

Xènia Roda Sánchez

Gut Microbiome Age as a Health Indicator: A Data-Driven Approach to Microbial Aging

MASTER'S THESIS

Supervisors: Dr. Andreu Paytuví-Gallart and Dra. Petia Ivanova Radeva

Master's Degree in Health Data Science



Barcelona, 2025

I would like to thank my supervisor at Sequentia, Dr. Andreu Paytuví-Gallart, for their guidance and support during this master thesis. I also appreciate the collaboration and assistance of my colleagues at Sequentia throughout this process.

My thanks to my academic advisor, Petia Ivanova Radeva, for their valuable advice and feedback.

Finally, I thank my parents for their support.

Abstract

The human gut microbiome plays a critical role in maintaining overall health, influencing digestion, immunity, and even mental health. While its composition has been extensively studied, robust methods to quantify its ‘biological age’—a potential biomarker for microbiome health—remain underexplored. This study bridges this gap by developing an interpretable machine learning framework to predict chronological age from gut microbiome composition in healthy individuals. We analyzed species-level abundance profiles from a multi-cohort dataset, employing four feature selection strategies to identify aging-related signatures. Among these, permutation feature importance (PFI) emerged as the most robust method, enhancing the model’s performance by identifying key contributors to the data. The optimal model (CatBoost regressor with PFI) achieved a mean absolute error of 9.17 years on held-out test data, demonstrating generalizability across diverse populations. Key findings revealed *Coprococcus* as a strong age predictor, with increasing abundance across adulthood suggesting microbiome resilience rather than dysbiosis in this age range. Additionally, the model’s interpretability, facilitated by SHAP values, highlights species-specific contributions to age prediction, allowing for a deeper understanding of microbial roles in aging. The findings also reveal age-related shifts in microbiota composition, with older individuals showing higher abundance of *Lachnospiraceae* and butyrate-producing bacteria, suggesting a more efficient production of beneficial short-chain fatty acids. This research lays the foundation for future microbiome-based diagnostic tools and targeted interventions to promote healthy aging.

Table of Contents

Abstract.....	3
1. Motivation and Background.....	8
1.1. Motivation.....	8
1.2. Background.....	8
1.2.1. The microbiome and its importance in health.....	8
The human gut microbiome and aging.....	9
Challenges in understanding the microbiome-aging link.....	11
1.2.2. Metagenomic and metatranscriptomics data analysis.....	11
Preprocessing Steps.....	12
Computing species abundance and diversity.....	12
1.2.3. Machine learning.....	13
1.2.4. Feature Importance.....	14
2. Objectives.....	16
3. State of the art.....	17
3.1. Machine learning with gut microbiome data.....	17
3.2. Machine learning for age prediction: Regression Models.....	17
3.3. Machine learning for age prediction: Classification Models.....	19
3.4. Comparative analysis: Regression vs. classification for gut.....	20
3.5. Key gut microbiome features predictive of age.....	20
4. Design and development.....	22
4.1. Dataset.....	22
4.2. Pre-Processing.....	22
4.2.1. Sample Quality Control and Trimming.....	22
4.2.2. Normalization and Filtering.....	22
4.3. Exploration Data Analysis (EDA).....	23
4.3.1. Rarefaction Curve.....	23
4.3.2. Alpha Diversity.....	23
4.3.3. Beta diversity analysis.....	24
4.4. Overview of the Machine Learning Pipeline.....	25
4.5. Machine Learning Methods.....	27
4.5.1. Regression vs. classification for age prediction.....	27
4.5.2. Model Evaluation: K-Fold Cross-Validation.....	27
4.5.3. Performance Metrics.....	28
4.6. Feature Selection.....	29
4.6.1. Method 1. Comparative Analysis of SHAP-values and Importance-Based Methods....	29
4.6.1.1. Combined SHAP and Built-in Ranks.....	31
4.6.2. Method 2. SelectKBest.....	33
4.6.3. Method 3. Boruta-SHAP.....	33

4.6.4. Method 4. Permutation Feature Importance (PFI).....	33
4.7. Hyperparameter Tunning.....	34
4.8. Model Evaluation.....	34
4.8.1. Final Model.....	35
4.9. Sample Clustering with UMAP.....	35
5. Experiments and results.....	36
5.1. Exploratory data analysis (EDA) results.....	36
5.2. Machine learning experiments.....	41
5.2.1. Baseline model performance.....	41
5.2.2. Feature selection analysis.....	42
Method 1: Comparative analysis of SHAP values and Important base method results.....	42
Method 2: SelectKBest results.....	44
Method 3: Boruta-SHAP results.....	46
Method 4: Permutation Feature Importance results.....	48
5.2.3. Summary of SHAP-based feature selection Insights.....	50
5.2.4. Interpretation of final model SHAP results.....	51
5.2.5. Hyperparameter optimization.....	54
5.3. Final model selection and performance.....	54
5.4. Identification of key features driving cluster separation.....	56
5.4.1. SHAP analysis for feature importance.....	56
5.4.2. Discriminative feature analysis via cluster centroids.....	56
5.4.3. UMAP visualization of top features.....	59
5.5. Age-related abundance patterns of key discriminative species.....	60
6. Discussion and conclusions.....	63
6.1. Discussion.....	63
6.1.1. Age-related differences in gut microbiota and the impact of diet and lifestyle.....	63
6.1.2. Regression analysis of gut microbiota as a predictor of chronological age.....	64
6.1.3. Biological interpretation of predictive features.....	65
Discussion on Coprococcus as a predictive feature in age-associated gut microbiota shifts.	65
Developing dynamics and core prevalence of Coprococcus.....	65
Age-associated abundance patterns in our model.....	66
6.1.4. Umap analysis interpretation.....	67
6.1.5. Limitations and Assumptions.....	68
6.2. Ethical-social impact, sustainability, and diversity.....	69
6.2.1. Ethical-social impact.....	69
6.2.2. Sustainability.....	69
6.2.3. Diversity.....	69
6.3. Conclusions.....	70
7. Future work.....	71
References.....	72

1. Motivation and Background

1.1. Motivation

The human gut microbiome, a complex ecosystem of microorganisms, plays a critical role in maintaining overall health. Recent studies have highlighted its influence on digestion, immunity, and even mental health. (H.-J. Wu, 2012) Understanding the dynamics of the gut microbiome can provide valuable insights into health and disease, for rationally manipulating them for therapeutic ends. (Schlomann & Parthasarathy, 2019).

While the gut microbiome's composition has been extensively studied, there is a lack of robust methods to quantify its 'biological age'—a metric that could serve as a factor for overall microbiome health. Current approaches often fail to account for the complexity of metagenomic data or lack interpretability, limiting their practical application in clinical settings.

This thesis aims to bridge this gap by developing a comprehensive framework for predicting microbiome age using advanced machine learning techniques. By collecting and preprocessing metagenomic data from a large, healthy population (O1), identifying relevant features (O2), and developing an optimized predictive model (O3), this research seeks to provide a reliable tool for assessing microbiome health. Furthermore, the model's interpretability (O6) will enable researchers and clinicians to identify key microbial species influencing microbiome age, paving the way for targeted interventions.

1.2. Background

1.2.1. The microbiome and its importance in health

The microbiome refers to the bacterial flora, the set of microorganisms present in our body. These microorganisms are mostly bacteria, archaea, viruses and eukaryotic microbes that reside in and on our bodies all of which live in symbiosis with us. The human intestine carries a vast and diverse microbial ecosystem, more than 500 different bacterial species, some of which have important health functions (Bengmark, 1998). In our microbiome, we can find both beneficial and harmful bacteria. Maintaining a healthy balance among these microorganisms is crucial for supporting bodily functions and overall well-being. For example, among the beneficial bacteria, *Bifidobacterium* and *Lactobacillus* play essential roles in food fermentation and the production of short-chain fatty acids, which are beneficial for intestinal health, while *Lactobacillus* also enhances the intestinal barrier by promoting mucus production, barrier-related proteins, and secreting antimicrobial compounds (Arboleya et al., 2016; Dempsey & Corr, 2022; O'Callaghan & van Sinderen, 2016).

On the other hand, harmful bacteria, when present in excess, can lead to health issues. For instance, *Escherichia coli* (*E. coli*) is generally beneficial in normal quantities, but certain strains can cause

severe gastrointestinal illnesses (Richter et al., 2018). Importantly, the majority of bacteria in our gut are neutral—neither inherently beneficial nor harmful. Additionally, the effects of these bacteria vary from person to person, depending on individual symptoms and the overall balance of their microbiome.

Significant interest has evolved on the gut microbiota in recent years as it has been associated with a large array of human diseases ranging from luminal diseases such as inflammatory bowel diseases (IBD) (Ferreira et al., 2014), metabolic diseases such as obesity, diabetes and atherosclerosis (Karlsson et al., 2013), allergic disease (Bisgaard et al., 2011) to developmental disorders (Mallick et al., 2024). Understanding of the relationship between gut bacteria and human health can be helpful for targeting new probiotic treatments and novel strategies in treating and managing a wide variety of human diseases.

From an immunological perspective, the gut microbiome is intrinsically important to the function of the host's innate immunity by influencing the development and function of antigen-presenting cells (APCs), neutrophils, and other innate immune cells (Choden, 2022). As the microbiome evolves throughout life—shaped by factors such as diet, growth, and environmental exposures—it plays a crucial role in the maturation of the immune system, gradually introducing a diverse array of bacteria that interact with and shape immune responses (P. Zhang, 2022).

The human gut microbiome and aging

The composition of the human gut microbiome has been observed to change over the course of an individual's life (Bosco & Noti, 2021). The first microbial colonization occurs at birth, when newborns are 'seeded' with their mothers' microbiota (Bäckhed et al., 2015). Early influences, such as the method of delivery and type of feeding (breast milk or formula), shape the initial composition and development of gut microbes (Zhuang et al., 2019).

The early-life development of the microbiome is of great interest due to its potential impact on the onset of diseases later in life. (Tamburini et al., 2016). Differences in the gut microbiome taxa have been noted in infants and young children who develop life-long diseases that include eczema, (Abrahamsson et al., 2012) asthma, (Arrieta et al., 2015) and type 1 diabetes (Knip & Siljander, 2016). Many childhood diseases are linked to immune system dysfunction, and the microbiome is believed to contribute to early immune development, potentially reducing the risk of autoimmune and inflammatory bowel diseases later in life. (Tamburini et al., 2016).

The adult microbiome appears to be highly personalized and stable and does not drastically change with short-term variation in diet (A. J. Johnson et al., 2019). In general, among adults, a diverse gut microbiome that contains a large number of species (high diversity) as well as high abundance of numerous species (high evenness) is healthy and robust (Lloyd-Price et al., 2016; Sommer et al., 2017). Microbiome diversity may indicate more than just good health, as studies have shown that elite athletes tend to have a more diverse gut microbiome (Clarke et al., 2014).

Although the established adult microbiome is observed to be fairly stable, external factors, likely primary diet, (Aya et al., 2023) affect the composition of the adult microbiome over the longer term. Several recent studies make this fact evident. For example, research on adult immigrants from Asia to

the USA compared the microbiome profiles of immigrants and their children to individuals from the same ethnic group in their country of origin (Vangay et al., 2018). This study found that the gut microbiome shifted in correlation with the length of time spent in the USA, showing a decline in overall diversity, a reduction in fiber-metabolizing *Prevotella* strains, and an increase in *Bacteroidetes*. While diet played a significant role in this shift, it was not the sole factor influencing these changes. Similarly, other studies highlight the strong influence of both individual nutrients and dietary patterns on gut microbiota. For instance, adherence to a Mediterranean diet has been associated with greater microbiota diversity and stability (Rinninella et al., 2019), further emphasizing the role of diet in shaping microbial communities.

During adulthood, specific taxa and overall diversity are associated with markers of good health. The genus *Christensenellaceae* has been correlated with good metabolic health including lower body mass index (Goodrich et al., 2014) and reduced adipose tissue. Several *Prevotella* species reside in the human gut, with *Prevotella copri* which is often linked with desirable health measures being the most common, found in approximately 40% of the global population. This species is more prevalent in non-Western populations, likely due to its association with high-fiber, low-fast diets. Additionally, *P. copri* is often linked to beneficial health outcomes, including lower visceral fat and improved glucose metabolism (Yeoh et al., 2022).

The older adult microbiome in healthy aging tend to have lower overall taxonomic diversity within their microbiomes, particularly in beneficial species such as *Bifidobacterium*, *Akkermansia*, and *Clostridium* (Almeida et al., 2022; Mancabelli et al., 2024) and have higher abundance of *Bacteroidetes* (Claesson et al., 2011; Haran & McCormick, 2021; Jeffery et al., 2016). These changes may affect immune function, nutrient absorption, and susceptibility to age-related conditions, such as frailty and chronic diseases (Almeida et al., 2022; Ragonnaud & Biragyn, 2021). In adults aged 90 and older, the composition of the gut microbiome appears to be strongly influenced by an individual's health status, with healthier, long-lived individuals exhibiting distinct microbiome profiles (Wilmanski et al., 2021). Research suggests that certain microbial taxa are linked to healthy aging, and their loss may serve as "taxonomic milestones" that either contribute to or reflect physiological decline and subsequent age-related disease (Ghosh et al., 2022a).

Age is not the only factor that drives to change our microbiome. Research is also advancing our understanding of the environment as a major driver of variability in the human microbiome. Metagenomic sequencing of microbiome samples from Europe, Africa, North and South America has revealed distinct microbial patterns specific to each geographic region throughout the lifespan (Mancabelli et al., 2024). While many studies report age-related shifts in microbial taxa that are consistent across different parts of the world, notable regional differences exist. For instance, research in Italy (Biagi et al., 2016) and Ireland (Claesson et al., 2012) has shown a decline in *Roseburia* with aging, whereas studies in Korea (Park et al., 2015) and China (X. Zhang et al., 2021) have reported an increase in this genus. Similarly, *Bacteroidetes* are generally thought to increase with age, yet this trend was reversed in healthy older individuals from Indonesia (Rahayu et al., 2019). A comparison of centenarians from India cohort, including lower *Bacteroidetes* and higher *Enterobacteriaceae* (Tuikhar et al., 2019). Additionally, while *Akkermansia* is often linked to healthy aging, it was instead associated with frailty in a group of older Chinese adults (Y. Xu et al., 2021). These findings highlight the complex interplay between geography, aging and microbiome composition.

Challenges in understanding the microbiome-aging link

As people age and develop age-related diseases, the variety of bacteria in their guts tends to decrease. This means there are fewer different types of microbes, and some beneficial ones may become less common (Haran et al., 2018; Jackson et al., 2016); but it is important to note that associations with certain taxa are not always consistent across studies. *Corprococcus* has generally been associated with healthy aging (Ghosh et al., 2022b; N. Wang et al., 2019) but also was observed to be decreased in a healthy centenarian cohort. (L. Wu et al., 2019). *Ruminococcus* is a genus that is known to be anti-inflammatory (Nagao-Kitamoto & Kamada, 2017) which was seen to be increased in centenarians, (Tuikhar et al., 2019), but also being associated with frailty (Haran et al., 2018; Verdi et al., 2018).

Since much of the research on aging and the gut microbiome relies on observational data, it is difficult to determine whether certain microbial taxa directly contribute to positive health outcomes or if other hidden factors are influencing these associations (Seo et al., 2023). Another relevant challenge is the assessment of relative abundance rather than absolute abundance, which could make difference the interpretations of compositional changes. Studies relying solely on relative abundance do not always accurately reflect changes in the total number of microorganisms, as the absolute cell count could be an important factor. Evaluating absolute abundance, such as the total number of cells or microbial biomass, could provide a more accurate picture of how changes in the microbiota impact health, especially in the context of aging (Bruijning et al., 2023). Despite these challenges, computational approaches such as machine learning are emerging as powerful tools to uncover hidden patterns in the gut microbiome. Unlike traditional studies, machine learning models can analyze large datasets, detect complex microbial signatures associated with aging, and improve our ability to predict biological age based on microbiome composition.

1.2.2. Metagenomic and metatranscriptomics data analysis

Metagenomics is the suite of genomics technologies and bioinformatics tools to directly access the genomic information of entire communities of organisms (Thomas et al., 2012).

This approach helps to gain ability to study complex microbial ecosystems, such as the human gut microbiome, by providing insights into their composition, function, and interactions with the host.

Metagenomic sequencing can be performed using two primary approaches: shotgun sequencing and 16S rRNA gene sequencing. Shotgun sequencing involves randomly fragmenting and sequencing DNA from all organisms within a sample, providing a comprehensive, high-resolution view of the entire microbial community's taxonomic composition and functional potential, often enabling species-level identification. In contrast, 16S rRNA sequencing specifically targets the gene encoding the small ribosomal subunit RNA. The taxonomic resolution achieved with this method depends significantly on the sequencing strategy; targeting specific hypervariable regions, as is common with short-read platforms like Illumina, typically allows for classification down to the genus level. However, sequencing the near full-length gene (~1500bp), achievable with long-read technologies such as Nanopore, can provide resolution down to the species level. While the enhanced resolution from full-length 16S sequencing is valuable, whole-genome shotgun (WGS) sequencing offers more

detailed taxonomic and functional information, making it a preferred method for in-depth studies of complex microbial ecosystems like the gut microbiome (Durazzi et al., 2021).

On the other hand, metatranscriptomics refers to the study of the collective transcriptomes of a given habitat (Ojala et al., 2023). It provides direct insights into the actively expressed genes within a microbial community. By focusing on RNA transcripts, researchers can identify genes that are actively involved in metabolic pathways, cellular processes, and environmental responses.

Metagenomics and metatranscriptomics are powerful tools for studying microbial communities; metagenomics produces a taxonomical profile of the sample, while metatranscriptomics help us to obtain a functional profile (Aguiar-Pulido et al., 2016).

Preprocessing Steps

Raw metagenomic data requires extensive preprocessing to ensure that it has good quality for downstream analysis (Wright et al., 2017). Key preprocessing steps include:

- **Quality Control:** Filtering out low-quality reads, adapter sequences, and host DNA contamination.
- **Taxonomic Classification:** Assigning sequences to their taxonomic group by using reference databases, such as NCBI in order to identify microbial species present in the sample.
- **Normalization:** Adjusting for variations in sequencing depth and sample size in order to perform sample comparisons. Techniques such as refraction or scaling to relative abundance are commonly used.

Computing species abundance and diversity

Once the data is preprocessed, species abundance and diversity can be computed to describe the microbial community. Species abundance refers to the relative proportion of each species within the sample, often calculated using counts of sequence reads mapped to specific taxa and it is a way to compare the prevalence of different species across samples. However, because metagenomic data is compositional (i.e., the abundances are relative and sum to a constant), additional transformations like the Centered Log Ratio (CLR) are often applied to make the data more suitable in the analysis. The CLR transformation normalizes the data by taking the algorithm of the ratio of each species' abundance to the geometric mean of all species abundance in the sample. For a composition vector x with positive components $[x_1, x_2, \dots, x_n]$. This approach preserves the relationships between species while addressing the challenges of compositionality (Zhou et al., 2023).

$$CLR(x_i) = \left[\log(x_1 / \sqrt{(x_1 x_2 \dots x_n)}), \log(x_2 / \sqrt{(x_1 x_2 \dots x_n)}), \dots, \log(x_n / \sqrt{(x_1 x_2 \dots x_n)}) \right]$$

Microbiome data are sparse and high dimensional, so effective visualization of these data requires dimensionality reduction. To visualize and interpret metagenomic data, the most commonly used method for dimensionality reduction in the microbiome is the calculation of between-sample microbial differences (beta diversity), followed by techniques such as Principal Component Analysis (PCA) and Principal Coordinate Analysis (PCoA). PCA is a linear dimensionality reduction method that transforms the data into a set of orthogonal components, ordered by the amount of variance they explain (Meng et al., 2016). On the other hand, PCoA is a more flexible technique that can incorporate various distance metrics, such as Bray-Curtis or UniFrac distances, to capture both linear and non-linear relationships in the data. PCoA is very useful for visualizing beta diversity, as it allows us to observe how samples tend to cluster together based on their microbial composition (Peeters et al., 2021). Unifold Manifold Approximation and Projection (UMAP) is an alternative method that can reduce the dimensionality of beta diversity distance matrices, which can improve visualization quality and correspondence with biological and technical variables of interest (Armstrong et al., 2021).

Diversity metrics, such as alpha diversity (within-sample diversity), provide insight into the richness and evenness of the microbial community. These metrics are often complemented by the use of rarefaction curves, which plot the number of observed species against sequencing depth. Rarefaction curves are essential for assessing whether the sequencing effort has adequately captured the microbial diversity within samples, ensuring that comparisons between samples are not biased by differences in sequencing depth (Galià-Camps et al., 2023). Together, relative abundance, CLR-transformed data, and diversity metrics are important for understanding the structure and relationships of the gut microbiome (Walters & Martiny, 2020).

1.2.3. Machine learning

Machine learning (ML) has emerged as a powerful tool for analyzing complex and high-dimensional datasets, making it particularly well-suited for studying the gut microbiome (Hernández Medina et al., 2022). Traditional methods often struggle to capture the relationships and patterns within such data. Thus, machine learning provides a diverse range of algorithms that are capable of overcoming these limitations, by identifying patterns, classifying samples, and predicting outcomes based on input features. Broadly speaking, we can categorize several ML approaches:

- Supervised Machine Learning (SML) refers to methods that learn associations between input data $x \in X$ and corresponding target data $y \in Y$ by identifying a mapping function $f : X \rightarrow Y$. This function can be either parameterized, where $y = f(x)$. Based on the nature of the target data Y , supervised learning problems are typically divided into two main categories:
 - Classification Tasks: These involve predicting discrete class levels. Common algorithms used for classification include Support Vector Machines (SVM), Decision Trees (DT), and Random Forest (RF).
 - Regression Tasks: These involve predicting continuous numerical values. Examples of regression techniques include linear regression, logistic regression, and polynomial regression.
- Unsupervised Machine Learning (UML) refers to methods that identify patterns or relationships within data $x \in X$ without any explicitly target data $y \in Y$. Unlike supervised learning, there is no supervision or feedback during the learning process. Instead the model

aims to find an optimal representation of the data. Common unsupervised learning problems include:

- Clustering: Identifying natural groupings or structures in unlabeled data (e.g., K-means clustering).
- Association Rules: Discovering relationships or patterns within large datasets (e.g., the Apriori algorithm).
- Semi-supervised Learning (SSL) combined elements of both supervised and unsupervised learning. In this approach, some input data $x_1, \dots, x_n \in X$ have corresponding target labels $y_1, \dots, y_n \in Y$, while other data points remain unlabeled. The unlabeled data can help the model detect underlying patterns or structures, which can improve the performance of supervised learning tasks. For example, in classification problems, unsupervised techniques might identify important features or patterns in the data, which are then used to enhance the supervised training process.

There are numerous algorithms and methods for achieving the goal of learning in the machine learning field. In this study, I focus on regression methods as my research resolves a regression problem, which is predicting a continuous numerical value—specifically, the age of healthy patients—based on their gut microbiome data.

1.2.4. Feature Importance

Having established the methods for building and evaluating our predictive models, we now seek to understand how these models arrive at their predictions. To gain insight into the model's decision-making process, it's important to analyze model explainability. Model explainability involves identifying the features that most significantly influence the model's predictions (Holzinger, 2018; Roscher et al., 2020).

Shapley Additive exPlanations (SHAP) is a game-theoretic approach used to explain the output of any machine learning model (Lundberg & Lee, 2017). SHAP computes Shapley values, which quantify the contribution of each feature to a particular prediction.

The Shapley value for a feature is calculated by assessing how the model's output changes when that feature is included versus excluded from various feature subsets. More formally, the Shapley value represents the average marginal contribution of a feature across all possible combinations of other features. This average contribution is weighted by the number of features in each subset.

In essence, SHAP provides a unified measure of feature importance, allowing us to understand which microbiome taxa are most influential in driving the model's age prediction.

$$\varphi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|! (|N| - |S| - 1)!}{|N|!} (f_{S \cup \{i\}}(x) - f_S(x))$$

Where,

- N is the set of all features.
- S is any subset of features excluding i .
- $f_s(x)$ is the model prediction using the features in the subset S .
- $f_{S \cup \{i\}}(x)$ is the model prediction using the features in the subset S , and feature i .
- $|S|$ is the number of features in subset S .
- $|N|$ is the total number of features.
- $|S|!$ is the number of permutations of features in subset S .
- $(|N| - |S| - 1)!$ is the number of permutations of features not present in subset S , also excluding feature i .

2. Objectives

To summarize, this thesis focuses on the following objectives:

O1. Data Collection and Preprocessing

- Gather and preprocess metagenomic gut microbiome data from a large, healthy population.
- Process and normalize the data, including species composition and abundances.

O2. Feature Engineering and Selection

- Identify and extract relevant features from metagenomic data that may be indicative of microbiome age.
- Experiment with various feature selection techniques to optimize model performance and reduce computational complexity.

O3. Model Development

- Evaluate and compare different machine learning algorithms to determine which best predicts microbiome age.
- Fine-tune hyperparameters and optimize model architecture to improve prediction accuracy and robustness.

O4. Model Validation and Evaluation

- Validate the model using a test dataset and evaluate its performance using relevant metrics.

O5. Analysis of Microbiome Age Prediction

- Analyze the predictions to identify patterns or correlations.
- Assess how microbiome age aligns with biological age and explore implications for future studies on aging and microbiome health.

O6. Interpretability and Practical Application

- Develop techniques to make the model interpretable and identify the specific microbial species and metadata variables most influential in determining microbiome age.
- Discuss potential applications of the model in clinical or research settings as a tool for health assessment and monitoring of gut microbiome age.

3. State of the art

3.1. Machine learning with gut microbiome data

The human gut microbiome, a vast and intricate community of microorganisms residing within the gastrointestinal tract, has emerged as a critical factor influencing various aspects of host health. This complex ecosystem plays a pivotal role in metabolic process, immune system regulation, and even neurodevelopmental pathways.

The concept of “biological age”, which reflects an individual's functional health status rather than just the number of years lived, is increasingly being explored in the context of gut microbiome. Given chronological age is merely a measure of time, and biological aging can vary significantly among individuals of the same chronological age, the potential of the gut microbiome to reflect this biological age is intriguing. If the microbiome indeed served as a reliable indicator of biological aging, the machine learning models trained to predict age from microbiome data could offer a non-invasive tool for assessing an individual's health status and their aging trajectory. Moreover, studies have shown remarkably consistent microbiome features associated with healthy aging across different geographical regions. This further extends the potential of the gut microbiome as a robust marker of the aging process (Dasari et al., 2025; Kim et al., 2024; H. Wang et al., 2025).

3.2. Machine learning for age prediction: Regression Models

Most research has focused on employing regression-based machine learning algorithms to predict the age of healthy individuals as a continuous variable. These studies aim to develop “microbiome aging clocks” capable of estimating an individual's age based on the composition or functional information of their gut microbiome community (*Age Prediction Using Microbiota*, 2025; Dasari et al., 2024).

One important study utilized the eXtreme Gradient Boosting (XGBoost) algorithm to predict the age of 568 healthy Korean individuals aged 20-80 years based on their fecal gut microbiota. This regression model, treating age as a continuous variable, achieved a mean absolute error (ME) of 5.48 years. Furthermore, the study demonstrated that incorporating urine metabolite data alongside the gut microbiome features improved the prediction accuracy, reducing the MAE to 4.93 years (Seo et al., 2023).

Conversely, another study employing Random Forest regression on a combined dataset from multiple publicly available studies found that the skin microbiome was a more accurate predictor of age (MAE 3.8 years) compared to the gut microbiome (MAE 11.5 years) in adults. This suggests that while the gut microbiome contains age-related information, other microbial communities within the human body exhibit even stronger correlations with the aging process (Huang et al., 2020).

Gaussian Process (GP) regression models have also been explored for chronological age prediction using gut microbiome data. One study applied GP regression incorporating microbial community

dissimilarities to the American Gut Project (AGP) dataset and reported Root Mean Square Error (RMSE) values for various kernel functions, with a multiple-kernel approach achieving an RMSE of 11.02 years (Adachi et al., 2025).

