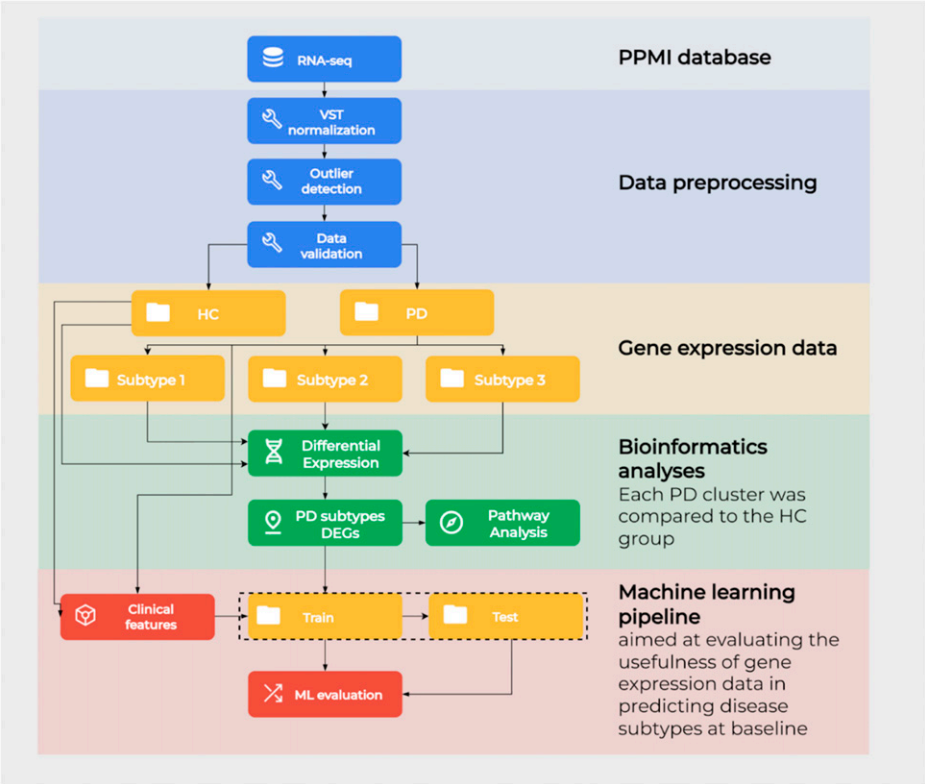

Figure 1. Schematic diagram outlining the analysis workflow.

Data

Data used in this study were obtained from the Parkinson Progression Marker Initiative (PPMI).[44] PPMI is one of the most important ongoing studies of PD progression markers, collecting data from multiple international sources and focusing on a diverse range of potential markers for tracking the progression of PD, including demographics, clinical variables, imaging data, cerebrospinal fluid, blood, DNA and, importantly to this study, RNA measures. The data were downloaded from the LONI Image and Data Archive (IDA) in April 2022. The cohort of study was defined using the PPMI Consensus Committee Analytic Dataset (RD: 2021-10-28). PPMI inclusion criteria for PD subjects include: age ≥30, Parkinson's disease diagnosis within the last 2 years, baseline Hoehn and Yahr Stage I–II, and no anticipated need for symptomatic treatment within 6 months of baseline.[25] Healthy controls (HC) inclusion criteria will include individuals without clinical signs suggestive of parkinsonism, no evidence of cognitive impairment, and no first-degree relative diagnosed with PD. To be included in this study cohort, subjects must have had a diagnosis of sporadic PD and available RNA-Seq data for multiple timepoints, as found in the LongRNA Transcriptome Sequencing of PPMI Samples (B38) study (RD: 2021-0402). The PPMI RNA Sequencing Project has generated overview transcriptomics data from raw sequencing reads of PPMI whole blood samples. The data were pre-analyzed and quality controlled from the PPMI group.[26]

The definition of the sample for this study follows that described in 20. Subjects that underwent disease progression subtyping were included, along with all available HC subjects. In brief, S1 starts with mild motor and non-motor symptoms, and motor impairment increases with a moderate rate over time; S2 has moderate motor and non-motor symptoms at baseline, with a slow progression rate; S3 has significant motor and non-motor deficits at baseline, and its impairment progresses rapidly over time, thus accounting for a worse prognosis. The IDs of the subjects assigned to disease progression subtypes were retrieved from.[20] To summarize, data analysis was performed on those subjects that had RNA-Seq data available and that were clustered into one PD subtype. This study cohort included a total number of 2085 RNA-Seq records for 4 years of longitudinal measures (starting from baseline, with constant time interval measures at 12 months) from 600 subjects (PD = 407, HC = 193) (S1 = 199; S2 = 52; S3 = 156), ready for the data preparation steps including outliers check and sex incompatibility check.

RNA-seq data preparation

To assess outliers, a Principal Component Analysis (PCA) was computed on variance stabilized and transformed (namely, vst from DESeq2) expression data of the top 20000 genes, and data points lying beyond the edges of the Highest Density Interval of the first principal component were deemed as outliers. The threshold was set to 0.99, thus considering as outliers all

observations outside the 99% Confidence Interval (CI).[27] A sex incompatibility check was performed to assess contamination due to abnormal transcription using t-SNE and DBSCAN on gene expression data from the following sex chromosome genes: *USP9Y*, *XIST*, *RPS4Y1*, *RPS4Y2*, *KDM5D*, *DDX3Y*. Subjects whose samples had inconsistent clustering between sex in metadata and sex from expression data were removed from the analysis (Supplemental Figure 1).

Differential expression analysis

Differentially expressed genes (DEGs) were identified using `DESeq2` R library v1.38.3 to perform a Likelihood ratio test (LRT). This experiment was designed as a time course analysis, thus the full model including group, time, and their interaction, was compared to a reduced model without the interaction. This analysis allowed us to identify those genes that at one or more time points after time zero showed a group-specific effect, thus excluding genes that moved up or down in time in the same way in both groups. Each PD cluster was compared to the HC group performing a separated LRT. For each comparison, `DESeq2` automatically estimated size factors based on the median ratio method, estimated dispersions, and performed the LRT for negative binomial GLMs.[28] Correction for multiple testing was performed using the False Discovery Rate (FDR) method, applying `DESeq2`'s default threshold for adjusted *P*-value <0.1.

Gene names and descriptions were retrieved using `g:Profiler` R package.[29]

Pathway analysis

To further investigate the differences in gene expression we performed a pathway analysis using `clusterProfiler` R library v4.6.2.[30] An Over-Representation Analysis (ORA) was performed on DEGs for all three comparisons on GO Biological Process (BP) domain, KEGG, and WikiPathways databases. Not to limit our pathway analysis to DEGs sets, we chose to investigate pathway modulation due to eventually small but coordinated changes in the expression levels of all genes, thus performing a Gene Set Enrichment Analysis (GSEA) for all three comparisons on GO BP, KEGG, and WikiPathways databases. To improve interpretability, GSEA results on GO were reduced to semantically similar terms using `rrvgo` R library v1.10.0.[31]

Machine learning model for subtype prediction at baseline

Data collected at the time of diagnosis (baseline) was used to predict the cluster, using a hierarchical machine learning approach. In this approach, we train multiple classifiers in a hierarchical structure, where each classifier is responsible for a specific task. This approach is useful here because the classification task can be broken down into simpler sub-tasks. As cluster 3 showed to be the most severe, the first step was to predict if the newly diagnosed PD subject belonged to S3. If not, the second step aimed to predict whether the subject was from S1 or S2 (Figure 2). The node near the root holded a classifier that was designed to discriminate between the most severe (S3) and the least severe (S1/2) phenotypes. The node at the leaf was then designed to distinguish between the least severe phenotypes (S1 vs S2).

