

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

Carlo Fabrizio  *¹, Andrea Termine  ¹, and Carlo Caltagirone  ²

¹Data Science Unit, Santa Lucia Foundation IRCCS, 00179 Rome, Italy

²Department of Clinical and Behavioral Neurology, Santa Lucia Foundation IRCCS, 00179 Rome, Italy

Abstract

Parkinson's Disease (PD) exhibits significant individual variability, and recent Artificial Intelligence advancements have identified three distinct progression subtypes, each with known clinical features but unexplored gene expression profiles. This study aimed to identify the transcriptomics characteristics of PD progression subtypes, and assess the utility of gene expression data in subtype prediction at baseline. Differentially expressed genes were subtype-specific, and not typically found in other PD studies. Pathway analysis showed distinct and shared features among subtypes. Two had opposing expression patterns for shared pathways, and the third had a more unique profile with respect to the others. Machine Learning 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. This study offers insights into PD subtypes transcriptomics, fostering precision medicine for improved diagnosis and prognosis.

Keywords: Parkinson's disease; RNA-Seq; precision medicine; subtyping

1 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

*c.fabrizio@hsantalucia.it

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

25 pathological feature is the accumulation of misfolded, aggregated α -synuclein in the substantia
26 nigra and other brain regions, which contributes to movement disorders like bradykinesia in
27 combination with either rest tremor, rigidity, or both [3, 4].

28 PD is a remarkably variable condition, characterized by a wide heterogeneity at individual level,
29 with variations in clinical features, dominant symptoms, and rate of progression [5]. This vari-
30 ability has prompted a number of studies investigating the existence of PD subtypes. To this
31 extent, PD is a well-suited model for precision medicine which, taking individual variability into
32 account, emphasizes fine-grained diagnostics to enhance treatment effectiveness [6]. One of
33 the challenges in PD research is to assign each affected individual to a specific disease cluster,
34 in order to find phenotypic subgroups that may have a particularly good response to specific
35 treatments [3].

36 While the majority of research concerning data-driven clustering in PD has centred on disease
37 subtyping using baseline cross-sectional data, mounting evidence suggests that PD has a highly
38 heterogeneous progression [7, 8]. Therefore, any static subtype defined at the baseline may not
39 well account for disease progression patterns. Accordingly, PD subtypes instability is partic-
40 ularly pronounced in the early stages of the disease [9, 10] and advanced PD patients exhibit
41 many clinical similarities despite early-stage heterogeneity [11, 12]. The hypothesis of heteroge-
42 neous progression in PD found further support in a 2021 study, where a predictive model found
43 that patients show non-sequential, overlapping disease progression trajectories over eight dis-
44 tinct disease states, finally suggesting that static subtype assignment might be ineffective at
45 capturing the full spectrum of PD progression [8].

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

53 Research on PD blood transcriptomics is consistently highlighting the association of inflam-
54 matory pathways, oxidative stress, and mitochondrial processes with the disease [15, 16], also
55 demonstrating that immune cell subtypes play a role in its transcriptomic changes [17]. Nonethe-

56 less, it was noted that RNA-Seq data is often ignored in Machine Learning studies of PD [18],
57 meaning that the potential of this data source remains to be fully exploited.

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

62 A pivotal study for PD subtyping employed clustering analysis at baseline and performed a lon-
63 gitudinal evaluation, but it was based on cross-sectional data analysis, thus overlooking the
64 temporal dimension [5]. The most recent attempt in 2022 introduced an intriguing approach,
65 combining NMF-reduced PD representations with Gaussian Mixture Model clustering; however,
66 it lacked a clear temporal framework, resulting in non-overlapping clusters for patients at the
67 latest time point [19]. Contrastingly, a 2019 study by Zhang et al. harnessed a Long Short Term
68 Memory (LSTM) model to identify three PD progression subtypes [20]. LSTM is an AI archi-
69 tecture specifically designed to handle sequential data, such as time series [21]. The analysis
70 of comprehensive clinical and biological data resulted in the identification of three distinct
71 subtypes: in brief, subtype I (S1) starts with mild motor and non-motor symptoms, and motor
72 impairment increases with a moderate rate over time; subtype II (S2) has moderate motor and
73 non-motor symptoms at baseline, with a slow progression rate; subtype III (S3) has significant
74 motor and non-motor symptoms at baseline, and its impairment progresses rapidly over time,
75 thus accounting for a worse prognosis [20]. An improved iteration of this approach, using an
76 LSTM coupled with a Deep Progression Embedding (DPE) model, was shared as a preprint in
77 2021, aligning with earlier findings but awaiting peer-review [22]. Other authors developed their
78 own algorithm for the identification of progression subtypes [23], but the heterogeneity in the
79 results dependent on the features selected for analysis, and unavailability of clustered subject
80 IDs, made us prone to focusing on PD progression subtypes identified in [20]. Not only the
81 latter is ongoing research, still needing peer-revision for its latest update [22], but has open
82 access to the full analysis code and tables through [GitHub](#).

83 **1.1 Aims**

84 To the best of our knowledge, RNA-Sequencing data has never been taken into account in PD
85 progression subtyping research. Although PD subtypes with distinctive progression phenotypes

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

96 **2 Results**

97 The data preparation process focused on determining which subjects included in the present
98 study (thus meeting the inclusion criterion of having available RNA data) had been clustered
99 into a disease progression subtypes by [20]. Out of the initial 466 PD subjects with assigned
100 subtypes (S1 = 201; S2 = 107; S3 = 158), a total of 407 PD subjects had RNA-Seq data available
101 (S1 = 199; S2 = 52; S3 = 156), and were included in downstream analyses. Outliers' detection
102 identified 19 records as outliers, and nine subjects showed sex inconsistencies (Supplementary
103 Figure 1). After their removal, the final dataset comprised 2,057 samples from 598 participants.
104 Finally, 58,780 genes were available for the analysis.

105 **2.1 Differentially Expressed Genes (DEGs)**

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

110 60 DEGs were found for S1 (41 up, 19 down), 34 for S2 (15 up, 19 down), and 32 for S3 (27 up, 5
111 down). The most part of these DEGs were distinctive of the subtypes, with just six of these DEGs
112 found as shared between two or more subtypes (Figure 1). A list of DEGs with gene names and
113 descriptions, along with the complete results tables from the differential expression analysis,
114 can be found in Supplementary Table 1.

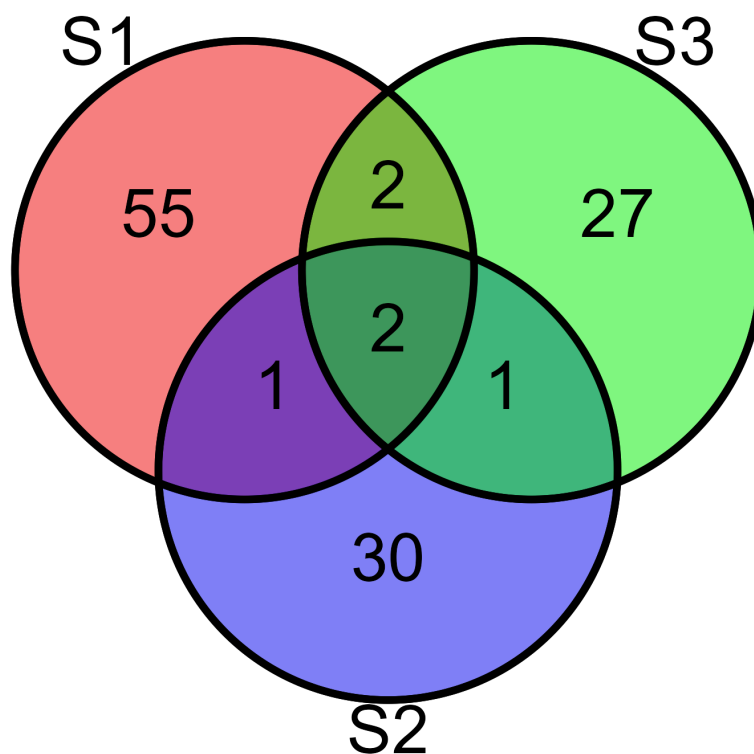


Figure 1: Venn diagram of DEGs for each subtype

115 2.2 Over Representation Analysis (ORA)

116 In order to understand the biological pathways associated with the DEGs, ORA was performed
117 on Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), and WikiPathways
118 databases. The full list of pathways from the ORA can be found in Supplementary Table 2.

119 2.2.1 Gene Ontology (GO)

120 The results include distinctive pathways characterizing each subtype (S1: 73; S2: 18; S3: 16),
121 with seven GO terms in common between S1 and S3, and three in common between S1 and S2
122 (Figure 2).

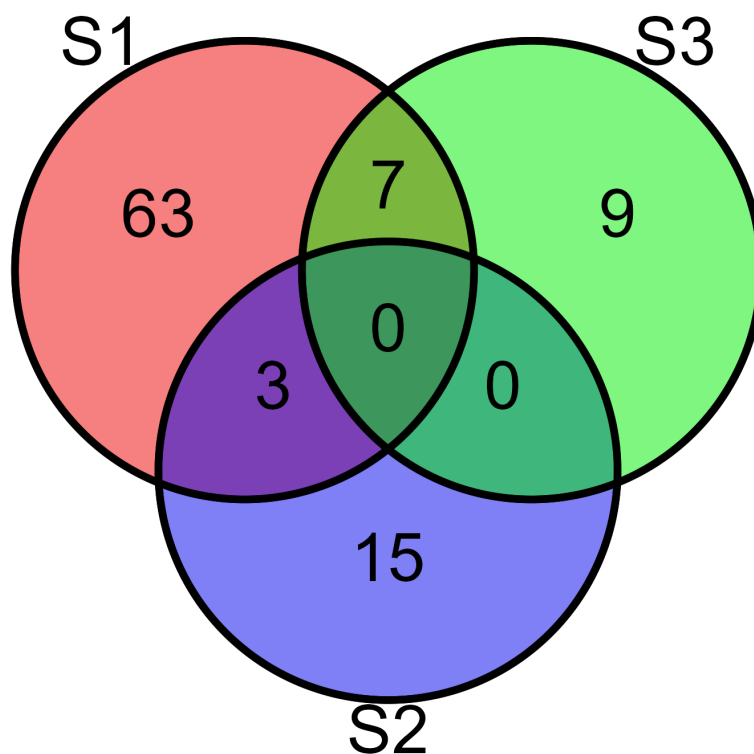


Figure 2: Venn diagram of GO terms for each subtype.

123 2.2.1.1 S1

124 The main theme of S1 biological pathways resulting from ORA encompassed cellular energy
125 metabolism, gene expression regulation, and cellular adaptation to various stressors. The pres-
126 ence of pathways associated with *oxidative phosphorylation* (GO:0006119, q-value: 5.54×10^{-7}),
127 *aerobic respiration* (GO:00099060, q-value: 3.74×10^{-6}), and *cellular respiration* (GO:0045333,
128 q-value: 1.38×10^{-5}) indicated a modulation of cellular energy derivation processes, mediated
129 by organic compounds oxidation. Additionally, the presence of pathways related to ATP synthe-
130 sis, including *mitochondrial ATP synthesis coupled electron transport* (GO:0042775, q-value:
131 5.54×10^{-7}), and *proton motive force-driven mitochondrial ATP synthesis* (GO:0042776, q-
132 value: 1.73×10^{-5}), highlighted the modulation of cellular energy metabolism in this disease sub-
133 type. There were several pathways associated with nucleotide metabolism, including *nucleotide*

134 *metabolic process* (GO:0009117, q-value: 1.54×10^{-2}) and *nucleoside phosphate metabolic pro-*
135 *cess* (GO:0006753, q-value: 1.58×10^{-2}), along with pathways associated with RNA splicing,
136 such as *RNA splicing* (GO:0008380, q-value: 1.14×10^{-3}), and *mRNA splicing, via spliceosome*
137 (GO:0000398, q-value: 4.13×10^{-3}). Cellular response to stress pathways were significantly en-
138 riched by the set of DEGs, including *cellular oxidant detoxification* (GO:0098869, q-value: 6.68
139 $\times 10^{-3}$), *response to reactive oxygen species* (GO:0000302, q-value: 1.26×10^{-2}), and *cellular*
140 *response to toxic substance* (GO:0097237, q-value: 1.26×10^{-2}).

