frontiers | Research Topics

Microbiome and machine learning, volume II

Edited by
Erik Bongcam-Rudloff, Marcus Claesson, Aldert Zomer, Randi Jacobsen Bertelsen, Isabel Moreno Indias and Domenica D'Elia

Published in
Frontiers in Microbiology

ISSN 1664-8714
ISBN 978-2-8325-5603-0
DOI 10.3389/978-2-8325-5603-0

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Microbiome and machine learning, volume II

Topic editors

Erik Bongcam-Rudloff — Swedish University of Agricultural Sciences, Sweden
Marcus Claesson — University College Cork, Ireland
Aldert Zomer — Utrecht University, Netherlands
Randi Jacobsen Bertelsen — University of Bergen, Norway
Isabel Moreno Indias — University of Malaga, Spain
Domenica D'Elia — Institute of Biomedical Technologies, National Research Council (CNR), Italy

Citation

Bongcam-Rudloff, E., Claesson, M., Zomer, A., Bertelsen, R. J., Moreno Indias, I., D'Elia, D., eds. (2024). *Microbiome and machine learning, volume II*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-5603-0

Table of contents

frontiers | Frontiers in Microbiology

TYPE Editorial
PUBLISHED 15 October 2024
DOI 10.3389/fmicb.2024.1499260

OPEN ACCESS

EDITED AND REVIEWED BY
John R. Battista,
Louisiana State University, United States

*CORRESPONDENCE
Domenica D'Elia
✉ domenica.delia@ba.itb.cnr.it

RECEIVED 20 September 2024
ACCEPTED 30 September 2024
PUBLISHED 15 October 2024

CITATION
D'Elia D, Zomer A, Moreno Indias I, Bongcam-Rudloff E, Bertelsen RJ and Claesson MJ (2024) Editorial: Microbiome and machine learning, volume II.
Front. Microbiol. 15:1499260.
doi: 10.3389/fmicb.2024.1499260

Editorial: Microbiome and machine learning, volume II

Domenica D'Elia[1]*, Aldert Zomer[2], Isabel Moreno Indias[3], Erik Bongcam-Rudloff[4], Randi Jacobsen Bertelsen[5] and Marcus Joakim Claesson[6]

[1]Department of Biomedical Sciences, National Research Council, Institute for Biomedical Technologies, Bari, Italy, [2]Department of Biomolecular Health Sciences (Infectious Diseases and Immunology), Faculty of Veterinary Medicine, Utrecht University, Utrecht, Netherlands, [3]Department of Endocrinology and Nutrition, Virgen de la Victoria University Hospital, The Biomedical Research Institute of Malaga and Platform in Nanomedicine (IBIMA-BIONAND Platform), University of Malaga, Malaga, Spain, [4]Swedish University of Agricultural Sciences, Department of Animal Breeding and Genetics, Uppsala, Sweden, [5]Department of Clinical Science, Faculty of Medicine, University of Bergen, Bergen, Norway, [6]School of Microbiology and APC Microbiome Ireland, University College Cork, Cork, Ireland

KEYWORDS

microbiome, machine learning, explainable artificial intelligence, standards, best practices

Editorial on the Research Topic
Microbiome and machine learning, volume II

Microbiomes play a crucial role in various biological processes, ranging from human and animal health to the functioning soil and marine ecosystems that support food production and biodiversity. Understanding how perturbations of these communities can impact their respective environments is essential for making new scientific discoveries and developing practical solutions to improve both human wellbeing and the health of our planet. However, encapsulating the sheer diversity of microbial communities and the intricate web of interactions they establish with other organisms results in vast and complex datasets. Traditional statistical methods often fall short in capturing both the nuances and global summary of these interactions. With its ability to process large datasets and identify intricate patterns, machine learning (ML) provides a powerful solution. Techniques such as neural networks and ensemble learning models are particularly well-suited for this task, enabling researchers to make sense of the multi-layered structures inherent in microbiome data. Nevertheless, the integration of ML in microbiome research has challenges, including input data standardization, heterogenous, noisy, and high-dimensional data as well as interpretability of ML models. Addressing these challenges requires a concerted effort from biologists, data scientists, and computational experts, fostering a collaborative environment where knowledge and techniques can be shared and refined. This is a exactly what we carried out as part of the COST Action ML4Microbiome (CA18131), which is best summarize by publications in the "Microbiome and Machine Learning" volumes in Frontiers in Microbiology. This second volume represents a significant step forward in harnessing the power of artificial intelligence to decode the complex world of microbiomes.

ML4Microbiome key achievements are summarized in D'Elia et al.. In this article, the authors also underscore the importance of ethical considerations when deploying machine learning in microbiome research. Ensuring data

privacy, avoiding biases in algorithmic predictions, and promoting transparency in model development are essential to maintaining public trust and maximizing the societal benefits of these technologies. Papoutsoglou et al. subsequently detailed the technical complexity of applying ML for microbiome research. The review identifies and addresses challenges such as preprocessing, feature selection, predictive modeling, performance estimation, and model interpretation, finally providing a set of recommendations on algorithm selection, pipeline creation, and evaluation to aid in decision-making processes related to microbiome research. An in-depth exploration of data preprocessing methods is provided by Ibrahimi et al.. This article aims to guide both established researchers and those new to the field in selecting appropriate transformation methods based on their research questions, objectives, and data characteristics.

To provide researchers with insights into specific ML resources facilitating microbiome analysis, Marcos-Zambrano et al. categorized ML tools based on the type of analysis they are designed for and the ML algorithms they employ. The focus spans various software tools for feature generation, taxonomic assignment, clustering, binning, and disease classification.

Kumar et al. emphasize the crucial role of metadata in interpreting and comparing microbiome datasets and highlight the need for standardized metadata protocols to fully leverage the potential of metagenomic data. In this paper microbiome data are classified into five types based on the methodology used for their production: shotgun sequencing, amplicon sequencing, metatranscriptomic sequencing, metabolomic measurements, and metaproteomic expression analysis. The significance of metadata in data interpretation and comparison and the challenges in collecting standardized metadata are thoroughly explored.

In the clinical domain, Chang et al. investigated the diagnostic classification and predictive power of four different ML methods for diagnostic screening in myasthenia gravis (MG) using gut microbiome data. The proposed ML model may serve as biomarkers for clinical use and can be applied for non-invasive screening of MG. Zhang et al. present a study that provides valuable insights into the potential impact of gut microbiota on carcinoid syndrome (CS). The article investigates the cause-and-effect relationship between gut microbiota abundance and carcinoid syndrome (CS) through a bidirectional Mendelian randomization study. Murovec et al. present a study aimed to compare microbiome profiles of patients with colorectal cancer (CRC) and colorectal adenomas (CRA) to healthy participants using metagenomic data. The methodology involved extensive analysis using the MetaBakery pipeline, integrating data matrices like microbial taxonomy, functional genes, enzymatic reactions, metabolic pathways, and predicted metabolites. By integrating all layers of information, the study showcased the development of robust prediagnostic methods for colorectal cancer detection.

To analyze microbiome data in the context of identifying biomarkers for colorectal cancer (CRC) Novielli et al. centered their study on leveraging explainable artificial intelligence (XAI). By employing ML techniques, the researchers aimed to classify a cohort of control subjects from those with CRC based on gut microbiota data and demographic information. The study underscored the potential of gut microbiota data within an XAI framework for precise CRC classification. Another study underscoring the importance of combining ML and XAI approaches is presented by Magarelli et al.. In this study, the researchers explored the use of ML algorithms, specifically the Random Forest (RF) classifier, to effectively classify the geographical origin of PDO Mozzarella di Bufala Campana based on microbiota data. The results showed that the RF classifier outperformed other algorithms, achieving high accuracy in discerning the origin of the samples. The study emphasized the critical role of microbiota analysis in ensuring the authenticity, quality, and safety of food products. Another innovative approach of using XAI is presented by Tangaro et al.. This article outlines a comprehensive study protocol for understanding the interplay among human microbiota, volatilome, and disease biomarkers in Behçet's disease (BD). The study design involves a three-phase approach, including a clinical study with control and experimental groups receiving a soluble fiber-based dietary supplement alongside standard therapy, followed by data collection and analysis using gas chromatography, mass spectrometry, and metagenetic analysis to examine microbiota and volatilome composition. The third phase introduces XAI to analyze collected data to identify markers associated with BD, dietary habits and the dietary supplement, aiming to establish correlations between microbiota, volatilome, and phenotypic characteristics. The results demonstrate how the use of XAI algorithms on multi-modal clinical data could revolutionize disease management.

The importance of practical applications of ML in industries, particularly in the fields of probiotics and pharmaceuticals is exemplified in the article by Liu et al., who were able to discriminate between *Bifidobacterium longum* subsp. infantis and subsp. longum by leveraging MALDI-TOF MS and ML techniques. Through the application of logistic regression, RF, and support vector machine, the researchers developed classification models to differentiate between the two subspecies. The RF model emerged as the most effective. Overall, this study underscores the potential of combining MALDI-TOF MS and ML for rapid and precise discrimination of *Bifidobacterium* subspecies essential for product development and quality control, paving the way for microbial identification and classification advancements.

While these comparative method evaluations are indisputably important, the development of new tools for analyzing microbiome data is also pivotal for aiding the rapidly evolving field of microbiome research. Bakir-Gungor et al. present microBiomeGSM that can identify taxonomic biomarkers from metagenomic data using a new grouping, scoring and modeling (GSM) approach. The tool incorporates pre-existing taxonomy information into a ML model to analyze metagenomic datasets associated with different diseases. By focusing on specific taxonomic levels (genus, family, and order), microBiomeGSM aims to identify their associations with diseases and facilitate disease diagnosis.

Another article by Ligeti et al. introduces the ProkBERT model family, a series of genomic language models developed for microbiome applications. By utilizing the novel Local Context-Aware tokenization technique, the ProkBERT models exhibit superior performance in various tasks such as promoter prediction and phage identification for both supervised and unsupervised

tasks. Importantly, the study emphasizes the significance of innovative approaches in leveraging the vast repositories of raw sequence data and navigating the complexities of labeling inconsistencies within the microbiology field.

Murovec et al. finally presents the development and utilization of MetaBakery, an integrated application designed as a framework for executing the bioBakery workflow on metagenomic sequencing data. MetaBakery streamlines the processing of paired or unpaired fastq files, with optional compression, using programs such as KneadData, MetaPhlAn, HUMAnN, and StrainPhlAn, along with integrated utilities. It includes MelonnPan for metabolite prediction and Mothur for calculating microbial alpha diversity. The development and utilization of MetaBakery provide a versatile and well-documented tool for microbiome analysis, offering efficient exploration of changing parameters and input datasets for various biostatistical and ML approaches.

In conclusion, as we continue to push the boundaries of what is possible at the intersection of microbiome science and ML, the potential applications are vast and varied. By bridging these two dynamic fields, we are paving the way for groundbreaking discoveries that have the potential to revolutionize science and improve lives. From enhancing our understanding of microbial ecology to developing novel diagnostic tools and treatments, the research showcased in this volume is a testament to the innovative and interdisciplinary nature of this field.

Author contributions

DD'E: Conceptualization, Supervision, Writing – original draft, Writing – review & editing, Validation. AZ: Writing – review & editing. IM: Writing – review & editing. EB-R: Writing – review & editing. RB: Writing – review & editing. MC: Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. The authors are grateful to all members of COST Action CA18131, Statistical and machine learning techniques in human microbiome studies for their contribution to the COST Action objectives and to COST (European Cooperation in Science and Technology) for the economic support (www.cost.eu).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

frontiers | Frontiers in Microbiology

TYPE Review
PUBLISHED 22 September 2023
DOI 10.3389/fmicb.2023.1261889

OPEN ACCESS

EDITED BY
Dimitris G. Hatzinikolaou,
National and Kapodistrian University of Athens, Greece

REVIEWED BY
Zhenglin Tan,
Hubei University of Economics, China
Eswarappa Pradeep Bulagonda,
Sri Sathya Sai Institute of Higher Learning (SSSIHL), India

*CORRESPONDENCE
Georgios Papoutsoglou
✉ papoutsoglou@csd.uoc.gr
Magali Berland
✉ magali.berland@inrae.fr

RECEIVED 19 July 2023
ACCEPTED 04 September 2023
PUBLISHED 22 September 2023

CITATION
Papoutsoglou G, Tarazona S, Lopes MB, Klammsteiner T, Ibrahimi E, Eckenberger J, Novielli P, Tonda A, Simeon A, Shigdel R, Béreux S, Vitali G, Tangaro S, Lahti L, Temko A, Claesson MJ and Berland M (2023) Machine learning approaches in microbiome research: challenges and best practices. *Front. Microbiol.* 14:1261889. doi: 10.3389/fmicb.2023.1261889

Machine learning approaches in microbiome research: challenges and best practices

Georgios Papoutsoglou[1,2]*, Sonia Tarazona[3], Marta B. Lopes[4,5], Thomas Klammsteiner[6,7], Eliana Ibrahimi[8], Julia Eckenberger[9,10], Pierfrancesco Novielli[11,12], Alberto Tonda[13,14], Andrea Simeon[15], Rajesh Shigdel[16], Stéphane Béreux[17,18], Giacomo Vitali[17], Sabina Tangaro[11,12], Leo Lahti[19], Andriy Temko[20], Marcus J. Claesson[9,10] and Magali Berland[17]*

[1]Department of Computer Science, University of Crete, Heraklion, Greece, [2]JADBio Gnosis DA S.A., Science and Technology Park of Crete, Heraklion, Greece, [3]Department of Applied Statistics and Operations Research and Quality, Polytechnic University of Valencia, Valencia, Spain, [4]Center for Mathematics and Applications (NOVA Math), NOVA School of Science and Technology, Caparica, Portugal, [5]Research and Development Unit for Mechanical and Industrial Engineering (UNIDEMI), Department of Mechanical and Industrial Engineering, NOVA School of Science and Technology, Caparica, Portugal, [6]Department of Ecology, Universität Innsbruck, Innsbruck, Austria, [7]Department of Microbiology, Universität Innsbruck, Innsbruck, Austria, [8]Department of Biology, University of Tirana, Tirana, Albania, [9]School of Microbiology, University College Cork, Cork, Ireland, [10]APC Microbiome Ireland, Cork, Ireland, [11]Department of Soil, Plant, and Food Sciences, University of Bari Aldo Moro, Bari, Italy, [12]National Institute for Nuclear Physics, Bari Division, Bari, Italy, [13]UMR 518 MIA-PS, INRAE, Paris-Saclay University, Palaiseau, France, [14]Complex Systems Institute of Paris Ile-de-France (ISC-PIF) - UAR 3611 CNRS, Paris, France, [15]BioSense Institute, University of Novi Sad, Novi Sad, Serbia, [16]Department of Clinical Science, University of Bergen, Bergen, Norway, [17]MetaGenoPolis, INRAE, Paris-Saclay University, Jouy-en-Josas, France, [18]MaIAGE, INRAE, Paris-Saclay University, Jouy-en-Josas, France, [19]Department of Computing, University of Turku, Turku, Finland, [20]Department of Electrical and Electronic Engineering, University College Cork, Cork, Ireland

Microbiome data predictive analysis within a machine learning (ML) workflow presents numerous domain-specific challenges involving preprocessing, feature selection, predictive modeling, performance estimation, model interpretation, and the extraction of biological information from the results. To assist decision-making, we offer a set of recommendations on algorithm selection, pipeline creation and evaluation, stemming from the COST Action ML4Microbiome. We compared the suggested approaches on a multi-cohort shotgun metagenomics dataset of colorectal cancer patients, focusing on their performance in disease diagnosis and biomarker discovery. It is demonstrated that the use of compositional transformations and filtering methods as part of data preprocessing does not always improve the predictive performance of a model. In contrast, the multivariate feature selection, such as the Statistically Equivalent Signatures algorithm, was effective in reducing the classification error. When validated on a separate test dataset, this algorithm in combination with random forest modeling, provided the most accurate performance estimates. Lastly, we showed how linear modeling by logistic regression coupled with visualization techniques such as Individual Conditional Expectation (ICE) plots can yield interpretable results and offer biological insights. These findings are significant for clinicians and non-experts alike in translational applications.

KEYWORDS

microbiome data analysis, machine learning methods, preprocessing, feature selection, predictive modeling, model selection, AutoML, colorectal cancer

1. Introduction

The microbiome is a highly diverse system that plays a significant role in human health. Its composition and function can vary widely among individuals, and can be influenced by several factors such as host age, lifestyle habits, environmental or nutritional factors. Dysbiosis, or an imbalance in the microbiome, has been linked to a variety of health conditions (Claesson et al., 2017). For example, the gut microbiome is involved in many important physiological processes, including digestion, immune function, and metabolism. Changes in gut microbiota have been linked to several diseases such as inflammatory bowel disease (Glassner et al., 2020), type 2 diabetes (Navab-Moghadam et al., 2017), and colorectal cancer (CRC) (Zeller et al., 2014), as well as to mental diseases such as schizophrenia through the gut-brain axis (Thirion et al., 2023). Microbiome science is now having important implications for drug development and personalized medicine (Behrouzi et al., 2019).

The microbiome research community has traditionally relied on bioinformatic methods in order to solve important challenges such as taxonomic classifications, metagenome assembly and phylogenetic binning (Claesson et al., 2017). The use of ML can further support clinical applications. The most common ML tasks in microbiome research involve disease diagnosis, prognosis or the response to treatment (Brouillette, 2023) based on the taxonomic or functional composition of samples (Ghannam and Techtmann, 2021). Another important task is to predict the response of the microbiome to drug treatments, different dietary interventions or environmental exposures based on its composition (Thirion et al., 2022). Moreover, ML can be used to discover diagnostic or prognostic biomarkers in the microbiome, that is, the informative features (i.e., genes, taxa or functions) that are most strongly associated with a disease, phenotype, environmental variable or treatment response. Biomarkers can, in turn, be used for early detection of a disease, patient stratification, and personalized medicine (Flemer et al., 2017; Cammarota et al., 2020; Ryan et al., 2020; Berland et al., 2023).

A comprehensive overview of the challenges and solutions associated with the application of statistical and ML techniques in human microbiome studies has recently, been provided by the ML4Microbiome COST action[1] (Moreno-Indias et al., 2021). A subsequent review of the applications of ML in human microbiome studies (Marcos-Zambrano et al., 2021) addressed the challenges of microbiome data analysis, and the importance of feature selection in the development of robust and interpretable models.

In this work, we continue in this direction by highlighting the specific issues pertaining to optimization and standardizing of state-of-the-art ML techniques for microbiome data predictive analysis. We define a set of initial Standard Operating Procedures (SOPs) in the form of practical advices, outline areas suitable for automation, and describe processes on how to integrate everything into pipelines. This will facilitate the translational usage of the developed models by clinicians and non-experts. We consider numerous aspects, ranging from tasks, algorithms or combinations of algorithms, hyper-parameters, to performance estimation protocols for disease prediction. We operationalize these pipelines using shotgun metagenomic datasets of gut microbiome and demonstrate the power of automated machine learning techniques (AutoML) in finding the optimal pipeline.

2. ML tasks and associated analysis steps

2.1. Biological, methodological, and technical constraints for data analysts

While predictive modeling using ML has the potential to provide valuable insights to the biology of the microbiome, several challenges and limitations need to be addressed (Table 1). Data preparation, for example, is an essential first step to enable predictive modeling. It consists of the bioinformatic analysis conversion of sequencing reads to tables that quantify genes, operational taxonomic units (OTU) or more recently Amplicon Sequence Variants (ASVs), metagenomic species (MSP), or functional modules. Two main sequencing methods are used to obtain microbiome data, 16S rRNA sequencing and shotgun metagenomics. Both of them have advantages and drawbacks. Profiling microbial communities using amplified 16S rRNA genes involves sequencing this specific gene, which is present in all bacteria, in order to identify and quantify the types of bacteria in a sample. It is a straightforward and cost-effective method to profile the taxonomic composition of a microbial community. The weaknesses of this methodology are (Větrovský and Baldrian, 2013; Poretsky et al., 2014; Tremblay et al., 2015; Khachatryan et al., 2020): (i) its relatively low taxonomic resolution due to the conservation of the target gene, (ii) imprecise taxa quantification due to the bias induced by the PCR amplification step and the variable gene copy number between and within microbial species, (iii) lack of functional information and intra-species and/or intra-population gene heterogeneity. Shotgun sequencing involves sequencing all extracted DNA in a microbiome sample, which allows a higher taxonomic resolution of the microbes species/strains, along with functional information (Brumfield et al., 2020; Durazzi et al., 2021). Analysis using metagenomic species reconstructed from non-redundant reference gene catalogs allows specific identification and quantification of the microbial species (Plaza Oñate et al., 2019). On the other hand, shotgun metagenomics sequencing is a much more expensive technique that generates large and complex datasets, which can be difficult to process, analyze, or interpret (Liu et al., 2021). Shotgun sequencing is also less suitable for samples with relatively low bacterial biomass (e.g., intestinal biopsies), where 16S rRNA sequencing is able to amplify these genes.

The specificity of the generated microbiome data has several implications which depend on the sequencing techniques used: (1) The total reads per sample (or depths of coverage) can vary by orders of magnitude within a single sequencing run. Comparison across samples with different depths of coverage requires specific adjustments that depend on the sequencing technique and the purpose of the analysis. (2) Microbiome data are sparse (excess of zeros in the feature tables) because (i) many species may be present in one individual and not in others (ii) species are present but sub-dominant and not found at the depth of coverage for a given sample. This feature is present in both 16S and shotgun data, but tends to be more severe in shotgun data. (3) This excess of zero renders the statistical distribution of the quantifications far from gaussian and thus hampers the use of

1 https://www.ml4microbiome.eu/

TABLE 1 List of challenges/constraints associated with applying machine learning (ML) approaches to microbiome data.

Challenge/Constraint	Description
Data acquisition and preparation	The process of acquiring and preparing microbiome data for predictive modeling involves bioinformatic analysis to convert raw sequencing reads into quantification (feature) tables. There are challenges associated with the sequencing methods used (16S rRNA sequencing or Shotgun metagenomics). Sequence data and accompanying metadata are often shared only with a bare minimum of detail, which is not always adequate for replication and further exploration.
Variability and sparsity of microbiome data	Microbiome data exhibits high variability in read depths per sample, sparsity (excess of zeros), non-Gaussian distributions and compositionality. The dependency structures among microbial species further complicates analysis.
Preprocessing tasks	Preprocessing tasks such as cleaning, normalization and batch effect correction are crucial for reducing technical biases and rendering data suitable for ML models. Challenges include choosing appropriate threshold filters for read quality and sparsity reduction, selecting normalization methods based on the model's assumptions, and accounting for experimental conditions.
Data dimensionality	Microbiome data is often high-dimensional, with more features (microbial genes or taxa) than samples. This can lead to overfitting and poor generalization, especially with small sample sizes. Feature filtering and selection methods are employed to reduce dimensionality, but different methods can yield different results, and correlated features can hinder selection.
Non-linearity	Several ML models assume a linear relation between response and predictors. Since non-linear relationships may exist both among and between features and the target, the selection of appropriate model is fundamental for analysis.
Interpretability of ML models	While ML models can identify predictive patterns, interpreting these patterns in a biological context can be challenging. Using inherently interpretable models (e.g., decision trees, linear regression) and integrating metadata, environmental data, or functional assays can enhance interpretability. Visualization techniques and explainable AI methods can also aid in understanding the relationships between features and outcomes. Nevertheless, there is usually an interpretability/performance trade-off, by which the most highly performing models are often harder to interpret.
Limited availability of methods and recommendations	There is a limited number of established methods and standardized approaches for tasks such as batch effect correction and feature selection. Further research and consensus are needed to address these limitations and provide more robust solutions.

modeling approaches which assume Gaussianity. (4) In high throughput sequencing, the total read count represents a fixed-size, random sample of the DNA/RNA molecules within the underlying habitat. It is crucial to note that this count is independent of the absolute number of molecules in the sample and is therefore subject to total sum constraints. Consequently, alterations in the abundance of one sequence necessitate compensatory changes in the abundance of other sequences. The mathematical framework for these data types is compositional analysis, however its application to microbiome data and the consequences on ML models is still an active research area. (5) Microbiome data has a complex inter-dependency structure, where the species may interact with each other in many ways, including mutualism, parasitism, commensalism, and competition. For shotgun data, sequenced genes might belong to the same species and as such strongly correlated. Some variables may be correlated which requires special attention for some ML algorithms.

Following microbiome quantification in the form of raw quantification (feature) tables, the first major challenge for predictive modeling is the preprocessing of the tables in order to reduce technical biases and render the data suitable for ML modeling. This is because differences in preprocessing can have a significant impact on the performance of the models and may introduce biases into the analysis. Typical preprocessing tasks involve normalization, cleaning, and batch effect correction. Normalization is needed for reducing technical biases, such as sequencing depth, and for making samples and features comparable. To the latter, the normalization strategy should consider the compositional nature of microbiome data and appropriate transformations should be applied to avoid misleading results (Li, 2015; Odintsova et al., 2017; Calle, 2019). Accordingly, incorrect or absence of scaling can lead to poor performance or even model failure. For example, when a distance metric is used like in Support Vector Machines (SVMs), scaling must be performed. Similarly, Linear Discriminant Analysis or Gaussian Naive Bayes are statistically effective if only the model errors are Gaussian. Modeling approaches based on decision trees, like CART, random forest, boosted decision trees, do not make such assumptions and work comfortably on raw unscaled data as well. Data cleaning, on the other hand, involves removing outlier samples or features with the aim to improve the quality of the data and reduce the impact of the noise in the modeling process. The identified outliers require careful examination before taking the decision to eliminate them. In addition, feature cleaning by low-abundance filtering often improves the performance of ML models and renders more interpretable signatures. However, there is no universal consensus of the threshold filter value to apply. Finally, batch effect correction, or including batch information as a covariate, can help in avoiding spurious associations between microbial features or phenotypes and unmasking true biological variation. This is particularly important in the case of extensive studies that involve samples analyzed at different time points or sequenced in separate runs, as well as meta-analyses comprising multiple independent studies (Goh et al., 2017). To this date, only a limited number of methods exist for this purpose, and there is a general lack of established recommendations for standardized approaches (Dai et al., 2019; Ling et al., 2022; Wang and Lê Cao, 2023).

Another major challenge is data dimensionality. Microbiome data is high-dimensional, meaning that there are often many more features than samples, which can lead to overfitting and poor generalization of performance. Feature selection and prevalence/abundance filtering methods can help to reduce the dimensionality of the data and select the most informative features for ML models. However, filtering methods do not remove redundant features. Similarly, different feature selection methods can optimize different objective functions, which

may be distinct from the objective functions used in the ML models, and can result in different sets of selected features. Highly correlated features also hamper the selection of the relevant features. All these factors may negatively impact the performance of the model. Furthermore, not all feature selection methods are able to scale up to the thousands of microbial genes or taxa present in different individuals. Along the same lines, not many ML algorithms can scale down to low sample sizes. Low sample size can limit the statistical power and generalizability of ML models. Special care must therefore be given to the performance estimation protocol used during training for the best predictive model. The bootstrap bias correction method, for example, is such an approach equipped to provide better results than traditional cross-validation methods particularly at low sample sizes, also reducing the variance in the estimates of model performance (Tsamardinos et al., 2018).

A final challenge is that although ML models can identify predictive patterns in the data, it is often difficult to interpret these patterns in a biological context. This can limit the utility of ML for generating hypotheses and guiding experimental research. One way to ensure interpretability is to choose predictive modeling algorithms that are inherently interpretable, such as decision trees, logistic regression or linear SVMs. These models have an intuitive connection between the input and the output making it easier to understand the relationship between the microbiome features and the outcome. However, there is usually a performance/interpretability trade-off in ML, by which more complex models (ensembles of trees, neural networks) show better predictive power, but their outputs are also harder to interpret. Another way to improve interpretability is by the combined use of feature selection and the integration between metadata, environmental data, or functional assays, to encourage the model to use a smaller number of features, making it more interpretable and at the same time provide a comprehensive understanding of the microbial community. Dimensionality reduction methods such as sparse Partial Least Squares regression (PLS) are highly interpretable and also provide a visual representation of the data and the model's predictions. Explainable AI techniques such as feature importance, partial dependence plots, and SHAP values can also help to explain the model's predictions and how they are influenced by the input features (Lê Cao et al., 2009).

2.2. ML steps, and appropriate algorithms to use

Once data has been collected and prepared for analysis, the typical process of building an ML model able to predict an outcome of interest consists of three consecutive steps: data preprocessing, feature selection and predictive modeling (Figure 1). For each of these steps there are several methods to consider so the optimal choice depends on the biological, methodological, and technical constraints of microbiome data (Table 2).

2.2.1. Data preprocessing

Regarding data preprocessing, one needs primarily to consider how to normalize the data to enable biologically meaningful comparisons between samples or features. Normalization methods try to eliminate the variability in sampling depth and the sparsity of the data. Rarefying has been a widely used normalization method, especially for 16S rRNA data, in cases where there are significant differences in the library sizes (e.g., more than 10-fold) (Pereira et al., 2018). However, rarefying may not always be an ideal choice since it can reduce statistical power depending on the amount of samples being removed and it does not address all challenges of compositional data (McMurdie and Holmes, 2014).

Alternatives to rarefying are scaling and transformation. However, these are not recommended to be used at the same time, as this practice can invalidate the data, e.g., rescaling may preserve the original distributions but transformation may not (Lovell et al., 2015). Scaling involves finding a sample-specific factor, i.e., a fixed value or proportion, to multiply the matrix counts. Transformation methods, on the other hand, will replace values with the normalized ones. Several scaling approaches have been proposed based on the total sum, trimmed mean (Robinson and Oshlack, 2010), geometric mean (Love et al., 2014), upper quartile or a data-driven threshold (Paulson et al., 2013). But choosing the most effective one is difficult (McMurdie and Holmes, 2014; Weiss et al., 2017; Pereira et al., 2018; Lin and Peddada, 2020) because of the possible over- or under-estimation of fraction of zero counts and distortion of feature correlations across samples due to the data sparsity and differences in sequencing depths. Similarly, there are several transformation methods for microbiome

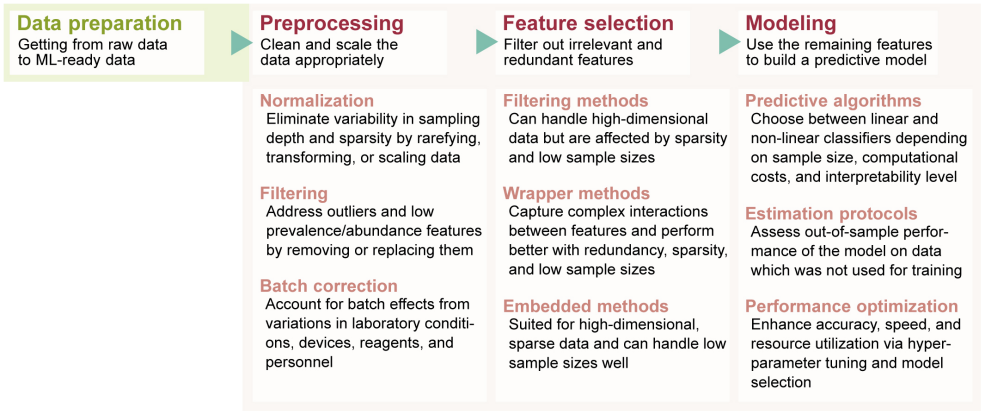

FIGURE 1
The typical process from data preparation to predictive model building, highlighting the methods to consider during each stage.

TABLE 2 Summary of machine learning (ML) algorithms for each workflow step.

Workflow step	Task	Algorithms
Data preprocessing	Handling outliers	Identify outliers by graphical methods (distribution or dimensionality reduction plots) or by statistical methods (Z-score).
		Investigate the cause of the outliers. if they are due to measurement errors or sample contamination, they should be removed.
	Filter out non-informative features	Threshold filtering, variance filtering or correlation-based filtering.
	Normalization	Rarefying.
		Scaling (different approaches: total sum, trimmed mean, geometric mean, upper quartile or data-driven threshold).
		Transformation (additive, centered or isometric log-ratio transformation).
	Batch correction	ComBat, limma, RUV, and PLSDA-batch.
Feature selection	Identify the most informative genes, taxa or functions	Filter methods: supervised (e.g., based on correlation, mutual information or ANOVA), unsupervised (e.g., based on dispersion and similarity measures).
		Wrapper methods: e.g., Recursive feature elimination (RFE), Statistically equivalent signatures (SES) or genetic algorithms.
		Embedded methods: feature selection during the model training process incorporating techniques such as Least absolute shrinkage and selection operator (LASSO) or Elastic net regularization.
Predictive modeling	Classification	Linear classifiers: logistic regression, linear discriminant analysis, partial least squares discriminant analysis (PLS-DA).
		Non-linear classifiers: SVMs, decision trees, random forests, artificial neural networks, gradient boosting, kernel PLS-DA.
	Performance estimation protocols: evaluate the quality of a predictive model	Holdout method: typically 70/30 split.
		K-fold Cross Validation protocol.
		Monte Carlo cross validation.
	Handling class imbalance	Stratified K-fold Cross Validation.
		Oversampling the minority class: random oversampling, synthetic oversampling.
		Undersampling the majority class: random undersampling, heuristic or learning models that try to find redundant examples for deletion.
		Class weighting.
	Optimization metrics	Threshold-independent measures: area under the receiver operating characteristic curve (AUROC), and area under the precision-recall curve (AUPRC).
		Threshold-dependent measures: accuracy, balanced accuracy, f1 score, Matthew's correlation coefficient (MCC).
Model selection	Hyper-Parameter Optimization (HPO) or Combined Algorithm Selection and HPO (CASH)	Optimization techniques: random search, grid search, Bayesian optimization, and evolutionary algorithms. Early stopping, model checkpoints
Model interpretability	Explainable artificial intelligence (XAI)	Global explainer: feature importance (e.g. permutation feature importance).
		Local explainer: Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP).
		Individual Conditional Expectation (ICE) plots.

data. Since microbiome datasets are essentially compositional, these methods follow Aitchison's methodology for compositional data to convert the feature counts to log-ratios within each sample by applying an additive, centered or isometric log-ratio transformation (Aitchison, 1982; Egozcue et al., 2003). Again, a reference feature (gene, taxon, or function) is sought to be used for eliminating the effect of the sampling fraction. However, one should be cautious on how to replace the zeros during calculations for which there is no clear consensus. For example, zeroes can be eliminated by either incorporating a pseudo count or imputing values using probability or zero-inflated models. However, both approaches pose certain challenges. The addition of a pseudo count may artificially inflate low-abundance features while imputation

introduces artificial values, which can distort the true distribution of the data and potentially obscure genuine biological variation (Kaul et al., 2017; Lubbe et al., 2021).

> *Practical advice*: *Normalization is problem- and algorithm-dependent. It is advisable to experiment with different types of normalization as part of the ML pipeline, and select the one who works the best. It is also important that some ML algorithms whose representation learning typically has a distance measure, might need specific types of normalization to work (e.g., SVMs require standardized data*).

Apart from normalization, cleaning the data from outliers and unnecessary features is an essential preprocessing step to consider. Handling outliers requires to identify them typically by graphical methods (e.g., distribution or dimensionality reduction plots) or statistical ones (e.g., Z-score). For example, sanity checks for outlier detection include Principal Component Analysis of microbiome data that have been normalized. Subsequent ML tasks might also encompass the use of robust methods downweighting outlying observations during the estimation procedure (Kurnaz et al., 2018; Monti and Filzmoser, 2022; Kurnaz and Filzmoser, 2023). The cause of their outlierness must then be investigated. Outliers may be due to measurement errors, sample contamination, or biological variation. Understanding the cause can help determine the appropriate approach to handling outliers. If outlier samples are due to measurement errors or sample contamination, it may be appropriate to remove them from the dataset.

In terms of non-informative features or taxa that are biologically irrelevant or known contaminants, filtering can be based on thresholds on their abundance/prevalence, variance or correlation. Low-abundance or prevalence filtering involves eliminating features present in less than, e.g., 10% of the samples (Cao et al., 2021). Variance filtering involves removing features of zero or low variance across the samples as they are less likely to contribute to the overall variation in the data and may be less informative. The threshold for variance filtering can be determined based on the distribution of the variance in the data. Finally, based on the correlation coefficient or the mutual information between features, one can detect and filter out those that are highly associated with each other as they are redundant and may not provide additional information.

> *Practical advice*: *Perform exploratory data analysis that includes visualization of observations in a reduced dimension subspace for the inspection of outliers, correlation maps for the identification of highly correlated features, and descriptive statistics for the inspection of missing values and non-informative taxa*.

After normalization and cleaning, one can perform batch correction to account for batch effects that may arise due to technical factors such as sequencing platform, library preparation, or batch processing. Several batch effect correction methods have been proposed in the literature, mainly for RNAseq and microarray data, such as ComBat, limma, and RUV that can be used to correct for batch effects in the microbiome domain (Wang and LêCao, 2020). Very recently, a new and effective approach for correcting batch effects called PLSDA-batch has been presented that can effectively correct for batch effects and improve the accuracy of downstream analyses (Wang and Lê Cao, 2023). Regardless of the chosen batch adjustment method, however, it is important to consider the statistical assumptions of the method, such as Gaussianity. It is possible that the data transformation applied prior to the batch adjustment may not satisfy these assumptions.

> *Practical advice*: *Before generating sequence data, make sure that samples are randomized so that whole study groups do not end up in separate batches. Visually inspect the post-sequencing impact of all possible batch effects on samples' distribution in a space of reduced dimension and in subsequent ML model performance, and if needed correct for this using recent appropriate tools*.

2.2.2. Feature selection

After preprocessing, feature selection is the next important step for microbiome data analysis in order to identify the most informative genes, taxa or functions. In principle, feature selection is the process of selecting a subset of features from a larger set of available features that are most important in predicting the outcome variable. The goal is to reduce the number of input features required to achieve good model performance, thereby improving the efficiency and interpretability of the model. This also helps to avoid overfitting, a common problem when analyzing high dimensional datasets where the model becomes too complex and starts to memorize the training data instead of learning general patterns.

Feature selection techniques can be broadly categorized into three types: filter methods, wrapper methods, and embedded methods. Filter methods can be either unsupervised (e.g., using dispersion and similarity measures) or supervised (e.g., based on correlation, mutual information or ANOVA), the latter evaluating the relevance of features to the outcome variable (Ferreira and Figueiredo, 2012), which can then be used to select the top-ranked features (Segata et al., 2011). Wrapper methods, such as recursive feature elimination, statistically equivalent signatures or genetic algorithms, employ statistical metrics too. However, they do so in combination with a predictive algorithm so as to select features based on their impact on the model's performance (Lagani et al., 2017; Sanz et al., 2018). Embedded methods perform feature selection during the model training process by incorporating regularization techniques, such as L1 or L2 regularization, that automatically penalize the less important features. Nevertheless, there is still no consensus on which feature selection method should be used (Marcos-Zambrano et al., 2021).

The choice of the appropriate feature selection method remains an open problem because microbiome data poses numerous analysis challenges, such as noise, high dimensionality and small sample sizes, sparsity, and intercorrelated or redundant features. Filter methods can handle high-dimensional data relatively well, but they may not perform well in the presence of sparsity or low sample size. For example, correlation-based methods may suffer from false positives when the correlation is driven by sparse features or may select one feature from a correlated pair, resulting in suboptimal feature selection. Similarly, ANOVA may have low statistical power with few samples and may select redundant features that do not contribute additional information beyond what is already provided by other features. Wrapper methods, on the other hand, can better capture the complex interactions between features and may perform better than filter methods in the presence of redundancy, sparsity or low sample

size. Their main drawback is that they can be computationally intensive and may not scale well to high-dimensional data. Embedded methods, such as least absolute shrinkage and selection operator (LASSO) or Elastic Net regularization, are well-suited for high-dimensional, sparse data and can handle low sample size relatively well (Tibshirani, 1996). These methods can perform both feature selection and regularization during model training, and can often identify a small number of highly relevant features that capture the underlying patterns in the data.

Regarding linearity, most filtering methods rely on linear statistical models to rank and select features based on their association with the response variable. Similarly, embedded methods typically use linear models to embed the features into a lower-dimensional space or to fit a regression model that selects the most informative features. To capture more complex relationships, mutual information is such a non-linear measure of the association between two random variables that can be used as a filter method. For example, the Minimum Redundancy Maximum Relevance (mRMR) selects features that have the highest mutual information with the target variable and the lowest mutual information with the previously selected features (Chen et al., 2016). Accordingly, any wrapper method that embeds a non-linear statistical metric can be used for capturing complex associations among features.

Lastly, certain feature selection techniques are inherently stochastic, implying that they may return different results in successive runs. Consequently, it is recommended to run each algorithm containing random elements multiple times, to obtain a more accurate understanding of its predictions. Alternatively, fixing the random seed in each run is guaranteed to provide consistent and deterministic results.

> *Practical advice*: *Consider testing a conservative filter method as a pre-screening stage in the feature selection task, or a more expensive multivariate method (e.g., embedded) to remove irrelevant and non-informative features in high-dimensional datasets. Another good general practice is to consider the objective function which is used in the feature selection and try to match it with the objective of the subsequently chosen modeling approach. For instance, feature selection based on Fisher score is suitable for linear discriminants, PCA is a good dimensionality reduction routine for Gaussian mixture models, recursive feature elimination is applicable for non-linear SVMs, etc.*

2.2.3. Predictive modeling

Lastly, the task of modeling involves selecting a predictive algorithm, a protocol for performance assessment, a protocol for model selection, and a metric for optimizing that performance. The choice of algorithm mainly depends on the problem type and the data characteristics. In the microbiome domain, classification problems are the most prevalent although efforts to address survival ones also exist. Regarding data characteristics there are several types of modeling algorithms, each having its strengths and weaknesses. Below, we explore methods that can handle challenges related to microbiome data such as scalability to high-dimensional data and small sample sizes, as well as interpretability. This will include both linear and nonlinear classifiers commonly employed and methods to estimate their performance.

Logistic regression and linear discriminant analysis, for example, are linear classifiers that can handle high-dimensional data, but they may be sensitive to overfitting when sample sizes are small (e.g., less than 100 per class). They may thus be good choices when the data is not too complex and the sample sizes are not too small. Partial Least Squares Discriminant Analysis (PLS-DA) is a good option for high-dimensional data with low sample sizes and benefits from multicollinearity, although care must be taken to avoid overfitting. On the other hand, SVMs are mainly nonlinear classifiers that can handle high-dimensional data but can be computationally expensive when the number of features is very large. Hence, they may be a good choice when the data is more complex, but the computational cost may be an issue. In contrast, decision trees and random forests are nonlinear classifiers that can handle high-dimensional data and are relatively robust to small sample sizes. However, they suffer from overfitting and instability when the trees are too deep or the data is noisy. Artificial neural networks and gradient boosting are also nonlinear classifiers that can handle high-dimensional data and are relatively robust to small sample sizes, but can be computationally expensive. Careful hyperparameter tuning is therefore important to avoid overfitting.

Unfortunately, due to the curse of dimensionality and the unknown patterns in the data, one cannot provide specific guidance on choosing a predictive modeling algorithm based on the number of features. Moreover, the so-called No Free Lunch Theorem in machine learning states that there is no single "best" method that can universally excel in solving all types of problems. The selection of an appropriate algorithm needs to consider the specific characteristics and constraints of the task at hand. Nevertheless, a combination of feature selection and a suitable performance estimation protocol can enhance a classifier's performance in a high-dimensional setting (Wolpert, 2002). If interpretability is an important consideration, logistic regression, PLS-DA or decision trees are highly interpretable, while SVMs and artificial neural networks may be less so. Essentially, if feature selection is performed and techniques such as feature importance measures and visualization are used, insights into the behavior of even the most complex models can be gained.

> *Practical advice*: *For high-dimensional scenarios, as is the case with microbiome data, the choice of the model must consider the sample size and the desired computational cost and interpretability level. Coupling with a feature selection algorithm may improve prediction accuracy.*

2.3. Building and evaluating ML workflows

2.3.1. Performance estimation protocols

Performance estimation protocols are methods used to evaluate the quality of a predictive model. Their main purpose is to estimate the performance of the model on new, unseen data called out-of-sample performance or generalization error–the error that the model will obtain if hypothetically tested on the unseen data of infinite size. Estimation of the performance should not be confused with improving the performance which is the purpose of the model selection routine. The simplest protocol for performance assessment is the holdout method which involves splitting the available data into two parts, a training set and a test set; typically, a 70/30 split is used. The model is

trained on the training set, and its performance is evaluated on the testing set. Holdout is suitable when the available data is sufficiently large, and the number of features is not too high relative to the number of samples. If the sample size is low, the performance on the test set will have a large variance. One way to reduce the variance is by repeating this protocol, each time by randomly assigning samples to training and test sets, and estimating the average model's performance.

The well-known K-fold Cross-Validation protocol can be used for that. It involves dividing the data into K mutually exclusive, equally sized sets, or folds. Each time the model is trained on all folds but one that is held out for estimating the performance. If the sample size is very low, this protocol may be repeated several times with a different partitioning to folds, to further reduce the estimation variance. A typical value for the number of folds is 5 or 10, but this can be adjusted depending on the size of the data. In a similar manner, Monte Carlo cross-validation can be suitable when the available data is sufficiently large, and the number of features is high relative to the number of samples. This method can be useful when the data is noisy or there is a high degree of variability in the data, as it allows for multiple random splits to be generated. Taking K-fold Cross-Validation to the extreme, one can perform leave-one-out (LOO) cross validation, where all but a single sample is used for training and the performance is assessed on the remaining sample and averaged across all samples. LOO is known to be an almost unbiased performance assessment routine (Vapnik, 2006). The main advantage is its repeatability that comes from the deterministic nature of the routine. However, it can be a time-consuming process when the number of samples is larger.

A common characteristic of all protocols is that they use a portion of the data to train a model and the rest to evaluate its out-of-sample performance. In cases where the samples are plenty, losing some part of the data to estimation is acceptable. If not, as in the microbiome case, finding the right balance between training and test data is essential. Obviously, the best predictive model is—on average, not always—the one trained on all available data. However, since there is no more data left, how does one estimate its performance? The answer is to use one of the aforementioned protocols, i.e., evaluate the model performance on some partitioning protocol but train the final model on all available data. This process is called the "Train-Test-Retrain" procedure and presents a big change in perspective because it uses the performance of a suboptimal model as a proxy for the performance of the full model (Tsamardinos et al., 2022). As a result, the estimate is conservative, which is better than being overly optimistic. Essentially, during performance estimation we are not evaluating a specific model instance, but the entire ML pipeline that produces the final model.

Lastly, a typical methodological problem in predictive modeling is that of data leakage which can lead to optimistic or entirely invalid models. Data leakage occurs when performing data preprocessing or feature selection on the whole dataset before applying cross-validation. For example, when standardizing the data using the mean and standard deviation of the entire dataset, the rescaling process gains knowledge of the full data distribution, introducing bias on the rescaled values that can affect the performance of the algorithms on the cross-validation test sets. To avoid data leakage, therefore, the preprocessing, feature selection and predictive modeling must be performed together within each fold of the cross-validation and only apply them to the test fold on each cycle, ensuring the integrity of the evaluation process.

> *Practical advice*: *Evaluate the entire ML pipeline with cross-validation. For small sample sizes (e.g., 100 per class) use a Stratified, Repeated K-fold Cross Validation, of 4–5 repeats, with retraining on all data to produce the final model with a maximum K the number of samples in the rarest class so that at least one sample from each class gets into each fold.*

2.3.2. Class imbalance

A data characteristic that often appears in the microbiome domain is class imbalance where the number of samples in one class is much smaller than the number of samples in the other classes. Class imbalance can be problematic and lead to biased models that underperform on the minority class. One technique to alleviate this is the stratification of samples to cross-validation folds, namely, stratified K-fold Cross-Validation. This entails partitioning the data with the extra constraint that the distribution of the outcome in each fold is close to the distribution of the outcome in all samples. Other ways to compensate for the class imbalance include oversampling the minority class, undersampling the majority class and class weighting. Oversampling methods include random oversampling, where instances from the minority class are randomly duplicated, and synthetic oversampling, where new instances are in-silico synthesized from existing ones of the minority class, referred to as data augmentation, e.g., SMOTE (Chawla et al., 2002). General concern with oversampling is the increase of likelihood of overfitting due to exact or synthetic copies of the existing data (Fernández et al., 2018). Undersampling methods include random undersampling, where instances from the majority class are randomly removed, and methods that involve heuristics or learning models trying to find redundant examples for deletion or useful examples for non-deletion. However, removing too many samples from the majority class can be a problem, especially if the dataset is small. Oversampling and undersampling techniques can potentially enhance model performance when applied either as preprocessing steps or as integral components of the model itself (Mihajlović et al., 2021).

On the other hand, class weighting regards the assignment of weights to the classes to balance their contributions to the loss function during training. By assigning higher weights to the minority class, the algorithm can redistribute its capacity to focus more on correctly predicting the minority class, thus improving the overall performance on the imbalanced dataset. However, finding the right weights can be quite challenging. This strategy, also known as cost-sensitive learning strategy encourages the model to focus on correctly predicting the minority class, as misclassifying instances of this class incurs a higher cost (Ling and Sheng, 2010).

Regardless of how class imbalance is approached when evaluating the performance of a model on an imbalanced dataset, it is important to use appropriate performance metrics that consider the imbalance (e.g., balanced accuracy; averaged versions of precision, recall, F1-score etc.).

> *Practical advice*: *Data stratification during performance estimation and appropriate choice of performance metric should be practiced. Test several over/under-sampling options is suggested but always validate the similarity between synthetic samples and actual data. Alternatively consider class weighing or cost-sensitive methods.*

2.3.3. Performance metrics

Performance measures play a crucial role in evaluating and quantifying the predictive capabilities of classifiers. For example, threshold-independent measures like Area Under the Receiver Operating Characteristic Curve (AUROC or AUC) and Area Under the Precision-Recall Curve (AUPRC) are advantageous in assessing overall classifier performance. The AUC quantifies the ability of a classifier to discriminate between positive and negative instances across all possible decision thresholds. In other words, it measures the classifier discriminative capacity. Intuitively, AUC denotes the probability that a randomly chosen positive instance is ranked by a classifier higher than a randomly chosen negative instance. AUC is robust to moderate class imbalance and useful when the relative costs of false positives and false negatives are equal (Bewick et al., 2004). In contrast, AUPR focuses on the precision-recall trade-off and is particularly useful in imbalanced datasets, when the positive class is of greater interest. It denotes the probability of correct detection of positive instances (Saito and Rehmsmeier, 2015). Threshold-dependent measures, on the other hand, assess classifier performance at a specific decision threshold between 0 and 1. Threshold tuning does not change the classifier quality but can improve the performance metric, while also being one of the simplest approaches to handle a severe class imbalance (Fernández et al., 2018). Accuracy, for example, calculates the proportion of correctly classified instances over the total number of instances. However, it can be misleading under class imbalance, as it may achieve a high accuracy score by simply predicting that all observations belong to the majority class (Akosa, 2017). In contrast, balanced accuracy measures the average accuracy obtained from both the minority and majority classes. However, it treats all misclassifications equally and does not provide information about the performance of the classifier on individual classes. The F1 score is defined as the harmonic mean of precision and recall, which considers both false positives and false negatives. Nevertheless, F1 score does not capture true negatives, which can be crucial in certain applications. In contrast, Matthew's Correlation Coefficient considers all four outcomes of a binary classification, true positive, true negative, false positive, and false negative rates. This is especially useful when the class distribution is imbalanced or when the costs associated with different types of errors vary (Chicco and Jurman, 2020).

> *Practical advice*: ***Selecting an appropriate performance metric depends on the specific requirements of the task, the prevalence of class imbalance, and the trade-offs between different types of classification errors. Although AUC is widely used, different metrics highlight different performance aspects. Using multiple ones may help in analysis and better understanding of the classifier performance.***

2.3.4. Hyperparameter tuning

Several different candidate algorithms should typically be tried for each of the analysis steps based on the aforementioned factors to find the optimal ML pipeline. Nonetheless, each algorithm comes with several settings, referred to as hyper-parameters, that need to be set before training. Examples of hyperparameters include the learning rate of a neural network, its early stopping or model checkpoint parameters, the regularization strength of a linear model, or the depth of a decision tree. Optimizing for these choices is called Tuning, or else, Hyper-Parameter Optimization (HPO) or Combined Algorithm Selection and HPO (CASH) (Thornton et al., 2013; Feurer et al., 2015). In a nutshell, HPO selects the best hyperparameter values to achieve optimal performance while CASH involves selecting the best machine learning algorithm and its hyperparameters. CASH aims to automate the process by searching over a large space of possible algorithm and hyperparameter combinations. This is particularly useful when there is no clear choice of algorithm, or when the performance of different algorithms is highly dependent on the choice of hyperparameters.

Both HPO and CASH require training and evaluating many different ML pipelines with different hyperparameters or algorithms. To this end, various optimization techniques have been proposed, such as random search, grid search, Bayesian optimization, and evolutionary algorithms, among others. These techniques aim to efficiently search the hyperparameter or algorithm space to find the best combination that optimizes the desired performance metric. Random and grid search are simple to implement and parallelize but can be inefficient for high-dimensional search spaces. Bayesian optimization and evolutionary algorithms are more efficient, can use past evaluations to guide the search and also handle non-continuous and non-convex search spaces. Evolutionary algorithms can also search for multiple optima but can be computationally expensive. The downside of Bayesian optimization is that it requires a well-defined prior over the search space and can be sensitive to the choice of function to determine the next set of hyperparameters to evaluate.

2.3.5. Model selection process while tuning

When trying multiple ML pipelines, it is tempting to select as best the one with the highest estimated performance. Practitioners sometimes confuse or mix up the error estimation process with the error reduction process. The performance assessment aims to estimate the error while model selection aims to reduce the error. When these procedures are mixed up a selection bias occurs leading to the respective performance estimate becoming compromised (usually over-optimistic). This problem is called the "winner's curse" and is conceptually equivalent to the multiple hypothesis testing problem in statistics (Jensen and Cohen, 2000). Essentially, this phenomenon occurs since each performance protocol simulates an ideal scenario by pretending that the test sets come from the future, but in reality, these test sets are used to select the winning model and thus the process that aimed to estimate the performance is now used to improve it. This problem becomes more pronounced in low sample sizes, where the optimism could be as much as 20 AUC points (Ding et al., 2014; Tsamardinos et al., 2014). Therefore, appropriate performance estimation protocols should be used to correct for the winner's curse.

The simplest solution to this problem is to hold out a second set of samples to be used for model selection. That is, extend the Train-Test protocol into the Train-Validate-Test protocol. The samples in the Validation may be used several times, but only for selecting the best model while those in the test set are used once, for performance estimation. Obviously, as before, this procedure is preferable when the sample size is large. In cases of low sample sizes, several alternatives have been proposed such as the nested cross validation (Salzberg, 1997), the Tibshirani-Tibshirani procedure (Tibshirani and Tibshirani, 2009) and the Bootstrap bias corrected cross validation (BBC-CV) (Tsamardinos et al., 2018) among others (Ding et al., 2014).

The nested cross validation involves a double loop procedure, where an inner cross-validation loop is run over the training data and

is used for hyperparameter tuning, and an outer one for estimating the performance. Although nested cross-validation is very useful when the dataset is small and the number of hyperparameters is large it is computationally very expensive. The Tibshirani and Tibshirani method does not employ a separate hold out set. Rather it employs traditional K-fold cross-validation estimates to calculate the bias and subtract it from the performance estimates. A similar, but computationally more efficient method that has smaller variance and bias, is the BBC-CV method that was recently presented (Tsamardinos et al., 2018). Here, in order to calculate the bias, bootstrap resampling is employed on the pooled out-of-sample estimates collected during cross-validation of multiple pipelines.

> *Practical advice*: *Combined Algorithm Selection and HPO allows finding the optimal ML pipeline when this combines different algorithms and hyperparameters. Start exploring the space by grid or random search, and always correct for the "winner's curse." If the sample size is sufficient, use nested-CV due to its simplicity of implementation, otherwise use BBC-CV.*

2.3.6. AutoML: challenges and best practices

The above information suggests that implementing a complete machine learning workflow typically requires a substantial amount of skilled manual effort. In addition to being time-consuming, it also requires an expert to make informed decisions about which methods to incorporate into the pipelines. However, the lack of such experts and the associated high costs have paved the way for the emergence of automated machine learning (AutoML) (Hutter et al., 2019). AutoML aims to automate various stages of the machine learning process, including data preprocessing, feature selection, model training, hyperparameter tuning, and model evaluation. By doing so, AutoML enables objective and data-driven analysis decisions, resulting in high-quality models that can be utilized even by inexperienced users (Xanthopoulos et al., 2020).

AutoML is frequently used synonymously with the aforementioned CASH and HPO approaches that focus on solving a particular optimization problem. However, these solely aim to deliver predictive models and do not encompass the entire machine learning workflow necessary for microbiome data analysis. While various AutoML systems such as the well-known auto-sklearn (Feurer et al., 2022) or GAMA (Gijsbers and Vanschoren, 2019) exist, only TPOT (Olson and Moore, 2019) and JADBio (Tsamardinos et al., 2022) have the capability to extend their functionality to include the feature selection step. Notably, JADBio goes even further by encompassing all the necessary steps, including the estimation of out-of-sample predictive performance, which most AutoML systems do not automate, thereby providing a comprehensive solution for the ML analysis of microbiome data.

While AutoML offers significant advantages by automating various steps of the machine learning workflow, it may also have certain challenges. Firstly, AutoML may lack transparency, making it challenging to understand and explain the underlying decisions made by the automated processes. This opacity can limit the ability to detect and address biases or errors. AutoML tools may also have limited customization options, as they are designed to cater to a wide range of users and tasks, restricting flexibility and domain-specific adaptations. Furthermore, it can increase computational cost due to extensive model exploration and lastly, relying solely on AutoML can diminish the essential role of human expertise and domain knowledge, which are crucial in understanding the context, interpreting results, and making informed decisions. It is therefore essential to strike a balance between the advantages of automation and the need for human involvement (Gijsbers et al., 2019; Romero et al., 2022).

> *Practical advice*: *AutoML is becoming increasingly popular, but most approaches primarily focus on solving the CASH problem to provide an optimal predictive model. As a result, researchers still need to decide on performance estimation methods and protect against the "winner's curse".*

2.4. Model interpretability and explainability of results

Model explainability involves understanding how algorithms learn the relationship between inputs and outputs. In classification models, there are three main goals: to create an accurate model, to accurately estimate how good the model is and interpretability. However, there is often a tradeoff between these objectives whereby linear models are interpretable but may underperform compared to nonlinear models. Complex nonlinear models achieve better performance but are less interpretable. This lack of interpretability limits their use in biomedical research where understanding the classification process is crucial.

For this reason, explainable artificial intelligence is a growing field that focuses on explaining the output or decisions of ML models (Carrieri et al., 2021; Lombardi et al., 2021, 2022; Bellantuono et al., 2022). One prominent technique in this respect is the measurement of feature importance. Feature importance methods aim to quantify the contribution of each feature to the model's predictions. Particularly, global methods provide an overall ranking of features while local methods try to explain the contribution of each feature to a specific prediction. For example, permutation importance is a global method that evaluates importance by disrupting the relationship between the feature and the true outcome. The underlying concept is simple: if permuting a feature's values results in higher prediction error, it indicates its importance. Conversely, if permuting the feature does not affect the error, it is classified as unimportant. Regarding local methods, Local Interpretable Model-agnostic Explanations (LIME) is a technique that approximates model behavior with an interpretable (linear) model at the neighborhood around each individual prediction (Ribeiro et al., 2016). Similarly, SHapley Additive exPlanations (SHAP) is a local explainer algorithm that uses a concept from game theory called Shapley values (Lundberg and Lee, 2017). Shapley values measure how much each feature contributes to the prediction by considering all possible combinations of features in a fair share manner. SHAP can work with any kind of model and can show the impact of each feature visually. Finally, some methods combine a stepwise forward strategy to identify a minimal subset of interpretable variables from a permutation-based score of importance (Genuer et al., 2015).

Individual Conditional Expectation (ICE) plots also provide a way to explore and understand the relationship between a specific input feature and the output of a model, while considering the influence of

other features (Tsamardinos et al., 2022). In an ICE plot, the x-axis represents the range of values for the chosen input feature. Each line in the plot corresponds to how the model prediction changes while varying all the remaining input features. In this way ICE plots help identify non-linear patterns, interactions, and heterogeneity in the model's behavior across instances, aiding in model interpretation at the individual level.

> *Practical advice*: ***Start with a simple, interpretable model; more complex models can be used to achieve better performance, for which model explanation techniques can be used, such as calculation of feature importances, LIME, SHAP values, and ICE plots.***

3. Comparative evaluation of ML approaches

To showcase the effectiveness of various ML approaches in enhancing predictive performance, we collected a set of CRC benchmark data on a two-class (healthy/cancer) classification problem. To this direction, we first evaluated the effect of typical preprocessing steps such as normalization and filtering. Then we used AutoML, namely JADBio, to find the best performing and best interpretable pipelines in terms of feature selection and predictive modeling.

3.1. Description of the data

This dataset (Barbet et al., 2023) gathers 2090 human stool samples characterized by shotgun metagenomic sequencing from 13 public cohorts spanning nine countries (Table 3). This data provides the gut microbiota composition in healthy controls and patients with adenoma or CRC.

Data were prepared as follows. Sequencing data was downloaded from the European Nucleotide Archive. Reads were quality trimmed and filtered from sequencing adapters using fastp. Remaining contamination by the host genome was filtered out by mapping reads against the human reference genome (T2T-CHM13v2.0) with bowtie2. Microbial species identification and quantification was estimated according to both human gut reference gene catalog (IGC2, 10.4 M genes, Wen et al., 2017) and human oral gene catalog (8.4 M genes, Le Chatelier et al., 2021) with the METEOR software (Pons et al., 2010), and clustered into Metagenomic Species Pangenomes taxonomically and functionally annotated (Plaza Oñate et al., 2019).

TABLE 3 Compilation of datasets from nine distinct countries, including 2,090 human stool samples characterized via shotgun metagenomic sequencing.

BioProject	Country	Nb all	Nb CRC
PRJDB4176	Japan	645	286
PRJEB10878	China	128	74
PRJEB12449	USA	104	52
PRJEB27928	Germany	82	22
PRJEB6070	France	156	53
PRJEB6070	Germany	43	38
PRJEB7774	Austria	156	46
PRJNA389927	USA	56	26
PRJNA389927	Canada	28	2
PRJNA397112	India	110	0
PRJNA447983	Italy	140	61
PRJNA531273	India	30	30
PRJNA608088	China	18	6
PRJNA429097	China	194	98
PRJNA763023	China	200	100
All cohorts	9 countries	**2090**	**894**

3.2. Evaluation of the preprocessing steps

We evaluated the effect of two typical preprocessing steps on the performance of various standard ML algorithms implemented in a caret workflow (Kuhn, 2015): RF—Random Forest, PLS—Partial least square, Earth—spline regression (can be applied to classification also), Pam—Partition around medoids (normally a clustering algorithm), Glmboost—Gradient Boosting with Component-wise Linear Models, Glmnet—Generalized linear model with elastic net penalty, GBM—Gradient boosting machine. The data were split in a training set (75%) used to tune the hyperparameters of the models and a test set (25%) used to evaluate the model performance. The split of the data has been repeated 100 times (Fromentin et al., 2021).

We first applied a fixed threshold on abundance values (retained features with a total abundance across samples >5e-06). A variable threshold of prevalence across samples in [0–0.5], with 0.05 steps was applied to remove features with low prevalence. Figure 2 shows the sensitivity and specificity results for the two best performing models: GBM and RF. We observed that a small filtering slightly improved the performances both on accuracy and computing resources criteria. However, it is noteworthy that no filtering on prevalence at all is also a valid option in terms of performances. As expected, strong filtering on prevalence (>0.15–0.2) decreases the sensitivity for GBM and the specificity for RF. Additional analyses of other microbiome datasets (Supplementary material 1 and Supplementary Figure S1), showed that performance was not affected by 0.2 prevalence filter with regard to 0 prevalence filter in RF models. However, other models such as PLS-DA got better classification error rates when 0.2 prevalence filter was applied. The results from these additional datasets indicate that the effect on performance of the low-abundance filter depends both on the ML model applied and on the characteristics of each dataset, being the level of sparsity of the database a key factor to consider. All in all, this fact highlights the importance of including the low-abundance filter as another hyperparameter to tune while training the model by cross-validation strategies.

Figure 3 shows the sensitivity and specificity results for all the models with or without the CLR logratio transformation before the modeling process. We observed that for the majority of the models, the CLR transformation decreased the sensitivity of the models, and it was particularly striking for the Glmnet and Glmboost models. It only improved the sensitivity for the PLS and Earth models. It improved the specificity of the PLS and Glmboost models, nevertheless, RF and GBM remained the top performing models.

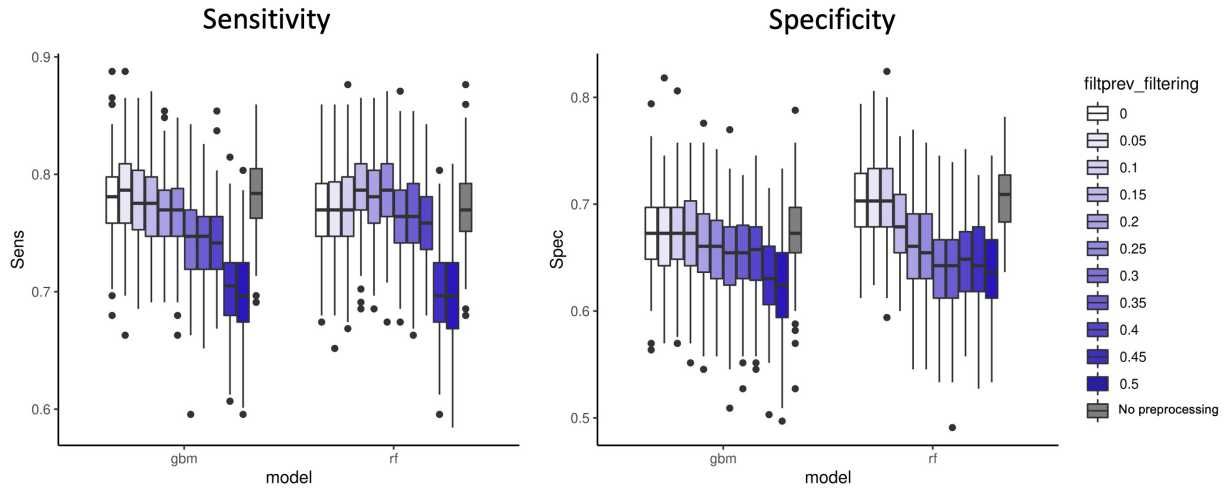

FIGURE 2
Sensitivity and specificity of the two best performing ML models (GBM and RF) on 100 data split repetitions applied on the CRC dataset with a range of filter on prevalence (shades of blue) or no filter on prevalence (gray).

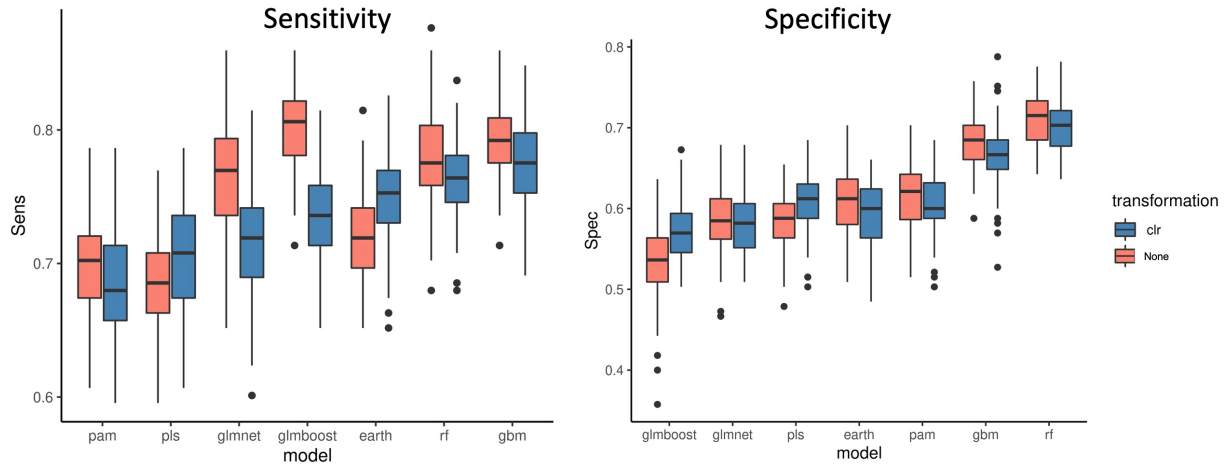

FIGURE 3
Sensitivity and specificity of 7 ML models across 100 data split repetitions applied on the CRC dataset with a CLR logratio transformation before (blue) or no transformation (red).

In contrast, results obtained from other datasets (Supplementary material 1 and Supplementary Figure S1) pointed out that, in general, CLR transformation rendered a better performance when compared to TSS normalization followed by logarithmic transformation. Again, RF was less affected by the type of normalization applied than PLS-DA. Taken together, as with threshold filtering, it is important to cross-validate transformation options in order to enhance the predictive performance and extract the best modeling pipeline.

3.3. Evaluation of feature selection and predictive modeling

To evaluate the performance of different ML pipelines we used the JADBio automl approach. JADBio is specifically designed for biomedical data and is able to fully automate the production of customizable ML pipelines that simultaneously integrate preprocessing, feature selection and predictive modeling algorithms (Table 4). Specifically, for preprocessing, we performed standardization. For feature selection we evaluated LASSO regularized regression (Tibshirani, 1996) and the Statistical Equivalent Signatures (SES) algorithm (Lagani et al., 2017). Both methods can handle the high-dimensionality of microbiome data. Regarding sample size and expected signature size, LASSO tends to perform better when sample sizes are larger but returns a greater number of features. SES, on the other hand, drawing inspiration from causal modeling theory, demonstrates better performance at low sample sizes and leads to smaller feature subsets at the expense of predictive performance. SES also produces multiple signatures that exhibit statistically indistinguishable predictive performances. For modeling, we employed well known linear/interpretable algorithms such as ridge

TABLE 4 Algorithms used for comparative evaluation of ML pipelines.

Algorithm	Used for
Standardization	Preprocessing
LASSO	(Single) feature selection
SES	(Multiple) feature selection
Decision trees	Predictive modeling
Ridge regression	Predictive modeling
Random forests	Predictive modeling
Support vector machines	Predictive modeling
Generalized cross-validation	Performance estimation
Grid search with heuristics	Configuration space search
BBC-CV	Performance correction

regression (Hoerl and Kennard, 1970) and decision trees (Breiman, 2017) as well as non-linear ones such as random forests (Breiman, 2001) and SVMs (Chang and Lin, 2011). To find the best model, we employed grid search and the Generalized Cross-validation approach, while the BBC-CV algorithm was used to correct for testing multiple ML pipelines. More details about the architecture can be found in Tsamardinos et al. (2022).

Table 5 summarizes the results from analyzing the collected shotgun data. Initially, our aim was to uncover any classification bias from technical or demographic covariates, through feature selection. Indeed, when we employed JADBio on the complete, all-cohorts, dataset we observed that the cohort ID/country exhibited predictive power, indicating inherent variations in gut species between different countries (Figure 4A). Furthermore, we found that the timepoint of measurement, the instrument model, and the westernization status of samples also possessed predictive value, pointing toward the existence of batch effects (the full description of the metadata features used in the models are provided in Supplementary Table S1).

To address these variabilities, we conducted a series of subsequent analyses by splitting the data into the different cohorts. Feature selection identified that the timepoint feature was predictive in the Japanese cohort, the instrument model in the German cohort and the westernization status in the Indian cohort (Table 5). Therefore, we repeated the analysis on the entire sample set after excluding the problematic samples coming from these cohorts. This time, the revised data we divided into two parts: one for training and the other for testing. The revised findings indicated the absence of latent variabilities, suggesting that our modifications successfully controlled for the previously observed effects.

The best performing pipeline on the revised data was a combination of SES and random forests, consistent with the majority of separate cohort analyses. A total of 596 different pipelines were evaluated by a repeated 10-fold CV approach (see Supplementary report for details). As shown in Figure 4B, pipelines incorporating SES for feature selection demonstrated higher average performance during training than those with LASSO. Among the predictive modeling algorithms tested, Random Forests exhibited the highest predictive performance, followed by Ridge Logistic Regression. Figure 5A illustrates the ROC curves of the best performing model. The achieved performance in terms of AUC on the test data was 0.758; on par with the training performance of 0.777 (C.I. [0.708, 0.822]). Figure 5B also presents the out-of-sample predictions during training.

In terms of feature selection, the best performing pipeline resulted in a signature comprising 70 features, primarily consisting of microbiome species, with the addition of gender. Figure 5C illustrates the importance of these features in predicting the outcome (see Supplementary Table S2 for the corresponding species names). While SES and RFs demonstrated superior performance in most of the analyses, Table 5 reveals the significant variation in predictive performances and generated signatures that was found. The detailed taxonomy of the species involved is provided in Supplementary Table S3. Variability in performance was also highlighted by Wirbel et al. (2019) where only the predictive performance on several cohorts was examined. This suggests the need for further investigation into the specificity of these microbiome signatures. Interestingly, however, 20 species present in the revised dataset's signature were also found in the signatures generated when analyzing each cohort independently, indicating their potential importance across diverse geographic communities.

Among the selected species in the revised dataset's signature, their relevance is in agreement with previous reports in the literature regarding their predictive role in CRC. In particular, considering the top five most important species identified for the revised dataset, excluding gender (Figure 5C), *Fusobacterium gonidiaformans* (msp_1081) was detected in colorectal carcinoma relative to normal colon (Castellarin et al., 2012; Kostic et al., 2012), and found to be enriched in adenomas (Gevers et al., 2014). Several Clostridium species (msp_0578) have been associated with CRC (i.e., *Clostridium symbiosum*, *Clostridium hylemonae*, and *Clostridium scindens*) (Zeller et al., 2014). In addition, an increased risk of CRC was found in patients with bacteremia from *Clostridium septicum*, *Clostridium perfringens* or other species, such as *Fusobacterium nucleatum* and Peptostreptococcus species (Kwong et al., 2018). Christensenellales (msp_0622) has shown to be associated with both host genetic status CRC and risk (Waters and Ley, 2019), while *Streptococcus thermophilus* (msp_0833) has been identified to be depleted in patients with colorectal cancer (Qing et al., 2021). Regarding *Fusobacterium nucleatum* subspecies animalis (msp_0610), also selected when independently analyzing the Austrian, French, German and Japanese cohorts, *Fusobacterium nucleatum* was associated with stages of colorectal neoplasia development, colorectal cancer and disease outcome (Flanagan et al., 2014).

Figure 5D visualizes how well the selected features separate the two classes on a low dimensional space representation. Furthermore, it indicates a few samples that could be considered as outliers and would need further investigation.

Figure 6A displays the ROC curves of the best interpretable model on both training and test sets. This model, based on Ridge Regression, demonstrates performance that is comparable to the best-performing model (training AUC 0.754, C.I. [0.693, 0.811], test AUC 0.731). The linear nature of the predictive algorithm enables direct interpretation of the generated model. In Figure 6B, the species selected by the interpretable model are showcased alongside their corresponding linear coefficient values in the log-odd formula.

For instance, the identified association of *Peptostreptococcus stomatis* (msp_1327) corroborates findings from the French cohort's original data publication (Zeller et al., 2014). Furthermore, while msp_0937 corresponds to an unclassified Duodenibacillus species, it is noteworthy that Duodenibacillus massiliensis is linked with treatment

TABLE 5 Summarized results from the analysis of the collected CRC shotgun datasets with 2014 features using JADBio.

Cohort	Samples	Training AUC	Validation AUC	Feature selection	Predictive algorithm	Report link	Features
Austrian	109	0.90		SES	RF	Report 1	msp_0041, msp_0610, msp_1600, msp_1721, msp_0304,msp_0376, msp_0831, msp_0417, msp_0869, msp_1017, msp_0350, msp_1101[a], msp_0717, msp_0215, msp_1176[b], msp_1587[a], msp_1195[b]
French	114	0.79		SES	RF	Report 2	msp_0024, msp_0554, msp_0006, msp_1158, msp_1327, msp_0610, msp_0800, msp_0168, msp_1402, msp_0350, msp_0835, msp_0317[a], msp_1,193, msp_0541[b], msp_1037, msp_1060c[a], msp_1213[b]
Chinese	128	0.74		SES	RF	Report 3	age, msp_0033[a], msp_0990, msp_1028c[b], msp_0468, msp_0044, msp_0457, msp_0713, msp_0235, msp_0178, msp_1206, msp_0236, msp_0318, msp_0126, msp_0542, msp_0639, msp_0864, msp_1603c, msp_0154, msp_1901[a], msp_1193[b]
Italian	113	0.63		SES	RF	Report 4	msp_1234, msp_0258, msp_0100, msp_0275, msp_1489c, msp_0562, msp_0199[a], msp_0338, msp_0340, msp_0125[b], msp_0369[aa], msp_0215[b], msp_0906[b]
Indian	140	1.00		LASSO	RF	Report 5	study_accession, age, msp_0027, msp_0128, msp_0258, msp_0585, msp_0841, msp_1459
German	125	0.98		SES	RF	Report 6	HQ_clean_read_count[a], msp_1234, msp_0722, instrument_model, msp_0610, msp_1018, msp_1428, mapped_read_count[a]
USA	104	0.64		SES	SVM	Report 7	msp_0147, msp_1293, msp_1522, msp_0679[a], msp_0035, msp_1,193, msp_0766, msp_0747, msp_1,038, msp_0083, msp_1850, msp_0566, msp_0180, msp_1069[b], msp_1621, msp_1241, msp_0845, msp_0854[a], msp_1110[b]
Japanese	577	0.69		SES	RF	Report 8	timepoint, msp_1327, msp_0003, msp_0749, msp_1315, msp_0132, msp_0935, msp_0436, msp_0574c, msp_0468, msp_0152, msp_0126, msp_1276, msp_1049, msp_1004, msp_1156, msp_0887, msp_0323, msp_0525, msp_0118, msp_1590c, msp_1028c, msp_0635, msp_0062, msp_0610
All cohorts	1,410	0.85		SES	RF	Report 9	study_accession[a], timepoint, age, msp_0610, msp_1327, msp_1112, msp_0454, msp_0668, msp_0910, msp_0129, msp_0128, msp_1,193, msp_0305, msp_0054, msp_0757, msp_0100, msp_1028c, msp_1682c, msp_0357, msp_1172, msp_0032, msp_0297, msp_0105, msp_1158, msp_0389, msp_0935, msp_1173c, msp_1946, msp_0546, msp_1234, msp_0574c, msp_0468, msp_0110, msp_0833, msp_0484, msp_1790, msp_1188, msp_0172, msp_0864, msp_1600, msp_0853c, msp_0831, msp_0258, msp_0077, msp_0126, msp_0062, msp_1156, msp_0204, msp_0034, msp_0542, instrument_model[a], westernised[a], country[a]

(Continued)

TABLE 5 (Continued)

Cohort	Samples	Training AUC	Validation AUC	Feature selection	Predictive algorithm	Report link	Features
Revised data (best perf.)	1,117	0.77	0.758	SES	RF	Report 10	msp_1081, gender, msp_0578, msp_0622, msp_0833, msp_0610, msp_0100, msp_1579c, msp_0676, msp_0236, msp_1010, msp_0317, msp_0757, msp_0910, msp_0496, msp_0574c, msp_1327, msp_1028c, msp_0938, msp_0126, msp_0129, msp_1188, msp_0172, msp_1069, msp_0257, msp_0835, msp_1324, msp_1682c, msp_0864, msp_1102, msp_1467, msp_1245, msp_0668, msp_1158, msp_0305, msp_0937, msp_1671c, msp_1790, msp_0110, msp_1754, msp_0062, msp_0814, msp_0853c, msp_1322, msp_1217, msp_1156, msp_1036, msp_0805, msp_1712, msp_1231, msp_0454, msp_0935, msp_1657, msp_1234, msp_0076, msp_1487, msp_1570, msp_1042, msp_0118, msp_1112, msp_0457, msp_1048, msp_0232, msp_0542, msp_0468, msp_0258, msp_1789, msp_1173c, msp_0347, msp_0089
Revised data (best inter.)	1,117	0.75	0.73	SES	LR	Report 11	msp_0100, msp_0118, msp_0126, msp_0129, msp_0172, msp_0257, msp_0258, msp_0317, msp_0468, msp_0542, msp_0574c, msp_0610, msp_0676, msp_0805, msp_0833, msp_0835, msp_0910, msp_0935, msp_0937, msp_1028c, msp_1112, msp_1156, msp_1158, msp_1188, msp_1231, msp_1245, msp_1327, msp_1570, msp_1754, msp_1789

Full description of the metadata and species features can be found in Supplementary Tables S1, S2. The variable health_status was set as the outcome to be predicted with classification (binary) as the analysis type. Samples from the same country were merged together. Detailed signatures, including the lists of selected features, can be accessed via the links to the respective JADBio analysis reports. Superscript letters denote statistical equivalence of features, i.e., replacing one feature with another feature labeled with the same superscript letter will, on average, yield the same predictive performance. RF, Random Forest; SVM, Support Vector Machine; LR, Logistic Regression.

response for patients with rectal cancer (Jang et al., 2020). Similarly, concerning the unknown msp_1245 (Parvimonas species), *Parvimonas micra* together with *Fusobacterium nucleatum* (msp_0574c), *Peptostreptococcus stomatis* (msp_1327), and *Akkermansia muciniphila* were found to be over-represented in CRC patients compared to non-CRC controls (Osman et al., 2021). In another confirmatory study *Peptostreptococcus anaerobius* (msp_0935) has been implicated in modulating colorectal carcinogenesis and tumor immunity. Additionally, *Prevotella intermedia* (msp_1028) and *Fusobacterium nucleatum* (msp_0574c) were found to act synergistically, enhancing the migration and invasion of CRC cells (Long et al., 2019; Lo et al., 2022).

The sign of the coefficient indicates whether the species is considered a risk factor or not by the model. For instance, Ruthenibacterium lactatiformans (msp_0172) has been previously identified as putative candidate non-invasive biomarkers in CRC patients (Trivieri et al., 2020). Figure 6C illustrates how its abundance influences the prediction. Specifically, the greater the abundance, the more the risk for a sample to be classified as a patient case (P). In contrast, species Clostridiales bacterium (msp_0835) is found to have a protective effect against CRC, as evidenced by its ICE plot (Figure 6D). The higher its abundance, the lower the probability to be in the patient class. Indeed, a recent study demonstrated the effectiveness of this species in both prophylactic and therapeutic contexts speculating its applicability to primary prevention in patient populations with a strong genetic predisposition or family history of CRC (Montalban-Arques et al., 2021). Taken together, combining feature selection results with interpretable modeling and visualization techniques, meaningful conclusions can be drawn about the predictive significance of different species.

4. Discussion

Our objectives in this work have been to: (1) review the challenges for an analyst when performing predictive modeling of microbiome data, (2) create a comprehensive set of practical advices, and (3) explore opportunities for automating various aspects of ML analysis to construct pipelines suitable for clinicians and non-experts in translational applications. To achieve these goals, we considered a typical ML workflow that starts after microbiome-related profiles are organized in a two-dimensional table format, such as OTUs, ASV, or MSP (metagenomic species) tables. This process involves multiple steps, including data preprocessing (e.g., normalization, filtering), feature selection, predictive modeling, and performance estimation. Our objective was to address the challenges associated with each of these steps considering diverse algorithms, their combinations, as well as our

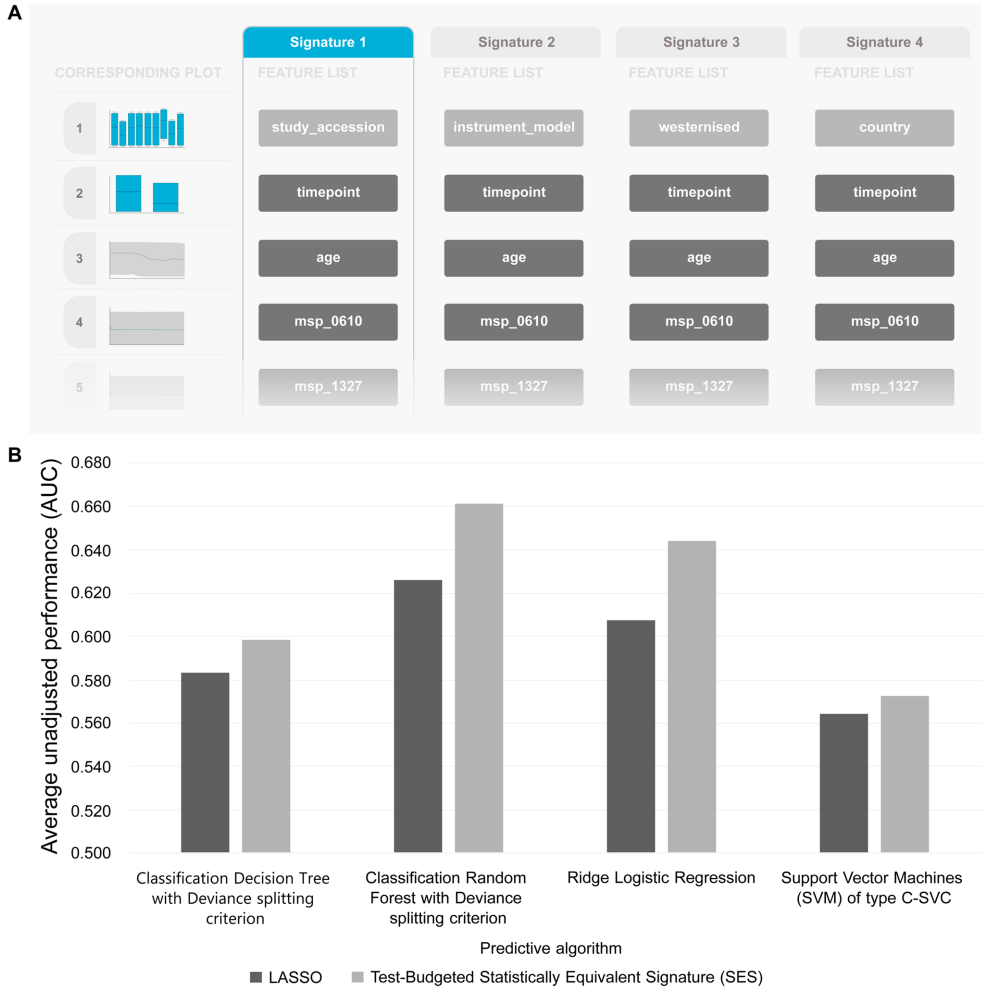

FIGURE 4
(A) Features detected by feature selection that generate classification bias. **(B)** Comparative evaluation of tested pipelines.

capacity to interpret and explain their results. Through the utilization of benchmark dataset(s) and automated machine learning techniques (AutoML), we were able to derive several noteworthy conclusions regarding the optimal utilization of ML methods toward disease diagnosis, prognosis, and biomarker discovery.

In the context of data preprocessing, a major challenge lies in selecting the appropriate normalization and filtering approaches due to variations in sampling depth, data sparsity (represented by an excess of zeros in the tables), and data compositionality. To mitigate sampling variability, rarefaction is used to remove samples. However, this may decrease statistical power and does not address compositionality (McMurdie and Holmes, 2014). Alternatively, researchers incorporate the sampled variation as covariates in data analysis. On the other hand, sparsity hampers models that rely on Gaussian assumptions. Certain algorithms, like decision trees and random forests, can handle sparsity, while others may fail. Filtering rare features and removing near-zero variance ones is a successful strategy, outperforming imputation methods in the context of logarithmic transformations that can introduce aberrant observations and depend on imputation algorithm quality. Finally, regarding normalization, contemporary sequencing cannot capture the total number of bacterial species, only their proportions. Compositional analysis is the appropriate mathematical framework, but its application and impact on ML models are still actively researched (Greenacre et al., 2021; Hron et al., 2021). From our observations, the CLR transformation seems to be useful for the PLS regression, although it was not in the top performing models. For the other models, the CLR transformation globally decreased the performances. However, these observations are based on the specific data set used in our experiments, and further evaluation will be necessary to assess their generalizability to other data sets before providing general recommendations regarding the choice of transformations.

For feature selection and predictive modeling, the primary challenges revolve around the high dimensionality of the data and the complex interactions inherent to microbial species, including co-occurrence and partial correlation. Building models that incorporate the thousands of microbiome features in a multivariate manner while maintaining predictive performance with limited sample sizes is undeniably demanding. It requires the utilization of scalable methods that account for the intricate dependency structure

10.3389/fmicb.2023.1261889

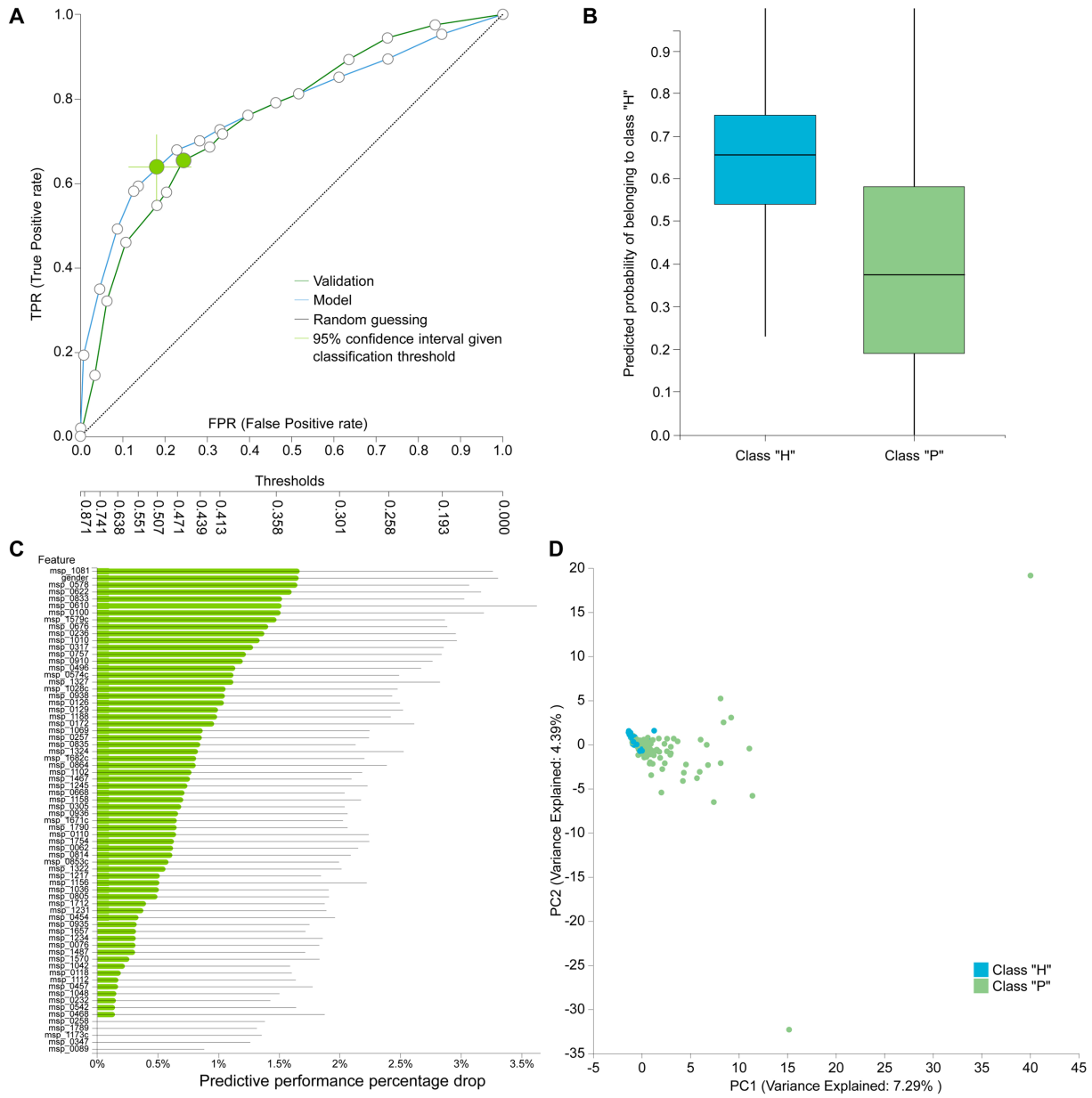

FIGURE 5
(A) Train (blue) and test (green) AUCROC after analyzing the revised dataset. ROC Curve considering CRC patients (P) as the positive class. **(B)** Out-of-sample (training) predictions for Healthy (H) and CRC patients class (P). **(C)** Feature importance defined as the percentage drop in predictive performance when the feature is removed from the model. Gray lines indicate 95% confidence intervals. **(D)** Supervised PCA on the selected features depicts the model performance in separating the two classes and also outlier samples.

of microbiome data, as well as appropriate performance estimation protocols to generate an optimal final model. Neglecting these considerations can result in overestimated conclusions and misleading insights. Using the JADBio autoML approach our observations indicate that multivariate feature selection methods such as the Statistically Equivalent Signatures algorithm combined with Random Forests can yield optimal balance between performance and results interpretability and explainability. These suggest a good starting point for an analyst.

However, it must be acknowledged that no single ML pipeline can universally accommodate all predictive modeling scenarios. As demonstrated here, there are several algorithms that account for the biological, methodological, and technical challenges in microbiome data. Additionally, different ML methods with different strengths and limitations exist for addressing the dimensionality and complexity of the problem and the underlying patterns in the data. Therefore, a highly advisable approach is to explore a diverse range of methods at each stage of the ML pipeline, and communicate the results according to the open science principles to facilitate transparency, verification, and reuse. Then, only through rigorous performance evaluation can the optimal predictive model and biomarkers be effectively identified, specifically tailored to address the particular microbiome problem at hand.

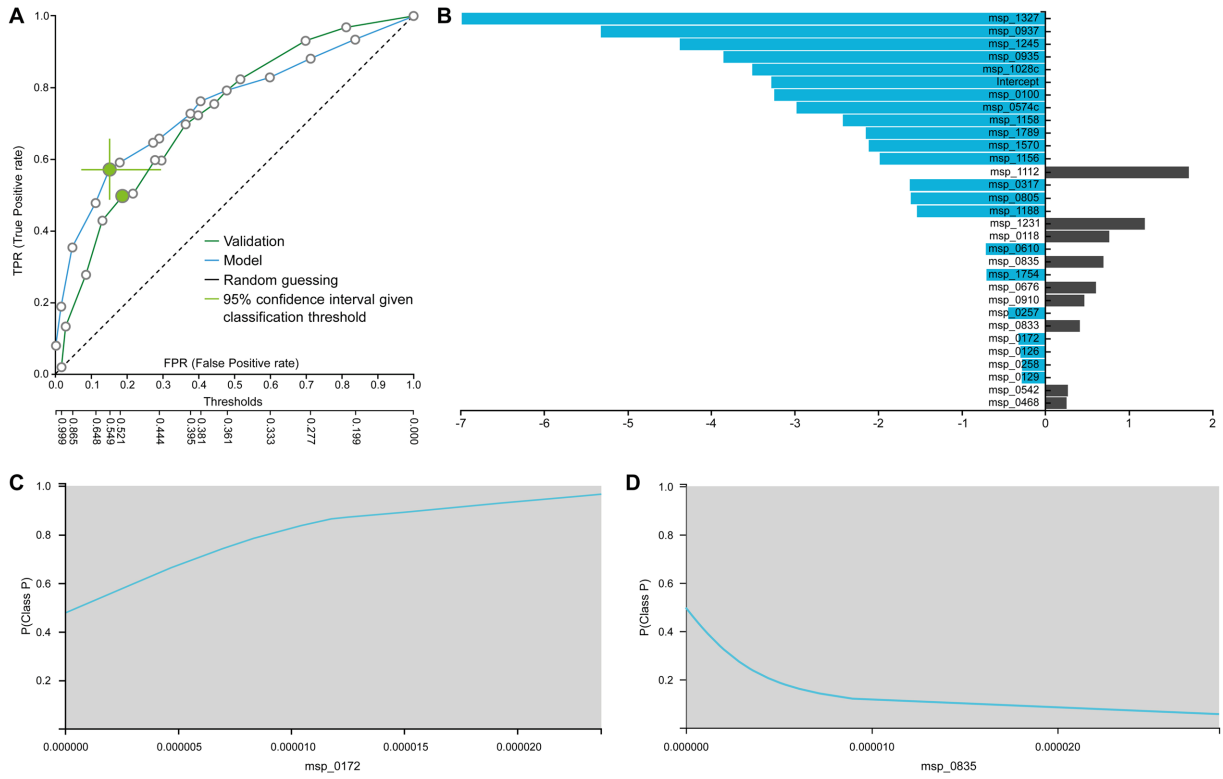

FIGURE 6
(A) ROC of the best interpretable model. **(B)** Contribution of each species to the prediction from logistic regression as the best interpretable model. Feature Interpretation using ICE plots with an example of a **(C)** risk factor (the higher the abundance, the higher the probability to be in the P (Patients) class) and a **(D)** protective factor (the higher the abundance, the lower the probability to be in the P (Patients) class).

Author contributions

GP: Conceptualization, Formal analysis, Project administration, Supervision, Visualization, Writing – original draft, Writing – review & editing. SoT: Conceptualization, Formal analysis, Project administration, Supervision, Visualization, Writing – original draft, Writing – review & editing. ML: Writing – review & editing, Methodology. TK: Writing – review & editing, Visualization. EI: Writing – review & editing, Methodology. JE: Writing – review & editing, Methodology. PN: Writing – original draft. ATo: Writing – review & editing. AS: Writing – review & editing. RS: Writing – review & editing. SB: Data curation, Writing – original draft. GV: Data curation, Writing – original draft. SaT: Writing – original draft. LL: Writing – review & editing. ATe: Writing – review & editing. MC: Funding acquisition, Writing – review & editing. MB: Conceptualization, Formal analysis, Project administration, Supervision, Visualization, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This study was based upon work from COST Action ML4Microbiome "Statistical and machine learning techniques in human microbiome studies" (CA18131), supported by COST (European Cooperation in Science and Technology, www.cost.eu). MB acknowledged support through the Metagenopolis grant ANR-11-DPBS-0001. ML acknowledged support by FCT - Fundação para a Ciência e a Tecnologia, I.P., with references UIDB/00297/2020 and UIDP/00297/2020 (NOVA Math), UIDB/00667/2020 and UIDP/00667/2020 (UNIDEMI), and CEECINST/00042/2021.

Acknowledgments

We greatly thank Emmanuelle Le Chatelier and Pauline Barbet (Université Paris-Saclay, INRAE, MetaGenoPolis, 78350, Jouy-en-Josas, France) for preparing the shotgun CRC benchmark dataset. We also thank Michelangelo Ceci (Department of Computer Science, University of Bari Aldo Moro, Bari, Italy) and Christian Jansen (Institute of Science and Technology, Austria) for their interim leadership of the Working Group 3 of the COST Action ML4Microbiome.

Conflict of interest

GP was directly affiliated with JADBio—Gnosis DA, S.A., which offers the JADBio service commercially.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2023.1261889/full#supplementary-material

References

Aitchison, J. (1982). The statistical analysis of compositional data. *J. R. Stat. Soc. B* 44, 139–160. doi: 10.1111/j.2517-6161.1982.tb01195.x

Akosa, J. (2017). *Predictive accuracy: a misleading performance measure for highly imbalanced data*. Availanble at: https://www.semanticscholar.org/paper/Predictive-Accuracy-%3A-A-Misleading-Performance-for-Akosa/8eff162ba887b6ed3091d5b6aa1a89e23342cb5c

Barbet, P., Almeida, M., Probul, N., Baumach, J., Pons, N., Plaza Onate, F., et al. (2023). Taxonomic profiles, functional profiles and manually curated metadata of human fecal metagenomes from public projects coming from colorectal cancer studies (version 5) [dataset]. *Recher. Data Gouv.* doi: 10.57745/7IVO3E

Behrouzi, A., Nafari, A. H., and Siadat, S. D. (2019). The significance of microbiome in personalized medicine. *Clin. Transl. Med.* 8:e16. doi: 10.1186/s40169-019-0232-y

Bellantuono, L., Monaco, A., Amoroso, N., Lacalamita, A., Pantaleo, E., Tangaro, S., et al. (2022). Worldwide impact of lifestyle predictors of dementia prevalence: an eXplainable artificial intelligence analysis. *Front. Big Data* 5:1027783. doi: 10.3389/fdata.2022.1027783

Berland, M., Meslier, V., Berreira Ibraim, S., Le Chatelier, E., Pons, N., Maziers, N., et al. (2023). Both disease activity and HLA–B27 status are associated with gut microbiome dysbiosis in spondyloarthritis patients. *Arthritis Rheumatol.* 75, 41–52. doi: 10.1002/art.42289

Bewick, V., Cheek, L., and Ball, J. (2004). Statistics review 13: receiver operating characteristic curves. *Crit. Care* 8, 508–512. doi: 10.1186/cc3000

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Breiman, L. (2017). *Classification and Regression Trees (eBook)*. Routledge.

Brouillette, M. (2023). Cancer debugged. *Nat. Biotechnol.* 41, 310–313. doi: 10.1038/s41587-023-01677-z

Brumfield, K. D., Huq, A., Colwell, R. R., Olds, J. L., and Leddy, M. B. (2020). Microbial resolution of whole genome shotgun and 16S amplicon metagenomic sequencing using publicly available NEON data. *PLoS One* 15:e0228899. doi: 10.1371/journal.pone.0228899

Calle, M. L. (2019). Statistical analysis of metagenomics data. *Genomics Inform* 17:e6. doi: 10.5808/GI.2019.17.1.e6

Cammarota, G., Ianiro, G., Ahern, A., Carbone, C., Temko, A., Claesson, M. J., et al. (2020). Gut microbiome, big data and machine learning to promote precision medicine for cancer. *Nat. Rev. Gastroenterol. Hepatol.* 17, 635–648. doi: 10.1038/s41575-020-0327-3

Cao, Q., Sun, X., Rajesh, K., Chalasani, N., Gelow, K., Katz, B., et al. (2021). Effects of rare microbiome taxa filtering on statistical analysis. *Front. Microbiol.* 11:607325. doi: 10.3389/fmicb.2020.607325

Carrieri, A. P., Haiminen, N., Maudsley-Barton, S., Gardiner, L.-J., Murphy, B., Mayes, A. E., et al. (2021). Explainable AI reveals changes in skin microbiome composition linked to phenotypic differences. *Sci. Rep.* 11:4565. doi: 10.1038/s41598-021-83922-6

Castellarin, M., Warren, R. L., Freeman, J. D., Dreolini, L., Krzywinski, M., Strauss, J., et al. (2012). *Fusobacterium nucleatum* infection is prevalent in human colorectal carcinoma. *Genome Res.* 22, 299–306. doi: 10.1101/gr.126516.111

Chang, C.-C., and Lin, C.-J. (2011). LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.* 2, 27:1–1:27. doi: 10.1145/1961189.1961199

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* 16, 321–357. doi: 10.1613/jair.953

Chen, L., Zhang, Y.-H., Huang, T., and Cai, Y.-D. (2016). Gene expression profiling gut microbiota in different races of humans. *Sci. Rep.* 6:23075. doi: 10.1038/srep23075

Chicco, D., and Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* 21:6. doi: 10.1186/s12864-019-6413-7

Claesson, M. J., Clooney, A. G., and O'Toole, P. W. (2017). A clinician's guide to microbiome analysis. *Nat. Rev. Gastroenterol. Hepatol.* 14, 585–595. doi: 10.1038/nrgastro.2017.97

Dai, Z., Wong, S. H., Yu, J., and Wei, Y. (2019). Batch effects correction for microbiome data with Dirichlet-multinomial regression. *Bioinformatics* 35, 807–814. doi: 10.1093/bioinformatics/bty729

Ding, Y., Tang, S., Liao, S. G., Jia, J., Oesterreich, S., Lin, Y., et al. (2014). Bias correction for selecting the minimal-error classifier from many machine learning models. *Bioinformatics* 30, 3152–3158. doi: 10.1093/bioinformatics/btu520

Durazzi, F., Sala, C., Castellani, G., Manfreda, G., Remondini, D., and De Cesare, A. (2021). Comparison between 16S rRNA and shotgun sequencing data for the taxonomic characterization of the gut microbiota. *Sci. Rep.* 11:3030. doi: 10.1038/s41598-021-82726-y

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., and Barceló-Vidal, C. (2003). Isometric Logratio transformations for compositional data analysis. *Math. Geol.* 35, 279–300. doi: 10.1023/A:1023818214614

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., and Herrera, F. (2018). *Learning from imbalanced data sets*. Springer Cham.

Ferreira, A. J., and Figueiredo, M. A. T. (2012). Efficient feature selection filters for high-dimensional data. *Pattern Recogn. Lett.* 33, 1794–1804. doi: 10.1016/j.patrec.2012.05.019

Feurer, M., Eggensperger, K., Falkner, S., Lindauer, M., and Hutter, F. (2022). Auto-Sklearn 2.0: hands-free AutoML via meta-learning. *J. Mach. Learn. Res.* 23, 1–61. doi: 10.48550/arXiv.2007.04074

Feurer, M., Klein, A., Eggensperger, K., Springenberg, J., Blum, M., and Hutter, F. (2015). Efficient and robust automated machine learning. Advances in Neural Information Processing Systems 28 Available at: https://proceedings.neurips.cc/paper_files/paper/2015/hash/11d0e6287202fced83f79975ec59a3a6-Abstract.html

Flanagan, L., Schmid, J., Ebert, M., Soucek, P., Kunicka, T., Liska, V., et al. (2014). *Fusobacterium nucleatum* associates with stages of colorectal neoplasia development, colorectal cancer and disease outcome. *Eur. J. Clin. Microbiol. Infect. Dis.* 33, 1381–1390. doi: 10.1007/s10096-014-2081-3

Flemer, B., Lynch, D. B., Brown, J. M. R., Jeffery, I. B., Ryan, F. J., Claesson, M. J., et al. (2017). Tumour-associated and non-tumour-associated microbiota in colorectal cancer. *Gut* 66, 633–643. doi: 10.1136/gutjnl-2015-309595

Fromentin, S., Oñate, F. P., Maziers, N., Berreira Ibraim, S., Gautreau, G., Gitton-Quent, O., et al. (2021). *Extensive benchmark of machine learning methods for quantitative microbiome data*. Available at: https://hal.science/hal-04163473

Genuer, R., Poggi, J.-M., and Tuleau-Malot, C. (2015). VSURF: an R package for variable selection using random forests. *R J.* 7, 19–33. doi: 10.32614/RJ-2015-018

Gevers, D., Kugathasan, S., Denson, L. A., Vázquez-Baeza, Y., van Treuren, W., Ren, B., et al. (2014). The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host Microbe* 15, 382–392. doi: 10.1016/j.chom.2014.02.005

Ghannam, R. B., and Techtmann, S. M. (2021). Machine learning applications in microbial ecology, human microbiome studies, and environmental monitoring. *Comput. Struct. Biotechnol. J.* 19, 1092–1107. doi: 10.1016/j.csbj.2021.01.028

Gijsbers, P., LeDell, E., Thomas, J., Poirier, S., Bischl, B., and Vanschoren, J. (2019). An open source AutoML benchmark (arXiv:1907.00909). *arXiv*. doi: 10.48550/arXiv.1907.00909

Gijsbers, P., and Vanschoren, J. (2019). GAMA: genetic automated machine learning assistant. *J. Open Sour. Softw.* 4:1132. doi: 10.21105/joss.01132

Glassner, K. L., Abraham, B. P., and Quigley, E. M. M. (2020). The microbiome and inflammatory bowel disease. *J. Allergy Clin. Immunol.* 145, 16–27. doi: 10.1016/j.jaci.2019.11.003

Goh, W. W. B., Wang, W., and Wong, L. (2017). Why batch effects matter in omics data, and how to avoid them. *Trends Biotechnol.* 35, 498–507. doi: 10.1016/j.tibtech.2017.02.012

Greenacre, M., Martínez-Álvaro, M., and Blasco, A. (2021). Compositional data analysis of microbiome and any-omics datasets: a validation of the additive logratio transformation. *Front. Microbiol.* 12:727398. doi: 10.3389/fmicb.2021.727398

Hoerl, A. E., and Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 12, 55–67. doi: 10.2307/1267351

Hron, K., Coenders, G., Filzmoser, P., Palarea-Albaladejo, J., Faměra, M., and Matys Grygar, T. (2021). Analysing pairwise Logratios revisited. *Math. Geosci.* 53, 1643–1666. doi: 10.1007/s11004-021-09938-w

Hutter, F., Kotthoff, L., and Vanschoren, J. (2019). *Automated machine learning: methods, systems, challenges*. Springer International Publishing.

Jang, B.-S., Chang, J. H., Chie, E. K., Kim, K., Park, J. W., Kim, M. J., et al. (2020). Gut microbiome composition is associated with a pathologic response after preoperative chemoradiation in patients with rectal cancer. *Int. J. Radiat. Oncol. Biol. Phys.* 107, 736–746. doi: 10.1016/j.ijrobp.2020.04.015

Jensen, D. D., and Cohen, P. R. (2000). Multiple comparisons in induction algorithms. *Mach. Learn.* 38, 309–338. doi: 10.1023/A:1007631014630

Kaul, A., Mandal, S., Davidov, O., and Peddada, S. D. (2017). Analysis of microbiome data in the presence of excess zeros. *Front. Microbiol.* 8:2114. doi: 10.3389/fmicb.2017.02114

Khachatryan, L., de Leeuw, R. H., Kraakman, M. E. M., Pappas, N., te Raa, M., Mei, H., et al. (2020). Taxonomic classification and abundance estimation using 16S and WGS—a comparison using controlled reference samples. *Forensic Sci. Int.: Genet.* 46:102257. doi: 10.1016/j.fsigen.2020.102257

Kostic, A. D., Gevers, D., Pedamallu, C. S., Michaud, M., Duke, F., Earl, A. M., et al. (2012). Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. *Genome Res.* 22, 292–298. doi: 10.1101/gr.126573.111

Kuhn, M. (2015). *The caret package*. Available at: https://topepo.github.io/caret/

Kurnaz, F. S., and Filzmoser, P. (2023). Robust and sparse multinomial regression in high dimensions. *Data Min. Knowl. Disc.* 37, 1609–1629. doi: 10.1007/s10618-023-00936-6

Kurnaz, F. S., Hoffmann, I., and Filzmoser, P. (2018). Robust and sparse estimation methods for high-dimensional linear and logistic regression. *Chemom. Intell. Lab. Syst.* 172, 211–222. doi: 10.1016/j.chemolab.2017.11.017

Kwong, T. N. Y., Wang, X., Nakatsu, G., Chow, T. C., Tipoe, T., Dai, R. Z. W., et al. (2018). Association between bacteremia from specific microbes and subsequent diagnosis of colorectal cancer. *Gastroenterology* 155, 383–390.e8. doi: 10.1053/j.gastro.2018.04.028

Lagani, V., Athineou, G., Farcomeni, A., Tsagris, M., and Tsamardinos, I. (2017). Feature selection with the R package MXM: discovering statistically equivalent feature subsets. *J. Stat. Softw.* 80, 1–25. doi: 10.18637/jss.v080.i07

Lê Cao, K.-A., Martin, P. G., Robert-Granié, C., and Besse, P. (2009). Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC Bioinformatics* 10:34. doi: 10.1186/1471-2105-10-34

Le Chatelier, E., Almeida, M., Plaza Oñate, F., Pons, N., Gauthier, F., Ghozlane, A., et al. (2021). A catalog of genes and species of the human oral microbiota (version 2) [dataset]. *Recher. Data Gouv.* doi: 10.15454/WQ4UTV

Li, H. (2015). Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annu. Rev. Stat. Appl.* 2, 73–94. doi: 10.1146/annurev-statistics-010814-020351

Lin, H., and Peddada, S. D. (2020). Analysis of microbial compositions: a review of normalization and differential abundance analysis. *NPJ Biofilms Microbiomes* 6:60. doi: 10.1038/s41522-020-00160-w

Ling, W., Lu, J., Zhao, N., Lulla, A., Plantinga, A. M., Fu, W., et al. (2022). Batch effects removal for microbiome data via conditional quantile regression. *Nat. Commun.* 13:5418. doi: 10.1038/s41467-022-33071-9

Ling, C. X., and Sheng, V. S. (2010). "Cost-sensitive learning" in *Encyclopedia of machine learning*. eds. C. Sammut and G. I. Webb (Berlin: Springer), 231–235.

Liu, Y.-X., Qin, Y., Chen, T., Lu, M., Qian, X., Guo, X., et al. (2021). A practical guide to amplicon and metagenomic analysis of microbiome data. *Protein Cell* 12, 315–330. doi: 10.1007/s13238-020-00724-8

Lo, C.-H., Wu, D.-C., Jao, S.-W., Wu, C.-C., Lin, C.-Y., Chuang, C.-H., et al. (2022). Enrichment of *Prevotella intermedia* in human colorectal cancer and its additive effects with *Fusobacterium nucleatum* on the malignant transformation of colorectal adenomas. *J. Biomed. Sci.* 29:88. doi: 10.1186/s12929-022-00869-0

Lombardi, A., Diacono, D., Amoroso, N., Biecek, P., Monaco, A., Bellantuono, L., et al. (2022). A robust framework to investigate the reliability and stability of explainable artificial intelligence markers of mild cognitive impairment and Alzheimer's disease. *Brain Informatics* 9:17. doi: 10.1186/s40708-022-00165-5

Lombardi, A., Diacono, D., Amoroso, N., Monaco, A., Tavares, J. M. R. S., Bellotti, R., et al. (2021). Explainable deep learning for personalized age prediction with brain morphology. *Front. Neurosci.* 15:674055. doi: 10.3389/fnins.2021.674055

Long, X., Wong, C. C., Tong, L., Chu, E. S. H., Ho Szeto, C., Go, M. Y. Y., et al. (2019). *Peptostreptococcus anaerobius* promotes colorectal carcinogenesis and modulates tumour immunity. *Nat. Microbiol.* 4, 2319–2330. doi: 10.1038/s41564-019-0541-3

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15:550. doi: 10.1186/s13059-014-0550-8

Lovell, D., Pawlowsky-Glahn, V., Egozcue, J. J., Marguerat, S., and Bähler, J. (2015). Proportionality: a valid alternative to correlation for relative data. *PLoS Comput. Biol.* 11:e1004075. doi: 10.1371/journal.pcbi.1004075

Lubbe, S., Filzmoser, P., and Templ, M. (2021). Comparison of zero replacement strategies for compositional data with large numbers of zeros. *Chemom. Intell. Lab. Syst.* 210:104248. doi: 10.1016/j.chemolab.2021.104248

Lundberg, S. M., and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Adv. Neural Inf. Proces. Syst.* 30, 4765–4774.

Marcos-Zambrano, L. J., Karaduzovic-Hadziabdic, K., Loncar Turukalo, T., Przymus, P., Trajkovik, V., Aasmets, O., et al. (2021). Applications of machine learning in human microbiome studies: a review on feature selection, biomarker identification, disease prediction and treatment. *Front. Microbiol.* 12:634511. doi: 10.3389/fmicb.2021.634511

McMurdie, P. J., and Holmes, S. (2014). Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput. Biol.* 10:e1003531. doi: 10.1371/journal.pcbi.1003531

Mihajlović, A., Mladenović, K., Lončar-Turukalo, T., and Brdar, S. (2021). Machine learning based metagenomic prediction of inflammatory bowel disease. *Stud. Health Technol. Inform.* 285, 165–170. doi: 10.3233/SHTI210591

Montalban-Arques, A., Katkeviciute, E., Busenhart, P., Bircher, A., Wirbel, J., Zeller, G., et al. (2021). Commensal clostridiales strains mediate effective anti-cancer immune response against solid tumors. *Cell Host Microbe* 29, 1573–1588.e7. doi: 10.1016/j.chom.2021.08.001

Monti, G. S., and Filzmoser, P. (2022). Robust logistic zero-sum regression for microbiome compositional data. *ADAC* 16, 301–324. doi: 10.1007/s11634-021-00465-4

Moreno-Indias, I., Lahti, L., Nedyalkova, M., Elbere, I., Roshchupkin, G., Adilovic, M., et al. (2021). Statistical and machine learning techniques in human microbiome studies: contemporary challenges and solutions. *Front. Microbiol.* 12:635781. doi: 10.3389/fmicb.2021.635781

Navab-Moghadam, F., Sedighi, M., Khamseh, M. E., Alaei-Shahmiri, F., Talebi, M., Razavi, S., et al. (2017). The association of type II diabetes with gut microbiota composition. *Microb. Pathog.* 110, 630–636. doi: 10.1016/j.micpath.2017.07.034

Odintsova, V., Tyakht, A., and Alexeev, D. (2017). Guidelines to statistical analysis of microbial composition data inferred from metagenomic sequencing. *Curr. Issues Mol. Biol.* 24, 17–36. doi: 10.21775/cimb.024.017

Olson, R. S., and Moore, J. H. (2019). "TPOT: a tree-based pipeline optimization tool for automating machine learning" in *Automated machine learning*. eds. F. Hutter, L. Kotthoff and J. Vanschoren (Cham: Springer International Publishing), 151–160.

Osman, M. A., Neoh, H., Mutalib, N.-S. A., Chin, S.-F., Mazlan, L., Ali, R. A. R., et al. (2021). *Parvimonas micra*, *Peptostreptococcus stomatis*, Fusobacterium nucleatum and *Akkermansia muciniphila* as a four-bacteria biomarker panel of colorectal cancer. *Sci. Rep.* 11:2925. doi: 10.1038/s41598-021-82465-0

Paulson, J. N., Stine, O. C., Bravo, H. C., and Pop, M. (2013). Differential abundance analysis for microbial marker-gene surveys. *Nat. Methods* 10, 1200–1202. doi: 10.1038/nmeth.2658

Pereira, M. B., Wallroth, M., Jonsson, V., and Kristiansson, E. (2018). Comparison of normalization methods for the analysis of metagenomic gene abundance data. *BMC Genomics* 19:274. doi: 10.1186/s12864-018-4637-6

Plaza Oñate, F., Le Chatelier, E., Almeida, M., Cervino, A. C. L., Gauthier, F., Magoulès, F., et al. (2019). MSPminer: abundance-based reconstitution of microbial pan-genomes from shotgun metagenomic data. *Bioinformatics* 35, 1544–1552. doi: 10.1093/bioinformatics/bty830

Pons, N., Batto, J.-M., Kennedy, S., Almeida, M., Boumezbeur, F., Moumen, B., et al. (2010). *METEOR -a plateform for quantitative metagenomic profiling of complex ecosystems*. Available at: https://forgemia.inra.fr/metagenopolis/meteor

Poretsky, R., Rodriguez-R, L. M., Luo, C., Tsementzi, D., and Konstantinidis, K. T. (2014). Strengths and limitations of 16S rRNA gene amplicon sequencing in revealing temporal microbial community dynamics. *PLoS One* 9:e93827. doi: 10.1371/journal.pone.0093827

Qing, L., Hu, W., Liu, W.-X., Zhao, L.-Y., Huang, D., Liu, X.-D., et al. (2021). *Streptococcus thermophilus* inhibits colorectal tumorigenesis through secreting β-galactosidase. *Gastroenterology* 160, 1179–1193.e14. doi: 10.1053/j.gastro.2020.09.003

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why should I trust you?": explaining the predictions of any classifier (arXiv:1602.04938). *arXiv*. doi: 10.48550/arXiv.1602.04938

Robinson, M. D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* 11:R25. doi: 10.1186/gb-2010-11-3-r25

Romero, R. A. A., Deypalan, M. N. Y., Mehrotra, S., Jungao, J. T., Sheils, N. E., Manduchi, E., et al. (2022). Benchmarking AutoML frameworks for disease prediction using medical claims. *BioData Mining* 15:15. doi: 10.1186/s13040-022-00300-2

Ryan, F. J., Ahern, A. M., Fitzgerald, R. S., Laserna-Mendieta, E. J., Power, E. M., Clooney, A. G., et al. (2020). Colonic microbiota is associated with inflammation and host epigenomic alterations in inflammatory bowel disease. *Nat. Commun.* 11:1512. doi: 10.1038/s41467-020-15342-5

Saito, T., and Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* 10:e0118432. doi: 10.1371/journal.pone.0118432

Salzberg, S. L. (1997). On comparing classifiers: pitfalls to avoid and a recommended approach. *Data Min. Knowl. Disc.* 1, 317–328. doi: 10.1023/A:1009752403260

Sanz, H., Valim, C., Vegas, E., Oller, J. M., and Reverter, F. (2018). SVM-RFE: selection and visualization of the most relevant features through non-linear kernels. *BMC Bioinformatics* 19:432. doi: 10.1186/s12859-018-2451-4

Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W. S., et al. (2011). Metagenomic biomarker discovery and explanation. *Genome Biol.* 12:R60. doi: 10.1186/gb-2011-12-6-r60

Thirion, F., Da Silva, K., Plaza Oñate, F., Alvarez, A.-S., Thabuis, C., Pons, N., et al. (2022). Diet supplementation with NUTRIOSE, a resistant dextrin, increases the abundance of *Parabacteroides distasonis* in the human gut. *Mol. Nutr. Food Res.* 66:e2101091. doi: 10.1002/mnfr.202101091

Thirion, F., Speyer, H., Hansen, T. H., Nielsen, T., Fan, Y., le Chatelier, E., et al. (2023). Alteration of gut microbiome in patients with schizophrenia indicates links between bacterial tyrosine biosynthesis and cognitive dysfunction. *Biol. Psychiatry Glob. Open Sci.* 3, 283–291. doi: 10.1016/j.bpsgos.2022.01.009

Thornton, C., Hutter, F., Hoos, H. H., and Leyton-Brown, K. (2013). Auto-WEKA: combined selection and hyperparameter optimization of classification algorithms. Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 847–855.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. B* 58, 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x

Tibshirani, R. J., and Tibshirani, R. (2009). A bias correction for the minimum error rate in cross-validation. *Ann. Appl. Stat.* 3, 822–829. doi: 10.1214/08-AOAS224

Tremblay, J., Singh, K., Fern, A., Kirton, E., He, S., Woyke, T., et al. (2015). Primer and platform effects on 16S rRNA tag sequencing. *Front. Microbiol.* 6:771. doi: 10.3389/fmicb.2015.00771

Trivieri, N., Pracella, R., Cariglia, M. G., Panebianco, C., Parrella, P., Visioli, A., et al. (2020). BRAFV600E mutation impinges on gut microbial markers defining novel biomarkers for serrated colorectal cancer effective therapies. *J. Exp. Clin. Cancer Res.* 39:285. doi: 10.1186/s13046-020-01801-w

Tsamardinos, I., Charonyktakis, P., Papoutsoglou, G., Borboudakis, G., Lakiotaki, K., Zenklusen, J. C., et al. (2022). Just add data: automated predictive modeling for knowledge discovery and feature selection. *Npj Precis. Oncol.* 6:38. doi: 10.1038/s41698-022-00274-8

Tsamardinos, I., Greasidou, E., and Borboudakis, G. (2018). Bootstrapping the out-of-sample predictions for efficient and accurate cross-validation. *Mach. Learn.* 107, 1895–1922. doi: 10.1007/s10994-018-5714-4

Tsamardinos, I., Rakhshani, A., and Lagani, V. (2014). "Performance-estimation properties of cross-validation-based protocols with simultaneous hyper-parameter optimization" in *Artificial intelligence: methods and applications*. eds. A. Likas, K. Blekas and D. Kalles (Cham: Springer International Publishing), 1–14.

Vapnik, V. (2006). *Estimation of dependences based on empirical data* Springer.

Větrovský, T., and Baldrian, P. (2013). The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. *PLoS One* 8:e57923. doi: 10.1371/journal.pone.0057923

Wang, Y., and Lê Cao, K.-A. (2023). PLSDA-batch: a multivariate framework to correct for batch effects in microbiome data. *Brief. Bioinform.* 24:bbac622. doi: 10.1093/bib/bbac622

Wang, Y., and LêCao, K.-A. (2020). Managing batch effects in microbiome data. *Brief. Bioinform.* 21, 1954–1970. doi: 10.1093/bib/bbz105

Waters, J. L., and Ley, R. E. (2019). The human gut bacteria Christensenellaceae are widespread, heritable, and associated with health. *BMC Biol.* 17:83. doi: 10.1186/s12915-019-0699-4

Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., et al. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* 5:27. doi: 10.1186/s40168-017-0237-y

Wen, C., Zheng, Z., Shao, T., Liu, L., Xie, Z., le Chatelier, E., et al. (2017). Quantitative metagenomics reveals unique gut microbiome biomarkers in ankylosing spondylitis. *Genome Biol.* 18:142. doi: 10.1186/s13059-017-1271-6

Wirbel, J., Pyl, P. T., Kartal, E., Zych, K., Kashani, A., Milanese, A., et al. (2019). Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat. Med.* 25, 679–689. doi: 10.1038/s41591-019-0406-6

Wolpert, D. H. (2002). "The supervised learning no-free-lunch theorems" in *Soft computing and industry: recent applications*. eds. R. Roy, M. Köppen, S. Ovaska, T. Furuhashi and F. Hoffmann (London: Springer Publishing), 25–42.

Xanthopoulos, I., Tsamardinos, I., Christophides, V., Simon, E., and Salinger, A. (2020). Putting the human Back in the AutoML loop. In A. Poulovassilis (Ed.), Proceedings of the Workshops of the EDBT/ICDT 2020 Joint Conference (Vol. 2578). CEUR. Available at: https://ceur-ws.org/Vol-2578/#ETMLP5

Zeller, G., Tap, J., Voigt, A. Y., Sunagawa, S., Kultima, J. R., Costea, P. I., et al. (2014). Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.* 10:766. doi: 10.15252/msb.20145645

frontiers | Frontiers in Microbiology

TYPE Perspective
PUBLISHED 25 September 2023
DOI 10.3389/fmicb.2023.1257002

OPEN ACCESS

EDITED BY
Richard Allen White III,
University of North Carolina at Charlotte, United States

REVIEWED BY
Himel Mallick,
Cornell University, United States

*CORRESPONDENCE
Domenica D'Elia
✉ domenica.delia@cnr.it

RECEIVED 11 July 2023
ACCEPTED 05 September 2023
PUBLISHED 25 September 2023

CITATION
D'Elia D, Truu J, Lahti L, Berland M, Papoutsoglou G, Ceci M, Zomer A, Lopes MB, Ibrahimi E, Gruca A, Nechyporenko A, Frohme M, Klammsteiner T, Pau EC-dS, Marcos-Zambrano LJ, Hron K, Pio G, Simeon A, Suharoschi R, Moreno-Indias I, Temko A, Nedyalkova M, Apostol E-S, Truică C-O, Shigdel R, Telalović JH, Bongcam-Rudloff E, Przymus P, Jordamović NB, Falquet L, Tarazona S, Sampri A, Isola G, Pérez-Serrano D, Trajkovik V, Klucar L, Loncar-Turukalo T, Havulinna AS, Jansen C, Bertelsen RJ and Claesson MJ (2023) Advancing microbiome research with machine learning: key findings from the ML4Microbiome COST action. *Front. Microbiol.* 14:1257002. doi: 10.3389/fmicb.2023.1257002

Advancing microbiome research with machine learning: key findings from the ML4Microbiome COST action

Domenica D'Elia[1]*, Jaak Truu[2], Leo Lahti[3], Magali Berland[4], Georgios Papoutsoglou[5,6], Michelangelo Ceci[7], Aldert Zomer[8], Marta B. Lopes[9,10], Eliana Ibrahimi[11], Aleksandra Gruca[12], Alina Nechyporenko[13,14], Marcus Frohme[14], Thomas Klammsteiner[15,16], Enrique Carrillo-de Santa Pau[17], Laura Judith Marcos-Zambrano[17], Karel Hron[18], Gianvito Pio[7], Andrea Simeon[19], Ramona Suharoschi[20], Isabel Moreno-Indias[21], Andriy Temko[22], Miroslava Nedyalkova[23], Elena-Simona Apostol[24], Ciprian-Octavian Truică[24], Rajesh Shigdel[25], Jasminka Hasić Telalović[26], Erik Bongcam-Rudloff[27], Piotr Przymus[28], Naida Babić Jordamović[29,30], Laurent Falquet[31], Sonia Tarazona[32], Alexia Sampri[33,34], Gaetano Isola[35], David Pérez-Serrano[17], Vladimir Trajkovik[36], Lubos Klucar[37], Tatjana Loncar-Turukalo[38], Aki S. Havulinna[39,40], Christian Jansen[41,42], Randi J. Bertelsen[43] and Marcus Joakim Claesson[44]

[1]Department of Biomedical Sciences, National Research Council, Institute for Biomedical Technologies, Bari, Italy, [2]Institute of Molecular and Cell Biology, University of Tartu, Tartu, Estonia, [3]Department of Computing, University of Turku, Turku, Finland, [4]Université Paris-Saclay, INRAE, MetaGenoPolis, Jouy-en-Josas, France, [5]JADBio Gnosis DA S.A., Science and Technology Park of Crete, Heraklion, Greece, [6]Department of Computer Science, University of Crete, Heraklion, Greece, [7]Department of Computer Science, University of Bari Aldo Moro, Bari, Italy, [8]Department of Biomolecular Health Sciences (Infectious Diseases and Immunology), Faculty of Veterinary Medicine, Utrecht University, Utrecht, Netherlands, [9]Center for Mathematics and Applications (NOVA Math), NOVA School of Science and Technology, Caparica, Portugal, [10]UNIDEMI, Department of Mechanical and Industrial Engineering, NOVA School of Science and Technology, Caparica, Portugal, [11]Department of Biology, University of Tirana, Tirana, Albania, [12]Department of Computer Networks and Systems, Silesian University of Technology, Gliwice, Poland, [13]Systems Engineering Department, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine, [14]Department of Molecular Biotechnology and Functional Genomics, Technical University of Applied Sciences Wildau, Wildau, Germany, [15]Department of Microbiology, Universität Innsbruck, Innsbruck, Austria, [16]Department of Ecology, Universität Innsbruck, Innsbruck, Austria, [17]Computational Biology Group, Precision Nutrition and Cancer Research Program, IMDEA Food Institute, CEI UAM+CSIC, Madrid, Spain, [18]Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacký University, Olomouc, Czechia, [19]BioSense Institute, University of Novi Sad, Novi Sad, Serbia, [20]Molecular Nutrition and Proteomics Research Laboratory, Department of Food Science, University of Agricultural Sciences and Veterinary Medicine of Cluj-Napoca, Cluj-Napoca, Romania, [21]Department of Endocrinology and Nutrition, Virgen de la Victoria University Hospital, the Biomedical Research Institute of Malaga and Platform in Nanomedicine (IBIMA-BIONAND Platform), University of Malaga, Malaga, Spain, [22]Department of Electrical and Electronic Engineering, University College Cork, Cork, Ireland, [23]Chemistry and Pharmacy Department, University of Sofia, Sofia, Bulgaria, [24]Computer Science and Engineering Department, Faculty of Automatic Control and Computers, University Politehnica of Bucharest, Bucharest, Romania, [25]Department of Clinical Science, University of Bergen, Bergen, Norway, [26]Department of Computer Science, University Sarajevo School of Science and Technology, Sarajevo, Bosnia and Herzegovina, [27]Swedish University of Agricultural Sciences, Department of Animal Breeding and Genetics, Uppsala, Sweden, [28]Nicolaus Copernicus University Torun, Torun, Poland, [29]Computational Biology, International Centre for Genetic Engineering and Biotechnology, Trieste, Italy, [30]Verlab Research Institute for Biomedical Engineering, Medical Devices and Artificial Intelligence, Sarajevo, Bosnia and Herzegovina, [31]University of Fribourg

and Swiss Institute of Bioinformatics, Fribourg, Switzerland, [32]Department of Applied Statistics and Operations Research and Quality, Universitat Politècnica de València, València, Spain, [33]British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom, [34]Victor Phillip Dahdaleh Heart and Lung Research Institute, University of Cambridge, Cambridge, United Kingdom, [35]Department of General Surgery and Surgical-Medical Specialties, School of Dentistry, University of Catania, Catania, Italy, [36]Ss. Cyril and Methodius University, Skopje, North Macedonia, [37]Institute of Molecular Biology, Slovak Academy of Sciences, Bratislava, Slovakia, [38]Faculty of Technical Sciences, University of Novi Sad, Novi Sad, Serbia, [39]Finnish Institute for Health and Welfare, Helsinki, Finland, [40]Institute for Molecular Medicine Finland, FIMM-HiLIFE, Helsinki, Finland, [41]Biome Diagnostics GmbH, Vienna, Austria, [42]Institute of Science and Technology Austria (ISTA), Klosterneuburg, Austria, [43]University of Bergen, Bergen, Norway, [44]School of Microbiology & APC Microbiome Ireland, University College Cork, Cork, Ireland

The rapid development of machine learning (ML) techniques has opened up the data-dense field of microbiome research for novel therapeutic, diagnostic, and prognostic applications targeting a wide range of disorders, which could substantially improve healthcare practices in the era of precision medicine. However, several challenges must be addressed to exploit the benefits of ML in this field fully. In particular, there is a need to establish "gold standard" protocols for conducting ML analysis experiments and improve interactions between microbiome researchers and ML experts. The Machine Learning Techniques in Human Microbiome Studies (ML4Microbiome) COST Action CA18131 is a European network established in 2019 to promote collaboration between discovery-oriented microbiome researchers and data-driven ML experts to optimize and standardize ML approaches for microbiome analysis. This perspective paper presents the key achievements of ML4Microbiome, which include identifying predictive and discriminatory 'omics' features, improving repeatability and comparability, developing automation procedures, and defining priority areas for the novel development of ML methods targeting the microbiome. The insights gained from ML4Microbiome will help to maximize the potential of ML in microbiome research and pave the way for new and improved healthcare practices.

KEYWORDS

microbiome, machine learning, artificial intelligence, standards, best practices

1. Introduction

In the recent decade, the human microbiome has been characterized in great detail in several large-scale studies as a critical player in many human diseases and conditions. As more associations between the microbiome and disease phenotypes are elucidated, the research focus is expected to shift towards identifying the microbiome-related biomarkers for disease diagnostics, prognostics, and therapeutics (Manor et al., 2020). Nevertheless, microbiome data analysis is challenging due to its intrinsic characteristics like compositional nature, high dimensionality (often more features than samples), technical variability, missing data, and integration needs. Another challenge in microbiome data analysis is the interpretation of statistical models, as microbiome data often contains many highly correlated variables. Machine Learning (ML) methods offer great potential to further progress microbiome science, but these obstacles first need to be mitigated. Thus, a dynamic collaboration between microbiome and ML researchers is pivotal. Some initiatives have made more general efforts to provide ML guidelines and standard recommendations for data management, preprocessing, analysis and integration, like the ELIXIR Machine Learning Focus Group[1] (Walsh et al., 2021) or the ISO committees (ISO/TC 276 Biotechnology; ISO/IEC JTC 1/SC 42 Artificial intelligence; ISO/IEC TS 4213:2022 Assessment of Machine Learning Classification Performance).[2]

Moreover, while not explicitly focused on ML, the ongoing International Human Microbiome Coordination and Support Action (IHMCSA[3]) maps the necessary steps for innovation and builds consensus on priorities and means for the future of microbiome science and its translation. This includes standardization of microbiome analysis methods, which in its extension, also includes ML. The adoption of FAIR principles (Findable, Accessible, Interoperable, Reproducible) by ML tools and

1 https://elixir-europe.org/focus-groups/machine-learning

2 https://standards.globalspec.com/std/14568212/ISO/IEC%20TS%204213#:~:text=ISO%2FIEC%20TS%204213%20October%201%2C%202022%20Information%20technology,performance%20of%20machine%20learning%20models%2C%20systems%20and%20algorithms

3 https://humanmicrobiomeaction.eu/

FIGURE 1
ML4Microbiome COST Action's Working Groups. The figure shows the organization of the COST Action ML4Microbiome in four Working Groups (WGs), each committed to specific objectives. WG1 evaluated the state-of-the-art ML methods and software applied in human microbiome studies to define priority areas for novel machine learning and statistics applications that better address the specific challenges of human microbiome analysis. WG2 aimed to collect (from external projects and repositories) datasets describing microbiomes and characteristics of the underlying cohorts to test which ML methods are most robust and comparable, to provide more optimized parameters for the use of these methods, to develop novel ML methodologies and to implement a DREAM Challange on clinical data. WG3 investigated opportunities for automating the established Standard Operating Procedures (SOPs) into pipelines for translational use by clinicians and non-experts. WG4 goal was to bridge existing gaps between ML (bioinformaticians, statisticians, computer-science scientists) and microbiome experts through the organization of meetings, workshops, conferences, training schools, dissemination and communication activities.

models is also being approached by FAIR4ML.[4] However, these ML-focused initiatives are general and do not consider microbiome data or their characteristics. Scientific fields for which the study of human microbiota is essential, such as health and nutrition, have highlighted the need to join forces in the standardization and interoperability to integrate microbiome data with ML tools (Walsh et al., 2021; Balech et al., 2022). The European Cooperation in Science and Technology (COST) Action ML4Microbiome[5] - Statistical and machine learning techniques in human microbiome studies (CA18131) - started in 2019 to create a productive symbiosis between discovery-oriented microbiome researchers and data-driven ML experts to prompt the optimization and standardization of the best practice use of ML techniques for human microbiome research. Up to now, ML4Microbiome has gathered researchers from 35 different European countries, attracted and trained a large number of young scientists and published various scientific articles. The following sections discuss the Action's network research topics, elaborate on their relevance to the research challenges, and briefly overview more relevant achievements.

