Maturitas 193 (2025) 108185

Contents lists available at ScienceDirect

Maturitas

journal homepage: www.elsevier.com/locate/maturitas

Original article

Understanding Parkinson's: The microbiome and machine learning approach

David Rojas-Velazquez [a,b,*], Sarah Kidwai [a], Ting Chia Liu [a], Mounim A. El-Yacoubi [e], Johan Garssen [a,c], Alberto Tonda [d], Alejandro Lopez-Rincon [a]

[a] Division of Pharmacology, Utrecht Institute for Pharmaceutical Sciences, Faculty of Science, Universiteitsweg 99, Utrecht 3508 TB, the Netherlands
[b] Department of Data Science, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Heidelberglaan 100, Utrecht, 3508 GA, the Netherlands
[c] Global Centre of Excellence Immunology, Danone Nutricia Research, Uppsalalaan 12, Utrecht 3584 CT, the Netherlands
[d] UMR 518 MIA-PS, INRAE, Universit'e Paris-Saclay, Institut des Syst'emes Complexes de Paris, Île-de-France (ISC-PIF) - UAR 3611 CNRS, 113 rue Nationale, Paris 75013, Paris, France
[e] SAMOVAR, Telecom SudParis, Institut Polytechnique de Paris, 91120 Palaiseau, Paris, France

ARTICLE INFO

Keywords:
Feature selection
Machine learning
Biomarker discovery
Deep learning

ABSTRACT

Objective: Given that Parkinson's disease is a progressive disorder, with symptoms that worsen over time, our goal is to enhance the diagnosis of Parkinson's disease by utilizing machine learning techniques and microbiome analysis. The primary objective is to identify specific microbiome signatures that can reproducibly differentiate patients with Parkinson's disease from healthy controls.
Methods: We used four Parkinson-related datasets from the NCBI repository, focusing on stool samples. Then, we applied a DADA2-based script for amplicon sequence processing and the Recursive Ensemble Feature Selection (REF) algorithm for biomarker discovery. The discovery dataset was PRJEB14674, while PRJNA742875, PRJEB27564, and PRJNA594156 served as testing datasets. The Extra Trees classifier was used to validate the selected features.
Results: The Recursive Ensemble Feature Selection algorithm identified 84 features (Amplicon Sequence Variants) from the discovery dataset, achieving an accuracy of over 80%. The Extra Trees classifier demonstrated good diagnostic accuracy with an area under the receiver operating characteristic curve of 0.74. In the testing phase, the classifier achieved areas under the receiver operating characteristic curves of 0.64, 0.71, and 0.62 for the respective datasets, indicating sufficient to good diagnostic accuracy. The study identified several bacterial taxa associated with Parkinson's disease, such as Lactobacillus, Bifidobacterium, and Roseburia, which were increased in patients with the disease.
Conclusion: This study successfully identified microbiome signatures that can differentiate patients with Parkinson's disease from healthy controls across different datasets. These findings highlight the potential of integrating machine learning and microbiome analysis for the diagnosis of Parkinson's disease. However, further research is needed to validate these microbiome signatures and to explore their therapeutic implications in developing targeted treatments and diagnostics for Parkinson's disease.

1. Introduction

Parkinson's disease (PD) is a progressive neurodegenerative disorder that predominantly affects movement. It manifests through symptoms such as tremors, rigidity, bradykinesia (slowness of movement), and postural instability. Additionally, non-motor symptoms like cognitive impairment, mental health disorders, sleep disturbances, and pain are common and can significantly impact quality of life [1]. The exact cause of PD remains unknown, but it is thought to result from a combination of genetic and environmental factors. The hallmark of PD is the loss of dopamine-producing neurons in the substantia nigra, a brain region that controls movement. This neuronal loss leads to the characteristic motor symptoms of the disease [2].

Diagnosis of PD is primarily clinical, based on medical history and

* Corresponding author at: Division of Pharmacology, Utrecht Institute for Pharmaceutical Sciences, Faculty of Science, Universiteitsweg 99, Utrecht 3508 TB, the Netherlands.
E-mail address: e.d.rojasvelazquez@uu.nl (D. Rojas-Velazquez).

https://doi.org/10.1016/j.maturitas.2024.108185
Received 11 August 2024; Received in revised form 16 December 2024; Accepted 20 December 2024
Available online 23 December 2024
0378-5122/© 2024 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

neurological examination. Although there is currently no cure for PD, various treatments, including medications like levodopa and dopamine agonists, can help to manage symptoms [3]. Non-pharmacological interventions, such as physical therapy and lifestyle changes, might play a crucial role in managing the disease as well. Recent advancements in Machine Learning (ML) and Artificial Intelligence (AI) are showing promise in enhancing the diagnosis and maybe even the management of PD [1]. The use of ML and AI in diagnosing and assessing PD is rapidly advancing. Recent research has highlighted the potential of various ML models, such as Convolutional Neural Networks (CNN) and Support Vector Machines (SVM), to analyze clinical, behavioral, and biometric data for PD diagnosis. These models have shown promise in refining diagnostic procedures by utilizing diverse data sources, including voice recordings, movement data, MRI, SPECT, PET, Cerebrospinal Fluid (CSF), and handwritten patterns [1].

Despite these advancements, challenges persist, particularly in data collection and the need for large datasets in clinical studies. Emphasizing transparency in reporting data collection, pre-processing protocols, and model implementation is crucial to improve the reproducibility of studies and support clinical decision-making [1]. ML algorithms can analyze various data types, such as neuroimaging, motor function tests, and even breathing patterns, to detect PD early. For instance, researchers at MIT have developed a neural network that can detect Parkinson's from nocturnal breathing patterns [4]. ML models can also classify the stages of PD with high accuracy. A recent study demonstrated that decision tree classification could distinguish between early and advanced stages of PD with an accuracy of 73.7% [5]. Furthermore, ML techniques are being used to develop non-invasive monitoring tools. For example, a device developed at MIT can assess the severity of PD by analyzing breathing patterns without requiring any interaction from the user [4]. Furthermore, ML models, such as SVM and random forest classifiers, are being implemented to improve the accuracy and efficiency of PD diagnosis [1,5]. These advancements highlight the potential of ML to transform the way PD is diagnosed and managed, offering hope for earlier detection and better patient outcomes.