141 **2.2.1.2 S2**

142 The significantly enriched pathways for this disease subtype mainly pointed to regulation of
143 gene expression and metabolic processes. The most significant term, with the lowest q-value of
144 7.43×10^{-7} , was *RNA processing* (GO:0006396). Along with this, several terms associated with
145 RNA metabolism and processing were found significant. These terms included *macromolecule*
146 *metabolic process* (GO:0043170, q-value: 2.38×10^{-2}), *RNA metabolic process* (GO:0016070,
147 q-value: 1.23×10^{-3}), and *nucleic acid metabolic process* (GO:0090304, q-value: 3.51×10^{-3}).
148 The term with the highest gene ratio found was *metabolic process* (GO:0008152, q-value: $2.92 \times$
149 10^{-2}), highlighting the modulation of metabolism. This subtype had seven GO terms in common
150 with S1: *RNA splicing, via transesterification reactions; RNA splicing, via transesterification*
151 *reactions with bulged adenosine as nucleophile; mRNA splicing, via spliceosome*.

152 **2.2.1.3 S3**

153 The main theme of S3 biological pathways resulting from the ORA was the response to ox-
154 idative stress and detoxification processes. The pathways with the lower q-values include *re-*
155 *sponse to hydrogen peroxide* (GO:0042542, q-value: 5.76×10^{-4}), *carbon dioxide transport*
156 (GO:0015670, q-value: 5.76×10^{-4}), and *oxygen transport* (GO:0015671, q-value: 5.76×10^{-4}).
157 The presence of these pathways suggests a cellular response to reactive oxygen species, in-
158 cluding the catabolic and metabolic processes of hydrogen peroxide. Additionally, pathways
159 related to detoxification processes were highlighted, such as *cellular oxidant detoxification*
160 (GO:0098869, q-value: 6.62×10^{-3}), *cellular detoxification* (GO:1990748, q-value: 8.13×10^{-3}),
161 and *detoxification* (GO:0098754, q-value: 1.08×10^{-2}). Furthermore, the pathways *cellular re-*
162 *sponse to toxic substances* (GO:0097237, q-value: 8.14×10^{-3}) and *reactive oxygen species*
163 *metabolic processes* (GO:0072593, q-value: 2.7×10^{-2}) further support the main theme of ox-

164 idative stress response and detoxification. This subtype had seven GO terms in common with
165 S1: *response to reactive oxygen species, response to hydrogen peroxide, hydrogen peroxide*
166 *metabolic process, cellular response to toxic substance, detoxification, cellular oxidant detoxi-*
167 *fication, cellular detoxification.*

168 2.2.2 Kyoto Encyclopedia of Genes and Genomes (KEGG)

169 ORA on the KEGG database showed pathways characterizing each subtype (S1: 16; S2: 1; S3:
170 2). Most of the pathways resulting from KEGG analysis are unique to the specific subtypes
171 (Figure 3).

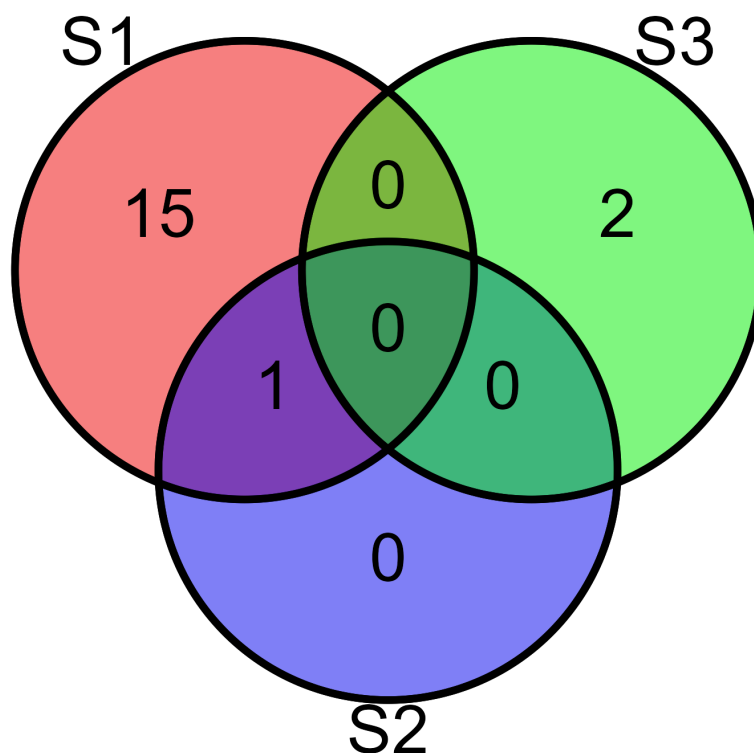


Figure 3: Venn diagram of KEGG terms for each subtype.

172 2.2.2.1 S1

173 The main theme of these ORA results on KEGG database is related to neurological diseases and

174 neurodegeneration, including *Parkinson's disease* (hsa05012, q-value: 1.85×10^{-5}), *Huntington*
175 *disease* (hsa05016, q-value: 2.96×10^{-4}), *prion disease* (hsa05020, q-value: 4.79×10^{-4}), *amy-*
176 *otrophic lateral sclerosis* (hsa05014, q-value: 2.83×10^{-3}), and *Alzheimer disease* (hsa05010,
177 q-value: 2.83×10^{-3}). Additionally, the presence of *oxidative phosphorylation* (hsa00190, q-
178 value: 5.72×10^{-8}) as the most significant resulting pathway further points to a modulation of
179 energy metabolism.

180 **2.2.2.2 S2**

181 This analysis yielded a single significant pathway, namely *Spliceosome* (hsa03040, q-value:
182 9.74×10^{-6}), which further points to the regulation of gene expression. S2 shares his sole
183 pathway with S1 ([Figure 3](#)).

184 **2.2.2.3 S3**

185 Here there are two significant pathways, namely *African trypanosomiasis* (hsa05143, q-value:
186 2.06×10^{-3}) and *Malaria* (hsa05144, q-value: 2.06×10^{-3}), suggesting the modulation of path-
187 ways involved in detoxification processes.

188 **2.2.3 WikiPathways**

189 ORA on the WikiPathways database showed pathways characterizing each subtype (S1: 5; S2:
190 2; S3: 19). Most of these pathways are unique to the subtypes ([Figure 4](#)).

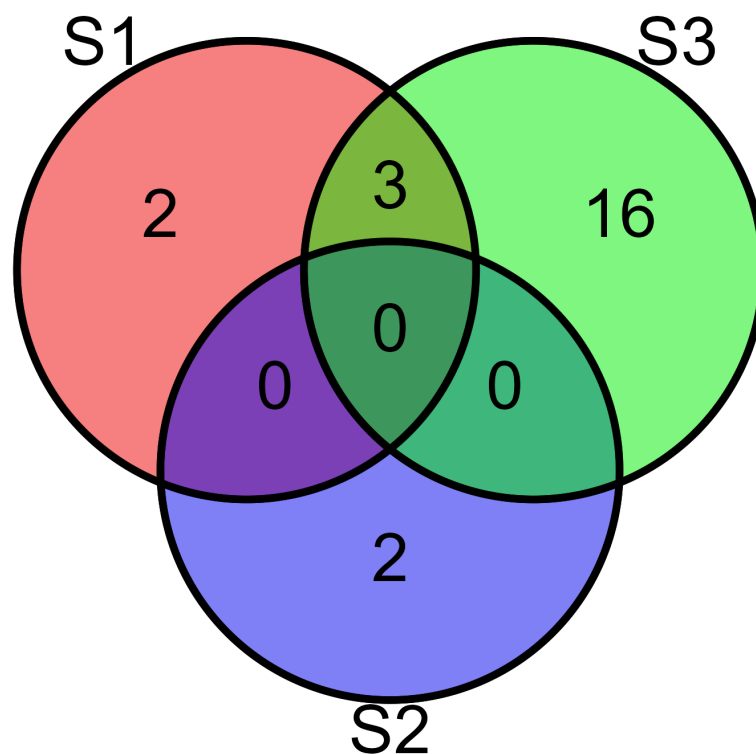


Figure 4: Venn diagram of WikiPathways terms for each subtype.

191 **2.2.3.1 S1**

192 Pathways regarding mitochondrial function and energy production represent the main theme in
193 these results. *Oxidative phosphorylation* (WP623, q-value: 5.68×10^{-3}) and *Electron transport*
194 *chain: OXPHOS system in mitochondria* (WP111, q-value: 3.02×10^{-6}) represent the modulation
195 of ATP generation pathways, along with *Nonalcoholic fatty liver disease* (WP4396, q-value: 1.89
196 $\times 10^{-3}$), associated with mitochondrial dysfunction and impaired energy metabolism.

197 **2.2.3.2 S2**

198 This set of results only includes two pathways, namely *Endoderm differentiation* (WP2853, q-
199 value: 1.85×10^{-2}) and *Mesodermal commitment pathway* (WP2857, q-value: 1.85×10^{-2}). Those
200 pathways are enriched by only one gene, namely *NCAPG2*.

201 **2.2.3.3 S3**

202 The main theme in this set of results is related to cellular signaling and metabolism. Many
203 of the terms are pathways involved in signaling processes associated with apoptosis, includ-
204 ing *Photodynamic therapy-induced NF- κ B survival signaling* (WP3617, q-value: 6.40×10^{-3}),
205 and *Apoptosis-related network due to altered Notch3 in ovarian cancer* (WP2864, q-value:
206 6.40×10^{-3}). Additionally, several pathways are involved in metabolism, including *Vitamin B12*
207 *metabolism* (WP1533, q-value: 6.40×10^{-3}), *Folate metabolism* (WP176, q-value: 6.40×10^{-3}),
208 and *Selenium micronutrient network* (WP15, q-value: 6.40×10^{-3}). S3 shares three pathways
209 with S1, namely *Oxidative phosphorylation* (WP623), *Electron transport chain: OXPHOS sys-*
210 *tem in mitochondria* (WP111), and *Mitochondrial complex I assembly model OXPHOS system*
211 (WP4324).

212 **2.3 Gene Set Enrichment Analysis (GSEA)**

213 The examination of overall gene expression levels was carried out through GSEA. This analysis
214 is not limited to the set of DEGs, as it accounts for variations in gene expression across all
215 analyzed genes. The complete results tables can be found in Supplementary Table 2.

216 **2.3.1 GO**

217 The number of enriched BP terms found in S1 was 1092, while 1070 enriched terms were found
218 in S2, and 134 enriched terms were found in S3. Venn plots show that most pathways were
219 shared between S1 and S2, while far less were shared with S3 ([Figure 5](#)).

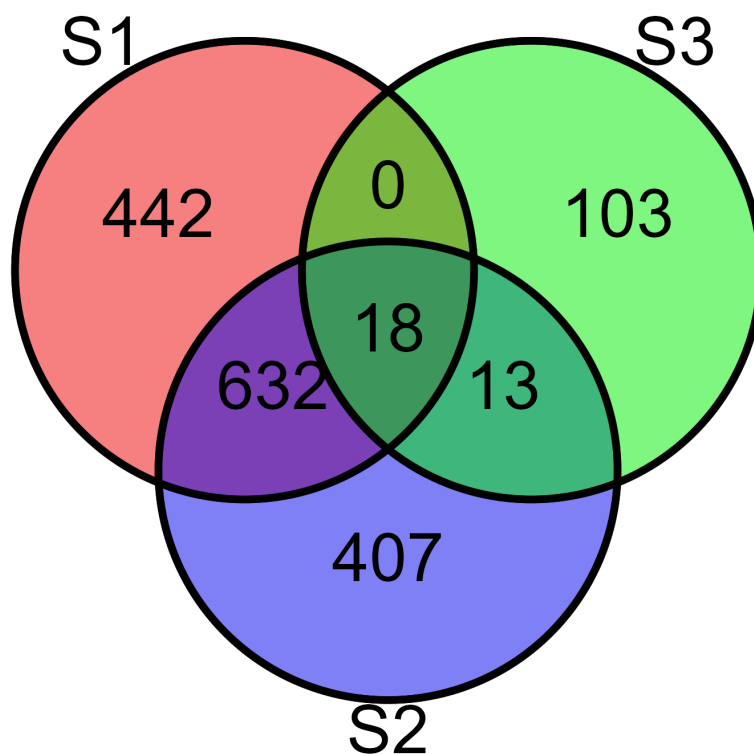


Figure 5: Venn diagram of GO terms for each subtype

220 2.3.1.1 S1

221 Gene Ontology BP domain pathways enrichment revealed main themes related to organismal
222 processes, cell signaling, and energy metabolism. A primary parent term was *Multicellular Or-*
223 *ganismal Processes* (GO:0032501, q-value: 5.49×10^{-50}). Also found as enriched with very high
224 significance there were *Nervous System Development* (GO:0007399, q-value: 9.00×10^{-38}),
225 and *Anatomical Structure Development* (GO:0048856, q-value: 6.30×10^{-35}). Cell signal-
226 ing and homeostatic processes were identified as enriched, as exemplified by the presence
227 of *Cell-Cell Signaling* (GO:0007267, q-value: 4.36×10^{-28}) and *Ion Transmembrane Transport*
228 (GO:0034220, q-value: 5.28×10^{-21}), along with *Cellular Processes* (GO:0009987, q-value: 7.54
229 $\times 10^{-12}$), and *Cell Communication* (GO:0007154, q-value: 8.86×10^{-20}). Energy metabolism was
230 mainly represented by *oxidative phosphorylation* (GO:0006119, q-value: 8.56×10^{-7}), *aerobic*