1.1. The ML4Microbiome action plan and challenges

To accomplish its goals, the ML4Microbiome network has designed an operational plan based on the coordinated and integrated work of four working groups (WGs), each addressing specific objectives (Figure 1). Several specific technical challenges have been identified (Moreno-Indias et al., 2021). Sequence-based microbiome studies use different types of data (16S rRNA gene or ITS amplicons/shotgun metagenomics or metatranscriptomics). Due to their different origin and types, separate modeling approaches are required. Moreover, microbiome data have large inter-individual variability and elevated noise levels, which Gaussian or log-normal models do not approximate well, providing challenges for traditional statistical methodologies (Voigt et al., 2015). There are more features than samples/observations (e.g., 100 studied humans may each have 1,000 microbial species and 1,000,000 microbial genes). This makes the application of ML methods challenging due to the curse of dimensionality, whereby huge data sparseness compromises the identification of data patterns or rules. Microbiome features often exhibit a complex dependency structure (taxonomic hierarchy or genes co-varying in abundance as encoded on the same genome, plasmid or phage). The relative abundance of each taxon is inherently related to the abundance of all other taxa, making it difficult to identify differentially abundant taxa (Weiss et al., 2017).

Microbial communities are also highly diverse, with many low-abundance taxa present only in a few samples. This can lead to high sparsity levels in the data, making it difficult to estimate the abundance of rare taxa accurately. Microbiome data is often compositional because most current studies have access only to the relative abundance of one microbial taxon (Gloor et al., 2017). In such cases, the abundance of one taxonomic group is constrained by the abundance of other taxonomic groups in the sample. Analyzing microbiome data as compositional data requires specific statistical approaches that account for this characteristic and address its unique challenges. Class sizes may be imbalanced (e.g., fewer disease samples than controls) (Ahlawat et al., 2021). An imbalanced class distribution coupled with high dimensional data poses a significant drawback for applying ML algorithms and results (Kim and Kim, 2018).

1.2. The current state of ML applications for microbiome data analysis

To assess the state-of-the-art of ML applications in microbiome data analysis, Working Group 1 (WG1) conducted a literature review accessible across the web application Machine Learning meTagenomic REsearch Scraper (MoLTRES[6]). The main aim of the tool is to provide

4 https://www.rd-alliance.org/groups/fair-machine-learning-fair4ml-ig

5 https://www.ml4microbiome.eu/

6 http://imdeafoodcompubio.com/index.php/moltres/

a user-friendly interface for centralized searching and storing ML studies on human microbiome data, encompassing feature selection, biomarker identification, disease prediction and treatment. The review highlighted a steady increase in the utilization of ML methods for human microbiome analysis in recent years. Most studies (>70%) using ML employed 16S rRNA gene amplicon sequencing data as the input data type, while 27% used only shotgun metagenome data. The most frequently used ML methods were random forest, logistic regression, and support vector machines. While the former method remained the most popular, the use of logistic regression and support vector machine algorithms has increased. These results were published by ML4Microbiome (Marcos-Zambrano et al., 2021), and subsequent updates by WG1 members were incorporated into MoLTRES.

1.3. Benchmark datasets and online repositories

When analyzing microbiome data, it is often helpful to create reference datasets to test existing or new ML tools, whether separate or combined. The importance of validation sets and gold standards is largely discussed in Papoutsoglou et al. (2023). Pasolli et al. (2016) have demonstrated that the performance of ML models may vary substantially depending on the disease addressed in the dataset. For this reason, Working Group 2 (WG2) and Working Group 3 (WG3) decided to establish a benchmark dataset based on a single disease for which a reasonable amount of public data was available. The choice has been made on colorectal cancer, for which 2090 human stool samples have been characterized by shotgun metagenomic sequencing from 13 public cohorts spanning nine countries. This data provides the gut microbiota composition in healthy controls and patients with adenoma or colorectal cancer. The shotgun dataset is publicly available (Barbet et al., 2022). To complement the shotgun-based benchmark dataset, a 16S rRNA gene sequencing dataset of samples from colorectal cancer patients and available metadata was curated by WG3 members, including $n = 709$ samples from previous studies (Zackular et al., 2014; Zeller et al., 2014; Baxter et al., 2016). The final curated dataset is available in the Zenodo repository (Marcos-Zambrano Judith, 2022). WG2 was also responsible for defining and evaluating the ML4Microbiome DREAM Challenge.[7] The challenge was designed to predict incident heart failure risk in a large population-based study of Finnish adults, FINRISK 2002 (Salosensaari et al., 2021), using a combination of gut microbiome data and clinical variables. The results of this DREAM Challenge, completed by 32 participants (seven teams), will be published separately (manuscript in preparation).

1.4. Optimization and standardization of machine learning methods - challenges and solutions

For the optimization and standardization of ML methods, WG3 considered a typical ML workflow that starts after microbiome-related profiles are organized in a two-dimensional table format of features, such as MSP (Metagenomic Species) or Amplicon Sequence Variants (ASV) tables for shotgun or 16S rRNA amplicon data, respectively. This process involves the following steps, (a) data preprocessing (e.g., normalization, filtering), (b) feature selection, (c) predictive modeling, and (d) performance estimation. Our objective was to address the challenges associated with each of these steps considering diverse algorithms, their combinations, and our capacity to interpret and explain their results. Although computational simulations may help estimate expectations and variability under uncertain situations (see, e.g., Gao et al., 2023), we explored benchmark data from the public domain spanning 16 different cohorts from nine countries and derived several noteworthy conclusions.

In data preprocessing, a major challenge lies in selecting the appropriate approaches due to variations in sampling depth, data sparsity (represented by an excess of zeros in the tables) and data compositionality. To first mitigate sampling variability, rarefaction is sometimes used to remove samples. However, this has remained a controversial practice since rarefaction reduces statistical power (McMurdie and Holmes, 2014), but it also provides the means to deal with uncertainties related to variations in read counts that are otherwise challenging to control (Schloss, 2023). Alternatively, researchers incorporate the differences in library size (number of reads per sample) as covariates in the models designed to consider offsets. Sparsity further hampers models that rely on Gaussian assumptions (e.g., linear models), while other models do not have distributional assumptions (e.g., Random Forests, Boosting models). In addition, this sparsity can lead to near-zero variance predictors that turn out to be zero variance predictors during the cross-validation process. Our results indicated that filtering out rare features and removing near-zero variance ones is a successful strategy, outperforming imputation methods in logarithmic transformations. Moreover, standard sequencing techniques cannot capture the total number of bacterial species but only their proportions. For this reason, compositional analysis is the appropriate mathematical framework (Gloor et al., 2017), but its application and impact on ML models are still actively investigated (Greenacre and Blasco, 2021). For example, we found that the CLR transformation can be useful; however, its generalizability to other data sets should be investigated. Therefore, due to the huge variability of approaches and frequently evolving methodologies, we are against giving precise and definitive recommendations.

For feature selection and predictive modeling, the primary challenges revolve around the high dimensionality of the data and the complex interactions inherent to microbial species, including co-occurrence and partial correlation. Building models that incorporate the thousands of microbiome features in a multivariate manner (e.g., principal component regression, partial least squares models) while maintaining predictive performance is undeniably challenging. Boosting or Random Forest models often provided the best performances. Interestingly, using the JADBio autoML approach, we observed that multivariate feature selection through the Statistically Equivalent Signatures algorithm combined with Random Forests could yield an optimal balance between performance and results interpretability and explainability (Tsamardinos et al., 2022). We also emphasize that appropriate performance estimation protocols are crucial to avoid overestimated conclusions and misleading insights. A summary of methods that can be used for each one of the steps of the ML workflow is reported in Table 2 of Papoutsoglou et al. (2023).

7 https://www.synapse.org/#!Synapse:syn27130803/wiki/616705

A novel multi-view learning method was developed based on boosting and multi-armed bandits. The goal was to simultaneously exploit (possibly incomplete) 16S and shotgun data about the same individuals, as well as the features identified through multiple preprocessing pipelines. The obtained results showed significant benefits towards an automated selection and exploitation of multiple views/pipelines for the analysis of microbiome data (manuscript submitted).

1.5. Community building, networking and training: the three key to success

The specific commitments of Working Group 4 (WG4) were to bring networking and training opportunities for emerging talents and thereby strengthen and build up an excellent scientific and technological community, including both ML and microbiome researchers. Providing people with opportunities (internal meetings, conferences and workshops) to discuss and present ideas and experiences was pivotal for establishing collaborations, developing new multidisciplinary interactions, attracting young researchers and providing them with opportunities for their scientific and professional career growth. Thanks to these activities, and despite the interference of the COVID-19 pandemic, the ML4Microbiome network expanded from the initial 24 member countries to 35 (55% from COST Inclusiveness Target Countries), and participants from 57 to 169, among which 48% represented by Young Researchers and Innovators (<40 years). Some could benefit from Short Term Scientific Mission (STMS) grants (16 in total) to work with research teams in different countries on ML4Microbiome-related projects with the view to publish the results of their activities in peer-reviewed journals.[8]

In terms of publication output, to date ML4Microbiome members have published work on specific ML applications for particular diseases, such as Cancer Diagnostics and Therapeutics (Cekikj et al., 2022), classification of patients with Celiac Disease (Arcila-Galvis et al., 2022), Coronary Artery Disease Risk Prediction (Vilne et al., 2022), novel paradigms in human gut microbiome metabolism (Bidkhori et al., 2021), Parkinson's disease (Rosario et al., 2021), Type 2 Diabetes (Ruuskanen et al., 2022), oral and related gut diseases (Di Stefano et al., 2023), along with systematic or scoping reviews on ML applications on microbiome data (Tonkovic et al., 2020; Marcos-Zambrano et al., 2021) and its challenges and solutions (Moreno-Indias et al., 2021) of which all are available from the complete list of the Action's publications on the ML4Microbiome website.

Training schools (TSs) were organized to provide young researchers with the proper background knowledge and hands-on training in MLs techniques applied to microbiome data. Four Training Schools were organized in four different countries, in which 19 trainers and 125 attendants participated over three-five days. Plenary blended learning sessions with keynote speakers were offered, along with high-level lectures covering specific ML-microbiome topics complemented by practical sessions and workshops. The different scientific and geographical backgrounds enhanced multidisciplinary discussions and promoted knowledge exchange between academics and industry participants, leading to scientific publications (Deutsch et al., 2021; Deutsch and Stres, 2021; Deutsch et al., 2022). This also helped trainers learn more about the real needs of young researchers in such a complex multidisciplinary research field, further sharpening the training methods for subsequent TSs. As a result, a syllabus was created, funded by one of Action's STMS, to incorporate ML for microbiome analysis into microbiome MSc courses at various institutes,[9] which previously only addressed read processing, clustering methods, diversity analysis and statistical analysis (manuscript in preparation). All the training material produced by ML4Microbiome, STMS reports, and presentations are freely available from the Action's website (see Footnote 5).

2. Discussion

Currently, microbiome research faces a new bottleneck: its translation into a clinical context, addressing risk, diagnosis/prognosis, and monitoring the effectiveness of therapy. The benefits of such applications involve better methodologies for current bioinformatics challenges, such as species identification from microbiome sequencing data, robust methods for microbiome-derived predictive models or statistical causal inference, and integration of microbiome data with other omics (Feldner-Busztin et al., 2023), among many others (and the possible impact of such applications in the clinic). Statistical modelling and analysis of microbiome-related omics data involve applying various techniques and ML algorithms, which ultimately aim to identify associations (and ideally causality) between microbial taxa, functional genes, metabolites, and host factors (e.g., omics and biochemical variables) with health and disease outcomes. We have outlined the challenges of such analysis and highlighted the importance of developing and optimizing statistical methods and pipelines to handle microbiome data's unique properties for accurate and reproducible microbiome research.

Somewhat disappointingly, albeit not unexpected, there is no unique ML approach to extract the hidden meaningful information beyond the massive microbiome data. Instead, combinations of ML tools seem to be the most promising approach coupled with knowledge of the parameters that need tuning. As we advance, the application of deep learning (DL), a particular component of ML, to microbiome analysis holds significant promise in understanding the intricate relationships between microbial communities and their functions, as well as their links to various diseases and phenotypes (Hernández Medina et al., 2022). We have, however, identified several challenges with implementing DL methods for microbiome data analysis, which can be extended to any ML model, that first need to be addressed. Firstly, the availability (abundance) and quality of microbiome samples and metadata currently limit the collection of large and diverse datasets for the training and validation of DL models, which are even more dependent on large sample sizes. Additionally, there is the issue of interpretability and explainability of DL models, which can restrict the biological insights and hypotheses that can be derived from them. Since many microbiome

8 https://www.ml4microbiome.eu/research-updates/publications-outputs/

9 https://microbiome.github.io/OMA/

analysis applications are related to healthcare, the interpretation of the ML models becomes a priority issue, especially for non-ML experts. Without understanding how the decision was made and the specific reasons for the outcome, many physicians would hesitate to trust the ML results, which could have ethical or legal consequences. In response, Explainable AI (XAI) methods such as SHAP (Shapley Additive exPlanations), DeepSHAP, DeepLIFT, CXplain, and LIME (Lipton, 2016; Chen et al., 2022; Molnar, 2022) have been widely used in recent years. Analysis of microbiome data, such as personalized biomarker identification (Rynazal et al., 2023) and accurate predictions of phenotypes (Carrieri et al., 2021), have also been used to improve the understanding of disease mechanisms and microbiome associations. Nevertheless, XAI has some limitations as many of its models are highly complex and possess many parameters, making it difficult to define the factors that affect the explanation. A tradeoff between explainability and accuracy, which depends on the application area, within which it is determined how critical the accuracy of the model is for the end user.

As ML advances, it is also crucial to consider its ethical implications, particularly its use in clinical practice. One significant ethical consideration in ML and microbiome research is the potential for biased or discriminatory algorithms. It is imperative to ensure that the data sets used to train ML models are diverse and representative of the studied population (Mehrabi et al., 2021). Additionally, the sensitive nature of microbiome data, including health and genetic information and their associated metadata, raises privacy concerns and the need for informed consent (Shabani and Borry, 2018). Therefore, ethical guidelines for data collection, storage, and usage must be implemented to protect individual rights and maintain the integrity and validity of the research (Knoppers and Chadwick, 2005). As such, ML-enabled microbiome research must be conducted responsibly and ethically to ensure that the benefits are equitable, sustainable, and safe (Anomaly, 2017). The outcomes generated by numerous studies have already impacted the microbiome research community. Nevertheless, further advancing the field requires increasing collaborative efforts between microbiologists and ML experts, including stakeholders in non-governmental organizations, health sectors and industry once more standardized ML-microbiome applications start to become available. The main objective of the COST Action ML4Microbiome has significantly improved these opportunities. Thanks to this initiative, we have sown the seeds for a dynamic, interconnected, cross-disciplinary community that has already contributed to advancing research in the field, but with more to come.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

DD'E: Conceptualization, Supervision, Writing – original draft, Writing – review & editing, Visualization. JaaT: Writing – review & editing. LL: Writing – review & editing. MB: Writing – original draft, Writing – review & editing. GeP: Writing – review & editing. MiC: Writing – review & editing. AZ: Writing – review & editing. ML: Writing – original draft, Writing – review & editing. EI: Writing – original draft, Writing – review & editing. AG: Writing – review & editing. AN: Writing – original draft, Writing – review & editing. MF: Writing – review & editing. TK: Visualization, Writing – review & editing. EP: Writing – review & editing. L-MZ: Writing – original draft, Writing – review & editing. KH: Writing – review & editing. GiP: Writing – review & editing. AnS: Writing – review & editing. RamS: Writing – review & editing. IM-I: Writing – review & editing. AT: Writing – review & editing. MN: Writing – review & editing. E-SA: Writing – review & editing. C-OT: Writing – review & editing. RajS: Writing – review & editing. JasT: Writing – review & editing. EB-R: Writing – review & editing. PP: Writing – review & editing. NJ: Writing – review & editing. LF: Writing – review & editing. ST: Writing – review & editing. AlS: Writing – review & editing. GI: Writing – review & editing. DP-S: Writing – review & editing. VT: Writing – review & editing. LK: Writing – review & editing. TL-T: Writing – review & editing. AH: Writing – review & editing. CJ: Writing – review & editing. RB: Writing – review & editing. MaC: Funding acquisition, Project administration, Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This study is based upon work from COST Action ML4Microbiome "Statistical and machine learning techniques in human microbiome studies" (CA18131), supported by COST (European Cooperation in Science and Technology), www.cost.eu. MB acknowledges support through the Metagenopolis grant ANR-11-DPBS-0001. IM-I acknowledges support by the "Miguel Servet Type II" program (CPII21/00013) of the ISCIII-Madrid (Spain), co-financed by the FEDER.

Acknowledgments

The authors are grateful to all COST Action CA18131 "Statistical and machine learning techniques in human microbiome studies" members for their contribution to the COST Action objectives, and to COST (European Cooperation in Science and Technology) for the economic support, www.cost.eu. WG2 and WG3 thank Emmanuelle Le Chatelier and Pauline Barbet (Université Paris-Saclay, INRAE, MetaGenoPolis, 78350, Jouy-en-Josas, France) for preparing the shotgun CRC benchmark dataset.

Conflict of interest

CJ is employed by Biome diagnostics GmbH.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

References

Ahlawat, K., Chug, A., and Singh, A. P. (2021). A novel hybrid sampling algorithm for solving class imbalance problem in big data. *Adv. Data Sci. Adapt. Anal.* 13:2150005. doi: 10.1142/S2424922X21500054

Anomaly, J. (2017). Ethics, antibiotics, and public policy. *Geo. JL Pub. Pol'y* 15, 999–1016.

Arcila-Galvis, J. E., Loria-Kohen, V., Ramírez de Molina, A., Carrillo de Santa Pau, E., and Marcos-Zambrano, L. J. (2022). A comprehensive map of microbial biomarkers along the gastrointestinal tract for celiac disease patients. *Front Microbiol.* 13:956119. doi: 10.3389/fmicb.2022.956119

Balech, B., Brennan, L., Carrillo de Santa Pau, E., Cavalieri, D., Coort, S., D'Elia, D., et al. (2022). The future of food and nutrition in ELIXIR [version 1; peer review: 1 approved with reservations]. *F1000Research* 11:978. doi: 10.12688/f1000research.51747.1

Barbet, P., Almeida, M., Probul, N., Baumbach, J., Pons, N., Plaza Onate, F., et al. (2022). Taxonomic profiles, functional profiles and manually curated metadata of human fecal metagenomes from public projects coming from colorectal cancer studies. *Recherche Data Gouv*, V5, UNF:6:Hif6zWkvCjqmOEJh2lhq0g== [fileUNF]. doi: 10.57745/7IVO3E

Baxter, N. T., Ruffin, M. T., Rogers, M. A., and Schloss, P. D. (2016). Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. *Genome Med.* 8:37. doi: 10.1186/s13073-016-0290-3

Bidkhori, G., Lee, S., Edwards, L. A., Chatelier, E. L., Almeida, M., Ezzamouri, B., et al. (2021). The Reactobiome unravels a new paradigm in human gut microbiome metabolism. bioRxiv 2021.02.01.428114 [Preprint]. Available at: https://www.biorxiv.org/content/10.1101/2021.02.01.428114v1 (Accessed June 28, 2023).

Carrieri, A. P., Haiminen, N., Gardiner, L., Murphy, B., Mayes, A. E., Paterson, S., et al. (2021). Explainable AI reveals changes in skin microbiome composition linked to phenotypic differences. *Sci. Rep.* 11, 1–18. doi: 10.1038/s41598-021-83922-6

Cekikj, M., Jakimovska Özdemir, M., Kalajdzhiski, S., Özcan, O., and Sezerman, O. U. (2022). Understanding the role of the microbiome in cancer diagnostics and therapeutics by creating and utilizing ML models. *Appl. Sci.* 12:4094. doi: 10.3390/app12094094

Chen, H., Lundberg, S. M., and Lee, S. (2022). Explaining a series of models by propagating Shapley values. *Nat. Commun.* 13, 1–15. doi: 10.1038/s41467-022-31384-3

Deutsch, L., Debevec, T., Millet, G. P., Osredkar, D., Opara, S., Šket, R., et al. (2022). (2022) urine and fecal 1H-NMR metabolomes differ significantly between pre-term and full-term born physically fit healthy adult males. *Meta* 12:536. doi: 10.3390/metabo12060536

Deutsch, L., Osredkar, D., Plavec, J., and Stres, B. (2021). Spinal muscular atrophy after Nusinersen therapy: improved physiology in pediatric patients with no significant change in urine, serum, and liquor 1H-NMR metabolomes in comparison to an age-matched, healthy cohort. *Meta* 11:206. doi: 10.3390/metabo11040206

Deutsch, L., and Stres, B. (2021). The importance of objective stool classification in fecal 1H-NMR metabolomics: exponential increase in stool crosslinking is mirrored in systemic inflammation and associated to fecal acetate and methionine. *Metabolites* 11:172. doi: 10.3390/metabo11030172

Di Stefano, M., Santonocito, S., Polizzi, A., Mauceri, R., Troiano, G., Lo Giudice, A., et al. (2023). A reciprocal link between Oral, gut microbiota during periodontitis: the potential role of probiotics in reducing Dysbiosis-induced inflammation. *Int. J. Mol. Sci.* 24:1084. doi: 10.3390/ijms24021084

Feldner-Busztin, D., Firbas Nisantzis, P., Edmunds, S. J., Boza, G., Racimo, F., Gopalakrishnan, S., et al. (2023). Dealing with dimensionality: the application of machine learning to multi-omics data. *Bioinformatics* 39:2. doi: 10.1093/bioinformatics/btad021

Gao, Y., Şimşek, Y., Gheysen, E., Borman, T., Li, Y., Lahti, L., et al. (2023). *miaSim*: an R/Bioconductor package to easily simulate microbial community dynamics. *Methods Ecol. Evol.* 14, 1967–1980. doi: 10.1111/2041-210X.14129

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., and Egozcue, J. J. (2017). Microbiome datasets are compositional: and this is not optional. *Front. Microbiol.* 8:2224. doi: 10.3389/fmicb.2017.02224

Greenacre, M., and Blasco, A. (2021). Compositional data analysis of microbiome and any-omics datasets: a validation of the additive Logratio transformation. *Front. Microbiol.* 12:727398. doi: 10.3389/fmicb.2021.727398

Hernández Medina, R., Kutuzova, S., Nielsen, K. N., Johansen, J., Hansen, L. H., Nielsen, M., et al. (2022). Machine learning and deep learning applications in microbiome research. *ISME Commun.* 2, 1–7. doi: 10.1038/s43705-022-00182-9

Kim, J., and Kim, J. (2018). The impact of imbalanced training data on machine learning for author name disambiguation. *Scientometrics* 117, 511–526. doi: 10.1007/s11192-018-2865-9

Knoppers, B. M., and Chadwick, R. (2005). Human genetic research: emerging trends in ethics. *Nat. Rev. Genet.* 6, 75–79. doi: 10.1038/nrg1505

Lipton, Z. C. (2016). The mythos of model interpretability. *ArXiv*. doi: 10.48550/arXiv.1606.03490 [Epub ahead of preprint].

Marcos-Zambrano, L. J., Karaduzovic-Hadziabdic, K., Loncar Turukalo, T., Przymus, P., Trajkovik, V., Aasmets, O., et al. (2021). Applications of machine learning in human microbiome studies: a review on feature selection, biomarker identification, disease prediction and treatment. *Front. Microbiol.* 12:634511. doi: 10.3389/fmicb.2021.634511

Manor, O., Dai, C. L., Kornilov, S. A., Smith, B., Price, N. D., Lovejoy, J. C., et al. (2020). Health and disease markers correlate with gut microbiome composition across thousands of people. *Nat. Commun.* 11, 1–12. doi: 10.1038/s41467-020-18871-1

Marcos-Zambrano Judith, L. (2022). 16S rRNA sequencing gene datasets for CRC data (1.0.0) [data set]. *Zenodo*. doi: 10.5281/zenodo.7382814

McMurdie, P. J., and Holmes, S. (2014). Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput. Biol.* 10:e1003531. doi: 10.1371/journal.pcbi.1003531

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Comput. Surv.* 54, 1–35. doi: 10.1145/3457607

Molnar, C. (2022). *Interpretable machine learning: a guide for making black box models explainable*. 2nd Edn Available at: https://christophm.github.io/interpretable-ml-book/.

Moreno-Indias, I., Lahti, L., Nedyalkova, M., Elbere, I., Roshchupkin, G., Adilovic, M., et al. (2021). Statistical and machine learning techniques in human microbiome studies: contemporary challenges and solutions. *Front. Microbiol.* 12:635781. doi: 10.3389/fmicb.2021.635781

Papoutsoglou, G., Tarazona, S., Lopes, M. B., Klammsteiner, T., Ibrahimi, E., Eckenberger, J., et al. (2023). Machine learning approaches in microbiome research: challenges and best practices. *Front. Microbiol. Sec. Systems Microbiol.* 14. doi: 10.3389/fmicb.2023.1261889

Pasolli, E., Truong, D. T., Malik, F., Waldron, L., and Segata, N. (2016). Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. *PLoS Comput. Biol.* 12:e1004977. doi: 10.1371/journal.pcbi.1004977

Rosario, D., Bidkhori, G., Lee, S., Bedarf, J., Hildebrand, F., Le Chatelier, E., et al. (2021). Systematic analysis of gut microbiome reveals the role of bacterial folate and homocysteine metabolism in Parkinson's disease. *Cell Rep.* 34:108807. doi: 10.1016/j.celrep.2021.108807

Ruuskanen, M. O., Erawijantari, P. P., Havulinna, A. S., Liu, Y., Méric, G., Tuomilehto, J., et al. (2022). Gut microbiome composition is predictive of incident type 2 diabetes in a population cohort of 5,572 Finnish adults. *Diabetes Care* 45, 811–818. doi: 10.2337/dc21-2358

Rynazal, R., Fujisawa, K., Shiroma, H., Salim, F., Mizutani, S., Shiba, S., et al. (2023). Leveraging explainable AI for gut microbiome-based colorectal cancer classification. *Genome Biol.* 24:21. doi: 10.1186/s13059-023-02858-4

Salosensaari, A., Laitinen, V., Havulinna, A. S., Meric, G., Cheng, S., Perola, M., et al. (2021). Taxonomic signatures of cause-specific mortality risk in human gut microbiome. *Nat. Commun.* 12, 1–8. doi: 10.1038/s41467-021-22962-y

Schloss, P. D. (2023) Rarefaction is currently the best approach to control for uneven sequencing effort in amplicon sequence analyses. *bioRxiv* [Epub ahead of preprint]. doi: 10.1101/2023.06.23.546313

Shabani, M., and Borry, P. (2018). Rules for processing genetic data for research purposes in view of the new EU general data protection regulation. *Eur. J. Hum. Genet.* 26, 149–156. doi: 10.1038/s41431-017-0045-7

Tonkovic, P., Kalajdziski, S., Zdravevski, E., Lameski, P., Corizzo, R., Pires, I. M., et al. (2020). Literature on applied machine learning in metagenomic classification: a scoping review. *Biology* 9:453. doi: 10.3390/biology9120453

Tsamardinos, I., Charonyktakis, P., Papoutsoglou, G., Borboudakis, G., Lakiotaki, K., Zenklusen, J. C., et al. (2022). Just add data: automated predictive modeling for knowledge discovery and feature selection. *NPJ Precision Oncol.* 6:38. doi: 10.1038/s41698-022-00274-8

Vilne, B., Ķibilds, J., Siksna, I., Lazda, I., Valciņa, O., and Krūmiņa, A. (2022). Could artificial intelligence/machine learning and inclusion of diet-gut microbiome interactions improve disease risk prediction? Case study: coronary artery disease. *Front. Microbiol.* 13:627892. doi: 10.3389/fmicb.2022.627892

Voigt, A. Y., Costea, P. I., Kultima, J. R., Li, S. S., Zeller, G., Sunagawa, S., et al. (2015). Temporal and technical variability of human gut metagenomes. *Genome Biol.* 16:73. doi: 10.1186/s13059-015-0639-8

Walsh, I., Fishman, D., Titma, T., Pollastri, G., Harrow, J., Psomopoulos, F. E., et al. (2021). DOME: recommendations for supervised machine learning validation in biology. *Nat. Methods* 18, 1122–1127. doi: 10.1038/s41592-021-01205-4

Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., et al. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* 5:27. doi: 10.1186/s40168-017-0237-y

Zackular, J. P., Rogers, M. A., Ruffin, M. T. 4th, and Schloss, P. D. (2014). The human gut microbiome as a screening tool for colorectal cancer. *Cancer Prev. Res. (Phila.)* 7, 1112–1121. doi: 10.1158/1940-6207.CAPR-14-0129

Zeller, G., Tap, J., Voigt, A. Y., Sunagawa, S., Kultima, J. R., Costea, P. I., et al. (2014). Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.* 10:766. doi: 10.15252/msb.20145645

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 27 September 2023
DOI 10.3389/fmicb.2023.1227300

OPEN ACCESS

EDITED BY
Isabel Moreno Indias,
University of Malaga, Spain

REVIEWED BY
Zachary R. Stromberg,
Pacific Northwest National Laboratory (DOE), United States
Dhiraj Kumar,
National Eye Institute (NIH), United States
Arturo Ortega,
Center for Research and Advanced Studies of the National Polytechnic Institute, Mexico

*CORRESPONDENCE
Wei-Ning Lin
✉ 081551@mail.fju.edu.tw

RECEIVED 23 May 2023
ACCEPTED 06 September 2023
PUBLISHED 27 September 2023

CITATION
Chang C-C, Liu T-C, Lu C-J, Chiu H-C and Lin W-N (2023) Machine learning strategy for identifying altered gut microbiomes for diagnostic screening in myasthenia gravis. *Front. Microbiol.* 14:1227300. doi: 10.3389/fmicb.2023.1227300

Machine learning strategy for identifying altered gut microbiomes for diagnostic screening in myasthenia gravis

Che-Cheng Chang[1,2,3], Tzu-Chi Liu[4], Chi-Jie Lu[4,5,6], Hou-Chang Chiu[7,8] and Wei-Ning Lin[3]*

[1]PhD Program in Nutrition and Food Science, Fu Jen Catholic University, New Taipei City, Taiwan, [2]Department of Neurology, Fu Jen Catholic University Hospital, Fu Jen Catholic University, New Taipei City, Taiwan, [3]Graduate Institute of Biomedical and Pharmaceutical Science, Fu Jen Catholic University, New Taipei City, Taiwan, [4]Graduate Institute of Business Administration, Fu Jen Catholic University, New Taipei City, Taiwan, [5]Artificial Intelligence Development Center, Fu Jen Catholic University, New Taipei City, Taiwan, [6]Department of Information Management, Fu Jen Catholic University, New Taipei City, Taiwan, [7]School of Medicine, Fu Jen Catholic University, New Taipei City, Taiwan, [8]Department of Neurology, Taipei Medical University, Shuang-Ho Hospital, New Taipei City, Taiwan

Myasthenia gravis (MG) is a neuromuscular junction disease with a complex pathophysiology and clinical variation for which no clear biomarker has been discovered. We hypothesized that because changes in gut microbiome composition often occur in autoimmune diseases, the gut microbiome structures of patients with MG would differ from those without, and supervised machine learning (ML) analysis strategy could be trained using data from gut microbiota for diagnostic screening of MG. Genomic DNA from the stool samples of MG and those without were collected and established a sequencing library by constructing amplicon sequence variants (ASVs) and completing taxonomic classification of each representative DNA sequence. Four ML methods, namely least absolute shrinkage and selection operator, extreme gradient boosting (XGBoost), random forest, and classification and regression trees with nested leave-one-out cross-validation were trained using ASV taxon–based data and full ASV–based data to identify key ASVs in each data set. The results revealed XGBoost to have the best predicted performance. Overlapping key features extracted when XGBoost was trained using the full ASV–based and ASV taxon–based data were identified, and 31 high-importance ASVs (HIASVs) were obtained, assigned importance scores, and ranked. The most significant difference observed was in the abundance of bacteria in the *Lachnospiraceae* and *Ruminococcaceae* families. The 31 HIASVs were used to train the XGBoost algorithm to differentiate individuals with and without MG. The model had high diagnostic classification power and could accurately predict and identify patients with MG. In addition, the abundance of *Lachnospiraceae* was associated with limb weakness severity. In this study, we discovered that the composition of gut microbiomes differed between MG and non-MG subjects. In addition, the proposed XGBoost model trained using 31 HIASVs had the most favorable performance with respect to analyzing gut microbiomes. These HIASVs selected by the ML model may serve as biomarkers for clinical use and mechanistic study in the future. Our proposed ML model can identify several taxonomic markers and effectively discriminate patients with MG from those without with a high accuracy, the ML strategy can be applied as a benchmark to conduct noninvasive screening of MG.

KEYWORDS

myasthenia gravis, amplicon sequence variants, gut microbiota, machine learning, extreme gradient boosting, leave one out cross validation

1. Introduction

Myasthenia gravis (MG) is a neuromuscular junction disorder that occurs when autoantibodies bind to components of the postsynaptic muscle membrane. The most easily observed symptom is fluctuating skeletal muscle weakness (Gilhus, 2016). The development of immunomodulating treatments has significantly improved the prognosis for patients with MG (Farrugia and Goodfellow, 2020; Narayanaswami et al., 2021). Although well-established management options for MG are widely available, MG can be difficult to identify because its clinical symptoms often vary considerably and may overlap with those of other neurological disorders. Furthermore, antibody testing, which is crucial for confirming a diagnosis of MG, can be expensive, time-consuming, and not readily available and has a high rate of false negatives (Gilhus, 2016). In addition, relapse-related symptoms and their severity can vary greatly by person to person (Hehir and Silvestri, 2018). Otherwise, the severity of MG can be difficult to assess in patients with positive for acetylcholine receptor antibodies because no clear association has been established between the antibody titer and disease severity (Berrih-Aknin and Le Panse, 2014). No marker of MG has been discovered that can assist in the diagnosis, follow-up, therapy response monitoring, and clinical variability determination of the disease.

Research revealed that gut microbiomes may contain biomarkers that can be used to evaluate several neurological diseases, such as Parkinson's disease (Lin et al., 2019). A growing body of evidence indicates that gut microbiota may be associated with immune function dysregulation, which can result in several autoimmune diseases (Pianta et al., 2017; Shahi et al., 2017; Gopalakrishnan et al., 2018; Qiu et al., 2018). Evidence also indicates that T-regulatory cells are present in large quantities in the intestinal mucosa and that microbial components and their metabolites may be involved in maintaining the homeostasis of the immune system (Chen and Tang, 2021). While several studies have demonstrated dysbiosis in autoimmune diseases, there remains a limited focus on neuromuscular disorders. Recently, there has been growing attention to the disturbance of microbiome composition and gut dysbiosis in MG, as well as its comorbidity with anxiety (Zhang et al., 2022; Kapoor et al., 2023). However, how gut microbiota alterations affect the course of such diseases remains unclear, and no method for identifying key features in gut microbiota has been discovered.

Machine learning (ML) methods, as a strategy of artificial intelligence (AI), that can successfully recognize patterns in clinical data, it can be efficiently used for triage, screening, diagnosis, and biomarker identification, and the joint use of manual and ML evaluations can offer more efficient and accurate results than the use of one method alone (Liu et al., 2019). Numerous studies have applied ML techniques to collect and analyze human microbiome data to elucidate the diverse taxonomies and functions of microbial communities and their effects on human health. However, no one-size-fits-all ML technique is available for analyzing gut microbiomes or determining which bacteria is most associated with MG. The identification of a simple screening test for the early detection of MG would allow for a timely diagnosis and the initiation of prompt treatment intervention.

Some studies have reported that the microbiota composition in the fecal samples of MG groups differed from those of healthy control groups (Moris et al., 2018; Qiu et al., 2018). Gut microbiota has been proposed as a potential diagnostic biomarker for MG therapies and early detection of progression (Kang et al., 2022; Thye et al., 2022). However, few studies have compared the feasibility and potential accuracy of applying an ML strategy to evaluate the gut microbiomes of individuals with MG. Our study hypothesized that the compositions of the gut microbiomes of individuals with and without MG would differ and that supervised ML models could be trained using gut microbiota data to provide diagnostic screening results for MG and predict clinical severity. Our study tested several ML analysis methods to identify the most favorable strategy for identifying MG. The results indicate that ML-based strategies can aid in identifying how microbiomes change in relation to MG and that the tree-based method extreme gradient boosting (XGBoost) performs the best (Chen and Guestrin, 2016). In addition, an ML-based support tool for measuring gut microbial populations was developed.

2. Materials and methods

2.1. Human subjects and sample/data collection

In this prospective study, 19 individuals with MG and 10 individuals without were consecutively recruited from Fu-Jen Catholic University Hospital. Individuals were enrolled in the MG group if they (1) were given a diagnosis of MG on the basis of having the combination of symptoms and signs that are characteristic of muscle weakness with diurnal changes and either (2a) had a positive test result for specific autoantibodies or (2b) had a positive electrophysiological diagnosis obtained using single-fiber electromyography and repetitive nerve stimulation (Rousseff, 2021). None of the participants had received any abdominal chirurgic intervention; consumed antibiotics, probiotics, or antacids during the previous 6 months; or reported gastrointestinal symptoms during the previous year. This study was approved by the Research Ethic Committee of Fu-Jen Catholic University Hospital and written informed consent was obtained from each participant (No. FJUH109042). All experiments were completed in accordance with the Declaration of Helsinki's Ethical Principles for Medical Research Involving Human Subjects and under a set of approved guidelines and regulations. The severity of MG was determined using quantitative MG (QMG), MG activities of daily living (MG-ADL), MG composition (MGC), and MG quality of life (MG-QoL) scores (Jaretzki et al., 2000). Using the categories of the QMG and MGC scales, we categorized the scores on these scales into ocular, bulbar, and limb groups. Figure 1 summarizes the overall study workflow.

2.2. Sample collection and processing

After the participants have completed the informed consent form and agreed to participate in the study, fecal samples from each volunteer were collected after enrollment. Volunteers self-collected Fresh stool samples after defecation in the hospital and immediately transferred the samples to a laboratory freezer at −80°C for cryopreservation.

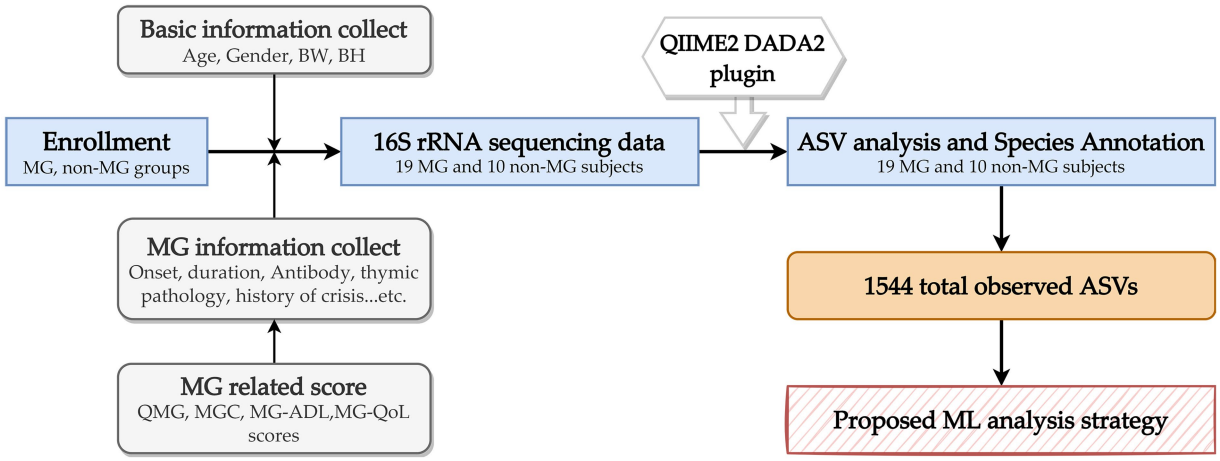

FIGURE 1
Overall study workflow. MG, myasthenia gravis; BW, body weight; BH, body height; QMG, quantitative MG score; MGC, MG composite score; MG-ADL, MG activities of daily living score; ASV, amplicon sequence variants; ML, machine learning.

Each stage in the process, including the sample testing and polymerase chain reaction (PCR) and library creation and sequencing, can affect the quality of the data, and the accuracy of analytical findings is directly influenced by the quality of data. Therefore, quality control measures were implemented at each stage of the process to ensure data accuracy.

2.3. DNA extraction and 16S metagenomics sequencing

Genomic DNA was extracted from the samples using the EasyPrep Stool Genomic DNA Kit (Biotools, New Taipei City, Taiwan). The DNA concentration was determined and adjusted to 5 ng/μL for subsequent processing. In accordance with the 16S Metagenomic Sequencing Library Preparation protocol (Illumina), the specific primer set 341F: 5'-CCTACGGGNGGCWGCAG-3′, 806R: 5'-GACTACHVGGGTATCTAATCC-3′ was employed to amplify the variable regions V3 and V4 of the 16S rRNA gene. A PCR was conducted using KAPA HiFi HotStart ReadyMix (Roche) and 12.5 ng of genomic DNA (gDNA) under the following conditions: 95°C for 3 min, 25 cycles of 95°C for 30 s, 55°C for 30 s, 72°C for 30 s, and a final extension of 72°C for 5 min. The reaction was subsequently maintained at 4°C. The products of the PCR were evaluated using 1.5% agarose gel, and samples with a bright main strip at approximately 500 bp were selected for further library preparation. The selected samples were purified using AMPure XP beads.

A sequencing library was prepared using the 16S Metagenomic Sequencing Library Preparation procedure (Illumina). To summarize, the 16S rRNA V3–V4 region PCR amplicon was subjected to a secondary PCR, which was conducted using the Nextera XT Index Kit with dual indices and Illumina sequencing adapters from Illumina. The indexed PCR product was evaluated for quality by using the Qubit 4.0 Fluorometer (Thermo Scientific) and a Qsep100™ system. The indexed PCR products were mixed in equal amounts to create a sequencing library. The library was sequenced on an Illumina MiSeq platform, which generated 300-bp paired reads.

2.4. Microbial community analysis and statistical analysis

Amplicon sequencing was performed using 300-bp paired-end raw reads, and each sample was demultiplexed on the basis of their barcode identification. Primer and adapter sequences were removed from the paired-end reads by using the QIIME2 cutadapt plugin (Martin, 2011). To construct amplicon sequence variants (ASVs), a denoising pipeline was applied using the QIIME2 DADA2 plugin (v2021.4) to implement quality filtering, dereplication, dataset-specific error model learning, denoising, paired-end-read joining, and chimera removal (Callahan et al., 2016). Trimming and filtering were performed with a maximum of two expected errors per read (maxEE = 2). The DADA2 algorithm was used to solve the problem of exact merged paired-end reads with an overlapping 12-base pair near-zero error rate. The feature-classifier and algorithm of QIIME2 was employed to annotate the taxonomic classification of each representative sequence on the basis of information retrieved from the Silva database (Bokulich et al., 2018). To analyze the sequence similarities among the ASVs, multiple sequence alignment was conducted, with the QIIME2 alignment MAFFT used against the Silva database (Katoh and Standley, 2013). A QIIME2 phylogeny fast tree was used to construct a phylogenetic tree with a set of sequences representative of the ASVs (Price et al., 2010).

2.5. Taxonomic analysis

The taxa that significantly differed between the MG and non-MG samples were identified, and an analysis of the overlap between the taxa of these samples was conducted. Significant biomarkers were identified through Linear discriminant analysis effect size (LEfSe) analysis (Segata et al., 2011). Subsequently, linear discriminant analysis (LDA) is applied for the bacterial taxa identified as significantly different to determine the effect size of each differentially abundant taxon. In the present study, taxa with an LDA score > 2 were considered significant.

2.6. Supervised ML modeling and proposed ML analytical strategy

This study applied four ML methods, namely least absolute shrinkage and selection operator (Lasso), XGBoost, random forest (RF), and classification and regression trees (CART). Because taxonomy or ASVs-based ML approaches provide different types of information, the present study proposed an ML analytical strategy that combines the benefits and valuable information of each approach that can be used to effectively screen key taxon features. Figure 2 presents the proposed ML analytical strategy. In the strategy, two sets of data obtained using different approaches, namely ASV taxon–based data and full ASV–based data, are prepared. Four ML methods (Lasso, CART, XGBoost, and RF) and nested leave-one-out cross-validation (LOOCV) are applied to complete ML model building for each data set, and the model with the highest performance is selected. The key features of each data set are extracted, and the overlapping key features of the data sets are screened to obtain a final set of key features.

LOOCV was executed for the construction of each ML model. In essence, LOOCV is similar to k-fold cross-validation. The primary difference between the two is that k-fold cross-validation involves validation with one of several equally sized folds that have been randomly divided from the data whereas LOOCV involves using a single subset of the data for all rounds of the validation process (Vabalas et al., 2019). Figure 3 illustrates the nested LOOCV process used in this study.

The performance of the model was evaluated on the basis of its accuracy (ACC), precision (PRE), sensitivity (SEN), specificity (SPE) and area under the receiver operating characteristic curve (AUC). The study experiments were conducted using Python (version 3.8.8) and Jupyter Notebook (version 6.3.0) softwares (Van Rossum and Drake, 1995; Kluyver et al., 2016). XGBoost was implemented using the XGBoost package (version 1.3.3) (Chen and Guestrin, 2016), and Lasso, CART, and RF were implemented using the scikit-learn package application programming interfaces (API) (version 0.24.2) (Pedregosa et al., 2011; Buitinck et al., 2013). LOOCV and hyperparameter tuning were implemented using the scikit-learn API (Pedregosa et al., 2011).

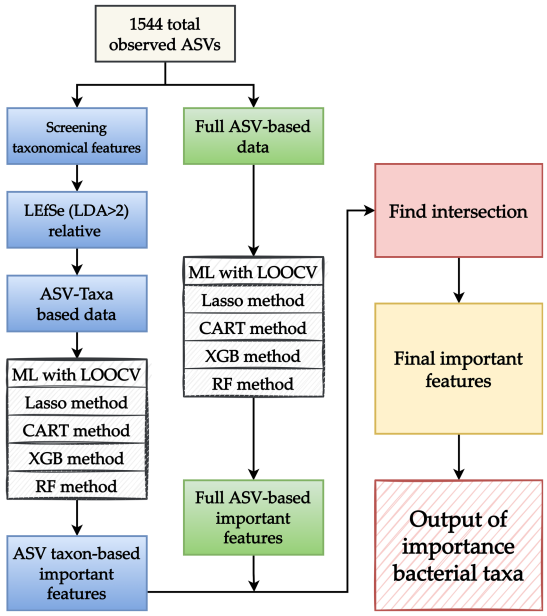

FIGURE 2
Proposed machine learning (ML) analytical strategy. After amplicon sequence variants (ASVs) are obtained, two sets of data obtained through different approaches, namely an ASV taxon–based data set (blue boxes) and full ASV–based data set (green boxes), are prepared. For the ASV taxon–based data, taxon analysis is used to screen 1,544 observed ASVs for key features. LEfSe is applied to identify taxonomic features with a linear discriminant analysis value of >2 to develop an ASV taxon–based data set. For the full ASV–based data, the 1,544 ASVs are directly used without any modifications. After the ASV taxon–based and full ASV–based data sets are created, four ML methods (Lasso, CART, XGBoost, and RF) and leave-one-out cross-validation (LOOCV) are applied, and the model with the best performance is selected. Key features are identified by applying the selected model to the key features identified in the aforementioned data sets. Because each approach provides different types of information, the overlapping key features identified in the two data sets are screened and collected to obtain a final set of key features. The output of key bacterial taxonomic features was used to identify the taxa associated with myasthenia gravis.

3. Results

Individuals who met the criteria for a diagnosis of MG were included in the present study. The mean age at enrollment was 51.5 years, and the majority of the participants were women (68%). The mean disease duration was 59.2 months. In addition, 36% of the patients with MG had a history of an MG crisis, and 21% had experienced life-threatening events at the onset of the disease. The clinical characteristics of the 19 individuals in the MG groups and 10 individuals in the non-MG group were obtained from their medical records (Table 1). The two groups did not significantly differ with respect to their age, sex, body weight, and height. To investigate the bacterial gut microbiota associated with MG, we conducted high-throughput sequencing of the V3–V4 region of the 16S ribosomal RNA gene. We obtained 1,544 ASV observations and used these ASVs to extract taxonomic information from the samples obtained from the MG and non-MG groups. A Venn diagram of the results that revealed 766 and 332 ASVs to be specific to individuals with and without MG, respectively, and 446 ASVs to be shared by individuals with and without MG (Figure 4). We also created cumulative bar charts for each taxonomic class (Supplementary Figure S1).

3.1. Differences in bacterial taxa between the MG and non-MG groups

To identify the significant differences in the gut microbiota between the MG and non-MG groups, we used LEfSe to identify eight taxonomic features with notable significant differences between the two groups (LDA > 2; Figures 5A,B). At the genus level, *Roseburia*, *Oscillospira*, and *Mitsuokella* were more abundant in the non-MG group (Figure 5A); at the class level, *Coriobacteriia* was more abundant in the MG group; and at the order level, *Coriobacteriales* was more abundant in the MG group. The abundances of several major bacterial taxa in the MG and non-MG groups and their phylogenetic relationships are presented in a cladogram in Figure 5B. The abundance of many species in the gut microbiomes of the MG and non-MG groups significantly differed. Figures 5C,D presents representative examples of the bacterial abundance at the family- and genus-levels in the two groups. These results support the hypothesis that the composition of gut microbiota of the MG and non-MG groups differed considerably.

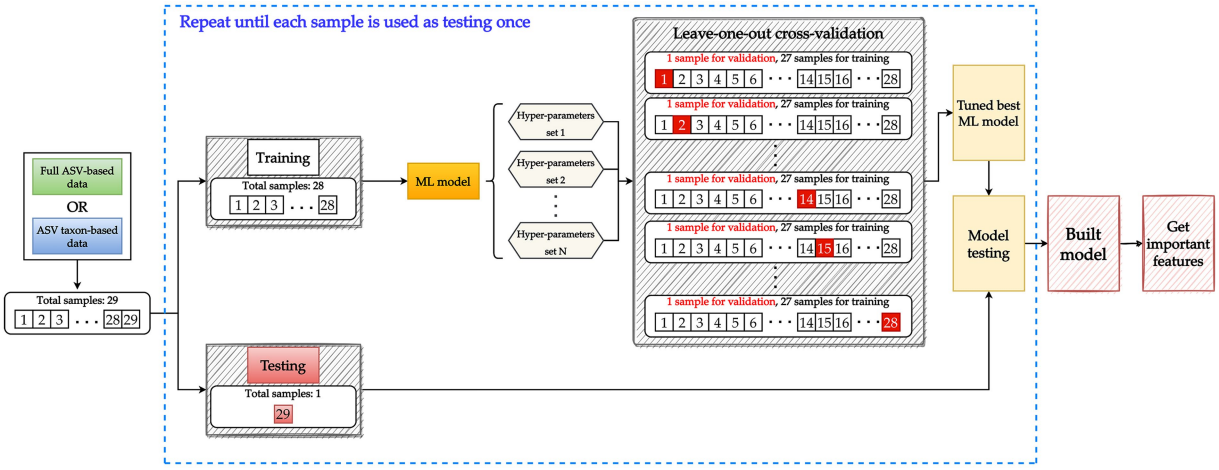

FIGURE 3
Model building involving a nested leave-one-out cross-validation (LOOCV) structure. Data are split into 28 samples for training and 1 sample for testing. To tune the hyperparameters of the ML model, the training data set with 28 samples is further split into 27 samples for training and 1 sample for validation. Each sample is used once for validation until all 28 samples have been used to validate all potential hyperparameter sets. A tuned model is then constructed using the training data (28 samples) and the best hyperparameter set. The testing data are used to evaluate the performance of the tuned model. The aforementioned process constitutes one iteration and is repeated until each sample has been used once as testing data. Key features are then extracted from the tuned model.

TABLE 1 Characteristics of subjects with MG and non-MG groups.

Characteristic	MG ($n = 19$)	Non-MG ($n = 10$)	p-value
Sex Female, *n* (%)	13 (68)	8 (80)	0.8212
Age (year)	51.5 ± 14.4	49.8 ± 13.9	0.7731
Height (cm)	161.4 ± 7.9	161.4 ± 4.7	0.9941
Weight (kg)	64.9 ± 15.9	63.7 ± 12.4	0.8432
BMI (kg/m^2)	26.7 ± 4.4	24.3 ± 3.3	0.8141
Age at onset (age)	45.6 ± 14.9	–	–
Disease duration (month)	59.2 ± 77.8	–	–
Serology of AchR antibody, *n* (%)	18 (95)		
History of MG crisis, *n* (%)	7 (36)	–	–
Life threatening at onset, *n* (%)	4 (21)	–	–
Thymic pathology		–	–
Thymoma, *n* (%)	8 (42)		
Thymic hyperplasia, *n* (%)	1 (5)		
Previous Thymectomy	6 (32)	–	–
MGFA clinical class, *n* (%)		–	–
Class II	12 (63)		
Class III	4 (21)		
Class IV	3 (16)		
Daily Pyridostigmine dose (mg)	192 ± 114	–	–
PSL dose per day (mg)	9.2 ± 10.5		
IS usage, *n* (%)	2 (11)	–	–
MGQOL score	12.8 ± 13.7	–	–
QMGS	10.3 ± 4.2	–	–
QMGS – ocular	1.7 ± 1.4	–	–
QMGS – bulbar	2.4 ± 2.2	–	–
QMGS – limbs	5.1 ± 2.8	–	–
MGC	8.5 ± 8.7	–	–
MGC – ocular	1.7 ± 1.6	–	–
MGC – bulbar	5.8 ± 6.9	–	–
MGC – limbs	0.9 ± 1.6	–	–
MG-ADL	4.74 ± 4.33	–	–
Antibody titer (Nmol/L)	81.2 ± 70.8	–	–

Anti-AChR Ab, antibody against acetylcholine receptor; Anti-MuSK Ab, antibody against Muscle-specific tyrosine kinase; dSN, double seronegative; AZA, treatment with azathioprine; MMF, treatment with mycophenolate; OT, treatment with tacrolimus; IVIG, treatment with intravenous immunoglobulin; PP, treatment with plasmapheresis; PSL, prednisolone.

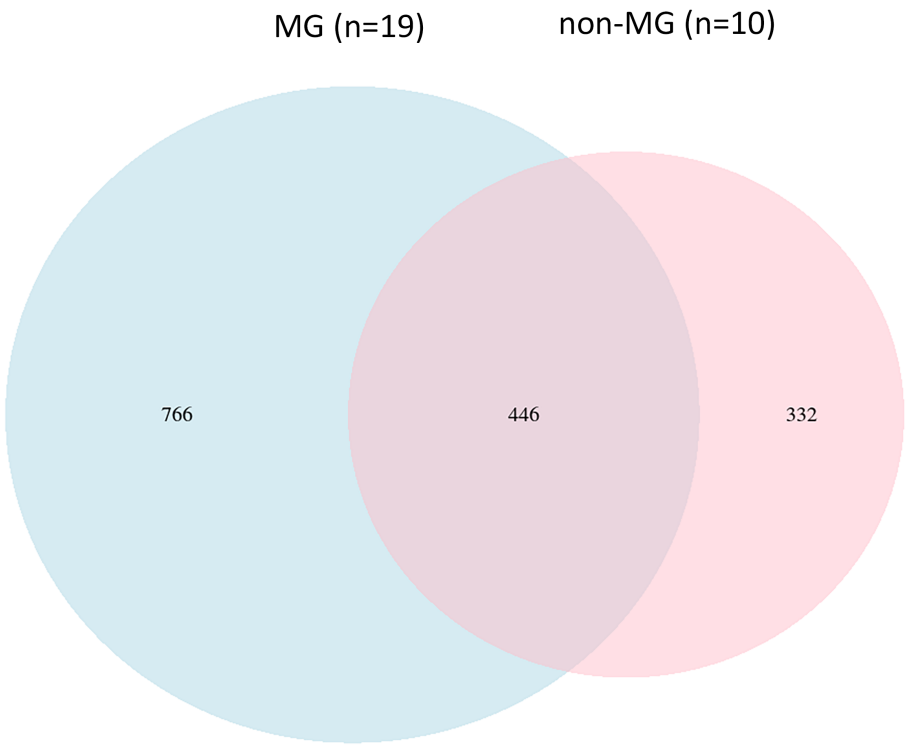

FIGURE 4
Comparison of the gut microbial composition among the two groups at ASV levels. A Venn diagram demonstrated a total of 1,544 ASVs, 446 were detected in both groups and 766, and 332 were unique to participants with (blue circle, $n = 19$) and without (pink circle, $n = 10$) MG, respectively.

3.2. Supervised ML analysis using enriched taxonomic features

To investigate the performance of ML methods based on different datasets, we trained supervised ML models with the taxonomic or ASV features for predictive classification and diagnostics of MG and non-MG. When enriched taxonomic features (ASV taxon–based data) were used for training, the four ML models were trained using eight taxonomic features (described above) to complete predictive classification and diagnosis of MG. Table 2 presents the performance results for the four ML models trained with ASV taxon–based data. As indicated in the table, XGBoost had the highest AUC (90.00), followed by RF (75.26), Lasso (67.89), and CART (35.26). Precision was used to measure the overall correctness of predictions of positive cases. The XGBoost model had a precision score of 100, indicating that a positive prediction by XGBoost is most likely correct. Overall, XGBoost had the highest performance when ASV taxon–based data were used for training and is thus promising as a means of correctly predicting positive cases.

3.3. Supervised ML analysis using ASV features

The four ML models were trained with all 1,544 ASV features (full ASV–based data) to investigate the effectiveness of diagnostic classifications made on the basis of all ASVs. Table 3 presents the results. Similar to the ASV taxon–based models, the full ASV–based models were such that XGBoost had the highest AUC score (87.89), followed by RF (63.68), Lasso (56.32), and CART (46.32). In the full ASV–based model, XGBoost had a promising precision score of 100. The results indicated that XGBoost had the highest performance when the full ASV–based data were used. A comparison of the AUCs of XGBoost when ASV taxon–based data (AUC = 90.00) and full ASV–based data (AUC = 87.89) were used was conducted using the Delong test. The results revealed no statistical difference between the two ($p = 0.43$), indicating that XGBoost performed well regardless of which data set was used. Through the combination of two distinct datasets analysis, XGBoost emerges as the superior ML method for effectively distinguishing between MG and non-MG subjects. This robust outcome underscores the promising potential of ML methods in disease diagnosis within gut microbiomes.

3.4. XGBoost performance higher than RF on training data with enrich taxonomic and ASV features

To further assess the performance of XGBoost compared to traditional machine learning methods, we utilized the receiver operating characteristic (ROC) curve for additional verification. The performance of XGBoost remained similar when different forms of data were used as inputs (Figure 6). For purposes of comparison, RF was also included because it is commonly used in gut microbiome–related studies (Lee and Rho, 2022). The comparison of the XGBoost and RF models when different types of

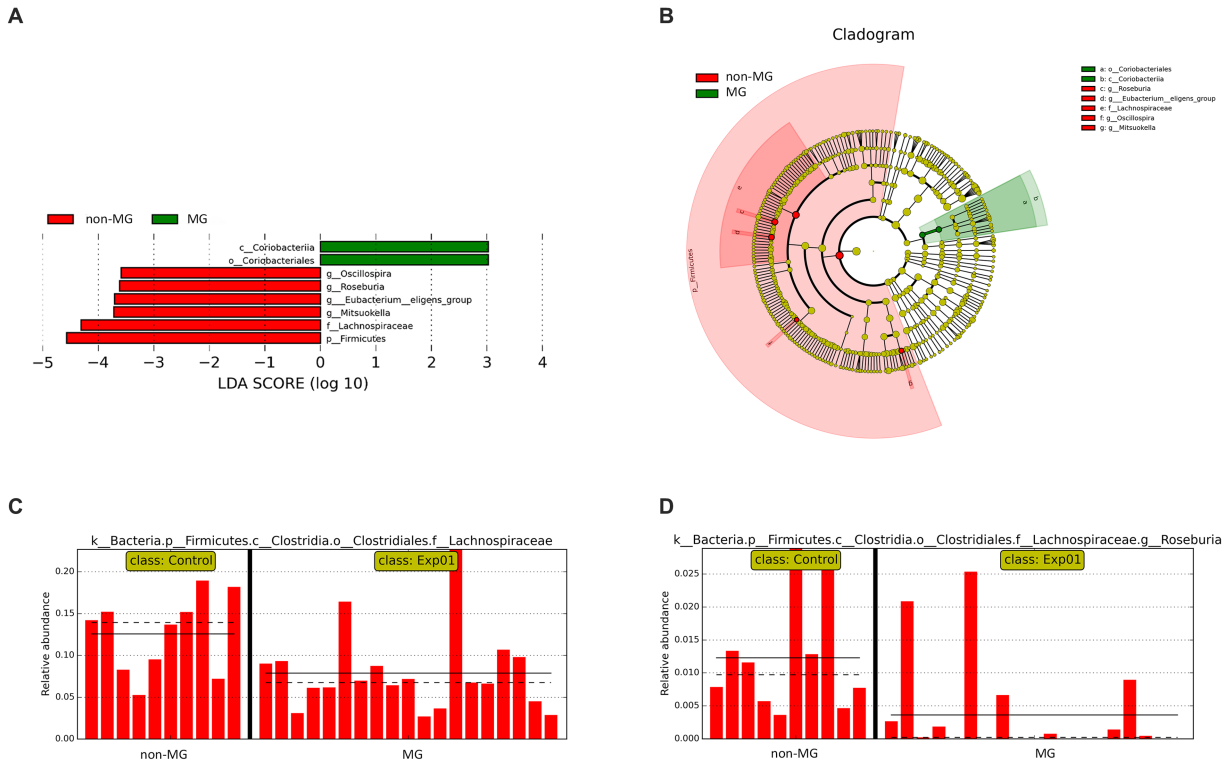

FIGURE 5
Taxonomic differences between the fecal microbiota of the MG and non-MG groups. **(A)** Cladogram created using linear discriminant analysis effect size (LEfSe) and presenting the phylogenetic distribution of the fecal microbiota of individuals with and without MG. **(B)** Linear discriminant analysis (LDA) and LEfSe revealed significant differences in the fecal microbiota of the MG (positive LDA score) and non-MG groups (negative LDA score). LDA scores (log10) > 2 are presented. **(C,D)** Representative examples of the relative abundances of *Lachnospiraceae* and *Roseburia* in individuals with and without MG, with each bar representing the abundance in a given sample. Solid and dashed lines indicate the mean and median, respectively.

TABLE 2 ML analysis using taxonomic features (ASV taxon–based ML analysis).

Method	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	AUC
Lasso	75.86	80.00	84.21	60.00	67.89
CART	41.38	66.67	21.05	80.00	35.26
XGboost	82.76	100	73.68	100	90.00
RF	75.86	100	63.16	100	75.26

AUC, area under the curve.

TABLE 3 ML results when full ASV–based data (full ASV–based ML analysis).

Method	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	AUC
Lasso	72.41	76.19	84.21	50.00	56.32
CART	65.52	71.43	78.95	40.00	46.32
XGboost	86.21	100	78.95	100	87.89
RF	58.62	100	36.84	100	63.68

AUC, area under the curve.

data were used (ASV taxon–based and full ASV–based data) revealed that XGBoost had a higher AUC than RF did, and the results were similar when the full ASV–based and ASV taxon–based data were used (Figure 6). In summary, XGBoost demonstrates high performance when trained using both general ASV data and key taxonomy features, making it a reliable tool for screening and diagnosing MG.

3.5. ML models trained with a combination of taxonomic and ASV features able to identify markers of MG

To improve the diagnostic classification performance of the ML model, we integrated the results obtained from both the full ASV–based and ASV taxon–based datasets. The overlapping key features

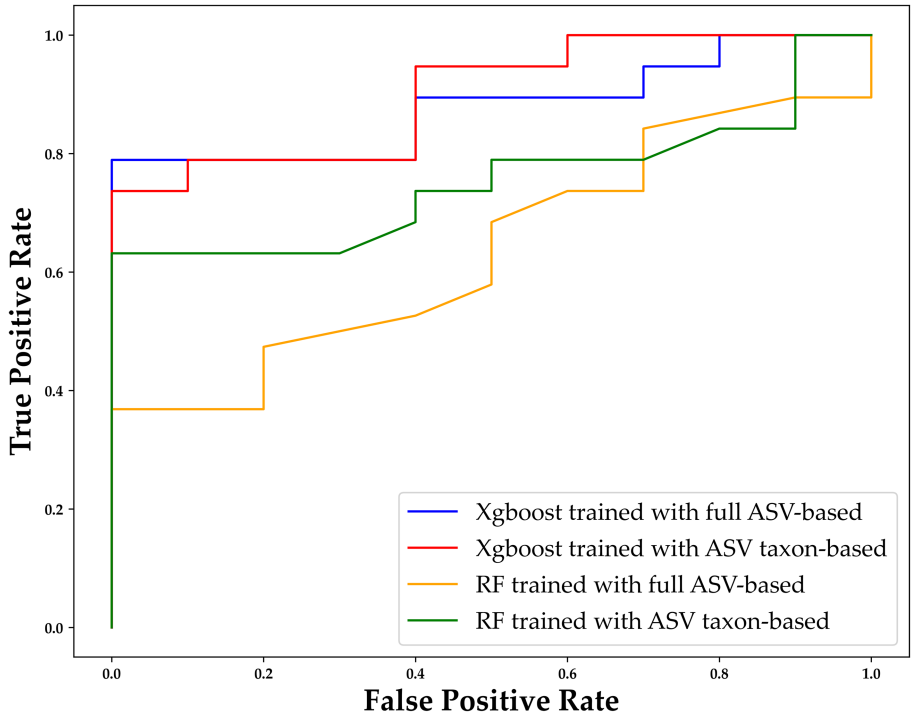

FIGURE 6
ROC curve of XGBoost and random forest (RF) with different types of data. The horizontal axis indicates the false positive rate (1–SPE), and the vertical axis indicates the true positive rate (SEN). The results for XGBoost trained with the full ASV–based and ASV taxon–based data are indicated in blue and red, respectively, and the results for RF trained with the full ASV–based and ASV taxon–based data are indicated in orange and green, respectively. ASV, amplicon sequence variant; XGBoost, extreme gradient boosting; RF, random forest.

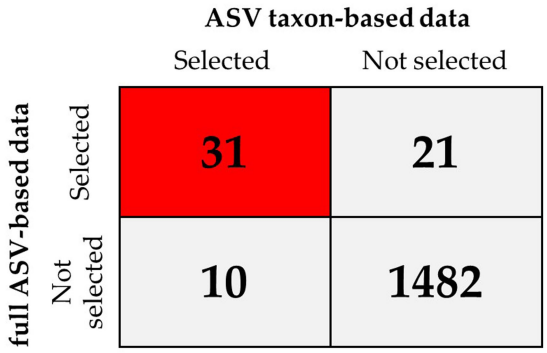

FIGURE 7
XGBoost feature selection results when the model was trained using full ASV–based and using ASV taxon–based data for comparison. The results revealed that of the 1,544 ASVs in total, 31 were selected by XGBoost when it was trained using the full ASV–based and ASV taxon–based data (red square), which indicated these were high-importance ASVs (HIASVs).

extracted when XGBoost was trained using the full ASV–based and ASV taxon–based data were identified and are presented in Figure 7. Thirty-one high-importance ASVs (HIASVs) were identified in the ML model when the full ASV–based and ASV taxon–based data were used. The HIASVs were assigned variable importance scores and ranked (Figure 8; Supplementary Tables S1, S2). All of the overlapping microorganisms belonged to the phylum Firmicutes. The findings revealed that the most significant difference between the gut microbiota of the individuals with and without MG was in the abundance of bacteria in the *Lachnospiraceae* and *Ruminococcaceae* families. The XGBoost algorithm was reapplied with the 31 HIASVs used to differentiate individuals with and without MG. In the XGBoost trained with the HIASVs, the dimensionality of the feature space was reduced, and the model had the highest AUC (90.53) and performed slightly better than the other ML models (Figure 9; Supplementary Table S3). The ML strategy we developed provided compelling evidence supporting our hypothesis, as it demonstrated high diagnostic classification power and generated accurate diagnostic screening results for MG.

3.6. Associations between gut microbiota and clinical characteristics of MG

To investigate the potential links between gut microbiome disruptions and MG clinical symptoms, a correlation analysis was conducted with a focus on the taxa of Firmicutes, *Lachnospiraceae*, *Roseburia*, and *Eubacterium*, the abundance of which was determined to significantly differ between the MG and non-MG groups. A heat map was used to present the spearman's rank correlation coefficients of the 4 significant taxa and results on 22 clinical indices. We discovered that the abundance of *Lachnospiraceae* was generally associated with the severity of limb weakness, that is, with the limb portion of the QMG (Figure 10). These findings demonstrate that certain gut microbiota levels are associated with clinical parameters

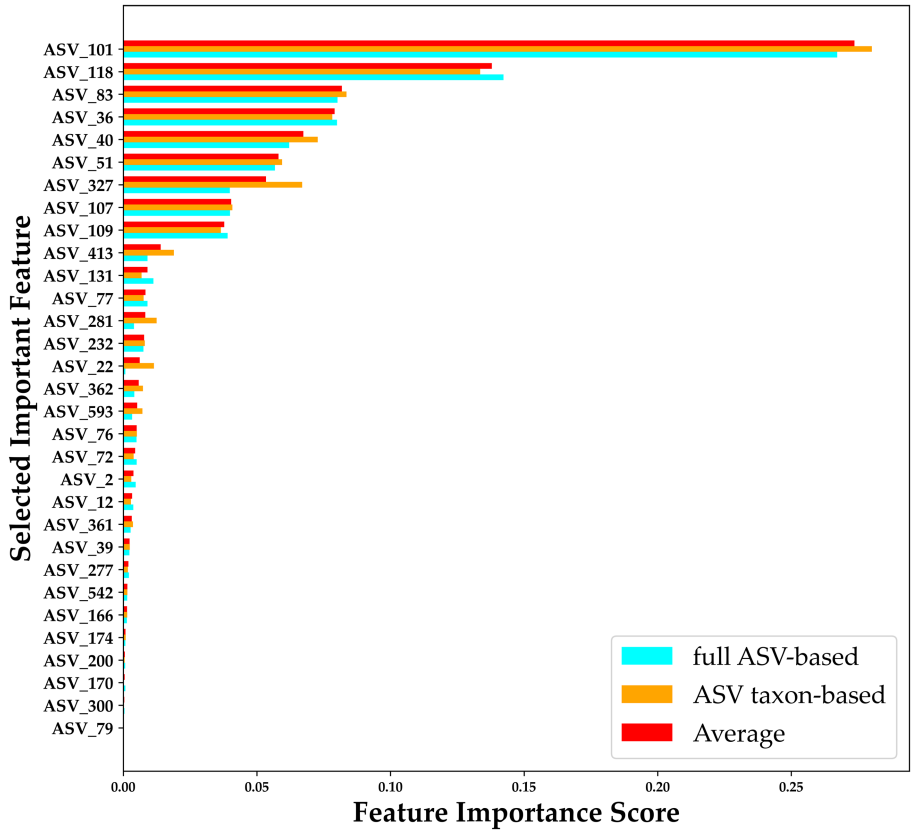

FIGURE 8
Importance scores for 31 HIASVs for classifying the presence and absence of myasthenia gravis. A comparison of the ASV feature importance score is presented in the figure, with blue indicating an importance score assigned when XGBoost trained with full ASV–based data was used, orange indicating an importance score assigned when XGBoost trained with ASV taxon–based data was used, and red indicating the average of the importance scores assigned by the Full ASV–trained and ASV taxon–trained XGBoost models. The average score was used to rank the ASVs. ASV, amplicon sequence variant.

and have the potential to serve as valuable tools for assessing disease severity in the future.

4. Discussion

In this study, we discovered that the structures and composition of the gut microbiome were differed between MG and non-MG subjects. Among our research participants with MG, 21% had experienced a life-threatening episode upon diagnosis resulting in more severe morbidity. Additionally, 36% of the patients had a history of myasthenic crisis, indicating a potential risk of clinical deterioration in MG. Antibody titers are traditionally used to support MG evaluations but not directly correlation with clinical symptoms (Berrih-Aknin and Le Panse, 2014). Therefore, biomarkers to support MG diagnosis and disease severity screening must be identified. In the present study, the supervised ML model, XGBoost, was determined to have better performance with respect to analyzing gut microbiomes. This study's use of LOOCV somewhat mitigated the study's limitation of a small sample size and improved the reliability and generalizability of our findings. Our proposed ML model, which identifies several taxonomic markers, was able to effectively discriminate patients with MG from those without. Therefore, this approach has potential as a new form of ML analysis strategy for screening MG. In addition, we identified overlapping ASVs that were identified when the ML model was trained using full ASV–based and using ASV taxon–based data to select 31 HIASVs. When the model was trained using these HIASVs, the AUC was better than it was when each data set alone was used for training. Our results reveal that microbiota in the families of *Lachnospiraceae* and *Ruminococcaceae* were the most abundant in individuals with MG. We also identified microbiota potentially associated with symptoms of MG severity, that is, with limb weakness. The findings indicate that the proposed ML model based on microbiome data offers advantages and has high accuracy in identifying markers. Therefore, the model can be a potential benchmark diagnostic tool that can identify the presence of MG and gut microbiota associated with MG's severity through noninvasive analysis.

Changes in gut microbial composition were demonstrated to affect the immunology systems that regulate bodily function. Our study revealed the differences between the microbiomes of individuals with and without MG by determining the abundance of several microbiota. The microbiota of the family *Lachnospiraceae*, a member of the phylum Firmicutes and order Clostridiales, were determined to be significantly depleted (t test, $p < 0.05$). Our ML models based on

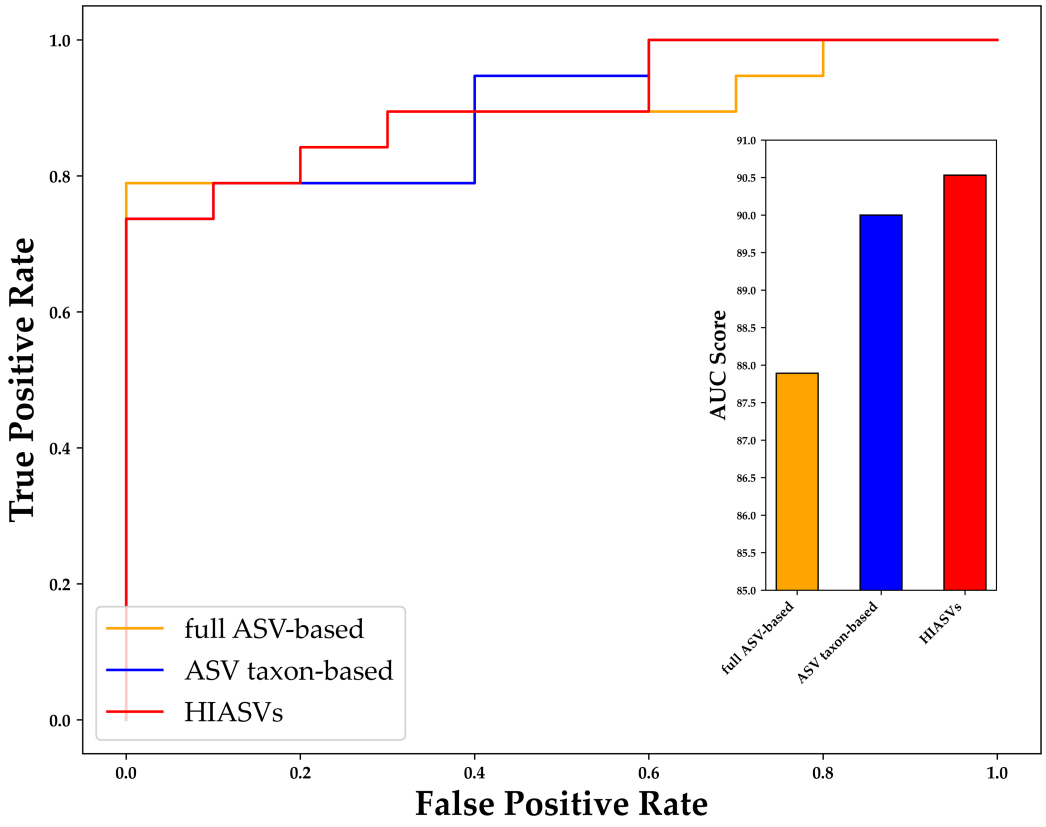

FIGURE 9
Receiver operating characteristic curve for comparing variants of XGBoost trained using different data sets. After 31 ASVs were identified as important by both XGBoost models (i.e., the model trained using the full ASV–based and that trained using ASV taxon–based data), these high-importance ASVs were used to train XGBoost, and were determined to be able to distinguish individuals with and without MG with an AUC of 90.53 (red bar), which was higher than the AUCs of the XGBoost models trained using only full ASV–based and only ASV taxon–based data. MG, myasthenia gravis; ASV, amplicon sequence variant; XGBoost, extreme gradient boosting; AUC, area under the curve.

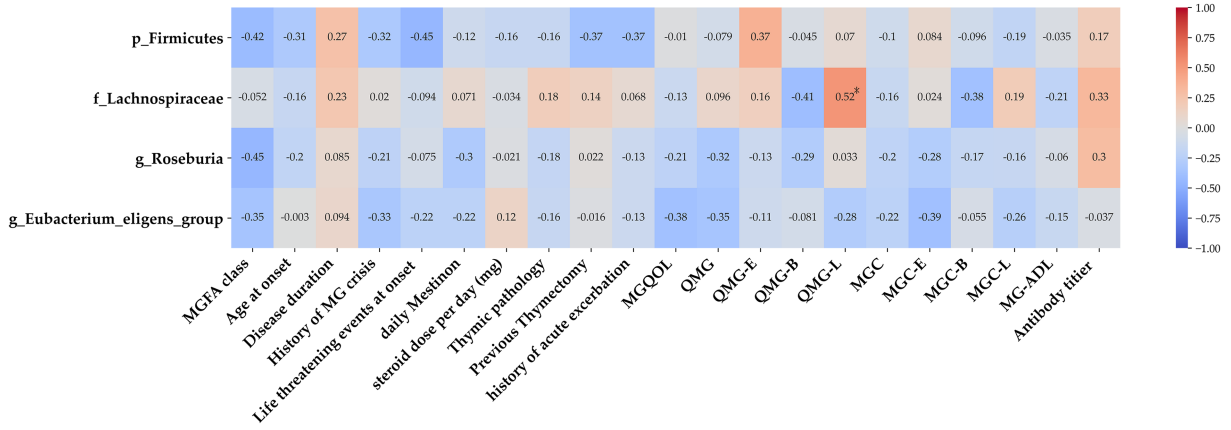

FIGURE 10
Association between gut microbiota and clinical indices of MG. Heat map of the Spearman's rank correlation coefficient of 4 significant taxa as well as 22 clinical indices. Red squares indicate positive associations between microbial species and clinical indices; blue squares indicate negative associations. Statistical significance is indicated within the squares (*$p < 0.05$). The family *Lachnospiraceae* was associated with several clinical parameters. MG, myasthenia gravis; IS, immunosuppressant; MGQOL, MG quality of life; MGC, MG composite; QMGS, quantitative myasthenia gravis score; MG-ADL, MG activities of daily living.

different ASVs verified this finding, and feature selection revealed that the family *Lachnospiraceae* was the most crucial with respect to MG. Genera from the family *Ruminococcaceae* and *Lachnospiraceae* were determined to be the most crucial for determining a diagnosis of MG when the model was trained using the HIASVs. *Lachnospiraceae* and *Ruminococcaceae* were discovered to be the two most abundant

families of Clostridiales and have been reported to be associated with the maintenance of gut health and the production of short chain fatty acids (SCFAs) (Gopalakrishnan et al., 2018; Vojinovic et al., 2019). The two families are highly abundant in gut microbiota and were reported to be depleted in the gut environments of individuals with different autoimmune diseases (Biddle et al., 2013).

Lachnospiraceae has been indicated to potentially influence healthy gut activity, and literature reviews have revealed that different members of this family are associated with different diseases. *Lachnospiraceae* was reported to be involved in autoimmune disorders, such as multiple sclerosis and inflammatory bowel diseases (Baumgart et al., 2007; Shahi et al., 2017). However, the mechanisms underlying *Lachnospiraceae's* regulation of immune responses and disease course remain unclear. A potential mechanism is the metabolism and production of SCFAs (Furusawa et al., 2013). This SCFA activity can modify the host immune system and function by lowering inflammatory marker levels and promoting regulatory T (Treg) cell accumulation (Atarashi et al., 2013). MG is an autoimmune condition because its pathogenesis involves disequilibrium between B cells and Treg cells, and patients with MG have a markedly lower abundance of Treg cells in their peripheral blood (Thiruppathi et al., 2012). The literature indicates that the abundance of *Ruminococcaceae* and *Lachnospiraceae* is negatively associated with these diseases (Biddle et al., 2013). A decrease in the abundance of *Lachnospiraceae* may lead to a reduction in Treg accumulation. New therapeutic strategies for treating MG should involve interventions focused on restoring *Lachnospiraceae* levels and thereby increasing Treg cell populations.

Many ML methods have been utilized in microbiota studies. ML can be used to perform numerous tasks, such as tracking phenotyping, classifying features, and identifying interactions and changes between microbiomes and other clinical variables (Gupta and Gupta, 2021; Marcos-Zambrano et al., 2021). Traditional ML models, including linear regression with Lasso and elastic nets, have been demonstrated to have higher performance in analyzing gut microbiome data and predicting dysbiosis (Pasolli et al., 2016; Lee and Rho, 2022). RF have also been used in microbiota studies. In RF models, trees are constructed to assist with decision-making and to group data into categories. In the current study, widely used ML models were used to select strategies for identifying the factors that influence MG risk (Lee and Rho, 2022). We applied XGBoost, an ensemble ML algorithm based on the decision tree method that can effectively match predicted outcomes (Chen and Guestrin, 2016). In XGBoost, many weak decision trees are integrated to form a model with strong predictive power. According to a study that compared common ML models, XGBoost, RF, and elastic nets have comparable performance when trained using microbiome data sets (Wang and Liu, 2020). In addition, XGBoost was reported to outperform a random model with respect to its cross-validation performance and to be able to forecast responses based on baseline microbiome data (Klimenko et al., 2022). Our finding that the optimal data set for training XGBoost involved both taxonomic and ASV feature data related to MG is comparable to the findings of many other studies that have investigated the characteristics that predict risk. Our results increase the depth of the understanding of the ML-XGBoost algorithm's potential for clinically supporting disease diagnosis on the basis of gut bacterial data. The proposed XGBoost-based model may be more useful as tool for identifying the features microbiomes features and have a better accuracy and AUC than RF and Lasso models. In the future, as the number of participants increases, we can persistently substantiate this hypothesis. XGBoost could be a potential useful method in ML-based microbiomes studies.

The ML model that was trained using different taxonomic features (i.e., the ASV taxon–based data) had the same performance as that trained using the full ASV–based data. We identified the overlapping key features selected by these models to improve the ML model's prediction power. Incorporating two sets of data to train an ML model using 31 HIASVs led to the model having the most accurate prediction. Most microbiome studies have used key operational taxonomic units to distinguish between study groups or used LDS-based taxonomic feature extraction to identify significantly different relative abundances between target groups. Our study combined genetic information (i.e., ASVs) and biological information (i.e., taxonomic features) to achieve more accurate prediction results. LOOCV was also applied and ensured that an unbiased estimate of the model's performance was obtained because every instance in the data set is used for both training and validation. LOOCV is also more computationally expensive and particularly useful when the size of a data set is small. It allows for the data to be used to the fullest, for both training and validation (Cheng et al., 2017). Our use of LOOCV enabled us to improve the accuracy of the model's performance and our ability to generalize our data. Furthermore, LOOCV can provide clear and interpretable results, which reduces study limitations.

Our findings are consistent with those of previous studies reporting a link between abnormalities in the gut microbiota and several autoimmune disorders (Qin et al., 2010; Chen et al., 2016; Zhou et al., 2018). Nevertheless, many autoimmune diseases do not have similar patterns of microbial dysbiosis, and therefore, the changes in the microbiota of patients with MG may not be generalizable to other autoimmune diseases. Studies have discovered that changes in gut microbiome composition can lead to inflammation that considerably affects immune responses in MG. A cohort study revealed that the gut microbiota of patients with MG was considerably altered, exhibiting a sharp decrease in the abundance of the bacterial taxa *Clostridium* correlated with a decrease in SCFA (Qiu et al., 2018). Zheng et al. demonstrated that individuals with MG often have significantly disturbed gut microbiomes and that this disturbance is associated with disease severity (Zheng et al., 2019). Another analysis revealed that MG is associated with a lower abundance of *Verrucomicrobiaceae* and *Bifidobacteriaceae* and an increased abundance of Bacteroidetes and *Desulfovibrionaceae* (Moris et al., 2018). Specially, Huang et al. found that AChR positive MG patients also experience changes in their oral microbiota (Huang et al., 2022). Our study identified bacterial genera for which the abundance differed in individuals with and without MG and applied two microbiomes-based ML models to identify key bacterial taxa. The findings may assist in improving the predictive outcomes of MG. In addition, LOOCV was used to improve the ML prediction performance. Most studies have used only OUTs or taxonomy data sets. A study reported that an ML model trained with OUTs to identify metabolite and microbiome markers was used to predict MG and that the model achieved an AUC of 0.76 (Moris et al., 2018). The model developed in our study achieved an AUC of 0.90 after being trained only with stool gut microbiome data. Stool gut microbiome data can be more easily and less expensively obtained than that of gut metabolites and metabolomes. Our findings demonstrate the potential of our proposed microbiome-based ML model as diagnostic support for identifying MG. The model can be further calibrated and the predictive capability can be improved by including more samples from different sources or

stratifying particular forms of MG and data from medical records in addition to gut microbiome data. Furthermore, the significant bacterial taxonomic features identified in our study may serve as novel biomarkers for clinical use and mechanistic study in the future.

ML has shown promise in predicting outcomes and identifying biomarkers for MG. A national study used an explainable ML-based model to accurately predict short-term outcomes in MG using various clinical parameters (Zhong et al., 2023). The SHapley Additive exPlanations (SHAP) method allowed for assessing the impact of each factor on the outcome, making the results more interpretable and quantification. Supervise ML, the multinomial model has also successfully identified diagnostic biomarkers for neurological disorders, including MG, using big biological data such as genotyping, blood, and urine biochemistry data (Lam et al., 2022). During the COVID-19 pandemic, ML algorithms were utilized for telemedicine in MG, analyzing eye or body motions and vocalization for standardized data acquisition and real-time feedback (Garbey et al., 2023). In contrast to the present work, the purpose of this study was aimed to investigate fecal specimens as a simple method for MG diagnostic screening despite the absence of patient blood or genetic data and the non-use of visual computing programs, these limitations did not impact the primary objectives of the research. Although interpretability ML was not utilized to assess the impact of various microorganisms on the outcomes, the study results still hold the potential to provide valuable information for MG diagnosis. Future studies may consider increasing the number of participants, incorporating blood and genetic data, and exploring the use of interpretable machine learning models to gain deeper insights into the influence of microbiota on MG.

Our study has some limitations. First, the numbers of recruited subjects were small and only from a single geographic region with lack of ancestry data, which limiting our ability to analyze potential confounding factors. Although we applied LOOCV to improve our model's prediction, additional large, multi-national and multi-center cohort studies should be conducted to validate our results. Second, the medication status of the recruited patients with MG differed, which could have affected the microbial compositions of their guts. Third, we did not analyze the metabolome of the stool sample. Gut microbiotas changes cannot provide the total necessary quantitative functional state of the microbiomes (Zierer et al., 2018). Forth, we did not record the dietary status of the participants. Based on the literature review, dietary is indeed a crucial factor influencing gut microbiota composition (Leeming et al., 2019; Zhang, 2022). Therefore, future research should incorporate participants' dietary records as a basis. Fifth, the proportion of males (32%) was relatively fewer in number. MG has been known to affect females more prominently (Jayam Trouth et al., 2012). The peaks was around at age 30 and 50 (Carr et al., 2010). Therefore, most of the research on MG and gut microbiota is based on female populations (Zheng et al., 2019; Tan et al., 2020). However, the limited number of male samples can be considered a limitation in the search for biomarkers. Finally, our study did not determine whether dysbiosis is the consequence, cause, or both of MG. Future longitudinal, multi-center, large cohort studies should be conducted, combing the recording of dietary and the ancestry data with a focus on the pathophysiology of bacterial taxa involved in MG. Additional research should be performed to identify the specific microbial species associated with MG and their corresponding metabolites to assist in defining targets for MG therapy.

5. Conclusion

Our study is the first to demonstrate the potential for using artificial intelligence through ML modeling to complete convenient diagnostic screening of MG on the basis of fecal microbiota composition. Our gut microbiome-based ML strategy can be used as a screening method to support the diagnosis and progression of MG. In addition, the combination ML-based feature selection approaches expand the knowledge on the biomarkers of MG. XGboost-based feature selection identified of HIASVs not only reduced the computational complexity of the ML model but also improved its diagnostic classification performance. These HIASVs may serve as novel biomarkers for clinical and mechanistic study in the future. Taken together, our findings provided a novel and user-friendly ML-based algorithm for explore critical microbiomes and diagnostic tools in MG. Future studies should prioritize conducting longitudinal, multi-center research to deepen the understanding of the mechanisms involved in the interactions of ASVs with hosts, which will aid in defining targets for MG therapy.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: NCBI – https://datadryad.org/stash/share/GewdUVu1bh5x0KNldA2E9qlN9ryGurFOCOdV-pKpLzk.

Ethics statement

The studies involving humans were approved by Research Ethic Committee of Fu-Jen Catholic University Hospital. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

C-CC and W-NL was involved in the study design, conducted the experiments, and writing the first draft of the paper. C-CC and H-CC were responsible for data collection. W-NL was responsible for proofreading and paper revision. C-CC, T-CL, and C-JL conducted the experiments, analyzed and interpreted the data. All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported by grants from the Fu Jen Catholic University Hospital (PL-202008012-V).

Acknowledgments

The authors are grateful to patients and families for the interest and generous participation in our research effort.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2023.1227300/full#supplementary-material

References

Atarashi, K., Tanoue, T., Oshima, K., Suda, W., Nagano, Y., Nishikawa, H., et al. (2013). Treg induction by a rationally selected mixture of Clostridia strains from the human microbiota. *Nature* 500, 232–236. doi: 10.1038/nature12331

Baumgart, M., Dogan, B., Rishniw, M., Weitzman, G., Bosworth, B., Yantiss, R., et al. (2007). Culture independent analysis of ileal mucosa reveals a selective increase in invasive *Escherichia coli* of novel phylogeny relative to depletion of Clostridiales in Crohn's disease involving the ileum. *ISME J.* 1, 403–418. doi: 10.1038/ismej.2007.52

Berrih-Aknin, S., and Le Panse, R. (2014). Myasthenia gravis: a comprehensive review of immune dysregulation and etiological mechanisms. *J. Autoimmun.* 52, 90–100. doi: 10.1016/j.jaut.2013.12.011

Biddle, A., Stewart, L., Blanchard, J., and Leschine, S. (2013). Untangling the genetic basis of Fibrolytic specialization by Lachnospiraceae and Ruminococcaceae in diverse gut communities. *Diversity* 5, 627–640. doi: 10.3390/d5030627

Bokulich, N. A., Kaehler, B. D., Rideout, J. R., Dillon, M., Bolyen, E., Knight, R., et al. (2018). Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome* 6:90. doi: 10.1186/s40168-018-0470-z

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., et al. (2013). API design for machine learning software: experiences from the scikit-learn project arXiV preprint arXiv:1309.0238.

Callahan, B. J., Mcmurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J., and Holmes, S. P. (2016). DADA2: high-resolution sample inference from Illumina amplicon data. *Nat. Methods* 13, 581–583. doi: 10.1038/nmeth.3869

Carr, A. S., Cardwell, C. R., Mccarron, P. O., and Mcconville, J. (2010). A systematic review of population based epidemiological studies in myasthenia gravis. *BMC Neurol.* 10:46. doi: 10.1186/1471-2377-10-46

Chen, T., and Guestrin, C. (2016). XGBoost. 785-794.

Chen, P., and Tang, X. (2021). Gut microbiota as regulators of Th17/Treg balance in patients with myasthenia gravis. *Front. Immunol.* 12:803101. doi: 10.3389/fimmu.2021.803101

Chen, J., Wright, K., Davis, J. M., Jeraldo, P., Marietta, E. V., Murray, J., et al. (2016). An expansion of rare lineage intestinal microbes characterizes rheumatoid arthritis. *Genome Med.* 8:43. doi: 10.1186/s13073-016-0299-7

Cheng, H., Garrick, D. J., and Fernando, R. L. (2017). Efficient strategies for leave-one-out cross validation for genomic best linear unbiased prediction. *J. Anim. Sci. Biotechnol.* 8:38. doi: 10.1186/s40104-017-0164-6

Farrugia, M. E., and Goodfellow, J. A. (2020). A practical approach to managing patients with myasthenia gravis-opinions and a review of the literature. *Front. Neurol.* 11:604. doi: 10.3389/fneur.2020.00604

Furusawa, Y., Obata, Y., Fukuda, S., Endo, T. A., Nakato, G., Takahashi, D., et al. (2013). Commensal microbe-derived butyrate induces the differentiation of colonic regulatory T cells. *Nature* 504, 446–450. doi: 10.1038/nature12721

Garbey, M., Joerger, G., Lesport, Q., Girma, H., Mcnett, S., Abu-Rub, M., et al. (2023). A digital telehealth system to compute the myasthenia gravis Core examination metrics. *JMIR Neurotechnol.* 2:e43387. doi: 10.2196/43387

Gilhus, N. E. (2016). Myasthenia gravis. *N. Engl. J. Med.* 375, 2570–2581. doi: 10.1056/NEJMra1602678

Gopalakrishnan, V., Spencer, C. N., Nezi, L., Reuben, A., Andrews, M. C., Karpinets, T. V., et al. (2018). Gut microbiome modulates response to anti-PD-1 immunotherapy in melanoma patients. *Science* 359, 97–103. doi: 10.1126/science.aan4236

Gupta, M. M., and Gupta, A. (2021). Survey of artificial intelligence approaches in the study of anthropogenic impacts on symbiotic organisms – a holistic view. *Symbiosis* 84, 271–283. doi: 10.1007/s13199-021-00778-0

Hehir, M. K., and Silvestri, N. J. (2018). Generalized myasthenia gravis: classification, clinical presentation, natural history, and epidemiology. *Neurol. Clin.* 36, 253–260. doi: 10.1016/j.ncl.2018.01.002

Huang, C., Gao, F., Zhou, H., Zhang, L., Shang, D., Ji, Y., et al. (2022). Oral microbiota profile in a Group of Anti-AChR antibody-positive myasthenia gravis patients. *Front. Neurol.* 13:938360. doi: 10.3389/fneur.2022.938360

Jaretzki, A., Barohn, R. J., Ernstoff, R. M., Kaminski, H. J., Keesey, J. C., Penn, A. S., et al. (2000). Myasthenia gravis: recommendations for clinical research standards. Task force of the medical scientific advisory Board of the Myasthenia Gravis Foundation of America. *Ann. Thorac. Surg.* 70, 327–334. doi: 10.1016/S0003-4975(00)01595-2

Jayam Trouth, A., Dabi, A., Solieman, N., Kurukumbi, M., and Kalyanam, J. (2012). Myasthenia gravis: a review. *Autoimmune Dis.* 2012:874680. doi: 10.1155/2012/874680

Kang, Y., Li, L., Kang, X., Zhao, Y., and Cai, Y. (2022). Gut microbiota and metabolites in myasthenia gravis: early diagnostic biomarkers and therapeutic strategies. *Clin. Immunol.* 245:109173. doi: 10.1016/j.clim.2022.109173

Kapoor, B., Gulati, M., Gupta, R., and Singla, R. K. (2023). Microbiota dysbiosis and myasthenia gravis: do all roads lead to Rome? *Autoimmun. Rev.* 22:103313. doi: 10.1016/j.autrev.2023.103313

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Klimenko, N. S., Odintsova, V. E., Revel-Muroz, A., and Tyakht, A. V. (2022). The hallmarks of dietary intervention-resilient gut microbiome. *NPJ Biofilms Microb.* 8:77. doi: 10.1038/s41522-022-00342-8

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B. E., Bussonnier, M., Frederic, J., et al. (2016). "Jupyter notebooks - a publishing format for reproducible computational workflows", in: International Conference on Electronic Publishing.

Lam, S., Arif, M., Song, X., Uhlen, M., and Mardinoglu, A. (2022). Machine learning analysis reveals biomarkers for the detection of neurological diseases. *Front. Mol. Neurosci.* 15:889728. doi: 10.3389/fnmol.2022.889728

Lee, S. J., and Rho, M. (2022). Multimodal deep learning applied to classify healthy and disease states of human microbiome. *Sci. Rep.* 12:824. doi: 10.1038/s41598-022-04773-3

Leeming, E. R., Johnson, A. J., Spector, T. D., and Le Roy, C. I. (2019). Effect of diet on the gut microbiota: rethinking intervention duration. *Nutrients* 11. doi: 10.3390/nu11122862

Lin, C. H., Chen, C. C., Chiang, H. L., Liou, J. M., Chang, C. M., Lu, T. P., et al. (2019). Altered gut microbiota and inflammatory cytokine responses in patients with Parkinson's disease. *J. Neuroinflammation* 16:129. doi: 10.1186/s12974-019-1528-y

Liu, Y., Chen, P. C., Krause, J., and Peng, L. (2019). How to read articles that use machine learning: Users' guides to the medical literature. *JAMA* 322, 1806–1816. doi: 10.1001/jama.2019.16489

Marcos-Zambrano, L. J., Karaduzovic-Hadziabdic, K., Loncar Turukalo, T., Przymus, P., Trajkovik, V., Aasmets, O., et al. (2021). Applications of machine learning in human microbiome studies: a review on feature selection, biomarker identification disease prediction and treatment. *Front. Microbiol.* 12:634511. doi: 10.3389/fmicb.2021.634511

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* 17:3. doi: 10.14806/ej.17.1.200

Moris, G., Arboleya, S., Mancabelli, L., Milani, C., Ventura, M., De Los Reyes-Gavilan, C. G., et al. (2018). Fecal microbiota profile in a group of myasthenia gravis patients. *Sci. Rep.* 8:14384. doi: 10.1038/s41598-018-32700-y

Narayanaswami, P., Sanders, D. B., Wolfe, G., Benatar, M., Cea, G., Evoli, A., et al. (2021). International consensus guidance for Management of Myasthenia Gravis. *Neurology* 2020, 114–122. doi: 10.1212/WNL.0000000000011124

Pasolli, E., Truong, D. T., Malik, F., Waldron, L., and Segata, N. (2016). Machine learning Meta-analysis of large metagenomic datasets: tools and biological insights. *PLoS Comput. Biol.* 12:e1004977. doi: 10.1371/journal.pcbi.1004977

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830. doi: 10.48550/arXiv.1201.0490

Pianta, A., Arvikar, S. L., Strle, K., Drouin, E. E., Wang, Q., Costello, C. E., et al. (2017). Two rheumatoid arthritis-specific autoantigens correlate microbial immunity with autoimmune responses in joints. *J. Clin. Invest.* 127, 2946–2956. doi: 10.1172/JCI93450

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* 5:e9490. doi: 10.1371/journal.pone.0009490

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464, 59–65. doi: 10.1038/nature08821

Qiu, D., Xia, Z., Jiao, X., Deng, J., Zhang, L., and Li, J. (2018). Altered gut microbiota in myasthenia gravis. *Front. Microbiol.* 9:2627. doi: 10.3389/fmicb.2018.02627

Rousseff, R. T. (2021). Diagnosis of myasthenia gravis. *J. Clin. Med.* 10. doi: 10.3390/jcm10081736

Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W. S., et al. (2011). Metagenomic biomarker discovery and explanation. *Genome Biol.* 12:R60. doi: 10.1186/gb-2011-12-6-r60

Shahi, S. K., Freedman, S. N., and Mangalam, A. K. (2017). Gut microbiome in multiple sclerosis: the players involved and the roles they play. *Gut Microbes* 8, 607–615. doi: 10.1080/19490976.2017.1349041

Tan, X., Huang, Y., Chai, T., Zhao, X., Li, Y., Wu, J., et al. (2020). Differential gut microbiota and fecal metabolites related with the clinical subtypes of myasthenia gravis. *Front. Microbiol.* 11:564579. doi: 10.3389/fmicb.2020.564579

Thiruppathi, M., Rowin, J., Li Jiang, Q., Sheng, J. R., Prabhakar, B. S., and Meriggioli, M. N. (2012). Functional defect in regulatory T cells in myasthenia gravis. *Ann. N. Y. Acad. Sci.* 1274, 68–76. doi: 10.1111/j.1749-6632.2012.06840.x

Thye, A. Y., Law, J. W., Tan, L. T., Thurairajasingam, S., Chan, K. G., Letchumanan, V., et al. (2022). Exploring the gut microbiome in myasthenia gravis. *Nutrients* 14. doi: 10.3390/nu14081647

Vabalas, A., Gowen, E., Poliakoff, E., and Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PLoS One* 14:e0224365. doi: 10.1371/journal.pone.0224365

Van Rossum, G., and Drake, F. L. (1995). Python reference manual. Centrum voor Wiskunde en Informatica Amsterdam.

Vojinovic, D., Radjabzadeh, D., Kurilshikov, A., Amin, N., Wijmenga, C., Franke, L., et al. (2019). Relationship between gut microbiota and circulating metabolites in population-based cohorts. *Nat. Commun.* 10:5813. doi: 10.1038/s41467-019-13721-1

Wang, X. W., and Liu, Y. Y. (2020). Comparative study of classifiers for human microbiome data. *Med. Microecol.* 4:100013. doi: 10.1016/j.medmic.2020.100013

Zhang, P. (2022). Influence of foods and nutrition on the gut microbiome and implications for intestinal health. *Int. J. Mol. Sci.* 23. doi: 10.3390/ijms23179588

Zhang, H., Li, Y., Zheng, P., Wu, J., Huang, Y., Tan, X., et al. (2022). Altered metabolism of the microbiota-gut-brain Axis is linked with comorbid anxiety in fecal recipient mice of myasthenia gravis. *Front. Microbiol.* 13:804537. doi: 10.3389/fmicb.2022.804537

Zheng, P., Li, Y., Wu, J., Zhang, H., Huang, Y., Tan, X., et al. (2019). Perturbed microbial ecology in myasthenia gravis: evidence from the gut microbiome and fecal metabolome. *Adv. Sci.* 6:1901441. doi: 10.1002/advs.201901441

Zhong, H., Ruan, Z., Yan, C., Lv, Z., Zheng, X., Goh, L. Y., et al. (2023). Short-term outcome prediction for myasthenia gravis: an explainable machine learning model. *Ther. Adv. Neurol. Disord.* 16:311549. doi: 10.1177/17562864231154976

Zhou, Y., Xu, Z. Z., He, Y., Yang, Y., Liu, L., Lin, Q., et al. (2018). *Gut microbiota offers universal biomarkers across ethnicity in inflammatory bowel disease diagnosis and infliximab response prediction*, vol. *3* doi: 10.1128/mSystems.00188-17

Zierer, J., Jackson, M. A., Kastenmüller, G., Mangino, M., Long, T., Telenti, A., et al. (2018). The fecal metabolome as a functional readout of the gut microbiome. *Nat. Genet.* 50, 790–795. doi: 10.1038/s41588-018-0135-7

frontiers | Frontiers in Microbiology

TYPE Mini Review
PUBLISHED 05 October 2023
DOI 10.3389/fmicb.2023.1250909

OPEN ACCESS

EDITED BY
Babak Momeni,
Boston College, United States

REVIEWED BY
Sam Ma,
Chinese Academy of Sciences (CAS), China
FengLong Yang,
Fujian Medical University, China

*CORRESPONDENCE
Eliana Ibrahimi
✉ eliana.ibrahimi@fshn.edu.al
Laura Judith Marcos-Zambrano
✉ judith.marcos@imdea.org

RECEIVED 30 June 2023
ACCEPTED 22 September 2023
PUBLISHED 05 October 2023

CITATION
Ibrahimi E, Lopes MB, Dhamo X, Simeon A, Shigdel R, Hron K, Stres B, D'Elia D, Berland M and Marcos-Zambrano LJ (2023) Overview of data preprocessing for machine learning applications in human microbiome research.
Front. Microbiol. 14:1250909.
doi: 10.3389/fmicb.2023.1250909

Overview of data preprocessing for machine learning applications in human microbiome research

Eliana Ibrahimi[1]*, Marta B. Lopes[2,3], Xhilda Dhamo[4], Andrea Simeon[5], Rajesh Shigdel[6], Karel Hron[7], Blaž Stres[8,9,10,11], Domenica D'Elia[12], Magali Berland[13] and Laura Judith Marcos-Zambrano[14]*

[1]Department of Biology, Faculty of Natural Sciences, University of Tirana, Tirana, Albania, [2]Department of Mathematics, Center for Mathematics and Applications (NOVA Math), NOVA School of Science and Technology, Caparica, Portugal, [3]UNIDEMI, Department of Mechanical and Industrial Engineering, NOVA School of Science and Technology, Caparica, Portugal, [4]Department of Applied Mathematics, Faculty of Natural Sciences, University of Tirana, Tirana, Albania, [5]BioSense Institute, University of Novi Sad, Novi Sad, Serbia, [6]Department of Clinical Science, University of Bergen, Bergen, Norway, [7]Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacký University Olomouc, Olomouc, Czechia, [8]Department of Catalysis and Chemical Reaction Engineering, National Institute of Chemistry, Ljubljana, Slovenia, [9]Faculty of Civil and Geodetic Engineering, Institute of Sanitary Engineering, Ljubljana, Slovenia, [10]Department of Automation, Biocybernetics and Robotics, Jožef Stefan Institute, Ljubljana, Slovenia, [11]Department of Animal Science, Biotechnical Faculty, University of Ljubljana, Ljubljana, Slovenia, [12]Department of Biomedical Sciences, National Research Council, Institute for Biomedical Technologies, Bari, Italy, [13]INRAE, MetaGenoPolis, Université Paris-Saclay, Jouy-en-Josas, France, [14]Computational Biology Group, Precision Nutrition and Cancer Research Program, IMDEA Food Institute, Madrid, Spain

Although metagenomic sequencing is now the preferred technique to study microbiome-host interactions, analyzing and interpreting microbiome sequencing data presents challenges primarily attributed to the statistical specificities of the data (e.g., sparse, over-dispersed, compositional, inter-variable dependency). This mini review explores preprocessing and transformation methods applied in recent human microbiome studies to address microbiome data analysis challenges. Our results indicate a limited adoption of transformation methods targeting the statistical characteristics of microbiome sequencing data. Instead, there is a prevalent usage of relative and normalization-based transformations that do not specifically account for the specific attributes of microbiome data. The information on preprocessing and transformations applied to the data before analysis was incomplete or missing in many publications, leading to reproducibility concerns, comparability issues, and questionable results. We hope this mini review will provide researchers and newcomers to the field of human microbiome research with an up-to-date point of reference for various data transformation tools and assist them in choosing the most suitable transformation method based on their research questions, objectives, and data characteristics.

KEYWORDS

human microbiome, data preprocessing, machine learning, compositionality, normalization, metagenomics data

1. Introduction

In recent decades, next-generation sequencing technologies have significantly impacted human microbiome research, allowing for a better understanding and characterization of microbiome-host interactions (Hadrich, 2020). Numerous 16S rRNA sequencing datasets are extended further by metagenomic sequencing of the whole microbial genome. The staggering increase in publications and datasets with an ever-increasing number of samples increased the need for more performant analysis approaches, such as advanced statistical methods and machine learning (ML) algorithms that can handle large-scale microbiome datasets and extract meaningful patterns, relationships, and associations. Before entering ML analysis microbiome raw data is preprocessed through several steps shown in Supplementary Figure S1.

ML models can be trained to predict the composition of microbial communities based on various input factors such as host genetics, diet, and environmental factors, which can help us understand the factors influencing microbial composition and its relation to human health (Gupta and Gupta, 2021; Hernández Medina et al., 2022). Despite the advantages, ML analysis of microbiome data is challenging due to inherent microbiome data characteristics (i.e., sparsity, compositionality, high dimensionality, dispersion), and new techniques are requested to address these challenges (Moreno-Indias et al., 2021; D'Elia et al., 2023).

Microbiome data is zero-inflated, which can be due to the sequencing depth (i.e., sampling zeros) or the real absence of taxa (i.e., true zeros) (Silverman et al., 2020). Furthermore, variations in the abundance of one taxon affect all other taxa due to the constraint that the total counts equal the library size. Hence, the raw counts observed do not directly indicate the absolute abundances of individual taxa (Weiss et al., 2017; Lloréns-Rico et al., 2021; Swift et al., 2023), giving rise to compositional data. As a result, transforming microbiome sequencing data is essential in preparing the data for analysis and applying ML algorithms.

This mini review aims to provide a comprehensive overview of the preprocessing methods used in recent human microbiome studies to transform microbiome sequencing data before ML analysis. To collect information, we conducted a scoping review based on the methodology outlined by Arksey and O'Malley (2005), combined with manual and automated literature searches following the approach outlined by Marcos-Zambrano et al. (2021). Papers included in the final review were published in peer-reviewed journals from January 2011 to January 2022 and specifically analyzed human microbiome 16S rRNA and shotgun metagenomic data through ML algorithms. As of December 2022, 3 reviewers had extracted findings on data preprocessing and transformation techniques from 95 published studies (Supplementary Table S1). In the subsequent sections, we present and discuss the findings and outcomes of our investigation.

2. Sequence preprocessing

Microbiome analysis starts with raw DNA sequencing reads or microbial taxa tables at different taxonomic resolutions, from Domain (i.e., Bacteria, Archaea, Eucarya) to strain and genome variants. Microbial taxa tables are created by processing raw sequences, known as *sequence preprocessing*. Both 16S rRNA sequencing and shotgun metagenomic sequencing generally involve preprocessing steps such as quality checking, trimming, filtering, removing, and merging (Travisany et al., 2015; Ryan et al., 2020). The key differences lie in the amplification of specific gene regions for 16S rRNA sequencing and the sequencing of entire genomes for shotgun metagenomics. The sequence preprocessing steps generally depend on the origin of the DNA sequences, sequence orientation, and sequencer type.

Quality scores are used for the recognition and removal of low-quality regions of sequence (trimming) or low-quality reads (filtration) and the determination of accurate consensus sequences (merging) (Bokulich et al., 2013). A widely adopted quality metric is the Phred quality score (Q) (Galkin et al., 2020). Then, leading, and trailing trimming are applied at the position of the read where the average score drastically changes and falls below the given threshold (Bolger et al., 2014). Typical sequence preprocessing techniques are: (1) reads filtering, if overall quality is very low (Amir et al., 2017); (2) minimal length filtering, for reads below a specified length; (3) barcode and adapter-trimming (Martin, 2011); (4) chimera filtering (Edgar et al., 2011); (5) phiX reads, commonly present in marker gene of Illumina sequence data (Callahan et al., 2016). A frequently used tool for shotgun aligning and taxonomic profiling is MetaPhlAn (Thomas et al., 2019; Blanco-Míguez et al., 2023). Shotgun metagenomics preprocessing generally requires a complex sequence of programs merged into pipelines to be used since there is no one-in-all software solution yet. The solution is usually found in automated pre-defined bioBakery Workflows (Beghini et al., 2021) or Bbtools, namely, BBMerge and BBDuk (Bushnell et al., 2017; Galkin et al., 2020).

Before entering the feature selection step, additional filtering is performed on the raw data to reduce noise while keeping the most relevant taxa. In this step, microbiome low abundance features (e.g., <500 reads) and/or prevalence (e.g., <10%) per sample group or in the entire sample, are filtered out. Based on the resulting count matrix, the taxonomic level under consideration (i.e., family, genus, species) can be chosen at this stage, considering that going down to the species level would lead to strong zero inflation.

Feature selection is approached by many studies through predictive feature selection strategies that encompass statistical methods for assessing the significance of the associations between the microbiome features and the disease condition. These methods include univariate and multivariate statistical methods, and different ML algorithms (Chen et al., 2021; Jiang et al., 2022). Network-based methods have also been employed for selecting hub strains from co-occurrence networks before entering the ML task (Xu et al., 2021). It is crucial to keep in mind that when using these predictive feature selection methods, if the training dataset is not kept distinct from the test dataset throughout all preprocessing, modeling, and assessment phases, the model gains access to test set information prior to performance evaluation, resulting in data leakage (Kapoor and Narayanan, 2022). The most common ML solution for this problem is applying a cross-validation procedure, where the initial dataset is split into several folds, and in each split, different folds are proclaimed as learning or testing folds.

3. Transformation techniques

Typically, the ML analysis of microbiome data is performed after transformations are applied to raw reads to address statistical

challenges mainly associated with sparsity and the proportional nature of the generated sequencing data (Lloréns-Rico et al., 2021). Based on our review, the most common data transformation methods applied in recent human microbiome studies, in both 16 s RNA sequences and shotgun data, are the relative and normalization-based methods followed by compositional transformations such as Centered log-ratio (CLR), and Isometric log-ratio (ILR). Many reviewed publications (i.e., 28%) lack sufficient details about the data preprocessing techniques that have been applied or fail to mention if any preprocessing has been carried out leading to reproducibility issues and questionable results. In Figure 1, we present a TreeMap chart illustrating the frequencies of transformation methods applied across the analyzed papers.

Within the reviewed studies, a subset dedicated to problems of disease diagnosis and risk prediction (Fabijanić and Vlahoviček, 2016; Wu et al., 2020; Ruuskanen et al., 2021; Liu et al., 2022). Data analyzed in these studies, 16S rRNA sequencing data and shotgun data, are transformed through relative abundance, log transformations, z-score normalization, and CLR. In the following subsections, we briefly discuss the normalization-based and compositional methods applied to microbiome data before ML analysis across the reviewed papers.

3.1. Normalization methods

Two predominant transformation methods applied to deal with uneven library sizes in sequencing microbiome data are relative abundance (Statnikov et al., 2013; Ning and Beiko, 2015; Wu et al., 2018, 2021; Bogart et al., 2019; Gupta et al., 2019; Lo and Marculescu, 2019; Vangay et al., 2019; Yachida et al., 2019; Fernández-Edreira et al., 2021; Lloréns-Rico et al., 2021), and rarefaction (Stämmler et al., 2016; Weiss et al., 2017; Baksi et al., 2018), used to solve the problem of different sequencing depths (Murovec et al., 2021).

Other normalization-based methods applied frequently to microbiome data in the reviewed studies are: Log transformation, preferred when the data is heavily skewed (Lahti et al., 2013; Fabijanić and Vlahoviček, 2016; Eck et al., 2017; Tap et al., 2017; Flemer et al., 2018; Wirbel et al., 2019; Hughes et al., 2020; Ryan et al., 2020; Fouladi et al., 2021; Jiang et al., 2021; Zhu et al., 2022). Total Sum Scaling (TSS) (Lê Cao et al., 2016; Lloréns-Rico et al., 2021) which divides each taxa count by the total number of counts in each individual sample; Minimum-Maximum normalization, used to retain the relationships between the original input data (Mulenga et al., 2021; Jiang et al., 2022); Z-score normalization (Wirbel et al., 2019; Jiang et al., 2021; Mulenga et al., 2021) which transforms the data with mean zero and unit variance; the Square Root that can be successfully applied to count data that follow a Poisson distribution (Liu et al., 2011; Holmes et al., 2012); Inverse-Rank normalization used to normalize signals to approximate a normal distribution after removing the quality control sample (Ni et al., 2021).

3.2. Compositional transformations

Our review reveals a noticeable rise in the utilization of ML techniques within human microbiome research over recent years, while the adoption of compositional transformations in handling microbiome data remains relatively constrained. Nevertheless, an encouraging increasing trend in the application of compositional approaches between 2016 and 2021 is observed, as visually represented in Supplementary Figure S2. The following paragraphs delve into compositional transformations that have been employed in recent human microbiome studies, while in Table 1 we provide an overview

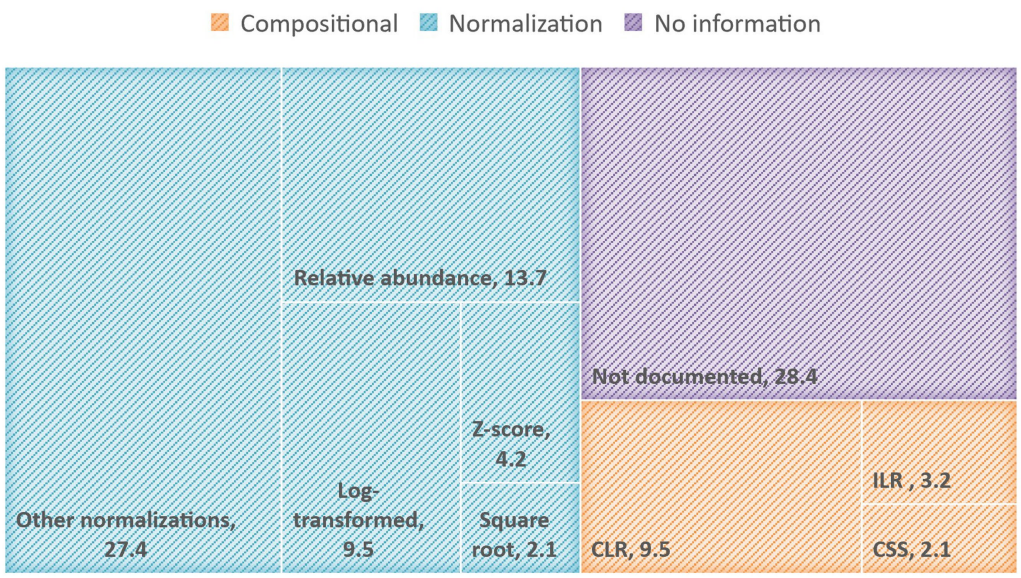

FIGURE 1
TreeMap chart illustrating the percentage of reviewed papers that applied normalization-based or compositional transformation methods, as well as the papers without clear information on preprocessing or data transformation. The other-normalization category comprises inverse-rank normalization, Box-Cox transformation, rarefaction, minimum-maximum transformation, scaling by standard deviation, normalization by total read depth, etc.

TABLE 1 Compositional transformations that are applied to human microbiome 16S rRNA and shotgun data.

Method	Bioconductor/R package	Literature
Additive log-ratio	Compositions	Aitchison (1982, 1986) and van den Boogaart and Tolosana-Delgado (2008)
Centered log-ratio	Compositions	Pawlowsky-Glahn et al. (2015) and van den Boogaart and Tolosana-Delgado (2008)
Isometric log-ratio	Compositions	Egozcue et al. (2003) and van den Boogaart and Tolosana-Delgado (2008)
Geometric mean of pairwise ratios	GMPR	Chen et al. (2018)
Trimmed mean of M-values	edgeR	Robinson et al. (2010)
Relative log expression (RLE)	edgeR	Robinson et al. (2010)
Variance-stabilizing (VST)	DESeq2	Love et al. (2014)

of the relevant literature and software tools necessary for the successful implementation of these methods.

Compositional data can be represented in a simplex space and analyzing them as absolute data with standard statistical techniques may lead to inappropriate results (Gloor et al., 2016; Quinn et al., 2018). Aitchison (1982) first proposed the additive log-ratio transformation (ALR), to address compositionality then also the centered log-ratio (CLR) (Aitchison, 1986). His followers proposed further the isometric log-ratio (ILR) (Egozcue et al., 2003; Pawlowsky-Glahn et al., 2015) and pivot log-ratio (PLR) (Filzmoser et al., 2018) transformations. The CLR transformation is applied more frequently in microbiome studies (Fabijanić and Vlahoviček, 2016; Lê Cao et al., 2016; Wirbel et al., 2019; Fukui et al., 2020; Reiman et al., 2021; Ruuskanen et al., 2021; Liu et al., 2022) than the ILR transformation (Kubinski et al., 2022), while the ALR was not applied in any of the studies included in the review.

Other compositional transformations that can be applied in microbiome data are: Cumulative Sum Scaling (CSS) (Dhungel et al., 2021; Lloréns-Rico et al., 2021), a particular representation of the relative information based on median-like quantiles; the Geometric mean of pairwise ratios (GMPR) transformation (Chen et al., 2018); the Trimmed mean of M-values (TMM) (Robinson et al., 2010); the Relative log expression (RLE) method (Robinson et al., 2010); the Variance-stabilizing transformation (VST) (Love et al., 2014).

4. Discussion

Transformations are essential for appropriately handling microbiome sequencing data, rectifying compositional issues, reducing noise, adhering to statistical assumptions, and enabling meaningful analysis and interpretation. The choice of transformation should depend on the specific characteristics of the data and the goals of the analysis. This mini review revealed substantial gaps in the process of microbiome data transformation. Relative transformations and other normalization-based methods that lead to or do not solve compositional issues (Lloréns-Rico et al., 2021) are frequently applied in recent human microbiome research.

Unlike compositional approaches (i.e., log ratios), normalization-based methods do not retrieve absolute scale from the relative data (Quinn et al., 2018). Nevertheless, when the raw data contains zero values, like in microbiome data, taking the logarithm results in negative infinity, distorting the data, and leading to invalid statistical inferences. To mitigate this issue, a pseudocount (i.e., small positive constant, ε) can be added to zero values before taking the logarithm. Selecting the right pseudocount in relation to the data's scale holds significant importance when applying log transformations (Thorsen et al., 2016). The scale of the ε, relative to the total read counts, should remain consistent across different data transformation methods applied (McKnight et al., 2019) and should be based on the context of the research problem and the scale of the data because the choice of ε can affect the results (Costea et al., 2014). Thus, it is essential to be mindful of the trade-offs between numerical stability and introducing additional bias due to the choice of ε.

Compositional transformations, ALR, CLR, and ILR log-ratio transformations, have different properties. The ALR transformation does not preserve distances because it is not isometric (Egozcue and Pawlowsky-Glahn, 2005), while CLR transformation keeps the distance, but the covariance and correlation matrix are singular because of the zero-sum of the transformed vectors (Quinn et al., 2018). In addition, aggregation of all components into the geometric mean can, in general, lead to the occurrence of false positives (Filzmoser and Walczak, 2014), so identifying the original components with the corresponding CLR variables has some limitations, which could possibly be overcome by a proper weighting strategy (Štefelová et al., 2021). Recent studies suggest that for high-dimensional compositional data, the ALR transformation should be a preferred choice for transforming variables because the interpretation of ALRs is easier than the ILR and CLR transformations (Greenacre et al., 2021). Besides log ratios, other transformations such as VST and ranked-based methods have been reported to successfully address microbiome data statistical specificities (Jeganathan and Holmes, 2021; Lloréns-Rico et al., 2021). When working with spatial human microbiome data, which can reflect the microbial composition and abundance within specific locations in the body (Adade et al., 2021), transformations for compositional spatial data that would improve ML techniques' performance when dealing with this data can be considered. Greenacre (2010, 2011) explored a power transformation that converges toward the Aitchison log-ratio transformation when the power parameter becomes 0, while Clarotto et al. (2022) propose the Isometric α-transformation (α-IT), which, unlike the ILR transformation, can successfully deal with zeros in the data.

Kubinski et al. (2022) investigated the impact of various transformation techniques on the model's predictive performance using gut microbiome data and highlighted the need to transform 16S

rRNA data using compositional transformation techniques. Among the available options, the CLR transformation was identified as the most suitable, as it enables the assessment of each feature's importance in the decision-making process of ML models. Another study by McKnight et al. (2019) examined the impact of log transformations commonly employed in normalization procedures. The authors demonstrated that log transformations could distort community comparisons by suppressing significant differences in common taxa while amplifying subtle differences in rare taxa.

Thus, despite the advantages, log-ratio approaches have their limitations and drawbacks and are not the only way to deal with compositionality. Quantitative transformations such as Quantitative Microbiota Profiling (QMP) (Vandeputte et al., 2017) and Absolute Counts Scaling (ACS) (Props et al., 2017; Jian et al., 2020) offer experimental approaches to address microbiome data proportional nature. QMP involves rarefying samples to achieve an even sampling depth and scaling them based on estimated microbial loads. On the other hand, ACS directly scales the relative sequencing counts using estimated microbial loads. Lloréns-Rico et al. (2021) investigated the impact of computational and experimental techniques in addressing the issues arising from microbiome data features (i.e., compositionality and sparsity). They concluded that quantitative approaches outperform computational methods in addressing compositionality and sparsity. Authors claim that the quantitative approaches improve the identification of true positive associations while reducing the occurrence of false positives. The same study reports that when adopting quantitative methods is not feasible, computational methods that address compositionality perform better than relative methods. There are other examples in the literature where compositional methods are employed to transform microbiome data where the reader can find more details (Quinn and Erb, 2020; Yang and Zou, 2020; Greenacre et al., 2021; Yang et al., 2021; Papoutsoglou et al., 2023).

It is important to mention that in many cases the analysis of microbiome data can be performed on raw read counts rather than in transformed data. Zero-inflated negative binomial and Dirichlet-multinomial models can fit microbiome raw data quite well (Xia et al., 2018). For example, Zhang et al. (2017) applied on raw read counts a negative binomial mixed model that enables the identification of connections between the host, environmental variables, and the microbiome.

Finally, the lack of adequate information on data preprocessing and high reporting heterogeneity among papers highlight the need for standardized reporting guidelines, as also suggested by Mirzayi et al. (2021), where recommendations and guidelines are provided to help microbiome researchers properly report their findings through the 'Strengthening The Organization and Reporting of Microbiome Studies' (STORMS), composed of a 17-item checklist each related with the typical sections of a scientific paper. The omission of preprocessing and transformations applied to the data can have several significant consequences such as reproducibility concerns, misinterpretation, comparability issues, and questionable results. To mitigate these consequences, it is essential for researchers to provide thorough documentation of their data preprocessing procedures in publications. Researchers should also consider sharing their code, scripts, or workflows used for data preprocessing, which can greatly enhance transparency and reproducibility.

5. Conclusions and final remarks

Our short review shows that the utilization of data transformations that address the proportional nature of microbiome sequencing data in human microbiome studies remains limited, with many researchers primarily opting for relative and normalization-based methods that do not specifically address microbiome data characteristics. There is a lack of transparency and clear explanations regarding data preprocessing and the choice of transformation methods among the reviewed papers while it is crucial to adhere to best practices and provide a detailed methodology for developing machine learning pipelines, particularly regarding data preprocessing.

This mini review does not intend to provide unequivocal recommendations in favor of one approach over another, instead, we encourage researchers to consider the characteristics of their data carefully and whether a particular transformation method is suitable for addressing their research questions and data characteristics.

Author contributions

EI: conceptualization, investigation, writing the draft and the final manuscript. ML: investigation and writing the draft and final manuscript. XD and AS: writing the draft manuscript. RS: investigation. KH, BS, DD'E, and MB revised the draft manuscript, provided comments and writing the final manuscript. LM-Z: conceptualization, investigation, and writing the draft and final manuscript. All authors contributed to the article and approved the submitted version.

Funding

This article is based upon work from COST Action ML4Microbiome "Statistical and machine learning techniques in human microbiome studies" (CA18131), supported by COST (European Cooperation in Science and Technology), www.cost.eu. ML acknowledges support by FCT - Fundação para a Ciência e a Tecnologia, I.P., with references UIDB/00297/2020 and UIDP/00297/2020 (NOVA Math), UIDB/00667/2020 and UIDP/00667/2020 (UNIDEMI), and CEECINST/00042/2021. KH acknowledges support through the HiTEc Cost Action CA21163 and the project PID2021-123833OB-I00 provided by the Spanish Ministry of Science and Innovation (MCIN/AEI/10:13039/501100011033) and ERDF A way of making Europe. MB acknowledges support through the Metagenopolis grant ANR-11-DPBS-0001. LM-Z is supported by Juan de la Cierva Grant (IJC2019-042188-I) from the Spanish State Research Agency of the Spanish Ministerio de Ciencia e Innovación y Ministerio de Universidades.

Acknowledgments

The authors are grateful to all the ML4Micorbiome members for the discussions and comments on this work during the ML4Microbiome meetings.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2023.1250909/full#supplementary-material

References

Adade, E. E., Al Lakhen, K., Lemus, A. A., and Valm, A. M. (2021). Recent progress in analyzing the spatial structure of the human microbiome: Distinguishing biogeography and architecture in the oral and gut communities. *Curr. Opin. Endocr. Metab. Res.* 18, 275–283. doi: 10.1016/j.coemr.2021.04.005

Aitchison, J. (1982). The statistical analysis of compositional data (with discussion). *J R Stat Soc Series B.* 44, 139–177.

Aitchison, J. (1986). *The statistical analysis of compositional data*. London: Chapman & Hall.

Amir, A., McDonald, D., Navas-Molina, J. A., Kopylova, E., Morton, J. T., Zech Xu, Z., et al. (2017). Deblur rapidly resolves single-nucleotide community sequence patterns. *MSystems* 2:e00191-16. doi: 10.1128/mSystems.00191-16

Arksey, H., and O'Malley, L. (2005). Scoping studies: towards a methodological framework. *Int. J. Soc. Res. Methodol.* 8, 19–32. doi: 10.1080/1364557032000119616

Baksi, K. D., Kuntal, B. K., and Mande, S. S. (2018). 'TIME': a web application for obtaining insights into microbial ecology using longitudinal microbiome data. *Front. Microbiol.* 9:36. doi: 10.3389/fmicb.2018.00036

Beghini, F., McIver, L. J., Blanco-Míguez, A., Dubois, L., Asnicar, F., Maharjan, S., et al. (2021). Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *elife* 10:e65088. doi: 10.7554/eLife.65088

Blanco-Míguez, A., Beghini, F., Cumbo, F., McIver, L. J., Thompson, K. N., Zolfo, M., et al. (2023). Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nat. Biotechnol.* 1–12. doi: 10.1038/s41587-023-01688-w

Bogart, E., Creswell, R., and Gerber, G. K. (2019). MITRE: inferring features from microbiota time-series data linked to host status. *Genome Biol.* 20:186. doi: 10.1186/s13059-019-1788-y

Bokulich, N. A., Subramanian, S., Faith, J. J., Gevers, D., Gordon, J. I., Knight, R., et al. (2013). Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nat. Methods* 10, 57–59. doi: 10.1038/nmeth.2276

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Bushnell, B., Rood, J., and Singer, E. (2017). BBMerge – Accurate paired shotgun read merging via overlap. *PLoS One* 12:e0185056. doi: 10.1371/journal.pone.0185056

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J., and Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* 13, 581–583. doi: 10.1038/nmeth.3869

Chen, L., Reeve, J., Zhang, L., Huang, S., Wang, X., and Chen, J. (2018). GMPR: A robust normalization method for zero-inflated count data with application to microbiome sequencing data. *PeerJ* 6:e4600. doi: 10.7717/peerj.4600

Chen, Y., Wu, T., Lu, W., Yuan, W., Pan, M., Lee, Y.-K., et al. (2021). Predicting the role of the human gut microbiome in constipation using machine-learning methods: a meta-analysis. *Microorganisms* 9:2149. doi: 10.3390/microorganisms9102149

Clarotto, L., Allard, D., and Menafoglio, A. (2022). A new class of α-transformations for the spatial analysis of compositional data. *Spat. Stat.* 47:100570. doi: 10.1016/j.spasta.2021.100570

Costea, P. I., Zeller, G., Sunagawa, S., and Bork, P. (2014). A fair comparison. *Nat. Methods* 11:359. doi: 10.1038/nmeth.2897

D'Elia, D., Truu, J., Lahti, L., Berland, M., Papoutsoglou, G., Ceci, M., et al. (2023). Advancing microbiome research with machine learning: key findings from the ML4Microbiome COST action. *Front. Microbiol.* 14:1257002. doi: 10.3389/fmicb.2023.1257002

Dhungel, E., Mreyoud, Y., Gwak, H.-J., Rajeh, A., Rho, M., and Ahn, T.-H. (2021). MegaR: an interactive R package for rapid sample classification and phenotype prediction using metagenome profiles and machine learning. *BMC Bioinformatics* 22:25. doi: 10.1186/s12859-020-03933-4

Eck, A., Zintgraf, L. M., de Groot, E. F. J., de Meij, T. G. J., Cohen, T. S., Savelkoul, P. H. M., et al. (2017). Interpretation of microbiota-based diagnostics by explaining individual classifier decisions. *BMC Bioinformatics* 18:441. doi: 10.1186/s12859-017-1843-1

Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., and Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* 27, 2194–2200. doi: 10.1093/bioinformatics/btr381

Egozcue, J. J., and Pawlowsky-Glahn, V. (2005). Groups of parts and their balances in compositional data analysis. *Math. Geol.* 37, 795–828. doi: 10.1007/s11004-005-7381-9

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., and Barceló-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Math. Geol.* 35, 279–300. doi: 10.1023/A:1023818214614

Fabijanić, M., and Vlahoviček, K. (2016). Big data, evolution, and metagenomes: predicting disease from gut microbiota codon usage profiles. *Methods Mol. Biol.* 1415, 509–531. doi: 10.1007/978-1-4939-3572-7_26

Fernández-Edreira, D., Liñares-Blanco, J., and Fernandez-Lozano, C. (2021). Machine Learning analysis of the human infant gut microbiome identifies influential species in type 1 diabetes. *Expert Syst. Appl.* 185:115648. doi: 10.1016/j.eswa.2021.115648

Filzmoser, P., Hron, K., and Templ, M. (2018). *Applied compositional data analysis*. Cham: Springer International Publishing.

Filzmoser, P., and Walczak, B. (2014). What can go wrong at the data normalization step for identification of biomarkers? *J. Chromatogr. A* 1362, 194–205. doi: 10.1016/j.chroma.2014.08.050

Flemer, B., Warren, R. D., Barrett, M. P., Cisek, K., Das, A., Jeffery, I. B., et al. (2018). The oral microbiota in colorectal cancer is distinctive and predictive. *Gut* 67, 1454–1463. doi: 10.1136/gutjnl-2017-314814

Fouladi, F., Carroll, I. M., Sharpton, T. J., Bulik-Sullivan, E., Heinberg, L., Steffen, K. J., et al. (2021). A microbial signature following bariatric surgery is robustly consistent across multiple cohorts. *Gut Microbes* 13:1930872. doi: 10.1080/19490976.2021.1930872

Fukui, H., Nishida, A., Matsuda, S., Kira, F., Watanabe, S., Kuriyama, M., et al. (2020). Usefulness of machine learning-based gut microbiome analysis for identifying patients with irritable bowels syndrome. *J. Clin. Med.* 9:2403. doi: 10.3390/jcm9082403

Galkin, F., Mamoshina, P., Aliper, A., Putin, E., Moskalev, V., Gladyshev, V. N., et al. (2020). Human gut microbiome aging clock based on taxonomic profiling and deep learning. *IScience* 23:101199. doi: 10.1016/j.isci.2020.101199

Gloor, G. B., Wu, J. R., Pawlowsky-Glahn, V., and Egozcue, J. J. (2016). It's all relative: analyzing microbiome data as compositions. *Ann. Epidemiol.* 26, 322–329. doi: 10.1016/j.annepidem.2016.03.003

Greenacre, M. (2010). Log-ratio analysis is a limiting case of correspondence analysis. *Math. Geosci.* 42, 129–134. doi: 10.1007/s11004-008-9212-2

Greenacre, M. (2011). Measuring subcompositional incoherence. *Math. Geosci.* 43, 681–693. doi: 10.1007/s11004-011-9338-5

Greenacre, M., Martínez-Álvaro, M., and Blasco, A. (2021). Compositional data analysis of microbiome and any-omics datasets: a validation of the additive logratio transformation. *Front. Microbiol.* 12:727398. doi: 10.3389/fmicb.2021.727398

Gupta, A., Dhakan, D. B., Maji, A., Saxena, R., P K, V. P., Mahajan, S., et al. (2019). Association of *Flavonifractor plautii*, a flavonoid-degrading bacterium, with the gut microbiome of colorectal cancer patients in India. *MSystems* 4:e00438-19. doi: 10.1128/mSystems.00438-19

Gupta, M. M., and Gupta, A. (2021). Survey of artificial intelligence approaches in the study of anthropogenic impacts on symbiotic organisms – a holistic view. *Symbiosis* 84, 271–283. doi: 10.1007/s13199-021-00778-0

Hadrich, D. (2020). New EU projects delivering human microbiome applications. *Fut. Sci. OA* 6:FSO474. doi: 10.2144/fsoa-2020-0028

Hernández Medina, R., Kutuzova, S., Nielsen, K. N., Johansen, J., Hansen, L. H., Nielsen, M., et al. (2022). Machine learning and deep learning applications in microbiome research. *ISME Commun.* 2:98. doi: 10.1038/s43705-022-00182-9

Holmes, I., Harris, K., and Quince, C. (2012). Dirichlet Multinomial Mixtures: Generative Models for Microbial Metagenomics. *PLoS One* 7:e30126. doi: 10.1371/journal.pone.0030126

Hughes, D. A., Bacigalupe, R., Wang, J., Rühlemann, M. C., Tito, R. Y., Falony, G., et al. (2020). Genome-wide associations of human gut microbiome variation and implications for causal inference analyses. *Nat. Microbiol.* 5, 1079–1087. doi: 10.1038/s41564-020-0743-8

Jeganathan, P., and Holmes, S. P. (2021). A statistical perspective on the challenges in molecular microbial biology. *J. Agric. Biol. Environ. Stat.* 26, 131–160. doi: 10.1007/s13253-021-00447-1

Jian, C., Luukkonen, P., Yki-Järvinen, H., Salonen, A., and Korpela, K. (2020). Quantitative PCR provides a simple and accessible method for quantitative microbiota profiling. *PLoS One* 15:e0227285. doi: 10.1371/journal.pone.0227285

Jiang, Z., Li, J., Kong, N., Kim, J.-H., Kim, B.-S., Lee, M.-J., et al. (2022). Accurate diagnosis of atopic dermatitis by combining transcriptome and microbiota data with supervised machine learning. *Sci. Rep.* 12:290. doi: 10.1038/s41598-021-04373-7

Jiang, S., Xiao, G., Koh, A. Y., Kim, J., Li, Q., and Zhan, X. (2021). A Bayesian zero-inflated negative binomial regression model for the integrative analysis of microbiome data. *Biostatistics* 22, 522–540. doi: 10.1093/biostatistics/kxz050

Kapoor, S., and Narayanan, A. (2022). Leakage and the reproducibility crisis in ML-based science. Available at: http://arxiv.org/abs/2207.07048.

