ESCI upf.
School of International Studies

MEMÒRIA DEL TREBALL DE FI DE GRAU DEL GRAU (ESCI-UPF)

Machine learning on gut microbiome reveals potential biomarkers for Parkinson's diagnosis

AUTOR/A: Xènia Roda Sánchez	**NIA:** 105513
GRAU: Bioinformàtica	

CURS ACADÈMIC: 2021 - 2022

DATA: 21/06/2022

TUTOR/S: Andreu Paytuví

Bachelor's Degree in Bioinformatics (UPF-UPC-UB-UAB)
Final Grade Project

Machine learning on gut microbiome reveals potential biomarkers for Parkinson's diagnosis

Xènia Roda Sánchez

Scientific director: Andreu Paytuví[1]

[1]Sequentia Biotech SL, Barcelona, España.

Abstract

Motivation: Parkinson's disease (PD) is the second most common neurodegenerative disorder after Alzheimer's disease. It develops when nerve cells die or become impaired, losing the ability to produce an important chemical called dopamine. Gut microbiota have been studied in relation to the pathophysiology of Parkinson's disease (PD) due to the early gastrointestinal symptomatology and the presence of α-synuclein pathology in the enteric nervous system, hypothesized to ascend via the vagal nerve to the central nervous system. Recent studies report Bacteriophages to the list of possible factors associated with the development of PD, showing shifts of the phage/bacteria ratio in lactic acid bacteria known to produce dopamine and regulate intestinal permeability.

Results: The objective of this study was to discover biomarkers through differences between the gut microbiome of controls and PD patients, by identifying candidate taxa, gene families and pathways to get insight of some possible variables that could become important for the early detection of the disease. Here, shotgun metagenomic data is analyzed with the read-based approach in order to compare the microbiome compositions of 20 control subjects and 20 PD patients using different metagenomics programs and machine learning algorithms. The most relevant features found were an increased abundance of *Lactococcus* phage and an overexpression of the Myo-chiro and scyllo-inositol degradation pathway in patients with PD.

1. Introduction

Parkinson's disease (PD) is the second most common neurodegenerative disease worldwide that afflicts about 1%-2% of the population aged over 65 years.[1,2] Growing lines of evidence suggest genetic and environmental risk factors play important roles in the pathogenesis of PD.[3–5] The disease is characterized by pathological accumulation of the protein α-synuclein (αS) leading to the slow and progressive degeneration of dopaminergic neurons in the *pars compacta* of the *substantia nigra* (SN).[6,7] Despite the huge amount of research about the disease, the cause of the neural loss in Parkinson's disease is poorly understood, and includes an emerging body of evidence suggesting that activation of neuro-inflammatory mechanisms contribute to the neurodegenerative process.[8] Evidences suggest that the accumulation of α-synuclein starts in the gut years before affecting the central nervous system (CNS); this is coupled with the dysbiosis state in the gut, where a change in the bacteria population likely leads to an inflammatory reaction, causing abnormal permeability of the intestinal epithelium,[9] allowing bacterial products enter to the circulatory system and go to the CNS, via the sympathetic nervous system, the glossopharyngeal nerve and the vagus nerve.[6]

Once α-synuclein reaches the CNS; it is believed that there is a spread in a prior-like fashion transferring from affected to unaffected cells acting as a template, promoting misfolding of the normal α-synuclein in the host. This process leads to the formation of larger aggregates, neural dysfunction and neurodegeneration. Indeed, recent reports demonstrate that a single intracerebral inoculation of misfolded α-synuclein can induce Lewy-like pathology in cells that can spread from affected to unaffected regions and can induce neurodegeneration with motor disturbances in both transgenic and normal mice.[10] Importantly, there is a strong bidirectional interaction between the gut microbiota and the central nervous system (CNS), a connection termed ‘microbiota-gut-brain-axis’. Dysregulation of the brain-gut-microbiota axis in PD may be associated with gastrointestinal manifestations frequently preceding motor symptoms, as well as with the pathogenesis of PD itself, supporting the hypothesis that the pathological process is spread from the gut to the brain *(Figure 1)*.[11]

However, the factors promoting alterations of gut bacteria in neurodegenerative diseases remain unexplored. Therefore, understanding the mechanisms underlying shifts in the intestinal bacterial community that may trigger pathogenic pathways leading to PD is essential for the development of new approaches to prevent, treat and diagnose this incurable disease.

The microbial community of the human GI tract is composed of bacteria, archaea, fungi, and viruses, including bacteriophages; this highly diverse and complex ecosystem is characterized by dynamic stability. Bacteriophages are the most abundant group outnumbering other viral as well as bacterial species, and are considered important regulators of microbiota stability. [12] However, bacteriophages as possible agents that may negatively affect mammalian health have attracted scientific attention only recently.[12,13]

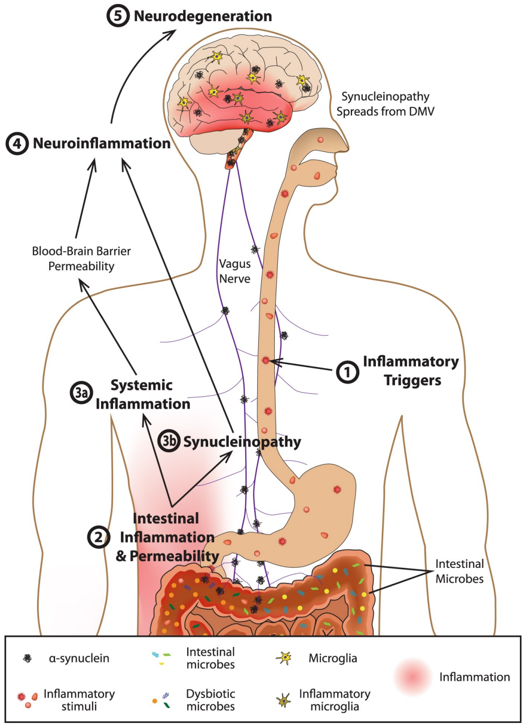

Figure 1: Model of gut-originating, inflammation-driven PD pathogenesis. *In a susceptible individual, inflammatory triggers (1) initiate immune responses in the gut that deleteriously impact the microbiota, increase intestinal permeability, and induce increased expression and aggregation of αSYN (2). Synucleinopathy may be transmitted from the gut to the brain via the vagus nerve (3b), and chronic intestinal inflammation and permeability promote systemic inflammation, which, among other things, can increase blood-brain barrier permeability (3a). Intestinal inflammation, systemic inflammation, and synuclein pathology in the brain all promote neuroinflammation (4) which drives the neurodegeneration that characterizes PD (5). Figure extracted from “The gut-brain axis: is intestinal inflammation a silent driver of Parkinson's disease pathogenesis?”, by Houser, M.C., Tansey, M.G. (2017).*

