

Metabolomic Signatures in Late-Onset Parkinson's Disease: A Cross-Sectional Analysis of Serum Biomarkers in the Tracking Parkinson's cohort



Natacha Loison
Wolfson College
University of Oxford

A thesis submitted for the degree of
MSc by Research in Psychiatry

Trinity term, 2025, Word count: 19,668

Supervised by Laura Winchester and Alejo Nevado-Holgado

Abstract

Parkinson’s disease (PD) is the second most common neurodegenerative disease, affecting 1% of the population over 60 years of age (Xiao et al. 2025). Currently, there is no single accepted diagnostic test for PD, and the need for accurate, earlier diagnostic biomarkers remains critical (Ben-Shlomo et al. 2024). Metabolomics offers a promising approach for biomarker studies, as the metabolome represents the final downstream product of biological processes and provides a detailed snapshot of an organism’s biochemical state (Clish 2015). This study analyzed baseline serum metabolomic data from the Tracking PD cohort (Malek et al. 2015), comprising 1,482 late-onset PD patients (disease duration <3.5 years, diagnosed ≥ 50 years) and 203 genetically-related unaffected sibling controls, with the aims of characterising metabolic alterations in PD serum samples, developing and validating metabolite-based classification models, and identifying specific metabolites and biological pathways most significantly implicated in PD pathology. Multiple analytical approaches were employed, including unsupervised clustering (Principal Component Analysis [PCA] and Uniform Manifold Approximation and Projection [UMAP]), generalised linear modelling (GLM), differential abundance analysis (Linear Models for Microarray Data [LIMMA]), and machine learning classification with SHapley Additive exPlanations (SHAP) feature importance analysis. Twenty-one metabolites were consistently altered across multiple statistical methods (LIMMA, GLM, SHAP), with fatty acid amides and tyrosine metabolism-associated metabolites playing central roles in distinguishing PD from healthy control metabolic profiles. Machine learning models achieved an area under the receiver operating characteristic curve (AUC) of 0.9884 for PD classification. Primary fatty acid amides (PFAMs) were elevated in both treated and untreated PD patients, suggesting these alterations reflect underlying disease mechanisms rather than treatment effects. These findings

identified widespread pathway-level disruptions in PD serum samples, with PFAMs representing a potential therapeutic target. Limitations of this project lie in the large class imbalance inherent to the samples, and future directions would be to further investigate PFAMs in other PD and related neurodegenerative cohorts.

Acknowledgements

I would like to thank my supervisors, friends and family for their enduring patience with me.

Contents

List of Tables	11
1 Introduction	14
1.1 Parkinson's disease	14
1.1.1 Pathology	14
1.1.2 Symptoms	15
1.1.3 Incidence and Prevalence	16
1.1.4 Treatment and Prevention	16
1.1.5 PD causes	16
1.2 Need for a Biomarker	18
1.2.1 Current Biomarkers in PD	20
1.3 Scope of this project	21
1.3.1 The Metabolome as a Biomarker	21
1.3.2 Serum as a biofluid	22
1.3.3 Current PD Serum Metabolite studies	22
1.3.4 Tracking Parkinson's Cohort	23
1.3.5 Samples and Subgroup Used in This Analysis	24
1.4 Aims	24
1.5 Treatment of Confounding Factors	25
2 Method	27
2.1 Data Collection and Generation	28
2.1.1 Metabolite Data	28
2.1.2 Clinical data	30
2.2 Data Preprocessing	31

Contents

2.2.1	Data Overview	31
2.2.2	Exclusion Criteria	32
2.3	Imputation	34
2.4	Clustering Analysis	37
2.4.1	Rationale	37
2.4.2	Principal Component Analysis (PCA)	38
2.4.3	Uniform Manifold Approximation and Projection (UMAP)	39
2.5	Linear Models	40
2.6	Rationale	40
2.6.1	LIMMA (Linear Models for Microarray Data) Statistical Framework	40
2.6.2	Generalized Linear Models (GLMs)	42
2.7	Classification	44
2.7.1	Algorithms and Setup	44
2.7.2	Pipeline	47
2.7.3	Pathway Enrichment	49
2.7.4	SHAP (SHapley Additive exPlanations) Analysis	51
2.7.5	Sensitivity Analysis	51
2.7.6	Metabolite Panel: Progressively adding feature from 1-5	54
2.7.7	Removing the Top 6 Amides and Top 4 L-DOPA Associated Molecules	54
2.8	Histogram Visualisation of the Top 5 SHAP Metabolites	55
2.9	Multi-Method Overlap Analysis	55
3	Results of Aim 1 Exploratory: Characterise metabolic alterations in PD serum samples	57
3.1	Data Overview	58
3.1.1	Sample Demographics	58
3.2	Metabolites Overview	60
3.3	Demographic Analysis: Results from Clustering	60
3.3.1	Principal Component Analysis (PCA)	60

Contents

3.3.2	UMAP	64
3.4	Differential Analysis: Results from LIMMA Modelling	68
3.5	Confounding Factors Analysis: Results from General Linear Model Analysis	69
3.5.1	Case-Control General Linear Models	69
3.5.2	LEDD General Linear Models	70
3.6	Summary	74
4	Results of Aim 2 Validation: Develop and validate a metabolite-based classification of PD	75
4.1	Developing PD Classification: Results from Machine Learning Pipeline	76
4.1.1	Data Split	76
4.1.2	Boruta Feature Selection Results	76
4.1.3	Evaluation of Model-Feature Combinations	77
4.2	Validating PD Classification: Results from best-performing algorithm	79
4.2.1	Group Sensitivity - Drug treatment as a confounding factor	81
4.3	Sensitivity Analysis	82
4.4	Summary	84
5	Results of aim 3 Mechanistic: Explore underlying disease pathways associated with metabolites most altered in PD samples	85
5.1	Pathway enrichment results	86
5.2	SHapley Additive exPlanations (SHAP) analysis	89
5.2.1	Misclassification - exploring what metabolites defined each subgroup	89
5.3	Metabolite Panel: Progressively Adding Features from 1-5	95
5.3.1	Group Sensitivity	96
5.4	Removing top metabolites - exploring model prediction dependence	98
5.4.1	Removing 4-LDOPA-associated metabolites	98
5.4.2	Removing 6 Fatty Acid Amide Metabolites	99

Contents

5.4.3	Both L-DOPA-associated metabolites and Fatty Acid Amides Removed	100
5.4.4	Summary	102
5.5	Histograms of the top 5 SHAP metabolites	103
5.6	Overlap results between Fold Change Analysis, GLM and SHAP feature importance	106
5.6.1	Three-way overlap metabolites	107
5.6.2	Super Pathway Analysis	110
5.6.3	Sub-Pathways	110
5.7	Summary	112
6	Discussion	113
6.1	Main findings	113
6.1.1	Widespread Metabolic Alterations Distinguish Parkinson’s Disease from Healthy Controls	113
6.1.2	Metabolic alterations persisted across drug treatment . . .	117
6.2	Biological and Mechanistic Insight	119
6.2.1	Fatty Acid Amide Pathways in PD	119
6.2.2	Tyrosine Metabolism in Dopamine Medication and PD . . .	122
6.2.3	Fibrinogen Cleavage Peptide	125
6.3	Strengths of the study	126
6.4	Limitations	127
6.5	Implications	130
6.6	Future directions	130
Appendices		
A	Appendix A	133
A.1	Subpathways provided by Metabolon	133
A.1.1	Super Pathways	133
A.1.2	Sub Pathways	134

Contents

A.2	Confounding Factors Analysis: Results from General Linear Model	
	Analysis	135
A.2.1	GLM1: PD $\sim$ metabolite	135
A.2.2	GLM4: LEDD $\sim$ metabolite	137
A.3	Removing Metabolites Analysis	139
A.3.1	Removing L-DOPA associated Metabolites	139
A.3.2	Removing top 6 Fatty Acid Amides	139
A.3.3	Removing Both	140
	Works Cited	142

List of Tables

3.1	Descriptive statistics for demographic and clinical characteristics of the study cohort.	59
3.2	Top 10 Differentially Expressed Metabolites from LIMMA Modelling Ranked by Adjusted P-Value	69
3.3	Top 10 metabolites from GLM 2: Case/control $\sim$ metabolite + sex + age (logistic regression), ranked by adjusted p -value	71
3.4	Top 10 metabolites from GLM 5: LEDD (Levodopa Equivalent Daily Dose) $\sim$ metabolite + gender + age (linear regression), ranked by adjusted p -value	73
4.1	Distribution of samples across groups in training and test sets . . .	76
4.2	Boruta feature selection results	76
4.3	Top 10 confirmed Boruta features ranked by Random Forest importance	77
4.4	Feature subset testing experimental design	78
4.5	Validation scores from nested cross-validation on all model combinations	78
4.6	Optimal model configuration and test set characteristics	79
4.7	Final SVM model performance on training and test sets	80
4.8	Final SVM detailed classification report	81
4.9	SVM performance metrics by patient group on the test set	81
4.10	Detailed confusion matrix showing PD subtypes (misclassifications highlighted)	82
4.11	Negative Control Tests: Model Performance on Random Data . . .	82
4.12	Model Performance on Balanced Dataset	83
4.13	Cross-Group Generalisation: Model Performance Across Different Patient Subgroups	83

List of Tables

4.14	Robustness Analysis: Stability of Model Performance and Feature Selection	84
5.1	Pathway enrichment analysis for significant metabolites in GLM 2: Top 5 super pathways	87
5.2	Subpathway enrichment analysis for GLM 2 (Case/control ~ metabolite + gender + age)	88
5.3	Top 20 metabolites ranked by SHAP (SHapley Additive exPlanations) contribution values from the machine learning model for Parkinson's disease prediction	90
5.4	Progressive Metabolite Addition Analysis Results	96
5.5	Group Sensitivity for Progressive Metabolite Addition, scored with Balanced Accuracy	97
5.6	Top 10 Features by Random Forest Importance from Boruta Feature Selection (No L-DOPA Data)	99
5.7	Top 10 Features by Random Forest Importance from Boruta Feature Selection (No Amides Data)	101
5.8	Top 10 Features by Random Forest Importance from Boruta Feature Selection (No L-DOPA + No Amides Data)	101
5.9	Performance Comparison of SVM Models with Different Feature Exclusions	102
5.10	Overlap of significantly altered metabolites across SHAP, GLM2, and LIMMA analyses.	106
5.11	Summary of metabolite rankings by super pathway. Average ranks across SHAP, GLM2, and LIMMA feature selection methods are shown for metabolites within each super pathway. Lower average ranks indicate higher importance.	110
5.12	Summary of metabolite rankings by sub-pathway. Average ranks across SHAP, GLM2, and LIMMA feature selection methods are shown for metabolites within each sub-pathway. Lower average ranks indicate higher importance.	111

List of Tables

A.1	Top 10 metabolites from GLM 1: Case/control $\sim$ metabolite (logistic regression), ranked by adjusted p -value	136
A.2	GLM 1: LEDD $\sim$ metabolite (linear regression)	138
A.3	SVM Classification Performance (No L-DOPA Data)	139
A.4	SVM Classification Performance (No Amides Data)	140
A.5	SVM Classification Performance (No L-DOPA + No Amides Data)	140

1

Introduction

Contents

1.1	Parkinson's disease	14
1.1.1	Pathology	14
1.1.2	Symptoms	15
1.1.3	Incidence and Prevalence	16
1.1.4	Treatment and Prevention	16
1.1.5	PD causes	16
1.2	Need for a Biomarker	18
1.2.1	Current Biomarkers in PD	20
1.3	Scope of this project	21
1.3.1	The Metabolome as a Biomarker	21
1.3.2	Serum as a biofluid	22
1.3.3	Current PD Serum Metabolite studies	22
1.3.4	Tracking Parkinson's Cohort	23
1.3.5	Samples and Subgroup Used in This Analysis	24
1.4	Aims	24
1.5	Treatment of Confounding Factors	25

1.1 Parkinson's disease

1.1.1 Pathology

Parkinson's disease (PD) is the second most common neurodegenerative disease after Alzheimer's (Xiao et al. 2025). PD is primarily associated with the deposition

1. Introduction

of toxic aggregated alpha-synuclein (α -synuclein) in various neuronal populations throughout the nervous system, leading to their eventual cell death; with severe losses of dopaminergic neurons in the substantia nigra (see Figure 1.1) (Morris et al. 2024). Well known for its intradisease heterogeneity, PD pathological changes can be found across multiple systems, including the autonomic nervous system, enteric nervous system, brainstem, limbic structures, and olfactory regions (Morris et al. 2024; Horsager et al. 2020; Beach et al. 2009).

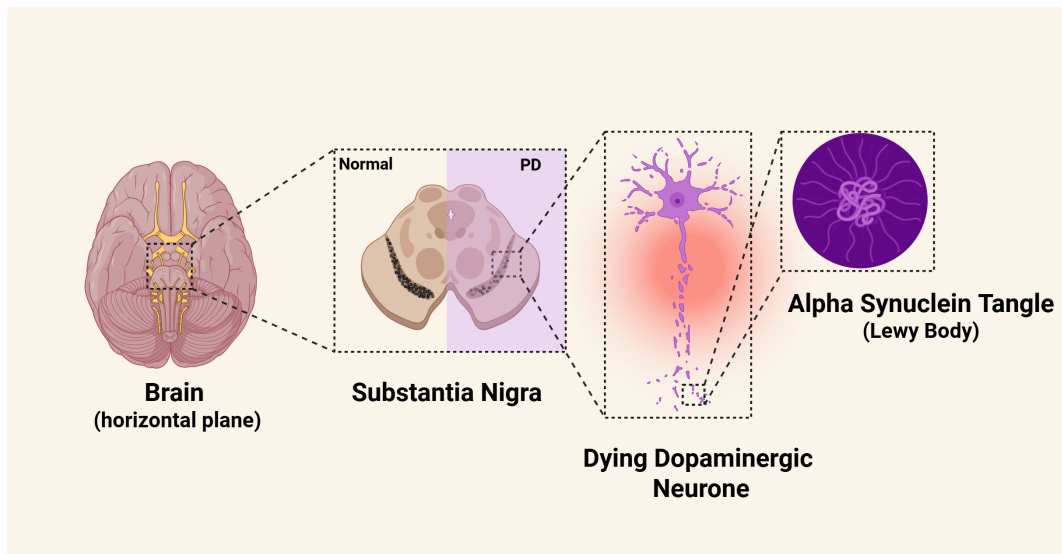


Figure 1.1: Neuropathological features of Parkinson's disease at multiple scales. Left: Horizontal section of the human brain showing the location of the substantia nigra (highlighted region) in the midbrain. Middle: Comparison of normal (left) and Parkinsonian substantia nigra, illustrating the characteristic loss of dopaminergic neurons and tissue degeneration. Right: Cellular pathology showing a dying dopaminergic neuron with abnormal alpha-synuclein protein tangles, which form Lewy bodies. Created in <https://BioRender.com>

1.1.2 Symptoms

Most commonly associated symptoms are motor-related, such as bradykinesia (slowness of movement), rigidity, gait disturbance and tremor. Less commonly associated yet still prevalent are non-motor symptoms, particularly in the prodromal stage of PD, including anxiety, depression, constipation, urinary and erectile dysfunction (autonomic). There is also a dementia like component of PD, found

1. Introduction

particularly in those with Parkinson’s with dementia, as well as in late-stage PD (Schrag et al. 2015).

1.1.3 Incidence and Prevalence

PD is reported to affect 1% of the population aged over the age of 60 (Xiao et al. 2025). Evidence for incidence increase is convoluted due to the lack of globally available statistics, but prevalence is increasing due to both the chronic nature of the disease and the increasing quality of care available (for both PD and co-morbidities). Inevitably, chronic neurodegenerative conditions such as PD place a heavy burden on the societal level (straining healthcare and economies) as well as on the personal level (with families acting as carers and reduced work capacity) (Ben-Shlomo et al. 2024).

1.1.4 Treatment and Prevention

Currently, PD has no disease-modifying treatment and no effective prevention program; there is a huge current and future patient base of people with great disability, who are without effective treatment (Foltynie et al. 2024). Although it should be noted that exercise prevention and intervention programs have shown promising results (Foltynie et al. 2024; Ben-Shlomo et al. 2024). Most interventions focus on treating symptoms of PD primarily by treating motor symptoms with dopamine derivatives (most commonly LDOPA) as well as pharmacologically, socially, and therapeutically treating non-motor symptoms (Foltynie et al. 2024).

1.1.5 PD causes

As the vast majority of PD cases are sporadic late onset , causes for PD are likely to be a convergence of several factors, including age, genetic risk, environmental factors and biological sex. By far the largest risk factor for neurodegenerative disease is age, and this holds for PD. While a small subset of PD patients will present with early-onset PD (diagnosis earlier than 50 years old), the majority of PD cases are seen in older populations, with the average age of onset being around 60 years (for late onset PD) (Hwang et al. 2025). Whether the neurone loss seen in PD is simply

1. Introduction

a sped-up process that occurs as an artefact of ageing or as a cumulative result of damage incurred from environmental stressors remains to be elucidated.

Parkinson's genetic heritability is set to be around 10-30% (Ben-Shlomo et al. 2024; Xiao et al. 2025), with well-established monogenic genes including: α -synuclein (SNCA), leucine-rich repeat kinase 2 (LRRK2), and vacuolar protein sorting 35 ortholog (VPS35) (autosomal dominant); and parkin RBR E3 ubiquitin protein ligase (PRKN), PTEN-induced kinase 1 (PINK1), and Parkinsonism-associated deglycase (DJ-1) (autosomal recessive) (Morris et al. 2024). The most common single rare variant in PD is glucocerebrosidase (GBA1), accounting for 5–10% of all cases (Lim et al. 2024). These genes have been shown to impact multiple cell functions, including mitochondrial health, lysosomal activity and protein aggregation.

While PD has a substantial genetic component, research has also highlighted the role that environment plays in PD. Most commonly cited are environmental toxins from herbicides, pesticides and organochlorines, with people living close to or working with these chemical agents showing a higher rate of PD (Ben-Shlomo et al. 2024). Specific lifestyle choices that decrease PD risk are smoking, a diet high in fruits and grains, and tea and coffee intake. Explaining the relationship between smoking status and reduced PD risk has been controversial. It's unclear whether it's directly causative (i.e., nicotine potentially has a protective effect on dopaminergic neurones) or correlational (i.e., genes associated with smoking liability also happen to be associated with decreasing PD risk) (Larsson and Burgess 2022). Dietary factors can be more easily understood as the gut is hypothesised to be the initial location of PD pathogenesis; therefore, it can be extrapolated that improving gut health would be protective to enteric neurones, ergo reducing PD risk (although multiple confounding factors affect these findings (Kim et al. 2018; Simon et al. 2020)). Men are more likely to have PD than women, in a 1.4:1 ratio (Wooten et al. 2004). This ratio is well preserved across different populations; however, there is a reversal in those with a LRRK2-associated PD with the Gly2019Ser mutation, where women are more likely to have PD (W. Chen et al. 2020).

1. Introduction

1.2 Need for a Biomarker

There is currently no single accepted diagnostic test for PD (Cannas et al. 2025), and considering the majority of PD patients are sporadic late-onset patients, there is a need for accurate and earlier diagnostic tests. Currently, PD is diagnosed primarily through clinical observation and diagnostic questionnaires. While these can achieve a high rate of accuracy (up to 84–90% for non-experts, up to 97% for experts) (Virameteekul et al. 2023), often by the time PD patients are diagnosed, they have lost more than 60% of their dopamine-producing neurones in their substantia nigra (Zhou et al. 2021). Neurones are post-mitotic cells and therefore do not replace themselves after death, leading to permanent damage. Therefore being able to diagnose the prodromal period of PD is of great importance.

While specialised centres and the use of the updated PD diagnostic criteria have increased PD diagnosis accuracy over the last decade (Virameteekul et al. 2023) in comparison to older cohorts (Rizzo et al. 2016), early PD has many overlapping symptoms with other forms of Parkinsonism (an umbrella term for disorders that share PD-like motor symptoms but have a different underlying pathology). This includes progressive supranuclear palsy, multiple system atrophy, corticobasal degeneration, dementia with Lewy bodies, and vascular Parkinsonism. As well as other motor disorders such as essential tremor (Shin et al. 2022).

Disease progression and staging also suffer from a lack of consensus, with no universally accepted framework for how Parkinson’s disease should be staged. Popular clinical methods include the Hoehn and Yahr scale, which classifies PD into five stages based primarily on motor symptom severity and bilateral involvement (Clarke et al. 2016), and the Movement Disorder Society-Unified Parkinson’s Disease Rating Scale (MDS-UPDRS), which provides a more comprehensive assessment encompassing motor and non-motor symptoms (Movement Disorder Society Task Force on Rating Scales for Parkinson’s Disease 2003). The Braak staging system offers a pathological perspective, proposing six stages based on the anatomical

1. Introduction

spread of Lewy body (alpha-synuclein) pathology through the brain in a caudal-to-rostral pattern (Braak et al. 2003).

Furthermore, there is currently no sub-typing in PD within routine clinical settings under the National Health Service (NHS) in the UK (United Kingdom of Great Britain and Northern Ireland) (NHS 2017). However, biologically, it has been well defined that PD shows large intra-disease heterogeneity (Morris et al. 2024). Patients show variable degrees of α -synuclein burden in the peripheral nervous system, and the site of initial α -synuclein aggregation differs between patients (e.g. olfactory, brainstem, enteric initiation) (Just et al. 2022). Patterns of neuronal loss within the brain are also heterogeneous. From a symptomatic perspective, patients will also show variable degrees of motor and non-motor symptoms as well as variable rates of disease progression, further supporting the argument of intradisease heterogeneity (Morris et al. 2024).

In addition, there is currently no established screening program for PD, and underdiagnosis remains common — particularly in older adults, with one screening study identifying undiagnosed cases in up to 36% of individuals aged 80–85 years (de Rijk et al. 1997). While clinical observation is of great value and accuracy, particularly when differentiating PD from other forms of Parkinsonism, mass screenings done via clinical observation and diagnostic questionnaire would be an immensely time-consuming task. Biological diagnostic tests on the other hand would be both quicker and more feasible as a screening process, particularly with easy, routine biofluids such as serum, plasma and urine (Yamashita et al. 2023).

There has also been a shift in the field towards looking for a biological definition of PD, with two proposed systems being released the same year (2024). This included Neuronal alpha-Synuclein Disease Integrated Staging System (NSD-ISS), which is a PD staging system based on α -synuclein pathology and functional impairment with a focus on aiding the creation of more homogeneous trial populations and specificity in diagnosing PD (Dam et al. 2024). On the other hand, the SynNeurGe classification framework uses three components (α -synuclein pathology, neurodegeneration, and genetic variants) that include both synuclein-positive and synuclein-negative forms

1. Introduction

of Parkinson’s disease. This framework focuses on capturing the largest amount of heterogeneity within the current defined population of PD patients, adapted to be used in a research context (Höglinger et al. 2024). While these frameworks are not utilised in this project due to them being relatively new additions to the field, they clearly demonstrate a desire to move away from a solely symptomatic diagnosis of PD and towards a more holistic, biomarker-driven diagnosis of PD.

1.2.1 Current Biomarkers in PD

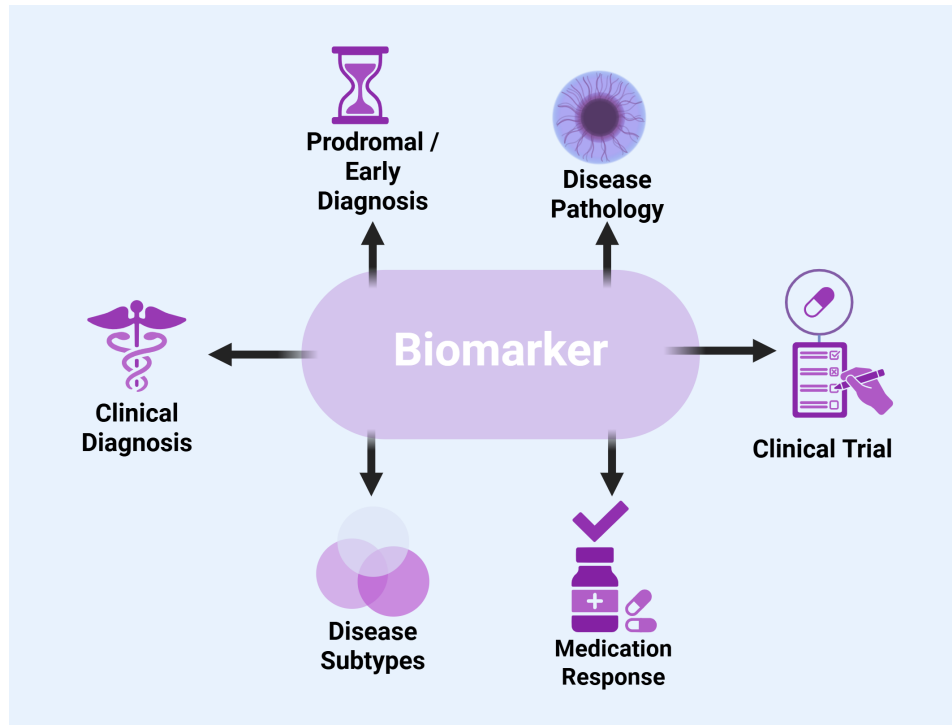


Figure 1.2: Clinical applications of biomarkers in Parkinson’s disease. Created in <https://BioRender.com>.

Biomarkers are measurable biological indicators of disease that serve multiple clinical and research purposes, including improving diagnostic accuracy, enabling earlier detection, elucidating disease pathology, classifying disease subtypes, predicting therapeutic response, and stratifying patients for clinical trials (see figure 1.2). Currently, the only biomarkers used in the NHS include brain scans (dopamine transporter single-photon emission computed tomography, DAT SPECT; magnetic resonance imaging, MRI) to rule out other Parkinsonian conditions (NHS 2017).

1. Introduction

However, recent research has produced many new potential PD biomarkers within the following modalities: plasma, serum, cerebrospinal fluid (CSF), and genetics. As well as imaging modalities such as structural and neuromelanin-sensitive magnetic resonance imaging (MRI), dopamine transporter single-photon emission computed tomography (DAT SPECT), and 18F-DOPA positron emission tomography (F-DOPA PET) (Yamashita et al. 2023). Particularly promising biomarkers include CSF α -synuclein seed amplification assays, achieving a sensitivity of 88% for PD patients and 96% specificity for healthy controls ($N = 1,123$, across eight countries) (Siderowf et al. 2023). As a validated biomarker, it has been used in both NSD-ISS and SynNeurGe frameworks (Dam et al. 2024; Höglinger et al. 2024). Limitations of this biomarker lie in the painful procedure of CSF collection via lumbar puncture, as well as its drop in sensitivity to 78% for PD patients without olfactory deficits (99% for those with olfactory deficits) and poor sensitivity for those with LRRK2-associated PD (Yamashita et al. 2023). As the scope of this project focuses on metabolic serum biomarkers for PD, further potential non-metabolic, non-serum biomarkers for PD can be found in the following reviews (Yamashita et al. 2023; Rangwala et al. 2025; Das and Ramteke 2025).

1.3 Scope of this project

1.3.1 The Metabolome as a Biomarker

The metabolome can be defined as the complete profiling of all metabolites (low-molecular-weight molecules, $< 1.5\text{kDa}$) produced by an organism at any given time (Clish 2015). Able to provide a detailed snapshot of the biochemical state of an organism, it can be considered the omic closest to the phenotyp. Metabolites represent the final downstream products of genomic, transcriptomic, and proteomic processes, directly reflecting the physiological state and response to genetic variations, environmental factors, and disease processes (Cannas et al. 2025), as illustrated in schematic figure 1.3. To note, this figure is a simplification of the complex cross talk between these systems and is largely for illustrative purposes.

1. Introduction

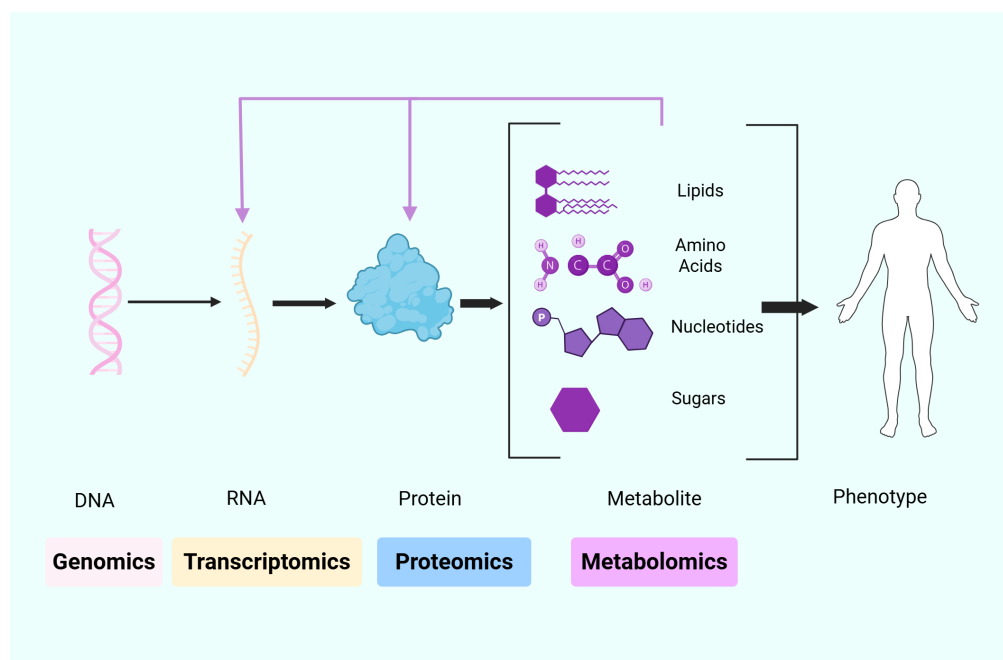


Figure 1.3: Schematic overview of the multi-omics cascade from genotype to phenotype. This simplified representation illustrates how genomic information is transcribed to RNA, translated to proteins, and metabolised to produce diverse metabolite classes (lipids, amino acids, nucleotides, and sugars), which collectively contribute to the observable clinical phenotype. Each layer represents a distinct omics approach for biomarker identification. Created in <https://BioRender.com>

1.3.2 Serum as a biofluid

Serum remains a popular biofluid due to the routine nature of blood sampling and the relative ease of mass spectrometry analysis (Morris et al. 2024). The serum metabolome reflects multiple factors, including the resting metabolic state of the individual, dietary intake, medication usage, hormonal status, age, and lifestyle choices (Clish 2015).

1.3.3 Current PD Serum Metabolite studies

Metabolite profiling of PD patients has shown, with great consensus, that there are metabolic differences between PD participants and healthy controls (Cannas et al. 2025; Luo et al. 2024). However, many of the metabolic changes are not reproduced in other studies, and some are in opposing directions (Luo et al. 2024). Part of the issue stems to (i) small sample sizes for metabolic studies, in a review by Cannas et al. (2025), they looked at papers from 2019-2024 and they stipulated

1. Introduction

that a study must include 50 PD patients to be included, which left them with 16 studies out of 200, (ii) difference in data collection, (i.e. mass spectrometry compared to NMR), data processing, and statistical handling and (iii) intra disease heterogeneity which is not addressed in patient stratification, (iv) lack of consensus how to handle confounding factors such as sex, age, co-morbidities, drug treatment, diet and (v) a distinct lack of longitudinal metabolite studies within PD.

However, some metabolite panels have managed to achieve very high Area Under the Curve scores, suggesting their clinical utility. For example, Dahabiyeh et al. (2024) achieved an AUC score of 0.977 from a seven-metabolite set, including cysteine-S-sulfate, 1-methylxanthine, vanillic acid, N-acetylaspartic acid, 3-N-acetyl tryptophan, 5-methoxytryptophol, and 13-hydroxyoctadecadienoic acid, namely 13-HODE. When following the several reviews (Cannas et al. 2025; Luo et al. 2024), key metabolic pathways that are highlighted include: amino acids, lipids, purine metabolism, energy metabolism, gut-derived microbacterial pathways and neurotransmitter metabolism (including drug-related dopamine metabolism).

A point that is highlighted by almost all reviews and studies in metabolomics is the idea that a metabolic biomarkers are not supposed to be implemented as the sole biomarker omics for PD. Rather, it is to be combined with other omics, such as genomics, transcriptomics, lipidomics, and proteomics, as each layer provides a different biological snapshot (Cannas et al. 2025). By mapping these omic changes together, not only does it increase accuracy, but it can also provide mechanistic information about PD. It would also help to address PD heterogeneity, mapping a genetic difference to a transcriptomic difference to a proteomic difference, which then relates to a metabolic and lipidomic differences.

1.3.4 Tracking Parkinson's Cohort

The cohort used in this thesis is the Tracking Parkinson's cohort; a multicentre, prospective, longitudinal study across 72 sites throughout the UK (Malek et al. 2015). It aims to define clinical variation in Parkinson's disease across different genotypes. The study recruited 2247 PD cases comprising 1987 recent onset patients

1. Introduction

(disease duration <3.5 years) and 260 young onset patients (diagnosed <50 years), along with a target of 840 unaffected siblings for comparison. Genetic siblings were used as to provide control samples with similar environmental influences as the disease samples. If a sibling control was found to become affected (with PD), their data was removed from the cohort. Genetic screening for LRRK2, GBA, PARK2, and PINK1 mutations was conducted, plus genome-wide genotyping using Illumina arrays. Serum samples are collected and stored in aliquots for future proteomic and metabolite analysis at specified intervals (study entry only for young onset, every 18 months for recent onset, every 3 years for siblings). The study collects clinical metadata, including motor assessments (MDS-UPDRS, Hoehn & Yahr, motor subtypes, LEDD), non-motor features (sleep, autonomic function, olfaction, impulse control), cognitive status (MoCA), quality of life (PDQ-8), plus detailed demographic data, family history, co-morbidities, medication history, and neuroimaging results where available. In this thesis, genetic data, demographic data, family history data, quality of life data and co-morbidity data was unavailable for analysis.

1.3.5 Samples and Subgroup Used in This Analysis

This project specifically analyses baseline serum metabolomic data from the recent-onset subgroup (disease duration <3.5 years), restricted to patients with late-onset disease (diagnosed ≥ 50 years of age, $n=1685$) compared to their genetically-related unaffected sibling controls ($n=203$). This selection excludes the young-onset cohort (diagnosed <50 years) and focuses only on serum samples from the initial visit. The choice to focus only on late onset sporadic patients was driven by the fact it remains as the most common type of PD (Ben-Shlomo et al. 2024). Since this is the initial exploration of the metabolic data, so the scope was restricted to baseline data.

1.4 Aims

This is the initial exploration of this metabolic data from the Tracking PD cohort, the aims are as follows:

1. Introduction

(i) **Exploratory: Characterise metabolic alterations in PD serum samples**

The first aim is to determine whether the metabolomic data demonstrate significant alterations between healthy control (HC) and PD samples. This analysis will also explore how the data are influenced by potential confounding factors, including sex, age, and pharmacological treatment.

(ii) **Validation: Develop and validate a metabolite-based classification of PD**

The second aim will approach the data using alternative statistical methods to determine whether PD and control serum samples can be discriminated via classification models. Classification performance may indicate the viability of metabolites for diagnostic prediction within this cohort. This analysis will also examine the confounding influence of drug treatment and assess whether the classifier maintains robust performance despite its presence.

(iii) **Mechanistic: Explore underlying disease pathways associated with metabolites most altered in PD samples**

The third aim is to identify which specific metabolites and biological pathways are most significantly implicated in this dataset, and to evaluate whether these alterations can be used to determine disease status and highlight any underlying mechanisms.

1.5 Treatment of Confounding Factors

As noted by Cannas et al. (2025), the treatment of confounding variables is frequently inadequately reported in metabolomic studies, despite their known influence on Parkinson’s disease metabolomic data. Key confounders identified include sex, age, disease onset type (early versus late), pharmacological treatments, genetic background, body mass index (BMI), fasting state at sample collection, and comorbidities. The present study addresses and controls for the following confounding variables: sex, age, late-onset disease presentation, and PD-specific drug treatment. It should be acknowledged that additional confounding factors

1. Introduction

remain unaccounted for due to data unavailability (even if collected in the initial cohort), representing a recognised limitation of this investigation.