Kubinski, R., Djamen-Kepaou, J.-Y., Zhanabaev, T., Hernandez-Garcia, A., Bauer, S., Hildebrand, F., et al. (2022). Benchmark of data processing methods and machine learning models for gut microbiome-based diagnosis of inflammatory bowel disease. *Front. Genet.* 13:784397. doi: 10.3389/fgene.2022.784397

Lahti, L., Salonen, A., Kekkonen, R. A., Salojärvi, J., Jalanka-Tuovinen, J., Palva, A., et al. (2013). Associations between the human intestinal microbiota, *Lactobacillus rhamnosus* GG and serum lipids indicated by integrated analysis of high-throughput profiling data. *PeerJ* 1:e32. doi: 10.7717/peerj.32

Lê Cao, K.-A., Costello, M.-E., Lakis, V. A., Bartolo, F., Chua, X.-Y., Brazeilles, R., et al. (2016). MixMC: A multivariate statistical framework to gain insight into microbial communities. *PLoS One* 11:e0160169. doi: 10.1371/journal.pone.0160169

Liu, W., Fang, X., Zhou, Y., Dou, L., and Dou, T. (2022). Machine learning-based investigation of the relationship between gut microbiome and obesity status. *Microbes Infect.* 24:104892. doi: 10.1016/j.micinf.2021.104892

Liu, Z., Hsiao, W., Cantarel, B. L., Drábek, E. F., and Fraser-Liggett, C. (2011). Sparse distance-based learning for simultaneous multiclass classification and feature selection of metagenomic data. *Bioinformatics* 27, 3242–3249. doi: 10.1093/bioinformatics/btr547

Liu, Y., Méric, G., Havulinna, A. S., Teo, S. M., Åberg, F., Ruuskanen, M., et al. (2022). Early prediction of incident liver disease using conventional risk factors and gut-microbiome-augmented gradient boosting. *Cell Metab.* 34, 719–730.e4. doi: 10.1016/j.cmet.2022.03.002

Lloréns-Rico, V., Vieira-Silva, S., Gonçalves, P. J., Falony, G., and Raes, J. (2021). Benchmarking microbiome transformations favors experimental quantitative approaches to address compositionality and sampling depth biases. *Nat. Commun.* 12:3562. doi: 10.1038/s41467-021-23821-6

Lo, C., and Marculescu, R. (2019). MetaNN: accurate classification of host phenotypes from metagenomic data using neural networks. *BMC Bioinformatics* 20:314. doi: 10.1186/s12859-019-2833-2

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15:550. doi: 10.1186/s13059-014-0550-8

Marcos-Zambrano, L. J., Karaduzovic-Hadziabdic, K., Loncar Turukalo, T., Przymus, P., Trajkovik, V., Aasmets, O., et al. (2021). Applications of machine learning in human microbiome studies: a review on feature selection, biomarker identification, disease prediction and treatment. *Front. Microbiol.* 12:634511. doi: 10.3389/fmicb.2021.634511

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal* 17:10. doi: 10.14806/ej.17.1.200

McKnight, D. T., Huerlimann, R., Bower, D. S., Schwarzkopf, L., Alford, R. A., and Zenger, K. R. (2019). Methods for normalizing microbiome data: An ecological perspective. *Methods Ecol. Evol.* 10, 389–400. doi: 10.1111/2041-210X.13115

Mirzayi, C., Renson, A., Furlanello, C., Sansone, S.-A., Zohra, F., Elsafoury, S., et al. (2021). Reporting guidelines for human microbiome research: the STORMS checklist. *Nat. Med.* 27, 1885–1892. doi: 10.1038/s41591-021-01552-x

Moreno-Indias, I., Lahti, L., Nedyalkova, M., Elbere, I., Roshchupkin, G., Adilovic, M., et al. (2021). Statistical and machine learning techniques in human microbiome studies: contemporary challenges and solutions. *Front. Microbiol.* 12:635781. doi: 10.3389/fmicb.2021.635781

Mulenga, M., Abdul Kareem, S., Qalid Md Sabri, A., Seera, M., Govind, S., Samudi, C., et al. (2021). Feature extension of gut microbiome data for deep neural network-based colorectal cancer classification. *IEEE Access* 9, 23565–23578. doi: 10.1109/ACCESS.2021.3050838

Murovec, B., Deutsch, L., and Stres, B. (2021). General unified microbiome profiling pipeline (GUMPP) for large scale, streamlined and reproducible analysis of bacterial 16S rRNA data to predicted microbial metagenomes, enzymatic reactions and metabolic pathways. *Metabolites* 11:336. doi: 10.3390/metabo11060336

Ni, Y., Lohinai, Z., Heshiki, Y., Dome, B., Moldvay, J., Dulka, E., et al. (2021). Distinct composition and metabolic functions of human gut microbiota are associated with cachexia in lung cancer patients. *ISME J.* 15, 3207–3220. doi: 10.1038/s41396-021-00998-8

Ning, J., and Beiko, R. G. (2015). Phylogenetic approaches to microbial community classification. *Microbiome* 3:47. doi: 10.1186/s40168-015-0114-5

Papoutsoglou, G., Tarazona, S., Lopes, M. B., Klammsteiner, T., Ibrahimi, E., Eckenberger, J., et al. (2023). Machine learning approaches in microbiome research: challenges and best practices. *Front. Microbiol.* 14:1261889. doi: 10.3389/fmicb.2023.1261889

Pawlowsky-Glahn, V., Egozcue, J. J., and Tolosana-Delgado, R. (2015). *Modelling and analysis of compositional data*. Chichester: John Wiley & Sons, Ltd.

Props, R., Kerckhof, F.-M., Rubbens, P., De Vrieze, J., Hernandez Sanabria, E., Waegeman, W., et al. (2017). Absolute quantification of microbial taxon abundances. *ISME J.* 11, 584–587. doi: 10.1038/ismej.2016.117

Quinn, T. P., and Erb, I. (2020). Interpretable log contrasts for the classification of health biomarkers: a new approach to balance selection. *MSystems* 5:e00230-19. doi: 10.1128/mSystems.00230-19

Quinn, T. P., Erb, I., Richardson, M. F., and Crowley, T. M. (2018). Understanding sequencing data as compositions: an outlook and review. *Bioinformatics* 34, 2870–2878. doi: 10.1093/bioinformatics/bty175

Reiman, D., Layden, B. T., and Dai, Y. (2021). MiMeNet: Exploring microbiome-metabolome relationships using neural networks. *PLoS Comput. Biol.* 17:e1009021. doi: 10.1371/journal.pcbi.1009021

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi: 10.1093/bioinformatics/btp616

Ruuskanen, M. O., Åberg, F., Männistö, V., Havulinna, A. S., Méric, G., Liu, Y., et al. (2021). Links between gut microbiome composition and fatty liver disease in a large population sample. *Gut Microbes* 13, 1–22. doi: 10.1080/19490976.2021.1888673

Ryan, F. J., Ahern, A. M., Fitzgerald, R. S., Laserna-Mendieta, E. J., Power, E. M., Clooney, A. G., et al. (2020). Colonic microbiota is associated with inflammation and host epigenomic alterations in inflammatory bowel disease. *Nat. Commun.* 11:1512. doi: 10.1038/s41467-020-15342-5

Silverman, J. D., Roche, K., Mukherjee, S., and David, L. A. (2020). Naught all zeros in sequence count data are the same. *Comput. Struct. Biotechnol. J.* 18, 2789–2798. doi: 10.1016/j.csbj.2020.09.014

Stämmler, F., Gläsner, J., Hiergeist, A., Holler, E., Weber, D., Oefner, P. J., et al. (2016). Adjusting microbiome profiles for differences in microbial load by spike-in bacteria. *Microbiome* 4:28. doi: 10.1186/s40168-016-0175-0

Statnikov, A., Henaff, M., Narendra, V., Konganti, K., Li, Z., Yang, L., et al. (2013). A comprehensive evaluation of multicategory classification methods for microbiomic data. *Microbiome* 1:11. doi: 10.1186/2049-2618-1-11

Štefelová, N., Palarea-Albaladejo, J., and Hron, K. (2021). Weighted pivot coordinates for partial least squares-based marker discovery in high-throughput compositional data. *Stat. Anal. Data Mining ASA Data Sci. J.* 14, 315–330. doi: 10.1002/sam.11514

Swift, D., Cresswell, K., Johnson, R., Stilianoudakis, S., and Wei, X. (2023). A review of normalization and differential abundance methods for microbiome counts data. WIREs. *Comput. Stat.* 15:e1586. doi: 10.1002/wics.1586

Tap, J., Derrien, M., Törnblom, H., Brazeilles, R., Cools-Portier, S., Doré, J., et al. (2017). Identification of an intestinal microbiota signature associated with severity of irritable bowel syndrome. *Gastroenterology* 152, 111–123.e8. doi: 10.1053/j.gastro.2016.09.049

Thomas, A. M., Manghi, P., Asnicar, F., Pasolli, E., Armanini, F., Zolfo, M., et al. (2019). Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nat. Med.* 25, 667–678. doi: 10.1038/s41591-019-0405-7

Thorsen, J., Brejnrod, A., Mortensen, M., Rasmussen, M. A., Stokholm, J., Al-Soud, W. A., et al. (2016). Large-scale benchmarking reveals false discoveries and count transformation sensitivity in 16S rRNA gene amplicon data analysis methods used in microbiome studies. *Microbiome* 4:62. doi: 10.1186/s40168-016-0208-8

Travisany, D., Galarce, D., Maass, A., and Assar, R. (2015). "Predicting the metagenomics content with multiple CART trees" in *Mathematical Models in Biology* (Cham: Springer International Publishing), 145–160.

van den Boogaart, K. G., and Tolosana-Delgado, R. (2008). "compositions": A unified R package to analyze compositional data. *Comput. Geosci.* 34, 320–338. doi: 10.1016/j.cageo.2006.11.017

Vandeputte, D., Kathagen, G., D'hoe, K., Vieira-Silva, S., Valles-Colomer, M., Sabino, J., et al. (2017). Quantitative microbiome profiling links gut community variation to microbial load. *Nature* 551, 507–511. doi: 10.1038/nature24460

Vangay, P., Hillmann, B. M., and Knights, D. (2019). Microbiome Learning Repo (ML Repo): A public repository of microbiome regression and classification tasks. *GigaScience* 8:giz042. doi: 10.1093/gigascience/giz042

Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., et al. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* 5:27. doi: 10.1186/s40168-017-0237-y

Wirbel, J., Pyl, P. T., Kartal, E., Zych, K., Kashani, A., Milanese, A., et al. (2019). Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat. Med.* 25, 679–689. doi: 10.1038/s41591-019-0406-6

Wu, H., Cai, L., Li, D., Wang, X., Zhao, S., Zou, F., et al. (2018). Metagenomics biomarkers selected for prediction of three different diseases in Chinese population. *Biomed. Res. Int.* 2018, 1–7. doi: 10.1155/2018/2936257

Wu, S., Chen, Y., Li, Z., Li, J., Zhao, F., and Su, X. (2021). Towards multi-label classification: Next step of machine learning for microbiome research. *Comput. Struct. Biotechnol. J.* 19, 2742–2749. doi: 10.1016/j.csbj.2021.04.054

Wu, T., Wang, H., Lu, W., Zhai, Q., Zhang, Q., Yuan, W., et al. (2020). Potential of gut microbiome for detection of autism spectrum disorder. *Microb. Pathog.* 149:104568. doi: 10.1016/j.micpath.2020.104568

Xia, Y., Sun, J., and Chen, D.-G. (2018). *Statistical Analysis of Microbiome Data with R*. Springer: Singapore.

Xu, C., Zhou, M., Xie, Z., Li, M., Zhu, X., and Zhu, H. (2021). LightCUD: a program for diagnosing IBD based on human gut microbiome data. *BioData Mining* 14:2. doi: 10.1186/s13040-021-00241-2

Yachida, S., Mizutani, S., Shiroma, H., Shiba, S., Nakajima, T., Sakamoto, T., et al. (2019). Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. *Nat. Med.* 25, 968–976. doi: 10.1038/s41591-019-0458-7

Yang, F., and Zou, Q. (2020). mAML: an automated machine learning pipeline with a microbiome repository for human disease classification. *Database* 2020:baaa050. doi: 10.1093/database/baaa050

Yang, F., Zou, Q., and Gao, B. (2021). GutBalance: a server for the human gut microbiome-based disease prediction and biomarker discovery with compositionality addressed. *Brief. Bioinform.* 22:bbaa436. doi: 10.1093/bib/bbaa436

Zhang, X., Mallick, H., Tang, Z., Zhang, L., Cui, X., Benson, A. K., et al. (2017). Negative binomial mixed models for analyzing microbiome count data. *BMC Bioinformatics* 18:4. doi: 10.1186/s12859-016-1441-7

Zhu, C., Wang, X., Li, J., Jiang, R., Chen, H., Chen, T., et al. (2022). Determine independent gut microbiota-diseases association by eliminating the effects of human lifestyle factors. *BMC Microbiol.* 22:4. doi: 10.1186/s12866-021-02414-9

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 22 November 2023
DOI 10.3389/fmicb.2023.1264941

OPEN ACCESS

EDITED BY
Domenica D'Elia,
National Research Council (CNR), Italy

REVIEWED BY
Tatjana Loncar-Turukalo,
University of Novi Sad Faculty of Technical Sciences, Serbia
Naida Babic Jordamovic,
International Centre for Genetic Engineering and Biotechnology, Italy

*CORRESPONDENCE
Mustafa Temiz
✉ mustafa.temiz@agu.edu.tr
Malik Yousef
✉ malik.yousef@gmail.com

RECEIVED 21 July 2023
ACCEPTED 08 November 2023
PUBLISHED 22 November 2023

CITATION
Bakir-Gungor B, Temiz M, Jabeer A, Wu D and Yousef M (2023) microBiomeGSM: the identification of taxonomic biomarkers from metagenomic data using grouping, scoring and modeling (G-S-M) approach.
Front. Microbiol. 14:1264941.
doi: 10.3389/fmicb.2023.1264941

microBiomeGSM: the identification of taxonomic biomarkers from metagenomic data using grouping, scoring and modeling (G-S-M) approach

Burcu Bakir-Gungor[1], Mustafa Temiz[2]*, Amhar Jabeer[1], Di Wu[3,4] and Malik Yousef[5,6]*

[1]Department of Computer Engineering, Faculty of Engineering, Abdullah Gul University, Kayseri, Türkiye, [2]Department of Electrical and Computer Engineering, Faculty of Engineering, Abdullah Gul University, Kayseri, Türkiye, [3]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States, [4]Division of Oral and Craniofacial Health Sciences, Adams School of Dentistry, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States, [5]Department of Information Systems, Zefat Academic College, Zefat, Israel, [6]Galilee Digital Health Research Center (GDH), Zefat Academic College, Zefat, Israel

Numerous biological environments have been characterized with the advent of metagenomic sequencing using next generation sequencing which lays out the relative abundance values of microbial taxa. Modeling the human microbiome using machine learning models has the potential to identify microbial biomarkers and aid in the diagnosis of a variety of diseases such as inflammatory bowel disease, diabetes, colorectal cancer, and many others. The goal of this study is to develop an effective classification model for the analysis of metagenomic datasets associated with different diseases. In this way, we aim to identify taxonomic biomarkers associated with these diseases and facilitate disease diagnosis. The microBiomeGSM tool presented in this work incorporates the pre-existing taxonomy information into a machine learning approach and challenges to solve the classification problem in metagenomics disease-associated datasets. Based on the G-S-M (Grouping-Scoring-Modeling) approach, species level information is used as features and classified by relating their taxonomic features at different levels, including genus, family, and order. Using four different disease associated metagenomics datasets, the performance of microBiomeGSM is comparatively evaluated with other feature selection methods such as Fast Correlation Based Filter (FCBF), Select K Best (SKB), Extreme Gradient Boosting (XGB), Conditional Mutual Information Maximization (CMIM), Maximum Likelihood and Minimum Redundancy (MRMR) and Information Gain (IG), also with other classifiers such as AdaBoost, Decision Tree, LogitBoost and Random Forest. microBiomeGSM achieved the highest results with an Area under the curve (AUC) value of 0.98% at the order taxonomic level for IBDMD dataset. Another significant output of microBiomeGSM is the list of taxonomic groups that are identified as important for the disease under study and the names of the species within these groups. The association between the detected species and the disease under investigation is confirmed by previous studies in the literature. The microBiomeGSM tool and other supplementary files are publicly available at: https://github.com/malikyousef/microBiomeGSM.

KEYWORDS

gut microbiome, metagenomics, type 2 diabetes, inflammatory bowel disease, colorectal cancer, machine learning, classification, feature selection

1 Introduction

A diverse community of trillions of microorganisms, including bacteria, archaea, viruses, as well as microbial eukaryotes like fungus, protozoa, and helminths, comprise the human microbiome. Human microbiome has an impact on overall human health and on homeostasis by influencing immunological function and by actively contributing to human metabolism (Marcos-Zambrano et al., 2021). Several disease-related conditions have been connected to a rupture in the stable interaction between gut epithelial cells and the gut microbiota (Petersen and Round, 2014). The number of microbiome-related studies has significantly risen in the last 10 years, and large population studies such as the American Gut Project (McDonald et al., 2018), the metagenomics of the Human Intestinal Tract (Qin et al., 2010), and the Human Microbiome Project (The Human Microbiome Project Consortium, 2012) have greatly expanded the amount of information currently accessible on the content and function of the human gut microbiome. The information from these studies is crucial for further research on host-microbiome linkages and how they relate to the commencement and evolution of many complicated diseases.

The community of microbes performs a variety of tasks for the host, including facilitating the uptake of nutrients (Martin et al., 2019), preserving homeostasis (Ohland and Jobin, 2015), fending off pathogens (Pickard et al., 2017), regulating immunological response (Mendes et al., 2019), among many others. Understanding these tasks and revealing the dialog between the bacterium and the host may help in developing plans for preserving the health status, treating diseases. In the last few decades, there has been an increased interest in researching microbial communities (and their associations) that live in various habitats, from the gut to the biosphere. Technological advancements lead to lower costs for 16S and metagenomic sequencing, greater sequencing resolution and depth (Levy and Myers, 2016). Synchronous development of brand-new techniques for high throughput characterization of different -omic data types, such as lipidomics, metabolomics, metagenomics, metatranscriptomics and metaproteomics (Muller, 2019) made this possible. However, it is a difficult task to experimentally detect the inter species microbe host associations due to several other difficulties relating to scale, scope, feasibility, and availability of samples for concurrent -omic readouts (Fritz et al., 2013). Computational approaches can circumvent some of these constraints, improving our knowledge of microbial associations (Dix et al., 2016).

The interactions between the host and the microbiome are critical factors affecting human health and disease. Therefore, recently there has been an exponential increase in microbiome studies. Many research efforts have been devoted to predicting disease based on taxonomic profiles derived from metagenomic sequencing data. In these studies, machine learning methods are used to predict the microbiome interactions associated with diseases. Beyond simply assessing their predictive capabilities using machine learning, these studies also highlight the importance of specific microbiomes as potential biomarkers for disease. In literature, there are numerous articles investigating microbiomes associated with three specific diseases: Colorectal Cancer (CRC), Type 2 Diabetes (T2D) and Inflammatory Bowel Disease (IBD). In particular, several studies aiming to uncover microbiomes related to T2D are summarized in Gao et al. (2018), Gurung et al. (2020), Cena et al. (2023), and Li R. et al. (2023). Microbiomes associated with CRC are reviewed in Huybrechts et al. (2020), Tabowei et al. (2022), Negrut et al. (2023), and Zwezerijnen-Jiwa et al. (2023). The studies of Soueidan and Nikolski (2016), LaPierre et al. (2019), Marcos-Zambrano et al. (2021), Lim et al. (2022), Hsu et al. (2023), and Mah et al. (2023) reviews the microbiomes associated with IBD.

More specifically, Deschênes et al. (2023) employed machine learning techniques to predict diseases by representing microbiomes using gene-based representations and taxonomic profiles. Through the creation of taxonomic profiles from shotgun metagenomic data, they identified significant taxa using their proposed methodology. They conducted experiments for five different diseases, namely type 2 diabetes, obesity, liver cirrhosis, colorectal cancer, and inflammatory bowel disease. For both IBD and CRC disease, the datasets used in Deschênes et al. (2023) are the same datasets used by the proposed approach in this study. In their study, they assessed the performance of nine distinct classifiers, including random forest, decision tree, two support vector machines with a linear kernel, random set coverage machine (rSCM), two logistic regressions, SVM with a radial basis function kernel (SVMrbf), and an ensemble algorithm derived from SCM (set coverage machine). For each dataset, they applied embedded feature selection techniques, such as random forest and ranking features based on resulting models, followed by machine learning model application. They reported improved classification performance for certain diseases by employing taxonomic profiling. The most effective results in taxonomic profiling were achieved using the random forest algorithm for liver cirrhosis, yielding an AUC of 88%. Their study demonstrated the effective use of converting microbiome data into taxonomic representation data for disease prediction. They reported that Lachnospiraceae microbiome is found as associated with T2D and it can be considered as a biomarker for this disease.

Sharma et al. (2020) predicted disease states using machine learning methods by examining related Operational Taxonomic Units (OTUs) at the same phylum taxonomic level, exploiting the connections among OTUs at this taxonomic rank. Their investigation focused on the relationship between disease and the microbiome, utilizing shotgun datasets for two distinct diseases, T2D and Cirrhosis. The dataset they chose for T2D analysis is the same as the dataset used by our proposed tool. They applied their proposed method, which they called "TaxoNN," to a dataset with 174 cases and 170 controls for T2D (Qin et al., 2012) and a dataset with 118 cases and 114 controls for cirrhosis (Qin et al., 2014). TaxoNN is a Deep Learning based multi-layered approach to group OTU information based on phylum clusters. It trains clusters containing OTUs that share the same phylum separately using Convolutional Neural Networks (CNNs). It combines features from each cluster to enhance prediction accuracy via an ensemble learning technique. Their proposed method was evaluated using six different classifiers, including Random Forest, Gaussian Bayes Classifier, Naive Bayes, Ridge Regression, Lasso Regression, and Support Vector Machines. The TaxoNN method yielded the highest result, achieving an AUC of 92% for cirrhosis and 75% for T2D. Moreover, TaxoNN identified microbiomes at the level of three dominant phyla (Firmicutes, Proteobacteria, and Actinobacteria) for both diseases, highlighting their impact on the diseases.

Giliberti et al. (2022) investigated the influence of the relative abundance of microbial taxa on host phenotype classification using human metagenomes. They employed machine learning methods to construct species-level taxonomic profiles and accurately detected the

presence of microbial taxa. In their evaluation scheme, they encompassed a total of 4,128 samples from 25 shotgun metagenomic datasets. Among the datasets used in their study, T2D dataset is same with the dataset used in this study. They also explored the effect on disease prediction using relative abundance values at three different taxonomic levels: genus, family, and order. Employing the Random Forest classification algorithm on species level dataset, they achieved the best performance for IBD dataset, across other datasets containing seven distinct disease categories (atherosclerotic cardiovascular disease, Alzheimer's disease, Behçet's disease, colorectal cancer, irritable bowel disease, type 1 diabetes, and type 2 diabetes). They identified statistically significant microbiomes for the diseases they identified. Among these microbiomes for these cases, the most significant result was obtained for Clostridium and this microbiome was followed by *Streptococcus* and *Ruthenibacterium*.

Pasolli et al. (2016) investigated the utility of microbiomes in disease prediction using metagenomic datasets for five different diseases: liver cirrhosis, CRC, IBD, obesity, and T2D. Among the datasets used in this study, T2D dataset is also utilized within this study. They conducted species-level prediction using microbiome profiles at the species level derived from metagenomic data. Their analysis encompassed a total of 2,424 shotgun metagenomic data samples from eight distinct studies. Employing cross-validation techniques, they compared classification outcomes using two widely employed classifiers in metagenomic data analysis, Random Forest and Support Vector Machine. In addition to these classifiers, they also evaluated the effectiveness of elastic network, neural network, and multiple regression methods. In addition to predicting diseases using microbiome data, they highlighted prominent microbiomes related to these diseases. Notably, they identified the Peptostreptococcus microbiome for colorectal cancer, the Streptococcus microbiome for T2D, and the Lachnospiraceae microbiome for IBD as influential microbiomes in disease prediction. Collectively, these papers advance our understanding for the potential role of the microbiome in these diseases using a variety of approaches and analyzes.

Identifying microbial taxa that may cause disease development and identifying microbial taxa whose impact varies depending on their abundance is one of the major goals of human microbiome studies. Uncovering the influence of taxons can help to the investigation of disease development processes and hence can contribute to the emergence of new approaches for prevention of these diseases (Zhang W. et al., 2022). Computational methods dealing with microbial relative abundances face several challenges in drawing meaningful conclusions due to their complex data structures and properties. Traditional computational methods are inadequate to assess microbiome population effects in isolation and to produce effective results without considering the diversity of the human microbiome. Recent research has used machine learning (ML) approaches to evaluate data from the human microbiome, more specifically to identify and understand the diversity of taxonomy and function within microbial communities, and to assess the impact of these factors on human health (Topçuoğlu et al., 2020). The use of ML in microbiome studies can be summarized as follows:

- ML models have been created to promote taxonomic representation and differentiation in microbiology.
- ML has been used for disease prediction by inferring host phenotypes.
- ML facilitates the characterization of disease-specific microbial signatures to classify patients based on microbial communities (Marcos-Zambrano et al., 2021).

In this paper, we present a novel approach, microBiomeGSM, to detect disease-associated taxonomic biomarkers by developing an efficient machine learning model based on the Grouping, Scoring and Modeling (G-S-M) approach. We have analyzed taxonomically transformed microbiome sequencing datasets with our proposed machine learning method. In this way, we aim to reveal the impact of the identified taxonomic biomarkers on specific diseases. To this end, our study contributes to the diagnosis and treatment of the disease under investigation. The proposed approach is applied on metagenomic datasets associated with 4 different datasets; and the taxonomic groups that have an impact on disease under study are identified. In the data preprocessing step, the MetaPhlAn tool developed by Ditzler et al. (2015) is used to extract taxonomic data from microbiome sequencing data. In the first component (grouping component) of microBiomeGSM, the species identified in a sample are grouped according to the level of taxa known to be associated with them. In the second component (scoring component) of microBiomeGSM, importance scores are assigned to taxon groups using inherent machine learning techniques. The score is a predictor of how well a sample can be classified based on the abundance values of the species included in that taxon group. In the final (modeling) component of microBiomeGSM, three different outputs are generated. The first output is the performance metrics of the developed machine learning model. The second output is the list of important taxa groups associated with the disease under study, and these taxonomic features can be considered as biomarkers. The third output is the species associated with the taxa groups. Performance evaluation of microBiomeGSM is assessed separately for each disease, and for 3 different taxonomic levels (genus, family, order). Feature selection algorithms are applied to the same dataset in order to comparatively evaluate the performance of microBiomeGSM. The biological relevance of the identified taxon groups at genus, family, order levels for different diseases is discussed with reference to existing knowledge in the literature.

2 Materials and methods

2.1 Dataset

The data used in this study are obtained from the NCBI Sequence Read Archive (SRA045646, SRA050230) provided by Qin et al. (2012) for T2D; accession number PRJNA398089 in the SRA for the Integrative Human Microbiome Project for IBDMDB (Beghini et al., 2021). IBD dataset is obtained from the MetaHit project (Marco-Ramell et al., 2018) (ERA000116). The CRC metagenomic dataset containing 1,262 samples was created by Beghini et al. (2021). Microbiome sequencing data is classified into disease states based on the metadata associated with them. To ensure data quality, we applied quality filtering to meet the standards outlined in the Human Microbiome Project Consortium SOP (2012), as referenced in Thomas et al. (2019). This procedure allowed us to categorize the raw sequencing data according to relevant disease states, enabling our subsequent analyzes. The microbiome samples were associated with

TABLE 1 The list of datasets used to test the model.

#	Dataset	# of Samples	# of positives	# of features (Species)	# of Groups (Genus)	# of Groups (Family)	# of Groups (Order)
1	CRC	1,262	600	912	261	100	49
2	IBDMDB	1,638	1,209	579	187	77	43
3	IBD	382	148	1,456	448	177	84
4	T2D	290	155	1,456	448	177	84

Number of samples who have positive class label are shown in the second column. The number of features/Species is shown in the third column. Number of groups created at Order, Family and Genus taxonomic levels are listed at the 4-6th columns, respectively.

TABLE 2 Statistical information about the numbers of features within a group, shown separately for each taxonomic level.

#	Dataset	Genus (avg/ max/min)	Family (avg/ max/min)	Order (avg/max/ min)
1	CRC	3.51 /52/1	9.16 /76/1	18.71/202/ 1
2	IBDMDB	3.09 /34/1	7.50 /64/1	13.44/163/ 1
3	IBD	3.24 /61/1	8.22 /65/1	17.32/195/1
4	Type 2 diabetes	3.24 /61/1	8.22 /65/1	17.32/195/ 1

Sen is the sensitivity, Spe is the specificity, AUC is the Area Under the Curve.

the microbial species of origin (taxa) using the MetaPhlAn tool, and the relative abundance composition for each taxon was generated accordingly. These taxa and their relative abundances serve as features or variables in our machine learning approaches. MetaPhlAn first assigns reads to microbial clusters using clade-specific genes for assignment. It then presents the relative abundance of microbial taxa based on these readings. In this study, the assignment to microbial species of origin (taxa) was determined for each DNA sequence using the MetaPhlAn tool. The relative abundance value is normalized by dividing the number of reads for each taxonomic level by the total number of reads for only one sample. In this way, the taxonomic abundance values are expressed as real numbers in the range [0,1] with a sum of 1 for each sample. Samples with less than 1 million total reads were not included in our study. For each sample, we determined the diversity of disease-relevant microbiomes, where diversity represents the presence and relative abundance of microorganisms (Alatawi et al., 2022).

The four microbiome datasets used to evaluate the microBiomeGSM tool are listed in Table 1. The table presents the number of samples in each dataset and the number of samples that are labeled as positive. Positive samples refer to patients, while negative samples refer to controls. Each dataset contains the abundance values of the species, which we consider as features. We have considered 3 taxonomic levels for creating the groups, i.e., genus, family, and order. For each dataset, the number of extracted groups is listed in the corresponding column, while '-' denotes missing information.

Statistical information regarding the numbers of features in each group is given in Table 2. For each data set and for each taxonomic level (genus, family, and order), the average, maximum, and minimum numbers of features within a group are given.

Supplementary Table S1 shows the distribution of the groups based on their sizes for the IBDMDB dataset. The numbers in the table indicate the number of groups that have the specified number of species for that specific taxonomic level. There are 187, 77, and 43 groups for genus, family and order levels, respectively. About 90% of the groups at the order level, about 90% of the groups at the family level, and about 97% of the groups at the genus level contain 20 or fewer species for the IBDMDB dataset.

2.2 microBiomeGSM

Our proposed method, microBiomeGSM, consists of three main components: Grouping, Scoring, and Modeling (G-S-M). The G-S-M approach has been used in other studies that consider the pre-existing biological knowledge (Yousef et al., 2019, 2021a,c, 2022a; Qumsiyeh et al., 2022; Yousef and Voskergian, 2022; Ersoz et al., 2023; Jabeer et al., 2023). Additionally it was modified to integrate two-omics datasets such as the miRcorrNet and miRModuleNet tools (Yousef et al., 2021a, 2022b); and even to integrate 3 omics datasets such as 3Mint tool (Unlu Yazici et al., 2023). Interested readers can find further details about those approaches in our recent reviews (Yousef et al., 2021b; Kuzudisli et al., 2023).

Utilizing the G-S-M approach, microBiomeGSM performs a search to identify the most important taxonomic groups in disease-associated metagenomic datasets. The relative abundance values of the species within the group can be checked for each sample; and the generated model decides whether the sample has the disease or not. By focusing on a specific taxonomic level, we can use the G component to find the most significant group for the disease under study. This approach provides the advantage of focusing on either the macroscopic or microscopic view of the most important group to distinguish between healthy samples and patient samples. An overview of the steps performed in microBiomeGSM is presented in Figure 1.

Let X be the two-class dataset consisting of the species in the columns, and samples in the rows including the class labels (1 denoting the disease state and 0 denoting the healthy state). To understand the approach in detail, let us assume that the taxonomic level is selected as "genus" for the "Select taxa rank" step in Figure 1. The input X_{abd} (abundance matrix) is first split into a training set (X_{train}) and a test set (X_{test}) with a ratio of 80:20 based on the class labels. Denote by S the feature space of all species in X_{abd} and by U_{genus} all unique genera for S. Grp{} denotes the selection function of each U_{genus} in S, grouping all species on the basis of similar genuses. Grp{U_{genus}^{i} for S} represents each genus in S, with all the species grouped by genus. For example, if we take Alistipes as one of the genus in U_{genus}, we get the following when we apply the Grp function.

Grp{U_{genus}^{i}}, where i = Alistipes and ∈ S.

Grp{Alistipes} = {alistipes_finegoldi, alistipes_indistinctus, alistipes_inops, alistipes_shahii}.

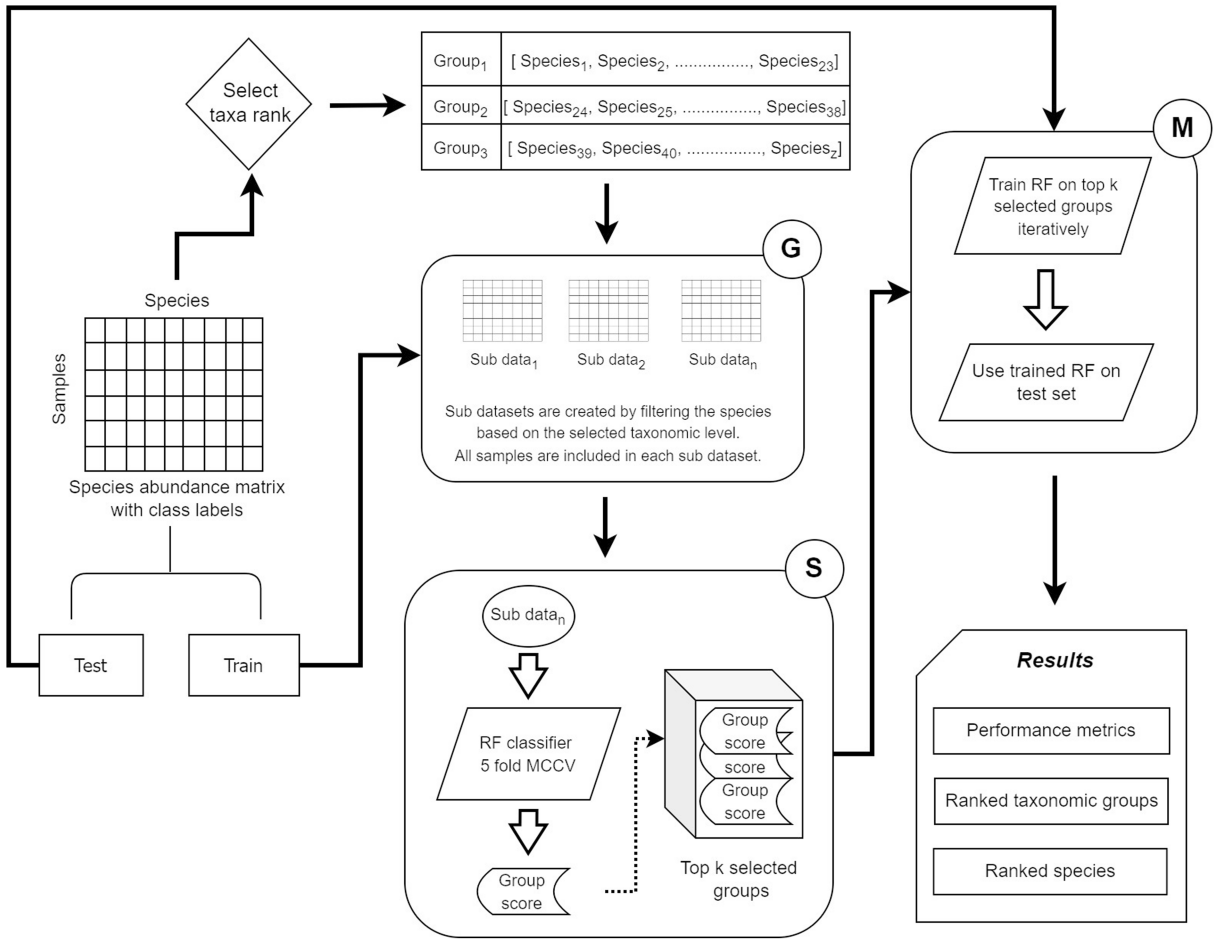

FIGURE 1
G-S-M approach in microBiomeGSM. MCCV denotes Monte Carlo Cross-Validation.

Similarly, this approach is applied to all genuses that are present in X_{abd}, and a list of genus groups is created, as shown in Figure 1 after the select taxa rank step. This is repeated for the three taxonomic levels identified.

When Figure 1 is examined, firstly, in the grouping component G, for all the groups of genus, we partition X_{train} into sub data denoted as sub_d_x. Following the earlier example of Alistipes, this group yields $sub\_d_{alistipes}$ which is created from X_{train}. The $sub\_d_{alistipes}$ contains the labels of the samples, but the feature space is restricted only to species within the Alistipes genus. This is applied to all different genera created in the prior step, so we have multiple subsets of data with a feature space specified by genus. Secondly, in the scoring step S, the generated sub_d is trained on a Random Forest classifier with 5-fold cross-validation with randomized stratified shuffling. Each sub_d is given a score equal to the mean of the accuracy over all foldings based on the prediction of the labels. Each sub_d is scored and then sorted based on the score. The top k groups with the highest score are used for the subsequent step. The value chosen for k is 10, but other values for k have been tested. Following the example of selecting genus as the taxonomic level, the top 10 genus groups that show strong discriminative ability are used to build the classification model. Thirdly, in the modeling component, the species from the top 10 genus groups are used to train a Random Forest model with 100-fold Monte Carlo Cross-Validation (MCCV). The top ranking set of species corresponding to the top ranked group is trained on X_{train} and then tested on X_{test}. Then, the second set of species corresponding to the second highest scoring group is aggregated with the top scoring set of species; and then used to train and test the model. This process is repeated until all species in the top 10 ranked genus groups are aggregated; and used to train and test the classifier. This whole process is repeated 100 times, stratifying the initial X_{abd} and randomly splitting it into X_{train} and X_{test} without replacement. The classification performance metrics are determined as the average of the metrics obtained in 100 folds. Similarly, the top ranked groups and the top ranked species are retained for each run.

2.3 Implementation of microBiomeGSM

The microBiomeGSM tool utilizes the pre-existing biological knowledge of the assignment of the species into different taxonomic levels, such as genus, family, and order. Experiments with the microBiomeGSM tool were conducted on the open-source KNIME platform (Berthold et al., 2009). This platform can handle a wide range of data types and operations. The user can configure the number of iterations, the rank function, and the number of iterations for MCCV. All rows with missing values are removed within the workflow.

2.4 Application of feature selection and classifiers using metagenomic data

In metagenomics research, it is observed that in studies using taxonomic features, the number of observations used for training data is higher than the number of observations used for testing data. This situation is undesirable if studies are to produce more effective results, and researchers are proposing various methods of resolution, particularly feature selection methods. Although the process of feature selection in disease prediction problems based on metagenome data has not been well studied, the literature suggests that this process may be as important as the choice of a classification method (LaPierre et al., 2019). The process of feature selection in metagenome-based disease prediction could help us learn more about disease development mechanisms. Therefore, further research in this direction is warranted. In metagenomics studies, in order to reduce the number of taxa, i.e., to select informative species (features), min Redundancy Max Relevance (mRMR) (Ding and Peng, 2005), Lasso (Tibshirani, 1996), Elastic Net (Zou and Hastie, 2005), and the iterative sure select algorithm (Duvallet et al., 2017) have been used extensively. Another feature selection method, called Fizzy, addresses the challenge of using classification techniques to identify important functional elements for downstream analysis (Ditzler et al., 2015). Oudah and Henschel presented an alternative taxonomy-based method for feature selection (Oudah and Henschel, 2018). Bakir-Gungor et al. (2021) applied CMIM (Fleuret and Ch, 2004), FCBF (Senliol et al., 2008), mRMR (Ding and Peng, 2005), and Select K best (SKB) (Pedregosa et al., 2011) to type 2 diabetes-associated metagenomics datasets and obtained powerful performance metrics (Bakir-Gungor et al., 2021). Jabeer et al. also proposed a robust classification method for evaluating colorectal cancer associated metagenomic datasets using a combination of feature selection methods and machine learning methods (Jabeer et al., 2022). Bakir-Gungor et al. (2022) also proposed a powerful method for IBD classification with fewer features by combining feature selection methods and machine learning methods (Bakir-Gungor et al., 2022). While these feature selection approaches have produced effective results in a variety of fields, they have only recently been applied to microbiome-based disease prediction problems.

In this study, we have comparatively evaluated microBiomeGSM with different classifiers and with different feature selection methods. As the feature selection methods, we have utilized Select K best (SKB), Fast Correlation Based Filter (FCBF), Extreme Gradient Boosting (XGBoost), Min Redundancy Max Relevance (mRMR), Information Gain (IG), and Conditional Mutual Information Maximization (CMIM). Wang and Liu (2020) compare the performance of classifiers with traditional methods and ensemble methods for disease prediction based on human microbiome data. They use Elastic Network and SVM as traditional methods and Random Forest and Extreme Gradient Boosting (XGBoost) as ensemble methods. In their study, they find that the XGBoost algorithm shows superior performance compared to other algorithms (Wang and Liu, 2020). In another study, Marcos-Zambrano et al. (2021) conducted an important review paper to reveal the links between the microbiome and diseases. In this study, which included information on the performance of machine learning methods, they found that the Support Vector Machines (SVM), Random Forest (RF), k-Nearest Neighbors (k-NN), and Logical Regression (LR) algorithms were widely used. They concluded that when selecting a machine learning algorithm, several factors should be considered such as the set of observations, the set of features, the type of data, and the quality of the data. They suggest using several different methods, comparing them, and choosing the one that provides the best performance value (Marcos-Zambrano et al., 2021).

2.5 microBiomeGSM model performance evaluation

Accuracy, F1 score, sensitivity, specificity, and AUC were used to evaluate the predictive performance of the proposed models. AUC score is a common measure for performance evaluation and a reliable metric for evaluating balanced datasets. Other metrics such as F1 score, sensitivity, specificity, and accuracy, were used to evaluate the performance of the created models because the dataset for this study has an uneven distribution of classes. When a balance between precision and recall is desired and there is an uneven distribution of classes, the F1 score is a good option among the performance metrics (many true negatives). Several classifiers report the probability values for their predictions, which can also be considered as confidence values for the prediction. The AUC often uses this information to figure out how often incorrect predictions occur at different confidence levels. In real life, test results from positive and negative examples overlap. AUC illustrates how the threshold or cut-off value for identifying positive examples affects the relationship between recall and precision. In this study, all of the above-mentioned metrics were calculated as the mean of 100 times MCCV. After each iteration, we obtain lists of significant taxonomic groups and species associated with these taxa groups for a given disease. To assign scores to the entities in the taxonomic groups list and in the species lists, a prioritization approach is used. For this purpose, we integrated the RobustRankAggreg algorithm (Kolde et al., 2012) and microBiomeGSM. RobustRankAggreg algorithm is available as an R package. Each entity (taxonomic group or species) in the lists is given a value of p by the RobustRankAggreg technique, indicating how highly ranked that entity. Using the RobustRankAggreg tool, microBiomeGSM outputs a list of species to which it has assigned a significance value (value of p) for a specific taxonomic group. Each taxa group is assigned a significance value and the species associated with that group are assigned the same value.

3 Results

The main objective of this study is to identify the microbial communities that are associated with specific diseases. In order to facilitate disease diagnosis, using metagenomic data we develop an efficient classification model based on taxonomic levels. In this section we present our findings for four different datasets. Here we also present comparative evaluation results against other existing methods.

3.1 Comparing varying group size for microBiomeGSM

One approach to evaluate model performance in the context of microBiomeGSM is to compare model performance between different values of the parameter k. k represents the number of groups (taxa) used in microBiomeGSM models. This approach can help researchers

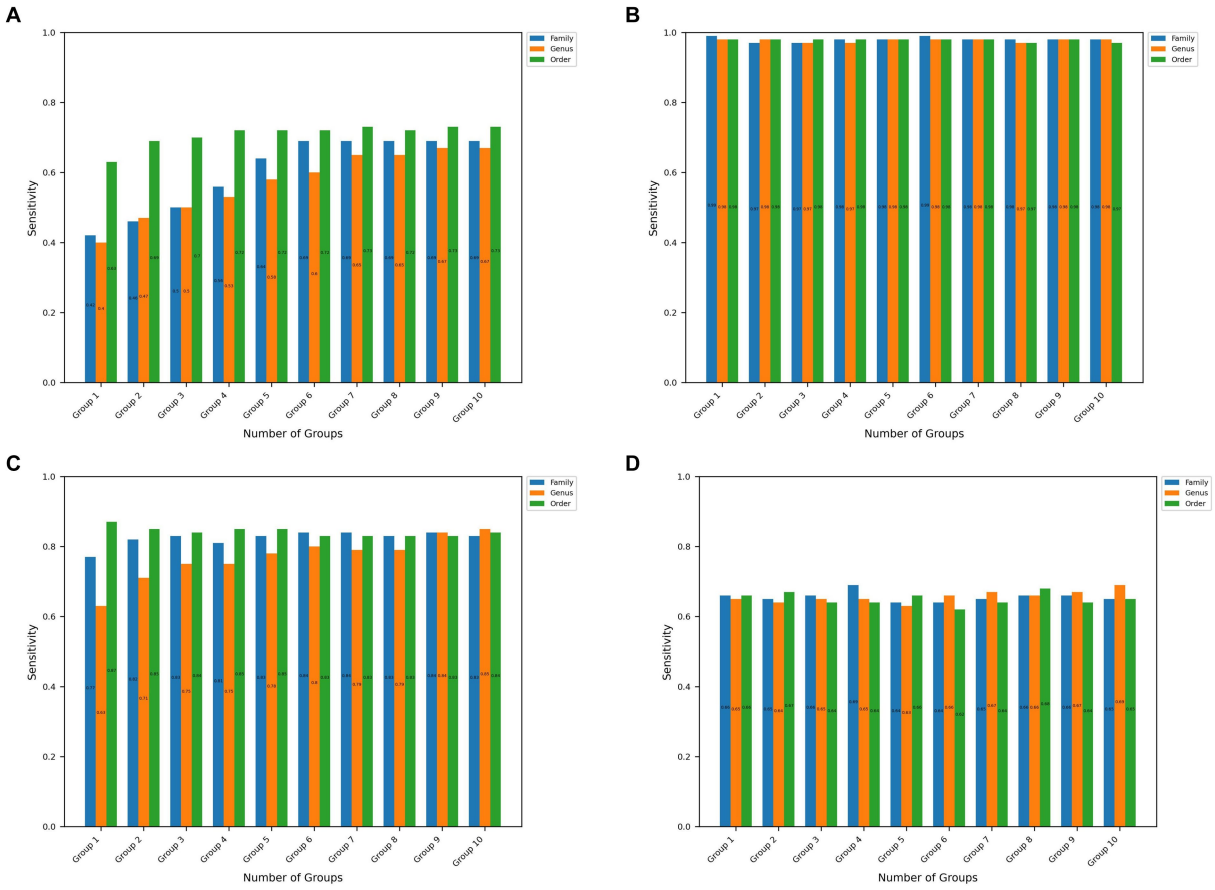

FIGURE 2
Sensitivity values obtained at the family, order, and genus taxon levels for the top 10 significant groups across all 4 datasets. **(A–D)** Represents the results obtained in CRC, IBDMDB, IBD, T2D datasets, respectively.

determine the optimal value of k that balances model complexity and predictive power, ultimately leading to more effective and interpretable models in microbiome-related research. It provides insight into how the inclusion or exclusion of specific taxa affects the overall performance of microBiomeGSM models.

Supplementary Table S2 shows the performance metrics obtained with 100-fold MCCV for the aggregated top 10 groups for four different datasets compared at three different taxonomic levels (genus, family, order) for grouping. For the IBDMDB dataset, microBiomeGSM achieved an AUC of 93% using the top 1 group at the family level. Performance metrics are shown for the top 2 groups via combining species from the first and second highest scoring groups. We obtained an AUC of 97% when the top 2 groups are combined at the family taxonomic level for the IBDMDB dataset. In this way, microBiomeGSM provides cumulative performance results for the top 10 highest scoring groups. For the IBDMDB dataset, the highest performance metric (an AUC of 98%) is obtained using the species from the top 10 groups at the order taxonomic level. For the IBD dataset, the highest performance metric (an AUC of 93%) is obtained using the species from the top 9 groups at the order taxonomic level. For the T2D dataset, the highest performance metric (an AUC of %74) is obtained using the species from the top 9 groups at the order taxonomic level. For the CRC dataset, the highest performance metric (an AUC of %83) is obtained using the species from the top 10 groups at the family taxonomic level. While examining other performance metrics (such as accuracy, sensitivity, specificity in Supplementary Table S2), it is noteworthy that satisfactory results are obtained with microBiomeGSM for each taxonomic level, especially for the IBDMDB dataset. The high sensitivity values that are reported for the CRC, IBDMDB, and IBD datasets display the success of the microBiomeGSM tool in terms of detecting the patient samples. In the CRC, IBDMDB, and IBD datasets, the strikingly high specificity values indicate that the microBiomeGSM tool correctly identifies the negative samples (i.e., individuals who do not have the disease). However, in the T2D dataset, the specificity rate appears to be relatively low compared to the other datasets. Nevertheless, the ability to detect negative samples remains at a reasonable level.

In addition, Figures 2, 3 show the sensitivity and specificity values obtained with the microBiomeGSM tool for all datasets. Figure 2 shows the sensitivity values obtained using the microBiomeGSM tool across all datasets. One can notice from Figure 2A that for the CRC data set the highest sensitivity value (73%) is obtained for the order taxon level using 10 cumulative groups. In particular, the sensitivity values calculated for the IBDMDB dataset were quite impressive, especially in group 1 and group 6, both at the family taxon level, reaching 99% sensitivity value, as shown in Figure 2B. Figure 2C shows another impressive set of results for the IBD data set. In Figure 2C, we observe high values for sensitivity, in particular 87%

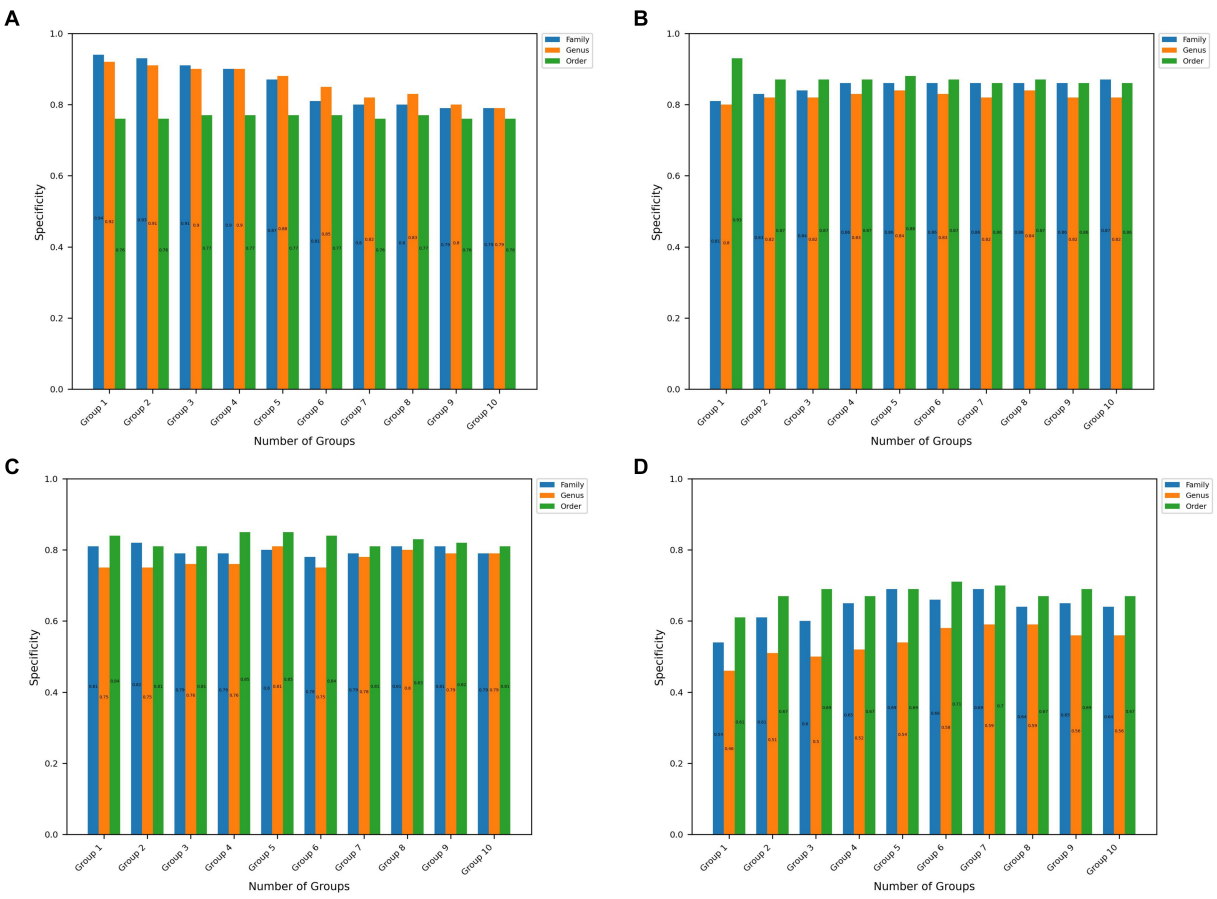

FIGURE 3
Specificity values at the order, genus, and family taxon level for the top 10 significant groups for all 4 disease datasets. **(A–D)** Represents the results obtained in CRC, IBDMDB, IBD, T2D datasets, respectively.

sensitivity at the taxon level in group 1. As shown in Figure 2D, the highest sensitivity value for the T2D data set is 69%. This result is obtained for the genus taxon level using 10 cumulative groups. A sensitivity value of 69% is also obtained for the family taxon level using 4 cumulative groups.

Figure 3 shows the specificity values obtained using the microBiomeGSM tool for all datasets. As shown in Figure 3A, the specificity value obtained for the CRC dataset is remarkable, reaching an impressive specificity value of 94% at the family taxon level for 1 group. Figure 3B depicts that the highest specificity value obtained for the IBDMDB dataset is 93% for 1 group at the order taxon level. As displayed in Figure 3C, the highest specificity value obtained for the IBD dataset is 85% for the 4 cumulative groups at the order taxon level. The same result is also obtained at the order taxon level for the 5 cumulative groups. One can notice in Figure 3D that the highest specificity value that is obtained for the T2D dataset is 71% for the 6 cumulative groups at the order taxon level.

The number of significant groups used to train the model could affect the performance of microBiomeGSM. Table 3 shows the influence of the number of groups and the number of species at family, genus and order levels on four datasets. Table 3 presents the performance of the top 10 cumulative groups and top 1 group for each taxonomic level on different tested datasets. For the IBDMDB dataset, for the family taxonomic level, one can observe that the AUC increases by 5% when we consider the top 10 significant groups cumulatively, while we increase the number of species from 34 to 205. On the same dataset, an increase of 8% in AUC score is observed at the Genus taxonomic level via increasing the number of species from 34 to 119. For the same dataset, a decrease of 1% is observed at the Order taxonomic level. Order taxonomic level using the top group that includes 98 species achieves the highest AUC success rate of 98% for the IBDMDB dataset. Similarly, family taxonomic level using the top 10 combined groups achieves 97% AUC on the IBDMDB dataset, but these 10 combined groups include a much higher number of species (205 species). For the IBD dataset, the highest AUC value of 91% was obtained using the microBiomeGSM tool. This value at the family taxonomic level was obtained by cumulatively combining 10 groups, using an average of 260.4 species. For the T2D dataset, the highest AUC value of 72% was obtained using the microBiomeGSM tool. This value, obtained at the order taxanomic level, was obtained by combining 10 groups cumulatively. For 1 group, an average of 138.28 species are used at the taxonomic level, while for 10 groups, an average of 596.99 species are used. For the CRC dataset, the highest AUC value of 87% was obtained using the microBiomeGSM tool. This value at the order taxanomic level was obtained by cumulatively combining 10 groups, using an average of 604 species.

microBiomeGSM reports important groups of features that are detected at different taxonomic levels for the disease under study.

TABLE 3 The effect of the number of groups that are generated at different taxonomic levels on performance metrics for all dataset.

CRC								
Taxonomic hierarchy	# of groups	Average # of species	Accuracy	Sen	Spe	F measure	AUC	Precision
Family	10	239.11	0.78	0.84	0.72	0.79	0.84	0.76
Family	1	16.17	0.67	0.91	0.43	0.73	0.69	0.62
Genus	10	102.06	0.77	0.82	0.71	0.78	0.84	0.76
Genus	1	7.16	0.66	0.88	0.44	0.72	0.71	0.63
Order	10	604	0.82	0.86	0.78	0.82	**0.87**	0.81
Order	1	154.25	0.76	0.82	0.71	0.78	0.81	0.76
IBDMDB								
Taxonomic hierarchy	# of Groups	Average # of species	Accuracy	Sen	Spe	F measure	AUC	Precision
Family	10	205.76	0.95	0.98	0.87	0.95	0.97	0.93
Family	1	34	0.93	0.98	0.81	0.94	0.93	0.9
Genus	10	119.4	0.92	0.98	0.82	0.95	0.97	0.93
Genus	1	34	0.92	0.98	0.80	0.94	0.91	0.91
Order	10	341.22	0.93	0.97	0.86	0.95	**0.98**	0.93
Order	1	98	0.96	0.98	0.93	0.97	0.98	0.95
IBD								
Taxonomic hierarchy	# of Groups	Average # of species	Accuracy	Sen	Spe	F measure	AUC	Precision
Family	10	260.24	0.82	0.85	0.79	0.82	**0.91**	0.81
Family	1	51.59	0.78	0.78	0.78	0.78	0.86	0.78
Genus	10	121.78	0.81	0.83	0.79	0.81	0.88	0.8
Genus	1	12.26	0.7	0.67	0.74	0.69	0.78	0.73
Order	10	608.27	0.82	0.82	0.81	0.82	0.9	0.82
Order	1	174.86	0.81	0.82	0.8	0.81	0.9	0.81
T2D								
Taxonomic hierarchy	# of groups	Average # of species	Accuracy	Sen	Spe	F measure	AUC	Precision
Family	10	321.16	0.65	0.68	0.63	0.66	0.71	0.65
Family	1	39.86	0.59	0.71	0.47	0.63	0.63	0.58
Genus	10	129.8	0.64	0.64	0.64	0.64	0.69	0.65
Genus	1	15.94	0.56	0.62	0.49	0.58	0.58	0.55
Order	10	596.99	0.65	0.65	0.64	0.64	**0.72**	0.65
Order	1	138.28	0.59	0.67	0.52	0.62	0.64	0.59

The results in bold in table represent the best AUC results.

Table 4 lists the top 10 important groups that are identified by microBiomeGSM for three different taxonomic levels on four different datasets. The identified features are ranked by their importance scores from high to low. The feature with the highest importance value is the strongest candidate to be announced as potential taxonomic biomarker for the disease under investigation.

The microBiomeGSM tool lists a number of associated species for each identified group. The species included in the top 5 significant groups are listed in Supplementary Tables S3–S5 for family, order, and genus taxonomic levels, respectively for four different datasets. All species for the family, order, and genus taxonomic levels for the T2D, IBDMDB and CRC datasets can be found in Supplementary Tables S6–S14, respectively.

For the IBDMDB dataset, the changes in the AUC score when the number of groups is increased from 1 to 10 are shown in Supplementary Figure S1. For the IBDMDB dataset, a high AUC score is obtained at the order taxonomic level. When the number of groups was increased, the AUC score decreased relatively, and no significant change was observed after 5 groups. At the genus and family taxonomic levels, there is a significant increase in the AUC

10.3389/fmicb.2023.1264941

TABLE 4 Top 10 groups identified by microBiomeGSM for different taxonomic levels, applied on all microbiome datasets.

CRC			
#	Taxonomic levels		
Rank	Family	Order	Genus
1	PEPTOSTREPTOCOCCACEAE	CLOSTRIDIALES	PARVIMONAS
2	PEPTONIPHILACEAE	TISSIERELLALES	PEPTOSTREPTOCOCCUS
3	FUSOBACTERIACEAE	BACTEROIDALES	FUSOBACTERIUM
4	BACILLALES_UNCLASSIFIED	FUSOBACTERIALES	GEMELLA
5	VEILLONELLACEAE	BACILLALES	DIALISTER
6	LACHNOSPIRACEAE	VEILLONELLALES	LACHNOCLOSTRIDIUM
7	ERYSIPELOTRICHACEAE	ERYSIPELOTRICHALES	PREVOTELLA
8	RUMINOCOCCACEAE	LACTOBACILLALES	STREPTOCOCCUS
9	PREVOTELLACEAE	ACTINOMYCETALES	PORPHYROMONAS
10	STREPTOCOCCACEAE	DESULFOVIBRIONALES	SOLOBACTERIUM

IBDMDB			
#	Taxonomic levels		
Rank	Family	Order	Genus
1	BACTEROIDACEAE	BACTEROIDALES	BACTEROIDES
2	LACHNOSPIRACEAE	CLOSTRIDIALES	ALISTIPES
3	RUMINOCOCCACEAE	FIRMICUTES_UNCLASSIFIED	EUBACTERIUM
4	RIKENELLACEAE	VEILLONELLALES	ROSEBURIA
5	FIRMICUTES_UNCLASSIFIED	BURKHOLDERIALES	FIRMICUTES_UNCLASSIFIED
6	TANNERELLACEAE	METHANOMASSILIICOCCALES	PARABACTEROIDES
7	EUBACTERIACEAE	DESULFOVIBRIONALES	RUMINOCOCCUS
8	CLOSTRIDIACEAE	ERYSIPELOTRICHALES	COPROCOCCUS
9	VEILLONELLACEAE	BIFIDOBACTERIALES	BLAUTIA
10	ODORIBACTERACEAE	EGGERTHELLALES	CLOSTRIDIUM

IBD			
#	Taxonomic levels		
Rank	Family	Order	Genus
1	LACHNOSPIRACEAE	CLOSTRIDIALES	BLAUTIA
2	BIFIDOBACTERIACEAE	CORIOBACTERIALES	BIFIDOBACTERIUM
3	CORIOBACTERIACEAE	BIFIDOBACTERIALES	EUBACTERIUM
4	RUMINOCOCCACEAE	ERYSIPELOTRICHALES	DOREA
5	ERYSIPELOTRICHACEAE	BACTEROIDALES	COLLINSELLA
6	CLOSTRIDIALES_FAMILY_XIII_INCERTAE_SEDIS	LACTOBACILLALES	PEPTOSTREPTOCOCCUS
7	EUBACTERIACEAE	SELENOMONADALES	COPROCOCCUS
8	PEPTOSTREPTOCOCCACEAE	VERRUCOMICROBIALES	ERYSIPELOTRICHACEAE_NONAME
9	CARNOBACTERIACEAE	CANDIDATUS_SACCHARIBACTERIA_NONAME	LACHNOSPIRACEAE_NONAME
10	CLOSTRIDIACEAE	BACILLALES	BACTEROIDES

T2D			
#	Taxonomic levels		
Rank	Family	Order	Genus
1	LACHNOSPIRACEAE	CLOSTRIDIALES	EUBACTERIUM
2	BIFIDOBACTERIACEAE	BIFIDOBACTERIALES	BIFIDOBACTERIUM
3	RUMINOCOCCACEAE	CORIOBACTERIALES	BLAUTIA
4	EUBACTERIACEAE	BACTEROIDALES	DOREA
5	CORIOBACTERIACEAE	LACTOBACILLALES	LACHNOSPIRACEAE_NONAME
6	CLOSTRIDIALES_FAMILY_XIII_INCERTAE_SEDIS	ERYSIPELOTRICHALES	RUMINOCOCCUS
7	ERYSIPELOTRICHACEAE	SELENOMONADALES	COPROCOCCUS
8	PEPTOSTREPTOCOCCACEAE	VERRUCOMICROBIALES	PEPTOSTREPTOCOCCUS
9	CARNOBACTERIACEAE	METHANOBACTERIALES	ERYSIPELOTRICHACEAE_NONAME
10	BACTEROIDACEAE	BACILLALES	GRANULICATELLA

TABLE 5 Area under the curve (AUC) results obtained using 100 features for different feature selection methods and classifiers for all dataset.

CRC						
Model	SKB	IG	XGB	FCBF	MRMR	CMIM
Adaboost	0.75 ± 0.02	0.71 ± 0.05	0.78 ± 0.04	0.71 ± 0.05	0.63 ± 0.06	0.77 ± 0.04
DT	0.67 ± 0.04	0.64 ± 0.04	0.69 ± 0.04	0.63 ± 0.06	0.61 ± 0.04	0.65 ± 0.05
Logitboost	0.76 ± 0.04	0.72 ± 0.05	0.78 ± 0.06	0.70 ± 0.04	0.64 ± 0.06	0.76 ± 0.05
RF	0.82 ± 0.03	0.79 ± 0.04	**0.85 ± 0.03**	0.77 ± 0.05	0.74 ± 0.04	0.80 ± 0.03

IBDMDB						
Model	SKB	IG	XGB	FCBF	MRMR	CMIM
Adaboost	0.89 ± 0.04	0.90 ± 0.03	0.89 ± 0.06	0.49 ± 0.08	0.51 ± 0.08	0.51 ± 0.08
DT	0.83 ± 0.03	0.82 ± 0.04	0.84 ± 0.03	0.46 ± 0.07	0.50 ± 0.07	0.50 ± 0.06
Logitboost	0.89 ± 0.04	0.91 ± 0.03	0.86 ± 0.06	0.50 ± 0.06	0.51 ± 0.08	0.49 ± 0.08
RF	0.96 ± 0.01	0.96 ± 0.01	**0.98 ± 0.01**	0.46 ± 0.1	0.54 ± 0.08	0.52 ± 0.07

IBD						
Model	SKB	IG	XGB	FCBF	MRMR	CMIM
Adaboost	0,90 ± 0.07	0,89 ± 0.03	0,91 ± 0.03	0,51 ± 0.06	0,51 ± 0.03	0,66 ± 0.08
DT	0,78 ± 0.08	0,70 ± 0.08	0,73 ± 0.07	0,53 ± 0.08	0,51 ± 0.04	0,56 ± 0.09
Logitboost	0,90 ± 0.04	0,90 ± 0.05	0,92 ± 0.05	0,55 ± 0.1	0,53 ± 0.05	0,59 ± 0.1
RF	**0,92 ± 0.03**	0,88 ± 0.06	0,91 ± 0.04	0,53 ± 0.09	0,55 ± 0.07	0,63 ± 0.11

T2D						
Model	SKB	IG	XGB	FCBF	MRMR	CMIM
Adaboost	0,56 ± 0.12	0,60 ± 0.05	0,64 ± 0.07	0,50 ± 0.10	0,5 ± 0.01	0,50 ± 0.12
DT	0,52 ± 0.08	0,52 ± 0.08	0,53 ± 0.05	0,41 ± 0.10	0,51 ± 0.02	0,49 ± 0.10
Logitboost	0,55 ± 0.10	0,58 ± 0.09	0,62 ± 0.10	0,48 ± 0.08	0,50 ± 0.01	0,51 ± 0.11
RF	0,62 ± 0.11	0,62 ± 0.07	**0,70 ± 0.06**	0,49 ± 0.08	0,51 ± 0.03	0,54 ± 0.10

The results in bold in table represent the best AUC results for the respective disease (CRC, IBDMDB, IBD, T2D).

score until 5 groups are combined and no significant change after 5 groups.

3.2 Comparing against traditional machine learning methods

Our Grouping-Scoring-Modeling (G-S-M) approach emerges as a paradigm shift from traditional feature selection methods. Instead of pinpointing individual informative features, the GSM methodology groups these features. These groups are then scored, and a classification model is built using these top-ranking feature conglomerates. The versatility of the GSM method, as detailed in our prior work (Yousef et al., 2021b), lies in its adaptability. Groups can be created either by computational/statistical methods or by using domain-specific knowledge. In order to use the GSM strategy for a given dataset, a deep domain expertise is required to skillfully define these groups, which makes each application different. The modifications required to tailor the G-S-M approach to the unique needs of microbiome research highlight the adaptability of the G-S-M method and the novelty of our current study.

We have comparatively evaluated the performance of microBiomeGSM against 4 different classifiers and 6 different feature selection methods using the same datasets. All algorithms are run with default parameters. The developed approach and feature selection methods were executed multiple times, and the results were averaged and shared. Table 5 shows the performance of the different feature selection algorithms and different classifiers on the same disease associated microbiome datasets. In these experiments, the number of features was set to 100. The best result for the IBDMDB dataset is obtained by using the XGBoost feature selection algorithm in combination with the Random Forest classification algorithm with 98% AUC. For the CRC dataset, the best result is obtained by using the XGBoost feature selection algorithm in combination with the Random Forest classification algorithm with an AUC of 85%. For the IBD dataset, the best result is obtained using the Random Forest classification algorithm with 92% AUC and the SKB feature selection algorithm. For the T2D dataset, the best result is obtained by using the XGBoost feature selection algorithm in combination with the Random Forest classification algorithm with 70% AUC.

We would like to note that the primary objective of microBiomeGSM is not to compete with other feature selection methods (FS). Even if microBiomeGSM's performance is on par with or slightly less favorable than other FS methods, its fundamental contribution lies in identifying the most informative microbiomes. These microbiomes play a pivotal role in aiding researchers in gaining

10.3389/fmicb.2023.1264941

TABLE 6 Evaluation metrics obtained with microBiomeGSM on four datasets for different taxonomic levels, compared with traditional classifiers using all features.

CRC						
Model	# of Species	Accuracy	Sensitivity	Specificity	Precision	AUC
AdaBoost	912	0.72 ± 0.06	0.79 ± 0.09	0.66 ± 0.17	0.7 ± 0.09	0.78 ± 0.04
DT	912	0.68 ± 0.09	0.75 ± 0.12	0.62 ± 0.26	0.66 ± 0.09	0.7 ± 0.04
LogitBoost	912	0.73 ± 0.06	0.78 ± 0.09	0.68 ± 0.18	0.71 ± 0.09	0.78 ± 0.04
RF	912	0.78 ± 0.05	0.82 ± 0.08	0.75 ± 0.14	0.76 ± 0.09	0.86 ± 0.03
microBiomeGSM: family	292.88 ± 16.09	0.74 ± 0.65	0.7 ± 0.39	0.77 ± 0.91	0.75 ± 0.83	0.81 ± 0.67
microBiomeGSM: genus	161.21 ± 5.17	0.74 ± 0.67	0.69 ± 0.41	0.79 ± 0.92	0.76 ± 0.84	0.8 ± 0.68
microBiomeGSM: order	607.5 ± 188.32	0.73 ± 0.69	0.72 ± 0.66	0.75 ± 0.73	0.74 ± 0.71	0.81 ± 0.77

IBDMDB						
Model	# of Species	Accuracy	Sensitivity	Specificity	Precision	AUC
AdaBoost	579	0.92 ± 0.02	0.97 ± 0.02	0.79 ± 0.1	0.93 ± 0.03	0.94 ± 0.01
DT	579	0.91 ± 0.02	0.94 ± 0.01	0.84 ± 0.05	0.94 ± 0.02	0.89 ± 0.02
LogitBoost	579	0.92 ± 0.01	0.98 ± 0.01	0.76 ± 0.07	0.92 ± 0.02	0.91 ± 0.04
RF	579	0.98 ± 0.01	1 ± 0	0.93 ± 0.06	0.98 ± 0.02	0.98 ± 0.01
microBiomeGSM: Family	205.76 ± 16.23	0.94 ± 0.02	0.98 ± 0.01	0.86 ± 0.05	0.93 ± 0.05	0.97 ± 0.02
microBiomeGSM: Genus	119.4 ± 15.87	0.93 ± 0.02	0.98 ± 0.01	0.85 ± 0.05	0.93 ± 0.05	0.97 ± 0.02
microBiomeGSM: Order	341.22 ± 15.6	0.93 ± 0.02	0.97 ± 0.02	0.86 ± 0.06	0.93 ± 0.06	0.98 ± 0.03

IBD						
Model	# of Species	Accuracy	Sensitivity	Specificity	Precision	AUC
AdaBoost	1,456	0.88 ± 0.04	0.85 ± 0.12	0.89 ± 0.05	0.84 ± 0.05	0.9 ± 0.04
DT	1,456	0.75 ± 0.05	0.72 ± 0.09	0.78 ± 0.06	0.67 ± 0.08	0.75 ± 0.06
LogitBoost	1,456	0.85 ± 0.04	0.81 ± 0.1	0.87 ± 0.07	0.8 ± 0.09	0.88 ± 0.04
RF	1,456	0.87 ± 0.05	0.91 ± 0.1	0.84 ± 0.05	0.78 ± 0.06	0.92 ± 0.05
microBiomeGSM: Family	260.24 ± 26.92	0.82 ± 0.06	0.85 ± 0.07	0.79 ± 0.1	0.81 ± 0.13	0.91 ± 0.07
microBiomeGSM: Genus	121.78 ± 27.83	0.81 ± 0.06	0.83 ± 0.06	0.79 ± 0.1	0.8 ± 0.12	0.88 ± 0.08
microBiomeGSM: Order	608.27 ± 24.22	0.82 ± 0.07	0.82 ± 0.08	0.81 ± 0.09	0.82 ± 0.15	0.9 ± 0.08

T2D						
Model	# of Species	Accuracy	Sensitivity	Specificity	Precision	AUC
AdaBoost	1,456	0.68 ± 0.08	0.91 ± 0.08	0.39 ± 0.26	0.67 ± 0.09	0.66 ± 0.1
DT	1,456	0.57 ± 0.05	0.98 ± 0.06	0.06 ± 0.19	0.57 ± 0.06	0.57 ± 0.09
LogitBoost	1,456	0.67 ± 0.08	0.93 ± 0.08	0.36 ± 0.24	0.65 ± 0.08	0.65 ± 0.1
RF	1,456	0.72 ± 0.09	0.91 ± 0.09	0.48 ± 0.29	0.71 ± 0.12	0.75 ± 0.1
microBiomeGSM: family	321.16 ± 36.31	0.65 ± 0.08	0.68 ± 0.09	0.63 ± 0.11	0.65 ± 0.15	0.71 ± 0.08
microBiomeGSM: genus	129.8 ± 35.03	0.64 ± 0.09	0.64 ± 0.1	0.64 ± 0.13	0.65 ± 0.18	0.69 ± 0.09
microBiomeGSM: order	596.99 ± 35.14	0.65 ± 0.08	0.65 ± 0.09	0.64 ± 0.12	0.65 ± 0.17	0.72 ± 0.09

a deeper understanding of the biological underpinnings of the disease under investigation. In essence, microBiomeGSM's value lies in its ability to contribute to the advancement of biological knowledge, rather than merely outperforming other feature selection techniques.

Table 6 shows the performance metrics of microBiomeGSM for each taxonomic level for four different datasets. The # of species column shows the number of species (features/variables) used to train and test the model. Since the number of species changes in each iteration of MCCV, we also report the standard deviation. Performance metrics are reported as the average of 100 iterations with the corresponding standard deviation. For the CRC dataset, among different classifiers the RF algorithm has the highest performance for all calculated metrics including the accuracy, sensitivity, specificity, precision, and AUC metric. The AdaBoost, LogitBoost and DT models

show lower performance compared to the RF model. The performance metrics of these three algorithms are similar but not as high as RF model. At the order taxonomic level, the mean values of the performance metrics are stable and the standard deviations are low. This indicates that the order level is a more appropriate choice for CRC classification. Comparing the RF model and the microBiomeGSM model, similar performance metrics are obtained for the CRC dataset, but it is worth mentioning that the number of features used in the proposed tool is lower. In other words, for the CRC dataset the microBiomeGSM model can accurately classify using fewer taxonomic features. For the IBDMDB dataset, among different classifiers the RF algorithm has the highest accuracy, sensitivity, specificity, precision, and AUC values. In particular, RF model achieved very high sensitivity and AUC values. For the IBDMDB dataset, the microBiomeGSM tool achieves an AUC of 98% for the order taxon level, the same performance metrics as obtained by the RF classification algorithm. However, the microBiomeGSM tool uses 341 features for the order taxon level, while the RF model uses 579 features. For IBD dataset, the RF algorithm generates the highest performance on several metrics, including accuracy, sensitivity, specificity, precision, and AUC. It performs particularly well on sensitivity and AUC. In our analysis, microBiomeGSM achieved an impressive AUC value of 91% at the family taxon level. Equally remarkable is the similar performance of the RF classification algorithm (an AUC of 92%) for the same task. However, it is important to highlight an important difference between these two approaches. For IBD dataset the RF classification algorithm achieved an AUC of 92% by using a much larger set of features (1,456 features) for the classification task. For the same dataset, the microBiomeGSM tool also showed remarkable performance (an AUC value of 91%). In stark contrast, microBiomeGSM achieved nearly equivalent AUC performance while using a much smaller set of features, only 260 features. This divergence in feature usage highlights the effectiveness and potential advantages of the microBiomeGSM tool in extracting meaningful information from microbiome data while optimizing computational resources. For T2D dataset, the RF classification algorithm outperforms other classification algorithms on several performance metrics including accuracy, sensitivity, specificity, precision and AUC. microBiomeGSM achieved an AUC value of 72% at the order taxon level. Interestingly, a similar level of performance is observed using the RF classification algorithm, which achieves an AUC value of 75%. However, it is important to note that the underlying mechanisms of these two methods are very different. The RF classification algorithm achieves this AUC value by incorporating a much larger set of features, 1,456 features, into its classification process. In contrast, the microBiomeGSM tool achieves comparable AUC metric by using a leaner set of 596 features. This difference in feature usage is worth highlighting as it shows that the microBiomeGSM tool is able to deliver competitive results with a lower computational load, making it an efficient and resource-efficient choice for the classification task at hand. These results highlight the nuanced trade-offs in selecting the appropriate tool or algorithm for the specific data analysis requirements.

As shown in Table 7, the performance of our proposed method varies depending on the taxonomic level considered. For the order taxonomic level, for all tested datasets, the proposed method outperforms other models in terms of the AUC score, except for the RF classifier. Similarly, for all datasets, at the family and genus taxonomic levels, the AUC values are also highly competitive, outperforming those of the other four machine learning algorithms used in this study, with the sole exception of the RF classifier. These results highlight the robust performance of our method across different taxonomic levels. A remarkable performance of our proposed method was observed when it is applied on the IBDMDB dataset. Here, we obtained an exceptionally high AUC value of 0.98 ± 0.03 at the order taxonomic level using a 100-fold MCCV approach. This remarkable result demonstrates the exceptional performance and the potential of the microBiomeGSM tool.

4 Discussion

The microbiome is considered as a crucial component of the human body and it is increasingly associated with numerous aspects of development and health. There is growing evidence that the microbiota is essential for understanding, diagnosing, and treating human diseases. In particular, alterations in the gut microbiome community have been linked to a variety of diseases, including CRC (Song et al., 2020), T2D (Salamon et al., 2018) and IBD (Alam et al., 2020). Several research efforts relied on sample-level feature abundance data to identify predictive microbiome biomarkers using machine learning. In this study, we proposed to perform more effective disease classification and prediction with fewer features. To this end, we developed microBiomeGSM to solve this problem compared to tools that perform predictions with a large amount of data. The success of microBiomeGSM can be explained with the following features of the G-S-M approach:

- For the grouping component of microBiomeGSM, only the features at the similar taxonomic levels are considered.
- microBiomeGSM uses efficient classifiers for the scoring component to identify the key groups for each taxonomic level;
- For the modeling component, significant taxonomic groups are considered cumulatively using effective classifiers.

Via analyzing metagenomic data, this study aims to solve the problem of disease diagnosis using existing taxonomic knowledge; and finally introduces a tool called microBiomeGSM. The proposed tool is based on the G-S-M (Grouping-Scoring-Modeling) approach and uses species-level information by grouping taxonomic features at different taxonomic levels such as genus, family, and order. The performance of microBiomeGSM on four different disease-associated metagenomic datasets was evaluated in comparison to other feature selection methods such as Fast Correlation Based Filter (FCBF), Select Best K (SKB), Extreme Gradient Boosting (XGB), Conditional Mutual Information Maximization (CMIM), Maximum Likelihood and Minimum Redundancy (MRMR), and Information Gain (IG).

The presented microBiomeGSM approach offers several advantages in the field of disease diagnosis via analyzing metagenomic datasets. One significant benefit is its ability to efficiently identify disease-associated taxonomic biomarkers through a robust machine learning model based on the Grouping, Scoring, and Modeling (G-S-M) methodology. Differently from existing approaches, microBiomeGSM identifies groups of important taxons and detects important species within that taxon for the disease under study. Hence, this innovative approach enables the extraction of valuable insights from microbiome data, shedding light on the influence of

TABLE 7 Comparative performance evaluation of microBiomeGSM and other machine learning approaches for different microbiome datasets.

Dataset		AdaBoost	DT	LogitBoost	RF	microBiomeGSM: family	microBiomeGSM: genus	microBiomeGSM: order
CRC	AUC	0.78 ± 0.14	0.70 ± 0.04	0.78 ± 0.04	**0.86** ± 0.03	0.81 ± 0.67	0.80 ± 0.68	**0.81** ± 0.77
	# of Species	912	912	912	912	292.88 ± 16.09	161.21 ± 5.17	607.5 ± 188.32
IBDMDB	AUC	0.94 ± 0.01	0.89 ± 0.02	0.91 ± 0.04	**0.98** ± 0.01	0.97 ± 0.02	0.97 ± 0.03	**0.98 ± 0.03**
	# of Species	579	579	579	579	205.76 ± 16.23	119.4 ± 15.87	341.22 ± 15.6
IBD	AUC	0.9 ± 0.04	0.75 ± 0.06	0.88 ± 0.04	**0.92** ± 0.05	**0.91** ± 0.07	0.88 ± 0.08	0.9 ± 0.08
	# of Species	1,456	1,456	1,456	1,456	260.24 ± 26.92	121.78 ± 27.83	608.27 ± 24.22
T2D	AUC	0.66 ± 0.1	0.57 ± 0.09	0.65 ± 0.1	**0.75** ± 0.1	0.71 ± 0.08	0.69 ± 0.09	**0.72** ± 0.09
	# of Species	1,456	1,456	1,456	1,456	321.16 ± 36.31	129.8 ± 35.03	596.99 ± 35.14

The results in bold in table represent the best AUC result among the 4 different traditional classifiers (AdaBoost, DT, LogitBoost and RF) and the best AUC result across taxon levels using microBiomeGSM.

specific taxonomic biomarkers on the disease under investigation. Furthermore, the performance evaluation across different diseases, different taxonomic levels (genus, family, order); and the comparative assessment with different feature selection algorithms exhibits the reliability of microBiomeGSM. Finally, the discussions on the biological relevance of the findings of the proposed approach, via drawing evidence from the existing literature, provide valuable context for the identified taxon groups for the disease under study, making microBiomeGSM an informative tool in disease research. Our tool's significance transcends its mere application; it holds the potential for pioneering discoveries. It is geared to discern not isolated microbial entities but entire assemblages of species, paving the way for profound biological interpretations. By spotlighting groups of bacteria and viruses in lieu of singular entities, our tool offers a holistic view, potentially identifying microbial communities implicated in specific diseases.

With this study, we would also like to motivate biologists and the microbiome community to redesign their grouping methods instead of using individual feature selection approaches. We envision that in the future, various biological datasets, including multi-omics, will be used to redefine the groupings. Such innovative grouping strategies, complemented by modeling, promise to provide profound insights into the molecular mechanisms of diseases and the role of microorganisms in disease development.

4.1 Biological interpretations of microBiomeGSM's findings

This section discusses the biological relevance of the features discovered by microBiomeGSM at different taxonomic levels for all tested datasets. T2D is a metabolic disease characterized by high glucose levels in blood and caused primarily by cellular resistance to the activity of insulin (Sedighi et al., 2017). There are several studies in the literature that have demonstrated the relation of different microorganisms at the genus, family, and order levels with T2D development. For the T2D dataset, the top 10 microbiomes identified by our method at the genus, family, order levels and the relevant literature can be summarized in Supplementary Table S15. On the other hand, inflammatory bowel diseases (IBDs), which include primarily ulcerative colitis and Crohn's disease, but also non-infectious inflammation of the bowel, have puzzled gastroenterologists and immunologists alike since their first modern descriptions around some 75–100 years ago (Ni et al., 2018; Bakir-Gungor et al., 2022). For the IBDMDB dataset, the top 10 microbiomes identified by our method at the genus, family, and order levels and the relevant literature can be summarized in Supplementary Table S15. CRC is a prevalent malignancy affecting the colon and rectum. It constitutes approximately 10% of all newly diagnosed cancer cases worldwide (Li X. et al., 2023). For the CRC dataset, the top 10 microbiomes identified by our method at the genus, family, and order levels and the relevant literature can be summarized in Supplementary Table S15.

Numerous studies have investigated the relationship between microbiomes and diseases like T2D, CRC, and IBD using similar datasets as used within this study. Upon examination of these studies, it becomes evident that while their experimental designs may vary, they consistently yield comparable results when it comes to identifying microbiomes linked to these diseases. These findings align with the

important microbiomes identified by microBiomeGSM for T2D, CRC, and IBD, showcasing the tool's effectiveness in accurately identifying relevant microbiomes associated with these diseases. These congruent findings reinforce the reliability and validity of the microbiome associations detected by the microBiomeGSM tool. It also underscores the tool's capacity to identify microbiomes that are consistently linked to specific diseases, providing valuable insights for disease characterization and prediction. Hassouneh et al. (2021) conducted a series of experiments aimed at uncovering microbiomes associated with IBD. In their analysis using the same dataset as used by the microBiomeGSM tool, they observed differences in Clostridium microbiota among IBD patients. Additionally, another microbiome identified for IBD in their study is *Ruminococcus*. Remarkably, these microbiomes align with the important microbiomes detected for the IBD disease by the microBiomeGSM tool. This correspondence in findings highlights the capacity of microBiomeGSM in identifying relevant microbiomes linked to IBD. Zhang Y. et al. (2022) conducted a study with the goal of identifying disease-associated microbiome species for Inflammatory Bowel Disease Microbiome Database (IBDMDB), employing the same dataset (PRJNA289734) as used in microBiomeGSM. In their research, they highlighted the significance of the Bacteroides microbiome. Interestingly, the Bacteroides microbiome is also identified as one of the important microbiomes by the microBiomeGSM tool proposed in our study. This alignment in findings underscores the effectiveness of microBiomeGSM in recognizing key microbiomes associated with diseases like IBD. Bai et al. (2022) conducted a series of experiments aimed at identifying microbiomes associated with T2D. In their research, they utilized the SRA4565 data for T2D and highlighted the significance of the methanobacteriales microbiome. Notably, methanobacteriales is among the top 10 microbiomes identified by the proposed microBiomeGSM tool. This convergence of findings underscores the effectiveness and utility of the proposed tool in uncovering microbiome associations with diseases like T2D. Forslund et al. (2015) conducted experiments utilizing the same T2D dataset employed by microBiomeGSM to investigate microbiomes associated with T2D. Upon close examination of their experiments, they underscored the significance of the Clostridiales microbiome in relation to T2D disease. Interestingly, Clostridiales also emerges as one of the important microbiomes identified by microBiomeGSM. This convergence in findings highlights the relevance and effectiveness of microBiomeGSM in identifying crucial microbiomes associated with T2D. Ma et al. (2021) conducted a study that investigated the microbiomes associated with CRC using the same dataset as in our study. Among the various microbiomes they examined, the Prevotella microbiome stood out as strongly linked to CRC. This association aligns with the findings of microBiomeGSM, underscoring the significance of the Prevotella microbiome in the context of characterizing CRC. Chen et al. (2023) conducted research using the same dataset to investigate microbiomes in the context of colorectal cancer, akin to the proposed microBiomeGSM tool. Similar to the findings of microBiomeGSM, their study also identified Peptostreptococcus, Fusobacterium, and Porphyromonas microbiomes as valuable and effective biomarkers for CRC. This convergence in results underscores the potential significance of these specific microbiomes in CRC characterization and their importance as potential biomarkers for the disease.

In summary, via analyzing the raw microbiome data of specific diseases, this study aims to identify taxonomic biomarkers that may have a role in the associated diseases. Three different taxon levels (genus, family, and order) are studied and disease prediction is performed by building effective machine learning models using the G-S-M approach. Four different datasets are analyzed and the identified microorganisms at genus, family and order levels are compared with the existing literature.

4.2 Limitation of the study

The quality and the scope of our study have been significantly influenced by several primary limiting factors. These factors encompass the nature of the data set, the tools employed for data preprocessing, the specific taxon groups considered, and the overall volume of data under examination. First and foremost, the data set itself plays a pivotal role in shaping the outcomes and conclusions of our study. Its size, diversity, and representativeness directly impact the generalizability of our findings. Furthermore, the quality of data, its sources, and any potential biases within the dataset significantly affect the reliability of our results. Equally significant is the role of the tools employed for data preprocessing. The choices made in data cleaning, feature selection, and data transformation can introduce variability and influence the robustness of our analytical pipeline. It is paramount to acknowledge how these preprocessing steps can shape the study's outcomes. Additionally, our study's focus on specific taxon groups within the dataset should be considered. The selection of these taxonomic levels and the criteria used for their inclusion or exclusion has bearing on the granularity and relevance of our findings. Finally, the number of data points utilized in our analysis is another crucial factor. A larger dataset provides a broader and potentially more representative sample, which can enhance the reliability and statistical power of our results. Conversely, a smaller dataset may limit the generalizability of our conclusions. A comprehensive understanding of these limiting factors is essential for contextualizing our study's outcomes and conclusions.

5 Conclusion

Over the past two decades, the number of microbiome studies has increased rapidly thanks to the advances in next generation sequencing (NGS) technologies. Lower costs and increasing computational power have enabled us to obtain enormous amounts of data on the diversity and function of a host or habitat's microbiome. Identifying and accounting for effective taxons in microbiome and disease classification can accelerate disease diagnosis, prognosis, and treatment. Here, we use an efficient machine learning model to identify taxonomic biomarkers that can diagnose diseases. The microBiomeGSM enables researchers to explore the diversity of contributions to disease development by examining metagenomic data at different taxonomic levels. While analyzing microbiome datasets, the microBiomeGSM tool that we present in this study exploits the existing biological knowledge about the taxonomic hierarchy of the species at different levels, such as genus, family, and order. Our results showed that via analyzing different microbiome datasets associated with different diseases, microBiomeGSM builds

effective machine learning models to facilitate the diagnosis of diseases. It is anticipated that this study will be a guide for future studies and will guide and improve the studies to be conducted on this topic. With this study, we hope to highlight the importance of taxonomic groups in microbiome-based disease prediction and to facilitate the diagnosis of disease using these taxonomic groups.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding authors.

Author contributions

BB-G: Methodology, Software, Writing – review & editing, Project administration, Supervision. MT: Methodology, Software, Writing – original draft, Writing – review & editing, Investigation, Visualization. AJ: Data curation, Formal analysis, Methodology, Software, Writing – original draft. DW: Investigation, Methodology, Supervision, Writing – original draft, Project administration, Writing – review & editing. MY: Formal analysis, Funding acquisition, Methodology, Project administration, Software, Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. The work of BB-G has been supported by the L'Oréal-UNESCO Young Women Scientist Program and by the Abdullah Gul University Support Foundation (AGUV). The work of MY has been supported by the Zefat Academic College. This article is based upon work from COST Action ML4Microbiome (CA18131), supported by COST (European Cooperation in Science and Technology), www.cost.eu, which has played a pivotal role in advancing microbiome research and facilitating the expansion of these research endeavours.

Acknowledgments

We extend our gratitude to COST ML4Microbiome Action for the funding, which has played a pivotal role in advancing microbiome research and facilitating the expansion of these research endeavors. This research was made possible by the generous support of the L'Oréal-UNESCO Young Women Scientist Program. BB-G would like to express her gratitude for the L'Oréal-UNESCO Young Women Scientist Award, received in 2022.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2023.1264941/full#supplementary-material

References

Alam, M. T., Amos, G. C. A., Murphy, A. R. J., Murch, S., Wellington, E. M. H., and Arasaradnam, R. P. (2020). Microbial imbalance in inflammatory bowel disease patients at different taxonomic levels. *Gut Pathog.* 12:1. doi: 10.1186/s13099-019-0341-6

Alatawi, H., Mosli, M., Saadah, O. I., Annese, V., al-Hindi, R., Alatawy, M., et al. (2022). Attributes of intestinal microbiota composition and their correlation with clinical primary non-response to anti-TNF-α agents in inflammatory bowel disease patients. *Biomol. Biomed.* 22, 412–426. doi: 10.17305/bjbms.2021.6436

Bai, X., Sun, Y., Li, Y., Li, M., Cao, Z., Huang, Z., et al. (2022). Landscape of the gut archaeome in association with geography, ethnicity, urbanization, and diet in the Chinese population. *Microbiome* 10:147. Available at:. doi: 10.1186/s40168-022-01335-7

Bakir-Gungor, B., Bulut, O., Jabeer, A., Nalbantoglu, O. U., and Yousef, M. (2021). Discovering potential taxonomic biomarkers of type 2 diabetes from human gut microbiota via different feature selection methods. *Front. Microbiol.* 12:426. doi: 10.3389/fmicb.2021.628426

Bakir-Gungor, B., Hacılar, H., Jabeer, A., Nalbantoglu, O. U., Aran, O., and Yousef, M. (2022). Inflammatory bowel disease biomarkers of human gut microbiota selected via different feature selection methods. *PeerJ* 10:e13205. doi: 10.7717/peerj.13205

Beghini, F., McIver, L., Blanco-Míguez, A., Dubois, L., Asnicar, F., Maharjan, S., et al. (2021). Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *Elife* 10:e65088. doi: 10.7554/eLife.65088

Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., et al. (2009). KNIME–the Konstanz information miner: version 2.0 and beyond. *ACM SIGKDD Explor. Newsl.* 11, 26–31. doi: 10.1145/1656274.1656280

Cena, J. A., Reis, L. G., de Lima, A. K. A., Vieira Lima, C. P., Stefani, C. M., and Dame-Teixeira, N. (2023). Enrichment of acid-associated microbiota in the saliva of type 2 diabetes mellitus adults: a systematic review. *Pathogens* 12:404. Available at:. doi: 10.3390/pathogens12030404

Chen, F., Li, S., Guo, R., Song, F., Zhang, Y., Wang, X., et al. (2023). Meta-analysis of fecal viromes demonstrates high diagnostic potential of the gut viral signatures for colorectal cancer and adenoma risk assessment. *J. Adv. Res.* 49, 103–114. Available at:. doi: 10.1016/j.jare.2022.09.012

Deschênes, T., Tohoundjona, F. W. E., Plante, P. L., di Marzo, V., and Raymond, F. (2023). Gene-based microbiome representation enhances host phenotype classification. *mSystems* 8:e0053123. doi: 10.1128/msystems.00531-23

Ding, C., and Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. *J. Bioinforma. Comput. Biol.* 3, 185–205. doi: 10.1142/S0219720005001004

Ditzler, G., Polikar, R., and Rosen, G. (2015). Multi-layer and recursive neural networks for metagenomic classification. *IEEE Trans. Nanobioscience* 14, 608–616. doi: 10.1109/TNB.2015.2461219

Dix, A., Vlaic, S., Guthke, R., and Linde, J. (2016). Use of systems biology to decipher host–pathogen interaction networks and predict biomarkers. *Clin. Microbiol. Infect.* 22, 600–606. doi: 10.1016/j.cmi.2016.04.014

Duvallet, C., Gibbons, S. M., Gurry, T., Irizarry, R. A., and Alm, E. J. (2017). 'Meta-analysis of gut microbiome studies identifies disease-specific and shared responses', nature. *Communications* 8:1784. doi: 10.1038/s41467-017-01973-8

Ersoz, N. S., Bakir-Gungor, B., and Yousef, M. (2023). GeNetOntology: identifying affected gene ontology groups via grouping, scoring and modelling from gene expression data utilizing biological knowledge based machine learning. *Front. Genet.* 14:82. doi: 10.3389/fgene.2023.1139082

Fleuret, F., and Ch, E. (2004). Fast binary feature selection with conditional mutual information. *J. Mach. Learn. Res.* 5, 1531–1555.

Forslund, K., Hildebrand, F., Nielsen, T., Falony, G., Le Chatelier, E., Sunagawa, S., et al. (2015). Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. *Nature* 528, 262–266. doi: 10.1038/nature15766

Fritz, J. V., Desai, M. S., Shah, P., Schneider, J. G., and Wilmes, P. (2013). From meta-omics to causality: experimental models for human microbiome research. *Microbiome* 1:14. doi: 10.1186/2049-2618-1-14

Gao, R., Zhu, C., Li, H., Yin, M., Pan, C., Huang, L., et al. (2018). Dysbiosis signatures of gut microbiota along the sequence from healthy, young patients to those with overweight and obesity. *Obesity* 26, 351–361. doi: 10.1002/oby.22088

Giliberti, R., Cavaliere, S., Mauriello, I. E., Ercolini, D., and Pasolli, E. (2022). Host phenotype classification from human microbiome data is mainly driven by the presence of microbial taxa. *PLoS Comput. Biol.* 18:e1010066. doi: 10.1371/journal.pcbi.1010066

Gurung, M., Li, Z., You, H., Rodrigues, R., Jump, D. B., Morgun, A., et al. (2020). Role of gut microbiota in type 2 diabetes pathophysiology. *EBioMedicine* 51:51. doi: 10.1016/j.ebiom.2019.11.051

Hassouneh, S. A.-D., Loftus, M., and Yooseph, S. (2021). Linking inflammatory bowel disease symptoms to changes in the gut microbiome structure and function. *Front. Microbiol.* 12:632. doi: 10.3389/fmicb.2021.673632

Hsu, M., Tun, K. M., Batra, K., Haque, L., Vongsavath, T., and Hong, A. S. (2023). Safety and efficacy of fecal microbiota transplantation in treatment of inflammatory bowel disease in the pediatric population: a systematic review and Meta-analysis. *Microorganisms* 11:1272. doi: 10.3390/microorganisms11051272

Huybrechts, I., Zouiouich, S., Loobuyck, A., Vandenbulcke, Z., Vogtmann, E., Pisanu, S., et al. (2020). The human microbiome in relation to Cancer risk: a systematic review of epidemiologic studies. *Cancer Epidemiol. Biomark. Prev.* 29, 1856–1868. doi: 10.1158/1055-9965.EPI-20-0288

Jabeer, A., KoÇak, A., Akkaş, H., Yenisert, F., Nalbantoğlu, O. U., Yousef, M., et al. (2022). Identifying Taxonomic Biomarkers of Colorectal Cancer in Human Intestinal Microbiota Using Multiple Feature Selection Methods', in 2022 Innovations in Intelligent Systems and Applications Conference (ASYU). *IEEE* 2022, 1–6. doi: 10.1109/ASYU56188.2022.9925551

Jabeer, A., Temiz, M., Bakir-Gungor, B., and Yousef, M. (2023). miRdisNET: discovering microRNA biomarkers that are associated with diseases utilizing biological knowledge-based machine learning. *Front. Genet.* 13:1076554. doi: 10.3389/fgene.2022.1076554

Kolde, R., Laur, S., Adler, P., and Vilo, J. (2012). Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics* 28, 573–580. doi: 10.1093/bioinformatics/btr709

Kuzudisli, C., Bakir-Gungor, B., Bulut, N., Qaqish, B., and Yousef, M. (2023). Review of feature selection approaches based on grouping of features. *PeerJ* 11:e15666. doi: 10.7717/peerj.15666

LaPierre, N., Ju, C. J. T., Zhou, G., and Wang, W. (2019). MetaPheno: a critical evaluation of deep learning and machine learning in metagenome-based disease prediction. *Methods* 166, 74–82. doi: 10.1016/j.ymeth.2019.03.003

Levy, S. E., and Myers, R. M. (2016). Advancements in next-generation sequencing. *Annu. Rev. Genomics Hum. Genet.* 17, 95–115. doi: 10.1146/annurev-genom-083115-022413

Li, X., Feng, J., Wang, Z., Liu, G., and Wang, F. (2023). Features of combined gut bacteria and fungi from a Chinese cohort of colorectal cancer, colorectal adenoma, and post-operative patients. *Front. Microbiol.* 14:583. doi: 10.3389/fmicb.2023.1236583

Li, R., Shokri, F., Rincon, A. L., Rivadeneira, F., Medina-Gomez, C., and Ahmadizar, F. (2023). Bi-directional interactions between glucose-lowering medications and gut microbiome in patients with type 2 diabetes mellitus: a systematic review. *Genes* 14:1572. doi: 10.3390/genes14081572

Lim, H., Cankara, F., Tsai, C. J., Keskin, O., Nussinov, R., and Gursoy, A. (2022). Artificial intelligence approaches to human-microbiome protein–protein interactions. *Curr. Opin. Struct. Biol.* 73:102328. doi: 10.1016/j.sbi.2022.102328

Ma, Y., Zhang, Y., Xiang, J., Xiang, S., Zhao, Y., Xiao, M., et al. (2021). Metagenome analysis of intestinal Bacteria in healthy people, patients with inflammatory bowel disease and colorectal Cancer. *Front. Cell. Infect. Microbiol.* 11:734. doi: 10.3389/fcimb.2021.599734

Mah, C., Jayawardana, T., Leong, G., Koentgen, S., Lemberg, D., Connor, S. J., et al. (2023). Assessing the relationship between the gut microbiota and inflammatory bowel disease therapeutics: a systematic review. *Pathogens* 12:262. doi: 10.3390/pathogens12020262

Marco-Ramell, A., Palau-Rodriguez, M., Alay, A., Tulipani, S., Urpi-Sarda, M., Sanchez-Pla, A., et al. (2018). Evaluation and comparison of bioinformatic tools for the enrichment analysis of metabolomics data. *BMC Bioinformatics* 19:1. doi: 10.1186/s12859-017-2006-0

Marcos-Zambrano, L. J., Karaduzovic-Hadziabdic, K., Loncar Turukalo, T., Przymus, P., Trajkovik, V., Aasmets, O., et al. (2021). Applications of machine learning in human microbiome studies: a review on feature selection, biomarker identification, disease prediction and treatment. *Front. Microbiol.* 12:511. doi: 10.3389/fmicb.2021.634511

Martin, A. M., Yabut, J. M., Choo, J. M., Page, A. J., Sun, E. W., Jessup, C. F., et al. (2019). The gut microbiome regulates host glucose homeostasis via peripheral serotonin. *Proc. Natl. Acad. Sci.* 116, 19802–19804. doi: 10.1073/pnas.1909311116

McDonald, D., Hyde, E., Debelius, J. W., Morton, J. T., Gonzalez, A., Ackermann, G., et al. (2018). American gut: an open platform for citizen science microbiome research. *mSystems* 3:e00031. doi: 10.1128/mSystems.00031-18

Mendes, V., Galvão, I., and Vieira, A. T. (2019). Mechanisms by which the gut microbiota influences cytokine production and modulates host inflammatory responses. *J. Interf. Cytokine Res.* 39, 393–409. doi: 10.1089/jir.2019.0011

Muller, E. E. L. (2019). Determining microbial niche breadth in the environment for better ecosystem fate predictions. *mSystems* 4:19. doi: 10.1128/msystems.00080-19

Negrut, R. L., Cote, A., and Maghiar, A. M. (2023). Exploring the potential of Oral microbiome biomarkers for colorectal Cancer diagnosis and prognosis: a systematic review. *Microorganisms* 11:1586. doi: 10.3390/microorganisms11061586

Ni, Y., Mu, C., He, X., Zheng, K., Guo, H., and Zhu, W. (2018). Characteristics of gut microbiota and its response to a Chinese herbal formula in elder patients with metabolic syndrome. *Drug Discov. Ther.* 12, 161–169. doi: 10.5582/ddt.2018.01036

Ohland, C. L., and Jobin, C. (2015). 'Microbial activities and intestinal homeostasis: a delicate balance between health and disease', cellular and molecular. *Gastroenterol. Hepatol.* 1, 28–40. doi: 10.1016/j.jcmgh.2014.11.004

Oudah, M., and Henschel, A. (2018). Taxonomy-aware feature engineering for microbiome classification. *BMC Bioinformatics* 19:227. doi: 10.1186/s12859-018-2205-3

Pasolli, E., Truong, D. T., Malik, F., Waldron, L., and Segata, N. (2016). Machine learning Meta-analysis of large metagenomic datasets: tools and biological insights. *PLoS Comput. Biol.* 12:e1004977. doi: 10.1371/journal.pcbi.1004977

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). '*Scikit-learn: Machine learning in Python*', *Machine Learning in Python*.

Petersen, C., and Round, J. L. (2014). Defining dysbiosis and its influence on host immunity and disease. *Cell. Microbiol.* 16, 1024–1033. doi: 10.1111/cmi.12308

Pickard, J. M., Zeng, M. Y., Caruso, R., and Núñez, G. (2017). Gut microbiota: role in pathogen colonization, immune responses, and inflammatory disease. *Immunol. Rev.* 279, 70–89. doi: 10.1111/imr.12567

Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* 490, 55–60. doi: 10.1038/nature11450

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464, 59–65. doi: 10.1038/nature08821

Qin, N., Yang, F., Li, A., Prifti, E., Chen, Y., Shao, L., et al. (2014). Alterations of the human gut microbiome in liver cirrhosis. *Nature* 513, 59–64. doi: 10.1038/nature13568

Qumsiyeh, E., Showe, L., and Yousef, M. (2022). GediNET for discovering gene associations across diseases using knowledge based machine learning approach. *Sci. Rep.* 12:19955. doi: 10.1038/s41598-022-24421-0

Salamon, D., Sroka-Oleksiak, A., Kapusta, P., Szopa, M., Mrozińska, S., Ludwig-Słomczyńska, A. H., et al. (2018). Characteristics of the gut microbiota in adult patients with type 1 and 2 diabetes based on the analysis of a fragment of 16S rRNA gene using next-generation sequencing. *Pol. Arch. Intern. Med.* 128, 336–343. doi: 10.20452/pamw.4246

Sedighi, M., Razavi, S., Navab-Moghadam, F., Khamseh, M. E., Alaei-Shahmiri, F., Mehrtash, A., et al. (2017). Comparison of gut microbiota in adult patients with type 2 diabetes and healthy individuals. *Microb. Pathog.* 111, 362–369. doi: 10.1016/j.micpath.2017.08.038

Senliol, B., Gulgezen, G., Yu, L., and Cataltepe, Z.. (2008) *Fast correlation based filter (FCBF) with a different search strategy*. In: 2008 23rd international symposium on computer and information sciences. 2008 23rd international symposium on computer and information sciences, pp. 1–4.

Sharma, D., Paterson, A. D., and Xu, W. (2020). TaxoNN: ensemble of neural networks on stratified microbiome data for disease prediction. *Bioinformatics* 36, 4544–4550. doi: 10.1093/bioinformatics/btaa542

Song, M., Chan, A. T., and Sun, J. (2020). Influence of the gut microbiome, diet, and environment on risk of colorectal Cancer. *Gastroenterology* 158, 322–340. doi: 10.1053/j.gastro.2019.06.048

Soueidan, H., and Nikolski, M. (2016). Machine learning for metagenomics: methods and tools. *arXiv* 2016:621. doi: 10.48550/arXiv.1510.06621

Tabowei, G., Gaddipati, G. N., Mukhtar, M., Alzubaidee, M. J., Dwarampudi, R. S., Mathew, S., et al. (2022). Microbiota Dysbiosis a cause of colorectal Cancer or not? A systematic review. *Cureus* 14, 14:e30893. doi: 10.7759/cureus.30893

The Human Microbiome Project Consortium (2012). Structure, function and diversity of the healthy human microbiome. *Nature* 486, 207–214. doi: 10.1038/nature11234

Thomas, A. M., Manghi, P., Asnicar, F., Pasolli, E., Armanini, F., Zolfo, M., et al. (2019). Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nat. Med.* 25, 667–678. doi: 10.1038/s41591-019-0405-7

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Series B (Methodological)* 58, 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x

Topçuoğlu, B. D., Lesniak, N. A., Ruffin, M. T. IV, Wiens, J., and Schloss, P. D. (2020). A framework for effective application of machine learning to microbiome-based classification problems. *MBio* 11, e00434–e00420. doi: 10.1128/mBio.00434-20

Unlu Yazici, M., Marron, J. S., Bakir-Gungor, B., Zou, F., and Yousef, M. (2023). Invention of 3Mint for feature grouping and scoring in multi-omics. *Front. Genet.* 14:1093326. doi: 10.3389/fgene.2023.1093326

Wang, X.-W., and Liu, Y.-Y. (2020). Comparative study of classifiers for human microbiome data. *Med. Microecol.* 4:100013. doi: 10.1016/j.medmic.2020.100013

Yousef, M., Abdallah, L., and Allmer, J. (2019). maTE: discovering expressed interactions between microRNAs and their targets. *Bioinformatics* 35, 4020–4028. doi: 10.1093/bioinformatics/btz204

Yousef, M., Goy, G., and Bakir-Gungor, B. (2022b). miRModuleNet: detecting miRNA-mRNA regulatory modules. *Front. Genet.* 13:455. doi: 10.3389/fgene.2022.767455

Yousef, M., Goy, G., Mitra, R., Eischen, C. M., Jabeer, A., and Bakir-Gungor, B. (2021a). miRcorrNet: machine learning-based integration of miRNA and mRNA expression profiles, combined with feature grouping and ranking. *PeerJ* 9:e11458. doi: 10.7717/peerj.11458

Yousef, M., Kumar, A., and Bakir-Gungor, B. (2021b). Application of biological domain knowledge based feature selection on gene expression data. *Entropy* 23:2. doi: 10.3390/e23010002

Yousef, M., Ozdemir, F., Jaber, A., Allmer, J., and Bakir-Gungor, B. (2022a). PriPath: Identifying dysregulated pathways from differential gene expression via grouping, scoring and modeling with an embedded machine learning approach. *BMC Bioinformatics* 24:60. doi: 10.21203/rs.3.rs-1449467/v1

Yousef, M., Ülgen, E., and Uğur Sezerman, O. (2021c). CogNet: classification of gene expression data based on ranked active-subnetwork-oriented KEGG pathway enrichment analysis. *PeerJ Comput. Sci.* 7:e336. doi: 10.7717/peerj-cs.336

Yousef, M., and Voskergian, D. (2022). TextNetTopics: text classification based word grouping as topics and topics' scoring. *Front. Genet.* 13:893378. doi: 10.3389/fgene.2022.893378

Zhang, Y., Bhosle, A., Bae, S., McIver, L. J., Pishchany, G., Accorsi, E. K., et al. (2022). Discovery of bioactive microbial gene products in inflammatory bowel disease. *Nature* 606, 754–760. doi: 10.1038/s41586-022-04648-7

Zhang, W., Liu, A., Zhang, Z., Chen, G., and Li, Q. (2022). An adaptive direction-assisted test for microbiome compositional data. *Bioinformatics* 38, 3493–3500. doi: 10.1093/bioinformatics/btac361

Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Series B (Statistical Methodology)* 67, 301–320. doi: 10.1111/j.1467-9868.2005.00503.x

Zwezerijnen-Jiwa, F. H., Sivov, H., Paizs, P., Zafeiropoulou, K., and Kinross, J. (2023). A systematic review of microbiome-derived biomarkers for early colorectal cancer detection. *Neoplasia* 36:100868. doi: 10.1016/j.neo.2022.100868

frontiers | Frontiers in Microbiology

TYPE Review
PUBLISHED 22 November 2023
DOI 10.3389/fmicb.2023.1250806

OPEN ACCESS

EDITED BY
Anastasis Oulas,
The Cyprus Institute of Neurology and Genetics, Cyprus

REVIEWED BY
Yu-Wei Wu,
Taipei Medical University, Taiwan
Sergio Peignier,
Institut National des Sciences Appliquées de Lyon (INSA Lyon), France

*CORRESPONDENCE
Laura Judith Marcos-Zambrano
judith.marcos@imdea.org
Enrique Carrillo de Santa Pau
enrique.carrillo@imdea.org

†These authors have contributed equally to this work

RECEIVED 30 June 2023
ACCEPTED 11 September 2023
PUBLISHED 22 November 2023

CITATION
Marcos-Zambrano LJ, López-Molina VM, Bakir-Gungor B, Frohme M, Karaduzovic-Hadziabdic K, Klammsteiner T, Ibrahimi E, Lahti L, Loncar-Turukalo T, Dhamo X, Simeon A, Nechyporenko A, Pio G, Przymus P, Sampri A, Trajkovik V, Lacruz-Pleguezuelos B, Aasmets O, Araujo R, Anagnostopoulos I, Aydemir Ö, Berland M, Calle ML, Ceci M, Duman H, Gündoğdu A, Havulinna AS, Kaka Bra KHN, Kalluci E, Karav S, Lode D, Lopes MB, May P, Nap B, Nedyalkova M, Paciência I, Pasic L, Pujolassos M, Shigdel R, Susín A, Thiele I, Truică C-O, Wilmes P, Yilmaz E, Yousef M, Claesson MJ, Truu J and Carrillo de Santa Pau E (2023) A toolbox of machine learning software to support microbiome analysis.
Front. Microbiol. 14:1250806.
doi: 10.3389/fmicb.2023.1250806

A toolbox of machine learning software to support microbiome analysis

Laura Judith Marcos-Zambrano[1*], Víctor Manuel López-Molina[1], Burcu Bakir-Gungor[2†], Marcus Frohme[3†], Kanita Karaduzovic-Hadziabdic[4†], Thomas Klammsteiner[5†], Eliana Ibrahimi[6†], Leo Lahti[7†], Tatjana Loncar-Turukalo[8†], Xhilda Dhamo[9†], Andrea Simeon[10†], Alina Nechyporenko[3,11†], Gianvito Pio[12,13†], Piotr Przymus[14†], Alexia Sampri[15†], Vladimir Trajkovik[16†], Blanca Lacruz-Pleguezuelos[1], Oliver Aasmets[17,18], Ricardo Araujo[19], Ioannis Anagnostopoulos[20,21], Önder Aydemir[22], Magali Berland[23], M. Luz Calle[24,25], Michelangelo Ceci[12,13], Hatice Duman[26], Aycan Gündoğdu[27,28], Aki S. Havulinna[29,30], Kardokh Hama Najib Kaka Bra[31], Eglantina Kalluci[9], Sercan Karav[32], Daniel Lode[3], Marta B. Lopes[33,34], Patrick May[35], Bram Nap[36], Miroslava Nedyalkova[37], Inês Paciência[38,39], Lejla Pasic[40], Meritxell Pujolassos[24], Rajesh Shigdel[41], Antonio Susín[42], Ines Thiele[36,43], Ciprian-Octavian Truică[44], Paul Wilmes[45,46], Ercument Yilmaz[47], Malik Yousef[48,49], Marcus Joakim Claesson[43,50], Jaak Truu[31] and Enrique Carrillo de Santa Pau[1*] on behalf of ML4Microbiome

[1]Computational Biology Group, Precision Nutrition and Cancer Research Program, IMDEA Food Institute, Madrid, Spain, [2]Department of Computer Engineering, Abdullah Gül University, Kayseri, Türkiye, [3]Division Molecular Biotechnology and Functional Genomics, Technical University of Applied Sciences Wildau, Wildau, Germany, [4]Faculty of Engineering and Natural Sciences, International University of Sarajevo, Sarajevo, Bosnia and Herzegovina, [5]Department of Microbiology and Department of Ecology, University of Innsbruck, Innsbruck, Austria, [6]Department of Biology, University of Tirana, Tirana, Albania, [7]Department of Computing, University of Turku, Turku, Finland, [8]Faculty of Technical Sciences, University of Novi Sad, Novi Sad, Serbia, [9]Department of Applied Mathematics, Faculty of Natural Sciences, University of Tirana, Tirana, Albania, [10]BioSense Institute, University of Novi Sad, Novi Sad, Serbia, [11]Department of Systems Engineering, Kharkiv National University of Radioelectronics, Kharkiv, Ukraine, [12]Department of Computer Science, University of Bari Aldo Moro, Bari, Italy, [13]Big Data Lab, National Interuniversity Consortium for Informatics, Rome, Italy, [14]Faculty of Mathematics and Computer Science, Nicolaus Copernicus University, Toruń, Poland, [15]Victor Phillip Dahdaleh Heart and Lung Research Institute, University of Cambridge, Cambridge, United Kingdom, [16]Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Skopje, North Macedonia, [17]Institute of Genomics, Estonian Genome Centre, University of Tartu, Tartu, Estonia, [18]Department of Biotechnology, Institute of Molecular and Cell Biology, University of Tartu, Tartu, Estonia, [19]Nephrology and Infectious Diseases R & D Group, i3S—Instituto de Investigação e Inovação em Saúde; INEB—Instituto de Engenharia Biomédica, Universidade do Porto, Porto, Portugal, [20]Department of Informatics, University of Piraeus, Piraeus, Greece, [21]Computer Science and Biomedical Informatics Department, University of Thessaly, Lamia, Greece, [22]Department of Electrical and Electronics Engineering, Karadeniz Technical University, Trabzon, Türkiye, [23]INRAE, MetaGenoPolis, Université Paris-Saclay, Jouy-en-Josas, France, [24]Faculty of Sciences, Technology and Engineering, University of Vic – Central University of Catalonia, Vic, Barcelona, Spain, [25]IRIS-CC, Fundació Institut de Recerca i Innovació en Ciències de la Vida i la Salut a la Catalunya Central, Vic, Barcelona, Spain, [26]Department of Molecular Biology and Genetics, Çanakkale Onsekiz Mart University, Çanakkale, Türkiye, [27]Department of Microbiology and Clinical Microbiology, Faculty of Medicine, Erciyes University, Kayseri, Türkiye, [28]Metagenomics Laboratory, Genome and Stem Cell Center (GenKök), Erciyes University, Kayseri, Türkiye, [29]Finnish Institute for Health and Welfare - THL, Helsinki, Finland, [30]Institute for Molecular Medicine Finland, FIMM-HiLIFE, Helsinki, Finland, [31]Institute of Molecular and Cell Biology, University of Tartu, Tartu, Estonia,

[32]Department of Molecular Biology and Genetics, Çanakkale Onsekiz Mart University, Çanakkale, Türkiye, [33]Department of Mathematics, Center for Mathematics and Applications (NOVA Math), NOVA School of Science and Technology, Caparica, Portugal, [34]UNIDEMI, Department of Mechanical and Industrial Engineering, NOVA School of Science and Technology, Caparica, Portugal, [35]Bioinformatics Core, Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg, [36]School of Medicine, University of Galway, Galway, Ireland, [37]Department of Inorganic Chemistry, Faculty of Chemistry and Pharmacy, University of Sofia, Sofia, Bulgaria, [38]Center for Environmental and Respiratory Health Research (CERH), Research Unit of Population Health, University of Oulu, Oulu, Finland, [39]Biocenter Oulu, University of Oulu, Oulu, Finland, [40]Sarajevo Medical School, University Sarajevo School of Science and Technology, Sarajevo, Bosnia and Herzegovina, [41]Department of Clinical Science, University of Bergen, Bergen, Norway, [42]Mathematical Department, UPC-Barcelona Tech, Barcelona, Spain, [43]APC Microbiome Ireland, University College Cork, Cork, Ireland, [44]Computer Science and Engineering Department, Faculty of Automatic Control and Computers, National University of Science and Technology Politehnica, Bucharest, Romania, [45]Systems Ecology Group, Luxembourg Centre for Systems Biomedicine, Esch-sur-Alzette, Luxembourg, [46]Department of Life Sciences and Medicine, Faculty of Science, Technology and Medicine, University of Luxembourg, Belvaux, Luxembourg, [47]Department of Computer Technologies, Karadeniz Technical University, Trabzon, Türkiye, [48]Department of Information Systems, Zefat Academic College, Zefat, Israel, [49]Galilee Digital Health Research Center (GDH), Zefat Academic College, Zefat, Israel, [50]School of Microbiology, University College Cork, Cork, Ireland

The human microbiome has become an area of intense research due to its potential impact on human health. However, the analysis and interpretation of this data have proven to be challenging due to its complexity and high dimensionality. Machine learning (ML) algorithms can process vast amounts of data to uncover informative patterns and relationships within the data, even with limited prior knowledge. Therefore, there has been a rapid growth in the development of software specifically designed for the analysis and interpretation of microbiome data using ML techniques. These software incorporate a wide range of ML algorithms for clustering, classification, regression, or feature selection, to identify microbial patterns and relationships within the data and generate predictive models. This rapid development with a constant need for new developments and integration of new features require efforts into compile, catalog and classify these tools to create infrastructures and services with easy, transparent, and trustable standards. Here we review the state-of-the-art for ML tools applied in human microbiome studies, performed as part of the COST Action ML4Microbiome activities. This scoping review focuses on ML based software and framework resources currently available for the analysis of microbiome data in humans. The aim is to support microbiologists and biomedical scientists to go deeper into specialized resources that integrate ML techniques and facilitate future benchmarking to create standards for the analysis of microbiome data. The software resources are organized based on the type of analysis they were developed for and the ML techniques they implement. A description of each software with examples of usage is provided including comments about pitfalls and lacks in the usage of software based on ML methods in relation to microbiome data that need to be considered by developers and users. This review represents an extensive compilation to date, offering valuable insights and guidance for researchers interested in leveraging ML approaches for microbiome analysis.

KEYWORDS

microbiome, machine learning, software, feature generation, feature analysis, data integration, microbial gene prediction, microbial metabolic modeling

1 Introduction

The great development during the last decades in high-throughput technologies has allowed outstanding advances in different areas of knowledge like genomics (The 1000 Genomes Project Consortium et al., 2015), epigenomics (Stunnenberg et al., 2016), biodiversity (Lewin et al., 2018) or diseases (Boycott et al., 2019; Zhang et al., 2019). Microbiology has been paramount/highly integral here, in particular due to the reduction of costs and easy access have led to the creation of large volumes of data. Keystone microbiome projects like the Human Microbiome Project (The Human Microbiome Project Consortium, 2012), and the American

Gut Project (McDonald et al., 2018) have collected 16S rRNA gene sequences for more than 31,000 and 15,000 human microbiome samples, respectively (date: 08/05/2023), whereas other general microbiome sequencing data repositories like MGnify include more than 147,000 human samples (date: 08/05/2023). This enormous volume of data has allowed the application of machine learning (ML) techniques in human research to support the classification of microbial DNA sequences, microbiome-related stratification of subjects, and the inference of host phenotypes in disease prediction/severity (Goodswen et al., 2021; Marcos-Zambrano et al., 2021; Yadav and Chauhan, 2022). The technology can provide useful and hidden patterns of information from large, noisy complex data like the microbiome. However, a number of challenges in the application of ML techniques in microbiology need to be addressed in terms of data type and quality, model interpretability, high dimensionality, or standards in development and deployment of ML techniques that have been reviewed elsewhere (Goodswen et al., 2021; Moreno-Indias et al., 2021).

Microbiome data has a high level of individual variation and can be influenced by known and unknown host-related processes. Therefore, ML can typically detect informative and hidden patterns in the data that might be with limited prior knowledge of the system in question. These algorithms can be divided into different categories, including supervised, unsupervised, semi-supervised and reinforcement learning (Sarker, 2021), whereof supervised and unsupervised methods are the most applied in human microbiome studies (Ghannam and Techtmann, 2021; Goodswen et al., 2021; Marcos-Zambrano et al., 2021). Previous work by the COST (European Cooperation in Science and Technology) Action CA18131 on *Statistical and Machine Learning Techniques in Human Microbiome Studies* (ML4Microbiome) has outlined the existing ML algorithms relevant for microbiome analysis (Marcos-Zambrano et al., 2021).

The complexity of microbiome interactions with the host, health outcomes, and the environment can be approached with the integration of different ML techniques and the exponentially growing body of microbiome data for a wide variety of applications in humans (Marcos-Zambrano et al., 2021). This is leading to the development of a wide array of specific software and frameworks that integrate different ML methods considering the different typologies of microbiome data. Microbiologists and biomedical scientists have a huge collection of tools to get the most out of their microbiome data, however, these tools are fragmented and dispersed among different repositories and publications. Frameworks for ML methods do not cover all different steps for microbiome analysis and the user often needs to combine different methods into a data science workflow to complete the analysis. Therefore, selecting the software and tools for microbiome data analysis requires diving into multiple repositories and resources being a time-consuming task at the rate at which these developments are growing in recent years.

Here, our aim is to go beyond the application of ML techniques in the microbiome field, extensively reviewed in the last few years (Ghannam and Techtmann, 2021; Goodswen et al., 2021; Marcos-Zambrano et al., 2021), and focus on a scoping review of ML-based software and framework resources currently available for the analysis of microbiome data in human studies. A description of each software with examples of usage is provided including comments about pitfalls and lacks in the application of ML methods in relation to microbiome data that need to be considered in software development. For a better understanding, the different pieces of software are organized by the type of analysis for which they were developed and the ML methods implemented. As far as we know, this is the most extensive catalog to date that intends to help microbiologists and biomedical scientists who are starting or wish to go deeper into specialized resources that integrate ML techniques for the analysis of microbiome data.

1.1 Specific software for ML applications in microbiome studies

In Supplementary Table 1 we summarize the most commonly used ML software for microbiome data analysis including the applicability (one application or more), availability of source code, last version, number of citations based on the Scopus database (this gives an idea about the level of usage), type of tool (level of deployment) and availability (public/commercial) for all the software and tools included. Each publication has been associated with the URL (pointed in the text) to the software described therein.[1] Next, the software was evaluated in terms of the technologies used and the main ML tasks performed by the software. This allowed us to verify the most common ML tasks, the technologies used, and the change in the technologies used in recent years.

In Figure 1, we summarize the typical software stack used for microbiome tools over the years for given ML tasks. The thickness of the line indicates the number of publications divided into "year" - "programming language" and "programming language" - "ML task." In recent years there has been a significant increase in the popularity of solutions created in interpreted programming languages (mainly Python and R) in relation to compiled programming languages (such as C/C++ or Java). With the exception of solutions written in the Perl interpreter, which has lost its popularity significantly over the years. There is a growing number of solutions using tensorflow for deep learning in microbiome research.

It should be noted that tool authors moved away from publishing software only in compiled (closed source) form (this trend could be observed until 2013 in our data), as closed source distribution of scientific software made verification impossible and contradicted the ideas of open science.

The last remark concerns the availability of the software after years, most likely due to the academic funding and career structure. Our observations show that as much as 11.5% of projects created between 2005 and 2022 are no longer maintained[2] - and the software can only be found in the Internet Web archive.

In Figure 2 we present a series of specialized ML software and tools used to facilitate several microbiome research steps. These steps include feature generation, where raw 16s rRNA and shotgun sequencing data are processed and transformed into interpretable microbial units; data

1 Up to 11.5% of the URLs were pointing to non-existent or outdated pages - in this case, the link to the software was checked with the Internet Archive (https://web.archive.org) to find a page corresponding to the described software.

2 The url provided in the publication to the software points to non-existent resources, and there is no redirection to a new page.

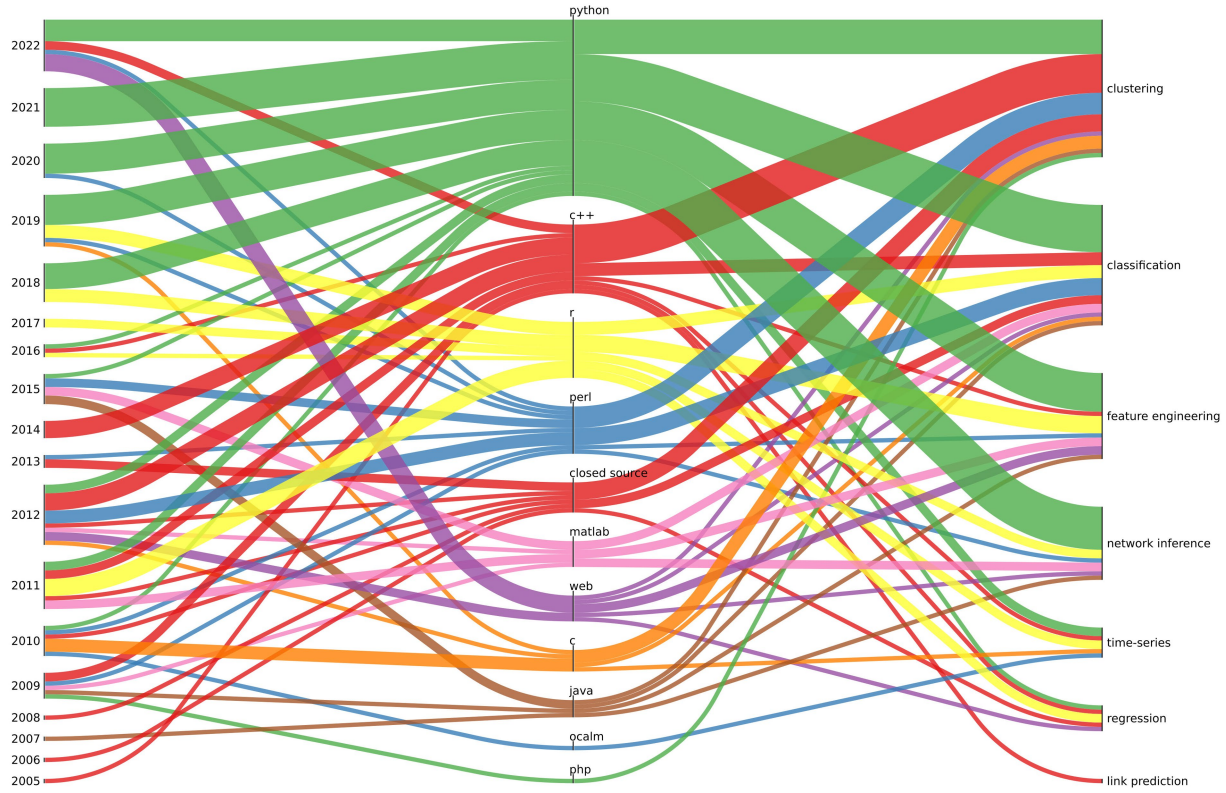

FIGURE 1
The relationship between the year of publication (left), programming language (centre), and ML task (right) is depicted for the most commonly used software in microbiome analysis. The thickness of the line represents the quantity of software projects associated with a particular relationship (a project may have multiple relationships of given kind i.e., a software may be written in C and Python).

integration, where disparate datasets are combined for comprehensive analysis; and feature analysis, where a variety of tools are employed to perform time series analysis, gene prediction, metabolic modeling, disease prediction, and comparative metagenomics. These software and tools, discussed in detail in the next sections, can empower researchers to uncover the intricate dynamics within microbiomes and advance their understanding of their roles in human health. The emphasis is on ML software, and hence quite a number of very popular software in microbiome studies (Metaphlan, KneadData, and Kraken2,) would not be mentioned, due to omitting ML approaches.

Furthermore, we provide a comprehensive interactive table in the Supplementary materials that summarizes available software and tools for analyzing different types of microbiome data, organized according to their primary application (code accessible at https://github.com/laurichi13/Toolbox-ML-software).

2 ML-software for feature generation

In microbiome analysis features are usually generated by using two learning approaches: clustering and classification. Clustering is an unsupervised approach (an approach without a teacher) where the system forms groups of inputs (or clusters) according to the explicit or implicit rule and given a particular set of patterns or cost function (Duda et al., 2001). On the other hand, classification involves learning from a set of patterns whose category is known (i.e., supervised approach) and applying it to a set of patterns with unknown category, without any grouping.

2.1 Feature generation and taxonomic assignment from 16S rRNA gene sequencing

Human (and environmental) microbial analyses are often performed using 16S rRNA gene sequencing. This is possible as the 16S rRNA gene is highly conserved and universally present across prokaryotes. The 16S rRNA gene analysis implies using primers to amplify the hypervariable regions of the 16S rRNA gene (ranging from V1 to V9; frequently targeted for bacteria are the V3, V4, and V3-V4 regions; Nguyen et al., 2016).

Amplicon Sequence Variants (ASVs) provide a precise resolution of sequence variations without imposing arbitrary dissimilarity limits, unlike Operational Taxonomic Units (OTUs), which are commonly used in 16S rRNA data processing (Eren et al., 2013). ASV techniques utilize Illumina-scale amplicon data and can identify sequence differences as small as one nucleotide. They infer the biological sequences in the sample while considering amplification and sequencing errors (Callahan et al., 2017). On the other hand, OTUs cluster sequences based on similarity and assign representative sequences to proxy microbial taxa (Westcott et al., 2017; Wei et al., 2021).

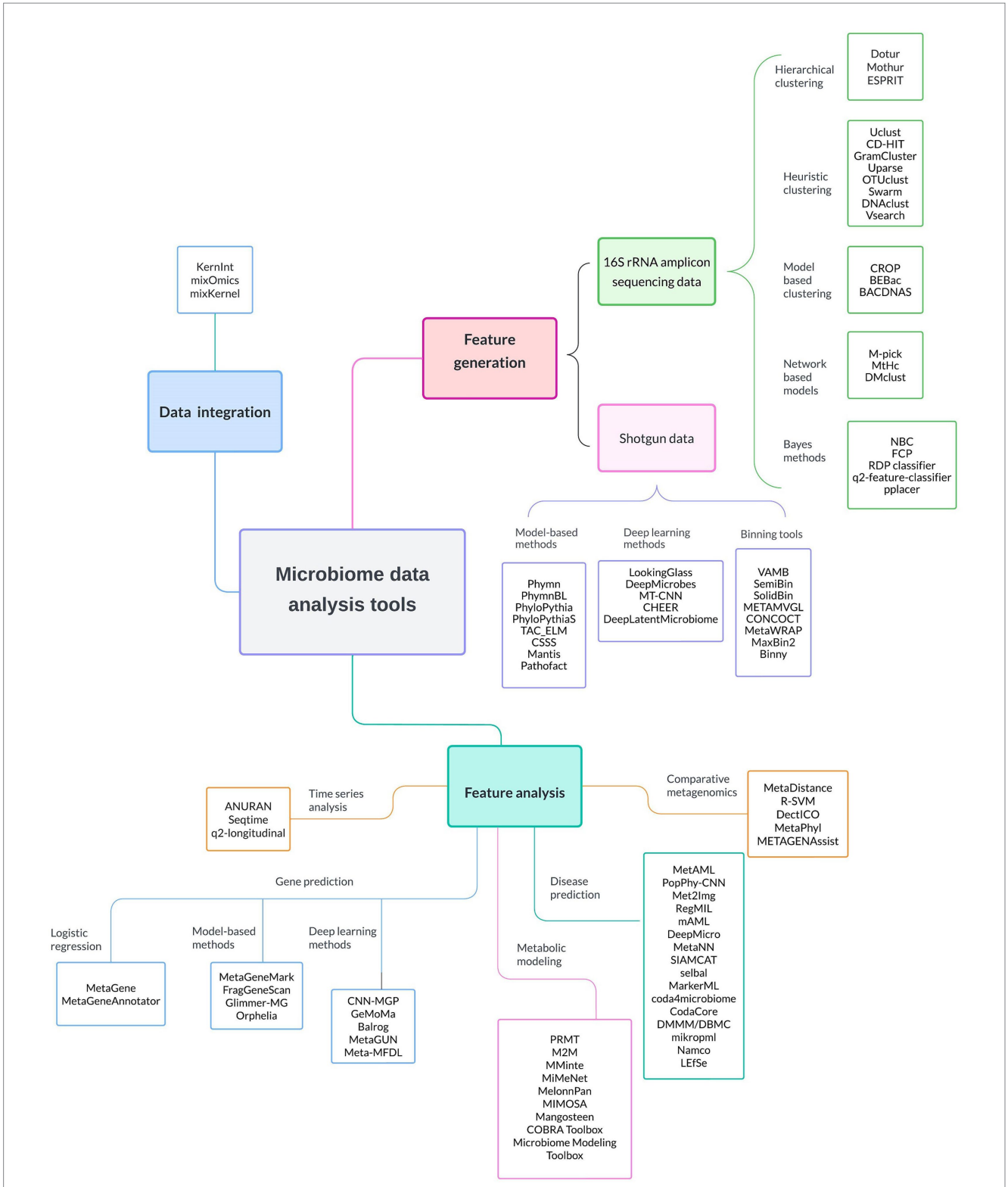

FIGURE 2
Comprehensive overview of the most commonly ML-based software applications employed in microbiome data analysis. These software tools are categorized based on their primary application into feature generation, feature analysis, and data integration. It is worth noting that numerous software options are applicable to both 16S rRNA gene sequencing data and shotgun metagenomics. Detailed descriptions of these software tools can be found in subsequent sections of the manuscript.