In the recent years, the intricate relationship between the gut microbiome and PD has been elucidated. Emphasizing how gastrointestinal dysfunction is a significant feature of PD, showing a bidirectional link between the brain and the gut. Over the past decade, advancements in high-throughput sequencing have shed light on the complex interactions within the microbiome–gut–brain axis, offering potential for new biomarkers and maybe even novel therapeutic targets [2].

The gap between model development and clinical application needs to be addressed, with further validation of ML approaches in clinical settings being crucial. Uniform reporting standards in ML studies related to PD are also necessary to ensure consistency and reliability in research outcomes. This study aims to explore these aspects, focusing on the integration of ML techniques in the clinical diagnosis with microbiome across multiple datasets, with the goal of enhancing early detection and improving patient outcomes.

2. Methods

We used the methodology outlined in detail by Rojas et al. (2024) [6], which integrates a DADA2-based script with the Recursive Ensemble Feature Selection (REFS) approach[1]. DADA2, an open-source R package, facilitates the amplicon workflow on 16s rRNA sequences, encompassing filtering, dereplication, sample inference, chimera identification, and merging of paired-end reads [7]. REFS, a robust and reliable biomarker discovery method, identifies the most effective features for differentiating between groups in datasets, achieving high accuracy with a minimal number of features [8].

REFS is an ensemble composed of eight algorithms (Stochastic Gradient Descent on linear models, SVM, Gradient Boosting, Random Forest, Logistic Regression, Passive Aggressive classifier, Ridge Classifier and Bagging) that analyze, rank, and select a set of features based on obtaining the highest accuracy in differentiating between case and control groups using the smallest number of features [6,8]. Additionally, we used a validation module to test this reduced set of features using five algorithms that are not part of the ensemble (AdaBoost, Extra Trees, KNeighbors, Multilayer Perceptron, and LassoCV) to assess its effectiveness in differentiating between case and control groups. This evaluation process helps to avoid bias selection and to provide assurance that the selected features are robust [6,8].

The methodology described in [6] consist in four phases:

- Dataset selection criteria: involves choosing, downloading, and extracting relevant information from metadata. There must be at least three datasets, one for discovery and two for testing. The datasets must be 16s rRNA amplicon sequencing, belong to the same disease or disorder, and include at least two groups (control and case) with a minimum of 10 samples each. Documentation should clearly specify group assignments, and all samples should come from the same source, such as tissue, feces, or mucosa.
- Raw data processing: involves performing amplicon workflow using the DADA2-based script on the raw data in the selected datasets and generate ASVs (features).
- The feature selection: selects the reduced set of features from the discovery dataset using REFS and they are validated applying the validation module. The discovery dataset is chosen based on the shortest sequence length after raw data processing.
- Testing: the selected features are searched in each testing datasets and its accuracy is validated in differentiating between the groups (cases and controls) using the validation module.

Additionally, we used a web tool called MicrobiomeAnalyst [9] which is a user-friendly, web-based platform designed for comprehensive statistical, functional, and integrative analysis of microbiome data. It supports various types of analyses, including: Marker Data Profiling, Shotgun Data Profiling, Taxon Set Analysis, Microbiome Metabolomics Profiling, Statistical Metaanalysis, Raw Data Processing, and Report Generation & Batch Processing. In this research work, we used the Taxon Set Analysis that identifies taxonomic signatures to analyze how the resulting taxa can be related with different diseases and symptoms. Although the use of this tool is not part of the methodology we follow [6], the results obtained provide relevant information that complements the literature analysis.

After following the *dataset selection criteria* phase defined in [6], we selected four Parkinson-related datasets downloaded from the NCBI repository:

1. **PRJEB14674** which contains 348 samples from feces, the cohort is from San Diego CA, USA, the age range is from 36 to 94 years old, the ethnicities are Hispanic/Latino and not Hispanic/Latino, 152 female samples, 196 male samples, and 28 not reported. However, only 345 are included in this study, due to samples with identifier ERR1513720, ERR1513792, and ERR1513897 being damaged. This dataset consists of 134 healthy-control samples (labeled as 0) and 211 case samples (labeled as 1).
2. **PRJNA742875** which contains 356 samples from different sources including feces, oral, and nasopharyngeal swab from Seoul, Korea. There is no additional information about gender, age and/or ethnicity. We considered only the 172 samples from feces, consisting of 84 healthy control samples (labeled as 0) and 88 case samples (labeled as 1).
3. **PRJEB27564** which contains 266 samples from feces, of which 130 healthy-control samples (labeled as 0) and 136 case samples (labeled

[1] Open source algorithm available on GitHub, https://github.com/steppenwolf0/ MicrobiomeREFS.

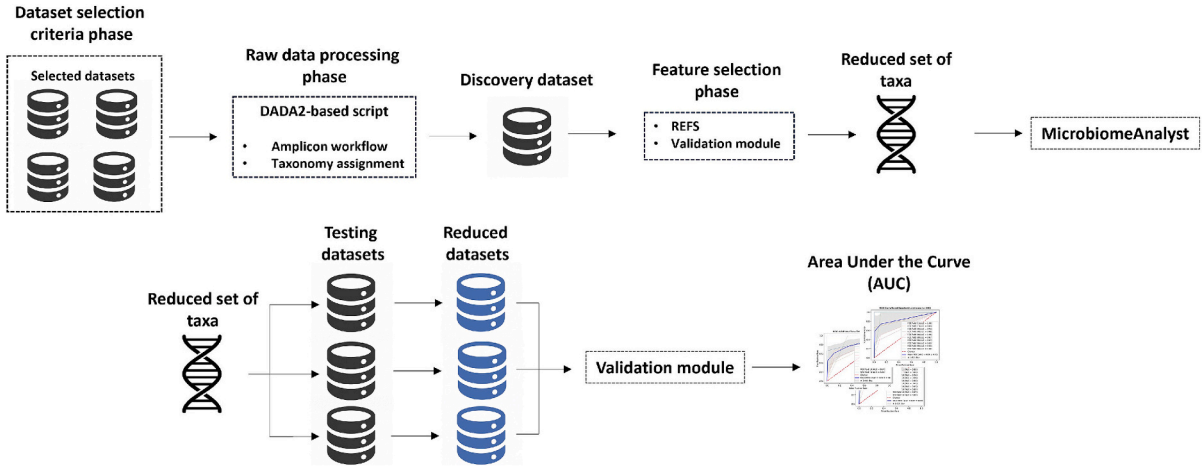

Fig. 1. General overview of the experiments. The upper part represents the dataset selection criteria, raw data processing, and feature selection phases including the MicrobiomeAnalyst analysis. The bottom part corresponds to the testing phase.

as 1). The cohort is from Finland, and there is no additional information about gender, age and/or ethnicity.