It seems self-evident that if phages have the potential to modulate the gut microbiota, then in turn they can have an indirect but important impact on host-microbe interactions and thus on host health. Furthermore, phages can translocate through the gut mucosa to local lymph nodes and internal organs, leading to intimate interactions with the host immune system. An unanswered question is to what extent the composition and flux in the phageome could be used as biomarkers of the microbiota, and thus as indirect biomarkers of health or disease in the host.[14]

Various clinical studies have shown evidence indicating the occurrence of intestinal dysbiosis in

PD patients compared with healthy controls, and the compositions of both fecal and mucosal microorganisms have been reported to change in PD patients.[15,16] These differences become particularly pronounced at the family, genus, and operational taxonomic unit (OTU) levels. At taxonomic level butyrate-producing bacteria were reported to be much more abundant in fecal samples from controls than in those from PD patients.[16] One study reported that the putative cellulose-degrading bacteria, from the genera *Blautia*, *Faecalibacterium*, and *Ruminococcus*, were significantly decreased, whereas the putative pathobionts from the general *Escherichia-Shigella*, *Streptococcus Proteus*, and *Enterococcus* were significantly increased, in PD subjects compared with healthy controls.[17]

1.1 Objectives

Considering the limitations of statistical testing methods, we are interested in performing predictive analysis from microbiome data. In this study, we used next-generation sequencing to analyze and compare gut microbiota data from stool samples from control subjects and PD patients submitted by the National Taiwan University. We have employed statistical machine learning methods to make predictions based on taxonomic and pathway abundance on the microbial samples collected. We hypothesize that the fecal microbiome of PD patients differs from the controls in terms of taxonomic composition and pathway abundance. The results may help the establishment of the association between gut microbiota and PD etiology and to elucidate important biomarkers for PD making use of artificial intelligence techniques.

2. Methods

2.1 Data collection

The first part of the project was intended to search for WGS metagenomic data from the gut microbiome of healthy patients and patients who had the disease in the NCBI database. This was accomplished by the data from the Sequence Read Archive (SRA), available through multiple cloud providers and NCBI servers, which is the largest publicly available data repository of high throughput sequencing data. SRA stores raw sequencing data to enhance reproducibility and facilitate new discoveries through data analysis.

To obtain the needed data for our analysis we searched in the NCBI SRA webpage (https://www.ncbi.nlm.nih.gov/sra) using the keyword "parkinson gut metagenome". We used the data from the accession PRJNA762484 which came from PD patients and healthy controls. The data was submitted by the National Taiwan University; Illumina NovaSeq 6000 was the instrument used and were paired-end reads.

In this work, 40 samples of sequenced-reads from the gut microbiota of PD and healthy patients (> 41 years) have been downloaded, having 20 for each group. Afterwards, for downloading the samples we used the SRA toolkit, which is connected with the NCBI.

2.2 Quality control

A quality check with FastQC (v0.11.5) was made to check the quality of the sequencing. After that, we used the program bbduk (v38.96) to remove the bad-quality portions of the reads.[18] The forward and reverse files were our input for bbduk, also specifying some parameters such as the number of threads or CPU's, which was four: threads=4, then a minimum length of 50, with the argument: minlength=50; then we define the start of the trimming, which was on the right (3'): qtrim=r. Finally, we define the minimum Phred score as trimq=25. Then, we executed again FastQC on the trimmed data.

2.3 Taxonomic classification

The assembly-free approach was performed in order to do the taxonomic classification step with Centrifuge (v1.0.4_beta)[19] in order to assign taxonomic labels to the sequencing reads. We performed the analysis with R-studio (v.4.1.2), a free environment software for graphics and statistics computations, having as input the output files generated from the program. Only the rows having the "species" annotation in the column taxRank of the file were selected. After that we merged all sample files into one. A filtering was done across all samples whose species had only one read support, in order to remove putative false positives. Moreover, we removed the species *Homo sapiens* as it was a contamination of the host. Subsequently, we made a stacked bar plot showing the abundances of each species in each sample collected. Remark that for this analysis a VM instance in Google Cloud was created with a 120GB of RAM, 50 GB of swap memory and 32 vCPU due to the high computational cost.

To complement our taxonomic analysis, we studied the results with another program, which was Gaia (v2.0), a metagenomics analysis tool developed by Sequentia Biotech. Gaia is an online metagenomics integrated suite which is able to perform metagenomics analysis for both amplicon and WGS metagenomics, as well as metatranscriptomics.[20]

2.4 Functional classification

Metagenomes and metatranscriptomes were functionally profiled using HUMAnN3 (v3.0.1)[21] to quantify genes and pathways. A concatenated file was created by merging the forward and reverse files to serve as our input for the program. Three output files for each input file were generated, one containing the abundance of each gene family in the community, which abundance was reported in RPK (reads per kilobase), the second file created will contain details for the abundance of each pathway in the community, and the last one is the pathway coverage file, which provides an alternative description of the presence (1) and absence (0) of pathways in a community, independent of their quantitative abundance. We performed the analysis with the first two files: the gene family and the pathway abundance files. The two files were normalized in order to facilitate the comparisons between samples with different sequencing depths. The HUMAnN 3.0 renorm table tool was used to compute the normalized abundances, which was converted to "copies per million" (CPM) units. In addition, the data was filtered by removing the UNINTEGRATED values.