2.2 Clustering of sequences (reads) for OTU/ASV assignment

Several clustering methods have been proposed, and several reviews are available with a solid methodological overview, limitations, performance comparison, and guidance in the selection of an appropriate clustering algorithm (Chen et al., 2013; Nguyen et al., 2016; Wei et al., 2021). Without the intention to provide a thorough evaluation of different OTU clustering methods, we here provide available tools for the generation of OTU tables, aiming to indicate the advantages and limitations of clustering approaches and resulting OTU features in general.

In contrast to the clustering-based OTU approach, the generation of ASVs can be described as a denoising method (Chiarello et al., 2022), where the algorithm gathers exact sequence variants *de novo* with little room for mismatches and determines their abundance. Based on the inferred ASVs, an error model is calculated for the dataset to compare highly similar reads in order to statistically exclude sequencing errors. This is based on the assumption that true biological sequences occur in higher frequencies than sequences emerging from sequencing errors. Moreover, unlike *de novo* clustered OTU, the identity of an ASV keeps its validity outside of the data set from which it was derived, thereby also simplifying meta-analyses of multiple data sets (Callahan et al., 2017). However, some limitations inherent to OTU-based methods such as multiple copies of the target region within an organism (e.g., 16S rRNA gene copy numbers) and the restricted information content of short reads also apply to ASV-based methods and should be considered in the interpretation of results.

2.2.1 Hierarchical clustering

Creating clusters of data with similar characteristics is an approach to finding structure in data. Hierarchical clustering is an unsupervised learning technique for grouping similar objects into clusters. It creates a hierarchy of clusters based on similarity features within the data. Hierarchical clustering can be divided into two types: agglomerative (bottom-up) and divisive (top-down). The dendrogram construction depends on the type of linkage (i.e., the definition of distance between the clusters) used. The typical choices for OTU clustering are single linkage (which calculates the distance between the two closest objects belonging to each cluster, or nearest neighbor), complete linkage (which in turn is based on the distance between the two most distant objects, or furthest neighbor) and average linkage (unweighted-pair group), which is a compromise between the nearest neighbor logic of single linkage (Zhang et al., 2013). Once a hierarchical tree is constructed, the meaningful clusters can be defined by cutting the tree at a user-specified similarity threshold and merging all the sequences with higher similarity in the same OTU. Among these methods, the most familiar ones are Dotur (Schloss and Handelsman, 2005), based on Multiple Sequence Alignments, Mothur (Schloss et al., 2009), based on Needleman-Wunsch alignments against a pre-aligned reference database and ESPRIT (Sun et al., 2009), which implements a complete-linkage hierarchical clustering and minimizes the memory usage by adopting a k-mer distance for faster identification of very similar sequence pairs, producing sparse distance matrix. In hierarchical approaches, the number of sequences to be compared (N) determines the computational complexity [$O(N^2)$], which usually renders these approaches more intensive as stated by the authors.

2.2.2 Heuristic clustering of sequences

Heuristic clustering attempts to improve speed and scalability, avoiding exhaustive pairwise distance computation, and using a greedy strategy to form clusters based on an initial set of cluster seeds (Wei et al., 2021). Given a set of sequences, a subsequence is selected as a seed of a new OTU cluster. This subsequence is then compared to all remaining sequences of the given set of sequences. All sequences at the distance below the threshold with respect to any of the seeds are added to the corresponding OTU and removed from the sequence set. If no similar seed is found, a new cluster seed is formed from the query sequence. The performance of these methods is as well related to the selection of seeds. Some representative examples are Uclust (Edgar, 2010) and CD-HIT (Li et al., 2001; Li and Godzik, 2006). GramCluster (Russell et al., 2010) indexes the input dataset by a suffix tree for efficiency. Uparse (Edgar, 2013), an improvement of USEARCH (Edgar, 2010) and OTUCLUST (Albanese et al., 2015) rely on high quality sequences only, including steps for quality filtering, trimming, and chimera filtering. Swarm (Mahé et al., 2014) uses an agglomerative, unsupervised, single-linkage clustering algorithm that avoids the use of a global threshold. Each amplicon can be seen as a point in the discrete amplicon space, where its nearest neighbours have one nucleotide difference. User set parameter d is considered a tolerable similarity threshold, so that d-neighbours in the amplicon space are all amplicons with d nucleotide differences. Clustering amplicons starts from a seed, collecting all of its d-neighbours, and continues iteratively from these subseeds until natural cluster limits are reached, where no d-neighbours of any subseed can be added. In such a discrete amplicon space, amplicon clusters (OTUs) should be clearly separated contiguous regions, and the procedures ensures that all similar amplicons (i.e., amplicons close in the space) belong to the same cluster. DNACLUST (Ghodsi et al., 2011) adopts a greedy approach but improves the speed using filtering based on k-mers. There is an open-source 64-bit program VSEARCH (Rognes et al., 2016) which can be used instead of USEARCH, for which the source code and 64-bit versions are not publicly available.

2.2.3 Model-based clustering

These methods attempt to circumvent the overestimation of OTUs due to the limitations of choosing an *a priori* similarity threshold (Chroneos, 2010; Huse et al., 2010). Setting a (hard) similarity threshold value directly affects clustering process and the resulting sequences' partition, while using the probabilistic distance description fits better the nature of real data. The model-based methods, such as CROP (Hao et al., 2011) for example, tend to use Gaussian probabilistic distribution, indirectly targeting a certain similarity threshold, but being more flexible and thus more robust to sequencing errors and sequence variations. Moreover, the model based approaches imply very careful selection of model parameters, which is usually given as an optimization problem limiting the probabilistic parameter search to the parameter subspace in which the clustering results correspond to the desired partitions and to real number of OTUs (Hao et al., 2011). Other methods are BEBaC (Cheng et al., 2012), which is based on the calculation of an unnormalized posterior probability for an arbitrary partition of the reads, and BACDNAS (Jääskinen et al., 2014), which models sequences by Markov chains.

2.2.4 Network-based models

They start from a graph construction which requires a full distance matrix between sequences, which involves computational

burden, both memory and time consumption. Given this distance matrix, a weighted network is constructed and then a graph-based clustering method, based on the modularity community detection method, can be used for OTU picking (Wei et al., 2021). Some representative methods are: M-pick (Wang et al., 2013), MtHc (Wei and Zhang, 2015), and DMclust (Wei et al., 2017).

All of the clustering methods rely on similarity metrics and similarity thresholds used, which impact the output and quality of clustering. The selection of similarity measures is crucial, and research evidence indicates lots of criticism towards using percent sequence similarity in the OTU picking process (White et al., 2010; Schloss and Westcott, 2011). The reader is referred to Nguyen et al. (2016) for more insight into the problems of using sequence similarity for defining OTUs, which analyzes results obtained using three different dissimilarity metrics.

2.3 Taxonomic assignment of OTU/ASV

The procedures mentioned above for OTU/ASV clustering do not focus on species that constitute a sample. This is the goal of diversity profiling and taxonomic assignment. Diversity profiling aims to investigate the microbial community structure by providing an abundance of different taxa. The taxonomic assignment focuses on knowing which taxon belongs to each read or assembled contig. We can find two main kinds of software concerning these objectives: Naïve Bayes and Bayesian methods.

2.3.1 Bayesian methods

The RDP classifier (Wang et al., 2007; Cole et al., 2009) relies on a reference sequence database that contains relevant species, and then assigns a class label to each read by the naïve Bayesian algorithm based on k-mer occurrence. Moreover, we can find NBC (Rosen et al., 2011) and the classifier FCP (Parks et al., 2011), which also implement a naïve Bayesian framework. pplacer (Matsen et al., 2010), is a software package for phylogenetic placement and subsequent visualization, which offers a full probabilistic and Bayesian framework to locate a query sentence in a reference phylogeny so that a taxon identifier can be assigned to the query sequence.

Through QIIME2 (Bolyen et al., 2019) plugin q2-feature-classifier (Bokulich et al., 2018a), it is now also possible to train an almost arbitrary classifier from the Python library Scikit-learn and use it to predict the taxonomy. The real shift in taxonomic assignment came with (Kaehler et al., 2019), when the increase in the species-level classification accuracy is achieved by incorporating environment-specific taxonomic abundance information. Classifiers for amplicon sequences, like Naive Bayes, assume that all species in the reference database are equally likely to be observed (Kaehler et al., 2019). However, in practice, the equal probabilities (or the uniform weights) assumption is not fulfilled resulting in reduced accuracy. As the authors explain (Kaehler et al., 2019), the accuracy is less if weight distribution is closer to uniform than if it is further. In QIIME2 it is implemented as a preprocessing step through its plugin q2-clawback. The plugin is used for assembling taxonomic weights, which are further used as input into taxonomic classification.

There are a few analysis methods for microbiome amplicon data that analyze the obtained data without having to pre-process the raw reads generated by sequencing to create feature tables of ASVs. Read2Pheno is a deep learning framework to predict phenotype from all the reads in a set of biological samples (Zhao et al., 2021). The software performs alignment-free microbial 16S rDNA sequence analysis to achieve read- and sample-level environmental prediction and extracts interesting sequence features using convolutional neural networks (CNN), recurrent neural networks, and attention mechanisms.

2.4 Feature table generation from microbiome shotgun sequencing data

In contrast to amplicon sequencing (e.g., of 16S rRNA genes), shotgun metagenomics involves sequencing of all or most microbial DNA in a sample. The DNA is cut into short fragments which are separately sequenced as compared to amplifying a particular genomic region, resulting in a large set of short DNA sequences (i.e., reads) that originates from different chromosomal regions from numerous genomes. Some of these reads are from genomic loci of taxonomic significance (like the 16S rRNA gene), while others are of coding sequences that reveal information about the biological processes encoded in the genome (Sharpton, 2014).

The analysis of metagenomic sequencing data involves numerous challenges. First, metagenomic data is relatively complex and large, rendering the processing more difficult. Furthermore, reads only partially reflect most genomes because most communities are too diverse. Because of the massive quantity of genomic information examined, metagenomic analysis typically requires a large volume of data to get relevant conclusions. This requirement may cause computing issues (both in terms of space and time). Fortunately, these algorithms are continuously advancing, making metagenomic analysis more accessible and efficient.

2.5 Taxonomic classification of short sequence reads

There are different types of ML methods used for the taxonomic classification of short sequence reads in metagenomic sequencing data. Model-based methods include Phymm and PhymmBL (Brady and Salzberg, 2009), which use interpolated Markov models to phylogenetically classify short sequence fragments. PhyloPythia and PhyloPythiaS (McHardy et al., 2007; Patil et al., 2012) use support vector machine classifiers based on k-mer frequencies to assign reads to pre-existing taxa. The CSSS method (Borozan et al., 2015) applies the nearest neighbor algorithm to assign taxonomic ranks to both bacterial and viral communities.

Deep learning models based on artificial neural networks that add several hidden layers and several neurons within each layer, are also used for taxonomic classification of short sequence reads in metagenomic sequencing data. These models are computationally expensive but often have high accuracy, and are good at capturing complex biological systems. TAC-ELM (Rasheed and Rangwala, 2012) is a composition-based method that uses a neural network-based model. LookingGlass (Hoarfrost et al., 2022) is a deep learning biological language model designed to capture the functional diversity of the microbial world by encoding contextually aware representations of short DNA reads. The model takes into account the order in which sequences appear and thus produces contextually relevant embeddings

of biological sequences from microbial communities. Generated embeddings are able to differentiate sequences with different molecular functions, identify homologous sequences and differentiate sequences from disparate environmental contexts. Furthermore, LookingGlass may be fine-tuned by transfer learning to perform a variety of different tasks such as to identify novel oxidoreductases, to predict enzyme optimal temperature, and to recognize the reading frames of DNA sequence fragments. Liang and colleagues (Liang et al., 2020) developed a deep learning-based framework, DeepMicrobes, for taxonomic classification of short metagenomics sequencing reads that identifies potential uncultured species signatures in inflammatory bowel disease. This model achieved comparable accuracy in abundance estimation at the genus level when compared to state-of-the-art tools. The pipeline developed by Ma et al. (2021; MT-CNN) is based on a multi-task learning model that can perform both taxonomic assignment and estimation of genomic region for assigned reads for human viruses, together with a naïve Bayesian network which takes into consideration both the taxonomic assignments and the genomic coverage for the ranking of likely human viruses from sequence data. Ren et al. (2020) and Tampuu et al. (2019) proposed other deep learning-based approaches for classifying viruses from metagenomic reads. Shang and Sun (2021) presented CHEER, a tree-structure CNN pipeline for taxonomic classification of viral metagenomic data. PathoFact (de Nies et al., 2021) uses hidden Markov models and a random forest model in combination with the deep learning based DeepARG (Arango-Argoty et al., 2018) to predict virulence factors and antimicrobial resistance genes, while Mantis (Queirós et al., 2021) is a protein function annotation tool that uses database identifiers intersection and natural language processing based on text mining of protein function descriptions to integrate knowledge from multiple reference data sources into a single consensus-driven annotation.

2.6 Binning metagenome-assembled genomes

Binning is the computational process of assigning each read to a group called a bin, where each bin is expected to contain reads from the same taxon. Despite the existence of some alignment-based techniques (not covered in this review), the majority of computational tools for binning are currently in use in sequence *k*-mer composition. In fact, even when only dinucleotides (dimers) are taken into account, the distribution of k-mer composition is stable across a single genome and varies between genomes, as noted by Kariin and Burge (1995).

Binning is frequently used in environmental and human studies with the aim of establishing the taxonomic profile of a given sample. We distinguish between binning and taxonomic classification of amplicon sequences primarily based on the input data: whereas the latter is used in targeted studies, binning deals with assembled contigs from metagenomic reads from any genomic region of any sampled genome. Thus, binning is the method of choice for analyzing complex communities to determine near complete metagenome-assembled genomes (MAGs). However, almost all currently used techniques were created for bacterial communities, with MetaVir (Roux et al., 2011) being a notable exception as it focuses on the analysis of viromes. Other communities, like fungi, are frequently analyzed using *ad hoc* techniques or software tools intended for bacteria [see, for example, (Lindahl et al., 2013; Orellana, 2013)].

There are several binning tools available that use different methods as reviewed by Yang et al. (2021). For instance, VAMB (Nissen et al., 2021) uses deep learning in the form of variational autoencoders, while SemiBin (Pan et al., 2022) uses deep siamese neural networks in a semi-supervised approach. SolidBin (Wang et al., 2019) is based on semi-supervised spectral clustering, and METAMVGL (Zhang and Zhang, 2021) is a multi-view graph-based metagenomic contig binning algorithm. MetaDecoder (Liu C. -C. et al., 2022) is using a two-layer model based on Gaussian mixture models. Binny (Hickl et al., 2022) uses k-mer composition and coverage by metagenomic reads for iterative, nonlinear dimension reduction of genomic signatures as well as subsequent automated contig clustering with cluster assessment using lineage-specific marker gene sets. MaxBin2 (Wu et al., 2016) and CONCOCT (Alneberg et al., 2014) employ tetranucleotide frequencies (TNFs) and read depths to group together scaffolds. MaxBin2 utilizes an expectation–maximization algorithm to estimate the distances between scaffolds, while CONCOCT leverages Gaussian mixture models to cluster the scaffolds. However, there is no one-size-fits-all solution for metagenome binning, and ensemble-based tools like the binning module in MetaWRAP (Uritskiy et al., 2018) offer a promising approach to amalgamating binning results from various tools.

3 Analysis of features derived from amplicon or shotgun metagenomics:

3.1 Comparative metagenomics

This section includes techniques that label entire samples by examining features derived from each amplicon or shotgun DNA fragment from the sample (k-mers or OTU/ASV frequencies), sometimes supplemented with additional information (e.g., metadata, phylogenetics, class labels etc.). A common application of this classification in biomedical settings is phenotype analysis based on metagenomic fragments (Soueidan and Nikolski, 2016).

MetaPhyl (Tanaseichuk et al., 2014) is a two-phase heuristic algorithm for separating short paired-end reads from different genomes in a metagenomic dataset. The algorithm is based on the observation that most of the *l*-mers belong to unique genomes when *l* is sufficiently large. In the first stage of the algorithm, groups of l-mers are produced, each of which is associated with a single genome. Clusters are combined based on information from l-mer repeats during the second phase. Read assignments are made using these final clusters. The algorithm can handle very short reads and sequencing errors.

The study by Cui and Zhang (2013) employed R-SVM, which utilized generalized recursive Support vector machines (SVMs) to conduct feature selection and discrimination of human metagenome samples from control and inflammatory bowel disease patients. This alignment-free supervised classification approach can effectively differentiate between metagenomic samples belonging to predefined categories by selecting distinctive sequence features. The authors demonstrated the potential of utilizing metagenomic sequence features of microbiomes in the human body to investigate particular health conditions through supervised ML techniques.

DectICO (Ding et al., 2015) is a feature extraction, and dynamic selection-based supervised metagenomic classification method that can correctly classify metagenomic samples without relying on known microbial genomes and reads alignment. The tool combines SVM as the learning algorithm, intrinsic correlation of oligonucleotides (ICO), which generalizes the k-mer frequencies to describe samples, and kernel partial least squares for feature selection. When long k-mers are considered, the authors contend that DectICO performs better than other sequence-composition-based classification methods.

METAGENassist (Arndt et al., 2012) is a web server to make comparative metagenomics accessible to microbiologists. Users can upload their bacterial census, either amplified 16S rRNA data or shotgun metagenomic data, along with metadata (e.g., environmental, culture, and host conditions). All statistical analyses are performed by combining and normalizing user-submitted taxonomic profile data and automatically mapped phenotypic information (e.g., oxygen requirements, temperature range, habitat, host type, pathogenicity, disease association etc.) from METAGENassist's phenotypic database. A variety of univariate methods are available for feature ranking regarding the significance of their changes due to the different conditions under study (e.g., fold change analysis, *t*-tests, Mann–Whitney tests, ANOVA, Kruskal–Wallis tests). Multivariate methods, namely, principal component analysis (PCA) and partial least squares discriminant analysis (PLS-DA), can be used for dimension reduction, visualization, classification, and feature identification. Hierarchical and partitional clustering methods are available to identify groups of samples regarding their feature abundance profiles, given their similarity based on a defined distance measure. For the prediction of attribute labels and the identification of important features (i.e., taxa or mapped phenotypes) METAGENAssist offers two methods, random forest and recursive SVM feature selection and sample classification (R-SVM). Mian (Jin et al., 2022) is another interactive web-based microbiome data table visualization and ML platform. Users can upload their metagenomic data as well as accompanying metadata, taxonomic mappings, phylogenetic tree or gene expression data. Mian allows users to preprocess their data, calculate alpha and beta diversity measures, apply feature selection methods and train ML models such as linear regressors, random forest or multilayer perceptrons. All tools are easy to tune and configure, and users will also be able to obtain common statistical measures as well as different plots for data visualization.

MetaDistance (Liu et al., 2011) is a MATLAB toolbox that comprehends the relationship between clinical phenotypes and microbiota profiles by developing new supervised learning tools. Instance-based [K-Nearest Neighbors (KNN)] and model-based (SVM) learning techniques have been combined to create the sparse distance learning approach (MetaDistance) that the authors have proposed for multi-class classification. The suggested approach is capable of class prediction and taxon identification in tandem. It can perform multi-class classification while not exacerbating any existing class imbalance. Additionally, this approach estimates only a few parameters, and specifically, the number of these parameters is equal to the number of features (input variables) in the dataset. This means that the model complexity is kept relatively low, which can be advantageous in scenarios with limited data or to prevent overfitting. It is very effective for metagenomic data issues, which frequently have small sample sizes, high dimensions, and unbalanced classifications with numerous classes.

3.2 Disease classification and feature prediction

The human microbiome is unique to each person and has been linked to various diseases, making it essential to associate the microbiome with the host's disease state (Yadav and Chauhan, 2022). The disease status may be influenced by the presence of specific microbe species, their abundance, phylogenetic relationships, intermicrobial interactions, and microbial metabolites. ML models can be useful for this task because they account for the complex dependencies between microbial community members and can identify disease profiles and microbial biomarkers with limited prior knowledge. Abundance values of microorganisms, functional annotations of metagenomes, and k-mer abundances from raw reads are common features used for disease prediction (Bakir-Gungor et al., 2022). Microbial abundance profiles are commonly used as a feature in disease classification. This field is still in its early stages, and several ML approaches have been developed for classification based on disease-associated microbiome composition data (Bakir-Gungor et al., 2021). Here, we present several ML approaches designed for classification purposes given the disease-associated data about microbiome composition.

MetAML (Metagenomic prediction Analysis based on Machine Learning) is a computational tool for disease detection using gut metagenomic data. Here, SVMs, RFs, Lasso, Elastic Net, and other classifiers are implemented in this ML software framework for metagenome-based prediction tasks (Pasolli et al., 2016). Cross-validation allows for quantitative evaluation of model precision and adaptability to the general population. Evaluation metrics commonly used to measure the model's performance include accuracy, sensitivity, specificity, precision, F1 score, AUC, among others (Table 1). MetAML has been tested on metagenomic case–control datasets from five different diseases, demonstrating potential for

TABLE 1 Commonly used metrics to assess the performance and effectiveness of machine learning models.

Metric	Definition
Accuracy	Measures the overall correctness of the predictions made by a model. It is the ratio of the correctly predicted instances to the total number of instances in the dataset.
Sensitivity (Recall or true positive rate)	Quantifies the proportion of actual positive instances that are correctly identified as positive by the model. It is the ratio of true positive predictions to the sum of true positives and false negatives.
Specificity	Represents the ability of a model to identify negative instances correctly. It is the ratio of true negative predictions to the sum of true negatives and false positives.
Precision	Indicates the proportion of correctly predicted positive instances out of the total instances predicted as positive by the model.
F1 score	Is the harmonic mean of precision and sensitivity and provides a balanced evaluation of a model's performance.
AUC (Area Under the ROC Curve)	The ROC curve plots the true positive rate against the false positive rate at various classification thresholds. AUC represents the area under this curve and is a measure of the model's ability to discriminate between positive and negative instances.

disease detection from gut metagenomic data. It has also been used in a study by Thomas et al. (2019), where ML models based on MetAML were developed to predict colorectal cancer using metagenome dataset. The models evaluated the prediction accuracies of the gut microbiome for colorectal cancer detection across populations and successfully identified consistent microbiome biomarkers and accurate disease-predictive models.

PopPhy-CNN (Reiman et al., 2020) is a convolutional neural network (CNN) that predicts the host's disease status using their microbiome samples. PopPhy-CNN involves transforming the phylogenetic tree and microbial abundance data into a structured matrix format. This matrix, enriched with evolutionary information, is then used as input for a CNN model to make predictions about the host's disease status. The incorporation of biological knowledge through this process contributes to the model's superior performance compared to other methods in binary classification and multi-class datasets. PopPhy-CNN models were more competitive than RF, SVMs, LASSO, 1D-CNN, MLPNN, and Ph-CNN models across nine moderately sized metagenomic datasets for binary classification (Qin et al., 2012, 2014; Karlsson et al., 2013; le Chatelier et al., 2013; Sokol et al., 2017). According to authors, PopPhy-CNN can deliver reliable performance with minimal training data and shows the best results for multi-class biological and synthetic datasets.

Met2Img (Hai Nguyen et al., 2019) is a disease prediction method that uses Synthetic Image Representations of Metagenomic data and CNN. The authors use a rectified linear unit (ReLU) activation function and transform each sample into an image containing coloured pixels representing the microbes and their relative quantities. The resulting images are subsequently used as features for the neural network. The authors evaluated the method using six metagenomic datasets, including five disease types and more than 1,000 samples. They reported encouraging results and held applicability across diverse omics data scenarios, including integrative contexts (i.e., taxonomic levels, CNN structure optimization, dimensionality reduction: effective colormaps, and GPU efficiency).

RegMIL is a Multiple Instance Learning (MIL) method that predicts phenotypes from metagenomic data. This approach employs a rapid, hash-based clustering technique referred as Canopy clustering to score instances in the training set. These scores estimate the contribution of an instance (sequence) to the disease. The instance scores of the training set are used to train a two-layer neural network-based regression model to score instances in the test set. In the end, one histogram-based bag-level feature representation by taking contributions of each instance to train a classifier (Rahman and Rangwala, 2018). RegMIL was shown to predict a person's health status with high accuracy when evaluated with liver cirrhosis and IBS datasets, outperforming other tools like MetAML (Rahman and Rangwala, 2018).

mAML is an automated ML tool specifically designed for classification tasks performed on metagenomic data. The tool was developed in Python and the entire pipeline can be run through a web server, although it is also available to download and run locally. mAML preprocesses the data, performs grid-search for hyperparameter tuning, and provides several performance metrics for the classification task set by the user. The web-based tool allows the user to personalize each of these tasks. The mAML pipeline exhibits various benefits: (i) it can effectively and automatically construct an optimized, interpretable and resilient model for a microbiome-based classification task; (ii) it is implemented on a web- based platform (the mAML web server); (iii) the pipeline can be employed for both binary and multiclass classification tasks; (iv) it is data- driven and can readily be extended to encompass multi-omics data or other data types, given the availability of domain specific datasets (Yang and Zou, 2020). The authors evaluated mAML on 13 different metagenomic datasets, including binary and multi-class data. The models generated by mAML outperformed other models such as Support Vector Classifiers or logistic regression (Fierer et al., 2010; Wu et al., 2011; Qin et al., 2014; Montassier et al., 2016), demonstrating the method's robustness. This method has been applied to predict carboxylate production from 16S rRNA gene dynamics (Liu B. et al., 2022).

DeepMicro is a deep learning method that is focused on the extraction of features from high dimensional microbiome data (more specifically extracted abundance and strain profile). It was shown to be more accurate than MetAML in transforming high-dimensional metagenomic data into a reliable low-dimensional representation for supervised or unsupervised learning (Curry et al., 2021). It was developed with disease prediction in mind, but has other applications. This approach could improve model performance for predictive problems using microbiome data, such as drug response prediction, forensic human identification, and food allergy prediction (Oh and Zhang, 2020).

DeepLatentMicrobiome which has an artificial neural network (ANN) architecture based on heterogeneous autoencoders (García-Jiménez et al., 2021), uses phenotypic features as well as environmental features (like temperature, precipitation, plant age, maize line and maize variety) to predict current or future microbiome compositions and can help scientists develop microbiome-engineering strategies with limited resources. Autoencoders are trained for each data source independently (thus acquiring heterogeneous autoencoders).

MetaNN (Lo and Marculescu, 2019) is a neural network-based technique that addresses challenges related to over-fitting and high dimensionality in metagenomic data, leading to improved classification accuracy. The method involves removing taxa that appear in less than 10% of the samples and generating additional samples using a negative binomial distribution to augment the training set. A neural network is then trained on the augmented dataset, resulting in superior performance compared to other ML models such as Random Forests, SVM and CNN, as demonstrated in evaluations by the authors Lo & Marculescu in 2019 using both synthetic and real datasets.

SIAMCAT is an R-based software that combines ML, statistical modeling, and advanced visualization approaches to enable comparative metagenomic studies. The tool provides normalization methods, cross-validation schemes, and implementation of various ML approaches such as LASSO (Tibshirani, 1996), Elastic Net (Zou and Hastie, 2005), and RF (Ho, 1995), among others. The trained models can then be used to make predictions based on the provided metagenomic data, and their performance can be measured using AUROC. According to Wirbel et al. (2021), SIAMCAT allows users to apply robust and verified ML models to their datasets, allowing pre-processing and normalization of the datasets depending on metagenomic data properties. It has been used in various studies, including those involving the classification of oral microbiome data

(de Jesus et al., 2021) and the assessment of the association between microbiome composition and clinical responses to immune checkpoint inhibitor treatment (Lee et al., 2022). In the study developed by Kartal et al. (2022), it was discussed if fecal and salivary microbiota could be used as predictors of pancreatic ductal adenocarcinoma.

Namco is an R Shiny application designed for microbiome research that provides a wide range of data analysis tasks, including raw data processing, basic statistics (distribution of dominant taxa among groups), creation of heatmaps using different ordination methods, diversity analysis, network analysis, and ML (Dietrich et al., 2022). Among the latter, Namco offers users the ability to develop classification models using random forest to predict outcomes such as disease state or treatment response. The most important features in the classification are identified as biomarker candidates. The tool also enables time-series analysis and clustering to investigate microbial changes in response to treatment across different host development stages or over time.

LEfSe is a method for identifying metagenomic biomarkers that can explain differences between phenotypic classes. This method uses linear discriminant analysis (LDA) effect size (LEfSe; Segata et al., 2011). It is based on the non-parametric factorial Kruskal-Wallis sum-rank test to determine the statistical significance of differences found across classes. Biological consistency is then assessed using the Wilcoxon rank-sum test, and the effect size of each differentially abundant feature is estimated via LDA. Firstly, the Kruskal-Wallis test is employed to scrutinize all features and determine if there are dissimilarities in their distribution among different classes. Subsequently, features that contravene the null hypothesis undergo further analysis using the Wilcoxon test. This test compares all pairwise combinations between subclasses in different classes to ascertain if they conform to the general trend of the class. The resultant subset of vectors is then employed to establish an LDA model that ranks the features based on their relative differences among classes. Ultimately, the output is a list of discriminative features that are in line with the subclass grouping within classes and are ranked based on their effect size in distinguishing between classes.

MarkerML is a web server that employs interpretable ML and statistical testing to discover important metagenomic features (Nagpal et al., 2022). Its main goal is to identify marker-features, which can contrast comparable states and help in decision-making. Model interpretability is achieved by incorporating Shapley Additive exPlanations (SHAP)-based (Lundberg and Lee, 2017) analyses to detect predictive marker features. MarkerML also implements statistical testing methods to contextualize marker-feature discovery in metagenomic datasets, such as ANCOM-BC (Lin and Peddada, 2020; Lin et al., 2022) or ALDEx2 (Fernandes et al., 2013, 2014; Gloor et al., 2016). It also offers features such as access to databases (e.g., Taxonomic, KEGG, COG, PFAM), normalization options, feature selection, and multiple ML algorithms (e.g., XGBoost, Random Forests, Logistic Regression; Nagpal et al., 2022). MarkerML relies on class comparison and prediction for biomarker discovery, achieved by analyzing differential abundance and ML techniques, respectively.

Selbal is an algorithm whose objective is to find a microbial signature, i.e., a model defined by a group of microbial taxa whose pattern of abundance is predictive or associated with an outcome variable of interest (Rivera-Pinto et al., 2018). It uses the Selbal model selection method to find two groups of taxa whose relative abundance (referred as "balance") sufficiently explains the target response variable (Rivera-Pinto et al., 2018). The algorithm iteratively runs multiple regressions while including a new taxon in the model each time. The two taxa whose balance is most closely connected to the response are the first ones that selbal selects. This approach has been used to differentiate between polycystic and non-polycystic ovary syndrome women (Lüll et al., 2021).

coda4microbiome (Calle et al., 2023) is an improved version of Selbal, which uses elastic-net penalization for joint variable selection in the all-pairs log-ratio model (i.e., the model that considers as explanatory variables all pairwise log-ratios of features). It outperforms Selbal by being more computationally efficient and allowing for different weights in the microbial signatures. While selbal uses forward selection, coda4microbiome applies elastic-net penalization on the "all-pairs log-ratio model" to perform joint variable selection. After reparameterization, the results are expressed as a microbial signature consisting of two taxa groups that are associated with the phenotype. coda4microbiome's signatures are more versatile than selbal's, as they allow different weights for taxa in each group, while selbal assigns the same weight to all taxa in each group. Coda4microbiome has also been implemented for both cross-sectional and longitudinal studies. The website of the project contains several tutorials.[3] Other log-ratio based approaches for analyzing microbiome data include *CodaCore* (Gordon-Rodriguez et al., 2021) and the R package *amalgam* (Quinn and Erb, 2020), which aim to identify predictive balances or amalgams in a stepwise additive fashion. Some log-ratio based approaches in microbiome data analysis try to improve predictive accuracy by considering log-ratios that can contain several original features. However, many methods rely on pairwise log-ratios or additive log-ratios, which only involve two features. For example, the *easyCoda* R package includes three options for choosing pairwise log-ratios in a regression setting (Coenders and Greenacre, 2022), while the *logratiolasso* R package proposes a log-ratio LASSO model that aims to produce a sparse model from the all-pairs log-ratio model (Bates and Tibshirani, 2019).

DMMM/DBMC is a Dirichlet Multinomial Mixture Model (DMMM) tool that can be used in both unsupervised and supervised settings to identify clusters in microbiome datasets and act as a Bayes classifier. It is implemented in the R package *DirichletMultinomial* (Holmes et al., 2012) and was extended by Gao et al. (2017) to include automatic feature selection, resulting in better classification accuracy than DMMM and random forest.

mikropml is an R package that follows best practices for machine learning, producing trained models, performance metrics, and feature importances (Topçuoğlu et al., 2021). It includes data preprocessing, model training, and selection, as well as hyperparameter tuning. The package has been used to classify colorectal cancer patients and identify variables associated with bacterial infections (Topçuoğlu et al., 2021). The tool has also been applied to test ML models for associations between microbiome composition and diseases like *Clostridium difficile* infections, producing significant results in

3 https://malucalle.github.io/coda4microbiome/

multiple studies (Lapp et al., 2021; Armour et al., 2022; Lesniak et al., 2022).

3.3 Gene prediction

Metagenomic studies aim to understand the metabolic and functional diversity of microbial communities and detect differences among them. However, establishing a complete geneset for each species in a sample is currently unfeasible. Gene prediction is a valuable tool in functional profiling, as it identifies patterns in DNA sequences that correspond to transcription and translation machinery. Here we present some of the most used algorithms including not-ML based prediction models.

Hidden Markov models (HMM) are commonly used in gene prediction, with several methods available. MetaGene (Noguchi et al., 2006) uses logistic regression models based on GC content and di-codon frequencies to differentiate between gene-coding and non-gene coding open reading frames (ORFs). MetaGeneAnnotator (Noguchi et al., 2008) extends this approach by integrating species-specific patterns of ribosome binding sites to improve translation start site prediction.

Model-based methods are commonly used in gene prediction, and there are several notable examples. MetaGeneMark (Zhu et al., 2010) is based on Hidden Markov models that are applicable to short DNA fragments. It uses training prokaryotic genomes to estimate polynomial and logistic approximations of oligonucleotide frequencies as a function of GC content. FragGeneScan (Rho et al., 2010) and Glimmer-MG (Kelley et al., 2012) both use Interpolated Markov Models to distinguish coding areas from non-coding DNA. Orphelia (Hoff et al., 2008, 2009) instead uses linear discriminants for mono-codon usage, di-codon usage, and translation initiation sites to extract characteristics from sequences, and also incorporates a neural network trained on random sub-sequences of genomes from the reference database to classify ORFs as protein-coding or not.

CNN-MGP (Al-Ajlan and El Allali, 2019) is a successful deep learning-based method for gene prediction. CNN-MGP avoids manual feature extraction and selection by predicting genes directly from raw DNA sequences. This method demonstrates the power of deep learning in accurate gene prediction. GeMoMa (Keilwagen et al., 2019) leverages evolutionary information from gene models in reference species to predict gene models in target species using amino acid sequence conservation, intron position conservation, and RNA-seq data. It is a homology-based gene prediction program.

Balrog (Bacterial Annotation by Learned Representation Of Genes; Sommer and Salzberg, 2021) is a model of prokaryotic genes based on a data-driven approach to gene finding with minimal hand-tuned heuristics. By training a single gene model on nearly all available high-quality prokaryotic gene data, this model matches the sensitivity of widely used gene finders.

ML-based methods have proven useful for metagenomic gene prediction. Meta-MFDL (Zhang et al., 2017) is a notable example that utilizes deep stacking networks to combine features such as monocodon usage, monoamino acid usage, ORF length coverage, and Z-curve features. This model has shown robustness and high accuracy in identifying metagenomic genes, outperforming other prediction models.

MetaGUN (Liu et al., 2013) is an ML-based method that uses SVM classifiers to identify protein-coding sequences in metagenomic fragments. MetaGUN uses entropy density profiles of codon usage, translation initiation site scores, and open reading frame length as input patterns.

3.4 Metabolic modeling

The metabolic activities carried out by the bacteria forming the gut microbiome are relevant for gut homeostasis and overall host health and physiology. These activities might not always be affected by taxonomic changes, and therefore it is essential to characterize microbiome-metabolome interactions. This will help to understand how shifts in the gut microbiome composition may affect host health, which in turn is crucial for the treatment and prevention of chronic diseases. In this section, we will describe methods that have been developed to characterize the metabolic activity of the microbiome.

Early modeling approaches focused on converting metagenomic features to metabolomic features due to the lack of comprehensive metabolomic profiles. The Predicted Relative Metabolic Turnover (PRMT) method (Larsen et al., 2011), originally developed for a marine metagenome, predicts metabolite consumption or production based on the enzymatic activities present in a metagenome. Briefly, it leverages information from KEGG and MG-RAST (reactions and EC numbers, respectively) to generate an environmental metabolomic matrix (EMM), estimates enzymatic activity based on number of sequences, and calculates a PRMT-score for each metabolite in the EMM (Larsen et al., 2011).

MIMOSA adapts this methodology in a multi-omic framework that combines taxonomic and metabolomic profiles in the context of the human microbiome (Noecker et al., 2016). This framework first infers community gene content based on taxonomic data and available and inferred genomic information. Then, making use of the PRMT method, it predicts the communitywide uptake or production of each metabolite, and estimates how species and genes might be contributing to these activities. Similarly to MIMOSA, Mangosteen is a metabolome prediction pipeline that relies on relationships between KEGG/BioCyc reactions and their associated molecular compounds (Yin et al., 2020).

However, with the increasing availability of both metagenomic and metabolomic data, numerous ML models have been developed to map metagenomic features to metabolites. These methods overcome the main limitation of reference-based methods, which are dependent on the quality of the queried databases. For instance, MelonnPan uses Elastic net regularization to predict community metabolomes from taxonomic profiles (Mallick et al., 2019). This approach has been used to predict metabolites in new microbial communities based on metagenomic data, shedding light on the functional role of microbiota in cardiovascular diseases (Liu et al., 2020).

Another ML-based approach, MiMeNet, is a multi-layer perceptron neural network that models microbe-metabolite relationships and the metabolomic profile of microbial communities from metagenomic taxonomic or functional features. This approach allows for scalability in handling large amounts of metagenomic and metabolomic features and leads to more robust predictive models by

simultaneously learning metabolites and enhancing the transfer of information (Reiman et al., 2021).

Metage2Metabo (M2M) is another software tool that simulates the metabolism of the gut microbiota and describes the metabolic relationships between the species' metabolic genes to establish how they complement each other in metabolic terms. M2M uses reference genomes or MAGs to construct genome-scale metabolic networks, which are then analyzed to detect metabolic capabilities and metabolic cooperation potential. Once this is carried out, M2M calculates the minimum number of species needed to perform a metabolic role of interest and the key species associated with that role (Belcour et al., 2020). M2M relies on the genome-scale metabolic network generating tool Pathway Tools (Karp et al., 2016).

Other approaches focus on constraint-based stoichiometric modeling using flux balance analysis (Orth et al., 2010) to determine the rate at which metabolites are being exchanged within the community (Thiele et al., 2013; Baldini et al., 2019; Heinken and Thiele, 2022). Constraint-based reconstruction and analysis (COBRA toolbox) is a software package for MATLAB, which allows for the creation and analysis of genome-scale metabolic models (Heirendt et al., 2019). It is reliant on the COBRA method which is a well-described set of strategies to employ when using metabolic modeling (Heirendt et al., 2019). Currently, the COBRA Toolbox is in its third edition and aims to simulate the relationship between genotype and phenotype through mathematical modeling (Heirendt et al., 2019). The Python COBRApy was developed as a framework allowing to model complex biological processes using COBRA methods (Ebrahim et al., 2013).

COBRA modeling has been used to create personalized human microbiome models and stratify them based on structure and function, which has been used to treat conditions such as inflammatory bowel disease and colorectal cancer (Heinken et al., 2021). It also supports other computational methods used for metabolome predictions with microbial data. For instance, MMinte (Mendes-Soares et al., 2016) relies on ModelSEED (Henry et al., 2010) and COBRApy (Ebrahim et al., 2013) for metabolic modeling and flux balance analysis (Mendes-Soares et al., 2016). This pipeline predicts metabolic interactions among microbial species in a community from 16S rRNA amplicon sequence data and association networks. It allows us to identify related genomes, reconstruct metabolic models, assess growth under specific metabolic conditions, analyze pairwise interactions, and generate a network of interactions (Mendes-Soares et al., 2016).

The COBRA method has also been used to construct organ-resolved whole-body human metabolic models, enabling simulations of both human and microbiome-human interactions (Heinken et al., 2020). In addition to the COBRA toolbox, the Microbiome Modeling Toolbox (Baldini et al., 2019) is a suite of MATLAB-based tools for building and analyzing microbe-microbe and personalized microbiome models. This toolbox generates, simulates, and interprets interactions between microbes and the host, as well as sample-specific microbial community models, using metagenomically derived data (Baldini et al., 2019). The updated version of the toolbox includes the mgPipe module, which facilitates the generation of personalized microbiome models from a vast collection of microbial metabolic reconstructions, such as the AGORA resource, containing over 7,000 microbial reconstructions (Magnúsdóttir et al., 2017; Heinken et al., 2020; Heinken and Thiele, 2022). The AGORA resource is also used by other tools, including the second version of MIMOSA (Noecker et al., 2016). Finally, MICOM is a customizable metabolic model of the human gut microbiome. Through COBRApy, it calculates growth rates based on metagenomic and dietary characteristics, allowing for the generation of personalized metabolic models for individual metagenomic samples (Diener et al., 2020).

3.5 Time-series analysis

Time-series data analysis is essential for understanding the structure and dynamics of microbial communities. However, it requires specialized statistical considerations distinct from those used in comparative microbiome studies to address ecological questions. To facilitate this, some software packages have been developed that use ML algorithms to analyze time-series data.

One such package is QIIME2 plugin q2-longitudinal (Bokulich et al., 2018b), designed for the analysis and visualization of longitudinal microbiome studies. This QIIME2 plugin incorporates various methods for paired difference and distance testing, linear mixed-effects models, nonparametric microbial interdependence, feature selection and volatility analysis, and interactive visualization. The feature-volatility action uses random forests to identify features that change over time and predict different states.

Another package is Seqtime, an R package that provides functions to analyze sequencing data time-series and simulate community dynamics (Faust et al., 2018). Additionally, the Anuran toolbox helps identify conserved or unique patterns across multiple networks over time, and whether biological networks have set operations that have different outcomes than expected by chance (Röttjers et al., 2021).

4 Data integration

The complexity and heterogeneity of the metagenomics datasets, which include various types, scales, and distributions, make it challenging to extract useful information from them in the context of omics data mining. One of the main obstacles to the successful use of ML techniques in metagenomics analysis is the integration of such a wide variety of heterogeneous data.

Picard et al. (2021) classified integration approaches into horizontal and vertical categories. Within the vertical integration strategies, further divisions include early, mixed, intermediate, late, and hierarchical approaches. Early and intermediate integration strategies enable the analysis of datasets within the context of their relationships with other datasets, leading to additional insights. However, early integration is challenging for most ML models, while intermediate integration often relies on unsupervised matrix factorization, which lacks the incorporation of pre-existing biological knowledge. Late integration involves applying ML models separately to each dataset and then combining their predictions. Hierarchical integration considers the interaction between different layers of omics data explicitly, but its implementation is currently in its early stages.

Most of the integration approaches implemented in software packages are based on mixed integration, which typically first modifies and transforms each dataset using different ML models. This enables

them to reduce data complexity and heterogeneity, as well as to facilitate subsequent integration and analysis of datasets. Here we collect some of the ML software used for metagenomics data integration:

There are several software packages available for metagenomics data integration. mixOmics (Rohart et al., 2017a) for example, is an R package that provides a wide range of multivariate methods for data exploration, sizing, and visualization, including integration platforms that investigate relationships between heterogeneous omics data (in terms of types, scales and distributions). Its multivariate projection-based methods are computationally efficient for processing large omics datasets and provide flexibility in analyzing biological datasets by using relaxed assumptions about the distribution of the data. MixOmics R includes both supervised and unsupervised frameworks as well as feature selection. Other frameworks, like DIABLO (Singh et al., 2019) and MINT (Rohart et al., 2017b), enable the integration of datasets to identify relevant relationships and significant patterns in heterogeneous data for better exploration of complex metagenomic data.

Kernel methods allow data scientists to model non-linear relationships between the data points with low computational complexity, thanks to the so-called 'kernel trick'. These have already been used to extend well-known algorithms such as PCA, linear DA and ridge regression (Cabassi and Kirk, 2020). A consensus multiple kernels is based on ideas similar to STATIS as an exploratory method designed to integrate multi-block datasets when the blocks are measured on the same samples (Mariette and Villa-Vialaneix, 2018). MixKernel (Mariette and Villa-Vialaneix, 2018) is another R package that offers methods for integrating heterogeneous types of data, focusing on kernel fusion methods for unsupervised exploratory analysis. Its kernel methods allow data scientists to model non-linear relationships between the data points with low computational complexity, thanks to the so-called kernel trick. KernInt (Ramon et al., 2021) is a kernel framework for integrating supervised and unsupervised analyses in spatiotemporal metagenomic datasets, using a kernel framework to unify supervised and unsupervised microbiome analyses, focusing on spatial and temporal integration, including the retrieval of microbial signatures.

4.1 General software for machine learning applications

A variety of ML software tools are available, with the majority being open source. Goodswen et al. (2021) and co-authors have compiled a brief list of general ML software tools to be applied in microbiome data. We have here extended this list in Supplementary Table 2 to include additional relevant general ML software for microbiology data analysis. These tools are primarily based on Python and R frameworks that contain collections of software libraries (packages) and require some basic programming knowledge for optimal use. However, some ML tools like WEKA, KNIME Analytics Platform, and Orange Data Mining, can be used through a GUI without extensive coding or programming expertise.

4.2 Commercial approaches and solutions

We identified more than 240 companies (in >350 locations) worldwide based on an online database of companies applying or offering microbiome analysis (Microbiome Employers, 2022) complemented with search engine results.

The companies' activities ranged from clinical research and the study of diagnostic and therapeutic effects in healthcare to the implementation of microbiome data analysis in agriculture, nutritional supplements and pharmaceuticals. The majority of these address microbiome analysis for therapeutics/pharmacy. Three typical examples are the discovery of novel molecules for therapeutics, agriculture, and nutrition (Adapsyn Bioscience, 2022), the prediction of viable biomarkers and therapeutic candidates against immunologic disorders (Pragmabio, 2022) and microbiome tests as a diagnostic application in medicine and cosmetics (Atlas Biomed, 2022).

For obvious reasons not to disclose proprietary knowledge or internal processes, the companies were mostly not willing to disclose details on their use of ML. With that said, 60 companies do apply ML according to stated keywords like 'Machine Learning', 'AI', or 'Deep Learning' in a given context on their websites. More detailed information about the used algorithms were, however, normally not available. The companies offering microbiome analyses and integrating ML methods either do this as part of a sequencing service (e.g., CosmosID, www.cosmosid.com) or consider microbiome analyses as a part of a more thorough analysis. Good examples of the latter with a "microbiome-subsection" in their product portfolio are Ardigen[4] with a precision medicine service or AstarteMedical[5] with their digital tools and diagnostics to improve pediatric outcomes. A more general approach is followed by EagleGenomics[6] which offers a platform-driven whole microbiome analysis ecosystem.

4.3 Challenges of ML to consider in software development for microbiome applications

4.3.1 Bias and variance

Almost all ML approaches introduce some bias (Quinn, 2021) in the training phase, i.e., assumptions on the model "shape" and on the data distribution made during the construction of the model. When such assumptions hold, the model tends to be highly accurate, both in the training set and in the testing set, but when such assumptions are violated, such bias can lead the method to miss, ignore or discard relevant relations between descriptive features and the target feature. Approaches that exhibit a high bias can therefore lead to *underfitting*.

On the other hand, ML approaches can also generate variance errors, specifically, when they are very sensitive to small fluctuations in the training set. This issue can ultimately push the algorithm to specifically model the random noise present in the training data. When this occurs, the learned model is very accurate on the training set but poorly generalizable to the unseen data of the testing set (*overfitting*). These phenomena, in the specific context of microbiome data, have been recently emphasized in some papers (Lin and Peddada, 2020; Nearing et al., 2021; Wirbel et al., 2021).

It is noteworthy that the above-mentioned phenomena occur in almost all the application domains, not only when analyzing

4 https://ardigen.com/
5 https://astartemedical.com/
6 www.eaglegenomics.com

microbiome data, and the possible solutions tend to be common to those generally adopted in other contexts. However, since the first attempts at the adoption of ML approaches to microbiome data analysis are very recent, the context is probably not mature enough for the adoption of methods with a high bias. Solutions like multi-view learning, semi-supervised learning and transfer learning can be profitably used to alleviate such problems.

4.4 Impact of dataset size on the model accuracy

In general, the availability of large amounts of data in available repositories such as NCBI,[7] METAHIT,[8] Human Microbiome Project,[9] ExperimentHub,[10] etc., increases the chance of learning accurate ML models, and the impact of the dataset size on the model accuracy depends on the data source. However, it varies on the basis of the specific problem at hand. For example, fewer data are required if there are clear patterns within the data, if they are easily separable (in the case of classification tasks), or if simple (e.g., linear) relationships can be identified between descriptive and target attributes (in the case of regression tasks). In addition, some ML algorithms inherently require huge amounts of data due to their complexity (e.g., the number of parameters to optimize): simpler methods, such as linear regression and decision trees, typically need less training examples than solutions based on deep learning.

In microbiome research, the number of available samples is usually very limited due to sequencing costs and logistical challenges of sample collection. This aspect limits the adoption of complex approaches, although very promising according to the results achieved in other contexts. A possible solution to alleviate this issue would consist in relying on approaches that are able to exploit the knowledge coming from other contexts with huge amounts of labeled examples, such as transfer learning methods (Pio et al., 2022), or that can exploit both labeled and unlabeled examples (which may be less expensive to gather) in a semi-supervised learning setting (Chapelle et al., 2010), also based on multi-view learning (Ceci et al., 2015).

4.5 Data quality

Even when large data sets are available, there is no guarantee that the available data sample represents the whole population, without (selection or other kinds of) biases. In addition, available data sets may also include examples with (i) incorrect labels, (ii) missing or wrong values in the descriptive features, possibly due to measurement errors, (iii) highly dimensional and very sparse representation, due to the usual scarce availability of individuals with respect to the large availability of (also incomplete) generated features. The presence of one or more of such issues requires the adoption of pre-processing techniques. However, general-purpose methods may introduce additional noise or remove/discard relevant information, which suggests the need to focus on specific approaches for handling the peculiarities of microbiome data.

Another possible solution would consist in integrating multiple data sources, or in combining multiple pre-processing methods, in an ensemble or multi-view fashion. This is also confirmed by Curry et al. (2021), who states "A major source of future advancement in phenotype-prediction would be the result of discovering new data sources or feature types that have complementary predictive power, then utilizing the appropriate model structures for leveraging additional information." This approach can turn out to be effective also in the case we use features generated using existing methods (such as OTU, ASV, Metagenome-profiling, etc.) since it provides an automatic and data-driven way to merge feature contributions.

5 Interpretability and explainability

The interpretability of the results of the analysis of microbiome data is a very difficult task (Feng et al., 2015; Yu et al., 2017). In order to support this activity, the ML community is recently giving attention to the problem of model interpretability, and explainability of the predictions. This is motivated by the fact that ML models are adopted in critical decision environments, like security, health and biology, which cannot generally accept a blind output of an automated system. The importance of such an issue has been perceived even more recently, due to the general spread of neural network architectures to solve several ML tasks, which are generally very accurate but inherently not interpretable. This issue is present also in the context of microbiome data (Carrieri et al., 2021), especially when they are adopted for diagnostics purposes. Therefore, together with the design and development of accurate ML methods, able to work with sparse, high-dimensional, and noisy data, the effort of the research community should focus on the design of methods able to learn explainable models, in order to generally increase their acceptance in the biomedical field.

6 Conclusion

ML techniques are powerful methods for analyzing the huge amount of data that is being generated in the human microbiome field (Marcos-Zambrano et al., 2021; Moreno-Indias et al., 2021). As discussed in this manuscript, its application is leading to a rapid growth of specific ML tools to support and facilitate the different steps in the analysis and interpretation of microbiome data. These software developments democratize access to ML techniques, making them more accessible and easier to use for a wide range of organizations and researchers. However, the shortcomings and challenges of the ML application in human data, reviewed extensively by the COST (European Cooperation in Science and Technology) Action CA18131 on *Statistical and Machine Learning Techniques in Human Microbiome Studies* (ML4Microbiome) in Marcos-Zambrano et al. (2021) and Moreno-Indias et al. (2021), along with the fragmentation and dispersion of the ML software and microbiome data require further efforts to create federated infrastructures and services, as stated by the European Open Science Cloud (European Commission Directorate General for

7 https://www.ncbi.nlm.nih.gov/

8 https://www.gutmicrobiotaforhealth.com/metahit/

9 https://hmpdacc.org/

10 https://bioconductor.org/packages/release/bioc/html/ExperimentHub.html

Research and Innovation. and EOSC Executive Board, 2021) or ELIXIR (Balech et al., 2022), to exploit complex human microbiome data accelerating innovation, and ensuring that the benefits of ML are distributed more broadly across society, these tools can help drive progress and create a more equitable and sustainable future. Hence, ML4Microbiome contributes to this aim with a very valuable resource to microbiologists and biomedical scientists identifying and cataloguing the ML software available, facilitating and supporting the analysis and interpretation of large human microbiome datasets. This paper is part of a series of publications that emerged from the efforts of COST Action ML4 Microbiome. Other articles will address challenges (ID 1257002), data transformation (ID 1261889, ID 1250909), and best practices. The primary focus of this particular article is to gather and present a comprehensive range of ML resources and tools that are available for metagenomic analysis. In the future, benchmarking efforts by the community will be required to evaluate the performance, accessibility and user experience of these tools to provide non ML expert users with easy, transparent, and trustable standards. As the availability of methods and the vast number of workflow choices spanning unique combinations of preprocessing, feature selection, ML algorithm, parameterization, optimization, and other technical details often have remarkable effects on the analysis outcomes, the field benefits from independent benchmarking of alternative machine learning approaches. Independent competitions and community challenges provide one route for this. A recent example of this is the Heart Failure Prediction Microbiome FINRISK DREAM challenge (FINRISK, 2022), which was organized by the ML4microbiome COST action to identify optimal strategies for microbiome-based prospective risk prediction for heart failure using large-scale population cohort data sets and which results will be published soon. In addition, It will be required that software developers follow Findable, Accessible, Interoperable and Reusable (FAIR) principles for a more efficient use of resources, get more accurate results and better decision-making.

Author contributions

LM-Z and EC: conceptualization, supervision, and writing – original draft. VL-M: investigation and writing – original draft. BB-G, MF, KK-H, TK, LL, TL-T, XD, ASi, AN, GP, ASa, and VT: investigation, validation, and writing – review and editing. EI and PP: visualization, investigation, and writing – review and editing, BL-P, OA, RA, IA, ÖA, MB, MC, HD, AG, AH, EK, SK, DL, ML, PM, BN, MN, IP, LP, MP, RS, ASu, IT, C-OT, PW, EY, MY, MC, and JT: investigation and writing – review and editing. MC, JT, and EC: funding acquisition. All authors contributed to the article and approved the submitted version.

Funding

This study was supported by COST Action CA18131 "Statistical and machine learning techniques in human microbiome studies." LM-Z is supported by Spanish State Research Agency Juan de la Cierva Grant IJC2019-042188-I (LM-Z). MB is supported by Metagenopolis grant ANR-11-DPBS-0001. MLC was partially supported by the Spanish Ministry of Economy, Industry and Competitiveness, Reference PID2019-104830RB-I00.

Acknowledgments

This article is based upon work from COST Action ML4Microbiome "Statistical and machine learning techniques in human microbiome studies," CA18131, supported by COST (European Cooperation in Science and Technology), www.cost.eu.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2023.1250806/full#supplementary-material

References

Adapsyn Bioscience (2022). Available at: https://adapsyn.com/.

Al-Ajlan, A., and El Allali, A. (2019). CNN-MGP: convolutional neural networks for metagenomics gene prediction. *Interdiscip. Sci. Comput. Life Sci.* 11, 628–635. doi: 10.1007/s12539-018-0313-4

Albanese, D., Fontana, P., de Filippo, C., Cavalieri, D., and Donati, C. (2015). MICCA: a complete and accurate software for taxonomic profiling of metagenomic data. *Sci. Rep.* 5:9743. doi: 10.1038/srep09743

Alneberg, J., Bjarnason, B. S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., et al. (2014). Binning metagenomic contigs by coverage and composition. *Nat. Methods* 11, 1144–1146. doi: 10.1038/nmeth.3103

Arango-Argoty, G., Garner, E., Pruden, A., Heath, L. S., Vikesland, P., and Zhang, L. (2018). DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome* 6:23. doi: 10.1186/s40168-018-0401-z

Armour, C. R., Topçuoğlu, B. D., Garretto, A., and Schloss, P. D. (2022). A goldilocks principle for the gut microbiome: taxonomic resolution matters for microbiome-based classification of colorectal cancer. *MBio* 13, e03161–e03121. doi: 10.1128/mbio.03161-21

Arndt, D., Xia, J., Liu, Y., Zhou, Y., Guo, A. C., Cruz, J. A., et al. (2012). METAGENassist: a comprehensive web server for comparative metagenomics. *Nucleic Acids Res.* 40, W88–W95. doi: 10.1093/nar/gks497

Atlas Biomed (2022). Available at: https://atlasbiomed.com/uk.

Bakir-Gungor, B., Bulut, O., Jabeer, A., Nalbantoglu, O. U., and Yousef, M. (2021). Discovering potential taxonomic biomarkers of Type 2 diabetes from human gut microbiota via different feature selection methods. *Front. Microbiol.* 12:628426. doi: 10.3389/fmicb.2021.628426

Bakir-Gungor, B., Hacılar, H., Jabeer, A., Nalbantoglu, O. U., Aran, O., and Yousef, M. (2022). Inflammatory bowel disease biomarkers of human gut microbiota selected via different feature selection methods. *PeerJ* 10:e13205. doi: 10.7717/peerj.13205

Baldini, F., Heinken, A., Heirendt, L., Magnusdottir, S., Fleming, R. M. T., and Thiele, I. (2019). The Microbiome Modeling Toolbox: from microbial interactions to personalized microbial communities. *Bioinformatics* 35, 2332–2334. doi: 10.1093/bioinformatics/bty941

Balech, B., Brennan, L., Carrillo de Santa Pau, E., Cavalieri, D., Coort, S., D'Elia, D., et al. (2022). The future of food and nutrition in ELIXIR. *F1000Res* 11:978. doi: 10.12688/f1000research.51747.1

Bates, S., and Tibshirani, R. (2019). Log-ratio lasso: Scalable, sparse estimation for log-ratio models. *Biom. Bull.* 75, 613–624. doi: 10.1111/biom.12995

Belcour, A., Frioux, C., Aite, M., Bretaudeau, A., Hildebrand, F., and Siegel, A. (2020). Metage2Metabo, microbiota-scale metabolic complementarity for the identification of key species. *elife* 9:e61968. doi: 10.7554/eLife.61968

Bokulich, N. A., Dillon, M. R., Zhang, Y., Rideout, J. R., Bolyen, E., Li, H., et al. (2018b). q2-longitudinal: longitudinal and paired-sample analyses of microbiome data. *mSystems* 3, e00219–e00218. doi: 10.1128/mSystems.00219-18

Bokulich, N. A., Kaehler, B. D., Rideout, J. R., Dillon, M., Bolyen, E., Knight, R., et al. (2018a). Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome* 6:90. doi: 10.1186/s40168-018-0470-z

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., al-Ghalith, G. A., et al. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* 37, 852–857. doi: 10.1038/s41587-019-0209-9

Borozan, I., Watt, S., and Ferretti, V. (2015). Integrating alignment-based and alignment-free sequence similarity measures for biological sequence classification. *Bioinformatics* 31, 1396–1404. doi: 10.1093/bioinformatics/btv006

Boycott, K. M., Hartley, T., Biesecker, L. G., Gibbs, R. A., Innes, A. M., Riess, O., et al. (2019). A Diagnosis for All Rare Genetic Diseases: The Horizon and the Next Frontiers. *Cells* 177, 32–37. doi: 10.1016/j.cell.2019.02.040

Brady, A., and Salzberg, S. L. (2009). Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nat. Methods* 6, 673–676. doi: 10.1038/nmeth.1358

Cabassi, A., and Kirk, P. D. W. (2020). Multiple kernel learning for integrative consensus clustering of omic datasets. *Bioinformatics* 36, 4789–4796. doi: 10.1093/bioinformatics/btaa593

Callahan, B. J., McMurdie, P. J., and Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J.* 11, 2639–2643. doi: 10.1038/ismej.2017.119

Calle, M. L., Pujolassos, M., and Susin, A. (2023). coda4microbiome: compositional data analysis for microbiome cross-sectional and longitudinal studies. *BMC Bioinform.* 24:82. doi: 10.1186/s12859-023-05205-3

Carrieri, A. P., Haiminen, N., Maudsley-Barton, S., Gardiner, L. J., Murphy, B., Mayes, A. E., et al. (2021). Explainable AI reveals changes in skin microbiome composition linked to phenotypic differences. *Sci. Rep.* 11:4565. doi: 10.1038/s41598-021-83922-6

Ceci, M., Pio, G., Kuzmanovski, V., and Džeroski, S. (2015). Semi-supervised multi-view learning for gene network reconstruction. *PLoS One* 10:e0144031. doi: 10.1371/journal.pone.0144031

Chapelle, O., Schölkopf, B., and Zien, A.. (2010). *Semi-Supervised Learning. 2nd* Edn Cambridge, Massachusetts. London, England: The MIT Press.

Chen, W., Zhang, C. K., Cheng, Y., Zhang, S., and Zhao, H. (2013). A Comparison of methods for clustering 16S rRNA sequences into OTUs. *PLoS One* 8:e70837. doi: 10.1371/journal.pone.0070837

Cheng, L., Walker, A. W., and Corander, J. (2012). Bayesian estimation of bacterial community composition from 454 sequencing data. *Nucleic Acids Res.* 40, 5240–5249. doi: 10.1093/nar/gks227

Chiarello, M., McCauley, M., Villéger, S., and Jackson, C. R. (2022). Ranking the biases: The choice of OTUs vs. ASVs in 16S rRNA amplicon data analysis has stronger effects on diversity measures than rarefaction and OTU identity threshold. *PLoS One* 17:e0264443. doi: 10.1371/journal.pone.0264443

Chroneos, Z. C. (2010). Metagenomics: Theory, methods, and applications. *Hum. Genomics* 4:282. doi: 10.1186/1479-7364-4-4-282

Coenders, G., and Greenacre, M. (2022). Three approaches to supervised learning for compositional data with pairwise logratios. *J. Appl. Stat.*, 1–22. doi: 10.1080/02664763.2022.2108007

Cole, J. R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R. J., et al. (2009). The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res.* 37, D141–D145. doi: 10.1093/nar/gkn879

Cui, H., and Zhang, X. (2013). Alignment-free supervised classification of metagenomes by recursive SVM. *BMC Genomics* 14:641. doi: 10.1186/1471-2164-14-641

Curry, K. D., Nute, M. G., and Treangen, T. J. (2021). It takes guts to learn: machine learning techniques for disease detection from the gut microbiome. *Emerg. Topics Life Sci.* 5, 815–827. doi: 10.1042/ETLS20210213

de Jesus, V. C., Khan, M. W., Mittermuller, B. A., Duan, K., Hu, P., Schroth, R. J., et al. (2021). Characterization of supragingival plaque and oral swab microbiomes in children with severe early childhood caries. *Front. Microbiol.* 12:683685. doi: 10.3389/fmicb.2021.683685

de Nies, L., Lopes, S., Busi, S. B., Galata, V., Heintz-Buschart, A., Laczny, C. C., et al. (2021). PathoFact: a pipeline for the prediction of virulence factors and antimicrobial resistance genes in metagenomic data. *Microbiome* 9:49. doi: 10.1186/s40168-020-00993-9

Diener, C., Gibbons, S. M., and Resendis-Antonio, O. (2020). MICOM: metagenome-scale modeling to infer metabolic interactions in the gut microbiota. *mSystems* 5, e00606–e00619. doi: 10.1128/mSystems.00606-19

Dietrich, A., Matchado, M. S., Zwiebel, M., Ölke, B., Lauber, M., Lagkouvardos, I., et al. (2022). Namco: a microbiome explorer. *Microb. Genom.* 8:mgen000852. doi: 10.1099/mgen.0.000852

Ding, X., Cheng, F., Cao, C., and Sun, X. (2015). DectICO: an alignment-free supervised metagenomic classification method based on feature extraction and dynamic selection. *BMC Bioinform.* 16:323. doi: 10.1186/s12859-015-0753-3

Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern classification. 2nd* Edn Hoboken, New Jersey, U.S.: Wiley.

Ebrahim, A., Lerman, J. A., Palsson, B. O., and Hyduke, D. R. (2013). COBRApy: COnstraints-based reconstruction and analysis for python. *BMC Syst. Biol.* 7:74. doi: 10.1186/1752-0509-7-74

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26, 2460–2461. doi: 10.1093/bioinformatics/btq461

Edgar, R. C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat. Methods* 10, 996–998. doi: 10.1038/nmeth.2604

Eren, A. M., Maignien, L., Sul, W. J., Murphy, L. G., Grim, S. L., Morrison, H. G., et al. (2013). Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods Ecol. Evol.* 4, 1111–1119. doi: 10.1111/2041-210X.12114

European Commission Directorate General for Research and Innovation. and EOSC Executive Board (2021). EOSC interoperability framework: report from the EOSC Executive Board Working Groups FAIR and Architecture. Publications Office.

Faust, K., Bauchinger, F., Laroche, B., de Buyl, S., Lahti, L., Washburne, A. D., et al. (2018). Signatures of ecological processes in microbial community time series. *Microbiome* 6:120. doi: 10.1186/s40168-018-0496-2

Feng, Q., Liang, S., Jia, H., Stadlmayr, A., Tang, L., Lan, Z., et al. (2015). Gut microbiome development along the colorectal adenoma–carcinoma sequence. *Nat. Commun.* 6:6528. doi: 10.1038/ncomms7528

Fernandes, A. D., Macklaim, J. M., Linn, T. G., Reid, G., and Gloor, G. B. (2013). ANOVA-Like differential expression (ALDEx) analysis for mixed population RNA-Seq. *PLoS One* 8:e67019. doi: 10.1371/journal.pone.0067019

Fernandes, A. D., Reid, J. N. S., Macklaim, J. M., McMurrough, T. A., Edgell, D. R., and Gloor, G. B. (2014). Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome* 2:15. doi: 10.1186/2049-2618-2-15

Fierer, N., Lauber, C. L., Zhou, N., McDonald, D., Costello, E. K., and Knight, R. (2010). Forensic identification using skin bacterial communities. *Proc. Natl. Acad. Sci. U. S. A.* 107, 6477–6481. doi: 10.1073/pnas.1000162107

FINRISK (2022). Heart failure and microbiome.

Gao, X., Lin, H., Dong, Q., Rho, M., and Wang, L. (2017). A dirichlet-multinomial bayes classifier for disease diagnosis with microbial compositions. *mSphere* 2, e00536–e00517. doi: 10.1128/mSphereDirect.00536-17

García-Jiménez, B., Muñoz, J., Cabello, S., Medina, J., and Wilkinson, M. D. (2021). Predicting microbiomes through a deep latent space. *Bioinformatics* 37, 1444–1451. doi: 10.1093/bioinformatics/btaa971

Ghannam, R. B., and Techtmann, S. M. (2021). Machine learning applications in microbial ecology, human microbiome studies, and environmental monitoring. *Comput. Struct. Biotechnol. J.* 19, 1092–1107. doi: 10.1016/j.csbj.2021.01.028

Ghodsi, M., Liu, B., and Pop, M. (2011). DNACLUST: accurate and efficient clustering of phylogenetic marker genes. *BMC Bioinform.* 12:271. doi: 10.1186/1471-2105-12-271

Gloor, G. B., Macklaim, J. M., and Fernandes, A. D. (2016). Displaying variation in large datasets: plotting a visual summary of effect sizes. *J. Comput. Graph. Stat.* 25, 971–979. doi: 10.1080/10618600.2015.1131161

Goodswen, S. J., Barratt, J. L. N., Kennedy, P. J., Kaufer, A., Calarco, L., and Ellis, J. T. (2021). Machine learning and applications in microbiology. *FEMS Microbiol. Rev.* 45:fuab015. doi: 10.1093/femsre/fuab015

Gordon-Rodriguez, E., Quinn, T. P., and Cunningham, J. P. (2021). Learning sparse log-ratios for high-throughput sequencing data. *Bioinformatics* 38, 157–163. doi: 10.1093/bioinformatics/btab645

Hai Nguyen, T., et al. (2019). "Disease Prediction Using Synthetic Image Representations of Metagenomic Data and Convolutional Neural Networks." in *2019 IEEE-RIVF International Conference on Computing and Communication Technologies (RIVF), Danang, Vietnam*. pp. 1–6

Hao, X., Jiang, R., and Chen, T. (2011). Clustering 16S rRNA for OTU prediction: a method of unsupervised Bayesian clustering. *Bioinformatics* 27, 611–618. doi: 10.1093/bioinformatics/btq725

Heinken, A., Basile, A., and Thiele, I. (2021). Advances in constraint-based modelling of microbial communities. *Curr. Opin. Syst. Biol.* 27:100346. doi: 10.1016/j.coisb.2021.05.007

Heinken, A., and Thiele, I. (2022). Microbiome Modelling Toolbox 2.0: efficient, tractable modelling of microbiome communities. *Bioinformatics* 38, 2367–2368. doi: 10.1093/bioinformatics/btac082

Heinken, A., Acharya, G., Ravcheev, D. A., Hertel, J., Nyga, M., Okpala, O. E., et al. (2020). AGORA2: Large scale reconstruction of the microbiome highlights wide-spread drug-metabolising capacities. *Syst. Biol.* doi: 10.1101/2020.11.09.375451

Heirendt, L., Arreckx, S., Pfau, T., Mendoza, S. N., Richelle, A., Heinken, A., et al. (2019). Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0. *Nat. Protoc.* 14, 639–702. doi: 10.1038/s41596-018-0098-2

Henry, C. S., DeJongh, M., Best, A. A., Frybarger, P. M., Linsay, B., and Stevens, R. L. (2010). High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat. Biotechnol.* 28, 977–982. doi: 10.1038/nbt.1672

Hickl, O., Queirós, P., Wilmes, P., May, P., and Heintz-Buschart, A. (2022). Binny: an automated binning algorithm to recover high-quality genomes from complex metagenomic datasets. *Brief. Bioinform.* 23:bbac431. doi: 10.1093/bib/bbac431

Ho, Tin Kam (1995). "Random decision forests." in *Proceedings of 3rd International Conference on Document Analysis and Recognition, Montreal, Que., Canada*. 1, pp. 278–282

Hoarfrost, A., Aptekmann, A., Farfañuk, G., and Bromberg, Y. (2022). Deep learning of a bacterial and archaeal universal language of life enables transfer learning and illuminates microbial dark matter. *Nat. Commun.* 13:2606. doi: 10.1038/s41467-022-30070-8

Hoff, K. J., Lingner, T., Meinicke, P., and Tech, M. (2009). Orphelia: predicting genes in metagenomic sequencing reads. *Nucleic Acids Res.* 37, W101–W105. doi: 10.1093/nar/gkp327

Hoff, K. J., Tech, M., Lingner, T., Daniel, R., Morgenstern, B., and Meinicke, P. (2008). Gene prediction in metagenomic fragments: a large scale machine learning approach. *BMC Bioinform.* 9:217. doi: 10.1186/1471-2105-9-217

Holmes, I., Harris, K., and Quince, C. (2012). Dirichlet multinomial mixtures: generative models for microbial metagenomics. *PLoS One* 7:e30126. doi: 10.1371/journal.pone.0030126

Huse, S. M., Welch, D. M., Morrison, H. G., and Sogin, M. L. (2010). Ironing out the wrinkles in the rare biosphere through improved OTU clustering: Ironing out the wrinkles in the rare biosphere. *Environ. Microbiol.* 12, 1889–1898. doi: 10.1111/j.1462-2920.2010.02193.x

Jääskinen, V., Parkkinen, V., Cheng, L., and Corander, J. (2014). Bayesian clustering of DNA sequences using Markov chains and a stochastic partition model. *Stat. Appl. Genet. Mol. Biol.* 13, 105–121. doi: 10.1515/sagmb-2013-0031

Jin, B. T., Xu, F., Ng, R. T., and Hogg, J. C. (2022). Mian: interactive web-based microbiome data table visualization and machine learning platform. *Bioinformatics* 38, 1176–1178. doi: 10.1093/bioinformatics/btab754

Kaehler, B. D., Bokulich, N. A., McDonald, D., Knight, R., Caporaso, J. G., and Huttley, G. A. (2019). Species abundance information improves sequence taxonomy classification accuracy. *Nat. Commun.* 10:4643. doi: 10.1038/s41467-019-12669-6

Kariin, S., and Burge, C. (1995). Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet.* 11, 283–290. doi: 10.1016/S0168-9525(00)89076-9

Karlsson, F. H., Tremaroli, V., Nookaew, I., Bergström, G., Behre, C. J., Fagerberg, B., et al. (2013). Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature* 498, 99–103. doi: 10.1038/nature12198

Karp, P. D., Latendresse, M., Paley, S. M., Krummenacker, M., Ong, Q. D., Billington, R., et al. (2016). Pathway Tools version 19.0 update: software for pathway/genome informatics and systems biology. *Brief. Bioinform.* 17, 877–890. doi: 10.1093/bib/bbv079

Kartal, E., Schmidt, T. S. B., Molina-Montes, E., Rodríguez-Perales, S., Wirbel, J., Maistrenko, O. M., et al. (2022). A faecal microbiota signature with high specificity for pancreatic cancer. *Gut* 71, 1359–1372. doi: 10.1136/gutjnl-2021-324755

Keilwagen, J., Hartung, F., and Grau, J. (2019). "GeMoMa: Homology-Based Gene Prediction Utilizing Intron Position Conservation and RNA-seq Data" in *Gene Prediction 1962*. ed. M. Kollmar (New York: Springer), 161–177.

Kelley, D. R., Liu, B., Delcher, A. L., Pop, M., and Salzberg, S. L. (2012). Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering. *Nucleic Acids Res.* 40:e9. doi: 10.1093/nar/gkr1067

Lapp, Z., Han, J. H., Wiens, J., Goldstein, E. J. C., Lautenbach, E., and Snitkin, E. S. (2021). Patient and microbial genomic factors associated with carbapenem-resistant *Klebsiella pneumoniae* extraintestinal colonization and infection. *mSystems* 6, e00177–e00121. doi: 10.1128/mSystems.00177-21

Larsen, P. E., Collart, F. R., Field, D., Meyer, F., Keegan, K. P., Henry, C. S., et al. (2011). Predicted Relative Metabolomic Turnover (PRMT): determining metabolic turnover from a coastal marine metagenomic dataset. *Microb. Informat. Exp.* 1:4. doi: 10.1186/2042-5783-1-4

le Chatelier, E., Nielsen, T., Qin, J., Prifti, E., Hildebrand, F., Falony, G., et al. (2013). Richness of human gut microbiome correlates with metabolic markers. *Nature* 500, 541–546. doi: 10.1038/nature12506

Lee, K. A., Thomas, A. M., Bolte, L. A., Björk, J. R., de Ruijter, L. K., Armanini, F., et al. (2022). Cross-cohort gut microbiome associations with immune checkpoint inhibitor response in advanced melanoma. *Nat. Med.* 28, 535–544. doi: 10.1038/s41591-022-01695-5

Lesniak, N. A., Schubert, A. M., Flynn, K. J., Leslie, J. L., Sinani, H., Bergin, I. L., et al. (2022). The gut bacterial community potentiates clostridioides difficile infection severity. *MBio* 13, e01183–e01122. doi: 10.1128/mbio.01183-22

Lewin, H. A., Robinson, G. E., Kress, W. J., Baker, W. J., Coddington, J., Crandall, K. A., et al. (2018). Earth BioGenome Project: Sequencing life for the future of life. *Proc. Natl. Acad. Sci. U. S. A.* 115, 4325–4333. doi: 10.1073/pnas.1720115115

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659. doi: 10.1093/bioinformatics/btl158

Li, W., Jaroszewski, L., and Godzik, A. (2001). Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* 17, 282–283. doi: 10.1093/bioinformatics/17.3.282

Liang, Q., Bible, P. W., Liu, Y., Zou, B., and Wei, L. (2020). DeepMicrobes: taxonomic classification for metagenomics with deep learning. *NAR Genom. Bioinform.* 2:lqaa009. doi: 10.1093/nargab/lqaa009

Lin, H., Eggesbø, M., and Peddada, S. D. (2022). Linear and nonlinear correlation estimators unveil undescribed taxa interactions in microbiome data. *Nat. Commun.* 13:4946. doi: 10.1038/s41467-022-32243-x

Lin, H., and Peddada, S. D. (2020). Analysis of compositions of microbiomes with bias correction. *Nat. Commun.* 11:3514. doi: 10.1038/s41467-020-17041-7

Lindahl, B. D., Nilsson, R. H., Tedersoo, L., Abarenkov, K., Carlsen, T., Kjøller, R., et al. (2013). Fungal community analysis by high-throughput sequencing of amplified markers – a user's guide. *New Phytol.* 199, 288–299. doi: 10.1111/nph.12243

Liu, C.-C., Dong, S. S., Chen, J. B., Wang, C., Ning, P., Guo, Y., et al. (2022). MetaDecoder: a novel method for clustering metagenomic contigs. *Microbiome* 10:46. doi: 10.1186/s40168-022-01237-8

Liu, Y., Guo, J., Hu, G., and Zhu, H. (2013). Gene prediction in metagenomic fragments based on the SVM algorithm. *BMC Bioinform.* 14:S12. doi: 10.1186/1471-2105-14-S5-S12

Liu, Z., Hsiao, W., Cantarel, B. L., Drábek, E. F., and Fraser-Liggett, C. (2011). Sparse distance-based learning for simultaneous multiclass classification and feature selection of metagenomic data. *Bioinformatics* 27, 3242–3249. doi: 10.1093/bioinformatics/btr547

Liu, B., Sträuber, H., Saraiva, J., Harms, H., Silva, S. G., Kasmanas, J. C., et al. (2022). Machine learning-assisted identification of bioindicators predicts medium-chain carboxylate production performance of an anaerobic mixed culture. *Microbiome* 10:48. doi: 10.1186/s40168-021-01219-2

Liu, S., Zhao, W., Liu, X., and Cheng, L. (2020). Metagenomic analysis of the gut microbiome in atherosclerosis patients identify cross-cohort microbial signatures and potential therapeutic target. *FASEB J.* 34, 14166–14181. doi: 10.1096/fj.202000622R

Lo, C., and Marculescu, R. (2019). MetaNN: accurate classification of host phenotypes from metagenomic data using neural networks. *BMC Bioinform.* 20:314. doi: 10.1186/s12859-019-2833-2

Lüll, K., Arffman, R. K., Sola-Leyva, A., Molina, N. M., Aasmets, O., Herzig, K. H., et al. (2021). The gut microbiome in polycystic ovary syndrome and its association with metabolic traits. *J. Clin. Endocrinol. Metab.* 106, 858–871. doi: 10.1210/clinem/dgaa848

Lundberg, S., and Lee, S.-I. (2017), A Unified Approach to Interpreting Model Predictions.

Ma, H., Tan, T. W., and Ban, K. H. K. (2021). A multi-task CNN learning model for taxonomic assignment of human viruses. *BMC Bioinform.* 22:194. doi: 10.1186/s12859-021-04084-w

Magnúsdóttir, S., Heinken, A., Kutt, L., Ravcheev, D. A., Bauer, E., Noronha, A., et al. (2017). Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. *Nat. Biotechnol.* 35, 81–89. doi: 10.1038/nbt.3703

Mahé, F., Rognes, T., Quince, C., de Vargas, C., and Dunthorn, M. (2014). Swarm: robust and fast clustering method for amplicon-based studies. *PeerJ* 2:e593. doi: 10.7717/peerj.593

Mallick, H., Franzosa, E. A., Mclver, L. J., Banerjee, S., Sirota-Madi, A., Kostic, A. D., et al. (2019). Predictive metabolomic profiling of microbial communities using amplicon or metagenomic sequences. *Nat. Commun.* 10:3136. doi: 10.1038/s41467-019-10927-1

Marcos-Zambrano, L. J., Karaduzovic-Hadziabdic, K., Loncar Turukalo, T., Przymus, P., Trajkovik, V., Aasmets, O., et al. (2021). Applications of Machine Learning in Human Microbiome Studies: A Review on Feature Selection, Biomarker Identification. *Front. Virol.* 12:634511. doi: 10.3389/fmicb.2021.634511

Mariette, J., and Villa-Vialaneix, N. (2018). Unsupervised multiple kernel learning for heterogeneous data integration. *Bioinformatics* 34, 1009–1015. doi: 10.1093/bioinformatics/btx682

Matsen, F. A., Kodner, R. B., and Armbrust, E. V. (2010). pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinform.* 11:538. doi: 10.1186/1471-2105-11-538

McDonald, D., Hyde, E., Debelius, J. W., Morton, J. T., Gonzalez, A., Ackermann, G., et al. (2018). American Gut: an Open Platform for Citizen Science Microbiome Research. *mSystems* 3, e00031–e00018. doi: 10.1128/mSystems.00031-18

McHardy, A. C., Martín, H. G., Tsirigos, A., Hugenholtz, P., and Rigoutsos, I. (2007). Accurate phylogenetic classification of variable-length DNA fragments. *Nat. Methods* 4, 63–72. doi: 10.1038/nmeth976

Mendes-Soares, H., Mundy, M., Soares, L. M., and Chia, N. (2016). MMinte: an application for predicting metabolic interactions among the microbial species in a community. *BMC Bioinform.* 17:343. doi: 10.1186/s12859-016-1230-3

Microbiome Employers (2022). Digital World Biology.

Montassier, E., al-Ghalith, G. A., Ward, T., Corvec, S., Gastinne, T., Potel, G., et al. (2016). Pretreatment gut microbiome predicts chemotherapy-related bloodstream infection. *Genome Med.* 8:49. doi: 10.1186/s13073-016-0301-4

Moreno-Indias, I., Lahti, L., Nedyalkova, M., Elbere, I., Roshchupkin, G., Adilovic, M., et al. (2021). Statistical and Machine Learning Techniques in Human Microbiome Studies: Contemporary Challenges and Solutions. *Front. Microbiol.* 12:635781. doi: 10.3389/fmicb.2021.635781

Nagpal, S., Singh, R., Taneja, B., and Mande, S. S. (2022). MarkerML – marker feature identification in metagenomic datasets using interpretable machine learning. *J. Mol. Biol.* 434:167589. doi: 10.1016/j.jmb.2022.167589

Nearing, J. T., Comeau, A. M., and Langille, M. G. I. (2021). Identifying biases and their potential solutions in human microbiome studies. *Microbiome* 9:113. doi: 10.1186/s40168-021-01059-0

Nguyen, N.-P., Warnow, T., Pop, M., and White, B. (2016). A perspective on 16S rRNA operational taxonomic unit clustering using sequence similarity. *NPJ Biofilms Microbiomes* 2:16004. doi: 10.1038/npjbiofilms.2016.4

Nissen, J. N., Johansen, J., Allesøe, R. L., Sønderby, C. K., Armenteros, J. J. A., Grønbech, C. H., et al. (2021). Improved metagenome binning and assembly using deep variational autoencoders. *Nat. Biotechnol.* 39, 555–560. doi: 10.1038/s41587-020-00777-4

Noecker, C., Eng, A., Srinivasan, S., Theriot, C. M., Young, V. B., Jansson, J. K., et al. (2016). Metabolic model-based integration of microbiome taxonomic and metabolomic profiles elucidates mechanistic links between ecological and metabolic variation. *mSystems* 1, e00013–e00015. doi: 10.1128/mSystems.00013-15

Noguchi, H., Park, J., and Takagi, T. (2006). MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Res.* 34, 5623–5630. doi: 10.1093/nar/gkl723

Noguchi, H., Taniguchi, T., and Itoh, T. (2008). MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. *DNA Res.* 15, 387–396. doi: 10.1093/dnares/dsn027

Oh, M., and Zhang, L. (2020). DeepMicro: deep representation learning for disease prediction based on microbiome data. *Sci. Rep.* 10:6026. doi: 10.1038/s41598-020-63159-5

Orellana, S. C. (2013). Assessment of fungal diversity in the environment using metagenomics:a decade in review. *Fungal Genom Biol* 3, 1–13. doi: 10.4172/2165-8056.1000110

Orth, J. D., Thiele, I., and Palsson, B. Ø. (2010). What is flux balance analysis? *Nat. Biotechnol.* 28, 245–248. doi: 10.1038/nbt.1614

Pan, S., Zhu, C., Zhao, X. M., and Coelho, L. P. (2022). A deep siamese neural network improves metagenome-assembled genomes in microbiome datasets across different environments. *Nat. Commun.* 13:2326. doi: 10.1038/s41467-022-29843-y

Parks, D. H., MacDonald, N. J., and Beiko, R. G. (2011). Classifying short genomic fragments from novel lineages using composition and homology. *BMC Bioinform.* 12:328. doi: 10.1186/1471-2105-12-328

Pasolli, E., Truong, D. T., Malik, F., Waldron, L., and Segata, N. (2016). Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. *PLoS Comput. Biol.* 12:e1004977. doi: 10.1371/journal.pcbi.1004977

Patil, K. R., Roune, L., and McHardy, A. C. (2012). The PhyloPythiaS Web server for taxonomic assignment of metagenome sequences. *PLoS One* 7:e38581. doi: 10.1371/journal.pone.0038581

Picard, M., Scott-Boyer, M. P., Bodein, A., Périn, O., and Droit, A. (2021). Integration strategies of multi-omics data for machine learning analysis. *Comput. Struct. Biotechnol. J.* 19, 3735–3746. doi: 10.1016/j.csbj.2021.06.030

Pio, G., Mignone, P., Magazzù, G., Zampieri, G., Ceci, M., and Angione, C. (2022). Integrating genome-scale metabolic modelling and transfer learning for human gene regulatory network reconstruction. *Bioinformatics* 38, 487–493. doi: 10.1093/bioinformatics/btab647

Pragmabio (2022). Available at: http://www.pragmabio.com/.

Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* 490, 55–60. doi: 10.1038/nature11450

Qin, N., Yang, F., Li, A., Prifti, E., Chen, Y., Shao, L., et al. (2014). Alterations of the human gut microbiome in liver cirrhosis. *Nature* 513, 59–64. doi: 10.1038/nature13568

Queirós, P., Delogu, F., Hickl, O., May, P., and Wilmes, P. (2021). Mantis: flexible and consensus-driven genome annotation. *GigaScience* 10:giab042. doi: 10.1093/gigascience/giab042

Quinn, T.P. (2021) Stool Studies Don't Pass the Sniff Test: A Systematic Review of Human Gut Microbiome Research Suggests Widespread Misuse of Machine LearningarXiv.

Quinn, T. P., and Erb, I. (2020). Amalgams: data-driven amalgamation for the dimensionality reduction of compositional data. *NAR Genom. Bioinform* 2:lqaa076. doi: 10.1093/nargab/lqaa076

Rahman, M.A., and Rangwala, H. (2018). "RegMIL: Phenotype Classification from Metagenomic Data." in *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, Washington DC USA*. pp. 145–154

Ramon, E., Belanche-Muñoz, L., Molist, F., Quintanilla, R., Perez-Enciso, M., and Ramayo-Caldas, Y. (2021). kernInt: A Kernel Framework for Integrating Supervised and Unsupervised Analyses in Spatio-Temporal Metagenomic Datasets. *Front. Microbiol.* 12:609048. doi: 10.3389/fmicb.2021.609048

Rasheed, Z., and Rangwala, H. (2012). Metagenomic taxonomic classification using extreme learning machines. *J. Bioinforma. Comput. Biol.* 10:1250015. doi: 10.1142/S0219720012500151

Reiman, D., Layden, B. T., and Dai, Y. (2021). MiMeNet: exploring microbiome-metabolome relationships using neural networks. *PLoS Comput. Biol.* 17:e1009021. doi: 10.1371/journal.pcbi.1009021

Reiman, D., Metwally, A. A., Sun, J., and Dai, Y. (2020). PopPhy-CNN: A phylogenetic tree embedded architecture for convolutional neural networks to predict host phenotype from metagenomic data. *IEEE J. Biomed. Health Inform.* 24, 2993–3001. doi: 10.1109/JBHI.2020.2993761

Ren, J., Song, K., Deng, C., Ahlgren, N. A., Fuhrman, J. A., Li, Y., et al. (2020). Identifying viruses from metagenomic data using deep learning. *Quant. Biol.* 8, 64–77. doi: 10.1007/s40484-019-0187-4

Rho, M., Tang, H., and Ye, Y. (2010). FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res.* 38:e191. doi: 10.1093/nar/gkq747

Rivera-Pinto, J., Egozcue, J. J., Pawlowsky-Glahn, V., Paredes, R., Noguera-Julian, M., and Calle, M. L. (2018). Balances: a New perspective for microbiome analysis. *mSystems* 3, e00053–e00018. doi: 10.1128/mSystems.00053-18

Rognes, T., Flouri, T., Nichols, B., Quince, C., and Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ* 4:e2584. doi: 10.7717/peerj.2584

Rohart, F., Eslami, A., Matigian, N., Bougeard, S., and Lê Cao, K. A. (2017b). MINT: a multivariate integrative method to identify reproducible molecular signatures across independent experiments and platforms. *BMC Bioinform.* 18:128. doi: 10.1186/s12859-017-1553-8

Rohart, F., Gautier, B., Singh, A., and Lê Cao, K. A. (2017a). mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Comput. Biol.* 13:e1005752. doi: 10.1371/journal.pcbi.1005752

Rosen, G. L., Reichenberger, E. R., and Rosenfeld, A. M. (2011). NBC: the Naive Bayes Classification tool webserver for taxonomic classification of metagenomic reads. *Bioinformatics* 27, 127–129. doi: 10.1093/bioinformatics/btq619

Röttjers, L., Vandeputte, D., Raes, J., and Faust, K. (2021). Null-model-based network comparison reveals core associations. *ISME Commun.* 1:36. doi: 10.1038/s43705-021-00036-w

Roux, S., Faubladier, M., Mahul, A., Paulhe, N., Bernard, A., Debroas, D., et al. (2011). Metavir: a web server dedicated to virome analysis. *Bioinformatics* 27, 3074–3075. doi: 10.1093/bioinformatics/btr519

Russell, D. J., Way, S. F., Benson, A. K., and Sayood, K. (2010). A grammar-based distance metric enables fast and accurate clustering of large sets of 16S sequences. *BMC Bioinform.* 11:601. doi: 10.1186/1471-2105-11-601

Sarker, I. H. (2021). Machine learning: algorithms, real-world applications and research directions. *SN Comput. Sci.* 2:160. doi: 10.1007/s42979-021-00592-x

Schloss, P. D., and Handelsman, J. (2005). Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Appl. Environ. Microbiol.* 71, 1501–1506. doi: 10.1128/AEM.71.3.1501-1506.2005

Schloss, P. D., and Westcott, S. L. (2011). Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis. *Appl. Environ. Microbiol.* 77, 3219–3226. doi: 10.1128/AEM.02810-10

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., et al. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* 75, 7537–7541. doi: 10.1128/AEM.01541-09

Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W. S., et al. (2011). Metagenomic biomarker discovery and explanation. *Genome Biol.* 12:R60. doi: 10.1186/gb-2011-12-6-r60

Shang, J., and Sun, Y. (2021). CHEER: HierarCHical taxonomic classification for viral mEtagEnomic data via deep leaRning. *Methods* 189, 95–103. doi: 10.1016/j.ymeth.2020.05.018

Sharpton, T. J. (2014). An introduction to the analysis of shotgun metagenomic data. *Front. Plant Sci.* 5:209. doi: 10.3389/fpls.2014.00209

Singh, A., Shannon, C. P., Gautier, B., Rohart, F., Vacher, M., Tebbutt, S. J., et al. (2019). DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics* 35, 3055–3062. doi: 10.1093/bioinformatics/bty1054

Sokol, H., Leducq, V., Aschard, H., Pham, H. P., Jegou, S., Landman, C., et al. (2017). Fungal microbiota dysbiosis in IBD. *Gut* 66, 1039–1048. doi: 10.1136/gutjnl-2015-310746

Sommer, M. J., and Salzberg, S. L. (2021). Balrog: a universal protein model for prokaryotic gene prediction. *PLoS Comput. Biol.* 17:e1008727. doi: 10.1371/journal.pcbi.1008727

Soueidan, H., and Nikolski, M. (2016). Machine learning for metagenomics: methods and tools. arXiv

Stunnenberg, H. G., Hirst, M., Abrignani, S., Adams, D., de Almeida, M., Altucci, L., et al. (2016). The international human epigenome consortium: a blueprint for scientific collaboration and discovery. *Cells* 167, 1145–1149. doi: 10.1016/j.cell.2016.11.007

Sun, Y., Cai, Y., Liu, L., Yu, F., Farrell, M. L., McKendree, W., et al. (2009). ESPRIT: estimating species richness using large collections of 16S rRNA pyrosequences. *Nucleic Acids Res.* 37:e76. doi: 10.1093/nar/gkp285

Tampuu, A., Bzhalava, Z., Dillner, J., and Vicente, R. (2019). ViraMiner: Deep learning on raw DNA sequences for identifying viral genomes in human samples. *PLoS One* 14:e0222271. doi: 10.1371/journal.pone.0222271

Tanaseichuk, O., Borneman, J., and Jiang, T. (2014). Phylogeny-based classification of microbial communities. *Bioinformatics* 30, 449–456. doi: 10.1093/bioinformatics/btt700

The 1000 Genomes Project ConsortiumAuton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., et al. (2015). A global reference for human genetic variation. *Nature* 526, 68–74. doi: 10.1038/nature15393

The Human Microbiome Project Consortium (2012). Structure, function and diversity of the healthy human microbiome. *Nature* 486, 207–214. doi: 10.1038/nature11234

Thiele, I., Heinken, A., and Fleming, R. M. T. (2013). A systems biology approach to studying the role of microbes in human health. *Curr. Opin. Biotechnol.* 24, 4–12. doi: 10.1016/j.copbio.2012.10.001

Thomas, A. M., Manghi, P., Asnicar, F., Pasolli, E., Armanini, F., Zolfo, M., et al. (2019). Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nat. Med.* 25, 667–678. doi: 10.1038/s41591-019-0405-7

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *J. R. Stat. Soc. Ser. B Methodol.* 58, 267–288.

Topçuoğlu, B., Lapp, Z., Sovacool, K., Snitkin, E., Wiens, J., and Schloss, P. (2021). mikropml: user-friendly R package for supervised machine learning pipelines. *JOSS* 6:3073. doi: 10.21105/joss.03073

Uritskiy, G. V., DiRuggiero, J., and Taylor, J. (2018). MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* 6:158. doi: 10.1186/s40168-018-0541-1

Wang, Q., Garrity, G. M., Tiedje, J. M., and Cole, J. R. (2007). Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. *Appl. Environ. Microbiol.* 73, 5261–5267. doi: 10.1128/AEM.00062-07

Wang, Z., Wang, Z., Lu, Y. Y., Sun, F., and Zhu, S. (2019). SolidBin: improving metagenome binning with semi-supervised normalized cut. *Bioinformatics* 35, 4229–4238. doi: 10.1093/bioinformatics/btz253

Wang, X., Yao, J., Sun, Y., and Mai, V. (2013). M-pick, a modularity-based method for OTU picking of 16S rRNA sequences. *BMC Bioinform.* 14:43. doi: 10.1186/1471-2105-14-43

Wei, Z.-G., and Zhang, S.-W. (2015). MtHc: a motif-based hierarchical method for clustering massive 16S rRNA sequences into OTUs. *Mol. BioSyst.* 11, 1907–1913. doi: 10.1039/C5MB00089K

Wei, Z.-G., Zhang, X. D., Cao, M., Liu, F., Qian, Y., and Zhang, S. W. (2021). Comparison of methods for picking the operational taxonomic units from amplicon sequences. *Front. Microbiol.* 12:644012. doi: 10.3389/fmicb.2021.644012

Wei, Z.-G., Zhang, S. W., and Zhang, Y. Z. (2017). DMclust, a Density-based Modularity Method for Accurate OTU Picking of 16S rRNA Sequences. *QSAR Comb. Sci.* 36:1600059. doi: 10.1002/minf.201600059

Westcott, S. L., Schloss, P. D., Watson, M., and Pollard, K. (2017). OptiClust, an improved method for assigning amplicon-based sequence data to operational taxonomic units. *mSphere* 2, e00073–e00017. doi: 10.1128/mSphereDirect.00073-17

White, J. R., Navlakha, S., Nagarajan, N., Ghodsi, M. R., Kingsford, C., and Pop, M. (2010). Alignment and clustering of phylogenetic markers - implications for microbial diversity studies. *BMC Bioinform.* 11:152. doi: 10.1186/1471-2105-11-152

Wirbel, J., Zych, K., Essex, M., Karcher, N., Kartal, E., Salazar, G., et al. (2021). Microbiome meta-analysis and cross-disease comparison enabled by the SIAMCAT machine learning toolbox. *Genome Biol.* 22:93. doi: 10.1186/s13059-021-02306-1

Wu, G. D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y. Y., Keilbaugh, S. A., et al. (2011). Linking long-term dietary patterns with gut microbial enterotypes. *Science* 334, 105–108. doi: 10.1126/science.1208344

Wu, Y.-W., Simmons, B. A., and Singer, S. W. (2016). MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* 32, 605–607. doi: 10.1093/bioinformatics/btv638

Yadav, M., and Chauhan, N. S. (2022). Role of gut-microbiota in disease severity and clinical outcomes. *Brief. Funct. Genomics*. 24:elac037. doi: 10.1093/bfgp/elac037

Yang, C., Chowdhury, D., Zhang, Z., Cheung, W. K., Lu, A., Bian, Z., et al. (2021). A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data. *Comput. Struct. Biotechnol. J.* 19, 6301–6314. doi: 10.1016/j.csbj.2021.11.028

Yang, F., and Zou, Q. (2020). mAML: an automated machine learning pipeline with a microbiome repository for human disease classification. *Database (Oxford)* 2020:baaa050. doi: 10.1093/database/baaa050

Yin, X., Altman, T., Rutherford, E., West, K. A., Wu, Y., Choi, J., et al. (2020). A comparative evaluation of tools to predict metabolite profiles from microbiome sequencing data. *Front. Microbiol.* 11:595910. doi: 10.3389/fmicb.2020.595910

Yu, J., Feng, Q., Wong, S. H., Zhang, D., Liang, Q., Qin, Y., et al. (2017). Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. *Gut* 66, 70–78. doi: 10.1136/gutjnl-2015-309800

Zhang, J., Bajari, R., Andric, D., Gerthoffert, F., Lepsa, A., Nahal-Bose, H., et al. (2019). The International cancer genome consortium data portal. *Nat. Biotechnol.* 37, 367–369. doi: 10.1038/s41587-019-0055-9

Zhang, S.-W., Jin, X. Y., and Zhang, T. (2017). Gene prediction in metagenomic fragments with deep learning. *Biomed. Res. Int.* 2017, 1–9. doi: 10.1155/2017/4740354

Zhang, Z., and Zhang, L. (2021). METAMVGL: a multi-view graph-based metagenomic contig binning algorithm by integrating assembly and paired-end graphs. *BMC Bioinform.* 22:378. doi: 10.1186/s12859-021-04284-4

Zhang, S.-W., Wei, Z.-G., Zhou, C., Zhang, Y.-C., and Zhang, T.-H. (2013). "Exploring the interaction patterns in seasonal marine microbial communities with network analysis." in *2013 7th International Conference on Systems Biology (ISB), Huangshan, China.* pp. 63–68.

Zhao, Z., Woloszynek, S., Agbavor, F., Mell, J. C., Sokhansanj, B. A., and Rosen, G. L. (2021). Learning, visualizing and exploring 16S rRNA structure using an attention-based deep neural network. *PLoS Comput. Biol.* 17:e1009345. doi: 10.1371/journal.pcbi.1009345

Zhu, W., Lomsadze, A., and Borodovsky, M. (2010). Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res.* 38:e132. doi: 10.1093/nar/gkq275

Zou, H., and Hastie, T. (2005). Regularization and Variable Selection Via the Elastic Net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 67, 301–320. doi: 10.1111/j.1467-9868.2005.00503.x

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 04 December 2023
DOI 10.3389/fmicb.2023.1297451

OPEN ACCESS

EDITED BY
Aldert Zomer,
Utrecht University, Netherlands

REVIEWED BY
Antonella Lupetti,
University of Pisa, Italy
Sercan Karav,
Çanakkale Onsekiz Mart University, Türkiye
Miriam Cordovana,
Bruker Daltonik GmbH, Germany

*CORRESPONDENCE
Jianguo Chen
✉ cij123@126.com
Weijie Wang
✉ weijiewang@ncst.edu.cn
Lida Xu
✉ lida.xu@hotgen.com.cn

†These authors have contributed equally to this work

RECEIVED 20 September 2023
ACCEPTED 16 November 2023
PUBLISHED 04 December 2023

CITATION
Liu K, Wang Y, Zhao M, Xue G, Wang A, Wang W, Xu L and Chen J (2023) Rapid discrimination of *Bifidobacterium longum* subspecies based on MALDI-TOF MS and machine learning.
Front. Microbiol. 14:1297451.
doi: 10.3389/fmicb.2023.1297451

Rapid discrimination of *Bifidobacterium longum* subspecies based on MALDI-TOF MS and machine learning

Kexin Liu[1,2†], Yajie Wang[3†], Minlei Zhao[4†], Gaogao Xue[2], Ailan Wang[2], Weijie Wang[1*], Lida Xu[2*] and Jianguo Chen[4*]

[1]College of Life Science, North China University of Science and Technology, Tangshan, China, [2]Beijing Hotgen Biotechnology Inc., Beijing, China, [3]Department of Clinical Laboratory, Beijing Ditan Hospital, Capital Medical, Beijing, China, [4]Beijing YuGen Pharmaceutical Co., Ltd., Beijing, China

Although MALDI-TOF mass spectrometry (MS) is widely known as a rapid and cost-effective reference method for identifying microorganisms, its commercial databases face limitations in accurately distinguishing specific subspecies of *Bifidobacterium*. This study aimed to explore the potential of MALDI-TOF MS protein profiles, coupled with prediction methods, to differentiate between *Bifidobacterium longum subsp. infantis* (*B. infantis*) and *Bifidobacterium longum subsp. longum* (*B. longum*). The investigation involved the analysis of mass spectra of 59 *B. longum* strains and 41 *B. infantis* strains, leading to the identification of five distinct biomarker peaks, specifically at m/z 2,929, 4,408, 5,381, 5,394, and 8,817, using Recurrent Feature Elimination (RFE). To facilate classification between *B. longum* and *B. infantis* based on the mass spectra, machine learning models were developed, employing algorithms such as logistic regression (LR), random forest (RF), and support vector machine (SVM). The evaluation of the mass spectrometry data showed that the RF model exhibited the highest performace, boasting an impressive AUC of 0.984. This model outperformed other algorithms in terms of accuracy and sensitivity. Furthermore, when employing a voting mechanism on multi-mass spectrometry data for strain identificaton, the RF model achieved the highest accuracy of 96.67%. The outcomes of this research hold the significant potential for commercial applications, enabling the rapid and precise discrimination of *B. longum* and *B. infantis* using MALDI-TOF MS in conjunction with machine learning. Additionally, the approach proposed in this study carries substantial implications across various industries, such as probiotics and pharmaceuticals, where the precise differentiation of specific subspecies is essential for product development and quality control.

KEYWORDS

Bifidobacterium longum* subspecies, MALDI-TOF MS, machine learning, identification, *B. longum*, *B. infantis

1 Introduction

Bifidobacterium longum subsp. infantis (*B. infantis*) and *Bifidobacterium longum subsp. longum* (*B. longum*), the most abundant *Bifidobacterium* species in the intestinal flora of infants, are essential for their immune development. Human breast milk contains a large amount of human milk oligosaccharides (HMOs), which cannot be digested by infant due to a lack of

necessary glucosidases. However, the positive effects of HMOs on newborns' health are attributed to the "beneficial" microorganisms that specialize in metabolizing HMOs. In contrast to *B. longum*, *B. infantis* typically harbors all the genes required for utilizing HMOs (Duar et al., 2020) and can digest various types of HMOs, including 3′-SL, 6′-SL, 2′-FL, 3′-FL, LNnT, and LacNAc (Zhang et al., 2022). The absence of *Bifidobacterium* and HMO utilization genes in the gut microbiota is associated with inflammation and immune imbalances in early life (Henrick et al., 2021). *B. infantis* is commonly found in breastfed infants in countries with a low prevalence of immune-mediated diseases, such as Bangladesh (Vatanen et al., 2022) and Malawi, but is rare in Europe and North America (Casaburi et al., 2021). However, supplementation with *B. infantis* EVC001, by remodelling the gut microbiome of breastfed infants, reduced intestinal inflammation (Henrick et al., 2019), decreased intestinal Th2 and Th17 cytokines and up-regulated IFNβ, favouring immune development in early life (Henrick et al., 2021). Therefore, accurate identification of *B. longum* and *B. infantis* is essential for efficient screening, functional studies and application development of *B. infantis*.

The current methods used to identify *Bifidobacteria* include PCR, SNP, cgMLST, and MALDI-TOF MS. MALDI-TOF MS is particularly advantageous due to its high throughput, fast speed, and low cost, making it widely used for identifying clinical pathogenic microorganisms and general microorganisms (Gato et al., 2021; Heilbronner and Foster, 2021; Wang H. Y. et al., 2022). However, the successful identification of bacteria using MALDI-TOF MS heavily relies on databases that contain spectra of known organisms and most of the biomarker peaks are in the range m/z 2,000–10,000 (Carvalho et al., 2022; Topić Popović et al., 2023). Most commercial databases only identify bacteria at the species level and lack the ability to accurately differentiate closely related subspecies, such as *B. longum* and *B. infantis*. Although six biomarker peaks have been reported to differentiate between *B. longum* and *B. infantis*, these peaks have not been commercially applied due to their high mass peaks (>15,000 m/z) (Sato et al., 2011), low reproducibility, and lack of availability in commercial databases. Recently, machine learning techniques have been used to accurately identify strains that cannot be distinguished using commercial databases by analyzing protein mass spectra obtained through MALDI-TOF MS (Weis et al., 2022; Kim et al., 2023).

Machine learning (ML) technology encompasses various algorithms such as random forest (RF), support vector machines (SVM), logistic regression (LR) and decision trees (DT) (Weis et al., 2020). ML enables rapid and precise identification of species-specific biomarkers from MALDI-TOF MS spectra, which has been widely implemented to analyze microbial signatures and construct classification models. Recently, the combination of MALDI-TOF MS and ML has gained popularity in classifying clinically pathogenic and drug-resistant bacteria, including *Escherichia coli* (van Oosten and Klein, 2020), *Staphylococcus aureus* (Rodríguez-Temporal et al., 2022), *Klebsiella pneumoniae* (Yu et al., 2023), *Brucella melitensis* (Dematheis et al., 2022), and *Campylobacter* spp. (Feucherolles et al., 2021). However, there is a lack of identification schemes for *Bifidobacterium* subspecies within a specific taxon in these studies. Hence, there is an urgent need to develop a combined machine learning and MALDI-TOF MS method for rapid and accurate identification of *Bifidobacterium* subspecies.

In the present study, we first screened for robust variations in subspecies-specific features between *B. longum* and *B. infantis* based on MALDI-TOF MS analysis and a combination of machine learning methods such as LR, SVM, and RF (Figure 1). The objective of this research was to develop a fast classification tool using Machine-learning-combined MALDI-TOF MS to accurately distinguish between *B. longum* and *B. infantis*.

2 Results

2.1 Molecular identification by PCR and phylogenetic analysis

Specific primers-based PCR could differentiate between *B. longum* and *B infantis*. Thus, this method was employed to confirm the taxonomic classification of all the strains in study. The specificity and sensitivity of the PCR assay using specific primers for distinguishing the two subspecies were confirmed by successfully differentiation of 11 representative strains. Out of the 89 isolates analyzed, 54 were identified as *B. longum* and 35 were identified as *B. infantis*. For additional confirmation, SNP information obtained from 100 genome sequences were utilized to construct a phylogenetic tree. The tree effectively separated the sequences into two distinct branches. The phylogenetic tree revealed that 59 *B. longum* strains, comprising five typical strains and 54 isolates, clustered together with a blue background, while 41 *B. infantis* strains formed a distinct group with a red background (Figure 2). These findings underscore the efficacy of using phylogenetic tree features for precise classification and identification of *B. longum* and *B. infantis*, which align with the outcomes obtained from specific PCR genotyping (Supplementary Table S1).

2.2 Identification of mass spectra for strains

Mass spectrometry results indicated the presence of numerous identical mass spectral peaks for both *B. longum* and *B. infantis*, making accurate differentiation challenging when relying solely on commercial databases (Figure 3A; Supplementary Table S1). However, further analysis unveiled six species-specific peaks that exhibited a high degree of conservation and could serve as potential biomarkers for identification. As shown in Figures 3B–D, peaks at m/z 4448.52 (94.9%, 56/59), 5394.35 (100.0%), and 8789.47 (100.0%) were exclusively found in the spectrogram of *B. longum*. Conversely, peaks at m/z 4408.42 (95.1%, 39/41), 5381.06, and 8817.28 (100.0%) were observed solely in the spectrogram of *B. infantis*. These findings reveal the potential of MALDI-TOF MS to differentiate between *B. longum* and *B. infantis* based on specific peaks with the protein fingerprint profile.

2.3 Discovery and identification of protein biomarkers by MALDI-TOF MS

To investigate the applicability of MALDI-TOF MS for discriminating *B. longum* and *B. infantis*, we performed redundant

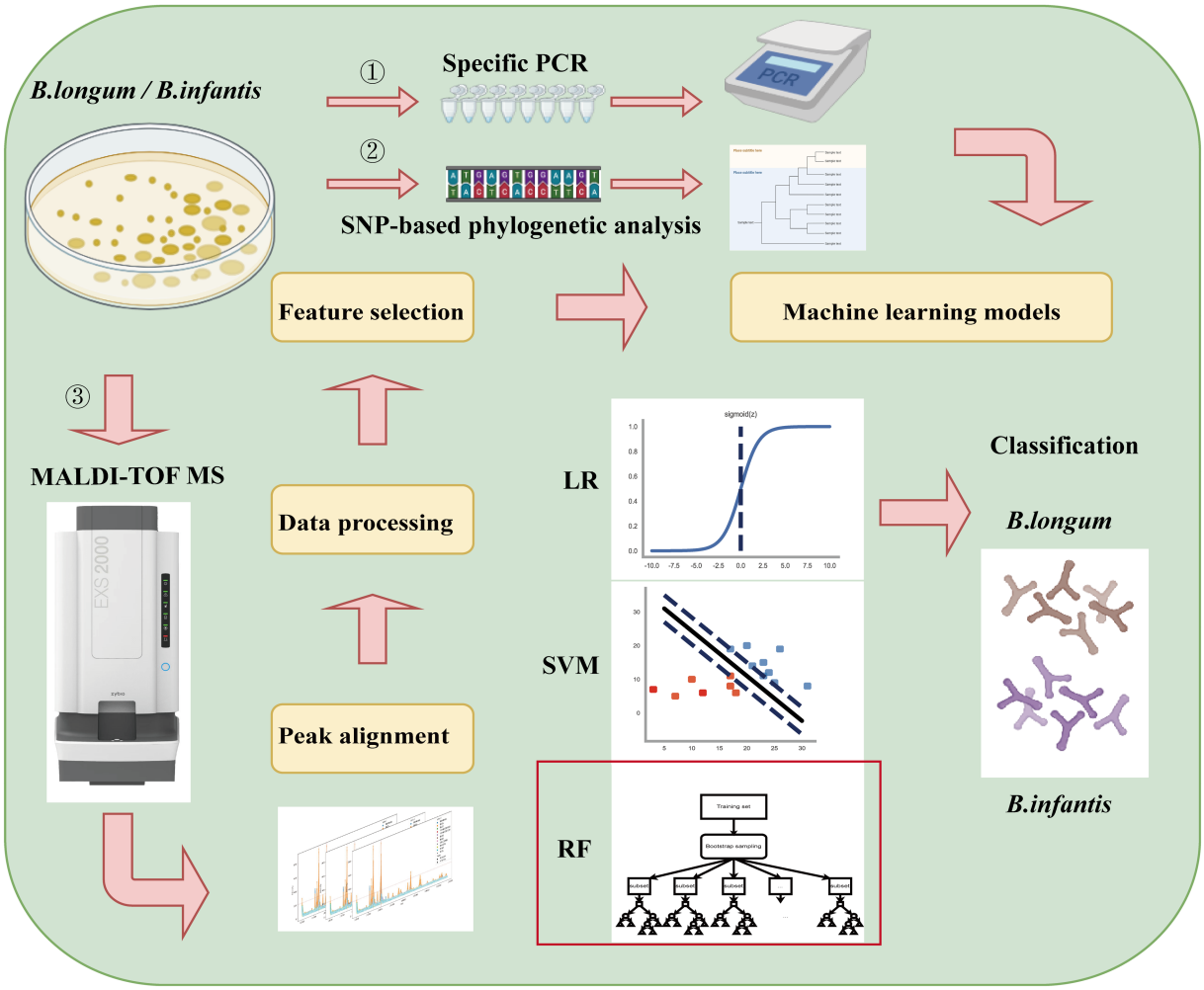

FIGURE 1
Experimental flow chart of *Bifidobacterium longum* subspecies discrimination based on MALDI-TOF MS and machine learning in this study.

removal, smoothing, and alignment of 400 spectra from 100 strains using OpenMS software. We identified some potential characteristic peaks and constructed a mass spectrometry data matrix for further analysis. To further investigate the distinguishing features, we performed a more specific heatmap clustering analysis of the mass spectrometry data matrix (Figure 4A). Then we performed the principal component analysis (PCA) of the mass spectrum data matrix obtained from the above method. The PCA plot clearly showed the distinct clustering patterns of the two subspecies (Figure 4B), indicating their potential for differentiation. Finally, 18 potential discriminatory peaks were identified, with 11 peaks specific to the *B. infantis*, including the 3,088 m/z, 3,573 m/z, 4,408 m/z, 5,338 m/z, 5,381 m/z, 6,820 m/z, 6,910 m/z, 8,131 m/z, 8,817 m/z, 9,963 m/z, 10,360 m/z. *B. longum* with seven specific peaks, respectively, are located at the 2,929 m/z, 3,152 m/z, 4,448 m/z, 4,479 m/z, 5,394 m/z, 7,051 m/z, 8,789 m/z. These discriminatory peaks are expected to serve as potential features for constructing the classifiers. Furthermore, to assess the importance of features, we analyzed between 18 feature peaks and drew bar graphs (Figure 4C) and found higher SHAP values for feature peaks at 4408 m/z, 5,381 m/z, 5,394 m/z and 8,817 m/z. This suggests that these peaks seem particularly well suited for building classifiers.

To gain insights into the identities of these characteristic peaks, we conducted a comparison between the experimental m/z values and genomic data. This analysis suggested that the ion peaks at m/z 5,381 and 5,394 corresponded to the 50S ribosomal protein L34. Additionally, peaks at m/z 8,817 and 7,051 were associated with 50S ribosomal proteins L27 and L30, respectively. The peak at 4408 m/z indicated the presence of the 30S ribosomal protein S5. Moreover, we identified matches with proteins from the DUF (domain of unknown function) family, including m/z 4,479, 8,789, and 9,963. Proteins belonging to the DUF family are characterized by a conserved EYA motif and a length ranging from 66 to 95 amino acids. However, their functional roles remain elusive due to the lack of annotation.

The 18 feature peaks obtained above were conducted recursive feature elimination using a logistic regression algorithm with cross-validation to determine the optimal feature set. Figure 5A illustrated that the highest cross-validation score of 0.945 was achieved when using five features. These five optimal features include m/z 2,929, 4,408, 5,381, 5,394, and 8,817. Among them, m/z 2,929 and 5,394 were characteristic peaks of *B. longum*, while the remaining peaks were specific to *B. infantis*. The significance of the five selected features was presented using a boxplot (Figure 5B), and the results indicated that

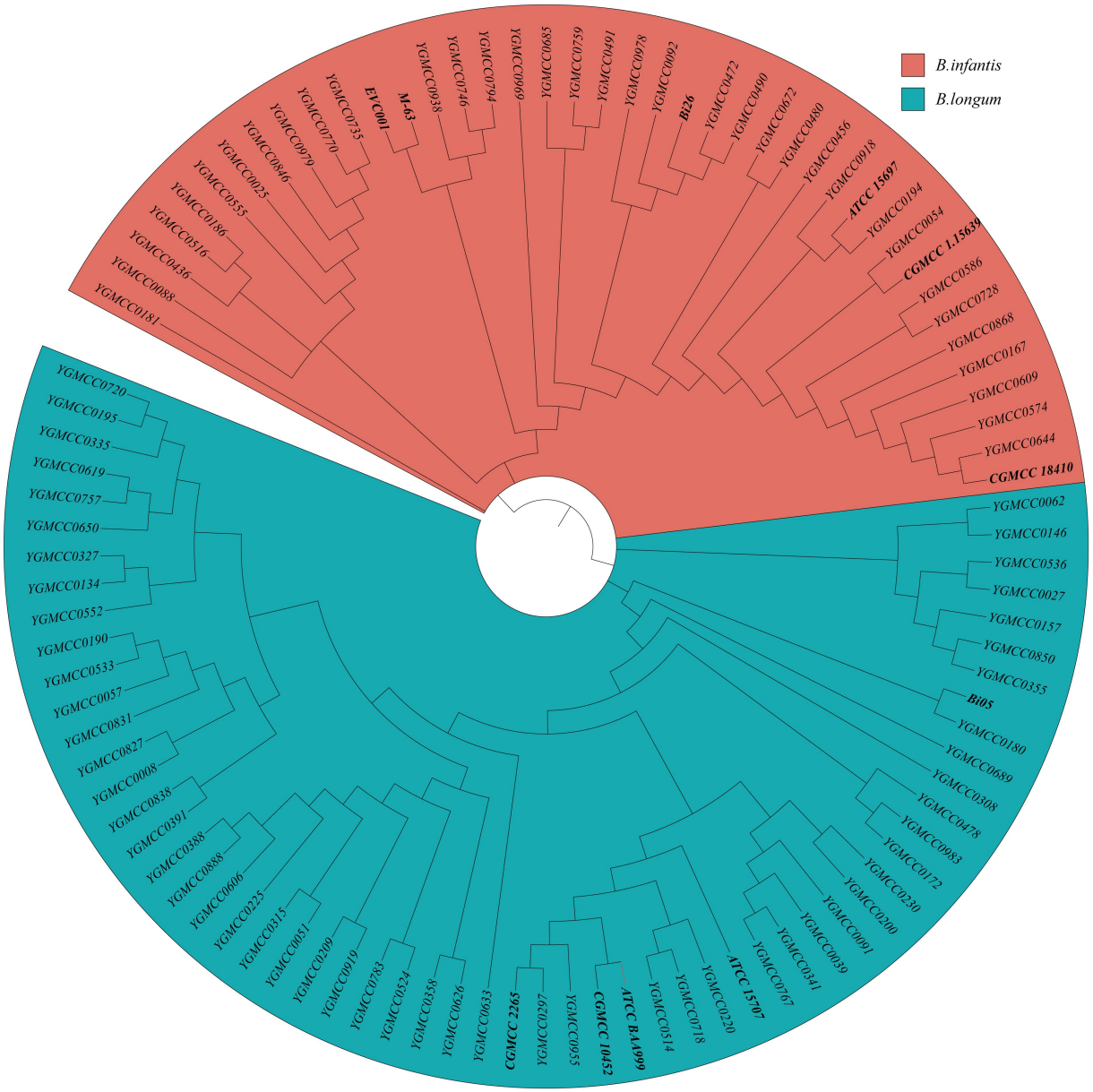

FIGURE 2
Identification of 41 *B. infantis* strains and 59 *B. longum* strains based on the phylogenetic analysis. The red and blue background represent *B. infantis* or *B. longum* strains, respectively.

the *p*-values of the five features, as determined by Fisher's exact test, were all less than 0.001. In addition, individual ROC curves were plotted for the five selected features (Figure 5C). The AUC values ranged from 0.777 for m/z 2,929 to 0.917 for m/z 5,381. It indicates that the features obtained after recursive elimination can contribute to achieving the best classification performance.

2.4 Construction of the machine learning models

We developed three commonly used machine learning models: LR, SVM, and RF, for microbial discrimination. The dataset utilized for model construction consisted of 100 strains, with their subspecies verified through PCR and phylogenetic analysis. This dataset was randomly divided into a training set for building the models and a test set for evaluation their performance. Based on the results obtained from the test set, we calculated performance metrics such as sensitivity, specificity, accuracy, Youden coefficient, and AUC value (see Table 1).

The classification performance parameters of the three models are shown in Table 2. Among them, RF achieved the highest accuracy, AUC, and Youden coefficient, all equal to 1.0, indicating its superior ability to discriminate between the two subspecies. The sensitivity of all three models was 1.0, which means that they could correctly identify all the positive cases. The RF model demonstrated the highest specificity with a value of 1.0, whereas the LR and SVM models exhibited a specificity of 0.931. The RF model also has the highest AUC value of 1.0, demonstrating excellent classification performance.

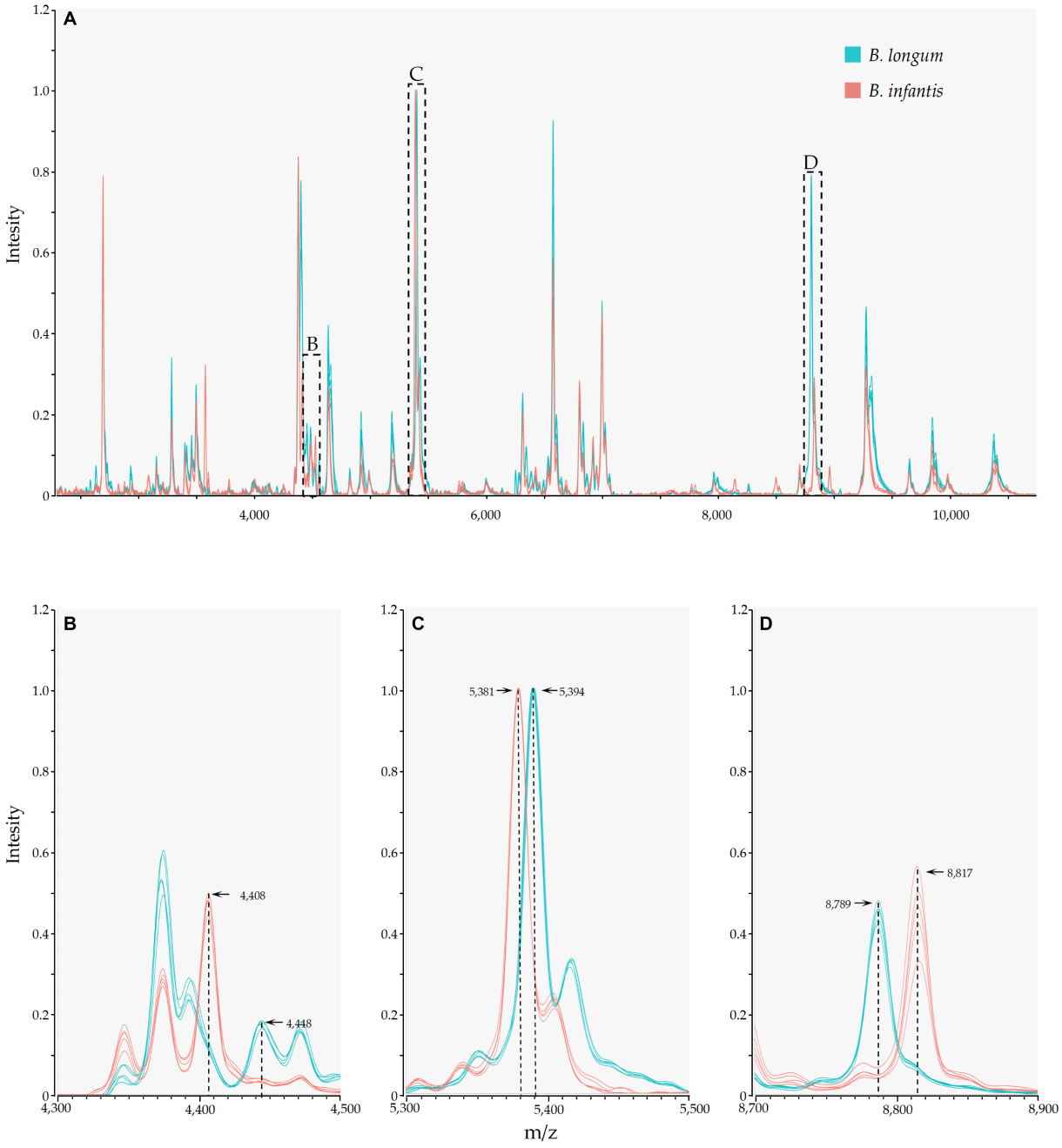

FIGURE 3
MALDI-TOF MS and species-specific peaks of *B. longum* (Orange) and *B. infantis* (Purple). The y-axis represents the intensity of the peaks, while the x-axis represents the m/z values; **(A)** depicts stowage diagram of *B. longum* and *B. infantis*; **(B–D)** display enlarged views of subspecies-specific peaks as depicted in **A**.

The SVM model's AUC was slightly better than that of the LR model, with values of 0.995, and 0.993, respectively. The Youden coefficient, reflecting the overall efficiency of the RF model, was 1.0, while for the SVM and LR models, it was 0.931.

2.5 Assessment of practical application of the machine learning model

An external dataset comprising 240 spectra obtained from 60 *Bifidobacterium longum* strains was collected. These isolates were obtained under identical experimental conditions. To validate the model's effectiveness, the three trained models were utilized to predict the subspecies of these 60 strains.

Among the three models, both LR and SVM model exhibited a specificity of 0.983, while it was 0.967 for the RF model. However, the LR model demonstrated a higher sensitivity (0.942) compared to the SVM model (0.883) and the RF model (0.900). Regarding accuracy, the RF model outperformed the SVM model and the LR model, achieving an accuracy rate of 0.954. To provide a more intuitive comparison of the models performance, we plotted the ROC curve (Figure 6A) and calculate the AUC values. All three models exhibited very similar AUC

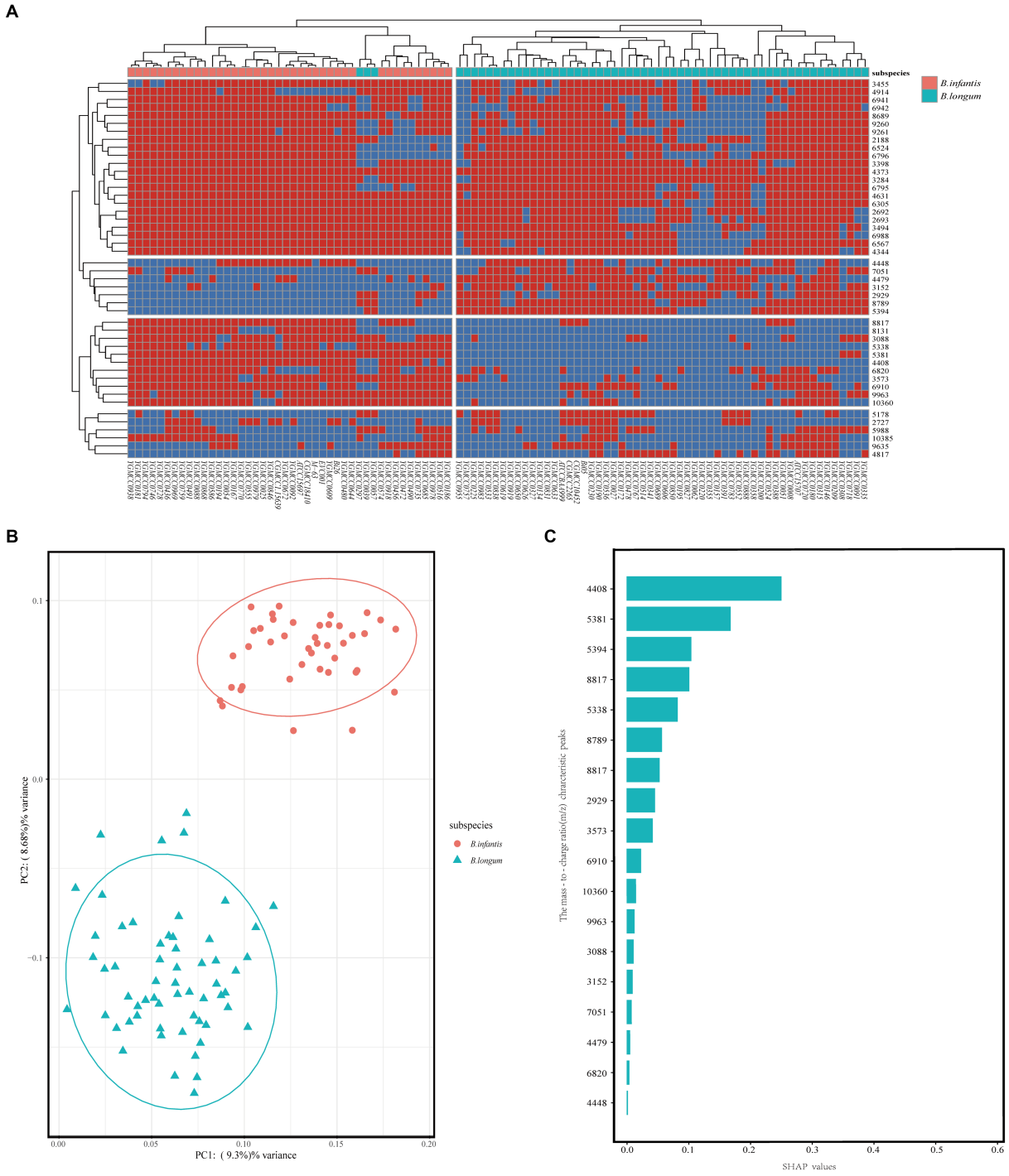

FIGURE 4
Unsupervised analysis and feature importance evaluation. **(A)** After mass spectral alignment, heat maps were plotted and clustered based on the absence/presence of common characteristic peaks in the top 50% of effective *p* values within subspecies. **(B)** PCA of *Bifidobacterium longum* subsp. Each dot on the PCA plot represents the average spectrum of each strain, blue for *B. longum,* and red for *B. infantis*. **(C)** Assessment of feature importance in a RF model for distinguishing between *B. longum* and *B. infantis*.

values, accurately measured at 0.984. The RF model had the highest Youden index (0.908), surpassing the SVM model (0.867) and the LR model (0.883). Figure 6B illustrated the distribution of prediction scores indicating the likelihood of being *B. infantis* strains for the two subspecies, as determined by the three models.

Based on the four data points results, we established the prediction conditions for the strain subspecies model. A confusion matrix for external strain identification was calculated based on the voting results (Figure 6C). Specific PCR test results and phylogenetic analysis results (Figure 6D) showed consistency. The results from specific PCR tests and phylogenetic analysis (Figure 6D) were consistent with these findings (Supplementary Table S1). Among them, in the LR model, the identification of *B. longum* was in line with PCR and phylogenetic results. However, for the *B. infantis*, specifically YGMCC0271,

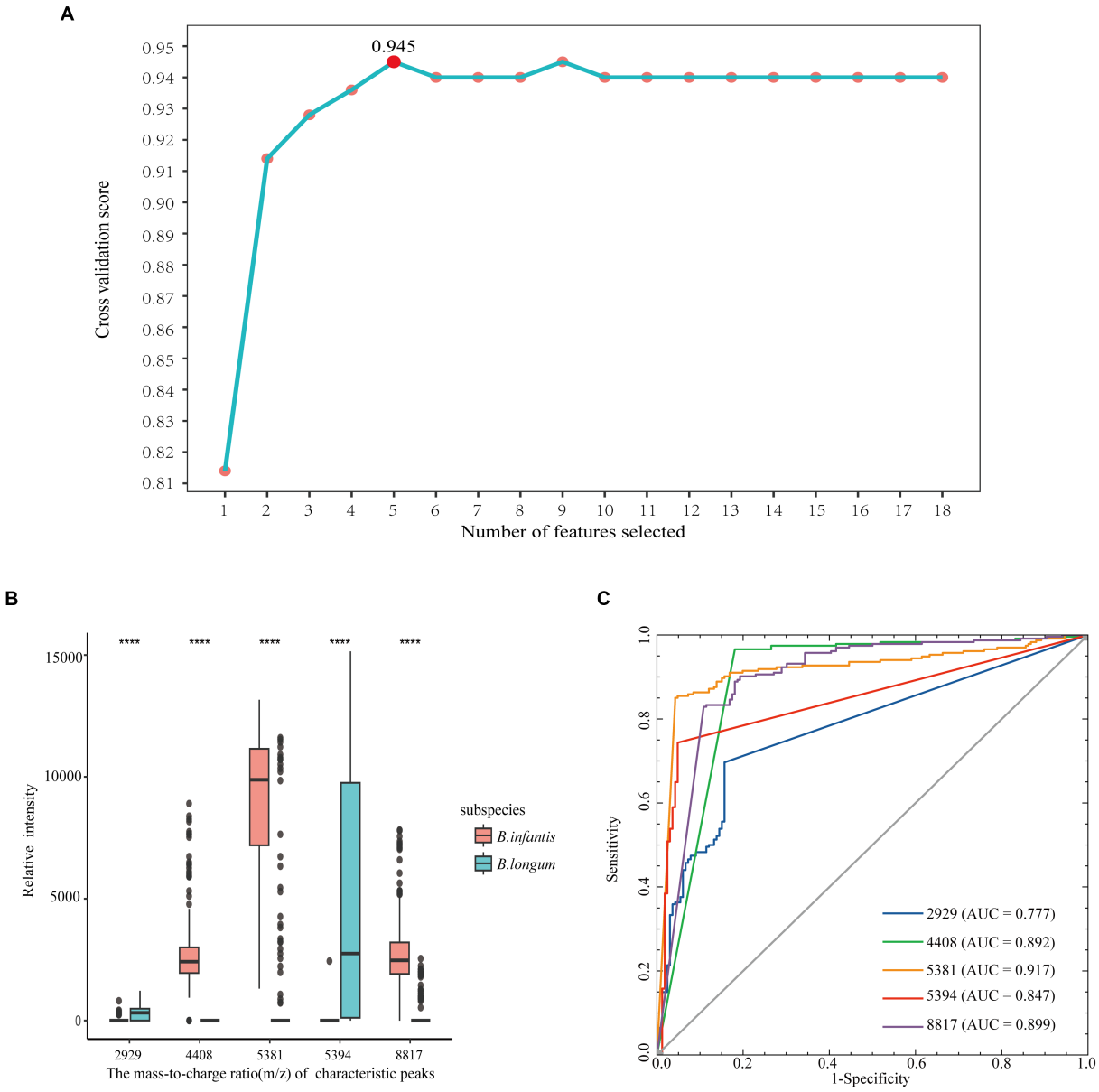

FIGURE 5
Recursive feature elimination. Line plot of 18 characteristic peaks and cross-validation fractions after REFCV **(A)**, and boxplot of mass-to-charge ratio and relative intensity of 5 optimal characteristic peaks between the two subspecies of *Bifidobacterium longum* (**** represents the *p* value of the difference < 0.0001). **(B)**. ROC curve and AUC value of the five optimal characteristic peaks **(C)**.