4. **PRJNA594156** which contains 300 samples from feces, of which 103 healthy-control samples (labeled as 0) and 197 case samples (labeled as 1). The cohort is from Vancouver, Canada, 128 female samples, 178 male samples, the age range is from 40 to 80 years old, there is no information about ethnicity.

We designated PRJEB14674 as the discovery dataset based on the eligibility criteria which states: "the discovery dataset is the one that contains the shortest sequence length after the raw data processing phase" [6]. Consequently, PRJNA742875, PRJEB27564, and PRJNA594156 were assigned as testing datasets. Following this, we proceeded with the methodology's feature selection and testing phases. Fig. 1 illustrates the experimental workflow.

3. Results

3.1. Raw data processing phase

After performing the full amplicon workflow on the raw data within the four datasets by executing the DADA2-based script, we generated the following Amplicon Sequence Variants (ASVs) with their respective filtering parameters:

- **PRJEB14674** parameter: trimLeft = 12; 7,269 ASVs generated.
- **PRJNA742875** parameter: trimLeft = 20; 6,275 ASVs generated.
- **PRJEB27564** parameter: trimLeft = 10, truncLen = c(250); 17,163 ASVs generated.
- **PRJNA594156** parameter: trimLeft = 13; 3,801 ASVs generated.

3.2. Feature selection phase

After applying the REFS algorithm to the discovery dataset, the resulting subset of features identified contained 84 out of the original 7,269 ASVs. This means, REFS achieved its highest accuracy (> 0.80) with the minimum number of features (84 features), Fig. 2a. Following the processes defined in this phase, after executing the validation module for the selected 84 features, the Extra Trees classifier had the best performance, with an Area under the Receiver Operating Characteristic Curve (AUC-ROC) of 0.74, Fig. 2b. According to [10], the resulting AUC-ROC is considered as *"good"* diagnostic accuracy. The

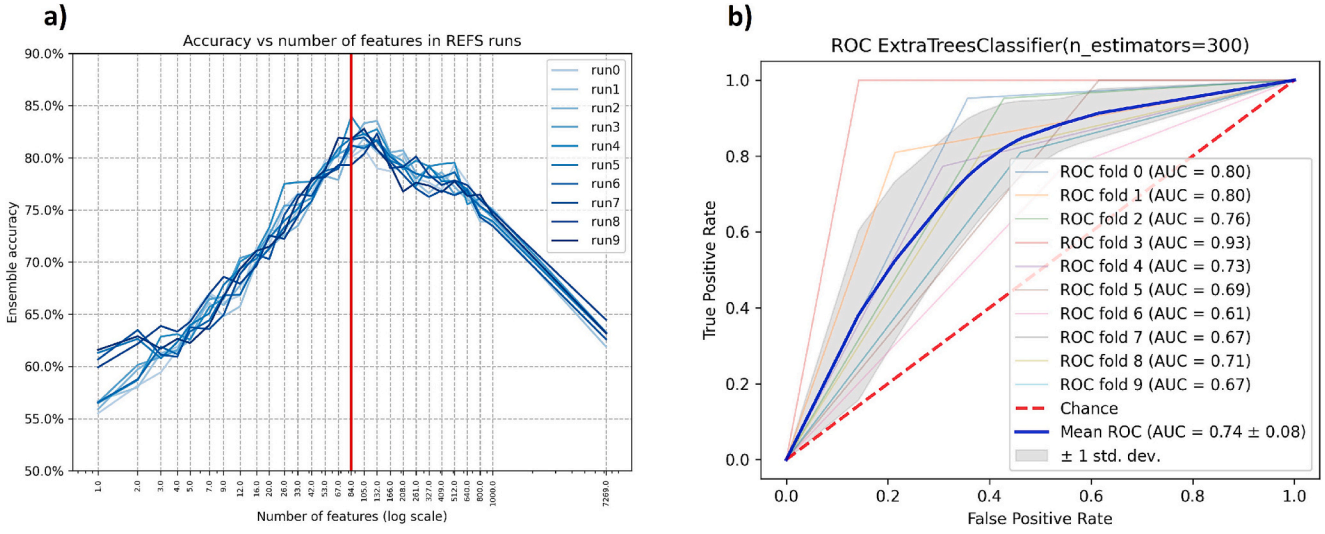

Fig. 2. a) Feature selection phase: the smallest subset of features that achieves the highest accuracy is identified by the red line; b) AUC-ROC for the classifier with the best performance in the validation module for discovery dataset PRJEB14674. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

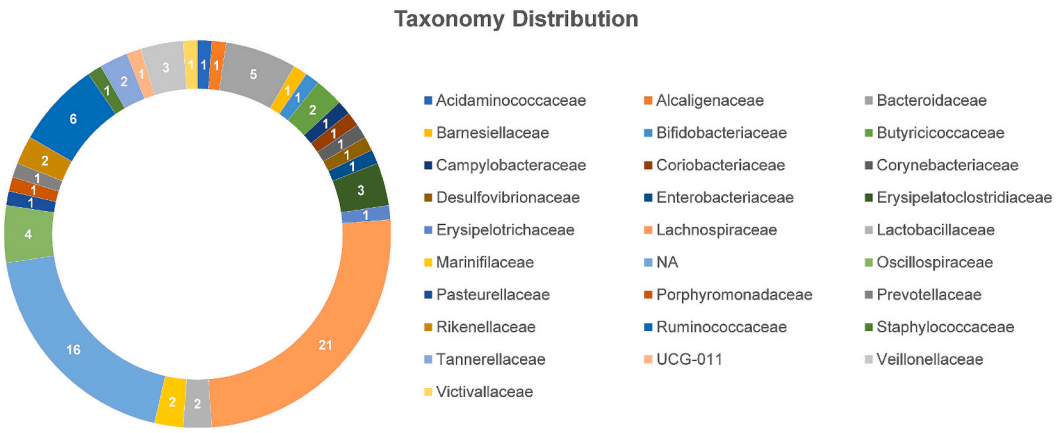

Fig. 3. Taxonomy distribution at family level.