231 *respiration* (GO:0009060, q-value: 1.31×10^{-5}), and *ATP biosynthetic process* (GO:0006754,
232 q-value: 6.13×10^{-6})

233 **2.3.1.2 S2**

234 The Gene Ontology analysis revealed biological pathways associated with organismal processes,
235 structures development, and cellular signaling. The most significant pathways were related
236 to *multicellular organismal processes* (GO:0032501, q-value: 5.12×10^{-68}), followed by devel-
237 opment pathways like *nervous system development* (GO:0007399, q-value: 2.26×10^{-47}) and
238 *anatomical structure morphogenesis* (GO:0009653, q-value: 8.43×10^{-41}). Signaling pathways
239 included *cell-cell signaling* (GO:0007267, q-value: 1.45×10^{-29}) and *G protein-coupled recep-*
240 *tor signaling pathway* (GO:0007186, q-value: 1.67×10^{-18}). Pathways related to *response to*
241 *stimulus* (GO:0050896, q-value: 5.62×10^{-10}), like *detection of stimulus involved in sensory*
242 *perception* (GO:0050906, q-value: 9.03×10^{-13}) were also found in this set of results. More-
243 over, this GSEA analysis yielded many pathways related to RNA metabolism and processing,
244 such as *positive regulation of transcription by RNA polymerase II* (GO:0045944, q-value: 2.03
245 $\times 10^{-11}$) and *positive regulation of RNA metabolic process* (GO:0051254, q-value: 1.11×10^{-3}).
246 Pathways for S1 and S2 were mostly shared and related to morphological changes (*nervous*
247 *system development, anatomical structure development, anatomical structure morphogenesis,*
248 *tissue development*). Interestingly, all pathways from S1 and S2 showed opposite enrichment
249 scores, indicating that these two groups were characterized by an opposite expression pattern
250 despite sharing most of their enriched pathways ([Figure 6](#)).

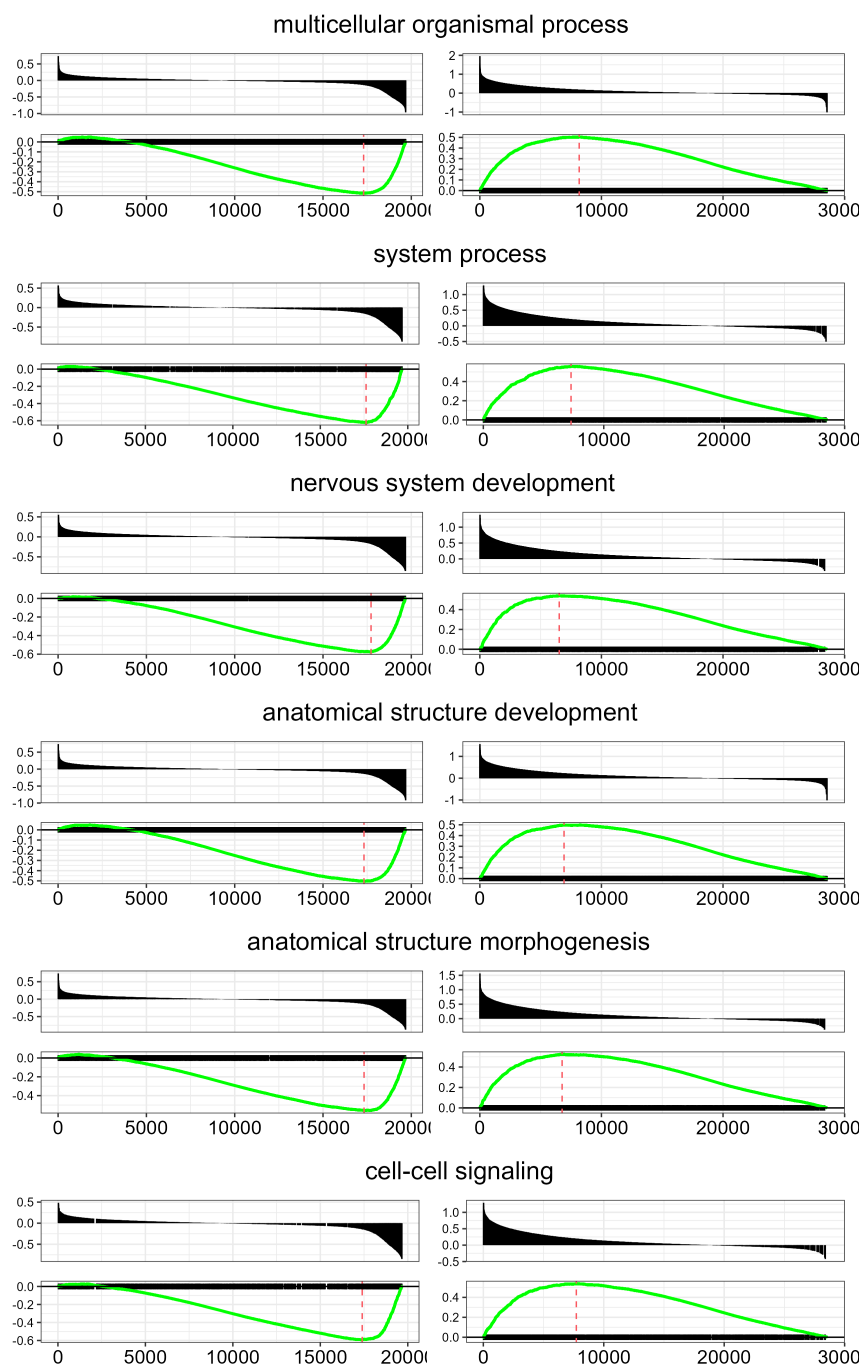


Figure 6: **Pathway Enrichment Analysis.** Visual representation of six distinct pathways, each labelled with its respective name as a section title. Within each section, there are two sets of plots: S1 on the left and S2 on the right. The upper plots illustrate the positions of gene set members on a 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.

251 **2.3.1.3 S3**

252 All enriched pathways in this set of results were distinctive of S3 (none was shared with the other
253 subtypes). One of the prominent themes identified in our analysis is related to sensory percep-
254 tion and signal transduction. Notably, the pathway *detection of chemical stimulus involved in*
255 *sensory perception of smell* (GO:0050911, q-value: 2.33×10^{-6}) and *detection of chemical stim-*
256 *ulus* (GO:0009593, q-value: 5.96×10^{-6}) were highly significant in this set of results. Another
257 important theme centers around cell signaling and regulation, with pathways like *positive regu-*
258 *lation of antigen receptor-mediated signaling pathway* (GO:0050857, q-value: 3.17×10^{-5}) and
259 *G protein-coupled receptor signaling pathway* (GO:0007186, q-value: 1.03×10^{-4}) were highly
260 enriched in this theme, indicating their crucial roles in modulating cellular responses and in-
261 tercellular communication. Furthermore, our analysis highlighted pathways associated with the
262 regulation of gene expression, such as *regulation of RNA export from nucleus* (GO:0046831,
263 q-value: 9.80×10^{-5}). Relevantly to Parkinson's disease, results included *cellular response to*
264 *misfolded protein* (GO:0071218, 5.92×10^{-3}) as an enriched pathway.

265 **2.3.2 KEGG**

266 This analysis resulted in 83 enriched terms found for S1, 15 for S2, and 3 for S3 ([Figure 7](#)).

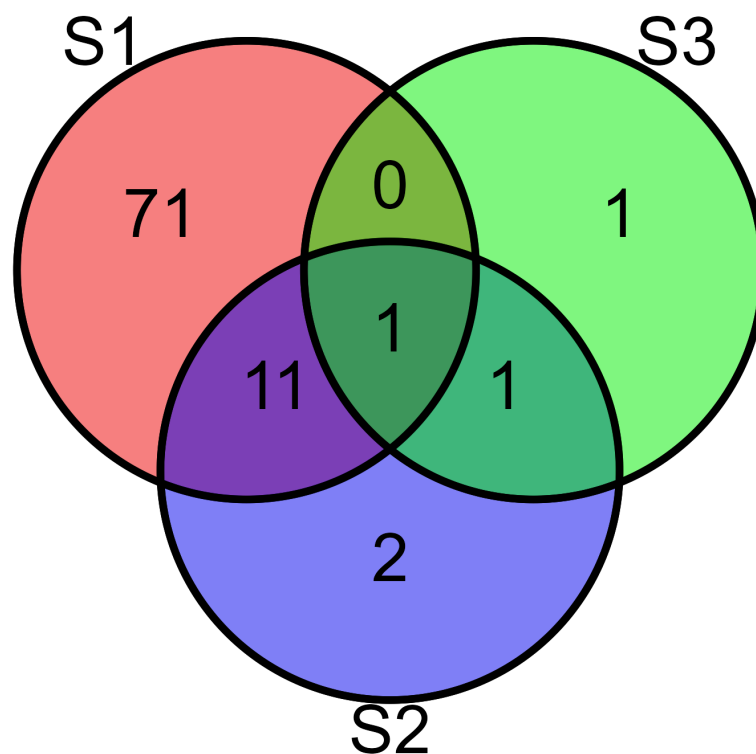


Figure 7: Venn diagram of KEGG terms for each subtype.

267 2.3.2.1 S1

268 The main theme of the pathways is the regulation of physiological processes and diseases, in-
269 cluding cell signaling and communication pathways, metabolism, disease pathways, along with
270 the regulation of the immune system. The upregulation of protein synthesis is highlighted by
271 the presence of the *Ribosome pathway* (hsa03010, q-value: 2.94×10^{-14}), and along with *neu-*
272 *trophil extracellular trap formation* (hsa04613, q-value: 1.46×10^{-13}) and *osteoclast differenti-*
273 *ation* (hsa04380, q-value: 1.10×10^{-6}) pathways, suggest cellular processes and immune sys-
274 tem dysregulation. Cell signaling pathways are also significant, such as *glutamatergic synapse*
275 (hsa04724, q-value: 6.77×10^{-7}), *GABAergic synapse* (hsa04727, q-value: 2.34×10^{-5}), and
276 *cholinergic synapse* (hsa04725, q-value: 8.77×10^{-4}) suggesting an implication of disrupted
277 neuronal signaling. Noteworthy, there were again pathways regarding metabolic and energy

278 production dysregulation, such as *oxidative phosphorylation* (hsa00190, q-value: 3.58×10^{-6})
279 and *protein digestion and absorption* (hsa04974, q-value: 2.75×10^{-9}). Finally, pathways related
280 to addiction were found, like *Nicotine addiction* (hsa05033, q-value: 4.38×10^{-5}) and *Morphine*
281 *addiction* (hsa05032, q-value: 2.18×10^{-3}).

282 **2.3.2.2 S2**

283 Pathways involved in cell communication and signal transduction in the nervous system are
284 found modulated in this set of results, including *Neuroactive ligand-receptor interaction* (hsa04080,
285 q-value: 5.93×10^{-14}), *Calcium signaling pathway* (hsa04020, q-value: 1.85×10^{-6}), and *Olfac-*
286 *tory transduction* (hsa04740, q-value: 2.09×10^{-8}). Cell development and connectivity was also
287 modulated, as indicated by the presence of pathways like *Wnt signaling pathway* (hsa04310,
288 q-value: 3.98×10^{-3}) and *Axon guidance* (hsa04360, q-value: 6.84×10^{-3}). Like in S1, pathways
289 regarding addiction processes were found, such as *Nicotine addiction* (hsa05033, q-value: 1.84
290 $\times 10^{-3}$) and *Morphine addiction* (hsa05032, q-value: 6.16×10^{-3}).