2

Method

Contents

2.1	Data Collection and Generation	28
2.1.1	Metabolite Data	28
2.1.2	Clinical data	30
2.2	Data Preprocessing	31
2.2.1	Data Overview	31
2.2.2	Exclusion Criteria	32
2.3	Imputation	34
2.4	Clustering Analysis	37
2.4.1	Rationale	37
2.4.2	Principal Component Analysis (PCA)	38
2.4.3	Uniform Manifold Approximation and Projection (UMAP)	39
2.5	Linear Models	40
2.6	Rationale	40
2.6.1	LIMMA (Linear Models for Microarray Data) Statistical Framework	40
2.6.2	Generalized Linear Models (GLMs)	42
2.7	Classification	44
2.7.1	Algorithms and Setup	44
2.7.2	Pipeline	47
2.7.3	Pathway Enrichment	49
2.7.4	SHAP (SHapley Additive exPlanations) Analysis	51
2.7.5	Sensitivity Analysis	51
2.7.6	Metabolite Panel: Progressively adding feature from 1-5	54
2.7.7	Removing the Top 6 Amides and Top 4 L-DOPA Associated Molecules	54
2.8	Histogram Visualisation of the Top 5 SHAP Metabolites	55
2.9	Multi-Method Overlap Analysis	55

2.1 Data Collection and Generation

The data collection and generation were undertaken by Professor Grosset, Metabolon (Inc.), and the authors of the Tracking Parkinson’s cohort study (Malek et al. 2015; Ryals et al. 2007).

2.1.1 Metabolite Data

Data Collection

Tracking Parkinson’s collected blood samples from control and PD cases at baseline across 72 sites in the UK upon entry to the study. All sites were either part of the UK National Health Service (NHS) or academic institutions linked to NHS sites (Malek et al. 2015).

Sample Preparation

For the metabolite branch of this study, a total of 1903 blood serum samples were sent to Metabolon (Inc.) for analysis. Sample preparation follows Metabolon’s standard proprietary protocol (Metabolon 2025a). First, internal recovery standards were added for quality control. Then, proteins were precipitated to release bound metabolites and to enrich for small molecules. Finally, the resulting serum metabolite extracts were then divided into fractions for analysis.

Untargeted Mass Spectrometry

Metabolite analysis was conducted by untargeted mass spectrometry (Metabolon 2025a). Samples were analysed using ultra-performance liquid chromatography coupled with tandem mass spectrometry (UPLC–MS/MS) across four methods. Two methods employed reverse-phase UPLC–MS/MS with positive electrospray ionisation, which detects molecules that easily form positive ions. These include

2. Method

hydrophobic and polar molecules (Y. Wang et al. 2015). One method used reverse-phase UPLC–MS/MS with negative electrospray ionisation, which detects molecules that easily form negative ions (Y. Wang et al. 2015). And a final method used hydrophilic interaction chromatography (HILIC) UPLC–MS/MS with negative electrospray ionisation, which detects polar molecules that aren’t as easily detected in the reverse-phase method (Gowda and Djukovic 2014). The combination of all four methods allowed for a broad range in metabolite detection.

Quality Control

Pooled replicates were used to assess technical variation; process blanks were used to check for background signals; and internal standards were used to monitor extraction and instrument performance. Variability was measured both at the instrument level and across the entire process. To minimise potential batch effects, samples were analysed in a randomised order, with quality control checks interspersed at regular intervals (Metabolon 2025a).

Metabolite Identification and Pathway Labelling

Data processing involved using automated peak detection, spectral deconvolution, and several layers of quality control, following approaches outlined on Metabolon’s webpages (Metabolon 2025a). Metabolite identification was performed in three stages: matching retention time, confirming mass within ± 10 ppm, and finally comparing MS/MS spectra against a comprehensive reference library (containing over 5,400 authenticated standards and approximately 7,000 recurrent unknown standards). Pathway analysis and further identification draw both on public resources (HMDB, KEGG, PubChem, ChEBI, Lipid Maps) and Metabolon’s own in-house databases (Metabolon 2025b).

Data Processing and Normalisation

Chromatographic peaks were quantified using the area under the curve method. To reduce variation introduced by running samples across multiple days, a correction algorithm was applied. The algorithm centred metabolite intensities to the daily

2. Method

median (set to 1.00) and scaled individual sample values accordingly, helping correct for instrument drift. Where needed, additional normalisation factors, such as osmolality, were included to account for differences in sample loading (Metabolon 2025a).

Data Missingness, Imputation and Transformation

In untargeted metabolomics, missing values usually arise because metabolite concentrations fall below the instrument’s detection limit (missing not at random), rather than from random technical errors (missing completely at random) or from data processing errors (missing at random). To retain statistical power, imputation was carried out by replacing missing values with the minimum observed value for each metabolite after batch normalisation. Then a natural log transformation was applied to the data, as metabolite data typically (except 0 values and very low values) displays a log-normal distribution, which is necessary for some statistical testing (Metabolon 2025a).

It is worth noting that for metabolites with a high proportion of missing values, replacing them all with the same minimum value can distort the distribution, leading to an artificial peak at the detection threshold and straying from the expected log-normal distribution (Faquih et al. 2020). Since this was the case for this dataset, a different method of imputation was used, which is discussed in the section ‘Data Overview’ below.

2.1.2 Clinical data

Tracking Parkinson’s collected clinical data at outpatient clinics for all PD cases at baseline. The following clinical variables are used in this analysis.

- Age: Continuous variable. Age (in years) of the participant when enrolled on the study.
- Sex: Categorical variable, male or female.
- Site: Categorical variable. Recruitment site for each participant.

2. Method

- Disease duration: Continuous variable. The time (in years) between Parkinson’s disease diagnosis and age at enrolment.
- LEDD: Continuous variable. Levodopa equivalent daily dose (LEDD) is a standardised measure of the total daily intake of dopaminergic medication.
- MoCA (adjusted for years of education): Continuous variable. The Montreal Cognitive Assessment (MoCA) was adjusted for education level. MoCA provides a global measure of cognition (0-30, higher scores indicate better performance). For interpretability, the scores are often grouped into categories (26 or above = normal, 21-25 = mild cognitive impairment, 20 or below = dementia) (Aiello et al. 2022).
- UPDRS Part III: Continuous variable. The motor sub score (part III) of the Unified Parkinson’s Disease Rating Scale (UPDRS), ranging from 0 (no motor symptoms) to 132 (severe motor impairment) (Movement Disorder Society Task Force on Rating Scales for Parkinson’s Disease 2003).

2.2 Data Preprocessing

While initial processing was performed using Metabolon’s proprietary pipeline, the following stages of data preprocessing were conducted by the author.

2.2.1 Data Overview

While exploring the initial dataset (batch-corrected, imputed and transformed metabolite data), it was observed that some metabolites showed a peak at very low values, greatly skewing their distribution; despite the natural log transformation applied (see Figure 2.1). This issue was unsurprising due to the type of imputation used (in section ‘Data Missingness, Imputation and Transformation’). Since the project’s plan involved statistical exploration, it was important that the data approximately follow a normal distribution. Therefore, the imputation was redone using the batch-corrected, unimputed and untransformed dataset provided by Metabolon.

2. Method

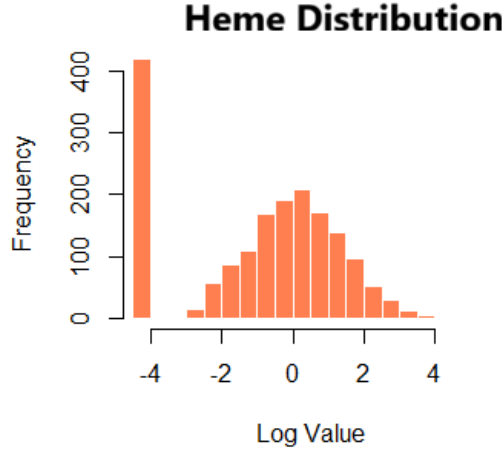


Figure 2.1: Histogram showing natural log-transformed heme values after missing data imputation using minimum observed values. The spike at low values (log -4) represents imputed data points, while the bell-shaped distribution is the unimputed data.

2.2.2 Exclusion Criteria

The scope of the project was determined to be late-onset PD patients, and the decision to redo the data's imputation resulted in the following exclusion criteria being used.

Exclusion Criteria for Samples (See Figure 2.1):

- Duplicate samples from one participant were excluded: to ensure all samples were independent.
- PD participants who were diagnosed with PD under the age of 50 were excluded: as they are classed as early-onset patients.
- Participants under the age of 50 at study entry (control and PD) were excluded: To age-match controls to late-onset PD participants, as age is a great source of variance in blood metabolite profiles. PD participants are forcibly already over the age of 50 due to the criteria above.
- Participants who had missing values for more than 50% of the recorded metabolites were excluded: This cut-off point was chosen as imputing with too few variables, the sample's data cannot be accurately estimated.

2. Method

- Participants who had their drug data (LEDD values) missing were excluded due to analyses relying on drug treatment status as a covariate.

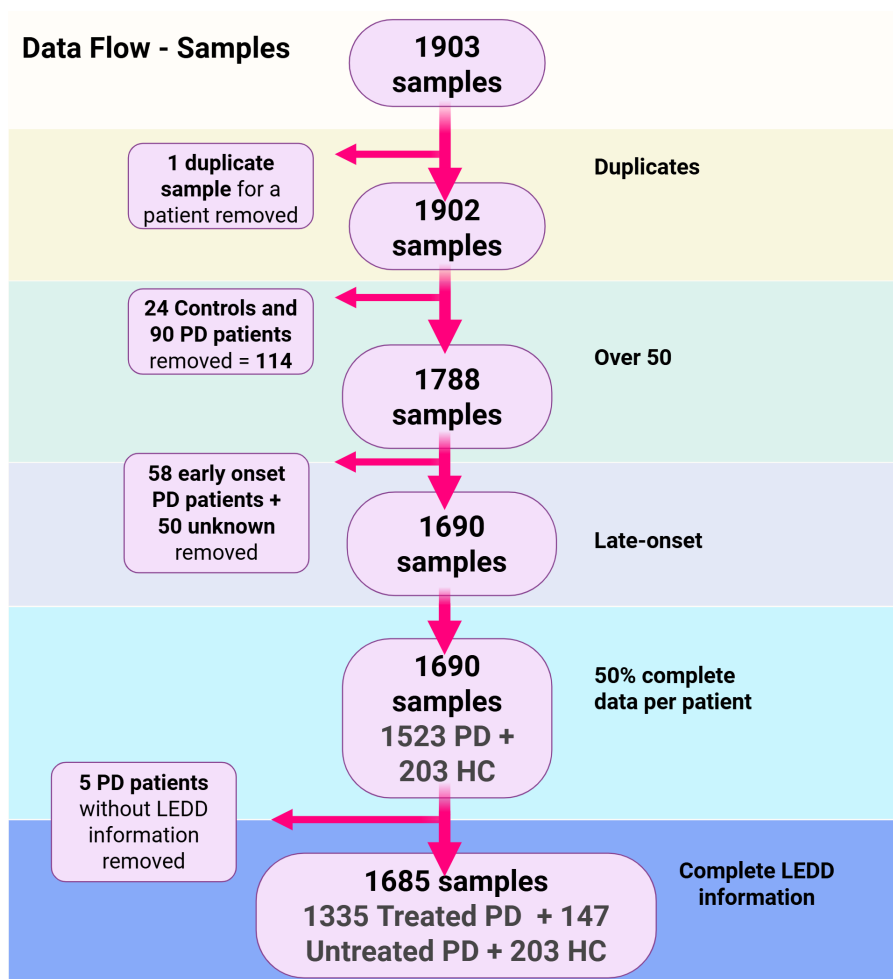


Figure 2.2: Data flow showing sample filtering for PD metabolomics study. From 1903 initial samples, sequential exclusions for duplicates, age (<50 years), and early-onset PD, and missing drug information (LEDD) resulted in 1685 samples (1482 PD + 203 controls) for analysis.

Exclusion Criteria for Metabolites (See Figure 2.3):

- Metabolites that were missing values for more than 80% of recorded samples were excluded: This is a standard cut-off point for imputation in metabolomics (Smilde et al. 2005).

2. Method

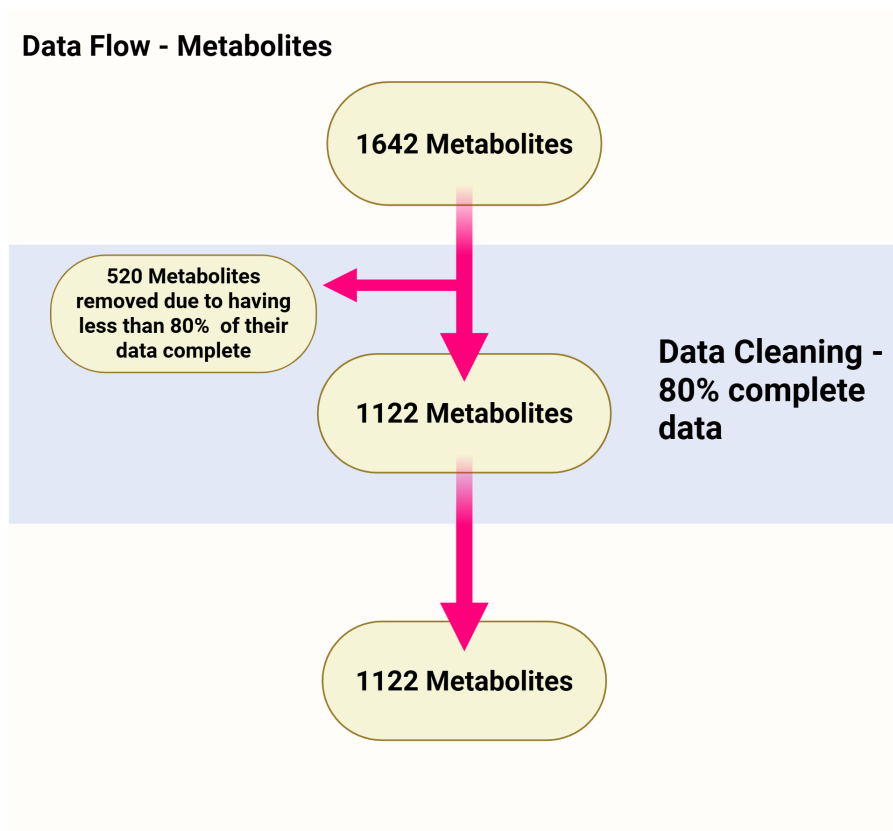


Figure 2.3: Data flow showing metabolite filtering based on data completeness. Of 1642 initial metabolites, 520 were removed for having <80% complete data, leaving 1122 metabolites for analysis.

The resulting dataset after this exclusion criteria was a dataset of 1,685 samples with 1,122 metabolites per sample.

2.3 Imputation

Imputation was carried out on the dataset described above, after applying the exclusion criteria (for reference: this dataset is batch-corrected and unimputed). A variety of imputation methods for metabolomic data were consulted when considering which method would be best for this project's dataset (Green et al. 2022; Do et al. 2018; Faquih et al. 2020; Wei et al. 2018).

2. Method

Transformation

A log10 transformation was applied to the dataset prior to imputation in accordance with established metabolomics preprocessing protocols (Wei et al. 2018; Do et al. 2018; Faquih et al. 2020). While the literature demonstrates variability in transformation timing—with some studies applying transformations post-imputation (Green et al. 2022) — the differences in analytical outcomes are generally considered negligible (Krutkin et al. 2025). Recent discourse in the field has questioned the appropriateness of log transformations for untargeted metabolomics data, particularly given the prevalence of low-abundance metabolites (due to the limits of detection). Power transformations have been proposed as alternative approaches to address these limitations (Sun and Xia 2023). However, to maintain methodological consistency with the selected imputation protocol and ensure reproducibility of results, the log10 transformation was retained.

Imputation Method

Some commonly used methods for metabolomics data imputation include: multiple imputation by chained equations (MICE), random forest (RF), and k-nearest neighbour (KNN) (Ou et al. 2025; Wei et al. 2018; Green et al. 2022; Faquih et al. 2020; Do et al. 2018). When deciding on a method, one must consider several factors; including dataset size, the proportion of missing data, the correlation structure among variables, and the computational resources available (Faquih et al. 2020). This project’s dataset can be considered medium-sized with 1,685 samples with 1,122 metabolites. Missingness thresholds were set at no more than 20% per metabolite and 50% per sample, resulting in relatively low overall missingness. Given the inherent correlations between metabolites, imputation methods that exploit these inter-variable relationships tend yield more accurate results (Do et al. 2018). Amongst the three methods, KNN is the least computationally intensive, followed by RF, with MICE being the most demanding (although all three methods tend be similarly intensive when used on extremely large datasets) (Faquih et al. 2020).

2. Method

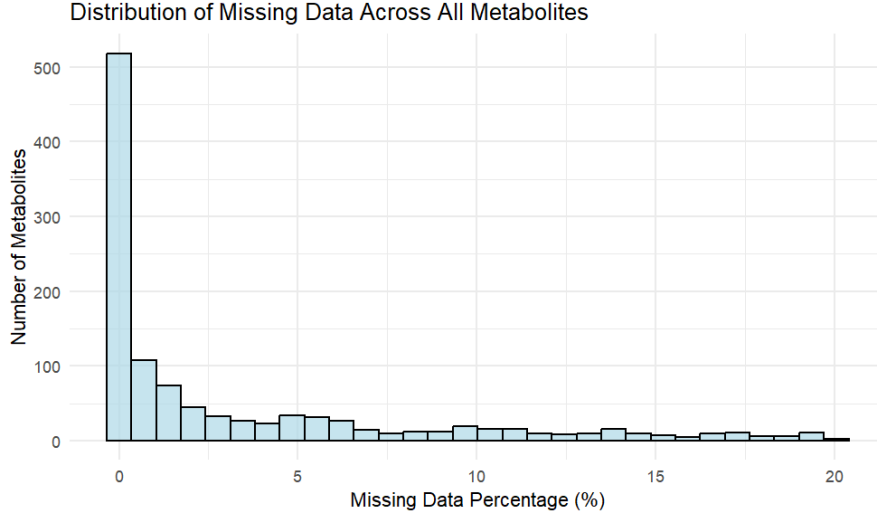


Figure 2.4: Frequency distribution of missing data percentages across metabolites. Most metabolites have minimal missing data (0-5%).

Given our dataset size, relatively low missingness (see Figure 2.4) and the need for efficient computation, KNN imputation was selected. KNN's also utilises 'neighbours' to find correlated metabolites when imputing data which is ideal. Do et al. evaluated different variants of KNN for metabolomics data: KNN-vars: finds metabolites with similar patterns to the missing one and uses their values, KNN-obs: finds samples with similar metabolite profiles and uses their values for the missing metabolite, KNN-obs-vars: first identifies related metabolites, then finds similar samples within this subset to impute the missing value. KNN-obs-vars consistently outperformed the others, preserving statistical power while maintaining a low type I error rate (reducing the likelihood of false positive associations) (Do et al. 2018). Based on these findings, the KNN-obs-vars approach was adopted.

The analysis was undertaken in R version [4.4.1], adapting R code from (Do et al. 2018). Given the absence of a dedicated R package for KNN-obs-sel, Do et al. (2018) developed a custom implementation of the KNN algorithm with base R functions. This project adapted the code from Do et al. (2018) with minor modifications; the algorithm was constructed using base R functions: `dist()`, `scale()`, `cor()`, and `order()`. The function's general structure is as follows: for each metabolite with missing values, 5–10 auxiliary metabolites were selected based on correlation strength

2. Method

($|r| \geq 0.2$), and missing values were imputed using a weighted average of the 10 nearest neighbour (samples) observations, with exponential distance weighting (e^{-d}).

Performance Evaluation

Performance evaluation involved visually comparing pre- and post-imputation histograms in 9 randomly selected metabolites with 10–20% missingness. At the dataset level, performance was evaluated by:

- using the Kolmogorov–Smirnov test to compare pre- and post-imputation distributions (Massey 1951),
- comparing metabolite medians before and after imputation using dot plots with overlaid medians in different colours,
- assessing global structure using Principal Component Analysis (PCA), and
- conducting t -tests on means, medians, and standard deviations ($\alpha = 0.05$) (Wei et al. 2018).

All tests indicated that the imputed data closely resembled the original data, with no evidence of systematic distortion.

2.4 Clustering Analysis

2.4.1 Rationale

Dimensionality reduction and clustering techniques were employed to explore the underlying structure within the metabolomics dataset and identify potential patient subgroups (Santos et al. 2024). This analysis had two objectives: (1) quality control assessment—clustering patterns were examined alongside clinical variables to determine whether groupings reflected known patient characteristics or technical artifacts, and (2) whether disease and control metabolic phenotypes were different enough to be separated via clustering.

2. Method

PCA was employed to assess global data structure and identify linear relationships, while UMAP was utilised to capture non-linear patterns and preserve local neighbourhood structure.

Clinical variables used to colour plots: disease status (PD vs control), age, gender, cognitive scores (MoCA), motor symptoms (UPDRS III), medication load (LEDD), and study site.

2.4.2 Principal Component Analysis (PCA)

Principal Component Analysis is a linear dimensionality reduction technique that transforms high-dimensional data into a lower-dimensional space while preserving maximum variance. PCA operates by computing the eigenvectors and eigenvalues of the covariance matrix, where each principal component represents an orthogonal direction of maximum variance in the data. The first principal component captures the largest amount of variance, the second captures the most remaining variance orthogonal to the first, and so forth (Groth et al. 2013). Principal Component Analysis was performed using the `prcomp()` function from base R version [4.4.1] (stats package) (RDocumentation 2025).

Data Preprocessing

PCA is particularly sensitive to outliers and scaling of the data, as it is reliant on variance as a measure of informativeness from features. Our data has already been both scaled and outliers have been checked for (Groth et al. 2013). Data preprocessing included centering (`center = TRUE`) and scaling (`scale = TRUE`) to ensure variables were on comparable scales prior to analysis.

Parameters and Software Implementation

All analyses were conducted in R version [4.4.1], and a random seed of 42 was set using `set.seed(42)` to ensure reproducibility of results.

2. Method

2.4.3 Uniform Manifold Approximation and Projection (UMAP)

Metabolomics data often exhibit non-linear relationships between metabolites and therefore these relational clusters may be better captured by UMAP, which is able to process non-linear structures (whereas PCA relies on a combination of linear variables) (Yang et al. 2021; Töpfer et al. 2015).

UMAP has gained significant traction in genomics initially but is now also used for other omics studies such as proteomics and metabolomics (Rather and Chachoo 2023), as it is able to preserve both local neighborhood structures (within cluster relationships) and global data topology (relationships between clusters). The algorithm constructs a weighted graph representation of the data in high-dimensional space, then optimizes a low-dimensional embedding that preserves the graph structure (Ghojogh et al. 2021). The following analysis was done with the umap package [0.2.10.0] from CRAN (Konopka 2025)..

Data Preprocessing

The same steps for preprocessing were undertaken for UMAP as for PCA.

Parameters and Software Implementation

UMAP's behaviour is primarily controlled by two parameters (Ghojogh et al. 2021):

- **n_neighbors**: Controls the number of neighbouring points considered when learning the local structure. Higher values preserve more global structure, while lower values emphasize local neighbourhood relationships.
- **min_dist**: Determines the minimum distance between points in the low-dimensional embedding, affecting cluster separation. Lower values create tighter, more separated clusters, while higher values produce more even distributions.

In this analysis default values for UMAP were used (**n_neighbors** = 15, **min_dist** = 0.1) with Euclidean distance as the metric (**metric** = "euclidean") (Konopka 2025). The analysis was implemented using the umap package version [umap

2. Method

0.2.10.0] in R version [4.4.1]. A random seed of 42 was set using `set.seed(42)` to ensure reproducibility (Ghojogh et al. 2021).

2.5 Linear Models

Linear models establish mathematical relationships between predictor variables (metabolite concentrations) and outcomes through linear combinations (Ip 2007).

2.6 Rationale

Linear models allow quantification of how changes in individual metabolites relate to disease status. In metabolomics-based disease diagnosis, linear models identify differentially abundant metabolites between patient groups, estimate effect sizes, and provide statistical significance testing to distinguish disease-associated metabolic signatures from normal variation (Klatt et al. 2021).

LIMMA was used to form volcano plots as a brief overview of differential analysis, whereas logistic regression models were used to explore metabolite changes more specifically while correcting for confounding factors.

2.6.1 LIMMA (Linear Models for Microarray Data) Statistical Framework

LIMMA is a statistical framework that fits linear models to each metabolite while employing empirical Bayes methods to borrow strength across features, thereby improving statistical power even at small sample sizes (Phipson et al. 2016). Originally developed for DNA microarray data, LIMMA has been successfully applied to RNA-seq data and metabolite intensities for identifying differentially expressed features across experimental conditions (Sánchez-Vinces et al. 2023). The analysis was conducted using the LIMMA package from R Bioconductor [3.23] (Ritchie et al. 2015).

2. Method

Data Preprocessing

Given that the metabolite mass spectrometry data was already $\log_{10}$ transformed during the preprocessing pipeline, the standard limma framework was implemented (without voom transformation (Law et al. 2014)). The log-transformed data approximated the normal distribution aids with many assumptions for LIMMA.

Methodology

The design matrix was constructed without an intercept (default) to model two groups: PD cases ($PD = 1$) and healthy controls ($PD = 0$). The specific contrast tested was “case - control”, comparing PD participants against healthy controls. LIMMA results were visualized using volcano plots to display the relationship between fold change and statistical significance across all metabolites.

Multiple Testing Correction

To address the multiple testing problem inherent in differential expression analysis, Benjamini-Hochberg false discovery rate (FDR) correction was applied. Statistical significance was defined as adjusted p-value < 0.05 and $|\log_2 \text{fold-change}| > 0.2$. These thresholds aim to correct for Type I error while still allowing patterns in the data to be observed.

Parameters and Software Implementation

Linear models were fitted using the following functions from the LIMMA package from R Bioconductor [3.23] (Ritchie et al. 2015). `lmFit()` with a design matrix specifying case-control group membership without an intercept term. Contrasts were defined using `makeContrasts()` to compare PD cases versus controls (case - control). Volcano plots were generated using the EnhancedVolcano package from Bioconductor [3.23].

2. Method

2.6.2 Generalized Linear Models (GLMs)

Generalized Linear Models (GLMs) provide a flexible statistical framework that extends ordinary linear regression to accommodate various types of outcome variables (Dobson 2002; Ip 2007). The core equation is given by:

$$y = X\beta + \varepsilon$$

where y represents the outcome variable, X denotes the predictor matrix, β are the coefficients of interest, and ε represents random error. GLMs utilize link functions to handle different outcome types, such as binary outcomes (logistic regression) and continuous outcomes (linear regression), within a unified framework. In metabolomics studies, GLMs are widely used due to their flexibility in modeling diverse outcome types and their straightforward implementation for testing associations between features (Klatt et al. 2021). Additionally, GLMs allow for the incorporation of confounding variables. In this study, logistic regression GLMs were employed to examine metabolite associations with Parkinson’s disease diagnosis (binary outcome), using R’s base function `glm()` (Rdocumentation 2025), from R [4.4.1].

Data Preprocessing

Parkinson’s disease (PD) status was encoded as 0 = healthy control and 1 = PD diagnosis. To ensure compatibility with GLM, which discards rows with missing data, missing Levodopa Equivalent Daily Dose (LEDD) values for controls were set to 0.

Methodology

Univariate Screening Approach A systematic univariate screening approach was implemented, fitting separate GLMs for each metabolite independently. The central question addressed was whether the abundance of a given metabolite was significantly altered between healthy controls and PD serum samples. For each model, the metabolite coefficient, standard error, test statistic, and p-value were extracted from the model’s summary output.

2. Method

Model Specifications Each metabolite was tested individually as a predictor of PD status across three model specifications, incorporating the following covariates:

- **GLM 1:** Case-Control PD diagnosis $\sim$ individual metabolite
- **GLM 2:** Case-Control PD diagnosis $\sim$ individual metabolite + age + gender
- **GLM 3:** Case-Control PD diagnosis $\sim$ individual metabolite + age + gender + LEDD
- **GLM 4:** LEDD $\sim$ individual metabolite
- **GLM 5:** LEDD $\sim$ individual metabolite + age + gender

Age and sex were included as covariates due to their well-known confounding effects in serum metabolomics (Darst et al. 2019). LEDD was added to isolate metabolites not directly linked to Levodopa (and equivalent drugs) metabolism, as these provide greater insight into the disease.

Multiple Testing Correction

To address multiple testing across hundreds of metabolites, the Benjamini-Hochberg false discovery rate (FDR) correction was applied to control Type I error. Significance levels were classified hierarchically based on FDR-adjusted p-values.

Parameters and Software Implementation

For logistic regression models (GLMs 1-3), the binomial family with logit link function was specified (`family = binomial()`).

Implementation Limitations

GLM 3 encountered convergence issues due to perfect separation, where LEDD values completely distinguished case-control status. Although a small number of drug-naïve PD patients also have an LEDD score of 0, all controls having an LEDD of 0 created perfect separation between the two classes. Consequently, covariates were unable to resolve drug correction. For the subsequent analysis, the results of GLM 2 (Case-Control PD diagnosis $\sim$ individual metabolite + age + gender) were used.

2.7 Classification

Classification is a supervised machine learning approach that assigns data samples to predefined categories based on learned patterns from training data (Xu et al. 2024; Liebal et al. 2020). In this project, binary classification was employed to distinguish between Parkinson’s disease patients and healthy controls using metabolomic abundances as input features. The classification process involves training algorithms on labelled metabolite data (training data) to identify discriminative patterns that can accurately predict disease status in unseen samples (testing data) (Galal et al. 2022).

2.7.1 Algorithms and Setup

Python Environment: Python was implemented in Visual Studio Code using Python version [v 3.13].

Cross Validation: The analysis utilized nested cross-validation with 5-fold stratified splits for both outer and inner loops. Nested cross-validation prevents selection bias by using separate data folds for hyperparameter optimization (inner loop) and performance evaluation (outer loop), ensuring that final performance estimates reflect true generalization ability rather than overfitting to the validation data used during model development (de Lope et al. 2024).

N_trials: Each model’s hyperparameter optimization used 100 Optuna trials with Tree-structured Parzen Estimator (TPE) sampling and median pruning for early trial termination. The number of trials is generally set to as high as computationally feasible (100 is a commonly chosen number) (LabOak 2025; Huckvale and Moseley 2024) and median pruning was added to aid in increasing n_trials without increasing computational burden (Optuna 2025).

Cores: All analyses utilized parallel processing with all available CPU cores to optimize computational efficiency

Class Imbalance: Due to class imbalance in the dataset, all algorithms were configured with balanced class weights or equivalent parameters. Performance metrics were monitored to assess the need for additional imbalance correction techniques, including upsampling, SMOTE, and bootstrap downsampling, which

2. Method

would be implemented if scores indicated insufficient class representation (very low test scores or the minority group having particularly low scores) (Fu et al. 2021).

Feature Selection

Boruta: Boruta feature selection uses an iterative random forest approach to identify features that perform significantly better than random noise variables (shadow features) (Labory et al. 2024). The implementation used the custom parameters of 500 estimators (to decrease computational burden) with balanced class weights (correct for class imbalance). Default parameters for the number of iterations (100) were chosen with default confidence threshold (100%) for feature confirmation (Learn 2025a).

Classification Algorithms

Logistic Regression with LASSO: Logistic Regression is a linear statistical method that models the probability of binary outcomes. LASSO (Least Absolute Shrinkage and Selection Operator) applies L1 regularisation to this framework, automatically performing feature selection by shrinking less important coefficients to zero (Hastie et al. 2009b). This is a common approach in metabolomics (Tong et al. 2020). Hyperparameter optimisation explored `C values` (0.001-10 on log scale), penalty set to 'l1', solver configured as 'liblinear' for L1 compatibility, and `max_iter` (100-2000) to ensure convergence. Class weights were set to `class_weight='balanced'` (Labory et al. 2024; Zhang et al. 2023; Learn 2025f).

Random Forest: Random Forest is an ensemble method that combines multiple decision trees through bootstrap aggregation to reduce overfitting and improve generalisation. A commonly used algorithm in metabolomics, it has been seen to do well (T. Chen et al. 2013; Mendez et al. 2019). Hyperparameter optimisation explored `n_estimators` (100-500) (Probst and Boulesteix 2017), `max_depth` (10, 15, 20, None), `min_samples_split` (2-20), `min_samples_leaf` (1-20), and `max_features` (sqrt, log2, None) (T. Chen et al. 2013). Class weights were set to `class_weight='balanced'`.

2. Method

Support Vector Machine: SVM finds an optimal hyperplane that maximizes the margin between classes, with kernel functions enabling non-linear classification (Learn 2025b). Another commonly used classification algorithm in metabolomics (Tong et al. 2020; Prajumwongs et al. 2025). The optimization tested linear, RBF, and polynomial kernels with `C values` (0.01-100), `gamma parameters` for non-linear kernels (1×10^{-4} to 10), and polynomial degrees (3, 5, 7) (Hastie et al. 2009b; Prajumwongs et al. 2025; Learn 2025b). Class weights were set to `class_weight='balanced'`.

XGBoost: XGBoost implements gradient boosting with regularization techniques and advanced tree pruning to prevent overfitting while maintaining high predictive performance (Hastie et al. 2009a). Not as commonly used in metabolomics, XGBoost was added as it has shown success in other omics such as proteomics (Behkamal et al. 2025). Parameters optimized included `n_estimators` (50–300), `max_depth` (3–10), `learning_rate` (0.01–0.3), `subsample` (0.6–1.0), `colsample_bytree` (0.3–1.0), and regularization terms `reg_α` and `reg_λ` (0–10) (Behkamal et al. 2025; Ardeshirifar 2024; XGBoost 2025). Class weights were set to `class_weight='balanced'`.

Hyperparameter Tuning

Optuna Hyperparameter Tuning: Optuna is a hyperparameter optimisation framework that uses Bayesian optimisation to efficiently explore parameter spaces through informed sampling rather than grid search (LabOak 2025). The implementation employed TPE sampling with median pruning (10 startup trials, 2 warmup steps) to terminate unpromising trials early, significantly reducing computational time while maintaining optimization quality (Optuna 2025).

Scoring Metrics

The following metrics are defined in greater detail in the following source (Learn 2025d) from the SciKit Learn python resource.

AUC: Area Under the ROC Curve measures the classifier’s ability to distinguish between PD and healthy controls across all classification thresholds, with values ranging from 0.5 (random performance) to 1.0 (perfect classification).

2. Method

ROC-AUC graphs: ROC curves plot true positive rate against false positive rate, providing visual assessment of classification performance and optimal threshold selection for clinical decision-making.

F1 Score: The harmonic mean of precision and recall, balancing the trade-off between correctly identifying PD cases and minimizing false positives, with values ranging from 0.0 (worst performance) to 1.0 (perfect precision and recall).

Balanced Accuracy: The average of sensitivity and specificity, providing equal weight to performance on both classes and accounting for potential class imbalances in the dataset, with values ranging from 0.0 (worst performance) to 1.0 (perfect classification of both classes).

2.7.2 Pipeline

2. Method

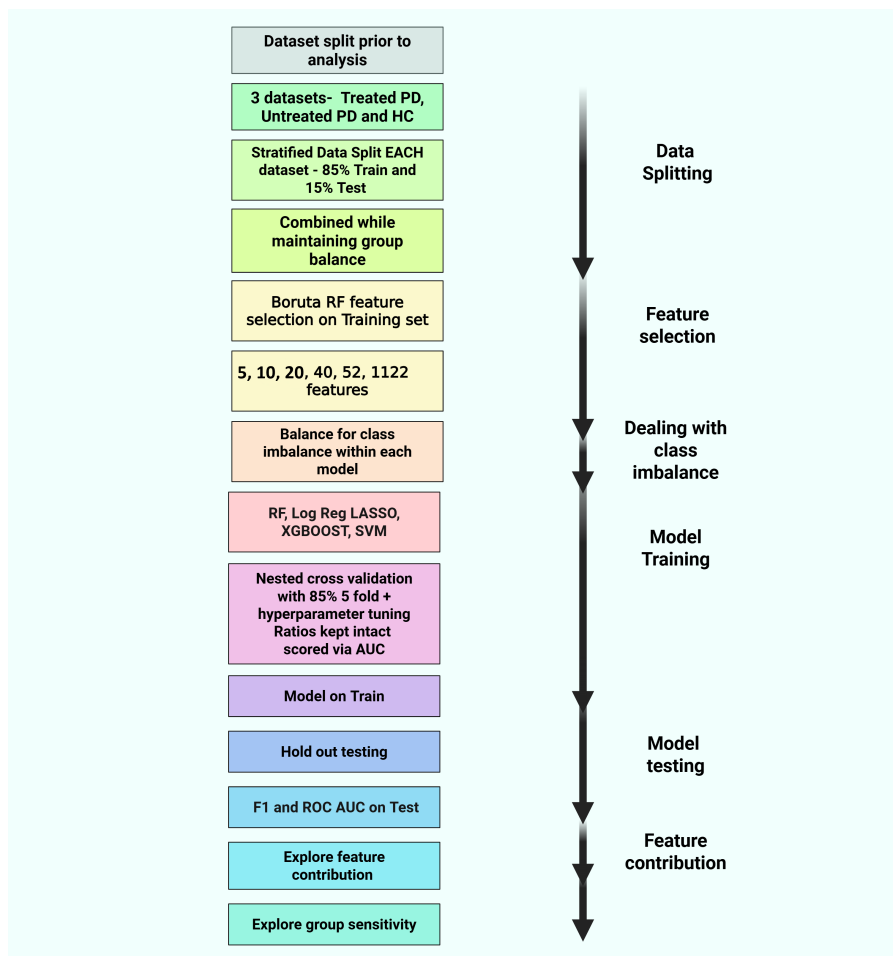


Figure 2.5: Overview of the machine learning classification pipeline, from balanced dataset splitting and feature selection to model training, validation, and interpretation of feature contributions. After feature selection, 52 confirmed features were found, hence numbers of features were staggered to 5, 10, 20 and 40. 1122 represents all features.