An ensemble model, combining multiple heterogeneous machine learning algorithms and utilizing both species-level taxonomic profiles and metabolic pathways profiles of the gut microbiome, was developed to predict age in a population ranging from 18 to 107 years. This model achieved an R-squared value of 0.599 and a MAE of 8.33 years, which demonstrates the importance of integrating again different types of microbiome data in the prediction (Y. Chen et al., 2022).

Deep neural networks (DNNs) have also been investigated for this task. One study using a curated dataset of 1,165 healthy individuals developed a human microbiome clock based on a DNN architecture, achieving a cross-validation MAE of 3.94 years (Min et al., 2024). This indicates the capability of complex machine learning models to capture intricate relationships between the gut microbiome and age.

Furthermore, a “microbiome clock” was constructed using Random Forest regression in a large longitudinal dataset of gut microbiome profiles from wild baboons with known ages, which provides the feasibility of predicting chronological age based on gut microbiome in a non-human primate model (Dasari et al., 2024).

Table 1: Summary of Microbiome-Based Age Prediction Studies

Study	Algorithm	Age	Features	Metric	Value	Population
Healthy Koreans	XGBoost	Yes	Gut micro-biome	MAE	5.48 years	568 (20-80 yrs)
Multi-study	Random Forest	Yes	Gut micro-biome	MAE	11.5 years	US/UK/China (4,434)
AGP data	Gaussian Process	Yes	OTUs/ASVs	RMSE	~11-12 yrs	USA (1,639)
Ensemble	Ensemble	Yes	Species + path-ways	R^2 , MAE	0.6, 8.33 yrs	2,604 (18-107 yrs)
Human clock	DNN	Yes	Gut micro-biome	MAE	3.94 years	1,165 (3,663 samples)
Baboons	Random Forest	Yes	16S rRNA	MAE	6-11 years	479 (14 yrs)
XGBoost+meta	XGBoost	Yes	Gut micro-biome + urine	MAE	4.93 years	568 (20-80 yrs)

3.3. Machine learning for age prediction: Classification Models

While regression models aim to predict a continuous age variable, classification models in this context typically focus on categorizing individuals into different age groups.

One study employed logistic regression models to analyze the association between different gut microbiome clusters and metabolic health in pre-adolescents and older adults. Although the primary goal was not direct age prediction, the study utilized age as a key factor in defining the study groups and analyzing the relationship between microbiome profiles and health outcomes (Malesu, 2025).

In a different approach, the LightGBM algorithm was used for classification to distinguish between “superagers” (older adults with cognitive abilities comparable to younger individuals) and “typical agers” based on their gut microbiome features. This study, involving 102 participants aged 60 years or older, achieved a high Area Under the Curve (AUC) of 0.861 and an accuracy of 0.762 on the test dataset. This classification target was categorical (superager or typical ager), and the model successfully identified distinct microbiome signatures associated with aging (Kim et al., 2024).

Advanced classification techniques, like phylogeny-based convolutional neural networks (PM-CNN), have been explored in microbiome research, showing potential for identifying complex patterns in health states. One study used gut and oral microbiome data to classify individuals with various diseases, suggesting future applications in age-related classification (Q. Wang et al., 2024). (Ghosh et al., 2020) demonstrated that the performance of gut microbiome-based disease classifiers is age-dependent. Using Random Forest models evaluated by the Area Under the Curve (AUC), they found that classification accuracy for several diseases (including IBD, T2D, and CRC) was generally higher when models were trained and tested within the same age group compared to when tested across different age groups, highlighting distinct microbiome signatures in younger versus older individuals.

Table 2: Summary of Classification Models for Gut Microbiome-Based Age Prediction

Study	Algorithm	Age Groups	Features	Metric	Value	Population	Target
Microbiome and metabolic health	Logistic Regression	Pre-ado. (9.8y), Older (62.7y)	Gut clusters	OR	1.10–1.65	Dutch cohorts	Metabolic health
Superagers study	LightGBM	Superagers, Typical agers	Genera	AUC, Acc.	0.86, 0.76	SK (n=102)	Cognitive aging
Multi-status classification	PM-CNN	Healthy vs diseases	Gut/oral	Accuracy	Median AUC 0.79	Multi-cohort (>2500 individuals, ages 20-89) Included IBD, T2D, CRC, Cirrhosis, Polyp patients and controls	Disease status classification (e.g., IBD vs. Control, T2D vs. Control)

3.4. Comparative analysis: Regression vs. classification for gut

The choice between employing regression or classification models for predicting age depends on the objectives of the research. Regression models are suited in scenarios where the aim is to predict a continuous variable, such as an individual's specific age in years or a measure of their biological age in a continuous scale. This approach is relevant when the goal is to develop a “microbiome aging clock” that can estimate an individual's age as a quantitative measure, allowing to see the aging process as a continuous trajectory.

Conversely, classification models are more appropriate when the research question involves distinguishing between distinct categories or groups. In the context of aging, this might involve categorizing individuals into broad age groups (e.g., young, middle-aged, old) for comparative analyses or, as seen in the study of superagers, classifying individuals based on specific aging-related phenotypes.

3.5. Key gut microbiome features predictive of age

Research employing both regression and classification has identified several key bacterial taxa and functional pathways within the gut microbiome that appear to be predictive of age in healthy individuals. These features offer potential biomarkers for the aging process.

Studies using regression models have consistently observed that microbial diversity, often measured by indices like richness and evenness, tends to increase in age. However, the diversity measure might decline in very advanced age. Certain bacterial species, such as *Faecalibacterium prausnitzii*, *Eubacterium rectale*, and *Rumunococcus torques*, have been reported to decrease in relative abundance with age. Conversely, other taxa, including *Finegoldia magna*, *Bifidobacterium dentium*, and *Clostridium clostridioforme*, have shown increased abundance in elderly individuals. At the functional level, metabolic pathways like L-valine biosynthesis and L-isoleucine biosynthesis have been found to be negatively correlated with age. Short-fatty acid (SCFA)-producing bacteria, such as *Bifidobacterium spp.*, *Akkermansia muciniphila*, and *Bacteroides spp.*, have also been highlighted as influential in age prediction models (*Age Prediction Using Microbiota*, 2025; Y. Chen et al., 2022; Malesu, 2025; Min et al., 2024; Seo et al., 2023; H. Wang et al., 2025).

In classification studies, particularly the one focused on superagers, specific bacterial genera were identified as significant features distinguishing superagers from typical agers (Kim et al., 2024). Key distinguishing genera included *Alistipes*, *PAC001137_g*, *PAC001138_g*, *Leuconostoc*, and *PAC001115_g*. Examining these features more closely reveals their taxonomic distribution. Within the Firmicutes phylum, the selected features included the genus *Leuconostoc* (family *Leuconostocaceae*). Several significant genera also belonged to the class *Clostridia* (also within Firmicutes), namely *PAC001115_g*, *PAC001137_g*, and *PAC001138_g*. Additional significant genera identified from *Clostridia* included *PAC000194_g*, *PAC001236_g*, and *Romboutsia*. From the *Bacteroidetes* phylum, the genus *Alistipes* was selected as a significant feature. Notably, this analysis identified no significant distinguishing features from the *Actinobacteria* or *Proteobacteria* phyla.

While there is some consistency in the identification of age-predictive microbial features across different studies, variation exists, likely due to differences in study populations, methodologies, and the specific machine learning algorithms utilized. For instance, while some studies report an increase in overall microbial diversity with age, others might find this trend only up to a certain age point. The specific bacterial taxa identified as age-related can also vary, potentially reflecting the influence of dietary habits, genetic backgrounds, and geographical locations of the studied cohorts. Different machine learning algorithms might also emphasize different sets of features based on their inherent mechanisms for pattern recognition.

The identified features encompass both well-known commensal bacteria and bacteria with established metabolic functions, such as the production of SCFAs and the metabolism of amino acids. This suggests that the aging process in the gut microbiome is characterized by both compositional shifts in the types of microbes present and functional changes in the metabolic activities carried out by these microbial communities.

Table 3. Key Gut Microbiome Features Predictive of Age Identified Across Studies

Feature (Taxon or Pathway)	Direction of Change with Age	Studies Reporting This Feature
Microbial Diversity (Richness & Evenness)	Increase (may decline in very old age)	(Malesu, 2025; Q. Wang et al., 2024)
Faecalibacterium prausnitzii	Decrease	(Wang et al., 2025)
Eubacterium rectale	Decrease	(Wang et al., 2025)
Ruminococcus torques	Decrease	(Wang et al., 2025)
L-valine biosynthesis	Decrease	(Wang et al., 2025)
L-isoleucine biosynthesis	Decrease	(Wang et al., 2025)
Finegoldia magna	Decrease	(Chen et al., 2022)
Bifidobacterium dentium	Decrease	(Chen et al., 2022)
Clostridium clostridioforme	Increase (in elderly)	(Chen et al., 2022)
SCFAs-producing bacteria (Bifidobacterium spp., Akkermansia muciniphila, Bacteroides spp.)	Increase (in elderly)	(Age Prediction Using Microbiota, 2025; Min et al., 2024)
Alistipes	Increase (may decline in very old age)	(Kim et al., 2024)
PAC001137_g	Increase (in elderly)	(Kim et al., 2024)
PAC001138_g	Increase (in elderly)	(Kim et al., 2024)
Leuconostoc	Increase (in elderly)	(Kim et al., 2024)
PAC001115_g	Increase (in elderly)	(Kim et al., 2024)

4. Design and development

4.1. Dataset

To examine the relationship between human gut taxonomic profiles and chronological age, we compiled full metagenomic sequences from 637 healthy individuals, sourced from 11 publicly available datasets accessed through the `curatedMetagenomicData` package in Bioconductor (version 3.12.0). Only samples from EU countries, accessed January 2023, and classified as healthy adults aged 19 to 66 years with available age metadata were included. We categorized this age distribution into 10 age ranges, each encompassing five years, in order to perform the exploratory analysis. This information is available in the supplementary file *metadata.xlsx*.

4.2. Pre-Processing

4.2.1. Sample Quality Control and Trimming

Raw sequencing reads were processed using the taxonomic profiler GAIA (v2.1.3) (Paytuví et al., 2019), developed by Sequentia Biotech. The GAIA workflow includes a module of Quality Check (QC) and a module of taxonomic profiling. Relative to the QC module, initial quality control assessment was performed using FastQC (v0.11.9). Subsequently, adapter sequences and low-quality bases were trimmed using BBDuk (v3.1.0), specifically employing the “`bbduk.sh`” script. Reads with an average base quality score below a Phred score of 20 were discarded. Reads meeting the quality threshold underwent right-end trimming to remove bases falling below the quality cutoff. Finally, reads shorter than 75 base pairs after trimming were removed from the dataset to ensure only high-quality, sufficiently long reads were retained for downstream analysis. Relative to the taxonomic module, GAIA performs taxonomic classification using a mapping-based approach based on BWA (H. Li & Durbin, 2009), retrieving the best alignments and applying a Lowest Common Ancestor (LCA) algorithm. The classification is based on a custom database derived from the NCBI BLAST “nt” database (release 2023) including human, prokaryote, fungi, and virus sequences.

4.2.2. Normalization and Filtering

Microbiome data is typically compositional, meaning it represents relative abundance rather than absolute counts. This compositionality arises from the technical limitations of sequencing, where the total number of reads obtained for a sample reflects sequencing depth rather than the actual biomass of each organism. As a result, the apparent abundance of a taxon is influenced by the sequencing depth, with higher sequencing depth potentially leading to more accurate estimates of relative abundance. Consequently, the abundance of one taxon can affect the apparent abundance of others.

We first apply a filtering step to remove low-abundance and low-prevalence taxa, which are more likely to be influenced by sequencing errors or represent technical noise. Specifically, we retained taxa that met both of the following criteria:

- Had a relative abundance of at least 0.05%.

This filtering is performed prior to the normalization step to ensure that we are working with biologically relevant taxa that are not likely to be dominated by noise from sequencing artifacts.

- Were present in at least 10% of the samples.

By making this filtering step we reduce the dimensionality of the data and we focus on more prevalent taxa, improving the robustness of downstream analyses.

Then we normalize the raw read counts to relative abundances with ‘rmick’ (version 1.9.0) an R package developed by Sequentia Biotech as the statistical basis of the MICK Human microbiome platform (<https://mick.bio/>). This transformation adjusts for varying sequencing depth across samples, converting raw counts into proportions where the sum of all taxa within a sample equals one. While this approach makes samples with different sequencing depths more comparable, it does not eliminate the inherent challenges of comparing low-abundance taxa between samples with widely varying read depths. Samples with very different sequencing depths may still lead to over- or under-representation of low-abundance taxa, which should be considered when interpreting results.

For subsequent machine learning (ML) modeling, we applied Min-Max scaling. While relative abundances are already bounded between 0 and 1, Min-Max scaling further standardizes the data. This scaling transforms all feature values to a common range of 0 to 1. This is crucial for many ML algorithms because it prevents features with larger variances from dominating the model training process. By ensuring that all features have the same dynamic range, Min-Max scaling contributes to a more stable and effective ML analysis.

4.3. Exploration Data Analysis (EDA)

4.3.1. Rarefaction Curve

To assess whether the sequencing depth adequately captured the microbial diversity within our samples, we performed rarefaction curve analysis. This technique involves plotting the number of observed species as a function of the number of sequencing reads in a sample. The shape of the resulting curve indicates whether sufficient sequencing depth was achieved. A curve that reaches an asymptote suggests that additional sequencing would likely yield few new taxa, indicating adequate sampling. Prior to this analysis, we filtered out taxa identified as unknown at the species level and removed singletons —taxa represented by only a single read across all samples— since these are likely to be sequencing artifacts or contaminants. Rarefaction curves were generated using the ‘vegan’ package in R (version 2.6-8).

4.3.2. Alpha Diversity

Alpha diversity, a measure of within-sample microbial diversity, was assessed using observed richness, which represents the total number of distinct taxa detected per sample.

To ensure comparability, unknown species were filtered to reduce noise. Counts were rarefied using the *subsampleCounts* function using the ‘mia’ (version 1.12.0), which downsamples all samples to the smallest library size, which was 589,485 reads. This is done to make sure that each sample contributes the same number of reads to the analysis, avoiding biases that could be introduced by samples with higher sequencing depth.

Alpha diversity indices (observed richness) was computed from the rarefied data using the *estimateRichness* function. To explore the association between host age and gut microbial richness, two different age grouping strategies were employed. First, samples were categorized into three broad age groups: ‘Young’ (18 - 32), ‘Middle’ (33 - 52), and ‘Old’ (53 - 66) (**Figure 8**). Second, a more granular analysis was performed using ten distinct age groups with narrower ranges, spanning from 18-22 years (Group 1) to 63-67 years (Group 10). Microbial richness distributions were visualized using violin and box plots, and statistical comparisons between groups were made using the Wilcoxon rank-sum test, a non-parametric method suitable for non-normally distributed data, which is common in microbiome studies. This test does not assume underlying data distributions and is robust to outliers. Only statistically significant differences were reported and visualized in the final plots. The potential linear relationship between age group and richness was also assessed.

4.3.3. Beta diversity analysis

- PCoA

To visualize differences in microbial community composition between samples (beta diversity), we performed Principal Coordinates Analysis (PCoA) based on Bray-Curtis dissimilarity. The Bray-Curtis dissimilarity matrix, which quantifies the dissimilarity in taxonomic abundance between samples, was calculated using the ‘scater’ (version 1.32.1) package. PCoA was then used to reduce the dimensionality of this matrix, allowing for visualization of sample relationships in a two-dimensional space. Statistical significance of differences in community composition between experimental groups (e.g., age ranges) was assessed using Permutational Multivariate Analysis of Variance (PERMANOVA) using the ‘adonis’ function in the ‘vegan’ (version 2.6-8) package in R. The PERMANOVA test was performed with 99 permutations. A p-value of <0.05 was considered statistically significant.

- PCA

PCA - Principal Component Analysis is a technique used to reduce the complexity of large datasets by identifying patterns and trends in the data. It does this by transforming the original variables into a new set of variables, called principal components, which are uncorrelated and ordered in decreasing order of variability. Prior to PCA, the raw read counts were transformed using a centered log-ratio (CLR) transformation. CLR transformation addresses the compositional nature of microbiome data, mitigating the effects of spurious correlations. Euclidean distance was used as a measure of dissimilarity. PCA was conducted using ‘factoextra’ (version 1.0.7) package in R. The resulting PCA plot visualizes samples based on their CLR-transformed microbial profiles, and the principal components (PCs) represent the axes of greatest variance in the data.

4.4. Overview of the Machine Learning Pipeline

The machine learning pipeline was developed and executed in Google Collab (version 1.2.0) environment as this platform facilitates the access to necessary libraries and computational resources.

The machine learning workflow employed in this study incorporated 10-fold cross-validation to ensure robust model training and evaluation. The key stages of the pipeline consists of four parts: feature selection, fitting the model with the important features selected, hyperparameter tuning, and the final testing.

Initially, seven algorithms (Random Forest Regressor, Linear Regression, Support Vector Regression, Gradient Boosting Regressor, XGBoost Regressor, LightGBM Regressor, and CatBoost Regressor) were trained and evaluated on the preprocessed data to provide the baseline metrics in order to compare in downstream analysis. After that, four different feature selection methods were applied to the training data:

- Method 1: Comparative analysis of SHAP values and Importance-Based methods
- Method 2: SelectKBest
- Method 3: Boruta-SHAP
- Method 4: Permutation Feature Importance (PFI)

For each feature selection method, evaluation metrics and visualizations were employed to identify the most informative features for training our regression models. Subsequently, hyperparameter tuning was performed with Optuna on the best-performing model (as determined by cross-validation) to optimize its predictive accuracy. Finally, the relative importance of the features in the final selected model was analyzed with SHAP values to gain insights into the underlying biological factors driving the prediction of age.

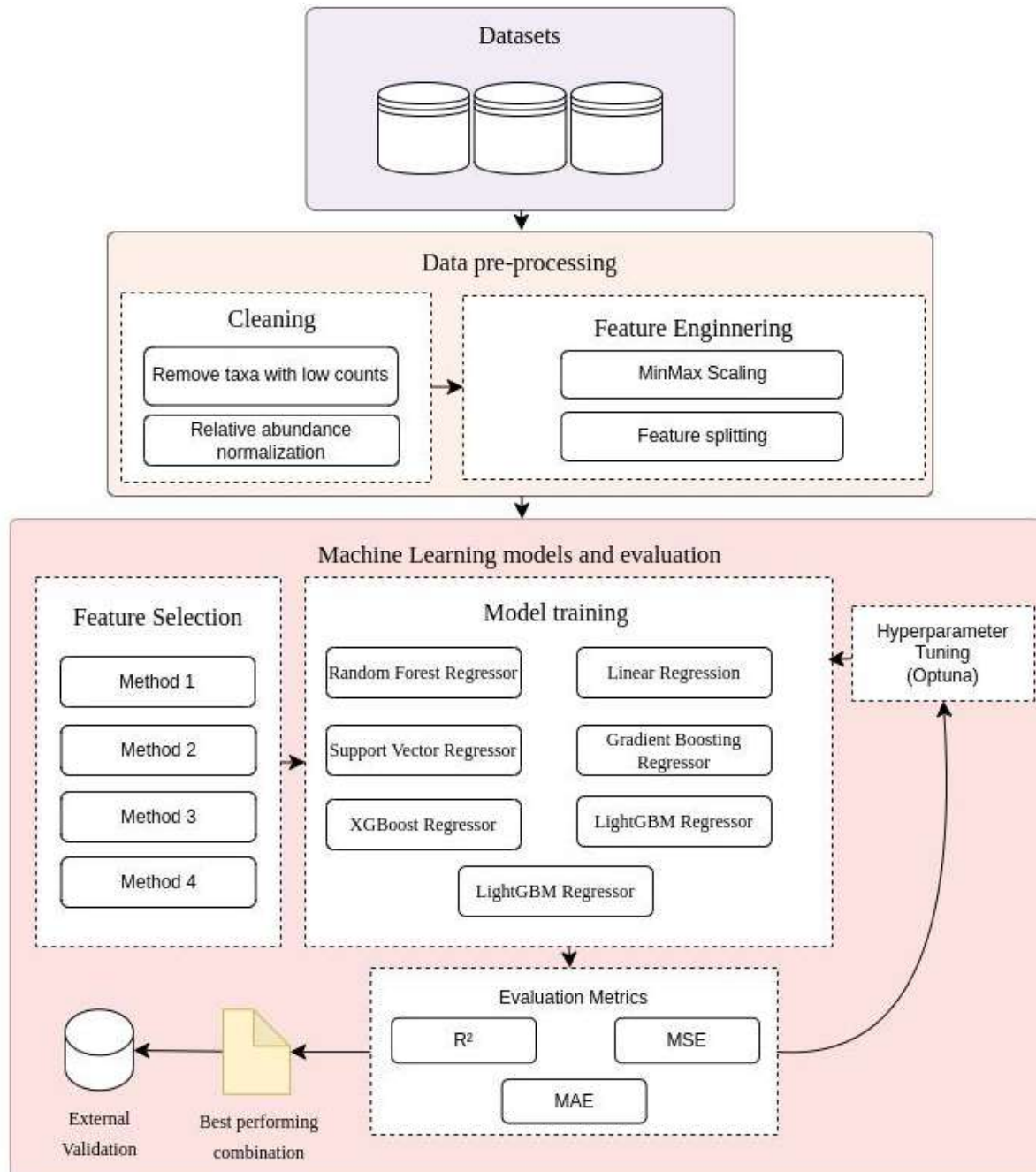


Figure 1. Overview of the Machine Learning Pipeline Schema

Workflow diagram illustrating the machine learning pipeline employed in this study. Raw datasets first undergo a pre-processing stage, which includes Cleaning (removal of taxa with low counts and relative abundance normalization) and Feature Engineering (MinMax scaling and feature splitting). Following pre-processing, one of four distinct Feature Selection methods is applied. The selected features are then used for Model Training, employing several regression algorithms: Random Forest Regressor, Linear Regression, Support Vector Regressor (SVR), Gradient Boosting Regressor, XGBoost Regressor, and LightGBM Regressor. Model Hyperparameters are tuned using Optuna. The performance of the trained models is assessed using Evaluation Metrics, specifically R-squared (R²), Mean Squared Error (MSE), and Mean Absolute Error (MAE). Finally, the best performing model combination is identified through External Validation.

4.5. Machine Learning Methods

This section details the machine learning methodologies employed to predict age from microbiome data. We begin by justifying the choice of regression over classification for this task.

4.5.1. Regression vs. classification for age prediction

The target variable in this study is age, a discrete numerical variable (e.g., 22, 34, 45 years). While age values are whole numbers, they represent a continuous result. Therefore, we choose to use regression, which is designed to predict numerical outputs.

Using classification would have required us to artificially categorize age into ranges (e.g., 20-25, 26-30 years). Such binning would lead to a loss of information and a reduction in the precision of our predictions. For instance, predicting an age range of “20-25 years” is less precise than predicting a specific age of “22 years.” Regression models allow us to maintain the full resolution of the age information, providing more precise and biologically accurate predictions.

4.5.2. Model Evaluation: K-Fold Cross-Validation

To evaluate the predictive performance of our regression model and ensure a robust assessment, we used k-fold cross-validation.

In k-fold cross-validation, the training dataset is partitioned into k equally sized subsets, or “folds”. The model is trained and evaluated k times. During each iteration:

1. One fold is held out as a temporary validation set.
2. The remaining k-1 folds are used to train the model.
3. The trained model is then used to predict the target variable for the held-out validation fold.
4. Performance metrics (R^2 , MSE, and MAE) are calculated by comparing the model’s predictions with the actual values in the validation fold.

This process is repeated k times, with each of the k folds serving as the validation set exactly once. Finally, performance metrics from the k iterations are averaged to provide an overall estimate of the model’s predictive accuracy.

For this study, we used 10-fold cross validation (k=10). This means the data was divided into ten folds, and the model was trained and evaluated 10 times.

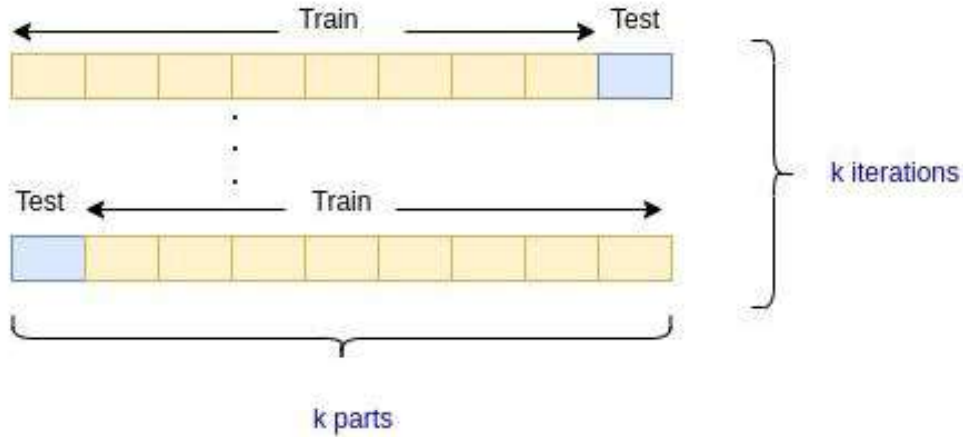


Figure 2. Diagram of *k*-fold cross-validation with *k* = 10

4.5.3. Performance Metrics

The performance of the model was evaluated using three metrics:

R² (Coefficient of Determination): This metric measures the proportion of variance in the independent variable (age) that can be explained by the model. Higher R² values indicates a better fit, with a maximum score of 1.

$$R^2 = 1 - (SS_{res} / SS_{tot})$$

Where,

- R² is the Coefficient of Determination
- SS_{res} is the Sum of squared residuals calculated as:

$$SS_{res} = \sum (y_i - \hat{y}_i)^2$$

- SS_{tot} = Total sum of squares, calculated as:

$$SS_{tot} = \sum (y_i - \bar{y})^2$$

- y_i is the actual value of the target variable (age) for the i-th sample.
- $\hat{y}_i$ is the predicted value of the target variable (age) for the i-th sample.
- $\bar{y}$ is the mean of the actual values of the target variable (age).
- $\sum$ denotes the summation over all samples.

Mean Squared Error (MSE): This metric calculates the average of the squared differences between the predicted and actual age values. Lower MSE values indicate better accuracy.

$$MSE = (1 / n) * \sum (y_i - \hat{y}_i)^2$$

Where,

- MSE is the Mean Squared Error.
- n is the total number of samples.
- y_i is the actual value of the target variable (age) for the i -th sample.
- $\hat{y}_i$ is the predicted value of the target variable (age) for the i -th prediction.
- Σ denotes the summation over all samples.

Mean Absolute Error (MAE): This metric calculates the average of the absolute difference between the predicted and actual age values. The MAE is expressed in the same units as the target variable (years), making it easily interpretable. For example, a MAE of 2.4 years indicates that, on average, the model's predictions are off by 2.4 years.

$$MAE = (1 / n) * \Sigma |y_i - \hat{y}_i|$$

Where,

- MAE is the Mean Absolute Error.
- n is the total number of samples.
- y_i is the actual value of the target variable (age) for the i -th sample.
- $\hat{y}_i$ is the predicted value of the target variable (age) for the i -th sample.
- $|\dots|$ denotes the absolute value.
- Σ denotes the summation over all samples.

4.6. Feature Selection

Feature selection in microbiome data analysis is particularly challenging due to the high dimensionality of the data and the confounding effects arising from complex and dynamic interactions among the host environment, and microbiome. These confounding factors can hinder the performance of variable selection methods, making it difficult to select important microbial features associated with the outcome.

This section describes the four feature selection methods employed to identify the most relevant microbiome features for predicting age. By combining multiple approaches, we aimed to achieve a robust and comprehensive selection process.