291 **2.3.2.3 S3**

292 Here three pathways are found as significantly enriched, indicating a positive regulation of
293 *olfactory transduction* (hsa04740, q-value: 6.02×10^{-10}) along with *Neuroactive ligand-receptor*
294 *interaction* (hsa 04080, q-value: 2.78×10^{-10}). Additionally, *Protein export pathway* (hsa03060,
295 q-value: 4.57×10^{-2}) was found enriched.

296 **2.3.3 WikiPathways**

297 Here 86 enriched terms were found in S1, 40 enriched terms were found in S2, 1 enriched term
298 was found in S3 ([Figure 8](#)).

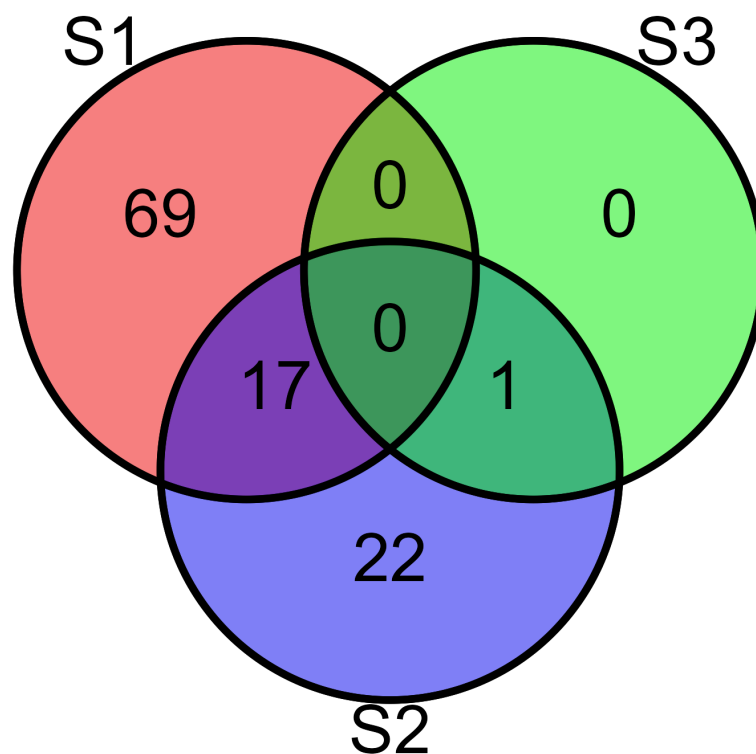


Figure 8: Venn diagram of WikiPathways terms for each subtype.

299 2.3.3.1 S1

300 In this set of results, the main theme was driven by enriched pathways related to protein syn-
301 thesis, cellular metabolism, neuronal signaling, and immune system response. Notably, the Cy-
302 *toplasmic ribosomal proteins pathway* (WP477, q-value: 1.81×10^{-14}) confirmed the modulation
303 of processes related to protein synthesis. Accordingly, the *Electron transport chain: OXPHOS*
304 *system in mitochondria pathway* (WP111, q-value: 3.58×10^{-7}) confirmed the importance of ox-
305 idative phosphorylation in energy production. The results also included pathways associated
306 with neuronal signaling, like *Phosphodiesterases in neuronal function* (WP4222, q-value: 1.50
307 $\times 10^{-4}$), *mBDNF and proBDNF regulation of GABA neurotransmission* (WP4829, q-value: 1.03
308 $\times 10^{-2}$), and *Neuroinflammation and glutamatergic signaling* (WP5083, q-value: 2.07×10^{-2}),
309 also pointing out to an involvement of the immune response, along with *IL-3 signaling pathway*

310 (WP286, q-value: 2.65×10^{-4}). Related to this, the *TYROBP causal network in microglia path-*
311 *way* (WP3945, q-value: 1.53×10^{-5}) highlighted the involvement of the regulatory mechanisms
312 within microglia. Another notable theme revolved around disease processes, as there was a
313 significant enrichment of pathways like *Nonalcoholic fatty liver disease* (WP4396, q-value: 1.05
314 $\times 10^{-3}$) and *Hepatitis B infection* (WP4666, q-value: 1.69×10^{-3}).

315 **2.3.3.2 S2**

316 This set of GSEA results revealed pathways encompassing signaling mechanisms, neuroge-
317 nesis, developmental processes, immune response, and disease processes. *GPCRs, class A*
318 *rhodopsin-like* (WP455, q-value: 4.52×10^{-6}) was the most significant, pointing out to signaling
319 along with *GPCRs, other* (WP117, q-value: 4.70×10^{-3}), and *GABA receptor signaling* (WP4159,
320 q-value: 4.02×10^{-2}). Pathways related to cellular differentiation and neurogenesis were also
321 present, such as *dopaminergic neurogenesis* (WP2855, q-value: 1.56×10^{-2}) and *Neural crest*
322 *differentiation* (WP2064, q-value: 9.45×10^{-4}). Developmental processes were represented by
323 pathways such as *cardiac progenitor differentiation* (WP2406, q-value: 2.40×10^{-4}), *osteoblast*
324 *differentiation and related diseases* (WP4787, q-value: 2.60×10^{-4}), and *neural crest differ-*
325 *entiation* (WP2064, q-value: 9.45×10^{-4}). Additionally, results included pathways associated
326 with immune response, such as *host-pathogen interaction of human coronaviruses - interferon*
327 *induction* (WP4880, q-value: 2.52×10^{-4}), *immune response to tuberculosis* (WP4197, q-value:
328 4.34×10^{-4}), and *SARS coronavirus and innate immunity* (WP4912, q-value: 4.26×10^{-2}). Finally,
329 we identified pathways associated with genetic disorders and syndromes, including *Prader-Willi*
330 *and Angelman syndrome* (WP3998, q-value: 1.09×10^{-2}), *MECP2 and associated Rett syndrome*
331 (WP3584, q-value: 1.35×10^{-2}), and *Cornelia de Lange Syndrome - SMC1/SMC3 role in DNA*
332 *damage* (WP5118, q-value: 1.86×10^{-2}).

333 **2.3.3.3 S3**

334 The GSEA here yielded only one significant pathway, namely *Interactome of polycomb repres-*
335 *sive complex 2 (PRC2)* (WP2916, q-value: 4.04×10^{-2}), indicating a modulation in gene expres-
336 sion regulation and chromatin organization.

337 2.4 Baseline prediction of disease progression subtype

338 A Machine Learning hierarchical classification approach was implemented to develop a predic-
339 tion system aimed at identifying the disease subtypes of a newly-diagnosed PD patient, namely
340 at the baseline. Data from multiple modalities were used, including demographics, motor, non-
341 motor, biospecimen, imaging (See [section 4.6, Table 2](#)). The first model in the hierarchy aimed
342 to predict whether the subject was from S3, which has the most distinctive phenotype and is
343 also the most severe. The model achieved a fair performance with 0.814 sensitivity, and 0.757
344 specificity, yielding an F-Score of 0.828 and a total AUROC of 0.877 ([Figure 9](#)).

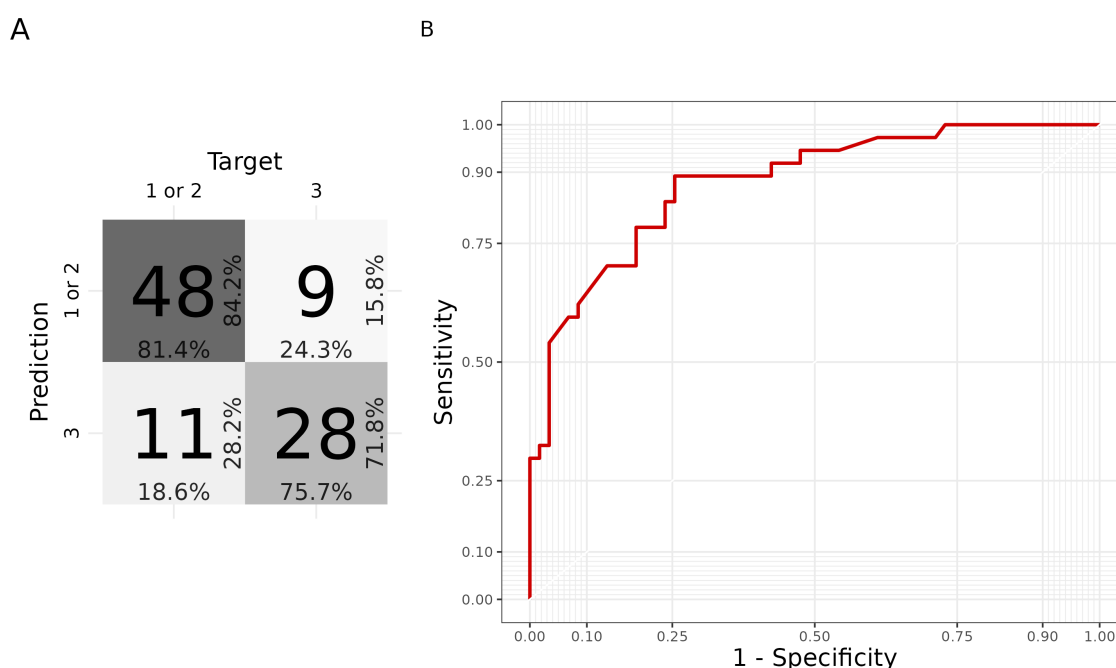


Figure 9: ROC curve and confusion matrix from the first model of the hierarchy.

345 Variable importance was investigated with the application of an explainable Artificial Intelli-
346 gence (XAI) method, namely SHAP values. These highlighted the score to MDS-UPDRS Part II
347 (disability evaluation) as the most important factor contributing to S3 identification. Among the
348 most important variables there are other clinical measures, along with a neuroimaging measure
349 (DaTScan Caudate R). Gene expression only had a marginal importance, with low absolute SHAP
350 values, giving little contribution to the final prediction (??).

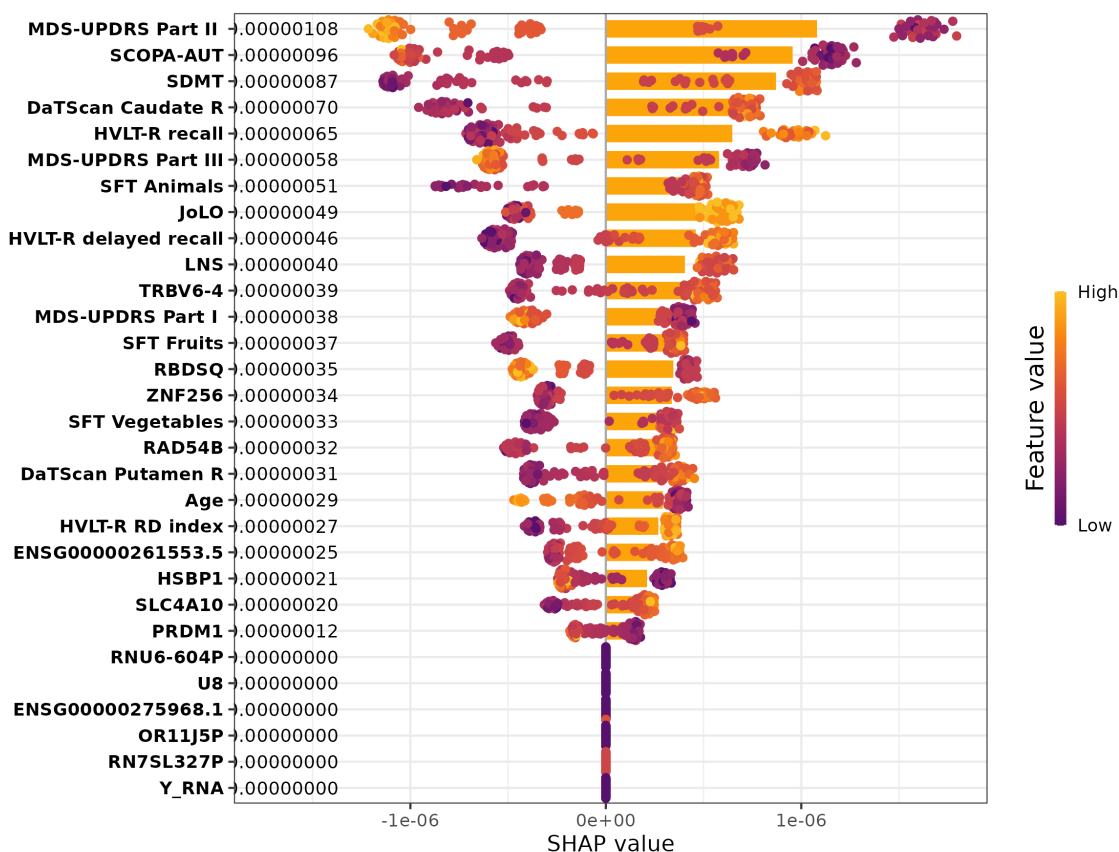


Figure 10: SHAP summary plot representing the contribution of each variable to the prediction of the model.