The hierarchical structure graph reported in Figure 2 represents the ideal workflow to follow in potential clinical use, and it is not intended to represent dependence between the models. The models were trained separately and tested independently, in order to use all of the available data and aiming to evaluate the usefulness of gene expression data in discriminating between the most and least severe phenotypes (S3 vs S1/2), and then in discriminating between the two least severe phenotypes (S1 vs S2). Ideally, a subject is evaluated by the first model in the hierarchy, which aims to identify subjects from S3. If a subject

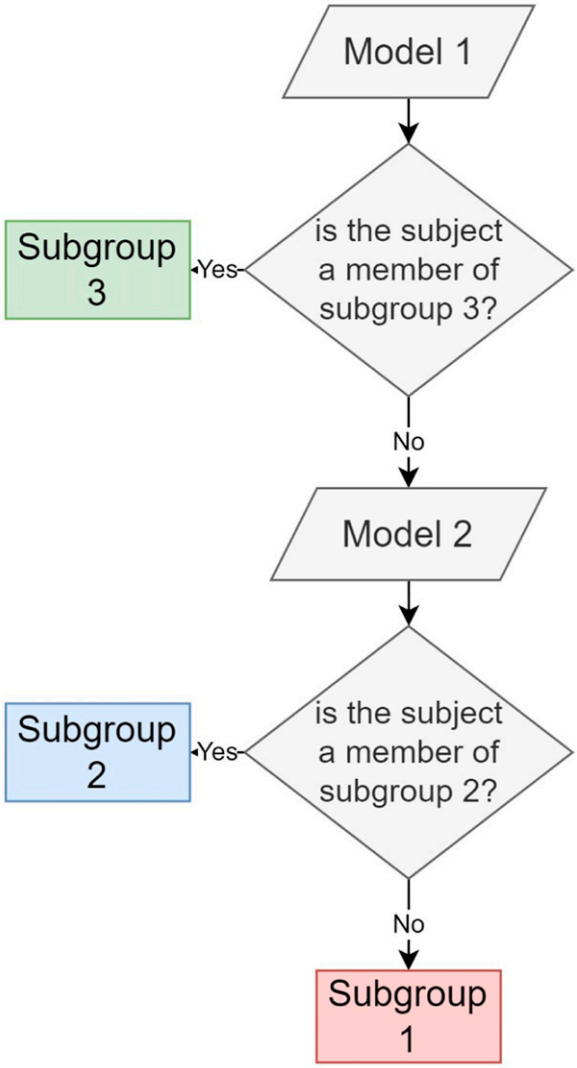

Figure 2. Schematic representation of the flow of the Hierarchical ML approach.

would be classified as S3, the pipeline would end. Otherwise, the subject would be evaluated by the second model, which aims to discriminate between S2 and S1.

Two XGBoost models were used in this pipeline.[32] The machine learning pipeline was developed using `tidymodels` R library v1.0.0. Train test split was performed at subject level, including 75% of the sample in the train set (Table 1). Data from multiple modalities were used, including demographics, motor, non-motor, biospecimen, imaging, and gene expression values (Table 2). Missing data were imputed with the mean value of the train set and rounded to integer value, thus respecting the original format of variables. All variables were transformed by applying a Box-Cox transformation[33] and feature selection was performed by univariate filtering with ANOVA on all three groups. Variables reporting an FDR-corrected *P*-value <0.05 were selected for training. Variables with an absolute Pearson's correlation value greater than 0.8 with other variables were removed. Synthetic minority oversampling technique (SMOTE) was used to address class imbalance before training.[34] The XGBoost models were trained using 10 Cross-Validation resamples to find the best combination of hyperparameters using a grid latin hypercube of values.[35] The best models resulting from cross-validation were tested on the test set and evaluation metrics were computed.

Variable importance and XAI

The importance of variables in contributing to the Machine Learning prediction of subtype at baseline was investigated using SHAP (SHapley Additive exPlanations) values.[36] As an XAI method,[37] SHAP values highlight the contribution of each feature to the final prediction, thus providing a measure to rank features importance. To calculate SHAP values and produce informative plots, `shapviz` R library functions[38] were applied to the XGBoost models.

Table 1. Number of observations in train and test splits.

SPLIT	SUBTYPE	N
Train	S1	141
Train	S2	33
Train	S3	108
Test	S1	47
Test	S2	12
Test	S3	37

Results

The data preparation process focused on determining which subjects included in the present study (thus meeting the inclusion criterion of having available RNA data) had been clustered into a disease progression subtypes by.[20] Out of the initial 466 PD subjects with assigned subtypes (S1 = 201; S2 = 107; S3 = 158), a total of 407 PD subjects had RNA-Seq data available (S1 = 199; S2 = 52; S3 = 156), and were included in downstream analyses. Outliers' detection identified 19 records as outliers, and nine subjects showed sex inconsistencies (Supplemental Figure 1). After their removal, the final dataset comprised 2057 samples from 598 participants.

Finally, 58,780 genes were available for the analysis.

Differentially expressed genes (DEGs)

Differential expression analysis was conducted to assess changes in gene expression attributable to the progression of the disease over a span of 4 years, thereby incorporating longitudinal measurements for a time course experiment analysis. In particular, each one of the three subtypes was compared to the HC group.

60 DEGs were found for S1 (41 up, 19 down), 34 for S2 (15 up, 19 down), and 32 for S3 (27 up, 5 down). The most part of these DEGs were distinctive of the subtypes, with just six of these DEGs found as shared between two or more subtypes (Figure 3). A list of DEGs with gene names and descriptions, along with the complete results tables from the differential expression analysis, can be found in Supplemental Table 1.

Over representation analysis (ORA)

In order to understand the biological pathways associated with the DEGs, ORA was performed on Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), and WikiPathways databases.

Almost all of the resulting pathway terms were distinctive of the subtypes, with very few terms in common among them (Figure 4).

Overall, significant pathways for S1 concurrently indicated a modulation of cellular energy metabolism, pointing to mitochondrial dysfunction, along with gene expression regulation, and stress response pathways. Moreover, the presence of pathways like *Parkinson's disease* (hsa05012, q-value: 1.85×10^{-5}) and *Alzheimer disease* (hsa05010, q-value: 2.83×10^{-3}) highlighted a significant involvement of S1 DEGs in neurological diseases and neurodegeneration.

The significantly enriched pathways for S2 mainly pointed to regulation of gene expression and modulation of metabolic processes, including several terms associated with RNA metabolism and processing.

Table 2. Full list of variables used for machine learning.

VARIABLE NAME	EXTENDED NAME	DESCRIPTION
AGE_AT_VISIT	Age	Age at the time of visit
REMSLEEP_tot	REM sleep behavior disorder questionnaire	Final score
SCOPAAUT_tot	Scales for outcomes in Parkinson's disease - autonomic dysfunction (SCOPA-AUT)	Final score
JLO_TOTRAW	Benton judgement of line orientation	Line orientation-sum 15 item
DVT_TOTAL_RECALL	Hopkins verbal learning test - revised	Derived-total recall T-score
DVT_DELAYED_RECALL	Hopkins verbal learning test - revised	Derived-delayed recall T-score
DVT_RECOG_DISC_INDEX	Hopkins verbal learning test - revised	Derived-recog. Discrim. Index T-Score
LNS_TOTRAW	Letter - number sequencing	LNS-sum questions 1-7
SDMTOTAL	Symbol digit modalities test	Symbol digit modalities total correct
VLTANIM	Modified semantic fluency	Total number of animals
VLTVEG	Modified semantic fluency	Total number of fruits
VLTFRUIT	Modified semantic fluency	Total number of vegetable
NP2PTOT	MDS-UPDRS	MDS-UPDRS part II total score
NP1PTOT	MDS-UPDRS	MDS-UPDRS part I (patient questionnaire) total score
NP3TOT	MDS-UPDRS	MDS-UPDRS part III total score
DATSCAN_CAUDATE_R	DATSCAN imaging	Striatal binding ratio of the right caudate small brain region of interest referenced to the occipital lobe
DATSCAN_PUTAMEN_R	DATSCAN imaging	Striatal binding ratio of the right putamen small brain region of interest referenced to the occipital lobe
ENSG00000144290.16	SLC4A10	Solute carrier family 4 member 10
ENSG00000248350.1	None	Heat shock factor binding protein 1 (HSBP1) pseudogene
ENSG00000057657.16	PRDM1	PR/SET domain 1
ENSG00000211713.3	TRBV6-4	T Cell receptor beta variable 6-4
ENSG00000212219.1	RNU6-604P	RNA, U6 small nuclear 604, pseudogene
ENSG00000197275.13	RAD54 B	RAD54 homolog B
ENSG00000239148.1	U8	U8 small nucleolar RNA
ENSG00000261553.5	None	Novel transcript
ENSG00000275968.1	None	None
ENSG00000258494.1	OR11J5P	Olfactory receptor family 11 subfamily J member 5 pseudogene
ENSG00000275992.1	RN7SL327P	RNA, 7SL, cytoplasmic 327, pseudogene
ENSG00000171649.11	ZIK1	Zinc finger protein interacting with K protein 1
ENSG00000152454.3	ZNF256	Zinc finger protein 256
ENSG00000199567.1	Y_RNA	Y RNA

The ORA analysis for S3 primarily identified pathways related to oxidative stress response and detoxification, along with response to reactive oxygen species and hydrogen peroxide metabolism. Another main theme in S3 pathways from ORA was the modulation of cellular signaling and metabolism, including apoptosis.

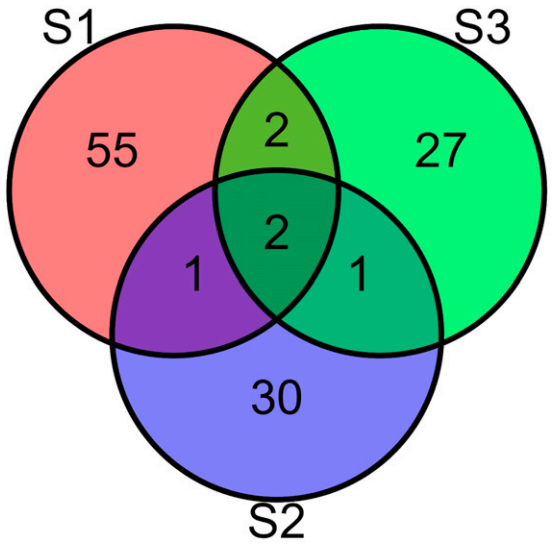

Figure 3. Venn diagram of DEGs for each subtype.