The following pipeline can be seen in Figure 2.5. Before the pipeline, the dataset was split into 3 datasets: PD treated (having a non 0 value for LEDD), consisting of 1335 samples, PD untreated (0 for LEDD), consisting of 147 samples and healthy controls, consisting of 203 samples.

1. Three datasets comprising untreated PD patients, treated PD patients, and healthy controls were individually stratified and split into training (80%) and test (20%) sets before being combined while maintaining group balance.
2. Boruta feature selection was applied to the combined training data using Random Forest to identify the most informative metabolomic features, which

2. Method

were subsequently ranked by importance scores.

3. Four machine learning algorithms (Random Forest, Lasso regression, Support Vector Machine, and XGBoost) were systematically evaluated across multiple feature subset sizes (5, 10, 20, 40 features, plus all confirmed features and the complete feature set) using nested 5-fold cross-validation with Optuna-based hyperparameter optimization using up to 100 trials.
4. The best-performing model-feature combination was selected based on cross-validation AUC scores, and consensus hyperparameters were derived from the optimal parameters across CV folds.
5. The final model was trained on the complete training set and evaluated on the holdout test set, with performance assessed through AUC, accuracy, F1-score, and confusion matrices.
6. Additional group-based sensitivity analysis examined model performance across patient subgroups, calculating clinically relevant metrics including sensitivity and specificity to assess the classifier’s ability to correctly identify both disease states and treatment effects.
7. All model artifacts, hyperparameters, and performance metrics were documented.

2.7.3 Pathway Enrichment

Rationale

A hypergeometric enrichment test was used to determine whether certain metabolic super-pathways contained more significantly altered metabolites than expected by chance from the results of the GLM analysis (Falcon and Gentleman 2008). This would allow for the evaluation of whether there were certain metabolic signatures that were broadly changed between disease and control. Metabolites that were significant (adjusted p value < 0.05) from the GLM2 (Case/control $\sim$ metabolite +

2. Method

sex + age, logistic regression)’s results were analysed in this section and only pathway *enrichment* was tested (one tailed test, pathway *depletion* was not investigated).

Definition

A one-sided hypergeometric test is equivalent to a one-sided Fisher’s exact test and is calculated using the equations defined below. Fold enrichment, calculated as the proportion of significant metabolites in a pathway divided by the proportion of all measured metabolites belonging to that pathway. This defines effect size, and only pathways with a fold enrichment greater than 1 were reported. P-values were Benjamini–Hochberg adjusted across all tested pathways and pathways with a p value < 0.05 were considered significant.

The following equations were used:

where N is the total number of metabolites in the background, K is the number of metabolites in the pathway of interest in the background, n is the total number of significant metabolites, k is the number of significant metabolites in the pathway of interest, and m is the number of pathways tested, with $p_{(1)} \leq \dots \leq p_{(m)}$ the ordered raw p-values.

Hypergeometric probability mass function

$$P(X = k) = \frac{\binom{K}{k} \binom{N-K}{n-k}}{\binom{N}{n}} \quad (2.7.3.1)$$

One-sided enrichment p-value (upper tail)

$$p = P(X \geq k) = \sum_{i=k}^{\min(n, K)} \frac{\binom{K}{i} \binom{N-K}{n-i}}{\binom{N}{n}} \quad (2.7.3.2)$$

Expected count under the null

$$E[X] = K \cdot \frac{n}{N} \quad (2.7.3.3)$$

Fold enrichment

$$\text{Fold Enrichment} = \frac{k/n}{K/N} \quad (2.7.3.4)$$

2. Method

Implementation

A script in R was produced using the equations defined above using packages tidyverse [2.0.0] and R base package stats from R [4.4.1].

2.7.4 SHAP (SHapley Additive exPlanations) Analysis

SHAP is an explainable post-hoc method for classification that quantifies each feature’s contribution to individual model predictions by calculating how much each feature shifts the output from a baseline expectation (Lundberg and S.-I. Lee 2017). SHAP analysis was performed on the optimal SVM model (40 features). The analysis used KernelExplainer (Ponce-Bobadilla et al. 2024) with 100 randomly sampled training instances as background and 500 maximum evaluations per test sample. Summary plots and misclassification plots of the most confidently mispredicted patients were produced.

2.7.5 Sensitivity Analysis

Sensitivity Analysis

Due to very high scores achieved by all the classifiers (including the optimal SVM model), a sensitivity analysis was conducted to evaluate the robustness and generalizability of the optimal SVM model. A comprehensive sensitivity analysis was conducted using multiple validation approaches. This analysis aimed to determine whether the observed model performance reflected genuine biological signal detection or potential artefacts such as overfitting, data leakage, or spurious correlations (Chuah et al. 2022).

Noise Testing

Gaussian noise testing was performed to establish baseline model performance on randomly generated data (Chuah et al. 2022). Synthetic datasets were created with identical dimensions to the original data (n =total samples from training and test sets, p =selected features) using normally distributed random variables ($\mu = 0$, $\sigma = 1$). Binary labels were assigned randomly with balanced class distribution (50% in each

2. Method

class). The trained SVM model, using identical hyperparameters and train-test split ratios as the original analysis, was evaluated on this noise data. This test validated whether the model could inappropriately detect patterns in pure noise, which would indicate potential overfitting or methodological issues (Chuah et al. 2022).

Label Shuffling Analysis

Label permutation testing assessed whether model performance depended on genuine biological associations between metabolite features and PD diagnosis (Ojala and Garriga 2010). The original metabolite feature matrix was retained while target labels (PD vs HC) were randomly shuffled, maintaining the original class distribution. This approach preserved the underlying data structure and feature correlations while eliminating any true biological signal. The SVM model was retrained and evaluated using the same methodology as the original analysis. Performance above chance levels on shuffled labels would suggest data leakage or inappropriate feature-target relationships (Ojala and Garriga 2010).

Cross-Group Generalisation Testing

Cross-group validation assessed model generalizability across PD patient populations using three scenarios: training on healthy controls with treated PD patients and testing on untreated patients (Scenario 1), training on controls with untreated patients and testing on treated patients (Scenario 2), and training on controls versus treated patients, then testing on controls versus untreated patients (Scenario 3). All scenarios used identical SVM hyperparameters and evaluation metrics (AUC, balanced accuracy, F1-score), with emphasis on balanced accuracy due to potential class imbalances.

Balanced Dataset Analysis

To address potential class imbalance effects (Owusu-Adjei et al. 2023; Salmi et al. 2024), balanced subsets were created by randomly sampling equal numbers of participants from each group (n=147 per group, representing the largest feasible balanced sample). This approach eliminated sample size disparities between groups

2. Method

that might confound model performance assessments. The SVM model was retrained and evaluated on these balanced datasets using an identical methodology to assess whether performance was influenced by group size differences rather than biological signal strength.

Hyperparameter Stability Assessment

Hyperparameter optimisation stability was evaluated by re-running Optuna optimisation (50 trials) on the balanced dataset using a different random seed. The Tree-structured Parzen Estimator sampler with median pruning was employed, maintaining identical optimisation parameters to the original analysis except for the random state. Optimal hyperparameters from this rerun were compared with the original optimal parameters to assess consistency of hyperparameter selection across different data configurations. Substantial differences would indicate unstable optimisation or data-dependent parameter selection (Learn 2025c).

Feature Selection Stability Testing

Boruta feature selection stability was assessed by re-running the algorithm using an alternative train-test split (`random_state=999` versus original `random_state=42`). All Boruta parameters remained identical (500 estimators, 100 iterations, 100% confidence threshold) to isolate the effect of different data splits on feature selection. This analysis validated whether the selected metabolite features represented robust biomarkers or were sensitive to specific data partitioning (Labory et al. 2024).

All sensitivity analyses employed consistent evaluation metrics, including area under the ROC curve (AUC), balanced accuracy, F1-score, and standard accuracy. Balanced accuracy was prioritised for subgroup analyses due to potential class imbalances. Model performance on sensitivity tests was compared against the original model performance and theoretical baseline expectations ($\text{AUC} \approx 0.5$ for random performance) to determine the extent to which observed performance reflected genuine biological signal versus methodological artefacts.

2. Method

2.7.6 Metabolite Panel: Progressively adding feature from 1-5

To determine the minimum number of metabolites required for accurate PD classification and assess the individual contribution of top-ranking features, a progressive feature addition analysis (Department of Electrical Engineering, Canadian University Dubai, Dubai, UAE et al. 2024) was conducted using the five most important metabolites identified through SHAP analysis. This approach systematically evaluated model performance when features were added cumulatively (1, 1+2, 1+2+3, 1+2+3+4, 1+2+3+4+5), using identical SVM hyperparameters and train-test splits as the original analysis.

The methodology recreated the complete dataset preprocessing pipeline to ensure comparability, then trained separate SVM models at each step using only the selected subset of metabolites. Performance was evaluated using AUC, balanced accuracy, accuracy, and F1-score, with group-specific sensitivity analysis conducted for healthy controls, treated PD patients, and untreated PD patients.

Results quantified the contribution of each additional metabolite, identified diminishing returns points, and assessed whether a minimal metabolite panel could achieve clinically acceptable performance compared to the full feature set. This analysis addressed whether a simpler biomarker panels could still produce accurate PD diagnosis in this cohort.

2.7.7 Removing the Top 6 Amides and Top 4 L-DOPA Associated Molecules

Given that classifier performance depends on features (Liebal et al. 2020), removing key features should result in decreased classification accuracy. Conversely, when performance remains stable after feature removal, identifying which metabolites maintain predictive capability can provide alternative predictive biomarkers (Learn 2025e).

This analysis used the original SVM hyperparameters to ensure fair comparison across different feature sets. Three metabolite removal scenarios were tested:

2. Method

L-DOPA pathway metabolites (3-methoxytyrosine, 3-methoxytyramine sulfate, m-tyramine sulfate, dopamine 3-O-sulfate), fatty acid amides (palmitoleamide, myristoleamide, myristamide, heptadecenamide, linolenamide, linoleamide), and combined removal of both groups. For each scenario, the remaining metabolites underwent Boruta feature selection using identical parameters as the original analysis, followed by SVM training with the original hyperparameters to isolate the effect of metabolite removal from hyperparameter optimization. These analyses were scored by AUC and balanced accuracy, to see if the removal of metabolites was also affecting different groups to different extents.

2.8 Histogram Visualisation of the Top 5 SHAP Metabolites

Histograms were generated to visualise the distribution of the top five metabolites identified through SHAP feature importance analysis. These visualisations compared metabolite abundance patterns between healthy controls and PD patients, as well as between treated and untreated PD subgroups. All histograms were created using base R packages (R version [4.4.1]) with 30 bins per plot. This approach enabled direct visual comparison of metabolite concentration patterns across participant groups (HC, PD treated, PD untreated).

2.9 Multi-Method Overlap Analysis

To identify robust biomarker candidates consistently detected across different analytical approaches, overlap analysis was performed between three feature selection methods: SHAP (SHapley Additive exPlanations) values from machine learning models, GLM2 regression analysis, and LIMMA differential expression analysis.

Metabolites from each method were filtered according to the following criteria: GLM2 metabolites with Benjamini-Hochberg adjusted p-value < 0.05 , LIMMA metabolites with absolute log fold change > 0.2 , and all SHAP-ranked metabolites were included. Three-way overlap identified metabolites that were significant

2. Method

across all three independent methods, representing the most consistently detected biomarker candidates.

For overlapping metabolites, a comparative ranking table was constructed showing each metabolite's rank and statistical metrics (average rank across methods and standard deviation across all methods) across all three methods. Additionally, metabolites were annotated with pathway information (super and sub-pathways) to identify enriched metabolic pathways among the consensus biomarkers. All analyses were conducted using Python [v 3.13] with pandas, numpy, and scipy libraries, and R data files were read using the pyreadr package.

3

Results of Aim 1 Exploratory: Characterise metabolic alterations in PD serum samples

Contents

3.1	Data Overview	58
3.1.1	Sample Demographics	58
3.2	Metabolites Overview	60
3.3	Demographic Analysis: Results from Clustering	60
3.3.1	Principal Component Analysis (PCA)	60
3.3.2	UMAP	64
3.4	Differential Analysis: Results from LIMMA Modelling	68
3.5	Confounding Factors Analysis: Results from General Linear Model Analysis	69
3.5.1	Case-Control General Linear Models	69
3.5.2	LEDD General Linear Models	70
3.6	Summary	74

The following section will provide an overview of the dataset, as well as exploring how confounding factors affect this dataset, and then finally explore if significant alterations in metabolite abundances can be seen between Parkinson's (PD) and healthy control (HC) serum samples.

3.1 Data Overview

The dataset used in this project is a matrix of 1,685 samples with 1,122 metabolite abundances per sample with no missing data.

3.1.1 Sample Demographics

The tracking PD cohort consists primarily of PD patients (1482) with a smaller class of healthy controls (203) (see, Figure 3.1), resulting in a ~ 1 HC: 7 PD ratio. There are more men overall (1068), and more men with PD (962), as compared to women in this dataset, which is in line with the literature (Ben-Shlomo et al. 2024). Most of the PD participants take some sort of PD medication, with only 8% of the PD participants being considered drug naive. The cohort itself (PD + HC) has an average age of 67, within a range of 50 - 99, see table 3.1. The average age of onset is 66 (range: 50 - 89), and all patients recruited had been diagnosed within 3 years before recruitment. As expected of a cohort who had been diagnosed within 3 years, this cohort displays mild symptomatic PD burden. The median total score for the MoCA (Montreal Cognitive Assessment) adjusted for years of education is 26 (range: 9–30), which is on the lower limit of normal cognition. The median score for UPDRS III (Unified Parkinson’s Disease Rating Scale Part III), which measures motor symptom burden, is 19 (range: 0 - 76). However, patients who take PD medication were likely measured while in an ‘ON’ medication state (as the study does not mention explicitly otherwise, (Malek et al. 2015)), which would decrease motor symptom burden. Most participants were being prescribed a relatively low dose of PD medication, with the median LEDD dosage being 300 mg/day (relative to early onset patients who were on average being prescribed 926 mg/day in the same cohort (Malek et al. 2015)).

3. Results of Aim 1 Exploratory: Characterise metabolic alterations in PD serum samples

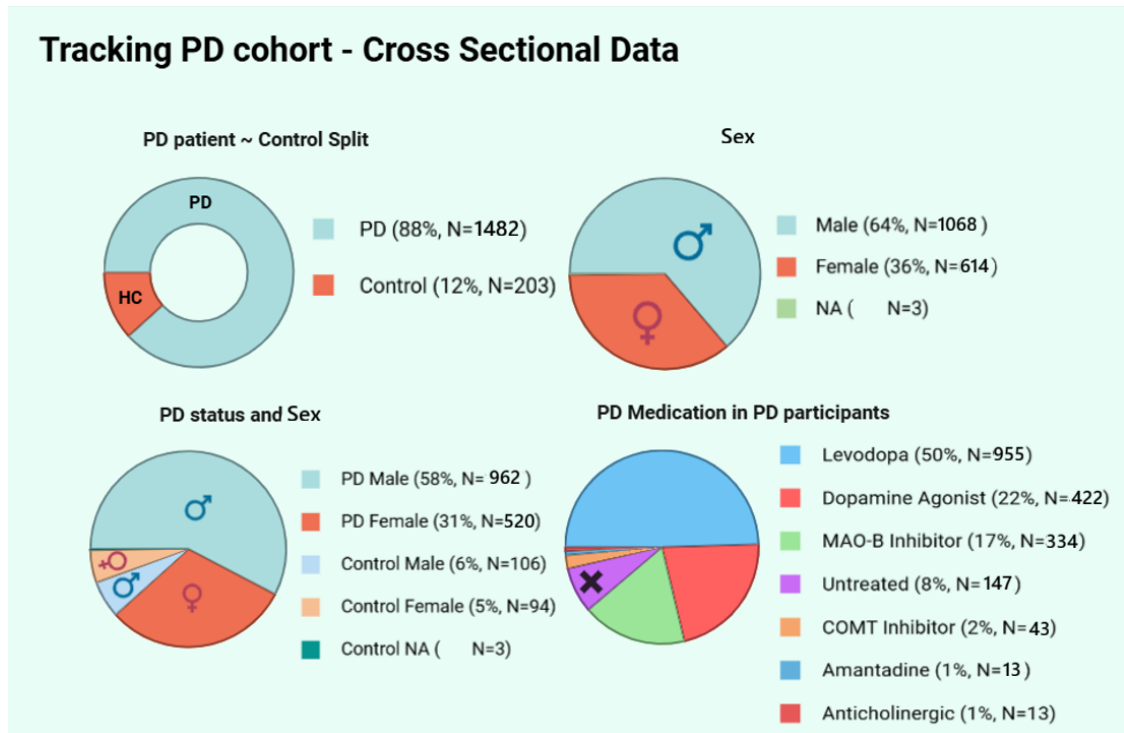


Figure 3.1: Summary Statistics for the late-onset Parkinson’s Disease (PD) participants and healthy controls for the Tracking PD cohort. There is a greater number of PD participants at 1482 than healthy controls at 203. Sex of the cohort reveals slightly more men than women, which is in line with the current consensus that PD affects men slightly more. The medication pie chart represents medical records as opposed participants. Most PD participants take multiple PD medications, however 147 participants are drug naïve. Created in <https://BioRender.com>

Table 3.1: Descriptive statistics for demographic and clinical characteristics of the study cohort.

Variable	Mean	Median	SD	Range
Age (years)	67.35	67.10	7.63	50.22–90.66
Age of Onset (years)	66.12	65.82	7.62	50.07–89.66
MoCA (adjusted) ^a	25.04	26.00	3.53	9–30
UPDRS III ^b	21.27	19.00	13.55	0–76
LEDD (mg/day) ^c	286.66	300.00	199.35	0–2049.5

Montreal Cognitive Assessment, adjusted for education level (26 or above = normal, 21-25 = mild cognitive impairment, 20 or below = dementia).

Unified Parkinson’s Disease Rating Scale Part III (motor examination); higher scores indicate greater motor impairment, maximum score = 132.

Levodopa Equivalent Daily Dose in milligrams per day.

3. Results of Aim 1 Exploratory: Characterise metabolic alterations in PD serum samples

3.2 Metabolites Overview

Metabolon’s untargeted mass spectrometry identified 1,642 metabolites in total: 1,339 known and 303 unknown metabolites for this dataset. The metabolite measures are reported as relative abundances of each metabolite. An additional dataset with chemical information on each metabolite was provided, along with a super and sub-pathway for each metabolite (custom categories created by Metabolon).

Super pathways: Amino Acid, Carbohydrates, Cofactors and Vitamins, Energy, Lipid, Nucleotide, Partially Characterised Molecules, Peptide, Xenobiotic, Blank (unknown)

Sub pathways: 109 sub pathways, see Appendix A.1

3.3 Demographic Analysis: Results from Clustering

Initial clustering was used to look at both the ability to separate PD and HC samples using unsupervised methods, but also as a quality control for confounding factors.

3.3.1 Principal Component Analysis (PCA)

PCA performed on the samples (each dot represents a sample) did not successfully separate the PD and HC sample groups, as seen in Figure 3.2. PC1 explains 7.8% of the dataset’s variation and PC2 explains 6%.

When colouring the same PCA by age (see 3.3), it can be observed that age (older - blue, younger - green-yellow) appears to shift across PC1 (x-axis), indicating that metabolites within PC1 may be capturing some of the variance in samples seen by age. When colouring the PCA by sex (see Figure 3.4), it can be seen that the two sexes seem to be separated across PC1, moving from more male (blue) samples to female (red) samples but not to a large degree. In figure 3.5, the PCAs show no colour-related clustering, which could indicate that LEDD and site location (site num) do not present as a major source of variation in the data.

3. Results of Aim 1 Exploratory: Characterise metabolic alterations in PD serum samples

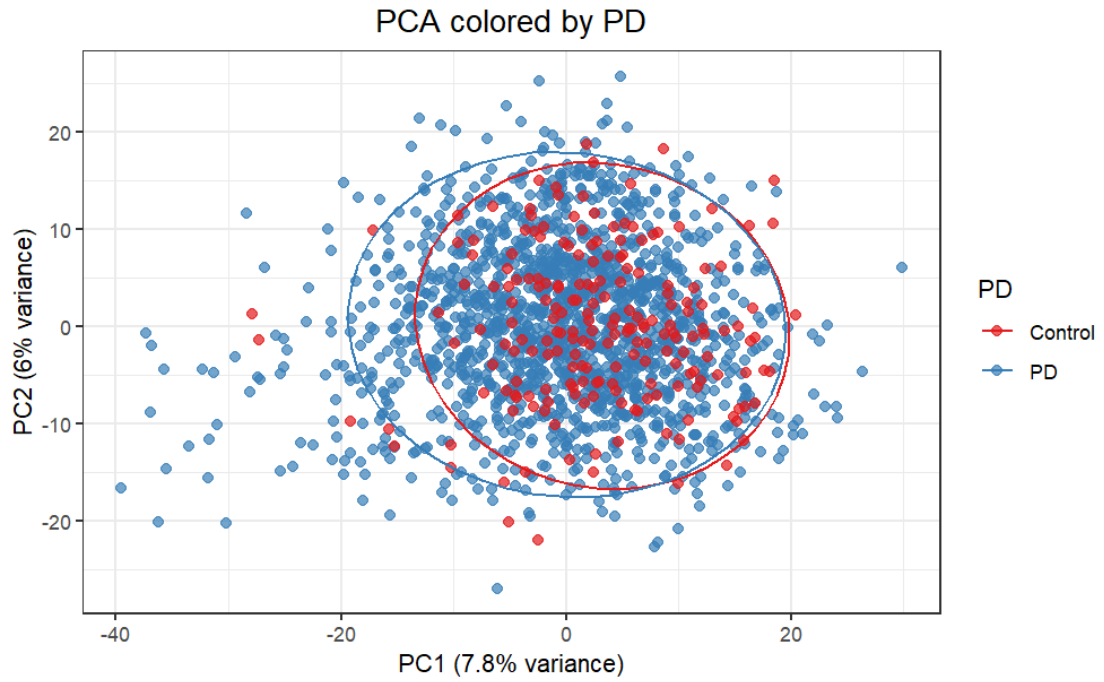


Figure 3.2: Figure: Principal Component Analysis colored by Parkinson's disease status. PCA scatter plot showing sample distribution across PC1 (7.8% variance) and PC2 (6% variance), with samples colored by disease status (red: control, blue: PD).

The samples (only PD) showed that disease severity (inferred from MoCA and UPDRS III scores) could not separate the data.

3. Results of Aim 1 Exploratory: Characterise metabolic alterations in PD serum samples

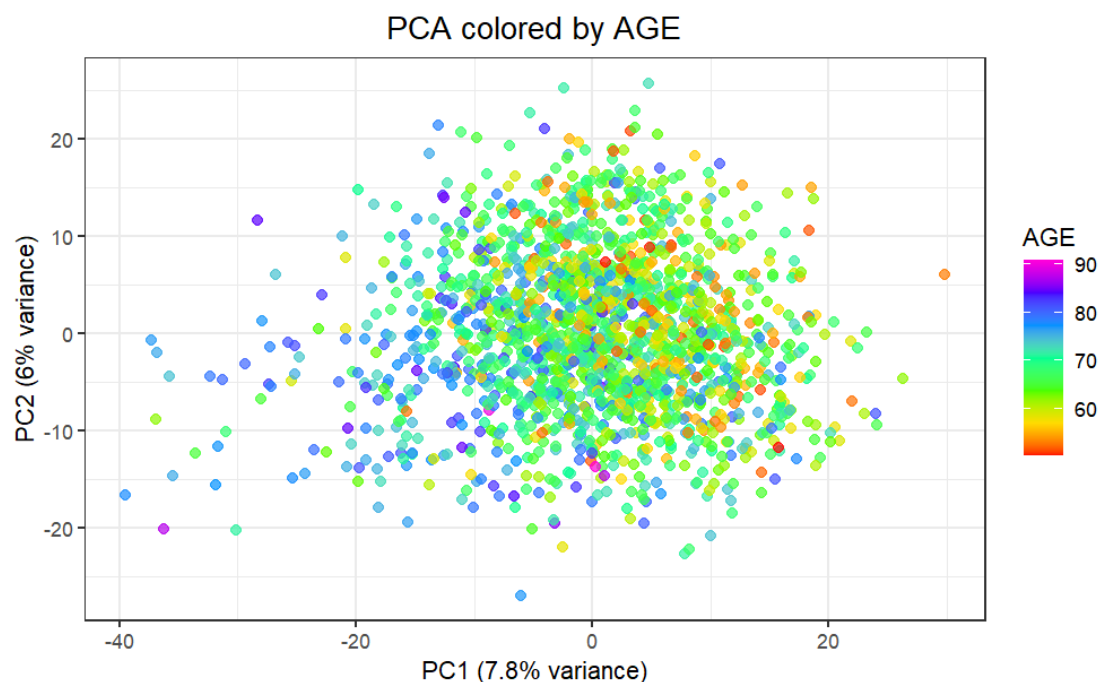


Figure 3.3: Figure: Principal Component Analysis of samples colored by age. PCA scatter plot showing the distribution of samples across the first two principal components (PC1: 7.8% variance, PC2: 6% variance). Points are coloured by age (orange/red: 50-60 years, yellow-green: 60-70 years, cyan-blue: 70-80 years, purple-magenta: 80-90 years) according to the colour scale shown.

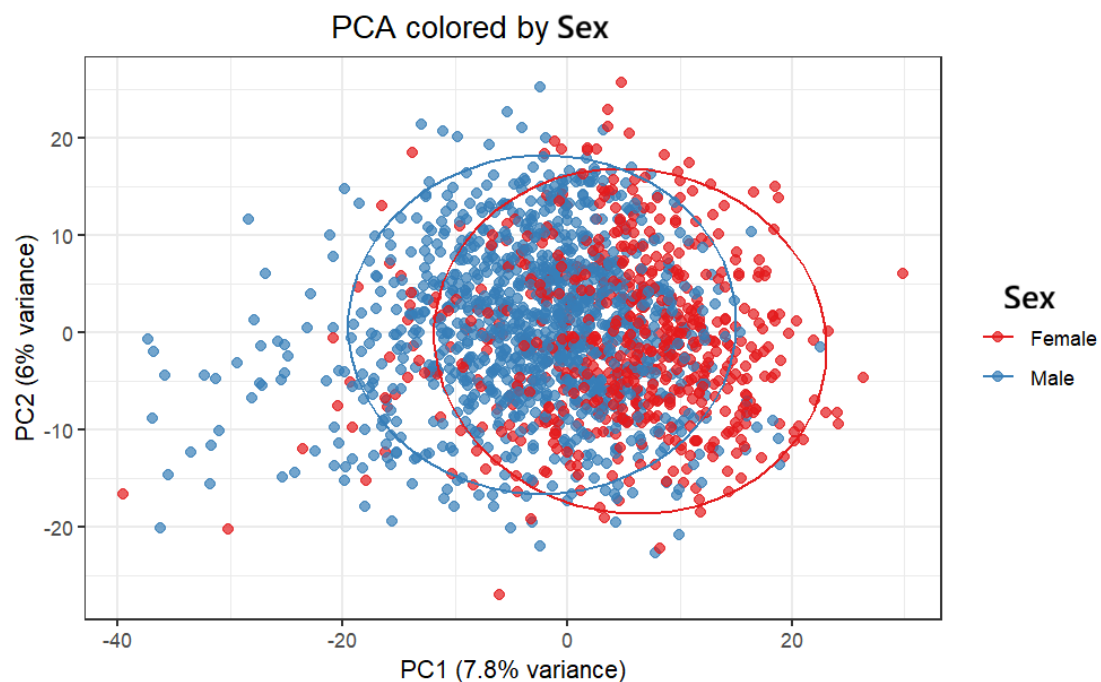


Figure 3.4: Figure: Principal Component Analysis coloured by sex. PCA scatter plot showing sample distribution across PC1 (7.8% variance) and PC2 (6% variance), with samples colored by gender (red: female, blue: male).

3. Results of Aim 1 Exploratory: Characterise metabolic alterations in PD serum samples

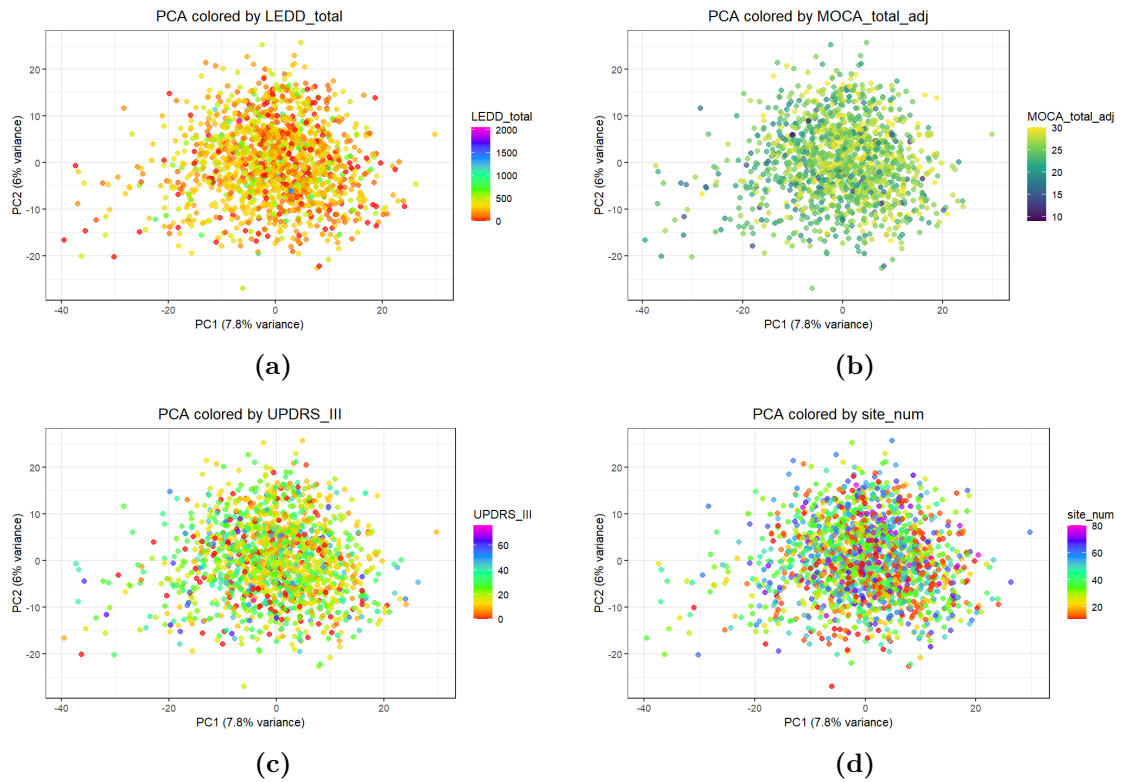


Figure 3.5: Figure: Principal Component Analysis colored by clinical variables and study site. PCA scatter plots showing sample distribution across PC1 (7.8% variance) and PC2 (6% variance), colored by: (a) LEDD_total (Levodopa Equivalent Daily Dose; 0-2000 mg), (b) MOCA_total_adj (Montreal Cognitive Assessment score; 0-30), (c) UPDRS_III (motor examination score; 0-70), and (d) site_num (study site identifier). .

3. Results of Aim 1 Exploratory: Characterise metabolic alterations in PD serum samples

3.3.2 UMAP

UMAP plots (see figure 3.6) of the samples show distinct clustering, showing between 2-3 PD clusters (blue) and 1 HC cluster (red), with some overlap with the top left PD cluster, with some healthy control samples. Ellipses are drawn around each group, indicating that the data form distinct clusters. In contrast to PCA, UMAP reveals more apparent structure in the data, with a distinct cluster of control samples visible in the upper right region. PD samples appear to cluster in 2 distinct groups (upper left and bottom cluster), with a potential third group (upper right), which shows overlap with the control group.

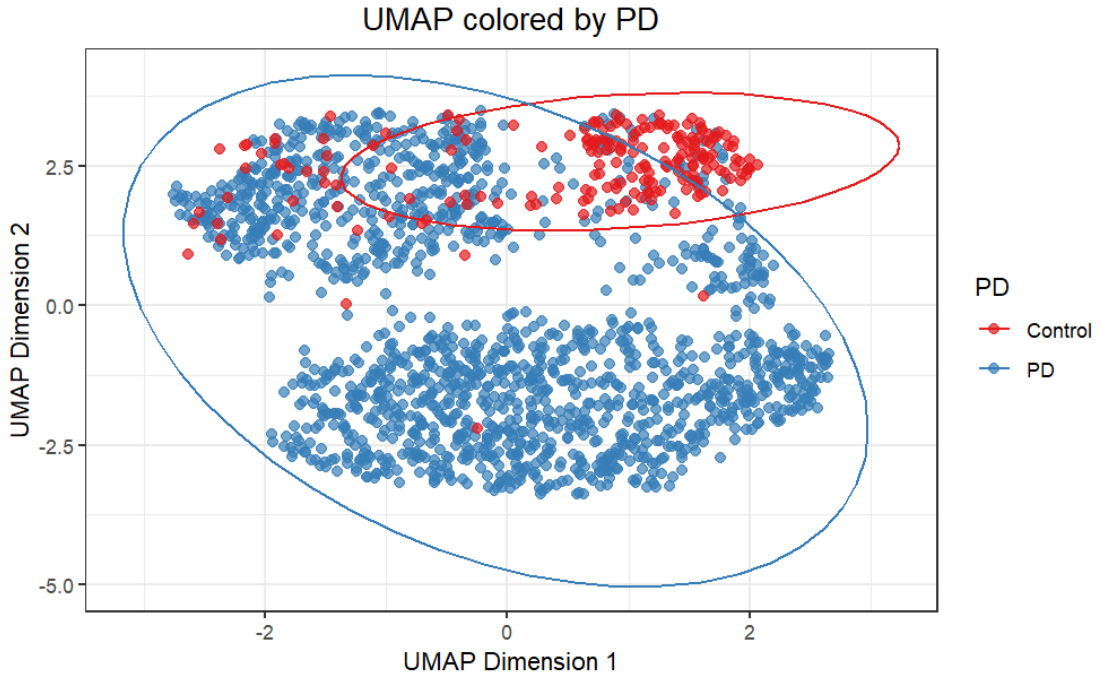


Figure 3.6: Figure: UMAP projection coloured by PD status. UMAP visualisation showing sample distribution across two dimensions, with samples colored by disease status (red: control, blue: PD).