??

351 For all those subjects that the model did not classify as S3, the second level of the hierarchy
 352 included a model aiming to predict whether the subject was from S1 or S2. It achieved a poor
 353 performance, with 0.745 sensitivity, 0.25 specificity, yielding a F-Score of 0.77 and a total AU-
 354 ROC of 0.576 (Figure 11).

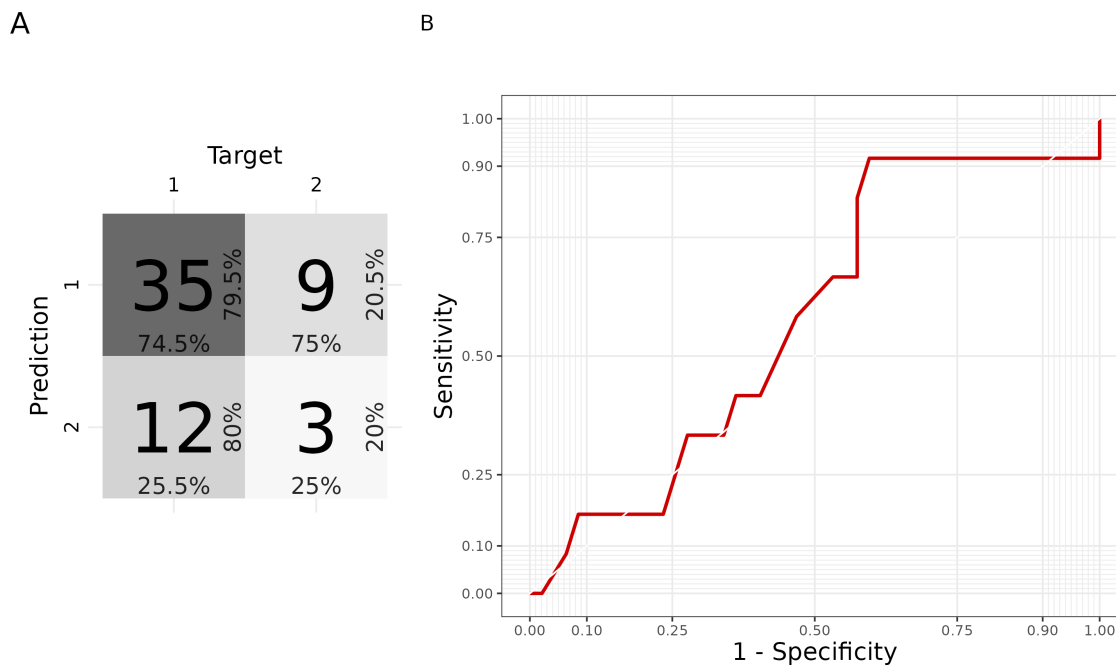


Figure 11: ROC curve and confusion matrix from the second model of the hierarchy.

355 SHAP values indicated that expression values of *U8*, *HSBP1*, *TRBV6-4*, and *SCL4A10*, along with
356 Benton Judgement of Line Orientation test score, were the most important factors to discrimi-
357 nate between S1 and S2 (Figure 12).

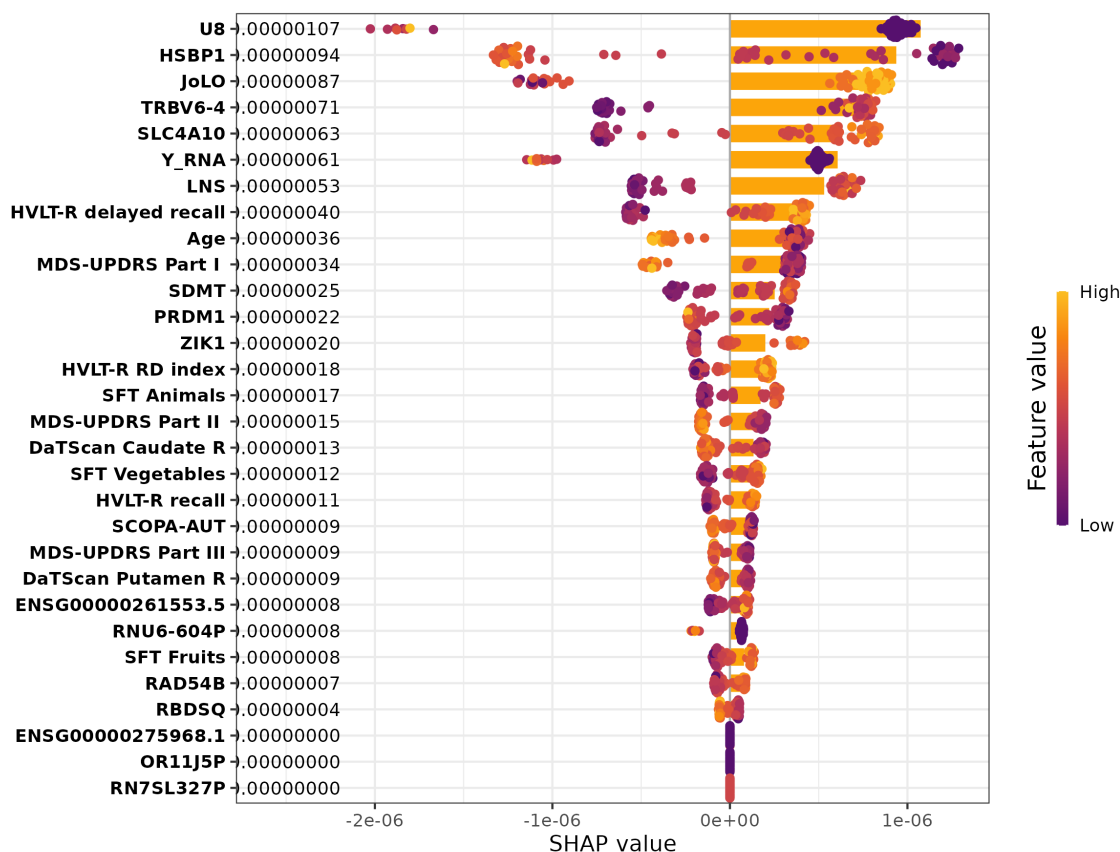


Figure 12: SHAP summary plot representing the contribution of each variable to the prediction of the model.

3 Discussion

The identification of progression subtypes is of extreme importance in order to attempt settling the heterogeneity of PD. Recent research has shown that people with PD can exhibit a variety of progression patterns from diagnosis onwards [5, 8, 19, 20, 22, 24]. 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 currently unfeasible due to the lack of validated nonclinical biomarkers of PD progression [25], 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 Artificial Intelligence algorithm that reliably takes into account time as a dimension. Briefly, S1 had mild motor and non-motor symptoms at baseline,

370 with a moderate rate of motor impairment increase and relatively stable cognitive abilities; S2
371 had moderate motor and non-motor symptoms at baseline, with a slow progression rate; and
372 S3 started with significant motor and non-motor symptoms, showing a rapid progression of
373 impairment, and thus reporting the worse prognosis among the three [20].

374 The DEGs identified in this study are unique to these progression subtypes, as none of the genes
375 that are commonly found as differentially expressed in PD studies are present in our results. As
376 a specific example, common transcriptomic markers such as *SYN1*, *ANKRD22*, and *SLC14A1*
377 [16] are absent from all our DEGs lists. This result is not surprising to us, as our experiment had
378 two main differences with other PD RNA studies. First, although based on transcriptomics of PD
379 subjects, we investigated progression subtypes as diagnostic classes, thus differences with a
380 classical PD group were expected. Second, our differential expression analysis was performed
381 as a time course experiment, in order to identify those genes that varied for expression values
382 as a result of the disease over time. This profoundly differs with previous PD transcriptomics
383 studies, which performed a cross-sectional analysis of gene expression, thus not taking time
384 into account. As a further note, there is general poor consensus between previous studies on
385 resulting DEGs from PD studies [15].

386 **3.1 S1**

387 All pathway analyses consistently highlighted the modulation of cellular energy metabolism, par-
388 ticularly pathways associated with oxidative phosphorylation, aerobic respiration, and cellular
389 respiration. Additionally, pathways related to ATP synthesis, mitochondrial dysfunction, and nu-
390 cleotide metabolism were commonly enriched across the ORA and GSEA over GO, KEGG and
391 WikiPathways databases. The modulation of energy metabolism is well known in PD, and it has
392 already been found from transcriptomics analyses both in blood and brain sample tissues [26,
393 27, 28]. Cellular response to stress pathways, including oxidant detoxification and response
394 to reactive oxygen species, were also consistently identified. Furthermore, the results consis-
395 tently pointed towards the involvement of neurological diseases and neurodegeneration, with
396 pathways associated with Parkinson's disease, Alzheimer's disease, Huntington disease, prion
397 disease, and amyotrophic lateral sclerosis consistently enriched.

398 Despite these similarities, there were also differences observed across the pathway analyses of
399 S1 data. In fact, there were different specific pathways within the broader common themes. For

400 instance, one analysis emphasized the significance of pathways related to ribosomal proteins
401 in protein synthesis, while another highlighted the importance of neuronal signaling pathways
402 and immune system dysregulation. Disease-related pathways such as nonalcoholic fatty liver
403 disease and hepatitis B infection were specifically enriched in one analysis. The involvement
404 of immune system response and processes related to oxidative stress are known in PD tran-
405 scriptomics [15, 17], and the observation of disease pathways enrichment is related to their
406 modulation.

407 The biological profile of S1 shares similarities with that of PD patients with *LRRK2* mutation,
408 which is involved in multiple biological functions, including mitochondrial activity and oxidative
409 pathways [29]. It is interesting to note that none of the patients included in this study had a
410 mutation in one of the risk loci known for PD, as this study was solely focused on idiopathic
411 PD. Nonetheless, it has already been observed that the patients with idiopathic PD or *LRRK2*
412 genetic PD show mostly overlapping phenotypes, and they are clinically difficult to distinguish
413 [30].

414 Cellular signaling pathways were also found enriched in the GSEA, confirming that signaling
415 mechanisms, often found among transcriptomics alterations from PD *post mortem* brain tissues
416 [31], can also emerge from the analysis of peripheral tissues, such as blood [32, 33].

417 **3.2 S2**

418 Pathway analyses consistently identified modulation of gene expression regulation and metabolic
419 processes. Specifically, pathways associated with RNA metabolism and processing emerged
420 among the most significant terms across all analyses. The implication of RNA metabolic pro-
421 cesses has been considered in the pathogenesis and disease course of PD, advancing that
422 these may be related to energy conservation, aggregated proteins modulation, and response to
423 cellular stress [34].

424 One notable difference lies in the number of pathways identified in each analysis, as some
425 analyses revealed a limited number of pathways. As an example, there were only two significant
426 pathways in the ORA on WikiPathways: Endoderm differentiation (WP2853) and Mesodermal
427 commitment pathway (WP2857). These were enriched by a single gene, *NCAPG2*. This gene
428 encodes for a regulatory subunit of the condensin II complex which, along with the condensin I
429 complex, plays a role in chromosome assembly and segregation during mitosis [35]. Alterations

430 of this gene have been associated with cancer and neurodevelopmental defects [36, 37], and
431 although its presence has already been observed in PD blood transcriptomics [38, 39], its role
432 in the disease is still unclear.

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

436 In GSEA results on the KEGG database, S2 exhibited pathways associated with addiction pro-
437 cesses, sharing this characteristic with S1. Pathways related to morphine addiction also emerged
438 in a recent evaluation of PD proteome from dopaminergic neurons in the substantia nigra (SN),
439 suggesting an involvement of potentially compromised GABA-related pathways [40].

440 **3.3 S3**

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

455 Metabolic pathways such as Vitamin B12 metabolism, Folate metabolism, and Selenium mi-
456 cronutrient network, were also found altered in this subtype. Recent studies have shown that
457 B12 deficiency is common in patients with neuropathies, and PD has B12 levels decline over the
458 course of the disease [41].