2.5 Machine learning

The next part of the project was focused on supervised machine learning methods, whose objective was to set up predictive models based on training samples already tagged, for later making predictions or inferences to samples not labeled. There were selected five different machine learning algorithms to evaluate how well our algorithms will perform in order to classify patients according to their condition, making use of the data obtained in the taxonomic and functional analysis. The used algorithms are:

- Decision Tree
- Random Forest
- Naive Bayes
- SVM or Support Vector Machines
- k-NN or k-Nearest Neighbor

2.6 Data preparation for ML

The data from the taxonomic and functional classification was transformed into a matrix containing as many rows as samples available in the experiment and as many columns as features (taxa/genes/pathways), and also an extra column containing the metadata belonging to the condition of the patients: healthy (without PD) or sick (with PD) is needed.

A pre-processing step of the data obtained from Centrifuge was made. We normalized the species abundances into a scale of rank from 0 to 100, to make sure that each variable contributes equally into the analysis.

For the functional gene family and the pathway matrix, we removed the pathways/genes classified for each species, leaving us only with the abundance of each unique pathway and gene family independently from the derived species. In addition, a transformation of the variable "condition" to a factor was done, since, unlike the rest of the variables, it is a categorical variable (PD/control). Once the 3 matrices were ready, they were transformed into CSV format in order to work with them in R.

2.7 Leave-One-Out Cross-Validation (LOOCV)

Since our dataset was small, this can lead to model overfitting during training and biased estimates of model performance. Leave-One-Out Cross-Validation (LOOCV) was the method used to train our ML algorithms in order to evaluate the performance of the model on our dataset. LOOCV is an extreme version of k-fold cross-validation that has the maximum computation cost. The method works by splitting the dataset into a training set and a testing set, using all but one observation as a part of the training set and evaluating the model on the test using all the observations in the train dataset. The process is repeated n times (where n is the total number of observations in the dataset), leaving out a different observation from the training set each time. The benefit of so many fit and evaluated models becomes a more robust estimate of model performance as each row of data is given an opportunity to represent the entire of the test dataset.[22–24]

2.8 Feature selection by Boruta algorithm

Some predictive modeling problems have a large number of variables that can slow the development and training of models;[25] in our case we initially had 5,657 species, 487 pathways and 735 gene families. Additionally, the performance of some models can degrade when including input variables that are not relevant to the target variable. Boruta was used as our feature selection algorithm which works as a wrapper algorithm around Random Forest, provided by the package "Boruta" v(7.0.0) in R and was applied on the three matrices: the gene family, the taxonomic and the pathway abundance.[26]

2.9 Project code

All code for this project is available at GitHub link **:** https://github.com/xeniaroda/ML_GutMicrobiome.git
Gaia analysis is available at this link: https://metagenomics.sequentiabiotech.com/shared/TaskFlow/667339ea-8a13-4e4d-826f-b204ab9da925/1f0a21ef-17b5-4ec0-917b-2e3b88e64c3e

3. Results

A total of 40 North-eastern Han Chinese included 20 healthy elders and 20 Parkinson's disease cases were included in the study. Detailed information about the clinical data was presented in *Supplementary Table 1*.

3.1 Taxonomic profiling of the fecal microbiota

Clean sequence reads were used by Centrifuge to classify accurately our reads for species quantification. *Bacteroides* (49.71% on average across all samples) were dominated but with high variation in terms of abundance across all samples. *Bacteroides* species found were: *Bacteroides cellulosilyticus, B. fragilis, B. ovatus, B. sp. I48 and B. thetaiotaomicron* with on-average abundances of 9.64%, 7.14%, 13.67%, 7.98%, 10.27%, respectively. The abundances of phylum *Firmicutes* was increased by 1.53% more in the PD patients than controls, being *Ruminococcus bicirculans* the one with more difference (1.23% in controls and 2.21% of abundances in the PD condition). In addition, the family *Bifidobacteriaceae* was also reported with a higher abundance, for *Bifidobacterium adolescentis* (0.97%) and *Bifidobacterium longum* (0.913%) in PD. *Prevotella intermedia* showed minor levels in controls (0.73%) compared to PD patients (0.97%), and *Prevotella enoeca* did not show major differences between control or PD patients, with abundances of 0.64% and 0.67%, respectively. What is more, *Proteobacteria* phylum showed high abundance in control subjects specifically regarding *Scherichia coli* (5.62%) compared to PD patients (1.9%) *(Figure 2)*.

The abundance of microbial metabolic pathways and gene families from the sequencing data was accurately and efficiently profiled by HUMAnN 3.0. Results were visualized as two heatmaps *(Figure 3)*.

A comparative analysis with Gaia was made with the aim to compare the similarities and differences in the microbiome composition of controls and PD patients. The program shows us a PCA for each taxonomic level. In general, in all PCA's made by the program, samples from both conditions overlap, thus not showing any significant differences between the microbiome composition from PD and controls. Having the control group as a reference, results show *Lactococcus* as the genus more over-represented among others followed by the genus *Staphylococcus* and *Acidiphillum* in control subjects. Regarding species, *Aeromonas veronii* was found to be under-represented in control subjects compared to PD patients and *Subdoligranulum sp., Lactobacillus plantarum* over expressed in controls *(Supplementary Figure 4)*.