For an extensive presentation of these results see Supplementary Results. The full list of pathways from the ORA can be found in Supplemental Table 2.

Gene Set Enrichment Analysis (GSEA)

The examination of the overall gene expression pattern was carried out through GSEA. This analysis is not limited to the set of DEGs, and it takes into account the variations in gene expression across all genes.

This analysis indicated the enrichment of thousands of pathways for GO, and tens of pathways for KEGG and WikiPathways. Most GO pathways were shared between S1 and S2, while far less were shared with S3 (Figure 5A). KEGG and WikiPathways terms were instead mostly distinctive of the subtypes, with few shared pathways between them (Figure 5A, and B).

The enriched pathways from GSEA on all databases for S1 highlighted key themes related to organismal processes like protein synthesis and energy metabolism, along with neuronal signaling and nervous system development. The results emphasized the modulation of homeostatic processes and metabolic dysregulation, likely related to the involvement of oxidative phosphorylation in energy production. Moreover, few disease pathways were found significant and, accordingly, pathways related to the immune system were also present. Additionally,

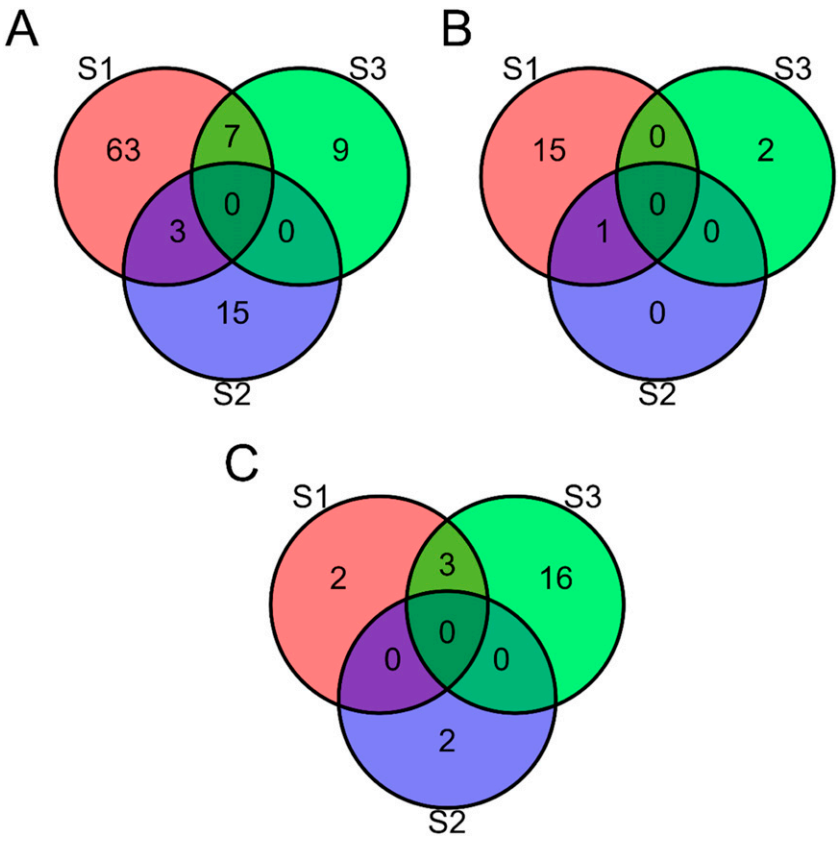

Figure 4. Venn diagram of GO (A), KEGG (B), and WikiPathways (C) terms for each subtype from the ORA.

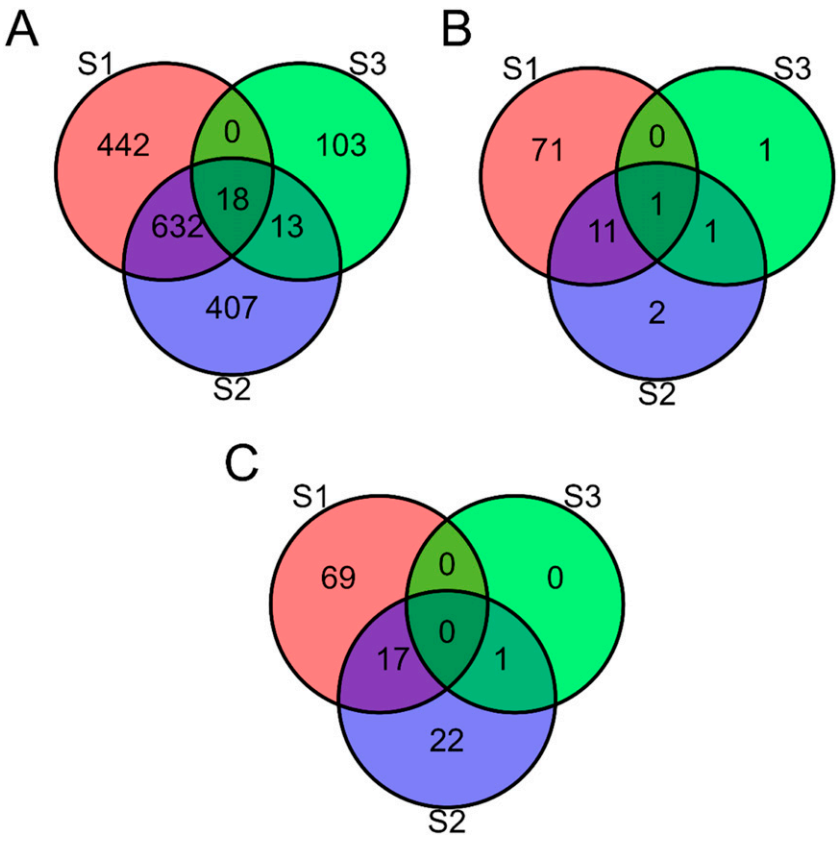

Figure 5. Venn diagram of GO (A), KEGG (B), and WikiPathways (C) terms for each subtype from the GSEA.

pathways associated with processes related to addiction were identified in S1.

Pathways from the GSEA for S2 revealed biological pathways associated with organismal processes and cellular signaling, along with structures development such as cell development and connectivity. Other terms revealed the modulation of the immune response and the involvement of disease processes which, among many others, included pathways associated with genetic disorders and syndromes. In the results for S2, pathways related to addiction processes were identified, similar to those found in S1.

GO pathways for S1 and S2 were mostly shared and related to morphological changes (*nervous system development*, *anatomical structure development*, *anatomical structure morphogenesis*, *tissue development*). Interestingly, all pathways from S1 and S2 showed opposite enrichment scores, indicating that these two groups were characterized by an opposite expression pattern despite sharing most of their enriched pathways (Figure 6).

GO pathways from the GSEA on S3 data revealed only unique pathways, with none shared with other subtypes. Key themes included sensory perception, signal transduction, cell signaling, and regulation of gene expression. Significant pathways involved the positive regulation of olfactory transduction, neuroactive ligand-receptor interaction, and protein export. Additionally, the Interactome of polycomb repressive complex 2 (PRC2) pathway highlighted modulation in gene expression and chromatin organization.

For an extensive presentation of these results see Supplementary Results. The complete results tables can be found in Supplemental Table 2.

Baseline prediction of disease progression subtype

A Machine Learning hierarchical classification approach was implemented to develop a prediction system aimed at identifying the disease subtypes of a newly-diagnosed PD patient, at the baseline. Data from multiple modalities were used, including demographics, motor, nonmotor, biospecimen, imaging (Table 2). The first model in the hierarchy aimed to predict whether the subject was from S3, which has the most distinctive phenotype and is also the most severe. The model achieved a fair performance with 0.814 sensitivity, and 0.757 specificity, yielding an F-Score of 0.828 and a total AUROC of 0.877 (Figure 7).