taxonomy distribution of the selected features at *"Family"* level is presented in Fig. 3. *NA* label means there is no taxonomy assigned at *"Family"* level, but in other levels (*"Order", "Class", or "Phylum"*). From the NA, only two features do not have taxonomy assigned to the second lowest level (*"Phylum"*) and two do not have taxonomy assigned to the lowest level (*"Domain"*). The taxonomy assignment at all levels and the corresponding sequence of the 84 features is detailed in Supplementary file 1.

Although we identified 84 features (individual taxa), our taxonomy distribution revealed only 28 bacterial groups (Fig. 3). This can be owed to the fact that we are using ASVs instead of Operational Taxonomy Units (OTUs). ASVs, representing individual taxa, can be tested and validated across independent datasets [11]. They differentiate sequences at the nucleotide level, meaning the taxonomic group remains the same, but the sequence representing each taxa differs [12]. In contrast, OTUs group similar sequences with an average error rate of 3% [13], potentially missing variations (mutations) in specific taxa, which could be crucial for medical applications.

3.3. Testing phase

After searching in each testing dataset for the 84 features identified in the discovery dataset, we find 75/84 in PRJNA742875, 77/84 in PRJEB27564, and finally 73/84 in PRJNA594156. The results after running the validation module were that the classifier with the best performance for the three testing datasets was the Extra Trees classifier, with AUC-ROCs of 0.64 for PRJEB14674, 0.71 for PRJEB27564, and 0.62 for PRJNA594156, see Fig. 4. According to [10], the AUC-ROCs with values of 0.64 and 0.62 are considered as *"sufficient"* diagnostic accuracy, while the AUC-ROC of 0.71 corresponds to a *"good"* diagnostic accuracy. It is important to mention that although AUC-ROC <0.7 could be considered on the edge of what is accepted, they can still be indicators of a reasonable discriminatory ability to diagnose patients with some disease/condition [14].

We conducted an abundance analysis of the 84 taxa by calculating the average abundance of each taxa in the healthy control group of each dataset; the corresponding average value for the Parkinson's group was then computed, and the taxa was classified as increased or decreased, compared to the average for the healthy control. The results of this process, summarized by the heatmap in Fig. 5, show that taxa abundance is not uniform across datasets, likely due to sequence quality and variations in sequencing equipment, a phenomenon known as the *batch effect* [15]. This variability may explain the accuracy results from the validation module in both feature selection and testing phases. To better isolate and study the results, we intentionally did not apply any batch correction process in this work. Correspondingly, we compared the taxa abundance reported in the literature with our findings. Fig. 5 illustrates these comparisons: taxa that are increased in the PD group are shown in dark green, those that are decreased are shown in light green, and taxa for which no information was found in the literature are indicated in black.

3.4. MicrobiomeAnalyst

Additionally to the last analysis, we use a tool called *"MicrobiomeAnalyst"*[2] to identify the relationship between the presence of the identified taxa and the occurrence of diseases [9], with results presented in Fig. 6. Based on the analysis of the provided data, the most numerous group ($p < 0.05$) of diseases falls under the category of Gastrointestinal Diseases. This category includes a variety of conditions such as Colorectal Neoplasms, Inflammatory Bowel Diseases, Constipation, Diarrhea, Enterocolitis, Short Bowel Syndrome, *Clostridium difficile* infection, Irritable Bowel Syndrome, and Dyspepsia. Each of these diseases has been observed to either increase or decrease in prevalence, highlighting the diverse nature of gastrointestinal health issues. This finding underscores the importance of focusing on gastrointestinal diseases in medical research and public health initiatives.

4. Discussion

In conclusion, the taxa identified using the methodology defined by Rojas et al. [6] have a direct relationship with PD. We were able to identify microbiome signatures across independent datasets that can differentiate PD from healthy control groups. Additionally, we discovered new taxa that could be related to PD: *Domain-Bacteria* (taxa 7 and 65), *Genus-Corynebacterium* (taxa 42), *Genus-Caproi ciproducens* (taxa 43), *Genus-Campylobacter* (taxa 78), *Genus-Bilophila* (taxa 26), *Genus-Butyricicoccus* (taxa 40 and 72), and the two unidentified taxa (32 and 66). Analyzing the abundance of these taxa in Fig. 5, we found that most exhibit consistent behavior (increased or decreased) across the four datasets, except for taxa 42 and 72, which were only found in the discovery dataset. This could warrant an in-depth analysis to study their relationship with PD.

In our literature analysis of the resulting taxa, we found that *GenusLactobacillus* (taxa 3 and 24), *Genus-Bifidobacterium* (taxa 1), *Genus-Butyricimonas* (taxa 25 and 51), and *Genus-Roseburia* (taxa 67) are increased in patients diagnosed with PD and are associated with cognitive ability and anxiety as a result of Parkinson's medical treatment [16–19]. *GenusUCG-005* (taxa 5), *Family-Oscillospiraceae* (taxa 6), and *Genus-Collinsella* (taxa 62) are increased in individuals with REM sleep behavior disorder (RBD) and early PD [20].