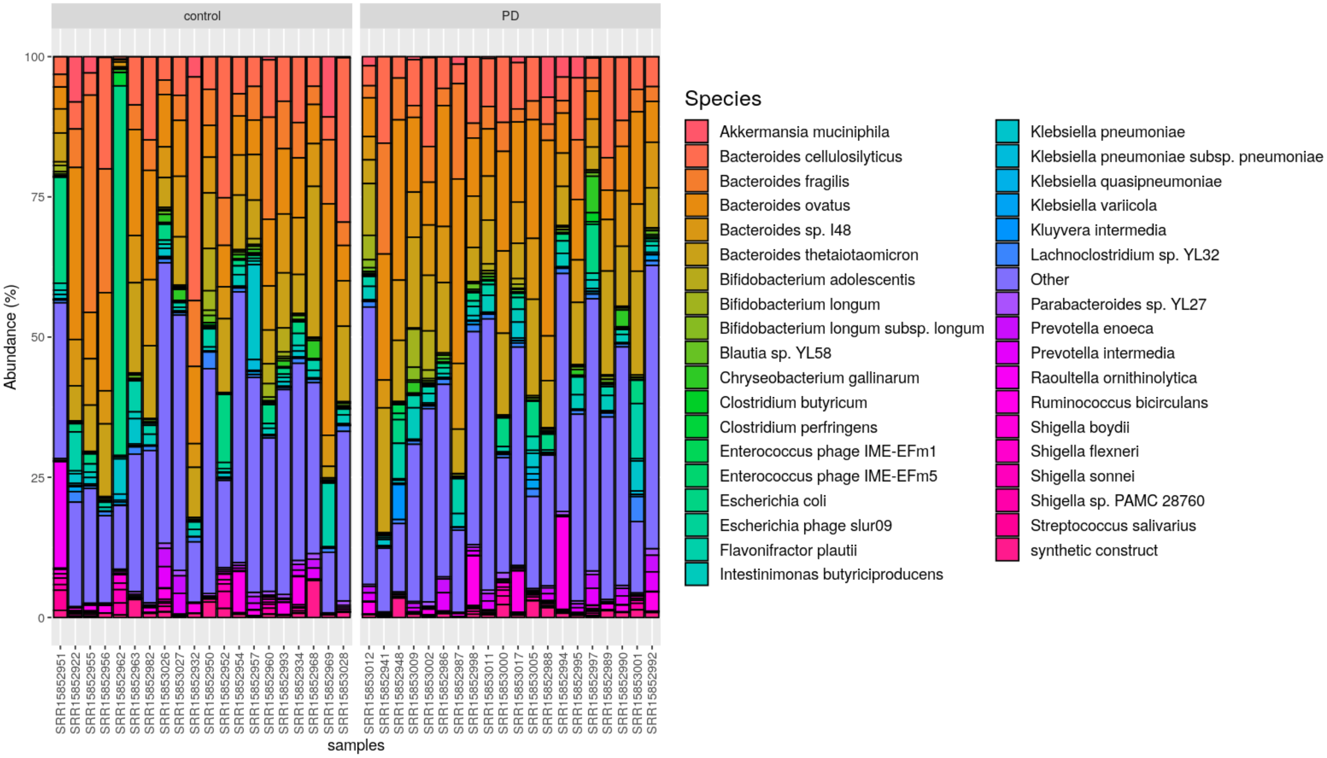

Figure 2: Stacked bar plot depicting relative abundances of all microbiota per sample. *Species abundances were computed, removing the ones that were below 1% for plotting purposes. Each vertical bar depicts the relative abundance for the PD or control condition for a given sample with a legend corresponding to its species classification. Label "Other" from the legend represents all those species removed from plotting purposes, having an abundance below 1%.*

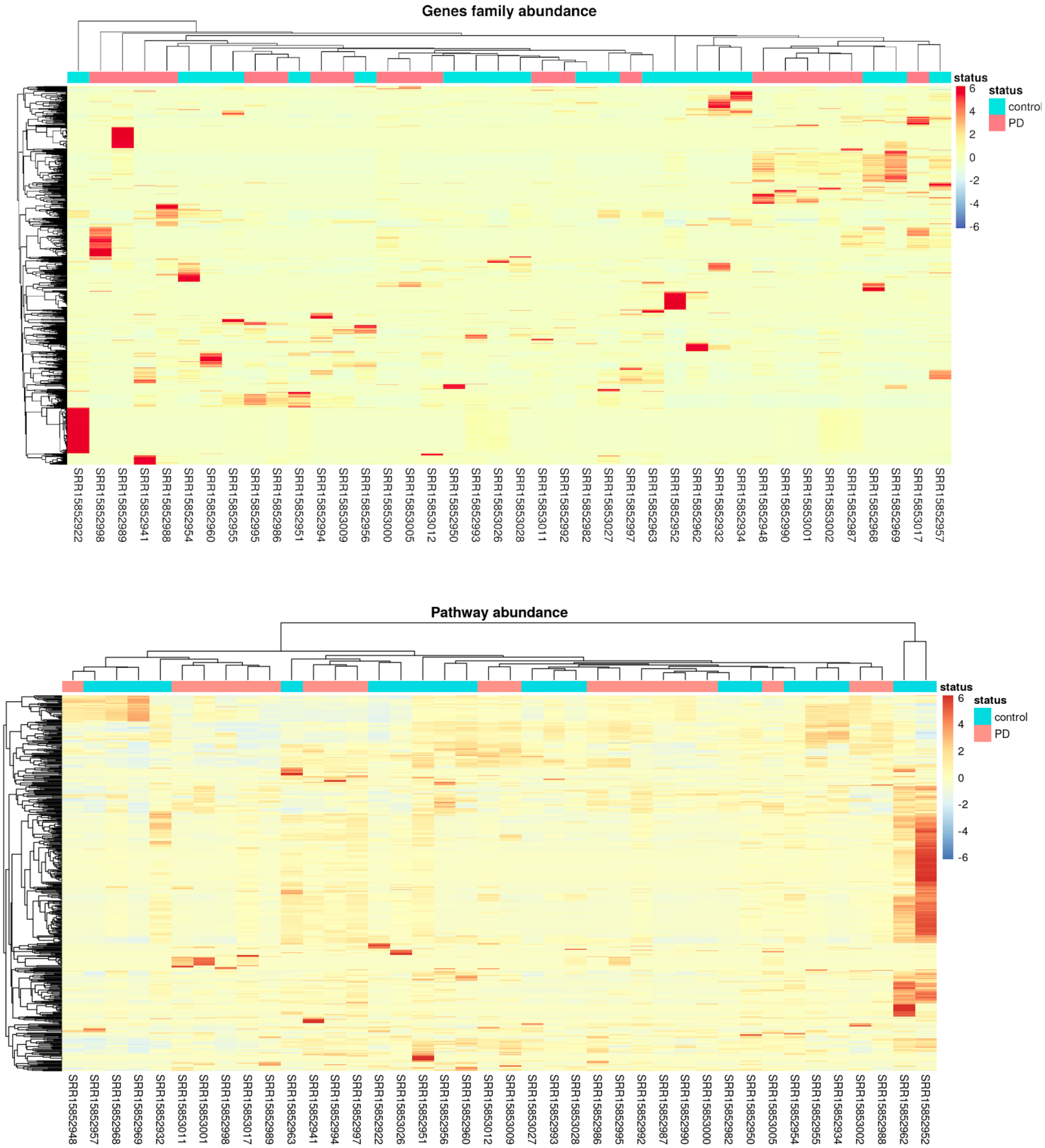

Figure 3: Gene family (top) and pathway abundance (bottom) heatmap. *Columns colored as blue represent the control condition, and columns colored in red represent Parkinson's condition. Results did neither show PD samples nor control samples clustered together.*