Variable importance was investigated with the application of an explainable Artificial Intelligence (XAI) method, namely SHAP values. These highlighted the score to MDS-UPDRS Part II (disability evaluation) as the most important factor contributing to S3 identification. Among the most important variables there are other clinical measures, along with a

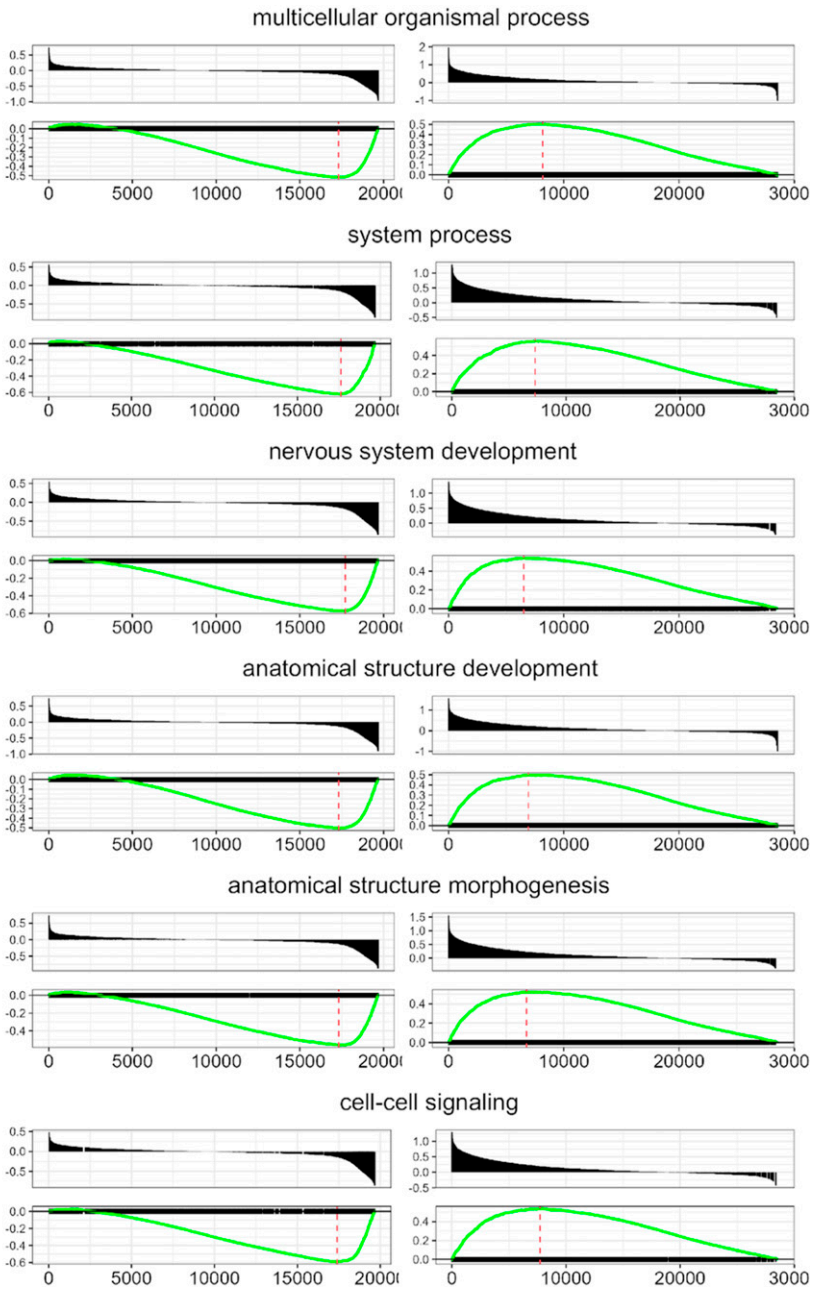

Figure 6. Visual representation of six distinct pathways from the enrichment analysis, each labelled with its respective name as a section title. Within each section, there are two sets of plots: S1 on the left and S2 on the right. The upper plots illustrate the positions of gene set members on a rank-ordered list, with the *x*-axis indicating position and the *y*-axis representing the ranked list metric. The lower plots display the enrichment scores, with a dashed line indicating the maximum rank of the enrichment score. It is clear to see that all of the represented pathways show opposite enrichment profiles between S1 and S2.

neuroimaging measure (DaTScan Caudate R). Gene expression only had a marginal importance, with low absolute SHAP values, giving little contribution to the final prediction (Figure 8).

The second level of the hierarchy held a model aiming to predict whether the subject was from S1 or S2. It achieved a poor performance, with 0.745 sensitivity, 0.25 specificity, yielding a F-Score of 0.77 and a total AUROC of 0.576 (Figure 9).

SHAP values indicated that expression values of *U8*, *HSBP1*, *TRBV6-4*, *and SCL4A10*, along with Benton Judgement of Line Orientation test score, were the most important factors to discriminate between S1 and S2 (Figure 10).

Discussion

The identification of progression subtypes is of extreme importance in order to attempt settling the heterogeneity of PD.

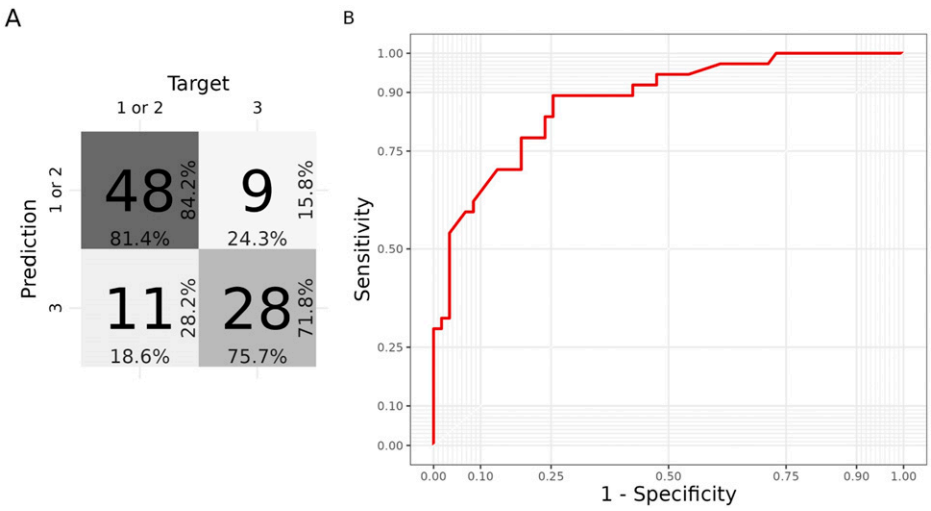

Figure 7. ROC curve and confusion matrix from the first model of the hierarchy.

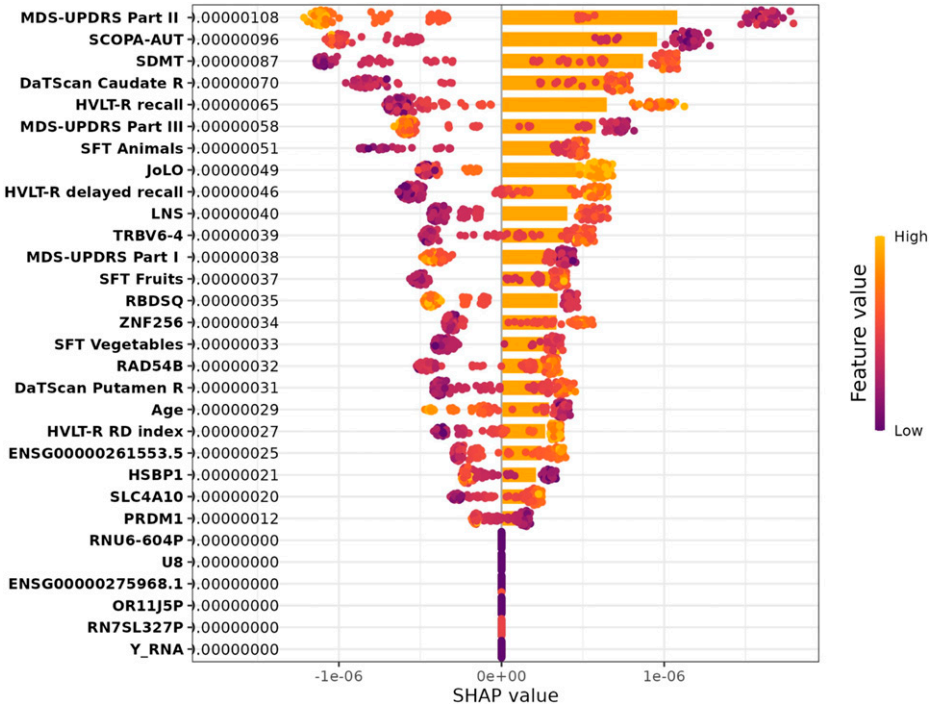

Figure 8. SHAP summary plot representing the contribution of each variable to the prediction of the model.

Recent research has shown that people with PD can exhibit a variety of progression patterns from diagnosis onwards.[5,8,19,20,22,39] The identification of disease modifying treatments can be fostered by finer-grained diagnoses and biomarkers identification, pursuing a precision medicine approach. Targeting specific biological processes is currentlyunfeasible due to the lack of validated nonclinical biomarkers of PD progression,[40] thus the importance of describing the biological profiles of progression subtypes is a paramount objective.