[2] Available at https://www.microbiomeanalyst.ca/MicrobiomeAnalyst/

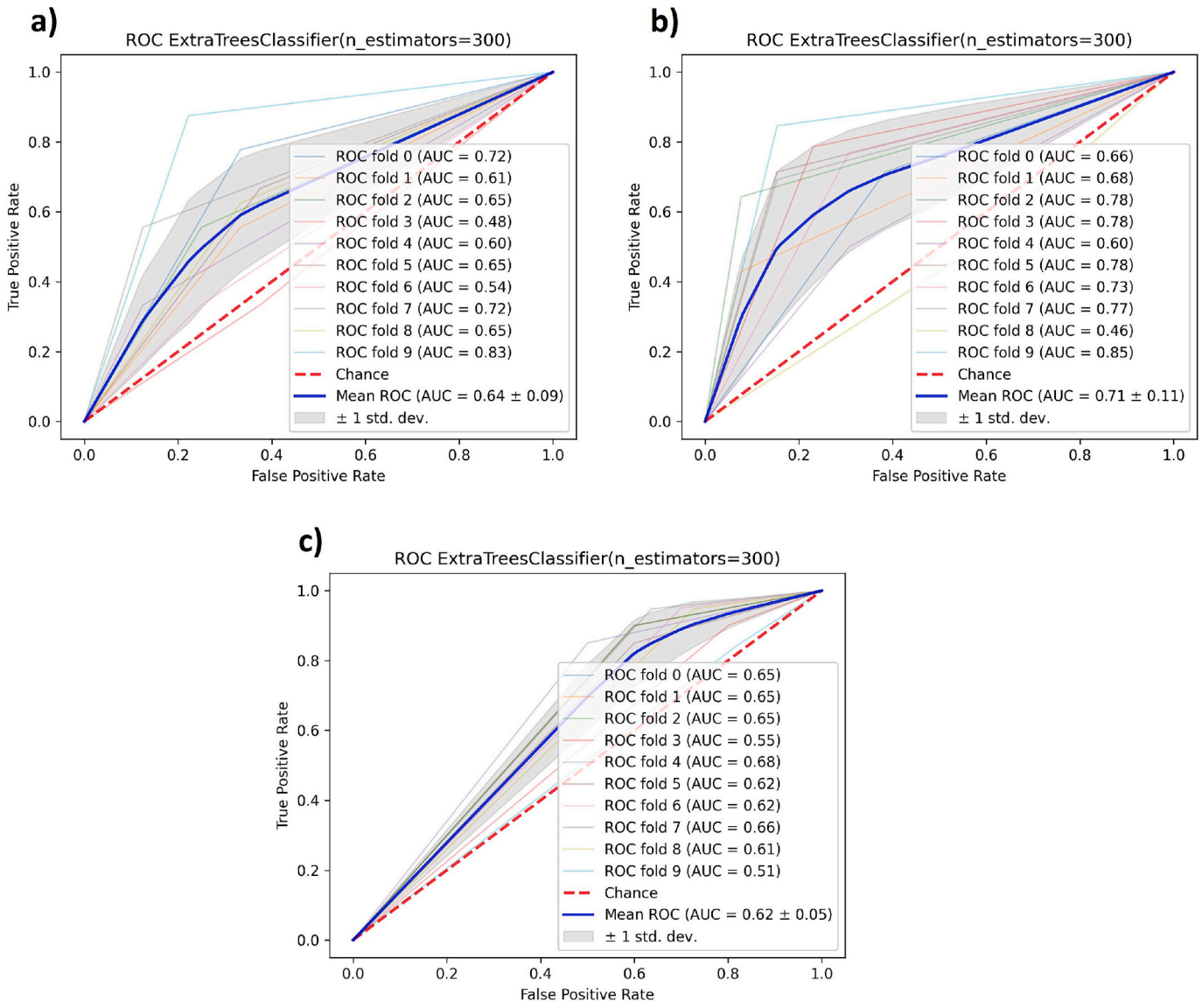

Fig. 4. AUC-ROC plots corresponding to the classifier with the best performance for the three testing datasets (Extra Trees): a) PRJNA742875, b) PRJEB27564, and c) PRJNA594156.

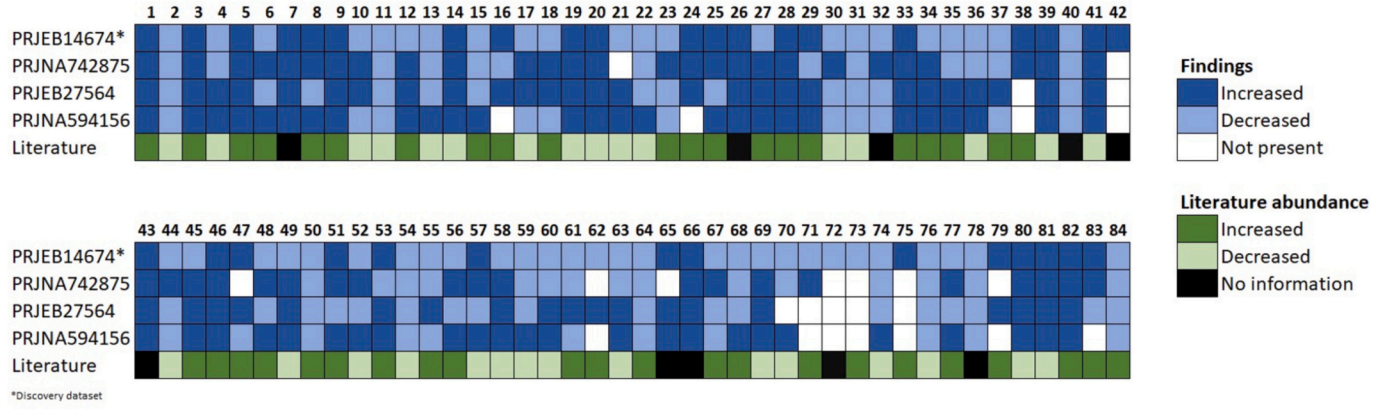

Fig. 5. Heatmap depicting the relative increase or decrease of each taxa in Parkinson's samples compared to control samples and compared to what is reported in the literature.