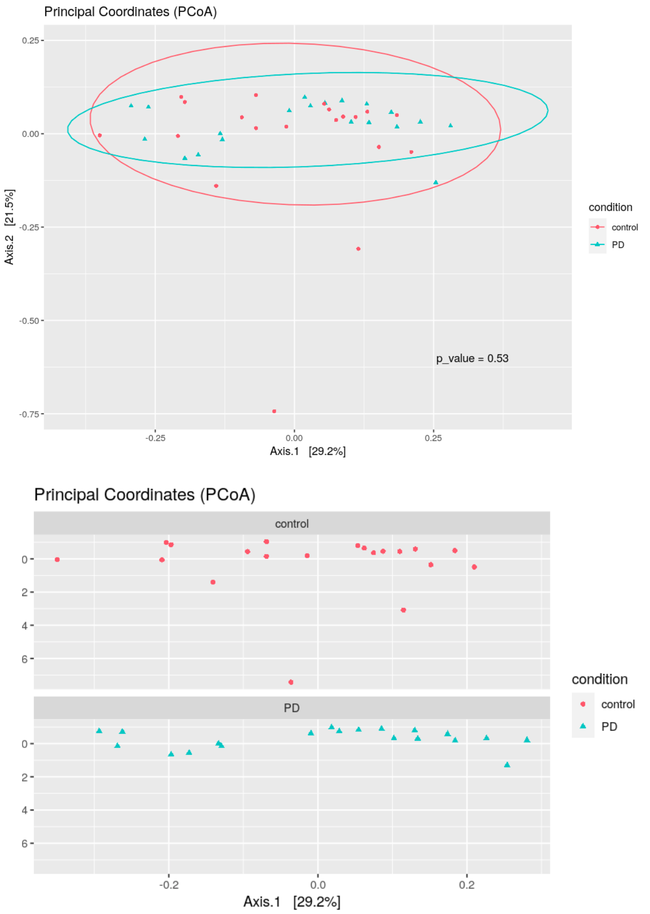

Figure 5: PCoA on taxonomic profiles. *Each point represents a single sample, and the distance between points represents how compositionally different the samples are from one another. The points are colored by health state, not showing a clear difference in the microbial community composition between diseased (blue) and healthy (red).*

PCoA was used as a method to explore and to visualize similarities or dissimilarities of our data. Ordination techniques, such as PCoA, reduce the dimensionality of microbiome data sets so that a summary of the beta diversity relationships can be visualized in two- or three-dimensional scatterplots.[27] A phyloseq object with the species and the metadata was created to perform the task with the package Phyloseq (v.1.26.1). PCoA data showed that individuals of the two groups partially overlapped but they still have their own trend to aggregate separately *(Figure 5)*. A PERMANOVA test was made together with PCoA, having a p-value>0.05, thus, accepting the null hypothesis that there were no significant differences in community composition between both conditions.

Furthermore, a differential abundance analysis with the R package DESeq (v.12.3) was made having bacteriophages as the organisms more represented in our result (*Supplementary Table 2*).

3.2 Selection of the best ML algorithm

As we show in the description of the microbiome, there are no big differences between both experimental groups (control and PD) as they share similar taxonomic profiles. In this context, we aimed to identify subtle differences that could discriminate between both groups by means of Machine Learning (ML) algorithms. For this purpose, we used taxonomic and functional (gene families and pathways) information, separately.

First of all, the five machine learning algorithms were trained with all the variables we had. *Supplementary Table 3* shows the resulting metrics after applying Leave-One-Out Cross-Validation (LOOCV) for each matrix. The best parameters selected by the method are presented in the gray column for each algorithm. In general the predictive power was very poor being decision trees the best classification algorithm for our dataset. However, regarding pathway and gene family the best algorithm with an AUC-ROC of 0.560 was the Naive Bayes. Even so, having those metrics we could not rely on decision trees or Naive Bayes to become our model of preference to perform the classification task. To overcome that, we extracted the most important variables for each feature we have (taxa/genes/pathways) by making use of the Boruta algorithm. Important differences in abundances were shown regarding phages, specially the phage *Lactococcus 50101*, *Escherichia* phage PBECO.4 and *Escherichia* phage vB_EcoM_Alf5 *(Figure 6)*; regarding pathways we found an important difference in abundance for Myo-chiro and scyllo-inositol degradation and RUMP formaldehyde oxidation I (*Supplementary Figure 7)*. Subsequently, we again trained the ML algorithms with the extracted features, with their metrics shown in the *Supplementary Table 4*, which showed better results. Random Forest seems to perform well in all 3 matrices, having a Kappa value of 0.80 and a AUC-ROC of 0.970. In addition, Support Vector Machine (SVM) has a great Kappa and AUC-ROC (0.80 and 0.910, respectively).

3.3 Machine Learning with Python: Classification script

With the aim to automate the prediction of a sample within the framework of personalized medicine, we aimed to create a Python script whose objective was to get the classification probability for a number of taxonomic samples from the gut microbiome classifying either for PD or controls together with the metrics of the trained algorithms, also showing the best one for our classification *(Figure 8).*

Out of the 3 matrices presented in our work, we selected the taxonomic to become our discrimination matrix for the classification task as it was shown to have more important variables selected and better accuracy metrics.