UMAP on samples, coloured by age (see figure 3.7), shows that the largest, bottom cluster is made up of more older participants (darker blue) than the other clusters, which are made up of younger participants (more yellow-green). However, the colour separation is not visually distinct and appears more like a gradient. Figure 3.8, coloured by LEDD, reveals clear clustering patterns. The largest bottom

3. Results of Aim 1 Exploratory: Characterise metabolic alterations in PD serum samples

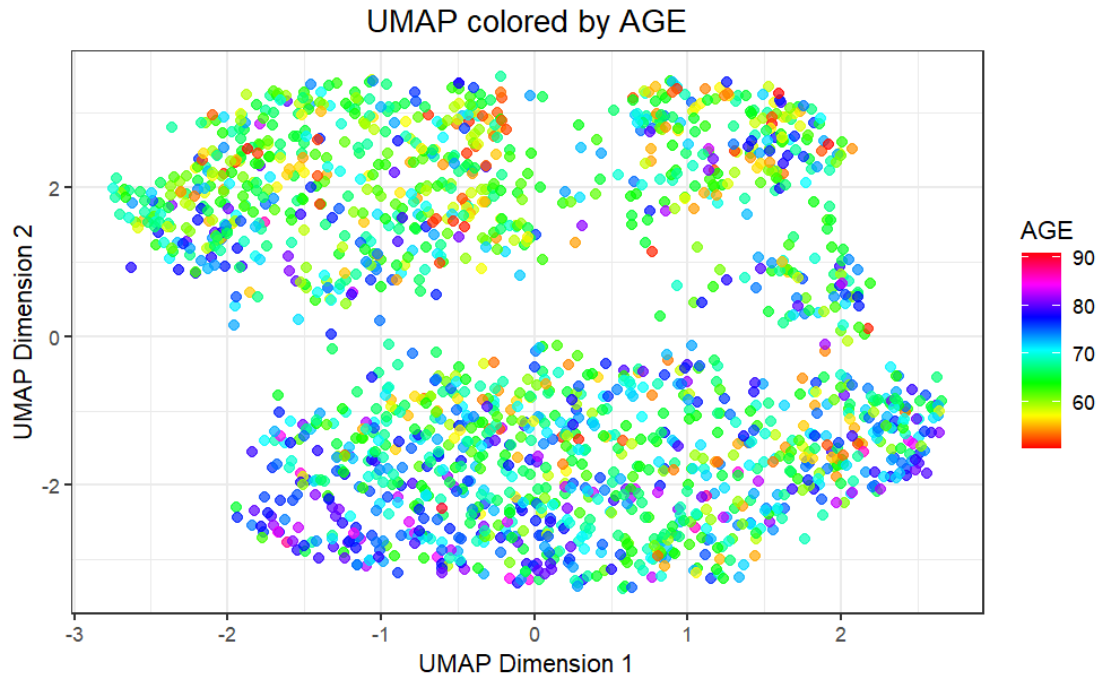


Figure 3.7: Figure: UMAP projection colored by age. UMAP visualisation showing sample distribution across two dimensions, with samples colored by participant age (50-90 years, orange to purple scale).

cluster exhibits higher doses (yellow-green), while the top clusters exhibit lower doses (0 mg/day), appearing red-orange. This visualisation includes only PD samples, as healthy controls have no LEDD values. Notably, some PD samples occupy the same region as healthy controls in Figure 3.6. Figure 3.9 shows the same UMAP coloured by other confounding factors, which show little distinct patterning. Site location, MoCA, and UPDRS appear to be randomly distributed throughout the clusters. Sex also does not separate clusters distinctly, but within clusters, samples of the same sex appear to cluster closer together.

3. Results of Aim 1 Exploratory: Characterise metabolic alterations in PD serum samples

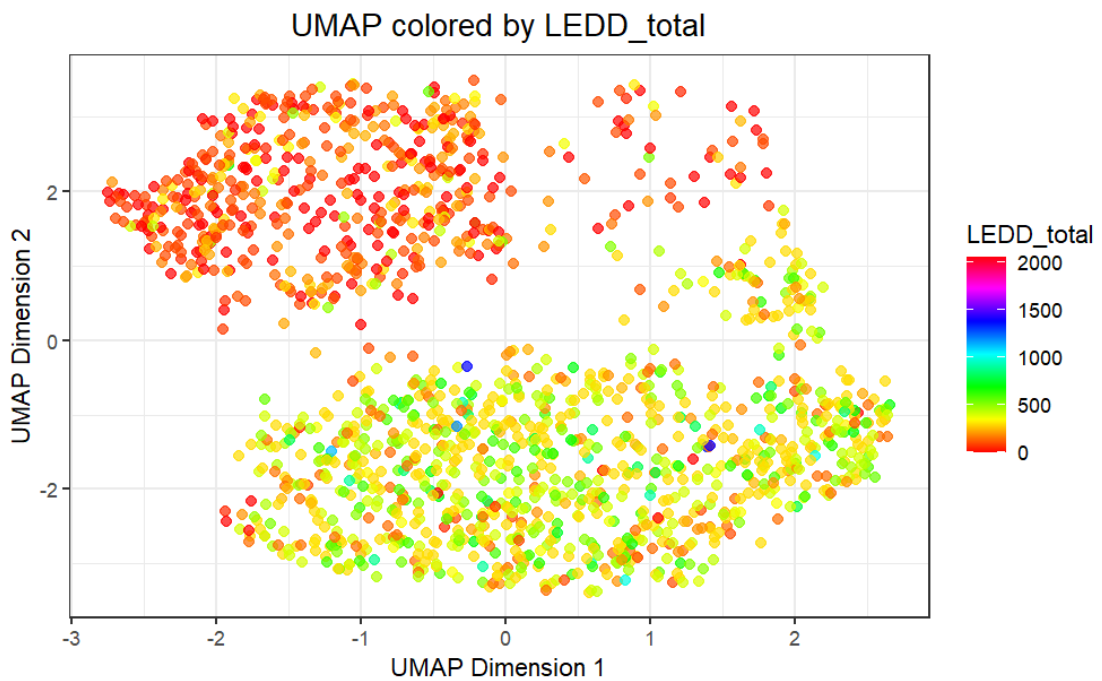


Figure 3.8: Figure: UMAP projection colored by levodopa equivalent daily dose. UMAP visualisation showing sample distribution across two dimensions, with samples colored by LEDD_total (Levodopa Equivalent Daily Dose; 0-2000 mg, red to green scale). A clear spatial segregation is evident, with patients on higher medication doses (red/orange) predominantly concentrated in the upper left region, while patients on lower doses or no medication (yellow/green) cluster in the lower right region.

3. Results of Aim 1 Exploratory: Characterise metabolic alterations in PD serum samples

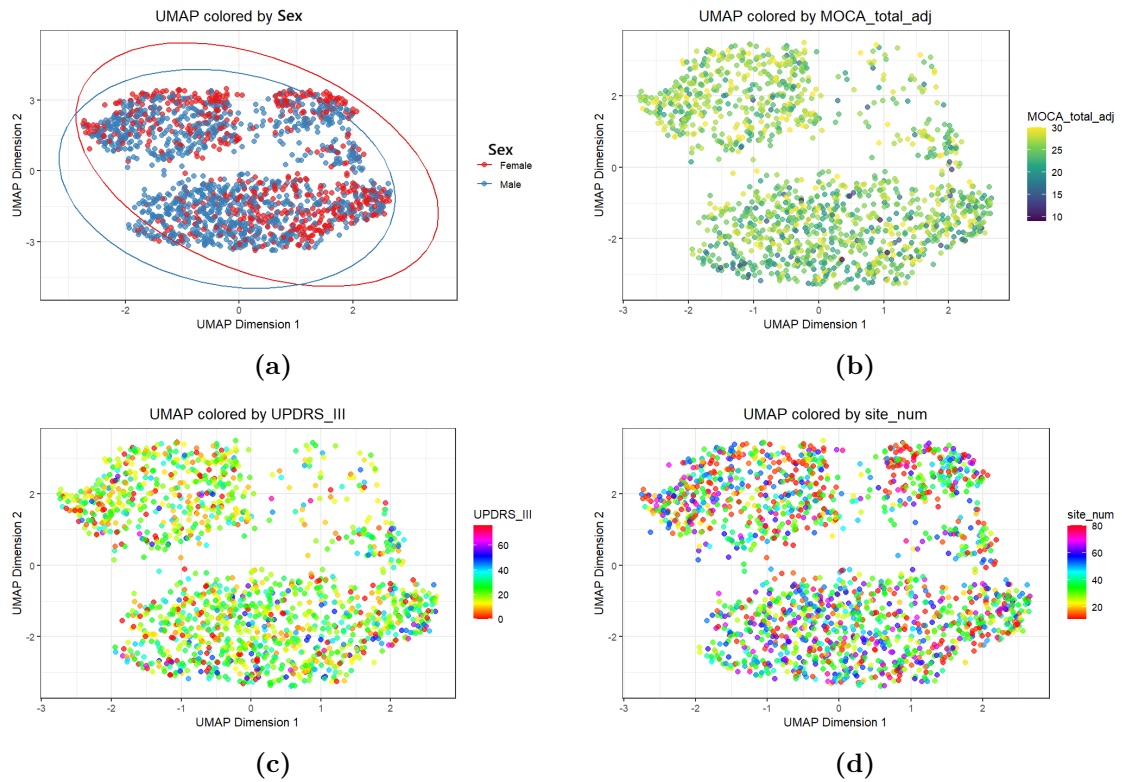


Figure 3.9: Figure: UMAP projections colored by clinical variables and study characteristics. UMAP visualizations showing sample distribution across two dimensions, colored by: (A) Sex (red: female, blue: male); (B) MOCA_total_adj (Montreal Cognitive Assessment score; 0-30); (C) UPDRS_III (motor examination score; 0-70); and (D) site_num (study site identifier).

3. Results of Aim 1 Exploratory: Characterise metabolic alterations in PD serum samples

3.4 Differential Analysis: Results from LIMMA Modelling

LIMMA differential abundance analysis was performed to compare PD and HC samples. Using thresholds of adjusted p-value < 0.05 ($-\log_{10}(\text{adjusted } p) > 1.3$) and ($|\log_2 \text{FC}| > 0.2$), a total of 30 metabolites were found to be significantly altered between the two groups (Figure 3.10, shown in red). A greater number of metabolites showed significant increases in PD samples compared to HC samples than showed significant decreases. Additionally, metabolites elevated in PD samples exhibited both larger fold changes and greater statistical significance than those elevated in HC samples.

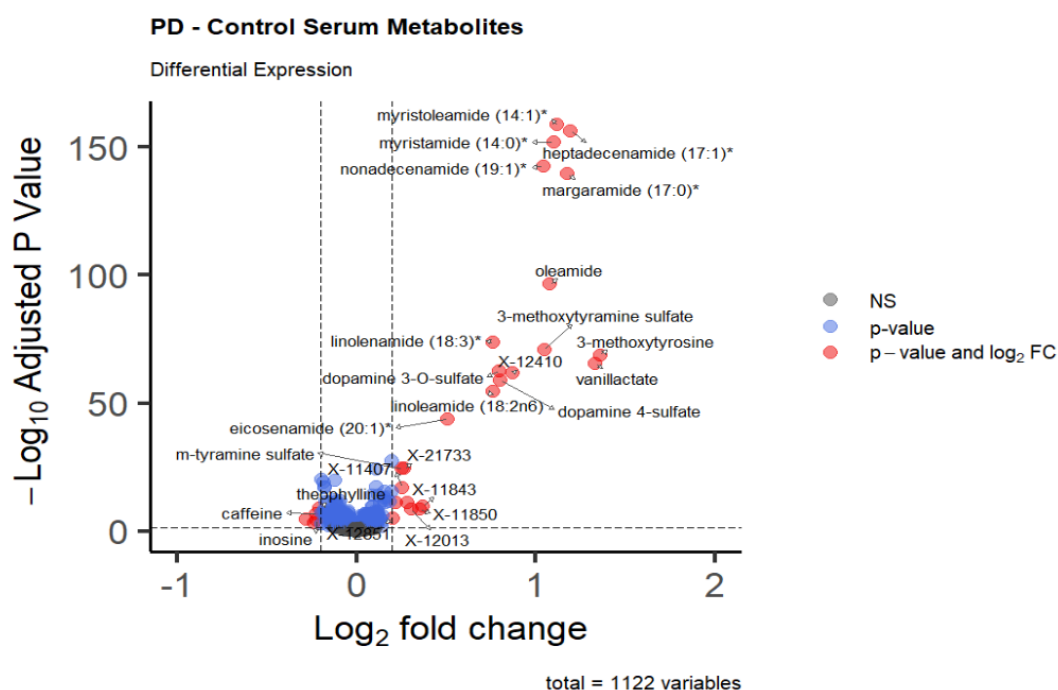


Figure 3.10: Volcano plot of differential serum metabolite expression between Parkinson’s disease (PD) patients and controls. The x-axis represents ($\log_2$) fold change in metabolite abundance, while the y-axis shows ($-\log_{10}$) adjusted p-values (Benjamini-Hochberg correction). Each point represents an individual metabolite ($n = 1122$). Metabolites are colored by significance: grey dots indicate non-significant changes (NS), blue dots represent metabolites meeting the p-value threshold (adjusted $p < 0.05$), and red dots denote metabolites meeting both statistical significance and biological relevance thresholds ($-\log_{10}(\text{adjusted } p) > 1.3$ and $|\log_2 \text{FC}| > 0.2$). Vertical dashed lines mark the fold change cut off of ± 0.2 , and the horizontal dashed line indicates the adjusted p-value threshold of 0.05. Only red dots are annotated.

3. Results of Aim 1 Exploratory: Characterise metabolic alterations in PD serum samples

The top 10 differentially expressed metabolites from LIMMA modelling (Table 3.2) were all significantly elevated in PD samples, as demonstrated by their positive log fold changes. These metabolites exhibited highly significant adjusted p-values ($\alpha = 0.05$), reflecting substantial abundance differences between PD and HC samples, with consistently higher levels observed in PD samples.

Table 3.2: Top 10 Differentially Expressed Metabolites from LIMMA Modelling Ranked by Adjusted P-Value

Metabolite	Log ₂ FC	Adj. P-Value	-Log ₁₀ Adj. P	Significance
palmitoleamide (16:1)*	1.351	6.02e-163	162.22	***
myristoleamide (14:1)*	1.122	1.84e-159	158.74	***
heptadecenamide (17:1)*	1.197	4.09e-157	156.39	***
myristamide (14:0)*	1.100	1.50e-152	151.82	***
nonadecenamide (19:1)*	1.044	3.47e-143	142.46	***
margaramide (17:0)*	1.179	2.56e-140	139.59	***
oleamide	1.081	3.29e-97	96.48	***
linolenamide (18:3)*	0.764	2.45e-74	73.61	***
3-methoxytyramine sulfate	1.048	1.40e-71	70.85	***
3-methoxytyrosine	1.359	2.08e-69	68.68	***

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Adj. p values = Benjamini-Hochberg correction applied

3.5 Confounding Factors Analysis: Results from General Linear Model Analysis

3.5.1 Case-Control General Linear Models

Two case-control general linear models (GLM) were run and analysed,

1. GLM1: Case/control $\sim$ metabolite (logistic regression)
2. GLM2: Case/control $\sim$ metabolite + sex + age (logistic regression)

The top 10 metabolites were compared between these two GLMs to assess whether correcting for age and sex would change which metabolites were most significantly altered, to observe how confounding factors (age and sex) affect the data. No change was observed in the identity and order of the top 10 metabolites

3. Results of Aim 1 Exploratory: Characterise metabolic alterations in PD serum samples

(ranked by adjusted p-value, Benjamini-Hochberg corrected), with the only notable difference being that the adjusted p-values decreased slightly in the model which corrected for age and sex. Given the similarity of the results, only GLM2 is included in this results section (see table 3.3); however, a table of the top 10 metabolites for GLM1 can be found in the appendix (A.2.1).

Examining GLM2 results (see table 3.3), many of the top 10 metabolites were similar to the top 10 metabolites seen from LIMMA results (see table 3.2). The most common super-pathway was Lipids, and the most common sub-pathway was Fatty Acid, Amide. The beta values were all positive with highly significant adjusted p-values ($\alpha = 0.05$), indicating that the top 10 metabolites were present at elevated concentrations within PD samples.

There was an attempt to run a third case-control GLM,

- GLM 3: Case/control $\sim$ individual metabolite + age + gender + LEDD

GLM 3 encountered convergence issues due to perfect separation, where LEDD values completely distinguished case-control status. Although a small number of drug-naïve PD patients also have an LEDD score of 0, all controls having an LEDD of 0 created perfect separation between the two classes. Consequently, covariates were unable to resolve drug correction. For the subsequent analysis, the results of GLM 2 (PD diagnosis $\sim$ individual metabolite + age + gender) were used.

3.5.2 LEDD General Linear Models

Two general linear models (GLM) with Levodopa Equivalent Daily Dose (LEDD) as an output were run and analysed,

1. GLM4: LEDD $\sim$ metabolite (linear regression)
2. GLM5: LEDD $\sim$ metabolite + sex + age (linear regression)

Similar to the case-control models, correcting for age and sex did not alter the identity or order of the top 10 metabolites. The adjusted p-values also decreased slightly when corrected for sex and age. The uncorrected LEDD model, GLM4, can

3. Results of Aim 1 Exploratory: Characterise metabolic alterations in PD serum samples

Table 3.3: Top 10 metabolites from GLM 2: Case/control $\sim$ metabolite + sex + age (logistic regression), ranked by adjusted p -value

Metabolite	Effect		Statistics		Pathway Information	
	β (SE)	p	p_{BH-adj}	Sig.	Super Pathway	Sub Pathway
palmitoleamide (16:1)*	2.035 (0.114)	9.86e-72	1.11e-68	***	Lipid	Fatty Acid, Amide
myristamide (14:0)*	2.472 (0.140)	6.68e-70	2.97e-67	***	Lipid	Fatty Acid, Amide
heptadecenamide (17:1)*	2.441 (0.138)	7.94e-70	2.97e-67	***	Lipid	Fatty Acid, Amide
myristoleamide (14:1)*	2.919 (0.172)	1.78e-64	5.00e-62	***	Lipid	Fatty Acid, Amide
margaramide (17:0)*	2.722 (0.161)	3.64e-64	8.18e-62	***	Lipid	Fatty Acid, Amide
nonadecenamide (19:1)*	2.958 (0.177)	1.11e-62	2.09e-60	***	Lipid	Fatty Acid, Amide
oleamide	1.435 (0.093)	2.23e-53	3.57e-51	***	Lipid	Fatty Acid, Amide
linolenamide (18:3)*	1.913 (0.129)	8.94e-50	1.26e-47	***	Lipid	Fatty Acid, Amide
dopamine 3-O-sulfate	1.988 (0.156)	2.73e-37	3.41e-35	***	Amino Acid	Tyrosine Metabolism
linoleamide (18:2n6)	1.250 (0.098)	3.54e-37	3.98e-35	***	Lipid	Fatty Acid, Amide

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Note: p_{BH-adj} = Benjamini-Hochberg correction applied

3. Results of Aim 1 Exploratory: Characterise metabolic alterations in PD serum samples

be found in the appendix A.2.2, and the corrected model GLM5 can be found in Table 3.4. These metabolites correlated with increases in LEDD dosage and may have been related to LEDD metabolism. The most common super-pathway was Amino Acid, and the most common sub-pathway was Tyrosine metabolism (which is linked to dopamine metabolism (Meiser et al. 2013)). Most metabolites exhibited large positive beta values, indicating that these metabolites increased alongside LEDD dosage increases. Several metabolites exhibited large negative beta values, indicating that these metabolites decreased as LEDD increased. Adjusted p-values were highly significant for the top 10 metabolites (compared to $\alpha = 0.05$).

3. Results of Aim 1 Exploratory: Characterise metabolic alterations in PD serum samples

Table 3.4: Top 10 metabolites from GLM 5: LEDD (Levodopa Equivalent Daily Dose) $\sim$ metabolite + gender + age (linear regression), ranked by adjusted p -value

Metabolite	Effect		Statistics		Pathway Information	
	β (SE)	p	$p_{\text{BH-adj}}$	Sig.	Super Pathway	Sub Pathway
3-methoxytyrosine	133.096 (3.864)	9.26e-193	1.04e-189	***	Amino Acid	Tyrosine Metabolism
vanillactate	133.440 (3.879)	2.19e-192	1.23e-189	***	Amino Acid	Tyrosine Metabolism
3-methoxytyramine sulfate	159.340 (5.384)	1.94e-152	7.26e-150	***	Amino Acid	Tyrosine Metabolism
X-12410	181.885 (6.262)	5.51e-148	1.54e-145	***	NA	NA
dopamine 4-sulfate	187.307 (6.704)	4.87e-139	1.09e-136	***	Amino Acid	Tyrosine Metabolism
dopamine 3-O-sulfate	193.160 (6.978)	5.83e-137	1.09e-134	***	Amino Acid	Tyrosine Metabolism
X-18779	420.154 (19.608)	3.04e-89	4.87e-87	***	NA	NA
N-acetyl-isoptreanine	-384.227 (22.235)	3.10e-61	4.34e-59	***	Amino Acid	Polyamine Metabolism
acisoga	-319.167 (22.323)	1.32e-43	1.64e-41	***	Amino Acid	Polyamine Metabolism
X-12726	194.817 (16.201)	6.94e-32	7.79e-30	***	NA	NA

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: $p_{\text{BH-adj}}$ = Benjamini-Hochberg correction applied

3. Results of Aim 1 Exploratory: Characterise metabolic alterations in PD serum samples

3.6 Summary

Results demonstrated that the data could be clustered using UMAP, and the most distinct colouration observed within UMAP clustering was LEDD, suggesting LEDD was likely an important confounding factor. Both LIMMA and GLM indicated substantial differences in metabolic profiles between PD and HC samples. LIMMA and GLM also identified the same metabolites as altered between the groups. Correcting for age and sex in GLMs did not change the identity or order of the top 10 metabolites, suggesting that age and sex may not have had large confounding effects on this data. LEDD GLMs showed that some metabolites were highly correlated with LEDD dose in participants, indicating that drug dosage did have a large confounding effect, as was seen in the UMAP.

4

Results of Aim 2 Validation: Develop and validate a metabolite-based classification of PD

Contents

4.1	Developing PD Classification: Results from Machine Learning Pipeline	76
4.1.1	Data Split	76
4.1.2	Boruta Feature Selection Results	76
4.1.3	Evaluation of Model-Feature Combinations	77
4.2	Validating PD Classification: Results from best-performing algorithm	79
4.2.1	Group Sensitivity - Drug treatment as a confounding factor	81
4.3	Sensitivity Analysis	82
4.4	Summary	84

This section presents the results from the second aim, which aimed to replicate the changes observed in the statistical models used in Aim 1 through a different analytical approach. A secondary aim was to identify metabolites that were altered regardless of treatment status. Classification also enabled assessment of how well these metabolites could be used in a diagnostic setting.

4. Results of Aim 2 Validation: Develop and validate a metabolite-based classification of PD

4.1 Developing PD Classification: Results from Machine Learning Pipeline

4.1.1 Data Split

As demonstrated in the previous results section 3, LEDD was shown to distinguish clusters of PD samples and was associated with metabolites that were highly dependent on LEDD dosage. Recognising that PD treatment status may generate distinct metabolic profiles, the data were stratified into three groups: HC, treated PD (LEDD > 0), and untreated PD (LEDD = 0, drug-naïve).

Table 4.1 shows that the total dataset was divided into three groups, with each group then split into training (80%) and testing (20%) subsets. The largest group was treated PD (1335), followed by HC (203), with untreated PD (147) representing the smallest group; this approximately followed a ratio of 9 treated PD:1.4 healthy controls:1 untreated PD. Following the split into training and testing subgroups, the combined training set contained 1347 samples and the testing set contained 338 samples.

Table 4.1: Distribution of samples across groups in training and test sets

Group	Total	Training	Test
Treated PD	1335	1068	267
Untreated PD	147	117	30
Healthy Control	203	162	41
Total	1685	1347	338

4.1.2 Boruta Feature Selection Results

Boruta was run on the training data (N=1347) with 1122 features. There were 52 confirmed features and 7 tentative features found by Boruta (Table 4.2). The

Table 4.2: Boruta feature selection results

Feature Category	Count	Percentage
Total Features Analyzed	1122	100%
Confirmed Features	52	4.6%
Tentative Features	7	0.6%
Rejected Features	1063	94.7%

4. Results of Aim 2 Validation: Develop and validate a metabolite-based classification of PD

top 10 confirmed (out of 52) features in Boruta (see table 4.3) are ranked by random forest importance and have some metabolites which overlap with the top 10 metabolites found in GLM and LIMMA. However, the rank (order) of the metabolites is different when compared to GLM and LIMMA.

Table 4.3: Top 10 confirmed Boruta features ranked by Random Forest importance

Rank	Feature	RF Importance
1	myristoleamide (14:1)*	0.121
2	margaramide (17:0)*	0.098
3	heptadecenamide (17:1)*	0.078
4	3-methoxytyramine sulfate	0.075
5	nonadecenamide (19:1)*	0.057
6	palmitoleamide (16:1)*	0.053
7	3-methoxytyrosine	0.052
8	X-12410	0.047
9	vanillactate	0.043
10	myristamide (14:0)*	0.042

X- denotes unidentified metabolites

4.1.3 Evaluation of Model–Feature Combinations

Given the lack of prior knowledge about which feature combinations would be most predictive for this dataset, a systematic evaluation of multiple features and models was conducted to identify the optimal classifier (see table 4.4). This totalled to 24 runs, across 4 methods (Random Forest, Lasso, SVM, XGBoost) and 6 different feature counts. All classifiers are binary classifiers predicting a sample’s PD status (regardless of being treated / untreated) based on a number of metabolite abundances.

Results of the top 10 models from the model-feature combinations are summarised in Table 4.5. All models demonstrated very high AUC scores, with 0.5 representing performance no better than random guessing and 1 representing a perfect predictor. Minimal differences were observed between the top-performing models. However, some models, particularly extreme gradient boosting (XGBoost), were computationally expensive, requiring up to approximately 18 hours for a

4. Results of Aim 2 Validation: Develop and validate a metabolite-based classification of PD

Table 4.4: Feature subset testing experimental design

Parameter	Value
Total features in dataset	1122
Feature counts tested	5, 10, 20, 40, 52 (Boruta confirmed), 1122 (all features)
Number of feature subsets	6
Machine learning models tested	Random Forest, Lasso, SVM, XGBoost
Hyperparameter tuning	Optuna, 100 trials per model
Feature ranking method	Random Forest importance from Boruta analysis
Total experimental runs	24 (6×4)

single combination to complete. The top-performing models tended to favour either 40 or 52 features as the optimal feature count. The best-performing model was a support vector machine (SVM) with 40 features (top 40 from Boruta), which achieved an AUC score of 0.9943 with a variation of ± 0.0031 across all nested cross-validation folds ($n = 5$). The data above is presented as a graph (see Figure

Table 4.5: Validation scores from nested cross-validation on all model combinations

Rank	Model	Features	AUC Score	Time (min)
1	SVM	40	0.9943 ± 0.0031	2.9
2	Lasso	40	0.9940 ± 0.0030	3.1
3	Lasso	52	0.9935 ± 0.0032	5.9
4	SVM	52	0.9933 ± 0.0036	2.8
5	XGBoost	52	0.9923 ± 0.0030	90.6
6	XGBoost	40	0.9918 ± 0.0028	80.2
7	XGBoost	1122	0.9907 ± 0.0038	1114.6
8	XGBoost	20	0.9882 ± 0.0058	54.0
9	Lasso	1122	0.9880 ± 0.0069	9.8
10	SVM	1122	0.9878 ± 0.0065	41.4

4.1), with each colour representing one of the 4 models across all feature counts. Support vector machine (SVM, red) and logistic regression with Lasso (Lasso, blue) perform the highest across all feature counts. XGBoost (yellow) performs slightly worse at the initial feature count of 5 and then scores as high as Lasso and SVM. Random Forest (RF, green) consistently performs worse at all feature counts compared to the other 3 models.

4. Results of Aim 2 Validation: Develop and validate a metabolite-based classification of PD

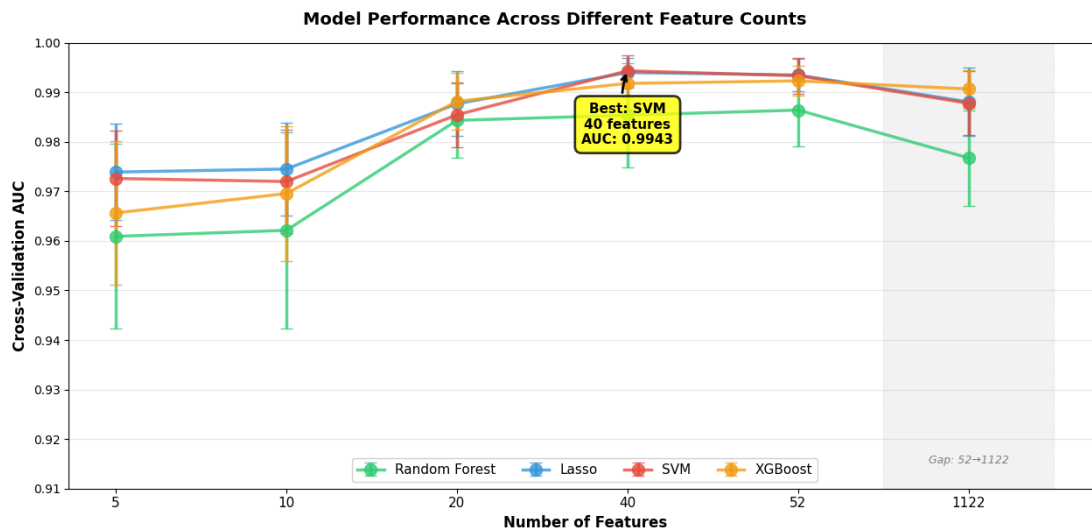


Figure 4.1: Cross-validation performance of four machine learning algorithms across different feature subset sizes selected by Boruta feature selection. Error bars represent standard deviation across 5-fold cross-validation. The optimal model is a SVM with 40 features, and an AUC = 0.9943. The shaded region indicates a discontinuity in the x-axis from 52 to 1122 features, representing the jump from Boruta-selected subsets to the complete feature set.

4.2 Validating PD Classification: Results from best-performing algorithm

The highest performing model was chosen, SVM with 40 features on the final hold-out testing set (338), see summary table 4.6.

Table 4.6: Optimal model configuration and test set characteristics

Parameter	Value
Selected model	Support Vector Machine (SVM) with a linear kernel
Number of features	40 (from Boruta-confirmed features)
Cross-validation performance	0.9943 $\pm$ 0.0031 AUC
Test set size	338 samples

Results on the testing set can be seen in the following Receiver Operating Characteristic (ROC) curve figure 4.2, and the final test scores can be seen summarised in the table 4.7. The final training curve is in blue and the testing curve is red, and as expected, the testing curve performs slightly worse than the

4. Results of Aim 2 Validation: Develop and validate a metabolite-based classification of PD

training curve. However, both curves show a high performance with the dotted line representing an AUC of 0.5 (classifier performs as well as random assignment) and following the left-handed corner of the graph would be an AUC of 1 (classifier always correctly predicts class). In the table 4.7, in the testing set, the AUC is slightly lower at 0.9884 than the training AUC at 0.9960, with a difference of less than 1%, indicating the model generalises well and is not overfitting. The F1 score (for PD =1, the dominant class) also reports a high test statistic, 0.9848. Accuracy drops slightly to 0.9734, but remains a high score.

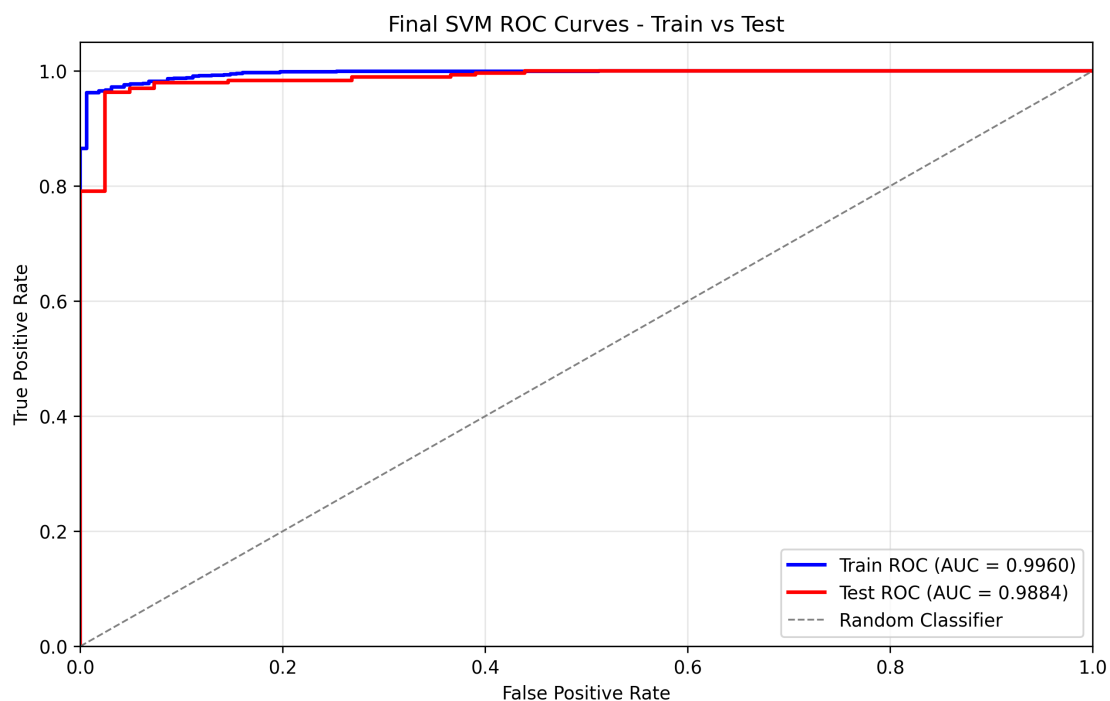


Figure 4.2: Receiver Operating Characteristic (ROC) curves for the final SVM model. The model has a training AUC of 0.9960 and test AUC of 0.9884. Random classification follows the diagonal dashed lines 0.5.

Table 4.7: Final SVM model performance on training and test sets

Dataset	Metric	Score
Training	AUC	0.9960
Test	AUC	0.9884
Test	Accuracy	0.9734
Test	F1 Score	0.9848

A detailed report of testing scores can be seen in table 4.8, and here we can see

4. Results of Aim 2 Validation: Develop and validate a metabolite-based classification of PD

that while the F1 score for PD samples (Class 1, both treated and untreated) is at 0.98, it drops for the HC class to 0.89. A similar pattern is seen for precision. The final balanced accuracy (taking into account class imbalance) for this model on the test set is 0.955.

Table 4.8: Final SVM detailed classification report

Class	Precision	Recall	F1-Score	Test set (N)
Class 0 (HC)	0.86	0.93	0.89	41
Class 1 (PD)	0.99	0.98	0.98	297
Accuracy		0.97		338
Balanced Accuracy		0.955		338
Macro avg	0.93	0.95	0.94	338
Weighted avg	0.97	0.97	0.97	338

4.2.1 Group Sensitivity - Drug treatment as a confounding factor

Using the predictions from the final SVM model, individual subgroup predictions were explored to see how well the model generalised across treatment groups. As expected, the model was best at predicting PD-treated samples with an accuracy of 0.9813, and then untreated PD at 0.9667 and worst at predicting HC at 0.9268 (see table 4.9).

Table 4.9: SVM performance metrics by patient group on the test set

Group	n	Accuracy	Sens./Spec.	Precision	F1 Score
Healthy Controls	41	0.9268	0.9268 ^a	—	—
PD Treated	267	0.9813	0.9813 ^b	1.0000	0.9905
PD Untreated	30	0.9667	0.9667 ^b	1.0000	0.9831

^aSpecificity (correct healthy identification)

^bSensitivity (correct PD detection)

The confusion matrix shows (see table 4.10), 7.3% of HC were misclassified as PD, while only 3.3% of untreated PD samples were misclassified as HC (despite having similar numbers to HC). Only 1.9% of treated-PD samples were misclassified as HC.

4. Results of Aim 2 Validation: Develop and validate a metabolite-based classification of PD

Table 4.10: Detailed confusion matrix showing PD subtypes (misclassifications highlighted)

	True Class	Predicted		Total	Misdiagnosed (%)
		Healthy (0)	PD (1)		
Actual	Healthy Control	38	3	41	7.3%
	PD Treated	5	262	267	1.9%
	PD Untreated	1	29	30	3.3%
Total		44	294	338	2.7%

Green highlighting indicates misclassified cases

4.3 Sensitivity Analysis

Given the unexpected high scores from the models, a sensitivity analysis was conducted. The sensitivity analysis can help validate the robustness of the classifier.

First, negative controls were tested, and the final SVM classifier (40 features, same hyperparameters) was tested on noise data and shuffled data (see table 4.11). If the classifier’s performance is truly based on the altered metabolites found in the data, it will perform poorly ($AUC \sim 0.5$) on the noisy and shuffled data (ie no metabolite patterns to detect). The SVM scored an AUC of 0.4815 for the noise data and 0.5250 on the shuffled data.