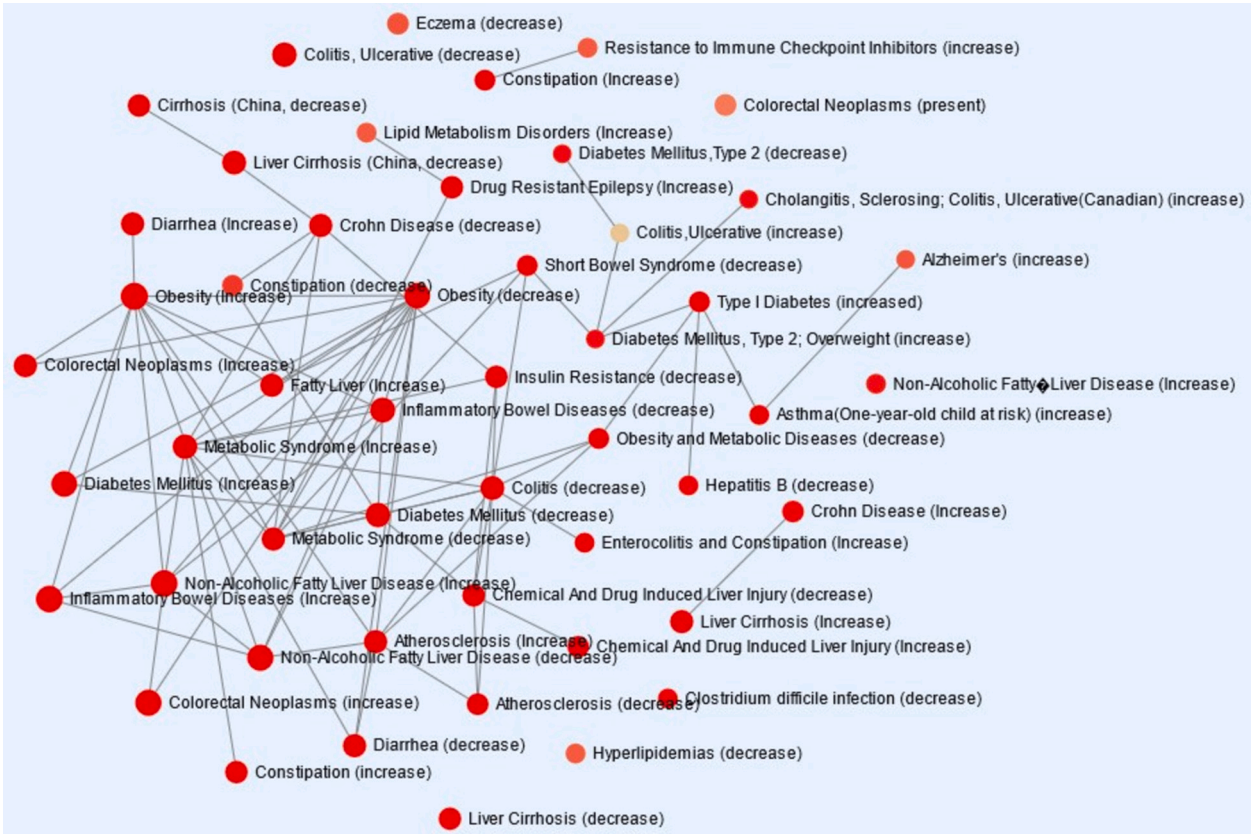

Fig. 6. Relationship between the resulting taxa and diseases using MicrobiomeAnalyst tool. Each node represents a taxa set, is colored based on the *p*-value, and the size depends on the number of elements within the node (hits). Two nodes are connected by an edge if they share >20% of the hits in their combined taxa.

Genus-Alistipes (taxa 8 and 34), *Genus-Barnesiella* (taxa 55), and *GenusRuminococcus* (taxa 84) are elevated in the PD group and are also related to cognitive ability [18,19]. *Genus-Phascolarctobacterium* (taxa 12) is increased in the PD group and has been used as a predictive feature in research involving machine learning algorithms and OTUs [21]. *Genus-Negativibacillus* (taxa 15) is increased in the PD group and has been found beneficial in alleviating disease progression and promoting cognitive function [22].

Genus-Porphyromonas (taxa 16) and *Genus-Parabacteroides* (taxa 33 and 47) have been found to be increased in the PD group [16,17]. *GenusBacteroides* (taxa 18, 37, 45, 53, and 68) is more abundant in patients with the non-tremor PD subtype compared to those with the tremor subtype. Additionally, its abundance correlates with motor symptom severity as defined by the Unified Parkinson's Disease Rating Scale (UPDRS) part III motor scores [17]. The UPDRS is one of the most evaluated, valid, and reliable scales currently available for the clinical study of PD [23]. Comparing the abundances of *Genus-Bacteroides* in Fig. 5, we observe high variability between increased and decreased abundances, possibly due to the lack of information about the PD subtypes analyzed in the dataset studies. *GenusPrevotella* (taxa 22) was found to be decreased in the PD group and has been used as a predictive feature to distinguish the PD group from controls using ML algorithms and OTUs [17,21].

Genus-Oscillospira (taxa 46) is increased in the PD group [18] and is associated with an increased risk of developing PD, indicating that specific microbiome changes can be used for early diagnosis [24]. As shown in Fig. 5, taxa 46 appears increased across all datasets.

Genus-Erysipelatoclostridium (taxa 48 and 83) and *Genus-Victivallis* (taxa 82) are reported as increased, with correlations to earlier age at PD onset and disease severity [18,25]. *Genus-DTU089* (taxa 50) is reported as increased in the PD group compared to the control group [26]. Low levels of *Genus-Barnesiella* at baseline and at 1-year follow-up are associated with worse clinical evolution of PD, with progression analysis indicating potential trends of lower *Genus-Barnesiella* levels in patients with faster disease progression [27].

Genus-Lachnospiraceae UCG-001 (taxa 61), *Class-Bacteroidia* (taxa 77), *Class-Gammaproteobacteria* (taxa 23 and 73), and *Family-Enterobacteriaceae* (taxa 64) are reported as increased in PD mouse models being analyzed for treatment. For example, rifaximin treatment reduced systemic inflammatory responses, prevented motor dysfunction and cognitive impairment, and alleviated neuroinflammation. Additionally, *Lachnospiraceae UCG-001* and *Class-Clostridia* (taxa 17, 19, 20, 39, 41, 44, 49, 57, 59, 69, 70, 80, and 81) were significantly enriched following UC-MSC treatment (protective model) [28]. *Genus-Holdemania* (taxa 79) was found to be increased in the PD group [29].