4.6.1. Method 1. Comparative Analysis of SHAP-values and Importance-Based Methods

To identify the most relevant microbiome features for predicting age, we employed a robust feature selection strategy that combines SHAP (Shapley Additive exPlanations) values with traditional feature importance scores derived from multiple machine learning models. Additionally, we combined

features by employing a rank-based method, which integrates the results from both approaches, ensuring a comprehensive selection of features that captures the most informative ones from both methods.

The feature selection process was performed exclusively on the training dataset to ensure that the test set remained completely independent to final model evaluation. To account for potential variability in feature rankings due to data splitting, we implemented a 10-fold cross-validation (CV) procedure. For each of the ten folds, we performed the following steps:

1. **Model training and feature ranking:** We trained seven different regression models on the training data within each fold: Random Forest Regressor, Linear Regression, Support Vector Regressor, Gradient Boosting Regressor, XGBoost Regressor, LightGBM Regressor, and CatBoost Regressor.
2. **SHAP value calculation and ranking:** For each trained model, we calculated SHAP values to assess each feature's contribution to individual predictions. We then ranked the features based on their SHAP values.
3. **Built-in feature importance ranking:** We also extracted the built-in feature importance scores from each of the seven regression models. These scores provide alternative measures of feature importance based on the specific algorithms. We then ranked the features based on these built-in scores.
4. **Rank aggregation:** We aggregated the feature rankings obtained from SHAP values and the built-in feature importance scores using a rank aggregation method. This step combines the information from both approaches to produce a consensus ranking of features. In the following section this method is explained.
5. **Feature evaluation and selection:** We evaluated the top 100 features from each of the three ranking methods (SHAP, Built-in, and Rank Aggregation) separately on the test set using Mean Absolute Error (MAE) as the evaluation metric again on the seven algorithms.
6. **Final feature selection:** The set of 100 features (either from SHAP, Built-in, or Rank Aggregation, the latter of which is explained in Section 4.6.1.1) that yielded the best (lowest) MAE on the test set was selected as the final set of features for the model.

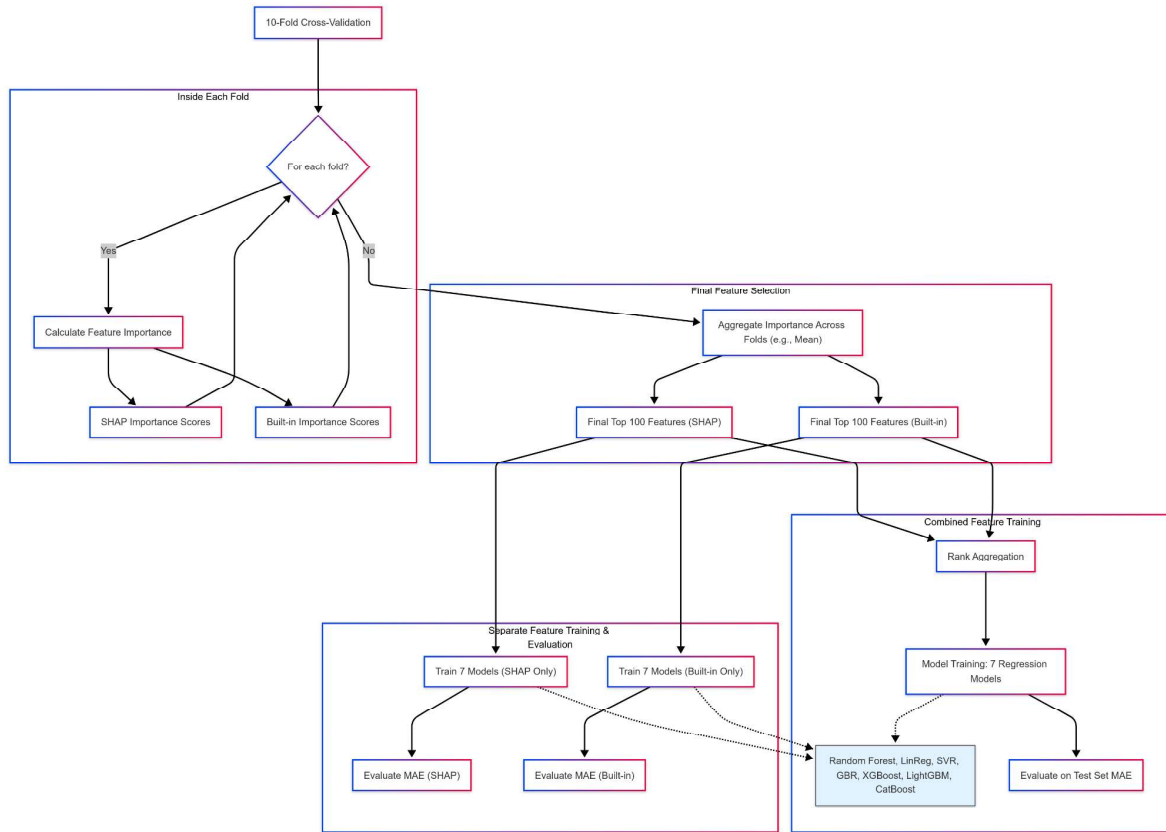


Figure 3. Comparative Analysis of SHAP-values and Importance-Based Method Scheme

This analysis compares two feature selection approaches integrated with 10-fold cross-validation. Within each fold, feature importance scores are calculated using both SHAP and a built-in method. After completing all folds, the importance scores for each feature are aggregated across the folds (e.g., using the mean score). Based on these aggregated scores, a final list of the top 100 features is determined for the SHAP method, and another final list of the top 100 features is determined for the built-in method. Subsequently, these two final feature sets are evaluated: first, each set is used separately to train seven different regression models; second, the features from both final lists are combined using a rank aggregation approach, and this consolidated set is used to train the same seven models. Model performance is evaluated on a test set using Mean Absolute Error (MAE).

4.6.1.1. Combined SHAP and Built-in Ranks

To put in common the information provided by SHAP values and built-in feature importance scores, we developed a combined rank method.

The combined rank method is calculated as follows:

1. **Aggregate fold rankings:** After completing the 10-fold cross-validation, we create two overall ranked lists: one by averaging the SHAP feature rankings across the ten folds, and another by averaging the built-in feature importance rankings across the ten folds for each of the seven algorithms trained on.
2. **Rank assignment:** We create dictionaries that map each feature to its average rank in the SHAP ranking (SHAP_rank) and the average rank in the built-in ranking (BuiltIn_rank).
3. **Rank combination:** We iterate through the unique set of features from both averaged ranked lists. For each feature:

- If the feature is present in both lists, its combined rank is calculated as the average of its SHAP rank and its built-in rank:

$$\text{Combined rank} = (\text{SHAP}_{\text{rank}} + \text{BuiltIn_rank}) / 2$$

- If the feature is present in only one list, it is assigned a default rank of 100 in the missing list before calculating the combined rank. This ensures that all features are considered, even if they are not identified as important by both methods.
4. **Final ranking:** We sort all features based on their combined rank, with lower ranks indicating higher importance.
 5. **Top 100 feature selection:** The top 100 features from the sorted list are selected and trained on the seven regression models.

Algorithm: Combined SHAP and Built-in feature ranking

Input:

CV_SHAP_Rankings: SHAP rankings from CV folds
 CV_Builtin_Rankings: Built-in rankings from CV folds
 default_rank: Default rank for missing features

Output:

Top_Features: Top-ranked features

Steps:

1. Overall_SHAP_Ranking = Average ranks from CV_SHAP_Rankings
2. Overall_Builtin_Ranking = Average ranks from CV_Builtin_Rankings
3. For each feature:
 SHAP_Rank = Overall_SHAP_Ranking[feature] or default_rank
 Builtin_Rank = Overall_Builtin_Ranking[feature] or default_rank
 Combined_Rank = (SHAP_Rank + Builtin_Rank) / 2
4. Sort features by Combined_Rank
5. Top_Features = Top features
6. Return Top_Features
7. **End.**//algorithm

Figure 4. Pseudocode of Feature Selection, Method I

4.6.2. Method 2. SelectKBest

The second feature selection method used was SelectKBest from the ‘scikit-learn’ (version: 1.1.3) package, a filter based feature selection method. Filter methods evaluate the relationship between each feature and the target variable independently of any machine learning model, relying on statistical measures to rank features importance (Bishop, C. M, 2006).

In this analysis, we used the $f_regression$ score function with SelectKBest. This function calculates the F-statistic, which measures the linear dependence between each feature and the target variable. The F-statistic is appropriate for regression tasks, as it assesses whether there is a statistically significant linear relationship between the feature and the continuous target variable (age).

We explored several values for k , the number of top-ranking features to select. Specifically, we trained and evaluated models using the top 100, 50, 40, 30, 20, 10 and 5 features, to assess the impact of feature set size on the model. The mean Mean Absolute Error (MAE) metric on the test set was used to compare the performance of models trained with different values of k . The results of this analysis are presented in Section 5.2.2, which illustrates the relationship between the number of selected features and model performance.

4.6.3. Method 3. Boruta-SHAP

The third method we selected was the combination of two powerful techniques: Boruta and SHAP. Each method offers distinct advantages, and their combination allows for a comprehensive evaluation of the most important features contributing to the model performance.

- **Boruta Algorithm:** Boruta is a feature selection method that works by comparing the importance of each feature in the dataset with the importance of random “shadow” features, which are created by permuting the values of the original features. Thus, features that have higher importance than the shadow features are retained, while those with lower importance are rejected (Kursa et al., 2010).

We used the BorutaShap (version 1.0.17) package, which is a wrapper method that combines the Boruta feature selection algorithm with Shapley values.

4.6.4. Method 4. Permutation Feature Importance (PFI)

We also employed Permutation Feature Importance (PFI) from the ‘scikit-learn’ (version: 1.1.3) package, a model-agnostic technique, to assess the contribution of individual microbial taxa to age prediction. PFI directly measures the impact of each feature on the model’s predictive performance by quantifying the increase in prediction error when a feature’s values are randomly shuffled (Breiman, 2001).

Figure 5 illustrates the computation of the importance of feature X3 in a small dataset made up of five features. First, the model error is computed using the trained model and the labeled dataset (Errorbase). Later, feature X3 in the labeled dataset is shuffled. By shuffling feature X3, the

relationship between the feature and model prediction is broken leading to a decrease in prediction performance or increase in model error. The model error of the new dataset with the shuffled feature is calculated ($ERROR_{x_shuffled}$) and the difference between the initial error and the shuffled error indicates the importance of the feature X3. The greater the difference, the more important the feature is.

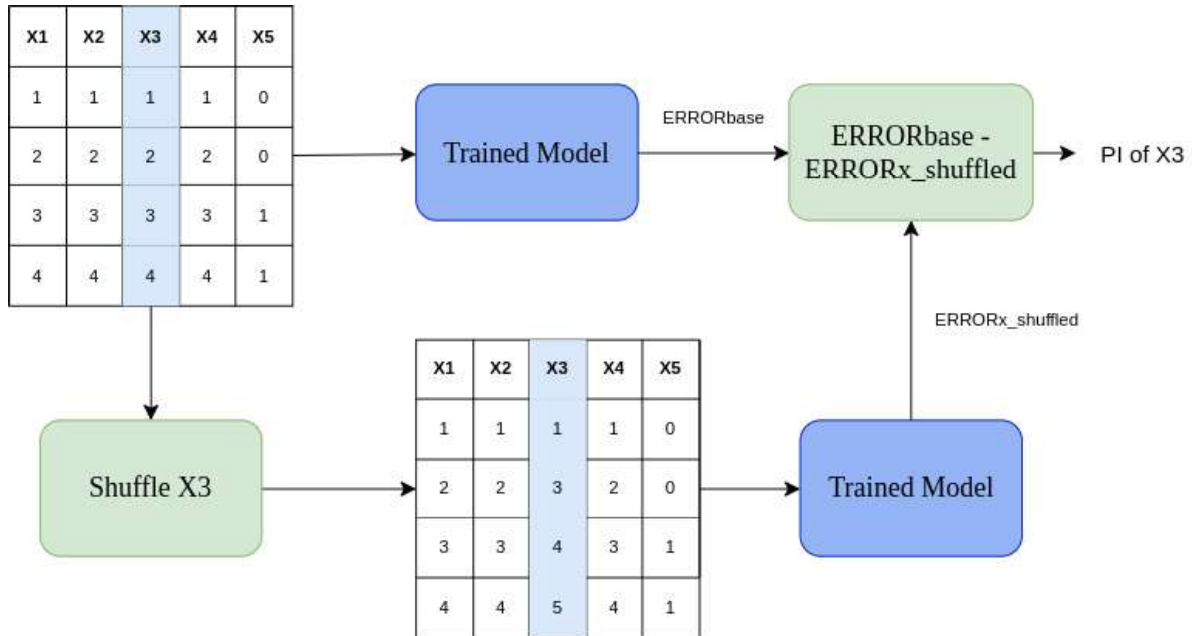


Figure 5. Schematic diagram of permutation importance algorithm feature score calculation.

4.7. Hyperparameter Tunning

To improve the performance of the machine learning models used in this study, hyperparameter tuning was conducted using Optuna (version: 4.2.1). Optuna automates the process of selecting the best hyperparameters by minimizing or maximizing an objective function, which in this case was the R^2 score on the test set.

4.8. Model Evaluation

The performance of the model was evaluated using a combination of metrics for the regression task, described in **Section 4.5.3**. These metrics were calculated on both the validation sets within the 10-fold cross-validation procedure, on the final, held-out test set, and on an independent, external validation dataset. The test set performance provides us a final unbiased evaluation of the model performance on data not used during training and the external validation dataset provides a further assessment of the model's generalization to a completely independent dataset, assessing a real-world applicability.

4.8.1. Final Model

In addition to the metrics described above, the final selected model was further analyzed to provide a deeper understanding of its behaviour and the importance of the selected features. Shapley Additive exPlanations (SHAP) values were used to assess the contribution of each feature to the final model's predictions. In addition, a plot showing the distribution of errors $|y_i - \hat{y}_i|$ during training. The visualization helps us understand in which range of age the model works well and in which the predictions are worst. Finally, the prevalence of the selected features was evaluated by calculating the frequency with which each feature appeared in the entire dataset. This was done to determine whether the selected features were consistently present and informative across the data, rather than being random or spurious.

4.9. Sample Clustering with UMAP

Following model development, further analyses and exploration was made on the features in our model in order to identify age-related patterns.

To visualize the relationships among samples based on the selected microbiome features, first was applied KMeans clustering to assign samples into distinct groups, where the optimal number of clusters was first determined by the Silhouette method. Then, each sample was assigned a cluster label. Next, SHAP classification was performed to identify the features that contributed most to the clustering. To achieve this, two complementary approaches were employed. First, an XGBoost classification model was trained to predict the sample cluster assignments based on the feature data. SHAP analysis was then applied to quantify the most influential features that contributed to the separation of clusters. The SHAP values allowed us to assess the relative importance of each feature in predicting cluster assignment. Second, the difference was calculated between the maximum and minimum values for each feature across the centroids of the identified clusters. This metric provided insight into the features with the greatest separation in abundance between clusters, helping us identify those that most strongly differentiate the clusters.

The top 5 most important features were selected based on their importance scores from SHAP values as well as their ability to distinguish clusters based on their maximum-minimum value differences across cluster centroids. These features were considered discriminative and were used to further explore the relationship between feature importance and the clustering results.

To visualize the relationships between the samples and their assigned clusters, Uniform Manifold Approximation and Projection (UMAP) was applied for dimensionality reduction. UMAP was used to reduce the data to high dimensional space to two dimensions to better visualize the clustering of the samples. The following parameters were used for UMAP: `n_neighbor=15`, `min_dist=0.1`, `n_components=2`, and the distance metric was set to euclidean for abundance data. The UMAP embedding was calculated on the full dataset of the features in our model, resulting in a 2D projection of the data that reflects the clustering of the samples.

The resulting two-dimensional UMAP embeddings were visualized using scatter plots, where data points were colored according to the normalized abundance values of the selected top five features.

Furthermore, the relative abundance of the top discriminative features across three age groups was examined: 'Young', 'Mid', and 'Old'. Violin plots and Krustal-Wallis tests were used to assess statistically significant differences in species abundance across the age groups, followed by Dunn's post-hoc test with Bonferroni correction.

5. Experiments and results

5.1. Exploratory data analysis (EDA) results

The individuals in our dataset ranged from 19 to 66 years of age, with a median age of 45 years (**Figure 6a**). We categorized the age distribution into 10 age ranges, each encompassing five years. Group 3 (28-35 years) had the highest sample count, with over 103 individuals, followed by Group 5 (38-42 years), which included just over 91 individuals, slightly fewer than Group 3. Group 1 (18-22 years) had the smallest sample size, with only 16 individuals.

The overall age distribution reveals a bimodal pattern, with two prominent peaks in the age range. The first peak occurs around 25-30 years, and the second peak appears around 55 years (**Supplementary Figure 1**), suggesting that our dataset is concentrated in these age ranges. The mean (red dashed line) is slightly to the right of center, around age 45, indicating a skew towards older individuals. The median (blue dotted line) is just to the right of the mean, reflecting the distribution of ages. The interquartile range (IQR), marked by green dot-dashed lines, spans from approximately ages 32 to 55, meaning that half of the samples fall within this age range, which constitutes the central portion of our dataset.

The gender distribution shows a higher representation of females (411) compared to males (233), with 0.5% of the samples classified as unknown gender (**Figure 6c**). The cohort's country distribution indicates that the majority of the samples are from the United Kingdom (222), Denmark (175), and Ireland (100). In contrast, Italy (14 samples) and Luxembourg (18 samples) are the least represented, with samples coming from a total of seven countries (**Figure 6b**).

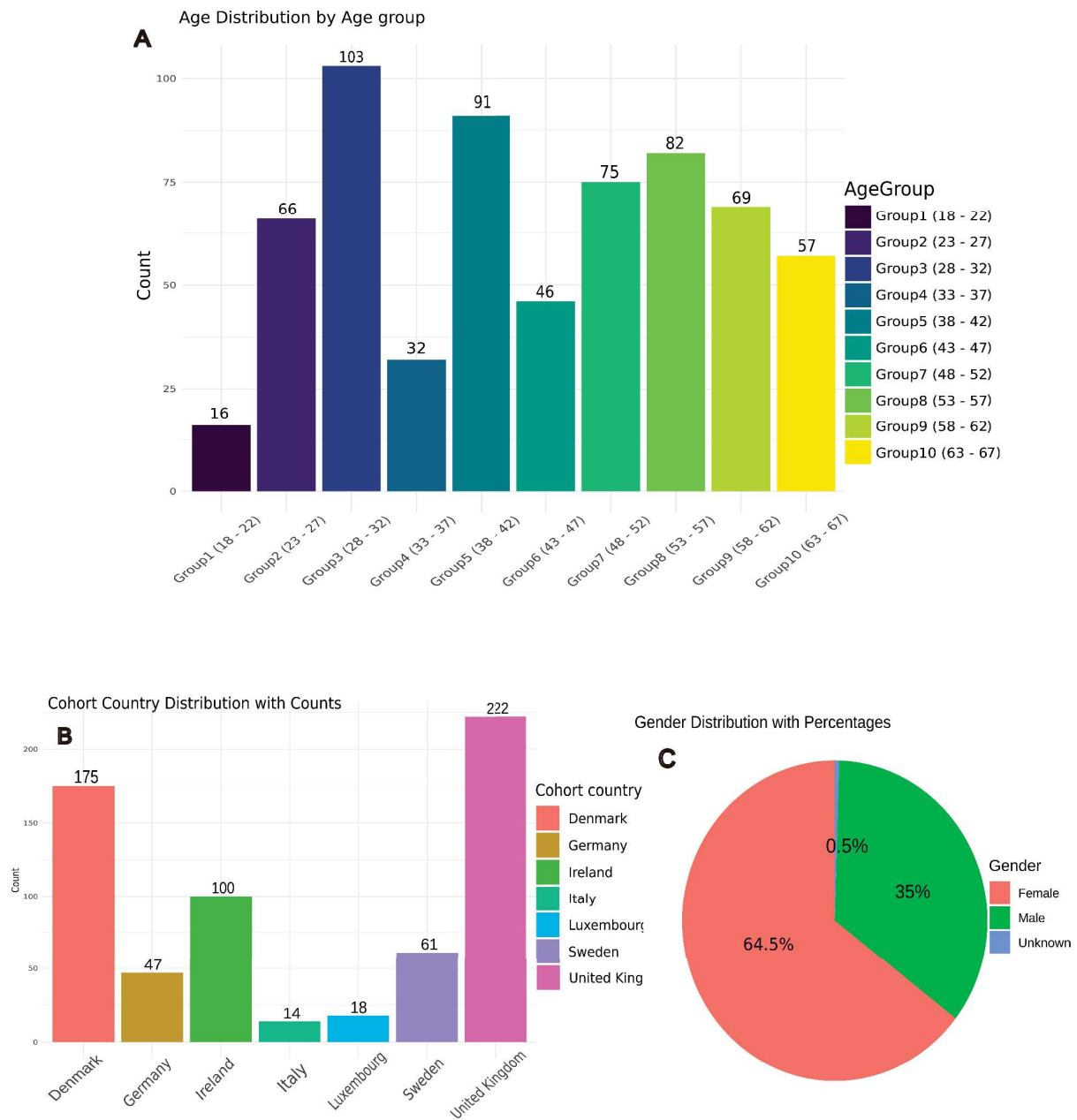


Figure 6. Demographic distribution of the study population.

(a) Age distribution of participants, ranging from 19 to 66 years. Individuals were grouped into ten 5-year intervals. Group 3 (28–35 years) had the highest number of individuals ($n = 103$), followed by Group 5 (38–42 years, $n = 91$), while Group 1 (18–22 years) had the fewest ($n = 16$).

(b) Country distribution of the cohort, with the majority of samples originating from the United Kingdom ($n = 222$), Denmark ($n = 175$), and Ireland ($n = 100$). Italy ($n = 14$) and Luxembourg ($n = 18$) were the least represented among the seven contributing countries.

(c) Gender distribution, showing a higher proportion of females ($n = 411$; 64.5%) compared to males ($n = 233$; 35%), with 0.5% of samples labeled as unknown gender.

After inspecting the dataset, we ended up with 637 samples. We normalized the raw counts to relative abundance, in order to remove the influence of sequencing depth (total reads per sample). This ensures that samples from varying depth are comparable. Following normalization, we retained taxa that had a relative abundance of at least 0.05% and present in at least 10% of the samples. Rarefaction curves showed that most of the samples had reached a saturation plateau (**Supplementary Figure 2**). We used the Wilcoxon rank-sum test to compare alpha diversity indices, highlighting if there were significant differences along the samples (**Figure 8**).

As a preliminary exploration of the microbiome composition, **Figure 7** presents a comparative analysis of the relative abundance of the top five most abundant genera found in the dataset: *Bacteroides*, *Phocaeicola*, *Prevotella*, an unclassified genus within the Lachnospiraceae family (unkn. Lachnospiraceae(f)), and an unclassified genus within the Oscillospiraceae family (unkn. Oscillospiraceae(f)). The analysis specifically contrasts two age groups: 'Young' individuals (19-33 years, shown in blue) and 'Elderly' individuals (34-66 years, shown in red).

Each panel displays boxplots illustrating the distribution of relative abundance for a specific genus within each age group. Statistical significance between the 'Young' and 'Elderly' groups for each genus was assessed using the Kruskal-Wallis test, with significance levels indicated by asterisks (* for $p < 0.05$, ** for $p < 0.01$).

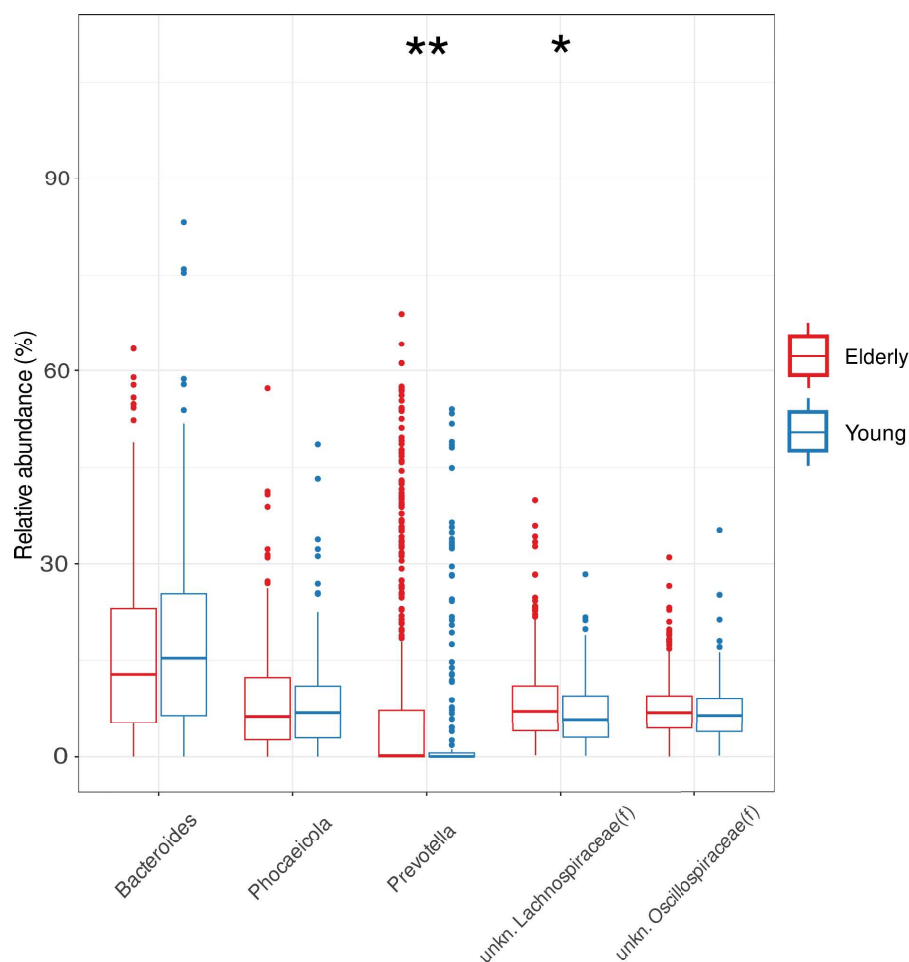


Figure 7. Comparison of relative abundance of key bacterial genera between 'Young' and 'Elderly' groups. Boxplots display the distribution of relative abundance (%) for *Bacteroides*, *Phocaeicola*, *Prevotella*, unknown Lachnospiraceae, and unknown Oscillospiraceae. Significant differences between age groups were determined using the Kruskal-Wallis test (* indicates $p < 0.05$, ** indicates $p < 0.01$).

Figure 8 illustrates the distribution of microbial richness across the three broad age groups revealing a statistically significant decrease in microbial richness from the 'Young' group compared to both the 'Middle' ($p = 0.0065$) and 'Old' groups ($p = 0.0044$). However, no significant difference in richness was observed between the 'Middle' and 'Old' groups ($p = 0.74$), suggesting a plateau in richness levels after young adulthood. A linear regression assessment yielded a very low coefficient of determination ($R^2 = 0.015$), indicating that this broad grouping explains only 1.5% of the variance in richness linearly.

The finer-grained analysis using ten age groups provided a more detailed view of richness dynamics across the lifespan studied. Age exhibited a weak but measurable association with gut microbial richness ($R^2 = 0.053$), accounting for only ~5% of the observed variation. Significant pairwise differences (e.g., $p < 0.05$) between select age groups, such as 18-22 vs. 23-27 years, suggest dynamic shifts in richness during early adulthood. However, the lack of consistent patterns across all groups implies that age-related changes are subtle and likely modulated by other factors.