YGMCC0192, and YGMCC0550, there was inconsistency, with an accuracy of 95%. In the SVM model, the identification of YGMCC0618, YGMCC0063, and YGMCC0038 did not align with PCR and phylogenetic results, resulting in an accuracy rate of 95%. Lastly, in the RF model, the identification of YGMCC0063 and YGMCC0120 differed from PCR and SNP results, achieving an accuracy rate of 96.67%. Based on the external strain identification results, the RF model emerged as the optimal choice.

3 Discussion

Genome-based taxonomy is a more standard method of classifying microorganisms than traditional methods (Parks et al., 2018). However, it is time-consuming, expensive, and labor-intensive, and fails to meet the demand for rapid and high-throughput identification of microorganisms. In recent years, MALDI-TOF MS has gained increasing importance in clinical microbial taxonomy as a fast, high-throughput, and robust method for microbial identification. It relies on the detection of microbial housekeeping and ribosomal proteins (Kim et al., 2022a; Haider et al., 2023). Nonetheless, while MALDI-TOF MS can identify bacteria at the species level, it struggles to accurately distinguish closely related species or subspecies. Machine learning algorithms have the capability to identify specific information in mass spectrometry data and analyze relationships among different features, enabling more precise analysis (Weis et al., 2020). By combining machine learning with MALDI-TOF MS, it becomes

TABLE 1 Frequencies and assignments of species-specific peaks for *B. longum* and *B. infantis*.

Experimental m/z	Presence of peak (%)		Theoretical m/z	Possible presence of protein
	B. longum	*B. infantis*		
2,929	77.97 (46/59)	7.32 (3/41)	2,932	Hypothetical protein
3,088	23.72 (14/59)	82.93 (34/41)	3,088	NAD(P)-binding domain-containing protein
3,152	69.49 (41/59)	12.20 (5/41)	3,150	Integrase partial
3,573	30.50 (18/59)	95.12 (39/41)	3,573	Restriction endonuclease
4,408	0.00 (0/59)	95.12 (39/41)	4,406	30S ribosomal protein S5 partial
4,448	55.93 (33/59)	29.27 (12/41)	4,447	50S ribosomal protein L9 partial
4,479	74.58 (44/59)	19.51 (8/41)	4,480	DUF600 family protein partial
5,338	6.78 (4/59)	80.49 (33/41)	5,338	Permease
5,381	10.17 (6/59)	100.0 (41/41)	5,377	50S ribosomal
5,394	81.36 (48/59)	0.00 (0/41)	5,391	Protein L34
6,820	28.81 (17/59)	78.05 (32/41)	6,822	Transporter drug/metabolite exporter family
6,910	38.98 (23/59)	97.56 (40/41)	6,910	Transposase
7,051	67.80 (40/59)	14.63 (6/41)	7,051	50S ribosomal protein L30
8,131	0.00 (0/59)	63.41 (26/41)	8,135	IS3 family transposase partial
8,817	13.56 (8/59)	87.80 (36/41)	8,816	50S ribosomal protein L27
8,789	79.66 (47/59)	2.44 (1/41)	8,789	DUF905 domain-containing protein
9,963	28.81 (17/59)	92.68 (38/41)	9,963	DUF4244 domain-containing protein
10,360	30.50 (18/59)	92.68 (38/41)	10,364	50S ribosomal protein L13 partial

TABLE 2 Model result metrics for three machine learning models in validation dataset.

Machine learning models	Specificity	Sensibility	Youden	AUC	Accuracy
LR	0.931	1.000	0.931	0.993	0.958
SVM	0.931	1.000	0.931	0.995	0.958
RF	1.000	1.000	1.000	1.000	1.000

possible to accurately identify closely related microorganisms at the subspecies level (De Bruyne et al., 2011; Rodríguez-Temporal et al., 2023). Recent studies have demonstrated the application of machine learning techniques in overcoming the limitations of mass spectrometry, such as detecting antibiotic-resistant microorganisms (Yoon and Jeong, 2021), analyzing antimicrobial resistance (Feucherolles et al., 2021), and distinguishing closely related species. By utilizing features obtained from MALDI-TOF MS, SVM algorithms have successfully differentiated clinically resistant strains of carbapenem, methicillin, and β-lactam antibiotics, as well as predicted resistance phenotypes with high accuracy (Ho et al., 2017; Wang J. et al., 2022). Furthermore, the combination of MALDI-TOF MS and machine learning is commonly used to distinguish closely related foodborne microorganisms. For example, an SVM-RBF model achieved a prediction accuracy of approximately 100% in accurately identifying *W. cibaria* and *W. confusa* (Kim et al., 2023).

In our research, we have found that distinguishing closely related species using MALDI-TOF MS can be challenging due to the similarities in their protein fingerprints. MALDI-TOF MS generates a report of the ten closest matches for an unknown species based on mass spectra and the consistency of reference strains in the database. However, when different species within the same genus or different subspecies within the same species have high scores among the top ten matches, accurately identifying the microorganism becomes difficult. Previous studies have attempted to distinguish between *Bifidobacterium longum* subspecies (Kim et al., 2022b) and *Bifidobacterium animalis* subspecies (Jahan et al., 2021) using MALDI-TOF MS. However, these studies had limitations in terms of sample size, unsystematic markers, and lack of validation data, and have not been commercially applied. In this study, our focus was specifically on identifying *B. longum* and *B. infantis* using MALDI-TOF MS. We discovered that commercial databases were unable to accurately differentiate between these two subspecies, which aligns with previous findings (Yahiaoui et al., 2020; Jahan et al., 2021; Kim et al., 2022b).

The aim of this study was to evaluate the ability of MALDI-TOF MS combined with machine-learning methods to rapidly and accurately discriminate between the closely related *B. longum* and *B. infantis*. We employed advanced machine learning algorithms and a larger sample size to enhance statistical significance. We ensured systematic biomarker collection and data analysis to improve the reliability and repeatability of our findings. We examined 400 mass spectra from 100 *Bifidobacterium longum* strains and used a logistic regression model with recursive feature elimination to identify the five most significant mass peaks. Among these peaks, the masses at 2929

10.3389/fmicb.2023.1297451

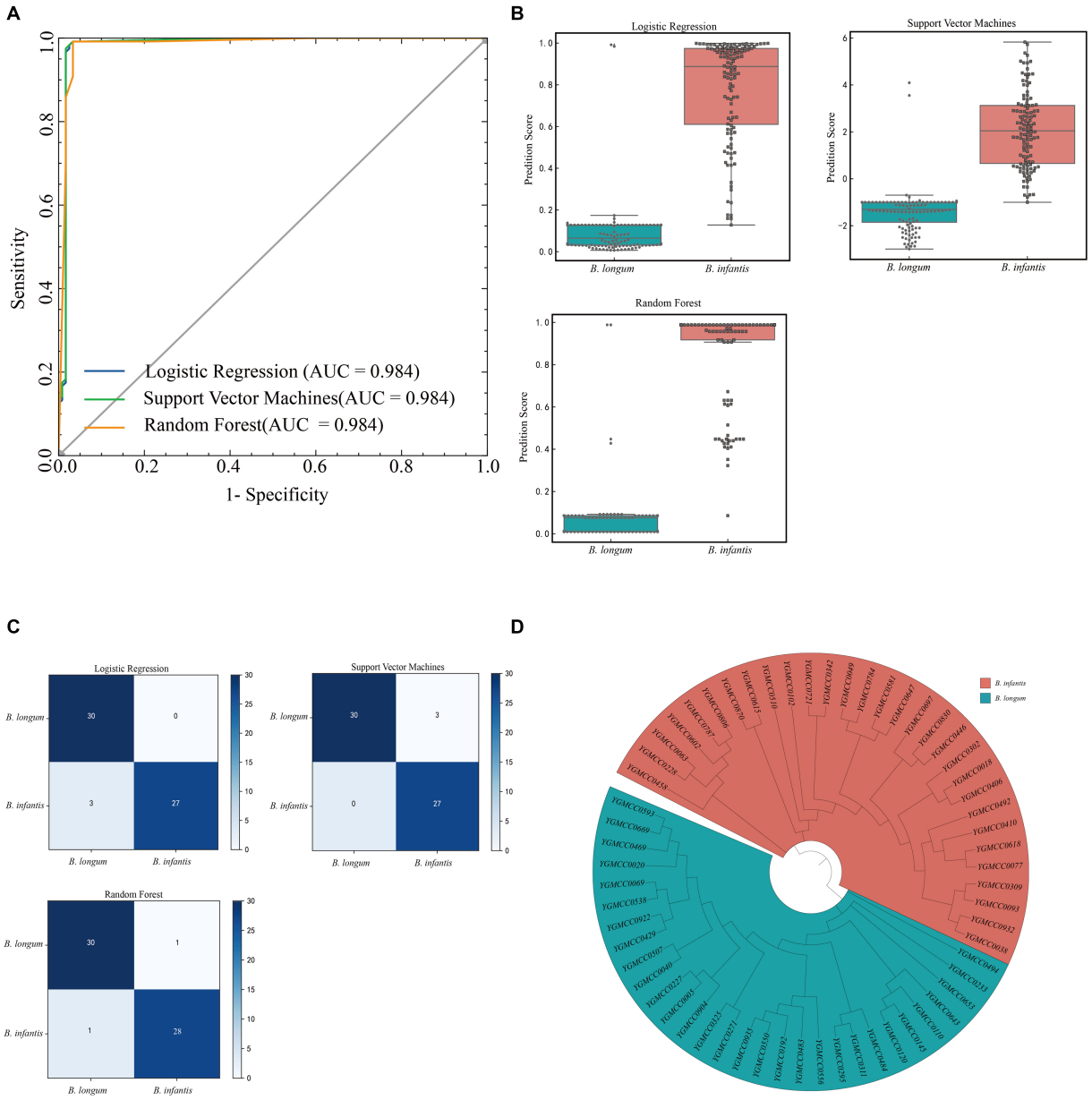

FIGURE 6
Model evaluation in test dataset. **(A)** ROC curves and AUC of three machine learning models in the test dataset. **(B)** Boxplot of the verification score of the three machine learning models on the spectral data of the external test set. **(C)** Confusion matrix of external strain identification results for three models. **(D)** Cluster analysis of isolated *B. infantis* or *B. longum* strains used for external validation set based on phylogenetic analysis. The red and blue background represnt *B. infantis* or *B. longum* strains, respectively.

and 5,394 m/z were specific to *B. longum*, while the masses at 4408, 5381, and 8,817 m/z were specific to *B. infantis*. These mass peaks can potentially serve as biomarkers for distinguishing between these two species. Using these biomarkers, we developed machine learning models employing LR, RF, and SVM algorithms. All three models exhibited excellent performance in identifying the spectrogram, with the RF model demonstrating high accuracy in discriminating between *B. longum* and *B. infantis*. Furthermore, after evaluating mass spectrum identification results through voting, the RF model achieved the highest accuracy in practical strain identification applications (see Table 3).

4 Materials and methods

4.1 Bacterial strains

Twelve reference strains and eighty-nine strains of *Bifidobacterium longum* subspecies, isolated at Beijing Yujing Pharmaceutical Co., Ltd., were selected to explore potential biomarkers (Table 4). The bifidobacterial strains were incubated for 48 h at 37°C under anaerobic conditions. *E. coli* ATCC 25922 incubated for 24 h at 37°C in Luria-Bertani (Solarbio, Beijing, China) agar was applied to external calibration of MALDI-TOF MS.

TABLE 3 Model result parameters for three machine learning models on an test dataset.

Machine learning models	Specificity	Sensibility	Youden	AUC	Accuracy
LR	0.983	0.900	0.883	0.984	0.942
SVM	0.983	0.883	0.867	0.984	0.933
RF	0.967	0.942	0.908	0.984	0.954

TABLE 4 Strain information used in this study.

Bacterial strains		Origins
Reference strains		
Bifidobacterium longum subsp. longum (*B. longum*)	ATCC 15707	[1]ATCC
	ATCC BAA999	
	CGMCC 10452	[2]CGMCC
	CGMCC 2265	
	Bi05	[3]IFF
Bifidobacterium longum subsp. infantis (*B. infantis*)	ATCC 15697	ATCC
	CGMCC 1.15639	CGMCC
	CGMCC 18410	
	Bi26	IFF
	EVC001	[4]Evolve
	M-63	[5]MORINAGA
Escherichia coli	ATCC 25922	ATCC
Isolates ([7]N)		
Bifidobacterium longum subspecies (149)		[6]YGMCC

[1]ATCC, American type culture collection;[2]CGMCC, China General Microbiological Culture Collection Center;[3]IFF, International Flavors & Fragrances Inc.;[4]Evolve, Infinant Health™;[5]MORINAGA, Morinaga Milk Industry Co., Ltd.;[6]YGMCC, Beijing Yujing Pharmaceutical Co., Ltd.;[7]N, Number of isolates.

4.2 MALDI-TOF MS analysis

Proteins from *B. longum* and *B. infantis* were extracted using the ethanol-formic acid-extraction method (Cuénod et al., 2023). Concisely, fresh bacterial culture was suspended in 300 μL of ddH_2O to which 900 μL of ethanol was added. The bacterial suspension was centrifuged at high speed (10,000× g) for 2 min, the supernatant was removed to completely discard the residual ethanol and recentrifuged. The resulting pellet was resuspended in 20 μL of 70% formic acid to which an equal volume of acetonitrile was added. After centrifugation at 10,000× g for 2 min, 1 μL of each supernatant was transferred to the 96-position MALDI-TOF target plate, allowed to air dry, and then overlaid with 1 μL of the matrix solution (10 mg/mL of α-cyano-4-hydroxy-cinnamic acid (HCCA) in 50% (v/v) acetonitrile with 2.5% (v/v) trifluoroacetic acid).

The mass spectra were acquired using an EXS2000 MALDI-TOF MS (Zybio Inc., Chongqing, China) equipped with a 200 Hz smart-beam solid-state laser and operated in positive linear mode (Xiong et al., 2023). Mass spectra were automatically recorded within a mass range of 2–20 kDa with a total of 200 laser shots. *E. coli* ATCC 25922 was used for mass calibration and instrument parameter optimization, with an average deviation of molecular weight less than 300 ppm after correction. MS data were analyzed using MDT Master (version 1.1). log scores ≥2.0 were accepted for the identification at the species level, and log scores <2.0 and ≥ 1.7 were used for identification at the genus level or the presumptive species level. Log scores below 1.7 were considered unreliable. For establishing stable machine learning models, four high-quality mass spectra (log scores ≥2.3, stable benchmarks, abundant protein peaks, and uniform distribution) were selected in each strain.

TABLE 5 Specific primer information used in this study.

Target	Primer	Sequence (5′–3′)	Size (bp)
B. longum	B.lon_831_F	TTCCAGTTGATCGCATGGTC	831
	B.lon_831_R	GGGAAGCCGTATCTCTACGA	
B. infantis	B.inf_832_F	TTCCAGTTGATCGCATGGTC	832
	B.inf_832_R	GGAAACCCCATCTCTGGGAT	

4.3 Species identification based on PCR and genomics sequences

For the identification of the isolates, genomic DNA was extracted using Easy Pure Bacteria Genomic DNA Kit (Trans, Beijing, China) in accordance with the manufacturer's instructions. Then, 1 μL of supernatant was used for the following PCR reaction, the reaction mixture contained 10 μL of SapphireAmp® Fast PCR Master Mix (TaKaRa, Beijing, China), 0.5 μL of each primer (10 μM), 1 μL of DNA template, and 8 μL of ddH2O. Specific primers were listed in Table 5.

PCR reactions were conducted as follows: one cycle of initial denaturation at 98°C for 3 min, followed by 35 cycles of 98°C for 10 s, 55°C for 10 s, and 72°C for 5 s, and a final extension at 72°C for 2 min. The PCR products were observed by an Agarose gel imaging system (Tanon, Shanghai, China).

Total 149 unknown *Bifidobacterium longum* strains were cultured anaerobically at 37°C for 24 h, then the cultured liquid (50 mL) was centrifuged at 12,000 × g and 4°C for 10 min to collect the cell biomass. Genomic DNA of 149 unknown *Bifidobacterium longum* strains were extracted using a Wizard® Genomic DNA Purification Kit (Promega, United States). Purified genomic DNA was quantified using a TBS-380 fluorometer (Turner BioSystems Inc., Sunnyvale, CA, United States). High-quality DNA (OD260/280 = 1.8–2.0, ≥10 μg) was used for further research. Genomic DNA was sequenced using Illumina sequencing (Illumina, Inc.). The data generated from Illumina platforms were used for bioinformatics analysis.

The phylogenetic analysis included the comparison of genomic sequences from 5 standard strains of *B. infantis*, 6 standard strains of *B. longum*, and an additional 149 unknown *B. longum* strains from our laboratory. These sequences were compared with the genomic sequence of ASM19655v1, which served as the reference genome. The analysis was performed using the Parsnp software, focusing on the core genome (Treangen et al., 2014; Wang et al., 2023). The iTOL (Interactive Tree of Life) tool was utilized to visualize and explore the phylogenetic tree (Letunic and Bork, 2019; Pereira et al., 2023), facilitating the identification and classification of *B. longum* subspecies based on their phylogenetic positions.

4.4 Genomic data mining and identification of biomarker proteins

To investigate the significance of using unique peaks from mass spectrum data as biomarkers, we conducted genomic data mining using publicly available databases. The genome sequences of *B. longum* and *B. infantis* were obtained from the National Center for Biotechnology Information (NCBI) database. To annotate the selected protein biomarkers, the web-based ProtParam tool[1] was utilized to calculate their theoretical molecular weights based on the translated amino acid sequences. Subsequently, a custom script was employed to filter and align the selected proteins, identifying the most relevant proteins enriched in the vicinity of the characteristic peaks.

4.5 Model construction and verification

4.5.1 Data preprocess

The MS data obtained using openMS (v2.8) software exhibited high quality, allowing for alignment of peaks obtained from different batches. The processed peak map data matrix was subjected to PCA to access the potential of the features. In addition, a heatmap was drawn for cluster analysis using the R language (v4.2.2). After obtaining the cluster branches of the potential feature peaks, the importance parameters of the features and evaluate the importance of the features.

The dataset consisting of 400 spectra from 59 *B. longum* and 41 *B. infantis* was randomly divided into 70% training and 30% test datasets. The data of subspecies type was binarized, with 0 representing the long subspecies and 1 representing the infant subspecies. All peaks (features) were scaled using Min-Max scalar to ensure variables at different scales contributed equally to the model fitting process.

4.5.2 Classifier model construction

Firstly, feature selection was carried by a meta-converter approach based on a logistic regression classifier with scikit-learn (v1.3.0). Recursive feature elimination with 5x cross-validation (RFECV) was applied to discard irrelevant features and improve the model's generalization ability.

Secondly, SHAP (SHapley Additive exPlanations) was used to interpret predictions. SHAP is a unified framework that assigns importance values to each feature for a specific prediction and identifies which feature is most important, facilitating the understanding of a machine learning model's decision-making process (Lundberg and Lee, 2017).

Thirdly, three machine learning algorithms including random forest (RF), logistic regression (LR), and support vector machine (SVM) were used to construct the distinguishing models using the scikit-learn package. The performances of the models were evaluated by generating the confusion matrix on the test dataset. The ROC curve was plotted using the Matplotlib package, and the area under the subject operating characteristic curve (AUROC) was calculated as a measure of classifier performance. The Youden index was utilized to determine the optimal cutoff threshold and calculate the sensitivity, specificity, and accuracy metrics for the model.

To assess the practical applicability of the model in strain identification, we performed an external validation using a new dataset. Each strain in this dataset was accompanied by four mass spectra collected under identical experimental conditions. Subsequently, we compared the identification outcomes with those obtained through specific PCR detection and phylogenetic analysis.

5 Conclusion

In our research, we successfully demonstrated the effectiveness of combining MALDI-TOF-MS with machine learning to accurately discriminate between *B. longum* and *B. infantis*. We identified everything from protein fingerprints to potential biomarkers, and developed three spectral map identification models using the ML algorithm, and finally evaluated the various performance metrics and voted to find the optimal algorithm. The algorithm is highly reliable and accurate in distinguishing the two subspecies. This approach has the potential to be applied in various industries, such as the food or pharmaceutical industry, for rapid and cost-effective identification of *B. longum* and *B. infantis*. Furthermore, the identification strategy presented in this study can also be extended to other closely related species.

1 https://web.expasy.org/protparam/

Data availability statement

The names of the repository/repositories and accession number(s) can be found below: NCBI; PRJNA1020989.

Author contributions

KL: Investigation, Methodology, Validation, Writing – original draft. YW: Investigation, Methodology, Validation, Writing – original draft. MZ: Data curation, Investigation, Methodology, Validation, Writing – original draft. GX: Software, Validation, Writing – original draft. AW: Data curation, Writing – review & editing. WW: Supervision, Validation, Writing – review & editing. LX: Data curation, Methodology, Supervision, Writing – review & editing. JC: Project administration, Supervision, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

KL, GX, AW, and LX were employed by Hotgen Biotechnology Inc. MZ and JC were employed by Beijing YuGen Pharmaceutical Co., Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2023.1297451/full#supplementary-material

References

Carvalho, M., Sands, K., Thomson, K., Portal, E., Mathias, J., Milton, R., et al. (2022). Antibiotic resistance genes in the gut microbiota of mothers and linked neonates with or without sepsis from low- and middle-income countries. *Nat. Microbiol.* 7, 1337–1347. doi: 10.1038/s41564-022-01184-y

Casaburi, G., Duar, R. M., Brown, H., Mitchell, R. D., Kazi, S., Chew, S., et al. (2021). Metagenomic insights of the infant microbiome community structure and function across multiple sites in the United States. *Sci. Rep.* 11:1472. doi: 10.1038/s41598-020-80583-9

Cuénod, A., Aerni, M., Bagutti, C., Bayraktar, B., Boz, E. S., Carneiro, C. B., et al. (2023). Quality of MALDI-TOF mass spectra in routine diagnostics: results from an international external quality assessment including 36 laboratories from 12 countries using 47 challenging bacterial strains. *Clin. Microbiol. Infect.* 29, 190–199. doi: 10.1016/j.cmi.2022.05.017

De Bruyne, K., Slabbinck, B., Waegeman, W., Vauterin, P., De Baets, B., and Vandamme, P. (2011). Bacterial species identification from MALDI-TOF mass spectra through data analysis and machine learning. *Syst. Appl. Microbiol.* 34, 20–29. doi: 10.1016/j.syapm.2010.11.003

Dematheis, F., Walter, M. C., Lang, D., Antwerpen, M., Scholz, H. C., Pfalzgraf, M.-T., et al. (2022). Machine learning algorithms for classification of MALDI-TOF MS spectra from phylogenetically closely related species *Brucella melitensis*, *Brucella abortus* and *Brucella suis*. *Microorganisms* 10:8. doi: 10.3390/microorganisms10081658

Duar, R. M., Henrick, B. M., Casaburi, G., and Frese, S. A. (2020). Integrating the ecosystem services framework to define Dysbiosis of the breastfed infant Gut: the role of *B. infantis* and Human Milk Oligosaccharides. *Front. Nutr.* 7:33. doi: 10.3389/fnut.2020.00033

Feucherolles, M., Nennig, M., Becker, S. L., Martiny, D., Losch, S., Penny, C., et al. (2021). Combination of MALDI-TOF mass spectrometry and machine learning for rapid antimicrobial resistance screening: the case of Campylobacter spp. *Front. Microbiol.* 12:804484. doi: 10.3389/fmicb.2021.804484

Gato, E., Constanso, I. P., Candela, A., Galán, F., Rodiño-Janeiro, B. K., Arroyo, M. J., et al. (2021). An improved matrix-assisted laser desorption ionization-time of flight mass spectrometry data analysis pipeline for the identification of Carbapenemase-producing *Klebsiella pneumoniae*. *J. Clin. Microbiol.* 59:e0080021. doi: 10.1128/JCM.00800-21

Haider, A., Ringer, M., Kotroczó, Z., Mohácsi-Farkas, C., and Kocsis, T. (2023). The current level of MALDI-TOF MS applications in the detection of microorganisms: a short review of benefits and limitations. *Microbiol. Res.* 14, 80–90. doi: 10.3390/microbiolres14010008

Heilbronner, S., and Foster, T. J. (2021). *Staphylococcus lugdunensis*: a skin commensal with invasive pathogenic potential. *Clin. Microbiol. Rev.* 34:2. doi: 10.1128/CMR.00205-20

Henrick, B. M., Chew, S., Casaburi, G., Brown, H. K., Frese, S. A., Zhou, Y., et al. (2019). *Infantis* EVC001 modulates enteric inflammation in exclusively breastfed infants. *Pediatr. Res.* 86, 749–757. doi: 10.1038/s41390-019-0533-2

Henrick, B. M., Rodriguez, L., Lakshmikanth, T., Pou, C., Henckel, E., Arzoomand, A., et al. (2021). Bifidobacteria-mediated immune system imprinting early in life. *Cells* 184, 3884–3898.e11. doi: 10.1016/j.cell.2021.05.030

Ho, P.-L., Yau, C.-Y., Ho, L.-Y., Chen, J. H. K., Lai, E. L. Y., Lo, S. W. U., et al. (2017). Rapid detection of cfiA metallo-β-lactamase-producing *Bacteroides fragilis* by the combination of MALDI-TOF MS and CarbaNP. *J. Clin. Pathol.* 70, 868–873. doi: 10.1136/jclinpath-2017-204335

Jahan, N. A., Godden, S. M., Royster, E., Schoenfuss, T. C., Gebhart, C., Timmerman, J., et al. (2021). Evaluation of the matrix-assisted laser desorption ionization time of flight mass spectrometry (MALDI-TOF MS) system in the detection of mastitis pathogens from bovine milk samples. *J. Microbiol. Methods* 182:106168. doi: 10.1016/j.mimet.2021.106168

Kim, E., Yang, S.-M., Cho, E.-J., and Kim, H.-Y. (2022b). Evaluation of matrix-assisted laser desorption/ionization time-of-flight mass spectrometry for the discrimination of *Lacticaseibacillus* species. *Food Microbiol.* 107:104094. doi: 10.1016/j.fm.2022.104094

Kim, E., Yang, S.-M., Jung, D.-H., and Kim, H.-Y. (2023). Differentiation between *Weissella cibaria* and *Weissella confusa* using machine-learning-combined MALDI-TOF MS. *Int. J. Mol. Sci.* 24:11009. doi: 10.3390/ijms241311009

Kim, E., Yang, S.-M., Kim, H.-J., and Kim, H.-Y. (2022a). Differentiating between *Enterococcusfaecium* and *Enterococcuslactis* by matrix-assisted laser desorption ionization time-of-flight mass spectrometry. *Foods* 11:7. doi: 10.3390/foods11071046

Letunic, I., and Bork, P. (2019). Interactive tree of life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* 47, W256–W259. doi: 10.1093/nar/gkz239

Lundberg, S., and Lee, S. I., A unified approach to interpreting model predictions. 31st Conference on Neural Information Processing Systems (2017), 1–10.

Parks, D. H., Chuvochina, M., Waite, D. W., Rinke, C., Skarshewski, A., Chaumeil, P.-A., et al. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* 36, 996–1004. doi: 10.1038/nbt.4229

Pereira, C. R., Neia, R. C., Silva, S. B., Williamson, C. H. D., Gillece, J. D., O'Callaghan, D., et al. (2023). Comparison of *Brucella abortus* population structure based on genotyping methods with different levels of resolution. *J. Microbiol. Methods* 211:106772. doi: 10.1016/j.mimet.2023.106772

Rodríguez-Temporal, D., Díez, R., Díaz-Navarro, M., Escribano, P., Guinea, J., Muñoz, P., et al. (2022). Determination of the ability of matrix-assisted laser desorption ionization time-of-flight mass spectrometry to identify high-biofilm-producing strains. *Front. Microbiol.* 13:1104405. doi: 10.3389/fmicb.2022.1104405

Rodríguez-Temporal, D., Herrera, L., Alcaide, F., Domingo, D., Héry-Arnaud, G., van Ingen, J., et al. (2023). Identification of *Mycobacterium abscessus* subspecies by MALDI-TOF mass spectrometry and machine learning. *J. Clin. Microbiol.* 61:e0111022. doi: 10.1128/jcm.01110-22

Sato, H., Teramoto, K., Ishii, Y., Watanabe, K., and Benno, Y. (2011). Ribosomal protein profiling by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry for phylogenety-based subspecies resolution of *Bifidobacterium longum*. *Syst. Appl. Microbiol.* 34, 76–80. doi: 10.1016/j.syapm.2010.07.003

Topić Popović, N., Kazazić, S., Bojanić, K., Strunjak-Perović, I., and Čož-Rakovac, R. (2023). Sample preparation and culture condition effects on MALDI-TOF MS identification of bacteria: a review. *Mass. Spectrom. Rev.* 42, 1589–1603. doi: 10.1002/mas.21739

Treangen, T. J., Ondov, B. D., Koren, S., and Phillippy, A. M. (2014). The harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biol.* 15:524. doi: 10.1186/s13059-014-0524-x

van Oosten, L. N., and Klein, C. D. (2020). Machine learning in mass spectrometry: a MALDI-TOF MS approach to phenotypic antibacterial screening. *J. Med. Chem.* 63, 8849–8856. doi: 10.1021/acs.jmedchem.0c00040

Vatanen, T., Ang, Q. Y., Siegwald, L., Sarker, S. A., Le Roy, C. I., Duboux, S., et al. (2022). A distinct clade of *Bifidobacterium longum* in the gut of Bangladeshi children thrives during weaning. *Cells* 185, 4280–4297.e12. doi: 10.1016/j.cell.2022.10.011

Wang, H. Y., Kuo, C. H., Chung, C. R., Lin, W. Y., Wang, Y. C., Lin, T. W., et al. (2022). Rapid and accurate discrimination of *Mycobacterium abscessus* subspecies based on matrix-assisted laser desorption ionization-time of flight Spectrum and machine learning algorithms. *Biomedicine* 11:45. doi: 10.3390/biomedicines11010045

Wang, J., Xia, C., Wu, Y., Tian, X., Zhang, K., and Wang, Z. (2022). Rapid detection of Carbapenem-resistant *Klebsiella pneumoniae* using machine learning and MALDI-TOF MS platform. *Infect. Drug Resist.* 15, 3703–3710. doi: 10.2147/IDR.S367209

Wang, Y. Y., Xie, L., Zhang, W. Z., Du, X. L., Li, W. G., Bia, L. L., et al. (2023). Application of a core genome sequence typing (cgMLST) pipeline for surveillance of *Clostridioides difficile* in China. *Front. Cell. Infect. Microbiol.* 13:1109153. doi: 10.3389/fcimb.2023.1109153

Weis, C., Cuénod, A., Rieck, B., Dubuis, O., Graf, S., Lang, C., et al. (2022). Direct antimicrobial resistance prediction from clinical MALDI-TOF mass spectra using machine learning. *Nat. Med.* 28, 164–174. doi: 10.1038/s41591-021-01619-9

Weis, C. V., Jutzeler, C. R., and Borgwardt, K. (2020). Machine learning for microbial identification and antimicrobial susceptibility testing on MALDI-TOF mass spectra: a systematic review. *Clin. Microbiol. Infect* 26, 1310–1317. doi: 10.1016/j.cmi.2020.03.014

Xiong, L., Long, X., Ni, L., Wang, L., Zhang, Y., Cui, L., et al. (2023). Comparison of autof Ms1000 and EXS3000 MALDI-TOF MS platforms for routine identification of microorganisms. *Infect. Drug Resist.* 16, 913–921. doi: 10.2147/IDR.S352307

Yahiaoui, R. Y., Goessens, W. H., Stobberingh, E. E., and Verbon, A. (2020). Differentiation between *Streptococcus pneumoniae* and other viridans group *streptococci* by matrix-assisted laser desorption/ionization time of flight mass spectrometry. *Clin. Microbiol. Infect.* 26, 1088.e1–1088.e5. doi: 10.1016/j.cmi.2019.11.024

Yoon, E.-J., and Jeong, S. H. (2021). MALDI-TOF mass spectrometry technology as a tool for the rapid diagnosis of antimicrobial resistance in Bacteria. *Antibiotics (Basel)* 10:982. doi: 10.3390/antibiotics10080982

Yu, J., Lin, Y.-T., Chen, W.-C., Tseng, K.-H., Lin, H.-H., Tien, N., et al. (2023). Direct prediction of carbapenem-resistant, carbapenemase-producing, and colistin-resistant *Klebsiella pneumoniae* isolates from routine MALDI-TOF mass spectra using machine learning and outcome evaluation. *Int. J. Antimicrob. Agents* 61:106799. doi: 10.1016/j.ijantimicag.2023.106799

Zhang, B., Li, L.-Q., Liu, F., and Wu, J.-Y. (2022). Human milk oligosaccharides and infant gut microbiota: molecular structures, utilization strategies and immune function. *Carbohydr. Polym.* 276:118738. doi: 10.1016/j.carbpol.2021.118738

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 22 December 2023
DOI 10.3389/fmicb.2023.1291699

OPEN ACCESS

EDITED BY
Erik Bongcam-Rudloff,
Swedish University of Agricultural Sciences, Sweden

REVIEWED BY
Renuka Dahiya,
University at Buffalo, United States
Zhe-Sheng Chen,
St. John's University, United States
Tao Yang,
Guizhou University of Traditional Chinese Medicine, China

*CORRESPONDENCE
Haibo Zhang
✉ haibozh@gzucm.edu.cn

RECEIVED 10 September 2023
ACCEPTED 07 December 2023
PUBLISHED 22 December 2023

CITATION
Zhang Z, Li D, Xie F, Muhetaer G and Zhang H (2023) The cause-and-effect relationship between gut microbiota abundance and carcinoid syndrome: a bidirectional Mendelian randomization study. *Front. Microbiol.* 14:1291699. doi: 10.3389/fmicb.2023.1291699

The cause-and-effect relationship between gut microbiota abundance and carcinoid syndrome: a bidirectional Mendelian randomization study

Zexin Zhang[1], Dongting Li[2], Fengxi Xie[3], Gulizeba Muhetaer[1] and Haibo Zhang[4,5,6,7]*

[1]The Second Clinical College of Guangzhou University of Chinese Medicine, Guangzhou, China, [2]The Affiliated Guangzhou Hospital of TCM of Guangzhou University of Chinese Medicine, Guangzhou, China, [3]Maoming Hospital, Guangzhou University of Chinese Medicine, Guangzhou, China, [4]The Second Affiliated Hospital of Guangzhou University of Chinese Medicine, Guangzhou, China, [5]Guangdong Key Laboratory of Clinical Research of Chinese Medicine, Guangzhou, China, [6]Guangdong Joint Laboratory of Guangdong, Hong Kong and Macao Chinese Medicine and Immune Diseases, Guangzhou, China, [7]State Key Laboratory of Wet Certificate of Chinese Medicine Jointly Built by the Province and the Ministry, Guangzhou, China

Objective: Carcinoid syndrome (CS) commonly results from neuroendocrine tumors. While active substances are recognized as the main causes of the typical symptoms such as diarrhea and skin flush, the cause-and-effect relationship between gut microbiota abundance and CS remains unclear.

Methods: The Single Nucleotide Polymorphisms (SNPs) related to gut microbiota abundance and CS were obtained from the GWAS summary data. The inverse variance weighted (IVW) method was used to assess the causal relationship between gut microbiota abundance and CS. Additionally, the MR-Egger, Weighted Median model, and Weighted model were employed as supplementary approaches. The heterogeneity function of the TwoSampleMR package was utilized to assess whether SNPs exhibit heterogeneity. The Egger intercept and Presso test were used to assess whether SNPs exhibit pleiotropy. The Leave-One-Out test was employed to evaluate the sensitivity of SNPs. The Steiger test was utilized to examine whether SNPs have a reverse causal relationship. A bidirectional mendelian randomization (MR) study was conducted to elucidate the inferred cause-and-effect relationship between gut microbiota abundance and CS.

Results: The IVW results indicated a causal relationship between 6 gut microbiota taxa and CS. Among the 6 gut microbiota taxa, the genus Anaerofilum (IVW OR: 0.3606, 95%CI: 0.1554–0.8367, *p*-value: 0.0175) exhibited a protective effect against CS. On the other hand, the family Coriobacteriaceae (IVW OR: 3.4572, 95%CI: 1.0571–11.3066, *p*-value: 0.0402), the genus Enterorhabdus (IVW OR: 4.2496, 95%CI: 1.3314–13.5640, *p*-value: 0.0146), the genus Ruminiclostridium6 (IVW OR: 4.0116, 95%CI: 1.2711–12.6604, *p*-value: 0.0178), the genus Veillonella (IVW OR: 3.7023, 95%CI: 1.0155–13.4980, *p*-value: 0.0473) and genus Holdemanella (IVW OR: 2.2400, 95%CI: 1.0376–4.8358, *p*-value: 0.0400) demonstrated

a detrimental effect on CS. The CS was not found to have a reverse causal relationship with the above 6 gut microbiota taxa.

Conclusion: Six microbiota taxa were found to have a causal relationship with CS, and further randomized controlled trials are needed for verification.

KEYWORDS

cause-and-effect relationship, gut microbiota abundance, carcinoid syndrome, active substances, Mendelian randomization study

Background

Carcinoid syndrome (CS) refers to a series of symptoms mediated by various biologically active substances secreted by neuroendocrine tumors (NETs), which mainly located in the gastrointestinal tract and lungs. The two most common manifestations of this syndrome are diarrhea and facial flushing (Vitale, et al., 2023). While some researchers have uncovered that the release of active substances such as serotonin (5-hydroxytryptamine, 5-HT), histamine, kinins, prostaglandins, and tachykinins was a significant factor in causing CS, the mechanisms behind the occurrence of CS remain unclear (Gade et al., 2020). According to relevant reports, the frequency of CS in NETs patients has increased from 11 to 19% (Halperin et al., 2017). Among the population of patients experiencing CS, those who experience diarrhea and facial flushing can reach as high as 80 to 85% (von der Ohe et al., 1993). Diarrhea is typically the initial symptom in patients with CS, sometimes occurring dozens of times per day. It is often the most distressing symptom experienced by CS patients, significantly reducing their quality of lives and increasing healthcare costs (Kimbrough et al., 2019; Perrier et al., 2023). Therefore, early management and intervention for CS are important.

Both 5-HT and the 5-HT pathway play a crucial role in the pathogenesis of CS (Fanciulli et al., 2020). Most CS patients exhibit alterations in tryptophan metabolism, which typically results in elevated concentrations of 5-HT, thereby activating the 5-HT pathway (Kvols and Reubi, 1993). Consequently, telotristat ethyl, an inhibitor of 5-HT synthesis, has been approved for treating refractory diarrhea in CS, highlighting the significance of the 5-HT pathway in CS (Fanciulli et al., 2020). In CS patients, 5-HT can stimulate intestinal motility and secretion, leading to increased bowel frequency and reduced stool viscosity (Hendrix et al., 1957; von der Ohe et al., 1993). Additionally, other bioactive substances such as prostaglandins also induce intestinal motility and enhance fluid secretion in the gastrointestinal tract, causing diarrhea (Metz et al., 1981). Researchers indicated that substances like prostaglandins, histamine, and substance P can disrupt intestinal secretion and motility, leading to the release of gastrin from enterochromaffin cells in the small intestine. Elevated levels of gastrin can contribute to the cyclic nature of diarrhea (de Celis Ferrari et al., 2018). Moreover, histamine and substance P can cause vasodilation of skin capillaries, resulting in flushing of the skin (Grahame-Smith, 1987). In observations of skin flushing in CS, Schaffalitzky De Muckadell et al. (1986) found increasing concentrations of neurokinin A, neurokinin K, and tachykinin-like peptides. This finding underscored the role of tachykinins in CS. Researchers have reported the presence of substance P, a potent vasodilator, in carcinoid tumor tissue (Ratzenhofer et al., 1981). Evidence also suggested that injecting substance P into healthy individuals can cause transient facial flushing (Schaffalitzky De Muckadell et al., 1986). This finding implied that substance P might be one of the underlying factors contributing to skin flushing in CS. Currently, tryptophan hydroxylase inhibitors and somatostatin analogs are widely used for CS treatment. However, drug resistance and poor tolerability are frequently reported (de Celis Ferrari et al., 2018; Gade et al., 2020). Therefore, there is an urgent need to establish the potential causative relationships in CS, to offer more comprehensive strategies for its treatment.

The intestinal microbiota refers to the community of bacteria, viruses, archaea, fungi, and protozoa that inhabit in the gastrointestinal tract. Numerous studies indicated that tryptophan metabolites play a significant role in regulating gastrointestinal function (Bosi et al., 2020). On the other hand, the intestinal microbiota plays a crucial role in promoting the production of 5-HT. Research has found that metabolites of Clostridium species can upregulate the expression of tryptophan hydroxylase (Tph) gene in enterochromaffin cells, thereby promoting the production of 5-HT. Microbiota-specific metabolites such as short-chain fatty acids, alpha-tocopherol, tyramine, and p-aminobenzoate can stimulate the expression of TPH1 and the release of 5-HT (Yano et al., 2015). In addition, microbes within the human intestines can also produce and degrade histamine (Sanchez-Perez et al., 2022). The gut microbiota and its metabolic product, acetate, can activate the innate immune pathway in intestinal endocrine cells, thereby increasing the secretion of endocrine peptides such as tachykinins (Tumurkhuu et al., 2018). All these findings indicated a potential correlation between gut microbiota and CS, yet currently, there is scarce research focused on this aspect.

Mendelian randomization (MR) study is an analytical method that utilizes genetic variations associated with exposure as instrumental variables to assess potential causal relationships between exposures and outcomes. MR takes advantage of alleles that are randomly segregated during meiotic gamete formation. Since genetic variations precede disease progression and are not influenced by postnatal lifestyle and environmental factors, MR can minimize the impact of confounding factors to a great extent (Sekula et al., 2016).

In this study, a large-scale genome-wide association study (GWAS) dataset was employed to conduct a bidirectional MR analysis, investigating the potential causal relationship between gut microbiota and CS. This approach addresses the existing research gaps in the field and the results will offer novel strategies for the treatment of CS.

Materials and methods

Data sources

The gut microbiota abundance data in relation to CS were sourced from the IEU Open GWAS project database,[1] a database of 246,376,709,462 genetic associations from 42,351 GWAS summary datasets, for querying or download. An exploration was undertaken within the GWAS summary data, utilizing the search query "gut microbiota abundance," which yielded a total of 211 outcomes. After excluding 15 records that were categorized as "unknown," 196 relevant results were finally used.

The study of large-scale association analyses identified host factors influencing human gut microbiome composition was curated and analyzed by MiBioGen consortium. This study included genome-wide genotypes and 16S fecal microbiome data from 18,340 individuals (24 cohorts). This study included a total of 211 bacterial taxonomic units, involving 131 genera, 35 families, 20 orders, 16 classes, and 9 phyla (Kurilshikov et al., 2021; MiBioGen consortium, 2023).

The summary data for Genome-Wide Association Study (GWAS) on CS was obtained from the FinnGen biobank analysis round 5. The dataset comprised 16,380,446 SNPs, with 211,123 controls and 161 cases, as reported in the 2021 publication. The participants included individuals of European descent, encompassing both males and females.

Screening of instrumental variables

In the context of MR study, it is generally required to adhere to three foundational prerequisite assumptions, specifically: (1) the assumption of associativity, (2) the presumption of independence, and (3) the principle of exclusivity (Smith and Ebrahim, 2003).

The assumption of associativity entails that the selected instrumental variables are closely correlated with the exposure of interest, allowing us to confidently employ them as substitutes for the exposure. Typically, we use criteria such as $p < 1e^{-05}$, $r^2 = 0.001$, and Kb = 10,000 as three fundamental thresholds (Sanna et al., 2019). Furthermore, in order to ensure the reliability of these screened instrumental variables, the application of an *F*-test can be employed to eliminate weak instruments. Weak instruments are commonly defined by an F-statistic value of less than 10. The formula to calculate *F*-statistic value is as follow:

$$F = \left(\frac{Beta}{SE}\right)^2$$ (Casas et al., 2006).

Among them, Beta refers to the effect size of the SNP on the exposure, and SE (Standard error) refers to the standard error of Beta. The assumption of independence in MR refers to the genetic variants (genotypes or genetic variations) being unrelated to other factors that could potentially affect the outcomes when they are randomly allocated. The assumption of independence necessitates that the distribution of genotypes among participants is random and not influenced by other possible confounding factors.

The assumption of exclusivity refers to the genetic variants (genotypes or genetic variations) being allocated among participants in a mutually exclusive manner, with each participant being assigned to a specific genotype only. This assumption ensures that genotypes do not overlap or coexist among participants, thereby allowing the association between genotypes and exposure to be accurately interpreted and assessed.

To ensure the enforcement of the aforementioned-assumptions, we subjected the selected instrumental variables (IVs) to tests for horizontal pleiotropy and heterogeneity. For the assessment of pleiotropy, we utilized the Egger intercept and MR Presso test. Heterogeneity assessment was carried out using the heterogeneity function of the TwoSampleMR package in R language 4.3.1. Furthermore, we conducted Steiger test to ascertain the exclusion of SNPs with reverse causal relationships. Leave-one-out sensitivity analysis was employed to evaluate the stability of each SNP's influence on the outcome.

Statistical analysis

IVW is used as the primary method to assess the causal relationship between gut microbiota abundance and CS. The strength of IVW lies in its ability to provide a more robust outcome; if a SNP in the instrumental variables is invalid, it can introduce bias to the results (Burgess et al., 2016). Additionally, we employed three alternative methods: MR-Egger regression, weighted median model, and weighted mode. MR-Egger regression is a technique that refines the IVW method. It takes into consideration the intercept term in the regression model to detect and correct for pleiotropy effects. It relies on the assumptions of the InSIDE (Instrument Strength Independent of Direct Effect) and NOME (No Measurement Error) principles (Bowden et al., 2015). Weighted median model and weighted mode share similarities in using the reciprocal of outcome variance as weights. The difference lies in their methods of aggregation. The weighted median model employs a weighted median approach, while the weighted mode employs a weighted mode approach.

Results

Characteristics of SNPs

According to the filtering criteria of $p < 1e^{-05}$, $r^2 = 0.001$, and Kb = 10,000, a total of 224, 434, 512, 486, 498, 280, and 125 SNPs were, respectively, obtained from the class, family, genus1, genus2, genus3, order, and phylum of the gut microbiota. Additionally, from the outcome "Carcinoid symptom," 179, 339, 413, 395, 391, 217, and 103 SNPs were extracted for analysis. The *F*-values of all instrumental variables were greater than 10, indicating no weak instrumental variables in this study. The detail of the characteristics of SNPs were shown in Supplementary Table S1.

Mendelian randomization analysis

The IVW results indicated that a total of 8 types of gut microbiota are associated with CS. As the IVW results of Gut

1 https://gwas.mrcieu.ac.uk/

microbiota abundance (class Coriobacteriia id.809), Gut microbiota abundance (family Coriobacteriaceae id.811), and Gut microbiota abundance (order Coriobacteriales id.810) were consistent, for a more precise outcome, we retained only the lowest taxonomic level, Gut microbiota abundance (family Coriobacteriaceae id.811), for presentation. This yielded a total of 6 gut microbiota types that are correlated with CS.

Specifically, we found that genus Anaerofilum (IVW odds ratio [OR] = 0.3606, 95% confidence interval [CI]: 0.1554–0.8367, $p = 0.0175$) had a protective effect on CS. While family Coriobacteriaceae (IVW OR = 3.4572, 95%CI: 1.0571–11.3066, $p = 0.0402$), genus Enterorhabdus (IVW OR = 4.2496, 95%CI: 1.3314–13.5640, p value: 0.0146), genus Ruminiclostridium6 (IVW OR: 4.0116, 95%CI: 1.2711–12.6604, p value: 0.0178), genus Veillonella (IVW OR: 3.7023, 95%CI: 1.0155–13.4980, p value: 0.0473) and genus Holdemanella (IVW OR: 2.2400, 95%CI: 1.0376–4.8358, p value: 0.0400) demonstrated a detrimental effect on CS (Figure 1 and Supplementary Table S2). The forest plot displayed the odds ratio (OR) and 95% confidence interval for each SNP, followed by the aggregation of all SNPs using IVW and MR Egger (Supplementary Figure S1). The scatter plot illustrated the effect distribution of all SNPs, demonstrating trends for four different MR analysis methods (Supplementary Figure S2).

The heterogeneity test indicated that there was no heterogeneity in the above results for the SNPs. The funnel plot showed that all SNPs are distributed evenly on both sides of a straight line, which confirms this observation (Supplementary Table S3 and Supplementary Figure S3). The Egger intercept and MR presso test indicated the absence of pleiotropy in the above results for the SNPs, demonstrating the reliability of this study (Supplementary Tables S4, S5). The Leave-one-out sensitivity analysis was conducted by excluding individual SNPs to assess the overall effect change, and no single SNP was found to have a significant impact on the outcome (Supplementary Figure S4). Lastly, we did not observe a reverse causal relationship between CS and the afore-mentioned gut microbiota.

Discussion

This study represented the inaugural attempt to investigate the causal relationship between gut microbiota and carcinoid syndrome (CS) utilizing a bidirectional MR analysis. The findings indicated the presence of causal associations between 6 specific gut microbiota taxa and CS, while no reverse causal relationship exists between CS and gut microbiota. At present, there was no prior research delved into the relationship between gut microbiota and CS.

CS is a collection of clinical symptoms caused by excessive secretion of mediators such as serotonin (5-HT), substance P, histamine, and prostaglandins (Yano et al., 2015). Research has shown a close correlation between gut microbiota and certain bioactive substances, but the relationship between gut microbiota and CS remains unclear. A prior study indicated that microbial metabolites in the gut, such as short-chain fatty acids, α-tocopherol, tyramine, and p-aminobenzoate, can stimulate the release of 5-HT (Yano et al., 2015). Moreover, the human gut microbiota can both produce and degrade histamine (Sanchez-Perez et al., 2022). These studies highlighted a potential correlation between gut microbiota and CS; however, further evidence is needed to confirm this association. This study employed MR analysis to reveal causal relationships between 6 specific gut microbiota taxa and CS, addressing the gaps in existing research and contributing to the refinement of therapeutic approaches for CS.

Our study revealed the causal relationships between a total of 6 specific gut microbiota taxa and CS. Among them, genus Anaerofilum was the only protective bacterial group identified in CS. Research indicated that the expression of functional genes in genus Anaerofilum effectively promoted the tryptophan-indole metabolic pathway in the intestines (Sun M. et al., 2020; Sun X. Z. et al., 2020). *In vitro* experiments have demonstrated that adding a certain concentration of indole induces the expression of tight junction proteins in intestinal epithelial cells, thereby restoring intestinal barrier function (Sun M. et al., 2020; Sun X. Z. et al., 2020). Intestinal barrier function is primarily provided by the tight junctions of adjacent epithelial cells (Camilleri et al., 2017), and disruption of tight junction function has been observed to lead to diarrhea in animal models (Halliez et al., 2016). Additionally, aside from the indole pathway, tryptophan also participates in the kynurenine pathway and the serotonin (5-HT) pathway. Prior studies showed that nearly all CS patients experience abnormal tryptophan metabolism, leading to a significant increase in blood 5-HT concentrations. 5-HT and its metabolites are believed to play a crucial role in the development of typical symptoms in CS patients

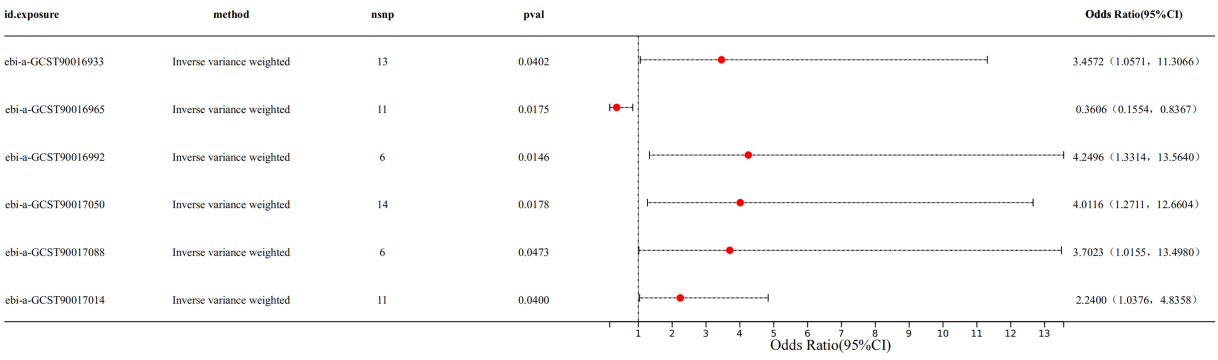

FIGURE 1
The IVW results of 6 gut microbiota against CS.

(Fanciulli et al., 2020). These symptoms include diarrhea (von der Ohe et al., 1993; Boutzios and Kaltsas, 2015), intestinal obstruction, and others (Connolly and Pellikka, 2006; Hannah-Shmouni et al., 2016). Therefore, the protective effect of genus Anaerofilum in CS may primarily exist in two aspects: On the one hand, genus Anaerofilum directly reduces the occurrence of diarrhea by promoting the tryptophan-indole metabolic pathway to repair the intestinal barrier; On the other hand, the active tryptophan-indole pathway effectively inhibits the tryptophan-5-HT pathway, reducing intestinal motility and secretory reflex, thereby indirectly improving diarrhea symptoms.

In addition to the genus Anaerofilum, we identified 5 other gut microbial populations as risk factors for CS. In a murine model, it was discovered that genus Enterorhabdus showed a positive correlation with tryptophan levels and inhibited the indoleamine pathway of tryptophan metabolism (Deng et al., 2021). As mentioned above, inhibiting the indoleamine pathway of tryptophan metabolism can indirectly increase the concentration of 5-HT. Additionally, the genus Enterorhabdus is also associated with intestinal barrier function. A study indicated that the genus Enterorhabdus can increase the production of lyso-phosphatidylcholine, thereby promoting the release of pro-inflammatory cytokines and damaging the intestinal epithelial barrier of murine (Tang et al., 2021).

The family Coriobacteriaceae was a significant risk factor for inducing CS. Research has shown that Coriobacteriaceae UCG-002 can produce cytotoxic compounds such as phenol and *p*-cresol, consequently altering epithelial permeability and reducing epithelial barrier function (Saito et al., 2018; Yu et al., 2023). Simultaneously, Tian et al. (2023) also observed an increased relative abundance of Coriobacteriaceae UCG-002 in cases of intestinal damage, and a positive correlation between Coriobacteriaceae UCG-002 and the inflammatory cytokine TNF-α. TNF-α, a type of tumor necrosis factor (TNF), inhibits the Wnt/β-catenin pathway, thereby compromising the stability of intestinal epithelium (Wu et al., 2020). Therefore, the family Coriobacteriaceae may potentially exacerbate certain symptoms in patients with carcinoid syndrome by inducing the production of various harmful mediators that damage the epithelial barrier function.

In a MR study exploring the relationship between gut microbiota and asthma, genus Ruminiclostridium 6 was found to be associated with the incidence of moderate to severe asthma (Li et al., 2023). Bronchial asthma is a heterogeneous disease characterized by chronic inflammation of the airways (Kaczynska et al., 2021). Certain cells such as eosinophils, neutrophils, and endogenous inflammatory mediators like leukotrienes and histamine participate in the inflammatory processes in the airways (Barnes, 2008). Additionally, many regulatory peptides such as kinins are shown to be involved in the regulation of asthma-related inflammation and airway hyperresponsiveness (Kaczynska et al., 2018; Pavon-Romero et al., 2021). Interestingly, about 20% of CS patients also experience bronchoconstriction mediated by kinins and bradykinins (Cunningham et al., 2008).

In addition to the previously mentioned 5-HT and bradykinin, histamine also plays a significant role in CS. Researchers suggested that histamine might be a potential mediator for the facial flushing and bronchospasm symptoms observed in patients with colorectal CS (de Celis Ferrari, et al., 2018). The pathogenic bacterium genus Veillonella was confirmed to have a strong ability to induce mast cells to release histamine (Nygren and Dahlen, 1981). Furthermore, in fecal samples from individuals with higher asthma frequencies, genus Veillonella was found to be enriched, and metabolic profiling indicated the importance of histidine metabolism in the asthma process, which leads to the formation of histamine upon histidine decarboxylation (Lee-Sarwar et al., 2022). Therefore, genus Veillonella might exacerbate the development of CS by inducing histamine production.

The genus Holdemanella is also identified as a risk factor for CS. However, at present, there are no reports linking this genus to endocrine mediators associated with CS. Hu et al. (2020) discovered that the expression of tight junction proteins in the ileum was significantly increased, and the relative abundance of genus Holdemanella in the gut microbiota was reduced in a weaned piglet model supplemented with catechin. Therefore, we speculated that genus Holdemanella might exacerbate diarrhea symptoms in patients with CS by potentially affecting intestinal mucosal barrier function.

Conclusion

In conclusion, this study utilized a bidirectional mendelian randomization analysis and identified 6 gut microbial populations that are causally associated with carcinoid syndrome. This research represented the first instance of uncovering a causal link between gut microbiota and CS, offering a novel strategy for its treatment. Further validation through additional randomized controlled trails are warranted in the future to solidify these findings.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

ZZ: Conceptualization, Methodology, Writing – original draft. DL: Writing – original draft. FX: Writing – original draft. GM: Investigation, Writing – original draft. HZ: Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was funded by the Guangdong Provincial Science and Technology Innovation Strategy Special Fund (Guangdong-Hong Kong-Macau Joint Lab, grant number 2020B1212030006) and Guangzhou University of Traditional Chinese Medicine's "Double First Class" and high-level university discipline collaborative innovation team: Traditional Chinese Medicine Reverse Lung Cancer Resistance Innovation Team (2021xk60).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2023.1291699/full#supplementary-material

References

Barnes, P. J. (2008). The cytokine network in asthma and chronic obstructive pulmonary disease. *J. Clin. Invest.* 118, 3546–3556. doi: 10.1172/JCI36130

Bosi, A., Banfi, D., Bistoletti, M., Giaroni, C., and Baj, A. (2020). Tryptophan metabolites along the microbiota-gut-brain Axis: an Interkingdom communication system influencing the gut in health and disease. *Int. J. Tryptophan. Res.* 13:1178646920928984. doi: 10.1177/1178646920928984

Boutzios, G., and Kaltsas, G. (2015). Clinical syndromes related to gastrointestinal neuroendocrine neoplasms. *Front. Horm. Res.* 44, 40–57. doi: 10.1159/000382053

Bowden, J., Davey Smith, G., and Burgess, S. (2015). Mendelian randomization with invalid instruments: effect estimation and bias detection through egger regression. *Int. J. Epidemiol.* 44, 512–525. doi: 10.1093/ije/dyv080

Burgess, S., Dudbridge, F., and Thompson, S. G. (2016). Combining information on multiple instrumental variables in Mendelian randomization: comparison of allele score and summarized data methods. *Stat. Med.* 35, 1880–1906. doi: 10.1002/sim.6835

Camilleri, M., Sellin, J. H., and Barrett, K. E. (2017). Pathophysiology, evaluation, and Management of Chronic Watery Diarrhea. *Gastroenterology* 152, 515–532.e2. doi: 10.1053/j.gastro.2016.10.014

Casas, J. P., Shah, T., Cooper, J., Hawe, E., McMahon, A. D., Gaffney, D., et al. (2006). Insight into the nature of the CRP-coronary event association using Mendelian randomization. *Int. J. Epidemiol.* 35, 922–931. doi: 10.1093/ije/dyl041

Connolly, H. M., and Pellikka, P. A. (2006). Carcinoid heart disease. *Curr. Cardiol. Rep* 8, 96–101. doi: 10.1007/s11886-006-0019-9

Cunningham, J. L., Janson, E. T., Agarwal, S., Grimelius, L., and Stridsberg, M. (2008). Tachykinins in endocrine tumors and the carcinoid syndrome. *Eur. J. Endocrinol.* 159, 275–282. doi: 10.1530/EJE-08-0196

de Celis, R., Ferrari, A. C., Glasberg, J., and Riechelmann, R. P. (2018). Carcinoid syndrome: update on the pathophysiology and treatment. *Clinics* 73:e490s. doi: 10.6061/clinics/2018/e490s

Deng, Y., Zhou, M., Wang, J., Yao, J., Yu, J., Liu, W., et al. (2021). Involvement of the microbiota-gut-brain axis in chronic restraint stress: disturbances of the kynurenine metabolic pathway in both the gut and brain. *Gut Microbes* 13, 1–16. doi: 10.1080/19490976.2020.1869501

Fanciulli, G., Ruggeri, R. M., Grossrubatscher, E., Calzo, F. L., Wood, T. D., Faggiano, A., et al. (2020). Serotonin pathway in carcinoid syndrome: clinical, diagnostic, prognostic and therapeutic implications. *Rev. Endocr. Metab. Disord.* 21, 599–612. doi: 10.1007/s11154-020-09547-8

Gade, A. K., Olariu, E., and Douthit, N. T. (2020). Carcinoid syndrome: a review. *Cureus* 12:e7186. doi: 10.7759/cureus.7186

Grahame-Smith, D. G. (1987). What is the cause of the carcinoid flush? *Gut* 28, 1413–1416. doi: 10.1136/gut.28.11.1413

Halliez, M. C., Motta, J. P., Feener, T. D., Guerin, G., LeGoff, L., Francois, A., et al. (2016). Giardia duodenalis induces paracellular bacterial translocation and causes postinfectious visceral hypersensitivity. *Am. J. Physiol. Gastrointest. Liver Physiol.* 310, G574–G585. doi: 10.1152/ajpgi.00144.2015

Halperin, D. M., Shen, C., Dasari, A., Xu, Y., Chu, Y., Zhou, S., et al. (2017). Frequency of carcinoid syndrome at neuroendocrine tumour diagnosis: a population-based study. *Lancet Oncol.* 18, 525–534. doi: 10.1016/S1470-2045(17)30110-9

Hannah-Shmouni, F., Stratakis, C. A., and Koch, C. A. (2016). Flushing in (neuro) endocrinology. *Rev. Endocr. Metab. Disord.* 17, 373–380. doi: 10.1007/s11154-016-9394-8

Hendrix, T. R., Atkinson, M., Clifton, J. A., and Ingelfinger, F. J. (1957). The effect of 5-hydroxytryptamine on intestinal motor function in man. *Am. J. Med.* 23, 886–893. doi: 10.1016/0002-9343(57)90298-x

Hu, R., He, Z., Liu, M., Tan, J., Zhang, H., Hou, D. X., et al. (2020). Dietary protocatechuic acid ameliorates inflammation and up-regulates intestinal tight junction proteins by modulating gut microbiota in LPS-challenged piglets. *J. Anim. Sci. Biotechnol.* 11:92. doi: 10.1186/s40104-020-00492-9

Kaczynska, K., Zajac, D., Wojciechowski, P., and Jampolska, M. (2021). Regulatory peptides in asthma. *Int. J. Mol. Sci.* 22:13656. doi: 10.3390/ijms222413656

Kaczynska, K., Zajac, D., Wojciechowski, P., Kogut, E., and Szereda-Przestaszewska, M. (2018). Neuropeptides and breathing in health and disease. *Pulm. Pharmacol. Ther.* 48, 217–224. doi: 10.1016/j.pupt.2017.12.001

Kimbrough, C. W., Beal, E. W., Dillhoff, M. E., Schmidt, C. R., Pawlik, T. M., Lopez-Aguiar, A. G., et al. (2019). Influence of carcinoid syndrome on the clinical characteristics and outcomes of patients with gastroenteropancreatic neuroendocrine tumors undergoing operative resection. *Surgery* 165, 657–663. doi: 10.1016/j.surg.2018.09.008

Kurilshikov, A., Medina-Gomez, C., Bacigalupe, R., Radjabzadeh, D., Wang, J., Demirkan, A., et al. (2021). Large-scale association analyses identify host factors influencing human gut microbiome composition. *Nat. Genet.* 53, 156–165. doi: 10.1038/s41588-020-00763-1

Kvols, L. K., and Reubi, J. C. (1993). Metastatic carcinoid tumors and the malignant carcinoid syndrome. *Acta Oncol.* 32, 197–201. doi: 10.3109/02841869309083912

Lee-Sarwar, K., Dedrick, S., Momeni, B., Kelly, R. S., Zeiger, R. S., O'Connor, G. T., et al. (2022). Association of the gut microbiome and metabolome with wheeze frequency in childhood asthma. *J. Allergy Clin. Immunol.* 150, 325–336. doi: 10.1016/j.jaci.2022.02.005

Li, R., Guo, Q., Zhao, J., Kang, W., Lu, R., Long, Z., et al. (2023). Assessing causal relationships between gut microbiota and asthma: evidence from two sample Mendelian randomization analysis. *Front. Immunol.* 14:1148684. doi: 10.3389/fimmu.2023.1148684

Metz, S. A., McRae, J. R., and Robertson, R. P. (1981). Prostaglandins as mediators of paraneoplastic syndromes: review and up-date. *Metabolism* 30, 299–316. doi: 10.1016/0026-0495(81)90156-6

MiBioGen consortium. (2023). MiBioGen. Available at: https://mibiogen.gcc.rug.nl/ (Accessed September 16, 2022)

Nygren, H., and Dahlen, G. (1981). Complement-dependent histamine release from rat peritoneal mast cells, induced by lipopolysaccharides from *Bacteroides oralis*, fusobacterium nucleatum and *Veillonella parvula*. *J. Oral Pathol.* 10, 87–94. doi: 10.1111/j.1600-0714.1981.tb01253.x

Pavon-Romero, G. F., Serrano-Perez, N. H., Garcia-Sanchez, L., Ramirez-Jimenez, F., and Teran, L. M. (2021). Neuroimmune pathophysiology in asthma. *Front.Cell.Dev.Biol.* 9:663535. doi: 10.3389/fcell.2021.663535

Perrier, M., Mouawad, C., Gueguen, D., Thome, B., Lapeyre-Mestre, M., and Walter, T. (2023). Health care resource use and costs among patients with carcinoid syndrome in France: analysis of the National Health Insurance Database. *Clin. Res. Hepatol. Gas* 47:102177. doi: 10.1016/j.clinre.2023.102177

Ratzenhofer, M., Gamse, R., Hofler, H., Aubock, L., Popper, H., Pohl, P., et al. (1981). Substance P in the argentaffin carcinoid of the caecum: biochemical and biological characterization. *Virchows Arch. A Pathol. Anat. Histol.* 392, 21–31. doi: 10.1007/BF00430545

Saito, Y., Sato, T., Nomoto, K., and Tsuji, H. (2018). Identification of phenol- and p-cresol-producing intestinal bacteria by using media supplemented with tyrosine and its metabolites. *FEMS Microbiol. Ecol.* 94:fiy125. doi: 10.1093/femsec/fiy125 Erratum in: *FEMS Microbiol Ecol.* 2019 Apr 1;95: Erratum in: *FEMS Microbiol Ecol.* 96.

Sanchez-Perez, S., Comas-Baste, O., Duelo, A., Veciana-Nogues, M. T., Berlanga, M., Latorre-Moratalla, M. L., et al. (2022). Intestinal Dysbiosis in patients with histamine intolerance. *Nutrients* 14:1774. doi: 10.3390/nu14091774

Sanna, S., van Zuydam, N. R., Mahajan, A., Kurilshikov, A., Vich Vila, A., Vosa, U., et al. (2019). Causal relationships among the gut microbiome, short-chain fatty acids and metabolic diseases. *Nat. Genet.* 51, 600–605. doi: 10.1038/s41588-019-0350-x

Schaffalitzky De Muckadell, O. B., Aggestrup, S., and Stentoft, P. (1986). Flushing and plasma substance P concentration during infusion of synthetic substance P in normal man. *Scand. J. Gastroenterol.* 21, 498–502. doi: 10.3109/00365528609015169

Sekula, P., Del Greco, M. F., Pattaro, C., and Kottgen, A. (2016). Mendelian randomization as an approach to assess causality using observational data. *J. Am. Soc. Nephrol.* 27, 3253–3265. doi: 10.1681/ASN.2016010098

Smith, G. D., and Ebrahim, S. (2003). 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidemiol.* 32, 1–22. doi: 10.1093/ije/dyg070

Sun, M., Ma, N., He, T., Johnston, L. J., and Ma, X. (2020). Tryptophan (Trp) modulates gut homeostasis via aryl hydrocarbon receptor (AhR). *Crit. Rev. Food Sci. Nutr.* 60, 1760–1768. doi: 10.1080/10408398.2019.1598334

Sun, X. Z., Zhao, D. Y., Zhou, Y. C., Wang, Q. Q., Qin, G., and Yao, S. K. (2020). Alteration of fecal tryptophan metabolism correlates with shifted microbiota and may be involved in pathogenesis of colorectal cancer. *World J. Gastroenterol.* 26, 7173–7190. doi: 10.3748/wjg.v26.i45.7173

Tang, X., Wang, W., Hong, G., Duan, C., Zhu, S., Tian, Y., et al. (2021). Gut microbiota-mediated lysophosphatidylcholine generation promotes colitis in intestinal epithelium-specific Fut2 deficiency. *J. Biomed. Sci.* 28:20. doi: 10.1186/s12929-021-00711-z

Tian, S., Zhao, Y., Qian, L., Jiang, S., Tang, Y., and Han, T. (2023). DHA-enriched phosphatidylserine alleviates high fat diet-induced jejunum injury in mice by modulating gut microbiota. *Food Funct.* 14, 1415–1429. doi: 10.1039/d2fo03019e

Tumurkhuu, G., Dagvadorj, J., Porritt, R. A., Crother, T. R., Shimada, K., Tarling, E. J., et al. (2018). *Chlamydia pneumoniae* hijacks a host autoregulatory IL-1β loop to drive foam cell formation and accelerate atherosclerosis. *Cell Metab.* 28, 432–448.e4. doi: 10.1016/j.cmet.2018.05.027

Vitale, G., Carra, S., Alessi, Y., Campolo, F., Pandozzi, C., Zanata, I., et al. (2023). Carcinoid syndrome: preclinical models and future therapeutic strategies. *Int. J. Mol. Sci.* 24:3610. doi: 10.3390/ijms24043610

von der Ohe, M. R., Camilleri, M., Kvols, L. K., and Thomforde, G. M. (1993). Motor dysfunction of the small bowel and colon in patients with the carcinoid syndrome and diarrhea. *N. Engl. J. Med.* 329, 1073–1078. doi: 10.1056/NEJM199310073291503

Wu, H., Xie, S., Miao, J., Li, Y., Wang, Z., Wang, M., et al. (2020). *Lactobacillus reuteri* maintains intestinal epithelial regeneration and repairs damaged intestinal mucosa. *Gut Microbes* 11, 997–1014. doi: 10.1080/19490976.2020.1734423

Yano, J. M., Yu, K., Donaldson, G. P., Shastri, G. G., Ann, P., Ma, L., et al. (2015). Indigenous bacteria from the gut microbiota regulate host serotonin biosynthesis. *Cells* 161, 264–276. doi: 10.1016/j.cell.2015.02.047

Yu, Z., Li, D., and Sun, H. (2023). Herba Origani alleviated DSS-induced ulcerative colitis in mice through remolding gut microbiota to regulate bile acid and short-chain fatty acid metabolisms. *Biomed. Pharmacother.* 161:114409. doi: 10.1016/j.biopha.2023.114409

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 12 January 2024
DOI 10.3389/fmicb.2023.1331233

OPEN ACCESS

EDITED BY
Domenica D'Elia,
National Research Council (CNR), Italy

REVIEWED BY
Jan Zrimec,
National Institute of Biology (NIB), Slovenia
Gianvito Pio,
University of Bari Aldo Moro, Italy

*CORRESPONDENCE
Balázs Ligeti
ligeti.balazs@itk.ppke.hu

RECEIVED 31 October 2023
ACCEPTED 11 December 2023
PUBLISHED 12 January 2024

CITATION
Ligeti B, Szepesi-Nagy I, Bodnár B, Ligeti-Nagy N and Juhász J (2024) ProkBERT family: genomic language models for microbiome applications. *Front. Microbiol.* 14:1331233. doi: 10.3389/fmicb.2023.1331233

ProkBERT family: genomic language models for microbiome applications

Balázs Ligeti[1]*, István Szepesi-Nagy[1], Babett Bodnár[1], Noémi Ligeti-Nagy[2] and János Juhász[1,3]

[1]Faculty of Information Technology and Bionics, Pázmány Péter Catholic University, Budapest, Hungary, [2]Language Technology Research Group, HUN-REN Hungarian Research Centre for Linguistics, Budapest, Hungary, [3]Institute of Medical Microbiology, Semmelweis University, Budapest, Hungary

Background: In the evolving landscape of microbiology and microbiome analysis, the integration of machine learning is crucial for understanding complex microbial interactions, and predicting and recognizing novel functionalities within extensive datasets. However, the effectiveness of these methods in microbiology faces challenges due to the complex and heterogeneous nature of microbial data, further complicated by low signal-to-noise ratios, context-dependency, and a significant shortage of appropriately labeled datasets. This study introduces the ProkBERT model family, a collection of large language models, designed for genomic tasks. It provides a generalizable sequence representation for nucleotide sequences, learned from unlabeled genome data. This approach helps overcome the above-mentioned limitations in the field, thereby improving our understanding of microbial ecosystems and their impact on health and disease.

Methods: ProkBERT models are based on transfer learning and self-supervised methodologies, enabling them to use the abundant yet complex microbial data effectively. The introduction of the novel Local Context-Aware (LCA) tokenization technique marks a significant advancement, allowing ProkBERT to overcome the contextual limitations of traditional transformer models. This methodology not only retains rich local context but also demonstrates remarkable adaptability across various bioinformatics tasks.

Results: In practical applications such as promoter prediction and phage identification, the ProkBERT models show superior performance. For promoter prediction tasks, the top-performing model achieved a Matthews Correlation Coefficient (MCC) of 0.74 for *E. coli* and 0.62 in mixed-species contexts. In phage identification, ProkBERT models consistently outperformed established tools like VirSorter2 and DeepVirFinder, achieving an MCC of 0.85. These results underscore the models' exceptional accuracy and generalizability in both supervised and unsupervised tasks.

Conclusions: The ProkBERT model family is a compact yet powerful tool in the field of microbiology and bioinformatics. Its capacity for rapid, accurate analyses and its adaptability across a spectrum of tasks marks a significant advancement in machine learning applications in microbiology. The models are available on GitHub (https://github.com/nbrg-ppcu/prokbert) and HuggingFace (https://huggingface.co/nerualbioinfo) providing an accessible tool for the community.

KEYWORDS

genomic language models, language models, promoter, phage, BERT, transformer models, LCA tokenization, machine learning in microbiology

1 Introduction

Numerous tasks in bioinformatics involve classifying or labeling sequence data such as predicting genes (Lukashin and Borodovsky, 1998; Delcher et al., 1999; Sommer and Salzberg, 2021), annotating sequence features (Aziz et al., 2008; Seemann, 2014; Tatusova et al., 2016; Meyer et al., 2019), etc. A significant challenge in this field is deriving efficient vector representations from these sequences (Zhang et al., 2023). Classification tasks related to sequences—like classifying assembled contigs into MAGs (metagenome-assembled-genomes) or analyzing AMR-associated genes—are often addressed by initially categorizing the data into bins or using simple composition-based representations, such as k-mer frequency distributions. A common method involves converting sequences into a basic presence-absence vector, indicating whether a particular genome contains specific sequence features like mutations, motifs, or other patterns. However, a drawback of this method is that proximity in this representation space doesn't always imply semantic similarity. Another prevalent representation uses hidden Markov models (Durbin et al., 1998), where the model parameters encapsulate the essential properties of the sequences. Yet, integrating such models with machine learning algorithms like support vector machines or random forests can be complex. Despite this, hidden Markov models have demonstrated their effectiveness in classification tasks and provide highest quality annotations (Zdobnov and Apweiler, 2001; Cantalapiedra et al., 2021).

Neural network-based representations have distinct advantages, primarily their compatibility with a wide range of machine-learning tools, including autoML and statistical frameworks. Past research has highlighted the effectiveness of neural network representations for sequences, with a variety of classification tasks addressed using networks such as CNNs and RNNs (Min et al., 2017). These networks have been employed in areas like motif discovery, gene-expression prediction (Kelley et al., 2018) splicing site recognition (Ji et al., 2021), and promoter identification, as detailed in several comprehensive reviews (Min et al., 2017; Sapoval et al., 2022; Zhang et al., 2023). However, convolutional neural networks face challenges, like the need for extensive labeled sequence data. They are also task-specific, limiting their applicability to other scenarios outside their training focus. A significant bottleneck in integrating neural networks into bioinformatics has been the scarcity of adequate labeled data. Recent advancements in machine learning, inspired by breakthroughs in natural language processing, image analysis (Han et al., 2022), and protein structure prediction (Alipanahi et al., 2015; Jumper et al., 2021), have introduced new paradigms. Transformer-based architectures, especially large language models (Devlin et al., 2019; Brown et al., 2020a; Raffel et al., 2020), offer versatile representations—often termed "reusable" or "fundamental models." Among the recent training approaches is the fine-tuning paradigm, which divides the training process into two phases: pretraining and fine-tuning. Pretraining demands vast amounts of self-labeled data, while fine-tuning can, in some instances, operate with minimal, or even no examples.

In bioinformatics, there exists a paradoxical challenge. On one hand, there's an abundance of sequence data available, especially in public repositories like the SRA (sequence read arhive). The volume of this data is expanding exponentially, and as sequencing and other data-producing technologies become more affordable, this growth trend is likely to persist. These data repositories are akin to hidden treasures. Yet, they remain under-analyzed and underprocessed. Researchers often focus primarily on specific mutations, neglecting other valuable aspects of the data. Conversely, while there's an abundance of raw sequence data, there's a scarcity of labeled data. The accompanying metadata is frequently limited, and given the high cost of experiments, only a handful of samples, typically ranging from 3–15, are available within a specific group or stratum. It's also worth noting that labeling criteria can differ significantly across projects.

Recognizing these challenges, there is a compelling need for innovative methods that can harness the vast repositories of raw sequence data and navigate the complexity of labeling inconsistencies. It is in this context that our research contributes a novel solution. The development and application of our genomic language model family aims to address the mentioned issues, providing a robust, adaptable, and efficient tool for sequence classification.

While the concept of pretrained models isn't new, several have emerged recently, such as DNABERT (Ji et al., 2021; Zhou et al., 2023), Nucleotide Transformer (Dalla-Torre et al., 2023), and LookingGlass (Hoarfrost et al., 2022). However, a common limitation among these methods is their primary focus on human sequences or their restricted context size.

In the pretraining phase, the objective is to derive a general representation that captures the semantic relationships between objects, which in this context means obtaining a nuanced representation of sequence data. Typically, achieving this requires billions of samples, yet the volume of available sequence data far surpasses this number. We trained our genomic language models on an extensive corpus of available sequence data, encompassing bacteria, archaea, viruses, and fungi. Subsequently, we fine-tuned our models to tackle specific classification tasks, including the recognition of promoters and phages.

The ProkBERT family encompasses a series of models tailored to meet the intricate demands of microbial sequence classification, analysis, and visualization. The versatility of the ProkBERT models is manifested through their diverse applications:

1. Zero-shot learning: This approach allows for clustering of sequences by leveraging the embeddings directly produced by the model, eliminating the necessity for explicit fine-tuning.
2. Sequence classification: ProkBERT models can be seamlessly fine-tuned, whether for token-specific or comprehensive sequence-based classification tasks.

With these capabilities, the ProkBERT family aims to bridge the current gaps in the field, offering a robust toolset for diverse bioinformatics challenges.

2 Materials and methods

In this study, we used the transfer-learning paradigm for sequence classification based on transformer-based architectures.

The first phase involves pretraining on a large amount of sequence data, allowing the model to learn general sequence patterns. Once this foundation is established, we move to the fine-tuning phase where the model is adapted to specific tasks or datasets. The following sections provide a step-by-step description of our methods, from preparing raw sequence data to the specifics of both pretraining and fine-tuning. Figure 1 illustrates the training process.

In the development of the ProkBERT family, the initial step involves pretraining the model on a vast corpus of data. During this pretraining phase, the model aims to tackle the Masked Language Modeling task. In this task, specific portions of the sequence are masked, and the model's objective is to predict these masked sections, optimizing the likelihood of the missing parts using cross-entropy as the loss function. The model typically receives input in the form of a vectorized representation of the sequence. A notable constraint of standard transformers is their limited input size. Though various solutions have been suggested to address this limitation, the maximum token size is typically restricted up to 4kb, significantly smaller than the average bacterial genome, but much larger than an average gene.

Fine-tuning nucleotide sequences is a technique used to adapt pre-trained models to specialized tasks or specific datasets. The first step involves segmenting raw sequences into chunks, usually ranging from 0.1–1kb in size, to optimize the model's learning capability (Pan and Yang, 2009). Using weights from a pre-trained model, the system benefits from the knowledge obtained from comprehensive training on extensive datasets (Vaswani et al., 2017; Devlin et al., 2019). This initialization helps in quicker convergence and improved performance. After this initialization, the model undergoes training on the desired dataset, adjusting to its specific patterns and details. The outcome of this procedure allows the model to produce labeled sequences or tokens, which can be used for various annotation or prediction purposes (Brown et al., 2020b).

2.1 Sequence data

2.1.1 Sequence segmentation and tokenization

The first step is processing the sequence data. While there are many parallels between sequence data processing and natural language processing, drawing direct analogies can be challenging. For instance, determining what constitutes a 'sentence' in the realm of nucleotide and protein sequences doesn't have a direct counterpart in natural language. Additionally, the input size for neural networks is inherently limited. Figure 2 illustrates the strategy employed to vectorize the sequences.

Initially, the input sequence is segmented into smaller chunks. We employed two approaches for this:

1. Contiguous sampling, where contigs are divided into multiple non-overlapping segments; and
2. Random sampling, which involves fragmenting the input sequence into various segments at random.

Following segmentation, the next phase is encoding the sequence into a simpler vector format. The primary question revolves around defining the fundamental building block for a token. Various solutions have been suggested, the most widely strategy is applying one-hot-encoding (Sapoval et al., 2022), but DNA-BERT (Ji et al., 2021) applies the maximal overlapping k-mer strategy, meanwhile others relies on nucleotide level mapping (Dalla-Torre et al., 2023).

This phase is termed tokenization. We introduce a method termed *Local Context-Aware* tokenization (LCA), where individual elements consist of overlapping k-mers. Two principal parameters dominate this approach: k-mer size and shift. For $k = 1$, the tokenization resorts to a basic character-based approach, with a typical example illustrated in Figure 2. Employing overlapping k-mers can lead to enhanced classification performance. A greater shift value allows the model to use a broader context while reducing computational demands, while having the information-rich local context as well.

As an example for LCA tokenization, let's take the sequence {*AAGTCCAGGATCAAGATT*} and a k-mer size of 6, and shift=1 as LCA parameters [see Figure 2C (b)]. In that particular case the tokens will be the following: {*AAGTCC*, *AGTCCA*, *GTCCAG*, *TCCAGG*, ..., *AAGATT*}. The k-mers are then mapped into numerical ids, which will be the input for ProkBERT. As another example with $k = 6$ and shift=2, the tokenized segments will be the following: {*AAGTCC*, *GTCCAG*, *CCAGGA*, ..., *AAGATT*}. If the sequence length is odd, then the last charcter won't be used. One of the main advantages of the approach is that with the same number of tokens it is possible to cover a larger context, therefore it is possible to considerably reduce the computational and memory requirements, which is the typical bottleneck of the transformer architecture.

In this study, we propose models with a k-mer size of 6 (termed `ProkBERT-mini`), k-mer size of 1 (dubbed `ProkBERT-mini-c`), and a variant supporting a larger context window, named `ProkBERT-mini-long`, which relies on a k-mer size of 6 with a shift = 2.

2.1.2 Training data

The dataset was retrieved from the NCBI RefSeq database (O'Leary et al., 2016; Li et al., 2021) on January 6th, 2023. It included reference or representative genomes from bacteria, viruses, archaea, and fungi. After filtering, the sequence database consisted of 976,878 unique contigs derived from 17,178 assemblies. These assemblies represent 3,882 distinct genera, amounting to approximately 0.18 petabase pairs. The segment databases was created by sampling fixed lengths of [256, 512, 1024] or, in other instances, variable lengths aiming for an approximate coverage of 1.

Tokenization was performed using various k-mer sizes and shift parameters. The compiled database was then stored in the Hierarchical Data Format (HDF). Collectively, the training database held roughly 200 billion tokens for each segmented dataset.

For transparency and further research, all training data is available at zenodo 10.5281/zenodo.10057832.

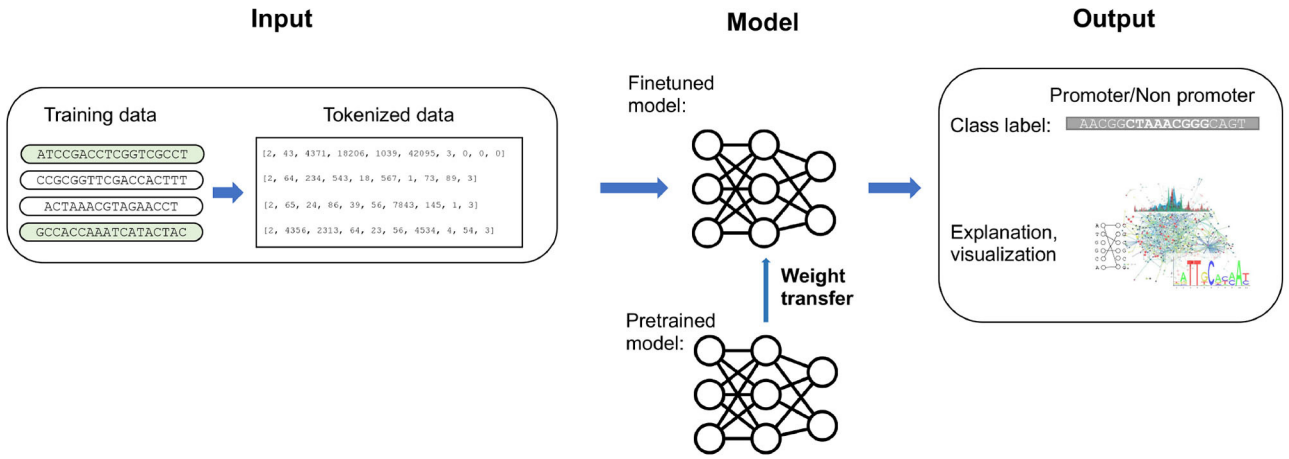

FIGURE 1
A schematic overview of the training process. Starting with raw sequence data, it undergoes preprocessing and vectorization. The model is then fine-tuned, beginning with weights from a pretrained model, to address the specific classification task. The output showcases classified sequences or tokens, predicted labels, scores, and a visualization highlighting underlying sequence patterns and explanations.

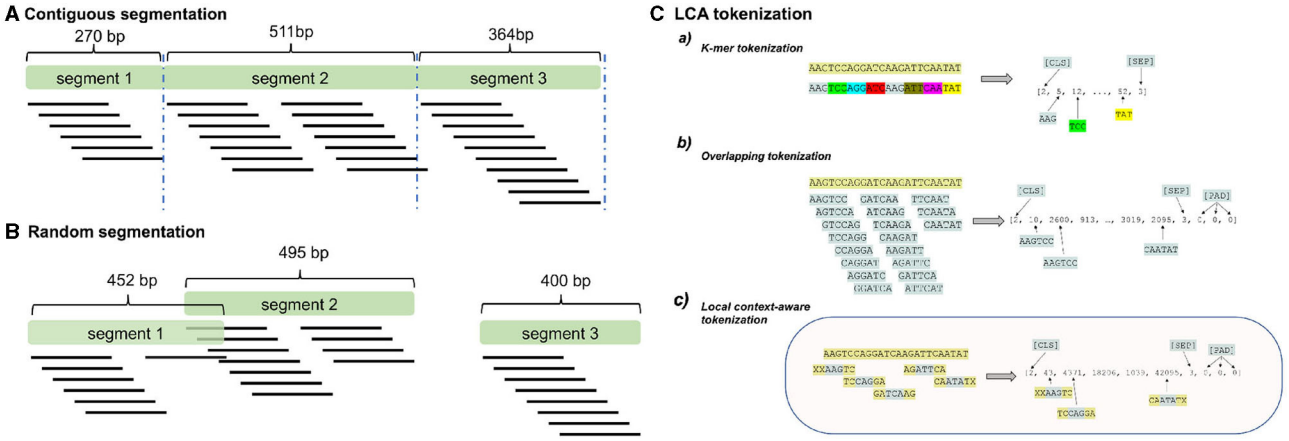

FIGURE 2
Preprocessing of sequences. The sequences are initially segmented into chunks ranging between 0–1 kb. Two segmentation strategies can be employed: **(A)** *Contiguous segmentation* where the sequence is split into non-overlapping parts; **(B)** *Random segmentation* where segments of varying lengths are randomly sampled from the original contig. The third part **(C)** outlines the tokenization process of the segments: (a) Splitting segments into non-overlapping tokens; (b) Creating maximally overlapping k-mers; (c) Generating partially overlapping k-mers by shifting with a fixed size.

2.2 Pretraining and learning sequence representations

2.2.1 Transformer model selection and parameters

In our study, we employed the MegatronBert model (Shoeybi et al., 2019), a variant of the BERT architecture (Devlin et al., 2019), optimized for large-scale training. The architecture overview is presented in Supplementary Figure S1. The key attributes of our models can be seen in Table 1. The `mini` and `mini-long` models share a common vocabulary of 4,101 k-mers. In contrast, the `mini-c` model is distinct, using a smaller set comprising only 9 items, including special tokens (i.e., [CLS], [SEP]) and nucleotides (A, C, T, G). All models employ a learnable relative key-value positional embedding, which maps input vectors into a 384-dimensional space. The `mini` and `mini-long` models support maximum sequence lengths of 1024 bp and 2048 bp, respectively. Across all models, the intermediate layers of the encoder use the GELU activation function, expanding the input dimensions to 3,072 before compressing them back to 384 dimensions. The Masked Language Modeling (MLM) head, a standard component in each model, decodes from 384 to 4,101 dimensions, adapted to the varying vocabulary sizes. To ensure efficient parallel computations, we encapsulated the entire architecture within a DataParallel wrapper, thus optimizing GPU utilization. For implementation, all models were developed using the PyTorch version 2.01 framework and the Hugging Face library version 4.33.2.

TABLE 1 A comprehensive overview of model parameters across varied configurations.

	Mini	Mini-c	Mini-long
Parameters	20,6 m	24,9 m	26,6 m
Tokenizer	6-mer, shift=1	1-mer	6-mer, shift=2
Layers	6	6	6
Attention heads	6	6	6
Max. context size (bp)	1027 nt	1022 nt	4096 nt
Training data	206,65 billion	206,65 billion	206,65 billion

2.2.2 Training process

2.2.2.1 Masked Language Modeling objective modifications

While Masked Language Modeling (MLM) acts as the primary pre-training objective for BERT models (Bidirectional Encoder Representations from Transformers) as established by Devlin et al. (2019), our implementation has slight variations. In the traditional BERT approach, a certain percentage of input tokens are randomly masked, and the model predicts these based on their context. Typically, about 15% of tokens undergo masking. However, due to our usage of overlapping k-mers, masking becomes more intricate. If a k-mer of size $k = 6$ is masked, we need to ensure at least six tokens are also masked to prevent trivial restoration from context and locality.

For an input sequence of tokens $\mathbf{x}$ and a binary mask vector $\mathbf{m}$—where 1 indicates a masked token and 0 indicates an unmasked token—the model outputs predicted vectors $\mathbf{y}$. As for the noise application on masked tokens, probabilities p_1, p_2, and p_3 define different noise strategies. In our model, when a token is masked, it is substituted with the special `[MASK]` token with a probability of p_1. Alternatively, with a probability p_2, it can be replaced with a random k-mer from our vocabulary. Lastly, there's a p_3 chance that the masked k-mer will remain as it is. Conventionally, these probabilities are set at 0.8, 0.1, and 0.1, respectively.

The MLM objective aims to minimize the negative log likelihood over all masked positions, as described by the equation:

$$\mathcal{L}_{\text{MLM}}(\mathbf{x}, \mathbf{m}, \mathbf{l}) = -\sum_{i\,:\,\mathbf{m}_i=1} \log \mathbf{y}_i[\mathbf{l}_i]$$

Where $\mathbf{y}_i[\mathbf{l}_i]$ denotes the predicted probability of the true label $\mathbf{l}_i$ for the masked position i. This objective, coupled with the noise injection strategy, ensures that the model learns bidirectional representations, thus becomes capable of understanding and generating contextually relevant tokens.

When dealing with overlapping k-mers, simple token masking becomes insufficient. If a single k-mer token is masked, all overlapping k-mers related to that token must also be masked. This is crucial because when a k-mer is not masked and subsequently restored, it might inadvertently provide contextual information about its neighbors. Such a situation would enable the trivial restoration of adjacent masked k-mers. In essence, one unmasked k-mer could potentially "leak" enough information to unmask its neighboring tokens. For examples, as presented in Figure 2C (*Overlapping tokeniization*), if only the second token "AGTCCA" is masked, it can be fully restored from its neighboring tokens: "AAGTCC" and "GTCCAG."

This overlapping nature of k-mers posed unique challenges. As a result, we had to dynamically adjust the MLM parameters and the lengths of sequence segments during the pretraining phase. Additionally, when multiple contiguous k-mers were masked together, the probability associated with the MLM had to be recalibrated. This was necessary to ensure that the actual proportion of the sequence being masked was consistent with our intended masking ratio.

2.2.2.2 Training phases and configuration

Initially, we employed parameters that allowed complete sequence restoration (k-mer of $k = 6$) by masking only five continuous tokens (with $p_1 = 0.9$) and manipulating 15% of the tokens. Once a loss threshold of 1 was attained, the MLM parameters were adjusted to heighten the masking complexity. We implemented various masking lengths, such as 2 nucleotides for k-mer of $k = 6$ and 2 characters for $k = 1$. Training data in the first phase had a fixed length of 128nt segments. The succeeding phase used variable-length datasets: with a probability of 0.5 a full-length segments, and with a probability of 0.5 a segment between 30–512 bp was selected into the the batch. The termination criterion for training was no further improvement or performance decrease, in both the MLM and promoter tasks. Models underwent training for roughly one batch each. We opted for batch sizes that spanned around 0.5–2 million bp sequences. Computations were executed on HPC-VEGA and Komondor platforms with Nvidia-A100 GPUs, leveraging slurm, pytorch distributed, and multiple GPU nodes.

2.2.3 Evaluating the pretrained model

We evaluted the masking performance of the models using the ESKAPE pathogens, namely *Enterococcus faecium* (GCF_009734005.1), *Staphylococcus aureus* (GCF_000013425.1), *Klebsiella pneumoniae* (GCF_000240185.1), *Acinetobacter baumannii* (GCF_008632635.1), *Pseudomonas aeruginosa PAO1* (GCF_000006765.1), and *Escherichia coli str. K-12* (GCF_000005845.2), because of their high clinical importance. First we investigated how the genomic structure is reflected in the embeddings, on different sequence features (i.e. CDS, intergenic, pseudo-genes, etc.). Next we measured how well the models can perform in masking.

2.2.4 Analysis of encoder outputs

In deep learning, an encoder typically processes input data (such as a sequence of tokens) and produces a dense vector representation for each token. These dense vectors, often referred to as embeddings or encoded vectors, capture the semantic information of the input tokens.

Given an input sequence S with T tokens, i.e.,

$$S = \{s_1, s_2, \ldots, s_T\}$$

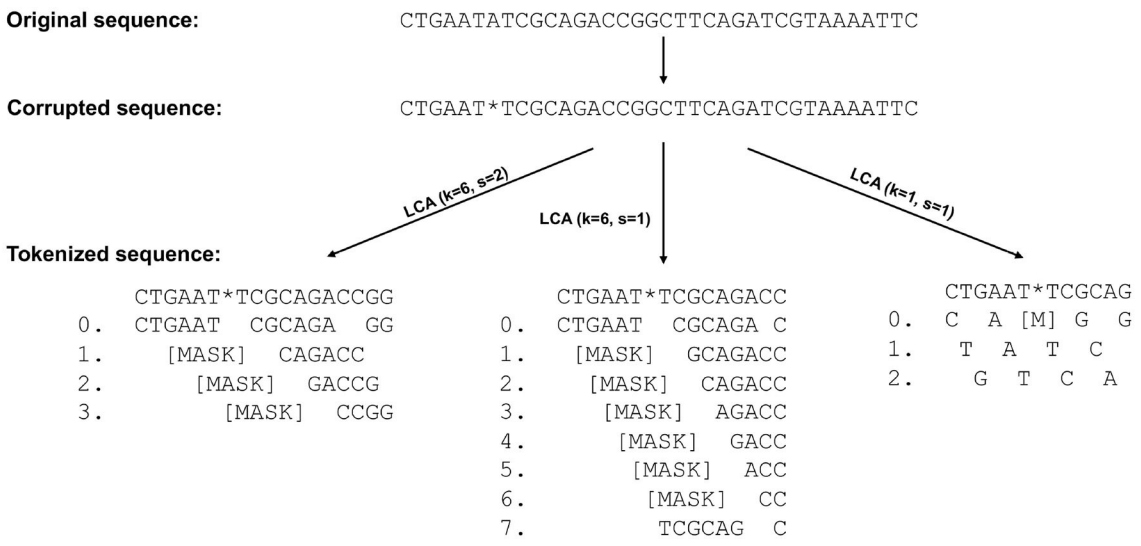

FIGURE 3
LCA Tokenization and corrupted sequence restoration. The figure illustrates how the corruption of a character at sequence level affects the initial vector representations of the seqment with respect to the different tokenization methods. The 7th nucleotide is unknown or masked, and those k-mers that overlap with that position are masked. As a result, when $k = 6$ and $s = 2$ not only the 7th character is hidden, but the 8th as well.

the encoder produces a sequence of vectors:

$$E = \{e_1, e_2, \ldots, e_T\}$$

where e_i represents the embedded vector for the token s_i. In case of multiple inputs or batches, if we have a batch of size B with each sequence containing T tokens, the encoder's output would be a 3D tensor of shape (B, T, D) where D is the dimensionality of the embeddings.

Once we have the encoded vectors, there are several ways to aggregate or pool them to get a single representation for the entire sequence as shown in Supplementary Figure S1. Here are some common pooling methods:

- **Mean Pooling:** Average the vectors: $e_{\text{mean}} = \frac{1}{T}\sum_{i=1}^{T} e_i$.
- **Sum Pooling:** Sum the vectors: $e_{\text{sum}} = \sum_{i=1}^{T} e_i$.
- **Max Pooling:** Max value per dimension: $e_{\max}[j] = \max_{i=1}^{T} e_i[j]$.
- **Min Pooling:** Min value per dimension: $e_{\min}[j] = \min_{i=1}^{T} e_i[j]$.

For batches, these pooling operations are applied independently for each input sequence in the batch. The provided NCBI annotations were preprocessed and extended. Intergenic regions were defined as non-annotated genomic features with respect to the strand. We retained the CDS, intergenic, pseudogenes, ncRNA features, while the rare or infrequently used features (such as riboswitch, binding_site, tmRNA, etc.) were excluded from the analysis. This was followed by sampling segments of various lengths from each genomic region. We sampled a maximum of 2000 sequence features from each contig, considering the strand, to evaluate strand-specific biases as well.

Then, we randomly corrupted a segment 10,000 times, i.e., a character was replaced with "*" and tokens containing "*" were mapped to the [MASK] token as illustrated on Figure 3.

The sampled segment database is available at Zenodo 10.5281/zenodo.10057832.

2.3 Application I: bacterial promoter prediction

The first task our models were evaluated on involved distinguishing between promoter and non-promoter sequences in bacteria. A sequence is labeled "1" if identified as a promoter and "0" otherwise. The next section gives an overview of the dataset structure and details about its constructions.

2.3.1 Dataset overview

The known promoters, referred to as positive samples, are primarily drawn from the Prokaryotic Promoter Database (PPD, Su et al., 2021), which contains experimentally validated promoter sequences from 75 organisms. Figure 4 illustrates the composition and source of our dataset, segregating prokaryotic promoters from non-promoters and including an independent test set based on *E.coli* sigma70 promoters.

2.3.1.1 Data partitioning and utilization

To ensure comprehensive evaluation, the dataset was split into three parts, divided randomly into training, validation, and testing datasets.

1. Training set: Constitutes 80% of the total data and is pivotal for initial model development and training.

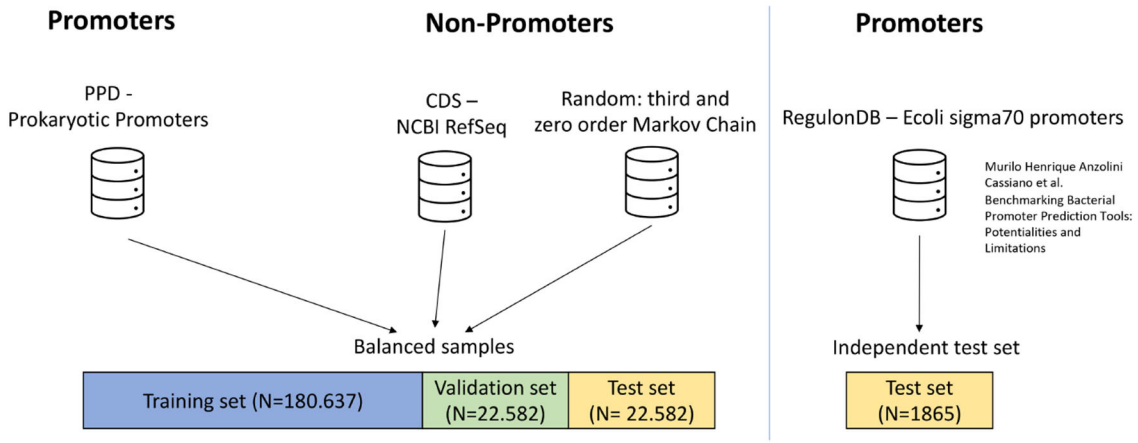

FIGURE 4
Schematic of the promoter dataset. This figure provides a visual representation of the sequence sources and their distribution within the study. The dataset comprises known promoter sequences from 75 organisms, retrieved from the Prokaryotic Promoter Database (PPD), alongside non-promoter sequences obtained from the NCBI RefSeq database (specifically sampled from CDS regions). It also includes non-promoter sequences constructed via higher and zero-order Markov chains that mirror compositional characteristics of known promoters. Additionally, an independent test set, focusing on *E. coli* sigma70 promoters, was employed, curated by Cassiano and Silva-Rocha (2020). A balanced distribution approach was adopted to even out the number of positive and negative samples, with the dataset being systematically divided into training, validation, and test subsets. This stratification underpins a thorough evaluation of the model efficacy.

2. Validation set: Comprises 10% of the data, aiding in fine-tuning model parameters and preventing overfitting.
3. Test set: Forms the remaining 10% of the data, crucial for unbiased model performance evaluation.

2.3.1.2 Dataset construction for multispecies train, test and validation sets

The prokaryotic promoter sequences are typically 81 bp long, ensuring compatibility with most tools' input prerequisites, particularly around the putative TSS region interval $[-60, +20]$. Our positive dataset encompasses promoter sequences from various species, predominantly found on both chromosomes and plasmids. Promoters included in the independent test set, based on exact match, were excluded from the training data. Species and contigs were mapped to NCBI assembly and sequence accessions. To curate comprehensive non-promoter sequences (negative samples), we employed three strategies:

1. Using non-promoter sequences (CDS–Coding Sequences).
2. Random sequences generated with a 3rd-order Markov chain.
3. Pure random sequences (0-order Markov chain) as proposed by Cassiano and Silva-Rocha (2020).

The distribution of this composite dataset was 40% CDS, 40% Markov-derived random sequences, and 20% pure random sequences (0-order Markov chain). One practical application of promoter detection in coding sequences is to check whether an unintentional promoter is injected or can be located inside a modified or designed coding sequence region, causing disruption. To cover this use-case, we incorporated the coding regions into our training and evaluation dataset. The CDS sequences were extracted from the genomic sequences of contigs, based on annotations from NCBI. The 81 bp long CDS region samples were selected based on the NCBI-provided annotations for the available contigs with respect to the underlying species. The promoter regions often contain AT-rich sequences, i.e., TATA box. To capture and model the AT-rich regions, we applied 3rd and 0 order Markov chains to generate sequence examples that reflect the compositional property of known promoters.

A 3rd-order Markov chain predicts the next nucleotide in a sequence based on the states of the previous three nucleotides. Formally, the probability of observing a nucleotide x_i given the nucleotides at positions x_{i-3}, x_{i-2}, and x_{i-1} is:

$$P(x_i|x_{i-3}, x_{i-2}, x_{i-1})$$

For DNA sequences, this yields $4^4 = 256$ possible nucleotide combinations. Such higher-order modeling can more effectively capture intricate sequence patterns and dependencies than lower-order models (Durbin et al., 1998). However, estimating transition probabilities requires extensive data due to the increased number of states (Koski and Noble, 2001). We determined these probabilities using promoter sequences, to which we added the reverse complement of each promoter. Subsequently, random promoter sequences were generated using these models.

We have a second, independent test for assessing model performance and referred to Cassiano and Silva-Rocha (2020)'s dataset comprising *E. coli* sigma70 sequences. The positive, well-recognized samples came from Regulon DB (Santos-Zavaleta et al., 2019). Cassiano and Silva-Rocha (2020) evaluated various tools using an experimentally validated *E. coli* K-12 promoter set dependent on sigma70, sourced from Regulon DB 10.5 (Santos-Zavaleta et al., 2019). Given the extensive documentation of sigma70-dependent promoters in bacteria, only these were considered. They used a positive dataset of 865 high-evidence sequences from Regulon DB and a negative set of 1,000 sequences mimicking the nucleotide distribution of the natural sequences. We ensured no overlap existed within the promoter datasets.

The promoter dataset is available as a Zenodo and Hugging Face dataset.

2.3.2 Training for promoter prediction

We employed a fine-tuning paradigm to evaluate our model. Our proposed binary classification model extends the Megatron BERT architecture (Shoeybi et al., 2019), tailored specifically for binary classification tasks. Let $\mathbf{X}$ represent the sequence of input embeddings, with $f_{\text{BERT}}(\mathbf{X})$ denoting the transformation by Megatron BERT. Given an input sequence of length T, this model transforms $\mathbf{X}$ into a sequence output $\mathbf{S}$ with dimensions $T \times$ hidden_size, where $\mathbf{S} = f_{\text{BERT}}(\mathbf{X})$. Unlike the conventional BERT model, which classifies sequences based on the special `[CLS]` token representing the "sentence," our approach emphasizes integrating representations of all tokens using a weighting scheme as shown in Supplementary Figure S1.

To obtain a fixed-size representation from the variable-length sequence $\mathbf{S}$, we devised a weighting mechanism. The sequence $\mathbf{S}$ undergoes a transformation through a linear layer to yield a sequence of weights $\mathbf{W}$:

$$\mathbf{W} = \text{softmax}(W_1\mathbf{S}^T + b_1)$$

Here, W_1 is a matrix sized hidden_size $\times$ 1 and b_1 is a bias term. The softmax operation ensures $\mathbf{W}$ forms a valid probability distribution over sequence positions. The model then computes a weighted sum of the sequence representations:

$$\mathbf{P} = \sum_{i=1}^{T} w_i s_i$$

Where w_i and s_i represent the weight and the sequence representation at the i^{th} position, respectively. Subsequently, $\mathbf{P}$ is processed by a dropout layer with a probability of hidden_dropout_prob to produce $\mathbf{P}'$. This results in the final classification logits $\mathbf{L}$.

Datasets, comprising training, validation, and testing subsets, were appropriately tokenized and adapted for ProkBERT processing. For optimization, the AdamW variant was chosen with parameters $\alpha \in \{0.0001, 0.0004, 0.0008\}$, $\beta_1 = 0.95$, $\beta_2 = 0.98$, and $\epsilon = 5 \times 10^{-5}$. A linear learning rate scheduler with warmup was utilized. The model underwent training for two epochs, with a batch size of 128 per GPU (NVIDIA A100-40GB GPUs) using the pytorch data distributed framework (nvcc). Additional configurations included a weight decay of 0.01.

2.4 Application II: phage sequence analysis

Bacteriophages have a significant role in the microbiome, influencing host dynamics and serving as essential agents for horizontal gene transfer (De la Cruz and Davies, 2000). Through this mechanism, they aid in the transfer of antibiotic resistance and virulence genes, promoting evolutionary processes. Understanding the diversity of phages is crucial for tackling challenges like climate change and diseases (Jansson and Wu, 2023). These phages exhibit distinct patterns in both healthy and diseased microbiomes (Yang et al., 2023). The correlation between the human virome and various health conditions, such as cancer, inflammatory bowel diseases, and diabetes, has been documented (Zhao et al., 2017; Han et al., 2018; Nakatsu et al., 2018; Fernandes et al., 2019; Liang et al., 2020; Zuo et al., 2022). However, deeper research is needed to discern causality and their impact on microbial and host biological processes.

Despite the abundance of phages (Bai et al., 2022a), accurately quantifying and characterizing them remains a challenge. One primary limitation is the restricted number of viral sequences in databases like NCBI RefSeq. Additionally, the categorization of viral taxonomy is still a topic of discussion (Walker et al., 2022). Though there have been recent efforts to expand databases (Zhang et al., 2022; Camargo et al., 2023), the overall understanding of viral diversity is still not complete (Yan et al., 2023). We have assembled a unique phage sequence database using recently published genomic data.

Another challenge is the life cycle of phages; temperate phages might integrate their genomes into bacterial chromosomes and are often annotated as bacterial genomes, leading to potential misidentification. Current databases also show biases toward certain genera (Schackart III et al., 2023), which can skew benchmarking and the evaluation of different methods. To address this, we used a balanced benchmarking approach, ensuring each viral group corresponds to their predicted host genus, minimizing bias. We also compared viral genomes to their respective hosts, a more demanding classification task, such as distinguishing a *Salmonella* phage from its host genome compared to marine *cyanobacteria*. For our study, we selected a specific number of phages for testing, ensuring there is no overlap between training and testing sets at the species level.

2.4.1 Phage dataset description

To train and assess our prediction models, we assembled a comprehensive phage sequence database from diverse sources. As of 9th July, 2023, we procured viral sequences and annotations from the RefSeq database (O'Leary et al., 2016; Li et al., 2021). By isolating entries labeled "phage," we obtained 6,075 contigs. Our database was further enriched with the inclusion of Ren et al. (2020), a dataset validated through the TemPhD method (Zhang et al., 2022), adding another 192,326 phage contigs extracted from 148,229 assemblies.

To address sequence redundancy present in both the RefSeq and TemPhD databases, we applied the CD-HIT algorithm (Li and Godzik, 2006; Fu et al., 2012) (using CD-HIT-EST with a default word size of 5). While several clustering thresholds (0.99, 0.95, 0.90) were experimented with and found to produce similar outcomes, we settled on a threshold of 0.99. This process resulted in a refined set of 40,512 distinct phage sequences, with an average length of approximately 43,356 base pairs, culminating in a total of 3.5 billion base pairs. Notably, these sequences target a wide spectrum of 660 bacterial genera. Subsequent to sequence curation, phage sequences were mapped to their respective bacterial hosts to formulate a balanced training dataset, ensuring equitable representation between phages and their hosts. This step is imperative, given the distinct distributions observed between bacterial sequences

and their phage counterparts. In numerous instances, due to ambiguities in species-level identification or gaps in taxonomic data, host mapping was executed at broader taxonomic strata, predominantly at the genus level.

In our examination of bacteriophage-host associations at the genus level, several bacterial genera stood out, showcasing pronounced phage interactions. *Salmonella*, a main cause of food-related sicknesses (Popoff et al., 2004), stands out with an impressive association of 24,182 phages, spanning a cumulative length of over a billion base pairs (1,026,930,954 bp) and an average phage length of 42,467 bp. Following closely, the common gut bacterium, *Escherichia* (Tenaillon et al., 2012), is linked with 8,820 phages, accumulating a total length of 408,866,394 bp. The genus Klebsiella, notorious for its role in various infections (Paczosa and Mecsas, 2016), associates with 4,904 phages. Genera such as *Listeria* (Vázquez-Boland et al., 2011), *Staphylococcus* (Lowy, 1998), and *Pseudomonas* (Driscoll et al., 2007), each with distinct clinical significance, exhibit rich phage interactions. Notably, *Mycobacterium* (Cole et al., 1998), consisting of pathogens like the tuberculosis-causing bacterium, shows associations with 2,156 phages. Many of these bacterial genera are benign and even beneficial under normal conditions, they also include species that can cause severe diseases in humans, especially when there's an imbalance in the body's natural flora or when antibiotic resistance develops. Monitoring phage interactions with these bacteria offers potential pathways for therapeutic interventions and a deeper understanding of microbial ecology in human health.

Additionally, balanced databases were created, stratified by the host genus level, to mitigate the effect of underrepresented or overrepresented phages, such as Salmonella. The reverse-complement sequences were included. The final dataset encompasses a total of 660 unique bacterial genera. Undersampling was performed with a threshold of 20,027,298 bp for 25 genera, while the others were upsampled with a maximum coverage of 5x, obtaining random samples of shorter fragments from the contigs. Random segmentation and sampling were carried out as previously described. The bacterial assemblies were randomly selected from the NCBI database, prioritizing higher-quality assemblies. Many of them were not included in the pretraining dataset. Subsequently, we constructed a database with various sequence lengths: 256, 512, 1024, and 2048 bps. The train-test-validation split was executed in a 0.8, 0.1, and 0.1 proportion at the phage sequence level.

For comparison with alternative methods and tools, we had to subsample our test set ($N = 10,000$) to conduct the evaluation within a reasonable timeframe.

2.4.2 Model training for phage sequence analysis

The task was formulated as binary classification, similarly to the promoters. Phage sequence classification was approached in a manner analogous to the promoter training. Given the extensive size of the dataset, preprocessing was conducted beforehand, segmenting sequences into various lengths: 256, 512, 1,024, and 2,048 bps. For both `mini` and `mini-c` models, the training process was partitioned into three distinct phases. An initial grid search was executed to optimize learning rates, and base models were trained for an hour. The parameter yielding the highest Matthews Correlation Coefficient (MCC) was selected. The model was then trained using segment lengths of 256 bps for half an epoch, followed by 512 bps for another half epoch, and concluding with two epochs for 1024 bps segments. The training regimen for the `mini-long` model was similar, albeit commencing with 512 bps segments, then transitioning to 1024 bps, and finally to 2048 bps segments. Model optimization employed the settings delineated previously.

2.5 Applied metrics

MCC (Matthews Correlation Coefficient): Used for binary classifications and defined as:

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

where TP is true positives, TN is true negatives, FP is false positives, and FN is false negatives. The coefficient ranges from -1 (total disagreement) to 1 (perfect agreement).

F1 Score: The harmonic mean of precision and recall, given by:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

with

$$\text{Precision} = \frac{TP}{TP + FP}$$

and

$$\text{Recall (Sensitivity)} = \frac{TP}{TP + FN}$$

Accuracy: Represents the proportion of correctly predicted instances to the total, defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Sensitivity (Recall): The proportion of actual positives correctly identified:

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

Specificity: The proportion of actual negatives correctly identified:

$$\text{Specificity} = \frac{TN}{TN + FP}$$

ROC-AUC (Receiver Operating Characteristic - Area Under Curve): Evaluates the model's discriminative ability between positive and negative classes. It's the area under the ROC curve, which plots Sensitivity against 1 − Specificity for various thresholds.

The silhouette score is a measure used to calculate the goodness of a clustering algorithm. It indicates how close each sample in one cluster is to the samples in the neighboring clusters, with values ranging from –1 to 1, where a high value indicates that the sample is well matched to its own cluster and poorly matched to neighboring clusters (Rousseeuw, 1987).

Equation for the silhouette score $s(i)$ for a single sample:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Where:

- $a(i)$ is the average distance from the i-th sample to the other samples in the same cluster.
- $b(i)$ is the smallest average distance from the i-th sample to samples in a different cluster, minimized over clusters.

3 Results and discussion

3.1 ProkBERT's learned representations capture genomic structure and phylogeny

We assessed the zero-shot capabilities of our models by examining their proficiency in predicting genomic features based solely on embedding vectors, in a manner akin to Nucleotide Transformers and related methodologies. Figure 5 presents the UMAP projection of these embedded vector representations. Employing the UMAP technique, we reduced the dimensionality of genomic segments and derived embeddings. These were then evaluated using silhouette scores across the three models: `ProkBERT-mini`, `ProkBERT-mini-c`, and `ProkBERT-mini-long`.

Our primary objective was to discern if the representations of sequence segments from ESKAPE pathogens could be distinctly categorized. Indeed, Figure 5 exhibits clear delineation among known genomic features, including CDS (coding sequences), intergenic regions, ncRNA, and pseudogenes. It's important to note that these models were not explicitly trained to differentiate these sequence features; the representations were solely derived through pretraining. For the critical genomic comparison between "intergenic" and "CDS" regions, the silhouette scores obtained were 0.4925, 0.5766, and 0.3352 across the respective models, emphasizing a consistent and clear distinction between these features. Regarding non-coding RNA representations, the silhouette scores for "ncRNA" vs. "CDS" were 0.1537, 0.2935, and 0.2192, while for "ncRNA" vs. "intergenic," they were 0.1648, 0.1302, and 0.3109, further affirming the assertion that ncRNAs cluster distinctly. Pseudogenes, as anticipated, exhibited some overlap with 'CDS', notably in the `ProkBERT-mini` model with a score of −0.0358. Yet, when compared with 'ncRNA', a distinct separation was observed, as evidenced by scores of 0.1630, 0.2365, and 0.1636.

This analysis aligns with biological knowledge, where pseudogenes are expected to be more similar to CDS, while ncRNAs, which have different functions and characteristics, form distinct clusters from CDS and intergenic regions. All three models appear to produce similar clustering results for the given pairs of genomic features.

The embeddings prominently display the genomic intricacies of ESKAPE pathogens. Notably, *Klebsiella pneumoniae* and *Escherichia coli*, both members of the *Enterobacteriaceae* family, exhibit close proximity in the embedding space, echoing potential genomic kinship or shared evolutionary paths. This observation is further corroborated by the low silhouette scores across the models. In contrast, species like *Pseudomonas aeruginosa* manifest as more distinct clusters, emphasizing their genetic disparities. Intriguing overlaps, such as those between differently labeled *Acinetobacter baumannii* entities, highlight potential challenges in the data or shared genomic features. Combined, the UMAP visualizations and silhouette scores provide a profound insight into species-specific genomic embeddings, revealing both shared and distinct genomic signatures.

3.2 ProkBERT can efficiently recover corrupted sequences

In evaluating the models' capabilities in the masking task, we used random masking across various genomic segments, such as CDS, ncRNA, intergenic, and pseudogenes, detailed in Table 2. We measured performance with metrics like ROC-AUC and average reference rank. However, a direct model comparison presents challenges. Notably, `ProkBERT-mini-c` boasts a significantly smaller vocabulary size (9) in comparison to `ProkBERT-mini` and `ProkBERT-mini-long` (4101) This allows `ProkBERT-mini-c` to achieve higher rankings, like top3, with relative ease as it encompasses nearly the entire vocabulary (there are 4 nucleotides). Also, the local context's representation in `ProkBERT-mini-long` is less dense, making the restoration of the masked nucleotides harder in contrast to the others.

For sequences spanning 1,024 nucleotides, `ProkBERT-mini` exhibited a commendable AUC of 0.9998, accompanied by top 1 and top 3 prediction accuracies of 51.69% and 92.27%, respectively. Concurrently, `ProkBERT-mini-c` achieved an AUC of 0.9586, with top 1 and top 3 accuracies at 51.28% and 92.22%. However, `ProkBERT-mini-long` reported slightly subdued figures, with an AUC of 0.9992 and top 1 and top 3 accuracies of 27.68% and 55.89%. This underscores the efficacy of the ProkBERT model family in handling genomic tasks. A salient observation from our analysis is that a model's prediction proficiency is intrinsically tied to the contextual size.

In our next assessment some performance nuances became evident across various genomic regions. The `prokbert-mini` model consistently stood out, especially within the Coding Sequence (CDS) and Intergenic domains. For these regions, it achieved an unmatched ROC-AUC of 0.9998. Specifically, within the CDS region, the model attained a Top1 accuracy of 50.33%, a Top3 accuracy of 91.87%, and an average reference rank of 0.811. In the Intergenic sections, these figures were 48.97%, 91.12%, and 0.843, respectively. The `prokbert-mini-c` model also exhibited commendable performance. Within the CDS regions, this model reached a Top1 accuracy of 50.65%, a Top3 accuracy of 91.91%, and an average reference rank of 0.802. For the Intergenic regions, the metrics were 48.84%, 91.39%, and 0.839 respectively. Despite the achievements of the aforementioned models, challenges persisted across all models in the non-coding RNA (ncRNA) domains. Even the top-performing `prokbert-mini` saw its Top1 accuracy drop to 32.46%, with an average reference rank increasing to 1.202. Contrastingly, the `prokbert-mini-long`, despite its detailed design, exhibited reduced accuracies, with

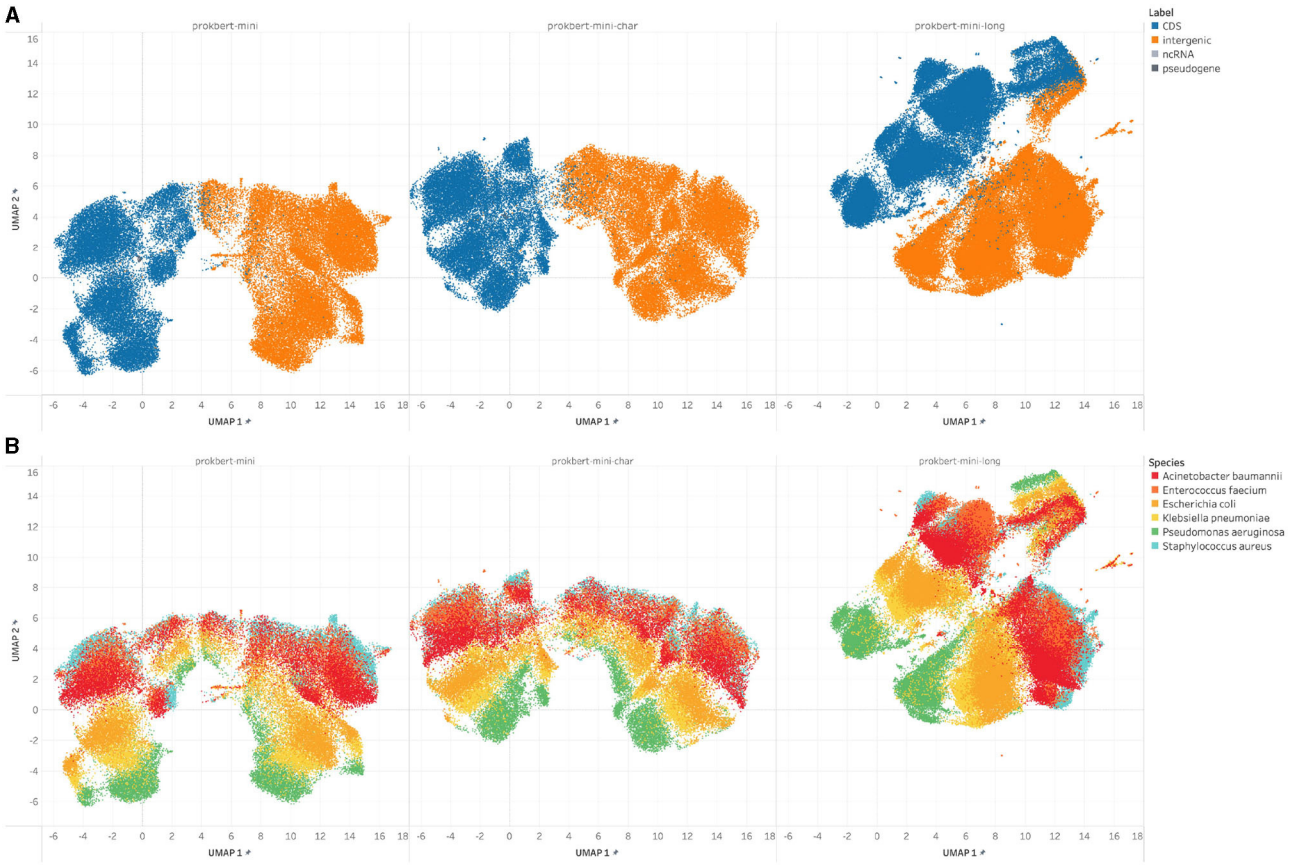

FIGURE 5
UMAP embeddings of genomic segment representations. The figure presents the two-dimensional UMAP projections of embedded vector representations for various genomic features, derived from the `ProkBERT-mini`, `ProkBERT-mini-c`, and `ProkBERT-mini-long` models. The distinct clusters highlight the models' ability to differentiate between features such as CDS (coding sequences), intergenic regions, ncRNA, and pseudogenes, even without explicit training for feature differentiation. **(B)** The segments are colored according to species, indicating that cluster structure reflects the phylogenic similarities. **(A)** Sequence embeddings of the different regions. **(B)** Sequence embeddings of the different species of ESKAPE pathogens.

TABLE 2 **Masking performance of the ProkBERT family.**

Model	L	Avg. Ref. Rank	Avg. Top1	Avg. Top3	Avg. AUC
`ProkBERT-mini`	128	**0.9315**	**0.4497**	0.8960	**0.9998**
`ProkBERT-mini-c`	128	0.9429	0.4391	**0.8965**	0.9504
`ProkBERT-mini-long`	128	3.9432	0.2164	0.4781	0.9991
`ProkBERT-mini`	256	0.8433	0.4848	0.9130	**0.9998**
`ProkBERT-mini-c`	256	**0.8262**	**0.4928**	**0.9151**	0.9565
`ProkBERT-mini-long`	256	3.5072	0.2470	0.5258	0.9992
`ProkBERT-mini`	512	0.8098	0.5056	0.9179	**0.9998**
`ProkBERT-mini-c`	512	**0.7983**	**0.5116**	**0.9203**	0.9580
`ProkBERT-mini-long`	512	3.3026	0.2669	0.5435	0.9992
`ProkBERT-mini`	1024	**0.7825**	**0.5169**	**0.9227**	**0.9998**
`ProkBERT-mini-c`	1024	0.7868	0.5128	0.9222	0.9586
`ProkBERT-mini-long`	1024	3.2082	0.2768	0.5589	0.9992

Bold numbers indicate the best results per category.

TABLE 3 Evaluation of promoter prediction tools on *E-coli* sigma70 dataset (Transposed).

Tool	Accuracy	MCC	Sensitivity	Specificity
`ProkBERT-mini`	**0.87**	**0.74**	0.90	0.85
`ProkBERT-mini-c`	**0.87**	0.73	0.88	0.85
`ProkBERT-mini-long`	**0.87**	**0.74**	0.89	0.85
CNNProm	0.72	0.50	0.95	0.51
iPro70-FMWin	0.76	0.53	0.84	0.69
70ProPred	0.74	0.51	0.90	0.60
iPromoter-2L	0.64	0.37	0.94	0.37
Multiply	0.50	0.05	0.81	0.23
bTSSfinder	0.46	-0.07	0.48	0.45
BPROM	0.56	0.10	0.20	0.87
IBPP	0.50	-0.03	0.26	0.71
Promotech	0.71	0.43	0.49	**0.90**
Sigma70Pred	0.66	0.42	0.95	0.41
iPromoter-BnCNN	0.55	0.27	**0.99**	0.18
MULTiPly	0.54	0.19	0.92	0.22

Bold numbers indicate the best results per category.

Top1 and Top3 accuracies of 25.18% and 52.66% across all labels, hinting at potential inefficiencies or overfitting. Collectively, these findings underscore the importance of tailored model architectures for genomic sequences and highlight the complexities of various genomic regions, laying a foundation for future targeted deep learning strategies in genomics.

3.3 ProkBERT performs accurately and robustly in promoter sequence recognition

Identifying promoters, which are crucial in initiating the transcription process, is fundamental to understanding gene regulation in bacteria. Our initial fine-tuning task focused on the identification of these genomic regions, primarily through a binary classification approach that distinguishes sequences as either promoters or non-promoters. Although this method is widely used, various alternative strategies have been explored. A significant limitation of current techniques, as highlighted by Chevez-Guardado and Peña-Castillo (2021), is their reliance on training with a limited range of species, mainly *E. coli*, but also including *Bacillus subtilis* and a few other key species.

As illustrated in Figure 1, our training began with a pretrained model followed by training using cross-entropy loss minimization. We evaluated the training outcomes on two datasets: a test set curated by Cassiano and Silva-Rocha (2020), and another one comprising mixed species. The models' performance on the first dataset can be seen in Table 3.

Cassiano and Silva-Rocha (2020) had previously gauged the efficacy of several well-established tools, including BPROM (Salamov and Solovyevand, 2011), bTSSfinder (Shahmuradov et al., 2017), BacPP (de Avila e Silva et al., 2011), CNNProm (Umarov and Solovyev, 2017), IBBP (Wang et al., 2018), Virtual Footprint, iPro70-FMWin (Rahman et al., 2019), 70ProPred (He et al., 2018), iPromoter-2L (Liu et al., 2018), and MULTiPly (Zhang et al., 2019). Additionally, we incorporated newer tools like Promotech (Chevez-Guardado and Peña-Castillo, 2021) and iPromoter-BnCNN (Amin et al., 2020). These tools encompass a broad spectrum of techniques. For instance, BPROM and bTSSfinder exploit conserved and promoter element motifs. BacPP and CNNProm use neural networks for promoter predictions in *E. coli* and other bacteria based on transformed nucleotide sequences. IBBP adopts a unique image-based approach combined with logistic regression and various sequence-based features. Tools like 70ProPred, iPro70-FMWin, MULTiPly, and iPromoter-2L leverage SVM, logistic regression, and random forest methodologies, drawing upon extracted sequence features such as physicochemical properties and k-mer compositions.

The results are presented in Table 3. The ProkBERT family models exhibit remarkably consistent performance across the metrics assessed. With respect to accuracy, all three tools achieve an impressive score of 0.87, marking them among the top performers in promoter prediction. This suggests that, regardless of the specific version, the underlying methodology used in the `mini` series is robust and effective.

When evaluating the balance between true and false predictions using MCC both `ProkBERT-mini` and `ProkBERT-mini-long` slightly edge out `ProkBERT-mini-c` with an MCC of 0.74 compared to 0.73 for `mini-c`. Although the difference is marginal, it might indicate subtle refinements in the `mini-long` approach. In terms of sensitivity, which focuses on the ability to correctly identify promoters, `ProkBERT-mini` leads with a score of 0.90, closely followed by `ProkBERT-mini-long` at 0.89 and `ProkBERT-mini-c` at 0.88. This hierarchy, albeit with small differences, highlights the minute improvements

achieved in the mini and mini-long versions. Lastly, for specificity, all three versions achieve an identical score of 0.85. This uniformity underscores the consistency in their ability to correctly identify non-promoters. In summary, while the performance across the mini versions is largely comparable, ProkBERT-mini and ProkBERT-mini-long display marginal advantages in certain metrics, hinting at potential refinements in these versions.

The Promotech tool demonstrates a mixed performance across the metrics. With an accuracy of 0.71, it correctly predicts the presence or absence of promoters 71% of the time. While this accuracy is lower than the top-performing tools like ProkBERT-mini and its variants, it is significantly better than the lower-performing tools such as Multiply and bTSSfinder. Sensitivity for Promotech is 0.49, suggesting that it correctly identifies nearly half of the actual promoters. However, its most remarkable performance metric is its specificity, with a score of 0.90. This means Promotech is adept at identifying non-promoters, correctly classifying them 90% of the time.

Among the methods assessed, CNNProm, Sigma70Pred, iPromoter-BnCNN, and iPromoter-2L exhibit notably high sensitivity scores, signifying their pronounced ability to correctly identify promoters. Specifically, iPromoter-BnCNN leads with an exceptional sensitivity of 0.99, closely trailed by Sigma70Pred at 0.95, CNNProm at 0.95, and iPromoter-2L at 0.94. Such high sensitivity scores indicate these models' potential in minimizing false negatives, which is crucial in applications where missing an actual promoter can have significant implications. However, it's vital to interpret these results with caution. The high sensitivity scores, especially of iPromoter-BnCNN and Sigma70Pred, come at the expense of specificity. For instance, iPromoter-BnCNN has a notably low specificity of 0.18, implying a substantial rate of false positives. Similarly, Sigma70Pred has a specificity of 0.41. This suggests that while these models are adept at identifying promoters, they often misclassify non-promoters as promoters. An essential factor to consider in this evaluation is the training data. Given that these models were trained on *E. coli* data, their performance might be biased when evaluated on the same or closely related datasets. This lack of independence between training and testing data can lead to overly optimistic performance metrics, as the models might merely be recalling patterns they've already seen, rather than generalizing to novel, unseen data.

Next, we evaluated our models' performance on a test set encompassing a broad mix of promoters, extending beyond just *E. coli*. The results are shown in Figure 6.[1]

The trio of tools in the ProkBERT family – mini, mini-c, and mini-long – consistently exhibited strong performance across the metrics analyzed. In terms of accuracy, all three achieved scores between 0.79 and 0.81, solidifying their position among leading promoter prediction tools. This uniformity in results points to a reliable methodology underlying the ProkBERT family. Using the Matthews Correlation Coefficient (MCC) as a measure of prediction balance, ProkBERT-mini and ProkBERT-mini-long both slightly outperformed ProkBERT-mini-c with MCC values of 0.63 and 0.62 respectively, against the 0.57 of mini-c. Considering sensitivity, ProkBERT-mini achieved the highest score of 0.81, with ProkBERT-mini-long and ProkBERT-mini-c trailing at 0.79 and 0.75, respectively. This order reiterates the nuanced enhancements in the models. With regard to specificity, ProkBERT-mini-long stood out with a score of 0.83, whereas ProkBERT-mini and ProkBERT-mini-c both scored 0.82, reflecting their adeptness at accurate non-promoter classification.

Of the tools assessed, both Sigma70Pred and iPromoter-BnCNN show moderate performance in sensitivity, with iPromoter-BnCNN taking the lead at 0.66 and Sigma70Pred following at 0.52. Promotech displayed a varied metric performance. With an accuracy rate of 61%, it identifies promoters correctly in a majority of instances. Its sensitivity value of 0.29 signifies its capability to detect roughly one-third of true promoters. Yet, its high specificity of 0.93 reveals its proficiency at negating non-promoters.

Promoter prediction is an intricate task that requires a balance between sensitivity and specificity. The consistently strong performance of the ProkBERT family highlights their reliability in this domain. Yet, the selection of a tool should be made after weighing the potential implications of both false positives and negatives.

3.4 ProkBERT swiftly and accurately identifies phage sequences, even in challenging settings

Various tools have addressed phage sequence identification, each employing distinct strategies. These methods can be categorized into: (i) homology or marker-based tools like VirSorter2 (Guo et al., 2021) and VIBRANT (Kieft et al., 2020), (ii) alignment-free methods, for instance, DeepVirFinder (Ren et al., 2020) and INHERIT (Bai et al., 2022b). The first category leans on existing annotations, databases, and sequences. In contrast, alignment-free methods are less influenced by existing knowledge, offering broader applicability and greater reliability with imperfect sequence data (Wu et al., 2023). We assessed our classification accuracy against INHERIT, VirSorter2, and DeepVirFinder (Ren et al., 2020). Notably, INHERIT employs a DNABert architecture for classification, akin to ours, drawing inspiration from DNABert (Ji et al., 2021).

In genomic studies, discerning phage-related segments becomes increasingly challenging as the segment length diminishes (Guo et al., 2021). This study rigorously evaluates six distinct phage classification methodologies over a range of sequence lengths, leveraging the accuracy and MCC as primary performance metrics.

For the shortest fragments (256bp), VirSorter was unable to process the test set. Among the evaluated methods, the ProkBERT models—mini, mini-c, mini-long—consistently emerged as top performers across varying lengths, as depicted in Figure 7. Specifically, ProkBERT-mini excels with shorter sequences, achieving the highest accuracy for 256 bp fragments. This high accuracy does not come at the cost

1 The selection of competitors for the second test set took into account the larger size of the dataset, which posed practical challenges for established methods optimized for smaller sequences, resulting in processing issues and longer evaluation times.