In this study we investigated the transcriptomic profile of three disease progression subtypes, which were identified in 20 with an AI algorithm that reliably takes into account time as a dimension. Briefly, S1 had mild motor and non-motor symptoms at baseline, with a moderate rate of motor impairment increase and relatively stable cognitive abilities; S2 had moderate motor and non-motor symptoms at baseline, with a slow progression rate; and S3 started with significant motor and non-motor symptoms, showing a rapid progression of impairment, and thus reporting the worse prognosis among the three.[20]

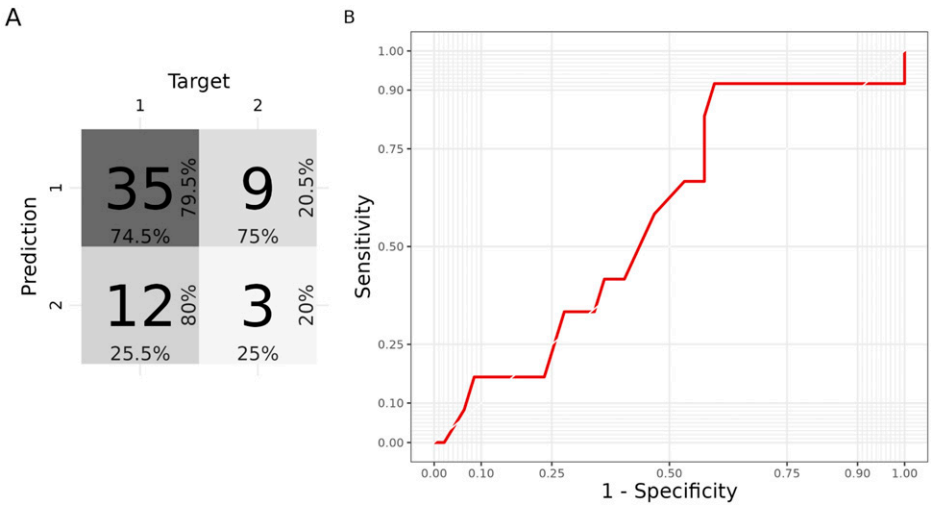

Figure 9. ROC curve and confusion matrix from the second model of the hierarchy.

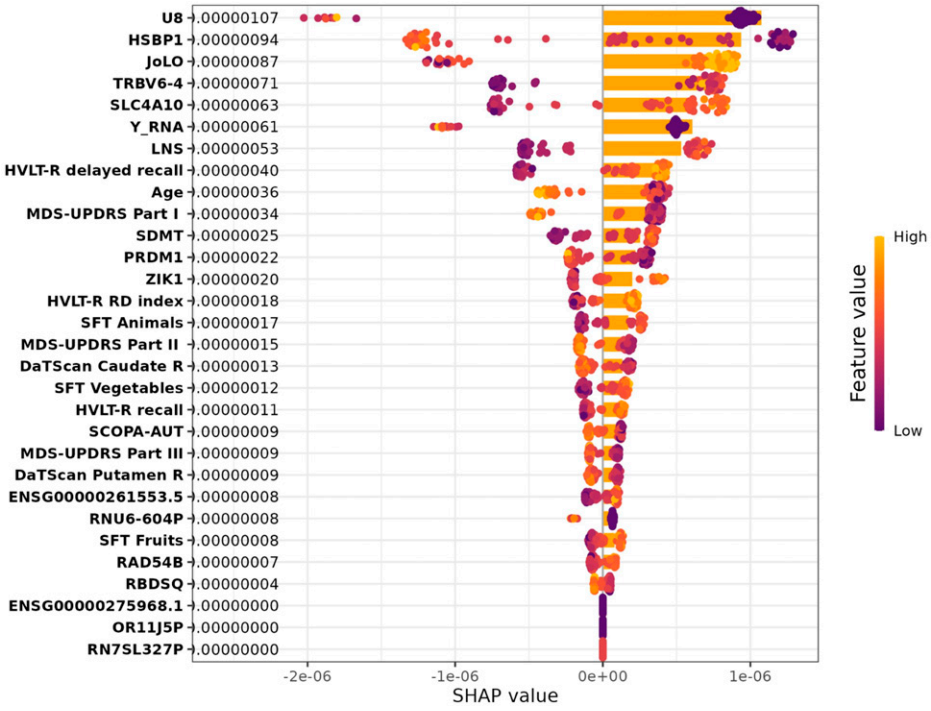

Figure 10. SHAP summary plot representing the contribution of each variable to the prediction of the model.

The DEGs identified in this study are unique to these progression subtypes, as none of the genes that are commonly found as differentially expressed in PD studies are present in our results. As a specific example, common transcriptomic markers such as *SYN1*, *ANKRD22*, and *SLC14A1*[16] are absent from all our DEGs lists. This result is not surprising to us, as our experiment had two main differences with other PD RNA studies. First, although based on transcriptomics of PD subjects, we investigated progression subtypes as diagnostic classes, thus differences with a classical PD group were expected. Second, our differential expression analysis was performed as a time course experiment, in order to identify those genes that varied for expression values as a result of the disease over time. This profoundly differs with previous PD transcriptomics studies, which performed a cross-sectional analysis of gene expression, thus not taking time into account. As a further note, there is general poor consensus between previous studies on resulting DEGs from PD studies.[15]

To provide a better understanding of the significance of the DEGs associated with the PD progression subtypes, and their impact on biological pathways, we summarized the results from the pathway analyses in (Tables 3–5). To better understand subtype-specific mechanisms due to gene expression alterations, we believe that there is a need for further in-depth studies of a

Table 3. Main indications extracted from the pathway analysis of the gene expression patterns for S1. The analysis determining pathway involvement indicates whether the enrichment is from DEGs (ORA) or the overall gene expression pattern (GSEA).

SUBTYPE	PATHWAYS RELATED TO	INDICATION DERIVED FROM	ALSO FOUND IN
S1	Cellular energy metabolism and mitochondrial dysfunction	ORA	-
S1	Neurological diseases and neurodegeneration	ORA	-
S1	Stress response	ORA	S3
S1	Regulation of gene expression and protein synthesis	ORA	S2, S3
S1	Cellular (neuronal) signaling and transduction	GSEA	S2, S3
S1	Nervous system and cell development	GSEA	S2
S1	Immune system	GSEA	S2
S1	Addiction	GSEA	S2

Table 4. Main indications extracted from the pathway analysis of the gene expression patterns for S2. The analysis determining pathway involvement indicates whether the enrichment is from DEGs (ORA) or the overall gene expression pattern (GSEA).

SUBTYPE	PATHWAYS RELATED TO	INDICATION DERIVED FROM	ALSO FOUND IN
S2	Regulation of gene expression and protein synthesis	ORA	S1, S3
S2	Metabolic processes	ORA	S3
S2	Disease processes (genetic disorders)	GSEA	-
S2	Cellular (neuronal) signaling and transduction	GSEA	S1, S3
S2	Nervous system and cell development	GSEA	S1
S2	Immune system	GSEA	S1
S2	Addiction	GSEA	S1

Table 5. Main indications extracted from the pathway analysis of the gene expression patterns for S3. The analysis determining pathway involvement indicates whether the enrichment is from DEGs (ORA) or the overall gene expression pattern (GSEA).

SUBTYPE	PATHWAYS RELATED TO	INDICATION DERIVED FROM	ALSO FOUND IN
S3	Cellular metabolism	ORA	-
S3	Cellular (neuronal) signaling and transduction	ORA	S1, S2
S3	Stress response	ORA	S1
S3	Regulation of gene expression and protein synthesis	GSEA	S1, S2
S3	Response to misfolded proteins	GSEA	-
S3	Apoptosis	GSEA	-

purely biological approach. Our results may serve as useful knowledge for the generation of hypotheses to test in such studies.

S1 transcriptomics profile

The transcriptomic profile of S1 was characterized by a significant modulation of cellular energy metabolism. In particular, we identified alterations in oxidative phosphorylation, aerobic respiration, and cellular respiration pathways. Moreover, pathways related to ATP synthesis, mitochondrial dysfunction, and nucleotide metabolism were commonly enriched across the ORA and GSEA over GO, KEGG and WikiPathways databases. The modulation of energy metabolism is well known in PD, and it has already been found from transcriptomics analyses both in blood and brain sample tissues.[41–43] Similarly, cellular

response to stress pathways, including oxidant detoxification and response to reactive oxygen species, were also consistently identified. These results confirm that metabolic alterations are a common background in neurological diseases and neurodegeneration, with our DEGs significantly enriching pathways associated with Parkinson's disease, Alzheimer's disease, Huntington disease, prion disease, and amyotrophic lateral sclerosis.

While all pathway analyses of S1 data similarities revealed common key themes, ORA and GSEA also identified unique pathways specific to each of the three databases. For instance, one analysis emphasized the significance of pathways related to ribosomal proteins in protein synthesis, while another highlighted the importance of neuronal signaling pathways and immune system dysregulation. Disease-related pathways such as nonalcoholic fatty liver disease and hepatitis B infection were specifically enriched in one analysis. The involvement of immune system response and processes related to oxidative stress are known in PD transcriptomics,[15,17] and the observation of disease pathways enrichment is related to their modulation.