Family-Lachnospiraceae (taxa 2, 4, 10, 11, 13, 14, 21, 30, 31, 36, 52, 54, 58, 60, 63, and 74) is found to be decreased in the PD group and identified as a relevant feature for distinguishing PD patients from the control group using machine learning [30]. *Family-Ruminococcaceae* (taxa 35 and 56) is increased in PD patients [22]. *Family-Pasteurellaceae* (taxa 76) is decreased in the PD group, especially in early-onset PD [18]. *Phylum-Firmicutes* (taxa 9, 27, 28, 29, 38, 71, and 75) was found to be increased in the PD group [18,19].

From Fig. 6, interestingly, there is a significant connection between gastrointestinal diseases and PD. Gastrointestinal dysfunctions are among the most common and troublesome non-motor symptoms in PD. Constipation, for instance, affects up to 70% of people with Parkinson's and often begins before the onset of the disease's characteristic motor symptoms. This early manifestation of gastrointestinal issues is thought to be linked to the accumulation of alpha-synuclein in the enteric nervous system, which mirrors the pathological changes seen in the brain of Parkinson's patients.

Moreover, conditions like gastroparesis, where the stomach empties its contents slowly, can significantly impact the effectiveness of PD

medications.

This highlights the intricate relationship between gastrointestinal health and the management of PD. Understanding and addressing gastrointestinal symptoms in Parkinson's patients is crucial for improving their overall quality of life and treatment outcomes [31].

Although the strengths of this work are the use of a reproducible standard methodology, validations in independent datasets from different ethnic groups, and robust results that are aligned with the literature, this type of research is limited by the availability of datasets, incomplete information in the metadata, batch effect, small number of samples, the quality of the samples, among others.

As seen in Fig. 5, factors such as sex, age, constipation, gastrointestinal discomfort, geography, and diet could also explain the variations in the taxa abundance. Not many studies consider these factors in their findings, although some research carried out in different ethnic groups such as Germany, Finland, Russia and Japan have shown that the composition of the intestinal microbiome of patients with PD it is altered, but does not depend on these types of factors [16,17], a deep analysis of the impact of these factors is required.

Although the results of this study are promising, further research is needed on these microbiome signatures to develop robust and reliable diagnostic techniques and explore their potential therapeutic implications in targeted medicine. We are planning to make a follow-up study to see the progression of the found taxa at different timepoints. We consider that these founded taxa combined by the timepoints study could be a key for personalized medicine and for a deeper understanding of PD and its progression. Additionally, it is planned to analyze how the microbiome composition can be affected by factors such as specific dietary pattern, other comorbidities, ethnicity, gender, sex, among others.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.maturitas.2024.108185.

Contributors

David Rojas-Velazquez contributed to study concept and design, data collection from the NCBI repository, the design and implementation of the tools used in this work, data analysis and drafting and editing of the paper.

Sarah Kidwai contributed to the biological validation of the information presented in the discussion section and reviewed the draft paper.

Ting Chia Liu contributed to the search for biological information in the literature presented in the discussion section and reviewed the text.

Mounim A. El-Yacoubi contributed to the biological validation of the information presented in the discussion section and reviewed the draft paper.

Johan Garssen contributed to the biological validation of the information presented in the discussion section and reviewed the draft paper.

Alberto Tonda contributed to the design and implementation of the tools used in this work, and reviewed the draft paper.

Alejandro Lopez-Rincon contributed to study concept and design, project supervision, the design and implementation of the tools used in this work, data analysis and drafting and editing of the paper.

All authors approved the final version and no other person made a substantial contribution to the paper.

Ethical approval

We adhere to the ethical guidelines in the handling, processing and storage dictated by the institutions involved in this manuscript.

Provenance and peer review

This article was commissioned and was externally peer reviewed.

Funding

No funding from an external source was received for this study.

Data sharing and collaboration

All datasets used in this research are available in the NCBI public repository.

Declaration of competing interest

The authors declare that they have no competing interest.

Acknowledgements

This work used the Dutch national e-infrastructure with the support of the SURF Cooperative using grant no. EINF-6769.