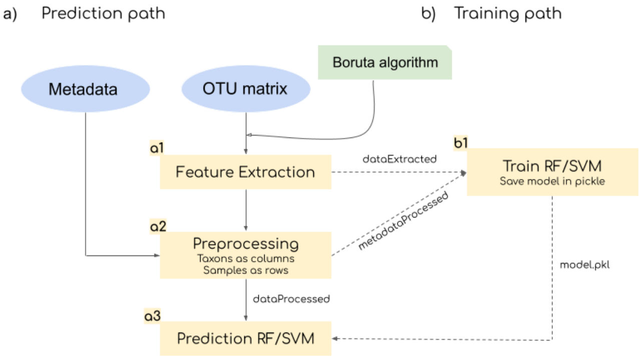

Figure 8: Python script workflow. *Workflow of Machine Learning Classifier in Python, splitted in two main paths: (a) having as input a taxonomic matrix, (a1) extracting rows containing the important features selected by Boruta algorithm. (a2) A preprocessing step taking as input the matrix with the extracted features and the metadata. (a3) Finally making the prediction step with Random Forest (RF) and Support Vector Machine (SVM) taking as input the taxonomic matrix preprocessed and (b1) the model in pickle format generated for RF and SVM. If we wanted to classify one sample, we would skip (b) path and perform only (a) path without inputting metadata as we will already have our model saved in Pickle format.*

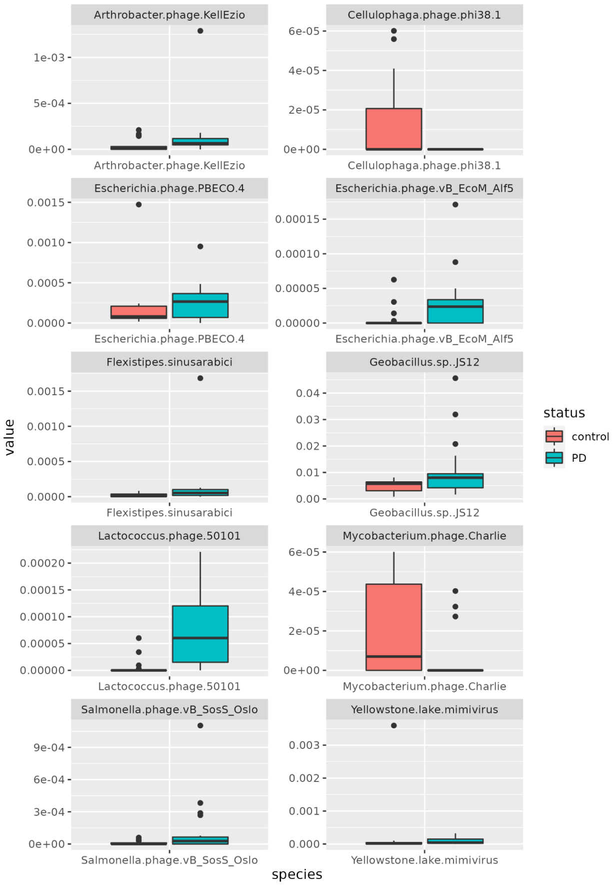

Figure 6: Boxplot diagram of the relative abundance of species selected by Boruta algorithm. *Ten species were selected as important by Boruta when doing the feature selection step, showing phages as the organisms more abundant.*

4. Discussion

The human gastrointestinal tract (GIT) harbors a complex and dynamic population of microorganisms, the gut microbiota, which exert a marked influence on the host during homeostasis and disease. The effects of species and pathways involved may promote intestinal bacterial overgrowth and disturb the gut homeostasis which is detrimental to the host. Dysbiosis in the gut microbiome may cause the systemic and/or central nervous system inflammation.[28] In addition, bacterial proteins could cross-react with human antigens and induce an adaptive immune response. Gut microbes may send signals to the brain via the vagus nerve by the direct stimulation of afferent neurons of the enteric nervous system.[29] Gut microbiota could promote α-synuclein aggregation and induce misfolding in both the central nervous system and the enteric nervous system in PD.[30]

In the present study we have compared the composition of the whole fecal microbiome between PD patients and control subjects. We have analyzed the metagenomic data generated in the study submitted by the National Taiwan University, which included 20 patients with Parkinson and 20 healthy controls. Our aim was to compare species, pathways and gene families for both conditions to reveal possible biomarkers that could initiate the progression of PD.

4.1 Parkinson's disease and Gut Microbiota

Taxonomic analysis composition of gut microbiota from patients with PD and control groups showed differences at phylum and species level. Firstly, we found that members of the genus *Bacteroides* were the ones who account for a major fraction in our samples. Congruent with previous studies the family *Bifidobacteriaceae* was reported with a higher abundance in PD.[31] On the contrary, a study reported a significantly reduced abundance of *Prevotellaceae* of 77.6% in patients with PD in comparison with controls, but in our case *Prevotella* with species *P. intermedia* and *P. enoeca* showed minor levels in controls.[32] What is more, a high difference in abundance of *Scherichia coli* was found to be low in PD patients.

The main findings of our study were the comparative evaluation of the microbiome composition that was made with artificial intelligence techniques, which reveals major differences in abundances regarding bacteriophages, also agreeing with the differential abundance analysis done. One of the biomarkers identified, *Lactococcus* phage 50101, in the current literature it is also supported by a study carried out by (GeorgeTetz, et al. 2018)[12], in which they described significant alterations in the representation of certain bacteriophages in the phagobiota of PD patients. These bacteria are considered as an important source of microbiota-derived neurochemicals, including dopamine which they produce in appreciable physiological amounts.[33] This loss of dopamine-producing *Lactococcus* may be, on the one hand, associated with early gastrointestinal symptoms of PD and, on the other, involved in triggering the neurodegenerative cascade of the disease.[34] In addition to be *Lactococcus* phage one of the possible biomarkers for the disease, *Escherichia* phage PBECO.4 and *Escherichia* phage vB_EcoM_Alf5 was found to be increased in PD patients; it may be related to the the decrease levels of *Escherichia coli* found in PD, however Boruta algorithm did not select *Escherichia Coli* as one of the most important features. In addition, GAIA analysis seems to match our results, being *Lactococcus* an important variable increased in healthy patients. Moreover a new variable was found significant in PD patients, which was *Aeromonas veronii;* considered to be pathogenic in humans, causing both gastrointestinal and extraintestinal infectious diseases.[35] Despite other species found to be significant in one condition or the other, results did not show any information regarding phages.