Table 4.11: Negative Control Tests: Model Performance on Random Data

Test	AUC	Balanced Accuracy
Gaussian Noise	0.4815	0.5118
Shuffled Labels	0.5250	0.5000

Note: Performance near chance level (0.5).

Given the highly imbalanced dataset, the hypothesis tested whether class imbalance was biasing the model. The final SVM (same features and same hyperparameters), after being retrained on a balanced dataset, scored highly on the test set with an AUC of 0.9887 (Table 4.12).

To assess whether the classifier was overly reliant on the treated PD class during training, different combinations of the subgroups were used to retrain the final SVM model, which was then tested on a subgroup excluded from training (Table 4.13). The balanced accuracy for the final SVM model trained on all data was

4. Results of Aim 2 Validation: Develop and validate a metabolite-based classification of PD

Table 4.12: Model Performance on Balanced Dataset

Metric	Value
Samples per Group	147
Total Samples	441
Train/Test Split	352/88
Test AUC	0.9887
Balanced Accuracy	0.9331

Note: Balanced dataset created by sampling 147 subjects from each group (HC, PD Treated, PD Untreated).

0.955 (Table 4.8). When trained on healthy controls and treated PD and tested on untreated PD, the model achieved a balanced accuracy of 0.9184. Training on HC and untreated PD, then testing on treated PD, yielded a balanced accuracy of 0.9603. Finally, training on both PD groups and testing on HC and untreated PD achieved a balanced accuracy of 0.9296. None of these combinations resulted in a substantial decrease in balanced accuracy.

Table 4.13: Cross-Group Generalisation: Model Performance Across Different Patient Subgroups

Test	Train → Test	Accuracy/Balanced Accuracy	F1 Score
Test 1	HC + Treated → Untreated (n=1538 → 147)	0.9184 ^b	NA
Test 2	HC + Untreated → Treated (n=350 → 1335)	0.9603 ^b	NA
Test 3	HC + Treated → HC + Untreated (n=1538 → 350)	0.9296 ^a	0.9184

Note: HC = Healthy Controls. ^bBalanced Accuracy. ^aAccuracy.

A general robustness analysis was undertaken to see if classifier performance could be explained by a certain data split. Data was re-split, then Boruta feature selection was rerun and hyperparameters were retuned, to see if the classifier was still able to perform as well (difference of -0.0097 in AUC), and Boruta-selected features appear stable (90% of the features overlap) (see table 4.14).

4. Results of Aim 2 Validation: Develop and validate a metabolite-based classification of PD

Table 4.14: Robustness Analysis: Stability of Model Performance and Feature Selection

Test Component	Original	Alternative
<i>Hyperparameter Sensitivity</i>		
Cross-validation AUC	0.9943	0.9847
C parameter	0.1580	2.5436
Gamma parameter	0.0916	0.0004
Kernel	linear	linear
AUC Difference		−0.0097
<i>Data Split Sensitivity</i>		
Random state	42	999
Training samples	1348	1348
Test samples	337	337
<i>Feature Selection Stability</i>		
Confirmed features	40	36
Tentative features	0	2
Feature overlap	36 (90.0%)	
Features only in original		4
Features only in alternative		0

Note: Alternative configuration uses different hyperparameters and data split (random_state=999 vs 42).

4.4 Summary

In this section, classifiers were able to predict PD status with very high accuracy, validating the findings from Aim 1 that metabolite profiles were distinctive between PD and HC samples. This metabolic distinction was preserved regardless of PD treatment status, indicating that drug treatment may not impede PD classification prediction, even for untreated PD patients. Classification demonstrated promise for predicting PD status, achieving very high AUC scores. Sensitivity analyses provided evidence that the classifier’s performance was not entirely an artefact of the dataset characteristics.

5

Results of aim 3 Mechanistic: Explore underlying disease pathways associated with metabolites most altered in PD samples

Contents

5.1	Pathway enrichment results	86
5.2	SHapley Additive exPlanations (SHAP) analysis . . .	89
5.2.1	Misclassification - exploring what metabolites defined each subgroup	89
5.3	Metabolite Panel: Progressively Adding Features from 1-5	95
5.3.1	Group Sensitivity	96
5.4	Removing top metabolites - exploring model prediction dependence	98
5.4.1	Removing 4-LDOPA-associated metabolites	98
5.4.2	Removing 6 Fatty Acid Amide Metabolites	99
5.4.3	Both L-DOPA-associated metabolites and Fatty Acid Amides Removed	100
5.4.4	Summary	102
5.5	Histograms of the top 5 SHAP metabolites	103
5.6	Overlap results between Fold Change Analysis, GLM and SHAP feature importance	106
5.6.1	Three-way overlap metabolites	107
5.6.2	Super Pathway Analysis	110
5.6.3	Sub-Pathways	110
5.7	Summary	112

This section presents the results of aim 3 to look at the specific metabolites and pathways which are most altered between healthy controls and Parkinson’s samples. As well as exploring how these metabolites interact with drug status.

5.1 Pathway enrichment results

A hypergeometric enrichment test was performed on the significantly altered metabolites from GLM2 (Case/control $\sim$ metabolite + sex + age, logistic regression) to determine whether certain metabolic super-pathways contained more significantly altered metabolites than expected by chance. Table 5.1 presents the results for the top five super-pathways, including the number of significantly altered metabolites, enrichment p-values, and the average adjusted p-value and beta coefficient for significant metabolites within each pathway.

Only two super-pathways demonstrated significant enrichment ($\alpha = 0.05$): Xenobiotics and Peptides. These pathways contained a higher proportion of significantly altered metabolites between PD and healthy control samples than would be expected based on random chance alone, given the total number of metabolites within each pathway. The Xenobiotics pathway (exogenous metabolites) showed stronger enrichment (lower enrichment p-value) compared to the Peptides pathway. The average significant xenobiotic metabolite had a negative beta value (-0.511) and a small adjusted p-value, indicating that xenobiotic metabolites were predominantly decreased in PD samples compared to controls, with high statistical significance. In contrast, peptide metabolites exhibited an average positive beta value (0.602) and a slightly larger (yet still significant) adjusted p-value, indicating that peptide metabolites were predominantly elevated in PD samples, though with somewhat lower statistical significance compared to xenobiotics.

The hypergeometric test for sub-pathways revealed that 3 sub-pathways’ metabolites were significantly ($\alpha = 0.05$) enriched: xanthine metabolism, fatty acid amide,

5. Results of aim 3 Mechanistic: Explore underlying disease pathways associated with metabolites most altered in PD samples

Table 5.1: Pathway enrichment analysis for significant metabolites in GLM 2: Top 5 super pathways

Super Pathway	Sig.	Exp.	Total	Fold Enr.	Adj. p-val	Signif.	Avg. p (BH)	Avg. Beta
Xenobiotics	53	31.46	111	1.68	3.7e-05	***	3.9e-03	-0.511
Peptide	22	12.75	45	1.72	1.1e-02	*	1.1e-02	0.602
Energy	5	3.12	11	1.60	5.8e-01	ns	1.3e-02	-0.643
Carbohydrate	7	6.24	22	1.12	9.4e-01	ns	1.1e-02	1.153
Amino Acid	55	54.13	191	1.02	9.4e-01	ns	9.1e-03	0.570
<i>+ 5 other super-pathways (not significant)</i>								

Sig. = Significant count; Exp. = Expected count; Enr. = Enrichment; Adj. = Adjusted Background: N=1122 metabolites; Significant: n=318; Enriched pathways: fold enrichment >1

Adjusted p-values, Benjamin-Hochberg multiple testing correction applied

Significance: *** p<0.001, ** p<0.01, * p<0.05, ns = not significant

and phosphatidylethanolamine (PE) pathways. Xanthine and fatty acid amide pathways are more significantly (have smaller adjusted p-values) enriched compared to phosphatidylethanolamine (PE). Examining the average adjusted p-value and average beta coefficient for significantly altered metabolites within each sub-pathway, the Fatty Acid, Amide sub-pathway exhibited a very small average adjusted p-value (7.5×10^{-33}) and a positive beta value (2.150), indicating that metabolites in this sub-pathway were, on average, highly significantly elevated in PD samples. In contrast, the Xanthine and PE (phosphatidylethanolamine) sub-pathways had negative beta values and larger adjusted p-values, indicating that metabolites in these sub-pathways were, on average, significantly decreased in PD samples, though with less statistical significance compared to the Fatty Acid, Amide sub-pathway.

Table 5.2: Subpathway enrichment analysis for GLM 2 (Case/control $\sim$ metabolite + gender + age)

Sub-Pathway	Sig. Count	Exp. Count	Total	Fold Enr.	Adj. p-val		Avg. p (BH)	Avg. Beta
Xanthine Metabolism	12	3.68	13	3.26	1.7e-04	***	5.8e-03	-1.226
Fatty Acid, Amide	10	2.83	10	3.53	1.7e-04	***	7.5e-03	2.150
Phosphatidylethanolamine (PE)	11	4.53	16	2.43	3.1e-02	*	6.0e-03	-1.533
Polyamine Metabolism	6	1.98	7	3.02	7.3e-02	ns	1.7e-03	0.779
Monoacylglycerol	8	3.40	12	2.35	1.4e-01	ns	1.1e-03	-1.866
Phosphatidylinositol (PI)	5	1.70	6	2.94	1.5e-01	ns	1.4e-02	-1.377
Chemical	13	7.09	25	1.83	1.6e-01	ns	8.5e-03	-0.339
Ceramides	5	1.98	7	2.52	3.0e-01	ns	1.1e-02	-1.672
Acetylated Peptides	4	1.42	5	2.82	3.0e-01	ns	1.2e-02	1.064
Fatty Acid Metab. (Acyl Choline)	5	2.27	8	2.21	4.9e-01	ns	2.6e-02	0.961
<i>+ 34 other sub-pathways (not significant)</i>								

Background: N=1122; Significant: n=318; Enriched pathways: fold enrichment >1

Significance: *** p<0.001, ** p<0.01, * p<0.05, ns = not significant

5. Results of aim 3 Mechanistic: Explore underlying disease pathways associated with metabolites most altered in PD samples

5.2 SHapley Additive exPlanations (SHAP) analysis

SHAP analysis was conducted on the final SVM model (40 features) described in the previous section 4 , to identify which metabolites (features) were most influential in determining the predicted class of each sample. The SHAP summary plot (Figure 5.1) displays all 40 features ranked by feature importance to the model. Several of the most important metabolites were also identified in the LIMMA and GLM analyses (see section 3). More detailed data of the top 20 metabolites is provided in the table 5.3, with super and sub pathway information. Similar to the hypergeometric test, fatty acid amides appear to be a common sub-pathway within the top 20 SHAP metabolites.

5.2.1 Misclassification - exploring what metabolites defined each subgroup

Several samples were misclassified in the classification analysis (see section 4). Examining representative examples of misclassification from each subgroup could provide insight into which metabolites characterise each subgroup. The most confidently misclassified sample (the sample for which the classifier had the highest confidence in its incorrect prediction) was selected for analysis from each subgroup.

False Positive: Misclassifying Healthy Control as Parkinson's

In this healthy control sample (Figure 5.2), elevated levels of palmitoleamide (SHAP value: +1.21), myristoleamide (+1.01), m-tyramine sulfate (+0.56), and myristamide (+0.20) contributed toward a PD prediction (red bars). Conversely, decreased levels of linoleamide (SHAP value: -0.48) and oleamide (-0.42) contributed toward an HC prediction (blue bars), but were insufficient to counteract the cumulative effect of the elevated metabolites. Palmitoleamide, myristoleamide and myristamide are fatty acid amides, and m-tyramine sulfate is a metabolite related to tyrosine metabolism

5. Results of aim 3 Mechanistic: Explore underlying disease pathways associated with metabolites most altered in PD samples

Table 5.3: Top 20 metabolites ranked by SHAP (SHapley Additive exPlanations) contribution values from the machine learning model for Parkinson’s disease prediction

Rank	Metabolite	Model Importance		Pathway Information	
		Mean	SHAP	Super Pathway	Sub Pathway
1	palmitoleamide (16:1)*	2.443		Lipid	Fatty Acid, Amide
2	myristoleamide (14:1)*	1.768		Lipid	Fatty Acid, Amide
3	3-methoxytyrosine	1.444		Amino Acid	Tyrosine Metabolism
4	3-methoxytyramine sulfate	1.423		Amino Acid	Tyrosine Metabolism
5	myristamide (14:0)*	0.476		Lipid	Fatty Acid, Amide
6	cyclo(leu-pro)	0.415		Peptide	Dipeptide
7	X-21733	0.414		Unknown	Unknown
8	m-tyramine sulfate	0.395		Amino Acid	Tyrosine Metabolism
9	margaramide (17:0)*	0.385		Lipid	Fatty Acid, Amide
10	perfluorooctanesulfonate (PFOS)	0.357		Xenobiotics	Chemical
11	heptadecenamide (17:1)*	0.292		Lipid	Fatty Acid, Amide
12	linolenamide (18:3)*	0.284		Lipid	Fatty Acid, Amide
13	linoleamide (18:2n6)	0.248		Lipid	Fatty Acid, Amide
14	fibrinopeptide B (1-13)**	0.227		Peptide	Fibrinogen Cleavage Peptide
15	dopamine 3-O-sulfate	0.224		Amino Acid	Tyrosine Metabolism
16	X-12410	0.193		Unknown	Unknown
17	threonate	0.178		Cofactors and Vitamins	Ascorbate and Aldarate Metabolism
18	biliverdin	0.159		Cofactors and Vitamins	Hemoglobin and Porphyrin Metabolism
19	p-cresol glucuronide*	0.129		Amino Acid	Tyrosine Metabolism
20	acetoacetate	0.124		Lipid	Ketone Bodies

Note: SHAP values represent the mean absolute contribution of each metabolite to the model’s predictions. Higher values indicate greater importance in distinguishing between Parkinson’s disease cases and controls. Metabolites marked with * or ** indicate compounds with tentative identification by Metabolon, with ** being more confident than *.

5. Results of aim 3 Mechanistic: Explore underlying disease pathways associated with metabolites most altered in PD samples

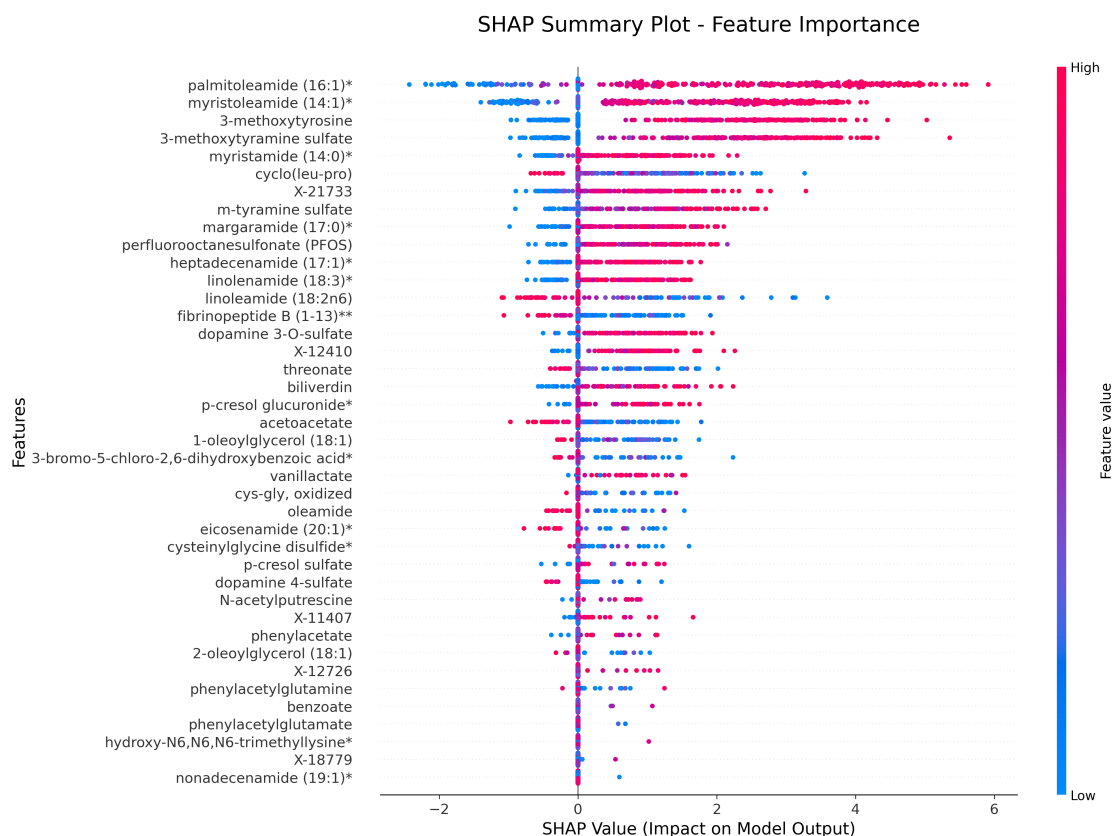


Figure 5.1: Top 40 metabolites from SHAP (SHapley Additive exPlanations) Summary Plot (explaining the final SVM model) illustrating feature importance and impact on model predictions. Each row represents a feature, ordered from most to least important (top to bottom). Each point represents a sample from the dataset, with its position along the x-axis indicating the SHAP value (magnitude and direction of impact on model output). Positive SHAP values (right of zero) indicate the feature pushes the prediction higher, while negative values (left of zero) push it lower. Point colours represent the feature value for that sample: magenta/pink indicates high feature values, while blue indicates low feature values.

False Negative: Misclassification of Treated Parkinson's as Healthy Control

In this treated Parkinson's sample, the sample exhibited lower levels of all metabolites used to classify samples as PD, particularly in levels of 3-methoxytyrosine (-0.71), myristoleamide (-0.63), linoleamide (-0.57), palmitoleamide (-0.5) and 3-methoxytyramine sulfate (-0.35), blue bars are factors which push the model towards a HC prediction. The model predicted HC with a confidence of 0.887. 3-methoxytyrosine and 3-methoxytyramine sulfate are metabolites related to tyrosine metabolism. Myristoleamide, linoleamide and palmitoleamide are fatty acid amides.

5. Results of aim 3 Mechanistic: Explore underlying disease pathways associated with metabolites most altered in PD samples

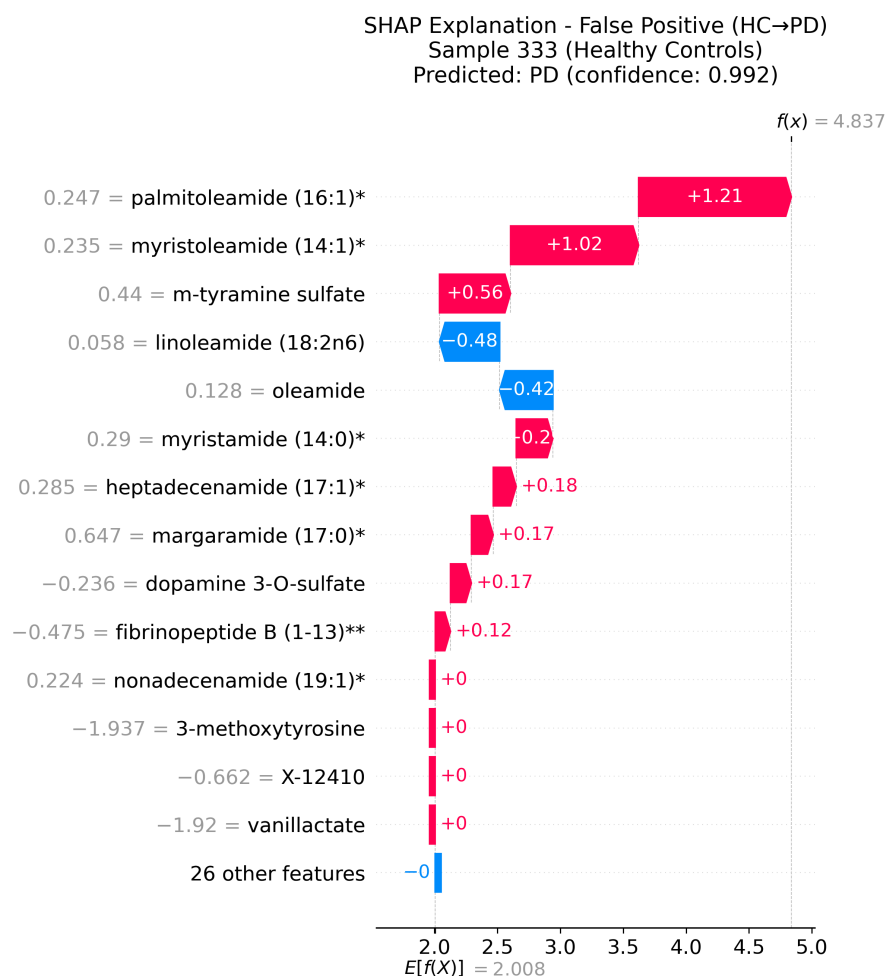


Figure 5.2: SHAP waterfall plot for false positive misclassification (Healthy Control → PD). Sample 333 (99.2% confidence) was incorrectly predicted as Parkinson’s disease despite a confirmed healthy status. Red bars indicate features pushing toward PD prediction; blue bars indicate features pushing toward healthy classification. The baseline value $f(x)$ represents the model’s expected value (2.008), with feature contributions cumulatively shifting the prediction to the final output $E[f(X)]$.

False Negative: Misclassifying Untreated Parkinson’s as a Healthy Control

This untreated PD patient shows a reduction in palmitoleamide (-1.42), myristoleamide (-0.83) and myristamide (-0.54). However shows elevated linoleamide (+1.2), which is the only metabolite which pushes the sample towards its true class PD. This profile is also the opposite profile to the misclassified healthy control. Palmitoleamide, myristoleamide, myristamide and linoleamide are all fatty acid amides.

5. Results of aim 3 Mechanistic: Explore underlying disease pathways associated with metabolites most altered in PD samples

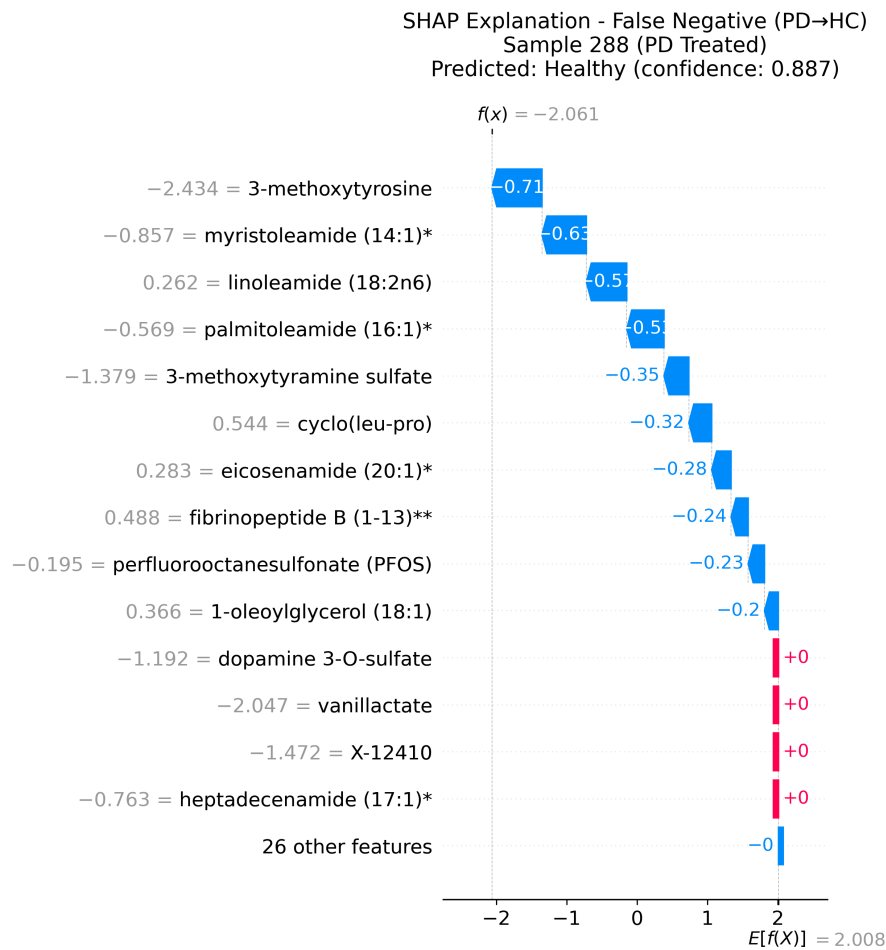


Figure 5.3: SHAP waterfall plot for false negative misclassification (PD → Healthy). Sample 288 (88.7% confidence), a treated PD patient, was incorrectly predicted as a healthy control. Blue bars indicate features pushing toward healthy prediction; red bars indicate features pushing toward PD classification. The baseline value $f(x)$ represents the model's expected value (2.008), with feature contributions cumulatively shifting the prediction to the final output $E[f(X)]$.

5. Results of aim 3 Mechanistic: Explore underlying disease pathways associated with metabolites most altered in PD samples

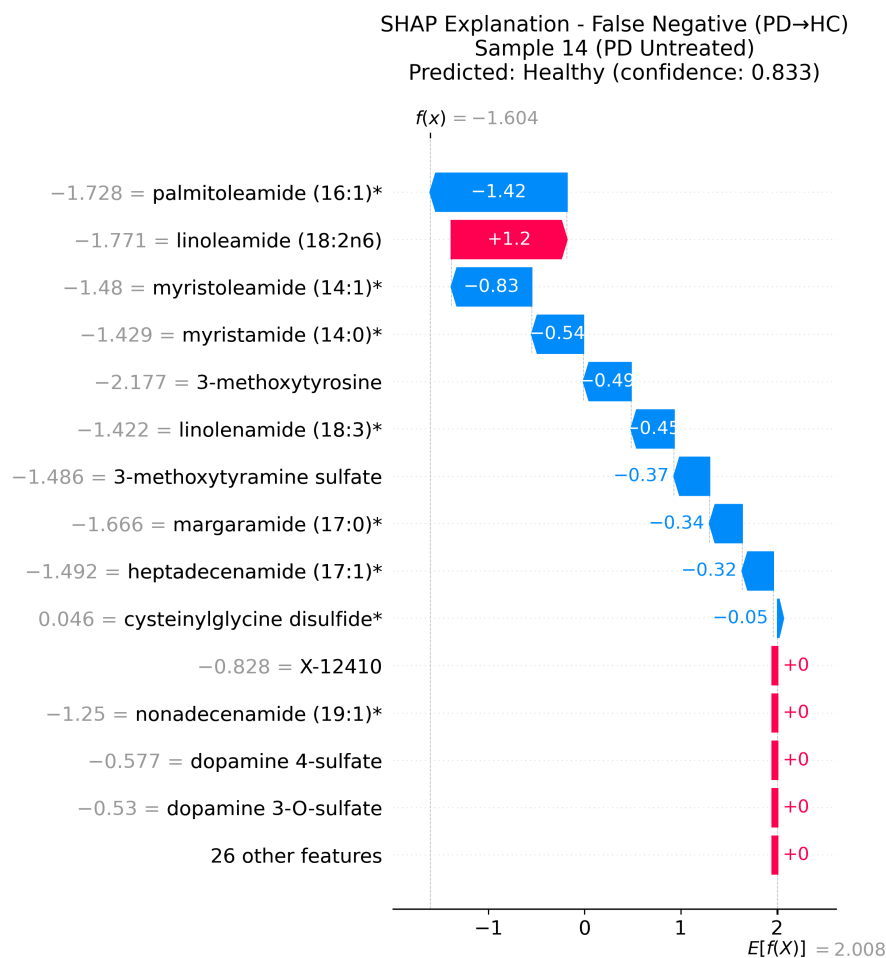


Figure 5.4: SHAP waterfall plot for untreated PD false negative (Sample 14). An untreated Parkinson's disease patient was incorrectly predicted as healthy with 83.3% confidence. Blue bars indicate features pushing toward healthy prediction; red bars indicate features pushing toward PD classification. The baseline value $f(x)$ represents the model's expected value (2.008), with feature contributions cumulatively shifting the prediction to the final output $E[f(X)]$.

5. *Results of aim 3 Mechanistic: Explore underlying disease pathways associated with metabolites most altered in PD samples*

5.3 Metabolite Panel: Progressively Adding Features from 1-5

The misclassification analysis revealed that the top five SHAP-ranked metabolites were central to classification decisions; their performance as a biomarker panel was evaluated. To assess whether these metabolites alone could achieve comparable classification performance—a particularly relevant question for clinical biomarker development—they were added incrementally to an SVM model using the same hyperparameters as the final model as discussed in section 4. The model was evaluated on the same testing set used in the final model assessment.

The 5 metabolites and the order they were added:

1. palmitoleamide (16:1)*
2. myristoleamide (14:1)*
3. 3-methoxytyrosine
4. 3-methoxytyramine sulfate
5. myristamide (14:0)*

Results are summarised in the following figure 4.1, several scoring methods are explored in different colours. AUC (blue), accuracy (green) and F1 (yellow) all show similar high scores across all metabolite feature counts. However, balanced accuracy (red), which can treat classes equally despite imbalance, shows a lower score until metabolite features are increased to 3 metabolites (3-methoxytyrosine). The ROC curves show a similar pattern, with step 1 and step 2 performing worse than steps 3-4 (see figure 5.6. This is more clearly seen in the summary table 5.4, the performance of palmitoleamide (16:1)* only is an AUC of 0.9141, but a balanced accuracy of 0.7989 (suggesting that palmitoleamide might not generalise well). The balanced accuracy only increases to 0.9242 with the inclusion of 3-methoxytyrosine; both the additions of step 2 and steps 4 lead to decreases in balanced accuracy (compared to the step before). However, even at the highest balanced accuracy (0.9242), this is still lower than its balanced accuracy with 40 features (0.955).

5. Results of aim 3 Mechanistic: Explore underlying disease pathways associated with metabolites most altered in PD samples

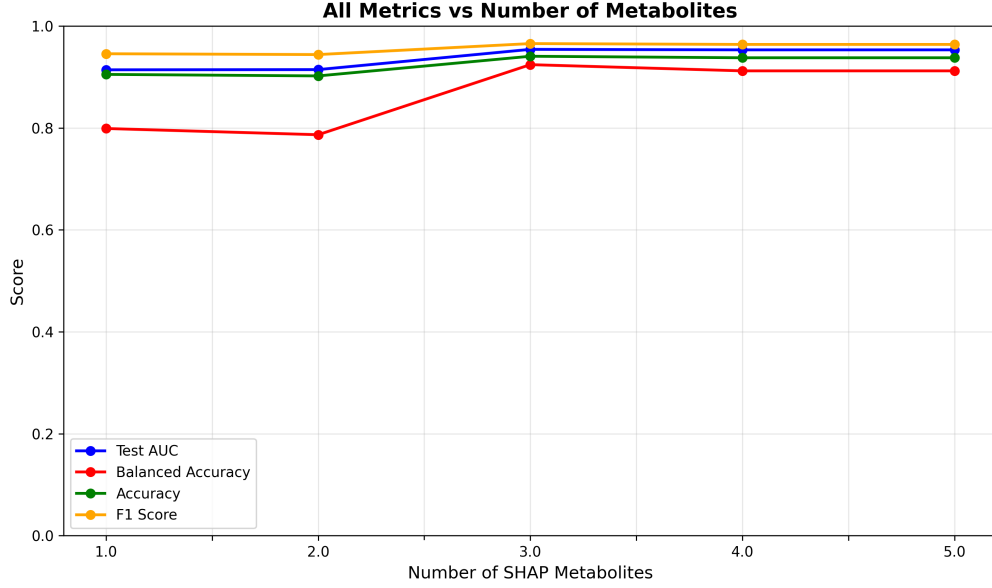


Figure 5.5: Performance metrics across progressive addition of SHAP-selected metabolites. Test AUC (blue), Balanced Accuracy (red), Accuracy (green), and F1 Score (orange) are shown as a function of the number of metabolites included in the model. Most notably, balanced accuracy increases with the addition of the third metabolite, 3-methoxytyrosine.

Table 5.4: Progressive Metabolite Addition Analysis Results

Step	New Metabolite	Test AUC	AUC Gain	Bal. Acc.	Bal. Acc. Gain
1	palmitoleamide (16:1)*	0.9141	0.9141	0.7989	0.7989
2	myristoleamide (14:1)*	0.9145	0.0004	0.7867	−0.0122
3	3-methoxytyrosine	0.9542	0.0397	0.9242	0.1375
4	3-methoxytyramine sulfate	0.9532	−0.0010	0.9120	−0.0122
5	myristamide (14:0)*	0.9532	0.0000	0.9120	0.0000

5.3.1 Group Sensitivity

To assess how the addition of each metabolite affected prediction performance for each group, classification metrics were calculated (Table 5.5). Both healthy controls and treated PD achieved their highest balanced accuracy scores with the addition of the third metabolite, 3-methoxytyrosine: treated PD reached 0.9551, exceeding healthy controls at 0.9024. Notably, the balanced accuracy for untreated PD remained stable at 0.8621 from the inclusion of the first metabolite (palmitoleamide (16:1)*) onwards. Further analysis revealed that the same samples were consistently misclassified despite the addition of metabolites, suggesting that

5. Results of aim 3 Mechanistic: Explore underlying disease pathways associated with metabolites most altered in PD samples

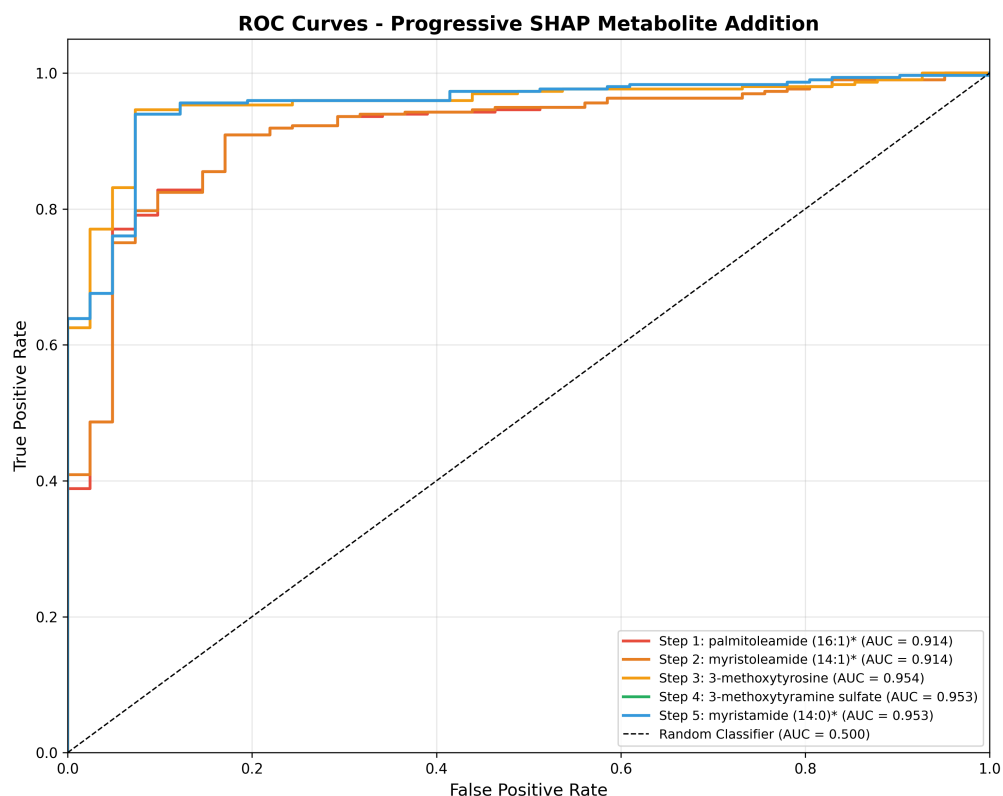


Figure 5.6: ROC curves for progressive SHAP metabolite addition. Each curve represents the model performance with cumulative addition of metabolites: Step 1 - palmitoleamide (16:1)*, Step 2 - myristoleamide (14:1)*, Step 3 - 3-methoxytyrosine, Step 4 - 3-methoxytyramine sulfate, Step 5 - myristamide (14:0)*. The dashed line represents a random classifier (AUC = 0.500). Steps 1-2 are noticeably worse curves compared to steps 3-5. Step 4 (green) and step 5 (blue) follow the same ROC curve (as they have the same AUC score); therefore, only step 5 is visible.

these five metabolites were insufficient to correctly classify all samples in this dataset.