References

[1] J. Mei, C. Desrosiers, J. Frasnelli, Machine learning for the diagnosis of parkinson's disease: a review of literature, Front. Aging Neurosci. 13 (2021) 633752.
[2] A.H. Tan, S.Y. Lim, A.E. Lang, The microbiome–gut–brain axis in parkinson disease—from basic research to the clinic, Nat. Rev. Neurol. 18 (8) (2022) 476–495.
[3] A.B. Singleton, M.J. Farrer, V. Bonifati, The genetics of p arkinson's disease: Progress and therapeutic implications, Mov. Disord. 28 (1) (2013) 14–23.
[4] A. Ouyang, Artificial intelligence model can detect parkinson's from breathing patterns, MIT News (2022). https://news.mit.edu/2022/artificial-intelligence-can-detect-parkinsons-f.
[5] J.M. Templeton, C. Poellabauer, S. Schneider, Classification of parkinson's disease and its stages using machine learning, Sci. Rep. 12 (1) (2022) 1–12. https://www.nature.com/articles/s41598-022-18015-z.pdf.
[6] D. Rojas-Velazquez, S. Kidwai, A.D. Kraneveld, A. Tonda, D. Oberski, J. Garssen, A. Lopez-Rincon, Methodology for biomarker discovery with reproducibility in microbiome data using machine learning, BMC bioinformatics 25 (1) (2024) 26.
[7] B.J. Callahan, P.J. McMurdie, M.J. Rosen, A.W. Han, A.J.A. Johnson, S.P. Holmes, Dada2: high-resolution sample inference from illumina amplicon data, Nat. Methods 13 (7) (2016) 581–583.
[8] A. Lopez-Rincon, M. Martinez-Archundia, G.U. Martinez-Ruiz, A. Schoenhuth, A. Tonda, Automatic discovery of 100-mirna signature for cancer classification using ensemble feature selection, BMC bioinformatics 20 (2019) 1–17.
[9] A. Dhariwal, J. Chong, S. Habib, I. King, L. Agellon, J. Xia, Microbiomeanalyst - a web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data, Nucleic Acids Res. 45 (W1) (2017) W180–W188, https://doi.org/10.1093/nar/gkx295. https://www.microbiomeanalyst.ca/.
[10] A.-M. Simundić, Measures of diagnostic accuracy: basic definitions, ejifcc 19 (4) (2009) 203.
[11] Y.-H. Zhou, P. Gallins, A review and tutorial of machine learning methods for microbiome host trait prediction, Front. Genet. 10 (2019) 579.
[12] B.J. Callahan, P.J. McMurdie, S.P. Holmes, Exact sequence variants should replace operational taxonomic units in marker-gene data analysis, ISME J. 11 (12) (2017) 2639–2643.
[13] J.T. Jeske, C. Gallert, Microbiome analysis via otu and asv-based pipelines—a comparative interpretation of ecological data in wwtp systems, Bioengineering 9 (4) (2022) 146.
[14] J.N. Mandrekar, Receiver operating characteristic curve in diagnostic test assessment, J. Thorac. Oncol. 5 (9) (2010) 1315–1316.
[15] A.L. Rincon, A.D. Kraneveld, A. Tonda, Batch correction of genomic data in chronic fatigue syndrome using cma-es, in: Proceedings of the 2020 Genetic and Evolutionary Computation Conference Companion, 2020, pp. 277–278.
[16] Z.D. Wallen, M. Appah, M.N. Dean, C.L. Sesler, S.A. Factor, E. Molho, C. P. Zabetian, D.G. Standaert, H. Payami, Characterizing dysbiosis of gut microbiome in pd: evidence for overabundance of opportunistic pathogens, npj Parkinson's Disease 6 (1) (2020) 11.
[17] C.-H. Lin, C.-C. Chen, H.-L. Chiang, J.-M. Liou, C.-M. Chang, T.P. Lu, E.Y. Chuang, Y.-C. Tai, C. Cheng, H.-Y. Lin, et al., Altered gut microbiota and inflammatory cytokine responses in patients with parkinson's disease, J. Neuroinflammation 16 (2019) 1–9.
[18] Z. Li, H. Liang, Y. Hu, L. Lu, C. Zheng, Y. Fan, B. Wu, T. Zou, X. Luo, X. Zhang, et al., CNS Neurosci. Ther. 29 (1) (2023) 140–157.
[19] T. Ren, Y. Gao, Y. Qiu, S. Jiang, Q. Zhang, J. Zhang, L. Wang, Y. Zhang, L. Wang, K. Nie, Gut microbiota altered in mild cognitive impairment compared with normal cognition in sporadic parkinson's disease, Front. Neurol. 11 (2020) 137.
[20] B. Huang, S.W. Chau, Y. Liu, J.W. Chan, J. Wang, S.L. Ma, J. Zhang, P.K. Chan, Y. K. Yeoh, Z. Chen, et al., Gut microbiome dysbiosis across early parkinson's disease, rem sleep behavior disorder and their first-degree relatives, Nat. Commun. 14 (1) (2023) 2501.
[21] J.R. Bedarf, F. Hildebrand, L.P. Coelho, S. Sunagawa, M. Bahram, F. Goeser, P. Bork, U. Wüllner, Functional implications of microbial and viral gut

metagenome changes in early stage l-dopa-naïve parkinson's disease patients, Genome Med. 9 (2017) 1–13.
[22] D. Kwon, K. Zhang, K.C. Paul, A.D. Folle, I. Del Rosario, J.P. Jacobs, A.M. Keener, J. M. Bronstein, B. Ritz, Diet and the gut microbiome in patients with parkinson's disease, npj Parkinson's Disease 10 (1) (2024) 89.
[23] C. Ramaker, J. Marinus, A.M. Stiggelbout, B.J. Van Hilten, Systematic evaluation of rating scales for impairment and disability in parkinson's disease, Movement disorders: official journal of the Movement Disorder Society 17 (5) (2002) 867–876.
[24] X. Zhang, B. Tang, J. Guo, Parkinson's disease and gut microbiota: from clinical to mechanistic and therapeutic studies, Translational Neurodegeneration 12 (1) (2023) 59.
[25] D. Rosario, G. Bidkhori, S. Lee, J. Bedarf, F. Hildebrand, E. Le Chatelier, M. Uhlen, S.D. Ehrlich, G. Proctor, U. Wüllner, et al., Systematic analysis of gut microbiome reveals the role of bacterial folate and homocysteine metabolism in parkinson's disease, Cell Rep. 34 (9) (2021).
[26] K. Zhang, K.C. Paul, J.P. Jacobs, H.-C.L. Chou, A. Duarte Folle, I. Del Rosario, Y. Yu, J.M. Bronstein, A.M. Keener, B. Ritz, Parkinson's disease and the gut microbiome in rural California, J. Parkinsons Dis. 12 (8) (2022) 2441–2452.
[27] A. Varesi, L.I.M. Campagnoli, F. Fahmideh, E. Pierella, M. Romeo, G. Ricevuti, M. Nicoletta, S. Chirumbolo, A. Pascale, The interplay between gut microbiota and parkinson's disease: implications on diagnosis and treatment, Int. J. Mol. Sci. 23 (20) (2022) 12289.
[28] Z. Sun, P. Gu, H. Xu, W. Zhao, Y. Zhou, L. Zhou, Z. Zhang, W. Wang, R. Han, X. Chai, et al., Human umbilical cord mesenchymal stem cells improve locomotor function in parkinson's disease mouse model through regulating intestinal microorganisms, Frontiers in Cell and Developmental Biology 9 (2022) 808905.
[29] Y. Qian, X. Yang, S. Xu, C. Wu, Y. Song, N. Qin, S.-D. Chen, Q. Xiao, Alteration of the fecal microbiota in chinese patients with parkinson's disease, Brain Behav. Immun. 70 (2018) 194–202.
[30] D. Pietrucci, A. Teofani, V. Unida, R. Cerroni, S. Biocca, A. Stefani, A. Desideri, Can gut microbiota be a good predictor for parkinson's disease? A machine learning approach, Brain Sci. 10 (4) (2020) 242.
[31] Parkinson's Foundation, Constipation and other gastrointestinal problems in parkinson's disease, 2024. https://www.parkinson.org/library/fact-sheets/constipation.