4.2 Myo-chiro and scyllo-inositol degradation

β-amyloid(Aβ) and α-synuclein(α-syn) are aggregation-prone proteins typically associated with two distinct neurodegenerative disorders: Alzheimer's disease (AD) and Parkinson's disease (PD). Basic research has begun to show that Aβ and α-syn may act synergically to promote the accumulation of each other.[36] While the exact mechanism by which these proteins interact remain unclear, growing evidence suggests that Aβ may drive α-syn by impairing protein clearance, activating inflammation, enhancing phosphorylation or directly promoting aggregation.[37] Myo-chiro and scyllo-Inositol degradation was one of the important pathways selected by our algorithm, shown to be increased in PD patients. Inositol (1,2,3,4,5,6-cyclohexanehexol) has nine possible stereoisomers which *myo*-Inositol is more abundant in nature than *scyllo*-Inositol. Accumulating evidence suggests *scyllo*-Inositol as a promising therapeutic agent for Alzheimer's disease, as it prevents the accumulation of beta-amyloid deposits, which is a hallmark of AD and PD. *Scyllo*-Inositol interacts with the beta-amyloid peptide and blocks the development of fivers, alleviating memory deficits and other symptoms associated with beta-amyloid accumulation.[38] Therefore, the hypothesis regarding that the starting point of PD could be in the gut, may be true as the degradation of the *scyllo*-Inositol could provoke the accumulation of beta-amyloid deposits, perhaps originated by dysbiosis in the gut microbiome.

4.3 RUMP formaldehyde oxidation I

Susceptibility for PD is modulated by various environmental factors, genetic predisposition or risk factors. Exposure to pesticides and industrial agents has been associated with an increased risk for PD, but to date none of these agents have been consistently identified as a causal factor for PD.[39] RUMP formaldehyde oxidation I was increased in PD patients, maybe suggesting a high abundance of the toxin formaldehyde by the environmental exposure to this pollutant, which an excess of this compound could induce alterations in brain metabolism and oxidative stress may contribute to the pathological progression of neurodegenerative disorders.[5,40,41]

5. Conclusions

Our findings shed a new light on previous reports regarding the gut microbiome involvement in PD, as a result of using artificial intelligence techniques in our analysis. Investigating whether the abundances of microbes and pathways implicated may be involved in the onset of the disease due to dysbiosis in the gut microbiome. Based on our analysis, low fecal abundance of *Lactococcus* phage and *Escherichia coli* phage could be a useful biomarker to exclude PD. Moreover, some important pathways could bring us some light about the starting point of the disease, which due to dysbiosis or external factors could overactivate the already mentioned pathways. The addition of more samples to the analysis may increase accuracy, and further exploring the potential of fecal microbiome analysis as a biomarker for PD seems worthwhile.

6. Supplementary material

The supplementary material generated in this project is available at the following link: https://drive.google.com/file/d/1MWIveieLBWodms3SULyKPtJ7PCEi0Zou/view?usp=sharing

7. Acknowledgements

I would like to express my thanks to my supervisor Andreu Paytuví for the guidance throughout all the steps of this project and for giving me the knowledge, ability and opportunity to undertake this study and complete it satisfactorily. In addition, I would like to extend my thanks to all the members of Sequentia Biotech for giving me the opportunity to work and learn with all of them.

8. References

1. Liu C-C, Li C-Y, Lee P-C, Sun Y. Variations in Incidence and Prevalence of Parkinson's Disease in Taiwan: A Population-Based Nationwide Study. Park Dis 2016; 2016: 1–8.
2. Löhle M, Storch A, Reichmann H. Beyond tremor and rigidity: non-motor features of Parkinson's disease. J Neural Transm 2009; 116: 1483–92.
3. Lai BCL, Marion SA, Teschke K, Tsui JKC. Occupational and environmental risk factors for Parkinson's disease. Parkinsonism Relat Disord 2002; 8: 297–309.
4. Li F, Wang P, Chen Z, Sui X, Xie X, Zhang J. Alteration of the fecal microbiota in North-Eastern Han Chinese population with sporadic Parkinson's disease. Neurosci Lett 2019; 707: 134297.
5. Caudle WM, Guillot TS, Lazo CR, Miller GW. Industrial toxicants and Parkinson's disease. NeuroToxicology 2012; 33: 178–88.
6. Baizabal-Carvallo JF, Alonso-Juarez M. The Link

between Gut Dysbiosis and Neuroinflammation in Parkinson's Disease. Neuroscience 2020; 432: 160–73.