Table 5.5: Group Sensitivity for Progressive Metabolite Addition, scored with Balanced Accuracy

Step	New Metabolite	Healthy Controls	PD Treated	PD Untreated
1	palmitoleamide (16:1)*	0.6585	0.9476	0.8621
2	myristoleamide (14:1)*	0.6341	0.9476	0.8621
3	3-methoxytyrosine	0.9024	0.9551	0.8621
4	3-methoxytyramine sulfate	0.8780	0.9551	0.8621
5	myristamide (14:0)*	0.8780	0.9551	0.8621

Values represent balanced accuracy for each group.

5. Results of aim 3 Mechanistic: Explore underlying disease pathways associated with metabolites most altered in PD samples

5.4 Removing top metabolites - exploring model prediction dependence

The top SHAP-ranked metabolites belonged primarily to two metabolic super-pathways: Fatty Acid, Amide and Tyrosine Metabolism. Dopamine and L-DOPA metabolism are key players in the tyrosine metabolism pathway (Meiser et al. 2013; Elroby et al. 2012). To assess the model’s reliance on these pathways, classification performance was evaluated when top metabolites from each super-pathway were systematically excluded from the feature set.

3 different filtered datasets were classified using an SVM

1. Removing the top 4 LDOPA-associated metabolites
2. Removing the top 6 fatty acid amide metabolites
3. Removing both top 4 LDOPA-associated metabolites + top 6 fatty acid amide metabolites

The SVM model employed was identical to the final model (linear kernel with the same hyperparameters), except that Boruta feature selection was rerun independently for each filtered dataset, and all confirmed features were retained.

5.4.1 Removing 4-LDOPA-associated metabolites

The LDOPA-associated metabolites were chosen to be removed by using a combination of the metabolites that appeared in the results of GLM 5: LEDD (Levodopa Equivalent Daily Dose) $\sim$ metabolite + gender + age and surrounding literature (Meiser et al. 2013), within the first 20 SHAP metabolites.

Removing 4 L-DOPA metabolites:

- 3-methoxytyrosine
- 3-methoxytyramine sulfate
- m-tyramine sulfate

5. Results of aim 3 Mechanistic: Explore underlying disease pathways associated with metabolites most altered in PD samples

- dopamine 3-O-sulfate

Original features: 1122

Filtered features: 1118 (removed 4)

Boruta Feature Selection with L-DOPA

Boruta feature selection was rerun on the dataset (Table 5.6), and within the top 10 features, seven fatty acid amides appeared (myristoleamide (14:1)*, margaramide (17:0)*, heptadecenamide (17:1)*, palmitoleamide (16:1)*, nonadecenamide (19:1)*, myristamide (14:0)*, and linolenamide (18:3)*), as well as two dopamine-related metabolites (vanillactate and dopamine 4-sulfate), which initially did not appear in the top 10 selection in the initial Boruta analysis.

Table 5.6: Top 10 Features by Random Forest Importance from Boruta Feature Selection (No L-DOPA Data)

Rank	Feature	Importance
1	Myristoleamide (14:1)*	0.1158
2	Margaramide (17:0)*	0.1113
3	Vanillactate	0.0805
4	Heptadecenamide (17:1)*	0.0687
5	Palmitoleamide (16:1)*	0.0662
6	X-12410	0.0607
7	Nonadecenamide (19:1)*	0.0595
8	Myristamide (14:0)*	0.0499
9	Dopamine 4-sulfate	0.0473
10	Linolenamide (18:3)*	0.0246

Note: Boruta analysis identified 57 confirmed features and 6 tentative features. Asterisks (*) indicate specific lipid notations.

5.4.2 Removing 6 Fatty Acid Amide Metabolites

The first 6 fatty acid amides were chosen from the top 20 SHAP metabolites, Removing top 6 Fatty Acid Amides:

- palmitoleamide (16:1)*

5. Results of aim 3 Mechanistic: Explore underlying disease pathways associated with metabolites most altered in PD samples

- myristoleamide (14:1)*
- myristamide (14:0)
- heptadecenamide (17:1)*
- linolenamide (18:3)*
- linoleamide (18:2n6)

Original features: 1122

Filtered features: 1116 (removed 6)

Boruta Feature Selection with Fatty Acid Amides Removed

Boruta feature selection was rerun with the top six fatty acid amides excluded (Table 5.7). Within the top 10 features, four alternative fatty acid amides were selected: margaramide (17:0)*, nonadecenamide (19:1)*, oleamide, and eicosenamide (20:1)*. Additionally, five dopamine-related metabolites appeared: 3-methoxytyramine sulfate, vanillactate, 3-methoxytyrosine, dopamine 3-O-sulfate, and dopamine 4-sulfate. This represented a greater proportion of tyrosine metabolism metabolites compared to the initial Boruta selection.

5.4.3 Both L-DOPA-associated metabolites and Fatty Acid Amides Removed

Removed all 10 metabolites, 4 L-DOPA-associated metabolites and 6 fatty acid amides, and then Boruta was re-run.

Boruta Feature Selection with both L-DOPA-associated Metabolites and Fatty Acid Amides Removed

Boruta feature selection was rerun with both the top L-DOPA-associated metabolites and top fatty acid amides excluded (Table 5.8). Within the top 10 features, four fatty acid amides were selected (margaramide (17:0)*, nonadecenamide (19:1)*, oleamide, and eicosenamide (20:1)*) alongside two dopamine-related metabolites

5. Results of aim 3 Mechanistic: Explore underlying disease pathways associated with metabolites most altered in PD samples

Table 5.7: Top 10 Features by Random Forest Importance from Boruta Feature Selection (No Amides Data)

Rank	Feature	Importance
1	Margaramide (17:0)*	0.1492
2	Nonadecenamide (19:1)*	0.1321
3	3-Methoxytyramine sulfate	0.0779
4	Vanillactate	0.0672
5	3-Methoxytyrosine	0.0612
6	X-12410	0.0463
7	Oleamide	0.0429
8	Dopamine 3-O-sulfate	0.0384
9	Dopamine 4-sulfate	0.0310
10	Eicosenamide (20:1)*	0.0297

Note: Boruta analysis identified 66 confirmed features and 3 tentative features. Asterisks (*) indicate specific lipid notations.

(vanillactate and dopamine 4-sulfate). Notably, three unidentified metabolites appeared (X-12410, X-11407, and X-18779), and for the first time, a metabolite from outside the fatty acid amide and tyrosine metabolism pathways was selected: 2-oleoylglycerol (18:1), a monoacylglycerol from lipid metabolism. This suggested that when key metabolites from the two dominant pathways were removed, the model compensated by selecting alternative metabolites from these pathways as well as expanding to other metabolic pathways.

Table 5.8: Top 10 Features by Random Forest Importance from Boruta Feature Selection (No L-DOPA + No Amides Data)

Rank	Feature	Importance
1	Margaramide (17:0)*	0.1629
2	Nonadecenamide (19:1)*	0.1065
3	Vanillactate	0.0888
4	X-12410	0.0812
5	Dopamine 4-sulfate	0.0584
6	Oleamide	0.0516
7	Eicosenamide (20:1)*	0.0404
8	X-11407	0.0233
9	2-Oleoylglycerol (18:1)	0.0193
10	X-18779	0.0192

Note: Boruta analysis identified 71 confirmed features and 5 tentative features. Asterisks (*) indicate specific lipid notations.

5. Results of aim 3 Mechanistic: Explore underlying disease pathways associated with metabolites most altered in PD samples

5.4.4 Summary

Detailed scores, group sensitivity metrics, and Boruta feature selection data are provided in Appendix A.3. A summary of these results is presented in Table 5.9. The removal of any subset of metabolites had minimal impact on classifier performance, except a modest decrease in balanced accuracy in the dataset excluding the top six amides (balanced accuracy decreased by 0.0362 compared to the original model). The balanced accuracy for the original model was 0.955, and the dataset that achieved the closest balanced accuracy of 0.9465 was the one excluding the top four L-DOPA-associated metabolites. The dataset excluding both L-DOPA-associated metabolites and fatty acid amides achieved a balanced accuracy of 0.9448. The lowest balanced accuracy was observed in the dataset excluding the top six fatty acid amides, with a score of 0.9188.

However, the number of features confirmed by Boruta increased progressively with each exclusion. In the original dataset, Boruta confirmed 52 features as important for predicting class, with 40 features yielding the best-performing model. After removing the top four L-DOPA-associated metabolites, Boruta confirmed 57 metabolites. After removing the top six fatty acid amide metabolites, Boruta confirmed 66 metabolites. When both L-DOPA-associated metabolites and fatty acid amide metabolites were removed, Boruta confirmed 71 metabolites. This increase in the number of confirmed features indicated that Boruta relied on weaker metabolite signals to compensate for the loss of more discriminatory metabolites.

Table 5.9: Performance Comparison of SVM Models with Different Feature Exclusions

Model	Features	AUC	Bal. Acc.	Δ AUC	Δ Bal. Acc.
Original	40	0.9884	0.955	—	—
No L-DOPA	57	0.9889	0.9465	+0.0005	−0.0085
No Amides	66	0.9843	0.9188	−0.0041	−0.0362
No L-DOPA + Amides	71	0.9857	0.9448	−0.0028	−0.0102

Δ values calculated relative to the Original model. Positive values indicate improvement.

5. Results of aim 3 Mechanistic: Explore underlying disease pathways associated with metabolites most altered in PD samples

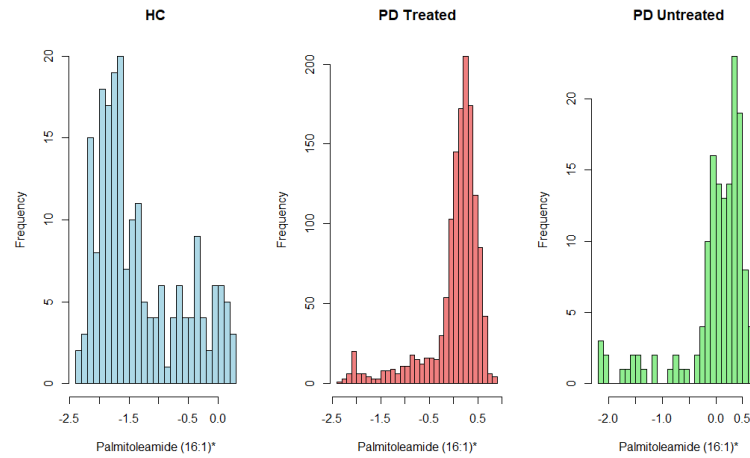
5.5 Histograms of the top 5 SHAP metabolites

Given that the dataset contained a large proportion of treated PD patients, it was anticipated that some drug metabolites would be present in the study. Dopamine is a central component of the tyrosine metabolism pathway (Meiser et al. 2013), and the top five SHAP-ranked metabolites included two linked to tyrosine metabolism (3-methoxytyrosine and 3-methoxytyramine) and three linked to fatty acid amide pathways (Palmitoleamide, Myristoleamide and Myristamide), all of which were elevated in PD samples according to the SHAP summary plot (see figure 5.1). To distinguish between drug-related and disease-related metabolites, the abundances of these metabolites were compared across treatment groups (treated PD, untreated PD) and healthy controls. Metabolites elevated exclusively in treated PD patients, but not in untreated PD or healthy controls, were considered potential drug metabolites. Conversely, metabolites elevated in both treated and untreated PD patients were considered potential disease markers.

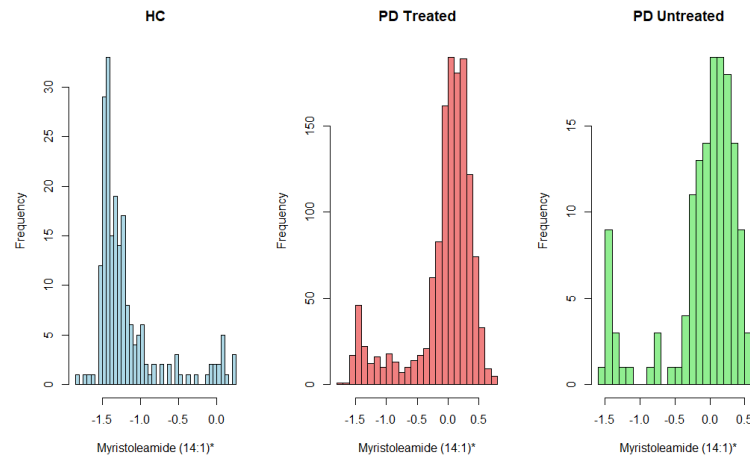
Figure 5.7, shows 3 fatty acid amides, all of which were elevated in PD-treated and PD-untreated histograms compared to controls, peaks centring around 0.5. Whereas in healthy controls, peaks centre around -2. These 3 fatty acid amides are elevated across both PD groups regardless of treatment status, when compared to controls.

Figure 5.8, tyrosine metabolism-related metabolites were elevated in PD-treated compared to both untreated and healthy controls. Peaks centre at 0 for both metabolites for treated PD, and around -2 (3-methoxytyrosine) and -1.5 (3-methoxytyramine sulfate) for untreated and healthy controls. These two tyrosine metabolism-related metabolites appear to be selectively increased in treated PD samples.

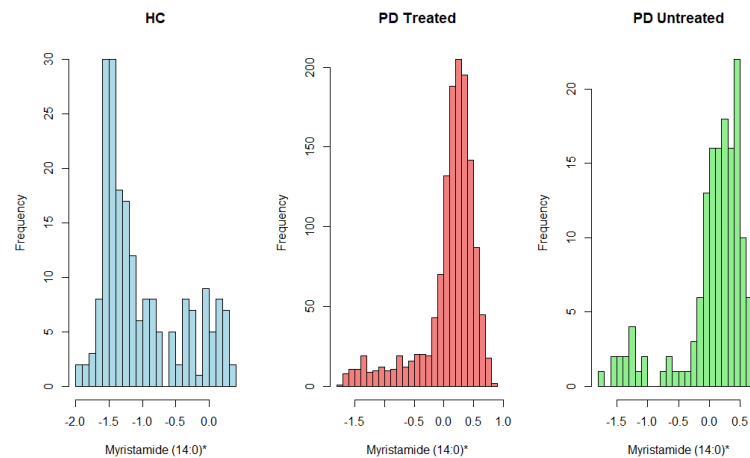
5. Results of aim 3 Mechanistic: Explore underlying disease pathways associated with metabolites most altered in PD samples



(a) Palmitoleamide (16:1)*



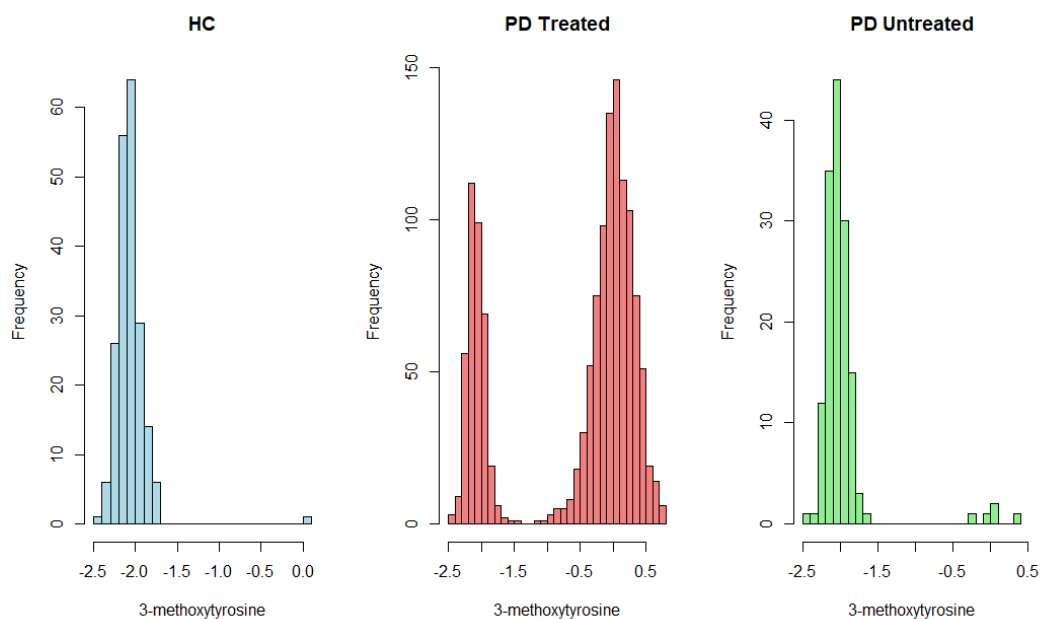
(b) Myristoleamide (14:1)*



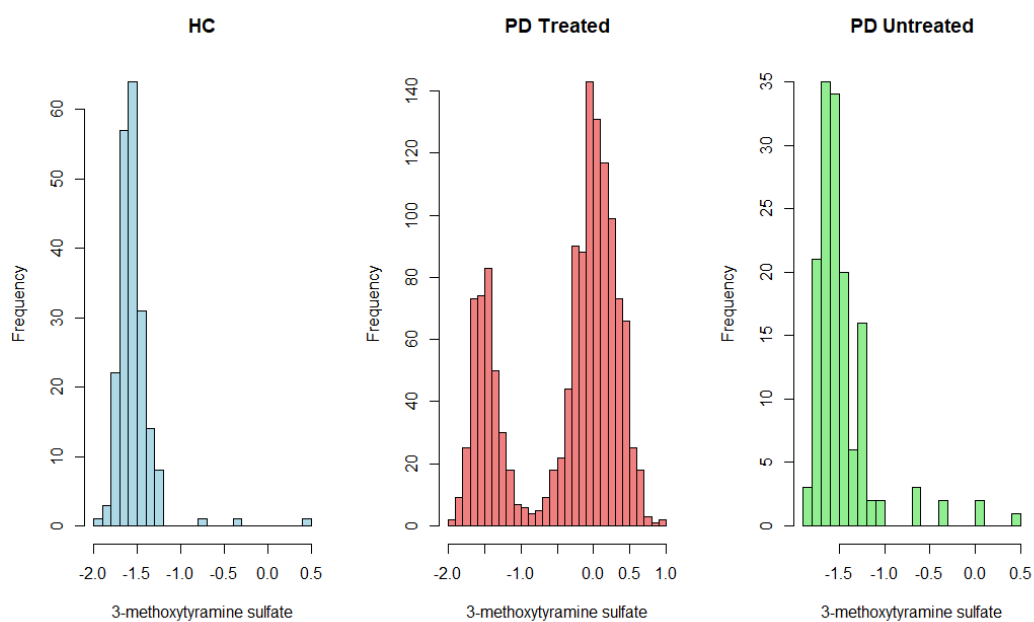
(c) Myristamide (14:0)*

Figure 5.7: Distribution of (a) Palmitoleamide, (b) Myristoleamide and (c) Myristamide abundances across HC (blue), PD Treated (red), and PD Untreated (green) groups. The x-axis represents log₁₀-transformed metabolite abundance, and the y-axis represents the frequency (number of samples). Note that the y-axis ranges differ across groups due to differing sample sizes, with substantially more treated PD patients in the dataset.

5. Results of aim 3 Mechanistic: Explore underlying disease pathways associated with metabolites most altered in PD samples



(a) 3-methoxytyrosine



(b) 3-methoxytyramine sulfate

Figure 5.8: Distribution of (a) 3-methoxytyrosine and (b) 3-methoxytyramine sulfate abundances across HC (blue), PD Treated (red), and PD Untreated (green) groups. The x-axis represents log₁₀-transformed metabolite abundance, and the y-axis represents the frequency (number of samples). Note that the y-axis ranges differ across groups due to differing sample sizes, with substantially more treated PD patients in the dataset.

5. Results of aim 3 Mechanistic: Explore underlying disease pathways associated with metabolites most altered in PD samples

5.6 Overlap results between Fold Change Analysis, GLM and SHAP feature importance

Across the three analyses—differential abundance analysis (LIMMA), generalised linear modelling (GLM), and SHAP feature importance—many of the same metabolites were identified as important. Exploring the overlap would help validate which metabolites were consistently most useful for identifying a disease signal. Each method assessed metabolites differently: differential abundance analysis ranked metabolites by fold change between groups and evaluated statistical significance (LIMMA), GLM ranked metabolites by statistical significance (adjusted p-values) while accounting for covariates, and SHAP ranked metabolites by their contribution to model predictions.

The 40 SHAP-ranked metabolites were compared against significantly altered metabolites identified by GLM2 (adjusted p-value < 0.05) and LIMMA (absolute fold change > 0.2 and adjusted p-value < 0.05). A three-way overlap was found across the methods (see table 5.10, with 21 metabolites in common).

Table 5.10: Overlap of significantly altered metabolites across SHAP, GLM2, and LIMMA analyses.

Category	N Metabolites
<i>Total metabolites per method (after filtering):</i>	
SHAP	40
GLM2	318
LIMMA	30
<i>Three-way overlap:</i>	
$\text{SHAP} \cap \text{GLM2} \cap \text{LIMMA}$	21
<i>Two-way overlaps:</i>	
$\text{SHAP} \cap \text{GLM2}$ only	19
$\text{SHAP} \cap \text{LIMMA}$ only	0
$\text{GLM2} \cap \text{LIMMA}$ only	9
<i>Unique to each method:</i>	
SHAP only	0
GLM2 only	269
LIMMA only	0

5. Results of aim 3 Mechanistic: Explore underlying disease pathways associated with metabolites most altered in PD samples

5.6.1 Three-way overlap metabolites

Table 5.6.1 contains all 21 overlapped metabolites ranked by their average rank across all three methods: SHAP (ranked by SHAP contribution), LIMMA (ranked by fold change) and GLM (ranked by adjusted p-value). Some summary notes from the table 5.6.1:

Most consistent across all methods (lowest rank variation):

- Palmitoleamide (16:1)*: SHAP = 1, GLM2 = 1, LIMMA = 2 (std = 0.6)
- Myristoleamide (14:1)*: SHAP = 2, GLM2 = 4, LIMMA = 6 (std = 2.0)
- Margaramide (17:0)*: SHAP = 9, GLM2 = 5, LIMMA = 5 (std = 2.3)
- Myristamide (14:0)*: SHAP = 5, GLM2 = 2, LIMMA = 7 (std = 2.5)
- Linoleamide (18:2n6): SHAP = 13, GLM2 = 10, LIMMA = 15 (std = 2.5)

Generally, all the top fatty acid amides were ranked highly across all the methods.

Biggest disagreements across methods (highest rank variation):

- Fibrinopeptide B (1-13)**: SHAP = 14, GLM2 = 64, LIMMA = 20 (std = 27.3)
- p-Cresol glucuronide*: SHAP = 19, GLM2 = 58, LIMMA = 21 (std = 22.0)
- Nonadecenamide (19:1)*: SHAP = 40, GLM2 = 6, LIMMA = 10 (std = 18.6)
- 3-Methoxytyrosine: SHAP = 3, GLM2 = 23, LIMMA = 1 (std = 12.2)
- Oleamide: SHAP = 25, GLM2 = 7, LIMMA = 8 (std = 10.1)

Two metabolites demonstrated particularly interesting ranking patterns: 3-methoxytyrosine and oleamide. 3-Methoxytyrosine showed strong agreement between SHAP and LIMMA, with both methods ranking it highly (SHAP rank 3, LIMMA rank 1), yet it was ranked substantially lower by GLM (rank 23). Oleamide exhibited relative agreement between GLM and LIMMA rankings but was ranked considerably lower by SHAP. These discrepancies likely reflect fundamental differences

5. Results of aim 3 Mechanistic: Explore underlying disease pathways associated with metabolites most altered in PD samples

in how each analytical approach identifies important metabolites: GLM detects population-wide statistical associations, LIMMA identifies metabolites with the largest relative abundance changes between groups, and SHAP quantifies which metabolites most effectively define the classification boundary between PD and HC samples. Consequently, metabolites may rank highly in one method but not others, depending on whether they exhibit strong statistical significance, large fold changes, or high discriminatory power for classification.

Metabolite	Super Path	Sub Path	SHAP Rank	SHAP Val	GLM2 Rank	GLM2 adj p	LIMMA Rank	LIMMA logFC	FC Dir	LIMMA adj p	Avg Rank	Rank Std
palmitoleamide (16:1)*	Lipid	Fatty Acid, Amide	1	2.44	1	1.1e-68	2	1.35	↑	6.0e-163	1.3	0.6
myristoleamide (14:1)*	Lipid	Fatty Acid, Amide	2	1.77	4	5.0e-62	6	1.12	↑	1.8e-159	4.0	2.0
myristamide (14:0)*	Lipid	Fatty Acid, Amide	5	0.48	2	3.0e-67	7	1.10	↑	1.5e-152	4.7	2.5
heptadecenamide (17:1)*	Lipid	Fatty Acid, Amide	11	0.29	3	3.0e-67	4	1.20	↑	4.1e-157	6.0	4.4
margaramide (17:0)*	Lipid	Fatty Acid, Amide	9	0.38	5	8.2e-62	5	1.18	↑	2.6e-140	6.3	2.3
3-methoxytyrosine	Amino Acid	Tyrosine Metabolism	3	1.44	23	1.1e-16	1	1.36	↑	2.1e-69	9.0	12.2
3-methoxytyramine sulfate	Amino Acid	Tyrosine Metabolism	4	1.42	16	5.3e-22	9	1.05	↑	1.4e-71	9.7	6.0
linolenamide (18:3)*	Lipid	Fatty Acid, Amide	12	0.28	8	1.3e-47	14	0.76	↑	2.5e-74	11.3	3.1
dopamine 3-O-sulfate	Amino Acid	Tyrosine Metabolism	15	0.22	9	3.4e-35	13	0.80	↑	2.8e-63	12.3	3.1
linoleamide (18:2n6)	Lipid	Fatty Acid, Amide	13	0.25	10	4.0e-35	15	0.76	↑	2.7e-55	12.7	2.5
X-12410	NA	NA	16	0.19	11	5.3e-34	11	0.87	↑	1.8e-62	12.7	2.9
vanillactate	Amino Acid	Tyrosine Metabolism	23	0.06	14	1.0e-22	3	1.33	↑	2.8e-66	13.3	10.0
oleamide	Lipid	Fatty Acid, Amide	25	0.05	7	3.6e-51	8	1.08	↑	3.3e-97	13.3	10.1
X-21733	NA	NA	7	0.41	15	2.4e-22	22	0.27	↑	4.7e-25	14.7	7.5
m-tyramine sulfate	Amino Acid	Tyrosine Metabolism	8	0.40	19	1.1e-17	23	0.26	↑	4.1e-25	16.7	7.8
dopamine 4-sulfate	Amino Acid	Tyrosine Metabolism	29	0.03	12	1.7e-33	12	0.80	↑	1.1e-59	17.7	9.8
eicosenamide (20:1)*	Lipid	Fatty Acid, Amide	26	0.04	13	7.9e-32	16	0.51	↑	2.0e-44	18.3	6.8
nonadecenamide (19:1)*	Lipid	Fatty Acid, Amide	40	0.00	6	2.1e-60	10	1.04	↑	3.5e-143	18.7	18.6
X-11407	NA	NA	31	0.03	21	1.7e-17	24	0.25	↑	8.5e-18	25.3	5.1
fibrinopeptide B (1-13)**	Peptide	Fibrin. Cleave. Pept.	14	0.23	64	1.4e-06	20	-0.28	↓	2.8e-05	32.7	27.3
p-cresol glucuronide*	Amino Acid	Tyrosine Metabolism	19	0.13	58	3.5e-07	21	0.28	↑	1.1e-11	32.7	22.0

Note: Fibrin. Cleave. Pept. = Fibrinogen Cleavage Peptide

5. Results of aim 3 Mechanistic: Explore underlying disease pathways associated with metabolites most altered in PD samples

5.6.2 Super Pathway Analysis

To identify which metabolic super pathways were most important across all analyses, the average rankings of metabolites within each super pathway were compared across SHAP, GLM2, and Fold change (Limma analysis) feature selection methods (Table 5.11). The Lipid super pathway contained the most metabolites (10) and demonstrated the lowest overall average rank of 9.7. The Amino Acid super pathway followed with 7 metabolites and an overall average rank of 15.9. The Peptide super pathway was represented by a single metabolite with an overall average rank of 32.7.

Table 5.11: Summary of metabolite rankings by super pathway. Average ranks across SHAP, GLM2, and LIMMA feature selection methods are shown for metabolites within each super pathway. Lower average ranks indicate higher importance.

Super Pathway	N Metabolites	Avg SHAP Rank	Avg GLM2 Rank	Avg LIMMA Rank	Overall Avg Rank
Lipid	10	14.4	5.9	8.7	9.7
Amino Acid	7	14.4	21.6	11.7	15.9
Peptide	1	14.0	64.0	20.0	32.7

5.6.3 Sub-Pathways

Examining the sub-pathway composition (Table 5.12), Fatty Acid, Amide contained the most metabolites (10) with an overall average rank of 9.7. Tyrosine Metabolism followed with 7 metabolites and an overall average rank of 15.9. Fibrinogen Cleavage Peptide was represented by a single metabolite with an overall average rank of 32.7.

Table 5.12: Summary of metabolite rankings by sub-pathway. Average ranks across SHAP, GLM2, and LIMMA feature selection methods are shown for metabolites within each sub-pathway. Lower average ranks indicate higher importance.

Sub Pathway	Super Pathway	N Met.	Avg SHAP Rank	Avg GLM2 Rank	Avg LIMMA Rank	Overall Avg Rank
Fatty Acid, Amide	Lipid	10	14.4	5.9	8.7	9.7
Tyrosine Metabolism	Amino Acid	7	14.4	21.6	11.7	15.9
Fibrinogen Cleavage Peptide	Peptide	1	14.0	64.0	20.0	32.7

5. Results of aim 3 Mechanistic: Explore underlying disease pathways associated with metabolites most altered in PD samples

5.7 Summary

Throughout this analysis, a consistent pattern emerged: fatty acid amides and tyrosine metabolism-associated metabolites played central roles in distinguishing PD from HC metabolic profiles. While methodological differences existed between the pathway enrichment analysis from GLM2 and the super-pathway analysis of the overlap, all three methods (SHAP, GLM2, and LIMMA) identified 21 metabolites as significantly altered, though with varying rankings. Five metabolites appeared consistently across analyses: palmitoleamide (16:1)*, myristoleamide (14:1)*, myristamide (14:0)*, 3-methoxytyrosine, and 3-methoxytyramine sulfate.

6

Discussion

Contents

6.1 Main findings	113
6.1.1 Widespread Metabolic Alterations Distinguish Parkinson's Disease from Healthy Controls	113
6.1.2 Metabolic alterations persisted across drug treatment	117
6.2 Biological and Mechanistic Insight	119
6.2.1 Fatty Acid Amide Pathways in PD	119
6.2.2 Tyrosine Metabolism in Dopamine Medication and PD	122
6.2.3 Fibrinogen Cleavage Peptide	125
6.3 Strengths of the study	126
6.4 Limitations	127
6.5 Implications	130
6.6 Future directions	130

6.1 Main findings

6.1.1 Widespread Metabolic Alterations Distinguish Parkinson's Disease from Healthy Controls

The PD patients in this study appeared to represent a relatively early disease phenotype, with the majority exhibiting mild to moderate motor (low UPDRS scores) and minimal cognitive impairment (high MoCA scores). However, UPDRS scores did show a large range, indicating some of the cohort experienced a far more severe

6. Discussion

motor dysfunction compared to others. PD-related medication was nearly universal among patients (92% of PD participants), though dosing varied substantially (LEDD range: 0 - 2049.5 mg/day, see table 3.1). This cohort composition has important implications for the interpretation of metabolic alterations: first, the absence of severe motor or cognitive burden may explain why metabolic differences may not be as pronounced as in other cohort studies, also likely reflecting a different set of metabolites which corresponds to this cohort PD stage. Second, the prevalence of levodopa treatment means that observed metabolic signatures likely reflect a combination of disease pathology and pharmacological effects, which cannot be fully disentangled in this study. Finally, apparent subgroup structure was observed in dimensionality reduction analyses (UMAP), though this clustering appeared to be primarily driven by treatment status rather than distinct disease subtypes. Given the recognised heterogeneity of PD, it remains possible that underlying pathophysiological subgroups exist but are obscured by the dominant treatment-related metabolic signal. While not checked extensively, age and sex did not seem to separate the data substantially in the clustering analyses, indicating that these demographic factors did not play as large of a confounding effect as drug treatment status, though this observation should be interpreted with caution, given that other metabolomic biomarker studies have reported more pronounced age and sex effects in their respective cohorts (Cannas et al. 2025; Ben-Shlomo et al. 2024).

Twenty-one metabolites were found in common across multiple analytical approaches (fold change analysis, GLM, and SHAP), separating PD from HC metabolic profiles, suggesting that there are robust biological differences between the two groups. Different analytical approaches appeared to capture different aspects of metabolic dysregulation. Population-level statistical models (Fold change and GLM) tended to identify metabolites showing consistent directional changes across the majority of PD patients, while classifier-based feature importance metrics (SHAP values) emphasised metabolites most discriminative at decision boundaries, potentially including those with a more heterogeneous profile.

6. Discussion

The overlap analysis revealed that the majority of discriminative metabolites belonged to two sub-pathways: Fatty Acid Amide metabolism (a lipid pathway) and Tyrosine metabolism (an amino acid pathway). Seventeen of the 21 metabolites belonged to one of these two sub-pathways, suggesting that metabolic alterations in PD affect these systems in a coordinated manner rather than representing isolated changes in individual metabolites. The interpretation of these metabolomic changes presents inherent challenges. Unlike genomic or proteomic alterations, individual metabolites typically participate in multiple biosynthetic and catabolic routes, may undergo reversible modifications, and can engage in cyclical pathways. Therefore, upregulation of a precursor metabolite does not necessarily result in proportional increases in all downstream products, further complicating inference of causal relationships (K. S. Lee et al. 2025). While the super-pathway and sub-pathway classifications provided by Metabolon facilitate the organisation of metabolites, these labels may oversimplify the cross-talk between metabolites and obscure the fact that many metabolites frequently participate in multiple interconnected pathways.

The final SVM classifier was able to confidently predict PD status in both treated and untreated PD samples (AUC = 0.988, balanced accuracy = 0.955), providing evidence of substantial metabolic differences between PD and HC groups, as well as evidence for the clinical utility of using serum metabolites for PD diagnosis. Further sensitivity analyses suggested that this discrimination was primarily driven by metabolite profiles rather than class imbalance, as the model performed equally well on an undersampled balanced population, as well as on different combinations of testing and training sets made up from different combinations of the subgroups. However, these sensitivity analyses were constrained by the limited sample sizes available for balanced resampling, particularly for untreated PD and HC groups (i.e. the model keeps seeing the same metabolite profiles for HC and untreated PD). A larger, balanced cohort would enable a more comprehensive assessment of metabolic heterogeneity within the wider population and provide greater confidence in the model's generalisability. While the observed performance indicates substantial

6. Discussion

metabolic alterations between PD and HC, the interpretation is nevertheless held back by the inherent class imbalance found in the data.

When removing the top metabolites from the classifier, this revealed redundancy within the top features. Elimination of top-ranked fatty acid amides and L-DOPA-associated metabolites resulted in minimal performance degradation (the largest AUC reduction is -0.0041, and the largest balanced accuracy reduction was -0.0362), with the classifier appearing to compensate through alternative metabolites within the same sub-pathway. This presents an interpretive challenge: while the top-ranked SHAP metabolites are presumably biologically relevant given their discriminative capacity, the classifier’s lack of dependence on these specific features suggests widespread metabolic alterations rather than isolated changes in a few key metabolites. This compensatory capacity may reflect genuine pathway-level dysregulation in PD, where disruption affects multiple related metabolites that can serve redundant classificatory roles. Alternatively, given the known heterogeneity of PD, the specific metabolites exhibiting the largest alterations may vary across individuals or cohorts, even while the affected pathways remain consistent. Thus, while these particular metabolites demonstrated the greatest discriminative value in this cohort, other studies might identify different but biochemically related metabolites as most prominent, depending on cohort characteristics and disease subtypes represented. This makes the top metabolites less suitable as a clinical biomarker panel.

The misclassification analysis revealed heterogeneity in metabolic profiles across subgroups. Although overall misclassification rates were low, errors were disproportionately concentrated among untreated PD and HC individuals. This is partly due to the smaller sample sizes of these subgroups, of which the model had less data to learn from. However, this does reinforce the idea that to create a classifier capable of generalising well, even larger and more balanced study cohorts are needed. In fact, one of the strengths of this study is the Tracking PD cohort’s sample size

6. Discussion

in comparison to other similar serum biomarker studies. Yet clinically applicable biomarker panels ultimately require substantially larger cohorts that can capture the full spectrum of metabolic heterogeneity within patient subgroups.