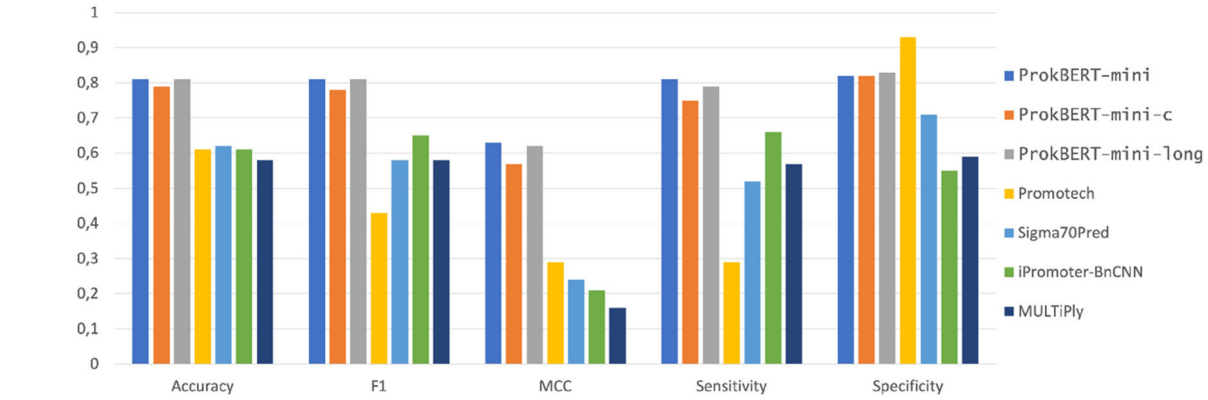

FIGURE 6
Promoter prediction performance metrics on a diverse test set. A comparative analysis of various promoter prediction tools, showcasing their performance across key metrics including accuracy, F1 score, MCC, sensitivity, and specificity. The tools evaluated include `ProkBERT-mini`, `ProkBERT-mini-c`, `ProkBERT-mini-long`, Promotech, Sigma70Pred, iPromoter-BnCNN, and MULTiPly.

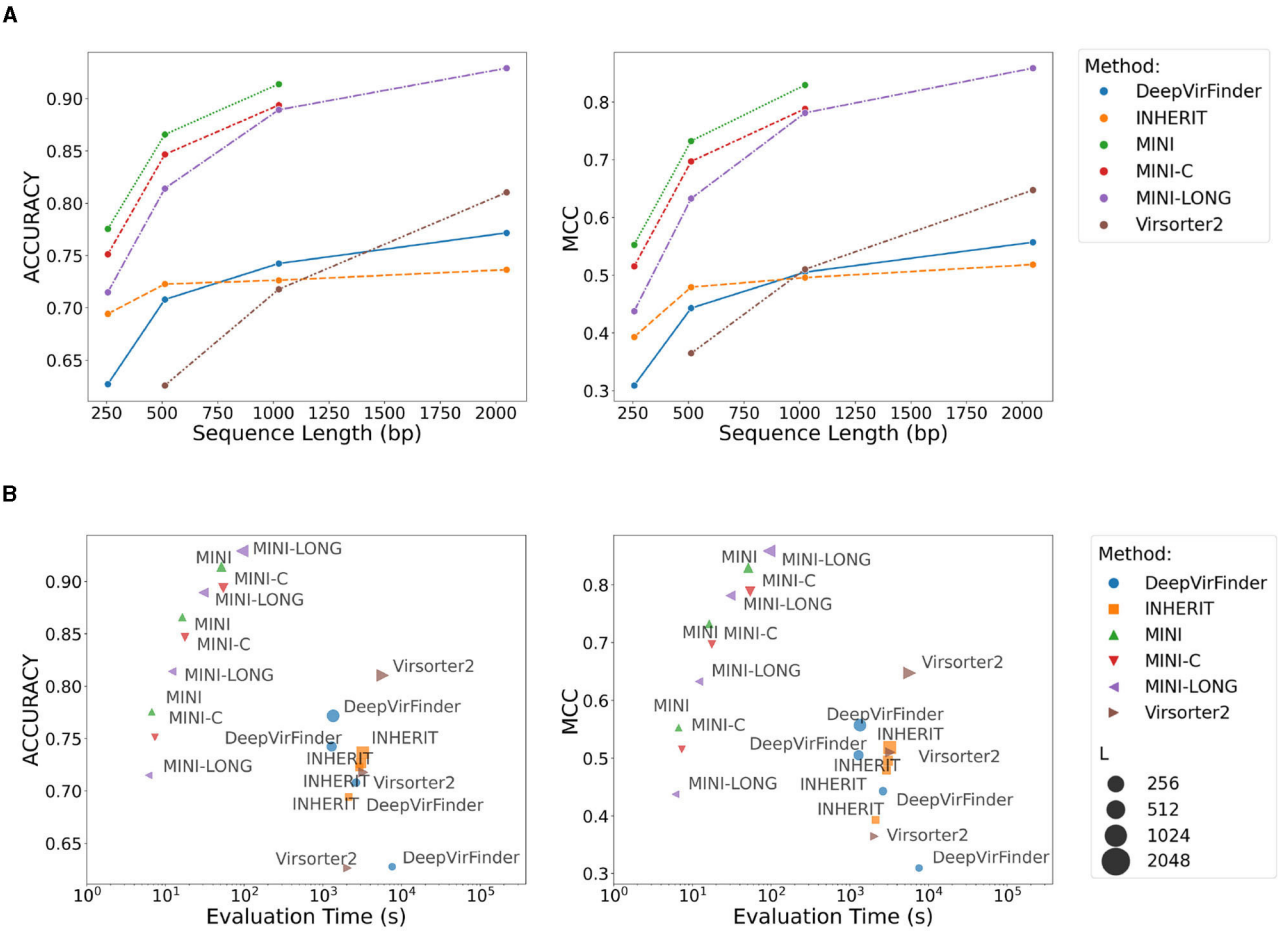

FIGURE 7
ProkBERT identifies phage sequences accurately and rapidly. **(A)** Method comparison over varying sequence lengths based on two essential performance metrics: accuracy and MCC. **(B)** Scatter plots illustrating the relationship between evaluation time (on a logarithmic scale) and the mentioned performance metrics. The size of each point signifies the sequence length. Evaluation time encompasses model loading, sequence preprocessing, and inference phases. **(A)** Comparison of methods across different sequence lengths. **(B)** Comparison of evaluation time and performance metrics.

of increased false positives or negatives, as evidenced by its comparable MCC values. In contrast, DeepVirFinder, ranking fifth, indicates potential optimization areas for such short sequences. While `ProkBERT-mini` consistently ranks highest for lengths up to 1,024 bps, `ProkBERT-mini-c` closely follows, signifying its stability and reliability. Notably, the maximum sequence length that `ProkBERT-mini` and `ProkBERT-mini-c` can process is limited to 1024bps, introducing the specialized `ProkBERT-mini-long` for extended sequences. This model showcases its prowess with 2kb sequences, achieving an accuracy of 92.90% and an MCC of 0.859. Virsorter2, despite initial struggles with shorter sequences, exhibits significant improvements for longer fragments. However, both DeepVirFinder and INHERIT show limited enhancements with increased sequence lengths, suggesting these methods might not capitalize on the additional information longer sequences provide as effectively as their counterparts. In conclusion, `ProkBERT-mini` and `ProkBERT-mini-long` clearly stand out as top-performing models across various sequence lengths. While other methods may have their merits, they simply don't match the consistency and robustness offered by the ProkBERT models.

In phage classification, sensitivity signifies the proportion of actual phage sequences that are correctly identified. Conversely, specificity represents the proportion of non-phage sequences accurately discerned. A method exhibiting high sensitivity effectively identifies most phage sequences, while high specificity indicates minimal misclassification of non-phage sequences as phage-related. Supplementary Figure S2 presents the comparative results of the models in terms of specificity and sensitivity. Interestingly, longer sequences tend to decrease the specificity for VirSorter2. This trend suggests that VirSorter2 might misclassify non-phage sequences more frequently as the sequence length increases. A concurrent analysis of sensitivity and specificity reveals nuances in method performance. For example, `ProkBERT-mini` consistently achieves top ranks in sensitivity but displays variable results in specificity. On the other hand, Virsorter2, despite its strong specificity, especially with extended sequences, requires enhancements in its sensitivity. Notably, several methods, including DeepVirFinder, `ProkBERT-mini`, `ProkBERT-mini-long`, and `ProkBERT-mini-c`, consistently maintain high specificity. Their narrow interquartile ranges around upper values underscore their consistent, reliable performance.

Next, we scrutinized the relationship between evaluation time and prediction performance. It's important to note that the evaluation time encompasses not just the prediction interval but also includes sequence preprocessing and model loading durations. The ProkBERT family shines in terms of both swiftness and efficacy. These methods, regardless of sequence length, consistently register evaluation durations under 10 seconds, making them invaluable for applications necessitating real-time predictions. Specifically, for 2kb sequences, `ProkBERT-mini-long` records a commendable accuracy of nearly 92.9%. Its Matthews Correlation Coefficient (MCC), a reliable metric of prediction prowess, stands at approximately 0.859 for the same sequence length. In contrast, both VirSorter2 and DeepVirFinder manifest protracted evaluation phases, with the latency amplifying as sequences lengthen. Remarkably, VirSorter2 demands an evaluation span surpassing 1,000 seconds for 2kb sequences. While assessing accuracy, DeepVirFinder exhibits suboptimal performance, especially with succinct sequences like 256 bp, where it achieves a mere 75%. However, it's essential to acknowledge that VirSorter2 extends beyond mere classification; it offers comprehensive annotations, a process inherently time-intensive.

In essence, the ProkBERT family represents a synergy of rapidity and reliability. Concurrently, other contenders like VirSorter2, DeepVirFinder, and INHERIT unveil distinct advantages, coupled with potential avenues for refinement.

4 Conclusion

In bioinformatics, there has always been a keen interest in developing tools that can offer precise and context-sensitive interpretations of sequences. Meeting this demand, we introduced the ProkBERT model family. These innovative models benefit from transfer learning (Pan and Yang, 2009), a method showing promise in a variety of applications. A standout feature of ProkBERT is its ability to harness vast amounts of unlabeled sequence data through self-supervised learning (He et al., 2020). This approach equips ProkBERT to handle challenges like limited labeled data, a problem that has often hindered traditional models such as CNNs, RNNs, and LSTMs (Cho et al., 2014; LeCun et al., 2015). Another strength of ProkBERT is its adaptability; it performs well in different scenarios, from those with sparse data to classic supervised learning tasks (Snell et al., 2017). When we compare ProkBERT to older models that largely depend on expansive datasets, it's clear that ProkBERT ushers in a more adaptable approach for sequence analysis in prokaryotic microbiome studies.

Our results affirm the robust generalization capabilities of the ProkBERT family. The learned representations are not only consistent but also harmonize well with established biological understanding. Specifically, the embeddings effectively delineate genomic features such as coding sequences (CDS), intergenic regions, and non-coding RNAs (ncRNA). Beyond capturing genomic attributes, the embeddings also encapsulate phylogenetic relationships. A case in point is the close proximity in the embedding space between *Klebsiella pneumoniae* and *Escherichia coli*, both belonging to the *Enterobacteriaceae* family.

We validated the versatility of the ProkBERT model family by applying it to two challenging problems: promoter sequence prediction and phage identification. Promoters play an instrumental role in transcriptomic regulation. Leveraging the transfer-learning paradigm, ProkBERT adeptly addressed the promoter prediction challenge, even when fine-tuned on multi-species datasets. This adaptability addresses a significant gap, as many conventional bioinformatics tools tend to be species-specific, often overlooking microbial diversity. In comprehensive benchmarks against prominent tools, including Multiply, Promotech, and i-Promoter2L, ProkBERT consistently outclassed both traditional machine learning and deep learning counterparts. For instance, in *E. coli* promoter recognition, it achieved an accuracy of 0.87 and an MCC of 0.74, and even in a mixed-species context, the accuracy was 0.81 with an MCC of 0.62. Additionally,

our findings underscore the robustness of the training, with the `ProkBERT-mini` variant demonstrating resilience against variations in optimization parameters, such as learning rate.

Our evaluations demonstrate the prowess of ProkBERT in classifying phage sequences. Remarkably, it achieves high sensitivity and specificity even in challenging cases where available sequence information is limited. However, this exercise also highlights an inherent limitation of ProkBERT, and more broadly of transformer models: the restricted context window size. While transformer architectures are adept at capturing long-range interactions (Lin et al., 2022), they typically have a limited view of only a few kilobases. In comparative benchmarks with varying sequence lengths, ProkBERT consistently surpassed established tools like VirSorter2 and DeepVirFinder. For instance, it attained an accuracy of 92.90% and an MCC of 0.859 in multiple benchmark studies. Intriguingly, ProkBERT even outperformed a DNA-BERT-based model, which employs a BERT architecture and vectorization strategy similar to ours.

Discussing model variants, both `ProkBERT-mini` and `ProkBERT-mini-c` have a maximum context size of 1kb, while `ProkBERT-mini-long` extends this to 2kb. Notably, `ProkBERT-mini-long` manages to use longer sequence information without compromising on prediction performance or demanding additional computational resources, thanks to the LCA tokenization strategy. Our results indicate that the local context information offered by `ProkBERT-mini-long` and `ProkBERT-mini` enhances robustness, giving them an edge over `ProkBERT-mini-c`.

ProkBERT's superiority is not limited to prediction accuracy; it also excels in terms of inference speed. Variants such as `ProkBERT-mini`, `ProkBERT-mini-long`, and `ProkBERT-mini-c` consistently deliver outstanding performance, both in terms of evaluation speed and accuracy. Regardless of the sequence length, these models typically complete evaluations in under 10 seconds, making them exceptionally suited for real-time applications (Vaswani et al., 2017).

The vector representations generated by ProkBERT can be seamlessly integrated with traditional machine learning tools, paving the way for innovative hybrid methodologies. Being an encoder architecture, ProkBERT's ability to produce embeddings for nucleotide sequences enables the direct incorporation of sequence information into more complex classifiers. This fusion of traditional and deep learning methods represents a promising frontier in bioinformatics. Furthermore, insights from natural language processing research suggest that the most informative representations may not always emerge from the final layer of a model (Rae et al., 2021). This underscores the need for future studies to delve deeper into the optimal layers for sequence representation extraction in bioinformatics models.

ProkBERT distinguishes itself by being both compact and powerful, embodying a blend of efficiency and accessibility. One prevailing challenge with contemporary large language models like GPT (Radford et al., 2019), BERT (Devlin et al., 2019), and T5 (Raffel et al., 2019) is their enormity. Models with hundreds of millions or even billions of parameters not only demand substantial computational resources but also complicate training and hyperparameter optimization processes. In stark contrast, ProkBERT is designed with a lean parameter count of approximately 20 million. This design choice ensures that it can comfortably fit within the memory constraints of modest GPUs. As a result, even researchers without access to high-performance computing setups or top-tier GPUs can utilize ProkBERT. Platforms like Google Colab, which offer free but limited GPU computation, become viable environments for training and evaluation tasks with ProkBERT.

As we present the findings of our study, it's important to recognize certain limitations and identify areas for future enhancement. These include: (i) creation of larger models: The effectiveness of our current models can be further improved by scaling up. Larger models are likely to capture more complex patterns, which is particularly beneficial for handling diverse and extensive datasets. (ii) Increasing context size: Expanding the context size in our models could lead to a better understanding of longer sequence dependencies. This enhancement is crucial for the accurate interpretation of biological sequences. (iii) Building new datasets: The development of new, comprehensive datasets is an ongoing necessity. These datasets should not only be larger in size but also more diverse, ensuring the robustness and wide applicability of our models. (iv) Diversity in sequencing applications: Despite our progress, the question of diversity in sequence applications remains. This includes broadening the range of sequences our models can recognize and applying them to a variety of biological phenomena. (v) Further applications and descriptions: Future research should also aim to add and describe additional applications. This would involve applying our models to new sequence analysis tasks, expanding the scope and utility of our work. Each of these points represents a critical area for improvement and further research. Addressing these limitations will enable us to develop more comprehensive and versatile tools in the field of bioinformatics.

In essence, our findings highlight ProkBERT's capability to learn detailed and adaptable vector representations for sequences. These representations hold promise not only for current analytical challenges but also for emergent and unforeseen sequence classification tasks in the future. Amidst the challenges of understanding microbial communities, ProkBERT stands as a transformative tool, elucidating the complex interplay of genes and organisms in the microbiome with remarkable precision.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://github.com/nbrg-ppcu/prokbert.

Author contributions

BL: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

IS-N: Investigation, Software, Validation, Writing – original draft. BB: Conceptualization, Data curation, Investigation, Methodology, Software, Writing – original draft. NL-N: Conceptualization, Formal analysis, Methodology, Writing – original draft, Writing – review & editing. JJ: Data curation, Methodology, Validation, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by grants of the Hungarian National Development, Research and Innovation (NKFIH) Fund, OTKA PD (138055). The work was also supported by EHPC-DEV-2022D10-001 (Development).

Acknowledgments

Thanks are due to Prof. S. Pongor (PPCU, Budapest) for help and advice. The authors extend their gratitude to all members of ML4Microbiome for their valuable discussions and feedback on this research during the ML4Microbiome meetings. The authors gratefully acknowledge the HPC RIVR consortium (www.hpc-rivr.si) and EuroHPC JU (eurohpc-ju.europa.eu) for funding this research by providing computing resources of the HPC system Vega at the Institute of Information Science (www.izum.si) as well as to HPC-KIFU Komondor.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2023.1331233/full#supplementary-material

References

Alipanahi, B., Delong, A., Weirauch, M. T., and Frey, B. J. (2015). Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nat. Biotechnol.* 33, 831–838. doi: 10.1038/nbt.3300

Amin, R., Rahman, C. R., Ahmed, S., Sifat, M. H. R., Liton, M. N. K., Rahman, M. M., et al. (2020). iPromoter-BnCNN: a novel branched CNN-based predictor for identifying and classifying sigma promoters. *Bioinformatics* 36, 4869–4875. doi: 10.1093/bioinformatics/btaa609

Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., et al. (2008). The RAST Server: rapid annotations using subsystems technology. *BMC Gen.* 9, 1–15. doi: 10.1186/1471-2164-9-75

Bai, G.-H., Lin, S.-C., Hsu, Y.-H., and Chen, S.-Y. (2022a). The human virome: viral metagenomics, relations with human diseases, and therapeutic applications. *Viruses* 14, 278. doi: 10.3390/v14020278

Bai, Z., Zhang, Y.-Z., Miyano, S., Yamaguchi, R., Fujimoto, K., Uematsu, S., et al. (2022b). Identification of bacteriophage genome sequences with representation learning. *Bioinformatics* 38, 4264–4270. doi: 10.1093/bioinformatics/btac509

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020a). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020b). "Language models are few-shot learners," in *Advances in Neural Information Processing Systems* 33.

Camargo, A., et al. (2023). IMG/VR v4: an expanded database of uncultivated virus genomes 782 within a framework of extensive functional, taxonomic, and ecological metadata. *Nucleic 783 Acids Res.* 51, D733–D743. doi: 10.1093/nar/gkac1037

Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P., and Huerta-Cepas, J. (2021). eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Molec. Biol. Evol.* 38, 5825–5829. doi: 10.1093/molbev/msab293

Cassiano, M. H. A., and Silva-Rocha, R. (2020). Benchmarking bacterial promoter prediction tools: Potentialities and limitations. *Msystems* 5, e00439. doi: 10.1128/mSystems.00439-20

Chevez-Guardado, R., and Pe na-Castillo, L. (2021). Promotech: a general tool for bacterial promoter recognition. *Genome Biol.* 22, 1–16. doi: 10.1186/s13059-021-02514-9

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., et al. (2014). "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* 1724–1734. doi: 10.3115/v1/D14-1179

Cole, S., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., et al. (1998). Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature* 393, 537–544. doi: 10.1038/31159

Dalla-Torre, H., Gonzalez, L., Mendoza-Revilla, J., Carranza, N. L., Grzywaczewski, A. H., Oteri, F., et al. (2023). The nucleotide transformer: Building and evaluating robust foundation models for human genomics. *bioRxiv, 2023–01*. doi: 10.1101/2023.01.11.523679

de Avila e Silva, S., Echeverrigaray, S., and Gerhardt, G. J. (2011). BacPP: bacterial promoter prediction–a tool for accurate sigma-factor specific assignment in enterobacteria. *J. Theor. Biol.* 287, 92–99. doi: 10.1016/j.jtbi.2011.07.017

De la Cruz, F., and Davies, J. (2000). Horizontal gene transfer and the origin of species: lessons from bacteria. *Trends Microbiol.* 8, 128–133. doi: 10.1016/S0966-842X(00)01703-0

Delcher, A. L., Harmon, D., Kasif, S., White, O., and Salzberg, S. L. (1999). Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.* 27, 4636–4641. doi: 10.1093/nar/27.23.4636

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186.

Driscoll, J., Brody, S., and Kollef, M. (2007). Pseudomonas aeruginosa: pathogenesis and pathogenic mechanisms. *Int. J. Med. Microbiol.* 297, 277–289. doi: 10.5539/ijb.v7n2p44

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511790492

Fernandes, M. A., Verstraete, S. G., Phan, T. G., Deng, X., Stekol, E., LaMere, B., et al. (2019). Enteric virome and bacterial microbiota in children with ulcerative colitis and Crohn's disease. *J. Pediatr. Gastroenterol. Nutr*. 68, 30. doi: 10.1097/MPG.0000000000002140

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28, 3150–3152. doi: 10.1093/bioinformatics/bts565

Guo, J., Bolduc, B., Zayed, A. A., Varsani, A., Dominguez-Huerta, G., Delmont, T. O., et al. (2021). VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* 9, 1–13. doi: 10.1186/s40168-020-00990-y

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., et al. (2022). A survey on vision transformer. *IEEE Trans. Patt. Anal. Mach. Intell*. 45, 87–110. doi: 10.1109/TPAMI.2022.3152247

Han, M., Yang, P., Zhong, C., and Ning, K. (2018). The human gut virome in hypertension. *Front. Microbiol*. 9, 3150. doi: 10.3389/fmicb.2018.03150

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. (2020). "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9726–9735. doi: 10.1109/CVPR42600.2020.00975

He, W., Jia, C., Duan, Y., and Zou, Q. (2018). 70ProPred: a predictor for discovering sigma70 promoters based on combining multiple features. *BMC Syst. Biol*. 12, 99–107. doi: 10.1186/s12918-018-0570-1

Hoarfrost, A., Aptekmann, A., Farfa nuk, G., and Bromberg, Y. (2022). Deep learning of a bacterial and archaeal universal language of life enables transfer learning and illuminates microbial dark matter. *Nat. Commun*. 13, 2606. doi: 10.1038/s41467-022-30070-8

Jansson, J. K., and Wu, R. (2023). Soil viral diversity, ecology and climate change. *Nat. Rev. Microbiol*. 21, 296–311. doi: 10.1038/s41579-022-00811-z

Ji, Y., Zhou, Z., Liu, H., and Davuluri, R. V. (2021). DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics* 37, 2112–2120. doi: 10.1093/bioinformatics/btab083

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589. doi: 10.1038/s41586-021-03819-2

Kelley, D. R., Reshef, Y. A., Bileschi, M., Belanger, D., McLean, C. Y., and Snoek, J. (2018). Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Res*. 28, 739–750. doi: 10.1101/gr.227819.117

Kieft, K., Zhou, Z., and Anantharaman, K. (2020). VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome* 8, 1–23. doi: 10.1186/s40168-020-00867-0

Koski, T., and Noble, J. M. (2001). A review of Bayesian networks and structure learning. *Mathem. Appl*. 29, 9–36. doi: 10.14708/ma.v40i1.278

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Li, W., and Godzik, A. (2006). CD-HIT: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659. doi: 10.1093/bioinformatics/btl158

Li, W., O'Neill, K. R., Haft, D. H., DiCuccio, M., Chetvernin, V., Badretdin, A., et al. (2021). RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. *Nucleic Acids Res*. 49, D1020–D1028. doi: 10.1093/nar/gkaa1105

Liang, G., Conrad, M. A., Kelsen, J. R., Kessler, L. R., Breton, J., Albenberg, L. G., et al. (2020). Dynamics of the stool virome in very early-onset inflammatory bowel disease. *J. Crohn's Colitis* 14, 1600–1610. doi: 10.1093/ecco-jcc/jjaa094

Lin, T., Wang, Y., Liu, X., and Qiu, X. (2022). A survey of transformers. *AI Open* 3, 111–132. doi: 10.1016/j.aiopen.2022.10.001

Liu, B., Yang, F., Huang, D.-S., and Chou, K.-C. (2018). iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC. *Bioinformatics* 34, 33–40. doi: 10.1093/bioinformatics/btx579

Lowy, F. D. (1998). Staphylococcus aureus infections. *New England J. Med*. 339, 520–532. doi: 10.1056/NEJM199808203390806

Lukashin, A. V., and Borodovsky, M. (1998). GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res*. 26, 1107–1115. doi: 10.1093/nar/26.4.1107

Meyer, F., Bagchi, S., Chaterji, S., Gerlach, W., Grama, A., Harrison, T., et al. (2019). MG-RAST version 4-lessons learned from a decade of low-budget ultra-high-throughput metagenome analysis. *Brief. Bioinform*. 20, 1151–1159. doi: 10.1093/bib/bbx105

Min, S., Lee, B., and Yoon, S. (2017). Deep learning in bioinformatics. *Brief. Bioinform*. 18, 851–869. doi: 10.1093/bib/bbw068

Nakatsu, G., Zhou, H., Wu, W. K. K., Wong, S. H., Coker, O. O., Dai, Z., et al. (2018). Alterations in enteric virome are associated with colorectal cancer and survival outcomes. *Gastroenterology* 155, 529–541. doi: 10.1053/j.gastro.2018.04.018

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucl. Acids Res*. 44, D733–D745. doi: 10.1093/nar/gkv1189

Paczosa, M. K., and Mecsas, J. (2016). Klebsiella pneumoniae: going on the offense with a strong defense. *Microbiol. Molec. Biol. Rev*. 80, 629–661. doi: 10.1128/MMBR.00078-15

Pan, S. J., and Yang, Q. (2009). A survey of transfer learning. *J. Mach. Lear. Res*. 22, 1–40. doi: 10.1109/TKDE.2009.191

Popoff, M. Y., Bockemuhl, J., and Gheesling, L. L. (2004). Supplement 2002 (no. 46) to the Kauffmann-White scheme. *Res. Microbiol*. 155, 568–570. doi: 10.1016/j.resmic.2004.04.005

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog* 1, 9.

Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., et al. (2021). Scaling language models: methods, analysis insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., et al. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Lear. Res*. 21, 5485–5551.

Rahman, M. S., Aktar, U., Jani, M. R., and Shatabda, S. (2019). iPro70-FMWin: identifying Sigma70 promoters using multiple windowing and minimal features. *Molec. Genet. Genom*. 294, 69–84. doi: 10.1007/s00438-018-1487-5

Ren, J., Song, K., Deng, C., Ahlgren, N. A., Fuhrman, J. A., Li, Y., et al. (2020). Identifying viruses from metagenomic data using deep learning. *Quant. Biol*. 8, 64–77. doi: 10.1007/s40484-019-0187-4

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Mathem*. 20, 53–65. doi: 10.1016/0377-0427(87)90125-7

Salamov, V. S. A., and Solovyevand, A. (2011). "Automatic annotation of microbial genomes and metagenomic sequences," in *Metagenomics and its Applications in Agriculture, Biomedicine and Environmental Studies*, 61–78.

Santos-Zavaleta, A., Salgado, H., Gama-Castro, S., Sánchez-Pérez, M., Gómez-Romero, L., Ledezma-Tejeida, D., et al. (2019). RegulonDB v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in E. coli K-12. *Nucleic Acids Res*. 47, D212–D220. doi: 10.1093/nar/gky1077

Sapoval, N., Aghazadeh, A., Nute, M. G., Antunes, D. A., Balaji, A., Baraniuk, R., et al. (2022). Current progress and open challenges for applying deep learning across the biosciences. *Nat. Commun*. 13, 1728. doi: 10.1038/s41467-022-29268-7

Schackart, I. I. I., K. E., Graham, J. B., Ponsero, A. J., and Hurwitz, B. L. (2023). Evaluation of computational phage detection tools for metagenomic datasets. *Front. Microbiol*. 14, 1078760. doi: 10.3389/fmicb.2023.1078760

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30, 2068–2069. doi: 10.1093/bioinformatics/btu153

Shahmuradov, I. A., Mohamad Razali, R., Bougouffa, S., Radovanovic, A., and Bajic, V. B. (2017). bTSSfinder: a novel tool for the prediction of promoters in cyanobacteria and Escherichia coli. *Bioinformatics* 33, 334–340. doi: 10.1093/bioinformatics/btw629

Shoeybi, M., Patwary, M., Puri, R., LeGresley, P., Casper, J., and Catanzaro, B. (2019). Megatron-LM: training multi-billion parameter language models using model parallelism. *CoRR, abs/1909.08053*.

Snell, J., Swersky, K., and Zemel, R. (2017). "Prototypical networks for few-shot learning," in *Advances in Neural Information Processing Systems* 4077–4087.

Sommer, M. J., and Salzberg, S. L. (2021). Balrog: A universal protein model for prokaryotic gene prediction. *PLoS Comput. Biol*. 17, e1008727. doi: 10.1371/journal.pcbi.1008727

Su, W., Liu, M.-L., Yang, Y.-H., Wang, J.-S., Li, S.-H., Lv, H., et al. (2021). PPD: a manually curated database for experimentally verified prokaryotic promoters. *J. Molec. Biol*. 433, 166860. doi: 10.1016/j.jmb.2021.166860

Tatusova, T., DiCuccio, M., Badretdin, A., Chetvernin, V., Nawrocki, E. P., Zaslavsky, L., et al. (2016). NCBI prokaryotic genome annotation pipeline. *Nucl. Acids Res*. 44, 6614–6624. doi: 10.1093/nar/gkw569

Tenaillon, O., Rodríguez-Verdugo, A., Gaut, R., McDonald, P., Bennett, A., Long, A., et al. (2012). The molecular diversity of adaptive convergence. *Science* 335, 457–461. doi: 10.1126/science.1212986

Umarov, R. K., and Solovyev, V. V. (2017). Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks. *PLoS ONE* 12, e0171410. doi: 10.1371/journal.pone.0171410

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Advances in Neural Information Processing Systems* 30.

Vázquez-Boland, J. A., Kuhn, M., Berche, P., Chakraborty, T., Domínguez-Bernal, G., Goebel, W., et al. (2011). Listeria monocytogenes: survival and adaptation in the gastrointestinal tract. *Front. Cell. Infect. Microbiol.* 1, 3. doi: 10.1128/CMR.14.3.584-640.2001

Walker, P. J., Siddell, S. G., Lefkowitz, E. J., Mushegian, A. R., Adriaenssens, E. M., Alfenas-Zerbini, P., et al. (2022). Recent changes to virus taxonomy ratified by the International Committee on Taxonomy of Viruses (2022). *Arch. Virol.* 167, 2429–2440. doi: 10.1007/s00705-022-05516-5

Wang, S., Cheng, X., Li, Y., Wu, M., and Zhao, Y. (2018). Image-based promoter prediction: a promoter prediction method based on evolutionarily generated patterns. *Scient. Rep.* 8, 17695. doi: 10.1038/s41598-018-36308-0

Wu, L.-Y., Pappas, N., Wijesekara, Y., Piedade, G. J., Brussaard, C. P., and Dutilh, B. E. (2023). Benchmarking bioinformatic virus identification tools using real-world metagenomic data across biomes. *bioRxiv, 2023–04*. doi: 10.1101/2023.04.26.538077

Yan, M., Pratama, A. A., Somasundaram, S., Li, Z., Jiang, Y., Sullivan, M. B., et al. (2023). Interrogating the viral dark matter of the rumen ecosystem with a global virome database. *Nat. Commun.* 14, 5254. doi: 10.1038/s41467-023-41075-2

Yang, K., Wang, X., Hou, R., Lu, C., Fan, Z., Li, J., et al. (2023). Rhizosphere phage communities drive soil suppressiveness to bacterial wilt disease. *Microbiome* 11, 1–18. doi: 10.1186/s40168-023-01463-8

Zdobnov, E. M., and Apweiler, R. (2001). InterProScan-an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 17, 847–848. doi: 10.1093/bioinformatics/17.9.847

Zhang, M., Li, F., Marquez-Lago, T. T., Leier, A., Fan, C., Kwoh, C. K., et al. (2019). MULTiPly: a novel multi-layer predictor for discovering general and specific types of promoters. *Bioinformatics* 35, 2957–2965. doi: 10.1093/bioinformatics/btz016

Zhang, S., Fan, R., Liu, Y., Chen, S., Liu, Q., and Zeng, W. (2023). Applications of transformer-based language models in bioinformatics: a survey. *Bioinform. Adv.* 3, vbad001. doi: 10.1093/bioadv/vbad001

Zhang, X., Wang, R., Xie, X., Hu, Y., Wang, J., Sun, Q., et al. (2022). Mining bacterial NGS data vastly expands the complete genomes of temperate phages. *NAR Genom. Bioinform.* 4, lqac057. doi: 10.1093/nargab/lqac057

Zhao, G., Vatanen, T., Droit, L., Park, A., Kostic, A. D., Poon, T. W., et al. (2017). Intestinal virome changes precede autoimmunity in type I diabetes-susceptible children. *Proc. Natl. Acad. Sci.* 114, E6166–E6175. doi: 10.1073/pnas.1706359114

Zhou, Z., Ji, Y., Li, W., Dutta, P., Davuluri, R., and Liu, H. (2023). DNABERT-2: Efficient foundation model and benchmark for multi-species genome. *arXiv preprint arXiv:2306.15006*.

Zuo, W., Michail, S., and Sun, F. (2022). Metagenomic analyses of multiple gut datasets revealed the association of phage signatures in colorectal cancer. *Front. Cell. Infect. Microbiol.* 12, 918010. doi: 10.3389/fcimb.2022.918010

frontiers | Frontiers in Microbiology

TYPE Methods
PUBLISHED 12 February 2024
DOI 10.3389/fmicb.2024.1341152

OPEN ACCESS

EDITED BY
Domenica D'Elia,
National Research Council (CNR), Italy

REVIEWED BY
Donato Cascio,
University of Palermo, Italy
Ramona Suharoschi,
University of Agricultural Sciences and Veterinary Medicine of Cluj-Napoca, Romania

*CORRESPONDENCE
Sabina Tangaro
✉ sabina.tangaro@uniba.it

RECEIVED 19 November 2023
ACCEPTED 31 January 2024
PUBLISHED 12 February 2024

CITATION
Tangaro S, Lopalco G, Sabella D, Venerito V, Novielli P, Romano D, Di Gilio A, Palmisani J, de Gennaro G, Filannino P, Latronico R, Bellotti R, De Angelis M and Iannone F (2024) Unraveling the microbiome-metabolome nexus: a comprehensive study protocol for personalized management of Behçet's disease using explainable artificial intelligence.
Front. Microbiol. 15:1341152.
doi: 10.3389/fmicb.2024.1341152

Unraveling the microbiome-metabolome nexus: a comprehensive study protocol for personalized management of Behçet's disease using explainable artificial intelligence

Sabina Tangaro[1,2]*, Giuseppe Lopalco[3], Daniele Sabella[3], Vincenzo Venerito[3], Pierfrancesco Novielli[1,2], Donato Romano[1,2], Alessia Di Gilio[4], Jolanda Palmisani[4], Gianluigi de Gennaro[4], Pasquale Filannino[1], Rosanna Latronico[1], Roberto Bellotti[2,5], Maria De Angelis[1] and Florenzo Iannone[3]

[1]Dipartimento di Scienze del Suolo, della Pianta e degli Alimenti, Università degli Studi di Bari Aldo Moro, Bari, Italy, [2]Istituto Nazionale di Fisica Nucleare, Sezione di Bari, Bari, Italy, [3]Dipartimento di Medicina di Precisione e Rigenerativa e Area Jonica, Università degli Studi di Bari Aldo Moro, Bari, Italy, [4]Dipartimento di Bioscienze, Biotecnologie e Ambiente, Università degli Studi di Bari Aldo Moro, Bari, Italy, [5]Dipartimento Interateneo di Fisica 'M. Merlin', Università degli Studi di Bari Aldo Moro, Bari, Italy

The presented study protocol outlines a comprehensive investigation into the interplay among the human microbiota, volatilome, and disease biomarkers, with a specific focus on Behçet's disease (BD) using methods based on explainable artificial intelligence. The protocol is structured in three phases. During the initial three-month clinical study, participants will be divided into control and experimental groups. The experimental groups will receive a soluble fiber-based dietary supplement alongside standard therapy. Data collection will encompass oral and fecal microbiota, breath samples, clinical characteristics, laboratory parameters, and dietary habits. The subsequent biological data analysis will involve gas chromatography, mass spectrometry, and metagenetic analysis to examine the volatilome and microbiota composition of salivary and fecal samples. Additionally, chemical characterization of breath samples will be performed. The third phase introduces Explainable Artificial Intelligence (XAI) for the analysis of the collected data. This novel approach aims to evaluate eubiosis and dysbiosis conditions, identify markers associated with BD, dietary habits, and the supplement. Primary objectives include establishing correlations between microbiota, volatilome, phenotypic BD characteristics, and identifying patient groups with shared features. The study aims to identify taxonomic units and metabolic markers predicting clinical outcomes, assess the supplement's impact, and investigate the relationship between dietary habits and patient outcomes. This protocol contributes to understanding the microbiome's role in health and disease and pioneers an XAI-driven approach for personalized BD management. With 70 recruited BD patients, XAI algorithms will analyze multi-modal clinical data, potentially revolutionizing BD management and paving the way for improved patient outcomes.

KEYWORDS

explainable artificial intelligence, microbiome, volatilome, Behçet's disease, gut, butyrate, innate immunity, autoinflammatory disease

1 Introduction

Human microbiome is the set of all the microorganisms that live in symbiosis with the human body, including bacteria, fungi, viruses and archaea. It has been found that, in a standard 70 kg male, bacteria are as numerous as somatic cells (Sender et al., 2016), but, due to their small dimensions, they contribute only 3% of the whole human body weight (Flint, 2012). Nevertheless, microbial communities are essential to keep the human body healthy. They synthesize some vitamins that our genes are not able to LeBlanc et al. (2013), help in the digestive processes (McConnell et al., 2008), teach the immune system how to recognize pathogens or cancer cells and even produce anti-inflammatory or anti-cancer compounds to defeat them (Nakkarach et al., 2021). The study of the human microbiome has demonstrated that microbial cell gene number in the human body are 150 times larger than our own genome (Zhu et al., 2010; Grice and Segre, 2012) and radically different collections of microbes have been found between different people. Scarce knowledge about what are the causes of these variations and what regulates them has been achieved. A very impactful issue is that no understanding on how the human microbiome modification has influence on wellness, conservation of health, starting and rise of diseases has been reached (Gilbert et al., 2018; Mandrioli et al., 2019). However, a correlation between changes in the microbiome, its metabolome and interaction with the immune, endocrine and nervous systems and the appearance of a wide spectrum of diseases [e.g., inflammatory bowel disease (Frank et al., 2007; Gevers et al., 2014; Ni et al., 2017), cancer (Kostic et al., 2013) or depressive disorders (Jiang et al., 2015; Zheng et al., 2016)] has been detected. This finding indicates the possibility of treating this kind of illness by manipulation of such a microbial community. Variations in human oral or intestinal microbiome and its volatilome can mirror host lifestyle and affect the levels of diseases biomarkers (Vernocchi et al., 2020). The comprehension of the relationships between host microbiome and phenotypes is of fundamental importance to understand health or disease states. Similarly, chemical characterization of human breath and the identification of volatile organic compounds (VOCs) patterns linked to a specific disease, can provide information on the health state of a patient and allow early diagnosis of chronic diseases or the monitoring of the patient's health state along therapeutic follow-up. In fact, VOCs are final products of cellular metabolic processes and their nature and/or concentration in human breath change along with metabolic pathways when a pathologic state onsets (Mozdiak et al., 2019).

Data from human microbiome and breath are inherently complex, noisy and highly variable because several factors such as diet, sex, hormonal status, drugs, habits, etc. could affect them. So, non-standard analytical methodologies are needed to extract their clinical and scientific potential. Nowadays, a lot of Artificial Intelligence (AI) methods, such as Machine Learning (ML) or complex networks, are available to catch this complexity. In particular, AI methods use several layers of linear and/or non-linear calculating units to understand the data they manipulate and to learn "patterns" from the same data. This learning can be used to classify the observations or to make predictions on them (Hassabis et al., 2017; Amodeo et al., 2021). The specific AI model to be used is chosen according to its capability to maximize prediction accuracy but requires, on the other hand, an increased complexity of the model itself, that makes it less interpretable (Shaban-Nejad et al., 2021) (e.g., "black boxes"). To overcome these drawbacks, coming from more complex models, and to adapt ML utilization to clinical contexts, eXplainable Artificial Intelligence (XAI) techniques have been introduced, that provide explanations for decisions the algorithm takes and for the risk scores calculated for every subject studied. Such a gain in interpretability for the chosen model is converted in the possibility to understand the main reasons standing behind a prediction and to point out the factors that majorly affect clinical risk scores at individual level. This approach is perfectly placed in an innovative concept of Personalized Medicine that requires the help of AI techniques.

The target of the proposed study is the Behçet Disease (BD), also known as Silk Road disease, a rare, complex and multi-systemic chronic vasculitis, characterized by mucocutaneous, articular, vascular and ocular lesions and also by central nervous system (CNS) symptoms. The most recurring signs of this disease are relapsing genital and oral aphthae (that can also spread in the whole digestive tract), ocular pathologies (>50% of cases), arthralgia and/or arthritis (45% of cases), venous system vasculitis and thrombosis. If thromboses occur in the arterial system, they usually involve pulmonary vessels. Neurological signs (neuro-BD) are frequent (>20%); they often occur 1–10 years after the first symptoms, and include headache, hemiparesis, behavior alterations and sphincter dysfunctions. Nowadays, BD etiology is still not clear and cannot be traced back to a single root cause: the overactivation of the innate immune system, typical of this disease, seems to be caused by an altered T-cells homeostasis, but it is common thought that also some components of the human microbiome can promote an abnormal adaptive immune response, in presence of a favorable genetic background (Rodrìguez-Carrio et al., 2021). In fact, several studies have linked BD to an intestinal or oral microbiota dysbiosis: in particular, a decrease in number of butyrate-producing bacteria, associated to a lower level of butyrate in fecal samples of patients has been noted (Consolandi et al., 2015). As concerning gut, butyrate is involved in regulatory T cells differentiation (Furusawa et al., 2013) and in the release inhibition of pro-inflammatory cytokines (Weng et al., 2007). Low production of butyrate in patients suffering from BD may cause both reduced T-reg responses and T-cells immune-pathological responses activation, as suggested by the prevalence of T helper cells Th1 and Th17 in patients affected by BD (Alpsoy, 2016). Influencing intestinal microbiota, with factors such as the diet, can have a role in correcting intestinal dysbiosis and in reducing the severity of BD active phases. The evidence collected in the last decade highlight that adhering to dietary patterns which include high content of fibers can be linked to a better intestinal microbiota equilibrium; such a condition is favorable for

short chain fatty acid (SCFA) producer bacteria and unfavorable for bacteria species associated to a pro-inflammatory pattern (Fu et al., 2020). Microbiota associated with dietary patterns rich in fibers was found to be positively correlated with high levels of SCFA (acetate, propionate, butyrate). Intestinal microbiota produces SCFA during indigestible polysaccharides (fibers) fermentation; these acid compounds have a well-documented protective role against several pathologies (Ho et al., 2018). To the best of our knowledge, a well-defined diet plan for BD does not exist, and the general advice is to follow a balanced diet and to maintain an ideal weight. Nevertheless, the just mentioned studies allow us to speculate that following a diet rich in fiber can correct intestinal dysbiosis, which is involved in the BD pathogenetic mechanism, and stimulate butyrate endogenous production from intestinal microbiota, bringing to a potential improvement of clinical manifestations.

Keeping all these evidence in mind, the proposed study is aimed to: (i) establish correlations between oral and intestinal microbiota, fecal and salivary volatilome, breath and phenotypic features of human hosts, affected by BD, active and/or in remission; (ii) identify, through cluster analysis methods of metabolites, different groups of patients affected by BD; (iii) identify some taxonomic units of oral and fecal microbiota and metabolic markers that majorly contribute to the prediction of different clinical outcomes (e.g., number of active mucosal lesions, remission following the Behçet Disease Current Activity Form (BDCAF)); (iv) identify, with XAI methods (Bellantuono et al., 2022), some personalized metabolic markers that, for each patient, contribute to the prediction of his/her clinical outcome (personalized medicine); (v) evaluate the effects of soluble fiber intake (inulin) on eubiosis/dysbiosis conditions of oral and intestinal microbiota and on endogenous production of butyrate; and (vi) establish correlations between eating habits and clinical outcome of patients.

2 Methods and analysis

2.1 Study design

The project we are going to propose will be performed in three different sub-activities. The first sub-activity includes a two-arm randomized study (duration: 3 months): patients in the control arm will keep on assuming the standard therapy while patients in the treatment arm will assume soluble inulin-type fructans (inulin 90% from *Cichorium intybus L.*; Farmalabor S.r.l., Canosa di Puglia, Italy), along with the standard therapy. At the starting of study and 3 months later, for each patient, the following samples and data will be collected:

i samples for the assessment of oral/fecal microbiota;
ii breath samples;
iii clinical data such as Body Mass Index (BMI), disease duration, clinical phenotype and ocular, articular or mucocutaneous involvement;
iv laboratory data such as Erythrocyte Sedimentation Rate (ESR) and C-reactive protein (CRP);
v information on breath components;
vi information about eating habits, inviting patients to keep a food diary that can provide detailed descriptions on type and quantity of food and beverages consumed.

Furthermore, the second sub-activity will consist of:

i analysis of volatilome in breath samples and microbiota in saliva and fecal samples;
ii analysis of bacterial community taxonomic composition in fecal and saliva samples;
iii chemical characterization of breath samples.

Volatile metabolites (volatilome) chemical characterization in breath samples will be determined through gas-chromatography coupled with mass spectrometry (GC–MS). For quality assurance in sampling phase and avoid any environmental contamination of breath samples, the end-tidal fraction of the exhaled breath will be collected by an automated device named Mistral (Predict srl) and directly transfer onto suitable adsorbent cartridges (Bio-monitoring steel tube, Markes International Ltd., UK) that will be preconditioned at 330°C for 30 min with pure helium (99.999%), analyzed to verify VOCs background level and properly stored at 4°C until use. Once collected onto the adsorbent cartridges, VOCs will thermally desorb and analyze by means a thermal desorber (UNITY-2, Markes International Ltd.) coupled with a gas chromatograph (GC 7890, Agilent Technologies) and a mass selective detector (MS 5975, Agilent Technologies). The analytical methodology for VOCs characterization in breath samples has been already optimized and validated in previously published studies (Di Gilio et al., 2020a,b). With the purpose to emphasize the chemical information related to human metabolomics and identify the most part of endogenous VOCs of interest (not exclusively those included in standard mix) a semi-quantitative analysis based on compound abundances will be performed. More specifically, the GC–MS chromatograms will be analyzed using the GC–MS post-run analysis software (Agilent Mass Hunter Qualitative Analysis-Agilent Technologies Ltd., Santa Clara, USA) integrating only the peaks with intensity higher than 5 times than baseline and VOCs compounds will be identified through spectral library matching (Compounds library of the National Institute of Standards and Technology, Gaithersburg, MD 20899–1070, USA) and through comparison with GC–MS chromatograms obtained by analysis of standard solutions of 44 VOCs (Ultra Scientific Cus-5997). Microbiota composition study will be performed through metagenetic analysis of rRNA16S gene (V3 and V4 regions). A negative control for sequencing will be included in the workflow of 16S amplification and library preparation, consisting of all the reagents included in the sample processing and without the sample, to ensure that no contamination took place. Libraries will be quantified using a Qubit fluorometer (Invitrogen Co., Carlsbad, CA, USA) and pooled, including the Phix control library, to an equimolar amount (4 nM final concentration). FastQ file quality will be assessed by using FastQC software and analyzed by using the QIIME2 dedicated pipeline[1] microbiome platform (version 2020.8). Denoising will be computed with the q2-deblur QIIME plugin. Taxonomy will be inferred with the QIIME-compatible database Silva v.138 SSU, using an amplicon sequence variant (ASV) table based on error-corrected reads (Calabrese et al., 2022; Vacca et al., 2022, 2023). Finally, the last sub-activity is devoted to the implementation of XAI methods: the

1 https://qiime2.org

data obtained with the previous sub-activities will be analyzed with innovative AI methods. The aim will be to evaluate the conditions of eubiosis/dysbiosis and to identify potential microbial and metabolic markers linked to BD, to eating habits of patients and to a soluble fiber dietary supplement administration. The estimated project duration should be 18 months, including the enrollment time.

2.2 Study population

The study will be conducted on patients with BD, active or in remission according to BCDAF, aged from 18 to 65, after having signed the informed consent for participating in the study and for assuming inulin. Exclusion criteria will include pregnancy and breastfeeding, serious concomitant diseases or instability conditions (such as autoimmune diseases, chronic viral infections, malignant cancers), recent myocardial infarction (MI), chronic liver diseases and inflammatory bowel diseases (IBD) and recent (last 6 months) or current participation to slimming programs or assumption of weight loss drugs.

2.3 Interventional method

The fiber dietary supplement will be administered randomly to half of the study patients, in open-label mode. The BD patients will receive either inulin supplementation or placebo. The participants were recommended to consume the powder during the breakfast by mixing it to 150 mL of warm water and then stirring up the powder until dissolved.

At the starting point and 3 months later, for each BD patients will be collected: samples for the assessment of oral/fecal microbiota, breath samples, BMI, disease duration, clinical phenotype and ocular, articular or mucocutaneous involvement and information about eating habits. Patients in the treatment arm will assume 5 g per day of inulin in addition to their ordinary diet and in a randomized order. The 5 g dose was chosen after considering the amounts of prebiotics that would be sufficient to induce positive and significant changes in the gut microbiota, but low enough to avoid adverse effects and minimize gastrointestinal discomfort (Bouhnik et al., 1999; Kolida et al., 2007).

All data obtained, will be analyzed with innovative AI methods, in order to evaluate the conditions of eubiosis/dysbiosis and to identify potential microbial and metabolic markers linked to BD, to eating habits of patients and to a soluble fiber dietary supplement administration.

2.4 Sample size estimation

To evaluate the differences, in terms of beta-diversity, in the whole microbial population, calculating the mean presence of operative taxonomic units (OTUs) between two groups with $\alpha = 0.05$, $1\text{-}\beta = 0.80$, final effect size = 0.80, the enrollment of 26 patients is needed. Taking into account a 20% dropout rate, an amount of 35 patients for each group is needed, with a total number of 70 patients for the whole study. For the univariate logistic regression with significance level $1\text{-}\beta = 0.80$ and $\alpha = 0.05$, the target is to detect a shift of the probability (P_0) ($Y = 1$) from the value of 0.10 regarding the mean value of X to the value of 0.30 when X is increased by a standard deviation above its mean value. This outcome corresponds to an Odd Ratio (OR) of about 3.80, which requires a total sample size of 90 patients to provide a two-tail significance test. In the end, a total of 70 patients has been taken into account as the minimum number necessary for the study, because it will be needed to implement multivariate models for the adjustments. In fact, considering an expected squared multiple correlation coefficient between the covariates of about 0.30, to be included into the multivariate models, the minimum sample dimension increases to 70 patients for the two-tail significance test. Finally, a group of 70 patients with BD, classified according to ISG and/or ICBD criteria, will be selected for this study. The features of this cohort are the following: 15/70 patients with mucocutaneous involvement (active or in remission, according to Behçet's Disease Current Activity Form criteria) and 55/70 patients with articular involvement (active or in remission).

2.5 Outcome measure

In the initial phase of our study, our primary focus lies in a data-driven analysis designed to distinguish, at the 3-month period, two distinct patient groups based on microbiota and volatilome profiles. The first group undergoes traditional treatment with soluble fiber intake (inulin), while the other receives only traditional treatment. This outcome is propelled by the application of Explainable Artificial Intelligence (XAI) techniques, aiming to uncover the pivotal features contributing to the differentiation between the two groups. Our investigation extends to understanding the global and local importance of these features, providing insights into the personalized metabolic responses to treatment.

The outcome measures considered are summarized in Table 1. Integrating these biological and clinical parameters using a data-driven approach, our objective is to paint a comprehensive picture of the personalized metabolic markers associated with Behçet's disease. This dual-phase evaluation not only enriches our understanding of microbiome and metabolome nexus with the disease but also lays the groundwork for targeted interventions and more detailed treatment strategies.

2.6 Adverse events

Symptoms relating to gastrointestinal discomfort (abdominal discomfort, diarrhea, constipation, bloating, and flatulence) are

TABLE 1 Biological and clinical outcome measures considered in the presented protocol study.

Biological outcome measures	Erythrocyte sedimentation rate (ESR)
	C-reactive protein (CRP)
Clinical outcome measures	Behçet's Disease Current Activity Form (BDCAF) – measure of disease activity
	Krause Total Severity Score – measure of disease severity
	Short-form (SF)-36 quality of life (QoL) scale – measure of disease QoL

widely reported in human prebiotic feeding studies, but they remain very mild at recommended intakes (Rumessen et al., 1990; Gibson et al., 1995). Based on the literature, 16 g of inulin-type fructans per day induces no or only minor gastrointestinal symptoms in healthy or diseased adults (Cani et al., 2009; Birkeland et al., 2020). Taking potential side effects into consideration, 5 g dose was preferred over higher doses due to a precautionary principle.

2.7 Data recording and data monitoring

Follow-up assessments and data collection will be undertaken at the U.O.C. Reumatologia Universitaria of the Policlinico Hospital, Bari, Italy, by trial personnel.

2.8 Data analysis

Data collected by investigators will include volatilome, oral/fecal microbiota, body mass index (BMI), disease duration, clinical phenotype and ocular, articular or mucocutaneous involvement.

The microbiota can be characterized in three different ways: alpha diversity metrics, relative abundance of phylotypes for each specimen and community state types (CST). Alpha diversity metrics, which represent the variety and richness of organisms in a specimen, and relative abundance of microbes will be analyzed through supervised machine learning algorithms as Random Forest or XGBoost classifiers. Supervised machine learning is a category of machine learning where the algorithm is trained on a labeled dataset, which means that each example in the training data is associated with the correct output or target. The algorithm learns to make predictions or decisions based on input data by generalizing from the labeled examples it has seen during training. Moreover, the XAI algorithm "SHapley Additive exPlanations" (SHAP) will be used to detect for each patient, which features are more important for the ML algorithm in its classification (Bellantuono et al., 2023; Novielli et al., 2023). SHAP is an algorithm used in machine learning to explain the predictions made by complex models, particularly for models like XGBoost, Random Forest, neural networks, and others. It provides interpretable explanations for individual predictions, helping users understand why a particular prediction was made. The third characterization, i.e., CST, which groups samples according to the composition of the microbiota, will be analyzed through the application of complex networks (CN). This mathematical method, also known as complex systems or complex networks theory, is a branch of network science that studies systems characterized by a large number of interconnected components or nodes, and the patterns and properties that emerge from these connections. In our case, interactions between microbiome and its host are complex phenomena, and to better understand this kind of complex interactions and to map microbiome behavior is of fundamental importance to have the possibility to model these interactions through CN. Modules of this complex biological network are key organizational elements for the network itself. To detect modular organizational structures of a complex network, community detection unsupervised algorithms will be used.

2.9 Comprehensive methodology for data challenges

To ensure a robust evaluation of our models, we will implement a cross-validation strategy. Cross-validation involves partitioning the dataset into subsets, training the model on some of these subsets, and testing it on the remaining subset. This process will be repeated multiple times, and the performance metrics will be averaged. This approach ensures that our models generalize well and helps prevent overfitting.

To handle the possible presence of missing values, we will adopt a two-fold approach:

1 *Variable Selection*: Variables with a relatively low percentage of missing values (below a defined threshold, e.g., 30%) will be considered to maintain data quality.
2 *Imputation Techniques*: For variables exceeding the threshold, established imputation techniques will be employed. Additionally, we will use imputation methods such as replacing missing values with the mean or maximum of the respective variable. Importantly, these techniques will be applied separately to the training and testing datasets to prevent data leakage and ensure model generalization to unseen data.

We would also like to highlight the utility of the XGBoost algorithm, which inherently handles missing values in tree algorithms by learning branch directions during training.

To handle the potential limitation in the number of available patterns compared to the number of features considered, which could lead to overfitting, we will address the issue through the implementation of two robust techniques: data augmentation and feature reduction.

1 *Data Augmentation*: The data augmentation strategy aims to artificially amplify the quantity of training samples for deep learning models, emulating the distribution of the original dataset. This becomes especially advantageous when confronted with the constraint of a limited size in the training dataset. By introducing more diverse instances, it facilitates the model in generalizing more effectively, tackling the challenge posed by smaller training datasets. Essentially, it functions as a preprocessing technique and a type of regularization, significantly enhancing model performance and mitigating the risk of overfitting. Furthermore, the integration of Generative Adversarial Networks (GANs) into data augmentation further expands its capabilities. GANs can be employed to simulate data, generating synthetic instances that closely resemble real data. This innovative use of GANs not only augments the dataset but also introduces a layer of complexity and realism, ultimately contributing to the model's ability to generalize and perform effectively across diverse scenarios (Creswell et al., 2018).
2 *Feature Reduction*: Feature reduction is a crucial aspect of our approach. Techniques such as Principal Component Analysis (PCA) (Song et al., 2010) and wrapper methods like Boruta (Kursa et al., 2010; Bellantuono et al., 2023) will be employed. These methods effectively reduce the dimensionality of the feature space, allowing us to train models even with a limited

number of instances. This not only aids in computational efficiency but also contributes to model interpretability.

2.10 Ethics approval

This study has been approved by Comitato Etico Indipendente, Azienda Ospedaliero-Universitaria 'Consorziale Policlinico' on February 2023 (prot. n. 0023249|09/03/2023).

3 Discussion

3.1 Choice of treatment

Behçet's disease is a rare, chronic, autoimmune disorder that can affect blood vessels throughout the body. It is named after the Turkish dermatologist, Hulusi Behçet, who first described the condition in 1937. This disease primarily involves inflammation of blood vessels (vasculitis) and can affect various parts of the body. The overactivation of the innate immune system, typical of this disease, seems to be caused by an altered T-cells homeostasis, but it is common thought that also some components of the human microbiome can promote an abnormal adaptive immune response, in presence of a favorable genetic background. Behçet's disease is more common in certain regions, such as the Mediterranean, Middle East, and Asia, but it can affect people of any ethnicity. Diagnosis is often based on clinical symptoms and may require ruling out other similar conditions. Treatment typically focuses on managing symptoms and reducing inflammation.

The gut microbiome has been a subject of extensive research in the context of immunological diseases. A recent study showed that a peculiar dysbiosis of the GM is present also in individuals with BS, mainly represented by a depletion of SCFA-producing bacteria, especially of butyrate (Pagliai et al., 2020). Several trials previously showed that inulin-type fructans supplemented in doses varying between 5 and 30 g per day may increase the SCFA levels and enrich microbial diversity in healthy and diseased people (Gibson et al., 1995; Ramirez-Farias et al., 2008; Calabrese et al., 2022; Vacca et al., 2023). Thus, the aim of the present project is to conduct a trial to investigate whether a supplement of inulin could be beneficial for the gut microbiome and metabolome to the amelioration of the clinical symptoms and disease severity in individuals with BS. In support, a previous proof-of-concept study demonstrated that butyrate-enriched diets modulate the redox state of the blood and promote fibrin degradation, which is impaired by a neutrophil-dependent mechanism in BS (Becatti et al., 2016). However, the same study reported no significant effects on gut microbiota composition and SCFA production, suggesting that more effective dietary interventions are needed (Emmi et al., 2021).

3.2 Anticipated results

This will be the first study that tries to understand the complex relationships between diet, intestinal microbiota and human breath in patients affected by BD through an innovative approach based on AI methods (Golob et al., 2023; Novielli et al., 2023; Papoutsoglou et al., 2023). Such an understanding can represent a significant step forward toward the comprehension of pathogenetic mechanism at the basis of BD onset and the identification of microbial, metabolic and immunological factors and therapeutic biomarkers able to control treatment outcome and to better understand how the such a treatment can modify microbiome. In fact, intestinal dysbiosis has been linked to inflammatory diseases (Douzandeh-Mobarrez and Kariminik, 2019) and recent studies have demonstrated that therapeutic treatment in rare rheumatological diseases can modify subclinical intestinal inflammation and dysbiosis (Manasson et al., 2020), highlighting the bidirectional nature of this correspondence. Furthermore, this study will evaluate for the first time with multivariate models if microbiome and breath modulation through the diet can improve disease activity in patients with BD under treatment. This analysis could enable us to find valuable markers to identify responders and non responders, allowing treatment optimization and a personalized therapeutic approach. This study could be also useful to analyze diet effects on BD activation and/or remission. Going into details, network approach thought for this study is aimed to catch functional structure of dynamic processes happening between microbiome and human host, to identify the coexistence of different microorganisms, to trace relationships between microorganisms and to identify cohesive groups that play fundamental roles in maintaining functional relationships in the global network during the treatment. Identification and quantification of some of the topological properties of the network modules can provide important information on microbiome interactions and on their relationship with possible disorders and anomalies in inflammatory and pathological states. Specifically, co-occurrence patterns and identified polymicrobial interactions will be related with other clinical and phenotypical data to detect correlations between network functional and structural properties and biological and pathological profiles in different starting conditions. This integrative approach is completely innovative, since it will allow to highlight some connectivity patterns linked to inflammatory states, pathologies, etiological agents and even the organisms responsible for pathology transmission.

In our study protocol, we propose groundbreaking methodologies for personalized understanding of Behçet's disease. One avenue of exploration involves the utilization of breath analysis to identify distinct Volatile Organic Compounds (VOCs) patterns in exhaled breath (Di Gilio et al., 2020a). By harnessing the capabilities of artificial intelligence algorithms, we aim to explore the nexus between microbiome and metabolome offering a non-invasive and efficient approach for Behçet's disease management. Here, machine learning takes center stage, enabling us to unravel complex patterns within the oral microbiome. The goal is to uncover unique microbiome signatures associated with Behçet's disease, laying the groundwork for a personalized medicine approach. This exploration promises not only a deeper understanding of the disease but also the potential for tailored interventions based on individualized oral microbiome and metabolome profiles (Bellando-Randone et al., 2021).

In the third facet of our study, we introduce the application of explainable artificial intelligence to analyze microbiome and volatilome data related. This innovative approach addresses the limitations of traditional machine learning methods, offering a clear and interpretable understanding of disease-associated microbiome and metabolome biomarkers. By incorporating local explanation embeddings and an unsupervised clustering method, we could anticipate the identification of distinct subgroups among subjects (Novielli et al., 2023). These perspectives open the door to personalized

interventions, marking a significant stride toward a more nuanced and effective treatment paradigm for Behçet's disease.

4 Conclusion

The protocol presents a promising and innovative approach to understanding BD, with potential implications for personalized treatment strategies, using eXplainable Artificial Intelligence.

The versatility of the selected analysis methods makes it possible to apply this approach to other types of complex diseases.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

ST: Conceptualization, Funding acquisition, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing. GL: Conceptualization, Methodology, Supervision, Writing – review & editing. DS: Methodology, Writing – review & editing. VV: Methodology, Writing – review & editing. PN: Methodology, Writing – original draft, Writing – review & editing. DR: Methodology, Writing – review & editing. AG: Methodology, Writing – review & editing. JP: Methodology, Writing – review & editing. GG: Methodology, Supervision, Writing – review & editing. PF: Conceptualization, Methodology, Writing – review & editing. RL: Writing – review & editing. RB: Methodology, Writing – review & editing. MA: Conceptualization, Methodology, Supervision, Writing – review & editing. FI: Conceptualization, Funding acquisition, Methodology, Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was funded by the University of Bari, Poject XAI4Microbiome – Intelligenza Artificiale eXplainable per l'identificazione di marker metabolici personalizzati nella malattia di Behçet, code S30 – CUP H99J21017720005. The National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.4—Call for tender No. 3138 of 16 December 2021 of Italian Ministry of University and Research funded by the European Union—NextGenerationEU. Award Number: Project code: CN00000013, Concession Decree No. 1031 of 17 February 2022 adopted by the Italian Ministry of University and Research, CUP H93C22000450007, Project title: "National Centre for HPC, Big Data and Quantum Computing" support partially this project.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

References

Alpsoy, E. (2016). Behçet's disease: a comprehensive review with a focus on epidemiology, etiology and clinical features, and management of mucocutaneous lesions. *J. Dermatol.* 43, 620–632. doi: 10.1111/1346-8138.13381

Amodeo, I., de Nunzio, G., Raffaeli, G., Borzani, I., Griggio, A., Conte, L., et al. (2021). A ma chine and deep learning approach to predict pulmo nary hyperte NsIon in newborn S with congenital diaphragmatic hernia (CLANNISH): protocol for a retrospective study. *PLoS One* 16:e0259724. doi: 10.1371/journal.pone.0259724

Becatti, M., Emmi, G., Silvestri, E., Bruschi, G., Ciucciarelli, L., Squatrito, D., et al. (2016). Neutrophil activation promotes fibrinogen oxidation and thrombus formation in Behçet disease. *Circulation* 133, 302–311. doi: 10.1161/CIRCULATIONAHA.115.017738

Bellando-Randone, S., Russo, E., Venerito, V., Matucci-Cerinic, M., Iannone, F., Tangaro, S., et al. (2021). Exploring the oral microbiome in rheumatic diseases, state of art and future prospective in personalized medicine with an AI approach. *J. Pers. Med.* 11:625. doi: 10.3390/jpm11070625

Bellantuono, L., Monaco, A., Amoroso, N., Lacalamita, A., Pantaleo, E., Tangaro, S., et al. (2022). Worldwide impact of lifestyle predictors of dementia prevalence: an eXplainable artificial intelligence analysis. *Front. Big Data* 5:1027783. doi: 10.3389/fdata.2022.1027783

Bellantuono, L., Tommasi, R., Pantaleo, E., Verri, M., Amoroso, N., Crucitti, P., et al. (2023). An eXplainable artificial intelligence analysis of Raman spectra for thyroid cancer diagnosis. *Sci. Rep.* 13:16590. doi: 10.1038/s41598-023-43856-7

Birkeland, E., Gharagozlian, S., Birkeland, K. I., Valeur, J., Måge, I., Rud, I., et al. (2020). Prebiotic effect of inulin-type fructans on faecal microbiota and short-chain fatty acids in type 2 diabetes: a randomised controlled trial. *Eur. J. Nutr.* 59, 3325–3338. doi: 10.1007/s00394-020-02282-5

Bouhnik, Y., Vahedi, K., Achour, L., Attar, A., Salfati, J., Pochart, P., et al. (1999). Short-chain fructo-oligosaccharide administration dose-dependently increases fecal bifidobacteria in healthy humans. *J. Nutr.* 129, 113–116. doi: 10.1093/jn/129.1.113

Calabrese, F. M., Disciglio, V., Franco, I., Sorino, P., Bonfiglio, C., Bianco, A., et al. (2022). A low glycemic index Mediterranean diet combined with aerobic physical activity rearranges the gut microbiota signature in NAFLD patients. *Nutrients* 14:1773. doi: 10.3390/nu14091773

Cani, P. D., Lecourt, E., Dewulf, E. M., Sohet, F. M., Pachikian, B. D., Naslain, D., et al. (2009). Gut microbiota fermentation of prebiotics increases satietogenic and incretin gut peptide production with consequences for appetite sensation and glucose response after a meal. *Am. J. Clin. Nutr.* 90, 1236–1243. doi: 10.3945/ajcn.2009.28095

Consolandi, C., Turroni, S., Emmi, G., Severgnini, M., Fiori, J., Peano, C., et al. (2015). Behçet's syndrome patients exhibit specific microbiome signature. *Autoimmun. Rev.* 14, 269–276. doi: 10.1016/j.autrev.2014.11.009

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., and Bharath, A. A. (2018). Generative adversarial networks: an overview. *IEEE Signal Process. Mag.* 35, 53–65. doi: 10.1109/MSP.2017.2765202

Di Gilio, A., Catino, A., Lombardi, A., Palmisani, J., Facchini, L., Mongelli, T., et al. (2020a). Breath analysis for early detection of malignant pleural mesothelioma: volatile organic compounds (VOCs) determination and possible biochemical pathways. *Cancers* 12:1262. doi: 10.3390/cancers12051262

Di Gilio, A., Palmisani, J., Ventrella, G., Facchini, L., Catino, A., Varesano, N., et al. (2020b). Breath analysis: comparison among methodological approaches for breath sampling. *Molecules* 25:5823. doi: 10.3390/molecules25245823

Douzandeh-Mobarrez, B., and Kariminik, A. (2019). Gut microbiota and IL-17A: physiological and pathological responses. *Probiotic Antimicro. Prot.* 11, 1–10. doi: 10.1007/s12602-017-9329-z

Emmi, G., Bettiol, A., Niccolai, E., Ramazzotti, M., Amedei, A., Pagliai, G., et al. (2021). Butyrate-rich diets improve redox status and fibrin lysis in Behçet's syndrome. *Circ. Res.* 128, 278–280. doi: 10.1161/CIRCRESAHA.120.317789

Flint, H. J. (2012). The impact of nutrition on the human microbiome. *Nutr. Rev.* 70, S10–S13. doi: 10.1111/j.1753-4887.2012.00499.x

Frank, D. N., Amand, A. L. S., Feldman, R. A., Boedecker, E. C., Harpaz, N., and Pace, N. R. (2007). Molecular–phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc. Natl. Acad. Sci. U. S. A.* 104, 13780–13785. doi: 10.1073/pnas.0706625104

Fu, Y., Moscoso, D. I., Porter, J., Krishnareddy, S., Abrams, J. A., Seres, D., et al. (2020). Relationship between dietary fiber intake and short-chain fatty acid producing bacteria during critical illness: a prospective cohort study. *J. Parenter. Enter. Nutr.* 44, 463–471. doi: 10.1002/jpen.1682

Furusawa, Y., Obata, Y., Fukuda, S., Endo, T. A., Nakato, G., Takahashi, D., et al. (2013). Commensal microbe-derived butyrate induces the differentiation of colonic regulatory T cells. *Nature* 504, 446–450. doi: 10.1038/nature12721

Gevers, D., Kugathasan, S., Denson, L. A., Vazquez-Baeza, Y., Van Treuren, W., Ren, B., et al. (2014). The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host Microbe* 15, 382–392. doi: 10.1016/j.chom.2014.02.005

Gibson, G. R., Beatty, E. R., Wang, X. I. N., and Cummings, J. H. (1995). Selective stimulation of bifidobacteria in the human colon by oligofructose and inulin. *Gastroenterology* 108, 975–982. doi: 10.1016/0016-5085(95)90192-2

Gilbert, J. A., Blaser, M. J., Caporaso, J. G., Jansson, J. K., Lynch, S. V., and Knight, R. (2018). Current understanding of the human microbiome. *Nat. Med.* 24, 392–400. doi: 10.1038/nm.4517

Golob, J. L., Oskotsky, T. T., Tang, A. S., Roldan, A., Chung, V., Ha, C. W. Y., et al. (2023). Microbiome preterm birth DREAM challenge: crowdsourcing machine learning approaches to advance preterm birth research. *Cell Reports Medicine*. 5:101350. doi: 10.1016/j.xcrm.2023.101350

Grice, E. A., and Segre, J. A. (2012). The human microbiome: our second genome. *Annu. Rev. Genomics Hum. Genet.* 13, 151–170. doi: 10.1146/annurev-genom-090711-163814

Hassabis, D., Kumaran, D., Summerfield, C., and Botvinick, M. (2017). Neuroscience-Inspired Artificial Intelligence. *Neuron* 95, 245–258. doi: 10.1016/j.neuron.2017.06.011

Ho, L., Ono, K., Tsuji, M., Mazzola, P., Singh, R., and Pasinetti, G. M. (2018). Protective roles of intestinal microbiota derived short chain fatty acids in Alzheimer's disease-type beta-amyloid neuropathological mechanisms. *Expert. Rev. Neurother.* 18, 83–90. doi: 10.1080/14737175.2018.1400909

Jiang, H., Ling, Z., Zhang, Y., Mao, H., Ma, Z., and Yin, Y. (2015). Altered fecal microbiota composition in patients with major depressive disorder. *Brain Behav. Immun.* 48, 186–194. doi: 10.1016/j.bbi.2015.03.016

Kolida, S., Meyer, D., and Gibson, G. R. (2007). A double-blind placebo-controlled study to establish the bifidogenic dose of inulin in healthy humans. *Eur. J. Clin. Nutr.* 61, 1189–1195. doi: 10.1038/sj.ejcn.1602636

Kostic, A. D., Chun, E., Robertson, L., Glickman, J. N., Gallini, C. A., Michaud, M., et al. (2013). Fusobacterium nucleatum potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment. *Cell Host Microbe* 14, 207–215. doi: 10.1016/j.chom.2013.07.007

Kursa, M. B., Jankowski, A., and Rudnicki, W. R. (2010). Boruta–a system for feature selection. *Fundamenta Informat.* 101, 271–285. doi: 10.3233/FI-2010-288

LeBlanc, J. G., Milani, C., de Giori, G. S., Sesma, F., van Sinderen, D., and Ventura, M. (2013). Bacteria as vitamins suppliers to their host: a gut microbiota perspective. *Curr. Opin. Biotechnol.* 24, 160–168. doi: 10.1016/j.copbio.2012.08.005

Manasson, J., Wallach, D. S., Guggino, G., Stapylton, M., Badri, M. H., Solomon, G., et al. (2020). Interleukin-17 inhibition in spondyloarthritis is associated with subclinical gut microbiome perturbations and a distinctive interleukin-25-driven intestinal inflammation. *Arthritis Rheumatol.* 72, 645–657. doi: 10.1002/art.41169

Mandrioli, J., Amedei, A., Cammarota, G., Niccolai, E., Zucchi, E., D'Amico, R., et al. (2019). FETR-ALS study protocol: a randomized clinical trial of fecal microbiota transplantation in amyotrophic lateral sclerosis. *Front. Neurol.* 10:1021. doi: 10.3389/fneur.2019.01021

McConnell, E. L., Murdan, S., and Basit, A. W. (2008). An investigation into the digestion of chitosan (noncrosslinked and crosslinked) by human colonic Bacteria. *J. Pharm. Sci.* 97, 3820–3829. doi: 10.1002/jps.21271

Mozdiak, E., Wicaksono, A. N., Covington, J. A., and Arasaradnam, R. P. (2019). Colorectal cancer and adenoma screening using urinary volatile organic compound (VOC) detection: early results from a single-Centre bowel screening population (UK BCSP). *Tech. Coloproctol.* 23, 343–351. doi: 10.1007/s10151-019-01963-6

Nakkarach, A., Foo, H. L., Song, A. A.-L., Mutalib, N. E. A., Nitisinprasert, S., and Withayagiat, U. (2021). Anti-cancer and anti-inflammatory effects elicited by short chain fatty acids produced by Escherichia coli, isolated from healthy human gut microbiota. *Microb. Cell Factories* 20:36. doi: 10.1186/s12934-020-01477-z

Ni, J., Shen, T.-C. D., Chen, E. Z., Bittinger, K., Bailey, A., Roggiani, M., et al. (2017). A role for bacterial urease in gut dysbiosis and Crohn's disease. *Sci. Transl. Med.* 9:416. doi: 10.1126/scitranslmed.aah6888

Novielli, P., Romano, Donato, Magarelli, Michele, Bitonto, Pierpaolo D., Diacono, Domenico, Chiatante, Annalisa, et al. (2023). Explainable artificial intelligence for microbiome data analysis in colorectal cancer biomarker identification. Manuscript submitted to Frontiers in Microbiology]. 15:2024. doi: 10.3389/fmicb.2024.1348974

Pagliai, G., Dinu, M., Fiorillo, C., Becatti, M., Turroni, S., Emmi, G., et al. (2020). Modulation of gut microbiota through nutritional interventions in Behçet's syndrome patients (the MAMBA study): study protocol for a randomized controlled trial. *Trials* 21, 1–10. doi: 10.1186/s13063-020-04444-6

Papoutsoglou, G., Tarazona, S., Lopes, M. B., Klammsteiner, T., Ibrahimi, E., Eckenberger, J., et al. (2023). Machine learning approaches in microbiome research: challenges and best practices. *Front. Microbiol.* 14:1261889. doi: 10.3389/fmicb.2023.1261889

Ramirez-Farias, C., Slezak, K., Fuller, Z., Duncan, A., Holtrop, G., and Louis, P. (2008). Effect of inulin on the human gut microbiota: stimulation of Bifidobacterium adolescentis and Faecalibacterium prausnitzii. *Br. J. Nutr.* 101, 541–550. doi: 10.1017/S0007114508019880

Rodrìguez-Carrio, J., Nucera, V., Masala, I. F., and Atzeni, F. (2021). Behçet disease: from pathogenesis to novel therapeutic options. *Pharmacol. Res.* 167:105993:105593. doi: 10.1016/j.phrs.2021.105593

Rumessen, J. J., Bodé, S., Hamberg, O., and Gudmand-Høyer, E. (1990). Fructans of Jerusalem arichokes: intestinal transport, absorption, fermentation, and influence on blood glucose, insulin, and C-peptide responses in healthy subjects. *Am. J. Clin. Nutr.* 52, 675–681. doi: 10.1093/ajcn/52.4.675

Sender, R., Fuchs, S., and Milo, R. (2016). Are we really vastly outnumbered? Revisiting the ratio of bacterial to host cells in humans. *Cell* 164, 337–340. doi: 10.1016/j.cell.2016.01.013

Shaban-Nejad, A., Michalowski, M., Brownstein, J. S., and Buckeridge, D. L. (2021). Guest editorial explainable AI: towards fairness, accountability, transparency and trust in healthcare. *IEEE J. Biomed. Health Inform.* 25, 2374–2375. doi: 10.1109/JBHI.2021.3088832

Song, Fengxi, Guo, Zhongwei, and Mei, Dayong. (2010). Feature selection using principal component analysis. Proceedings of the 2010 international conference on system science, engineering design and manufacturing informatization, Yichang, China.

Vacca, M., Celano, G., Calabrese, F. M., Rocchetti, M. T., Iacobellis, I., Serale, N., et al. (2023). In vivo evaluation of an innovative synbiotics on stage IIIb-IV chronic kidney disease patients. *Front. Nutr.* 10:1215836. doi: 10.3389/fnut.2023.1215836

Vacca, M., Raspini, B., Calabrese, F. M., Porri, D., De Giuseppe, R., Chieppa, M., et al. (2022). The establishment of the gut microbiota in 1-year-aged infants: from birth to family food. *Eur. J. Nutr.* 61, 2517–2530. doi: 10.1007/s00394-022-02822-1

Vernocchi, P., Gili, T., Conte, F., Del Chierico, F., Conta, G., Miccheli, A., et al. (2020). Network analysis of gut microbiome and metabolome to discover microbiota-linked biomarkers in patients affected by non-small cell lung cancer. *Int. J. Mol. Sci.* 21, 8730–8749. doi: 10.3390/ijms21228730

Weng, M., Walker, W. A., and Sanderson, I. R. (2007). Butyrate regulates the expression of pathogen-triggered IL-8 in intestinal epithelia. *Pediatr. Res.* 62, 542–546. doi: 10.1203/PDR.0b013e318155a422

Zheng, P., Zeng, B., Zhou, C., Liu, M., Fang, Z., and Xu, X. (2016). Gut microbiome remodeling induces depressive-like behaviors through a pathway mediated by the host's metabolism. *Mol. Psychiatry* 21, 786–796. doi: 10.1038/mp.2016.44

Zhu, B., Wang, X., and Li, L. (2010). Human gut microbiome: the second genome of human body. *Protein Cells* 1, 718–725. doi: 10.1007/s13238-010-0093-z

frontiers | Frontiers in Microbiology

TYPE Review
PUBLISHED 13 February 2024
DOI 10.3389/fmicb.2024.1343572

OPEN ACCESS

EDITED BY
Aldert Zomer,
Utrecht University, Netherlands

REVIEWED BY
Balázs Ligeti,
Pázmány Péter Catholic University, Hungary
Jaak Truu,
University of Tartu, Estonia

*CORRESPONDENCE
Graziano Pesole
graziano.pesole@uniba.it;
graziano.pesole@cnr.it
Bruno Fosso
bruno.fosso@uniba.it

RECEIVED 23 November 2023
ACCEPTED 29 January 2024
PUBLISHED 13 February 2024

CITATION
Kumar B, Lorusso E, Fosso B and Pesole G (2024) A comprehensive overview of microbiome data in the light of machine learning applications: categorization, accessibility, and future directions. *Front. Microbiol.* 15:1343572. doi: 10.3389/fmicb.2024.1343572

A comprehensive overview of microbiome data in the light of machine learning applications: categorization, accessibility, and future directions

Bablu Kumar[1,2], Erika Lorusso[2,3], Bruno Fosso[2]* and Graziano Pesole[2,3]*

[1]Università degli Studi di Milano, Milan, Italy, [2]Department of Biosciences, Biotechnology and Environment, University of Bari A. Moro, Bari, Italy, [3]National Research Council, Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies, Bari, Italy

Metagenomics, Metabolomics, and Metaproteomics have significantly advanced our knowledge of microbial communities by providing culture-independent insights into their composition and functional potential. However, a critical challenge in this field is the lack of standard and comprehensive metadata associated with raw data, hindering the ability to perform robust data stratifications and consider confounding factors. In this comprehensive review, we categorize publicly available microbiome data into five types: shotgun sequencing, amplicon sequencing, metatranscriptomic, metabolomic, and metaproteomic data. We explore the importance of metadata for data reuse and address the challenges in collecting standardized metadata. We also, assess the limitations in metadata collection of existing public repositories collecting metagenomic data. This review emphasizes the vital role of metadata in interpreting and comparing datasets and highlights the need for standardized metadata protocols to fully leverage metagenomic data's potential. Furthermore, we explore future directions of implementation of Machine Learning (ML) in metadata retrieval, offering promising avenues for a deeper understanding of microbial communities and their ecological roles. Leveraging these tools will enhance our insights into microbial functional capabilities and ecological dynamics in diverse ecosystems. Finally, we emphasize the crucial metadata role in ML models development.

KEYWORDS

metagenome, shotgun sequencing, machine learning, metadata, disease prediction

1 Introduction

Human microbiome research has made significant progress in recent years, with a growing amount of metagenomic, metabolomic, and metaproteomic data that holds immense potential for hypothesis testing, meta-analyses, and disease diagnosis (Gilbert et al., 2018). However, several challenges hinder researchers from fully harnessing these resources, including the substantial time investments required, difficulties in accessing metadata, the demand for computational resources and bioinformatic expertise, and inconsistencies in annotation and formatting among individual studies (Pasolli et al., 2017).

Recently, several reviews and surveys have been published on the application of multi-omics approaches, particularly in the context of microbiome research. Marcos-Zambrano et al. (2021) focused on the application of machine learning (ML) techniques in human microbiome studies, covering topics such as features selection, biomarkers identification, disease prediction, and treatments. Hernández Medina et al. (2022) and Mathieu et al. (2022) overviewed how the latest microbiome studies harness the inductive prowess of ML and deep learning (DL) methods and considering how microbiome data peculiarities (i.e., compositionality, sparsity, and high-dimensionality)—necessitates adequate handling. Another noteworthy review article by Quince et al. (2017) emphasized best practices for shotgun metagenomic studies, discussed the identification and management of various technical limitations encountered during experimental approaches and provided an overview of implementing computational pipelines for shotgun data analysis. In a comprehensive discussion of experimental considerations for omics-based microbiome studies, Mallick et al. (2017) listed bioinformatics analysis tools tailored explicitly for metagenomics and metatranscriptomics and also touched upon the challenges associated with integrated multi-omic analyses. Nyholm et al. (2020) provided a perspective article that summarized the application of the holo-omics approach in biological research. They focused on holo-omics use cases in studies related to host-microbiota interactions, with an emphasis on exploring applications across various fields rather than engaging in a debate about available tools and methods. In a recent perspective Huttenhower et al. (2023) described how microbiome data sharing faces challenges due to its complexity and interdisciplinary nature. While best practices exist, they are not always widely adopted due to the effort involved. The need for microbiome-specific resources and recognition of data sharing efforts should be prioritized for progressing this field.

While these reviews and studies have significantly contributed to our understanding of microbiome research, there appears to be a noticeable gap in the public domain. Specifically, there seems to be a lack of comprehensive review articles that emphasize the critical importance of metadata in optimizing the implementation of ML and other advanced techniques within microbiome studies. Predictive models relying on artificial intelligence (AI) and ML tools have proven to be invaluable for gaining insights from the vast quantities of metagenomic data generated in laboratory. These tools also play a crucial role in unraveling the ecology and behavior of microbial taxa under study. AI and ML contribute to informed decision-making, effective management strategies, and conservation planning by providing a deeper understanding of microorganisms.

We aim to fill critical gaps in the existing microbiome research literature, with a specific focus on implementing machine learning (ML) techniques for microbiome classification while utilizing sample/raw metadata or disease metadata (pathological conditions) for each study and systematically reprocessing and reanalyzing the data. Unlike previous reviews, we highlight the importance of integrated metadata analysis, which involves discussing both experimental considerations (e.g., study design, sample collection, and sample processing steps) and bioinformatics considerations (e.g., managing diverse data types, assessing computational demands, selecting integration approaches, and analysis tools). We delve into the current landscape of metagenomic, metabolomic, and metaproteomic data analysis within microbial communities and concentrate on integrated metadata derived from metagenomic microbial community analyses. This review may be of interest to a broad range of researchers in the microbiome field, including those with expertise in ML, DL, and bioinformatics. We anticipate that our work will help to accelerate the development and implementation of advanced ML-based approaches for microbiome classification and disease diagnosis.

2 Exploring the diversity of microbiome data types and challenges in data analysis

2.1 "Omics" data types: understanding five distinct categories

Recent advances in next-generation sequencing (NGS) technology have enabled the generation of vast amounts of metagenomic data. Each of these data types provides unique insights into different aspects of the molecular world, and advances in high-throughput technologies and data science have made it increasingly possible to leverage all of these data types simultaneously (La Reau et al., 2023). Metagenomic sequences obtained with different sequencing strategies can be analyzed to answer a variety of questions: What is the relationship between the resolution of bacterial composition and the total number of obtained reads? To what extent do different sequencing methodologies selectively capture bacterial genera, resulting in exclusive identification by one strategy but not the other? To what degree do the sequencing approaches diverge in their capacity to explain relevant insights into specific experimental conditions? Moreover, other omics applications have been used to investigate the complexity in microbial communities, namely, metabolomics and metaproteomics. This wealth of data can be broadly categorized into five distinct types: shotgun sequencing, amplicon sequencing, metatranscriptomic data, metabolomic, and metaproteomic data.

2.1.1 DNA-metabarcoding: profiling microbial communities

The most commonly used approach to analyze microbiota is DNA-metabarcoding (also known as amplicon-based metagenomics). In metabarcoding, samples are characterized using reads obtained through the selective amplification of marker genes, like the evolutionarily conserved 16S rRNA gene or the ITS region. 16S rRNA gene profiling allow us to characterize the taxonomic composition of prokaryotic communities while ITS (ITS1 or ITS2) has been suggested for fungi (Santamaria et al., 2012, 2018; Tangaro et al., 2021). Nonetheless, there are three main limitations in Amplicon Sequencing: (I) Taxonomic resolution and the ability to profile non-bacterial members of the community, such as Eukaryotes in the environment. The conservation of the

16S rRNA gene and the length of the amplicon product restrict the achievable taxonomic resolution. This means that certain closely related taxa may be difficult to differentiate based solely on the 16S rRNA gene sequence. Approaches based on the long reads sequencing (e.g. Oxford Nanopore and Pacific Biosciences), able to cover the whole 16S rRNA and ITS regions, are promising in reach species level taxonomic resolution (Johnson et al., 2019; Notario et al., 2023). (II) Inherent limitations in functional profiling: this approach attempts to estimate functional capacity using the 16S rRNA gene, it inherently lacks the ability to directly analyze the functional potential of microbes or microbial genes. Tools exist able to infer functional capabilities based on the taxonomic profiles such as Tax4Fun2 (Asshauer et al., 2015; Wemheuer et al., 2020) and the phylogenetic investigation of communities by reconstruction of unobserved states with PICRUSt (Langille et al., 2013) and PICRUSt2 (Douglas et al., 2020), but the accuracy and resolution of these predictions are limited. (III) PCR amplification and its effects: PCR-based marker gene surveys are vulnerable to a multitude of factors that can introduce errors and bias into microbiome studies. These factors, extensively documented in the literature (Nearing et al., 2021), encompass: undersampling, differential extraction contamination, storage bias, amplification parameters and quality of the starting template. Undersampling refers to the risk to obtain an incomplete representation of the microbial community due to limited sampling. Contamination from DNA introduced during laboratory experiment through reagents and equipment, known as contaminating DNA from reagents, is another concern. The sample storage conditions under which samples are kept can significantly impact the quality and quantity of DNA. The amplification parameters employed in PCR, including enzyme choice, annealing temperature, amplification time, ramp rates, and cycle number, can introduce variability and errors. Variations in the starting template concentration can also affect the outcomes of amplification. Furthermore, DNA properties such as GC content and secondary structure, known as template properties, can influence amplification efficiency. Errors may be introduced by primer mismatches or degeneracies, where primer sequences may not perfectly match target sequences. Polymerase errors during DNA polymerization in the PCR process contribute to the issue (Berden et al., 2022). Challenges also arise from chimeric reads, which are formed from hybrid sequences originating from different templates during amplification (Haas et al., 2011). Random errors, unpredictable in nature, can emerge during the PCR process, while systematic PCR errors may be associated with specific primer pairs or conditions. It's crucial to recognize that sequencing itself introduces errors, with Illumina sequencing posing particular challenges due to its imaging-based nature (Pienaar et al., 2006). These potential sources of error and bias has led to concerns about the accuracy, reproducibility, and potential contamination in microbiome studies (Gohl et al., 2016). Nonetheless, despite the need for PCR amplification, 16S rRNA gene profiling requires a relatively low number of sequenced reads per sample (~100,000) to maximize the identification of rare taxa. This makes it a cost-effective alternative compared to shotgun metagenomic sequencing (Peterson et al., 2021).

2.1.2 Shotgun sequencing data: unveiling microbial abundance and functionality

Metagenomics experiments in the context of microbial communities employ a shotgun sequencing approach, which involves the isolation of DNA from the sample, its preparation for sequencing, and subsequent deep sequencing. Shotgun metagenomic (SM) data enable high resolution in estimation of taxon abundance from phylum (Sunagawa et al., 2013), to strain level (Scholz et al., 2016) within the original sample. In addition to taxonomic profiling, shotgun sequencing data is used for studying the functional potential of the human microbiome (Li et al., 2022). In the analysis of SM data, the sequencing depth serves as a crucial factor for understanding how it might affect the results. This impact is particularly evident when sequencing depth is insufficient, or the sample size is inadequate. A study by Li et al. (2022) reported that 15 million or higher depth as the optimal minimum sequencing to explore species level composition for metagenome-wide association studies (MWAS). The shotgun sequencing method has distinct advantages over targeted sequencing techniques, such as 16S rRNA gene sequencing. Shotgun sequencing is known for its relatively unbiased nature, making it a suitable choice for capturing the genomes of diverse species, regardless their phylogenetic origin (Lu et al., 2017). In addition, recent studies by La Reau et al. (2023) have revealed that shallow shotgun sequencing produced lower technical variation and higher taxonomic resolution than 16S rDNA sequencing at a much lower cost than deep SM sequencing.

There are several challenges and recommendations reported in SM sequencing: (I) Human DNA Contamination and Skewed Ratios: Challenges arise from shotgun sequencing approaches due to their propensity to generate reads in proportion to the relative concentrations of DNA within the sample. This often leads to an extremely skewed ratio of microbial to human DNA, resulting in human sequencing reads dominating within samples. For instance, stool samples typically consist of <10% human DNA, whereas samples obtained from sources like saliva, throat, buccal mucosa, and vaginal swabs can contain more than 90% of reads aligned to the human genome (Lloyd-Price et al., 2017). (II) Removing host-derived DNA for accurate microbial analysis: Host-derived reads should be removed from the metagenomic data before downstream analysis by using available bioinformatic tools to avoid bias in microbial quantification (Pereira-Marques et al., 2019). (III) Distinguishing active from inactive microbial populations: A major limitation of SM is that this technique does not allow distinguishing between active (alive) and inactive (dead) microbial populations and whether the predicted genes are actually expressed and under what conditions.

However, some potential sources of bias are common to both SM and meta barcoding. For instance, DNA extraction methods can significantly impact the results. In addition, in the case of SM, it is crucial to consider the differences in sequencing total DNA through a PCR-free or PCR-enriched protocol. In this case, PCR bias is also common to both strategies. These biases can influence the resolution of bacterial composition, the selective capture of bacterial genera, and the capacity to elucidate insights into specific experimental conditions using different sequencing methodologies. Understanding and addressing these biases are crucial for accurate

and reliable interpretation of metagenomic data (McLaren et al., 2019).

2.1.3 Metatranscriptomic insights: revealing microbial activity

Metatranscriptomics is the study of the transcriptional activity of microbes and microbial populations, which is particularly useful for functionally investigate the gut microbiota. It is a powerful tool for understanding the active states of microbes, their genes, and the different expressed pathways, as well as for detecting and understanding the microbial role in pathological conditions. We can gain insights into the gene expression patterns of pathogenic microorganisms and their interactions with the host by examining the RNA transcripts present in a host microbiome. This information can aid in the early detection and diagnosis of infectious diseases, as well as in monitoring treatment efficacy and disease outcomes (Bashiardes et al., 2016).

However, there are some limitations to metatranscriptomic analysis in disease detection. First, the complexity of the microbial community and the varying abundance of different transcripts can make it challenging to assess their source from pathogenic or commensal microorganisms. Additionally, technical biases and limitations in sequencing technologies (i.e. reads length) may affect the sensitivity and accuracy of detecting low-abundance transcripts. Furthermore, the interpretation of metatranscriptomic data in the context of disease requires careful consideration of various factors such as the host immune status, sample collection techniques, and potential confounding factors. Standardized protocols for sample collection, RNA extraction, and data analysis are essential to ensure reproducibility and reliability of results.

Despite these challenges, metatranscriptomic analysis holds great promise for understanding host-microbe interactions in disease (Bashiardes et al., 2016), discover novel microbial interactions (Bikel et al., 2015), detect regulatory antisense RNA (Bao et al., 2015), and track expression of genes and determine the relationship between viruses and their host (Moniruzzaman et al., 2017). Advancements in sequencing technologies, bioinformatics tools, and data integration approaches will continue to enhance our ability to harness metatranscriptomics for accurate and informative disease diagnosis and monitoring (Shakya et al., 2019).

2.1.4 Metabolomic signatures: unraveling interactions through metabolites

Metabolomics is an investigative approach focused on the analysis of small molecules (<1.5 kDa), commonly known as metabolites, within various biological samples such as urine, serum, plasma, feces, and saliva. It is challenging to differentiate between features originating from microbes and those from the host or environment, so it is crucial to have clear links between these features and the corresponding microbial profiles from the specimen. These data become most valuable when closely connected to the corresponding microbial profiles from the source specimen. Also, this method aims to identify and characterize metabolites in these samples, thereby enabling the development of distinctive metabolic profiles for individuals or populations. These profiles are reflective of a complex interplay between genetic, environmental, and microbial factors.

Metabolomics encompasses two key approaches targeted and untargeted. Targeted metabolomics focuses on specific known metabolites, commonly used for validating biomarkers or studying the effects of interventions like drug treatments or dietary changes. It offers high sensitivity and precision but is confined to the predetermined metabolites on the target list. Untargeted metabolomics aims to identify and quantify all metabolites present in a sample, enabling the discovery of new metabolites, biomarkers, and pathways. While less precise than targeted metabolomics, this method provides a wider coverage of metabolites, shedding light on complex biological interactions involving genetic, environmental, and microbial factors. Distinguishing between features from microbes, the host, or the environment is challenging, requiring clear associations between these features and the respective microbial profiles from the specimen for accurate interpretation (Bingol, 2018; Yang et al., 2019).

A noteworthy illustration of this concept can be found in the examination of bioactive microbial metabolites, specifically short-chain fatty acids (SCFAs), which includes propionate, butyrate, and acetate. These SCFAs have been implicated in the development and progression of several diseases, including inflammatory bowel disease (IBD) and colorectal cancer (Storr et al., 2013). Additionally, there are other metabolites like bile acids, sphingolipids, and tryptophan derivatives, all of which exhibit evidence of microbial interactions and bioactivity within the gut environment (Mallick et al., 2019).

Recent studies by Muller et al. (2021) have demonstrated that it is possible to differentiate between individuals with IBD and those without, as well as distinguish between specific subtypes of IBD (ulcerative colitis and Crohn's disease) by employing ML pipeline and metabolic profiling techniques. This highlights the potential of metabolomics in contributing to our understanding of the underlying metabolic alterations associated with various diseases and conditions. Notably, these alterations include metabolites closely associated with critical microbial pathways like bile acid transformations and polyamines metabolism.

Noteworthy, obtaining, processing, and comparing microbiome-metabolome datasets from multiple studies is typically a cumbersome, extremely challenging, and time-consuming process. Initial challenges include downloading the data associated with each study, which are often missing or incomplete, and linking microbiome, metabolome, and metadata sample identifiers in each study. While sharing raw and/or processed metagenomics data is common and relatively standardized in terms of formats and online open-access repositories, metabolomics data is much less standardized and often not being shared in microbiome studies. Once all the raw data have been obtained, they need to be jointly re-processed, which often requires additional expertise or the use of a variety of bioinformatic methods. Making sure taxon and metabolite identifiers can be mapped and compared across datasets is another critical challenge and may require careful and tedious curation efforts. Schorn et al. have recently addressed some of these challenges by releasing a community resource for linking raw genomic/metagenomic data with metabolomic data (Schorn et al., 2021), yet, this resource requires proficiency in

processing raw data sources and is targeted primarily at identifying and confirming novel links between biosynthetic gene clusters and metabolites (Muller et al., 2022). Regarding metabolomics raw data, the European repository MetaboloLights (Yurekten et al., 2023) currently contains 85 microbiome studies (out of 1,397, accessed 1/1/2024) and it is interesting to note how currently in the EMBL-EBI ENA (European Nucleotide Archive) repository are available 146,583 datasets, highlighting the limited amount of raw metabolomic data available (Yuan et al., 2023).

2.1.5 Metaproteomics: revealing the proteome complexity

The gut microbiome, a highly intricate ecosystem comprising trillions of microorganisms, presents a challenge for conventional DNA-based approaches (Li L. et al., 2023). These methods often fall short in elucidating the functional aspects of the microbiome, unable to confirm whether predicted genes are actively expressed, under what conditions, or to what extent (Park and Graveley, 2007; Verberkmoes et al., 2009). Moreover, the viability and activity status of the microbial cells remain uncertain. Meta-transcriptomics (described above), although offering a solution by assessing RNA expression as an indicator of gene activity, encounters challenges related to the fate of expressed RNAs, ranging from protein production to degradation or epigenetic silencing (Holoch and Moazed, 2015; Yang et al., 2016). These limitations can be overcome by directly assessing proteins.

Addressing these limitations, metaproteomic emerges as a promising avenue, utilizing liquid chromatography–tandem mass spectrometry (LC-MS/MS) to delve into protein functions. Unlike DNA and RNA methods, metaproteomic directly assesses proteins, providing insights into microbial diversity and dynamic host-microbiota interactions in the human gastrointestinal tract. This technique aids in unraveling molecular mechanisms associated with both homeostasis and disease pathogenesis (Lee et al., 2017). In other words, metaproteomic is a large-scale characterization of the entire protein complement and was initially used to study the microbial function of environmental samples, like soil, activated sludge, and acid mine drainage (Long et al., 2020).

Despite its potential, metaproteomic faces challenges, notably in the depth of analysis due to the absence of a suitable database. Taxonomic diversity calculators, commonly used in gut microbiome studies, prove insufficient in assessing functional states. The need for a functional perspective becomes evident, as diversity alone does not necessarily correlate with the microbiome's functionality (Li L. et al., 2023).

Among metaproteomic studies, a mass spectrometry-based shotgun proteomics approach is employed. This technique involves the detection and identification of all proteins in a cell mixture without gel-based separation or de novo sequencing. Peptides resulting from enzymatic digestion of the proteome are separated by liquid chromatography and analyzed through tandem mass spectrometers. The resulting information is then compared against peptide databases derived from genome sequences. While shotgun metaproteomic has shown success in studies involving microbial communities with low diversity, adapting this approach to more complex environments, such as the human gut microbiome, remains technically challenging. This method has been demonstrated in few studies, including those focused on acid mine drainage systems, endosymbionts, and sewage sludge water. Indeed, in the ProteomeXchange (Vizcaíno et al., 2014; Deutsch et al., 2017, 2022) repository, 211 studies out of 31,443 (0.7%, data accessed on 1/1/2024) regards microbiome investigations. However, challenges persist, and further advancements are needed to overcome technical limitations in analyzing complex microbial communities (Verberkmoes et al., 2009). The pursuit of a comprehensive understanding of metaproteomics is strongly recommended, with a key reference available in Xiong et al. (2015). Erickson et al. (2012) described the simultaneous application of SM and metaproteomics to identify potential functional signatures in Crohn Disease (CD).

Table 1 summarizes the advantages, disadvantages, capabilities, and recommended use of metagenomic data types.

3 Machine learning for microbiome data analysis

In microbiome studies, there is a wide range of questions yet to be solved; these question follows how microbial communities and specific microbes within those community's cause, respond to, or contribute to disease. Do various diseases exhibit unique gut microbiome alterations? Are some conditions associated with pathogen intrusion, while others demonstrate a decline in beneficial bacterial populations? Can we pinpoint microbial biomarkers consistently enriched or diminished in a given disorder across diverse patient populations? Several recent studies have highlighted the advantages of implementing the ML pipeline on SM data to understand microbial taxa, identify signatures for disease identification and diagnose complex medical conditions, particularly for gut microbiome-related diseases. These studies demonstrate the following key benefits: (I) Improved Classification Accuracy to taxa associated with IBD: Mihajlović et al. (2021) employed a random forest (RF) model to classify Inflammatory Bowel Disease (IBD), achieving an average F1 score of 91%. This underscores the robust connection between IBD and the gut microbiome, showcasing how ML can enhance diagnostic accuracy in complex diseases. (II) Access the microbial taxa signature from SM data: Liñares-Blanco et al. (2022) generated a metagenomic signature using RF, effectively identifying IBD from fecal samples. The model achieved AUC scores of 0.74 and 0.76 for different IBD subtypes, Ulcerative Colitis (UC) and Crohn's Disease (CD), respectively, highlighting the utility of ML in subtype-specific diagnosis. Bakir-Gungor et al. (2021) utilized machine learning, specifically the RF method, to develop a classification model for Type 2 Diabetes (T2D) diagnosis and revealing that a subset of 15 commonly selected features had a significant impact on minimizing the microbiota required for T2D diagnosis, thereby reducing time and cost, showcasing the efficiency of ML in biomarker selection. (III) Biomarker discovery and patient subgrouping: Another study by Bakir-Gungor et al. (2022) aimed to identify biomarkers associated with human gut microbiota during IBD. Supervised and unsupervised ML models were employed to (i) aid IBD diagnosis, (ii) discover IBD-associated biomarkers, and (iii) Identify patient subgroups using clustering approaches.

TABLE 1 Assessing metagenomic data types: advantages, disadvantages, capabilities, and recommended applications.

Data type	Definition	Capabilities*	Advantages	Disadvantage	Recommended use
Shotgun- metagenomics	Whole-genome sequencing of all genomes in a sample, including DNA from bacteria, fungi, viruses, and the host organism	• High resolution, • Moderate selectivity • High capacity	• Can identify all members of a microbial community, including novel and rare taxa. • Can be used to study gene expression and metabolic activity.	• Expensive, time-consuming, • May not be able to identify all bacterial genera at equal efficiency. • Difficult to assemble and analyze complex metagenomes. • May not be able to detect low-abundance taxa.	• Studying the diversity and composition of microbial communities, identifying new species and strains of microbes, • Investigating the functional potential of a microbial community
Amplicon- sequencing	Targeted sequencing of a specific gene or region of DNA from a sample	• Low resolution, • High selectivity • Medium capacity	• Can be used to target specific bacterial genera or genes. • Is relatively inexpensive and fast to generate	• Cannot identify all members of a microbial community • Biased toward certain bacterial genera	• Profiling the abundance of specific bacterial taxa in a community, Tracking changes in the microbial community over time, Identifying bacterial pathogens
Meta- transcriptomics	Whole-transcriptome sequencing of all RNA transcripts in a sample, including RNA from bacteria, fungi, viruses, and the host organism	• High resolution, • Moderate selectivity • High capacity	• Can be used to study gene expression and metabolic activity at a high resolution.	• Expensive, time-consuming, May not be feasible to identify all bacterial genera at equal efficiency. • Difficult to analyze, as it is not always clear which genes are being expressed by which bacteria	• Studying the functional potential of a microbial community, Identifying differentially expressed genes. • Investigating the response of a microbial community to environmental stimuli
Metabolomics	Identification and quantification of all metabolites in a sample	• Low resolution • Low selectivity • High capacity	• Can be used to study the metabolic activity of a microbial community • Can be used to identify novel metabolites.	• Cannot identify all members of a microbial community. • Biased toward certain metabolites. • Difficult to identify and quantify all of the metabolites present in a sample	• Studying the metabolic potential of a microbial community, Identifying biomarkers of disease • Analyze interaction between microbes and their environment
Metaproteomics	Study of the entire protein collection (proteome) of a microbial community	• Low resolution	• High-throughput, sensitive, quantitative	• Expensive, time-consuming, difficult to interpret results	• Study microbial communities, detect pathogens, and monitor environmental changes.

Capabilities*: - Resolution, The ability to distinguish between different microbes or genes; Selectivity, The ability to target specific microbes or genes for analysis; Capacity, The amount of data that can be generated and analyzed.

10.3389/fmicb.2024.1343572

Random Forest outperformed other classifiers, shedding light on potential microbiome-mediated mechanisms of IBD and offering insights for microbiome-based diagnostics. Another study by Zeller et al. (2014) aimed to detect early-stage colorectal cancer (CRC) by employing metagenomic sequencing of fecal samples to identify distinctive taxonomic markers distinguishing CRC patients from those without tumors. CRC-associated changes in the fecal microbiome reflected, at least in part, the microbial community composition within tumors, indicating potential tumor-related host-microbe interactions. The analysis also revealed a metabolic shift from fiber degradation in controls to host carbohydrate and amino acid utilization in CRC patients, accompanied by increased lipopolysaccharide metabolism. IV) Geospatial Microbial Provenance: In a recent study Bhattacharya et al. (2022) implemented ML to enable geospatial microbial provenance. Researchers delved into the assessment of geographical specificity within environmental metagenomes. Primary objective was to discern unique microbial signatures that could be attributed to specific cities, relying on taxonomic classifications as the basis for differentiation. The outcomes of this comprehensive analysis unveiled a remarkable level of accuracy in pinpointing the origin of metagenomic data. The accuracy rates for classifying samples by city ranged impressively from 85 to 89%, while continental classification exhibited an even higher accuracy level, fluctuating between 90 and 94%. Leung et al. (2022) proposed an integration of metagenomics, metabolomics, and clinical data to classify enrolled participants based on their NAFLD (nonalcoholic liver disease) status and liver fat accumulation, and reaching an overall AUROC score of about 93%.

Also, ML offers a significant advantage over traditional statistics in the field of microbial ecology, where conventional statistical methods have been the norm for data summarization, hypothesis testing, and interpreting interactions within microbial datasets. The primary objective is to predict specific phenotypes, such as disease states or age, based on microbiome data. One fundamental distinction between statistical models and ML lies in their primary objectives: statistical models aim to describe and infer relationships between variables, whereas ML is tailored to optimize predictive accuracy on external datasets. To illustrate, supervised ML typically employs a learning step on a training dataset with labeled data patterns associated with specific outcomes, while a separate test dataset with unlabeled data is used to evaluate the model's performance. Finally, a validation dataset could be employed to further evaluate the obtained model, when unseen data (i.e. data not used neither for training nor for testing) are used. In contrast, statistical models primarily focus on understanding how values relate to outcomes, often without the need to partition the data for performance evaluation. ML possesses several advantages over classical statistics in microbial ecology research. It excels in detecting subtle variations in microbial community structure and can pinpoint particular bacterial taxa that play a pivotal role in predicting specific outcomes. Additionally, ML can model complex, non-linear combinations of bacterial count data and environmental parameters, which closely resemble real-world systems. This obviates the need for intricate data transformations or preprocessing, which can be challenging when dealing with molecular data.

Widening this aspect, ML approaches emerge as tool for multi-omics data integration. The aim of multi-omics (or integrative omics) approaches is to extract substantial evidence from large-scale data by identifying, classifying, and quantifying different biological molecules involved in complex structure, such tissues or microbial communities (Vailati-Riboni et al., 2017). An interesting application of multi-omics approaches was proposed by Monteleone et al. (2021) in which they linked microbiota composition and metabolites in Anorexia Nervosa (AN). This condition in characterized by weight loss/regain cycles. Authors characterize both the microbiota and the metabolome in the underweight and regain phases, identifying a perturbation in gut microbiota of AN female's patients compared to healthy ones, and an association to specific metabolites.

3.1 Utilization of publicly available microbiome data in research studies

The rapid advancement of NGS technology has led to an exponential growth in the volume of data housed within publicly accessible repositories like the GenBank by the National Center for Biotechnology Information (NCBI), the Metagenomic Rapid Annotations using Subsystems Technology (MG-RAST), the European Nucleotide Archive (ENA), and the DNA Data Bank of Japan (DDBJ), among others. These repositories are invaluable resources that store vast amounts of DNA sequences (Eckert et al., 2020). Utilizing these raw sequences, made available to the public, enables the application of cutting-edge ML and DL techniques for extensive data analysis. In this section, we aim to provide an insightful overview of the current trends in metadata analysis through the use of publicly accessible raw data and associated metadata.

Pasolli et al. (2016) conducted an extensive analysis of metagenomic data, involving 2,424 publicly available datasets. They introduced an ML-based framework for predicting microbiome-phenotype associations, focusing on species-level abundances and strain-specific markers. Cross-validation revealed strong disease prediction capabilities, especially when using strain-specific markers. Interestingly, including "control" samples from other studies in training sets improved predictions. *Streptococcus anginosus* was identified as a potential marker for general microbiome dysbiosis rather than specific diseases. This work advances our understanding of microbial dysbiosis and provides a publicly accessible software framework and data.

Duvallet et al. (2017) gathered data from 28 published case-control 16S rDNA amplicon sequencing gut microbiome datasets, encompassing 10 different disease states. Their objective was to explore whether consistent and disease-specific alterations in gut microbial communities could be identified across various studies of the same disease. Notably, some diseases, like colorectal cancer (CRC), exhibited an abundance of disease-associated bacteria, while others, such as IBD, were characterized by a depletion of beneficial bacteria. Specific conditions like diarrhea displayed substantial shifts in the overall microbial community, often involving numerous associated microbes, while most conditions

showed only a few microbial associations this study identify unique patterns of dysbiosis shared across multiple disease states in the human gut microbiome, characterized by variations in the direction (i.e., the proportion of disease-enriched vs. disease-depleted genera) and the scope (i.e., the total number of genera showing differences between cases and controls) of disease-associated shifts. Pietrucci et al. (2022) investigated the possible association among gut-microbiome and Autism Spectrum Disorder (ASD) by using metabarcoding data from eight different project and 6 different geographical location. The applied several ML approaches and demonstrated their potential in overcoming limitation of classical statistical approaches and perform features selection in complex datasets.

Gupta et al. (2020) introduced the Gut Microbiome Health Index (GMHI), for assessing health status based on the species-level taxonomic profile of stool shotgun metagenome samples. GMHI evaluates the likelihood of disease presence, independently of clinical diagnosis, by comparing the relative abundances of microbial species associated with positive and negative health conditions. They implemented a mathematical index identified from a comprehensive dataset of 4,347 publicly available human stool metagenomes across various disease states. When they applied to large-scale dataset, GMHI effectively distinguishes between healthy and non-healthy groups, as compared to traditional ecological indices like Shannon diversity and richness. In Lam and Ye (2022) a network-based approach was implemented with aim to build a microbial association networks upon a subset of the Gupta et al. (2020) data. Additionally, they focused the more on analyzing diseases individually rather than a disease-agnostic approach, to better characterize microbial community traits in each disease. Lam and Ye (2022) by focusing on microbial community interactions in both healthy and diseased microbiomes, aimed at identifying factors for the stratification of disease states and the identification of potential microbial risk factors beyond individual species. Furthermore, to gain insights into community interactions across phenotypes, they also introduce a new metric called "module resilience" to study the retention of microbial community modules in microbial interaction networks.

Casimiro-Soriguer et al. (2022) performed a meta-analysis of 1,042 fecal metagenomic samples from seven publicly available studies. They applied ML pipeline based on functional profiles, instead of the conventional taxonomic profiles, to produce a highly accurate predictor of CRC with the aim to discriminate samples with adenoma, which makes this approach very promising for CRC prevention by detecting early stages in which intervention is easier and more effective. In addition, ML is used to extract features relevant to the classification, which reveals basic molecular mechanisms accounting for the changes undergone by the microbiome functional landscape in the transition from healthy gut to adenoma and CRC conditions.

Lugli et al. (2023) investigated the genetic diversity within bacterial taxa constituting the infant gut microbiome by utilizing the vast collection of publicly available shotgun metagenomic data and associated metadata from multiple global studies, encompassing infants from birth up to the age of 3 years. The extensive dataset, comprising 10,935 metagenomic profiles, enabled the identification of critical bacterial signatures within the infant microbiome, linked to distinct community-state types. Additionally, in the study metabolic reconstructions of these infant microbiomes shed light on the functional attributes of these predominant microorganisms during the early years of life, revealing potential correlations with health states from both metagenomic and metatranscriptomic perspectives.

Nelkner et al. (2023) conducted a meta-analysis using data from 16 primary studies, examining microbial communities in agricultural soils across Europe. They aimed to understand how European soil characteristics influence microbial community composition, particularly focusing on Thaumarchaeota members. Their analysis used publicly available metagenome sequencing data to assess microbial abundance at different taxonomic levels. This study highlights the significance of standardized metadata reporting and the benefits of open data sharing in the scientific community.

Key studies in microbiome research emphasize the significance of utilizing publicly available metagenomic data (Pasolli et al., 2016; Gupta et al., 2020; Lam and Ye, 2022; Lugli et al., 2023), which, when combined with metadata from different studies, facilitate the validation and confirmation of research findings. It also promotes data sharing, allowing scientists to build upon each other's work and develop comprehensive insights into complex phenomena.

3.2 Challenges to implementing machine learning

One key challenge is the interpretability of ML models, which often function as "black boxes" without clear mechanistic understanding. Interpretable ML approaches, such as deep forest algorithms and methods that incorporate prior knowledge like microbial interaction networks, are emerging to address this issue (Räz, 2024). The second barrier is the scarcity of large, high-quality, and correctly labeled microbiome datasets needed to train ML models effectively (Schloss, 2018). Additionally, ensuring data quality through techniques like deduplication, class balancing, outlier removal, and imputation is crucial. Lastly, selecting, evaluating, and tuning the right ML model for a specific task can be challenging, but a rich ecosystem of libraries and frameworks, as well as synthetic microbiome datasets, can aid in model development and benchmarking (Hernández Medina et al., 2022).

The challenges faced by ML in terms of metadata can be analogously compared to the complexities encountered in taxonomic annotation of bacteria, as discussed in the previous article by Mathieu et al. (2022). Definition and standardization of metadata: Over the past two decades, there has been a growing need for establishing not just standards for collecting and processing metagenomic data but also for developing well-defined methods for preparing metadata. This is essential to ensure the reusability of data and to train ML models for comprehensive and interdisciplinary microbiome analysis, as highlighted by Cernava et al. (2022). As bacterial species definitions are based on laboratory protocol and experiments, their relevant metadata including technical and analytical methods, must be well-defined

and standardized in ML. The lack of clear metadata definitions can lead to difficulties in classifying bacterial species and organizing raw read data to perform effective statistical tasks. Data heterogeneity: Similar to the high DNA heterogeneity observed in bacterial species, metadata can vary greatly across different datasets and sources. This data heterogeneity poses challenges in integrating and comparing information when metadata standards are inconsistent. Moreover, considering we've only accessed a fraction of bacterial diversity on Earth, metadata used in ML may be incomplete and fail to capture the full spectrum of information needed for robust model training. Datasets may lack essential metadata attributes, making it challenging to build accurate models. Data representation: Just as metagenomic assembled genomes (MAGs) may not resemble complete genomes, metadata representation can be inconsistent or not following a standard format. This can make it difficult to interpret and utilize metadata for ML purposes. Taxonomy and classification: Similarly, integrating MAGs into metagenomic classifiers is complex due to their ambiguous taxonomy affiliations. In machine learning, associating metadata with specific categories or labels can be challenging when dealing with data that doesn't neatly fit into predefined classes. Integration with Models: Just as MAGs are not fully integrated into taxonomy, metadata may not always seamlessly integrate with ML models. It requires careful preprocessing and feature engineering to incorporate metadata effectively into the modeling process.

Yilmaz et al. (2011) introduced minimum information standard about metagenomic sequence (MIMS) and the minimum information about marker gene sequence (MIMARKS). Those are two widely used standards for reporting metagenomic and DNA metabarcoding data. These standards provide checklists of essential information for sharing data, such as the sample type, collection method, sequencing platform, and data processing steps.

In addition to MIMS and MIMARKS, there are a number of other standards that can be used to report specific types of metadata, such as the environmental package (E-Package): a standard for reporting environmental metadata associated with metagenomic samples (Logares et al., 2012) and the human microbiome project (HMP) data analysis pipeline: A standard for reporting metadata associated with human microbiome studies (Huttenhower et al., 2012) and microbiome quality assurance (MQA) a protocol for reporting quality control metrics for metagenomic and DNA metabarcoding data (Lassalle et al., 2018). The adoption of these standards makes microbiome data findable, accessible and, reusable for other researchers. This is essential for accelerating progress in metagenomics and DNA metabarcoding research (ten Hoopen et al., 2017).

The technologic advancements in instrumentation toward high-throughput and high-resolution methods in metabolomics, have supported the accumulation of big data across laboratories that needs a support regarding data and metadata deposition (Haug et al., 2017). The Metabolomic Standard Initiative (MSI) and COSMOS (COordination of Standards in MetabOlomicS) (Salek et al., 2015) are constantly supporting the definition of minimum standards in metabolomic data deposition by implementing the MSI Core Information for Metabolomics Reporting (CIMR) (Sumner et al., 2007). Moreover, COSMOS is actively engaging publishers to promote the requirements for authors to deposit metabolomics results, as is required for other "omics" disciplines (Salek et al., 2013). As an outcome of the COSMOS initiative, in 2014 the MetabolomeXchange database and repository was launched. It aggregates data from the major providers, namely MetaboLights (Yurekten et al., 2023), Metabolomics Workbench (Sud et al., 2016), and Metabolomic Repository Bordeaux, to facilitate the access and reusability of metabolomic datasets and associated metadata (Ferry-Dumazet et al., 2011).

Similarly to what happened for NGS data, proteomics and metaproteomics data release (both raw and processed) was initially driven by journals guidelines, and resulting in a lack of minimal associated metadata (e.g. experimental design, peptide identification and quantification, protein identifications and protein ratios) (Olsen and Mann, 2011). In this context, the ProteomeXchange (Vizcaíno et al., 2014; Deutsch et al., 2017, 2022) international consortium aims to overcome data and metadata deposition issues by exploiting the cooperation of primary [PRIDE (Perez-Riverol et al., 2021) and PASSEL (Farrah et al., 2012)] and secondary [PeptideAtlas (Desiere et al., 2006) and UniProt (The UniProt Consortium, 2023)] resources, bioinformaticians, researchers and also representatives from journals active in the field, and offering a framework for consistent and user-friendly data deposition.

3.3 Limitation of ML/AI application to microbiome data analysis

Training by using a feature count table consisting of vectors composed of the relative representation of each taxon or MAGs in the sample is the most common approach to develop a predictive model (Figure 1), which is followed by normalizing the raw counts using an appropriate approach accounting for sparsity and data compositionality (Gloor et al., 2017; Casimiro-Soriguer et al., 2022). However, the implementation of ML comes with its own set of limitations, potential errors and common challenges associated with applying ML to this input data:

3.3.1 Data quality and pre-processing

Due to the high dimensionality, sparsity, and noise of metagenomic data, a significant challenge arises during the normalization process to feed into the ML model. Non-biological zeros are a prevalent phenomenon observed in both 16S rRNA and SM datasets (Jiang et al., 2021). The abundance distributions of taxa are distorted by these zeros, which can be attributed to three distinct categories: biological, technical, and sampling zeros (Brill et al., 2022). Biological zeros correspond to actual zero abundances of taxa that do not exist in the microbiome samples. In contrast, technical zeros and sampling zeros are non-biological zeros with distinct origins. Technical zeros result from pre-sequencing experimental artifacts, such as DNA degradation during library preparation and inefficient sequence amplification driven by factors like GC content bias (Silverman et al., 2020). On the other hand, sampling zeros stems from limitations in sequencing depths. Addressing the intricacies associated with these zero categories is imperative for robust ML model construction.

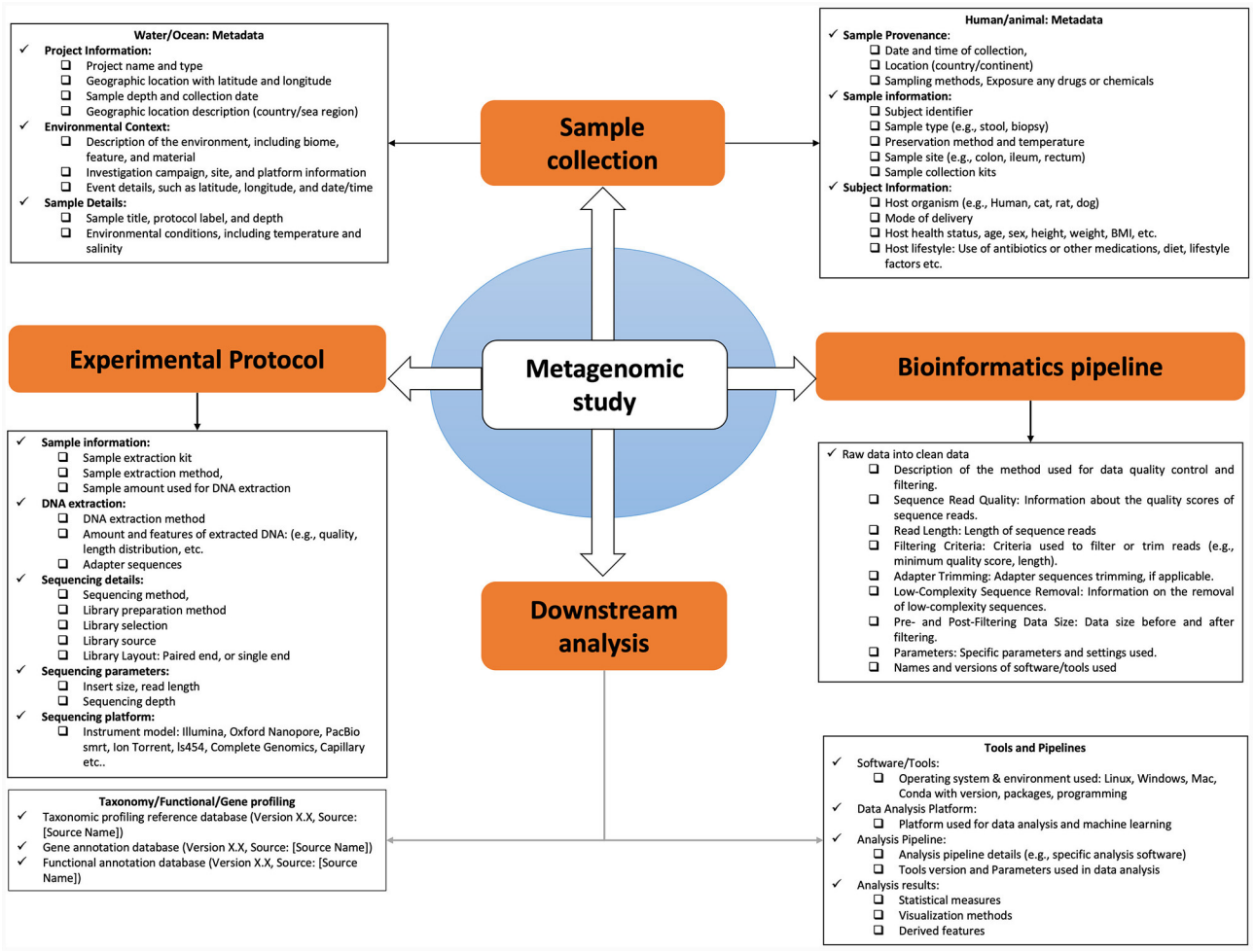

FIGURE 1
Comprehensive framework for metagenomic data documentation and metadata analysis.

In addition, a typical dataset may contain a few hundred training instances but thousands of OTUs/ASVs (i.e., features); this large number of features can greatly challenge the classification accuracy of any method and compound the problem of choosing the important features to focus on.

3.3.2 Biological complexity

The microbiome is variable between individuals and time. This biological variability can make it challenging to identify universal patterns or develop generalizable models (Kodikara et al., 2022; Vinciotti et al., 2023). Also, the taxonomic and functional variability of microbial communities can exhibit significant differences across different environments, making it difficult to establish consistent associations.

3.3.3 Interpretability

Complex machine learning models, such as deep neural networks, might lack interpretability, making it challenging to understand the biological significance of the learned patterns as these models may not be able to generalize to new, unseen data (Linardatos et al., 2021). Interpretable models are often preferred in microbiome research to gain insights into the relationships between microbial features and expected outcomes (Bengtsson-Palme, 2020).

3.3.4 Overfitting and generalization

Due to the high dimensionality of microbiome data, models may be overfitting to noise and contain many spurious correlations in the training data (Walsh et al., 2023). To prevent overfitting, we can use several techniques, such as early stopping, regularization, and data augmentation (Balestriero et al., 2022). Early stopping involves stopping the training process before the model has fully converged, while regularization involves adding a penalty term to the loss function that discourages the model from overfitting (Schmidt, 2023).

Imbalance dataset and cross-validation issues may lead to optimistic estimates of model performance. In this case recommended to use methods like stratified cross-validation

techniques to account for class imbalances in microbiome datasets (Gou et al., 2020; Casimiro-Soriguer et al., 2022; Watson, 2022).

3.3.5 Batch effects and confounding variables

Batch Effects are very common, and this often introduces systematic differences between the measurements of different batches of experimental such as sites/between laboratories, sample preservation protocols, storage conditions, DNA/RNA isolation methods and kits (Ling et al., 2022; Li Y. et al., 2023), sequencing methods can introduce batch effects, which may confound the true biological signals. Combining data from different batches without carefully removing batch effects can give rise to misleading interpretations of taxonomical classificational and ML model interpretations. Therefore, it is necessary to identify and remove the batch effects before proceeding to the downstream analysis and proper normalization and batch correction techniques are essential (Luo et al., 2010) and multiple approaches for batch effect removal have been reported (Alter et al., 2000; Benito et al., 2004; Ling et al., 2022).

Confounding Variables such as diet, medication, and lifestyle can influence the microbiome composition (Li Y. et al., 2023). Failure to account for confounding variables may lead to spurious associations (Al Bander et al., 2020).

Feature Selection and Dimensionality Reduction are used to face the sparsity of microbiome data issue, which makes it challenging to identify important features and patterns through the input data (Lee et al., 2023). Feature selection or dimensionality reduction techniques must be applied during model training.

3.3.6 Model validation and reproducibility

Lack of independent datasets for validation, testing, or failure to reproduce results can undermine the reliability of ML in microbiome analysis (Rojas-Velazquez et al., 2024).

Pammi et al. (2023) reviewed the use of artificial intelligence in integrating "multi-omic" and compared metagenomics analysis approaches, highlighting the effectiveness of statistically equivalent signatures for feature selection and random forest modeling in achieving accurate disease diagnosis and biomarker discovery in colorectal cancer patients.

4 Understanding metadata: data about data

Metadata is "data about data" (Cernava et al., 2022) refers to contextual information associated with metagenomic experimental data offering a comprehensive understanding of the sample's background. In microbiome research, metadata's definition varies based on the type of metagenomic sample under analysis. For instance, metadata for a human gut sample will differ from that of an ocean sample, yet both serve to contextualize the data. Metadata plays a pivotal role in providing context by describing various aspects of the sample, including collection time points, geographical location, biome type, environmental or experimental conditions, and sample pre-processing steps (Leipzig et al., 2021). The structure of metadata can vary by study, but it typically includes features such as chemical data (e.g., pH, salinity), physical data (e.g., temperature, incident light), sample collection time points, host condition (disease/healthy), diet variations, antibiotic exposure, and geographical location (Nassar et al., 2022). Moreover, metadata should encompass information on sampling methods, sample size, and sample preparation techniques. Precise metadata annotation is crucial for detailing the sample source, tissue collection methods, environmental characteristics, and additional specifics like DNA extraction protocols, sequencing library preparation methods, and sequencing depth. In essence, metadata enriches metagenomic data by providing the critical context needed for analysis and interpretation in microbiome result (Nassar et al., 2022).

4.1 The significance of comprehensive metadata in microbiome research

The collection and utilization of various metadata elements in microbiome research are of paramount importance. These elements encompass a wide array of information, from the characterization of the microbiome's natural environment (ecoregion) to the specific host organism (host microbiome) and even human-made environments (engineered microbiome). For a microbiome study, metadata exists at multiple stages along the path from sampling to analysis of omics data as shown (Figure 2). This metadata falls into two main categories: assay metadata, which encompass technical details like machine type, assay date, and reagent kits, and biological metadata, which describe experimental aspects like sample conditions, exposure to drugs, animal housing conditions, or host genetic information. The absence of such information may affect downstream statistical analysis and even qualitative interpretation challenging or impossible.

4.1.1 Sample metadata

Information about provenance and characteristics of the samples: when it was collected (e.g., date and time), where it was collected from (e.g., latitude, longitude, elevation/depth, site name, country, etc.), what kind of sample it was (e.g., soil, seawater, feces/stool), and the properties of the environment during collection (e.g., temperature, salinity, pH) or if sample is clinical then phenotypic condition (e.g., age, sex, disease state/normal) from which the sample was taken and the nature of the sample material itself all contribute valuable context to microbial studies (Wood-Charlson et al., 2020; Vangay et al., 2021).

4.1.2 Experimental metadata

It is subjected to preparation steps for nucleotide sequence analysis or metabolome/metaproteome. Information about experimental preparation of the original sample (Gohl et al., 2016; Vangay et al., 2021). A sample could be split (aliquoted) and processed through multiple preparation methods; therefore, there could be multiple sets of preparation metadata for a single set of samples such as controlled or treated. For DNA sequencing preparation metadata include the type of DNA, extraction

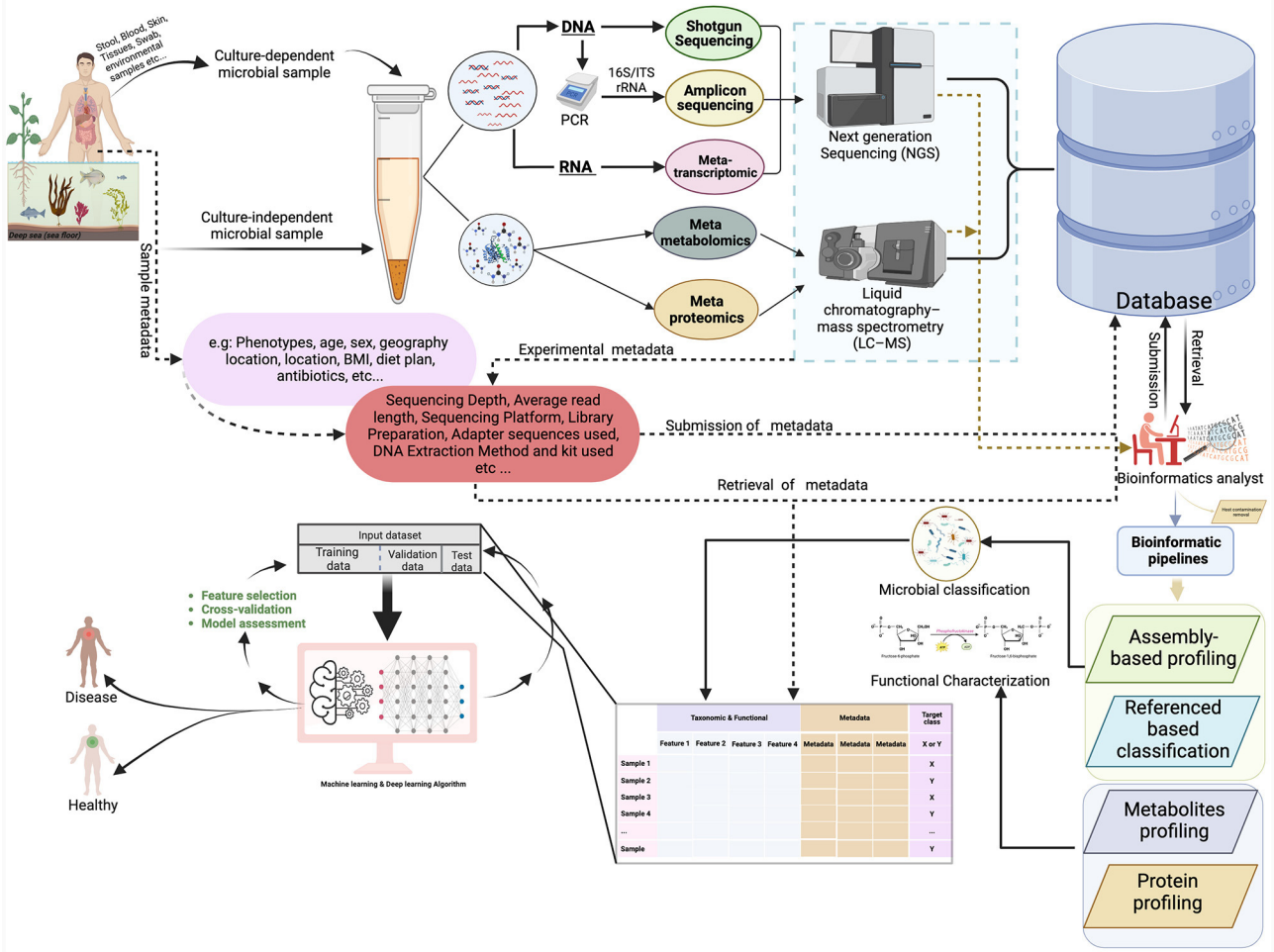

FIGURE 2
This figure provides an overview of the microbiome workflow for studying microbial communities using shotgun sequencing, 16S rRNA gene sequencing, metatranscriptomics, metaproteomics, and metabolomics. The figure illustrates the process of sample collection from various sites and then proceeds through different experimental procedures, bioinformatics pipelines, and ML analyses. The figure was created with https://www.biorender.com/.

protocol, conditions used for sequencing (e.g., primers, library kits, sequencing instrumentation, and parameters), and where the raw sequence data sets are accessible. The information required to properly describe metabolome and metaproteome data are even more complex and workflows profoundly change according to the used platforms and technologies (Rechenberger et al., 2019).

4.1.3 Data pre-processing metadata

Data about the properties and downstream processing of the raw reads data, including software/tools parameters and version. For example, if DNA sequences were generated, this could include the sequence properties (e.g., sequence lengths, sequences per sample, and total base pairs, total percentage of GC content, percentage of sequence duplication), quality control and filtering (e.g., sequencing depth, adapter trimming, quality trimming and filtering, dereplicating, and chimera sequence removal), assembly parameters (e.g., assembly tool, binning tool, and finishing strategy), reference genome used (version and source), gene annotation (e.g., gene calling tool and annotation database), and other processing parameters (Roy et al., 2018).

4.1.4 Feature metadata

Data about features detected in the raw data, rather than about the samples themselves. For example, if amplicon sequencing was performed, feature metadata might include information (e.g., taxonomy, reference genome sequences with version information and source, and sequence identifiers) about the OTUs or ASVs generated in the OTU-picking or denoising algorithm, respectively. If metabolomics analysis was done, feature metadata might include information (e.g., mass spectrometry (MS2) fragments produced or candidates for identification) about the metabolites detected. Obtaining key metadata from sample collection to data analysis would greatly improve reproducibility. For metaproteomics, it might include identified proteins and related pathways.