459 **3.4 Results comparison between subtypes**

460 The results of our transcriptomics analysis revealed a number of similarities between the three
461 PD subtypes (S1, S2, and S3). All three subtypes showed a significant modulation of pathways
462 related to the regulation of gene expression, metabolism, and cell signaling. Pathways associ-
463 ated with nervous system dysregulation were consistently found in all three subtypes. Although
464 expected when analyzing brain cells, we believe that when resulting in blood it's a confirmatory
465 result of appropriate transcriptomics findings, and this is also in line with previous works on pe-
466 ripheral tissues [15, 42]. We may consider this as a general alteration due to the disease state,
467 as these were also found in other PD transcriptomics experiments [16], and not distinctive of
468 any of the subtypes.

469 S1 and S2 had a few shared themes, including addiction pathways, structure development, im-
470 mune response alterations and disease processes. In fact, among the distinctive character-
471 istics for S1 we find neurological and neurodegenerative disease pathways. Moreover, S1 was
472 unique in its alteration of energy production and mitochondrial functions. Interestingly, all of
473 the shared pathways between S1 and S2 had opposite enrichment patterns in the GSEA (Figure
474 6). This demonstrates that S1 and S2 are distinct progression forms of the same disease. De-
475 spite sharing a few transcriptomic characteristics, these appear to be modulated in opposing
476 ways, and thus may be at the foundation of their different progression courses. S2 was unique
477 in its alteration of pathways related to developmental processes and neurogenesis. Moreover,
478 this subtype showed an alteration in olfactory transduction, as also observed in S3. S3 was
479 unique in its increased expression of genes involved in detoxification processes, and pathways
480 related to cellular stress response were altered in both S1 and S3. Interestingly, this was the
481 only subtype characterized by enrichment of response to misfolded proteins.

482 **3.5 Subtype prediction at baseline**

483 The machine learning classifier provided a reliable tool to predict disease progression subtypes
484 using baseline data. This tool could easily be implemented into a user-friendly software, to fi-
485 nally build a reliable Computer-Aided Diagnosis (CAD) tool to identify subjects with the most
486 severe prognosis. As resulting from the variable importance analysis, the contribution of gene
487 expression was marginal for the prediction of S3, not allowing for substantial discrimination
488 between disease subtypes in neither of the steps of the hierarchical ML approach. Clinical vari-

489 ables instead demonstrated high importance to identify S3 subjects, with perceived disability
490 (MDS-UPDRS Part II) being the most important predictor for a more severe prognosis. In fact,
491 S3 subjects were characterized by a faster progression and worse symptomatology, sharing
492 some similarities with the classical Posture Instability / Gait Difficulty (PIGD) subtype. Inter-
493 estingly, most of the S3 subjects were PIGD patients, and those that were Tremor Dominant
494 (TD) instead were likely to shift to PIGD over 6 years [20]. Although expression values resulted
495 as the most important factors to discriminate between S1 and S2, the model at the second level
496 of the hierarchy had a poor test performance. This made it unreliable and, as a consequence,
497 the evaluation of its behavior is meaningless. Considering that this hierarchical classification
498 model has 0.877 AUROC to detect the most severe subtype, this would give useful indication
499 for prognosis. As such, this ML model may foster precision medicine for PD, providing support
500 for a finer-grained diagnosis by applying the results of subtyping research. As all PD subjects
501 included in this study were newly diagnosed, and the classifier was trained and tested on base-
502 line data, it could be applied in clinical practice when evaluating a new PD patient. Additionally,
503 we would like to highlight that the model was trained on baseline data to predict a class de-
504 fined by disease progression, which involves the passage of time. Notably, it has a greater
505 ability to predict a subject's future compared to traditional PIGD/TD subtyping. This prediction
506 holds particular relevance for individuals whose phenotype aligns with the S3 subtype, where
507 this classification is more prone to change over time.

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

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

520 **3.6 Strengths and limitations**

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

529 The vastness of the results tables from the pathway analyses hindered results manageability. As
530 a group of researchers, we did our best to read the results table and report noteworthy results,
531 yet it is to be disclosed that a complete and accurate report was unfeasible. As a comment
532 to this, we would like to speculate that future technological development may help with the
533 interpretation of High Throughput Sequencing data analysis results: Large Language Models
534 (LLM), such as ChatGPT [43], are showing increasingly better ability to handle textual data, and
535 may one day be well-suited to summarize and expose these kinds of results. Potential future
536 analysis of our results by means of such methods is encouraged, and full results tables can be
537 found in Supplementary Tables 1-2.

538 **4 Methods**

539 **4.1 Workflow overview**

540 Data from the PPMI database were used for both of the objectives of this study: (1) to identify
541 the transcriptomics characteristics of the disease progression subtypes, and (2) to train the
542 ML model aimed at evaluating the usefulness of gene expression data in predicting disease
543 subtypes at baseline. First, data were gathered and the cohort of study was defined, as de-
544 scribed in [section 4.2](#). RNA-Seq data were preprocessed ([section 4.3](#)) and then a differential
545 expression analysis was performed as described in [section 4.4](#) The resulting DEGs were further
546 analyzed through pathway analyses, as described in [section 4.5](#) Following cohort definition, the
547 ML classifier was trained to predict the cluster at baseline, as described in [section 4.6](#), then
548 its behaviour was investigated using XAI methods ([section 4.7](#)). R code used to perform data

549 analysis can be found on GitHub (https://github.com/217c/parkinson_subtypes_rnaseq,
550 accessed on 25 September 2023).

551 **4.2 Data**

552 Data used in this study were obtained from the Parkinson Progression Marker Initiative (PPMI)
553 [44]. PPMI is one of the most important ongoing studies of PD progression markers, collecting
554 data from multiple international sources and focusing on a diverse range of potential markers
555 for tracking the progression of PD, including demographics, clinical variables, imaging data,
556 cerebrospinal fluid, blood, DNA and, importantly to this study, RNA measures. The data were
557 downloaded from the LONI Image and Data Archive (IDA) in April 2022. The cohort of study
558 was defined using the PPMI Consensus Committee Analytic Dataset (RD: 2021-10-28). PPMI
559 inclusion criteria for PD subjects include: age ≥ 30 , Parkinson's disease diagnosis within the last
560 2 years, baseline Hoehn and Yahr Stage I–II, and no anticipated need for symptomatic treatment
561 within 6 months of baseline [44]. Healthy controls (HC) inclusion criteria will include individuals
562 without clinical signs suggestive of parkinsonism, no evidence of cognitive impairment, and
563 no first-degree relative diagnosed with PD. To be included in this study cohort, subjects must
564 have had a diagnosis of sporadic PD and available RNA-Seq data for multiple timepoints, as
565 found in the LongRNA Transcriptome Sequencing of PPMI Samples (B38) study (RD: 2021-04-
566 02). The PPMI RNA Sequencing Project has generated overview transcriptomics data from
567 raw sequencing reads of PPMI whole blood samples. The data were pre-analyzed and quality-
568 controlled from the PPMI group [45].

569 The definition of the sample for this study follows that described in [20]. Subjects that un-
570 derwent disease progression subtyping were included, along with all available HC subjects. In
571 brief, S1 starts with mild motor and non-motor symptoms, and motor impairment increases with
572 a moderate rate over time; S2 has moderate motor and non-motor symptoms at baseline, with
573 a slow progression rate; S3 has significant motor and non-motor deficits at baseline, and its
574 impairment progresses rapidly over time, thus accounting for a worse prognosis. The IDs of the
575 subjects assigned to disease progression subtypes were retrieved from [20]. To summarize,
576 data analysis was performed on those subjects that had RNA-Seq data available and that were
577 clustered into one PD subtype. This study cohort included a total number of 2085 RNA-Seq
578 records for 4 years of longitudinal measures (starting from baseline, with constant time interval

579 measures at 12 months) from 600 subjects (PD = 407, HC = 193) (S1 = 199; S2 = 52; S3 =
580 156).

581 **4.3 RNA-Seq data preparation**

582 To assess outliers, a Principal Component Analysis (PCA) was computed on variance stabilized
583 and transformed (namely, vst from DESeq2) expression data of the top 20000 genes, and data
584 points lying beyond the edges of the Highest Density Interval of the first principal component
585 were deemed as outliers. The threshold was set to 0.99, thus considering as outliers all obser-
586 vations outside the 99% CI [46]. A sex incompatibility check was performed to assess con-
587 tamination due to abnormal transcription using t-SNE and DBSCAN on gene expression data
588 from the following sex chromosome genes: *USP9Y*, *XIST*, *RPS4Y1*, *RPS4Y2*, *KDM5D*, *DDX3Y*.
589 Subjects whose samples had inconsistent clustering between sex in metadata and sex from
590 expression data were removed from the analysis (Supplementary Figure 1).

591 **4.4 Differential Expression Analysis**

592 Differentially expressed genes (DEGs) were identified using DESeq2 R library v1.38.3 to perform
593 a Likelihood ratio test (LRT). This experiment was designed as a time course analysis, thus
594 the full model including group, time, and their interaction, was compared to a reduced model
595 without the interaction. This analysis allowed us to identify those genes that at one or more
596 time points after time 0 showed a group-specific effect, thus excluding genes that moved up
597 or down in time in the same way in both groups. Each PD cluster was compared to the HC
598 group performing a separated LRT. For each comparison, DESeq2 automatically estimated size
599 factors based on the median ratio method, estimated dispersions, and performed the LRT for
600 negative binomial GLMs [47]. Correction for multiple testing was performed using the False
601 Discovery Rate (FDR) method, applying DESeq2's default threshold for adjusted p-value < 0.1.
602 Gene names and descriptions were retrieved using *g:Profiler* R package [48].

603 **4.5 Pathway analysis**

604 To further investigate the differences in gene expression we performed a pathway analysis using
605 *clusterProfiler* R library v4.6.2 [49]. An Over-Representation Analysis (ORA) was performed
606 on DEGs for all three comparisons on GO Biological Process (BP) domain, KEGG, and WikiPath-

607 ways databases. Not to limit our pathway analysis to DEGs sets, we chose to investigate path-
608 way modulation due to eventually small but coordinated changes in the expression levels of all
609 genes, thus performing a Gene Set Enrichment Analysis (GSEA) for all three comparisons on
610 GO BP, KEGG, and WikiPathways databases. To improve interpretability, GSEA results on GO
611 were reduced to semantically similar terms using `rrvgo` R library v1.10.0 [50].

612 **4.6 Machine Learning model for subtype prediction at baseline**

613 Data collected at the time of diagnosis (baseline) was used to predict the cluster, using a hier-
614 archical machine learning approach. In this approach, we train multiple classifiers in a hierar-
615 chical structure, where each classifier is responsible for a specific task. This approach is useful
616 here because the classification task can be broken down into simpler sub-tasks. As cluster 3
617 showed to be the most severe, the first step was to predict if the newly diagnosed PD subject
618 belonged to S3. If not, the second step aimed to predict whether the subject was from S1 or S2
619 (Figure 13).

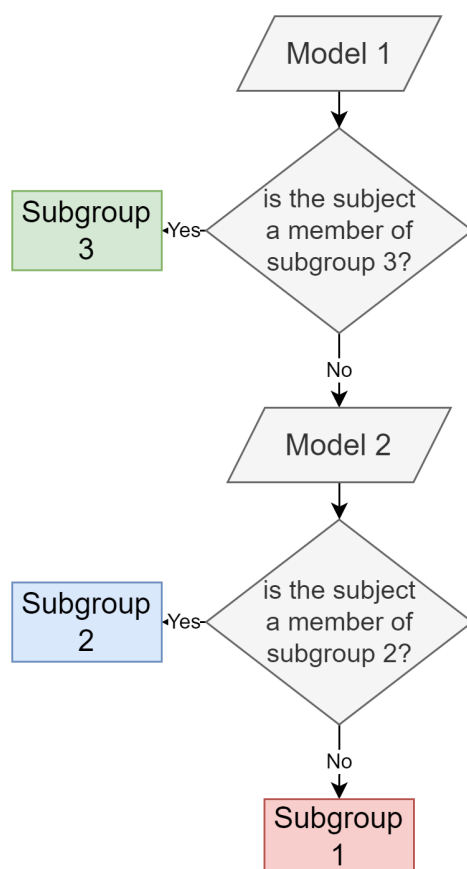


Figure 13: Schematic representation of the flow of the Hierarchical ML approach

620 Specifically, two XGBoost models were used in this pipeline [51]. Firstly, a subject is evaluated
621 by the first model in the hierarchy, which aims to identify subjects from S3. If the subject is
622 found to be from S3, the pipeline ends. If the subject is found not to be from S3, then the subject
623 is evaluated by the second model, which aims to discriminate between S2 and S1 subjects. The
624 machine learning pipeline was developed using `tidymodels` R library v1.0.0. Train test split
625 was performed at subject level, including 75% of the sample in the train set (Table 1). Data
626 from multiple modalities were used, including demographics, motor, non-motor, biospecimen,
627 imaging, and gene expression values (Table 2). Missing data were imputed with the mean value
628 of the train set and rounded to integer value, thus respecting the original format of variables.
629 All variables were transformed by applying a Box-Cox transformation [52] and feature selection
630 was performed by univariate filtering with ANOVA on all three groups. Variables reporting an
631 FDR-corrected p-value < 0.05 were selected for training. Variables with an absolute Pearson's
632 correlation value greater than 0.8 with other variables were removed. Synthetic minority over-
633 sampling technique (SMOTE) was used to address class imbalance before training [53]. The
634 XGBoost models were trained using 10 Cross-Validation resamples to find the best combination
635 of hyperparameters using a grid latin hypercube of values [54]. The best models resulting from
636 cross-validation were tested on the test set and evaluation metrics were computed.