To summarise, widespread metabolic changes were observed between HC and PD patients, and it's observed across many related metabolites as opposed to a few key metabolites. While this makes these metabolites less clinically useful as panel markers, it does provide an interesting pathway to explore through a pathogenic lens.

6.1.2 Metabolic alterations persisted across drug treatment

Drug treatment as an important confounding factor has been extensively covered in the literature (Cannas et al. 2025), and with treated PD patients making up a substantial percentage of the sample size, drug metabolites were also detected. The separation of clusters observed in the UMAP analysis appeared to be primarily driven by treatment status (LEDD dosage), and the GLM analysis with LEDD as an outcome variable revealed that many metabolites showed a strong correlation with the dose of LEDD administered. An inevitable fact is that drug metabolites are digested and enter the serum alongside disease-related metabolites. Drug metabolites can also interact with the disease pathology itself, influencing its metabolic pathways, alongside affecting normal metabolic processes.

An interpretive challenge arises from the fact that the majority of PD patients in this study were receiving PD medication: these individuals are, by the very nature of taking treatment, already diagnosed, and therefore the classifier is not truly predicting PD status but rather confirming an existing diagnosis. Due to sample size constraints, the inclusion of treated PD patients was necessary as these made up a majority of the PD samples. However, when combining untreated PD and treated PD, the aim becomes to find universal metabolic alterations. These alterations, if present in both untreated and treated PD samples, could potentially indicate underlying disease pathology rather than treatment effects. The decision to split

6. Discussion

the data by subgroup and employ a binary classification approach aimed to address this problem by examining how the classifier performs on both types of PD patients. A three-way classification scheme would have likely led to an underpowered analysis where treated PD was classified predominantly on the basis of drug metabolites, while untreated PD would be classified on potential disease signals.

Histograms exploring the top metabolite (SHAP analysis) across all treatment groups were able to show that for the top primary fatty acid amides (PFAMs), they were reliably increased in both treated and untreated PD (decreased in HC), and the top tyrosine metabolism metabolites (which is associated with dopamine metabolism and, consequently, with L-DOPA metabolism) were only increased in the treated PD group. This suggests that the fatty acid amide pathway is elevated regardless of treatment status, potentially representing an underlying disease pathway. The final SVM classifier also had a very high balanced accuracy (which takes into account the accuracy per subgroup) of 0.955, which indicates the model was able to predict untreated PD with great success despite it being the smallest subgroup in sample size. When removing L-DOPA-associated metabolites from the classifier, the model was still able to score very highly, with the AUC even marginally increasing when L-DOPA metabolites were removed (not large enough to suggest that L-DOPA was inhibiting correct prediction), indicating the lack of dependence the model had on these metabolites. When removing metabolite sets, balanced accuracy saw the smallest decrease when L-DOPA-associated metabolites were removed. Boruta (feature selection) from when L-DOPA-associated metabolites were removed had a larger proportion of PFAMs than on the whole dataset. Although Boruta also chose other tyrosine metabolism-associated metabolites to compensate as well. However, the prominence of drug metabolite effects cannot be discredited, as demonstrated by the fact that tyrosine metabolism emerged as the second most common sub-pathway in the overlap analysis.

6. Discussion

Drug metabolites are inevitably important to distinguish PD samples due to the high rates of PD medication within this cohort; however, PFAMs may provide a universal disrupted pathway between treated and untreated PD.

6.2 Biological and Mechanistic Insight

6.2.1 Fatty Acid Amide Pathways in PD

Within the overlapping 21 metabolites identified across the three analyses (see Table 5.6.1), 10 were classified as PFAMs, with all 10 showing elevated levels in PD patients. This suggests a coordinated pattern across multiple structurally related lipid species, although this lipid class remains relatively underexplored. PFAMs are amide derivatives of fatty acids in which the carboxylic (COOH) group is converted into an amide (CONH₂) (ScienceDirect 2025). Their basic structure is as follows: R1-CO-NH₂ (amide functional group, where a carbonyl carbon is bonded to a nitrogen atom). They represent a structurally diverse family of lipid-derived molecules that are generally associated with roles in cell signalling and neuromodulation (Waluk et al. 2014). Many PFAMs have been detected in the central nervous system (CNS) (Waluk et al. 2014). The pathways for the biosynthesis of these molecules have not yet been completely elucidated but two pathways has been suggested: direct conversion from fatty acyl-CoA thioesters via ammonolysis, and indirect synthesis through oxidative cleavage of N-fatty acylglycine intermediates. Multiple biosynthetic mechanisms likely exist, as has been demonstrated for related fatty acid amide precursors (Farrell et al. 2012; Bradshaw et al. 2009). These molecules are primarily degraded by fatty acid amide hydrolase (FAAH), the principal enzyme responsible for their hydrolysis, yielding the parent fatty acid and ammonia (Farrell et al. 2012).

The PFAMs identified in this analysis included palmitoleamide, myristoleamide, myristamide, heptadecenamide, margaramide, linolenamide, linoleamide, oleamide, eicosenamide, and nonadecenamide. Of these, palmitoleamide, myristoamide, myristoleamide, linolenamide, linoleamide, oleamide, and eicosenamide are known or

6. Discussion

predicted substrates of FAAH (Dale L Boger et al. 2000). This suggests a common degradation pathway may be affected in PD. However, the specific biological functions of most of these PFAMs remain largely uncharacterized in the literature.

Oleamide is among the best-characterized members of this lipid class. It has been implicated in multiple biological processes, including sleep regulation, neurotransmission, and cell communication (D. L. Boger et al. 1998). To date, no studies have specifically examined oleamide levels in PD cohorts. However, recent metabolomic analyses of Alzheimer’s disease (AD) have identified elevated oleamide concentrations in plasma and serum from patients with mild cognitive impairment (MCI) (Oka et al. 2024; Gómez-Pascual et al. 2024). The authors of these studies interpreted the elevation as potentially representing a compensatory neuroprotective response to AD. Consistent with this interpretation, experimental studies have demonstrated neuroprotective and anti-inflammatory properties of oleamide in animal models of epilepsy (Nam et al. 2017; R. M. Ibrahim et al. 2023). And a further study extracted oleamide from Akarkara root (*Anacyclus pyrethrum* L.) as a pharmacological agent which exhibited neuroprotective effects in cellular models (R. M. Ibrahim et al. 2023).

Palmitoleamide, while not directly linked to PD, was also found to be increased in AD MCI plasma and serum, with higher rates related to slower rates of conversion from MCI to AD (Oka et al. 2024), suggesting some role of neuroprotection. Myristamide administration has been reported to improve cognitive outcomes in a mouse model of high-fat diet-induced cognitive decline; however, this finding is based on a single study with limited sample sizes (n=5-11) and warrants cautious interpretation (Kawano et al. 2024). Linoleamide and eicosnamide (commonly known as erucamide) found within the plant Ethanolic extract of *Erigeron bonariensis*, improved cognition and reduced Alzheimer-like pathology in ovariectomized D-galactose-treated rats (W. W. Ibrahim et al. 2024). The remaining PFAMs identified in this analysis have not been characterized in neurodegenerative disease context.

6. Discussion

The existing literature on these amides is sparse and predominantly focused on Alzheimer’s disease rather than PD, limiting the conclusions that can be drawn.

Fatty acid amides are biosynthetically derived from their parent fatty acids through enzymatic pathways that remain incompletely characterized (Farrell et al. 2012) as mentioned above. However fatty acids have been implicated in past PD serum biomarker studies with large meta-analyses of PD serum biomarker studies noting that many of studies saw a significant difference in fatty acid levels between controls and PD (Luo et al. 2024; Cannas et al. 2025). Notably, Luo et al. (2024) reviewed metabolic biomarkers in PD and identified several studies (Schulte et al. 2016; Trupp et al. 2014; Vascellari et al. 2020; Gonzalez-Riano et al. 2021) demonstrating decreased oleic acid levels in PD patients. Similarly, linoleic acid was reported as decreased in PD samples (Schulte et al. 2016). However, this project did not observe corresponding decreases in parent fatty acids in this dataset. Several factors may contribute to this discrepancy. First, differences in disease stage, dopamine medication use, other medications, demographic characteristics, and comorbidities between cohorts may influence metabolite profiles. Second, this cohort had a notably imbalanced design with 203 controls and 1,482 PD samples, which may have limited statistical power to detect differences in parent fatty acids. Third, variations in analytical platforms, sample handling, and metabolite extraction protocols can influence detection and quantification of these lipids.

Importantly, the elevation of PFAMs in the absence of decreased parent fatty acids in this dataset suggests that amide accumulation in this cohort may not simply reflect enhanced biosynthesis from fatty acid precursors (which would be expected to produce a detectable decrease in serum fatty acid levels). This pattern could instead point toward alternative mechanisms such as impaired degradation or altered regulatory pathways. The primary degradation pathway for PFAMs involves hydrolysis by fatty acid amide hydrolase (FAAH), a membrane-bound enzyme belonging to the serine hydrolase family. FAAH is thought to play a

6. Discussion

key role in terminating PFAM signalling in both the central nervous system and periphery (Fezza et al. 2023). Given that most of the identified PFAMs are known or predicted FAAH substrates, altered FAAH activity represents one plausible mechanistic explanation for the observed elevations. Evidence from PD animal models has shown that FAAH inhibitors can ameliorate motor dysfunction, reduce glial activity, and prevent dopaminergic neuron loss (Fezza et al. 2023; Soti et al. 2022), though contradictory findings regarding the neuroprotective capacity of FAAH inhibition have also been reported (Fezza et al. 2023; Soti et al. 2022). Whether the elevated PFAMs observed in this study reflect a compensatory neuroprotective response, impaired FAAH activity, or alternative regulatory mechanisms cannot be determined from this cross-sectional dataset alone.

6.2.2 Tyrosine Metabolism in Dopamine Medication and PD

Out of the 21 metabolites identified across the three analyses (see Table 5.6.1), 7 were associated with tyrosine and dopamine metabolism, with all 7 showing elevated levels in PD patients. These metabolites included dopamine 3-O-sulfate, dopamine 4-O-sulfate, 3-methoxytyramine sulfate, 3-methoxytyrosine, vanillactate, m-tyramine sulfate, and p-cresol glucuronide. This pattern likely reflects a combination of therapeutic intervention effects and disease-related metabolic alterations, as dopaminergic medications are used exclusively by the PD cohort and would not be present in age-matched healthy controls. Of the 7 metabolites, 5 metabolites are most likely due to dopamine-related medication: dopamine 3-O-sulfate, dopamine 4-O-sulfate, 3-methoxytyramine sulfate, 3-methoxytyrosine, vanillactate.

In PD, progressive degeneration of dopaminergic neurons in the substantia nigra pars compacta results in striatal dopamine depletion and the characteristic motor symptoms of the disease (Morris et al. 2024). Because dopamine cannot cross the blood-brain barrier, exogenous L-DOPA, L-3,4-dihydroxyphenylalanine (commonly

6. Discussion

sold as Levodopa) is used as a primary medication for PD (Elroby et al. 2012). L-DOPA serves as the immediate precursor to dopamine in the catecholamine synthesis pathway. The canonical (endogenous) pathway proceeds from the amino acid L-tyrosine, which is hydroxylated by tyrosine hydroxylase (TH) to form L-DOPA, followed by decarboxylation by aromatic amino acid decarboxylase (AADC) to yield dopamine. Dopamine itself functions as both a neurotransmitter and as a precursor to norepinephrine and epinephrine (Meiser et al. 2013). The conversion of LDOPA to dopamine can occur outside of the nervous system, and can occur within serum which may be why they are being detected at such high levels.

When L-DOPA is administered, it is metabolised not only to dopamine, but also by other pathways that produce metabolite byproducts, including vanilactate (vanillic acid) (Wassenberg et al. 2010). Dopamine is metabolised through multiple enzymatic pathways, with catechol-O-methyltransferase (COMT) and monoamine oxidase (MAO-B) representing the two degradation pathways (Meiser et al. 2013). COMT catalyzes O-methylation of dopamine to form 3-methoxytyramine, which can then undergo sulfation to produce 3-methoxytyramine sulfate (Goldstein 2010) (elevated in this cohort). COMT can also act on tyrosine itself to produce 3-methoxytyrosine (elevated in this cohort), representing an alternative methylation pathway (Beauchamp et al. 2025). MAO-B catalyzes oxidative deamination of dopamine to produce 3,4-dihydroxyphenylacetic acid (DOPAC), which is subsequently O-methylated to homovanillic acid (HVA). Additionally, phase II conjugation reactions, particularly sulfation by sulfotransferases (SULT), represent important elimination pathways for dopamine and its metabolites. Direct sulfation of dopamine produces dopamine 3-O-sulfate and dopamine 4-O-sulfate (positional isomers) (Meiser et al. 2013), both of which were elevated in PD samples.

The PD participants were also taking other dopamine related medication aside from Levodopa, including dopamine agonists (which act on dopamine receptors), MAO-B inhibitors (which block oxidative deamination), COMT inhibitors (which

6. Discussion

block O-methylation) and amantadine (to increase endogenous dopamine biosynthesis pathways). The majority of patients in this cohort were likely receiving combinations of levodopa with these related agonists and inhibitors, leading to different levels of different LDOPA and dopamine degradation related metabolites. The disease state of these metabolites are hard to measure as they are inevitable changed due to drug medication, a further future analysis could be to observe these metabolite levels in drug naive participants.

Increases in levodopa metabolites in PD biological samples are very commonly reported. Luo et al. identified seven studies showing replicated increases in 3-methoxytyrosine levels in PD samples. While 3-methoxytyramine sulfate was not identified in any PD biomarker studies, its precursor 3-methoxytyramine was found to be elevated in two studies (Paul et al. 2023; Dahabiyeh et al. 2024) reviewed by Cannas et al. (2025). While vanillic acid, dopamine 3-O-sulfate, and dopamine 4-O-sulfate have not been reported in PD biomarker studies, the established pathways of L-DOPA and dopamine biosynthesis and degradation suggest these are plausible levodopa-related drug metabolites.

m-Tyramine (meta-tyramine, 3-hydroxyphenethylamine) is both an exogenous compound found in fermented foods such as cheese and wine, and an endogenous trace amine in the brain, with particularly high levels in the basal ganglia (the primary site of PD pathology (Morris et al. 2024)) and limbic system. The enzyme cytochrome P450 2D6 (CYP2D6) can convert m-tyramine to dopamine via 3-hydroxylation, although this is not considered a primary pathway for dopamine production (X. Wang et al. 2014). Additionally, m-tyramine may arise as a gut bacterial metabolite in levodopa-treated patients, where L-DOPA is sequentially converted to dopamine by gut bacteria and subsequently dehydroxylated to m-tyramine (Maini Rekdal et al. 2019). The presence of elevated m-tyramine in this PD cohort, which is primarily composed of L-DOPA-treated patients, strongly

6. Discussion

suggests that this finding represents a medication-related artifact rather than an endogenous feature of PD pathophysiology.

p-Cresol glucuronide is a conjugated metabolite derived from p-cresol (4-methylphenol), a volatile organic compound characterized by a phenolic ring with a methyl group at the para position. p-Cresol is primarily produced through the microbial metabolism of tyrosine by intestinal anaerobic bacteria and demonstrates the capacity to cross both the intestinal epithelium and blood-brain barrier, which accounts for its detection in peripheral circulation. p-Cresol glucuronide has been reported as significantly elevated in PD serum samples (Paul et al. 2023), as noted in the meta-analysis by Cannas et al. (2025). Additionally, a structurally related conjugate, p-cresol sulfate, has been reported as increased in one serum study (Schulte et al. 2016) and three plasma studies reviewed by Luo et al. (2024). The present analysis also identified elevated levels of this metabolite in PD serum samples, consistent with these prior findings. The mechanistic basis for elevated p-cresol conjugates in PD remains incompletely understood. Current hypotheses centre on the gut-brain axis in PD pathogenesis, with evidence suggesting that pathological processes for some patients may originate in the gastrointestinal tract and subsequently disrupt the local microbial ecosystem (Morris et al. 2024). This disruption may result in altered microbial metabolite profiles, including increased production or altered metabolism of p-cresol and its conjugates.

6.2.3 Fibrinogen Cleavage Peptide

Fibrinopeptide B (FPB) is a small peptide fragment cleaved from fibrinogen by thrombin during the coagulation cascade and circulating FPB levels reflect ongoing thrombin activity and fibrin formation (Meeks 2009). In this cohort, a decrease in FPB was found compared to healthy controls. Decreased fibrinopeptide B (FPB) levels likely reflect reduced thrombin-mediated cleavage of fibrinogen, which could arise through multiple mechanisms (Kattula et al. 2017). These include anticoagulant therapy or coagulation factor deficiencies that reduce thrombin

6. Discussion

generation, decreased fibrinogen substrate availability (as fibrinogen is synthesised primarily in hepatocytes and increases during acute inflammation), or overall reduced coagulation cascade activation (Kattula et al. 2017). Notably, this cohort may include patients taking anticoagulant medications such as dabigatran (a direct thrombin inhibitor), rivaroxaban (a Factor Xa inhibitor), and warfarin (a vitamin K antagonist), all of which would be expected to reduce thrombin generation or activity and consequently decrease FPB production. However, without detailed medication histories for PD patients or medication data for healthy controls, whether the observed FPB decrease represents anticoagulant effects, endogenous coagulation alterations, or both cannot be conclusively determined.

6.3 Strengths of the study

This study benefits from several methodological strengths that enhance the reliability and interpretability of its findings. The cohort represents one of the larger metabolomic investigations of PD reported in recent literature, with 1,482 PD samples and 203 healthy control samples. This sample size is notably larger, particularly than those reported in comparable serum metabolomic studies: the largest serum study reviewed by Cannas et al. (2025) included 642 PD samples and 277 healthy control samples (Paul et al. 2023), while the largest study analysed by Luo et al. (2024) included 82 PD samples and 1272 controls (Schulte et al. 2016). To put into context of other PD biomarker cohorts, the largest cohort of any biosample type reported by Luo et al. (2024) was a study which included 639 PD samples and 109,151 control samples (Shao et al. 2021). So while the Tracking PD cohort is not the largest PD blood metabolite cohort, it is the study with the most PD samples across all 90 studies reviewed by Cannas et al. (2025) and Luo et al. (2024). The increased sample size in the present study provides greater statistical power to detect metabolic differences and may better capture the heterogeneity inherent to PD populations, though this study has a very small control sample size.

6. Discussion

This project employed an integrated framework that progressed from exploratory analyses through validation to mechanistic interpretation, allowing for metabolites to be validated across multiple statistical and machine learning methods. The use of an overlap analysis to identify metabolites that were consistently discriminative across fold change analysis, generalised linear modelling, and SHAP feature importance provided a more robust set of metabolic changes. The combination of different methods, which have different areas of focus (population-wide to edge-case individuals), allows for a more holistic view of metabolic changes in serum samples.

Furthermore, the inclusion of both treated and untreated PD samples and the analyses that investigated their relationship allowed for the exploration of metabolic alterations that persist across treatment status. This allowed for the identification of potential disease-related metabolic signatures despite the treatment status of many of the participants. Removing all L-DOPA-associated metabolites before all analysis could have hidden certain relationships between the metabolites. Though the small size of the untreated PD subgroup means that any conclusions regarding treatment-independent effects must be drawn with appropriate caution.

6.4 Limitations

Several limitations must be acknowledged when interpreting the findings of this study. The substantial class imbalance between PD patients and healthy controls represents a significant constraint, with 1,482 PD samples compared to only 203 control samples. This imbalance may have limited the statistical power to detect certain metabolic differences in controls and could have influenced the classifier’s performance metrics, despite efforts to address this through class weighting and sensitivity analyses. It likely affected all the analyses, and it may be why, in the volcano plot (fold change) analysis and GLM, far more metabolites are upregulated in PD samples as opposed to the number of metabolites which were downregulated. This might be due to the small HC population not being representative enough of a wider healthy population, and therefore, potential downregulated (in PD)

6. Discussion

metabolites are not captured in this dataset or not reaching significance amongst the small sample size. However, it is also possible that within this PD group and potentially wider PD pathology, there is more of an accumulation of metabolites as opposed to depletion in metabolic dysfunction.

This study design does not include other neurodegenerative or neurological disease samples, which limits the ability to assess the specificity of the identified metabolic signatures to PD. Without such comparisons, it remains unclear whether the observed metabolic alterations are unique to PD pathology or reflect broader features of neurodegeneration or chronic disease states (such as chronic inflammation). Further studies using other disease groups as negative controls would be of great importance.

Although drug treatment interference has been addressed in earlier sections, it remains a fundamental limitation of this project: the majority of PD patients were already diagnosed and receiving treatment at the time of sample collection, meaning that the observed metabolic profiles likely reflect a complex mixture of disease pathology and pharmacological effects. The presence of L-DOPA and other dopaminergic medication metabolites in the serum introduces considerable interpretive challenges, as it becomes difficult to fully disentangle which metabolic changes represent underlying disease mechanisms and which are secondary to therapeutic intervention. While the inclusion of a small untreated PD subgroup provided some opportunity to explore treatment-independent effects, the limited sample size of this group restricts the confidence with which such inferences can be made. Other drug metabolite effects were also seen in the data, Fibrinopeptide B (FPB) was found to be consistently decreased in PD samples. While not conclusively linked to anti-coagulant use, understanding the co-morbidities that PD patients face (Ben-Shlomo et al. 2024), particularly late-onset PD patients; this is a likely possibility.

6. Discussion

This study also faced technical limitations related to metabolite annotation. Some metabolites in this dataset lacked complete biochemical identification (with the overlap analysis, 3 consistently upregulated metabolites were unidentified), and in other cases, metabolites may be referred to by different names across studies (eicosenamide is also known as erucamide) or may be incompletely characterised. This can complicate cross-study comparisons and pathway interpretations, particularly for less well-studied metabolite classes such as the primary fatty acid amides identified in this analysis.

This project focuses on the baseline data provided by the Tracking PD cohort, while not inherently a limitation, this means the results from this project cannot be used to suggest metabolic changes in disease progression, treatment response or even as a disease risk biomarker (as this is not a prodromal cohort). Additionally, genetic information on this cohort, while collected in the original study, was not available for this analysis, limiting the understanding of the potential underlying heterogeneity found within the PD samples. LRRK2 PD is thought to have a different disease phenotype than sporadic late-onset PD, and both PD forms can manifest past the age of 50 (definition of late-onset) (Morris et al. 2024). Further confounding factors such as the samples being non-fasting state blood samples, lack of body mass index (BMI) data, lack of comorbidity data and incomplete no-PD medication data, decrease the reliability of the findings and potentially translatability.

Finally, it should be noted that the Tracking Parkinson’s cohort was established using the Queen Square Brain Bank criteria and collected information based on PD literature from 2015, which predates the current emphasis in PD research on disease heterogeneity and subtype classification. Contemporary PD research increasingly recognises the importance of stratifying patients within PD cohorts, focusing on biological markers of synuclein, neurodegeneration patterns and other misfolded proteins (Höglinger et al. 2024; Dam et al. 2024), of which was not included into Tracking PD’s study design.

6.5 Implications

This project was able to find widespread pathway level disruptions of metabolites in PD serum samples, with 20 metabolites showing consistently elevated levels (1 showing consistently decreased levels) across multiple statistical methods. This suggests that rather than a panel of stand-alone clinical biomarkers, this research suggests potentially targeting primary fatty acid amides pathways as a therapeutic target. The PFAMs were elevated in both treated and untreated PD, indicating that these metabolites might reflect a disease mechanism, and likely a compensatory one. Some of these PFAMs (oleamide, linoleamide and eicosenamide) have already been studied as potential therapeutic agents in AD models which shows potential (R. M. Ibrahim et al. 2023; W. W. Ibrahim et al. 2024). Over half of the PFAMs found are all hydrolysed by the same enzyme FAAH and FAAH inhibition is already used in PD studies as a potential therapeutic target. While this project is solely a pilot project modelling the baseline metabolic data from the Tracking PD cohort, there appears to be decent potential for further research.

6.6 Future directions

First, smaller analyses which can support the work presented in this thesis include removing all PFAMs (particularly those known to be or predicted to be hydrolysed by FAAH) from the data and rerunning classification to see if model performance drops drastically once the entire pathway is removed. As well as running analyses to suggest thresholds for PFAMs levels which discriminate between PD and HC.

Larger, future studies include rerunning the analyses in other HC and PD cohorts as external validation to understand whether this pattern replicates in external cohorts. Then including serum sample data from other similar neurodegenerative diseases to see how specific the signal from PFAMs are to PD. If a strong signal is established, potentially investigating FAAH activity in PD models may also be of interest. Then to see if PFAMs are of use as diagnostic biomarker, running

6. Discussion

analyses within drug-naive/prodromal PD cohorts. Tracking PD cohort also contains longitudinal metabolic data and exploring how PFAMs correlate with disease progression would further be of scientific value. Finally, Tracking PD cohort also provides genetic and proteomic data on the same participants as the metabolic cohort, in which PFAMs signals could be integrated into changes seen at the protein and genetic level, to be able to provide a more robust clinical biomarker panel but also a coherent narrative of PD pathogenesis.

Appendices

A

Appendix A

Contents

A.1 Subpathways provided by Metabolon	133
A.1.1 Super Pathways	133
A.1.2 Sub Pathways	134
A.2 Confounding Factors Analysis: Results from General	
Linear Model Analysis	135
A.2.1 GLM1: PD $\sim$ metabolite	135
A.2.2 GLM4: LEDD $\sim$ metabolite	137
A.3 Removing Metabolites Analysis	139
A.3.1 Removing L-DOPA associated Metabolites	139
A.3.2 Removing top 6 Fatty Acid Amides	139
A.3.3 Removing Both	140

A.1 Subpathways provided by Metabolon

A.1.1 Super Pathways

Amino Acid, Carbohydrates, Cofactors and Vitamins, Energy, Lipid, Nucleotide,
Partially Characterised Molecules, Peptide, Xenobiotic, Blank (unknown)

A.1.2 Sub Pathways

Acetylated Peptides, Alanine and Aspartate Metabolism, Aminosugar Metabolism, Androgenic Steroids, Ascorbate and Aldarate Metabolism, Bacterial/Fungal, Benzoate Metabolism, Carnitine Metabolism, Ceramide PEs, Ceramides, Chemical, Corticosteroids, Creatine Metabolism, Diacylglycerol, Dihydroceramides, Dihydro-sphingomyelins, Dipeptide, Dipeptide Derivative, Disaccharides and Oligosaccharides, Docosanoid, Drug - Cardiovascular, Drug - Topical Agents, Eicosanoid, Endocannabinoid, Fatty Acid Metabolism (Acyl Carnitine, Dicarboxylate), Fatty Acid Metabolism (Acyl Carnitine, Hydroxy), Fatty Acid Metabolism (Acyl Carnitine, Long Chain Saturated), Fatty Acid Metabolism (Acyl Carnitine, Medium Chain), Fatty Acid Metabolism (Acyl Carnitine, Monounsaturated), Fatty Acid Metabolism (Acyl Carnitine, Polyunsaturated), Fatty Acid Metabolism (Acyl Carnitine, Short Chain), Fatty Acid Metabolism (Acyl Choline), Fatty Acid Metabolism (Acyl Glutamine), Fatty Acid Metabolism (Acyl Glycine), Fatty Acid Metabolism (also BCAA Metabolism), Fatty Acid, Amide, Fatty Acid, Amino, Fatty Acid, Branched, Fatty Acid, Dicarboxylate, Fatty Acid, Dihydroxy, Fatty Acid, Monohydroxy, Fibrinogen Cleavage Peptide, Food Component/Plant, Fructose, Mannose and Galactose Metabolism, Gamma-glutamyl Amino Acid, Glutamate Metabolism, Glutathione Metabolism, Glycerolipid Metabolism, Glycine, Serine and Threonine Metabolism, Glycogen Metabolism, Glycolysis, Gluconeogenesis, and Pyruvate Metabolism, Guanidino and Acetamido Metabolism, Hemoglobin and Porphyrin Metabolism, Hexosylceramides (HCER), Histidine Metabolism, Inositol Metabolism, Ketone Bodies, Lactosylceramides (LCER), Lactoyl Amino Acid, Leucine, Isoleucine and Valine Metabolism, Long Chain Monounsaturated Fatty Acid, Long Chain Polyunsaturated Fatty Acid (n3 and n6), Long Chain Saturated Fatty Acid, Lysine Metabolism, Lysophospholipid, Lysoplasmalogen, Medium Chain Fatty Acid, Methionine, Cysteine, SAM and Taurine Metabolism, Mevalonate Metabolism, Modified Peptides, Monoacylglycerol, Nicotinate and Nicotinamide Metabolism, Oxidative Phosphorylation, Pantothenate and CoA Metabolism, Partially Characterized Molecules, Pentose Metabolism, Phenylalanine Metabolism,

A. Appendix A

Phosphatidylcholine (PC), Phosphatidylethanolamine (PE), Phosphatidylinositol (PI), Phospholipid Metabolism, Plasmalogen, Polyamine Metabolism, Polypeptide, Pregnenolone Steroids, Primary Bile Acid Metabolism, Progestin Steroids, Purine and Pyrimidine Metabolism, Purine Metabolism, (Hypo)Xanthine/Inosine containing, Purine Metabolism, Adenine containing, Purine Metabolism, Guanine containing, Pyrimidine Metabolism, Cytidine containing, Pyrimidine Metabolism, Orotate containing, Pyrimidine Metabolism, Thymine containing, Pyrimidine Metabolism, Uracil containing, Secondary Bile Acid Metabolism, Short Chain Fatty Acid, Sphingolipid Synthesis, Sphingomyelins, Sphingosines, Sterol, TCA Cycle, Tocopherol Metabolism, Tryptophan Metabolism, Tyrosine Metabolism, Urea cycle; Arginine and Proline Metabolism, Vitamin A Metabolism, Vitamin B6 Metabolism, Xanthine Metabolism.

A.2 Confounding Factors Analysis: Results from General Linear Model Analysis

A.2.1 GLM1: $PD \sim \text{metabolite}$

Table A.1: Top 10 metabolites from GLM 1: Case/control $\sim$ metabolite (logistic regression), ranked by adjusted p -value

Metabolite	Effect		Statistics		Pathway Information	
	β (SE)	p	$p_{\text{BH-adj}}$	Sig.	Super Pathway	Sub Pathway
palmitoleamide (16:1)*	2.037 (0.110)	3.52e-76	3.96e-73	***	Lipid	Fatty Acid, Amide
myristamide (14:0)*	2.472 (0.136)	3.45e-74	1.94e-71	***	Lipid	Fatty Acid, Amide
heptadecenamide (17:1)*	2.435 (0.134)	6.89e-74	2.58e-71	***	Lipid	Fatty Acid, Amide
myristoleamide (14:1)*	2.926 (0.168)	8.98e-68	2.52e-65	***	Lipid	Fatty Acid, Amide
margaramide (17:0)*	2.738 (0.158)	1.73e-67	3.89e-65	***	Lipid	Fatty Acid, Amide
nonadecenamide (19:1)*	2.952 (0.172)	8.08e-66	1.51e-63	***	Lipid	Fatty Acid, Amide
oleamide	1.451 (0.091)	3.22e-57	5.17e-55	***	Lipid	Fatty Acid, Amide
linolenamide (18:3)*	1.913 (0.125)	1.45e-52	2.03e-50	***	Lipid	Fatty Acid, Amide
dopamine 3-O-sulfate	2.074 (0.153)	4.48e-42	5.59e-40	***	Amino Acid	Tyrosine Metabolism
linoleamide (18:2n6)	1.269 (0.095)	2.20e-40	2.48e-38	***	Lipid	Fatty Acid, Amide

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

A. Appendix A

A.2.2 GLM4: LEDD $\sim$ metabolite

Table A.2: GLM 1: LEDD $\sim$ metabolite (linear regression)

Metabolite	Effect		Statistics		Pathway Information	
	β (SE)	p	$p_{\text{BH-adj}}$	Sig.	Super Pathway	Sub Pathway
3-methoxytyrosine	123.519 (3.801)	1.91e-176	2.14e-173	***	Amino Acid	Tyrosine Metabolism
vanillactate	121.763 (3.799)	9.60e-173	5.39e-170	***	Amino Acid	Tyrosine Metabolism
3-methoxytyramine sulfate	150.360 (5.284)	3.27e-143	1.22e-140	***	Amino Acid	Tyrosine Metabolism
X-12410	164.703 (6.064)	7.84e-133	2.20e-130	***	NA	NA
dopamine 4-sulfate	172.774 (6.523)	1.75e-127	3.93e-125	***	Amino Acid	Tyrosine Metabolism
dopamine 3-O-sulfate	179.763 (6.808)	7.36e-127	1.38e-124	***	Amino Acid	Tyrosine Metabolism
X-18779	407.901 (19.335)	6.81e-87	1.09e-84	***	NA	NA
N-acetyl-isoptureanine	-382.032 (22.215)	1.02e-60	1.43e-58	***	Amino Acid	Polyamine Metabolism
acisoga	-315.292 (22.283)	8.43e-43	1.05e-40	***	Amino Acid	Polyamine Metabolism
X-12726	186.958 (15.759)	4.08e-31	4.57e-29	***	NA	NA

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

A.3 Removing Metabolites Analysis

A.3.1 Removing L-DOPA associated Metabolites

When using this dataset, the SVM model managed to score a very high AUC at 0.9889, which is similar to the full dataset score .

Table A.3: SVM Classification Performance (No L-DOPA Data)

Overall Performance Metrics	
Test AUC	0.9889
Balanced Accuracy	0.9465
Number of Features	57
Group-wise Performance	
Group (n)	Performance
Healthy Controls (n=41) ^a	0.9268
PD Treated (n=267) ^b	0.9738
PD Untreated (n=29) ^b	0.8966

^aSpecificity (correct HC identification)

^bSensitivity (correct PD detection)

A.3.2 Removing top 6 Fatty Acid Amides

Applying amides filter... Original features: 1122 - palmitoleamide (16:1)* - myristoleamide (14:1)* - myristamide (14:0)* - heptadecenamide (17:1)* - linolenamide (18:3)* - linoleamide (18:2n6)

Table A.4: SVM Classification Performance (No Amides Data)

Overall Performance Metrics	
Test AUC	0.9843
Balanced Accuracy	0.9188
Number of Features	66
Group-wise Performance	
Group (n)	Performance
Healthy Controls (n=41) ^a	0.8780
PD Treated (n=267) ^b	0.9700
PD Untreated (n=29) ^b	0.8621

^aSpecificity (correct HC identification)^bSensitivity (correct PD detection)

A.3.3 Removing Both

Table A.5: SVM Classification Performance (No L-DOPA + No Amides Data)

Overall Performance Metrics	
Test AUC	0.9857
Balanced Accuracy	0.9448
Number of Features	71
Group-wise Performance	
Group (n)	Performance
Healthy Controls (n=41) ^a	0.9268
PD Treated (n=267) ^b	0.9738
PD Untreated (n=29) ^b	0.8621

^aSpecificity (correct HC identification)^bSensitivity (correct PD detection)

A. Appendix A

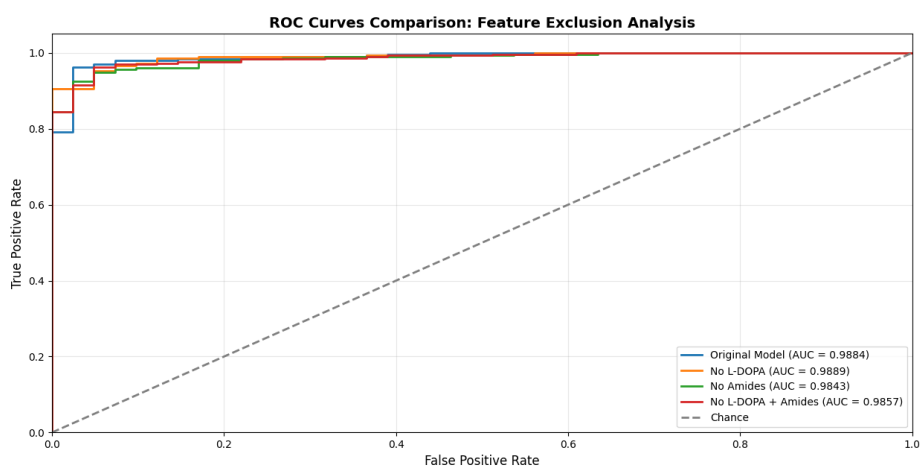


Figure A.1: ROC curves comparing model performance with different feature exclusions. The original model achieves an AUC of 0.9884. Exclusion of L-DOPA features slightly improves performance (AUC = 0.9889), while removal of amide-related features decreases performance (AUC = 0.9843). Combined exclusion of both L-DOPA and amide features yields intermediate performance (AUC = 0.9857).