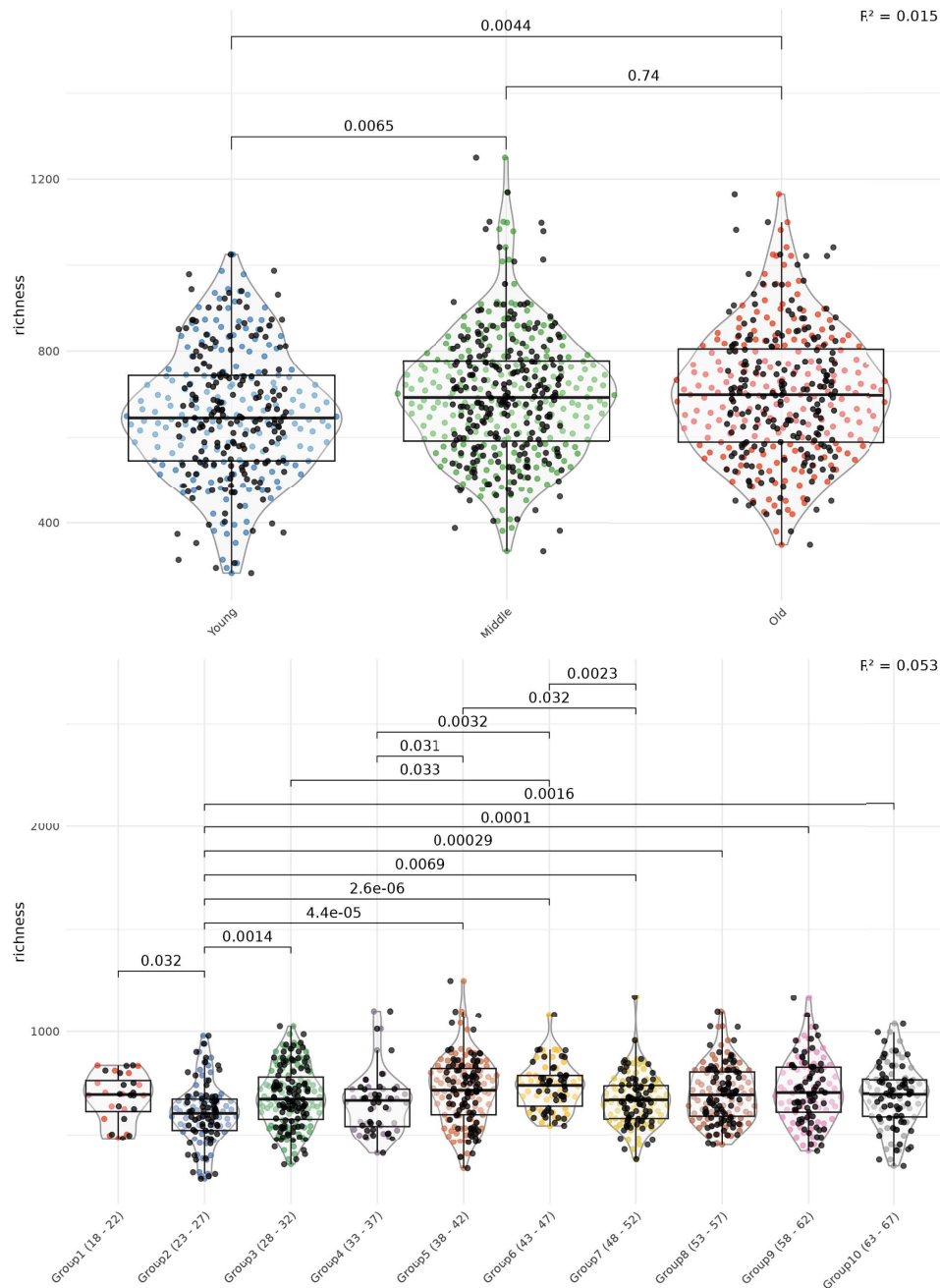


Figure 8. Alpha Diversity (Richness) Across Age Groups.

(Top) Analysis of three broad age groups reveals negligible correlation with microbial richness ($R^2 = 0.015$).
 (Bottom) Analysis across 10 finer age groups demonstrates a weak positive trend ($R^2 = 0.053$).

Principal Coordinate Analysis (PCoA) based on Bray-Curtis dissimilarity was performed to visualize differences in taxonomic composition across samples. The PCoA plot (**Supplementary Figure 3a**) revealed no distinct clustering of samples by experimental groups, suggesting overlapping microbial communities. However, PERMANOVA analysis indicated statistically significant differences between groups ($p = 0.01$), despite the lack of a clear separation in the space. This implies that while group-level differences exist, they may be driven by subtle variations in low abundance taxa.

Principal Component Analysis (PCA) Biplot (**Supplementary Figure 3b**) was used to explore relationships between samples and microbial taxa. The biplot displays the sample distribution, which is similar to PCoA; samples positioned closer to each other share similar taxonomic profiles. Vectors (arrows) represent taxa (e.g., *Ruminococcus bicirculans*) that strongly influence the principal components. Longer vectors indicate taxa with higher variance, while their direction reflects positive/negative correlations with each axis. For instance, *Clostridium sp. M62/1* was negatively correlated with PC1, which suggests a potential role in differentiating samples. Interestingly, *Clostridium clostridioforme*, a related species, has been shown to have increased abundance in elderly individuals, which may imply that *Clostridium sp. M62/1* could similarly be associated with aging or age-related microbiome shifts, further supporting its role in distinguishing these samples (Y. Chen et al., 2022).

5.2. Machine learning experiments

The machine learning pipeline was implemented to assess the predictive performance of microbial profiles in estimating host age. The analysis followed a structured approach: (1) baseline evaluation of regression algorithms, (2) comparative assessment of feature selection methods, (3) hyperparameter optimization of the top-performing model, and (4) interpretability analysis using SHAP values. Key results from each stage are presented below.

5.2.1. Baseline model performance

Seven regression algorithms—Random Forest (RF), Linear Regression (LR), Support Vector Regression (SVR), Gradient Boosting (GB), XGBoost (XGB), LightGBM (LGBM), and CatBoost (CB)—were trained using 10-fold cross-validation to establish baseline performance metrics (MAE, MSE, and R^2). As shown in **Table 4**, CatBoostRegressor demonstrated the strongest predictive capability, achieving the lowest MAE in train (10.39 ± 0.98) and the highest R^2 (0.10). The MAE achieved on the test set for CatBoostRegressor was 11.22.

Table 4: Initial Training (637 total samples and 1436 features)

Regression Model	R^2 CV, train		MSE CV, train		MAE CV, train		R^2 (Test)	MSE (Test)	MAE (Test)
	std	Mean	std	Mean	std	Mean			
Random Forest	0.08639	0.103104	18.753706	153.215815	0.6805	10.565386	0.093701	177.036301	11.386172
LinearRegression	0.120885	-0.681799	22.551131	286.214144	0.555959	13.291843	-9.25907	2011.037559	23.643418
Support Vector Machine	0.021883	0.005773	9.761189	169.646341	0.354222	11.431808	0.016933	192.032107	12.215015
Gradient Boosting	0.087119	0.053474	18.362843	161.898176	0.69163	10.816552	0.056717	184.260802	11.410172
XGBoost	0.149311	0.003514	27.407645	170.230496	0.896959	10.85582	-0.042228	203.588731	12.075024
LightGBM	0.106235	0.027539	21.793206	166.075271	0.87001	10.857259	0.044886	186.571764	11.43499
CatBoostRegressor	0.136276	0.115233	26.1497	151.171503	0.982842	10.390237	0.101515	175.509981	11.221978

5.2.2. Feature selection analysis

To improve the model's predictive performance and gain insights into the key microbial features associated with age, a feature selection analysis was conducted. Four different feature selection methods were employed.

Method 1: Comparative analysis of SHAP values and Important base method results

To identify the most relevant microbiome features for predicting age, we employed a feature selection strategy that combines SHAP (Shapley Additive exPlanations) values with traditional feature importance scores derived from multiple machine learning models. We also combined features by employing a rank-based method, which integrates the results from both approaches. The feature selection process was performed on the training dataset using 10-fold cross-validation. For each fold, we trained five regression models (Random Forest Regressor, Gradient Boosting Regressor, XGBoost Regressor, LightGBM Regressor, and CatBoost Regressor), which were selected based on preliminary baseline performance. Linear Regression and Support Vector Regression were excluded due to their comparatively poor performance in the baseline metrics. After training, SHAP values and built-in feature importance scores were calculated, and these rankings were then aggregated. A high degree of overlap was observed between the top 100 features identified by SHAP and built-in importance across the different algorithms, as illustrated in **Supplementary Figure 4**. This consistency suggests that both methods reliably identify a core set of predictive features.

We evaluated the top 100 features from the SHAP, Built-in, and Rank Aggregation methods on the test set using Mean Absolute Error (MAE). The set of 100 features that yielded the best (lowest) MAE on the test set were selected as the final set of features for the model to be trained on the seven regression algorithms (Random Forest Regressor, Linear Regression, Support Vector Regressor, Gradient Boosting Regressor, XGBoost Regressor, LightGBM Regressor, and CatBoost Regressor).

The results of this analysis are summarized in **Table 5**. Detailed metrics for each of the five trained algorithms are available in the Excel file *Feature_Selection_metrics_method1.xlsx* in the Supplementary Materials.

Table 5: Regression Model Performance with Different Feature Selection Methods

Feature Selection Method	Model	Cross-Validation			Test		
		R ²	MSE	MAE	R ²	MSE	MAE
XGBoost	LightGBM (SHAP)	0.1900 (0.1270)	135.44 (24.41)	9.526 (0.970)	0.1015	175.50	10.789
GradientBoosting	LightGBM (SHAP)	0.2475 (0.1331)	125.75 (24.69)	9.220 (1.065)	0.0676	182.14	11.108
LightGBM	LightGBM (SHAP)	0.2124 (0.1195)	131.60 (22.65)	9.424 (0.918)	0.0563	184.35	11.086
CatBoostRegressor	LightGBM (Built-in)	0.2243 (0.1311)	129.35 (22.01)	9.343 (0.851)	0.0898	177.79	10.899
Random Forest Regressor	LightGBM (SHAP)	0.1756 (0.1218)	138.34 (26.15)	9.736 (1.053)	0.0576	184.09	11.133

Note: Values in parentheses represent standard deviations for cross-validation metrics. Model: LightGBM (Method: SHAP for all except CatBoost which uses Built-in)

As highlighted in **Table 5**, the best performing combination of feature selection and model training was achieved by extracting the top 100 most important features as determined by SHAP values from the XGBoost model, and then training a LightGBM regressor using those features. This resulted in a MAE of 10.78 on the test dataset.

The performance of this LightGBM model is further detailed in **Figure 9a** which shows the predicted age versus chronological age, demonstrating a moderate correlation ($R^2 = 0.19$) and a mean absolute error of 9.53 years. **Figure 9b** displays the SHAP values for the top 20 most influential features (ordered by importance), illustrating the magnitude and direction of each feature's impact on the predicted age. As shown in **Figure 9b**, low abundance of *Sutterella wadsworthensis* is associated with a decrease in predicted age, while high abundance of *Siphoviridae sp. ctAjZ17* is associated with an increase in predicted age. These findings suggest that specific microbial taxa may serve as important biomarkers for age prediction.

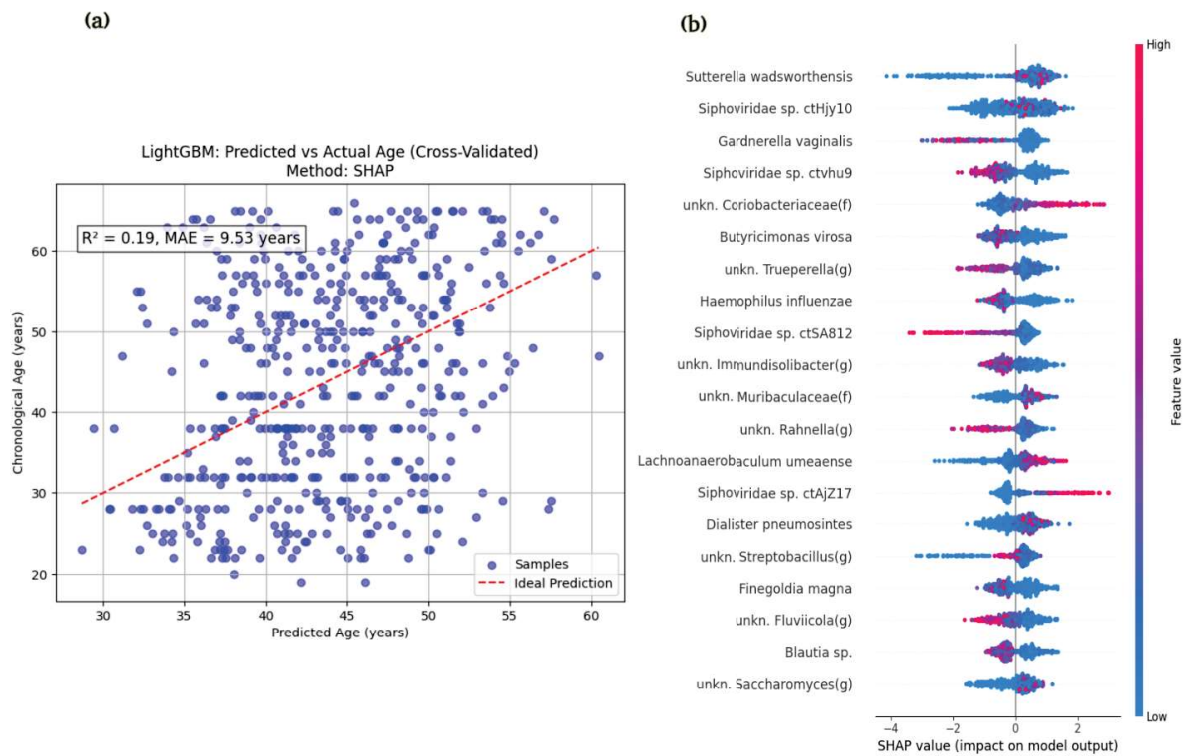


Figure 9. Performance and Feature Importance of the LightGBM Age Prediction Model

(a) Predicted age versus chronological age for the LightGBM model trained on the top 100 features selected based on SHAP values from XGBoost. The dashed line represents the ideal prediction. The model exhibits a moderate correlation between predicted and actual age ($R^2 = 0.19$) with a mean absolute error (MAE) of 9.53 years.

(b) SHAP (Shapley Additive exPlanations) values for the 20 most influential microbial features in the LightGBM age prediction model. Features are ordered by decreasing importance. The color gradient represents feature values (blue: low, red: high), and the x-axis indicates the SHAP value, representing the feature's impact on the predicted age, while high abundance of *Siphoviridae sp. ctAjZ17* is associated on an increase in predicted age.

Method 2: SelectKBest results

Next, we evaluated the performance of models trained using feature sets selected by SelectKBest. As described in the Methods section, SelectKBest is a filter method that selects the top k features based on their statistical relationships with the target variable. We explored values of k ranging from 5 to 100 to assess the impact of feature set size on model performance. The results of this analysis are illustrated in **Figure 10**, which shows the relationship between the number of selected features and test set MAE for each of the seven regression algorithms.

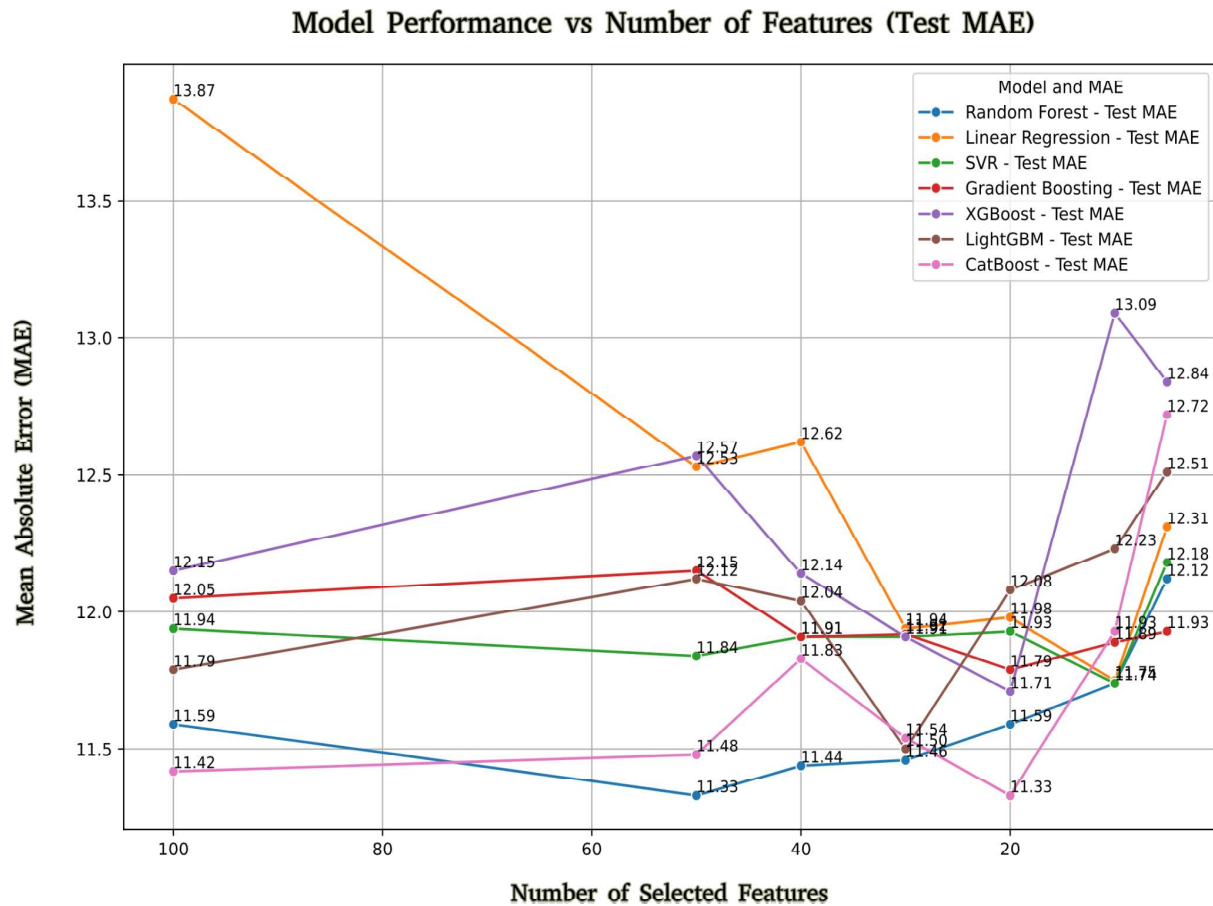


Figure 10. Model Performance vs. Number of Selected Features (SelectKBest)

Test set Mean Absolute Error (MAE) as a function of the number of selected features (k) for seven different regression algorithms (Random Forest, Linear Regression, SVR, Gradient Boosting, XGBoost, LightGBM, and CatBoost). Features were selected using SelectKBest with the $f_regression$ score function. The plot illustrates that model performance varies across algorithms and feature set sizes. For several algorithms, MAE initially decreases as k decreases, reaching a minimum around $k=20$, and then increases as k decreases further.

Figure 10 reveals that the relationship between the number of selected features and model performance is complex and varies across algorithms. For several algorithms, MAE initially decreases as we decrease k from 100 towards 20, reaching a minimum around $k=20$, and then increases as k decreases further. This indicates that initially removing less informative features improves the model's ability to predict age, but removing too many relevant features (below 20) leads to underfitting and a

decline in performance. In contrast, the performance of Random Forest Regressor and SVR is relatively flat across different values of k , suggesting that these models are less sensitive to the number of features.

As shown in **Figure 10**, model performance varied across algorithms and feature set sizes. The lowest MAE on the test was achieved by Random Forest and CatBoostRegressor models trained with 50 and 20 features, respectively, achieving a MAE score of 11.33. However, it is slightly worse than the MAE of 10.78 achieved using the SHAP-based feature selection in the previous section, suggesting that the features selected by SHAP may be more relevant for this age-prediction task.

The performance of the CatBoostRegressor model with the top 20 features selected by SelectKBest is further illustrated in **Figure 11a**, which shows the predicted age versus chronological age. The model's predictions show a weak correlation with actual age ($R^2 = 0.01$) and a mean absolute error of 10.86 years. The SHAP values for the top 20 most influential features in this model are shown in **Figure 11b**.

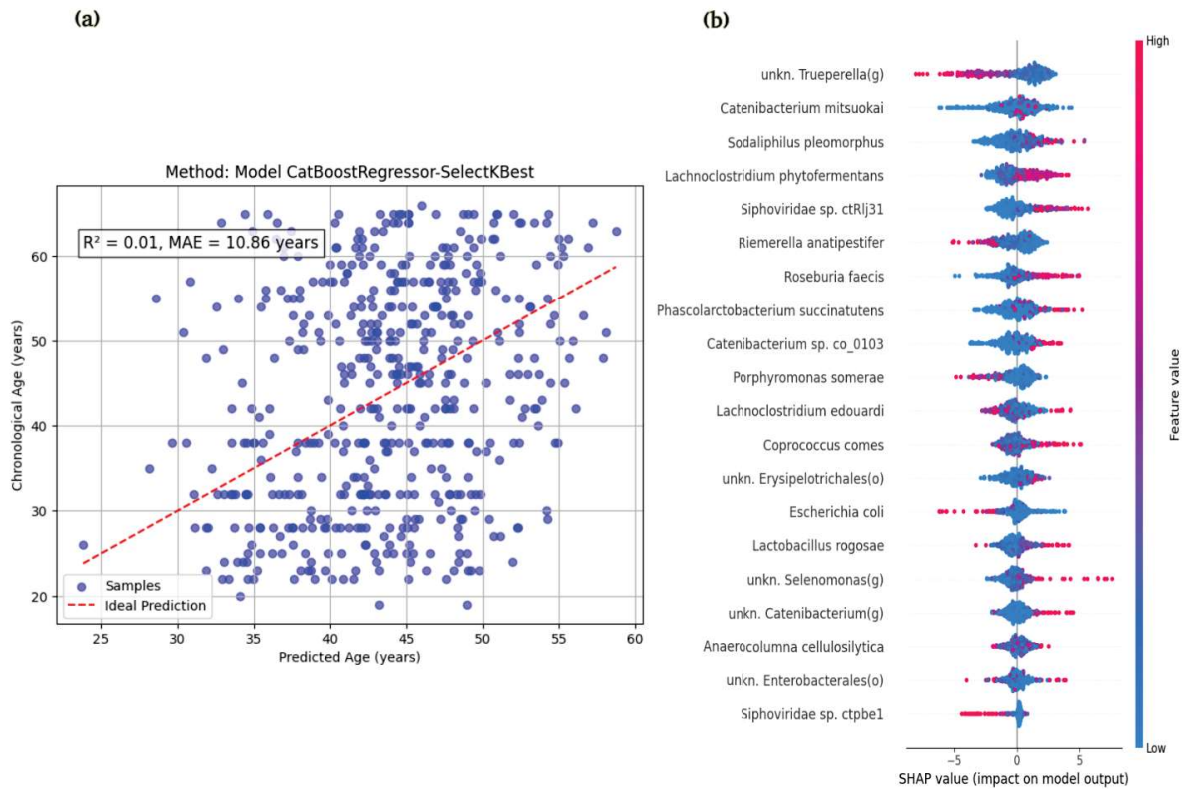


Figure 11. Performance and Feature Importance of the CatBoostRegressor Age Prediction Model

(a) Predicted age versus chronological age for the CatBoostRegressor model trained on the top 20 features selected based on SelectKBest method. The dashed line represents the ideal prediction. The model exhibits a moderate correlation between predicted and actual age ($R^2 = 0.01$) with a mean absolute error (MAE) of 10.86 years.

(b) SHAP (Shapley Additive exPlanations) values for the 20 most influential microbial features in the CatBoostRegressor age prediction model. Features are ordered by decreasing importance. The color gradient represents feature values (blue: low, red: high), and the x-axis indicates the SHAP value, representing the feature's impact on the predicted age.

Method 3: Boruta-SHAP results

The third feature selection method employed was Boruta-SHAP, a wrapper method that combines the Boruta algorithm with Shapley values. As described in the Methods section, Boruta compares the importance of original features to that of shadow features, while SHAP values provide a more granular understanding of each feature's contribution to the model's predictions. This approach identified 7 features as important for predicting age. The feature importance ranking is shown in **Figure 12a**, where features with higher Z-scores are considered more important. As seen in **Figure 12a**, the selected features ('Siphoviridae sp. ctHjy10', 'unkn. Trueperella (g)', 'Sutterella wadsworthensis', 'Siphoviridae sp. ctRlj31', 'Alistipes dispar', and 'unkn. Chryseobacterium (g)') have higher Z-scores than the shadow features (Max_Shadow, etc.). Notably, while some shadow features occasionally achieved higher mean importance in the summary plot (**Figure 12a**), the selected features demonstrated lower variance and stable superiority in per-iteration comparisons, justifying their retention.

The performance of the CatBoosRegressor model trained using 7 features selected by Boruta-SHAP is shown in **Figure 12b**. The model achieved an R^2 of -0.08 and a MAE of 10.88 years. The SHAP values for these 7 features are displayed in **Figure 12c**, illustrating their relative importance and direction of effect. Although Boruta-SHAP identified a smaller subset of features compared to previous methods, the model's performance (MAE=10.88) is comparable to that achieved with the top 100 SHAP features from XGBoost (MAE =10.78), suggesting that some of these features may be relevant for age prediction.

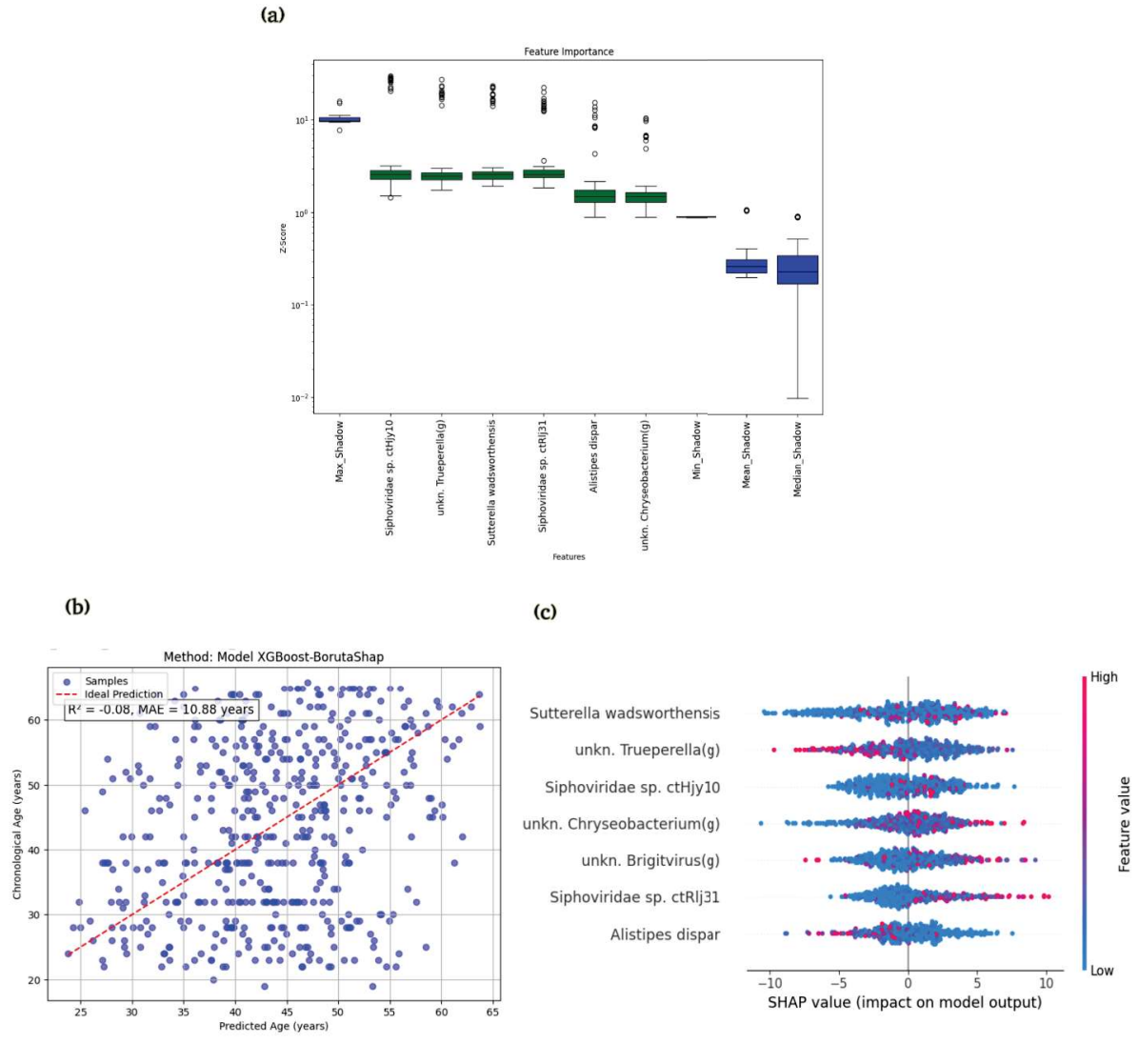


Figure 12. Boruta-SHAP Feature Selection Results

(a) Feature importance ranking as determined by Boruta-SHAP. The y-axis represents the Z-score of feature importance (logarithmic scale). Features with higher Z-scores are considered more important. Selected features have higher Z-scores than shadow features.