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

637

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 - Auto- nomic 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 Cor- rect
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 Ques- tionnaire) Total...
NP3TOT	MDS-UPDRS	MDS-UPDRS Part III Total Score

Continued on next page

Variable Name	Extended Name	Description
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	RAD54B	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

639 **4.7 Variable importance and XAI**

640 The importance of variables in contributing to the Machine Learning prediction of subtype at
641 baseline was investigated using SHAP (SHapley Additive exPlanations) values [55]. As an XAI
642 method [56], SHAP values highlight the contribution of each feature to the final prediction,
643 thus providing a measure to rank features importance. To calculate SHAP values and produce
644 informative plots, `shapviz` R library functions [57] were applied to the XGBoost models.

645 **Data availability statement**

646 Data used in the preparation of this article were obtained from the Parkinson's Progression
647 Markers Initiative (PPMI) database

648 (www.ppmi-info.org/access-data-specimens/download-data), RRID:SCR_006431.

649 The PPMI IDs of the subjects in the disease subtypes were obtained from the GitHub repository
650 related to [20].

651 **Acknowledgements**

652 PPMI – a public-private partnership – is funded by the Michael J. Fox Foundation for Parkin-
653 son's Research and funding partners, including 4D Pharma, Abbvie, AcureX, Allergan, Amathus
654 Therapeutics, Aligning Science Across Parkinson's, AskBio, Avid Radiopharmaceuticals, BIAL,
655 Biogen, Biohaven, BioLegend, BlueRock Therapeutics, Bristol-Myers Squibb, Calico Labs, Cel-
656 gene, Cerevel Therapeutics, Coave Therapeutics, DaCapo Brainscience, Denali, Edmond J. Safra
657 Foundation, Eli Lilly, Gain Therapeutics, GE HealthCare, Genentech, GSK, Golub Capital, Handl
658 Therapeutics, Insitro, Janssen Neuroscience, Lundbeck, Merck, Meso Scale Discovery, Mis-
659 sion Therapeutics, Neurocrine Biosciences, Pfizer, Piramal, Prevail Therapeutics, Roche, Sanofi,
660 Servier, Sun Pharma Advanced Research Company, Takeda, Teva, UCB, Vanqua Bio, Verily, Voy-
661 ager Therapeutics, the Weston Family Foundation and Yumanity Therapeutics.

662 For up-to-date information on the study, visit www.ppmi-info.org.

663

664 Funding

665 This work was partially supported by Ricerca Corrente grants (Italian Ministry of Health) from
666 the Santa Lucia Foundation IRCCS—Linea di Ricerca A: Neurologia Clinica e Comportamen-
667 tale.

668 References

- 669 [1] E. Ray Dorsey et al. “Global, regional, and national burden of Parkinson’s disease, 1990–
670 2016: a systematic analysis for the Global Burden of Disease Study 2016”. English. In: *The*
671 *Lancet Neurology* 17.11 (Nov. 2018). Publisher: Elsevier, pp. 939–953. ISSN: 1474-4422,
672 1474-4465. DOI: [10.1016/S1474-4422\(18\)30295-3](https://doi.org/10.1016/S1474-4422(18)30295-3).
- 673 [2] WHO. *Parkinson disease: a public health approach: technical brief*. en. 2022.
- 674 [3] Bastiaan R. Bloem, Michael S. Okun, and Christine Klein. “Parkinson’s disease”. English.
675 In: *The Lancet* 397.10291 (June 2021). Publisher: Elsevier, pp. 2284–2303. ISSN: 0140-
676 6736, 1474-547X. DOI: [10.1016/S0140-6736\(21\)00218-X](https://doi.org/10.1016/S0140-6736(21)00218-X).
- 677 [4] Ronald B. Postuma et al. “MDS clinical diagnostic criteria for Parkinson’s disease”. en. In:
678 *Movement Disorders* 30.12 (2015). Number: 12_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/mds.26424>
679 pp. 1591–1601. ISSN: 1531-8257. DOI: [10.1002/mds.26424](https://doi.org/10.1002/mds.26424).
- 680 [5] Seyed-Mohammad Fereshtehnejad et al. “Clinical criteria for subtyping Parkinson’s dis-
681 ease: biomarkers and longitudinal progression”. en. In: *Brain* 140.7 (July 2017), pp. 1959–
682 1976. ISSN: 0006-8950, 1460-2156. DOI: [10.1093/brain/awx118](https://doi.org/10.1093/brain/awx118).
- 683 [6] Sara Riggare and Maria Hägglund. “Precision Medicine in Parkinson’s Disease - Explor-
684 ing Patient-Initiated Self-Tracking”. eng. In: *Journal of Parkinson’s Disease* 8.3 (2018),
685 pp. 441–446. ISSN: 1877-718X. DOI: [10.3233/JPD-181314](https://doi.org/10.3233/JPD-181314).
- 686 [7] Julia C. Greenland, Caroline H. Williams-Gray, and Roger A. Barker. “The clinical hetero-
687 geneity of Parkinson’s disease and its therapeutic implications”. en. In: *European Journal*
688 *of Neuroscience* 49.3 (2019). _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/ejn.14094>,
689 pp. 328–338. ISSN: 1460-9568. DOI: [10.1111/ejn.14094](https://doi.org/10.1111/ejn.14094).
- 690 [8] Kristen A Severson et al. “Discovery of Parkinson’s disease states and disease progres-
691 sion modelling: a longitudinal data study using machine learning”. en. In: *The Lancet*
692 *Digital Health* 3.9 (Sept. 2021), e555–e564. ISSN: 2589-7500. DOI: [10.1016/S2589-
693 7500\(21\)00101-1](https://doi.org/10.1016/S2589-7500(21)00101-1).

- 694 [9] R. Erro et al. “Comparing postural instability and gait disorder and akinetic-rigid sub-
695 typing of Parkinson disease and their stability over time”. en. In: *European Journal of*
696 *Neurology* 26.9 (Sept. 2019), pp. 1212–1218. ISSN: 1351-5101, 1468-1331. DOI: [10.1111/
697 ene.13968](https://doi.org/10.1111/ene.13968).
- 698 [10] Tanya Simuni et al. “How stable are Parkinson’s disease subtypes in de novo patients:
699 Analysis of the PPMI cohort?” en. In: *Parkinsonism & Related Disorders* 28 (July 2016),
700 pp. 62–67. ISSN: 13538020. DOI: [10.1016/j.parkreldis.2016.04.027](https://doi.org/10.1016/j.parkreldis.2016.04.027).
- 701 [11] Eduardo De Pablo-Fernández et al. “Prognosis and Neuropathologic Correlation of Clin-
702 ical Subtypes of Parkinson Disease”. en. In: *JAMA Neurology* 76.4 (Apr. 2019), p. 470.
703 ISSN: 2168-6149. DOI: [10.1001/jamaneurol.2018.4377](https://doi.org/10.1001/jamaneurol.2018.4377).
- 704 [12] P. A. Kempster et al. “Relationships between age and late progression of Parkinson’s
705 disease: a clinico-pathological study”. en. In: *Brain* 133.6 (June 2010), pp. 1755–1762.
706 ISSN: 0006-8950, 1460-2156. DOI: [10.1093/brain/awq059](https://doi.org/10.1093/brain/awq059).
- 707 [13] Andrew Siderowf et al. “Assessment of heterogeneity among participants in the Parkin-
708 son’s Progression Markers Initiative cohort using α -synuclein seed amplification: a cross-
709 sectional study”. en. In: *The Lancet Neurology* 22.5 (May 2023), pp. 407–417. ISSN:
710 14744422. DOI: [10.1016/S1474-4422\(23\)00109-6](https://doi.org/10.1016/S1474-4422(23)00109-6).
- 711 [14] Carlo Fabrizio et al. “Artificial Intelligence for Alzheimer’s Disease: Promise or Chal-
712 lenge?” en. In: *Diagnostics* 11.8 (Aug. 2021). Number: 8 Publisher: Multidisciplinary Digital
713 Publishing Institute, p. 1473. DOI: [10.3390/diagnostics11081473](https://doi.org/10.3390/diagnostics11081473).
- 714 [15] Genevie Borrageiro et al. “A review of genome-wide transcriptomics studies in Parkin-
715 son’s disease”. en. In: *European Journal of Neuroscience* 47.1 (Jan. 2018), pp. 1–16. ISSN:
716 0953816X. DOI: [10.1111/ejn.13760](https://doi.org/10.1111/ejn.13760).
- 717 [16] Seung Hyun Lee et al. “Parkinson’s Disease Subtyping Using Clinical Features and Biomark-
718 ers: Literature Review and Preliminary Study of Subtype Clustering”. en. In: *Diagnostics*
719 12.1 (Jan. 2022), p. 112. ISSN: 2075-4418. DOI: [10.3390/diagnostics12010112](https://doi.org/10.3390/diagnostics12010112).
- 720 [17] David W. Craig et al. “RNA sequencing of whole blood reveals early alterations in immune
721 cells and gene expression in Parkinson’s disease”. en. In: *Nature Aging* 1.8 (Aug. 2021).
722 Number: 8 Publisher: Nature Publishing Group, pp. 734–747. ISSN: 2662-8465. DOI: [10.
723 1038/s43587-021-00088-6](https://doi.org/10.1038/s43587-021-00088-6).

- 724 [18] Raphael T. Gerraty et al. “Machine learning within the Parkinson’s progression markers
725 initiative: Review of the current state of affairs”. In: *Frontiers in Aging Neuroscience* 15
726 (Feb. 2023), p. 1076657. ISSN: 1663-4365. DOI: [10.3389/fnagi.2023.1076657](https://doi.org/10.3389/fnagi.2023.1076657).
- 727 [19] Anant Dadu et al. “Identification and prediction of Parkinson’s disease subtypes and
728 progression using machine learning in two cohorts”. en. In: *npj Parkinson’s Disease* 8.1
729 (Dec. 2022), p. 172. ISSN: 2373-8057. DOI: [10.1038/s41531-022-00439-z](https://doi.org/10.1038/s41531-022-00439-z).
- 730 [20] Xi Zhang et al. “Data-Driven Subtyping of Parkinson’s Disease Using Longitudinal Clinical
731 Records: A Cohort Study”. en. In: *Scientific Reports* 9.1 (Jan. 2019), p. 797. ISSN: 2045-
732 2322. DOI: [10.1038/s41598-018-37545-z](https://doi.org/10.1038/s41598-018-37545-z).
- 733 [21] Sepp Hochreiter and Jürgen Schmidhuber. “Long Short-Term Memory”. en. In: *Neural
734 Computation* 9.8 (Nov. 1997), pp. 1735–1780. ISSN: 0899-7667, 1530-888X. DOI: [10.
735 1162/neco.1997.9.8.1735](https://doi.org/10.1162/neco.1997.9.8.1735).
- 736 [22] Chang Su et al. *Integrative analyses of multimodal clinical, neuroimaging, genetic, and
737 transcriptomic data identify subtypes and potential treatments for heterogeneous Parkin-
738 son’s disease progression*. en. preprint. *Neurology*, July 2021. DOI: [10.1101/2021.07.
739 18.21260731](https://doi.org/10.1101/2021.07.18.21260731).
- 740 [23] Sanjukta Krishnagopal et al. “Identifying and predicting Parkinson’s disease subtypes
741 through trajectory clustering via bipartite networks”. eng. In: *PLoS One* 15.6 (2020), e0233296.
742 ISSN: 1932-6203. DOI: [10.1371/journal.pone.0233296](https://doi.org/10.1371/journal.pone.0233296).
- 743 [24] Seyed-Mohammad Fereshtehnejad et al. “New Clinical Subtypes of Parkinson Disease
744 and Their Longitudinal Progression: A Prospective Cohort Comparison With Other Phe-
745 notypes”. In: *JAMA Neurology* 72.8 (Aug. 2015), pp. 863–873. ISSN: 2168-6149. DOI:
746 [10.1001/jamaneurol.2015.0703](https://doi.org/10.1001/jamaneurol.2015.0703).
- 747 [25] Tiago A. Mestre et al. “Parkinson’s Disease Subtypes: Critical Appraisal and Recom-
748 mendations”. In: *Journal of Parkinson’s Disease* 11.2 (Apr. 2021), pp. 395–404. ISSN:
749 18777171, 1877718X. DOI: [10.3233/JPD-202472](https://doi.org/10.3233/JPD-202472).
- 750 [26] Matthias Elstner et al. “Expression analysis of dopaminergic neurons in Parkinson’s dis-
751 ease and aging links transcriptional dysregulation of energy metabolism to cell death”.
752 en. In: *Acta Neuropathologica* 122.1 (July 2011), pp. 75–86. ISSN: 0001-6322, 1432-0533.
753 DOI: [10.1007/s00401-011-0828-9](https://doi.org/10.1007/s00401-011-0828-9).