The biological profile of S1 shares similarities with that of PD patients with *LRRK2* mutation, which is involved in multiple biological functions, including mitochondrial activity and oxidative pathways.[44] It is interesting to note that none of the patients included in this study had a mutation in one of the risk loci known for PD, as this study was solely focused on idiopathic PD. Nonetheless, it has already been observed that the patients with idiopathic PD or *LRRK2* genetic PD show mostly overlapping phenotypes, and they are clinically difficult to distinguish.[45]

Cellular signaling pathways were also found enriched in the GSEA, confirming that signaling mechanisms, often found among transcriptomics alterations from PD *post mortem* brain tissues,[46] can also emerge from the analysis of peripheral tissues, such as blood.[47,48]

A summarizing overview of S1 characteristics derived from the pathway analyses can be found in Table 3.

S2 transcriptomics profile

Pathway analyses consistently identified modulation of gene expression regulation and metabolic processes. Specifically, pathways associated with RNA metabolism and processing emerged among the most significant terms across all analyses. The implication of RNA metabolic processes has been considered in the pathogenesis and disease course of PD, advancing that these may be related to energy conservation, aggregated proteins modulation, and response to cellular stress.[49]

One notable characteristic of S2 pathway analyses results lies in the number of pathways identified in each analysis, as some analyses revealed a limited number of pathways. As an example, there were only two significant pathways in the ORA on WikiPathways: Endoderm differentiation (WP2853) and Mesodermal commitment pathway (WP2857). These were enriched by a single gene, *NCAPG2*. This gene encodes for a regulatory subunit of the condensin II complex which, along with the condensin I complex, plays a role in chromosome assembly and segregation during mitosis.[50] Alterations of this gene have been associated with cancer and neurodevelopmental defects,[51,52] and although its presence has already been observed in PD blood transcriptomics,[53,54] its role in the disease is still unclear.

Cell-cell communication was found modulated in the GSEA results on all databases. Relatedly, various pathways related to stimulus response emerged as modulated, indicating their involvement in this phenotype.

In GSEA results on the KEGG database, S2 exhibited pathways associated with addiction processes, sharing this characteristic with S1. Pathways related to morphine addiction also emerged in a recent evaluation of PD proteome from dopaminergic neurons in the substantia nigra (SN), suggesting an involvement of potentially compromised GABA-related pathways.[55]

A summarizing overview of S2 characteristics derived from the pathway analyses can be found in Table 4.

S3 transcriptomics profile

This subtype had significantly fewer shared terms with the other two, which in turn showed a much higher level of similarity. Pathways resulting from the ORA on DEGs indicate the involvement of response to oxidative stress and detoxification processes, aligning with findings in S1. Additionally, ORA on the KEGG database highlighted pathways related to diseases such as African trypanosomiasis and Malaria, implying a possible modulation of detoxification processes within this phenotype. Overall, these resulting pathways indicate a modulation of the processes associated with cellular adaptation and defense against oxidative stress and toxic substances. Cellular signaling was also found modulated in many of the results sets, and this is a shared alteration for all three subtypes. Accordingly, S3 results included sensory perception and signal transduction as prominent themes, with pathways related to the detection of chemical stimuli and smell perception. Also, the enrichment of pathways related to olfactory transduction, neuroactive ligand-receptor interaction, and protein export was observed. Furthermore, pathways associated with gene expression regulation and cellular response to misfolded proteins were significant, as also found in the other subtypes.

Metabolic pathways such as Vitamin B12 metabolism, Folate metabolism, and Selenium micronutrient network, were also found altered in this subtype. Recent studies have shown that B12 deficiency is common in patients with neuropathies, and PD has B12 levels decline over the course of the disease.[56]

A summarizing overview of S3 characteristics derived from the pathway analyses can be found in Table 5.

Comparison of the transcriptomics profiles among subtypes

The results of our transcriptomics analysis revealed a number of similarities between the three PD subtypes (S1, S2, and S3). All three subtypes showed a significant modulation of pathways related to the regulation of gene expression, metabolism, and cell signaling. Pathways associated with nervous system dysregulation were consistently found in all three subtypes. Although expected when analyzing brain cells, we believe that when resulting in blood it's a confirmatory result of appropriate transcriptomics findings, and this is also in line with previous works on peripheral tissues.[15,57] We may consider this as a general alteration due to the disease state, as these were also found in other PD transcriptomics experiments,[16] and not distinctive of any of the subtypes.

S1 and S2 had a few shared themes, including addiction pathways, structure development, immune response alterations and disease processes. In fact, among the distinctive characteristics for S1 we find neurological and neurodegenerative disease pathways. Moreover, S1 was unique in its alteration of energy production and mitochondrial functions. Interestingly, all of the shared pathways between S1 and S2 had opposite enrichment patterns in the GSEA (Figure 7). This demonstrates that S1 and S2 are distinct progression forms of the same disease. Despite sharing a few transcriptomic characteristics, these appear to be modulated in opposing ways, and thus may be at the foundation of their different progression courses. S2 showed an alteration in olfactory transduction, as also observed in S3. S3 was unique in its increased expression of genes involved in detoxification processes, and pathways related to cellular stress response were altered in both S1 and S3. Interestingly, this was the only subtype characterized by enrichment of response to misfolded proteins. Despite the diverse methodologies employed in PD subtyping research, our findings align with those of other authors who have compared transcriptomics profiles of PD subtypes. Specifically, nucleic acids metabolic processes, immune response, mitochondria, and cell metabolism emerged as the most significantly modulated pathways.[24] Additionally, a robust correlation was identified between gene expression changes in PD patients and cell models.[24] This observation suggests that peripheral tissues may serve as valuable indicators of critical disease-related mechanisms occurring in the brain, despite the inherent limitations associated with dissociated cell models.

Subtype prediction at baseline

The machine learning classifier provided a reliable tool to predict disease progression subtypes using baseline data. This tool could easily be implemented into a user-friendly software, to finally build a reliable Computer-Aided Diagnosis (CAD) tool to identify subjects with the most severe prognosis. As resulting from the variable importance analysis, the contribution of gene expression was marginal for the prediction of S3, not allowing for substantial discrimination between disease subtypes in neither of the steps of the hierarchical ML approach. Clinical variables instead demonstrated high importance to identify S3 subjects, with perceived disability (MDS-UPDRS Part II) being the most important predictor for a more severe prognosis. In fact, S3 subjects were characterized by a faster progression and worse symptomatology, sharing some similarities with the classical Posture Instability/Gait Difficulty (PIGD) subtype. Interestingly, most of the S3 subjects were PIGD patients, and those that were Tremor Dominant (TD) instead were likely to shift to PIGD over 6 years.[20] Although expression values resulted as the most important factors to discriminate between S1 and S2, the model at the second level of the hierarchy had a poor test performance. This made it unreliable and, as a consequence, the evaluation of its behavior is meaningless. Considering that this hierarchical classification model has 0.877 AUROC to detect the most severe subtype, this would give useful indication for prognosis. As such, this ML model may foster precision medicine for PD, providing support for a finer-grained diagnosis by applying the results of subtyping research. As all PD subjects included in this study were newly diagnosed, and the classifier was trained and tested on baseline data, it could be applied in clinical practice when evaluating a new PD patient. Additionally, we would like to highlight that the model was trained on baseline data to predict a class defined by disease progression, which involves the passage of time. Notably, it has a greater ability to predict a subject's future compared to traditional PIGD/TD subtyping. This prediction holds particular relevance for individuals whose phenotype aligns with the S3 subtype, where this classification is more prone to change over time.

In the replication study of the PD progression subtype identification, it has been found that the most severe subtype (S3) had distinctive clinical features when compared to the two less severe subtypes (S1 and S2). Moreover, it was observed that there was limited signal in baseline variables to discriminate between the less severe subtypes.[22] These observations are in line with our results, as the performance of our classifier is poor in discriminating between S1 and S2 (0.576 AUROC). Additionally, our analysis revealed that not even transcriptomics assessment was useful to discriminate between S1 and S2 at baseline.

The classifier trained in this study aimed to assess the usefulness of gene expression data to subtype prediction. It has shown that it may be best to use as few variables as possible to predict the subtype, and appropriate ML evaluation on these data would provide indications about the best-suited variables for this purpose. Nonetheless, this study was solely aimed at testing the usefulness of RNA-Seq data. Further research may aim at finding the minimum set of clinical variables that maximize prediction reliability.

As previously acknowledged, there is currently no standardized method for distinguishing PD progression subtypes.

However, identifying peripheral indicators of disease progression could greatly benefit PD clinicians. Despite the prior identification of DEGs between two PD progression subtypes, there has been a notable absence of attempts to leverage this data for classification using ML techniques.[24] Our research supports the idea that peripheral predictors for PD progression subtypes are not to be found in blood transcriptomics. Instead, we propose that baseline clinical data can be effectively utilized to enhance prognosis.