(b) Predicted age versus chronological age for the CatBoostRegressor model trained on the 7 features selected by Boruta-SHAP. The dashed red line represents the ideal prediction.

(c) SHAP (Shapley Additive exPlanations) values for the 7 most influential microbial features in the age prediction model. Features are ordered by decreasing importance. The color gradient represents feature values (blue: low, red: high), and the x-axis indicates the SHAP value, representing the feature's impact on the predicted age.

Method 4: Permutation Feature Importance results

The last feature selection method tested in our analysis to identify the most predictive microbial taxa in our dataset was Permutation Feature Importance (PFI), a model-agnostic approach that quantifies the contribution of each feature to the model's performance. As described in the Methods section, this method involves shuffling the values of a given feature and measuring the resulting increase in prediction error (Mean Absolute Error, MAE). Features with higher importance scores indicate a greater impact on model performance. We performed 50 permutations per feature to ensure robust estimates. The PFI scores for all features were calculated, and only features with a PFI score greater than 0.03 were considered for selection. Among these, the top 22 features with the highest PFI scores were selected for further analysis (**Figure 13**). This number was determined after multiple rounds of testing and evaluation, where we observed that selecting the top 22 features consistently yielded the lowest MAE in the test set indicating the best model performance compared to other feature set sizes. **Table 6** compares the performance metrics (R^2 , MAE, and MSE) of the seven regression algorithms trained using these 22 PFI-selected features. The model's predictions, using CatBoostRegressor, are shown in relation to chronological age in **Figure 14a**. The SHAP values for the top 20 most influential features are shown in **Figure 14b**.

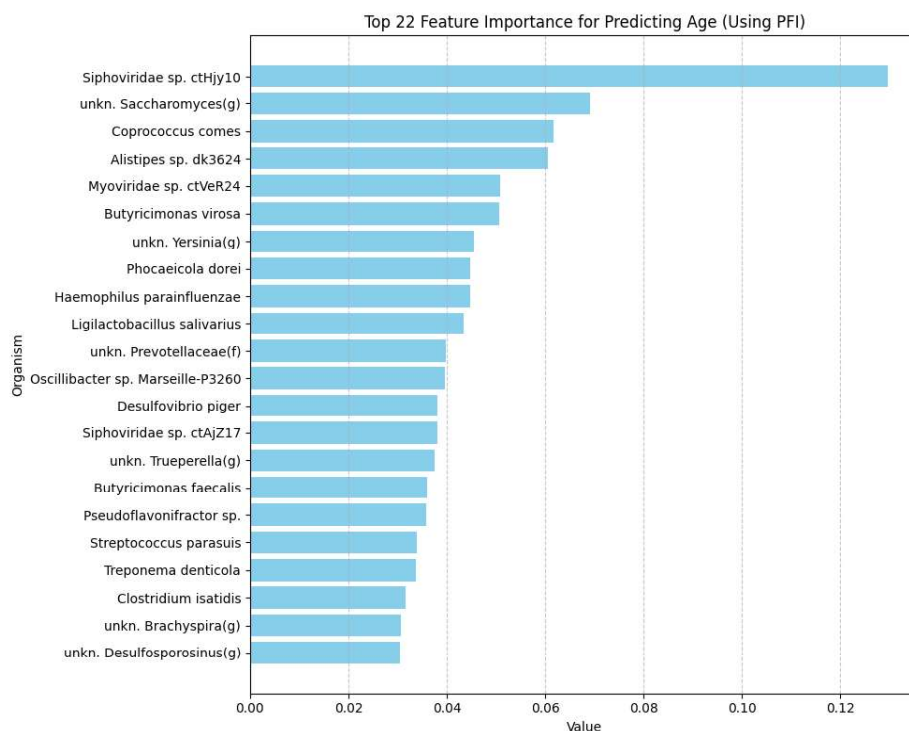


Figure 13. Permutation Feature Importance for the Top 22 Microbial Taxa

Bar plot showing the Permutation Feature Importance (PFI) scores for the 22 most important microbial taxa in predicting age. PFI scores represent the increase in prediction error (MAE) when the values of each feature are randomly permuted; higher values indicate greater importance.

Figure 13 displays the results of the Permutation Feature Importance (PFI) analysis, showing the 22 microbial taxa that were found to be most predictive of age in this dataset. The x-axis represents the

PFI score. This score indicates how much the model's prediction error (in MAE) increased when the values for that taxon are randomly shuffled. A larger PFI score means that the taxon has a greater influence on the model's ability to accurately predict age. The y-axis lists the 22 microbial taxa. These are ordered from the highest PFI score (most important) at the top to the lowest PFI score (least important) at the bottom. Finally, the bar length represents the PFI score for that taxon. Longer bars indicate higher importance.

From the figure we can observe that the taxon with the highest PFI score, and therefore the greatest importance for age prediction, is *Siphoviridae sp. ctHjy10*. Several other taxa also exhibit relatively high PFI scores, suggesting that contribute substantially to the model's predictive performance. These include *unkn. Saccharomyces(g)*, *Coprococcus comes*, and *Alistipes sp. Dk3624*. The remaining taxa in the plot have progressively lower PFI scores, indicating that they are less influential in predicting age compared to other taxa.

Table 6: Comparison of Regression Algorithms Performance Metrics

Model	CV R ²		CV MSE		CV MAE		Test R ²	Test MSE	Test MAE
	Mean	Std	Mean	Std	Mean	Std			
Random Forest	0.148	0.090	141.97	13.53	10.161	0.717	0.142	167.51	10.983
Linear Regression	-0.129	0.445	186.04	58.81	11.215	0.904	0.011	193.16	11.840
SVR	0.041	0.046	160.96	18.93	11.069	0.852	0.039	187.68	12.034
Gradient Boosting	0.128	0.124	144.03	11.56	10.005	0.377	0.068	182.00	11.187
XGBoost	0.051	0.137	157.75	18.85	10.410	0.708	0.099	176.00	11.023
CatBoostRegressor	0.148	0.117	141.46	11.89	9.994	0.455	0.218	152.76	10.189
LightGBM	0.118	0.176	145.53	18.22	9.898	0.633	0.135	168.87	10.737

Note: CatBoostRegressor (highlighted) shows overall strong performance, particularly in Test MAE (10.189). Cross-validation (CV) metrics include mean and standard deviation (Std) values.

Table 6 presents the comparison of the seven regression algorithms trained using the top 22 features selected by PFI. The table shows both cross-validation (CV) performance on the training data and performance in the held-out test set. Among algorithms, CatBoostRegressor demonstrated the strongest performance, achieving the highest test set R² (0.218) and the lowest test set MAE (10.189). This MAE of 10.189 represents the best performance achieved by any of the feature selection methods so far.

The performance of the CatBoostRegressor model trained with the PFI-selected features is further detailed in **Figure 14a**, which shows the predicted age versus chronological age. The model's prediction shows a moderate correlation with actual age ($R^2 = 0.15$) and a mean absolute error of 9.99 years. The SHAP values for the top 20 most influential features in this model are shown in **Figure 14b**. In the SHAP plot we can see that a high abundance of *Coprococcus comes* and *Siphoviridae sp. ctAjZ17* is associated with increased predicted age.

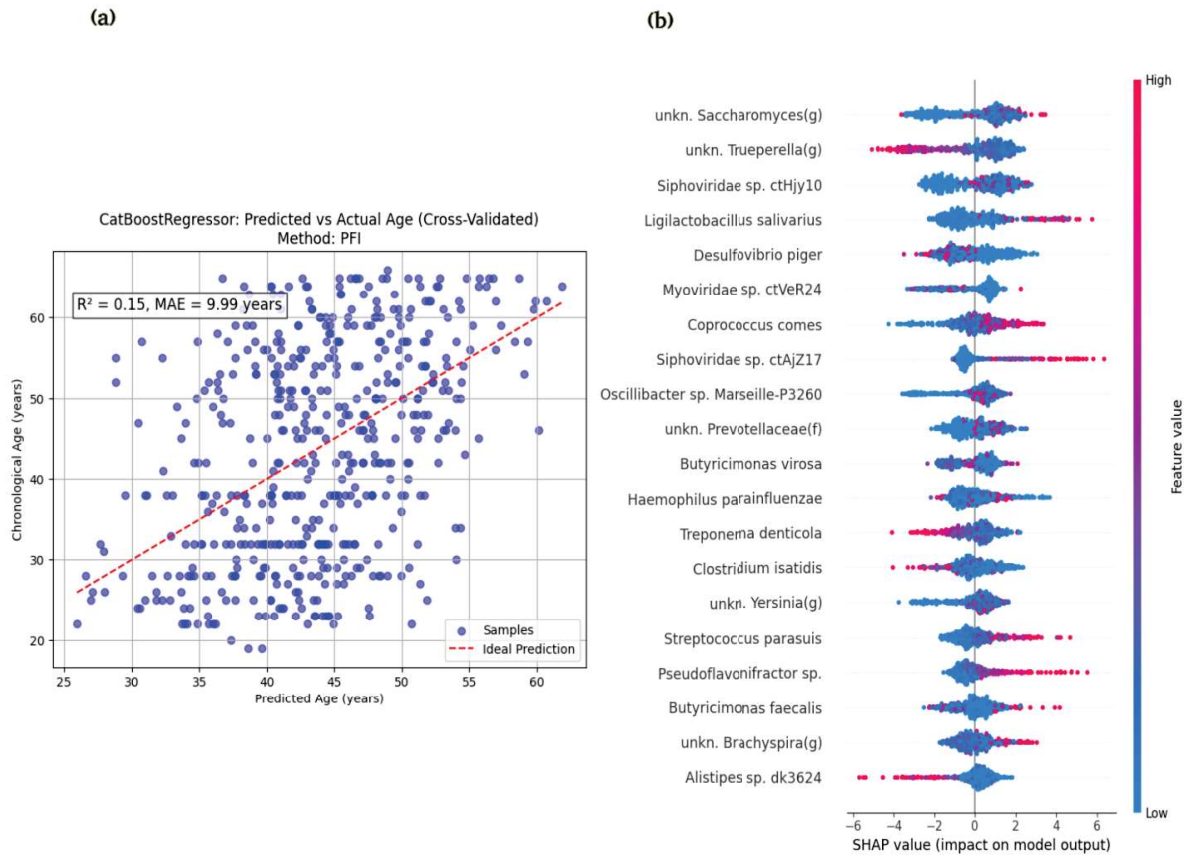


Figure 14. CatBoostRegressor Performance and SHAP analysis

(a) Predicted age versus chronological age for the CatBoostRegressor model trained on the top 22 features selected by Permutation Feature Importance (PFI). The dashed red line represents the ideal prediction. The model exhibits a moderate correlation between predicted and actual age ($R^2 = 0.15$) with a mean absolute error (MAE) of 9.99 years.

(b) SHAP (SHapley Additive exPlanations) values for the 20 most influential microbial features in the age prediction model. Features are ordered by decreasing importance. The color gradient represents feature values (blue: low, red: high), and the x-axis indicates the SHAP value, representing the feature's impact on the predicted age.

5.2.3. Summary of SHAP-based feature selection Insights

The SHAP (SHapley Additive exPlanations) analysis across four feature selection methods identified several robust microbial features associated with host age (**Table 7**). Notably, *Trueperella* (unknown species) was consistently selected by all four methods, exhibiting predominantly negative SHAP values for higher feature values (i.e., red points on the negative SHAP axis), suggesting a negative association with increasing age. *Siphoviridae* sp. ctHjy10 appeared in three out of four methods and showed a bimodal contribution, with both low (blue) and high (red) feature values associated with either negative or positive SHAP values, indicating a context-dependent influence on age prediction. *Saccharomyces* (unknown species), found in two methods, also displayed both positive and negative SHAP values for low feature levels (blue), implying variable effects on the model output.

Coprococcus comes, also present in two methods, showed positive SHAP values for high abundance (red) and negative values for low abundance (blue), indicating a potential positive link with age. *Siphoviridae* sp. ctAjZ17, identified in two methods, was associated with positive SHAP values at higher feature levels (red). *Butyricimonas virosa*, selected by two methods, displayed a symmetric contribution, with both low and high values (blue) linked to SHAP values across the positive and negative axes, suggesting a more nuanced role. Finally, *Haemophilus parainfluenzae*, although less consistently selected, contributed positively to the model at lower feature levels (blue).

This information is presented in **Table 7** as a summary of the most relevant features detected in the feature selection step.

Table 7. Summary of SHAP contributions for top microbial features across four feature selection methods

Feature	# Methods Selected	SHAP Contribution Pattern	Interpretation
<i>Trueperella</i> (unkn. sp.)	4	Red dots on negative SHAP axis	Strong negative association with host age
<i>Siphoviridae</i> sp. ctHjy10	3	Blue dots on both positive and negative SHAP axes	Context-dependent or nonlinear effect on age prediction
<i>Saccharomyces</i> (unkn. sp.)	2	Blue dots on both positive and negative SHAP axes	Variable influence, possibly feature interaction-dependent
<i>Coprococcus comes</i>	2	Red dots on positive , blue on negative SHAP axes	Positive association with age when abundant
<i>Siphoviridae</i> sp. ctAjZ17	2	Red dots on positive SHAP axis	Positive contribution at higher abundance levels
<i>Butyricimonas virosa</i>	2	Blue dots on both positive and negative SHAP axes	Ambiguous or dual-direction effect on model output
<i>Haemophilus parainfluenzae</i>	1	Blue dots on positive SHAP axis	Modest positive effect, especially at low abundance

5.2.4. Interpretation of final model SHAP results

After reviewing the generally important species across the different feature selection methods, we now turn our attention to understanding the final model's prediction based on SHAP. **Figure 15** presents the SHAP plot, offering insights into how the individual microbial features contribute to the age prediction of our best-performing model. Y-axis indicates the feature names in order of importance from top to bottom. X-axis represents the SHAP value for each feature and each sample, indicating how much that feature contributed to pushing the model output above or below the baseline prediction. Red dots represent a high abundance of the organism, while blue dots indicate a low abundance.

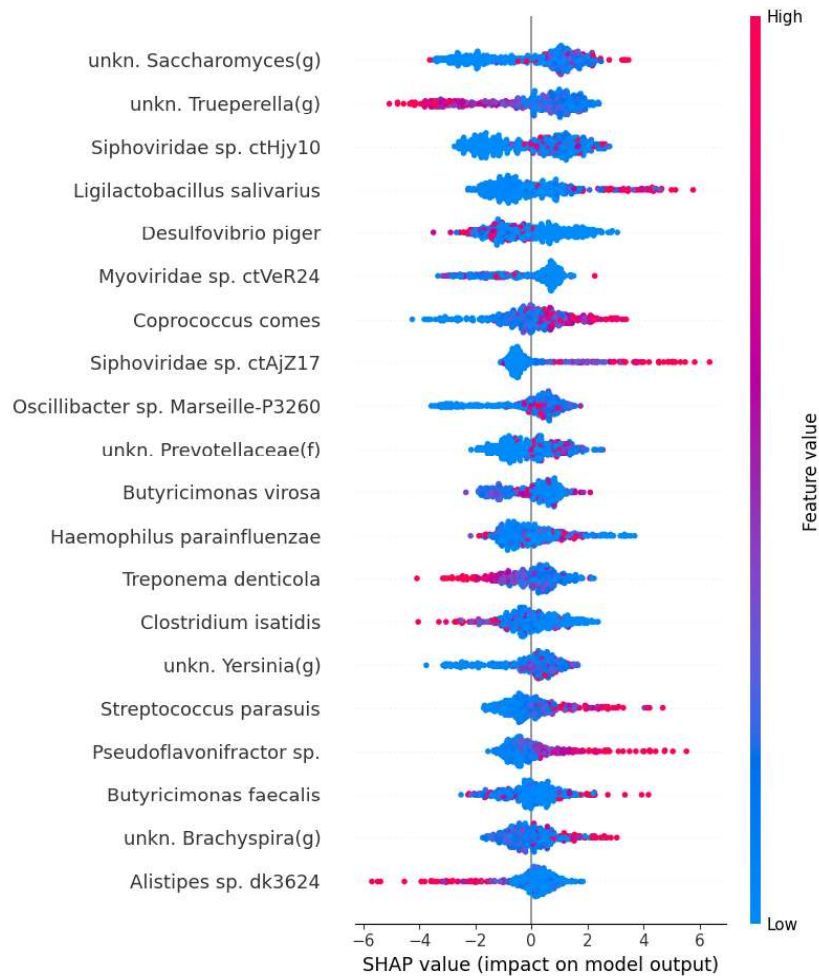


Figure 15. SHAP summary plot showing the impact of gut microbiome taxa on the prediction of chronological age

Each point represents a sample, where the x-axis denotes the SHAP value indicating the magnitude and direction of a feature's contribution to the model's predicted age. Taxa (features) are ordered by their overall importance (mean absolute SHAP value). The color gradient reflects the relative abundance of each taxon across samples, with red indicating higher abundance and blue indicating lower abundance. Positive SHAP values (right of zero) suggest that the presence of that taxon is associated with higher predicted age, while negative values suggest an association with younger predicted age.

The SHAP plot reveals distinct patterns in how different organisms influence the model's output. For example, the clustering of red dots for *Ligilactobacillus salivarius*, *Coprococcus comes*, *Siphoviridae* sp. ctAjZ17, *Streptococcus parasuis*, *Pseudoflavonifractor* sp., *Butyricimonas faecalis*, and an unknown species from the genus *Brachyspira* on the positive side of the SHAP value axis suggests that the presence or high abundance of these organisms generally contributes to an increased predicted age. Conversely, organisms such as *Alistipes* sp. Dk3624 and *Trueperella* (unknown species) exhibit a concentration of red dots on the negative side, indicating that their presence or high abundance is typically associated with a lower predicted age. These insights into the specific contributions of individual organisms provide valuable information for understanding the biological underpinnings of age prediction based on our model.

To investigate whether the most informative features in our regression model exhibited coherent age-related trends, we employed the `degPatterns` function from the `DEGreport` R package (v1.40.1). This tool enables the clustering and visualization of feature trajectories across a continuous

variable—in this case, age. We applied it to the Centered Log-Ratio (CLR) transformed abundances of the 22 microbial features from our best-performing model.

As part of the degPatterns workflow, the abundance values are internally standardized using the `scale()` function in R. This transformation converts each feature's CLR-transformed abundance into Z-scores, meaning values are centered to the mean and scaled by the standard deviation across samples for each feature. Thus, the y-axis in the resulting plot reflects standardized (Z-score) CLR abundances, enabling more direct comparisons of temporal trends across features with different abundance scales.

The analysis revealed a cluster of species whose abundance consistently increased with age (**Figure 16**), including *Siphoviridae* sp. *ctHjy10*, unclassified *Saccharomyces* (g), unclassified *Yersinia* (g), unclassified *Prevotellaceae* (f), *Oscillibacter* sp. *Marseille-P3260*, and *Streptococcus parasuis*. Notably, three of these species—unclassified *Yersinia* (g), *Oscillibacter* sp. *Marseille-P3260*, and *Streptococcus parasuis*—were also informative in the SHAP analysis. For *Yersinia* and *Oscillibacter*, lower abundance (blue dots) was associated with negative SHAP values, suggesting that their presence at lower levels predicts younger age. In contrast, *Streptococcus parasuis* showed red dots on the positive SHAP axis, indicating a direct positive contribution to the prediction of older age when abundant. The concordance between the degPatterns trajectory and the SHAP-based interpretation provides additional support for their biological interpretability of our model's predictions.

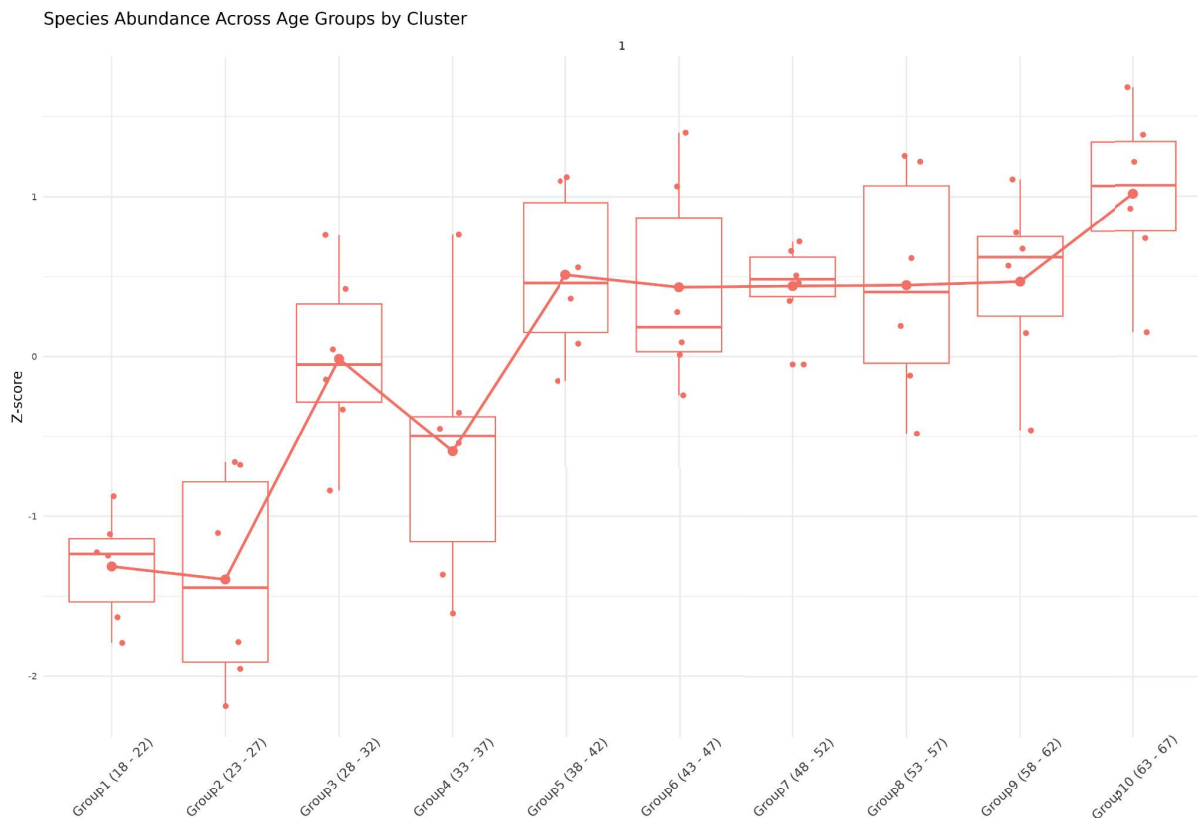


Figure 16. Age-related trajectory of microbial abundance for a representative cluster of species from the *degPatterns* analysis.

The plot displays Z-score-scaled CLR abundances of microbial features across age-stratified sample groups. The increasing trend indicates a cluster of taxa whose abundance rises with host age. Boxplots represent the distribution within each age group, and the overlaid line indicates the cluster average trajectory.

5.2.5. Hyperparameter optimization

The best-performing regression model in our analysis was a CatBoostRegressor configured with 200 iterations, a learning rate of 0.1, and a tree depth of 6. This configuration yielded the highest predictive accuracy among all models evaluated.

5.3. Final model selection and performance

In the final stage of evaluating our best-performing model, we examined the distribution of its prediction errors across different age groups within the held-out test set. **Figure 17a** presents a boxplot illustrating the absolute error, defined as the absolute difference between the true age and the age predicted by the model, for five distinct age categories (18-30, 31-40, 41-50, 51-60, and 61-66). Final model prediction metrics can be found in the Excel named “*Predictions_test_set.xlsx*”.

A notable trend observed in the boxplots is that the median absolute error appears to be higher for the youngest age group (18 - 30 years) at 10.64 years and the oldest age group (61 - 66 years) at 20.79 years. The model demonstrates improved accuracy in the middle age ranges, with the lowest median absolute error observed in the 41 - 50 year group (3.86 years), followed by the 31- 40 year group (6.39 years) and the 51 - 60 year group (8.95 years). For a detailed view of the individual predictions on the test set, including the true ages and the corresponding model predictions, please refer to the Excel named “*Predictions_test_set.xlsx*” in Supplementary Material. For context, we also calculated the median absolute error across the entire test set, which was found to be **9.17 years (Figure 17b)**.

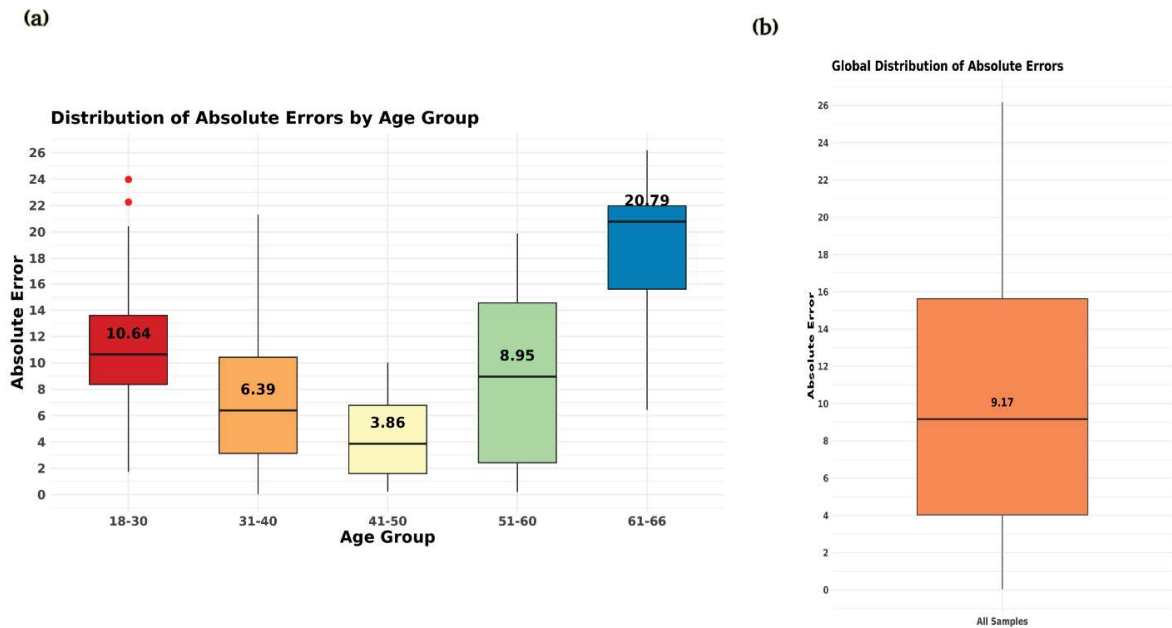


Figure 17. *Distribution of Absolute Errors*

(a) Distribution of absolute errors by age group, with boxplots for each of the five age categories (18-30, 31-40, 41-50, 51-60, and 61-66). The median absolute error for each group is displayed above the corresponding box.