4.2 FAIR data principles in metagenomics and machine learning

The FAIR Data Principles are a set of guidelines for making data more findable (F), accessible (A), interoperable (I), and reusable (R). These principles are important for both data sharing and machine learning, as they help to ensure that data is discoverable, accessible, and compatible with different machine learning algorithms and tools (Wilkinson et al., 2016). In the context of metagenomics and machine learning, the FAIR Data Principles can be applied to the following: Findability: Metagenomic data should be deposited in public databases, such as the NCBI Sequence Read Archive (SRA) or the European Nucleotide Archive (ENA). These databases provide unique identifiers and searchable metadata for each dataset, making it access the data they need. Accessibility: Metagenomic data should be accessible to researchers using standardized protocols, such as hypertext transfer protocol (HTTP) or file transfer protocol (FTP). This ensures that researchers can access the data regardless of their computing environment. Interoperability: Metagenomic data should be stored in a format that is compatible with different machine learning algorithms and tools. This allows researchers to easily use the data to train and evaluate machine learning models. Reusability: Metagenomic data should be released with clear and accessible data usage licenses. This consents researchers to reuse the data for their own research without having to concern about copyright or other restrictions (ten Hoopen et al., 2017; Vesteghem et al., 2020).

4.3 Metadata standardization: ensuring data accuracy

Despite the critical nature of metadata, metadata collection is often poorly standardized and error prone. Tabular formats (such as Microsoft Excel) continue to be popular options for metadata collection and record-keeping, yet freeform text entry without validation is prone to errors (e.g., misspellings, incorrect data, missing data, and inconsistent values) (Schloss, 2018). These issues can emerge within a single study and are even more likely across multiple studies. For example, with standardized metadata, experimental results from different labs can be grouped together for combined studies with a scope that can extend beyond what can be done from a single lab (Thompson et al., 2020). It also lays the foundation for researchers to quickly find previous experiments of interest to them. Situations may arise where obtaining precise coordinates for certain locations becomes a complex endeavor. These challenges can stem from various factors, including governmental restrictions imposed in specific countries or regions, intellectual property protection, or concerns related to data privacy and property rights. These issues are particularly prominent in datasets associated with potentially sensitive subjects, such as high levels of pathogens or antibiotic resistance genes (Serwecińska, 2020). In some cases, private landowners may be unwilling to disclose the exact locations of their facilities. They might wish to avoid negative associations with their business operations, especially in situations where their facilities are associated with research findings concerning pathogens or antibiotic resistance genes. Moreover, researchers in the industrial sector may be hesitant to make data on specific field sites publicly available. This averseness may be motivated by the fact that these sites are involved in testing new plant cultivars and breeding efforts. The proprietary nature of their work and the competitive landscape could drive this concern. In the realm of biological data and microbiome research, there is a growing awareness of the need to protect the collection coordinates of endangered species, including those listed on conservation red lists (Zhu et al., 2021). This keen concern is rooted in efforts to combat poaching and illegal collection of these species. As a result, there is an ongoing debate regarding how to balance the imperative of protecting these species with the need for scientific data sharing (Levesque, 2017). Lastly, governmental organizations may also have reservations about disclosing precise locations of sites deemed geopolitically important or contaminated. Such disclosures could have difficulties for national security, public safety, or environmental concerns.

4.4 Navigating metadata challenges in metagenome databases

4.4.1 Lack of Metadata

One major limitation of existing public repositories and specialized metagenomic databases (e.g., NCBI, ENA, SRA, MGnify, MG-RAST, NMDC, QIITA) is the often incomplete and inconsistent metadata associated with metagenomic samples. Metadata it is frequently missing or inadequately annotated, making it challenging to perform cross-study comparisons effectively. Lack of Standardization: Metagenome databases suffer from a lack of standardized metadata. Metadata across different studies and databases may use varying terminologies, formats, and ontologies, leading to difficulties in harmonizing and integrating data for meaningful analysis. Difficulty of Metadata Annotation: Manually annotating metadata for metagenomic samples is a labor-intensive and time-consuming process (Kasmanas et al., 2020). While some efforts have been made to standardize metadata using controlled vocabularies and ontologies, these approaches are not always comprehensive or flexible enough to capture the diversity of sample origins, particularly in engineered environments (Cernava et al., 2022). Inefficient Sample Retrieval: Retrieving samples of interest from existing metagenome databases can be incompetent and challenging. The lack of standardized metadata and user-friendly search interfaces makes it difficult for researchers to select relevant samples based on specific criteria, such as host characteristics or environmental factors (Clark et al., 2022). Limited Cross-Study Comparisons: The inconsistent and incomplete metadata in metagenome databases hinder the ability to perform meaningful cross-study comparisons (Nassar et al., 2022). This limitation restricts the potential for meta-analyses and the discovery of patterns or associations that may not be evident in individual studies. Dependence on Manual Annotation: Many existing efforts to improve metadata quality rely heavily on manual annotation, which is not scalable to handle the exponentially increasing volume of metagenomic data. This limitation can lead to delays in data availability and the inability

to keep up with the pace of data generation (Kasmanas et al., 2020). Complexity for Non-Bioinformaticians: Some databases that offer comprehensive metadata are not easily accessible to non-bioinformaticians. For example, metadata stored as *ExpressionSet* objects in R environments can create complexity for researchers who are not proficient in bioinformatics. Limited Support for Specific Environments: Hierarchical ontology relationships may not adequately describe diverse and specific environments, such as engineered environments. Existing controlled vocabularies and ontologies may lack the necessary granularity to capture the full range of sample origins. Inflexible Ontology Relationships: Some databases rely on hierarchical ontology relationships, which can be inflexible and may not accommodate the complexity and diversity of environmental descriptions adequately (Romano et al., 2011). The limitations of existing metagenome databases primarily revolve around the challenges related to metadata quality, standardization, and accessibility. These limitations hinder the full potential of metagenomic data analysis and the ability to perform comprehensive cross-study comparisons and meta-analyses. The development of automated methods for metadata extraction and more user-friendly interfaces is essential to address these limitations and unlock the full value of metagenomic datasets.

4.5 Root causes of annotation errors in public databases

Despite some notable progress in data-sharing policies and practices, accurate and reliable annotation of metagenomic data in public repositories is crucial for dry laboratory researchers and their subsequent applications. In public databases such as NCBI, European Nucleotide Archive (ENA) (Yuan et al., 2023), Sequence Read Archive (SRA) (Katz et al., 2022), MGnify (Richardson et al., 2023), MG-RAST (Meyer et al., 2008), and National Microbiome Data Collaborative (NMDC) (Wood-Charlson et al., 2020), the reliability of annotations heavily relies on the metadata provided by researchers during the submission of sequencing data. However, following are listed several root causes that have been identified that contribute to annotation errors within these databases. (i) User metadata submission errors: Researchers are responsible for submitting metadata that describes the characteristics of their raw/processed sequence, including the name of the model or host organism, pathological conditions (diseased/healthy), biomaterial provider, collection date and time, tissue or samples, developmental stage, and geographical location. However, if researchers make errors or inaccurately assign metadata, it can lead to miss-annotation of sequences and associated data. For example, if a researcher studying soybeans from soybean roots mistakenly assigns the organism's name as *Glycine max* instead of *Glycine soja*, all sequences tied to that metadata will be incorrectly labeled as *Glycine max*, leading to potential misinterpretation and inaccurate analyses (Nassar et al., 2022). (ii) Contamination errors in biological samples: During sample collection and processing, contamination from unintended sources can occur, resulting in the misidentification of organisms or genetic material. If such contamination goes unnoticed or unaddressed, it can lead to incorrect annotations in the public databases. For instance, if a sample intended for sequencing a specific organism becomes contaminated with genetic material from different organisms (usually microbials), the resulting sequences may be incorrectly labeled and associated with the wrong organism in the database (Schnoes et al., 2009). (iii) Bioinformatic tools inaccuracies can lead to erroneous annotations. Different bioinformatics tools and algorithms are utilized to process and annotate sequencing raw data. However, these methods can introduce errors or biases that propagate throughout the database. Imprecise algorithms or incomplete reference databases and versions can result in miss-annotations or missing annotations for specific sequences, further compromising the reliability of the database (Schnoes et al., 2009).

4.6 Challenges and debates in data release protocols: balancing recognition and access

Despite developments in data-sharing policies and practices, many genomic datasets remain restricted even after approval for public release. This conflicts with the terms of funding agencies, which support data dissemination for science and society progress. The lack of clear and comprehensive guidelines for data usage compounds the issue (Schnoes et al., 2009). Public domain data release protocols acknowledge the tension between unrestricted access and data producers desire for recognition through first publication rights. This conflict has led to multiple interpretations, fuelling an ongoing debate about how publicly available data should be used. The pressure to be the first to uncover significant discoveries can lead to data withholding until after publication, hindering broader dissemination (Tenopir et al., 2020). Even after publication, challenges persist, including time constraints in preparing data for sharing, legal and privacy considerations, and concerns about misinterpretation or misuse. Researchers often face difficulty locating the data they need, devoting up to 50-80% of their time to these obstacles (Eckert et al., 2020). Vangay et al. (2021) sustained identifying and addressing the root causes of annotation errors in public databases is essential for maintaining data integrity and ensuring the accuracy of downstream analyses and research applications. By taking into consideration of the factors that contribute to miss-annotations, efforts can be directed toward implementing quality control measures, improving metadata validation processes, enhancing contamination detection methods, and refining computational tools to minimize errors and improve the reliability of public databases.

4.7 Privacy concerns in metagenomics: uncovering personal information

The availability of open-access metagenomic datasets provides a valuable resource for studying health- and disease-associated signatures of microbial communities. However, an ongoing debate within microbiome research revolves around addressing privacy concerns to protection of personal information (Guccione et al., 2023). Franzosa et al. (2015) investigated the human microbiome by utilizing metagenomic codes. These metagenomic codes were

designed to identify individuals based on specific microbial taxa or genes that are distinct and consistent across different body sites. Combining insights from microbial ecology and computer science, researchers discovered that it is possible to distinguish individuals from groups of hundreds based solely on their microbiomes, with over 80% accuracy even after a year, particularly notable in the case of the gut microbiome (Franzosa et al., 2015). While this underscores the fascinating individuality of human microbial signatures, it also raises significant privacy concerns for participants in microbiome research projects, highlighting the need for robust privacy safeguards in the handling of such health data (Chuong et al., 2017).

In Japan Tomofuji et al. (2023), uncovered a potential concern about metagenomic data obtained from human fecal samples. Specifically, they achieved a remarkable 93.7% accuracy in predicting biological gender by analyzing the read depth of non-pseudo-autosomal regions of sex chromosomes. This report has significant effects, especially in the context of human microbiome studies, where it can help rectify mislabelled samples and contribute to the field of human genetics. However, the accurate prediction of genetic sex bearing privacy concerns, particularly for individuals who may not wish to disclose this information. This concern is especially relevant to transgender individuals, who may face varying degrees of legal protection worldwide. To address these privacy issues, methods for removing human DNA reads from metagenomic data were developed during the National Institutes of Health's Human Microbiome Project (Wagner et al., 2016). It is worth noting that sex prediction based on DNA extracted from fecal samples had previously been predominantly conducted for wild animals using PCR amplification of marker genes (Guccione et al., 2023).

Furthermore, another study demonstrates sensitivity in identifying matched genotype data and accurately predicted ancestral backgrounds in samples. Ancestral backgrounds were defined as American, European, African, East Asian, and South Asian (Tomofuji et al., 2023). These findings highlight the importance of considering the ethical implications and privacy concerns when utilizing open-source microbiome data.

4.8 Improving metadata quality in microbiome research

Metadata is essential for the interpretation, reproducibility, and reuse of microbiome data. However, metadata quality is often variable, which can hinder research progress. To improve metadata quality, we can consider employing Manual and Automated curation. The first one is the most accurate approach, but it is also the most time-consuming and expensive. The latter employs ML approaches and other techniques to extract metadata from raw sample data. It is the most scalable approach, but it can be less accurate than the first one. One example of an automated curation approach is the ML framework developed by Nassar et al. (2022) that automatically extracts important metadata from a vast number of metagenomics studies found in the Europe PMC literature repository. This integration allows for the continual enhancement of current metadata in ENA and MGnify metagenomics studies by sourcing information from research articles. As a result, the MGnify database now displays these annotations, providing information on metadata like health status, disease conditions, geographic locations, and sequencing methods. Gonçalves and Musen (2019) study shed light on the varying quality of metadata available in prominent databases such as NCBI's BioSample and the European Bioinformatics Institute's BioSamples. One of the contributing factors to this variability is the infrequent use of controlled vocabularies during the metadata submission process. Additionally, the allowance for the creation of user-defined attributes has resulted in a proliferation of heterogeneity within the metadata landscape. This diversity often poses challenges for researchers, making it difficult to harness the full potential of information within a specific dataset or across multiple datasets (Gonçalves and Musen, 2019).

Klie et al. (2021) aimed to enhance the metadata coverage of SRA BioSample entries using deep learning-based named entity recognition (NER). The study achieved high prediction accuracies for certain metadata categories when extracting information from sample titles (TITLEs). It is worthy to note, they processed all the available BioSample up to May 2018, and Genus/Specie and strains generally refers to processed samples. However, lower accuracies and the absence of predictions for other metadata categories underscored existing issues with the current metadata annotations in BioSample. These findings demonstrate the effectiveness of recurrent neural networks for NER-based metadata prediction and suggest the potential of such models to expand metadata coverage in BioSample, reducing the reliance on manual curation (Klie et al., 2021). Below some additional thoughts on the future directions of machine learning for metadata retrieval in metagenomics. Firstly, ML algorithms (De et al., 2022; Nassar et al., 2022; Raghavendra Nayaka and Ranjan, 2023) could be developed to extract metadata from scientific literature, abstracts, and environmental monitoring data. This would allow researchers to extract more reliable metadata with less effort. Secondly, ML algorithms could be used to develop new metadata standards that are tailored to specific research questions. This would help to ensure that metadata is collected in a way that is most useful for the scientific community.

5 Metadata exploitation for robust ML models

During development of ML-based classifiers, the incorporation of metadata emerges as a crucial factor for accurate predictions and robust model development. A series of studies mark the significance of considering host associated metadata elements, ranging from geographical location to dietary habits and perinatal factors, host genetic factor (Lopera-Maya et al., 2022; New et al., 2022) shedding light on microbial compositions. Below we have highlighted examples of why researchers should consider host associated factors to train supervised predictive ML model for better generalization capability on the unseen dataset.

5.1 Changes in the gut microbiome: from infancy to adulthood and beyond

Studies have shown that the gut microbiome of infants undergoes significant changes during the first 3 years of life, with differences observed between populations and influenced by factors, such as delivery mode. Yatsunenko et al. (2012) compared fecal samples from Amerindians in Venezuela and residents of U.S. metropolitan areas, finding that the gut microbiome exhibited similar functional maturation patterns across the initial 3 years of life across populations. Palmer et al. (2007) also, observed substantial variation in the composition of gut bacteria in infants during the first year of life, with reduced variation within twin pairs and decreased variation with age. Orrhage and Nord (1999) emphasized the impact of delivery mode on the infant microbiome (Fanaro et al., 2003; Penders et al., 2006; Yatsunenko et al., 2012). Studies have shown that cesarean section (CS) results in a different microbiota compared to vaginal delivery (VD) (Bennet and Nord, 1987; Hällström et al., 2004; Elovitz et al., 2019). Cheng et al. (2022) emphasized the importance of further investigation to comprehensively delineate the multifaceted factors shaping microbiota dynamics during maternal-neonatal interactions, extending beyond traditional perinatal considerations.

Gudnadottir et al. (2022) employed the network-meta-analysis method and revealed that the microbiome demonstrates predictive potential for preterm birth and emphasizes the significance of specific microbial compositions in the vaginal microbiome as potential indicators for the likelihood of preterm birth.

Odamaki et al. (2016) and Meng et al. (2022) delved into the alterations in gut microbiota across different age groups and their associations with gut inflammation, particularly during the sexual maturity stage in healthy individuals. As individuals progress in age, there is a significant increase in the relative abundance of Firmicutes, accompanied by a concurrent decrease in the relative abundance of Bacteroides. The study further identified a positive correlation between body weight and the Firmicutes:Bacteroides ratio, shedding light on potential associations between microbiota composition and physiological parameters.

In addition to the age-related patterns identified in gut microbiota, the investigation also observed variations in microbial compositions across different body sites, including the vagina, skin, oral cavity, and respiratory tract. Detailed information on these variations is available at (Hou et al., 2022).

Kim et al. (2020) outlined that gender constitutes a significant variable shaping the composition of the gut microbiota. Furthermore, an investigation involving male and female germ-free C57BL/6J mice, Wang et al. (2016) and Zhao et al. (2019) revealed distinctive microbial preferences in the intestines of male and female mice. Despite these findings highlighting the relevance of gender in microbiota dynamics, a comprehensive understanding of this association remains elusive.

Cheng et al. (2022) emphasized geographical location as a paramount variable influencing the overall structure of maternal and neonatal microbiota, especially evident in two distinct populations from Asia and Europe. Elsherbiny et al. (2022) in Egypt elucidated the impact of geographical location on the gut microbiota in children with Type-1 Diabetes Mellitus, revealing differences in alpha diversity between controls and diabetic groups.

The Chinese healthy gut project (Ren et al., 2023), outlined on the correlation between gut microbiota and various dietary and lifestyle factors among healthy individuals in China. Notably, lifestyle phenotypes, including sleep procrastination, negative mood, and drinking habits, exhibited substantial influence on gut microbiota composition, with these factors showing the largest effect sizes.

5.1.1 Role of diets

Noble et al. (2021) investigated the impact of sugar-sweetened beverage consumption during adolescence on the gut microbiome, which was linked to alterations in hippocampal function, as already demonstrated by David et al. (2014). Vujkovic-Cvijin et al. (2020) identified unexpected sources of gut microbiota variance, including alcohol consumption frequency and bowel movement quality. Singh and Mittal (2020) and Gacesa et al. (2022) comprehensively reviewed the profound impact of diet on the pathophysiology of mental disorders, highlighting its crucial role in shaping mental health outcomes. Ren et al. (2023) delved into the effects of dietary factors on the structure of the gut microbiota, while Manor et al. (2020) highlighted the composition-specific nature of host-microbe associations, providing insights into the intricate connections between microbiome composition, clinical markers, and lifestyle factors.

5.1.2 Medication and antibiotic exposure

BMI and insulin level: Bäckhed et al. (2004) has illuminated a substantial connection between the gut microbiota and the regulation of body weight. Also, Ridaura et al. (2013) demonstrated weight gain in germ-free mice following gut microbiota transplants from individuals with obesity. These findings highlight the intricate relationship between gut microbiota composition and its role in regulating body weight. Gupta et al. (2020) emphasized the use of BMI scores to classify underweight, overweight, or obese individuals. Evans et al. (2014) shows that physical activity could shifts in the composition of the gut microbiome in animal models (Kang et al., 2014) but the robustness of this association at population-level remains uncertain. Concerning antibiotics, two cohort studies, utilizing a difference-in-differences approach, demonstrated that antibiotic exposure in infancy altered the relative abundance of off-target species and antibiotic resistance genes (Ramirez et al., 2020; Ribeiro et al., 2020; Lebeaux et al., 2022; Patangia et al., 2022).

In the realm of machine learning challenges, MetAML, an ML-based classifier, revealed variable results between prediction tasks, cautioning against potential overestimation of disease prediction due to confounding factors like active antibiotic treatment (Pasolli et al., 2016).

Abdul Rahman et al. (2023) developed supervised and unsupervised ML models to predict colorectal cancer using global dietary data, encompassing both younger and older adults from seven major countries (Canada, India, Italy, South Korea, Mexico, Sweden, and the United States) and diverse sociodemographic

factors. Su et al. (2022) show that the limitation of using a combined public dataset did not specify the co-morbidities and antibiotics; thus, model performance depends on the exclusion of these metadata.

5.2 Future direction

Previous studies show that the composition of the human gut microbiome varies significantly among individuals. This variability suggests that incorporating metadata, including confounding factors and dietary information, into ML models is highly beneficial. Figure 1 illustrates a potential approach for integrating metadata information alongside microbiome features. This integrated analysis can lead to novel research questions, refine sample and feature selection, and improve the robustness of predictive statistical and ML models, e.g. develop ML model to predict the phenotype of a host organism. The interplay between ML and metadata is crucial for effective model implementation. Incorporating host metadata into microbiota studies can ensure that groups are well-matched, enhancing the reliability and reproducibility of studies investigating diseases or phenotypes associated with distinct pathological, physiological, lifestyle, or dietary traits.

6 Conclusion

Integrated metadata analysis is essential for maximizing the potential of ML and other advanced techniques in microbiome research. While recent advances in metagenomics, metabolomics, and metaproteomics have generated a wealth of publicly available data, its comprehensive utilization is hindered by several challenges, including the need for substantial time investments, accessibility issues with metadata, computational resource requirements, and the need for specialized bioinformatic expertise. As widely discussed in the previous sections, the inclusion of metadata information in ML models development is crucial to avoid erroneous outcomes. Metadata become essential to attenuate the negative impact of confounding factors, both technical and biological. Moreover, either when multi-omics data integration is considered, the inclusion of clinical metadata about enrolled subjects emerge as a source of knowledge leveraging the models accuracy, as demonstrated by Leung et al. (2022). Indeed, this review highlights the importance of integrated metadata analysis in microbiome research. By combining microbial data with sample-specific information, researchers can gain a deeper understanding of the microbial communities that inhabit the human body and their role in health and disease. This knowledge can be used to develop new diagnostic and therapeutic strategies. However, integrated metadata analysis is also challenging due to issues related to data management, computational demands, integration approaches, and the selection of appropriate analysis tools. To fully leverage the potential of integrated metadata analysis in microbiome research, it is essential to address these challenges through the development of new tools and resources, as well as the training of researchers in the necessary skills.

Author contributions

BK: Conceptualization, Writing—original draft, Writing—review & editing. EL: Writing—review & editing. BF: Writing—review & editing, Conceptualization, Supervision, Writing—original draft. GP: Conceptualization, Funding acquisition, Supervision, Writing—review & editing.

Funding

The author (s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by ELIXIR-IT, the Italian Node of the European research infrastructure for life-science data, CUP B53C22000690005. Moreover, this research was co-funded by the Complementary National Plan PNC-I.1 “Research initiatives for innovative technologies and pathways in the health and welfare sector” D.D. 931 of 06/06/20-2, DARE-DigitAl lifelong pRevEntion initiative, code PNC0000002, CUP B53C22006420001.

Acknowledgments

BK is a PhD student within the European School of Molecular Medicine (SEMM). We also thank Maria Rosa Mirizzi and Luigi Boccaccio for technical administrative assistance. LLM service chatGPT was used to grammatically check the manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher’s note

References

Abdul Rahman, H., Ottom, M. A., and Dinov, I. D. (2023). Machine learning-based colorectal cancer prediction using global dietary data. *BMC Cancer* 23, 144. doi: 10.1186/s12885-023-10587-x

Al Bander, Z., Nitert, M. D., Mousa, A., and Naderpoor, N. (2020). The gut microbiota and inflammation: an overview. *Int. J. Environ. Res. Public Health* 17, 7618. doi: 10.3390/ijerph17207618

Alter, O., Brown, P. O., and Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. U.S.A.* 97, 10101–10106. doi: 10.1073/pnas.97.18.10101

Asshauer, K. P., Wemheuer, B., Daniel, R., and Meinicke, P. (2015). Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data. *Bioinformatics* 31, 2882–2884. doi: 10.1093/bioinformatics/btv287

Bäckhed, F., Ding, H., Wang, T., Hooper, L. V., Koh, G. Y., Nagy, A., et al. (2004). The gut microbiota as an environmental factor that regulates fat storage. *Proc. Nat. Acad. Sci.* 101, 15718–15723. doi: 10.1073/pnas.0407076101

Bakir-Gungor, B., Bulut, O., Jabeer, A., Nalbantoglu, O. U., and Yousef, M. (2021). Discovering potential taxonomic biomarkers of type 2 diabetes from human gut microbiota via different feature selection methods. *Front. Microbiol.* 12, 628426. doi: 10.3389/fmicb.2021.628426

Bakir-Gungor, B., Hacilar, H., Jabeer, A., Nalbantoglu, O. U., Aran, O., and Yousef, M. (2022). Inflammatory bowel disease biomarkers of human gut microbiota selected via different feature selection methods. *PeerJ* 10, e13205. doi: 10.7717/peerj.13205

Balestriero, R., Bottou, L., and LeCun, Y. (2022). The effects of regularization and data augmentation are class dependent. *arXiv [Preprint]. arXiv:*2204.03632 doi: 10.48550/arXiv.2204.0363

Bao, W., Kojima, K. K., and Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* 6, 11. doi: 10.1186/s13100-015-0041-9

Bashiardes, S., Zilberman-Schapira, G., and Elinav, E. (2016). Use of metatranscriptomics in microbiome research. *Bioinform. Biol. Insights* 10, 19–25. doi: 10.4137/BBI.S34610

Bengtsson-Palme, J. (2020). Microbial model communities: To understand complexity, harness the power of simplicity. *Comput. Struct. Biotechnol. J.* 18, 3987–4001. doi: 10.1016/j.csbj.2020.11.043

Benito, M., Parker, J., Du, Q., Wu, J., Xiang, D., Perou, C. M., et al. (2004). Adjustment of systematic microarray data biases. *Bioinformatics* 20, 105–114. doi: 10.1093/bioinformatics/btg385

Bennet, R., and Nord, C. E. (1987). Development of the faecal anaerobic microflora after caesarean section and treatment with antibiotics in newborn infants. *Infection* 15, 332–336. doi: 10.1007/BF01647733

Berden, P., Wiederkehr, R. S., Lagae, L., Michiels, J., Stakenborg, T., Fauvart, M., et al. (2022). Amplification efficiency and template accessibility as distinct causes of rain in digital PCR: Monte Carlo modeling and experimental validation. *Anal. Chem.* 94, 15781–15789. doi: 10.1021/acs.analchem.2c03534

Bhattacharya, C., Tierney, B. T., Ryon, K. A., Bhattacharyya, M., Hastings, J. J. A., Basu, S., et al. (2022). Supervised machine learning enables geospatial microbial provenance. *Genes* 13, 1914. doi: 10.3390/genes13101914

Bikel, S., Valdez-Lara, A., Cornejo-Granados, F., Rico, K., Canizales-Quinteros, S., Soberón, X., et al. (2015). Combining metagenomics, metatranscriptomics and viromics to explore novel microbial interactions: towards a systems-level understanding of human microbiome. *Comput. Struct. Biotechnol. J.* 13, 390–401. doi: 10.1016/j.csbj.2015.06.001

Bingol, K. (2018). Recent advances in targeted and untargeted metabolomics by NMR and MS/NMR methods. *High Throughput* 7, 9. doi: 10.3390/ht7020009

Brill, B., Amir, A., and Heller, R. (2022). Testing for differential abundance in compositional counts data, with application to microbiome studies. *Ann. Appl. Stat.* 16, 2648–2671. doi: 10.1214/22-AOAS1607

Casimiro-Soriguer, C. S., Loucera, C., Peña-Chilet, M., and Dopazo, J. (2022). Towards a metagenomics machine learning interpretable model for understanding the transition from adenoma to colorectal cancer. *Sci. Rep.* 12, 450. doi: 10.1038/s41598-021-04182-y

Cernava, T., Rybakova, D., Buscot, F., Clavel, T., McHardy, A. C., Meyer, F., et al. (2022). Metadata harmonization–Standards are the key for a better usage of omics data for integrative microbiome analysis. *Environ. Microb.* 17, 33. doi: 10.1186/s40793-022-00425-1

Cheng, Y., Selma-Royo, M., Cao, X., Calatayud, M., Qi, Q., Zhou, J., et al. (2022). Influence of geographical location on maternal-infant microbiota: study in two populations from Asia and Europe. *Front. Cell. Infect. Microb.* 11, 663513. doi: 10.3389/fcimb.2021.663513

Chuong, K. H., Hwang, D. M., Tullis, D. E., Waters, V. J., Yau, Y. C. W., Guttman, D. S., et al. (2017). Navigating social and ethical challenges of biobanking for human microbiome research. *BMC Med. Ethics* 18, 1. doi: 10.1186/s12910-016-0160-y

Clark, S., Bleken, F. L., Stier, S., Flores, E., Andersen, C. W., Marcinek, M., et al. (2022). Toward a unified description of battery data. *Adv. Energy Mat.* 12, 2102702. doi: 10.1002/aenm.202102702

David, L. A., Maurice, C. F., Carmody, R. N., Gootenberg, D. B., Button, J. E., Wolfe, B. E., et al. (2014). Diet rapidly and reproducibly alters the human gut microbiome. *Nature* 505, 559–563. doi: 10.1038/nature12820

De, S., Moss, H., Johnson, J., Li, J., Pereira, H., and Jabbari, S. (2022). Engineering a machine learning pipeline for automating metadata extraction from longitudinal survey questionnaires. *IASSIST Quart.* 46. doi: 10.29173/iq1023

Desiere, F., Deutsch, E. W., King, N. L., Nesvizhskii, A. I., Mallick, P., Eng, J., et al. (2006). The PeptideAtlas project. *Nucleic Acids Res.* 34, D655–D658. doi: 10.1093/nar/gkj040

Deutsch, E. W., Bandeira, N., Perez-Riverol, Y., Sharma, V., Carver, J. J., Mendoza, L., et al. (2022). The ProteomeXchange consortium at 10 years: 2023 update. *Nucleic Acids Res.* 51, D1539–D1548. doi: 10.1093/nar/gkac1040

Deutsch, E. W., Csordas, A., Sun, Z., Jarnuczak, A., Perez-Riverol, Y., Ternent, T., et al. (2017). The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition. *Nucleic Acids Res.* 45, D1100–D1106. doi: 10.1093/nar/gkw936

Douglas, G. M., Maffei, V. J., Zaneveld, J. R., Yurgel, S. N., Brown, J. R., Taylor, C. M., et al. (2020). PICRUSt2 for prediction of metagenome functions. *Nat. Biotechnol.* 38, 685–688. doi: 10.1038/s41587-020-0548-6

Duvallet, C., Gibbons, S. M., Gurry, T., Irizarry, R. A., and Alm, E. J. (2017). Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nat. Commun.* 8, 1784. doi: 10.1038/s41467-017-01973-8

Eckert, E. M., Cesare, A. D., Fontaneto, D., Berendonk, T. U., Bürgmann, H., Cytryn, E., et al. (2020). Every fifth published metagenome is not available to science. *PLoS Biol.* 18, e3000698. doi: 10.1371/journal.pbio.3000698

Elovitz, M. A., Gajer, P., Riis, V., Brown, A. G., Humphrys, M. S., Holm, J. B., et al. (2019). Cervicovaginal microbiota and local immune response modulate the risk of spontaneous preterm delivery. *Nat. Commun.* 10, 1305. doi: 10.1038/s41467-019-09285-9

Elsherbiny, N. M., Ramadan, M., Faddan, N. H. A., Hassan, E. A., Ali, M. E., El-Rehim, A. S. E.-D. A., et al. (2022). Impact of geographical location on the gut microbiota profile in Egyptian children with type 1 diabetes mellitus: a pilot study. *IJGM* 15, 6173–6187. doi: 10.2147/IJGM.S361169

Erickson, A. R., Cantarel, B. L., Lamendella, R., Darzi, Y., Mongodin, E. F., Pan, C., et al. (2012). Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn's disease. *PLoS ONE* 7, e49138. doi: 10.1371/journal.pone.0049138

Evans, C. C., LePard, K. J., Kwak, J. W., Stancukas, M. C., Laskowski, S., Dougherty, J., et al. (2014). Exercise prevents weight gain and alters the gut microbiota in a mouse model of high fat diet-induced obesity. *PLoS ONE* 9, e92193. doi: 10.1371/journal.pone.0092193

Fanaro, S., Chierici, R., Guerrini, P., and Vigi, V. (2003). Intestinal microflora in early infancy: composition and development. *Acta Paediatr. Suppl.* 91, 48–55. doi: 10.1111/j.1651-2227.2003.tb00646.x

Farrah, T., Deutsch, E. W., Kreisberg, R., Sun, Z., Campbell, D. S., Mendoza, L., et al. (2012). PASSEL: the PeptideAtlas SRMexperiment library. *Proteomics* 12, 1170–1175. doi: 10.1002/pmic.201100515

Ferry-Dumazet, H., Gil, L., Deborde, C., Moing, A., Bernillon, S., Rolin, D., et al. (2011). MeRy-B: a web knowledgebase for the storage, visualization, analysis and annotation of plant NMR metabolomic profiles. *BMC Plant Biol.* 11, 104. doi: 10.1186/1471-2229-11-104

Franzosa, E. A., Huang, K., Meadow, J. F., Gevers, D., Lemon, K. P., Bohannan, B. J. M., et al. (2015). Identifying personal microbiomes using metagenomic codes. *Proc. Nat. Acad. Sci.* 112, E2930–E2938. doi: 10.1073/pnas.1423854112

Gacesa, R., Kurilshikov, A., Vich Vila, A., Sinha, T., Klaassen, M. A. Y., Bolte, L. A., et al. (2022). Environmental factors shaping the gut microbiome in a Dutch population. *Nature* 604, 732–739. doi: 10.1038/s41586-022-04567-7

Gilbert, J., Blaser, M. J., Caporaso, J. G., Jansson, J., Lynch, S. V., and Knight, R. (2018). Current understanding of the human microbiome. *Nat. Med.* 24, 392–400. doi: 10.1038/nm.4517

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., and Egozcue, J. J. (2017). Microbiome Datasets Are Compositional: And This Is Not Optional. *Frontiers in Microbiology* 8. Available at: https://www.frontiersin.org/articles/10.3389/fmicb.2017.02224 (accessed January 2, 2024).

Gohl, D. M., Vangay, P., Garbe, J., MacLean, A., Hauge, A., Becker, A., et al. (2016). Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. *Nat. Biotechnol.* 34, 942–949. doi: 10.1038/nbt.3601

Gonçalves, R. S., and Musen, M. A. (2019). The variable quality of metadata about biological samples used in biomedical experiments. *Sci Data* 6, 190021. doi: 10.1038/sdata.2019.21

Gou, W., Ling, C., He, Y., Jiang, Z., Fu, Y., Xu, F., et al. (2020). Interpretable machine learning framework reveals robust gut microbiome features associated with type 2 diabetes. *Diabetes Care* 44, 358–366. doi: 10.2337/dc20-1536

Guccione, C., McDonald, D., Fielding-Miller, R., Curtius, K., and Knight, R. (2023). You are what you excrete. *Nat Microbiol* 8, 1002–1003. doi: 10.1038/s41564-023-01395-x

Gudnadottir, U., Debelius, J. W., Du, J., Hugerth, L. W., Danielsson, H., Schuppe-Koistinen, I., et al. (2022). The vaginal microbiome and the risk of preterm birth: a systematic review and network meta-analysis. *Sci. Rep.* 12, 7926. doi: 10.1038/s41598-022-12007-9

Gupta, V. K., Kim, M., Bakshi, U., Cunningham, K. Y., Davis, J. M., Lazaridis, K. N., et al. (2020). A predictive index for health status using species-level gut microbiome profiling. *Nat. Commun.* 11, 4635. doi: 10.1038/s41467-020-18476-8

Haas, B. J., Gevers, D., Earl, A. M., Feldgarden, M., Ward, D. V., Giannoukos, G., et al. (2011). Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res.* 21, 494–504. doi: 10.1101/gr.112730.110

Hällström, M., Eerola, E., Vuento, R., Janas, M., and Tammela, O. (2004). Effects of mode of delivery and necrotising enterocolitis on the intestinal microflora in preterm infants. *Eur. J. Clin. Microbiol. Infect. Dis.* 23, 463–470. doi: 10.1007/s10096-004-1146-0

Haug, K., Salek, R. M., and Steinbeck, C. (2017). Global open data management in metabolomics. *Curr. Opin. Chem. Biol.* 36, 58–63. doi: 10.1016/j.cbpa.2016.12.024

Hernández Medina, R., Kutuzova, S., Nielsen, K. N., Johansen, J., Hansen, L. H., Nielsen, M., et al. (2022). Machine learning and deep learning applications in microbiome research. *ISME COMMUN.* 2, 1–7. doi: 10.1038/s43705-022-00182-9

Holoch, D., and Moazed, D. (2015). RNA-mediated epigenetic regulation of gene expression. *Nat. Rev. Genet.* 16, 71–84. doi: 10.1038/nrg3863

Hou, K., Wu, Z.-X., Chen, X.-Y., Wang, J.-Q., Zhang, D., Xiao, C., et al. (2022). Microbiota in health and diseases. *Sig Transduct Target Ther* 7, 1–28. doi: 10.1038/s41392-022-00974-4

Huttenhower, C., Finn, R. D., and McHardy, A. C. (2023). Challenges and opportunities in sharing microbiome data and analyses. *Nat Microbiol* 8, 1960–1970. doi: 10.1038/s41564-023-01484-x

Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J. H., Chinwalla, A. T., et al. (2012). Structure, function and diversity of the healthy human microbiome. *Nature* 486, 207–214. doi: 10.1038/nature11234

Jiang, R., Li, W. V., and Li, J. J. (2021). mbImpute: an accurate and robust imputation method for microbiome data. *Genome Biol.* 22, 192. doi: 10.1186/s13059-021-02400-4

Johnson, J. S., Spakowicz, D. J., Hong, B.-Y., Petersen, L. M., Demkowicz, P., Chen, L., et al. (2019). Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nat. Commun.* 10, 5029. doi: 10.1038/s41467-019-13036-1

Kang, S. S., Jeraldo, P. R., Kurti, A., Miller, M. E. B., Cook, M. D., Whitlock, K., et al. (2014). Diet and exercise orthogonally alter the gut microbiome and reveal independent associations with anxiety and cognition. *Mol. Neurodegener.* 9, 36. doi: 10.1186/1750-1326-9-36

Kasmanas, J. C., Bartholomäus, A., Corrêa, F. B., Tal, T., Jehmlich, N., Herberth, G., et al. (2020). HumanMetagenomeDB: a public repository of curated and standardized metadata for human metagenomes. *Nucleic Acids Res.* 49, D743–D750. doi: 10.1093/nar/gkaa1031

Katz, K., Shutov, O., Lapoint, R., Kimelman, M., Brister, J. R., and O'Sullivan, C. (2022). The sequence read archive: a decade more of explosive growth. *Nucleic Acids Res.* 50, D387–D390. doi: 10.1093/nar/gkab1053

Kim, Y. S., Unno, T., Kim, B.-Y., and Park, M.-S. (2020). Sex differences in gut microbiota. *World J. Mens. Health* 38, 48–60. doi: 10.5534/wjmh.190009

Klie, A., Tsui, B. Y., Mollah, S., Skola, D., Dow, M., Hsu, C.-N., et al. (2021). Increasing metadata coverage of SRA BioSample entries using deep learning–based named entity recognition. *Database* 2021, baab021. doi: 10.1093/database/baab021

Kodikara, S., Ellul, S., and and, L.ê, Cao, K.-A. (2022). Statistical challenges in longitudinal microbiome data analysis. *Briefings Bioinform.* 23, bbac273. doi: 10.1093/bib/bbac273

La Reau, A. J., Strom, N. B., Filvaroff, E., Mavrommatis, K., Ward, T. L., and Knights, D. (2023). Shallow shotgun sequencing reduces technical variation in microbiome analysis. *Sci. Rep.* 13, 7668. doi: 10.1038/s41598-023-33489-1

Lam, T. J., and Ye, Y. (2022). Meta-analysis of microbiome association networks reveal patterns of dysbiosis in diseased microbiomes. *Sci. Rep.* 12, 17482. doi: 10.1038/s41598-022-22541-1

Langille, M. G., Zaneveld, J., Caporaso, J. G., McDonald, D., Knights, D., Reyes, J. A., et al. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat. Biotechnol.* 31, 814–821. doi: 10.1038/nbt.2676

Lassalle, F., Spagnoletti, M., Fumagalli, M., Shaw, L., Dyble, M., Walker, C., et al. (2018). Oral microbiomes from hunter-gatherers and traditional farmers reveal shifts in commensal balance and pathogen load linked to diet. *Mol. Ecol.* 27, 182–195. doi: 10.1111/mec.14435

Lebeaux, R. M., Madan, J. C., Nguyen, Q. P., Coker, M. O., Dade, E. F., Moroishi, Y., et al. (2022). Impact of antibiotics on off-target infant gut microbiota and resistance genes in cohort studies. *Pediatr. Res.* 92, 1757–1766. doi: 10.1038/s41390-022-02104-w

Lee, P. Y., Chin, S.-F., Neoh, H., and Jamal, R. (2017). Metaproteomic analysis of human gut microbiota: where are we heading? *J. Biomed. Sci.* 24, 36. doi: 10.1186/s12929-017-0342-z

Lee, Y., Cappellato, M., and Di Camillo, B. (2023). Machine learning–based feature selection to search stable microbial biomarkers: application to inflammatory bowel disease. *GigaScience* 12, giad083. doi: 10.1093/gigascience/giad083

Leipzig, J., Nüst, D., Hoyt, C. T., Ram, K., and Greenberg, J. (2021). The role of metadata in reproducible computational research. *Patterns* 2, 100322. doi: 10.1016/j.patter.2021.100322

Leung, H., Long, X., Ni, Y., Qian, L., Nychas, E., Siliceo, S. L., et al. (2022). Risk assessment with gut microbiome and metabolite markers in NAFLD development. *Sci Transl Med* 14, eabk0855. doi: 10.1126/scitranslmed.abk0855

Levesque, R. J. R. (2017). Data sharing mandates, developmental science, and responsibly supporting authors. *J. Youth Adolesc.* 46, 2401–2406. doi: 10.1007/s10964-017-0741-1

Li, L., Wang, T., Ning, Z., Zhang, X., Butcher, J., Serrana, J. M., et al. (2023). Revealing proteome-level functional redundancy in the human gut microbiome using ultra-deep metaproteomics. *Nat. Commun.* 14, 3428. doi: 10.1038/s41467-023-39149-2

Li, L., Yang, K., Li, C., Zhang, H., Yu, H., Chen, K., et al. (2022). Metagenomic shotgun sequencing and metabolomic profiling identify specific human gut microbiota associated with diabetic retinopathy in patients with type 2 diabetes. *Front. Immunol.* 13, 943325. doi: 10.3389/fimmu.2022.943325

Li, Y., Xie, G., Zha, Y., and Ning, K. (2023). GAN-GMHI: a generative adversarial network with high discriminative power for microbiome-based disease prediction. *J. Genet. Genomics* 50, 1026–1028. doi: 10.1016/j.jgg.2023.03.009

Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. (2021). Explainable AI: a review of machine learning interpretability methods. *Entropy* 23, 18. doi: 10.3390/e23010018

Liñares-Blanco, J., Fernandez-Lozano, C., Seoane, J. A., and López-Campos, G. (2022). Machine learning based microbiome signature to predict inflammatory bowel disease subtypes. *Front. Microbiol.* 13, 872671. doi: 10.3389/fmicb.2022.872671

Ling, W., Lu, J., Zhao, N., Lulla, A., Plantinga, A. M., Fu, W., et al. (2022). Batch effects removal for microbiome data via conditional quantile regression. *Nat. Commun.* 13, 5418. doi: 10.1038/s41467-022-33071-9

Lloyd-Price, J., Mahurkar, A., Rahnavard, G., Crabtree, J., Orvis, J., Hall, A. B., et al. (2017). Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature* 550, 61–66. doi: 10.1038/nature23889

Logares, R., Haverkamp, T. H. A., Kumar, S., Lanzén, A., Nederbragt, A. J., Quince, C., et al. (2012). Environmental microbiology through the lens of high-throughput DNA sequencing: synopsis of current platforms and bioinformatics approaches. *J. Microbiol. Methods* 91, 106–113. doi: 10.1016/j.mimet.2012.07.017

Long, S., Yang, Y., Shen, C., Wang, Y., Deng, A., Qin, Q., et al. (2020). Metaproteomics characterizes human gut microbiome function in colorectal cancer. *NPJ Biofilms Microb.* 6, 1–10. doi: 10.1038/s41522-020-0123-4

Lopera-Maya, E. A., Kurilshikov, A., van der Graaf, A., Hu, S., Andreu-Sánchez, S., Chen, L., et al. (2022). Effect of host genetics on the gut microbiome in 7,738 participants of the Dutch Microbiome Project. *Nat. Genet.* 54, 143–151. doi: 10.1038/s41588-021-00992-y

Lu, J., Breitwieser, F. P., Thielen, P., and Salzberg, S. L. (2017). Bracken: estimating species abundance in metagenomics data. *PeerJ Comput. Sci.* 3, e104. doi: 10.7717/peerj-cs.104

Lugli, G. A., Mancabelli, L., Milani, C., Fontana, F., Tarracchini, C., Alessandri, G., et al. (2023). Comprehensive insights from composition to functional microbe-based biodiversity of the infant human gut microbiota. *NPJ Biofilms Microbiomes* 9, 1–13. doi: 10.1038/s41522-023-00392-6

Luo, J., Schumacher, M., Scherer, A., Sanoudou, D., Megherbi, D., Davison, T., et al. (2010). A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data. *Pharmacogenomics J.* 10, 278–291. doi: 10.1038/tpj.2010.57

Mallick, H., Franzosa, E. A., Mclver, L. J., Banerjee, S., Sirota-Madi, A., Kostic, A. D., et al. (2019). Predictive metabolomic profiling of microbial communities using amplicon or metagenomic sequences. *Nat. Commun.* 10, 3136. doi: 10.1038/s41467-019-10927-1

Mallick, H., Ma, S., Franzosa, E. A., Vatanen, T., Morgan, X. C., and Huttenhower, C. (2017). Experimental design and quantitative analysis of microbial community multiomics. *Genome Biol.* 18, 228. doi: 10.1186/s13059-017-1359-z

Manor, O., Dai, C. L., Kornilov, S. A., Smith, B., Price, N. D., Lovejoy, J. C., et al. (2020). Health and disease markers correlate with gut microbiome composition across thousands of people. *Nat. Commun.* 11, 5206. doi: 10.1038/s41467-020-18871-1

Marcos-Zambrano, L. J., Karaduzovic-Hadziabdic, K., Loncar Turukalo, T., Przymus, P., Trajkovik, V., Aasmets, O., et al. (2021). Applications of machine learning in human microbiome studies: a review on feature selection, biomarker identification, disease prediction and treatment. *Front. Microbiol.* 12, 634511. doi: 10.3389/fmicb.2021.634511

Mathieu, A., Leclercq, M., Sanabria, M., Perin, O., and Droit, A. (2022). Machine learning and deep learning applications in metagenomic taxonomy and functional annotation. *Front. Microbiol.* 13, 811495. doi: 10.3389/fmicb.2022.811495

McLaren, M. R., Willis, A. D., and Callahan, B. J. (2019). Consistent and correctable bias in metagenomic sequencing experiments. *Elife* 8, e46923. doi: 10.7554/eLife.46923.027

Meng, C., Feng, S., Hao, Z., Dong, C., and Liu, H. (2022). Changes in gut microbiota composition with age and correlations with gut inflammation in rats. *PLoS ONE* 17, e0265430. doi: 10.1371/journal.pone.0265430

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E., Kubal, M., et al. (2008). The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinform.* 9, 386. doi: 10.1186/1471-2105-9-386

Mihajlović, A., Mladenovi,ć, K., Lončar-Turukalo, T., and Brdar, S. (2021). Machine learning based metagenomic prediction of inflammatory bowel disease. *Stud. Health Technol. Inform.* 285, 165–170. doi: 10.3233/SHTI210591

Moniruzzaman, M., Wurch, L. L., Alexander, H., Dyhrman, S. T., Gobler, C. J., and Wilhelm, S. W. (2017). Virus-host relationships of marine single-celled eukaryotes resolved from metatranscriptomics. *Nat. Commun.* 8, 16054. doi: 10.1038/ncomms16054

Monteleone, A. M., Troisi, J., Fasano, A., Dalle Grave, R., Marciello, F., Serena, G., et al. (2021). Multi-omics data integration in anorexia nervosa patients before and after weight regain: a microbiome-metabolomics investigation. *Clin. Nutr.* 40, 1137–1146. doi: 10.1016/j.clnu.2020.07.021

Muller, E., Algavi, Y. M., and Borenstein, E. (2021). A meta-analysis study of the robustness and universality of gut microbiome-metabolome associations. *Microbiome* 9, 203. doi: 10.1186/s40168-021-01149-z

Muller, E., Algavi, Y. M., and Borenstein, E. (2022). The gut microbiome-metabolome dataset collection: a curated resource for integrative meta-analysis. *NPJ Biofilms Microb.* 8, 1–7. doi: 10.1038/s41522-022-00345-5

Nassar, M., Rogers, A. B., Talo', F., Sanchez, S., Shafique, Z., Finn, R. D., et al. (2022). A machine learning framework for discovery and enrichment of metagenomics metadata from open access publications. *GigaScience* 11, giac077. doi: 10.1093/gigascience/giac077

Nearing, J. T., Comeau, A. M., and Langille, M. G. I. (2021). Identifying biases and their potential solutions in human microbiome studies. *Microbiome* 9, 113. doi: 10.1186/s40168-021-01059-0

Nelkner, J., Huang, L., Lin, T. W., Schulz, A., Osterholz, B., Henke, C., et al. (2023). Abundance, classification and genetic potential of Thaumarchaeota in metagenomes of European agricultural soils: a meta-analysis. *Environ Microb.* 18, 26. doi: 10.1186/s40793-023-00479-9

New, F. N., Baer, B. R., Clark, A. G., Wells, M. T., and Brito, I. L. (2022). Collective effects of human genomic variation on microbiome function. *Sci. Rep.* 12, 3839. doi: 10.1038/s41598-022-07632-3

Noble, E. E., Olson, C. A., Davis, E., Tsan, L., Chen, Y.-W., Schade, R., et al. (2021). Gut microbial taxa elevated by dietary sugar disrupt memory function. *Transl. Psychiatry* 11, 1–16. doi: 10.1038/s41398-021-01309-7

Notario, E., Visci, G., Fosso, B., Gissi, C., Tanaskovic, N., Rescigno, M., et al. (2023). Amplicon-based microbiome profiling: from second- to third-generation sequencing for higher taxonomic resolution. *Genes* 14, 1567. doi: 10.3390/genes14081567

Nyholm, L., Koziol, A., Marcos, S., Botnen, A. B., Aizpurua, O., Gopalakrishnan, S., et al. (2020). Holo-omics: integrated host-microbiota multi-omics for basic and applied biological research. *iScience* 23, 101414. doi: 10.1016/j.isci.2020.101414

Odamaki, T., Kato, K., Sugahara, H., Hashikura, N., Takahashi, S., Xiao, J., et al. (2016). Age-related changes in gut microbiota composition from newborn to centenarian: a cross-sectional study. *BMC Microbiol.* 16, 90. doi: 10.1186/s12866-016-0708-5

Olsen, J. V., and Mann, M. (2011). Effective representation and storage of mass spectrometry-based proteomic data sets for the scientific community. *Sci. Signal.* 4, pe7. doi: 10.1126/scisignal.2001839

Orrhage, K., and Nord, C. E. (1999). Factors controlling the bacterial colonization of the intestine in breastfed infants. *Acta Paediatr. Suppl.* 88, 47–57. doi: 10.1111/j.1651-2227.1999.tb01300.x

Palmer, C., Bik, E. M., DiGiulio, D. B., Relman, D. A., and Brown, P. O. (2007). Development of the human infant intestinal microbiota. *PLoS Biol.* 5, e177. doi: 10.1371/journal.pbio.0050177

Pammi, M., Aghaeepour, N., and Neu, J. (2023). Multiomics, artificial intelligence, and precision medicine in perinatology. *Pediatr. Res.* 93, 308–315. doi: 10.1038/s41390-022-02181-x

Park, J. W., and Graveley, B. R. (2007). Complex alternative splicing. *Adv. Exp. Med. Biol.* 623, 50–63. doi: 10.1007/978-0-387-77374-2_4

Pasolli, E., Schiffer, L., Manghi, P., Renson, A., Obenchain, V., Truong, D. T., et al. (2017). Accessible, curated metagenomic data through ExperimentHub. *Nat. Methods* 14, 1023–1024. doi: 10.1038/nmeth.4468

Pasolli, E., Truong, D. T., Malik, F., Waldron, L., and Segata, N. (2016). Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. *PLoS Comput. Biol.* 12, e1004977. doi: 10.1371/journal.pcbi.1004977

Patangia, D. V., Anthony Ryan, C., Dempsey, E., Paul Ross, R., and Stanton, C. (2022). Impact of antibiotics on the human microbiome and consequences for host health. *Microbiologyopen* 11, e1260. doi: 10.1002/mbo3.1260

Penders, J., Thijs, C., Vink, C., Stelma, F. F., Snijders, B., Kummeling, I., et al. (2006). Factors influencing the composition of the intestinal microbiota in early infancy. *Pediatrics* 118, 511–521. doi: 10.1542/peds.2005-2824

Pereira-Marques, J., Hout, A., Ferreira, R. M., Weber, M., Pinto-Ribeiro, I., van Doorn, L.-J., et al. (2019). Impact of host DNA and sequencing depth on the taxonomic resolution of whole metagenome sequencing for microbiome analysis. *Front. Microbiol.* 10, 01277. doi: 10.3389/fmicb.2019.01277

Perez-Riverol, Y., Bai, J., Bandla, C., García-Seisdedos, D., Hewapathirana, S., Kamatchinathan, S., et al. (2021). The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Res.* 50, D543–D552. doi: 10.1093/nar/gkab1038

Peterson, D., Bonham, K. S., Rowland, S., Pattanayak, C. W., and Klepac-Ceraj, V. (2021). Comparative analysis of 16S rRNA gene and metagenome sequencing in pediatric gut microbiomes. *Front. Microbiol.* 12, 670336. doi: 10.3389/fmicb.2021.670336

Pienaar, E., Theron, M., Nelson, M., and Viljoen, H. (2006). A quantitative model of error accumulation during pcr amplification. *Comput. Biol. Chem.* 30, 102–111. doi: 10.1016/j.compbiolchem.2005.11.002

Pietrucci, D., Teofani, A., Milanesi, M., Fosso, B., Putignani, L., Messina, F., et al. (2022). Machine learning data analysis highlights the role of parasutterella and alloprevotella in autism spectrum disorders. *Biomedicines* 10, 2028. doi: 10.3390/biomedicines10082028

Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., and Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* 35, 833–844. doi: 10.1038/nbt.3935

Raghavendra Nayaka, P., and Ranjan, R. (2023). An efficient framework for algorithmic metadata extraction over scholarly documents using deep neural networks. *SN Comput. Sci.* 4:341. doi: 10.1007/s42979-023-01776-3

Ramirez, J., Guarner, F., Bustos Fernandez, L., Maruy, A., Sdepanian, V. L., and Cohen, H. (2020). Antibiotics as major disruptors of gut microbiota. *Front. Cell. Infect. Microb.* 10, 572912. doi: 10.3389/fcimb.2020.572912

Räz, T. (2024). ML interpretability: simple isn't easy. *Stud. Hist. Philos. Sci.* 103, 159–167. doi: 10.1016/j.shpsa.2023.12.007

Rechenberger, J., Samaras, P., Jarzab, A., Behr, J., Frejno, M., Djukovic, A., et al. (2019). Challenges in clinical metaproteomics highlighted by the analysis of acute leukemia patients with gut colonization by multidrug-resistant enterobacteriaceae. *Proteomes* 7, 2. doi: 10.3390/proteomes7010002

Ren, Y., Wu, J., Wang, Y., Zhang, L., Ren, J., Zhang, Z., et al. (2023). Lifestyle patterns influence the composition of the gut microbiome in a healthy Chinese population. *Sci. Rep.* 13, 14425. doi: 10.1038/s41598-023-41532-4

Ribeiro, C. F. A., Silveira, G. G. O. S., Cândido, E. S., Cardoso, M. H., Espínola Carvalho, C. M., Franco, O. L. (2020). Effects of antibiotic treatment on gut microbiota and how to overcome its negative impacts on human health. *ACS Infect. Dis.* 6, 2544–2559. doi: 10.1021/acsinfecdis.0c00036

Richardson, L., Allen, B., Baldi, G., Beracochea, M., Bileschi, M. L., Burdett, T., et al. (2023). MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Res.* 51, D753–D759. doi: 10.1093/nar/gkac1080

Ridaura, V. K., Faith, J. J., Rey, F. E., Cheng, J., Duncan, A. E., Kau, A. L., et al. (2013). Cultured gut microbiota from twins discordant for obesity modulate adiposity and metabolic phenotypes in mice. *Science* 341, 1241214. doi: 10.1126/science.1241214

Rojas-Velazquez, D., Kidwai, S., Kraneveld, A. D., Tonda, A., Oberski, D., Garssen, J., et al. (2024). Methodology for biomarker discovery with reproducibility in microbiome data using machine learning. *BMC Bioinform.* 25, 26. doi: 10.1186/s12859-024-05639-3

Romano, P., Giugno, R., and Pulvirenti, A. (2011). Tools and collaborative environments for bioinformatics research. *Brief. Bioinform.* 12, 549–561. doi: 10.1093/bib/bbr055

Roy, S., Coldren, C., Karunamurthy, A., Kip, N. S., Klee, E. W., Lincoln, S. E., et al. (2018). Standards and guidelines for validating next-generation sequencing bioinformatics pipelines: a joint recommendation of the association for molecular pathology and the College of American Pathologists. *J. Mol. Diag.* 20, 4–27. doi: 10.1016/j.jmoldx.2017.11.003

Salek, R. M., Neumann, S., Schober, D., Hummel, J., Billiau, K., Kopka, J., et al. (2015). COordination of Standards in MetabOlomicS (COSMOS):

facilitating integrated metabolomics data access. *Metabolomics* 11, 1587–1597. doi: 10.1007/s11306-015-0810-y

Salek, R. M., Steinbeck, C., Viant, M. R., Goodacre, R., and Dunn, W. B. (2013). The role of reporting standards for metabolite annotation and identification in metabolomic studies. *Gigascience* 2, 13. doi: 10.1186/2047-217X-2-13

Santamaria, M., Fosso, B., Consiglio, A., De Caro, G., Grillo, G., Licciulli, F., et al. (2012). Reference databases for taxonomic assignment in metagenomics. *Briefings Bioinform.* 13, 682–695. doi: 10.1093/bib/bbs036

Santamaria, M., Fosso, B., Licciulli, F., Balech, B., Larini, I., Grillo, G., et al. (2018). ITSoneDB: a comprehensive collection of eukaryotic ribosomal RNA Internal Transcribed Spacer 1 (ITS1) sequences. *Nucleic Acids Res.* 46, D127–D132. doi: 10.1093/nar/gkx855

Schloss, P. D. (2018). Identifying and overcoming threats to reproducibility, replicability, robustness, and generalizability in microbiome research. *mBio* 9, 10.1128/mbio.00525-18. doi: 10.1128/mBio.00525-18

Schmidt, J. (2023). Testing for Overfitting. *arXiv [Preprint]. arXiv*:2305.05792 doi: 10.48550/arXiv.2305.05792

Schnoes, A. M., Brown, S. D., Dodevski, I., and Babbitt, P. C. (2009). Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput. Biol.* 5, e1000605. doi: 10.1371/journal.pcbi.1000605

Scholz, M., Ward, D. V., Pasolli, E., Tolio, T., Zolfo, M., Asnicar, F., et al. (2016). Strain-level microbial epidemiology and population genomics from shotgun metagenomics. *Nat. Methods* 13, 435–438. doi: 10.1038/nmeth.3802

Schorn, M. A., Verhoeven, S., Ridder, L., Huber, F., Acharya, D. D., Aksenov, A. A., et al. (2021). A community resource for paired genomic and metabolomic data mining. *Nat. Chem. Biol.* 17, 363–368. doi: 10.1038/s41589-020-00724-z

Serwecińska, L. (2020). Antimicrobials and antibiotic-resistant bacteria: a risk to the environment and to public health. *Water* 12, 3313. doi: 10.3390/w12123313

Shakya, M., Lo, C.-C., and Chain, P. S. G. (2019). Advances and challenges in metatranscriptomic analysis. *Front. Genet.* 10, 904. doi: 10.3389/fgene.2019.00904

Silverman, J. D., Roche, K., Mukherjee, S., and David, L. A. (2020). Naught all zeros in sequence count data are the same. *Comput. Struct. Biotechnol. J.* 18, 2789–2798. doi: 10.1016/j.csbj.2020.09.014

Singh, A., and Mittal, M. (2020). Neonatal microbiome - a brief review. *J. Matern. Fetal Neonatal Med.* 33, 3841–3848. doi: 10.1080/14767058.2019.1583738

Storr, M., Vogel, H. J., and Schicho, R. (2013). Metabolomics: is it useful for inflammatory bowel diseases? *Curr. Opin. Gastroenterol.* 29, 378–383. doi: 10.1097/MOG.0b013e328361f488

Su, Q., Liu, Q., Lau, R. I., Zhang, J., Xu, Z., Yeoh, Y. K., et al. (2022). Faecal microbiome-based machine learning for multi-class disease diagnosis. *Nat. Commun.* 13, 6818. doi: 10.1038/s41467-022-34405-3

Sud, M., Fahy, E., Cotter, D., Azam, K., Vadivelu, I., Burant, C., et al. (2016). Metabolomics workbench: an international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Res.* 44, D463–470. doi: 10.1093/nar/gkv1042

Sumner, L. W., Amberg, A., Barrett, D., Beale, M. H., Beger, R., Daykin, C. A., et al. (2007). Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics* 3, 211–221. doi: 10.1007/s11306-007-0082-2

Sunagawa, S., Mende, D. R., Zeller, G., Izquierdo-Carrasco, F., Berger, S. A., Kultima, J. R., et al. (2013). Metagenomic species profiling using universal phylogenetic marker genes. *Nat. Methods* 10, 1196–1199. doi: 10.1038/nmeth.2693

Tangaro, M. A., Defazio, G., Fosso, B., Licciulli, V. F., Grillo, G., Donvito, G., et al. (2021). ITSoneWB: profiling global taxonomic diversity of eukaryotic communities on Galaxy. *Bioinformatics* 37, 4253–4254. doi: 10.1093/bioinformatics/btab431

ten Hoopen, P., Finn, R. D., Bongo, L. A., Corre, E., Fosso, B., Meyer, F., et al. (2017). The metagenomic data life-cycle: standards and best practices. *Gigascience* 6, 1–11. doi: 10.1093/gigascience/gix047

Tenopir, C., Rice, N. M., Allard, S., Baird, L., Borycz, J., Christian, L., et al. (2020). Data sharing, management, use, and reuse: practices and perceptions of scientists worldwide. *PLoS ONE* 15, e0229003. doi: 10.1371/journal.pone.0229003

The UniProt Consortium (2023). UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* 51, D523–D531. doi: 10.1093/nar/gkac1052

Thompson, L., Vangay, P., Blumberg, K., Christianson, D., Dundore-Arias, J. P., Hu, B., et al. (2020). Introduction to metadata and ontologies: everything you always wanted to know about metadata and ontologies (but were afraid to ask). Berkeley, CA: Lawrence Berkeley National Laboratory (LBNL). National Microbiome Data Collaborative (NMDC).

Tomofuji, Y., Sonehara, K., Kishikawa, T., Maeda, Y., Ogawa, K., Kawabata, S., et al. (2023). Reconstruction of the personal information from human genome reads in gut metagenome sequencing data. *Nat. Microbiol.* 8, 1079–1094. doi: 10.1038/s41564-023-01381-3

Vailati-Riboni, M., Palombo, V., and Loor, J. J. (2017). "What Are Omics Sciences?," in *Periparturient Diseases of Dairy Cows: A Systems Biology Approach*, ed. B. N. Ametaj (Cham: Springer International Publishing), 1–7.

Vangay, P., Burgin, J., Johnston, A., Beck, K. L., Berrios, D. C., Blumberg, K., et al. (2021). Microbiome metadata standards: report of the national microbiome data collaborative's workshop and follow-on activities. *mSystems* 6, e01194–20. doi: 10.1128/mSystems.01194-20

Verberkmoes, N. C., Russell, A. L., Shah, M., Godzik, A., Rosenquist, M., Halfvarson, J., et al. (2009). Shotgun metaproteomics of the human distal gut microbiota. *ISME J.* 3, 179–189. doi: 10.1038/ismej.2008.108

Vesteghem, C., Brøndum, R. F., Sønderkær, M., Sommer, M., Schmitz, A., Bødker, J. S., et al. (2020). Implementing the FAIR Data Principles in precision oncology: review of supporting initiatives. *Brief. Bioinform.* 21, 936–945. doi: 10.1093/bib/bbz044

Vinciotti, V., Wit, E., and Richter, F. (2023). Random Graphical Model of Microbiome Interactions in Related Environments. *arXiv [Preprint]. arXiv:* 2304.01956

Vizcaíno, J. A., Deutsch, E. W., Wang, R., Csordas, A., Reisinger, F., Ríos, D., et al. (2014). ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. Biotechnol.* 32, 223–226. doi: 10.1038/nbt.2839

Vujkovic-Cvijin, I., Sklar, J., Jiang, L., Natarajan, L., Knight, R., and Belkaid, Y. (2020). Host variables confound gut microbiota studies of human disease. *Nature* 587, 448–454. doi: 10.1038/s41586-020-2881-9

Wagner, J., Paulson, J. N., Wang, X., Bhattacharjee, B., and Corrada Bravo, H. (2016). Privacy-preserving microbiome analysis using secure computation. *Bioinformatics* 32, 1873–1879. doi: 10.1093/bioinformatics/btw073

Walsh, C., Stallard-Olivera, E., and Fierer, N. (2023). Nine (not so simple) steps: a practical guide to using machine learning in microbial ecology. *MBio* e02050–e02023. doi: 10.1128/mbio.02050-23. [Epub ahead of print].

Wang, J., Wang, J., Pang, X., Zhao, L., Tian, L., and Wang, X. (2016). Sex differences in colonization of gut microbiota from a man with short-term vegetarian and inulin-supplemented diet in germ-free mice. *Sci. Rep.* 6, 36137. doi: 10.1038/srep36137

Watson, D. S. (2022). Interpretable machine learning for genomics. *Hum. Genet.* 141, 1499–1513. doi: 10.1007/s00439-021-02387-9

Wemheuer, F., Taylor, J. A., Daniel, R., Johnston, E., Meinicke, P., Thomas, T., et al. (2020). Tax4Fun2: prediction of habitat-specific functional profiles and functional redundancy based on 16S rRNA gene sequences. *Environ. Microb.* 15, 11. doi: 10.1186/s40793-020-00358-7

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* 3, 160018. doi: 10.1038/sdata.2016.18

Wood-Charlson, E. M., Anubhav, A.uberry, D., Blanco, H., and Borkum, M. I., Corilo, Y. E., et al. (2020). The national microbiome data collaborative: enabling microbiome science. *Nat. Rev. Microbiol.* 18, 313–314. doi: 10.1038/s41579-020-0377-0

Xiong, W., Abraham, P. E., Li, Z., Pan, C., and Hettich, R. L. (2015). Microbial metaproteomics for characterizing the range of metabolic functions and activities of human gut microbiota. *Proteomics* 15, 3424–3438. doi: 10.1002/pmic.201400571

Yang, Q., Zhang, A., Miao, J., Sun, H., Han, Y., Yan, G., et al. (2019). Metabolomics biotechnology, applications, and future trends: a systematic review. *RSC Adv.* 9, 37245. doi: 10.1039/C9RA06697G

Yang, X., Coulombe-Huntington, J., Kang, S., Sheynkman, G. M., Hao, T., Richardson, A., et al. (2016). Widespread expansion of protein interaction capabilities by alternative splicing. *Cell* 164, 805–817. doi: 10.1016/j.cell.2016.01.029

Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., et al. (2012). Human gut microbiome viewed across age and geography. *Nature* 486, 222–227. doi: 10.1038/nature11053

Yilmaz, P., Kottmann, R., Field, D., Knight, R., Cole, J. R., Amaral-Zettler, L., et al. (2011). Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nat. Biotechnol.* 29, 415–420. doi: 10.1038/nbt.1823

Yuan, D., Ahamed, A., Burgin, J., Cummins, C., Devraj, R., Gueye, K., et al. (2023). The european nucleotide archive in 2023. *Nucleic Acids Res.* 52, D92–D97. doi: 10.1093/nar/gkad1067

Yurekten, O., Payne, T., Tejera, N., Amaladoss, F. X., Martin, C., Williams, M., et al. (2023). MetaboLights: open data repository for metabolomics. *Nucleic Acids Res.* 52, D640–D646. doi: 10.1093/nar/gkad1045

Zeller, G., Tap, J., Voigt, A. Y., Sunagawa, S., Kultima, J. R., Costea, P. I., et al. (2014). Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.* 10, 766. doi: 10.15252/msb.20145645

Zhao, S., Li, C., Li, G., Yang, S., Zhou, Y., He, Y., et al. (2019). Comparative analysis of gut microbiota among the male, female and pregnant giant pandas (*Ailuropoda Melanoleuca*). *Open Life Sci.* 14, 288–298. doi: 10.1515/biol-2019-0032

Zhu, L., Wang, J., and Bahrndorff, S. (2021). Editorial: the wildlife gut microbiome and its implication for conservation biology. *Front. Microbiol.* 12, 697499. doi: 10.3389/fmicb.2021.697499

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 15 February 2024
DOI 10.3389/fmicb.2024.1348974

OPEN ACCESS

EDITED BY
Domenica D'Elia,
National Research Council (CNR), Italy

REVIEWED BY
Inwoo Baek,
Advanced Radiation Technology Institute, Korea Atomic Energy Research Institute, Republic of Korea
Piotr Przymus,
Nicolaus Copernicus University in Toruń, Poland

*CORRESPONDENCE
Sabina Tangaro
✉ sabina.tangaro@uniba.it

RECEIVED 03 December 2023
ACCEPTED 24 January 2024
PUBLISHED 15 February 2024

CITATION
Novielli P, Romano D, Magarelli M, Bitonto PD, Diacono D, Chiatante A, Lopalco G, Sabella D, Venerito V, Filannino P, Bellotti R, De Angelis M, Iannone F and Tangaro S (2024) Explainable artificial intelligence for microbiome data analysis in colorectal cancer biomarker identification.
Front. Microbiol. 15:1348974.
doi: 10.3389/fmicb.2024.1348974

Explainable artificial intelligence for microbiome data analysis in colorectal cancer biomarker identification

Pierfrancesco Novielli[1,2], Donato Romano[1,2], Michele Magarelli[1], Pierpaolo Di Bitonto[1], Domenico Diacono[2], Annalisa Chiatante[1], Giuseppe Lopalco[3], Daniele Sabella[3], Vincenzo Venerito[3], Pasquale Filannino[1], Roberto Bellotti[2,4], Maria De Angelis[1], Florenzo Iannone[3] and Sabina Tangaro[1,2]*

[1]Dipartimento di Scienze del Suolo, della Pianta e degli Alimenti, Università degli Studi di Bari Aldo Moro, Bari, Italy, [2]Istituto Nazionale di Fisica Nucleare, Sezione di Bari, Bari, Italy, [3]Dipartimento di Medicina di Precisione e Rigenerativa e Area Jonica, Università degli Studi di Bari Aldo Moro, Bari, Italy, [4]Dipartimento Interateneo di Fisica M. Merlin, Università degli Studi di Bari Aldo Moro, Bari, Italy

Background: Colorectal cancer (CRC) is a type of tumor caused by the uncontrolled growth of cells in the mucosa lining the last part of the intestine. Emerging evidence underscores an association between CRC and gut microbiome dysbiosis. The high mortality rate of this cancer has made it necessary to develop new early diagnostic methods. Machine learning (ML) techniques can represent a solution to evaluate the interaction between intestinal microbiota and host physiology. Through explained artificial intelligence (XAI) it is possible to evaluate the individual contributions of microbial taxonomic markers for each subject. Our work also implements the Shapley Method Additive Explanations (SHAP) algorithm to identify for each subject which parameters are important in the context of CRC.

Results: The proposed study aimed to implement an explainable artificial intelligence framework using both gut microbiota data and demographic information from subjects to classify a cohort of control subjects from those with CRC. Our analysis revealed an association between gut microbiota and this disease. We compared three machine learning algorithms, and the Random Forest (RF) algorithm emerged as the best classifier, with a precision of 0.729 ± 0.038 and an area under the Precision-Recall curve of 0.668 ± 0.016. Additionally, SHAP analysis highlighted the most crucial variables in the model's decision-making, facilitating the identification of specific bacteria linked to CRC. Our results confirmed the role of certain bacteria, such as *Fusobacterium*, *Peptostreptococcus*, and *Parvimonas*, whose abundance appears notably associated with the disease, as well as bacteria whose presence is linked to a non-diseased state.

Discussion: These findings emphasizes the potential of leveraging gut microbiota data within an explainable AI framework for CRC classification. The significant association observed aligns with existing knowledge. The precision exhibited by the RF algorithm reinforces its suitability for such classification tasks. The SHAP analysis not only enhanced interpretability but identified specific bacteria crucial in CRC determination. This approach opens avenues for targeted

interventions based on microbial signatures. Further exploration is warranted to deepen our understanding of the intricate interplay between microbiota and health, providing insights for refined diagnostic and therapeutic strategies.

KEYWORDS

machine learning, explainable artificial intelligence, colorectal cancer, microbiome, biomarker identification, microbiota, precision medicine

1 Introduction

Colorectal cancer (CRC) stands as the third most prevalent cancer globally (Morgan et al., 2023), claiming a significant toll in cancer-related fatalities. The high mortality is due to the abnormal growth of cells with the capacity to invade tissues and spread to other parts of the body. Most colorectal cancers are due to lifestyle and advanced age and only a few cases are attributable to hereditary genetic diseases. Its incidence is constantly increasing, and in-depth understanding of the pathogenetic mechanisms, early diagnosis and innovative therapeutic options have become crucial imperatives to address this growing challenge. The complexity of colorectal cancer is highlighted by the diversity of pathological pathways involved and the variability in response to treatments. The prevailing gold standard for CRC diagnosis, colonoscopy, is burdened by invasiveness and discomfort. However, resistance to conventional treatments, post-surgical recurrence and the need to improve access to care, especially in disadvantaged communities make it necessary to open up to personalized therapies and more targeted management strategies. A non-standardized approach keep in mind the peculiar molecular characteristics of each tumor and the patient's individual responses to therapies. Hence, the pressing demand for non-invasive, cost-effective early detection methods persists. Non-invasive therapies take on particular relevance with a view to reducing physical and psychological stress on patients, reducing the recovery period and improving the quality of life post-treatment.

The gut microbiota, a complex community of microorganisms that colonize the gastrointestinal tract, has emerged as a critical player in the regulation of intestinal homeostasis and the modulation of local immune responses. In recent years, a growing body of scientific evidence has highlighted the critical role of the intestinal microbiota in the pathogenesis and development of colorectal cancer. The dynamic interactions between the microbiota and the intestinal mucosa play a key role in maintaining a physiological environment and preventing the onset of cellular alterations. However, dysbiosis or imbalances in the composition of the microbiota can contribute to carcinogenesis, promoting chronic inflammation, the production of carcinogenic metabolites and alteration of the mucosal barrier. Certain bacteria, like *Fusobacterium nucleatum* and *Parvimonas micra*, are notably more abundant in CRC patients, often linked to the disease's development (Yachida et al., 2019; Löwenmark et al., 2020; Wu et al., 2021). These findings drive the exploration of using fecal biomarkers for CRC diagnosis. Understanding the central role of the gut microbiota in the context of colorectal cancer could guide the development of personalized strategies for disease management, exploiting the therapeutic potential of microbial manipulation. Harnessing the power of machine learning (ML) (Amodeo et al., 2021; Bellando-Randone et al., 2021; Rynazal et al., 2023; Golob et al., 2024), our study crafts a comprehensive framework to scrutinize fecal microbiome data gleaned from both healthy subjects and those afflicted with CRC. This framework intricately involves data preprocessing, feature extraction, feature selection, and model construction, employing an array of ML algorithms. To ensure transparency and interpretability in our study, we embrace the principles of Explainable Artificial Intelligence (XAI) (Lombardi et al., 2021a,b; Bellantuono et al., 2023; Novielli et al., 2023). XAI not only enhances the trustworthiness of our models but also empowers clinicians to understand the rationale behind each prediction. This is particularly crucial in the context of personalized CRC management, where treatment decisions need to be aligned with the unique characteristics of each patient. The impact of gut microbiota on CRC analyzed through machine learning, coupled with transparent explanations afforded by XAI, holds the potential to develop how to diagnose and manage colorectal cancer, fostering a new era of precision medicine that is both effective and readily comprehensible.

TABLE 1 Summary table of the datasets used in the analysis.

Dataset	Control	CRC	Metadata
Baxter et al. (2016)	171	120	Gender, age, BMI, country
Zackular et al. (2014)	30	30	Gender, age, BMI, country
Zeller et al. (2014)	50	41	Gender, age, BMI, country
TOTAL	251	191	Gender, age, BMI, country

2 Materials

In this study, we used three different dataset of three different works (Zackular et al., 2014; Zeller et al., 2014; Baxter et al., 2016). For each of them, we considered the control patient (NC) and the CRC ones. These datasets collect 442 human stool samples characterized by 16S metagenomic sequencing of the V4 region of the 16S rRNA, from different countries: Canada (CA), France (FRA), United States of America (USA). These dataset provide information regarding the abundance of the gut microbiota in NC patients and CRC ones at genus level.

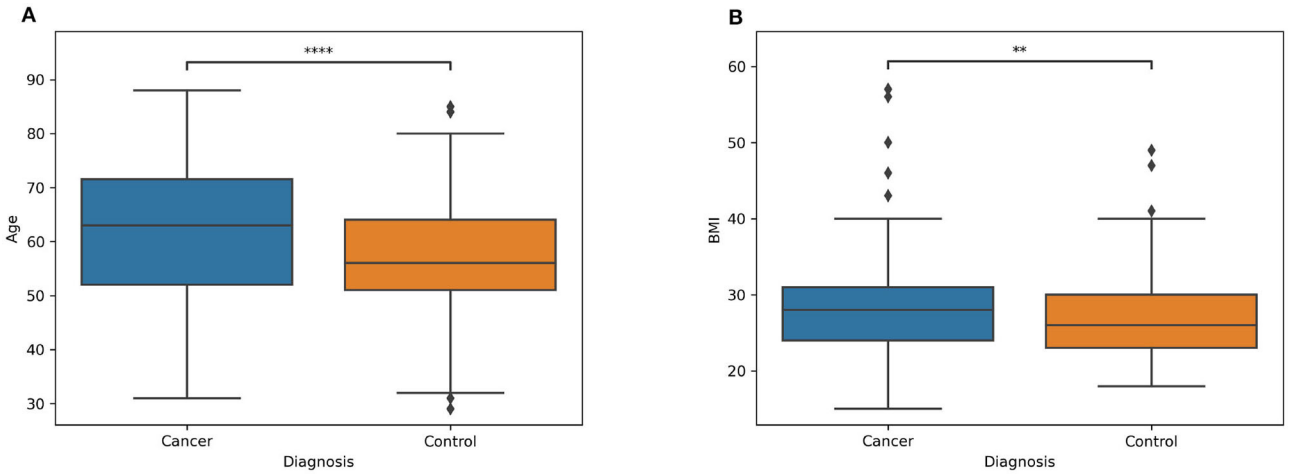

FIGURE 1
Boxplot of two classes (patients and controls) of the **(A)** age and **(B)** BMI. The symbol * denotes the significance level determined by the Mann-Whitney U rank test for comparing two distributions. **Stands for *p*-value less or equal then 0.01. ****stands for *p*-value less or equal then 0.0001.

TABLE 2 **Demographic characteristics of the study participants.**

	CRC (191)	Control (251)	p-value
Gender	114 M / 77 F	101 M / 150 F	< 0.01
Country	2 CA / 41 FRA / 148 USA	3 CA / 50 FRA / 198 USA	0.892

The Fisher's exact test was performed for gender and country.

Moreover, each of them is characterized with four metadata features: gender, age, body mass index (BMI), country, as reported in Table 1.

Information about the distribution of age and BMI for both patients and controls are showed respectively in Figures 1A, B, while the demographic characteristics of the entire dataset is reported in Table 2. In the Supplementary Table S1 is reported the information related to the metadata of each subject involved in the analysis.

3 Methods

The workflow begins with the preprocessing of microbiome data, followed by the construction of an explainable machine learning model. The performance of three classifiers—XGBoost, Random Forest, and Support Vector Machine—was rigorously compared through a 20-repeated 5-fold Stratified Cross Validation. Finally, we explore the functionality of the optimal classifier using the XAI approach. This includes collecting SHAP values for different (feature, prediction) pairs and averaging them across the 20 repetitions of the model CV. Figure 2 outlines the Artificial Intelligence procedure implemented in this study to develop a Machine Learning classifier for distinguishing between control and CRC samples.

3.1 Preprocessing of the microbiome samples

Preprocessing of microbiome data is a crucial step in the analysis pipeline (Ibrahimi et al., 2023; Papoutsoglou et al., 2023). The microbiome data undergo several preprocessing steps. Firstly, a filtration of taxonomic units is conducted, focusing on removing non-informative features or taxa that are biologically irrelevant or potential contaminants (Cao et al., 2021). This involves applying thresholds based on abundance/prevalence, variance, or correlation. In our case, low-abundance or prevalence filtering eliminates features present in <10% of the samples. The subsequent step involves normalization, aiming to address variability in sampling depth and data sparsity. One approach for data normalization is through transformation methods, wherein values are replaced with their normalized counterparts. Given that microbiome datasets are inherently compositional, these methods adhere to Aitchison's methodology for compositional data. They transform feature counts into log-ratios within each sample, utilizing an additive, centered log-ratio transformation (Aitchison, 1982; Egozcue et al., 2003).

3.2 Machine learning classifier

3.2.1 XGBoost

The XGBoost algorithm employs a collective of decision trees trained through an iterative gradient boosting process. This process involves addressing critical points within decision trees at each step through subsequent trees. Addressing the challenge of missing values, XGBoost employs sparsity-aware split finding (Chen and Guestrin, 2016). This technique leverages data sparsity patterns in a unified manner, determining the optimal direction in the event of a missing feature necessary for a split. In the quest for optimal performance in classification under cross-validation conditions, we explore various XGBoost parameters:

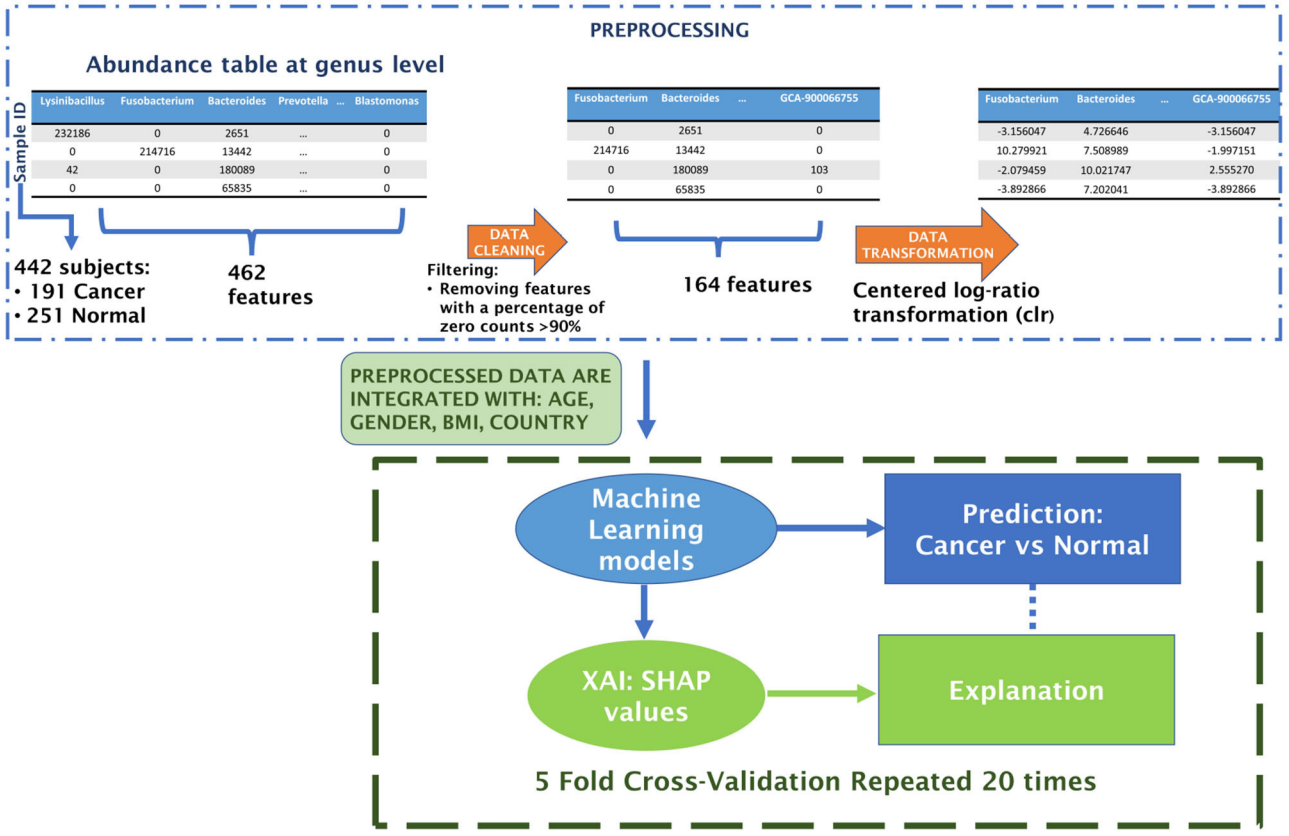

FIGURE 2
Schematic flowchart of the analysis.

- max depth ϵ {None, 3, 5},
- col sample bytree ϵ {0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9},
- n estimators ϵ {50, 100, 150, 200, 250}.

The implementation of the XGBoost algorithm utilizes the Python (version 3.11.5) package xgboost (version 2.0.2).

3.2.2 Random forest

The Random Forest (RF) algorithm entails an ensemble of decision trees derived through resampling the training dataset with repetitions (bootstrapping) (Breiman, 2001). This process, along with the randomization of features during training, ensures low mutual correlation between RF trees. Decision trees generate independent predictions for each observation, and their collective outcomes are aggregated through either averaging (for regression) or majority voting (for classification). Noteworthy characteristics of RF algorithms include easy tunability, a minimal number of parameters, resilience against overfitting, the ability to assess feature importance during training, and an unbiased estimation of generalization error. In this study, we aimed to optimize the control/crc classification in cross-validation mode by varying specific RF parameters, including:

- max depth {ϵ None, 3, 5},
- n estimators ϵ {50, 100, 150, 200, 250}.

The RF algorithm implementation utilized the Python (version 3.11.5) package scikit-learn (version 1.3.0) (Pedregosa et al., 2011).

3.2.3 Support vector machine

The Support Vector Machine (SVM) operates by determining the optimal boundary between two or more classes in the data space through the minimization of a loss function known as Hinge Loss, augmented with a penalty term (Cortes and Vapnik, 1995). In this algorithm, only a limited set of input observations, termed support vectors, actively contribute to delineating the boundary between classes. The SVM algorithm iterates by treating misclassified instances as support vectors, with their contribution to the loss being proportional to their distance from the boundary. This approach ensures that the loss is influenced solely by a subset of input observations, facilitating an efficient estimation of optimal parameters. For the optimization of control/CRC classification under cross-validation conditions, we vary the following SVM parameters:

- C ϵ {1, 5, 10, 20},
- Gamma ϵ {0.001, 0.01, 1}.

The SVM algorithm is implemented using the Python (version 3.11.5) package scikit-learn (version 1.3.0) (Pedregosa et al., 2011).

TABLE 3 Comparison between evaluation metrics of XGBoost (XGB), Random Forest (RF), and Support Vector Machine (SVM) classifiers.

	ACC	F1	PREC	AUC ROC	AUPRC
XGB	0.652 (0.017)	**0.567 (0.022)**	0.613 (0.022)	**0.701 (0.015)**	0.639 (0.021)
RF	**0.673 (0.015)**	0.507 (0.030)	**0.729 (0.038)**	0.699 (0.011)	**0.668 (0.016)**
SVM	0.633 (0.025)	0.478 (0.091)	0.613 (0.032)	0.663 (0.036)	0.597 (0.037)

The mean values accompanied by the standard deviation are shown. The highest values for each metric are indicated in bold, and the second-highest values are underscored.

3.3 Evaluation metrics

In the realm of classification machine learning, the selection of appropriate evaluation metrics is crucial for assessing the performance of models. These metrics provide quantitative measures of a model's ability to correctly classify instances and are essential tools for comparing and optimizing different algorithms. In order to obtain statistically robust results, a 5-fold cross-validation was applied to partition the dataset, where each fold was used as a test set while the remaining four as training ones (Schaffer, 1993). An hyperparameter tuning was conducted with a random search by using the RandomizedSearchCV function of the python library scikit-learn (Bergstra and Bengio, 2012), implemented with a nested 3-fold cross- validation to avoid bias in the estimation of test error (Varma and Simon, 2006). The entire process was repeated 20 times, by dividing the dataset with different partitions between each repetition.

The metrics used to evaluate the performance of models were (Venerito et al., 2022):

- Accuracy: The accuracy is the proportion of correct predictions (both true positives and true negatives) among the total number predictions.
- Recall: The recall is a metric evaluating the frequency with which a machine learning model accurately recognizes positive instances (true positives) among all the actual positive samples. It is calculated by dividing the number of true positives by the total number of elements that actually belong to the positive class.
- Precision: The precision is a metric assessing how often a machine learning model predicts the positive class. It is computed by dividing the number of accurate positive predictions (true positives) by the total instances predicted as positive by the model (sum of true positives and false positives).
- F1 score: The F1 score is the harmonic mean of the precision and recall.
- AUC ROC: The area under the Receiver Operating Characteristic (ROC) curve;
- AUPRC: The area under the Precision-Recall (PR) curve.

We considered as positive instances those ones belonging to the CRC class.

For the evaluation of the best classifier, the one with the highest AUPRC will be chosen. This metric is well-suited for assessing the discriminative power of a classifier in the presence of an imbalanced dataset, where the number of positive cases is greater than the number of negative cases (Ozenne et al., 2015).

3.4 SHAP algorithm

The eXplainable Artificial Intelligence (XAI) framework encompasses a variety of techniques united by their shared focus on informativeness, uncertainty estimation, generalization, and transparency. In this study, we employ the SHAP local explanation algorithm to uncover the significance of features in classifying control/CRC samples. Serving as a local, model-agnostic *post-hoc* explainer, the SHAP algorithm derives inspiration from Shapley (SHAP) values rooted in cooperative game theory (Lundberg and Lee, 2017; Lundberg et al., 2020). It constructs interpretable linear models for individual samples, highlighting the contribution of each feature to the sample's prediction. The computation of SHAP values involves assessing the difference in model output predictions with and without specific features, considering all conceivable feature subsets. As a result, the model requires retraining on all subsets F of the complete set S of features ($F \subseteq S$). The SHAP value for the *jth* feature of the instance x is determined by aggregating it across all possible subsets (Equation 1):

$$\Phi j(x) = \sum_{F \subseteq S - \{j\}} \frac{|F|!(|S| - |F| - 1)!}{|S|!} [fx(F \cup j) - fx(F)] \qquad (1)$$

where $|F|!$ represents the permutations of features in the subset F, $(|S| - |F| - 1)!$ the permutations of features in the subset S - ($F \subseteq \{j\}$) and $|S|!$ is the total number of feature permutations.

The SHAP value calculation is implemented in the Python (version 3.11.5) package shap (version 0.43.0). For RF and XGBoost models, we utilized the TreeExplainer function with the "feature perturbation" parameter set to "interventional." This approach is tailored to disrupt dependencies between features, aligning with the principles outlined in causal inference (Janzing et al., 2020). By adopting this parameter configuration, our objective was to alleviate the impact of highly correlated predictors, thereby mitigating potential misinterpretations and ensuring a more robust analysis.

4 Results

The objective of this study was to investigate changes in the gut microbiota among individuals with CRC in comparison to control subjects. To unveil these alterations, a machine learning-based classification model was employed, and the contribution of features was analyzed. Our attention will be directed toward the outcomes of the Artificial Intelligence workflow, specifically examining the classification performance of various Machine Learning algorithms and the prevalence of bacteroides that exerts the most significant influence on predictions.

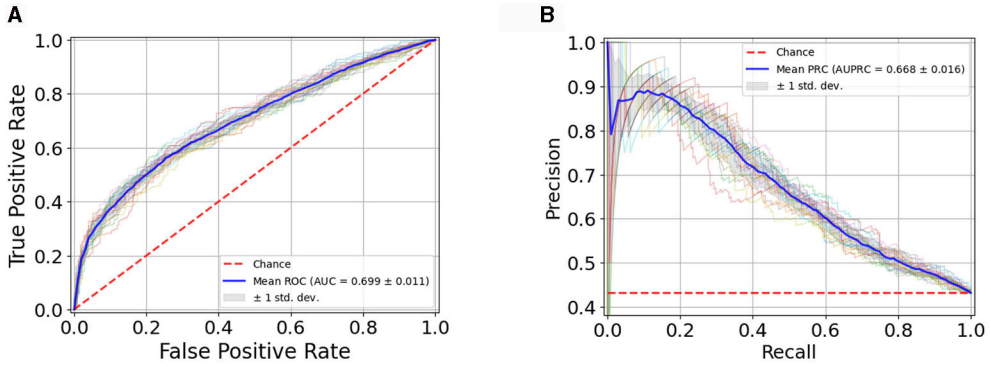

FIGURE 3
(A) Average ROC Curve with standard deviation over 20 model runs; **(B)** Average PR Curve with standard deviation over 20 model runs.

4.1 Feature engineering

The dataset utilized in this study consists of abundance tables representing microbial communities from the V4 region of the 16S rRNA, collected at the genus level. Starting with an initial dataset comprising 462 features (microbial communities), the data cleaning process, as described in the methods, reduced the total number of features to 164. Following the centered log-ratio transformation for each sample, additional variables were incorporated, including country, age, BMI, and gender. This resulting dataset served as the input for the machine learning classification framework.

4.2 Classification CRC/control

A comprehensive correlation analysis was conducted among all features considered as inputs to the ML classifier and the output target class. The outcomes of this analysis are presented in Supplementary Figure S1, where the top features are displayed in descending order based on their correlation coefficients with the target class. Despite observing statistically significant correlations among the features, it is noteworthy that the maximum correlation does not exceed 0.3. This implies that a univariate analysis approach for classifier creation is not suitable, necessitating a multivariate approach. The limited strength of individual feature correlations underscores the need for constructing multivariate ML classification models to capture the intricate relationships within the dataset and achieve a more comprehensive understanding of the predictive factors associated with the target class.

Within this study, the efficacy of three supervised machine learning algorithms—XGB, RF, and a SVM—was assessed. The optimal classifier emerged as the one exhibiting the highest AUPRC, averaged across the 20 repetitions of the 5-fold cross-validation. As outlined in Table 3, the RF model proved to be the most proficient, excelling in terms of accuracy, precision and area under the precision-recall curve.

Figure 3 illustrates the RF classification model's performance, assessed through the Receiver Operating Characteristic (ROC) curve (Figure 3A), showcasing an Area Under the Curve (AUC) value of 0.699 ± 0.011 and through the Precision-Recall (PR) curve (Figure 3B) with an AUC of 0.668 ± 0.016. The plots showcase the average curves derived from 20 repetitions of the Cross-Validation, accompanied by their standard deviation.

In Supplementary Figures S2–S4, we present the analysis of parameter stability during the tuning phase of nested cross-validation. These figures illustrate, across multiple repetitions, the frequency with which a particular parameter was selected as the best parameter for our models. This in-depth examination provides valuable insights into the robustness and consistency of the chosen parameters throughout the nested cross-validation process.

4.3 Explainability

Model explainability involves understanding how algorithms discern the relationship between inputs and outputs. While complex non-linear models achieve superior performance, their interpretability is often compromised. This lack of interpretability limits their application in biomedical research, where a thorough understanding of the classification process is crucial. Feature importance methods aim to quantify the contribution of each feature to the model's predictions. Global methods provide an overarching ranking of features, while local methods illuminate the contribution of each feature to a specific prediction. In Figure 4, global feature importance is illustrated using various methods.

In Figure 4A, the Random Forest embedded feature importance is presented. The importance of a feature is computed as the (normalized) total reduction of the criterion brought about by that feature, commonly referred to as the Gini importance.

Figure 4B showcases the feature importance based on SHAP values. Essentially, this method constructs an interpretable linear model around each test instance and estimates feature importance at the local level. The plot in Figure 4B reveals the most important features for classification according to the SHAP algorithm. Shapley values are calculated by averaging across all iterations of the algorithm for each subject, considering the 20 repetitions. This summary plot provides an insightful overview of each feature's relative impact on the model's predictions, contributing to a

10.3389/fmicb.2024.1348974

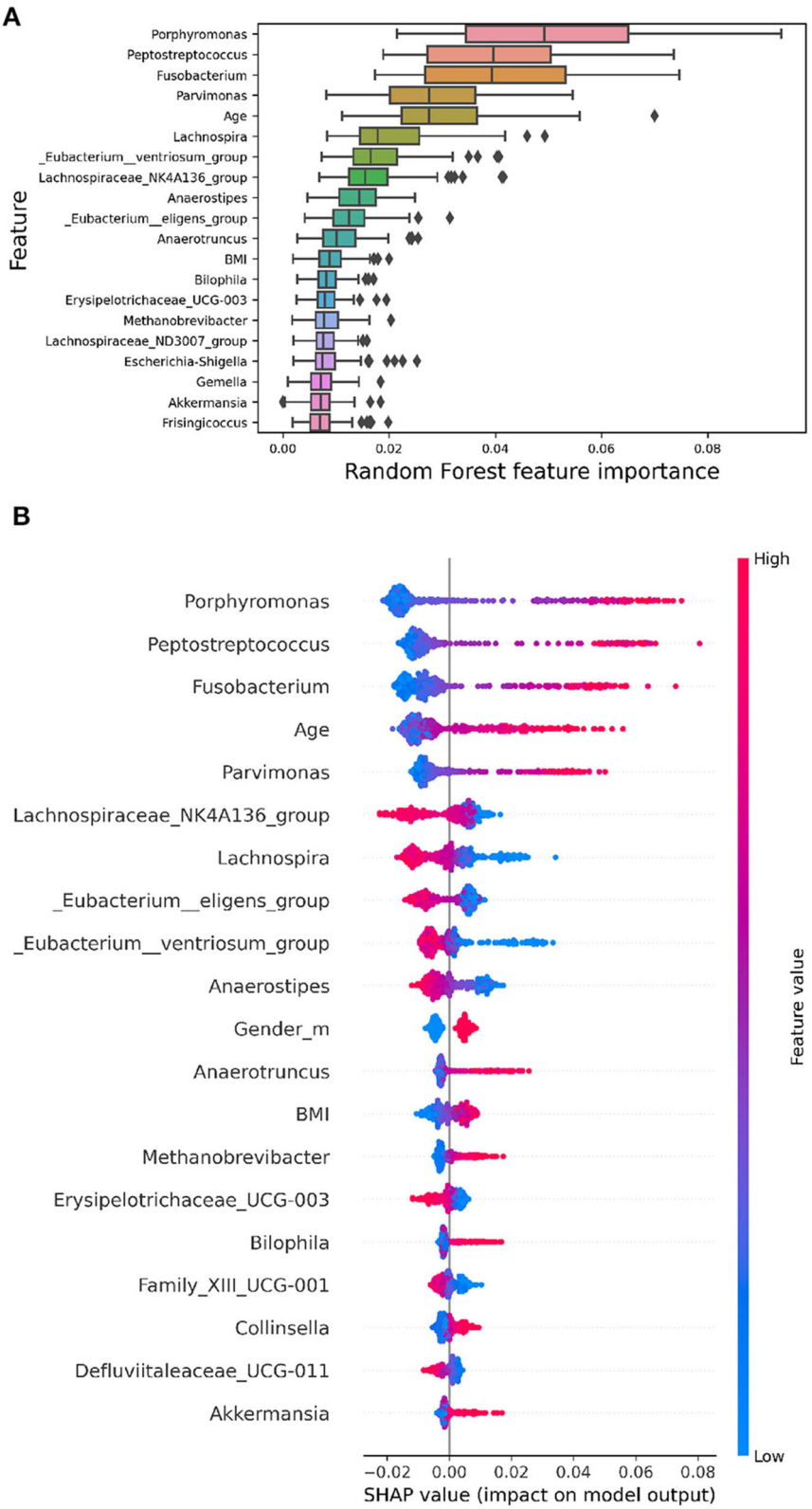

FIGURE 4
The images display the top 20 features ranked by their importance. **(A)** RF embedded feature importance. The boxplots represent the distributions of the feature importance coefficient calculated across all validation folds of the model. **(B)** SHAP summary plot depicting Shapley values for each feature. Each point represents a subject's Shapley value, with the y-axis indicating the corresponding feature and the x-axis representing the Shapley value. The color gradient reflects feature values, ranging from low to high, while features are ordered by mean importance, with more important features positioned toward the top.

thorough understanding of the overall importance and influence of different features in the analysis.

The Figure 4B indicates the presence of bacteria, such as *Porphiromonas*, with a high relative abundance (highlighted in red points on the summary plot) on the positive side of the x-axis, while a low relative abundance (highlighted in blue points) is more prevalent on the negative side. This suggests that a higher relative abundance of these bacteria is generally associated with a higher probability value for CRC, while a lower relative abundance is linked to a lower probability value for CRC. Conversely, bacteria like *Lachnospira* exhibit the opposite pattern, implying that a high abundance of this genus is correlated with a lower probability of CRC. These nuanced insights into the direction of effects are not discernible using global explanation methods like RF's built-in feature importance. Notably, the importance rankings of features obtained from both RF and SHAP values show substantial overlap (Jaccard Index = 0.67), highlighting the robustness and stability of the model. Furthermore, the SHAP summary plot highlights that among the top 20 most significant variables, Age, Gender, and BMI are included.

We have extended our explainability analysis to include the other two models (SVM and XGBoost). Due to computational constraints, we limited the number of repetitions for SVM to 5. The SHAP summary plots for these models are now available in the Supplementary Figure S5. Additionally the Table 4 illustrates the overlap coefficient (Vijaymeena and Kavitha, 2016) between the SHAP values of the three models. Notably, we observed a higher degree of overlap between the Shapley values of the two top-performing models, RF and XGBoost.

Figure 5 displays the dependence plots for the top two variables according to the SHAP summary plot. Notably, the dependence of marginal contributions for a specific variable varies with the fluctuations in the variable itself. Specifically, in the depicted dependence plots, an increase in the values of *Fusobacterium* (Figure 5A) or *Porphyromonas* (Figure 5B) corresponds to a rise in the associated SHAP values. Consequently, elevated values of these variables play a significant role in the algorithm's decision to classify an instance as CRC. Moreover, the color code represents the abundance of another bacterium. In Figures 5A, B can be observed the correlation of *Fusobacterium* with *Peptostreptococcus* and *Porphyromonas*, respectively.

TABLE 4 Overlap coefficient between the top 20 most important features, as determined by SHAP, across the three ML models.

RF	0.55	
XGBoost	0.40	0.75
	SVM	RF

5 Discussion

In our research, we have crafted an Artificial Intelligence workflow adept at deciphering the human microbiome within a cohort of control and CRC subjects, offering a highly dependable prediction of CRC outcomes. A notable strength lies in the entirely data-driven implementation of the classifier. Additionally, the preprocessing pipeline impartially eliminates less informative bacteria without relying on diagnostic labels associated with the microbiome. Beyond its precision, the top classifier yields predictions that are readily interpretable. XAI analysis results reveal a discernible pattern aligning with established knowledge, highlighting some bacterial genera among the 20 most significant features, known for their association with CRC in existing literature.

Among the foremost 20 features, *Fusobacterium*, *Porphyromonas*, *Peptostreptococcus*, and *Parvimonas* have emerged as potential microbiological markers that could significantly improve the accuracy of colorectal cancer (CRC) diagnoses (Chen et al., 2022).

Figure 5 offers insight into the connection between specific bacterial genera and CRC. The observed positive correlation between the relative abundance of well-documented bacteria like *Fusobacterium* and *Porphyromonas* and SHAP values suggests their influence on the model's predictions. This correlation hints at the biological relevance of these taxa in the context of CRC. Essentially, a higher abundance of these bacteria appears to positively impact the model's attribution of the positive class (cancer) during output explanation. The visual representation in Figure 5 aids in understanding the model's decision-making from a biological standpoint (Zhou et al., 2018; Koliarakis et al., 2019).

The recognition of abundant bacteria originating from the oral cavity, including *Fusobacterium*, *Peptostreptococcus*, and *Parvimonas*, indicates a dynamic symbiotic metacommunity intricately linked to the initiation of colorectal cancer (CRC). Within the human body, a symbiotic relationship with the microbiota exists, where polymicrobial communities inhabit cavities such as the oral and intestinal regions. Despite these areas being anatomically separated with distinct microbiota colonization, there are indications that bacteria from the oral cavity may migrate to the colon (Koliarakis et al., 2019). *Fusobacterium* has been associated with genetic and epigenetic abnormalities in colorectal cancer (CRC) tissues, including microsatellite instability (MSI). In the tumorigenesis and progression of CRC, Fusobacterium has the potential to enhance proliferation and metabolism, alter the immune microenvironment, and promote metastasis and chemoresistance. It may serve as a biomarker for identifying individuals at high risk for CRC (Wang and Fang, 2023).

According to our study, a high concentration of bacteria from the *Lachnospiraceae* family is associated with a lower likelihood of CRC. This spurious association has been observed in previous works, including (Hexun et al., 2023; Zhang et al., 2023), and this could be linked to the mechanism whereby a high concentration of these bacteria may promote heightened immune surveillance, thus controlling colorectal cancer progression and counteracting it.

Additionally, from the summary plot, we observe another pattern well-documented in the literature. There are studies indicating that certain bacteria of the *Clostridiales* order, including *Eubacterium eligens*, *Eubacterium ventriosum*, and *Anaerostipes*, are significantly reduced in CRC patients compared to control subjects (Montalban-Arques et al., 2021). This is evident in Figure 4B, where corresponding to these commensal bacteria, the high concentration of these bacteria (red points on the plot) is associated with negative SHAP values, indicating that

10.3389/fmicb.2024.1348974

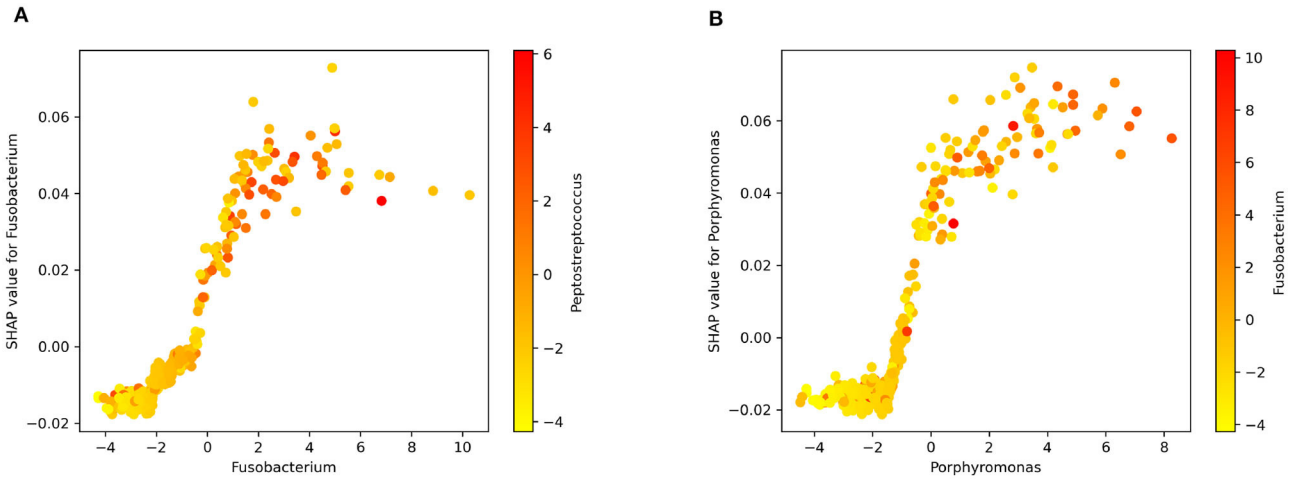

FIGURE 5
SHAP dependence plot for **(A)** *Fusobacterium* and *Peptostreptococcus*. **(B)** *Porphyromonas* and *Fusobacterium*.

the model assigns a low probability of classifying these subjects as CRC.

Regarding demographic descriptors, age, gender, and BMI have emerged as important features. Higher age, male gender, and elevated BMI appear to be positively associated with CRC. These findings are widely accepted and supported by scientific literature, where obesity is recognized as a factor associated with the development of this tumor, along with advancing age. Age exhibits a consistent trend with expected associations: longer lifespans correspond to a higher risk of having CRC (Murphy et al., 2011; Ye et al., 2020; Elangovan et al., 2021).

In addition to the strengths mentioned above, we performed a comprehensive analysis of explainability across the three models employed in our study. This analysis, as can be observed in Figure 4B and in Supplementary Figure S5, demonstrates the comparability of explainability results in terms of both the most important features and the correlation between feature values and their corresponding Shap values. Notably, the positive/negative correlations observed between SHAP values and the abundance of specific features persist consistently across all three models.

This consistency in the interpretability of our models enhances the robustness of our findings.

The presented study acknowledges certain limitations that we aim to address in future research efforts. While the classification performance provides valuable insights, there is the potential for further optimization. This could be attributed to the presence of other factors associated with colorectal cancer, such as hereditary factors and smoking, which were not considered in our analysis. Furthermore, the utilized database, obtained through 16S rRNA sequencing, provides a limited taxonomic resolution compared to Shotgun sequencing. A finer taxonomic resolution might have contributed to a more precise analysis and potentially identified stronger associations with the disease.

In the realm of CRC research, our study takes a distinctive approach by applying XAI techniques to unravel the intricate relationship between the human microbiome and CRC. Utilizing SHAP in microbiome research for predicting CRC outcomes enhances the transparency of our model and introduces a new perspective for the application of XAI in personalized medicine. Our identification of microbiological markers and taxonomic units associated with CRC risk contributes to the understanding of disease mechanisms and has the potential to inform diagnostic and therapeutic strategies. By acknowledging demographic descriptors alongside microbiome features, our work ensures a comprehensive approach that can be applicable across diverse patient populations. In recognizing the challenges and limitations of our study, we aim to guide future investigations, emphasizing our commitment to advancing both the scientific understanding of CRC and the practical applications of contemporary technologies.

6 Conclusion

This study has enabled the identification of bacteria that significantly influence the discrimination between healthy and diseased individuals through Explainable Artificial Intelligence (XAI), suggesting the identification of new disease biomarkers.

Additionally, the use of explainable artificial intelligence models can support making these models more transparent and interpretable, allowing for the appreciation, understanding, and utilization of the microbiota composition for each individual. By employing such the proposed method for each subject, an assessment of the microbiota can be conducted, with the aim of implementing actions to evaluate its modification, if necessary.

Data availability statement

Publicly available datasets were analyzed in this study. The datasets analyzed for this study can be found in the Zenodo repository (Marcos-Zambrano, 2022), and via https://github.com/pierfrancesco2021/XAI-for-Microbiome-Data-Analysis-in-CRC.

Author contributions

PN: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. DR: Data curation, Methodology, Writing – original draft. MM: Data curation, Methodology, Writing – original draft. PB: Writing – original draft, Writing – review & editing. DD: Writing – review & editing. AC: Formal analysis, Methodology, Writing – review & editing. GL: Writing – review & editing. DS: Writing – review & editing. VV: Writing – review & editing. PF: Writing – review & editing. RB: Supervision, Writing – review & editing. MD: Supervision, Writing – review & editing. FI: Writing – review & editing. ST: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Visualization, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was funded by the University of Bari, Poject XAI4Microbiome —Intelligenza Artificiale eXplainable per l'identificazione di marker metabolici personalizzati nella malattia di Behçet code S30—CUP H99J21017720005. The National Recovery and Resilience Plan (NRRP) Mission 4 Component 2 Investment 1.4—Call for tender no. 3138 of 16 December 2021 of Italian Ministry of University and Research funded by the European Union—NextGenerationEU. Award number: Project code: CN00000013, Concession decree no. 1031 of 17 February 2022 adopted by the Italian Ministry of University and Research, CUP H93C22000450007, Project title "National Center for HPC, Big Data and Quantum Computing" support this project. Authors would like to thank the resources made available by ReCaS, a project funded by the MIUR (Italian Ministry for Education, University and Research) in the "PON Ricerca e Competitivita' 2007-2013-Azione I-Interventi di rafforzamento strutturale" *PONa3* 00052, Avviso 254/Ric, University of Bari.

Acknowledgments

We greatly thank COST Action ML4Microbiome "Statistical and machine learning techniques in human microbiome studies" (CA18131), supported by COST (European Cooperation in Science and Technology, www.cost.eu).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2024.1348974/full#supplementary-material

References

Aitchison, J. (1982). The statistical analysis of compositional data. *J. Royal Stat. Soc. Series B* 44, 139–160. doi: 10.1111/j.2517-6161.1982.tb01195.x

Amodeo, I., De Nunzio, L., Raffaeli, G., Borzani, G., Griggio, I., Conte, A., et al. (2021). A machine and deep learning approach to predict pulmonary hypertension in newborns with congenital diaphragmatic hernia (clannish): protocol for a retrospective study. *Plos ONE* 16, 724. doi: 10.1371/journal.pone.0259724

Baxter, N. T., Ruffin, M. T., Rogers, M. A., and Schloss, P. D. (2016). Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. *Genome Med.* 8, 1–10. doi: 10.1186/s13073-016-0290-3

Bellando-Randone, S., Russo, E., Venerito, V., Matucci-Cerinic, M., Iannone, F., Tangaro, S., et al. (2021). Exploring the oral microbiome in rheumatic diseases, state of art and future prospective in personalized medicine with an ai approach. *J. Pers. Med.* 11, 625. doi: 10.3390/jpm11070625

Bellantuono, L., Tommasi, R., Pantaleo, E., Verri, M., Amoroso, N., Crucitti, P., et al. (2023). An explainable artificial intelligence analysis of raman spectra for thyroid cancer diagnosis. *Sci. Rep.* 13, 16590. doi: 10.1038/s41598-023-43856-7

Bergstra, J., and Bengio, Y. (2012). Random search for hyper-parameter optimization. *J. Mach. Learn. Res.* 13, 1–27.

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Cao, Q., Sun, X., Rajesh, K., Chalasani, N., Gelow, K., Katz, B., et al. (2021). Effects of rare microbiome taxa filtering on statistical analysis. *Front. Microbiol.* 11, 607325. doi: 10.3389/fmicb.2020.607325

Chen, H., Jiao, J., Wei, M., Jiang, X., Yang, R., Yu, X., et al. (2022). Metagenomic analysis of the interaction between the gut microbiota and colorectal cancer: a paired-sample study based on the gmrepo database. *Gut Pathogens* 14, 48. doi: 10.1186/s13099-022-00527-8

Chen, T., and Guestrin, C. (2016). "XGBoost: A scalable tree boosting system," in *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, CA), 785–794. doi: 10.1145/2939672.2939785

Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.* 20, 273–297. doi: 10.1007/BF00994018

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., and Barcelo-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematic. Geol.* 35, 279–300. doi: 10.1023/A:1023818214614

Elangovan, A., Skeans, J., Landsman, M., Ali, S. M., Elangovan, A. G., Kaelber, D. C., et al. (2021). Colorectal cancer, age, and obesity-related comorbidities: a large database study. *Dig. Dis. Sci.* 66, 3156–3163. doi: 10.1007/s10620-020-06602-x

Golob, J. L., Oskotsky, T. T., Tang, A. S., Roldan, A., Chung, V., Ha, C. W., et al. (2024). Microbiome preterm birth DREAM challenge: crowdsourcing machine learning approaches to advance preterm birth research. *Cell Rep. Med.* 5, 101350. doi: 10.1016/j.xcrm.2023.101350

Hexun, Z., Miyake, T., Maekawa, T., Mori, H., Yasukawa, D., Ohno, M., et al. (2023). High abundance of lachnospiraceae in the human gut microbiome is related to high immunoscores in advanced colorectal cancer. *Cancer Immunol. Immunother.* 72, 315–326. doi: 10.1007/s00262-022-03256-8

Ibrahimi, E., Lopes, M. B., Dhamo, X., Simeon, A., Shigdel, R., Hron, K., et al. (2023). Overview of data preprocessing for machine learning applications in human microbiome research. *Front. Microbiol.* 14, 1250909. doi: 10.3389/fmicb.2023.1250909

Janzing, D., and Minorics, L., and Blobaum, P. (2020). "Feature relevance quantification in explainable AI: a causal problem," in *International Conference on Artificial Intelligence and Statistics (PMLR)*. Breckenridge, CL, PMLR, 2907–2916

Koliarakis, I., Messaritakis, I., Nikolouzakis, T. K., Hamilos, G., Souglakos, J., Tsiaoussis, J., et al. (2019). Oral bacteria and intestinal dysbiosis in colorectal cancer. *Int. J. Mol. Sci.* 20, 4146. doi: 10.3390/ijms20174146

Lombardi, A., Diacono, D., Amoroso, N., Monaco, A., Tavares, J. M. R., Bellotti, R., et al. (2021a). Explainable deep learning for personalized age prediction with brain morphology. *Front. Neurosci.* 15, 578. doi: 10.3389/fnins.2021.674055

Lombardi, A., Tavares, J. M. R., and Tangaro, S. (2021b). Explainable artificial intelligence (xai) in systems neuroscience. *Front. Syst. Neurosci.* 15, 766980. doi: 10.3389/fnsys.2021.766980

Löwenmark, T., Löfgren-Burström, A., Zingmark, C., Eklöf, V., Dahlberg, M., Wai, S. N., et al. (2020). Parvimonas micra as a putative non-invasive faecal biomarker for colorectal cancer. *Sci. Rep.* 10, 15250. doi: 10.1038/s41598-020-72132-1

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., et al. (2020). From local explanations to global understanding with explainable ai for trees. *Nat. Mach. Int.* 2, 56–67. doi: 10.1038/s42256-019-0138-9

Lundberg, S. M., and Lee, S. I. (2017). A unified approach to interpreting model predictions. *Adv. Neural Inf. Proc. Syst.* 30, 1–14.

Marcos-Zambrano, L. J. (2022). *16S rRNA sequencing gene datasets for CRC data (1.0.0)* [Data set]. Zenodo. doi: 10.5281/zenodo.7382814

Montalban-Arques, A., Katkeviciute, E., Busenhart, P., Bircher, A., Wirbel, J., Zeller, G., et al. (2021). Commensal clostridiales strains mediate effective anti-cancer immune response against solid tumors. *Cell Host Microbe* 29, 1573–1588. doi: 10.1016/j.chom.2021.08.001

Morgan, E., Arnold, M., Gini, A., Lorenzoni, V., Cabasag, C., Laversanne, M., et al. (2023). Global burden of colorectal cancer in 2020 and 2040: Incidence and mortality estimates from globocan. *Gut* 72, 338–344. doi: 10.1136/gutjnl-2022-327736

Murphy, G., Devesa, S. S., Cross, A. J., Inskip, P. D., McGlynn, K. A., Cook, M. B., et al. (2011). Sex disparities in colorectal cancer incidence by anatomic subsite, race and age. *Int. J. Cancer* 128, 1668–1675. doi: 10.1002/ijc.25481

Novielli, P., Romano, D., Magarelli, M., Diacono, D., Monaco, A., Amoroso, N., et al. (2023). Personalized identification of autism-related bacteria in the gut microbiome using explainable artificial intelligence. *Preprint*. doi: 10.21203/rs.3.rs-3519546/v1

Ozenne, B., Subtil, F., and Maucort-Boulch, D. (2015). The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. *J. Clin. Epidemiol.* 68, 855–859. doi: 10.1016/j.jclinepi.2015.02.010

Papoutsoglou, G., Tarazona, S., Lopes, M. B., Klammsteiner, T., Ibrahimi, E., Eckenberger, J., et al. (2023). Machine learning approaches in microbiome research: challenges and best practices. *Front. Microbiol.* 14, 1261889. doi: 10.3389/fmicb.2023.1261889

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* 12, 2825–2830. Available online at: https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

Rynazal, R., Fujisawa, K., Shiroma, H., Salim, F., Mizutani, S., Shiba, S., et al. (2023). Leveraging explainable ai for gut microbiome-based colorectal cancer classification. *Genome Biol.* 24, 1–13. doi: 10.1186/s13059-023-02858-4

Schaffer, C. (1993). Selecting a classification method by cross-validation. *Mach. Learn.* 13, 135–143. doi: 10.1007/BF00993106

Varma, S., and Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinf.* 7, 1–8. doi: 10.1186/1471-2105-7-91

Venerito, V., Lopalco, G., Abbruzzese, A., Colella, S., Morrone, M., Tangaro, S., et al. (2022). A machine learning approach to predict remission in patients with psoriatic arthritis on treatment with secukinumab. *Front. Immunol.* 13, 3196. doi: 10.3389/fimmu.2022.917939

Vijaymeena, M., and Kavitha, K. (2016). A survey on similarity measures in text mining. *Mach. Learn. Appl. Int. J.* 3, 19–28. doi: 10.5121/mlaij.2016.3103

Wang, N., and Fang, J. Y. (2023). Fusobacterium nucleatum, a key pathogenic factor and microbial biomarker for colorectal cancer. *Trends Microbiol.* 31, 159–172. doi: 10.1016/j.tim.2022.08.010

Wu, Y., Jiao, N., Zhu, R., Zhang, Y., Wu, D., and Wang, A. J., et al. (2021). Identification of microbial markers across populations in early detection of colorectal cancer. *Nat. Commun.* 12, 3063. doi: 10.1038/s41467-021-23265-y

Yachida, S., Mizutani, S., Shiroma, H., Shiba, S., Nakajima, T., Sakamoto, T., et al. (2019). Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. *Nat. Med.* 25, 968–976. doi: 10.1038/s41591-019-0458-7

Ye, P., Xi, Y., Huang, Z., and Xu, P. (2020). Linking obesity with colorectal cancer: epidemiology and mechanistic insights. *Cancers* 12, 1408. doi: 10.3390/cancers12061408

Zackular, J. P., Rogers, M. A., and Ruffin, I. V. M. T., and Schloss, P. D. (2014). The human gut microbiome as a screening tool for colorectal cancer. *Cancer Prev. Res.* 7, 1112–1121. doi: 10.1158/1940-6207.CAPR-14-0129

Zeller, G., Tap, J., Voigt, A. Y., Sunagawa, S., Kultima, J. R., Costea, P. I., et al. (2014). Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.* 10, 766. doi: 10.15252/msb.20145645

Zhang, X., Yu, D., Wu, D., Gao, X., Shao, F., Zhao, M., et al. (2023). Tissue-resident lachnospiraceae family bacteria protect against colorectal carcinogenesis by promoting tumor immune surveillance. *Cell Host Microbe* 31, 418–432. doi: 10.1016/j.chom.2023.01.013

Zhou, Z., Chen, J., Yao, H., and Hu, H. (2018). Fusobacterium and colorectal cancer. *Front. Oncol.* 8, 371. doi: 10.3389/fonc.2018.00371

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 03 June 2024
DOI 10.3389/fmicb.2024.1393243

OPEN ACCESS

EDITED BY
Domenica D'Elia,
National Research Council (CNR), Italy

REVIEWED BY
Giovanni Luca Christian Masala,
University of Kent, United Kingdom
Maria Colomba Comes,
National Cancer Institute Foundation (IRCCS), Italy

*CORRESPONDENCE
Sabina Tangaro
✉ sabina.tangaro@uniba.it

RECEIVED 28 February 2024
ACCEPTED 13 May 2024
PUBLISHED 03 June 2024

CITATION
Magarelli M, Novielli P, De Filippis F, Magliulo R, Di Bitonto P, Diacono D, Bellotti R and Tangaro S (2024) Explainable artificial intelligence and microbiome data for food geographical origin: the Mozzarella di Bufala Campana PDO Case of Study. *Front. Microbiol.* 15:1393243. doi: 10.3389/fmicb.2024.1393243

Explainable artificial intelligence and microbiome data for food geographical origin: the Mozzarella di Bufala Campana PDO Case of Study

Michele Magarelli[1], Pierfrancesco Novielli[1,2], Francesca De Filippis[3], Raffaele Magliulo[3], Pierpaolo Di Bitonto[1], Domenico Diacono[2], Roberto Bellotti[2,4] and Sabina Tangaro[1,2]*

[1]Dipartimento di Scienze del Suolo, della Pianta e degli Alimenti, Università degli Studi di Bari Aldo Moro, Bari, Italy, [2]Istituto Nazionale di Fisica Nucleare, Sezione di Bari, Bari, Italy, [3]Dipartimento di Agraria, Università degli Studi di Napoli Federico II, Naples, Italy, [4]Dipartimento Interateneo di Fisica M. Merlin, Università degli Studi di Bari Aldo Moro, Bari, Italy

Identifying the origin of a food product holds paramount importance in ensuring food safety, quality, and authenticity. Knowing where a food item comes from provides crucial information about its production methods, handling practices, and potential exposure to contaminants. Machine learning techniques play a pivotal role in this process by enabling the analysis of complex data sets to uncover patterns and associations that can reveal the geographical source of a food item. This study aims to investigate the potential use of explainable artificial intelligence for identifying the food origin. The case of study of Mozzarella di Bufala Campana PDO has been considered by examining the composition of the microbiota in each samples. Three different supervised machine learning algorithms have been compared and the best classifier model is represented by Random Forest with an Area Under the Curve (AUC) value of 0.93 and the top accuracy of 0.87. Machine learning models effectively classify origin, offering innovative ways to authenticate regional products and support local economies. Further research can explore microbiota analysis and extend applicability to diverse food products and contexts for enhanced accuracy and broader impact.

KEYWORDS

explainable artificial intelligence, machine learning, microbiome, food origin, PDO

1 Introduction

With the burgeoning demand for high-quality, region-specific products, the need to ensure the origin and treceability of food products plays a pivotal role in ensuring authenticity, quality, and safety in the global food supply chain (Gallo et al., 2021). The concepts of food traceability and origin are closely interlinked and hold pivotal significance in ensuring food safety and transparency throughout the production process but also supports local economies and encourages sustainable agricultural practices. They are integral in guaranteeing that foods are safe, genuine, and adhere to quality standards. Traceability refers to the ability to follow the journey of a product along the entire supply chain, encompassing detailed information about its production, processing, packaging, distribution, and sale (del Rio-Lavín et al., 2023).

On the other hand, the origin of food products indicates the specific location where they were cultivated, manufactured, or processed. Understanding the origin of a food item is essential for various reasons, including ensuring its safety, quality, and sustainability. Presently, determining the origin of a food product relies on diverse methods and tools. Collaboration among producers, distributors, and other stakeholders in the supply chain is crucial to ensuring transparency and accuracy in disclosing the origin of food products (Corallo et al., 2020). Some food products may acquire origin certifications, such as the Protected Designation of Origin (PDO) in Europe or other regional certifications, which verify that the product originates from a specific geographical area and complies with designated standards (Badia-Melis et al., 2015). Analyzing the intricate ecosystem of microorganisms inhabiting food, known as the food microbiota, can be a useful tool for understanding the safety, quality, and characteristics of food products of foods. This diverse microbial community, comprising bacteria, fungi, and viruses, is influenced by various factors such as geographical location, production methods, and processing techniques. A fundamental aspect of harnessing the food microbiota for product origin lies in its dynamic composition, which reflects the unique environmental conditions and production practices of each food item. By scrutinizing the microbiota composition of food samples, distinctive microbial signatures indicative of their origin or production environment can be discerned. Recent advancements in molecular biology and sequencing technologies have revolutionized our ability to characterize the food microbiota with unprecedented precision and speed. High-throughput sequencing methods, including next-generation sequencing, facilitate rapid and accurate identification of microbial species present in food samples (Reuter et al., 2015). Comparative analysis of microbiota profiles among different food samples enables the identification of subtle variations that serve as valuable markers for product origin. Specific microbial strains or community structures may be linked to particular regions or production facilities, offering distinctive identifiers for food products. Moreover, the food microbiota serves as a sentinel for monitoring food quality and safety along the supply chain (Guidone et al., 2016). Alterations in microbial composition or abundance can signal potential contamination or spoilage incidents, enabling prompt interventions to mitigate risks and uphold food safety standards. In addition to conventional laboratory techniques, emerging methodologies such as metagenomics and metatranscriptomics provide comprehensive insights into the food microbiota. These cutting-edge approaches enable holistic analysis of all microbial genetic material within a sample, facilitating deeper understanding of microbial dynamics and functions (Cao et al., 2021). The use of machine learning in food classification and origin represents a significant step forward in ensuring the safety and authenticity of food products. Firstly, machine learning enables the development of predictive models that can differentiate between different types of foods based on specific characteristics. By leveraging machine learning algorithms, it becomes possible to process vast amounts of data, including information on production practices, environmental factors, and biochemical compositions, to accurately predict the origin of a food product. For example, using data from chemical, sensory, or genetic analyses, models can be trained to recognize the presence of contaminants or identify the geographical origin of a food. Furthermore, the application of machine learning to food classification offers numerous opportunities to enhance food safety, ensure product authenticity, and optimize the identification of food origin. The integration of machine learning and microbiota offers an innovative approach to understanding the complexity of interactions between the microbiome and food. By analyzing microbiome data using machine learning algorithms, it is possible to identify patterns and associations that can be valuable for enabling the develop preventive strategies to reduce risks and improve the nutritional quality of foods. The application of machine learning techniques in the field of food microbiota presents multiple opportunities to analyze large amounts of microbiological data, identify patterns and associations between microbial composition and food characteristics, predict food quality and safety, to understanding microbial dynamics and search for solutions to promote health (Bellantuono et al., 2023; Papoutsoglou et al., 2023). Through data analysis and the development of predictive models, crucial challenges in the food industry can be addressed, promoting greater transparency and trust among consumers. Explainable Artificial Intelligence (XAI) algorithms are useful to make artificial intelligence (AI) models understandable and interpretable to humans, because many machine learning and AI models often operate as "black boxes," making it difficult to understand how and why they produce certain predictions or decisions. The goal of XAI is to provide explanations and insights into the operation of AI models, enabling users to understand the reasons behind their predictions or decisions. This is particularly important in contexts where transparency, accountability, and trust in AI are crucial. In Explainable Artificial Intelligence (XAI), trustworthiness plays a role in ensuring the reliability and transparency of AI models. It refers to the degree of confidence and faith users have in the explanations provided by the model regarding its predictions and decision-making processes. XAI techniques may include SHapley Additive exPlanations (SHAP) analysis that seek to translate the internal workings of AI models into understandable human explanations (Novielli et al., 2024). This research delves into the crucial realm of preserving and authenticating the geographical origin of Mozzarella di Bufala Campana PDO, specifically focusing on the provinces of Salerno and Caserta. The characteristic that will be used for data analysis is the abundance of bacteria present in the microbiota of the samples. This information will be crucial for identifying any patterns or correlations between bacterial composition and the geographical origin of Mozzarella di Bufala PDO. By utilizing data analysis techniques such as machine learning (Monaco et al., 2021; Papoutsoglou et al., 2023), it will be possible to create predictive models capable of accurately classifying the geographical origin of each sample based on microbiota information. This approach will provide a trustworthy assessment of the mozzarella's origins, thereby contributing to food quality and safety.

2 Materials

The data utilized in this study, decripted in Table 1 stems from the microbiological analysis of the microbiome of 65 samples of Mozzarella di Bufala PDO originating from 30 dairies in the province of Salerno and 35 dairies in the province of Caserta. These samples underwent thorough examination in the laboratories of the Microbiology Division within the Department of Agricultural Sciences at the University of Naples Federico II. All dairies were located within the PDO area produced traditional Mozzarella di Bufala according to the PDO regulation. Total DNA was extracted using the Qiagen Power Soil Pro kit. Metagenomic libraries were prepared using the Nextera XT Index Kit (Illumina, San Diego, California, United States), then whole metagenome sequencing was performed on an Illumina NovaSeq platform, leading to 2 × 150 bp, paired-end reads. Reads were quality-checked and filtered through Prinseq-lite v. 0.20.4, using parameters "-trim_qual_right 5" and "-min_len 60." An average of 25 M of paired-end reads were obtained (2 × 150 bp) for each sample. Raw reads were pre-processed and filtered as previously described (De Filippis et al., 2021). Briefly, contamination from host reads was removed using the Human Sequence Removal pipeline developed within the Human Microbiome Project by using the Best Match Tagger (BMtagger) mapping reads against the *Bubalus bubalis* (Mediterranean breed) genome (accession number: GCA003121395.1). Then, non-host reads were quality-filtered using PRINSEQ v. 0.20.4 (Schmieder and Edwards, 2011). Bases having a Phred score <15 were trimmed and those <75 bp were discarded. High-quality reads were further processed to obtain microbiome taxonomic profiles using MetaPhlAn v. 4.0 (Blanco-Míguez et al., 2023).

Our analysis encompasses a diverse set of samples, reflecting the regional diversity of Mozzarella di Bufala PDO production across different dairies in the provinces of Salerno and Caserta. The 65 samples provide a robust dataset for investigating variations in microbial composition, offering valuable insights into the distinctive qualities of Mozzarella di Bufala PDO from different geographic origins. The species abundance data unveils the relative prevalence of microbial species, offering insights into the intricate microbiome of Mozzarella di Bufala PDO. This information is organized in a tabular format, where each row corresponds to a specific sample, and each column represents a distinct microbial species. To enhance our understanding of the origin of each Mozzarella di Bufala PDO sample, we include details about the respective cheese dairy, specifying both the dairy name and its geographic origin. Each sample presents 139 output variables, each representing the abundance of a specific bacterium. In the context of your analysis on the microbiome of Mozzarella di Bufala PDO, these output variables likely reflect the proportions or relative quantities of different types of bacteria present in each sample. The type of bacteria and their relative abundance in each sample could have significant implications for the quality and sensory characteristics of the product. Since many samples have abundance values equal to zero, indicating the absence of the bacteria, a preprocessing step was performed. In this pre-processing step, columns with more than 70% zero values were removed, reducing the total number of columns to 23. In order to conduct a robust analysis, the initial dataset has been strategically

TABLE 1 Description of samples and input variables.

Type of samples	Diary from Campania region
n samples from Salerno	30
n samples from Caserta	35
Type of input variables	Microbiome relative abundance
n input variables for each sample	139

partitioned into a validation dataset and a test dataset to. This partitioning is designed to ensure a representative and unbiased evaluation of the models developed during the study (Ibrahimi et al., 2023). The validation dataset consists of 22 samples from the province of Salerno and 33 samples from the province of Caserta. This division allows for the exploration of regional variations within the microbiome of Mozzarella di Bufala PDO, considering the distinctive characteristics of these geographical locations. The validation set was then used to assess three different classifiers through a five-fold cross-validation repeated 20 times (Schaffer, 1993), and the performance of the best classifier (Random Forest, RF) was analyzed. Following that, the trained model was tested on the test dataset, and its performance was evaluated on this separate set of samples.

The independent test dataset, on the other hand, comprises eight samples from Salerno and two samples from Caserta. Notably, these 10 test samples are collected on the same day from the same dairy as the samples present in the validation set. By adopting this partitioning strategy, we aim to develop a model that not only captures the nuances of the training dataset but also demonstrates robust predictive abilities when faced with previously unseen samples.

3 Methods

The main steps of our analysis are outlined in the flowcharts in Figure 1. It provides a comprehensive overview of the model's performance during both the training and validation phases, as well as in the subsequent testing phase, allowing for an overall evaluation of its predictive capabilities.

3.1 Machine learning based classification

To assess the classification of these samples, three distinct supervised machine learning methods were employed: Random Forest, XGBoost, and Multi-Layer Perceptron (MLP). The identification of the optimal classifier was based on both accuracy and Area Under the Curve (AUC).