- 754 [27] Ron Shamir et al. “Analysis of blood-based gene expression in idiopathic Parkinson dis-
755 ease”. en. In: *Neurology* 89.16 (Oct. 2017), pp. 1676–1683. ISSN: 0028-3878, 1526-632X.
756 DOI: [10.1212/WNL.0000000000004516](https://doi.org/10.1212/WNL.0000000000004516).
- 757 [28] Yanli Zhang et al. “Transcriptional analysis of multiple brain regions in Parkinson’s dis-
758 ease supports the involvement of specific protein processing, energy metabolism, and
759 signaling pathways, and suggests novel disease mechanisms”. en. In: *American Jour-
760 nal of Medical Genetics Part B: Neuropsychiatric Genetics* 137B.1 (Aug. 2005), pp. 5–16.
761 ISSN: 15524841, 1552485X. DOI: [10.1002/ajmg.b.30195](https://doi.org/10.1002/ajmg.b.30195).
- 762 [29] Daniel C. Berwick et al. “LRRK2 Biology from structure to dysfunction: research pro-
763 gresses, but the themes remain the same”. en. In: *Molecular Neurodegeneration* 14.1
764 (Dec. 2019), p. 49. ISSN: 1750-1326. DOI: [10.1186/s13024-019-0344-2](https://doi.org/10.1186/s13024-019-0344-2).
- 765 [30] C. Marras et al. “Phenotype in parkinsonian and nonparkinsonian LRRK2 G2019S muta-
766 tion carriers”. en. In: *Neurology* 77.4 (July 2011), pp. 325–333. ISSN: 0028-3878, 1526-
767 632X. DOI: [10.1212/WNL.0b013e318227042d](https://doi.org/10.1212/WNL.0b013e318227042d).
- 768 [31] Maoxin Huang et al. “Cell-Cell Communication Alterations via Intercellular Signaling
769 Pathways in Substantia Nigra of Parkinson’s Disease”. In: *Frontiers in Aging Neuro-
770 science* 14 (Feb. 2022), p. 828457. ISSN: 1663-4365. DOI: [10.3389/fnagi.2022.828457](https://doi.org/10.3389/fnagi.2022.828457).
- 771 [32] Krithi Irmady et al. “Blood transcriptomic signatures associated with molecular changes
772 in the brain and clinical outcomes in Parkinson’s disease”. en. In: *Nature Communications*
773 14.1 (July 2023), p. 3956. ISSN: 2041-1723. DOI: [10.1038/s41467-023-39652-6](https://doi.org/10.1038/s41467-023-39652-6).
- 774 [33] Lille Kurvits et al. “Transcriptomic profiles in Parkinson’s disease”. en. In: *Experimental
775 Biology and Medicine* 246.5 (Mar. 2021), pp. 584–595. ISSN: 1535-3702, 1535-3699.
776 DOI: [10.1177/1535370220967325](https://doi.org/10.1177/1535370220967325).
- 777 [34] Bingwei Lu, Stephan Gehrke, and Zhihao Wu. “RNA metabolism in the pathogenesis of
778 Parkinson’s disease”. eng. In: *Brain Research* 1584 (Oct. 2014), pp. 105–115. ISSN: 1872-
779 6240. DOI: [10.1016/j.brainres.2014.03.003](https://doi.org/10.1016/j.brainres.2014.03.003).
- 780 [35] Takao Ono et al. “Differential contributions of condensin I and condensin II to mitotic
781 chromosome architecture in vertebrate cells”. eng. In: *Cell* 115.1 (Oct. 2003), pp. 109–121.
782 ISSN: 0092-8674. DOI: [10.1016/s0092-8674\(03\)00724-4](https://doi.org/10.1016/s0092-8674(03)00724-4).
- 783 [36] Tahir N. Khan et al. “Mutations in NCAPG2 Cause a Severe Neurodevelopmental Syn-
784 drome that Expands the Phenotypic Spectrum of Condensinopathies”. eng. In: *Ameri-*

- 785 *can Journal of Human Genetics* 104.1 (Jan. 2019), pp. 94–111. ISSN: 1537-6605. DOI:
786 [10.1016/j.ajhg.2018.11.017](https://doi.org/10.1016/j.ajhg.2018.11.017).
- 787 [37] Qi Wang et al. “NCAPG2 could be an immunological and prognostic biomarker: From
788 pan-cancer analysis to pancreatic cancer validation”. In: *Frontiers in Immunology* 14
789 (Jan. 2023), p. 1097403. ISSN: 1664-3224. DOI: [10.3389/fimmu.2023.1097403](https://doi.org/10.3389/fimmu.2023.1097403).
- 790 [38] Jon Infante et al. “Identification of candidate genes for Parkinson’s disease through
791 blood transcriptome analysis in LRRK2-G2019S carriers, idiopathic cases, and controls”.
792 en. In: *Neurobiology of Aging* 36.2 (Feb. 2015), pp. 1105–1109. ISSN: 01974580. DOI:
793 [10.1016/j.neurobiolaging.2014.10.039](https://doi.org/10.1016/j.neurobiolaging.2014.10.039).
- 794 [39] Ester Pantaleo et al. “A Machine Learning Approach to Parkinson’s Disease Blood Tran-
795 scriptomics”. en. In: *Genes* 13.5 (Apr. 2022), p. 727. ISSN: 2073-4425. DOI: [10.3390/
796 genes13050727](https://doi.org/10.3390/genes13050727).
- 797 [40] Yura Jang et al. “Mass Spectrometry–Based Proteomics Analysis of Human Substantia Ni-
798 gra From Parkinson’s Disease Patients Identifies Multiple Pathways Potentially Involved
799 in the Disease”. en. In: *Molecular & Cellular Proteomics* 22.1 (Jan. 2023), p. 100452. ISSN:
800 15359476. DOI: [10.1016/j.mcpro.2022.100452](https://doi.org/10.1016/j.mcpro.2022.100452).
- 801 [41] Chadwick W. Christine et al. “Relationship of Cerebrospinal Fluid Vitamin B12 Status
802 Markers With Parkinson’s Disease Progression”. eng. In: *Movement Disorders: Official
803 Journal of the Movement Disorder Society* 35.8 (Aug. 2020), pp. 1466–1471. ISSN: 1531-
804 8257. DOI: [10.1002/mds.28073](https://doi.org/10.1002/mds.28073).
- 805 [42] Shuang Liu et al. “Gene expression profiling predicts pathways and genes associated
806 with Parkinson’s disease”. en. In: *Neurological Sciences* 37.1 (Jan. 2016), pp. 73–79. ISSN:
807 1590-1874, 1590-3478. DOI: [10.1007/s10072-015-2360-5](https://doi.org/10.1007/s10072-015-2360-5).
- 808 [43] OpenAI. *Introducing ChatGPT*. en-US. 2022.
- 809 [44] Kenneth Marek et al. “The Parkinson Progression Marker Initiative (PPMI)”. en. In: *Progress
810 in Neurobiology* 95.4 (Dec. 2011), pp. 629–635. ISSN: 03010082. DOI: [10.1016/j.
811 pneurobio.2011.09.005](https://doi.org/10.1016/j.pneurobio.2011.09.005).
- 812 [45] Elizabeth Hutchins et al. *Quality Control Metrics for Whole Blood Transcriptome Anal-
813 ysis in the Parkinson’s Progression Markers Initiative (PPMI)*. en. preprint. Genetic and
814 Genomic Medicine, Jan. 2021. DOI: [10.1101/2021.01.05.21249278](https://doi.org/10.1101/2021.01.05.21249278).

- 815 [46] Andrew Gelman and Sander Greenland. “Are confidence intervals better termed “uncer-
816 tainty intervals?”” en. In: *BMJ* (Sept. 2019), p. l5381. ISSN: 0959-8138, 1756-1833. DOI:
817 [10.1136/bmj.15381](https://doi.org/10.1136/bmj.15381).
- 818 [47] Michael I Love, Wolfgang Huber, and Simon Anders. “Moderated estimation of fold change
819 and dispersion for RNA-seq data with DESeq2”. en. In: *Genome Biology* 15.12 (Dec. 2014),
820 p. 550. ISSN: 1474-760X. DOI: [10.1186/s13059-014-0550-8](https://doi.org/10.1186/s13059-014-0550-8).
- 821 [48] Uku Raudvere et al. “g:Profiler: a web server for functional enrichment analysis and con-
822 versions of gene lists (2019 update)”. In: *Nucleic Acids Research* 47.W1 (July 2019), W191–
823 W198. ISSN: 0305-1048. DOI: [10.1093/nar/gkz369](https://doi.org/10.1093/nar/gkz369).
- 824 [49] Tianzhi Wu et al. “clusterProfiler 4.0: A universal enrichment tool for interpreting omics
825 data”. en. In: *The Innovation* 2.3 (Aug. 2021), p. 100141. ISSN: 26666758. DOI: [10.1016/
826 j.xinn.2021.100141](https://doi.org/10.1016/j.xinn.2021.100141).
- 827 [50] Sergi Sayols. “rrvgo: a Bioconductor package for interpreting lists of Gene Ontology
828 terms”. eng. In: *microPublication Biology* 2023 (2023). ISSN: 2578-9430. DOI: [10.17912/
829 micropub.biology.000811](https://doi.org/10.17912/micropub.biology.000811).
- 830 [51] Tianqi Chen and Carlos Guestrin. “XGBoost: A Scalable Tree Boosting System”. In: (2016).
831 Publisher: arXiv Version Number: 3. DOI: [10.48550/ARXIV.1603.02754](https://doi.org/10.48550/ARXIV.1603.02754).
- 832 [52] George EP Box and David R. Cox. “An analysis of transformations”. In: *Journal of the
833 Royal Statistical Society Series B: Statistical Methodology* 26.2 (1964). ISBN: 1369-7412
834 Publisher: Oxford University Press, pp. 211–243.
- 835 [53] N. V. Chawla et al. “SMOTE: Synthetic Minority Over-sampling Technique”. In: *Journal
836 of Artificial Intelligence Research* 16 (June 2002), pp. 321–357. ISSN: 1076-9757. DOI:
837 [10.1613/jair.953](https://doi.org/10.1613/jair.953).
- 838 [54] Delphine Dupuy, Céline Helbert, and Jessica Franco. “**DiceDesign** and **DiceEval** : Two R
839 Packages for Design and Analysis of Computer Experiments”. en. In: *Journal of Statistical
840 Software* 65.11 (2015). ISSN: 1548-7660. DOI: [10.18637/jss.v065.i11](https://doi.org/10.18637/jss.v065.i11).
- 841 [55] Scott M Lundberg and Su-In Lee. “A Unified Approach to Interpreting Model Predictions”.
842 In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30.
843 Curran Associates, Inc., 2017.
- 844 [56] David Gunning et al. “XAI—Explainable artificial intelligence”. en. In: *Science Robotics*
845 4.37 (Dec. 2019), eaay7120. ISSN: 2470-9476. DOI: [10.1126/scirobotics.aay7120](https://doi.org/10.1126/scirobotics.aay7120).
- 846 [57] Michael Mayer. *shapviz: SHAP Visualizations*. 2023.