(b) Global distribution of absolute errors, represented by a single boxplot showing the median absolute error across all samples, which is labeled as 9.17.

Finally, to evaluate the performance of our final model on unseen data, we applied it to an independent external dataset. The pdf file named “External_Validation_Metrics.pdf” in Supplementary Materials presents the result of this validation, displaying the true age (‘Age’ - blue bars), the model’s predicted age (‘Predicted_Age’ - red bars), and the absolute error (‘MAE’ -yellow bars) for each sample in the independent dataset, identified by their ‘Sample_Id’. For some samples, the predicted age closely aligns with the true age, resulting in a smaller absolute error (e.g., Sample_Id C with MAE 0.4 years). However for other samples, there are more substantial discrepancies between the predicted and true ages, leading to larger absolute errors (e.g., Sample_Id B with MAE 12 years). The average Mean Absolute Error (MAE) on this independent dataset was calculated to be **9.425 years**, calculated for individuals within the 19 to 66 years age range—consistent with the training set. This ensures that the performance metric reflects the model’s predictive ability within the domain it was optimized for.

When comparing the error distribution to that observed during internal testing, we found a similar trend: the MAE was lower in individuals aged approximately 39 to 48 years, while higher errors were observed in both younger and older individuals. This suggests that the model performs best within the central age range and tends to be less accurate at the boundaries of the age distribution, likely due to reduced representation or increased variability in those age groups.

5.4. Identification of key features driving cluster separation

To explore age-related structure within the samples based on the selected microbiome features, we first applied KMeans clustering. The optimal number of clusters was determined using the Silhouette method, which indicated the presence of 5 distinct clusters (**Figure 19**).

To understand which microbial features were most influential in defining the observed clusters, we employed two complementary approaches: SHAP feature importance and cluster centroid analysis.

5.4.1. SHAP analysis for feature importance

First, K-means clustering was applied to the abundance feature data to identify five distinct sample clusters (0 to 4). These clusters were then used as labels for training an XGBoost classification model, which aimed to predict cluster assignments based on 22 selected features. SHAP (Shapley Additive Explanations) analysis was subsequently performed to assess the importance of each feature in predicting the cluster assignments. The overall feature importance rankings, based on SHAP values, are presented in **Figure 18**, together with the individual predictions for each cluster. This analysis provides insights into how specific features drive the classification towards particular clusters. For example, *Coprococcus comes* was identified by SHAP as the most significant feature overall, particularly for distinguishing Cluster 3 (**Figure 18**, top panel). Examining the detailed SHAP beeswarm plot specifically for Cluster 3 (**Figure 18**, bottom right panel) confirms this observation. It shows that higher abundance values for *Coprococcus comes* (represented by red points) correspond to strongly positive SHAP values (points shifted to the right). This indicates that increased levels of this species are a key driver pushing the model to classify samples into Cluster 3.

Haemophilus parainfluenzae (second ranked) shows importance for cluster 0. High abundance (red points) strongly pushes the predictions towards this cluster (positive SHAP values). Additionally, for Cluster 0, species like *Siphoviridae* sp. ctHjy10 and *Myoviridae* sp. ctVeR24 also appears to drive predictions towards this cluster.

A similar pattern is observed for *Desulfovibrio piger* (third ranked), *unkn. Prevotellaceae(f)* and *Butyricimonas faecalis*. All three species show particular importance for Cluster 2, where high abundance corresponds to positive SHAP values.

Pseudoflavonifractor sp. (fourth ranked) primarily drives classification towards Cluster 4. Similar behaviour is observed for the species: *unkn. Desulfosporosinus(g)*, *Treponema denticola*, *unkn. Trueperella*, and *Oscillibacter* sp. Marseille-P3260. On the other hand, regarding Cluster 4, *Coprococcus comes* and *unkn. Prevotellaceae(f)* show the contrary behavior; high abundance (red points) corresponds to negative SHAP values, indicating that higher levels of these species push predictions away from this cluster.

5.4.2. Discriminative feature analysis via cluster centroids

In addition to SHAP-based feature importance, a discriminative feature analysis was performed by calculating the difference between the maximum and minimum mean abundance values of each

feature across the centroids of the five identified clusters. This approach helped identify features with the greatest variation in abundance between clusters, and the features were ranked based on this maximum difference. A comparison of the top features identified by both SHAP analysis and the cluster centroid difference method is provided in **Table 8**. Based on their high ranking in both analyses, indicating significant contribution to the cluster separation, five species – *Coprococcus comes*, *Haemophilus parainfluenzae*, *unkn. Prevotellaceae(f)*, *Desulfovibrio piger*, and *Pseudoflavonifractor sp.* – were selected for further investigation regarding their association with host age.

Table 8: Comparison of Top Discriminative Features from UMAP and SHAP Analysis

Top UMAP discriminative features	Top SHAP important features
<i>Coprococcus comes</i>	<i>Coprococcus comes</i>
<i>Haemophilus parainfluenzae</i>	<i>Haemophilus parainfluenzae</i>
unkn. Prevotellaceae(f)	<i>Desulfovibrio piger</i>
<i>Desulfovibrio piger</i>	<i>Pseudoflavonifractor sp.</i>
<i>Pseudoflavonifractor sp.</i>	unkn. Prevotellaceae(f)
<i>Myoviridae sp. ctVeR24</i>	unkn. <i>Desulfosporosinus(g)</i>
<i>Treponema denticola</i>	<i>Oscillibacter sp. Marseille-P3260</i>
unkn. <i>Desulfosporosinus(g)</i>	<i>Treponema denticola</i>
unkn. <i>Trueperella(g)</i>	<i>Syphoviridae sp. ctHjy10</i>
<i>Butyricimonas faecalis</i>	<i>Butyricimonas faecalis</i>

Note: The five species in bold (*Coprococcus comes*, *Haemophilus parainfluenzae*, unkn.

Prevotellaceae(f), *Desulfovibrio piger*, and *Pseudoflavonifractor sp.*) are the most discriminative features appearing in both analyses, with the first two being top-ranked in both lists.

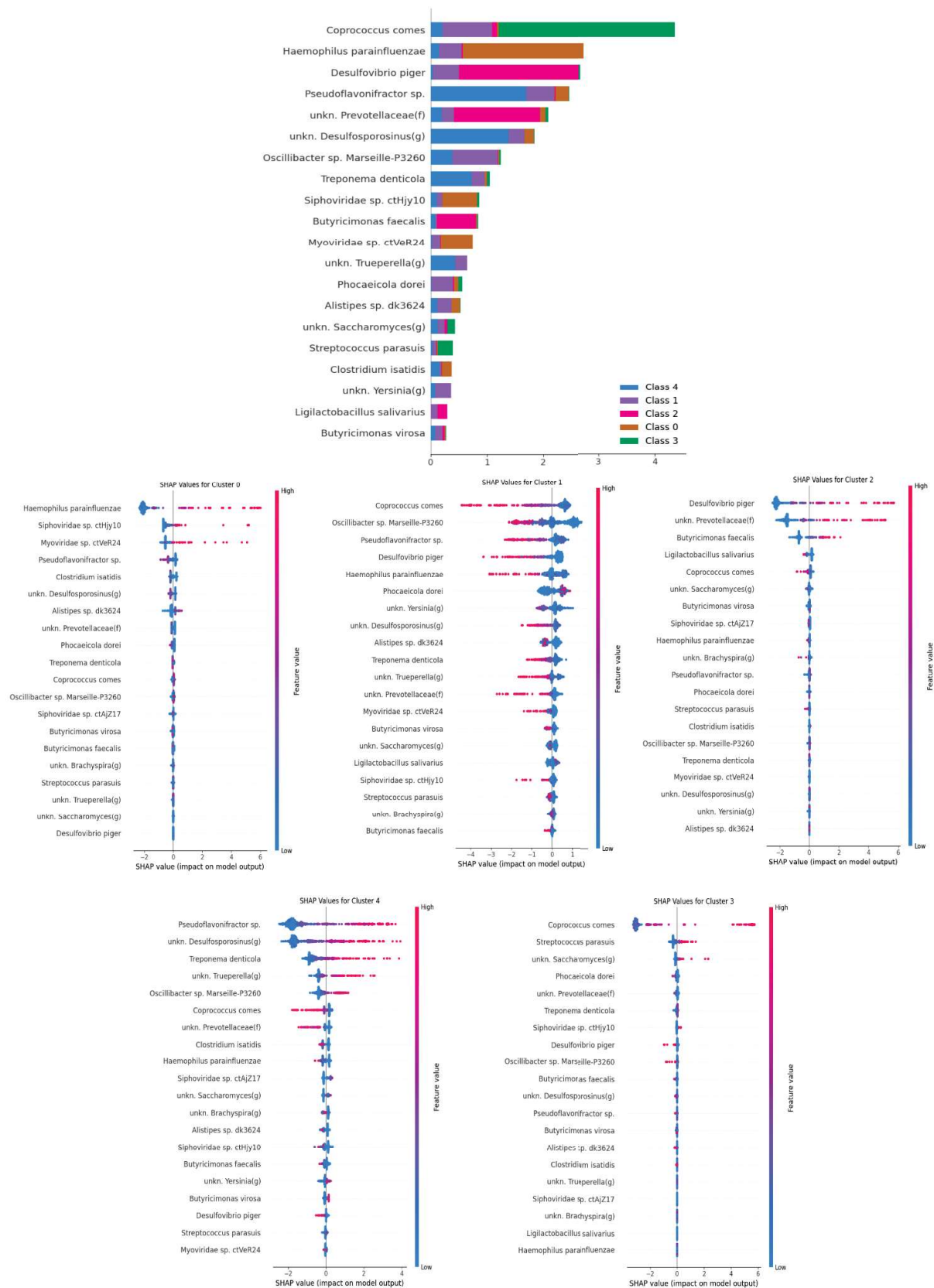


Figure 18. SHAP Analysis and Feature Importance Visualization for Cluster Assignment

The upper part of the figure shows a barplot of the initial predictions from the SHAP analysis, highlighting the most influential features in the cluster assignment. Below, individual beeswarm plots for each of the five identified clusters display the most important features driven the clusterization.

5.4.3. UMAP visualization of top features

To further investigate the features that most significantly contribute to the observed clustering patterns, we visualized the spatial distribution and relative abundance of these top 5 features identified by both approaches (*Coprococcus comes*, *Haemophilus parainfluenzae*, *unkn. Prevotellaceae(f)*, *Desulfovibrio piger*, and *Pseudoflavonifractor sp.*). Their abundance values were projected onto the UMAP embedding, as shown in **Figure 19**. Each panel in the figure displays the same UMAP projection of the samples, but the points are colored based on the relative abundance of one specific feature; ranging from low (blue) to high (red) abundance. This visualization allows for the inspection of spatial patterns, revealing whether samples with high or low abundance of a particular organism tend to co-localize within specific regions in the UMAP space, linking their associating with the underlying sample clusters.

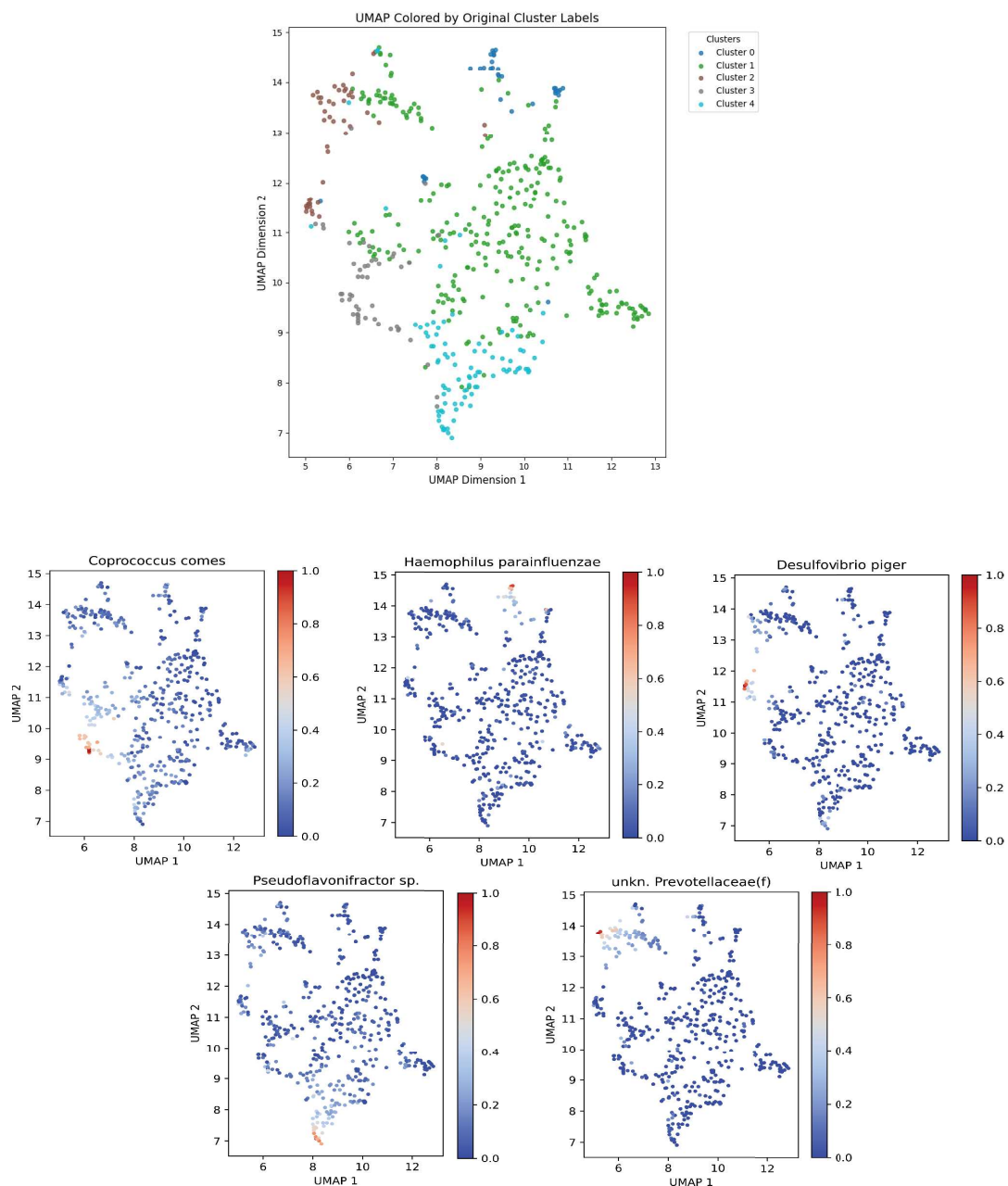


Figure 19. UMAP plots showing the normalized abundance/value of each of the top 5 SHAP features, together with the UMAP colored by cluster labels.

5.5. Age-related abundance patterns of key discriminative species

To investigate potential links between the identified microbial clusters and host age, we examined the relative abundance of the five most discriminative species —those ranked highest in both SHAP importance and cluster centroid difference—across three distinct age groups: 'Young' (19-34 years), 'Mid' (35-49 years), and 'Old' (50-66 years). Statistical significance of differences across the three groups was assessed using the Kruskal-Wallis test, followed by Dunn's post-hoc test with Bonferroni correction when the test indicated a significant overall difference ($p < 0.05$).

Of the five species examined, three showed statistically significant differences in relative abundance across the age groups: *Coprococcus comes*, unkn. Prevotellaceae(f), and *Desulfovibrio piger*. Violin plots illustrating the distribution of relative abundance for these three species are presented in **Figure 20**.

- For *Coprococcus comes*, a significant difference in abundance was observed across age groups (Kruskal-Wallis $p = 0.012$), with the 'Young' group showing a lower mean relative abundance (mean = 0.12) compared to the 'Mid' (mean = 0.16) and 'Old' (mean = 0.16) groups. Post-hoc analysis revealed significant differences between the 'Young' vs 'Mid' and 'Young vs 'Old'.
- For the unclassified family, unkn. Prevotellaceae(f), also showed a highly significant difference across age groups (Kruskal-Wallis $p = 3.17e-04$). Similar to *C. comes*, Dunn's test revealed significantly lower abundance in the 'Young' group compared to both 'Mid' ($p < 0.05$) and 'Old' groups ($p < 0.05$). Mean abundances were approximately 0.04, 0.06, and 0.06.
- *Desulfovibrio piger* was also found to be significantly different (Kruskal-Wallis $p = 0.031$). Post-hoc analysis indicated that the Mid group had significantly lower abundance compared specifically to the 'Old' group ($p < 0.05$), while other pairwise comparisons were not significant. Mean abundances were approximately 0.05, 0.04, and 0.05.

The remaining two species initially selected, *Haemophilus parainfluenzae* and *Pseudoflavonifractor sp.*, did not show statistically significant differences in abundance across the defined age groups (Kruskal-Wallis $p > 0.05$; data not shown).

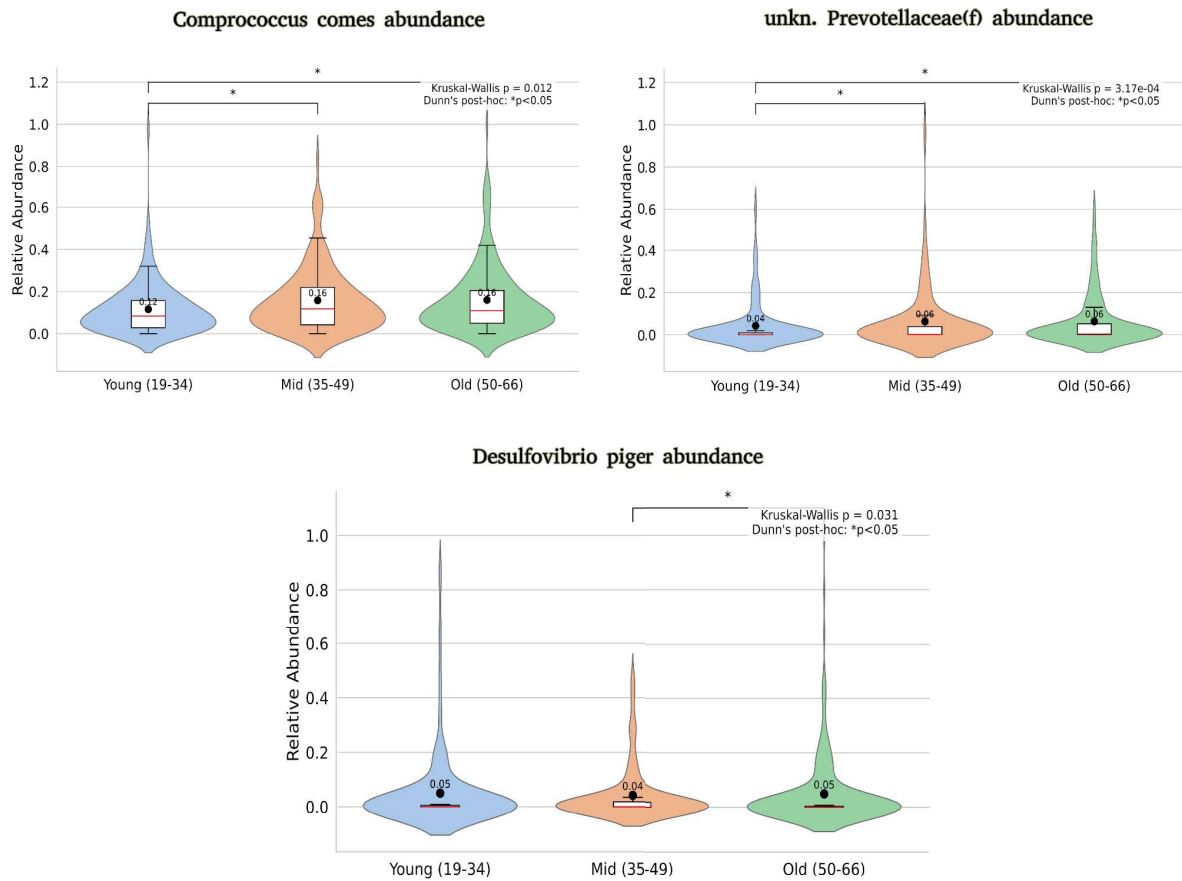


Figure 20. Age-related relative abundance of significantly different key species

Multi-panel figure showing violin plots for the 3 species with significant Kruskal-Wallis results: *Coprococcus comes* ($p=0.012$), unkn. Prevotellaceae(f) ($p=3.17e-04$), and *Desulfovibrio piger* ($p=0.031$).

5.6. Implementation

This thesis required extensive coding, with 6 Jupyter notebooks scripts and 3 R scripts. It is important to note that the code developed during this project cannot be shared publicly, as it has become the property of the collaborating company. Therefore, no link to the repository or code files is provided.

The description of each code file is detailed below:

R Scripts

- **select_samples.R**

This script selects a subset of the samples used in the analysis and includes visualizations of the associated metadata for initial inspection.

- **Preprocessing_report_WGS.qmd**
This Quarto file contains exploratory metagenomic analyses of the selected samples and serves as part of the exploratory data analysis (EDA).
- **degPatterns_analysis.R**
This script performs the age-related trajectory analysis using the degPatterns function from the DEGREport package.

Jupyter Notebooks

- **Feature_Selection_Methods.ipynb**
This notebook performs feature selection and visualization using four different methods to assess feature importance.
- **Evaluate_Feature_Prevalence.ipynb**
This notebook calculates the prevalence of the 22 microbial features included in the final model, offering a comprehensive view of how consistently these features appear across samples.
- **Validation_Test.ipynb**
This notebook validates the final model using external samples that were not included in the training or testing phases.
- **Clustering_optimalK.ipynb**
This script determines the optimal number of clusters for the UMAP-based clustering using methods like the Silhouette score.
- **Clustering.ipynb**
This notebook performs dimensionality reduction (UMAP) and clustering of the samples based on the selected microbial features.
- **Cluster_correlation.ipynb**
This script analyzes the five most discriminative bacterial species (identified via SHAP and centroid differences) and compares their abundance across three age groups: Young, Mid, and Old.

6. Discussion and conclusions

6.1. Discussion

6.1.1. Age-related differences in gut microbiota and the impact of diet and lifestyle

Our study revealed notable differences in the gut microbiota between young and old individuals. Specifically, microbiota richness was significantly higher in older adults compared to younger individuals ($p = 0.004$) (**Figure 8**). The richness of gut microbiota are acknowledged as a health biomarker, including the frailty of elderly people and some personality traits (K. V.-A. Johnson, 2020; Lei et al., 2019). Previous studies have shown that frail individuals tend to have lower microbiota richness (Jackson et al., 2016; Maffei et al., 2017).

Additionally, the *Prevotella* enterotype was predominant in oldest-old people, while the *Bacteroides* enterotype was predominant in young individuals (**Figure 7**). Enterotypes are widely recognized as important biomarkers for both human health and behaviour (Heiman & Greenway, 2016; K. V.-A. Johnson, 2020).

These findings are consistent with previous research comparing the gut microbiota of centenarians and younger populations. A fiber-rich, high-carbohydrate diet has been shown to correlate with a prevalence of *Prevotella* (Selber-Hnatiw et al., 2020). In contrast, the *Bacteroides* enterotype, along with reduced microbiota richness, is often associated with the modern Western lifestyle (Lin et al., 2013; Mosca et al., 2016).

Notably, our study also revealed that, in older individuals, butyrate-producing bacteria from the family *Lachnospiraceae* were found in significantly higher abundance than in the younger population. This is particularly noteworthy, as *Lachnospiraceae* bacteria play a crucial role in gut health through their production of short-chain fatty acids (SCFAs), such as butyrate. Butyrate is known for its anti-inflammatory properties and its positive impact on gut barrier function (Louis & Flint, 2017; Riedel et al., 2006; Vacca et al., 2020). The higher abundance of these bacteria in the older population suggests that the aging microbiome may be more efficient at producing butyrate, a compound that has been linked to multiple health benefits, including neuroprotective effects that influence brain functions. Notably, evidence suggests that butyrate could help mitigate neuropathological processes related to Alzheimer's disease (AD) pathogenesis (Chilton et al., 2024).

Over the past 50 years, dietary diversity has significantly decreased, and diets that exclude animal or plant-based foods may further limit microbiome diversity (Heiman & Greenway, 2016). The eating habits of the oldest-old population differed markedly from those of younger individuals, with a higher intake of potatoes and cereal products—foods rich in starch and fiber. These nutrients are known to have beneficial effects on gut microbiota and their metabolites (Bibi et al., 2019).

Furthermore, the past few decades have also seen a sharp rise in the incidence of various intestinal and extraintestinal diseases, including obesity, diabetes, and metabolic disorders. Initially observed in Western countries, this trend has recently spread to developing nations, where changes in childhood environmental and dietary conditions have occurred rapidly (Mosca et al., 2016).

6.1.2. Regression analysis of gut microbiota as a predictor of chronological age

In this study, we developed a machine learning model to predict chronological age using gut microbiome compositions in healthy individuals. Our approach leveraged species-level abundance profiles to train regression models and identify microbial taxa associated with aging.

We implemented four distinct feature selection strategies, including an innovative hybrid approach that combines SHAP (SHapley Additive exPlanations) values with traditional feature importance scores across multiple models. By integrating these metrics via rank-based method, we prioritise features that were consistently informative across both interpretability and predictive performance perspectives, thereby capturing robust, biologically relevant signatures. Model interpretability was further reinforced using SHAP to quantify species-specific contributions to predictions.

Among the four feature selection strategies tested, Permutation Feature Importance (PFI) yielded the best predictive performance. This approach directly quantifies the contribution of each feature by measuring the change in model error after randomly shuffling feature values, thereby evaluating how much the model relies on each feature to make accurate predictions. Unlike methods such as SelectKBest or traditional importance metrics—which may rank features based on univariate relationships or model specific-heuristics—PFI accounts for the actual impact of a feature within the full multivariate context of the model. Moreover, compared to SHAP-based and hybrid SHAP-Boruta methods, PFI is less sensitive to potential overfitting caused by interpreting internal model mechanisms, particularly in high dimensional datasets like microbiome profiles. Although SHAP provides valuable interpretability, its performance in feature selection can be influenced by redundancy among features or model-specific biases. In contrast, PFI more robustly identifies features whose perturbation meaningfully degrades predictive performance, likely making it more suitable for identifying microbial taxa with the strongest generalizable associations to age.

The optimal model—combining permutation feature importance with a CatBoostregressor model—achieved strong predictive performance, with a mean absolute error (MAE) of 9.99 years in cross-validated training, 9.17 years on the test set, and 9.43 years on an independent validation cohort. Although these values are slightly higher than those reported in some previous studies, it is important to note several methodological differences that may account for this discrepancy. Many state-of-the-art models have achieved lower MAEs by focusing on a broader age range (e.g., 20-80+ years) where extreme-age comparisons may amplify apparent predictive accuracy, our focused range (19-66 years) targets adults microbiome's core aging trajectory, yet excludes older adults (>66 years) where age-related microbiome shifts may become more pronounced. Additionally, many of the best performing models in the literature are based on single-cohort or single-country datasets, reducing population-level variability. Our dataset, by contrast, includes samples from multiple studies and diverse geographic origins, increasing biological heterogeneity. Although this may raise prediction error, it also improves the generability and external validity of the findings. Lastly, while some models incorporate additional host variables such as diet, BMI, or lifestyle, our approach relies exclusively on microbiome composition. This choice enables to isolate microbial contributions to aging, but may also constrain predictive power compared to multimodal models.

6.1.3. Biological interpretation of predictive features

The interpretation of the predictive features identified by our regression model is crucial to understanding the biological relevance of age-associated shifts in the gut microbiome. Using SHAP values, we quantified the individual contribution of each microbial species to the model's age predictions. Among the 22 most important features for the regression model, several microbial taxa stood out as key predictors of chronological age.