7. Stefanis L. -Synuclein in Parkinson's Disease. Cold Spring Harb Perspect Med 2012; 2: a009399–a009399.
8. Hirsch EC, Hunot S. Neuroinflammation in Parkinson's disease: a target for neuroprotection? Lancet Neurol 2009; 8: 382–97.
9. Schippa S, Conte M. Dysbiotic Events in Gut Microbiota: Impact on Human Health. Nutrients 2014; 6: 5786–805.
10. Olanow CW, Brundin P. Parkinson's Disease and Alpha Synuclein: Is Parkinson's Disease a Prion-Like Disorder?: PD, ALPHA SYNUCLEIN, AND PRION DISORDERS. Mov Disord 2013; 28: 31–40.
11. Scheperjans F. Can microbiota research change our understanding of neurodegenerative diseases? Neurodegener Dis Manag 2016; 6: 81–5.
12. Tetz G, Brown SM, Hao Y, Tetz V. Parkinson's disease and bacteriophages as its overlooked contributors. Sci Rep 2018; 8: 10812.
13. Ursell LK, Metcalf JL, Parfrey LW, Knight R. Defining the human microbiome. Nutr Rev 2012; 70: S38–44.
14. Dalmasso M, Hill C, Ross RP. Exploiting gut bacteriophages for human health. Trends Microbiol 2014; 22: 399–405.
15. Hasegawa S, Goto S, Tsuji H, Okuno T, Asahara T, Nomoto K, Shibata A, Fujisawa Y, Minato T, Okamoto A, Ohno K, Hirayama M. Intestinal Dysbiosis and Lowered Serum Lipopolysaccharide-Binding Protein in Parkinson's Disease. PLOS ONE 2015; 10: e0142164.
16. Unger MM, Spiegel J, Dillmann K-U, Grundmann D, Philippeit H, Bürmann J, Faßbender K, Schwiertz A, Schäfer K-H. Short chain fatty acids and gut microbiota differ between patients with Parkinson's disease and age-matched controls. Parkinsonism Relat Disord 2016; 32: 66–72.
17. Li W, Wu X, Hu X, Wang T, Liang S, Duan Y, Jin F, Qin B. Structural changes of gut microbiota in Parkinson's disease and its correlation with clinical features. Sci China Life Sci 2017; 60: 1223–33.
18. Kechin A, Boyarskikh U, Kel A, Filipenko M. cutPrimers: A New Tool for Accurate Cutting of Primers from Reads of Targeted Next Generation Sequencing. J Comput Biol 2017; 24: 1138–43.
19. Kim D, Song L, Breitwieser FP, Salzberg SL. Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Res 2016; 26: 1721–9.
20. Paytuví A, Battista E, Scippacercola F, Aiese Cigliano R, Sanseverino W. GAIA: an integrated metagenomics suite. Bioinformatics, 2019 [cited 2022 May 24]. . doi: 10.1101/804690
21. Beghini F, McIver LJ, Blanco-Míguez A, Dubois L, Asnicar F, Maharjan S, Mailyan A, Manghi P, Scholz M, Thomas AM, Valles-Colomer M, Weingart G, Zhang Y, Zolfo M, Huttenhower C, Franzosa EA, Segata N. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. eLife 2021; 10: e65088.
22. Brownlee J. LOOCV for Evaluating Machine Learning Algorithms [Internet]. Mach. Learn. Mastery 2020 Jul 26 [cited 2022 May 24]. Available from: https://machinelearningmastery.com/loocv-for-evaluating-machine-learning-algorithms/
23. ZACH. A Quick Intro to Leave-One-Out Cross-Validation (LOOCV) [Internet]. Statology 2020 Nov 3 [cited 2022 May 24]. Available from: https://www.statology.org/leave-one-out-cross-validation/
24. Wong T-T. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. Pattern Recognit 2015; 48: 2839–46.
25. Li J, Cheng K, Wang S, Morstatter F, Trevino RP, Tang J, Liu H. Feature Selection: A Data Perspective. ACM Comput Surv 2018; 50: 1–45.
26. Kursa MB, Rudnicki WR. Feature Selection with the **Boruta** Package. J Stat Softw 2010; 36. doi: 10.18637/jss.v036.i11
27. Goodrich JK, Di Rienzi SC, Poole AC, Koren O, Walters WA, Caporaso JG, Knight R, Ley RE. Conducting a Microbiome Study. Cell 2014; 158: 250–62.
28. Carding S, Verbeke K, Vipond DT, Corfe BM, Owen LJ. Dysbiosis of the gut microbiota in disease. Microb Ecol Health Dis 2015 Feb 2; 26. doi: 10.3402/mehd.v26.26191
29. Galland L. The Gut Microbiome and the Brain. J Med Food 2014; 17: 1261–72.
30. Chandra R, Hiniker A, Kuo Y-M, Nussbaum RL, Liddle RA. α-Synuclein in gut endocrine cells and its implications for Parkinson's disease. JCI Insight 2017; 2: e92295.
31. Petrov VA, Saltykova IV, Zhukova IA, Alifirova VM, Zhukova NG, Dorofeeva YuB, Tyakht AV, Kovarsky BA, Alekseev DG, Kostryukova ES, Mironova YuS, Izhboldina OP, Nikitina MA, Perevozchikova TV, Fait EA, Babenko VV, Vakhitova MT, Govorun VM, Sazonov AE. Analysis of Gut Microbiota in Patients with Parkinson's Disease. Bull Exp Biol Med 2017; 162: 734–7.
32. Scheperjans F, Aho V, Pereira PAB, Koskinen K, Paulin L, Pekkonen E, Haapaniemi E, Kaakkola S, Eerola‐Rautio J, Pohja M, Kinnunen E, Murros K, Auvinen P. Gut microbiota are related to Parkinson's

disease and clinical phenotype. Mov Disord 2015; 30: 350–8.

33. Kuley E, Özogul F. Synergistic and antagonistic effect of lactic acid bacteria on tyramine production by food-borne pathogenic bacteria in tyrosine decarboxylase broth. Food Chem 2011; 127: 1163–8.
34. Houser MC, Tansey MG. The gut-brain axis: is intestinal inflammation a silent driver of Parkinson's disease pathogenesis? Npj Park Dis 2017; 3: 3.
35. Vila J, Marco F, Soler L, Chacon M, Figueras MJ. In vitro antimicrobial susceptibility of clinical isolates of Aeromonas caviae, Aeromonas hydrophila and Aeromonas veronii biotype sobria. J Antimicrob Chemother 2002; 49: 701–2.
36. Finder VH, Glockshuber R. Amyloid-beta aggregation. Neurodegener Dis 2007; 4: 13–27.
37. Marsh SE, Blurton-Jones M. Examining the mechanisms that link β-amyloid and α-synuclein pathologies. Alzheimers Res Ther 2012; 4: 11.
38. Yamaoka M, Osawa S, Morinaga T, Takenaka S, Yoshida K. A cell factory of Bacillus subtilis engineered for the simple bioconversion of myo-inositol to scyllo-inositol, a potential therapeutic agent for Alzheimer's disease. Microb Cell Factories 2011; 10: 69.
39. Fujita KA, Ostaszewski M, Matsuoka Y, Ghosh S, Glaab E, Trefois C, Crespo I, Perumal TM, Jurkowski W, Antony PMA, Diederich N, Buttini M, Kodama A, Satagopam VP, Eifes S, del Sol A, Schneider R, Kitano H, Balling R. Integrating Pathways of Parkinson's Disease in a Molecular Interaction Map. Mol Neurobiol 2014; 49: 88–102.
40. Tulpule K, Dringen R. Formaldehyde in brain: an overlooked player in neurodegeneration? J Neurochem 2013; 127: 7–21.
41. Rana I, Rieswijk L, Steinmaus C, Zhang L. Formaldehyde and Brain Disorders: A Meta-Analysis and Bioinformatics Approach. Neurotox Res 2021; 39: 924–48.