Providing a tool for progression subtype prediction at baseline is pivotal to improve the application of subtyping research results into PD clinical practice. Not only this study provides a biological characterization of progression subtypes, but it also demonstrates that a hierarchical ML approach is suitable to detect the most severe subtype, with a potentially relevant impact on prognosis.

Strengths and limitations

This study provides a characterization of the transcriptomics profile for three PD subtypes identified in a data-driven manner, namely using AI to analyze the disease progression. A data-driven approach to disease subtyping is free from the biases due to the experimenter and is more precise, as no a priori choices based on medical expertise are made. PPMI has one of the largest PD cohorts to date, offering a consistently large group to identify disease subtypes with AI methods. As the identification of disease progression subtypes was performed using an LSTM,[20] the present study is hypothesis-free and aims to characterize the most reliable PD progression subtypes available in the literature.

This study utilized whole blood RNA-Sequencing data for transcriptomic analysis. While this approach has its advantages in terms of accessibility and potential clinical relevance, it's essential to note that blood transcriptomics may not fully capture the complexities of gene expression alterations specific to the brain, which is primarily affected in PD. Further studies on brain cells transcriptomics are needed to fully unravel the characteristics of PD progression subtypes.

Moreover, albeit PPMI is a multi-site data collection, this study is limited by the analysis of a single cohort. The extrapolation of these findings to clinical practice would need robust validation supported by further replication studies on multiple independent cohorts, eventually strengthened by the integration of other clinical and biological markers.

The vastness of the results tables from the pathway analyses hindered results manageability. As a group of researchers, we did our best to read the results table and report noteworthy results, yet it is to be disclosed that a complete and accurate report was unfeasible. As a comment to this, we would like to speculate that future technological development may help with the interpretation of High Throughput Sequencing data analysis results: Large Language Models (LLM), such as ChatGPT,[58] are showing increasingly better ability to handle textual data, and may one day be well-suited to summarize and expose these kinds of results. Potential future analysis of our results by means of such methods is encouraged, and full results tables can be found in Supplemental Tables 1-2.

Acknowledgements

PPMI – a public-private partnership – is funded by the Michael J. Fox Foundation for Parkinson's Research and funding partners, including 4D Pharma, Abbvie, AcureX, Allergan, Amathus Therapeutics, Aligning Science Across Parkinson's, AskBio, Avid Radiopharmaceuticals, BIAL, Biogen, Biohaven, BioLegend, BlueRock Therapeutics, Bristol-Myers Squibb, Calico Labs, Celgene, Cerevel Therapeutics, Coave Therapeutics, DaCapo Brainscience, Denali, Edmond J. Safra Foundation, Eli Lilly, Gain Therapeutics, GE HealthCare, Genentech, GSK, Golub Capital, Handl Therapeutics, Insitro, Janssen Neuroscience, Lundbeck, Merck, Meso Scale Discovery, Mission Therapeutics, Neurocrine Biosciences, Pfizer, Piramal, Prevail Therapeutics, Roche, Sanofi, Servier, Sun Pharma Advanced Research Company, Takeda, Teva, UCB, Vanqua Bio, Verily, Voyager Therapeutics, the Weston Family Foundation and Yumanity Therapeutics. For up-to-date information on the study, visit.

Author contributions

See CRediT data.

Ethical Statement

Ethics approval

The current institutional review board (IRB) for the Michael J. Fox Foundation's (MJFF) ongoing Parkinson's Progression Markers Initiative (PPMI) study is WCG IRB since 2022. The study is registered at ClinicalTrials.gov with IDs: NCT04477785, NCT01141023. The Parkinson's Progression Markers Initiative (PPMI) study was approved by the Institutional Review Boards of respective institutions, which includes more than 50 institutes around the world (the full list is available at the following link https://www.ppmi-info.org/about-ppmi/ppmi-clinical-sites). The authors confirm all relevant ethical guidelines have been followed, and any necessary ethics committee approvals have been obtained. PPMI sites received approval from an ethical standards committee on human experimentation before study initiation and obtained written informed consent for research from all participants in the study.

Informed Consent

Written informed consent were obtained from each participant at enrollment, in accordance with the Declaration of Helsinki. All methods were performed in accordance with the relevant guidelines and regulations.

All necessary patient/participant consent has been obtained and the appropriate institutional forms have been archived, and that patient/participant/sample identifiers included were not known to anyone (e.g., hospital staff, patients or participants themselves) outside the research group so cannot be used to identify individuals.

ORCID iDs

Carlo Fabrizio https://orcid.org/0000-0002-7824-8423
Andrea Termine https://orcid.org/0000-0003-4374-7430
Carlo Caltagirone https://orcid.org/0000-0002-0189-4515

Data Availability Statement

Data used in the preparation of this article were obtained from the Parkinson's Progression Markers Initiative (PPMI) database, RRID:SCR_006431. The PPMI IDs of the subjects in the disease subtypes were obtained from the GitHub repository related to 20.