Discussion on *Coprococcus* as a predictive feature in age-associated gut microbiota shifts

Coprococcus, a genus known for its significant role in gut health, was one of the most notable predictors in our model. *Coprococcus* is involved in various metabolic processes, including the fermentation of carbohydrates and the production of short-chain fatty acids (SCFAs) like butyrate, which are essential for maintaining gut homeostasis (Yu et al., 2024). It also exhibits anti-inflammatory properties and supports immune system regulation (Sandrini et al., 2024). Typically abundant in the gut microbiota of healthy individuals, *Coprococcus* correlates with a balanced gut ecosystem (Jiang et al., 2024), while its dysregulation has been linked to diseases like small intestinal bacterial overgrowth (SIBO) and cow milk protein allergy (J. Li et al., 2024; J. Xu et al., 2025). Furthermore, *Coprococcus* has been proposed as a potential biomarker due to its inverse correlation with neuropsychological and neurodegenerative disorders (Notting et al., 2023).

Developing dynamics and core prevalence of *Coprococcus*

The acquisition of *Coprococcus* in the gut microbiome follows a distinct trajectory from infancy to adulthood. Colonization begins early in life, with detectable levels in 87% of Canadian infants (3-4 months old) at a low relative abundance (~0.8%) (Azad et al., 2013). However, its prevalence and abundance increase with age, as seen in a Chinese cohort where 78.2% of 3-year-olds harbored *Coprococcus* at a mean abundance of 0.21% (Niu et al., 2020).

Geographical variations in early microbiome composition may further shape *Coprococcus* colonization patterns. For instance, infants from Western countries typically exhibit an *Actinobacteria*-dominant (A-type) microbiota, whereas Chinese and Ugandan infants often display a *Proteobacteria*-dominant (P-type) profile (Kort et al., 2021; Kuang et al., 2016). Despite these differences, *C. eutactus* prevalence consistently rises with age, as demonstrated in Uganda children, increasing from 62% at 24 months to 81% by 36 months (Kort et al., 2021).

By adulthood, *Coprococcus* becomes a hallmark of a mature gut microbiome. It is part of an adult-associated co-abundance group (CAG) that includes *Lachnospiraceae*, *Blautia*, and *Faecalibacterium* (Odamaki et al., 2016). However, *Coprococcus* abundance declines in advanced age, showing negative correlation with aging and significantly lower levels in centenarians compared to younger elderly populations (Odamaki et al., 2016; F. Wang et al., 2015).

Age-associated abundance patterns in our model

In our model (trained on individuals aged 19-66 years), SHAP value analysis revealed that higher abundances of *Coprococcus* were associated with older age, as indicated by positive SHAP values (**Figure 15**). This suggests that *Coprococcus* may serve as a biomarker of microbiome maturation in adulthood, at least within this age range. While some previous studies associate *Coprococcus* with general health and microbiome stability (Notting et al., 2023; Shetty et al., 2017), our findings highlight its progressive enrichment across early to late adulthood in healthy individuals—a pattern distinct from the decline observed in advanced aging populations (e.g., centenarians; (F. Wang et al., 2015)).

We propose two key interpretations:

1. **Health vs. Aging dynamics:** In our cohort (healthy adults up to age 66), *Coprococcus* abundance may reflect sustained microbiome resilience rather than incipient dysbiosis. Its positive accumulation with age could signify cumulative colonization or dietary adaptation (e.g., fiber intake), whereas its decline in older populations (>80 years) might mark a true senescent-related shift.
2. **Species-specific trends:** Most studies linking *Coprococcus* to “youthfulness” do not distinguish between species. The strain driving our model’s prediction (e.g., *C. comes*) might have distinct age-related behaviour compared to health-associated species like *C. eutactus* (Kort et al., 2021).

An unexpected result in our model was the association of *Trueperella*, an opportunistic bacterial pathogen, with younger individuals. *Trueperella* is known to cause infectious diseases, and its prevalence is typically higher in older adults due to their greater susceptibility to antibiotic-resistant bacteria (Tamai et al., 2022). However, our SHAP-based interpretation indicated that higher abundances of *Trueperella* were more common in younger individuals, which contradicts prior assumptions. This paradox may be explained by environmental or dietary factors, as *Trueperella* is primarily an animal-associated microbe, with potential exposure from animal-derived foods or livestock contact (Hu et al., 2025). It is also possible that age-related differences in immune function, gut physiology, or microbiome resilience influence the colonization potential of *Trueperella*. The classification of this bacterium as an “unknown species” further underscores the complexity of the microbiome and the need for more detailed taxonomic and functional characterization.

Our analysis also highlighted the role of **phages** in regulating the gut bacterial community. Phages play a crucial role in bacterial evolution and diversity, influencing microbial populations by transferring genes between bacteria (Emencheta et al., 2023; Rybicka & Kaźmierczak, 2025). Certain phages, such as those in the **Siphoviridae** family, are associated with beneficial bacteria like *Bifidobacterium* (Gregory et al., 2020; Y. Zhang et al., 2023). In our SHAP analysis, the presence of *Siphoviridae* sp. ctAjZ17 was identified as a predictor of older age, reflecting potential age-related changes in the gut virome or its links to host health status.

In contrast, some taxa analyzed did not show clear associations with age. For example, *Haemophilus parainfluenzae* was primarily discussed in relation to diseases like Crohn's disease and obesity,

without a clear link to age-related patterns (Sohn et al., 2023). Similarly, *Ligilactobacillus salivarius* (formerly *Lactobacillus salivarius*) has been noted for its probiotic and immunomodulatory properties, but age-related trends in its natural occurrence remain unclear (Yang et al., 2024).

Oscillibacter species (including *O. sp. Marseille-P3260*) have been mainly studied in disease models or animal systems, limiting conclusions about their relevance to human aging. Finally, *Desulfovibrio* species, particularly *D. piger* and *D. fairfieldensis*, exhibited a complex profile. While generally considered pathobionts, these species were enriched in centenarians and associated with longevity-related traits, suggesting that their role in aging may be context-dependent (S. Chen et al., 2024; Y.-R. Chen et al., 2019; Singh et al., 2023).

Overall, these findings illustrate the dynamic nature of the gut microbiome in aging and highlight the complex interactions between microbiota and host health. While some taxa show clear associations with age, others reveal intriguing exceptions that warrant further investigation to better understand their role in the aging process.

6.1.4. Umap analysis interpretation

Further exploration of the selected features using UMAP (Uniform Manifold Approximation and Projection) revealed distinct patterns in microbial community structure across different age groups. The UMAP visualization, when coupled with our SHAP-based feature importance analysis, provides a deeper understanding of the biological aging process as reflected in the gut microbiome.

Our combined approach—utilizing SHAP values to assess feature importance, centroid-based discriminative analysis, and UMAP visualization—offered a comprehensive framework to interpret the key drivers of microbiome community structure. Notably, taxa such as *Coprococcus comes*, **unkn. Prevotellaceae(f)**, and *Desulfovibrio piger* not only aligned with clear UMAP-defined spatial patterns but also demonstrated significant associations with host age, suggesting that age-related ecological shifts in the gut microbiome may underlie part of the observed clustering. For instance, an increased abundance of *Coprococcus comes* was aligned with older ages, while an unclassified **Prevotellaceae** family member demonstrated a high abundance in older individuals, in line with findings from our SHAP and depth-pattern analyses. Conversely, *Desulfovibrio piger* displayed a low abundance in older individuals, suggesting that this pathobiont may have a context-dependent relationship with aging, potentially influenced by host health status or environmental factors.

Interestingly, *Haemophilus parainfluenzae* and *Pseudoflavonifractor sp.*, although they were highly discriminative of UMAP clusters, did not show statistically significant variation with age. This lack of age correlation, despite clear differences in clustering, highlights the importance of integrating both unsupervised and supervised approaches when interpreting microbial variation. These taxa may be influenced by additional host-related factors—such as mucosal immunity, recent infections, or antibiotic exposure—that were not accounted for in this study. Given that *Haemophilus parainfluenzae* has been discussed in the context of diseases like Crohn's disease and obesity (Sohn et al., 2023) and *Pseudoflavonifractor sp.* has been linked to immune modulation and microbial ecology, these factors may confound the clear age-related trends that we observed in other taxa.

By combining UMAP analysis with our prior SHAP-based interpretation, we can appreciate how these microbial features interact within a broader ecological context. The insights from both analyses reveal the complexity of the microbiome in aging, suggesting that age-related shifts in the gut microbiome may not be solely driven by chronological age, but also by a host of other factors like immune function, diet, and environmental exposures.

6.1.5. Limitations and Assumptions

This study offers valuable insights into the gut microbiome's role in aging but is subject to several limitations. First, the sample size and population diversity were limited, focusing on adults aged 19-66 years, which reduces generalizability to older populations. The cross-sectional design further limits our understanding of longitudinal microbiome changes, highlighting the need for future studies to adopt a lifespan perspective. Our findings specifically underscore the need for lifespan-wide studies to dissect *Coprococcus*'s age-dependent roles. Notably, the exclusion of individuals beyond age 66 precludes analysis of its trajectory in late adulthood—a period marked by accelerated microbiome dysbiosis.

Additionally, while our model prioritized microbiome composition, it excluded host factors like diet, BMI, and lifestyle, which could enhance predictive accuracy if included. The reliance on taxonomic classification also introduces some uncertainty due to incomplete identification of certain taxa, such as *Trueperella* and *Siphoviridae* sp. *ctAjZ17*.

Moreover, microbiome composition is highly variable, influenced by geographic and dietary factors, which may impact the model's robustness. Our focus on the adult's microbiome, excluding individuals over 66, may have overlooked important microbiome shifts. Lastly, while our data spans multiple regions, inherent variability between cohorts suggest that further validation is necessary for model generalization.

Several assumptions underlie this study. It was assumed that the microbiome compositions in our dataset reflect the microbial dynamics of the targeted age group, and that species abundance profiles serve as reliable predictors of chronological age. Additionally, it was assumed that age-associated shifts in microbiome composition are consistent across geographical locations, though this assumption may need further validation.

By addressing these limitations and assumptions in future research we can refine our understanding of the microbiome's role in aging.

6.2. Ethical-social impact, sustainability, and diversity

6.2.1. Ethical-social impact

The development of a microbiome-based regression model for age prediction raises important ethical and social considerations. First, privacy and data sensitivity must be prioritized, as gut microbiome data can reveal highly personal information, including dietary habits, geographic origins, and potential health risks. Strict anonymization protocols and secure data storage are essential to protect participant confidentiality, particularly if the model is later applied in clinical or commercial settings.

Informed consent is another critical issue; participants should be fully aware of how their data will be used, including any future applications in age prediction or related health assessments. Transparency is key to maintaining public trust in microbiome research. Additionally, there is a risk of misuse, such as insurance companies or employers leveraging predicted biological age for discriminatory practices. Ethical guidelines must be established to prevent such exploitation, potentially through regulatory frameworks that restrict non-medical use of microbiome-derived predictions.

Finally, health communication plays an important role. Since microbiome science is complex, predictive results must be presented in an accessible manner to avoid misinterpretation.

6.2.2. Sustainability

From a sustainability perspective, this research could contribute to long-term healthcare efficiency. If microbiome-based age prediction proves reliable, it may serve as a non-invasive biomarker for early detection of accelerated aging or age-related diseases, enabling preventive interventions that reduce long-term medical costs.

The model also aligns with resource-efficient diagnostics, as advances in sequencing technologies continue to make microbiome profiling faster and more affordable. By shifting focus toward predictive, data-driven health monitoring, this approach could reduce reliance on more invasive or expensive procedures.

Furthermore, promoting open science practices; such as sharing de-identified datasets or open-sourcing the model, can enhance sustainability in microbiome research. Collaborative reuse of data minimizes redundant studies and accelerates scientific progress while adhering to ethical standards.

6.2.3. Diversity

A major limitation in microbiome research is population bias. Gut microbiota composition varies significantly across ethnicities, diets, and geographic regions. If the model is trained on a narrow demographic (e.g., Western, urban populations), its predictions may not generalize well to underrepresented groups. Acknowledging this limitation is crucial, and future work should prioritize

diverse sampling—including different age groups and ethnic backgrounds—to improve model fairness and applicability.

6.3. Conclusions

In this study, we developed a robust machine learning model to predict chronological age based solely on gut microbiome compositions in healthy individuals. By leveraging species-level abundance profiles and integrating multiple feature selection strategies, including an innovative hybrid approach combining SHAP values with traditional feature importance scores, we identified several microbial taxa that may be associated with aging. This multi-faced approach enabled us to search for important features more in depth that not only had strong predictive performance but also offered valuable biological insights into age-related dynamics of the microbiome.

The predictive model, which incorporated a CatBoost regressor alongside permutation feature importance, achieved strong performance with mean absolute errors (MAE) of 9.99 years in cross-validated training, 9.17 years on the test set, and 9.43 years on an independent validation cohort. While the MAE was slightly higher than some previous studies, this is likely due to the narrower age range (19-66) we focused on, compared to models that target broader age ranges. The diversity of our dataset, which includes samples from multiple studies and geographic locations, further enhances the generalizability and external validity of our findings, though it may increase prediction error.

Among the most important microbial taxa identified, *Coprococcus comes* stands out for its role in maintaining gut health and its involvement in various metabolic processes. This genus is abundant in the gut microbiota of healthy individuals, including older adults, and has been associated as a promising candidate for a mental health biomarker. Our model highlighted *Coprococcus* as a significant predictor of age, reinforcing its potential in aging-related microbiome shifts. This aligns with previous research demonstrating that *Coprococcus* contributes to the fermentation of carbohydrates, production of short-chain fatty acids (SCFAs), and regulation of immune responses, all of which are vital for maintaining gut health.

In addition to *Coprococcus*, taxa like *Desulfovibrio piger* and **Unclassified Prevotellaceae** demonstrated strong associations with age, further corroborated by our UMAP and SHAP, providing a comprehensive understanding of shifts in the gut microbiome.

These findings not only contribute to the growing body of knowledge on age-related changes in gut microbiota but also highlight the potential of microbiome-based biomarkers for biological age prediction.

7. Future work

To advance the findings of this study, future research should prioritise several key directions. First, expanding the age range of studied populations to include older adults (>66 years) would help clarify the full trajectory of *Coproccoccus* across the lifespan, particularly its decline in late-stage aging. Longitudinal study designs would further extract true age-related microbial succession from transient fluctuations caused by dietary changes, medications, or other factors.

Methodological advancements will also be important for future progress. High resolution sequencing or cultivation-based approaches could address current taxonomic uncertainties surrounding poorly characterised taxa like *Siphoviridae* sp. *ctAjZ17*. Developing predictive models that combine microbial data with host factors would improve both accuracy and biological interpretability. Finally, validating these findings across diverse geographic populations would test the generalizability of *Coproccoccus*-age associations while accounting for regional differences in diet and lifestyle. Together, these approaches would clarify whether our species in our model (e.g., *Coproccoccus comes*) acts as a driver, biomarker in healthy aging, and potentially reveal avenues for therapeutic intervention.

References

<https://doi.org/10.1016/j.gpb.2018.10.002>

Abrahamsson, T. R., Jakobsson, H. E., Andersson, A. F., Björkstén, B., Engstrand, L., & Jenmalm, M. C. (2012). Low diversity of the gut microbiota in infants with atopic eczema. *Journal of Allergy and Clinical Immunology*, 129(2), 434-440.e2.

<https://doi.org/10.1016/j.jaci.2011.10.025>

Adachi, A., Zhang, F., Kanaya, S., & Ono, N. (2025). Quantifying uncertainty in microbiome-based prediction using Gaussian processes with microbial community dissimilarities. *Bioinformatics Advances*, 5(1), vbaf045. <https://doi.org/10.1093/bioadv/vbaf045>

Age prediction using Microbiota. (2025).

<https://www.microbiota-ism.com/news-on-microbiota/831-age-prediction-using-microbiota.html>

Aguiar-Pulido, V., Huang, W., Suarez-Ulloa, V., Cickovski, T., Mathee, K., & Narasimhan, G. (2016). Metagenomics, Metatranscriptomics, and Metabolomics Approaches for Microbiome Analysis. *Evolutionary Bioinformatics Online*, 12(Suppl 1), 5.

<https://doi.org/10.4137/EBO.S36436>

Almeida, H. M., Sardeli, A. V., Conway, J., Duggal, N. A., & Cavaglieri, C. R. (2022). Comparison between frail and non-frail older adults' gut microbiota: A systematic review and meta-analysis. *Ageing Research Reviews*, 82, 101773.

<https://doi.org/10.1016/j.arr.2022.101773>

Arbolea, S., Watkins, C., Stanton, C., & Ross, R. P. (2016). Gut Bifidobacteria Populations in Human Health and Aging. *Frontiers in Microbiology*, 7.

<https://doi.org/10.3389/fmicb.2016.01204>

Armstrong, G., Martino, C., Rahman, G., Gonzalez, A., Vázquez-Baeza, Y., Mishne, G., & Knight, R. (2021). Uniform Manifold Approximation and Projection (UMAP) Reveals Composite Patterns and Resolves Visualization Artifacts in Microbiome Data. *mSystems*, 6(5),

10.1128/msystems.00691-21. <https://doi.org/10.1128/msystems.00691-21>

- Arrieta, M.-C., Stiemsma, L. T., Dimitriu, P. A., Thorson, L., Russell, S., Yurist-Doutsch, S., Kuzeljevic, B., Gold, M. J., Britton, H. M., Lefebvre, D. L., Subbarao, P., Mandhane, P., Becker, A., McNagny, K. M., Sears, M. R., Kollmann, T., Investigators, the C. S., Mohn, W. W., Turvey, S. E., & Finlay, B. B. (2015). Early infancy microbial and metabolic alterations affect risk of childhood asthma. *Science Translational Medicine*.
<https://doi.org/10.1126/scitranslmed.aab2271>
- Aya, V., Jimenez, P., Muñoz, E., & Ramírez, J. D. (2023). Effects of exercise and physical activity on gut microbiota composition and function in older adults: A systematic review. *BMC Geriatrics*, 23(1), 364. <https://doi.org/10.1186/s12877-023-04066-y>
- Azad, M. B., Konya, T., Maughan, H., Guttman, D. S., Field, C. J., Chari, R. S., Sears, M. R., Becker, A. B., Scott, J. A., Kozyrskyj, A. L., & CHILD Study Investigators. (2013). Gut microbiota of healthy Canadian infants: Profiles by mode of delivery and infant diet at 4 months. *CMAJ: Canadian Medical Association Journal = Journal de l'Association Médicale Canadienne*, 185(5), 385–394. <https://doi.org/10.1503/cmaj.121189>
- Bäckhed, F., Roswall, J., Peng, Y., Feng, Q., Jia, H., Kovatcheva-Datchary, P., Li, Y., Xia, Y., Xie, H., Zhong, H., Khan, M. T., Zhang, J., Li, J., Xiao, L., Al-Aama, J., Zhang, D., Lee, Y. S., Kotowska, D., Colding, C., ... Wang, J. (2015). Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life. *Cell Host & Microbe*, 17(5), 690–703.
<https://doi.org/10.1016/j.chom.2015.04.004>
- Bengmark, S. (1998). Ecological control of the gastrointestinal tract. The role of probiotic flora. *Gut*, 42(1), 2–7. <https://doi.org/10.1136/gut.42.1.2>
- Biagi, E., Franceschi, C., Rampelli, S., Severgnini, M., Ostan, R., Turrone, S., Consolandi, C., Quercia, S., Scurti, M., Monti, D., Capri, M., Brigidi, P., & Candela, M. (2016). Gut Microbiota and Extreme Longevity. *Current Biology: CB*, 26(11), 1480–1485.
<https://doi.org/10.1016/j.cub.2016.04.016>
- Bibi, S., Navarre, D. A., Sun, X., Du, M., Rasco, B., & Zhu, M.-J. (2019). Beneficial Effect of Potato Consumption on Gut Microbiota and Intestinal Epithelial Health. *American Journal of Potato*

- Research*, 96(2), 170–176. <https://doi.org/10.1007/s12230-018-09706-3>
- Bisgaard, H., Li, N., Bonnelykke, K., Chawes, B. L. K., Skov, T., Paludan-Müller, G., Stokholm, J., Smith, B., & Krogfelt, K. A. (2011). Reduced diversity of the intestinal microbiota during infancy is associated with increased risk of allergic disease at school age. *Journal of Allergy and Clinical Immunology*, 128(3), 646–652.e5. <https://doi.org/10.1016/j.jaci.2011.04.060>
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*.
<https://link.springer.com/book/9780387310732>
- Bosco, N., & Noti, M. (2021). The aging gut microbiome and its impact on host immunity. *Genes & Immunity*, 22(5), 289–303. <https://doi.org/10.1038/s41435-021-00126-8>
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.
<https://doi.org/10.1023/A:1010933404324>
- Bruijning, M., Ayroles, J. F., Henry, L. P., Koskella, B., Meyer, K. M., & Metcalf, C. J. E. (2023). Relative abundance data can misrepresent heritability of the microbiome. *Microbiome*, 11(1), 222. <https://doi.org/10.1186/s40168-023-01669-w>
- Chen, S., Zhang, Z., Liu, S., Chen, T., Lu, Z., Zhao, W., Mou, X., & Liu, S. (2024). Consistent signatures in the human gut microbiome of longevous populations. *Gut Microbes*, 16(1), 2393756. <https://doi.org/10.1080/19490976.2024.2393756>
- Chen, Y., Wang, H., Lu, W., Wu, T., Yuan, W., Zhu, J., Lee, Y. K., Zhao, J., Zhang, H., & Chen, W. (2022). Human gut microbiome aging clocks based on taxonomic and functional signatures through multi-view learning. *Gut Microbes*, 14(1), 2025016.
<https://doi.org/10.1080/19490976.2021.2025016>
- Chen, Y.-R., Zhou, L.-Z., Fang, S.-T., Long, H.-Y., Chen, J.-Y., & Zhang, G.-X. (2019). Isolation of *Desulfovibrio* spp. From human gut microbiota using a next-generation sequencing directed culture method. *Letters in Applied Microbiology*, 68(6), 553–561.
<https://doi.org/10.1111/lam.13149>
- Chilton, P. M., Ghare, S. S., Charpentier, B. T., Myers, S. A., Rao, A. V., Petrosino, J. F., Hoffman, K. L., Greenwell, J. C., Tyagi, N., Behera, J., Wang, Y., Sloan, L. J., Zhang, J., Shields, C. B., Cooper, G. E., Gobejishvili, L., Whittemore, S. R., McClain, C. J., & Barve, S. S. (2024).

- Age-associated temporal decline in butyrate-producing bacteria plays a key pathogenic role in the onset and progression of neuropathology and memory deficits in 3×Tg-AD mice. *Gut Microbes*, 16(1), 2389319. <https://doi.org/10.1080/19490976.2024.2389319>
- Choden, T. (2022). The gut microbiome and the immune system. *Exploration of Medicine*, 3(3), Article 3. <https://doi.org/10.37349/emed.2022.00087>
- Claesson, M. J., Cusack, S., O'Sullivan, O., Greene-Diniz, R., de Weerd, H., Flannery, E., Marchesi, J. R., Falush, D., Dinan, T., Fitzgerald, G., Stanton, C., van Sinderen, D., O'Connor, M., Harnedy, N., O'Connor, K., Henry, C., O'Mahony, D., Fitzgerald, A. P., Shanahan, F., ... O'Toole, P. W. (2011). Composition, variability, and temporal stability of the intestinal microbiota of the elderly. *Proceedings of the National Academy of Sciences of the United States of America*, 108 Suppl 1(Suppl 1), 4586–4591. <https://doi.org/10.1073/pnas.1000097107>
- Claesson, M. J., Jeffery, I. B., Conde, S., Power, S. E., O'Connor, E. M., Cusack, S., Harris, H. M. B., Coakley, M., Lakshminarayanan, B., O'Sullivan, O., Fitzgerald, G. F., Deane, J., O'Connor, M., Harnedy, N., O'Connor, K., O'Mahony, D., van Sinderen, D., Wallace, M., Brennan, L., ... O'Toole, P. W. (2012). Gut microbiota composition correlates with diet and health in the elderly. *Nature*, 488(7410), 178–184. <https://doi.org/10.1038/nature11319>
- Clarke, S. F., Murphy, E. F., O'Sullivan, O., Lucey, A. J., Humphreys, M., Hogan, A., Hayes, P., O'Reilly, M., Jeffery, I. B., Wood-Martin, R., Kerins, D. M., Quigley, E., Ross, R. P., O'Toole, P. W., Molloy, M. G., Falvey, E., Shanahan, F., & Cotter, P. D. (2014). Exercise and associated dietary extremes impact on gut microbial diversity. *Gut*, 63(12), 1913–1920. <https://doi.org/10.1136/gutjnl-2013-306541>
- Dasari, M. R., Roche, K. E., Jansen, D., Anderson, J., Alberts, S. C., Tung, J., Gilbert, J. A., Blekhman, R., Mukherjee, S., & Archie, E. A. (2024). Social and environmental predictors of gut microbiome age in wild baboons. *eLife*, 13. <https://doi.org/10.7554/eLife.102166.1>
- Dasari, M. R., Roche, K. E., Jansen, D., Anderson, J., Alberts, S. C., Tung, J., Gilbert, J. A., Blekhman, R., Mukherjee, S., & Archie, E. A. (2025). Social and environmental predictors of gut microbiome age in wild baboons. *eLife*, 13. <https://doi.org/10.7554/eLife.102166.2>

- Dempsey, E., & Corr, S. C. (2022). *Lactobacillus* spp. for Gastrointestinal Health: Current and Future Perspectives. *Frontiers in Immunology*, 13, 840245.
<https://doi.org/10.3389/fimmu.2022.840245>
- Durazzi, F., Sala, C., Castellani, G., Manfreda, G., Remondini, D., & De Cesare, A. (2021). Comparison between 16S rRNA and shotgun sequencing data for the taxonomic characterization of the gut microbiota. *Scientific Reports*, 11(1), 3030.
<https://doi.org/10.1038/s41598-021-82726-y>
- Emencheta, S. C., Olovo, C. V., Eze, O. C., Kalu, C. F., Berebon, D. P., Onuigbo, E. B., Vila, M. M. D. C., Balcão, V. M., & Attama, A. A. (2023). The Role of Bacteriophages in the Gut Microbiota: Implications for Human Health. *Pharmaceutics*, 15(10), 2416.
<https://doi.org/10.3390/pharmaceutics15102416>
- Ferreira, C. M., Vieira, A. T., Vinolo, M. A. R., Oliveira, F. A., Curi, R., & Martins, F. dos S. (2014). The Central Role of the Gut Microbiota in Chronic Inflammatory Diseases. *Journal of Immunology Research*, 2014(1), 689492. <https://doi.org/10.1155/2014/689492>
- Galià-Camps, C., Baños, E., Pascual, M., Carreras, C., & Turon, X. (2023). Multidimensional variability of the microbiome of an invasive ascidian species. *iScience*, 26(10), 107812.
<https://doi.org/10.1016/j.isci.2023.107812>
- Ghosh, T. S., Das, M., Jeffery, I. B., & O'Toole, P. W. (2020). Adjusting for age improves identification of gut microbiome alterations in multiple diseases. *eLife*, 9, e50240.
<https://doi.org/10.7554/eLife.50240>
- Ghosh, T. S., Shanahan, F., & O'Toole, P. W. (2022a). The gut microbiome as a modulator of healthy ageing. *Nature Reviews. Gastroenterology & Hepatology*, 19(9), 565–584.
<https://doi.org/10.1038/s41575-022-00605-x>
- Ghosh, T. S., Shanahan, F., & O'Toole, P. W. (2022b). Toward an improved definition of a healthy microbiome for healthy aging. *Nature Aging*, 2(11), 1054–1069.
<https://doi.org/10.1038/s43587-022-00306-9>
- Goodrich, J. K., Waters, J. L., Poole, A. C., Sutter, J. L., Koren, O., Blekhman, R., Beaumont, M., Van Treuren, W., Knight, R., Bell, J. T., Spector, T. D., Clark, A. G., & Ley, R. E. (2014). Human

- genetics shape the gut microbiome. *Cell*, 159(4), 789–799.
<https://doi.org/10.1016/j.cell.2014.09.053>
- Gregory, A. C., Zablocki, O., Zayed, A. A., Howell, A., Bolduc, B., & Sullivan, M. B. (2020). The Gut Virome Database Reveals Age-Dependent Patterns of Virome Diversity in the Human Gut. *Cell Host & Microbe*, 28(5), 724–740.e8. <https://doi.org/10.1016/j.chom.2020.08.003>
- Haran, J. P., Bucci, V., Dutta, P., Ward, D., & McCormick, B. (2018). The nursing home elder microbiome stability and associations with age, frailty, nutrition and physical location. *Journal of Medical Microbiology*, 67(1), 40–51. <https://doi.org/10.1099/jmm.0.000640>
- Haran, J. P., & McCormick, B. A. (2021). Aging, Frailty, and the Microbiome-How Dysbiosis Influences Human Aging and Disease. *Gastroenterology*, 160(2), 507–523.
<https://doi.org/10.1053/j.gastro.2020.09.060>
- Heiman, M. L., & Greenway, F. L. (2016). A healthy gastrointestinal microbiome is dependent on dietary diversity. *Molecular Metabolism*, 5(5), 317–320.
<https://doi.org/10.1016/j.molmet.2016.02.005>
- Hernández Medina, R., Kutuzova, S., Nielsen, K. N., Johansen, J., Hansen, L. H., Nielsen, M., & Rasmussen, S. (2022). Machine learning and deep learning applications in microbiome research. *ISME Communications*, 2(1), 1–7. <https://doi.org/10.1038/s43705-022-00182-9>
- Holzinger, A. (2018). From Machine Learning to Explainable AI. *2018 World Symposium on Digital Intelligence for Systems and Machines (DISA)*, 55–66.
<https://doi.org/10.1109/DISA.2018.8490530>
- Hu, J., An, X., Yang, P., Tan, R., Chen, T., Chen, J., Tao, Y., Li, X., Sun, R., Zhang, S., Liu, S., & Yang, L. (2025). Analysis of the Primary Pathogenic Bacteria in Abscess Disease of Musk Deer Using Metagenomic Approaches. *Animals: An Open Access Journal from MDPI*, 15(8), 1105. <https://doi.org/10.3390/ani15081105>
- Huang, S., Haiminen, N., Carrieri, A.-P., Hu, R., Jiang, L., Parida, L., Russell, B., Allaband, C., Zarrinpar, A., Vázquez-Baeza, Y., Belda-Ferre, P., Zhou, H., Kim, H.-C., Swafford, A. D., Knight, R., & Xu, Z. Z. (2020). Human Skin, Oral, and Gut Microbiomes Predict Chronological Age. *mSystems*, 5(1), 10.1128/msystems.00630-19.