Works Cited

- Aiello, Edoardo Nicolò et al. (Feb. 2022). ‘The Montreal Cognitive Assessment (MoCA): updated norms and psychometric insights into adaptive testing from healthy individuals in Northern Italy’. eng. In: *Aging Clinical and Experimental Research* 34.2, pp. 375–382.
- Ardeshirifar, Ramtin (2024). *Automated Classification of Dry Bean Varieties Using XGBoost and SVM Models*. Version Number: 1. URL: <https://arxiv.org/abs/2408.01244> (retrieved 22/9/2025).
- Beach, Thomas G. et al. (June 2009). ‘Unified staging system for Lewy body disorders: correlation with nigrostriatal degeneration, cognitive impairment and motor dysfunction’. eng. In: *Acta Neuropathologica* 117.6, pp. 613–634.
- Beauchamp, Leah C. et al. (2025). ‘Evidence of COMT dysfunction in the olfactory bulb in Parkinson’s disease’. In: *Acta Neuropathologica* 149.1, p. 21. URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC11872990/> (retrieved 9/10/2025).
- Behkamal, Bahareh et al. (July 2025). ‘A machine learning-based framework for predicting metabolic syndrome using serum liver function tests and high-sensitivity C-reactive protein’. en. In: *Scientific Reports* 15.1. Publisher: Nature Publishing Group, p. 21725. URL: <https://www.nature.com/articles/s41598-025-06723-1> (retrieved 22/9/2025).
- Ben-Shlomo, Yoav et al. (Jan. 2024). ‘The epidemiology of Parkinson’s disease’. In: *Lancet (London, England)* 403.10423, pp. 283–292. URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC11123577/> (retrieved 29/9/2025).
- Boger, D. L., S. J. Henriksen and B. F. Cravatt (Aug. 1998). ‘Oleamide: an endogenous sleep-inducing lipid and prototypical member of a new class of biological signaling molecules’. eng. In: *Current Pharmaceutical Design* 4.4, pp. 303–314.
- Boger, Dale L et al. (Dec. 2000). ‘Fatty acid amide hydrolase substrate specificity’. In: *Bioorganic & Medicinal Chemistry Letters* 10.23, pp. 2613–2616. URL: <https://www.sciencedirect.com/science/article/pii/S0960894X0000528X> (retrieved 7/10/2025).
- Braak, Heiko et al. (2003). ‘Staging of brain pathology related to sporadic Parkinson’s disease’. eng. In: *Neurobiology of Aging* 24.2, pp. 197–211.
- Bradshaw, Heather B et al. (May 2009). ‘The endocannabinoid anandamide is a precursor for the signaling lipid N-arachidonoyl glycine by two distinct pathways’. In: *BMC Biochemistry* 10, p. 14. URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC2689249/> (retrieved 9/10/2025).
- Cannas, Federico et al. (Sept. 2025). ‘Parkinson’s Disease Through the Lens of Metabolomics: A Targeted Systematic Review on Human Studies (2019–2024)’. In: *Journal of Clinical Medicine* 14.17, p. 6277. URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC12429492/> (retrieved 29/9/2025).

Works Cited

- Chen, Tianlu et al. (2013). ‘Random forest in clinical metabolomics for phenotypic discrimination and biomarker selection’. eng. In: *Evidence-Based Complementary and Alternative Medicine: eCAM* 2013, p. 298183.
- Chen, Weiyao et al. (Jan. 2020). ‘Gender differences in prevalence of LRRK2-associated Parkinson disease: A meta-analysis of observational studies’. eng. In: *Neuroscience Letters* 715, p. 134609.
- Chuah, Joshua et al. (Aug. 2022). ‘Framework for Testing Robustness of Machine Learning-Based Classifiers’. In: *Journal of Personalized Medicine* 12.8, p. 1314. URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC9409965/> (retrieved 22/9/2025).
- Clarke, Carl E. et al. (Aug. 2016). ‘Hoehn and Yahr stages’. en. In: *Clinical effectiveness and cost-effectiveness of physiotherapy and occupational therapy versus no therapy in mild to moderate Parkinson’s disease: a large pragmatic randomised controlled trial (PD REHAB)*. NIHR Journals Library. URL: <https://www.ncbi.nlm.nih.gov/books/NBK379751/> (retrieved 11/10/2025).
- Clish, Clary B. (Oct. 2015). ‘Metabolomics: an emerging but powerful tool for precision medicine’. In: *Cold Spring Harbor Molecular Case Studies* 1.1, a000588. URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC4850886/> (retrieved 11/10/2025).
- Dahabiyeh, Lina A. et al. (Apr. 2024). ‘Diagnosing Parkinson’s disease and monitoring its progression: Biomarkers from combined GC-TOF MS and LC-MS/MS untargeted metabolomics’. In: *Heliyon* 10.9, e30452. URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC11077040/> (retrieved 9/10/2025).
- Dam, Tien et al. (Sept. 2024). ‘Neuronal alpha-Synuclein Disease integrated staging system performance in PPMI, PASADENA, and SPARK baseline cohorts’. en. In: *npj Parkinson’s Disease* 10.1. Publisher: Nature Publishing Group, p. 178. URL: <https://www.nature.com/articles/s41531-024-00789-w> (retrieved 29/9/2025).
- Darst, Burcu F. et al. (Feb. 2019). ‘Longitudinal plasma metabolomics of aging and sex’. In: *Aging (Albany NY)* 11.4, pp. 1262–1282. URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC6402508/> (retrieved 31/8/2026).
- Das, Somdutta and Harshal Ramteke (2025). ‘A Comprehensive Review of the Role of Biomarkers in Early Diagnosis of Parkinson’s Disease’. In: *Cureus* 16.2 (), e54337. URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC10945043/> (retrieved 29/9/2025).
- De Lope, Elisa Gómez et al. (Mar. 2024). ‘Comprehensive blood metabolomics profiling of Parkinson’s disease reveals coordinated alterations in xanthine metabolism’. en. In: *npj Parkinson’s Disease* 10.1. Publisher: Nature Publishing Group, p. 68. URL: <https://www.nature.com/articles/s41531-024-00671-9> (retrieved 22/9/2025).
- De Rijk, M. C. et al. (Jan. 1997). ‘Prevalence of parkinsonism and Parkinson’s disease in Europe: the EUROPARKINSON Collaborative Study. European Community Concerted Action on the Epidemiology of Parkinson’s disease’. eng. In: *Journal of Neurology, Neurosurgery, and Psychiatry* 62.1, pp. 10–15.
- Department of Electrical Engineering, Canadian University Dubai, Dubai, UAE et al. (2024). ‘Forward feature selection: empirical analysis’. In: *Journal of Intelligent Systems and Internet of Things* 11.1, pp. 44–54. URL: <https://www.americaspg.com/articleinfo/18/show/2281> (retrieved 23/9/2025).
- Do, Kieu Trinh et al. (2018). ‘Characterization of missing values in untargeted MS-based metabolomics data and evaluation of missing data handling strategies’. In: *Metabolomics* 14.10, p. 128. URL:

Works Cited

- <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6153696/> (retrieved 6/9/2025).
- Dobson, Annette J. (2002). *An introduction to generalized linear models*. 2nd ed. Chapman & Hall/CRC texts in statistical science series. Boca Raton: Chapman & Hall/CRC.
- Elroby, Shabaan A. K. et al. (Apr. 2012). 'Toward the Understanding of the Metabolism of Levodopa I. DFT Investigation of the Equilibrium Geometries, Acid-Base Properties and Levodopa-Water Complexes'. In: *International Journal of Molecular Sciences* 13.4, pp. 4321–4339. URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC3344216/> (retrieved 9/10/2025).
- Falcon, S. and R. Gentleman (2008). 'Hypergeometric Testing Used for Gene Set Enrichment Analysis'. en. In: *Bioconductor Case Studies*. Ed. by Florian Hahne et al. New York, NY: Springer, pp. 207–220. URL: https://doi.org/10.1007/978-0-387-77240-0_14 (retrieved 3/10/2026).
- Faquihi, Tariq et al. (Nov. 2020). 'A Workflow for Missing Values Imputation of Untargeted Metabolomics Data'. In: *Metabolites* 10.12, p. 486. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7761057/> (retrieved 28/8/2025).
- Farrell, Emma K. et al. (Feb. 2012). 'Primary fatty acid amide metabolism: conversion of fatty acids and an ethanolamine in N18TG2 and SCP cells'. In: *Journal of Lipid Research* 53.2, pp. 247–256. URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC3269159/> (retrieved 9/10/2025).
- Fezza, F. et al. (Jan. 2023). 'Chapter 31 - Fatty acid amide hydrolase, anandamide, and neurological diseases'. In: *Neurobiology and Physiology of the Endocannabinoid System*. Ed. by Vinood B. Patel, Victor R. Preedy and Colin R. Martin. Academic Press, pp. 417–428. URL: <https://www.sciencedirect.com/science/article/pii/B9780323908771000401> (retrieved 7/10/2025).
- Foltynie, Tom et al. (Jan. 2024). 'Medical, surgical, and physical treatments for Parkinson's disease'. English. In: *The Lancet* 403.10423. Publisher: Elsevier, pp. 305–324. URL: [https://www.thelancet.com/journals/lancet/article/PIIS0140-6736\(23\)01429-0/fulltext#fig1](https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(23)01429-0/fulltext#fig1) (retrieved 28/9/2025).
- Fu, Guang-Hui et al. (June 2021). 'Feature Ranking and Screening for Class-Imbalanced Metabolomics Data Based on Rank Aggregation Coupled with Re-Balance'. en. In: *Metabolites* 11.6, p. 389. URL: <https://www.mdpi.com/2218-1989/11/6/389> (retrieved 22/9/2025).
- Galal, Aya, Marwa Talal and Ahmed Moustafa (Nov. 2022). 'Applications of machine learning in metabolomics: Disease modeling and classification'. English. In: *Frontiers in Genetics* 13. Publisher: Frontiers. URL: <https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2022.1017340/full> (retrieved 22/9/2025).
- Ghojogh, Benyamin et al. (Aug. 2021). *Uniform Manifold Approximation and Projection (UMAP) and its Variants: Tutorial and Survey*. arXiv:2109.02508 [cs]. URL: <http://arxiv.org/abs/2109.02508> (retrieved 21/9/2025).
- Goldstein, David S. (Dec. 2010). 'Catecholamines 101'. In: *Clinical autonomic research : official journal of the Clinical Autonomic Research Society* 20.6, pp. 331–352. URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC3046107/> (retrieved 9/10/2025).

Works Cited

- Gómez-Pascual, Alicia et al. (June 2024). ‘Paired plasma lipidomics and proteomics analysis in the conversion from mild cognitive impairment to Alzheimer’s disease’. eng. In: *Computers in Biology and Medicine* 176, p. 108588.
- Gonzalez-Riano, Carolina et al. (Aug. 2021). ‘Prognostic biomarkers of Parkinson’s disease in the Spanish EPIC cohort: a multiplatform metabolomics approach’. In: *NPJ Parkinson’s Disease* 7, p. 73. URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC8368017/> (retrieved 7/10/2025).
- Gowda, G.A. Nagana and Danijel Djukovic (2014). ‘Overview of Mass Spectrometry-Based Metabolomics: Opportunities and Challenges’. In: *Methods in molecular biology (Clifton, N.J.)* 1198, pp. 3–12. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4336784/> (retrieved 27/8/2025).
- Green, Rebecca et al. (2022). ‘Metabolic correlates of late midlife cognitive outcomes: findings from the 1946 British Birth Cohort’. eng. In: *Brain Communications* 4.1, fcab291.
- Groth, Detlef et al. (2013). ‘Principal Components Analysis’. en. In: *Computational Toxicology: Volume II*. Ed. by Brad Reisfeld and Arthur N. Mayeno. Totowa, NJ: Humana Press, pp. 527–547. URL: https://doi.org/10.1007/978-1-62703-059-5_22 (retrieved 21/9/2025).
- Hastie, Trevor, Robert Tibshirani and Jerome Friedman (2009a). ‘Boosting Fits an Additive Model’. In: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. New York: Springer. URL: <https://link.springer.com/book/10.1007/978-0-387-84858-7> (retrieved 22/9/2025).
- (2009b). *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY: Springer New York. URL: <http://link.springer.com/10.1007/978-0-387-84858-7> (retrieved 22/9/2025).
- Höglinger, Günter U. et al. (Feb. 2024). ‘A biological classification of Parkinson’s disease: the SynNeurGe research diagnostic criteria’. English. In: *The Lancet Neurology* 23.2. Publisher: Elsevier, pp. 191–204. URL: [https://www.thelancet.com/journals/lanneur/article/PIIS1474-4422\(23\)00404-0/fulltext#sec-4](https://www.thelancet.com/journals/lanneur/article/PIIS1474-4422(23)00404-0/fulltext#sec-4) (retrieved 29/9/2025).
- Horsager, Jacob et al. (Oct. 2020). ‘Brain-first versus body-first Parkinson’s disease: a multimodal imaging case-control study’. eng. In: *Brain: A Journal of Neurology* 143.10, pp. 3077–3088.
- Huckvale, Erik D. and Hunter N. B. Moseley (May 2024). ‘Predicting the Pathway Involvement of Metabolites Based on Combined Metabolite and Pathway Features’. en. In: *Metabolites* 14.5, p. 266. URL: <https://www.mdpi.com/2218-1989/14/5/266> (retrieved 22/9/2025).
- Hwang, Yun Su et al. (Jan. 2025). ‘Identification of Novel Genetic Loci Affecting Age at Onset of Parkinson’s Disease: A Genome-wide Association Study’. eng. In: *Movement Disorders: Official Journal of the Movement Disorder Society* 40.1, pp. 77–86.
- Ibrahim, Rana M. et al. (Nov. 2023). ‘Combinative effects of akarkara root-derived metabolites on anti-inflammatory and anti-alzheimer key enzymes: integrating bioassay-guided fractionation, GC-MS analysis, and in silico studies’. In: *BMC Complementary Medicine and Therapies* 23, p. 413. URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC10655324/> (retrieved 7/10/2025).

Works Cited

- Ibrahim, Weam W. et al. (2024). ‘Neuroprotective potential of *Erigeron bonariensis* ethanolic extract against ovariectomized/D-galactose-induced memory impairments in female rats in relation to its metabolite fingerprint as revealed using UPLC/MS’. In: *Inflammopharmacology* 32.2, pp. 1091–1112. URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC11006746/> (retrieved 9/10/2025).
- Ip, Edward H. (2007). ‘General Linear Models’. en. In: *Topics in Biostatistics*. Ed. by Walter T. Ambrosius. Totowa, NJ: Humana Press, pp. 189–211. URL: https://doi.org/10.1007/978-1-59745-530-5_10 (retrieved 19/9/2025).
- Just, Mie Kristine et al. (May 2022). ‘Alpha-Synuclein Strain Variability in Body-First and Brain-First Synucleinopathies’. English. In: *Frontiers in Aging Neuroscience* 14. Publisher: Frontiers. URL: <https://www.frontiersin.org/journals/aging-neuroscience/articles/10.3389/fnagi.2022.907293/full> (retrieved 29/9/2025).
- Kattula, Sravya, James R. Byrnes and Alisa S. Wolberg (Mar. 2017). ‘Fibrinogen and fibrin in hemostasis and thrombosis’. In: *Arteriosclerosis, thrombosis, and vascular biology* 37.3, e13–e21. URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC5324399/> (retrieved 10/10/2025).
- Kawano, Kohei et al. (Aug. 2024). ‘Fatty acid amides present in Camembert cheese improved cognitive decline after oral administration in mice’. eng. In: *Neuroscience Research* 205, pp. 34–39.
- Kim, Iris Y. et al. (May 2018). ‘Integration of risk factors for Parkinson disease in 2 large longitudinal cohorts’. eng. In: *Neurology* 90.19, e1646–e1653.
- Klatt, Stephan et al. (Oct. 2021). ‘A six-metabolite panel as potential blood-based biomarkers for Parkinson’s disease’. en. In: *npj Parkinson’s Disease* 7.1. Publisher: Nature Publishing Group, p. 94. URL: <https://www.nature.com/articles/s41531-021-00239-x> (retrieved 19/9/2025).
- Konopka, Tomasz (2025). *Package umap R*. URL: <https://cran.r-project.org/web/packages/umap/refman/umap.html> (retrieved 21/9/2025).
- Krutkin, Dennis D. et al. (Feb. 2025). ‘To Impute or Not To Impute in Untargeted Metabolomics—That is the Compositional Question’. In: *Journal of the American Society for Mass Spectrometry* 36.4, pp. 742–759. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11969646/> (retrieved 6/9/2025).
- LabOak (May 2025). *A Comprehensive Guide to Optuna: Next-Generation Hyperparameter Optimization*. en-US. URL: <https://laboak.com/a-comprehensive-guide-to-optuna-next-generation-hyperparameter-optimization/> (retrieved 22/9/2025).
- Labory, Justine, Evariste Njomgue-Fotso and Silvia Bottini (Dec. 2024). ‘Benchmarking feature selection and feature extraction methods to improve the performances of machine-learning algorithms for patient classification using metabolomics biomedical data’. eng. In: *Computational and Structural Biotechnology Journal* 23, pp. 1274–1287.
- Larsson, Susanna C. and Stephen Burgess (July 2022). ‘Appraising the causal role of smoking in multiple diseases: A systematic review and meta-analysis of Mendelian randomization studies’. In: *eBioMedicine* 82, p. 104154. URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC9278068/> (retrieved 29/9/2025).

Works Cited

- Law, Charity W. et al. (Feb. 2014). ‘voom: precision weights unlock linear model analysis tools for RNA-seq read counts’. In: *Genome Biology* 15.2, R29. URL: <https://doi.org/10.1186/gb-2014-15-2-r29> (retrieved 18/9/2025).
- Learn, Sci-kit (Sept. 2025a). *scikit-learn-contrib/boruta_py*. original-date: 2016-01-30T16:25:26Z. URL: https://github.com/scikit-learn-contrib/boruta_py (retrieved 22/9/2025).
- (2025b). *1.4. Support Vector Machines*. en. URL: <https://scikit-learn/stable/modules/svm.html> (retrieved 22/9/2025).
- (2025c). *3.2. Tuning the hyper-parameters of an estimator*. en. URL: https://scikit-learn/stable/modules/grid_search.html (retrieved 23/9/2025).
- (2025d). *3.4. Metrics and scoring: quantifying the quality of predictions* — *scikit-learn 1.7.2 documentation*. URL: https://scikit-learn.org/stable/modules/model_evaluation.html (retrieved 22/9/2025).
- (2025e). *5.2. Permutation feature importance*. en. URL: https://scikit-learn/stable/modules/permutation_importance.html (retrieved 23/9/2025).
- (2025f). *LogisticRegression*. en. URL: https://scikit-learn/stable/modules/generated/sklearn.linear_model.LogisticRegression.html (retrieved 22/9/2025).
- Lee, Kil Sun, Xiaoyang Su and Tao Huan (May 2025). ‘Metabolites are not genes - avoiding the misuse of pathway analysis in metabolomics’. eng. In: *Nature Metabolism* 7.5, pp. 858–861.
- Liebal, Ulf W. et al. (June 2020). ‘Machine Learning Applications for Mass Spectrometry-Based Metabolomics’. In: *Metabolites* 10.6, p. 243. URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC7345470/> (retrieved 22/9/2025).
- Lim, Shen-Yang et al. (Dec. 2024). ‘Uncovering the genetic basis of Parkinson’s disease globally: from discoveries to the clinic’. In: *The Lancet Neurology* 23.12, pp. 1267–1280. URL: <https://www.sciencedirect.com/science/article/pii/S1474442224003788> (retrieved 28/9/2025).
- Lundberg, Scott M. and Su-In Lee (Dec. 2017). ‘A unified approach to interpreting model predictions’. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS’17. Red Hook, NY, USA: Curran Associates Inc., pp. 4768–4777 (retrieved. 22/9/2025).
- Luo, Xi et al. (June 2024). ‘Identification of metabolites reproducibly associated with Parkinson’s Disease via meta-analysis and computational modelling’. In: *NPJ Parkinson’s Disease* 10, p. 126. URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC11217404/> (retrieved 7/10/2025).
- Maini Rekdal, Vayu et al. (June 2019). ‘Discovery and inhibition of an interspecies gut bacterial pathway for Levodopa metabolism’. In: *Science* 364.6445. Publisher: American Association for the Advancement of Science, eaau6323. URL: <https://www.science.org/doi/10.1126/science.aau6323> (retrieved 9/10/2025).
- Malek, Naveed et al. (2015). ‘Tracking Parkinson’s: Study Design and Baseline Patient Data’. eng. In: *Journal of Parkinson’s Disease* 5.4, pp. 947–959.
- Massey, Frank J. (1951). ‘The Kolmogorov-Smirnov Test for Goodness of Fit’. In: *Journal of the American Statistical Association* 46.253. Publisher: [American

Works Cited

- Statistical Association, Taylor & Francis, Ltd.], pp. 68–78. URL: <https://www.jstor.org/stable/2280095> (retrieved 7/9/2025).
- Meeks, Shannon L. (Jan. 2009). ‘Chapter 98 - Congenital Disorders of Fibrinogen’. In: *Transfusion Medicine and Hemostasis*. Ed. by Christopher D. Hillyer et al. San Diego: Academic Press, pp. 537–539. URL: <https://www.sciencedirect.com/science/article/pii/B9780123744326000981> (retrieved 10/10/2025).
- Meiser, Johannes, Daniel Weindl and Karsten Hiller (May 2013). ‘Complexity of dopamine metabolism’. In: *Cell Communication and Signaling* 11.1, p. 34. URL: <https://doi.org/10.1186/1478-811X-11-34> (retrieved 9/10/2025).
- Mendez, Kevin M., Stacey N. Reinke and David I. Broadhurst (2019). ‘A comparative evaluation of the generalised predictive ability of eight machine learning algorithms across ten clinical metabolomics data sets for binary classification’. In: *Metabolomics* 15.12, p. 150. URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC6856029/> (retrieved 22/9/2025).
- Metabolon, inc. (2025a). *Experimental Procedures*. en-US. URL: <https://www.metabolon.com/software/support/portal/experimental-procedures/> (retrieved 27/8/2025).
- (2025b). *Identification*. en-US. URL: <https://www.metabolon.com/why-metabolon/identification/> (retrieved 28/8/2025).
- Morris, Huw R. et al. (Jan. 2024). ‘The pathogenesis of Parkinson’s disease’. eng. In: *Lancet (London, England)* 403.10423, pp. 293–304.
- Movement Disorder Society Task Force on Rating Scales for Parkinson’s Disease (July 2003). ‘The Unified Parkinson’s Disease Rating Scale (UPDRS): status and recommendations’. eng. In: *Movement Disorders: Official Journal of the Movement Disorder Society* 18.7, pp. 738–750.
- Nam, Hye Yeon et al. (2017). ‘Antiepileptic and Neuroprotective Effects of Oleamide in Rat Striatum on Kainate-Induced Behavioral Seizure and Excitotoxic Damage via Calpain Inhibition’. eng. In: *Frontiers in Pharmacology* 8, p. 817.
- NHS (Oct. 2017). *Parkinson’s disease - Diagnosis*. en. Section: conditions. URL: <https://www.nhs.uk/conditions/parkinsons-disease/diagnosis/> (retrieved 29/9/2025).
- Ojala, Markus and Gemma C. Garriga (2010). ‘Permutation tests for studying classifier performance’. In: *Journal of Machine Learning Research* 11.62, pp. 1833–1863. URL: <http://jmlr.org/papers/v11/ojala10a.html>.
- Oka, Takaki et al. (Apr. 2024). ‘Multiomics analysis to explore blood metabolite biomarkers in an Alzheimer’s Disease Neuroimaging Initiative cohort’. eng. In: *Scientific Reports* 14.1, p. 6797.
- Optuna (2025). *optuna.pruners.MedianPruner — Optuna 4.5.0 documentation*. URL: <https://optuna.readthedocs.io/en/stable/reference/generated/optuna.pruners.MedianPruner.html> (retrieved 22/9/2025).
- Ou, Huiting et al. (Jan. 2025). ‘Imputation for Lipidomics and Metabolomics (ImpLiMet): a web-based application for optimization and method selection for missing data imputation’. In: *Bioinformatics Advances* 5.1, vbae209. URL: <https://doi.org/10.1093/bioadv/vbae209> (retrieved 6/9/2025).
- Owusu-Adjei, Michael et al. (Nov. 2023). ‘Imbalanced class distribution and performance evaluation metrics: A systematic review of prediction accuracy for determining model

Works Cited

- performance in healthcare systems’. In: *PLOS Digital Health* 2.11, e0000290. URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC10688675/> (retrieved 23/9/2025).
- Paul, Kimberly C. et al. (Dec. 2023). ‘Untargeted serum metabolomics reveals novel metabolite associations and disruptions in amino acid and lipid metabolism in Parkinson’s disease’. In: *Molecular Neurodegeneration* 18, p. 100. URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC10731845/> (retrieved 9/10/2025).
- Phipson, Belinda et al. (June 2016). ‘ROBUST HYPERPARAMETER ESTIMATION PROTECTS AGAINST HYPERVARIABLE GENES AND IMPROVES POWER TO DETECT DIFFERENTIAL EXPRESSION’. eng. In: *The Annals of Applied Statistics* 10.2, pp. 946–963.
- Ponce-Bobadilla, Ana Victoria et al. (Oct. 2024). ‘Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development’. In: *Clinical and Translational Science* 17.11, e70056. URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC11513550/> (retrieved 22/9/2025).
- Prajumwongs, Piya et al. (Apr. 2025). ‘Identification of serum metabolite biomarkers and metabolic reprogramming mechanisms to predict recurrence in cholangiocarcinoma’. In: *Scientific Reports* 15, p. 12782. URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC11997029/> (retrieved 22/9/2025).
- Probst, Philipp and Anne-Laure Boulesteix (2017). ‘To tune or not to tune the number of trees in random forest?’ In: Publisher: arXiv Version Number: 1. URL: <https://arxiv.org/abs/1705.05654> (retrieved 22/9/2025).
- Rangwala, Hussain Sohail et al. (Jan. 2025). ‘From Diagnosis to Treatment: A Comprehensive Review of Biomarkers and Therapeutic Advances in Parkinson’s Disease’. In: *Annals of Neurosciences* 32.1, pp. 51–57. URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC11863227/> (retrieved 29/9/2025).
- Rather, Arif Ahmad and Manzoor Ahmad Chachoo (Mar. 2023). ‘Robust correlation estimation and UMAP assisted topological analysis of omics data for disease subtyping’. In: *Computers in Biology and Medicine* 155, p. 106640. URL: <https://www.sciencedirect.com/science/article/pii/S0010482523001051> (retrieved 21/9/2025).
- RDocumentation (2025). *prcomp function - RDocumentation*. URL: <https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/prcomp> (retrieved 21/9/2025).
- Rdocumentation (2025). *glm function - RDocumentation*. URL: <https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/glm> (retrieved 19/9/2025).
- Ritchie, Matthew E. et al. (Apr. 2015). ‘limma powers differential expression analyses for RNA-sequencing and microarray studies’. In: *Nucleic Acids Research* 43.7, e47. URL: <https://doi.org/10.1093/nar/gkv007> (retrieved 18/9/2025).
- Rizzo, Giovanni et al. (Feb. 2016). ‘Accuracy of clinical diagnosis of Parkinson disease: A systematic review and meta-analysis’. eng. In: *Neurology* 86.6, pp. 566–576.
- Ryals, John et al. (July 2007). ‘Metabolon, Inc’. eng. In: *Pharmacogenomics* 8.7, pp. 863–866.
- Salmi, Mabrouka et al. (Sept. 2024). ‘Handling imbalanced medical datasets: review of a decade of research’. en. In: *Artificial Intelligence Review* 57.10, p. 273. URL: <https://doi.org/10.1007/s10462-024-10884-2> (retrieved 23/9/2025).
- Sánchez-Vinces, Salvador et al. (Feb. 2023). ‘Mass-Spectrometry-Based Lipidomics Discriminates Specific Changes in Lipid Classes in Healthy and Dyslipidemic Adults’.

Works Cited

- In: *Metabolites* 13.2, p. 222. URL:
<https://pmc.ncbi.nlm.nih.gov/articles/PMC9964724/> (retrieved 18/9/2025).
- Santos, Wanderleya T. et al. (Sept. 2024). 'Metabolomics Unveils Disrupted Pathways in Parkinson's Disease: Toward Biomarker-Based Diagnosis'. In: *ACS Chemical Neuroscience* 15.17. Publisher: American Chemical Society, pp. 3168–3180. URL:
<https://doi.org/10.1021/acscchemneuro.4c00355> (retrieved 21/9/2025).
- Schrag, Anette et al. (Jan. 2015). 'Prediagnostic presentations of Parkinson's disease in primary care: a case-control study'. eng. In: *The Lancet. Neurology* 14.1, pp. 57–64.
- Schulte, Eva C. et al. (Jan. 2016). 'Alterations in Lipid and Inositol Metabolisms in Two Dopaminergic Disorders'. In: *PLoS ONE* 11.1, e0147129. URL:
<https://pmc.ncbi.nlm.nih.gov/articles/PMC4726488/> (retrieved 7/10/2025).
- ScienceDirect (2025). *Fatty Acid Amide - an overview | ScienceDirect Topics*. URL:
<https://www.sciencedirect.com/topics/chemistry/fatty-acid-amide> (retrieved 7/10/2025).
- Shao, Yaping et al. (Jan. 2021). 'Comprehensive metabolic profiling of Parkinson's disease by liquid chromatography-mass spectrometry'. In: *Molecular Neurodegeneration* 16.1, p. 4. URL: <https://doi.org/10.1186/s13024-021-00425-8> (retrieved 9/10/2025).
- Shin, Hae-Won, Sang-Wook Hong and Young Chul Youn (May 2022). 'Clinical Aspects of the Differential Diagnosis of Parkinson's Disease and Parkinsonism'. eng. In: *Journal of Clinical Neurology (Seoul, Korea)* 18.3, pp. 259–270.
- Siderowf, Andrew et al. (May 2023). 'Assessment of heterogeneity among participants in the Parkinson's Progression Markers Initiative cohort using α -synuclein seed amplification: a cross-sectional study'. eng. In: *The Lancet. Neurology* 22.5, pp. 407–417.
- Simon, David K., Caroline M. Tanner and Patrik Brundin (Feb. 2020). 'Parkinson Disease Epidemiology, Pathology, Genetics and Pathophysiology'. In: *Clinics in geriatric medicine* 36.1, pp. 1–12. URL:
<https://pmc.ncbi.nlm.nih.gov/articles/PMC6905381/> (retrieved 29/9/2025).
- Smilde, Age K. et al. (Sept. 2005). 'Fusion of Mass Spectrometry-Based Metabolomics Data'. In: *Analytical Chemistry* 77.20, pp. 6729–6736. URL:
<https://doi.org/10.1021/ac051080y> (retrieved 31/8/2026).
- Soti, Monavareh et al. (Jan. 2022). 'Parkinson's disease related alterations in cannabinoid transmission'. In: *Brain Research Bulletin* 178, pp. 82–96. URL:
<https://www.sciencedirect.com/science/article/pii/S0361923021003294> (retrieved 7/10/2025).
- Sun, Jun and Yinglin Xia (July 2023). 'Pretreating and normalizing metabolomics data for statistical analysis'. In: *Genes & Diseases* 11.3, p. 100979. URL:
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10827599/> (retrieved 28/8/2025).
- Tong, Feng et al. (June 2020). 'Classification of the urinary metabolome using machine learning and potential applications to diagnosing interstitial cystitis'. In: *Bladder* 7.2, e43. URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC7401992/> (retrieved 22/9/2025).
- Töpfer, Nadine, Sabrina Kleessen and Zoran Nikoloski (Feb. 2015). 'Integration of metabolomics data into metabolic networks'. In: *Frontiers in Plant Science* 6, p. 49. URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC4330704/> (retrieved 21/9/2025).

Works Cited

- Trupp, Miles et al. (2014). ‘Metabolite and peptide levels in plasma and CSF differentiating healthy controls from patients with newly diagnosed Parkinson’s disease’. eng. In: *Journal of Parkinson’s Disease* 4.3, pp. 549–560.
- Vascellari, Sarah et al. (Sept. 2020). ‘Gut Microbiota and Metabolome Alterations Associated with Parkinson’s Disease’. In: *mSystems* 5.5, e00561–20. URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC7498685/> (retrieved 7/10/2025).
- Virameteekul, Sasivimol et al. (Apr. 2023). ‘Clinical Diagnostic Accuracy of Parkinson’s Disease: Where Do We Stand?’ eng. In: *Movement Disorders: Official Journal of the Movement Disorder Society* 38.4, pp. 558–566.
- Waluk, Dominik P. et al. (Jan. 2014). ‘Chapter 9 - Mammalian Fatty Acid Amides of the Brain and CNS’. In: *Omega-3 Fatty Acids in Brain and Neurological Health*. Ed. by Ronald Ross Watson and Fabien De Meester. Boston: Academic Press, pp. 87–107. URL: <https://www.sciencedirect.com/science/article/pii/B9780124105270000090> (retrieved 7/10/2025).
- Wang, Xiaoshuang et al. (Feb. 2014). ‘The endogenous substrates of brain CYP2D’. In: *European Journal of Pharmacology* 724, pp. 211–218. URL: <https://www.sciencedirect.com/science/article/pii/S0014299913009539> (retrieved 9/10/2025).
- Wang, Yang et al. (2015). ‘Current state of the art of mass spectrometry-based metabolomics studies – a review focusing on wide coverage, high throughput and easy identification’. en. In: *RSC Advances* 5.96, pp. 78728–78737. URL: <https://xlink.rsc.org/?DOI=C5RA14058G> (retrieved 27/8/2025).
- Wassenberg, T. et al. (Dec. 2010). ‘Urinary dopamine in aromatic L-amino acid decarboxylase deficiency: The unsolved paradox’. In: *Molecular Genetics and Metabolism* 101.4, pp. 349–356. URL: <https://www.sciencedirect.com/science/article/pii/S1096719210002994> (retrieved 9/10/2025).
- Wei, Runmin et al. (Jan. 2018). ‘Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data’. en. In: *Scientific Reports* 8.1. Publisher: Nature Publishing Group, p. 663. URL: <https://www.nature.com/articles/s41598-017-19120-0> (retrieved 6/9/2025).
- Wooten, G. F. et al. (Apr. 2004). ‘Are men at greater risk for Parkinson’s disease than women?’ eng. In: *Journal of Neurology, Neurosurgery, and Psychiatry* 75.4, pp. 637–639.
- XGBoost (2025). *XGBoost Parameters — xgboost 3.0.5 documentation*. URL: <https://xgboost.readthedocs.io/en/stable/parameter.html> (retrieved 22/9/2025).
- Xiao, Bin et al. (May 2025). ‘Pathogenesis of Parkinson’s Disease’. eng. In: *Neurologic Clinics* 43.2, pp. 185–207.
- Xu, Yudian et al. (Dec. 2024). ‘Integrating Machine Learning in Metabolomics: A Path to Enhanced Diagnostics and Data Interpretation’. eng. In: *Small Methods* 8.12, e2400305.
- Yamashita, Kyla Y. et al. (Dec. 2023). ‘Biomarkers in Parkinson’s disease: A state of the art review’. In: *Biomarkers in Neuropsychiatry* 9, p. 100074. URL: <https://www.sciencedirect.com/science/article/pii/S266614462300014X> (retrieved 29/9/2025).

Works Cited

- Yang, Yang et al. (July 2021). ‘Dimensionality reduction by UMAP reinforces sample heterogeneity analysis in bulk transcriptomic data’. In: *Cell Reports* 36.4, p. 109442. URL: <https://www.sciencedirect.com/science/article/pii/S2211124721008597> (retrieved 21/9/2025).
- Zhang, J. Diana et al. (May 2023). ‘Interpretable Machine Learning on Metabolomics Data Reveals Biomarkers for Parkinson’s Disease’. In: *ACS Central Science* 9.5, pp. 1035–1045. URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC10214508/> (retrieved 22/9/2025).
- Zhou, Jingheng et al. (Dec. 2021). ‘Dopamine Neuron Challenge Test for early detection of Parkinson’s disease’. eng. In: *NPJ Parkinson’s disease* 7.1, p. 116.