REFERENCES

1. Ray Dorsey E. "Global, regional, and national burden of Parkinson's disease, 1990– 2016: a systematic analysis for the Global Burden of Disease Study 2016". English. *Lancet Neurol*. 2018;17(11):939-953. doi:10.1016/S1474-4422(18)30295-3
2. WHO. *Parkinson disease: a public health approach: technical brief. en*. Geneva: WHO, 2022.
3. Bloem BR, Okun MS, Klein C. Parkinson's disease. *Lancet*. 2021;397:2284-2303. doi:10.1016/S0140-6736(21)00218-X
4. Postuma RB, Berg D, Stern M, et al. MDS clinical diagnostic criteria for Parkinson's disease. *Mov Disord*. 2015;30(12):1591-1601. doi:10.1002/mds.26424.
5. Fereshtehnejad S, Zeighami Y, Dagher A, Postuma RB. Clinical criteria for subtyping Parkinson's disease: biomarkers and longitudinal progression. *Brain*. 2017;140(7):1959-1976. doi:10.1093/brain/awx118
6. Riggare S, Hägglund M. Precision medicine in Parkinson's disease - exploring patient-initiated self-tracking. *J Parkinsons Dis*. 2018;8(3):441-446. doi:10.3233/JPD-181314
7. Greenland JC, Williams-Gray CH, Barker RA. The clinical heterogeneity of Parkinson's disease and its therapeutic implications. *Eur J Neurosci*. 2019;49(3):328-338. doi:10.1111/ejn.14094
8. Severson KA, Chahine LM, Smolensky LA, et al. Discovery of Parkinson's disease states and disease progression modelling: a longitudinal data study using machine learning. *Lancet Digit Health*. 2021;3(9):e555-e564.
9. Erro R, Picillo M, Amboni M, et al. Comparing postural instability and gait disorder and akinetic-rigid subtyping of Parkinson disease and their stability over time. *Eur J Neurol*. 2019;26(9):1212-1218. doi:10.1111/ene.13968
10. Simuni T, Caspell-Garcia C, Coffey C, et al. "How stable are Parkinson's disease subtypes in de novo patients: Analysis of the PPMI cohort?" en. *Parkinsonism Relat Disorders*. 2016;28:62-67. doi:10.1016/j.parkreldis.2016.04.027
11. De Pablo-Fernández E, J Lees A, L Holton J, T Warner T "Prognosis and Neuropathologic Correlation of Clinical Subtypes of Parkinson Disease". en. *JAMA Neurol*. 2019;76(4):470. doi:10.1001/jamaneurol.2018.4377
12. Kempster PA, O'Sullivan SS, Holton JL, Revesz T, Lees AJ. Relationships between age and late progression of Parkinson's disease: a clinico-pathological study. *Brain*. 2010;133(6):1755-1762. doi:10.1093/brain/awq059
13. Siderowf A, Concha-Marambio L, Lafontant DE, et al. Assessment of heterogeneity among participants in the Parkinson's Progression Markers Initiative cohort using α-synuclein seed amplification: a cross-sectional study. *Lancet Neurol*. 2023;22(5):407-417. doi:10.1016/S1474-4422(23)00109-6
14. Fabrizio C, Termine A, Caltagirone C, Sancesario G. "Artificial Intelligence for Alzheimer's Disease: Promise or Challenge?" en. *Diagnostics*. 2021;11:1473. doi:10.3390/diagnostics11081473
15. Borrageiro G, Haylett W, Seedat S, Kuivaniemi H, Bardien S. A review of genome-wide transcriptomics studies in Parkinson's disease. *Eur J Neurosci*. 2018;47(1):1-16. doi:10.1111/ejn.13760
16. Lee S, Park SM, Yeo SS, et al. Parkinson's disease subtyping using clinical features and biomarkers: literature review and preliminary study of subtype clustering. *Diagnostics*. 2022; 12: 112. doi: 10.3390/diagnostics12010112
17. Craig DW, Hutchins E, Violich I, et al. RNA sequencing of whole blood reveals early alterations in immune cells and gene expression in Parkinson's disease. *Nat Aging*. 2021;1:734-747.
18. Gerraty RT, Provost A, Li L, Wagner E, Haas M, Lancashire L. Machine learning within the Parkinson's progression markers initiative: review of the current state of affairs. *Front Aging Neurosci*. 2023;15:1076657.
19. Dadu A, Satone V, Kaur R, et al. Identification and prediction of Parkinson's disease subtypes and progression using machine learning in two cohorts. *NPJ Parkinsons Dis*. 2022;8:172. doi:10.1038/s41531-022-00439-z
20. Zhang X, Chou J, Liang J, et al. Data-driven subtyping of Parkinson's disease using longitudinal clinical records: a cohort study. *Sci Rep*. 2019;9:797. doi:10.1038/s41598-018-37545-z
21. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput*. 1997;9(8):1735-1780.
22. Su C, Hou Y, Brendel M, Henchcliffe C Integrative analyses of multimodal clinical, neuroimaging, genetic, and transcriptomic data identify subtypes and potential treatments for heterogeneous Parkinson's disease progression. en. preprint. *Neurology*. 2021;2023:764–773. doi:10.1101/2021.07.18.21260731
23. Krishnagopal S, Coelln R, Shulman LM, Girvan M. Identifying and predicting Parkinson's disease subtypes through trajectory clustering via bipartite networks. *PLoS One*. 2020;15(6):e0233296. doi:10.1371/journal.pone.0233296
24. Pinho R, Guedes LC, Soreq L, et al. Gene expression differences in peripheral blood of Parkinson's disease patients with distinct progression profiles. In: *PloS One*. San Francisco, CA USA: Public Library of Science; 2016.
25. Marek K. "The Parkinson Progression Marker Initiative (PPMI)". en. *Prog Neurobiol*. 2011;95(4):629-635. doi:10.1016/j.pneurobio.2011.09.005
26. Hutchins E, Craig D, Violich I, Alsop E, , et al. Quality Control Metrics for Whole Blood Transcriptome Analysis in the Parkinson's Progression Markers Initiative (PPMI) en. preprint. *Genetic and Genomic Medicine*. 2021. doi:10.1101/2021.01.05.21249278
27. Gelman A, Greenland S. "Are confidence intervals better termed "uncertainty intervals"?" en. *BMJ*. 2019;366:l5381. doi:10.1136/bmj.l5381
28. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15(12):550. doi:10.1186/s13059-014-0550-8
29. Raudvere U, Kolberg L, Kuzmin I, et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res*. 2019;47:W191–W198. doi:10.1093/nar/gkz369
30. Wu T, Hu E, Xu S, et al. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innovation*. 2021;2(3):100141. doi:10.1016/j.xinn.2021.100141
31. Sayols S. "rrvgo: a Bioconductor package for interpreting lists of Gene Ontology terms". eng. *Micropublication Biology 2023*. 2023;11:1240039.
32. Chen T, Guestrin C. *XGBoost: A Scalable Tree Boosting System*. Ithaca: arXiv Version Number: 3; 2016.
33. George EP, Cox D. An analysis of transformations. *J Roy Stat Soc B Stat Methodol*. 1964;26(2):211-243.
34. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res*. 2002;16:321-357. doi:10.1613/jair.953
35. Dupuy D, Helbert C, Franco J. "DiceDesign and DiceEval : Two *R* Packages for Design and Analysis of Computer Experiments". en. *J Stat Software*. 2015;65:11.
36. Lundberg SM, Su-In Lee. "A unified approach to interpreting model predictions". In: *Advances in Neural Information Processing Systems*. Red Hook: Curran Associates, Inc. Ed. by Guyon I, et al. Vol. 30. , 2017.
37. Gunning D, Stefik M, Choi J, Miller T, Stumpf S, Yang GZ. XAI-Explainable artificial intelligence. *Sci Robot*. 2019;4(37):eaay7120. doi:10.1126/scirobotics.aay7120
38. Scott M, Su-In L. A unified approach to interpreting model predictions. *Advances in neural information processing systems*. 2017;30:4765-4774.
39. Fereshtehnejad S, Romenets SR, Anang JBM, Latreille V, Gagnon JF, Postuma RB. New clinical subtypes of Parkinson disease and their longitudinal progression: a prospective cohort comparison with other phenotypes. *JAMA Neurol*. 2015;72(8):863-873. doi:10.1001/jamaneurol.2015.0703
40. Mestre TA, Fereshtehnejad SM, Berg D, et al. Parkinson's disease subtypes: critical appraisal and recommendations. *J Parkinsons Dis*. 2021;11(2):395-404. doi:10.3233/JPD-202472
41. Elstner M, Morris CM, Heim K, et al. Expression analysis of dopaminergic neurons in Parkinson's disease and aging links transcriptional dysregulation of energy metabolism to cell death. *Acta Neuropathol*. 2011;122(1):75-86. doi:10.1007/s00401-011-0828-9
42. Shamir R, Klein C, Amar D, et al. Analysis of blood-based gene expression in idiopathic Parkinson disease. *Neurology*. 2017;89(16):1676-1683. doi:10.1212/WNL.0000000000004516

43. Zhang Y, James M, Middleton FA, Davis RL "Transcriptional analysis of multiple brain regions in Parkinson's disease supports the involvement of specific protein processing, energy metabolism, and signaling pathways, and suggests novel disease mechanisms". en. *In: American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*. 2005;137:5-16. doi:10.1002/ajmg.b.30195
44. Berwick DC, Heaton GR, Azeggagh S, Harvey K. LRRK2 Biology from structure to dysfunction: research progresses, but the themes remain the same. *Mol Neurodegener*. 2019;14(1):49. doi:10.1186/s13024-019-0344-2
45. Marras C, Schüle B, Schuele B, et al. Phenotype in parkinsonian and non-parkinsonian LRRK2 G2019S mutation carriers. *Neurology*. 2011;77(4):325-333. doi:10.1212/WNL.0b013e318227042d
46. Huang M, Xu L, Liu J, Huang P, Tan Y, Chen S. Cell–cell communication alterations via intercellular signaling pathways in substantia nigra of Parkinson's disease. *Front Aging Neurosci*. 2022;14:828457. doi:10.3389/fnagi.2022.828457
47. Irmady K, Hale CR, Qadri R, et al. Blood transcriptomic signatures associated with molecular changes in the brain and clinical outcomes in Parkinson's disease. *Nat Commun*. 2023;14(1):3956. doi:10.1038/s41467-023-39652-6
48. Kurvits L, Lättekivi F, Reimann E, et al. Transcriptomic profiles in Parkinson's disease"en. *Exp Biol Med*. 2021;246(5):584-595. doi:10.1177/1535370220967325
49. Lu B, Gehrke S, Wu Z. RNA metabolism in the pathogenesis of Parkinson's disease. *Brain Res*. 2014;1584:105-115. doi:10.1016/j.brainres.2014.03.003
50. Ono T, Losada A, Hirano M, Myers MP, Neuwald AF, Hirano T. Differential contributions of condensin I and condensin II to mitotic chromosome architecture in vertebrate cells. *Cell*. 2003;115(1):109-121. doi:10.1016/s0092-8674(03)00724-4
51. Khan TN, Khan K, Sadeghpour A, et al. Mutations in NCAPG2 cause a severe neurodevelopmental syndrome that expands the phenotypic spectrum of condensinopathies. *Am J Hum Genet*. 2019;104(1):94-111. doi:10.1016/j.ajhg.2018.11.017
52. Wang Q, Li Z, Zhou S, et al. NCAPG2 could be an immunological and prognostic biomarker: from pan-cancer analysis to pancreatic cancer validation. *Front Immunol*. 2023;14:1097403. doi:10.3389/fimmu.2023.1097403
53. Infan-te J, Prieto C, Sierra M, et al. Identification of candidate genes for Parkinson's disease through blood transcriptome analysis in LRRK2-G2019S carriers, idiopathic cases, and controls. *Neurobiol Aging*. 2015;36(2):1105-1109. doi:10.1016/j.neurobiolaging.2014.10.039
54. Pantaleo E, Monaco A, Amoroso N, et al. A machine learning approach to Parkinson's disease blood transcriptomics. *In: Genes*. 2022;13(5):727. doi:10.3390/genes13050727
55. Jang Y, Pletnikova O, Troncoso JC, et al. Mass spectrometry-based proteomics analysis of human substantia nigra from Parkinson's disease patients identifies multiple pathways potentially involved in the disease. *Mol Cell Proteomics*. 2023;22(1):100452. doi:10.1016/j.mcpro.2022.100452
56. Christine CW, Auinger P, Saleh N, et al. Relationship of cerebrospinal fluid Vitamin B12 status markers with Parkinson's disease progression. *Mov Disord*. 2020;35(8):1466-1471. doi:10.1002/mds.28073
57. Liu S, Zhang Y, Bian H, Li X. Gene expression profiling predicts pathways and genes associated with Parkinson's disease. *Neurol Sci*. 2016;37(1):73-79. doi:10.1007/s10072-015-2360-5
58. OpenAI. *Introducing ChatGPT. En-US*. California: OpenAI. 2022.