<https://doi.org/10.1128/msystems.00630-19>

Jackson, M. A., Jeffery, I. B., Beaumont, M., Bell, J. T., Clark, A. G., Ley, R. E., O'Toole, P. W., Spector, T. D., & Steves, C. J. (2016). Signatures of early frailty in the gut microbiota.

Genome Medicine, 8(1), 8. <https://doi.org/10.1186/s13073-016-0262-7>

Jeffery, I. B., Lynch, D. B., & O'Toole, P. W. (2016). Composition and temporal stability of the gut microbiota in older persons. *The ISME Journal*, 10(1), 170–182.

<https://doi.org/10.1038/ismej.2015.88>

Jiang, H., Zeng, W., Zhang, X., Peng, A., Cao, D., & Zhu, F. (2024). Gut Microbiome variation in patients with early-stage mild-to-moderate intracerebral hemorrhage: A pilot study exploring therapeutic targets. *Journal of Stroke and Cerebrovascular Diseases: The Official Journal of National Stroke Association*, 33(11), 108001.

<https://doi.org/10.1016/j.jstrokecerebrovasdis.2024.108001>

Johnson, A. J., Vangay, P., Al-Ghalith, G. A., Hillmann, B. M., Ward, T. L., Shields-Cutler, R. R., Kim, A. D., Shmagel, A. K., Syed, A. N., Personalized Microbiome Class Students, Walter, J., Menon, R., Koecher, K., & Knights, D. (2019). Daily Sampling Reveals Personalized Diet-Microbiome Associations in Humans. *Cell Host & Microbe*, 25(6), 789-802.e5.

<https://doi.org/10.1016/j.chom.2019.05.005>

Johnson, K. V.-A. (2020). Gut microbiome composition and diversity are related to human personality traits. *Human Microbiome Journal*, 15, 100069. <https://doi.org/10.1016/j.humic.2019.100069>

Karlsson, F., Tremaroli, V., Nielsen, J., & Bäckhed, F. (2013). Assessing the human gut microbiota in metabolic diseases. *Diabetes*, 62(10), 3341–3349. <https://doi.org/10.2337/db13-0844>

Kim, H. E., Kim, B. R., Hong, S. H., Song, S. Y., Jeong, J. H., & Kim, G. H. (2024). Predicting superagers: A machine learning approach utilizing gut microbiome features. *Frontiers in Aging Neuroscience*, 16. <https://doi.org/10.3389/fnagi.2024.1444998>

Knip, M., & Siljander, H. (2016). The role of the intestinal microbiota in type 1 diabetes mellitus.

Nature Reviews. Endocrinology, 12(3), 154–167. <https://doi.org/10.1038/nrendo.2015.218>

Kort, R., Schlösser, J., Vazquez, A. R., Atukunda, P., Muhoozi, G. K. M., Wacoo, A. P., Sybesma, W. F. H., Westerberg, A. C., Iversen, P. O., & Schoen, E. D. (2021). Model Selection Reveals the

- Butyrate-Producing Gut Bacterium *Coprococcus eutactus* as Predictor for Language Development in 3-Year-Old Rural Ugandan Children. *Frontiers in Microbiology*, 12. <https://doi.org/10.3389/fmicb.2021.681485>
- Kuang, Y.-S., Li, S.-H., Guo, Y., Lu, J.-H., He, J.-R., Luo, B.-J., Jiang, F.-J., Shen, H., Papasian, C. J., Pang, H., Xia, H.-M., Deng, H.-W., & Qiu, X. (2016). Composition of gut microbiota in infants in China and global comparison. *Scientific Reports*, 6(1), 36666. <https://doi.org/10.1038/srep36666>
- Kursa, M. B., Jankowski, A., & Rudnicki, W. R. (2010). Boruta – A System for Feature Selection. *Fundamenta Informaticae*, 101(4), 271–285. <https://doi.org/10.3233/FI-2010-288>
- Lei, W., Xu, Y., Su, J., Chong, C.-M., Su, H.-X., Luo, J., Fang, E. F., Bao, Z., & Chen, G. (2019). Applications of high-throughput ‘omics’ data in the study of frailty. *Translational Medicine of Aging*, 3, 40–51. <https://doi.org/10.1016/j.tma.2019.04.002>
- Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 25(14), 1754–1760. <https://doi.org/10.1093/bioinformatics/btp324>
- Li, J., Tian, C., Feng, S., Cheng, W., Tao, S., Li, C., Xiao, Y., & Wei, H. (2024). Modulation of Gut Microbial Community and Metabolism by *Bacillus licheniformis* HD173 Promotes the Growth of Nursery Piglets Model. *Nutrients*, 16(10), 1497. <https://doi.org/10.3390/nu16101497>
- Lin, A., Bik, E. M., Costello, E. K., Dethlefsen, L., Haque, R., Relman, D. A., & Singh, U. (2013). Distinct Distal Gut Microbiome Diversity and Composition in Healthy Children from Bangladesh and the United States. *PLOS ONE*, 8(1), e53838. <https://doi.org/10.1371/journal.pone.0053838>
- Lloyd-Price, J., Abu-Ali, G., & Huttenhower, C. (2016). The healthy human microbiome. *Genome Medicine*, 8(1), 51. <https://doi.org/10.1186/s13073-016-0307-y>
- Louis, P., & Flint, H. J. (2017). Formation of propionate and butyrate by the human colonic microbiota. *Environmental Microbiology*, 19(1), 29–41. <https://doi.org/10.1111/1462-2920.13589>

- Lundberg, S., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions* (arXiv:1705.07874). arXiv. <https://doi.org/10.48550/arXiv.1705.07874>
- Maffei, V. J., Kim, S., Blanchard, E., IV, Luo, M., Jazwinski, S. M., Taylor, C. M., & Welsh, D. A. (2017). Biological Aging and the Human Gut Microbiota. *The Journals of Gerontology: Series A*, 72(11), 1474–1482. <https://doi.org/10.1093/gerona/glx042>
- Malesu, V. K. (2025, February 10). *Your gut bacteria shape your health from childhood to old age, study reveals*. News-Medical. <https://www.news-medical.net/news/20250210/Your-gut-bacteria-shape-your-health-from-childhood-to-old-age-study-reveals.aspx>
- Mallick, R., Basak, S., Das, R. K., Banerjee, A., Paul, S., Pathak, S., & Duttaroy, A. K. (2024). Roles of the gut microbiota in human neurodevelopment and adult brain disorders. *Frontiers in Neuroscience*, 18. <https://doi.org/10.3389/fnins.2024.1446700>
- Mancabelli, L., Milani, C., De Biase, R., Bocchio, F., Fontana, F., Lugli, G. A., Alessandri, G., Tarracchini, C., Viappiani, A., De Conto, F., Nouvenne, A., Ticinesi, A., Bussolati, O., Meschi, T., Cecchi, R., Turroni, F., & Ventura, M. (2024). Taxonomic and metabolic development of the human gut microbiome across life stages: A worldwide metagenomic investigation. *mSystems*, 9(4), e01294-23. <https://doi.org/10.1128/msystems.01294-23>
- Meng, C., Zeleznik, O. A., Thallinger, G. G., Kuster, B., Gholami, A. M., & Culhane, A. C. (2016). Dimension reduction techniques for the integrative analysis of multi-omics data. *Briefings in Bioinformatics*, 17(4), 628–641. <https://doi.org/10.1093/bib/bbv108>
- Min, M., Egli, C., & Sivamani, R. K. (2024). The Gut and Skin Microbiome and Its Association with Aging Clocks. *International Journal of Molecular Sciences*, 25(13), Article 13. <https://doi.org/10.3390/ijms25137471>
- Mosca, A., Leclerc, M., & Hugot, J. P. (2016). Gut Microbiota Diversity and Human Diseases: Should We Reintroduce Key Predators in Our Ecosystem? *Frontiers in Microbiology*, 7. <https://doi.org/10.3389/fmicb.2016.00455>
- Nagao-Kitamoto, H., & Kamada, N. (2017). Host-microbial Cross-talk in Inflammatory Bowel Disease. *Immune Network*, 17(1), 1–12. <https://doi.org/10.4110/in.2017.17.1.1>

- Niu, J., Xu, L., Qian, Y., Sun, Z., Yu, D., Huang, J., Zhou, X., Wang, Y., Zhang, T., Ren, R., Li, Z., Yu, J., & Gao, X. (2020). Evolution of the Gut Microbiome in Early Childhood: A Cross-Sectional Study of Chinese Children. *Frontiers in Microbiology*, 11. <https://doi.org/10.3389/fmicb.2020.00439>
- Notting, F., Pirovano, W., Sybesma, W., & Kort, R. (2023). The butyrate-producing and spore-forming bacterial genus *Coprococcus* as a potential biomarker for neurological disorders. *Gut Microbiome*, 4, e16. <https://doi.org/10.1017/gmb.2023.14>
- O’Callaghan, A., & van Sinderen, D. (2016). Bifidobacteria and Their Role as Members of the Human Gut Microbiota. *Frontiers in Microbiology*, 7, 925. <https://doi.org/10.3389/fmicb.2016.00925>
- Odamaki, T., Kato, K., Sugahara, H., Hashikura, N., Takahashi, S., Xiao, J., Abe, F., & Osawa, R. (2016). Age-related changes in gut microbiota composition from newborn to centenarian: A cross-sectional study. *BMC Microbiology*, 16(1), 90. <https://doi.org/10.1186/s12866-016-0708-5>
- Ojala, T., Kankuri, E., & Kankainen, M. (2023). Understanding human health through metatranscriptomics. *Trends in Molecular Medicine*, 29(5), 376–389. <https://doi.org/10.1016/j.molmed.2023.02.002>
- Park, S.-H., Kim, K.-A., Ahn, Y.-T., Jeong, J.-J., Huh, C.-S., & Kim, D.-H. (2015). Comparative analysis of gut microbiota in elderly people of urbanized towns and longevity villages. *BMC Microbiology*, 15, 49. <https://doi.org/10.1186/s12866-015-0386-8>
- Paytuví, A., Battista, E., Scippacercola, F., Cigliano, R. A., & Sanseverino, W. (2019). *GALA: An integrated metagenomics suite* (p. 804690). bioRxiv. <https://doi.org/10.1101/804690>
- Peeters, J., Thas, O., Shkedy, Z., Kodalcı, L., Musisi, C., Owokotomo, O. E., Dyczko, A., Hamad, I., Vangronsveld, J., Kleinewietfeld, M., Thijs, S., & Aerts, J. (2021). Exploring the Microbiome Analysis and Visualization Landscape. *Frontiers in Bioinformatics*, 1. <https://doi.org/10.3389/fbinf.2021.774631>
- Ragonnaud, E., & Biragyn, A. (2021). Gut microbiota as the key controllers of “healthy” aging of elderly people. *Immunity & Ageing*, 18(1), 2. <https://doi.org/10.1186/s12979-020-00213-w>

- Rahayu, E. S., Utami, T., Mariyatun, M., Hasan, P. N., Kamil, R. Z., Setyawan, R. H., Pamungkaningtyas, F. H., Harahap, I. A., Wiryohanjoyo, D. V., Pramesi, P. C., Cahyanto, M. N., Sujaya, I. N., & Juffrie, M. (2019). Gut microbiota profile in healthy Indonesians. *World Journal of Gastroenterology*, 25(12), 1478–1491. <https://doi.org/10.3748/wjg.v25.i12.1478>
- Richter, T. K. S., Michalski, J. M., Zanetti, L., Tennant, S. M., Chen, W. H., & Rasko, D. A. (2018). Responses of the Human Gut *Escherichia coli* Population to Pathogen and Antibiotic Disturbances. *mSystems*. <https://doi.org/10.1128/msystems.00047-18>
- Riedel, C. U., Foata, F., Philippe, D., Adolfsson, O., Eikmanns, B. J., & Blum, S. (2006). Anti-inflammatory effects of bifidobacteria by inhibition of LPS-induced NF- κ B activation. *World Journal of Gastroenterology*, 12(23), 3729–3735. <https://doi.org/10.3748/wjg.v12.i23.3729>
- Rinninella, E., Cintoni, M., Raoul, P., Lopetuso, L. R., Scaldaferri, F., Pulcini, G., Miggianno, G. A. D., Gasbarrini, A., & Mele, M. C. (2019). Food Components and Dietary Habits: Keys for a Healthy Gut Microbiota Composition. *Nutrients*, 11(10), Article 10. <https://doi.org/10.3390/nu11102393>
- Roscher, R., Bohn, B., Duarte, M. F., & Garcke, J. (2020). Explainable Machine Learning for Scientific Insights and Discoveries. *IEEE Access*, 8, 42200–42216. IEEE Access. <https://doi.org/10.1109/ACCESS.2020.2976199>
- Rybicka, I., & Kaźmierczak, Z. (2025). The human phageome: Niche-specific distribution of bacteriophages and their clinical implications. *Applied and Environmental Microbiology*, 0(0), e01788-24. <https://doi.org/10.1128/aem.01788-24>
- Sandrini, S., Perricone, V., Cremonesi, P., Castiglioni, B., Biscarini, F., Verdile, N., Brevini, T., Comi, M., Jiang, X., & Agazzi, A. (2024). Yeast mixture supplementation modulates faecal microbiota and ileum morphology of weaning pigs. *Animal: An International Journal of Animal Bioscience*, 18(9), 101275. <https://doi.org/10.1016/j.animal.2024.101275>
- Schlomann, B. H., & Parthasarathy, R. (2019). Timescales of gut microbiome dynamics. *Current Opinion in Microbiology*, 50, 56–63. <https://doi.org/10.1016/j.mib.2019.09.011>
- Selber-Hnatiw, S., Sultana, T., Tse, W., Abdollahi, N., Abdullah, S., Al Rahbani, J., Alazar, D.,

- Alrumhein, N. J., Aprikian, S., Arshad, R., Azuelos, J.-D., Bernadotte, D., Beswick, N., Chazbey, H., Church, K., Ciubotaru, E., D'Amato, L., Del Corpo, T., Deng, J., ... Gamberi, C. (2020). Metabolic networks of the human gut microbiota. *Microbiology*, 166(2), 96–119. <https://doi.org/10.1099/mic.0.000853>
- Seo, S.-H., Na, C.-S., Park, S.-E., Kim, E.-J., Kim, W.-S., Park, C., Oh, S., You, Y., Lee, M.-H., Cho, K.-M., Kwon, S. J., Whon, T. W., Roh, S. W., & Son, H.-S. (2023). Machine learning model for predicting age in healthy individuals using age-related gut microbes and urine metabolites. *Gut Microbes*, 15(1), 2226915. <https://doi.org/10.1080/19490976.2023.2226915>
- Shetty, S. A., Hugenholtz, F., Lahti, L., Smidt, H., & de Vos, W. M. (2017). Intestinal microbiome landscaping: Insight in community assemblage and implications for microbial modulation strategies. *FEMS Microbiology Reviews*, 41(2), 182–199. <https://doi.org/10.1093/femsre/fuw045>
- Singh, S. B., Carroll-Portillo, A., & Lin, H. C. (2023). Desulfovibrio in the Gut: The Enemy within? *Microorganisms*, 11(7), 1772. <https://doi.org/10.3390/microorganisms11071772>
- Sohn, J., Li, L., Zhang, L., Genco, R. J., Falkner, K. L., Tettelin, H., Rowsam, A. M., Smiraglia, D. J., Novak, J. M., Diaz, P. I., Sun, Y., & Kirkwood, K. L. (2023). Periodontal disease is associated with increased gut colonization of pathogenic Haemophilus parainfluenzae in patients with Crohn's disease. *Cell Reports*, 42(2), 112120. <https://doi.org/10.1016/j.celrep.2023.112120>
- Sommer, F., Anderson, J. M., Bharti, R., Raes, J., & Rosenstiel, P. (2017). The resilience of the intestinal microbiota influences health and disease. *Nature Reviews. Microbiology*, 15(10), 630–638. <https://doi.org/10.1038/nrmicro.2017.58>
- Tamai, I. A., Mohammadzadeh, A., Salehi, T. Z., Mahmoodi, P., & Pakbin, B. (2022). Expression of virulence factor genes in co-infections with *Trueperella pyogenes* isolates and other bacterial pathogens; an in vivo study. *Microbial Pathogenesis*, 164, 105435. <https://doi.org/10.1016/j.micpath.2022.105435>
- Tamburini, S., Shen, N., Wu, H. C., & Clemente, J. C. (2016). The microbiome in early life: Implications for health outcomes. *Nature Medicine*, 22(7), 713–722. <https://doi.org/10.1038/nm.4142>

- Thomas, T., Gilbert, J., & Meyer, F. (2012). Metagenomics—A guide from sampling to data analysis. *Microbial Informatics and Experimentation*, 2, 3. <https://doi.org/10.1186/2042-5783-2-3>
- Tuikhar, N., Keisam, S., Labala, R. K., Imrat, null, Ramakrishnan, P., Arunkumar, M. C., Ahmed, G., Biagi, E., & Jeyaram, K. (2019). Comparative analysis of the gut microbiota in centenarians and young adults shows a common signature across genotypically non-related populations. *Mechanisms of Ageing and Development*, 179, 23–35. <https://doi.org/10.1016/j.mad.2019.02.001>
- Vacca, M., Celano, G., Calabrese, F. M., Portincasa, P., Gobbetti, M., & Angelis, M. D. (2020). The Controversial Role of Human Gut Lachnospiraceae. *Microorganisms*, 8(4), 573. <https://doi.org/10.3390/microorganisms8040573>
- Vangay, P., Johnson, A. J., Ward, T. L., Al-Ghalith, G. A., Shields-Cutler, R. R., Hillmann, B. M., Lucas, S. K., Beura, L. K., Thompson, E. A., Till, L. M., Batres, R., Paw, B., Pergament, S. L., Saenyakul, P., Xiong, M., Kim, A. D., Kim, G., Masopust, D., Martens, E. C., ... Knights, D. (2018). US Immigration Westernizes the Human Gut Microbiome. *Cell*, 175(4), 962–972.e10. <https://doi.org/10.1016/j.cell.2018.10.029>
- Verdi, S., Jackson, M. A., Beaumont, M., Bowyer, R. C. E., Bell, J. T., Spector, T. D., & Steves, C. J. (2018). An Investigation Into Physical Frailty as a Link Between the Gut Microbiome and Cognitive Health. *Frontiers in Aging Neuroscience*, 10, 398. <https://doi.org/10.3389/fnagi.2018.00398>
- Walters, K. E., & Martiny, J. B. H. (2020). Alpha-, beta-, and gamma-diversity of bacteria varies across habitats. *PLoS ONE*, 15(9), e0233872. <https://doi.org/10.1371/journal.pone.0233872>
- Wang, F., Yu, T., Huang, G., Cai, D., Liang, X., Su, H., Zhu, Z., Li, D., Yang, Y., Shen, P., Mao, R., Yu, L., Zhao, M., & Li, Q. (2015). Gut Microbiota Community and Its Assembly Associated with Age and Diet in Chinese Centenarians. 25(8), 1195–1204. <https://doi.org/10.4014/jmb.1410.10014>
- Wang, H., Chen, Y., Feng, L., Lu, S., Zhu, J., Zhao, J., Zhang, H., Chen, W., & Lu, W. (2025). A gut aging clock using microbiome multi-view profiles is associated with health and frail risk. *Gut Microbes*, 16(1), 2297852. <https://doi.org/10.1080/19490976.2023.2297852>

- Wang, N., Li, R., Lin, H., Fu, C., Wang, X., Zhang, Y., Su, M., Huang, P., Qian, J., Jiang, F., Wang, H., Jiang, L., Yu, X., Liu, J., Chen, Y., & Jiang, Q. (2019). Enriched taxa were found among the gut microbiota of centenarians in East China. *PloS One*, *14*(10), e0222763. <https://doi.org/10.1371/journal.pone.0222763>
- Wang, Q., Fan, X., Wu, S., & Su, X. (2024). PM-CNN: Microbiome status recognition and disease detection model based on phylogeny and multi-path neural network. *Bioinformatics Advances*, *4*(1), vbae013. <https://doi.org/10.1093/bioadv/vbae013>
- Wilmanski, T., Diener, C., Rappaport, N., Patwardhan, S., Wiedrick, J., Lapidus, J., Earls, J. C., Zimmer, A., Glusman, G., Robinson, M., Yurkovich, J. T., Kado, D. M., Cauley, J. A., Zmuda, J., Lane, N. E., Magis, A. T., Lovejoy, J. C., Hood, L., Gibbons, S. M., ... Price, N. D. (2021). Gut microbiome pattern reflects healthy ageing and predicts survival in humans. *Nature Metabolism*, *3*(2), 274–286. <https://doi.org/10.1038/s42255-021-00348-0>
- Wright, M. N., Gola, D., & Ziegler, A. (2017). Preprocessing and Quality Control for Whole-Genome Sequences from the Illumina HiSeq X Platform. *Methods in Molecular Biology (Clifton, N.J.)*, *1666*, 629–647. https://doi.org/10.1007/978-1-4939-7274-6_30
- Wu, H.-J. (2012). The role of gut microbiota in immune homeostasis and autoimmunity. *Gut Microbes*, *3*(1), 4–14. <https://doi.org/10.4161/gmic.19320>
- Wu, L., Zeng, T., Zinellu, A., Rubino, S., Kelvin, D. J., & Carru, C. (2019). A Cross-Sectional Study of Compositional and Functional Profiles of Gut Microbiota in Sardinian Centenarians. *mSystems*, *4*(4), 10.1128/msystems.00325-19. <https://doi.org/10.1128/msystems.00325-19>
- Xu, J., Sheikh, T. M. M., Shafiq, M., Khan, M. N., Wang, M., Guo, X., Yao, F., Xie, Q., Yang, Z., Khalid, A., & Jiao, X. (2025). Exploring the gut microbiota landscape in cow milk protein allergy: Clinical insights and diagnostic implications in pediatric patients. *Journal of Dairy Science*, *108*(1), 73–89. <https://doi.org/10.3168/jds.2024-25455>
- Xu, Y., Wang, Y., Li, H., Dai, Y., Chen, D., Wang, M., Jiang, X., Huang, Z., Yu, H., Huang, J., & Xiong, Z. (2021). Altered Fecal Microbiota Composition in Older Adults With Frailty. *Frontiers in Cellular and Infection Microbiology*, *11*, 696186. <https://doi.org/10.3389/fcimb.2021.696186>

- Yang, Y., Song, X., Wang, G., Xia, Y., Xiong, Z., & Ai, L. (2024). Understanding *Ligilactobacillus salivarius* from Probiotic Properties to Omics Technology: A Review. *Foods*, *13*(6), 895. <https://doi.org/10.3390/foods13060895>
- Yeoh, Y. K., Sun, Y., Ip, L. Y. T., Wang, L., Chan, F. K. L., Miao, Y., & Ng, S. C. (2022). *Prevotella* species in the human gut is primarily comprised of *Prevotella copri*, *Prevotella stercorea* and related lineages. *Scientific Reports*, *12*(1), 9055. <https://doi.org/10.1038/s41598-022-12721-4>
- Yu, M., Zhao, J., Xie, Q., Deng, J., Zhu, Y., Chen, J., Xiang, Z., Zhang, T., Liu, G., Xia, C., Shi, L., Wu, B., Gouvinhas, I., & Barros, A. N. (2024). Golden Flower Tibetan Tea Polysaccharides Alleviate Constipation in Mice by Regulating Aquaporins-Mediated Water Transport System and Gut Microbiota. *Foods (Basel, Switzerland)*, *13*(17), 2749. <https://doi.org/10.3390/foods13172749>
- Zhang, P. (2022). Influence of Foods and Nutrition on the Gut Microbiome and Implications for Intestinal Health. *International Journal of Molecular Sciences*, *23*(17), 9588. <https://doi.org/10.3390/ijms23179588>
- Zhang, X., Zhong, H., Li, Y., Shi, Z., Ren, H., Zhang, Z., Zhou, X., Tang, S., Han, X., Lin, Y., Yang, F., Wang, D., Fang, C., Fu, Z., Wang, L., Zhu, S., Hou, Y., Xu, X., Yang, H., ... Ji, L. (2021). Sex- and age-related trajectories of the adult human gut microbiota shared across populations of different ethnicities. *Nature Aging*, *1*(1), 87–100. <https://doi.org/10.1038/s43587-020-00014-2>
- Zhang, Y., Sharma, S., Tom, L., Liao, Y.-T., & Wu, V. C. H. (2023). Gut Phageome—An Insight into the Role and Impact of Gut Microbiome and Their Correlation with Mammal Health and Diseases. *Microorganisms*, *11*(10), 2454. <https://doi.org/10.3390/microorganisms11102454>
- Zhou, R., Ng, S. K., Sung, J. J. Y., Goh, W. W. B., & Wong, S. H. (2023). Data pre-processing for analyzing microbiome data – A mini review. *Computational and Structural Biotechnology Journal*, *21*, 4804–4815. <https://doi.org/10.1016/j.csbj.2023.10.001>
- Zhuang, L., Chen, H., Zhang, S., Zhuang, J., Li, Q., & Feng, Z. (2019). Intestinal Microbiota in Early Life and Its Implications on Childhood Health. *Genomics, Proteomics & Bioinformatics*, *17*(1), 13–25. <https://doi.org/10.1016/j.gpb.2018.10.002>