3.1.1 Random forest classifier

The Random Forest Classifier represents a sophisticated ensemble learning algorithm within the realm of machine learning (Chaudhary et al., 2016). Envisioned as a confluence of decision

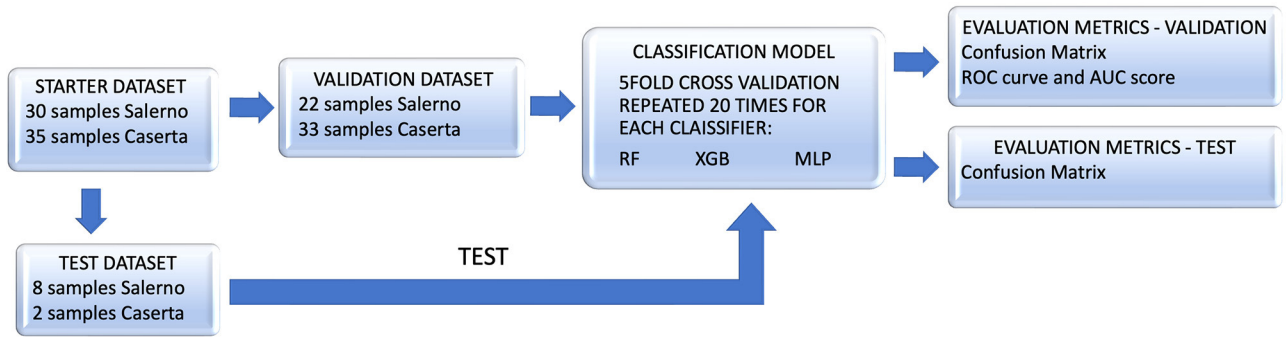

FIGURE 1
The flowchart outlines the steps of the conducted analysis. The validation set was used to assess three different classifiers through a five-fold repeated 20 times cross-validation, and the performance of the best classifier (Random Forest, RF) was analyzed. Following that, the trained model was tested on the test dataset, and its performance was evaluated on this separate set of samples.

trees, it operates on the principle of aggregating predictions from diverse models to augment stability and overall performance. The ensemble is constituted by an assembly of decision trees, each meticulously trained on a distinct subset of the training dataset through the lens of bootstrap sampling a method characterized by its sampling with replacement. The algorithm's efficacy is derived from the varied nature of decision trees. This diversity, arising from the differential subsets of data upon which each tree is trained, mitigates the risk of overfitting, fostering a robust model. In the predictive phase, each decision tree contributes its prediction, and the final class is determined through a majoritarian consensus. This collective decision-making process amplifies the model's resilience and generalization capabilities (Breiman, 2001).

3.1.2 EXtreme gradient boosting classifier

EXtreme Gradient Boosting (XGBoost) is a widely-used machine learning algorithm for regression and classification problems renowned for its prowess in diverse applications, particularly excelling in the realm of structured or tabular data and supervised learning scenarios (Shwartz-Ziv and Armon, 2022). XGBoost has been extensively used in data science and machine learning competitions due to its ability to achieve excellent performance on a wide range of problems and datasets. It's also known for its flexibility and ability to handle large amounts of data. Positioned within the domain of ensemble learning, XGBoost elevates traditional gradient boosting algorithms to new heights. XGBoost typically builds an ensemble of decision trees, where each tree contributes to the final prediction. The combination of multiple trees enhances the model's predictive capabilities. XGBoost supports built-in cross-validation, enabling robust model evaluation and parameter tuning for optimal performance. XGBoost has an high predictive accuracy. By constructing an ensemble of models, each correcting the errors of the others, it can provide more accurate predictions compared to many other algorithms. It also incorporates regularization techniques that help manage the issue of overfitting, keeping the model general and adaptable to new data (Chen and Guestrin, 2016).

3.1.3 Multi-layer perceptron classifier

The Multi-Layer Perceptron (MLP) stands as a sophisticated architecture within the domain of artificial neural networks, prominently featured in the landscape of machine learning. It is distinguished by its layered composition, comprising an input layer, one or more hidden layers, and an output layer. Each layer encompasses interconnected nodes, or artificial neurons, where the transmission of information follows a feedforward trajectory, progressing from the input layer through the hidden layers and culminating in the output layer. In a Multi-Layer Perceptron (MLP), input nodes constitute the initial layer of the neural network and serve as the units through which data is introduced into the system. Each input node represents a specific feature or variable from the dataset intended for model training. The hidden layers are intermediary layers between the input and output layers, responsible for capturing and learning complex patterns and representations within the input data. These layers contribute to the model's ability to discern intricate relationships that may not be immediately apparent in the raw features. Output nodes constitute the final layer of the neural network and are responsible for producing the model's predictions or outcomes. The configuration and characteristics of the output layer depend on the nature of the task, whether it involves classification, regression, or other specific objectives (Ruck et al., 1990).

3.2 Evaluation metrics

Evaluation metrics are crucial tools for assessing the performance and effectiveness of machine learning models (Ferrer, 2022). These metrics provide quantitative measures that help quantify how well a model is performing on a given task. The choice of evaluation metrics depends on the nature of the problem (classification, regression, etc.) and the specific goals of the analysis. Here are some commonly used evaluation metrics:

- Accuracy:

The proportion of correctly classified instances among the total instances

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

- Sensitivity:
  The fraction of true positive predictions out of all actual positive instances

$$SENS = \frac{TP}{TP + FN} \quad (2)$$

- Specificity:
  Specificity is the proportion of actual negatives correctly identified by the model out of the total number of actual negatives.

$$SPEC = \frac{TN}{FP + TN} \quad (3)$$

- Precision:
  The fraction of true positive predictions out of all positive predictions

$$PREC = \frac{TP}{TP + FP} \quad (4)$$

- Area Under the ROC Curve (AUC-ROC):
  The Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC) are assessment tools employed to gauge the effectiveness of a binary classification model. The ROC curve presents a graphical depiction of how sensitivity (true positives) and specificity (true negatives) change across various classification thresholds. Essentially, it illustrates the balance between accurately identifying positive and negative instances by the model. The AUC quantifies the overall performance of the model by measuring the area under the ROC curve: a value closer to 1 signifies superior model performance, while a value around 0.5 suggests random classification. In summary, these metrics are vital for evaluating and contrasting the classification ability of binary models (Ozenne et al., 2015).

3.3 Explainable artificial intelligence methods

Explainable Artificial Intelligence (XAI) is a crucial aspect in the development of AI systems, focused on making artificial intelligence (AI) models understandable and interpretable to humans. A specific method employed for XAI is the SHapley Additive exPlanations (SHAP) (Arrieta et al., 2020). SHAP values are used to evaluate the impact of individual features on the model's performance, particularly on a validation set. Mathematically, the SHAP value for a specific feature (j) is calculated based on the inclusion or exclusion of that feature from the model as:

$$\Phi_j(x) = \sum_{F \subseteq S - \{j\}} \frac{|F|!(|S| - |F| - 1)!}{|S|!} [f_x(F \cup j) - f_x(F)] \quad (5)$$

where $\Phi_j(x)$ represents the SHAP value of feature j for the prediction of the model f given the input x, S is the set of all features, $F \subseteq S - \{j\}$ represents all possible subsets of features excluding feature j, $\frac{|F|!(|S|-|F|-1)!}{|S|!}$ is a weight parameter that multiplies all of the permutations of S! by the potential permutations of the remaining class that doesn't belong to S, while $f_x(F \cup j)$ and $f_x(F)$ denote respectively the model's prediction when feature j is added to the subset F and when it is absent (Lundberg and Lee, 2017). We also averaged the ten realizations of SHAP values in order to obtain a single representative SHAP vector.

The SHAP value measures how much including feature j changes the model's prediction compared to the prediction without feature j, averaged over all possible combinations of features. Positive SHAP values indicate that the feature contributes positively to the prediction, while negative values indicate a negative contribution. The SHAP values provide a quantitative measure of the contribution of each feature to the model's output, enabling a more interpretable understanding of how individual features influence the algorithm's decision-making process. This transparency is crucial for building trust in AI systems and facilitating their use in various real-world applications where interpretability is essential (Janzing et al., 2020). This approach contributes to the trustworthiness and applicability of our findings, enhancing the overall validity of the study's outcomes in the context of Mozzarella di Bufala PDO from Salerno and Caserta.

4 Results

This study aims to investigate the potential use of explainable artificial intelligence for identifying the food origin. The case of study of Mozzarella di Bufala Campana PDO has been considered by examining the composition of the microbiota in 65 samples.

This study involved evaluating the effectiveness of three supervised machine learning algorithms, namely XGBoost, Random Forest, and a complex Multi-Layer Perceptron network. The analysis revealed that the Random Forest classifier outperformed the others, demonstrating the highest Area Under the Curve (AUC) value of 0.93 ± 0.10 and the top accuracy score of 0.87 ± 0.11. Table 2 provides a comprehensive comparison of the three models based on their AUC and accuracy scores.

4.1 Machine learning analysis

The results are illustrated in the confusion matrix in Table 3, obtained following a five-fold repeated 20 times cross-validation procedure on the validation set. This methodology allows us

TABLE 2 Comparison between evaluation metrics of XGBoost (XGB), Random Forest (RF), and Multi-Layer Perceptron (MLP) classifiers.

Classifier	Accuracy	AUC
XGB	0.82 ± 0.12	0.87 ± 0.11
RF	0.87 ± 0.11	0.93 ± 0.10
MLP	0.68 ± 0.13	0.78 ± 0.11

to assess the effectiveness of our algorithm in a robust and reliable manner. In Figure 2 it is possible to observe the boxplot displaying the trend evaluation metrics, including accuracy (Equation 1), specificity (Equation 3), sensitivity (Equation 2) and precision (Equation 4), obtained through a five-fold repeated cross-validation scheme.

TABLE 3 Confusion matrix depicts predicted values against actual values.

Actual class		Predicted class	
		Caserta	Salerno
	Caserta	29	4
	Salerno	3	19

In this instance, 29 samples from Caserta and 19 from Salerno are correctly classified, while four samples from Caserta and three from Salerno are misclassified.

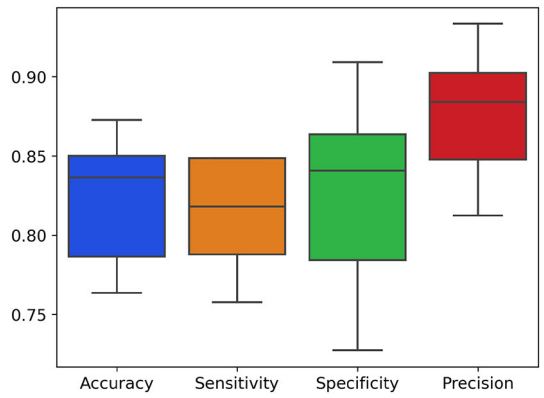

FIGURE 2
Boxplot of the distributions of evaluation metrics (accuracy, specificity, sensitivity and precision) following five-fold cross-validation repeated 20 times.

The confusion matrix highlights the algorithm's ability to correctly classify observations based on the geographical origin of the samples, divided between the Salerno and Caserta areas. We observe that the algorithm achieved an accuracy of 87.87% in correctly identifying samples from the Salerno area and 86.36% for those from the Caserta area. These results indicate a good capability of our machine learning model in distinguishing the geographical origin of Mozzarella di Bufala Campana PDO based on the microbiota structure. The accuracy in both cases is quite high, suggesting that the model generalizes well to new data and could be used as a supportive tool in determining the geographical origin of unknown samples.

The Receiver Operating Characteristic curve in the Figure 3 defines AUC score, measuring the area under this curve, is 0.93 ± 0.10 and it suggests a high accuracy in classifying samples based on their geographical origin, affirming the robustness of the model's performance.

After conducting cross-validation, the outcomes were then utilized to compute feature importance employing SHapley Additive exPlanations (SHAP), as expressed in Equation (5). The SHAP ranking plot is a graph that displays the importance of features in machine learning models using SHAP and features are arranged along the y-axis based on their importance, with the most important features at the top and the least important ones at the bottom. Each colored point represents a single data instance, and the horizontal position of the point indicates the value of the shap for that specific instance. The color of the point indicates the value of the feature: higher values are represented in warm colors (red), while lower values are represented in cool colors (blue). Through a SHAP analysis, the 20 most important feature were identified, deriving from the analysis of the microbiota 65 samples. In the

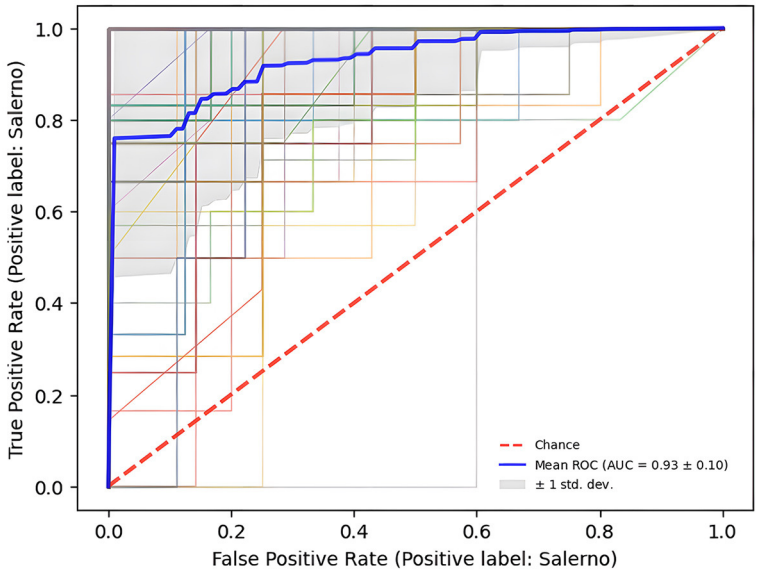

FIGURE 3
ROC curve depicts the classification model's ability to vary the trade-off between sensitivity (True Positive Rate) and specificity (1 – False Positive Rate).

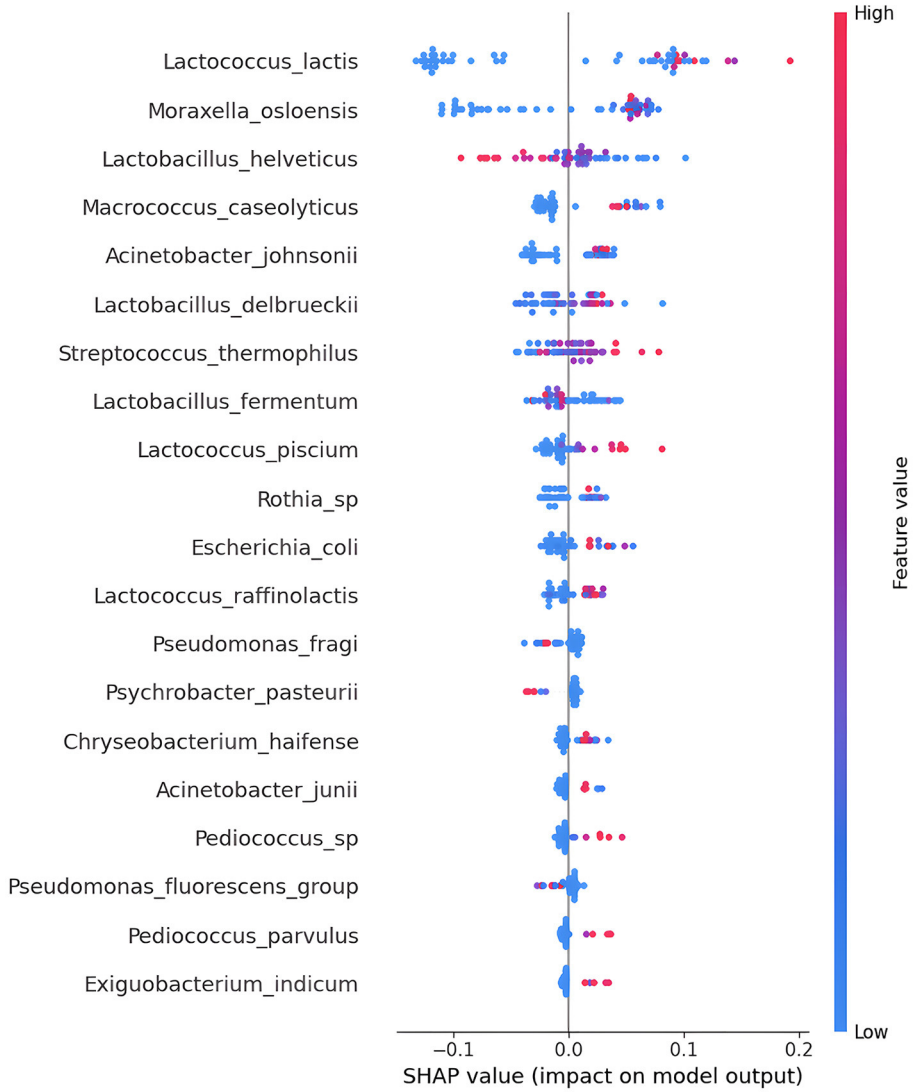

FIGURE 4
The SHapley Additive exPlanations (SHAP) summary plot provides an overview of the importance of features in contributing to model predictions. In this type of plot, each point represents a data instance, and the horizontal position of the point indicates how much the effect of a specific feature contributes to the change in prediction compared to the model's average prediction. The color of the point represents the value of the feature, with darker colors indicating higher values.

SHAP plot in Figure 4 it is evident how certain features, such as *Lactococcus lactis* and *Moraxella osloensis*, contribute significantly to the model's prediction. The feature *Lactobacillus helveticus* is important for the model's interpretability, as the colored points are well distinguished, and red points indicate that high values of that bacterium have influenced Salerno class, and vice versa. This suggests that these elements play a crucial role in the geographical discrimination of the samples.

The results of the Shap analysis highlight the fact that two Phyla are most represented (Firmicutes and Proteobacteria). The taxonomy of each sample of SHAP analysis is descripted in Table 4. Lactobacillaceae is represented by five bacteria, Moraxella family is represented by four bacteria, while Lactococcaceae family is represented by three bacteria. Starting from the taxonomic group of the genus, it can be seen that there is a significant diversity of microbes, even if the *Lactococcus* genus and *Lacotbacillus* genus is represented three times each other.

A possible application of the classification model is to execute it on the previously selected test dataset. In testing the model, a dataset consisting of 10 samples from the same study was utilized, including two from Caserta and eight from Salerno. These samples were previously excluded during the model training phase. The confusion matrix of the test, depicted in the figure, provides a detailed overview of the model's performance on this specific test dataset. It is particularly noteworthy that all samples from Caserta were correctly classified by the model. On the other hand, only one sample from Salerno was misclassified. This result suggests a significant accuracy in the model's ability to discriminate between the two production locations, with a particularly high success rate for samples from Caserta. The confusion matrix in Table 5 offers

TABLE 4 Classification of the first 20 bacteria deriving from the Shap analysis.

Phylum	Class	Order	Family	Genus	Species
Firmicutes	*Bacilli*	*Lactobacillales*	*Lactococcaceae*	*Lactococcus*	*Lactococcus lactis*
Proteobacteria	*Gammaproteobacteria*	*Pseudomonadales*	*Moraxellaceae*	*Moraxella*	*Moraxella osloensis*
Firmicutes	*Bacilli*	*Lactobacillales*	*Lactobacillaceae*	*Lactobacillus*	*Lactobacillus helveticus*
Firmicutes	*Bacilli*	*Bacillales*	*Staphylococcaceae*	*Macrococcus*	*Macrococcus caseolyticus*
Proteobacteria	*Gammaproteobacteria*	*Pseudomonadales*	*Moraxellaceae*	*Acinetobacter*	*Acinetobacter johnsonii*
Firmicutes	*Bacilli*	*Lactobacillales*	*Lactobacillaceae*	*Lactobacillus*	*Lactobacillus delbrueckii*
Firmicutes	*Bacilli*	*Lactobacillales*	*Streptococcaceae*	*Streptococcus*	*Streptococcus thermophilus*
Firmicutes	*Bacilli*	*Lactobacillales*	*Lactobacillaceae*	*Lactobacillus*	*Lactobacillus fermentum*
Firmicutes	*Bacilli*	*Lactobacillales*	*Lactococcaceae*	*Lactococcus*	*Lactococcus piscium*
Actinobacteria	*Actinobacteria*	*Micrococcales*	*Micrococcaceae*	*Rothia*	*Unclassified bacterium*
Proteobacteria	*Gammaproteobacteria*	*Enterobacterales*	*Enterobacteriaceae*	*Escherichia*	*Escherichia coli*
Firmicutes	*Bacilli*	*Lactobacillales*	*Lactococcaceae*	*Lactococcus*	*Lactococcus raffinolactis*
Proteobacteria	*Gammaproteobacteria*	*Pseudomonadales*	*Pseudomonadaceae*	*Pseudomonadaceae*	*Pseudomonadaceae fragi*
Proteobacteria	*Gammaproteobacteria*	*Pseudomonadales*	*Moraxellaceae*	*Psychrobacter*	*Psychrobacter pasteurii*
Bacteroidetes	*Flavobacteriia*	*Flavobacteriales*	*Weeksellaceae*	*Chryseobacterium*	*Chryseobacterium haifense*
Proteobacteria	*Gammaproteobacteria*	*Pseudomonadales*	*Moraxellaceae*	*Acinetobacter*	*Acinetobacter junii*
Firmicutes	*Bacilli*	*Lactobacillales*	*Lactobacillaceae*	*Pediococcus*	*Unclassified bacterium*
Proteobacteria	*Gammaproteobacteria*	*Pseudomonadales*	*Pseudomonadaceae*	*Pseudomonadaceae*	*Pseudomonadaceae fluorescens*
Firmicutes	*Bacilli*	*Lactobacillales*	*Lactobacillaceae*	*Pediococcus*	*Pediococcus parvulus*
Firmicutes	*Bacilli*	*Bacillales*	*Bacillaceae*	*Exiguobacterium*	*Exiguobacterium indicum*

The Phylum, Class, Order, Family, Genus and Species columns indicate the classification of each bacteria.

TABLE 5 Confusion matrix depicts predicted values against actual values.

Actual class		Predicted class	
		Caserta	Salerno
	Caserta	2	0
	Salerno	1	7

In this instance, seven samples from Salerno and two from Caserta are correctly classified, while only one sample from Salerno is misclassified.

a detailed assessment of the model's performance on the specific test dataset.

5 Discussion

Mozzarella di Bufala Campana PDO is a designation that certifies the mozzarella is produced in the Campania region, Italy, and follows traditional production methods and established quality standards to preserve its authenticity and excellence. The PDO protects the product name from imitations and assures buyers that they are purchasing a genuine product produced according to the traditional specifications of the designated area. Recognizing the correct origin is crucial to preserving the diversity and excellence of local productions. Protection against imitations and counterfeits, guaranteed by the PDO, helps maintain the product's reputation and preserves its cultural history. Ultimately, correctly identifying the origin of PDO mozzarella not only ensures product quality but also contributes to preserving the cultural and gastronomic heritage associated with this unique Italian specialty.

Indeed, the integration of machine learning (ML) and explainable artificial intelligence (XAI) techniques holds significant value in various contexts, including the analysis of biological data such as microbiota and metabolomics. Machine learning facilitates the creation of accurate predictive models based on microbiological data, aiding in the authentication and protection of PDO products like Mozzarella di Bufala Campana. XAI techniques ensure transparency and interpretability, reinforcing trust among consumers, regulators, and industry stakeholders. This combination not only enhances the certification of food origin but also strengthens the preservation of cultural and gastronomic heritage associated with traditional foods. Overall, microbiota analysis plays a vital role in ensuring the authenticity, quality, and safety of food products like Mozzarella di Bufala Campana PDO. In this study, each sample exhibits a relative abundance of various microbial species, which are not present in all samples. The most prevalent genera are *Pseudomonas*, *Lactobacillus*, *Streptococcus*, and *Acinetobacter*. The cheese-making process of Mozzarella di Bufala Campana is a combination of high-quality ingredients and specific procedures, with particular attention to the crucial role played by natural whey containing thermophilic lactic bacteria. The presence

of thermophilic lactic bacteria is interesting because they survive at high temperatures during the processing, thus contributing to the uniqueness of Mozzarella di Bufala Campana (Levante et al., 2023). The ecological complexity of these thermophilic lactic bacteria is an aspect that can be studied in detail to better understand the fermentation process and the production of this traditional cheese. Research conducted has shown that, despite ecological complexity, only certain thermophilic lactic acid bacteria (LAB), namely *Streptococcus thermophilus*, *Lactobacillus delbrueckii*, and *Lactobacillus helveticus*, are the main players in the curd fermentation. This is one of the peculiarities that helps preserve the unique characteristics of the cheese and protects local producers from imitations and counterfeits. It also assures buyers that they are purchasing an authentic and high-quality product, respecting the long history and reputation of Mozzarella di Bufala Campana as a traditional and artisanal product (Pisano et al., 2016).

6 Conclusion

This paper is an example of how an XAI analysis can be applied with trustworthiness in the context of discriminating the geographical origin of PDO Mozzarella di Bufala Campana based on microbiota bacterial abundance. This validates the approach employed in our study and confirms that certain bacteria can be considered reliable indicators of geographical origin. The predictive models developed using machine learning techniques have proven to be effective in classifying the geographical origin of mozzarella samples. These results provides strong support for food traceability, enabling consumers to make informed choices and ensuring that products are authentic and safe. The results obtained have significant implications for the food industry as they offer an innovative and reliable method to authenticate and protect high-quality regional products. This can contribute to strengthening consumer confidence in food products and supporting local economies through the promotion of sustainable agricultural practices. Further research could delve deeper into microbiota analysis and assess the effectiveness of other analytical techniques in improving the accuracy of predictions regarding the geographical origin of food products. Machine learning facilitates the creation of robust predictive models capable of accurately identifying the origin of food products based on microbiological data. Furthermore, XAI techniques provide transparency and interpretability, enabling stakeholders to understand how these models arrive at their conclusions. This combination not only ensures the trustworthiness of predictions but also fosters trust among consumers, regulators, and industry professionals. Moving forward, further research could delve deeper into microbiota analysis and explore the effectiveness of additional analytical techniques in enhancing the accuracy of predictions regarding the geographical origin of food products. Additionally, investigating the application of these approaches in diverse contexts and food products would expand the scope and applicability of our findings, driving continual advancements in food traceability and quality assurance practices.

Data availability statement

The data presented in the study are deposited in the Sequence Read Archive (SRA) database of the NCBI, accession numbers PRJNA1084214 and PRJNA997821.

Author contributions

MM: Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Formal analysis. PN: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Conceptualization. FD: Writing – review & editing, Validation, Investigation, Data curation. RM: Writing – review & editing, Data curation. PD: Writing – review & editing, Validation. DD: Writing – review & editing, Validation. RB: Writing – review & editing, Validation. ST: Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. METROFOOD-IT project has received funding from the European Union—NextGenerationEU, PNRR—Mission 4 "Education and Research" Component 2: from research to business, Investment 3.1: Fund for the realization of an integrated system of research and innovation infrastructures - IR0000033 (D.M. Prot. n.120 del 21/06/2022).

Acknowledgments

Authors would like to thank the resources made available by ReCaS, a project funded by the MIUR (Italian Ministry for Education, University and Re- 270 search) in the "PON Ricerca e Competitivit'a 2007–2013-Azione I-Interventi di rafforzamento strutturale" PONa3 00052, Avviso 254/Ric, University of Bari.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

References

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., et al. (2020). Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inform. Fusion* 58, 82–115. doi: 10.1016/j.inffus.2019.12.012

Badia-Melis, R., Mishra, P., and Ruiz-García, L. (2015). Food traceability: new trends and recent advances. A review. *Food Control* 57, 393–401. doi: 10.1016/j.foodcont.2015.05.005

Bellantuono, L., Tommasi, R., Pantaleo, E., Verri, M., Amoroso, N., Crucitti, P., et al. (2023). An explainable artificial intelligence analysis of Raman spectra for thyroid cancer diagnosis. *Sci. Rep*. 13:16590. doi: 10.1038/s41598-023-43856-7

Blanco-Míguez, A., Beghini, F., Cumbo, F., McIver, L. J., Thompson, K. N., Zolfo, M., et al. (2023). Extending and improving metagenomic taxonomic profiling with uncharacterized species using metaphlan 4. *Nat. Biotechnol*. 41, 1633–1644. doi: 10.1038/s41587-023-01688-w

Breiman, L. (2001). Random forests. *Mach. Learn*. 45, 5–32. doi: 10.1023/A:1010933404324

Cao, Q., Sun, X., Rajesh, K., Chalasani, N., Gelow, K., Katz, B., et al. (2021). Effects of rare microbiome taxa filtering on statistical analysis. *Front. Microbiol*. 11:607325. doi: 10.3389/fmicb.2020.607325

Chaudhary, A., Kolhe, S., and Kamal, R. (2016). An improved random forest classifier for multi-class classification. *Inf. Process. Agric*. 3, 215–222. doi: 10.1016/j.inpa.2016.08.002

Chen, T., and Guestrin, C. (2016). "Xgboost: a scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (New York, NY: ACM), 785–794. doi: 10.1145/2939672.2939785

Corallo, A., Latino, M. E., Menegoli, M., and Striani, F. (2020). The awareness assessment of the italian agri-food industry regarding food traceability systems. *Trends Food Sci. Technol*. 101, 28–37. doi: 10.1016/j.tifs.2020.04.022

De Filippis, F., Valentino, V., Alvarez-Ordóñez, A., Cotter, P. D., and Ercolini, D. (2021). Environmental microbiome mapping as a strategy to improve quality and safety in the food industry. *Curr. Opin. Food Sci*. 38, 168–176. doi: 10.1016/j.cofs.2020.11.012

del Rio-Lavín, A., Monchy, S., Jiménez, E., and Pardo, M. Á. (2023). Gut microbiota fingerprinting as a potential tool for tracing the geographical origin of farmed mussels (*Mytilus galloprovincialis*). *PLoS ONE* 18:e0290776. doi: 10.1371/journal.pone.0290776

Ferrer, L. (2022). Analysis and comparison of classification metrics. *arXiv* [Preprint]. arXiv:2209.05355. doi: 10.48550/arXiv.2209.05355

Gallo, A., Accorsi, R., Goh, A., Hsiao, H., and Manzini, R. (2021). A traceability-support system to control safety and sustainability indicators in food distribution. *Food Control* 124:107866. doi: 10.1016/j.foodcont.2021.107866

Guidone, A., Zotta, T., Matera, A., Ricciardi, A., De Filippis, F., Ercolini, D., et al. (2016). The microbiota of high-moisture mozzarella cheese produced with different acidification methods. *Int. J. Food Microbiol*. 216, 9–17. doi: 10.1016/j.ijfoodmicro.2015.09.002

Ibrahimi, E., Lopes, M. B., Dhamo, X., Simeon, A., Shigdel, R., Hron, K., et al. (2023). Overview of data preprocessing for machine learning applications in human microbiome research. *Front. Microbiol*. 14:1250909. doi: 10.3389/fmicb.2023.1250909

Janzing, D., Minorics, L., and Blöbaum, P. (2020). Feature relevance quantification in explainable AI: a causality problem. *arXiv [Preprint]. arXiv*:1910.13413.

Levante, A., Bertani, G., Marrella, M., Mucchetti, G., Bernini, V., Lazzi, C., et al. (2023). The microbiota of Mozzarella di Bufala Campana PDO cheese: a study across the manufacturing process. *Front. Microbiol*. 14:1196879. doi: 10.3389/fmicb.2023.1196879

Lundberg, S. M., and Lee, S.-I. (2017). "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*, eds. I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc), 4765–4774.

Monaco, A., Pantaleo, E., Amoroso, N., Lacalamita, A., Giudice, C. L., Fonzino, A., et al. (2021). A primer on machine learning techniques for genomic applications. *Comput. Struct. Biotechnol. J*. 19, 4345–4359. doi: 10.1016/j.csbj.2021.07.021

Novielli, P., Romano, D., Magarelli, M., Bitonto, P. D., Diacono, D., Chiatante, A., et al. (2024). Explainable artificial intelligence for microbiome data analysis in colorectal cancer biomarker identification. *Front. Microbiol*. 15:1348974. doi: 10.3389/fmicb.2024.1348974

Ozenne, B., Subtil, F., and Maucort-Boulch, D. (2015). The precision-recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. *J. Clin. Epidemiol*. 68, 855–859. doi: 10.1016/j.jclinepi.2015.02.010

Papoutsoglou, G., Tarazona, S., Lopes, M. B., Klammsteiner, T., Ibrahimi, E., Eckenberger, J., et al. (2023). Machine learning approaches in microbiome research: challenges and best practices. *Front. Microbiol*. 14:1261889. doi: 10.3389/fmicb.2023.1261889

Pisano, M. B., Scano, P., Murgia, A., Cosentino, S., and Caboni, P. (2016). Metabolomics and microbiological profile of Italian mozzarella cheese produced with buffalo and cow milk. *Food Chem*. 192, 618–624. doi: 10.1016/j.foodchem.2015.07.061

Reuter, J. A., Spacek, D. V., and Snyder, M. P. (2015). High-throughput sequencing technologies. *Mol. Cell* 58, 586–597. doi: 10.1016/j.molcel.2015.05.004

Ruck, D. W., Rogers, S. K., and Kabrisky, M. (1990). Feature selection using a multilayer perceptron. *J. Neural Netw. Comput*. 2, 40–48.

Schaffer, C. (1993). Selecting a classification method by cross-validation. *Mach. Learn*. 13, 135–143. doi: 10.1007/BF00993106

Schmieder, R., and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 27, 863–864. doi: 10.1093/bioinformatics/btr026

Shwartz-Ziv, R., and Armon, A. (2022). Tabular data: deep learning is not all you need. *Inform. Fusion* 81, 84–90. doi: 10.1016/j.inffus.2021.11.011

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 30 July 2024
DOI 10.3389/fmicb.2024.1426465

OPEN ACCESS

EDITED BY
Domenica D'Elia,
National Research Council (CNR), Italy

REVIEWED BY
Laura Judith Marcos Zambrano,
IMDEA Food Institute, Spain
Eglantina Kalluci,
University of Tirana, Albania
Thomas Klammsteiner,
University of Innsbruck, Austria

*CORRESPONDENCE
Blaž Stres
✉ blaz.stres@ki.si

RECEIVED 01 May 2024
ACCEPTED 16 July 2024
PUBLISHED 30 July 2024

CITATION
Murovec B, Deutsch L, Osredkar D and Stres B (2024) MetaBakery: a Singularity implementation of bioBakery tools as a skeleton application for efficient HPC deconvolution of microbiome metagenomic sequencing data to machine learning ready information.
Front. Microbiol. 15:1426465.
doi: 10.3389/fmicb.2024.1426465

MetaBakery: a Singularity implementation of bioBakery tools as a skeleton application for efficient HPC deconvolution of microbiome metagenomic sequencing data to machine learning ready information

Boštjan Murovec[1], Leon Deutsch[2,3], Damjan Osredkar[4,5] and Blaž Stres[2,6,7,8]*

[1]University of Ljubljana, Faculty of Electrical Engineering, Ljubljana, Slovenia, [2]University of Ljubljana, Department of Animal Science, Biotechnical Faculty, Ljubljana, Slovenia, [3]The NU, The Nu B.V., Leiden, Netherlands, [4]Department of Pediatric Neurology, University Children's Hospital, University Medical Centre Ljubljana, Ljubljana, Slovenia, [5]University of Ljubljana, Medical Faculty, Ljubljana, Slovenia, [6]D13 Department of Catalysis and Chemical Reaction Engineering, National Institute of Chemistry, Ljubljana, Slovenia, [7]University of Ljubljana, Faculty of Civil and Geodetic Engineering, Ljubljana, Slovenia, [8]Department of Automation, Biocybernetics and Robotics, Jožef Stefan Institute, Ljubljana, Slovenia

In this study, we present MetaBakery (http://metabakery.fe.uni-lj.si), an integrated application designed as a framework for synergistically executing the bioBakery workflow and associated utilities. MetaBakery streamlines the processing of any number of paired or unpaired fastq files, or a mixture of both, with optional compression (gzip, zip, bzip2, xz, or mixed) within a single run. MetaBakery uses programs such as KneadData (https://github.com/bioBakery/kneaddata), MetaPhlAn, HUMAnN and StrainPhlAn as well as integrated utilities and extends the original functionality of bioBakery. In particular, it includes MelonnPan for the prediction of metabolites and Mothur for calculation of microbial alpha diversity. Written in Python 3 and C++ the whole pipeline was encapsulated as Singularity container for efficient execution on various computing infrastructures, including large High-Performance Computing clusters. MetaBakery facilitates crash recovery, efficient re-execution upon parameter changes, and processing of large data sets through subset handling and is offered in three editions with bioBakery ingredients versions 4, 3 and 2 as versatile, transparent and well documented within the MetaBakery Users' Manual (http://metabakery.fe.uni-lj.si/metabakery_manual.pdf). It provides automatic handling of command line parameters, file formats and comprehensive hierarchical storage of output to simplify navigation and debugging. MetaBakery filters out potential human contamination and excludes samples with low read counts. It calculates estimates of alpha diversity and represents a comprehensive and augmented re-implementation of the bioBakery workflow. The robustness and flexibility of the system enables efficient exploration of changing parameters and input datasets, increasing its utility for microbiome analysis. Furthermore, we have shown that the MetaBakery tool can be used in modern biostatistical and machine learning approaches including large-scale microbiome studies.

KEYWORDS

microbial metagenomics, bioinformatics pipeline, machine learning, human gut microbiome, sequence processing, non-communicable diseases, singularity, bioBakery

1 Introduction

Numerous decisions are made by health care providers in medicine on the basis of a multivariate set of descriptors estimating probability that a specific disease is present in an individual (diagnostic context) or a specific condition is going to occur in the near future (prognostic context). In the former diagnostic case the probability that a particular disease may be present is useful for directing patients for further testing or start of immediate treatment next to exclusion of certain causes of observed symptoms. In the latter prognostic context predictions can be utilized to plan therapeutic decisions based on the risk for developing medical condition within specific timeframe and to stratify participants in intervention trials (Collins et al., 2015; Moons et al., 2015). In either context, the combined information from multiple predictors observed and measured in an individual sample are utilized due to the fact that information from a single predictor is often insufficient to provide reliable estimates of diagnostic or prognostic value. Therefore multivariable models are being developed, validated with the aim to assist doctors and individuals in estimating probabilities and potentially guide their decision making (Collins et al., 2015; Moons et al., 2015).

However, recently the quality of reporting of prediction model studies was shown to be poor, therefore several initiatives such as TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis Initiative) (Collins et al., 2015), SPIRIT-AI (Standard Protocol Items: Recommendations for Interventional Trials-Artificial Intelligence) (Cruz Rivera et al., 2020a,b), CONSORT-AI (Consolidated Standards of Reporting Trials-Artificial Intelligence) (Liu et al., 2020a,b) were initiated to name a few. In addition, FAIR guiding principles for research software (Findable, Accessible, Interoperable, Reusable) were introduced in 2022 (Barker et al., 2022; Loftus et al., 2022). This marked a significant milestone for the research community, acknowledging the growing importance of research software globally. These principles also established guidelines outlining minimum requirements for reporting algorithms in healthcare, emphasizing qualities such as explainability, dynamism, precision, autonomy, fairness, and reproducibility (Loftus et al., 2022).

Finally, good data management is the key leading to knowledge discovery and innovation, data integration and reuse by the community after the publication process. FAIR guiding principles for scientific data management (Wilkinson et al., 2016) put specific emphasis on enhancing the ability of machines to automatically use the data and support its reuse by the community to maximize the added value. These principles also take into consideration sharing conditional on privacy considerations (GDPR), claims of proprietary control, practical constraints, access privileges, and the quality of accompanying metadata (Boeckhout et al., 2018).

Recently, two larger scale reports were published describing fecal microbiome-based machine learning for multi-class disease diagnosis (Gupta et al., 2020; Su et al., 2022) utilizing species-level gut microbiome information layer derived metagenomics sequencing runs. Detecting early signs of disease before specific diagnostic symptoms appear is crucial, particularly using biological samples that allow detailed characterization and can be collected noninvasively and regularly. This presents a promising opportunity for developing straightforward prescreening tests to aid both doctors and individuals in decision-making. However, these connections between human health and the accompanying microbiome must be based on real-world conditions observed in the population, ensuring reliability and robustness across various human subjects, conditions, sub-populations, and other factors.

In addition to scientific research, also the industry for (human) microbiome-targeted products is faced with several challenges related to reproducibility and scientific rigor, which can impact the reliability and validity of research findings and the development of microbiome-based products. The primary challenges in microbiome research include the absence of standardized methods and protocols for sample collection, processing, sequencing, and data analysis. Variability in samples affected by host genetics, environmental factors, diet, lifestyle, and other confounding factors all add to complexity. Additionally, limited data sharing and transparency, including controlled access to organized raw data, metadata, and analysis pipelines with respective hyperparameters hinder independent validation of results and the advancement of scientific rigor in this field (Pray et al., 2013; Sinha et al., 2015; Ma et al., 2018; D'Elia et al., 2023; Ruxton et al., 2023).

Broad data sharing policies now enforce the repurposing of existing data from published studies. This serves as real-world data for discovering widely applicable principles and methodologies, generating hypotheses, and validating results. By integrating diverse large datasets from thousands of participants across numerous countries, this approach offers a holistic view at a scale that surpasses single publication datasets.

Existing methods are designed based on the strong assumption that the data with sufficient sample size and accurate and detailed metadata information is available to design groups or train models. The current metadata of a considerable number of sequencing samples is incomplete, misleading, or not publicly available (Kumar et al., 2024), which may lead to these methods being infeasible or causing bias in biomarker inference. Moreover, their intrinsic design in using known phenotype information makes them incapable of revealing new subtypes or stages of diseases (Liu et al., 2022). The taxonomic analysis alone may induce spurious biomarkers since diverse microbial communities from different patients can perform remarkably similar functional capabilities as shown before.

Identification of biomarkers at the level of taxonomy although utilizing species information does not make use of all other layers of information derived from metagenomics, namely alpha diversity, functional genes, enzymatic reactions, metabolic pathways, metabolites that hence remain unexplored. In addition, the gap

between analyses of data using various generations of the same software remains underappreciated source of additional error, as textual information remains cited throughout the published literature while the underlying data re-analyses utilizing different versions of software and underlying databases may support advanced conclusions. Finally, the overall complexity of programs and the supporting databases constitutes another barrier for their deployment on high performance computing (HPC) or cloud computing. To fill this gap, we provide advances on many fronts, by (i) building a reproducible, stable, HPC ready, singularity image integrating the necessary plethora of heavy duty tools from bioBakery, mothur and MelonPann origin (Schloss et al., 2009; Segata et al., 2012; Truong et al., 2015; Pasolli et al., 2017; Franzosa et al., 2018; McIver et al., 2018; Mallick et al., 2019; Schloss, 2020; Beghini et al., 2021), (ii) analyzing previously utilized datasets (Gupta et al., 2020) in conjunction with not yet integrated datasets of clinical relevance (Youngblut and Ley, 2021), (iii) extending the analyses to novel layers of information (functional genes, enzymatic reactions, metabolic pathways, metabolites), (iv) assembling metadata from various studies, and (v) organizing the data into a complete machine learning dataset amenable for 70% of data for training and unseen 30% for validation. Finally, (vi) the meta integration of bioBakery v2, v3 and v4 versions of workflows of original publications enables anyone to back-map the mismatch between the original publications and advancement of algorithms and databases. In total, 4,976 publicly available samples pooled across multiple studies exploring 17 disease types in relation to healthy cohorts reported from 15 countries before, were analyzed. The wealth of data, rigorous analytical approach in data deconvolution and ML provide significant novel insight and actionable models for recognition of medical conditions over a large international dataset.

2 Materials and methods

2.1 Multi-study integration of human gut metagenomes

Data collection was commenced as described and detailed before (Gupta et al., 2020; Supplementary Table S1). In short, published studies with publicly available WGS metagenome data of human stool (gut microbiome) and corresponding subject metadata were included. Also, where multiple samples were taken per individual across different time-points only the baseline first or so-called baseline samples reported in the original study were utilized. To keep up with the same stringency as in the original study, studies reporting on diet or medical interventions or children (<10 years of age) were excluded, in addition to samples collected from disease controls but not marked as healthy or without diagnose assignment in the original study. The primary criteria for data selection included the number of samples, comparable sequencing depth, the quality of QC-ed sequences, and availability of corresponding metadata.

Metadata were synchronized for Healthy group across complete dataset with respect to their BMI and assigned the following categories, irrespective of their initial classification in the original studies: underweight (BMI < 18.5), overweight (BMI ≥ 25 and < 30), or obese (BMI ≥ 30). Consequently, stool metagenome data were renamed as underweight, overweight, or obese in our analysis. In addition, the .fastq files from the following additional projects were included: (i) a subset of the Flemish Gut Flora Project dataset was acquired to explore the efficiency of fecal microbiome data layers in classification of depression based on fecal metagenomic data and metadata (age, sex, BMI, BSS, RAND) of 150 subjects (M = 50, SD = 12,96, 38% male) – 80 with depression and 70 healthy controls (Valles-Colomer et al., 2019); (ii) samples of the PreTerm project (n = 24) (Deutsch et al., 2022a); (iii) samples of the PlanHab project (n = 54) (Šket et al., 2017a,b, 2018, 2020); and (iv) 22 wildcard users (volunteers providing their own .fastq files and necessary metadata; utilized for validation).

Raw sequence files (.fastq files) were downloaded from the EBI (European bioinformatics Institute) next to NCBI Sequence Read Archive and European Nucleotide Archive databases (Gupta et al., 2020) (Supplementary Table S1). Flemish Gut Flora Project data were requested from the Lifelines cohort study[1] following the prescribed standard protocol for data access. Shotgun sequencing data and metadata are available at the EGA (accession no. EGAS00001003298). Subsequent requests for access to data need to be directed to Flemish Gut Flora consortium.

2.2 Sequence data analysis

All datasets were preprocessed utilizing Slovenian HPC cluster SLING/VEGA infrastructure[2, 3] (accessed 28.2.2024) and Austrian HPC MACH2[4] (accessed 28.2.2024.) running Singularity-integrated MetaBakery V3. In total, 1.5 million CPU-hours were utilized to perform quality trimming and deconvolute the sequence information into taxonomy, diversity, functional gene, enzymatic reaction and metabolic pathway data layers next to relaxation network predicted metabolites (Figure 1).

In this study we prepared MetaBakery[5, 6] as a skeleton application for a synergistic execution of the bioBakery worklow of programs (McIver et al., 2018)[7] along with their supporting utilities. Arbitrary number of paired or unpaired fastq files or intermixed serves as input for MetaBakery, either uncompressed or compressed (gzip, zip, bzip2, xz, or mixed) within a single MetaBakery run. The fastq inputs are preprocessed using the KneadData[8] or skipped for already preprocessed data. The inputs are then subjected to the main analyzing programs: MetaPhlAn (Truong et al., 2015; Blanco-Míguez et al., 2023), HUMAnN (Beghini et al., 2021) and StrainPhlAn (Truong et al., 2015; Beghini et al., 2021) along with their supporting utilities (count feature, regroup table, renorm table and join tables). The original bioBakery functionality was enriched by the integration of MelonnPan (Mallick et al., 2019) for metabolite prediction and Mothur (Schloss et al., 2009) for calculation of microbial alpha diversity. The entire pipeline is executed in a nearly single-click way once input files are put in a directory; a config file may optionally

1 https://lifelines.nl/lifelines-research/access-to-lifelines
2 https://en-vegadocs.vega.izum.si/
3 https://www.sling.si/en/sling-2/
4 https://www.uibk.ac.at/zid/systeme/hpc-systeme/mach2/
5 http://metabakery.fe.uni-lj.si
6 http://metabakery.fe.uni-lj.si/metabakery_manual.pdf
7 https://huttenhower.sph.harvard.edu/biobakery_workflows/
8 https://github.com/bioBakery/kneaddata

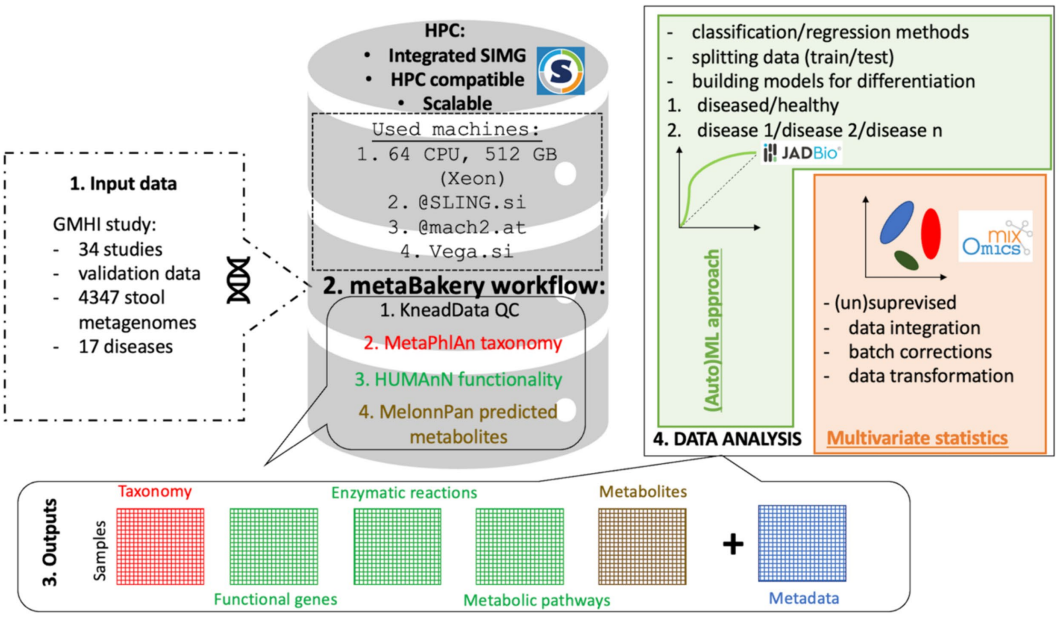

FIGURE 1
Basic schematic representation of the MetaBakery approach. Highlights: all integrated programs and databases are fully preconfigured; external databases may be used instead of the built-in ones; efficient utilization of computing resources; suitable for autonomous and batch execution; suitable for High-Performance Computing facilities; automatic crash recovery; possibility of splitting large datasets into manageable chunks and processing them separately [on different computers and/or high performance computing (HPC) systems]; transparent handling of paired and unpaired reads (possibly intermixed); transparent handling of major compression formats (.gz, .zip, .bz2, .xz), possibly intermixed; automatic handling of command-line parameters for included programs; experienced users can prescribe custom parameters; efficient restarts with changed parameters and input sets; complete screen and configuration dumping for easy documentation; easy access to command lines, exit codes and messages of programs; versions V4, V3 and V2 of bioBakery programs; only meaningful output files are presented to a user.

be specified to tailor the execution. The pipeline automatically inspects the computer's configuration to tune for an efficient execution (Supplementary Figure S1).

The skeleton application within MetaBakery is written in the Python 3 programming language and consists of more than twenty thousand lines of Python code, as well as some utilities written in the C++ programming language for increased efficiency. To achieve efficient running of a number of interdependent programs, an entirely new underlying framework called ExeFlow was developed building from the GUMPP skeleton application (Murovec et al., 2021). To enable its direct adoption for large HPC clusters MetaBakery was packed as Singularity container (Kurtzer et al., 2017; Sochat, 2017; Sochat et al., 2017) to integrate and preconfigure all embedded programs along with their and our own supporting utilities and the relevant databases (Table 1).

Singularity technology was shown to be far better suited for running on high-performance computing facilities compared to other container technologies, like, e.g., Docker (Dirk, 2014) in addition to the fact that it is often the only supported container technology on such large systems.

In addition to improved usability and performance, MetaBakery offers additional benefits (Supplementary Figure S2). The results of all intermediate steps are stored in a specially crafted repository (on a local disk), where each result is associated with its full context, which includes the results of its predecessors and the full set of relevant parameters. On one hand, this enables crash recovery and prompt continuation of processing in the case of a workflow termination (operating system crash, power failure, full hard disk); this feature is offered by the bioBakery (Beghini et al., 2021) workflows as well. In addition, MetaBakery enables efficient re-execution of the workflow with different parameters and/or extended or reduced input data sets. Upon MetaBakery's re-execution, the available results from an arbitrary number of previous runs are instantly retrieved from the repository. Only new steps are subjected to actual processing. This system opens up the possibility to efficiently experiment with modified parameters or input datasets to observe their effects on the final results. Reuse of the past results is completely automatic and transparent. For example, if after a complete MetaBakery's run, a user inspects the results and wants to alter some parameters of the HUMAnN step, then results of previous KneadData, MetaPhlAn and StrainPhlAn runs are instantly retrieved from the repository. This does not hold only for the next-to-the-last run, but for an arbitrary number of past runs. In a similar way, subsets of inputs (paired-end or single-end fastq files) may be freely added or removed between different MetaBakery runs, and only the affected processing steps are recalculated.

MetaBakery also provides a crucial feature for processing large human, non-human or environmental metagenomics projects (consisting of hundreds of fastq files or more). Such datasets can only be processed in a reasonable amount of time on HPC platforms. However, HPC policies often prohibit, or at least penalize tasks with long wall times required to process such large input sets. To alleviate this difficulty, MetaBakery provides the ability to split an input dataset into an arbitrary number of subsets (by means of grouping files, not by splitting individual fastq files).

TABLE 1 MetaBakery ingredients by its edition enabling comparison of results obtained from various versions of the same utilities.

	MetaBakery V2	MetaBakery V3	MetaBakery V4
	KneadData 0.12	KneadData 0.12	KneadData 0.12
	human_hg38_RefMrna (default)	human_hg38_RefMrna (default)	human_hg38_RefMrna (default)
	hg37dec_v0.1 (default)	hg37dec_v0.1 (default)	hg37dec_v0.1 (default)
	mouse_C57BL_6NJ	mouse_C57BL_6NJ	mouse_C57BL_6NJ
Program databases	SILVA_128_LSUParc_SSUParc_ribosomal_RNA	SILVA_128_LSUParc_SSUParc_ribosomal_RNA	SILVA_128_LSUParc_SSUParc_ribosomal_RNA
	MetaPhlAn 2.7.7	MetaPhlAn 3.1	MetaPhlAn 4.0.6
Program database	v20_m200	v31_CHOCOPhlAn_201901	vJan21_CHOCOPhlAnSGB 202,103
	HUMAnN 2.8.1	HUMAnN 3.1.1	HUMAnN 3.6.1
	CHOCOPhlAn 0.1.1	CHOCOPhlAn 201901b	CHOCOPhlAn_201901_v31
	UniRef90 1.1 (both, full and EC filtered)	UniRef90 201901b (both, full and EC filtered)	UniRef90 201901b (both, full and EC filtered)
Program databases	UniRef50 1.1 (both, full and EC filtered)	UniRef50 201901b (both, full and EC filtered)	UniRef50 201901b (both, full and EC filtered)
Program	StrainPhlAn 1.2.0	StrainPhlAn 3.1.0	StrainPhlAn 4.0.6
Program	MelonnPan	MelonnPan	MelonnPan
Program	Mothur 1.46.1	Mothur 1.46.1	Mothur 1.46.1

The only restriction is that in the case of paired reads, the associated R1.fastq and R2.fastq files remain in the same subset. In the extreme case, each subset may consist of only a single unpaired fastq file or a single R1_R2 fastq pair. These subsets can be processed separately on different computers or HPC nodes, even in different parts of the world. The collected partial results can be subjected to MetaBakery by activating its special mode of operation, in which the final results are reconstructed from the partial ones as if the entire input set had been processed in a single MetaBakery run. The reconstruction consists of all post-processing steps, such us: count feature, regroup table, renorm table and join tables, as well as extended features like Mothur calculations and prediction of metabolites with MelonnPan. In addition to bioBakery enabled databases, a custom built STRUO2 database (Youngblut and Ley, 2021) can be utilized as an external component metaBakery.

MetaBakery is offered in three editions. The first edition contains version 4 of the BioBakery programs (MetaPhlAn 4, HUMAnN 3.6 – to be replaced by version 4 when available, StrainPhlAn 4, along with associated utilities and appropriate databases). The second edition contains version 3 of the BioBakery programs (MetaPhlAn 3, HUMAnN 3, StrainPhlAn 3, with appropriate utilities and databases) (Suzek et al. 2007, 2015). The third edition consists of version 2 of the BioBakery programs (MetaPhlAn V2.7.7, HUMAnN 2.8.1, StrainPhlAn 1.2.0, together with the associated utilities and databases).

In summary, MetaBakery is suitable for standalone execution on both commodity hardware and high-performance computing facilities. All command-line parameters and intermediate file formats are handled automatically by the system, so the end user does not have to deal with these technical details. Nevertheless, experienced users can, if they wish, specify their own parameters for each included program to fine-tune its execution. To facilitate documentation of analyses and subsequent review of executions, MetaBakery stores an exact verbatim copy of its screen output as part of a final report. In addition, the actual command lines, standard output streams (stdout), standard error streams (stderr), and exit codes for each program are stored hierarchically on a disk for ease of navigation, review and debugging. The analysis setup is assisted by optional configuration files, where a complete workflow configuration is prescribed, which also aids in documenting a particular run. All features and mentioned use cases are explained in a user-friendly MetaBakery Users' Manual[9] and configuration file template.[10] MetaBakery highlights are summarized in Table 2.

The following additional decision steps were taken in analogy with Gupta et al. (2020) when processing datasets with MetaBakery: (i) potential human contamination was filtered by removing reads that aligned to the human genome (reference genome hg19), in addition to repetitive elements; (ii) stool metagenome samples of low read count after quality filtration (<1 M reads) were excluded from our analysis; (iii) the alpha diversity estimates ($n = 35$) were calculated from biome formatted taxonomy profiles in mothur (Schloss et al., 2009). As a result of all the extended additions, MetaBakery acts as re-implementation of the BioBakery workflow (https://huttenhower.sph.harvard.edu/biobakery_workflows/) integrating three versions of tools (V2, V3 and V4) to deliver various microbiome layers of information: (i) taxonomy (Bacteria, Archaea, Fungi, Protozoa, and Viruses), (ii) alpha diversity estimates; (iii) functional genes, (iv) enzymatic reactions, (v) metabolic pathways, and (vi) predicted metabolites, that are utilized next to subject (patient or healthy) metadata.

9 http://metabakery.fe.uni-lj.si/metabakery_manual.pdf

10 http://metabakery.fe.uni-lj.si/config_template.txt

TABLE 2 MetaBakery highlights.

All integrated programs and databases are fully preconfigured.
External databases may be used instead of the built-in ones (not for V2).
Efficient utilization of computing resources.
Suitable for autonomous and batch execution.
Suitable for High-Performance Computing (HPC) facilities.
Automatic crash recovery.
Possibility of splitting large datasets into manageable chunks and processing them separately (possibly on different computers and/or HPC systems).
Transparent handling of paired and unpaired reads (possibly intermixed).
Transparent handling of major compression formats (gz, zip, bz2, xz), possibly intermixed.
Automatic handling of programs' command-line parameters.
Experienced users can prescribe custom parameters.
Efficient restarts with changed parameters and input sets.
Complete screen and configuration dumping for easy documentation.
Easy access to command lines, exit codes and messages of programs.
V4, V3 and V2 versions of BioBakery programs.
Only meaningful output files are presented to a user.

2.3 Data content

The entire pipeline was used on two different datasets focusing on human microbiome studies: (i) smaller dataset [depression data; [(Valles-Colomer et al., 2019); accession no. EGAS00001003298] consisting of $n = 80$ samples from patients with depression and $n = 70$ healthy controls] and (ii) larger dataset ($n = 4{,}976$ samples - healthy controls and patients with different diseases such as ACVD, ankylosing spondylitis, colorectal adenoma, colorectal cancer, Crohn's disease, impaired glucose tolerance, IBD, obesity, liver cirrhisos, NAFLD, overweight, rheumatoid arthiritis, type 2 diabetes, symptomatic atherosclerosis, ulcerative colitis and underweight) (Gupta et al., 2020; Deutsch et al., 2022a). Both datasets were previously published in scientific journals to ensure the comparability and efficiency of the MetaBakery tool.

In total, 4,976 samples were processed in this study within 1.5 mio CPU-hours at SLING/VEGA HPC cluster[11] (accessed 28.2.2024).

The resulting six data matrices (taxonomy, diversity, functional genes, enzymatic reactions, metabolic pathways and predicted metabolites) were matched with the corresponding human subject metadata matrix and prepared for subsequent machine learning step.

The analyses were run on complete data. Sequences for 4,976 individuals with different diseases and healthy cohorts as control group were downloaded. Bioinformatics was completed with our Singularity implemented pipeline and produced the following information tables: (i) taxonomy table (2,408 variables, file size 0.03 Gb); (ii) gene families (11,451,445 variables, file size 134 Gb); (iii) enzymatic reactions (622,447 variables, file size 8 Gb); (iv) metabolic pathways (47,536 variables, file size 0.6 Gb); (v) predicted metabolites (80 variables, 0.008 Gb); (vi) diversity estimates (35 variables, file size 0.005 Gb); (vii) participant metadata (10 variables, 0.003 Gb).

The compilation of all these variables for almost 5,000 samples produced a matrix with 13 million rows, exhibiting all of the characteristics of microbiome data (Marcos-Zambrano et al., 2021, 2023; Moreno-Indias et al., 2021; Ibrahimi et al., 2023; Papoutsoglou et al., 2023). Contrary to previous approaches (Gupta et al., 2020; Su et al., 2022) that involved significant data reduction steps using arbitrary assumptions (i.e., average OTU abundance <0.15, prevalence >5%) we did not involve such steps as there is no previous guidance on how to set the values in other information layers (diversity, functional gene, enzymatic reactions, metabolic pathways, predicted metabolites) or whether the same settings are transferable between information layers or which variables represent noise within or between multiclass categories.

Benjamini–Hochberg correction was used to control for multiple testing, and results were considered significant at false discovery rate (FDR) < 0.05 as described before in our past studies (Šket et al., 2017a,b, 2018, 2020; Murovec et al., 2020, 2021; Deutsch et al., 2021, 2022a,b; Deutsch and Stres, 2021).

2.4 Machine learning

Automated machine learning, Just Add Data Bio (JADBio), an Amazon cloud based machine learning platform for analyzing potential biomarkers (Tsamardinos et al., 2022), was used to search for biomarkers on both datasets. The JADBIO platform was developed for predictive modeling and providing high-quality predictive models for diagnostics using state-of-the-art statistical and machine learning methods. Personal analytic biases and methodological statistical errors were eliminated from the analysis by autonomously exploring different settings in the modeling steps, resulting in more convincing discovered features to distinguish between different groups. JADBIO with extensive tuning effort and six CPUs was used to model different dataset choices in addition to the features observed in samples of all groups from different projects by splitting the total data into a training set and a test set in a 70:30 ratio. The training set was used to train the model and the test set was used to evaluate the model (Deutsch et al., 2022a). The modeling step was evaluated using 12 different performance metrics (AUC, mean average precision, accuracy, F1 score, Matthews correlation, precision, true-positive rate, specificity, true-positive, true-negative, false-positive, and false-negative). In all cases, 10-fold cross-validation without drop (with a maximum of 20 repeats) was performed. 1,000–3,000 different model configurations (with different feature selection and predictive algorithms with different hyperparameters) were used and up to 100,000 different models were trained per each of the six datasets. The largest dataset representing the gene family data set was reduced to obtain rows with less than 25% zeros per row.

3 Results and discussion

3.1 MetaBakery development, streamlining and large-scale utilization

MetaBakery represents an integrated ready-made system that shortcuts the nontrivial need for technical details of installing and configuring the included programs, libraries and databases. Nevertheless, the high level of flexibility is retained as the integrated

11 https://en-vegadocs.vega.izum.si/

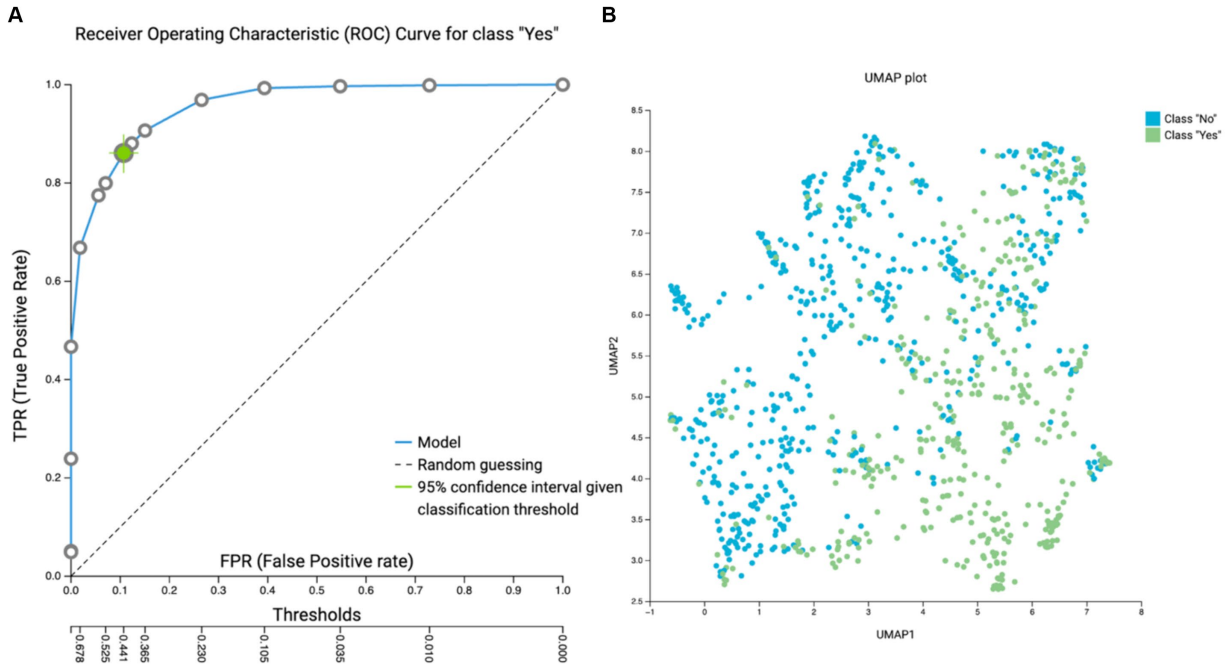

FIGURE 2
(A) Receiver operating characteristic (ROC) curve (AUC = 0.959) for class "Yes" = diseased. **(B)** Uniform manifold approximation and projection (UMAP) attempts to learn the high-dimensional manifold on which the original data lays, and then maps it down to two dimensions. UMAP plots provides a visual aid for assessing relationships among samples.

databases can be freely substituted by advanced users, amended with configuration setting options available to them[12, 13] (Schloss et al., 2009; Segata et al., 2012; Truong et al., 2015; Pasolli et al., 2017; Franzosa et al., 2018; McIver et al., 2018; Mallick et al., 2019; Schloss, 2020; Beghini et al., 2021).

The pipeline handles parallelism differently than the bioBakery as CPUs are always allocated to all running tasks guided by performance parameters (determined by empirical measurements in this study) that indicate the use of CPUs and disk by individual programs to execute as many tasks as possible in parallel without overloading the underlying hardware. Single-threaded or less efficiently parallelized programs no longer take up an entire group of CPUs for themselves, since they are executed evenly on all CPUs in parallel with other processing steps. Better resource utilization thus results from the simultaneous execution of multiple programs on the same set of CPUs which is of special importance when dealing with short HPC wall times. The built-in performance parameters are fully configurable although MetaBakery's default settings were determined by empirical measurements on various pieces of hardware: (i) HPC nodes with varying numbers of CPUs from 256 down to 16, (ii) a desktop computer with dual XEON processor with 64 hyper-threaded processors, and (iii) less powerful desktop computers with 12 and 8 CPUs. Hence, based on the test results our MetaBakery was programmed to tune itself to perform out-of-the-box on the entire hardware spectrum (Supplementary Figure S2).

MetaBakery is offered in three editions. The first edition contains version 4 of the BioBakery programs (MetaPhlAn 4, HUMAnN 3.6 – to be replaced by version 4 when available, StrainPhlAn 4, along with associated utilities and appropriate databases). The second edition contains version 3 of the BioBakery programs (MetaPhlAn 3, HUMAnN 3, StrainPhlAn 3, with appropriate utilities and databases). The third edition consists of version 2 of the BioBakery programs (MetaPhlAn V2.7.7, HUMAnN 2.8.1, StrainPhlAn 1.2.0, together with the associated utilities and databases).

3.2 Large scale computing results: 4976 taxonomy layers

Our data integration resulted in utilization of 4,976 samples encompassing healthy and 16 disease states from 35 studies of 15 countries. In our first data analysis we focused on delineation between the two groups, namely the healthy on one side and a group of disease states on the other. Overall taxonomy classification efficiency enabled us to build a relatively simple and effective model without any specific filtering as also deployed before in the past studies (Gupta et al., 2020) based on taxonomy information only. In essence, we were able to utilize taxonomy information to clearly separate healthy from the diseased states (Figure 2 and Supplementary Figure S3).

In our second analysis we focused on multiclass problem of distinguishing various disease states among themselves. Classification models for many of the disease states based on taxonomy only utilizing rather modest numbers of samples also showed the clear need for larger cohorts on the one side, however clearly provided the necessary information that the signal can readily be detected in such

12 http://metabakery.fe.uni-lj.si/metabakery_manual.pdf
13 http://metabakery.fe.uni-lj.si/config_template.txt

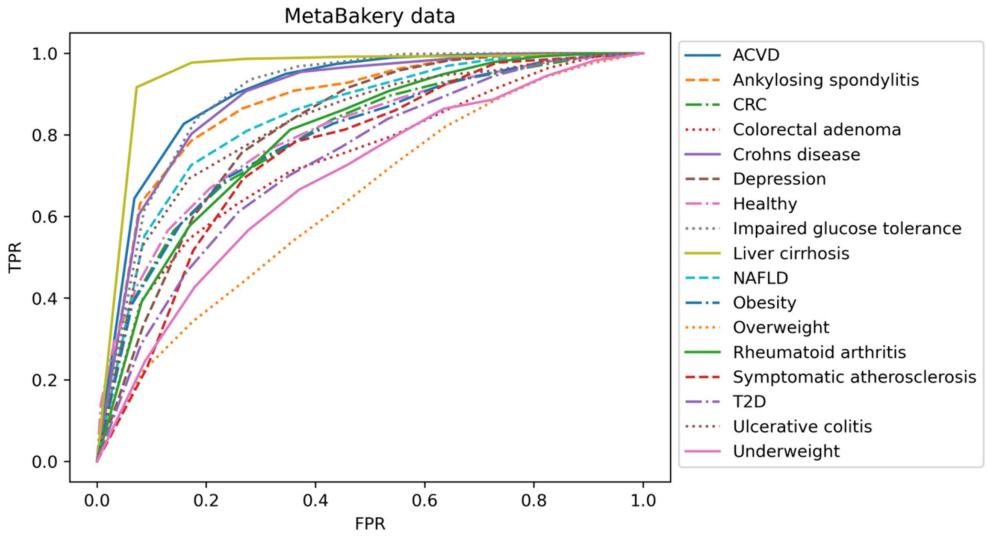

FIGURE 3
Representation of multiclass classification based on AUC curves utilizing taxonomy data layer only and different diseases. TPR, true positive rate; FPR, false positive rate.

small size data as well, guiding future larger-scale data integration (Figure 3).

Diversity metrics utilizing 35 indices were integrated as one of the outputs of the MetaBakery pipeline. For this purpose, the standard diversity calculators from Mothur (Schloss et al., 2009) were integrated into the MetaBakery pipeline, which combine the entire analytical concept of modern microbiology in one pipeline (Supplementary Figure S4), extending the so far amplicon centered approach to metagenomics in a streamlined way.

3.3 Large scale computing results: depression dataset

In our third analysis we focused on depression dataset, utilizing data integration of taxonomy, diversity, functional genes, enzymatic reactions, metabolic pathways and metabolites. Overall, variables were tested for information content that would separate healthy from the clinically depressed participants. We took a two-step approach to model the depression data. In the first step, taxonomy data (852 variables), gene family data (596,146 variables), enzymatic reactions (237,025 variables), metabolic pathways (14,525 variables), and predicted metabolites (80 variables) were modeled individually. In the second step, only the most important features were then modeled on the merged dataset (97 variables). In addition, taxonomy data from 3 different MetaPhlAn versions were also modeled (MetaPhlAn 2.0–972 variables, MetaPhlAn 3.0–859 variables, and MetaPhlAn 4.0–4,249 variables) (Supplementary Table S2). A binary classification was used to distinguish between healthy and depressed individuals.

At the taxonomy level, 23 features (MetaBakery version 2.0), 22 features (Metbakery version 3.0), and 25 features (Metbakery version 4.0) were found to be the most significant in distinguishing depression patients from healthy individuals (Supplementary Figure S5). Because the AUC was highest in MetaBakery 3.0, the corresponding functional data were used to build more successful models at the functional fingerprint level (gene families, enzymatic reactions, metabolic pathways, predicted metabolites). Nine genes, 25 enzymatic reactions, 16 metabolic pathways, and 25 predicted metabolites were discovered in each corresponding data set using JADBio ML (Supplementary Figure S6). In the last step, a subset of the significant features from the first step was used to improve the model. And the logistic ridge model with an AUC of 0.967 was constructed to distinguish patients with depression from healthy individuals (Figure 4).

4 Conclusion

In this study, we presented MetaBakery,[14] an integrated application designed as a framework for synergistically executing the bioBakery workflow (Franzosa et al., 2018; McIver et al., 2018; Beghini et al., 2021) and associated utilities. MetaBakery streamlines the processing of any number of paired or unpaired fastq files, or a mixture of both, with optional compression (gzip, zip, bzip2, xz, or mixed) within a single run. MetaBakery uses programs such as KneadData,[15] MetaPhlAn, HUMAnN and StrainPhlAn as well as integrated utilities and extends the original functionality of bioBakery. In particular, it includes MelonnPan for the prediction of metabolites and Mothur for calculation of microbial alpha diversity. Written in Python 3 and C++, this near single-click pipeline encapsulated as Singularity container leverages the ExeFlow framework for efficient execution on various computing infrastructures, including large High-Performance Computing (HPC) clusters. MetaBakery facilitates crash recovery, efficient re-execution upon parameter changes, and processing of large data sets through subset handling. MetaBakery is offered in three

14 http://metabakery.fe.uni-lj.si
15 https://github.com/bioBakery/kneaddata

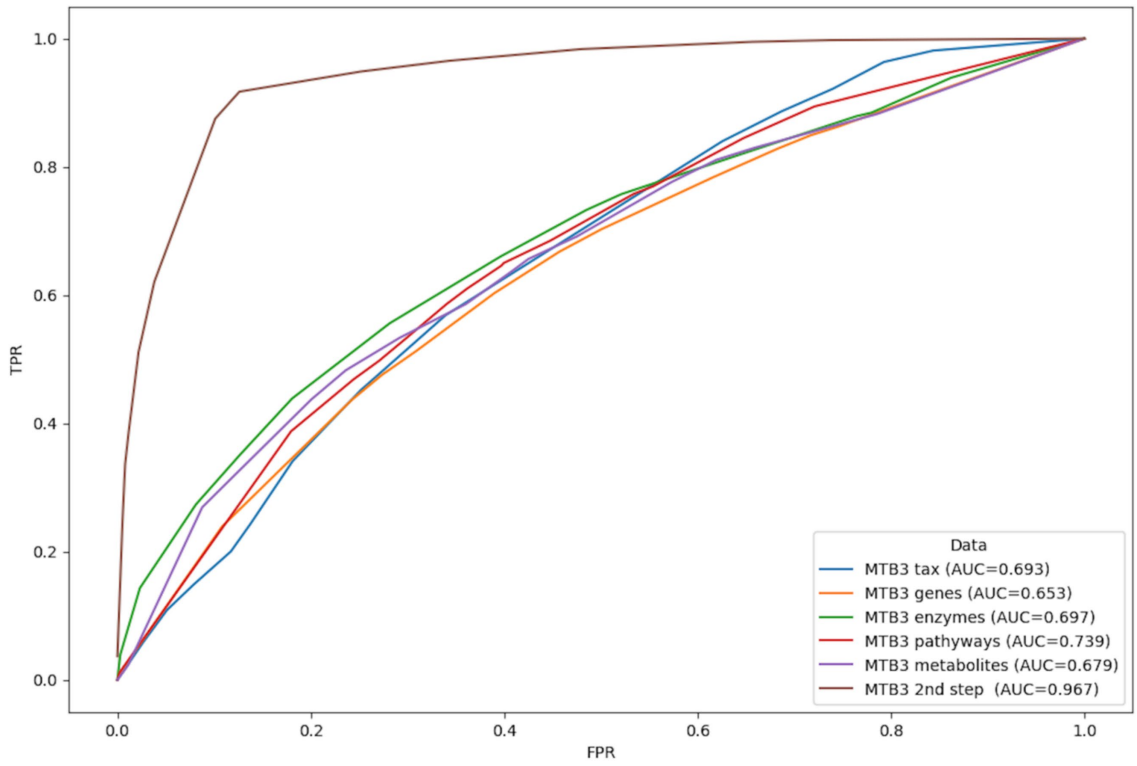

FIGURE 4
Representation of classification based on AUC curves between healthy individuals and patients with depression based on individual information layers: taxonomy (blue), functional genes (orange), enzymatic reactions (green), metabolic pathways (red) and predicted metabolites (purple) calculated with MetaBakery3. Brown line represents the most successful model utilizing the collected features detected as the most important in all data matrices in one analysis. TPR, true positive rate; FPR, false positive rate.

editions with bioBakery ingredients versions 4, 3 and 2. MetaBakery is versatile, transparent and well documented, with functions described in the MetaBakery Users' Manual.[16] It provides automatic handling of command line parameters, file formats and comprehensive hierarchical storage of output to simplify navigation and debugging. MetaBakery filters out potential human contamination and excludes samples with low read counts. It calculates estimates of alpha diversity and represents a comprehensive and augmented re-implementation of the bioBakery workflow. The robustness and flexibility of the system enables efficient exploration of changing parameters and input datasets, increasing its utility for microbiome analysis. Furthermore, we have shown that MetaBakery tool can be used in modern biostatistical and machine learning approaches including large-scale microbiome studies, potentially providing completely new insights into the microbial world.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary material.

Ethics statement

Ethical approval was not required for the study involving humans in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and the institutional requirements.

Author contributions

BM: Writing – original draft, Writing – review & editing, Data curation, Formal analysis, Methodology, Software. LD: Writing – original draft, Writing – review & editing, Data curation, Formal analysis, Investigation, Methodology, Software, Validation. DO: Writing – original draft, Writing – review & editing, Conceptualization, Funding acquisition, Resources. BS: Writing – original draft, Writing – review & editing, Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization.

16 http://metabakery.fe.uni-lj.si/metabakery_manual.pdf

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. BM was partially supported by Slovenian Research and Innovation Agency (SRA/ARRS) program P2-0095 (Parallel and distributed systems). LD acknowledges the MR+ support of the Slovenian Research and Innovation Agency (SRA R#51867) awarded to BS. BS was in part supported by the Slovenian Research and Innovation Agency program P2-0180 (Tools and methods for process analysis, simulation and technology development) and the project J7-50230 Building Efficient Noncommunicable-disease Early Warning Tools (BE NEWt). The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Acknowledgments

The support of National Institute of Chemistry, D13 Department of Catalysis and Chemical Reaction Engineering is gratefully acknowledged. The authors gratefully acknowledge the HPC RIVR consortium (www.hpc-rivr.si) and EuroHPC JU (eurohpc-ju.europa.eu) for funding this research by providing computing resources of the HPC system Vega at the Institute of Information Science (www.izum.si). The computational results presented have been achieved (in part) using the HPC infrastructure of the University of Innsbruck" using Leo3 and Leo4e (https://www.uibk.ac.at/zid/systeme/hpc-systeme/). The ongoing support from the side of Heribert Insam (Ret.), Department of Microbiology, University of Innsbruck, Austria, is gratefully acknowledged. The COST Action ML4Microbiome (CA18131) and the research network therein is kindly acknowledged for fruitful discussions that brought our attention to this topic and prompted us to extend our work.

Conflict of interest

LD was employed by The Nu B.V.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2024.1426465/full#supplementary-material

References

Barker, M., Chue Hong, N. P., Katz, D. S., Lamprecht, A. L., Martinez-Ortiz, C., Psomopoulos, F., et al. (2022). Introducing the FAIR principles for research software. *Sci Data* 9:622. doi: 10.1038/S41597-022-01710-X

Beghini, F., McIver, L. J., Blanco-Míguez, A., Dubois, L., Asnicar, F., Maharjan, S., et al. (2021). Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *eLife* 10. doi: 10.7554/ELIFE.65088

Blanco-Míguez, A., Beghini, F., Cumbo, F., McIver, L. J., Thompson, K. N., Zolfo, M., et al. (2023). Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nat. Biotechnol.* 41, 1633–1644. doi: 10.1038/s41587-023-01688-w

Boeckhout, M., Zielhuis, G. A., and Bredenoord, A. L. (2018). The FAIR guiding principles for data stewardship: fair enough? *Eur. J. Hum. Genet.* 26, 931–936. doi: 10.1038/S41431-018-0160-0

Collins, G. S., Reitsma, J. B., Altman, D. G., and Moons, K. G. M. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ* 350. doi: 10.1136/BMJ.G7594

Cruz Rivera, S., Liu, X., Chan, A. W., Denniston, A. K., Calvert, M. J., Ashrafian, H., et al. (2020a). Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Lancet Digit Health* 2, e549–e560. doi: 10.1016/S2589-7500(20)30219-3

Cruz Rivera, S., Liu, X., Chan, A. W., Denniston, A. K., Calvert, M. J., Darzi, A., et al. (2020b). Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nat. Med.* 26, 1351–1363. doi: 10.1038/S41591-020-1037-7

D'Elia, D., Truu, J., Lahti, L., Berland, M., Papoutsoglou, G., Ceci, M., et al. (2023). Advancing microbiome research with machine learning: key findings from the ML4Microbiome COST action. *Front. Microbiol.* 14:1257002. doi: 10.3389/fmicb.2023.1257002

Deutsch, L., Debevec, T., Millet, G. P., Osredkar, D., Opara, S., Šket, R., et al. (2022a). Urine and Fecal1H-NMR metabolomes differ significantly between pre-term and full-term born physically fit healthy adult males. *Meta* 12:536. doi: 10.3390/metabo12060536

Deutsch, L., Osredkar, D., Plavec, J., and Stres, B. (2021). Spinal muscular atrophy after nusinersen therapy: improved physiology in pediatric patients with no significant change in urine, serum, and liquor 1h-nmr metabolomes in comparison to an age-matched, healthy cohort. *Meta* 11:206. doi: 10.3390/metabo11040206

Deutsch, L., Sotiridis, A., Murovec, B., Plavec, J., Mekjavic, I., Debevec, T., et al. (2022b). Exercise and Interorgan communication: short-term exercise training blunts differences in consecutive daily Urine1H-NMR Metabolomic signatures between physically active and inactive individuals. *Meta* 12:473. doi: 10.3390/metabo12060473

Deutsch, L., and Stres, B. (2021). The importance of objective stool classification in fecal 1H-NMR metabolomics: exponential increase in stool crosslinking is mirrored in systemic inflammation and associated to fecal acetate and methionine. *Meta* 11:172. doi: 10.3390/metabo11030172

Dirk, M. (2014). Docker: lightweight Linux containers for consistent development and deployment. *Linux J.* 76–91. doi: 10.5555/2600239.2600241

Franzosa, E. A., McIver, L. J., Rahnavard, G., Thompson, L. R., Schirmer, M., Weingart, G., et al. (2018). Species-level functional profiling of metagenomes and metatranscriptomes. *Nat. Methods* 15, 962–968. doi: 10.1038/S41592-018-0176-Y

Gupta, V. K., Kim, M., Bakshi, U., Cunningham, K. Y., Davis, J. M., Lazaridis, K. N., et al. (2020). A predictive index for health status using species-level gut microbiome profiling. *Nat. Commun.* 11:4635. doi: 10.1038/S41467-020-18476-8

Ibrahimi, E., Lopes, M. B., Dhamo, X., Simeon, A., Shigdel, R., Hron, K., et al. (2023). Overview of data preprocessing for machine learning applications in human microbiome research. *Front. Microbiol.* 14:1250909. doi: 10.3389/fmicb.2023.1250909

Kumar, B., Lorusso, E., Fosso, B., and Pesole, G. (2024). A comprehensive overview of microbiome data in the light of machine learning applications: categorization, accessibility, and future directions. *Front. Microbiol.* 15:1343572. doi: 10.3389/fmicb.2024.1343572

Kurtzer, G. M., Sochat, V., and Bauer, M. W. (2017). Singularity: scientific containers for mobility of compute. *PLoS One* 12:e0177459. doi: 10.1371/JOURNAL.PONE.0177459

Liu, X., Cruz Rivera, S., Moher, D., Calvert, M. J., Denniston, A. K., Ashrafian, H., et al. (2020a). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Lancet Digit Health* 2, e537–e548. doi: 10.1016/S2589-7500(20)30218-1

Liu, X., Rivera, S. C., Moher, D., Calvert, M. J., and Denniston, A. K.SPIRIT-AI and CONSORT-AI Working Group (2020b). Reporting guidelines for clinical trial reports

for interventions involving artificial intelligence: the CONSORT-AI extension. *BMJ* 370:m3164. doi: 10.1136/bmj.m3164

Liu, Z., Wang, Q., Ma, A., Chung, D., Zhao, J., Ma, Q., et al. (2022). Inference of disease-associated microbial gene modules based on metagenomic and metatranscriptomic data. *bioRxiv*. doi: 10.1101/2021.09.13.460160

Loftus, T. J., Tighe, P. J., Ozrazgat-Baslanti, T., Davis, J. P., Ruppert, M. M., Ren, Y., et al. (2022). Ideal algorithms in healthcare: explainable, dynamic, precise, autonomous, fair, and reproducible. *PLOS digital health* 1:e0000006. doi: 10.1371/JOURNAL.PDIG.0000006

Ma, Y., Chen, H., Lan, C., and Ren, J. (2018). Help, hope and hype: ethical considerations of human microbiome research and applications. *Protein Cell* 9, 404–415. doi: 10.1007/S13238-018-0537-4

Mallick, H., Franzosa, E. A., Mclver, L. J., Banerjee, S., Sirota-Madi, A., Kostic, A. D., et al. (2019). Predictive metabolomic profiling of microbial communities using amplicon or metagenomic sequences. *Nat. Commun.* 10:3136. doi: 10.1038/S41467-019-10927-1

Marcos-Zambrano, L. J., Karaduzovic-Hadziabdic, K., Loncar Turukalo, T., Przymus, P., Trajkovik, V., Aasmets, O., et al. (2021). Applications of machine learning in human microbiome studies: a review on feature selection, biomarker identification, disease prediction and treatment. *Front. Microbiol.* 12:634511. doi: 10.3389/FMICB.2021.634511/BIBTEX

Marcos-Zambrano, L. J., López-Molina, V. M., Bakir-Gungor, B., Frohme, M., Karaduzovic-Hadziabdic, K., Klammsteiner, T., et al. (2023). A toolbox of machine learning software to support microbiome analysis. *Front. Microbiol.* 14:1250806. doi: 10.3389/fmicb.2023.1250806

McIver, L. J., Abu-Ali, G., Franzosa, E. A., Schwager, R., Morgan, X. C., Waldron, L., et al. (2018). bioBakery: a meta'omic analysis environment. *Bioinformatics* 34, 1235–1237. doi: 10.1093/BIOINFORMATICS/BTX754

Moons, K. G. M., Altman, D. G., Reitsma, J. B., Ioannidis, J. P. A., Macaskill, P., Steyerberg, E. W., et al. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. *Ann. Intern. Med.* 162, W1–W73. doi: 10.7326/M14-0698

Moreno-Indias, I., Lahti, L., Nedyalkova, M., Elbere, I., Roshchupkin, G., Adilovic, M., et al. (2021). Statistical and machine learning techniques in human microbiome studies: contemporary challenges and solutions. *Front. Microbiol.* 12:635781. doi: 10.3389/fmicb.2021.635781

Murovec, B., Deutsch, L., and Stres, B. (2020). Computational framework for high-quality production and large-scale evolutionary analysis of metagenome assembled genomes. *Mol. Biol. Evol.* 37, 593–598. doi: 10.1093/molbev/msz237

Murovec, B., Deutsch, L., and Stres, B. (2021). General unified microbiome profiling pipeline (Gumpp) for large scale, streamlined and reproducible analysis of bacterial 16s rrna data to predicted microbial metagenomes, enzymatic reactions and metabolic pathways. *Meta* 11:336. doi: 10.3390/metabo11060336

Papoutsoglou, G., Tarazona, S., Lopes, M. B., Klammsteiner, T., Ibrahimi, E., Eckenberger, J., et al. (2023). Machine learning approaches in microbiome research: challenges and best practices. *Front. Microbiol.* 14:1261889. doi: 10.3389/fmicb.2023.1261889

Pasolli, E., Schiffer, L., Manghi, P., Renson, A., Obenchain, V., Truong, D. T., et al. (2017). Accessible, curated metagenomic data through ExperimentHub. *Nat. Methods* 14, 1023–1024. doi: 10.1038/NMETH.4468

Pray, L., Pillsbury, L., and Tomayko, E. (Eds.) (2013). The human microbiome, diet, and health: Workshop summary. Washington, DC: The National Academies Press.

Ruxton, C. H. S., Kajita, C., Rocca, P., and Pot, B. (2023). Microbiota and probiotics: chances and challenges – a symposium report. *Gut Microbiome* 4:e6. doi: 10.1017/GMB.2023.4

Schloss, P. D. (2020). Reintroducing mothur: 10 years later. *Appl. Environ. Microbiol.* 86. doi: 10.1128/AEM.02343-19

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., et al. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* 75, 7537–7541. doi: 10.1128/AEM.01541-09

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., and Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods* 9, 811–814. doi: 10.1038/NMETH.2066

Sinha, R., Abnet, C. C., White, O., Knight, R., and Huttenhower, C. (2015). The microbiome quality control project: baseline study design and future directions. *Genome Biol.* 16, 1–6. doi: 10.1186/S13059-015-0841-8/METRICS

Šket, R., Debevec, T., Kublik, S., Schloter, M., Schoeller, A., Murovec, B., et al. (2018). Intestinal metagenomes and metabolomes in healthy young males: inactivity and hypoxia generated negative physiological symptoms precede microbial dysbiosis. *Front. Physiol.* 9:198. doi: 10.3389/fphys.2018.00198

Šket, R., Deutsch, L., Prevoršek, Z., Mekjavić, I. B., Plavec, J., Rittweger, J., et al. (2020). Systems view of deconditioning during spaceflight simulation in the PlanHab project: the departure of urine 1 H-NMR metabolomes from healthy state in young males subjected to bedrest inactivity and hypoxia. *Front. Physiol.* 11:1550. doi: 10.3389/fphys.2020.532271

Šket, R., Treichel, N., Debevec, T., Eiken, O., Mekjavic, I., Schloter, M., et al. (2017a). Hypoxia and inactivity related physiological changes (constipation, inflammation) are not reflected at the level of gut metabolites and butyrate producing microbial community: the PlanHab study. *Front. Physiol.* 8:250. doi: 10.3389/fphys.2017.00250

Šket, R., Treichel, N., Kublik, S., Debevec, T., Eiken, O., Mekjavić, I., et al. (2017b). Hypoxia and inactivity related physiological changes precede or take place in absence of significant rearrangements in bacterial community structure: the PlanHab randomized trial pilot study. *PLoS One* 12:e0188556. doi: 10.1371/journal.pone.0188556

Sochat, V. (2017). Singularity registry: open source registry for singularity images. *J Open Source Softw* 2:426. doi: 10.21105/JOSS.00426

Sochat, V. V., Prybol, C. J., and Kurtzer, G. M. (2017). Enhancing reproducibility in scientific computing: metrics and registry for singularity containers. *PLoS One* 12:e0188511. doi: 10.1371/JOURNAL.PONE.0188511

Su, Q., Liu, Q., Lau, R. I., Zhang, J., Xu, Z., Yeoh, Y. K., et al. (2022). Faecal microbiome-based machine learning for multi-class disease diagnosis. *Nat. Commun.* 13. doi: 10.1038/S41467-022-34405-3

Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R., and Wu, C. H. (2007). UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* 23, 1282–1288. doi: 10.1093/BIOINFORMATICS/BTM098

Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., and Wu, C. H. (2015). UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* 31, 926–932. doi: 10.1093/BIOINFORMATICS/BTU739

Truong, D. T., Franzosa, E. A., Tickle, T. L., Scholz, M., Weingart, G., Pasolli, E., et al. (2015). MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* 12, 902–903. doi: 10.1038/NMETH.3589

Tsamardinos, I., Charonyktakis, P., Papoutsoglou, G., Borboudakis, G., Lakiotaki, K., Zenklusen, J. C., et al. (2022). Just add data: automated predictive modeling for knowledge discovery and feature selection. *NPJ Precis. Oncol.* 6, 38–17. doi: 10.1038/s41698-022-00274-8

Valles-Colomer, M., Falony, G., Darzi, Y., Tigchelaar, E. F., Wang, J., Tito, R. Y., et al. (2019). The neuroactive potential of the human gut microbiota in quality of life and depression. *Nat. Microbiol.* 4, 623–632. doi: 10.1038/S41564-018-0337-X

Wilkinson, M. D., Dumontier, M., Aalbersberg, I., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* 3:160018. doi: 10.1038/SDATA.2016.18

Youngblut, N. D., and Ley, R. E. (2021). Struo2: efficient metagenome profiling database construction for ever-expanding microbial genome datasets. *PeerJ* 9:e12198. doi: 10.7717/PEERJ.12198

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 26 August 2024
DOI 10.3389/fmicb.2024.1426407

OPEN ACCESS

EDITED BY
Domenica D'Elia,
National Research Council (CNR), Italy

REVIEWED BY
Balázs Ligeti,
Pázmány Péter Catholic University, Hungary
Bruno Fosso,
University of Bari Aldo Moro, Italy

*CORRESPONDENCE
Blaž Stres
✉ blaz.stres@ki.si

RECEIVED 01 May 2024
ACCEPTED 09 August 2024
PUBLISHED 26 August 2024

CITATION
Murovec B, Deutsch L and Stres B (2024) Predictive modeling of colorectal cancer using exhaustive analysis of microbiome information layers available from public metagenomic data.
Front. Microbiol. 15:1426407.
doi: 10.3389/fmicb.2024.1426407

Predictive modeling of colorectal cancer using exhaustive analysis of microbiome information layers available from public metagenomic data

Boštjan Murovec[1], Leon Deutsch[2,3] and Blaž Stres[2,4,5,6]*

[1]Faculty of Electrical Engineering, University of Ljubljana, Ljubljana, Slovenia, [2]Department of Animal Science, Biotechnical Faculty, University of Ljubljana, Ljubljana, Slovenia, [3]The NU, The NU B.V., Leiden, Netherlands, [4]D13 Department of Catalysis and Chemical Reaction Engineering, National Institute of Chemistry, Ljubljana, Slovenia, [5]Faculty of Civil and Geodetic Engineering, Institute of Sanitary Engineering, Ljubljana, Slovenia, [6]Department of Automation, Biocybernetics and Robotics, Jožef Stefan Institute, Ljubljana, Slovenia

This study aimed to compare the microbiome profiles of patients with colorectal cancer (CRC, *n*=380) and colorectal adenomas (CRA, *n*=110) against generally healthy participants (*n*=2,461) from various studies. The overarching objective was to conduct a real-life experiment and develop a robust machine learning model applicable to the general population. A total of 2,951 stool samples underwent a comprehensive analysis using the in-house MetaBakery pipeline. This included various data matrices such as microbial taxonomy, functional genes, enzymatic reactions, metabolic pathways, and predicted metabolites. The study found no statistically significant difference in microbial diversity among individuals. However, distinct clusters were identified for healthy, CRC, and CRA groups through linear discriminant analysis (LDA). Machine learning analysis demonstrated consistent model performance, indicating the potential of microbiome layers (microbial taxa, functional genes, enzymatic reactions, and metabolic pathways) as prediagnostic indicators for CRC and CRA. Notable biomarkers on the taxonomy level and microbial functionality (gene families, enzymatic reactions, and metabolic pathways) associated with CRC were identified. The research presents promising avenues for practical clinical applications, with potential validation on external clinical datasets in future studies.

KEYWORDS

gut microbiome, machine learning, colorectal cancer, colorectal adenoma, metagenomics, functional microbiome

1 Introduction

The prevalence of colorectal carcinoma (CRC) as the third most common nongender-related cancer and its associated mortality after lung cancer is of great concern (Sung et al., 2021). With an aging population leading to an expected 80% increase in global incidence over the next two decades, understanding sporadic colorectal cancers has become increasingly important (Karsa et al., 2010). These non-hereditary colorectal cancers account for 70–87% of cases, with genetics accounting for only a fraction of disease incidence (Frank et al., 2017). The lack of a clear genetic link underscores the potential influence of other factors, including

lifestyle and environmental components, as co-determinants of disease (Siegel et al., 2014). Certain risk factors such as age, tobacco and alcohol use, physical inactivity, increased body weight, and dietary habits have been associated with CRC, but clarification of these associations remains an ongoing challenge (Huxley et al., 2009; Johnson et al., 2013).

The human gut microbiome, which encompasses the microbial communities in the intestinal tract, is becoming increasingly important because of its role in human disease (Pasolli et al., 2016). Supported by evidence that bacterial organisms trigger carcinogenic mechanisms, the role of the gut microbiome in the development of CRC has been proposed (Wong and Yu, 2023). The association of *Fusobacterium nucleatum* with CRC was revealed by amplicon sequencing of the 16S ribosomal RNA (rRNA) gene and later confirmed as causative in animal models CRC (Kostic et al., 2012, 2013; Rubinstein et al., 2013). While 16S rRNA gene studies revealed such associations, metagenomic sequencing studies revealed a smaller number of CRC-associated microbial species and functional activities. However, the consistency and prognostic potential of these high-resolution microbial signatures across different cohorts and study designs remain uncertain. Although the use of the gut microbiome for CRC diagnostics has been proposed, its validation in multiple independent studies is still pending (Zackular et al., 2014; Zeller et al., 2014; Feng et al., 2015; Baxter et al., 2016; Yu et al., 2017).

Therefore, there remains a need to establish and validate links between the human gut microbiome and CRC across different populations, cohorts, and microbiome tools. While some cross-cohort studies have been based on 16S rRNA gene studies, this technique has its own limitations (Durazzi et al., 2021). The advent of whole-metagenome shotgun datasets for CRC cohorts facilitates a comprehensive exploration of the CRC-associated microbiome that includes strain-level precision and meta-analytic prediction strategies. Therefore, extensive cross-cohort studies are essential for an unbiased and robust assessment of the relationship between CRC and the gut microbiome.

While sequencing of gene amplicons for microbial identification, especially 16S rRNA sequencing, remains a priority, metagenomic analysis by genome-wide shotgun sequencing is becoming increasingly important. It was shown before that with shotgun sequencing entire microbial community can be screened (including viruses, fungi), especially the less abundant taxa, which can also be biologically important. On the other hand, with shotgun sequencing, microbial genes and metabolic pathways can be detected. In contrast, amplicon sequencing only allows for the prediction of microbial genes and metabolic pathways (Durazzi et al., 2021). Shotgun sequencing integrates function, taxonomy and phylogeny and provides insights into the structure and function of the microbial community. It allows us to identify not only taxonomic units, but also genes, enzymatic reactions and metabolic pathways involved in microbial functionality. Given that there are 150 times more microbial genes than human genes, shotgun sequencing will soon enable us to understand the mechanisms behind the association of the microbiota with various diseases, including CRC (Qin et al., 2010; Wang et al., 2015).

The aim of this study was to compare the microbiome of patients with colorectal cancer and colorectal adenomas with that of generally healthy participants from different studies. With this goal in mind, we sought to conduct a real-life experiment and create a robust machine learning model that can be applied to the general population.

In a typical procedure for building a disease classifier, a certain number of individuals with and without a disease are sampled by some research group in order to obtain data for machine learning. The pool of sampled individuals is necessarily limited, by means of which their diversity is less than satisfactory. Hence, the resulting machine-learning model is necessarily overfitted to the very participants in a study. In contrast, the study in this article was conduct on as large dataset as it was possible to constellate from available sampled data from all over the world. The aim was to incorporate as rich diversity of a broad population into the resulting machine learning model. With this regard, it is reasonable to expect that at least some confounding factors are removed from the obtained disease classifier.

2 Methods

2.1 Data

Paired read sequences from 2,461 healthy participants, 380 CRC patients and 110 CRA individuals were downloaded from publicly available datasets studying different associations of different diseases and healthy controls. The main data selection criteria were the number of samples, depth of sequencing, the quality of resulting QC-ed sequences and the availability of metadata. Healthy individuals were defined as those who were reported as not having any overt disease not adverse symptoms at the time of the original study. The list of available datasets used in this study is available in Supplementary Table S1. The same dataset was used in study representing gut microbiome health index (Gupta et al., 2020). With a larger, healthy cohort, the aim was to consider the substantial variability of the human gut microbiome among healthy individuals (He et al., 2018).

2.2 Sequence processing

Paired-end reads were obtained from publicly available datasets using download procedures of European Nucleotide Archive[1] (Supplementary Table S1; Supplementary material: Extended discussion) and analyzed using our custom metagenomics sequence processing pipeline MetaBakery (currently in preparation, Deutsch et al., 2022a). MetaBakery is a new implementation of the BioBakery workflow (Beghini et al., 2021) and includes tools such as KneadData v0.12.0[2] with contaminant databases human_hg38_refMrna and hg37dec_v0.1 for quality control, MetaPhlAn 3.1.0 with database mpa_v31_CHOCOPhlAn_201901 for taxonomic analysis (for bacteria, archaea, fungi, protozoa and viruses) (Beghini et al., 2021) and HUMAnN 3.1.1 (Beghini et al., 2021) with databases full_chocophlan.v201901_v31 and uniref90_201901b_full for inferring functional genes, enzymatic reactions and metabolic pathways. In addition, MelonnPan 0.99.0 (Mallick et al., 2019) was used for the prediction of microbial metabolites. MetaBakery is

1 https://ena-docs.readthedocs.io/en/latest/retrieval/file-download.html

2 https://huttenhower.sph.harvard.edu/kneaddata/, accessed October 10, 2023.

containerized as a Singularity image and optimized for high performance clustering processing of large numbers of samples. For diversity assessment, Mothur 1.46.1 was integrated as part of MetaBakery pipeline utilizing biome format for diversity calculators ($n = 35$) (Schloss et al., 2009; Schloss, 2020). For this study no hand-crafted command-line parameters were used for executing the above-mentioned programs. If not instructed differently, the MetaBakery pipeline executes each program with its default parameters, as they apply to execution within the bioBakery workflow.

Minor steps of the analyses with MetaBakery were performed on a dual Xeon system with 32 CPU cores (64 hyperthreads), 512 GB RAM and 6 TB SATA hard disk at the Faculty of Electrical Engineering, University of Ljubljana, Slovenia. HPC system Vega at the Institute of Information Science[3] and the HPC infrastructure Leo3, Leo4e of the University of Innsbruck, Austria, were utilized for heavy duty processing. In total, 980,000 CPUh were consumed.

2.3 Statistical analysis

Python 3.9[4] (Van Rossum and Drake, 2009) served as the basis for our statistical analysis. We used the non-parametric Mann–Whitney test integrated in the scipy.stats library (Virtanen et al., 2020) to accurately determine the statistical significance between groups in terms of diversity and the features identified in the auto machine-learning (autoML) analysis. These features were selected by an automatic machine learning analysis based on taxonomic signatures, gene families, enzymatic reactions, metabolic pathways and predicted metabolites in the different groups (CRC, CRA, healthy). We used the Python libraries matplotlib (Hunter, 2007) and seaborn (Waskom, 2021) to visualize our results. The scikit-learn library (Pedregosa et al., 2011) in Python facilitated the linear discriminant analysis (LDA), while the preprocessing was done using the StandardScaler method. Using the LDA method, we visualized and interpreted the differences between three different clusters: CRC, CRA and healthy participants. These observations were based on taxonomic signatures, gene families, enzymatic reactions, metabolic pathways and predicted metabolites, leading to a comprehensive understanding of the data. In addition UMAP clustering was performed using JADBIO machine learning (Tsamardinos et al., 2022).

2.4 Automated machine learning

The web-based machine learning platform "Just Add Data Bio" (JADBIO, Ver. 1.4.105) was used to investigate potential biomarkers (Tsamardinos et al., 2022). A two-stage methodology was used for the analysis. First, the models were trained individually for each component of the data matrix, i.e., for taxonomy, functional genes, enzymatic reactions, metabolic pathways and predicted metabolites. Subsequently, an integration step was performed in which all significant features were merged, and the model was retrained. JADBIO was developed for predictive modeling and uses advanced statistical and machine learning techniques to create robust diagnostic predictive models. The analysis was systematically performed to rule out personal analytical bias and methodological statistical errors by autonomously examining different modeling settings (Deutsch and Stres, 2021; Murovec et al., 2021; Deutsch, 2022; Deutsch et al., 2022a,b). This process led to the identification of key features that allow effective discrimination between different groups. Using considerable computational resources and careful parameter tuning, JADBIO was used to model different dataset variations. The data was preprocessed to retain all rows (representing taxonomical features, gene families, enzymatic reactions and metabolic pathways) with at least 1,250 non-zero values, aiming to exclude the influence of large proportion of zeroes in the dataset. More than 2000 different model configurations were used to find the best possible model per every data matrix (Supplementary Table S2). All steps involving machine learning were used as implemented in JADBIO. Different model configurations were tested with different preprocessing steps, feature selectors, feature selection hyperparamters, predictive algorithms and hyperparameters were tested (Supplementary Table S2; Supplementary material: Extended discussion). The analysis included features extracted from samples of different projects and groups, with the data split 70:30 into training and test datasets. The training dataset was used to develop the model, while the test dataset evaluated its performance (Deutsch and Stres, 2021; Murovec et al., 2021; Deutsch, 2022; Deutsch et al., 2022a,b). Receiver operating characteristic curves (ROC) were generated for all groups studied to evaluate the model. These curves graphically represented the trade-off between the rate of true-positive findings (sensitivity) and the rate of false-positive findings (1-specificity). Individual conditional expectation plots (ICE) were used for depth to illustrate the differential contribution of each feature to the predictive power of the model. Progressive feature inclusion plots were also created to provide insight into the impact of feature inclusion on model performance.

3 Results

3.1 Diversity

The in-house analytical pipeline MetaBakery (in preparation, Deutsch et al., 2022a) was used to preprocess the sequence data with integrated tool KneadData[5] and to analyze the sequences at the level of taxonomy [MetaPhlAn3 (Beghini et al., 2021)], diversity [Mothur (Schloss et al., 2009)], functional genes, enzymatic reactions and metabolic pathways [HUMAnN3 (Beghini et al., 2021)] and predicted metabolites [MelonnPan (Mallick et al., 2019)]. Sequences from 2,461 healthy individuals, 380 CRC patients and 110 individuals with confirmed CRA were used for the analysis. A total of 1839 taxonomic units (kingdoms, phyla, clades, orders, families, genera and species) including archaea, bacteria, protozoa and viruses, 80,372 gene families, 34,008 enzymatic reactions, 31,555 metabolic pathways and 81 predicted metabolites were identified and analyzed in the human gut microbiota. 19 different diversity metrics were used to compare all

3 www.izum.si

4 https://www.python.org/, accessed October 10, 2023.

5 https://huttenhower.sph.harvard.edu/kneaddata/, accessed October 10, 2023.

three groups and determine the presence of differences. Although in most cases the diversity metrics were higher in the CRC and CRA groups, these differences were not significant, including the Shannon diversity index (Figure 1) as determined by the Mann–Whitney test (Supplementary Table S3; Supplementary Figure S1).

3.2 LDA analysis

Using the scikit-learn Python library, linear discriminant analysis (LDA) was used to explore potential differences between healthy individuals, CRA and CRC patients in the five data matrices (taxonomy, functional genes, enzymatic reactions, metabolic pathways and predicted metabolites). As shown in Figure 2, LDA clustering effectively discriminates between CRC, CRA and healthy individuals based on four different metagenomic fingerprints (taxonomy in Figure 2A, functional genes in Figure 2B, enzymatic reactions in Figure 2C and metabolic pathways in Figure 2D). However, no clear LDA cluster separation was observed for the predicted metabolites (Supplementary Figure S2). In addition, UMAP analysis was performed using JADBIO (Supplementary Figure S3).

3.3 Machine learning results

Although clear separation was observed in only four datasets (taxonomy, genes, enzymatic reactions and metabolic pathways), all five metagenomics data matrices (taxonomy data, functional genes, enzymatic reactions, metabolic pathways and predicted metabolites) were used for automatic machine learning using the JADBIO web-based tool. All matrices were prepared such that rows with at least 1,250 non-zero entries were retained in the dataset.

Based on the 1839 categories describing the taxonomic data of four different kingdoms (Archaea, Bacteria, Protozoa and Viruses), the models were trained using extensive tuning effort in search of biologically meaningful distinguishing features between all three groups. All important features were representative of the Bacteria kingdom and the best performing model was Classification Random Forest training 1,000 trees with deviance splitting criterion, minimum leaf size = 2, splits = 1, alpha = 1 and variables to split = 1.0 sqrt (nvars) according to JADBIO, after testing more than 2000 different configurations. More than 25 features were selected as the most appropriate to achieve the best possible differentiation between all three groups (AUC = 0.817), but the first ten taxonomic units can achieve more than 95% successful performance for differentiation (Figure 3A; Supplementary Figure S4; Supplementary Table S4). This model was tested with all 25 selected features using test data and achieved a performance of AUC = 0.787.

HUMAnN3 (Beghini et al., 2021), integrated in our MetaBakery pipeline, was used to assess the functional potential of the microbiome. Functional genes were determined using the UniRef database (Suzek et al., 2007, 2015). 80.372 functional genes were discovered in the samples and 70% of the total dataset was used to find the best possible model. The best possible model was Classification Random Forest training 1,000 trees with deviance splitting criterion, minimum leaf size = 3, splits = 1, alpha = 1 and variables to split = 0.577 sqrt (nvars) with an area under the curve value of 0.815 (Figure 3B). From the entire pool of genes, 25 of them were selected as the most important features for differentiation. However, a classification performance of 100% was achieved with the first 15 of them (Supplementary Figure S5).

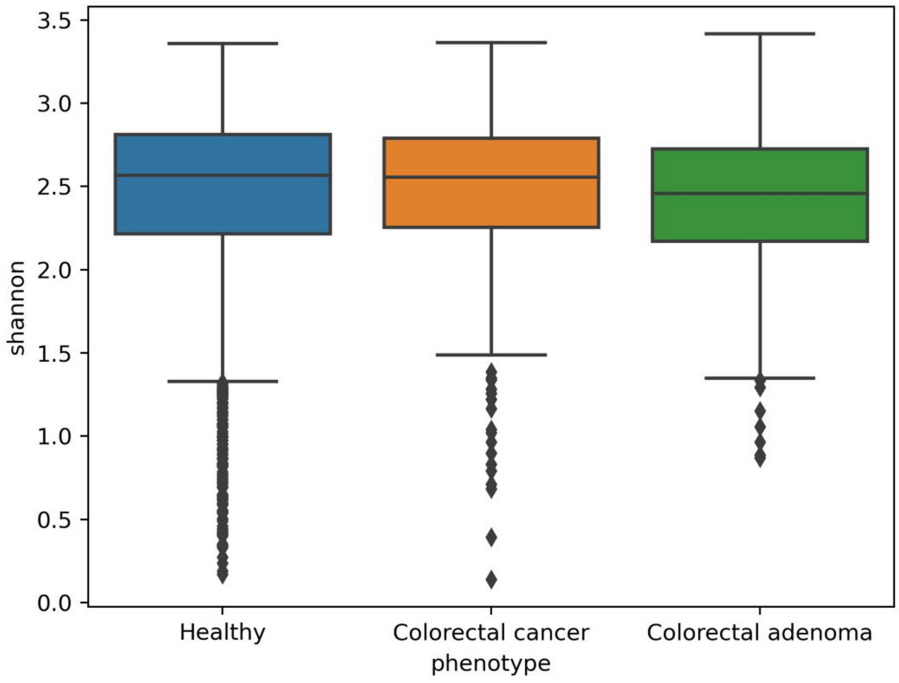

FIGURE 1
Boxplots representing Shannon diversity metrics for healthy individuals and patients with colorectal cancer or colorectal adenoma.

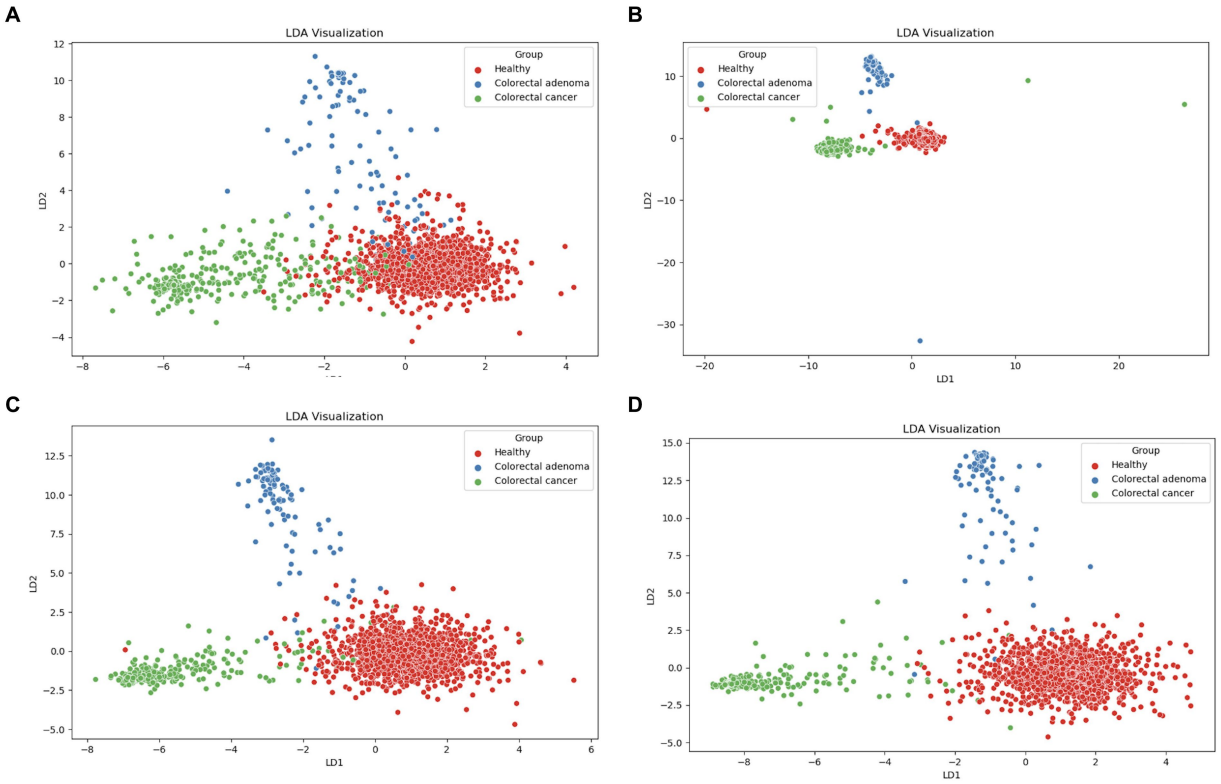

FIGURE 2
LDA scores plots of components one and two for healthy (red), patients with CRC (green) and CRA (blue): **(A)** taxonomy, **(B)** gene families, **(C)** enzymatic reactions, and **(D)** metabolic pathways.

The model was tested on 30% of the entire dataset and achieved an accuracy of AUC = 0.822.

The aggregation of functional gene information into enzymatic reactions (Figure 3C) led us to model 34,008 enzymatic reactions. The best model was Classification Random Forest training 1,000 trees with deviance splitting criterion, minimum leaf size = 1, splits = 1, alpha = 1 and variables to split = 0.577 sqrt (nvars), with an Area under the Curve (AUC) value of 0.825. 25 different features were identified as the most important for discrimination and the first 18 of them can achieve a prediction performance of 100% (Supplementary Figure S5; Supplementary Table S4). The model was tested and achieved a performance with an AUC value of 0.812.

The aggregation of enzymatic reactions into metabolic pathways (Figure 3D) led to the modeling of 31,555 metabolic pathways. The best model was Classification Random Forest training 100 trees with deviance splitting criterion, minimum leaf size = 2, splits = 1, alpha = 1 and variables to split = 0.577 sqrt (nvars), with an area under the curve (AUC) value of 0.799.25 different features were identified as the most important for discrimination and the first 13 of them can reach a prediction performance of 100% (Supplementary Figure S6; Supplementary Table S4). The model was tested on the test dataset and achieved a performance with an AUC value of 0.768.

The LDA analysis and clustering visualizations have already shown that the lowest expected performance can be obtained when modeling the predicted metabolite data obtained with the MelonnPan tool (Mallick et al., 2019). This was also confirmed with Classification Random Forest training 1,000 trees with deviance splitting criterion, minimum leaf size = 2, splits = 1, alpha = 1 and variables to split = 1.0 sqrt (nvars) as the best prediction algorithm based on 81 predicted metabolites. However, the performance of this model was low (AUC = 0.621). The performance on the test dataset was even lower (AUC = 0.606) (Supplementary Figures S7, S8; Supplementary Table S4).

All features identified by JADBIO through automatic machine learning were also tested using the Mann–Whitney statistics to check correctness and significance between groups for each feature. Most comparisons for each feature in the areas of taxonomy, functional genes, enzymatic reactions, and metabolic pathways were statistically significant, especially when comparing CRC and healthy controls. Comparisons of CRA and healthy controls on the one hand or CRC and CRA on the other were less significant. The differences in the selected predicted metabolites were not significant (Supplementary Table S5).

In the final step of the machine learning analysis, the most important features were integrated into a data set and the machine learning was repeated on this reduced data set. Classification Random Forest trained 1,000 trees with deviance splitting criterion, minimum leaf size = 3, splits = 1, alpha = 1 and variables to split 0.816 sqrt was selected as the most successful for aggressive feature selection and 25 out of 120 features were selected as the most important for classification (5 belong to taxonomy–kingdom bacteria, 12 to gene families, 5 to enzymatic reactions and 3 to metabolic pathways). None of the predicted metabolites from the first step were selected in the second step. The final performance of this model was 0.87 (AUC).

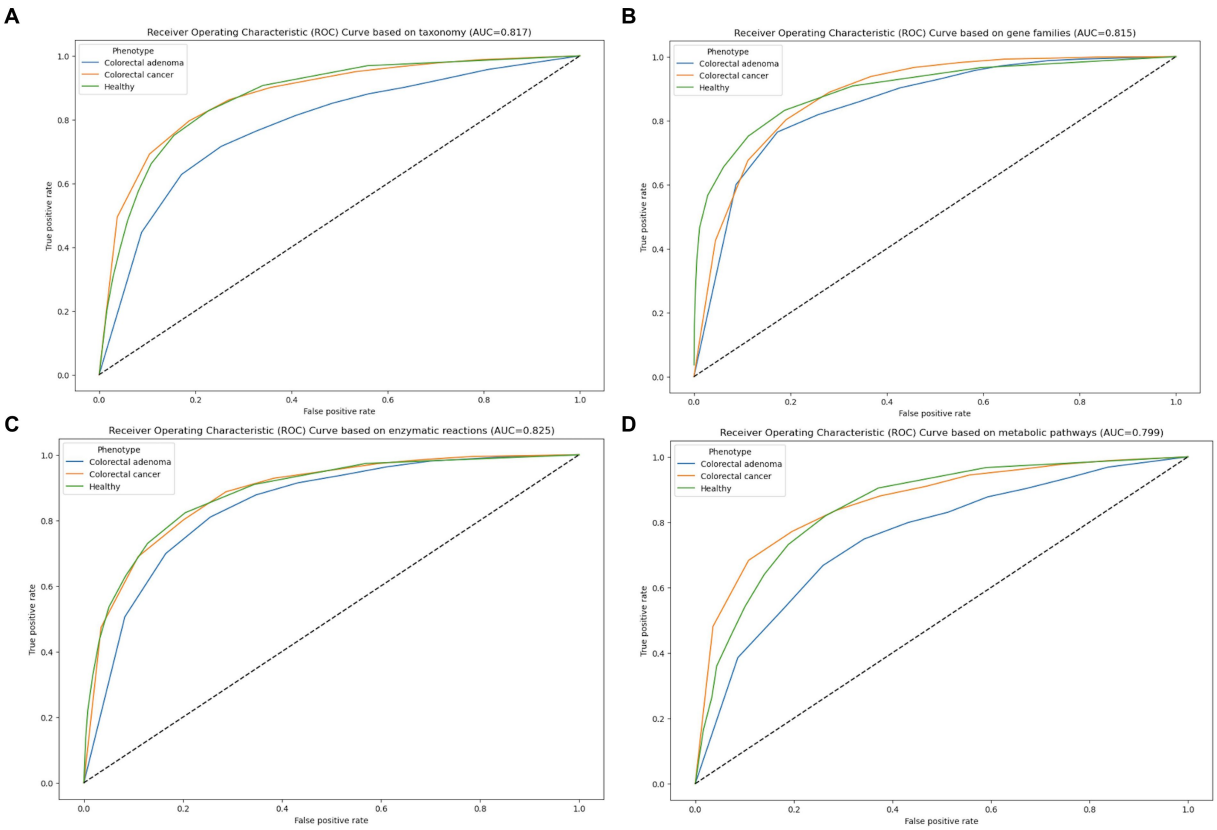

FIGURE 3
ROC plots for classification between healthy individuals (green), CRC (orange) and CRA (blue) patients based on taxonomy **(A)**, functional genes **(B)**, enzymatic reactions, **(C)** and metabolic pathways **(D)**.

4 Discussion

A total of 2,951 stool samples from different studies, including healthy individuals as well as those with CRC and CRA, were subjected to comparative analysis. Our MetaBakery pipeline was used for sequence processing. Comprehensive data matrices were used that included various features such as microbial taxonomy (1839 taxonomic units), functional genes (80,372 genes), enzymatic reactions (34,008 enzymes), metabolic pathways (31,555 metabolic pathways), and predicted metabolites (81 metabolites). In addition, we integrated 19 different diversity matrices calculated using methods consistent with Mothur's approach.

We showed that there is no statistically significant difference in microbial diversity in patients with colorectal cancer (CRC). These results are consistent with some other studies suggesting that microbial diversity and richness may increase in colorectal cancer patients (Feng et al., 2015; Thomas et al., 2019; Qi et al., 2022; Liu J. et al., 2023). To further investigate possible differences, we first performed a comprehensive analysis of the entire dataset using linear discriminant analysis (LDA) to identify possible clusters. Significant differences emerged in four different metagenomic data matrices (taxonomy, functional genes, enzymatic reactions and metabolic pathways), which formed separate clusters for each group (healthy, CRC, CRA). A clear difference was seen between the healthy and CRC patient groups. However, the CRA patients were consistently positioned between the healthy controls and the CRC patients, emphasizing that CRA represents a closer step to the development of CRC in terms of the composition of the microbiome. CRA is considered as a stage 0 in development of intramucosal carcinoma and can progress into malignant forms, which is also known as an adenoma-carcinoma sequence. The most important question here is whether the change in the microbiome is the consequence of the development of the disease or whether the disease is a consequence of the change in the microbiome. Given the obvious differences observed in LDA analysis between healthy microbiomes, CRC and CRA samples, machine learning (ML) analysis was performed. Datasets from different studies were used to represent real-world scenarios and achieve a level of variability that corresponds to natural conditions rather than exerting excessive control.

We obtained consistent model performance with AUC values around 0.8 for all data inputs. In this study, we present several groups of microbial taxa, functional genes, enzymatic reactions and metabolic pathways that offer potential for the prediagnostic evaluation of CRC and CRA that represent an early stage in the development of CRC. Several CRC biomarker species were independently identified in the different studies by univariate statistics (Segata et al., 2011): *Fusobacterium nucleatum*, *Solobacterium moorei*, *Porphyromonas asaccharolytica*, *Parvimonas micra*, *Peptostreptococcus stomatis* and *Parvimonas* ssp. (Kostic et al., 2012, 2013; Thomas et al., 2019; Mizutani et al., 2020; Qi et al., 2022). In our study different groups of taxa, from phylum to genera, were identified important for distinguishing between different conditions (health, CRC or CRA).

Many previous studies focused exclusively on a binary classification including only colorectal cancers and healthy individuals, which may have introduced bias. The detection of individuals with CRA, a precursor of CRC, is important from a diagnostic point of view.

In recent years, research into the functionality of the microbiome has become increasingly important. The emergence of microbial metagenomics has highlighted that data modeling must also be approached from the perspective of microbial functionality, as the ratio of human to microbial genes is 1:150 (Qin et al., 2010). This shift is crucial as it provides a better understanding of overall microbial functionality rather than microbial taxonomy (Deschênes et al., 2023). Furthermore, it promises to reveal why certain components of the microbiome may be associated with the occurrence of various diseases. With this in mind, our investigations extend to microbial functional potential, which includes functional genes, enzymatic reactions, metabolic pathways and predicted metabolites.

Our initial focus on functional genes, enzymatic reactions and metabolic pathways has led to promising results and moderate classification accuracy. Based on the UniRef database (Suzek et al., 2007, 2015), 15 different gene families were discovered that are important for classification between all three groups. Most of the discovered gene families belong to the human gut microbiota. Moreover, for example, the gene family A0A015S3B6|unclassified belongs to the protein of *Bacteroides fragilis*, which has also been previously mentioned as one of the biomarker candidates for CRC (Pandey et al., 2023). The gene family A0A078RCV9 belongs to *Phocaeicola vulgatus*, (formerly *Bacteroides vulgatus*, which was already associated with CRC in 1995) (Moore and Moore, 1995; Lucas et al., 2017; Vu et al., 2022). The gene families A0A174XNP7 (belonging to *Flavonifractor plautii*) and A0A174Q9G9 (*Bacteroides intestinalis*) have been associated with colorectal cancer patients in India (Gupta et al., 2019).

The most important enzymatic reaction is 3.5.1.88-RXN according to feature selection, which belongs to *Holdemanella biformis*, one of the species that can act anti-oncogenically through the production of SCFAs (Zagato et al., 2020). Reaction 3.4.21.92-RXN belongs to *Lawsonibacter asaccharolyticus*, previously associated with acetate, a potential therapeutic agent in the treatment of colorectal cancer (Marques et al., 2013; Sahuri-Arisoylu et al., 2021; Dong et al., 2023). Reatcion 3.2.1.1-RXN belongs to *Clostridium* sp. *CAG_58*, the most important taxon from the taxonomic data feature selection, was previously associated with adiposity. Higher obesity has generally been associated with an increased likelihood of CRC (Bull et al., 2020; Asnicar et al., 2021). Reaction 2.5.1.64-RXN belongs to *Klebsiella oxytoca*, another microbial species that has been isolated from patients with CRC and is one of the reasons for the increased inflammation in these patients due to biofilm formation (Abbas et al., 2020). One of the most interesting features discovered in the enzymatic reactions was 2.3.1.180-RXN belonging to *Fusobacterium nucleatum*, which, as mentioned above, was one of the most important species-level biomarkers observed in other studies (Kostic et al., 2012, 2013). Even though we did not observe this species at the taxonomic level, we did observe this reaction. Reaction 2PGADEHYDRAT-RXN was also identified and belongs to *Collinsella aerofaciens*, a microbe observed in the stool of patients with elevated blood levels (Chénard et al., 2020).

MetaCyc (Caspi et al., 2020) metabolic pathways were also identified as important features for classification. The most important feature in this regard was ARO-PWY: chorismate biosynthesis I. Chorismate is also a precursor of tryptophan. It was observed that the reduction in the amount of tryptophan is proportional to the poor quality of life of colorectal cancer patients (Zhang et al., 2019). The next metabolic pathway was ARGSYN-PWY: L-arginine biosynthesis I. It was observed that supplementation with L-arginine can alleviate intestinal inflammation. Increased intestinal inflammation was observed to be associated with the initiation and progression of CRC (Zhang et al., 2021; Liu Y. et al., 2023). Arginine was also observed to have significant diagnostic value for CRC patients (Yi et al., 2023).

However, the AUC values for the predicted metabolites were lower compared to other data matrices. Pantothenate was observed to be the most important feature. Pantothenate was previously observed as an important metabolite for the diagnosis of CRC patients (Yi et al., 2023). Putrescine, the second most important feature, is a polyamine that is basically involved in all steps of tumorigenesis (Sánchez-Alcoholado et al., 2021).

Although there are still no definitive explanations for many discovered genes, enzymes and metabolic pathways, this uncertainty will decrease over time. For example, it is expected that questions about the significance of a particular metabolic pathway for the classification of a particular disease will be clarified. We have also ventured into the prediction of metabolites using relaxation networks such as those included in MelonnPan. Although the results were statistically insignificant, it is plausible that subsequent iterations of this tool or similar tools could improve the prediction of metabolites. This potential breakthrough could facilitate the linking of metabolite predictions with results from fecal or blood metabolome analyses (Šket et al., 2020; Deutsch et al., 2022a). Such an integrated approach could reveal new dimensions in the understanding of microbe-host relationships, enriching our knowledge and potentially paving the way for practical clinical applications. With the approach outlined in this study, we have shown that it is possible to develop robust prediagnostic methods for colorectal cancer detection based on microbial fingerprints (Cammarota et al., 2020; Su et al., 2022; Zhou et al., 2024) integrating all layers of information (taxonomy, diversity, functional genes, enzymatic reactions, metabolic pathways, metabolites). One of the limitations mirroring the current status of the research in this fields and of our study is the lack of external clinical datasets of sufficient high quality of sequences and metadata to validate our models. However, with the advent of novel datasets the models created in this study could be used in larger studies in the future to evaluate the results obtained. Nevertheless, the research presented here provides one of the first important steps toward efficient, reproducible and tractable classification of CRC and CRA samples in a form of prediagnostic informative tool.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary material.

Ethics statement

Ethical approval was not required for the study involving humans in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was not required

from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and the institutional requirements.

Author contributions

BM: Writing – review & editing, Writing – original draft, Software, Methodology, Formal analysis, Data curation. LD: Writing – review & editing, Writing – original draft, Validation, Project administration, Methodology, Investigation, Formal analysis, Data curation. BS: Writing – review & editing, Writing – original draft, Visualization, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. BM was partially supported by Slovenian Research and Innovation Agency (SRA/ARRS) program P2-0095 (Parallel and distributed systems) and project J7-50230 (Building Efficient Noncommunicable-disease Early Warning Tool). LD acknowledges the MR+ support of the Slovenian Research and Innovation Agency (SRA R#51867) awarded to BS. BS was in part supported by P2-0180 (Tools and methods for process analysis, simulation and technology development). The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Acknowledgments

The support of National Institute of Chemistry, D13 Department of Catalysis and Chemical Reaction Engineering is gratefully acknowledged. The authors gratefully acknowledge the HPC RIVR consortium (www.hpc-rivr.si) and EuroHPC JU (eurohpc-ju.europa.eu) for funding this research by providing computing resources of the HPC system Vega at the Institute of Information Science (www.izum.si). "The computational results presented have been achieved (in part) using the HPC infrastructure of the University of Innsbruck" using Leo3 and Leo4e (https://www.uibk.ac.at/zid/systeme/hpc-systeme/). The ongoing support from the side of prof. Heribert Insam (Ret.), Department of Microbiology, University of Innsbruck, Austria, is gratefully acknowledged. The COST Action ML4Microbiome (CA18131) and the research network therein is kindly acknowledged for fruitful discussions that brought our attention to this topic and prompted us to extend our work.

Conflict of interest

LD was employed by the NU B.V.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2024.1426407/full#supplementary-material

SUPPLEMENTARY TABLE S1
Accession numbers of used studies.

SUPPLEMENTARY TABLE S2
The list of tested model configurations.

SUPPLEMENTARY TABLE S3
Mann–Whitney statistic of diversity metrics.

SUPPLEMENTARY TABLE S4
Important features identified by JADBIO Auto-ML approach.

SUPPLEMENTARY TABLE S5
Mann–Whitney statistics of important features, extended discussion—limitations of ML approaches in microbiome studies, extended discussion—rationale of using paired read sequences.

References

Abbas, A. F., Al-Saadi, A. G. M., and Alkhudhairy, M. K. (2020). Biofilm formation and virulence determinants of *Klebsiella oxytoca* clinical isolates from patients with colorectal cancer. *J. Gastrointest. Cancer* 51, 855–860. doi: 10.1007/S12029-019-00317-7

Asnicar, F., Berry, S. E., Valdes, A. M., Nguyen, L. H., Piccinno, G., Drew, D. A., et al. (2021). Microbiome connections with host metabolism and habitual diet from 1, 098 deeply phenotyped individuals. *Nat. Med.* 27, 321–332. doi: 10.1038/S41591-020-01183-8

Baxter, N. T., Ruffin, M. T., Rogers, M. A. M., and Schloss, P. D. (2016). Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. *Genome Med.* 8:37. doi: 10.1186/S13073-016-0290-3

Beghini, F., McIver, L. J., Blanco-Míguez, A., Dubois, L., Asnicar, F., Maharjan, S., et al. (2021). Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bio bakery 3. *eLife* 10:65088. doi: 10.7554/ELIFE.65088

Bull, C. J., Bell, J. A., Murphy, N., Sanderson, E., Davey Smith, G., Timpson, N. J., et al. (2020). Adiposity, metabolites, and colorectal cancer risk: Mendelian randomization study. *BMC Med.* 18:396. doi: 10.1186/S12916-020-01855-9

Cammarota, G., Ianiro, G., Ahern, A., Carbone, C., Temko, A., Claesson, M. J., et al. (2020). Gut microbiome, big data and machine learning to promote precision medicine for cancer. *Nat. Rev. Gastroenterol. Hepatol.* 17, 635–648. doi: 10.1038/S41575-020-0327-3

Caspi, R., Billington, R., Keseler, I. M., Kothari, A., Krummenacker, M., Midford, P. E., et al. (2020). The Meta Cyc database of metabolic pathways and enzymes – a 2019 update. *Nucleic Acids Res.* 48, D445–D453. doi: 10.1093/NAR/GKZ862

Chénard, T., Malick, M., Dubé, J., and Massé, E. (2020). The influence of blood on the human gut microbiome. *BMC Microbiol.* 20, 1–10. doi: 10.1186/S12866-020-01724-8/TABLES/2

Deschênes, T., Tohoundjona, F. W. E., Plante, P.-L., Di Marzo, V., and Raymond, F. (2023). Gene-based microbiome representation enhances host phenotype classification. *mSystems*. doi: 10.1128/MSYSTEMS.00531-23

Deutsch, L. (2022). Bioinformatics integration of microbiome and metabolomics data in the translational context: Doctoral dissertation. University of Ljubljana, Ljubljana, Slovenia. Available at: https://repozitorij.uni-lj.si/IzpisGradiva.php?id=142851&lang=eng

Deutsch, L., Debevec, T., Millet, G. P., Osredkar, D., Opara, S., Šket, R., et al. (2022a). Urine and Fecal1H-NMR metabolomes differ significantly between pre-term and full-term born physically fit healthy adult males. *Meta* 12:536. doi: 10.3390/metabo12060536

Deutsch, L., Sotiridis, A., Murovec, B., Plavec, J., Mekjavic, I., Debevec, T., et al. (2022b). Exercise and Interorgan communication: short-term exercise training blunts differences in consecutive daily Urine1H-NMR Metabolomic signatures between physically active and inactive individuals. *Meta* 12:473. doi: 10.3390/metabo12060473

Deutsch, L., and Stres, B. (2021). The importance of objective stool classification in fecal 1H-NMR metabolomics: exponential increase in stool crosslinking is mirrored in systemic inflammation and associated to fecal acetate and methionine. *Meta* 11:172. doi: 10.3390/metabo11030172

Dong, Y., Zhang, K., Wei, J., Ding, Y., Wang, X., Hou, H., et al. (2023). Gut microbiota-derived short-chain fatty acids regulate gastrointestinal tumor immunity: a novel therapeutic strategy? *Front. Immunol.* 14:1158200. doi: 10.3389/FIMMU.2023.1158200

Durazzi, F., Sala, C., Castellani, G., Manfreda, G., Remondini, D., and De Cesare, A. (2021). Comparison between 16S rRNA and shotgun sequencing data for the taxonomic characterization of the gut microbiota. *Sci. Rep.* 11, 1–10. doi: 10.1038/s41598-021-82726-y

Feng, Q., Liang, S., Jia, H., Stadlmayr, A., Tang, L., Lan, Z., et al. (2015). Gut microbiome development along the colorectal adenoma-carcinoma sequence. *Nat. Commun.* 6:6528. doi: 10.1038/NCOMMS7528

Frank, C., Sundquist, J., Yu, H., Hemminki, A., and Hemminki, K. (2017). Concordant and discordant familial cancer: familial risks, proportions and population impact. *Int. J. Cancer* 140, 1510–1516. doi: 10.1002/IJC.30583

Gupta, A., Dhakan, D. B., Maji, A., Saxena, R., Vishnu Prasoodanan, P. K., Mahajan, S., et al. (2019). Association of *Flavonifractor plautii*, a flavonoid-degrading bacterium, with the gut microbiome of colorectal Cancer patients in India. *mSystems* 4:438. doi: 10.1128/MSYSTEMS.00438-19

Gupta, V. K., Kim, M., Bakshi, U., Cunningham, K. Y., Davis, J. M., Lazaridis, K. N., et al. (2020). A predictive index for health status using species-level gut microbiome profiling. *Nat. Commun.* 11:4635. doi: 10.1038/S41467-020-18476-8

He, Y., Wu, W., Zheng, H. M., Li, P., McDonald, D., Sheng, H. F., et al. (2018). Regional variation limits applications of healthy gut microbiome reference ranges and disease models. *Nat. Med.* 24, 1532–1535. doi: 10.1038/s41591-018-0164-x

Hunter, J. D. (2007). Matplotlib: a 2D graphics environment. *Comput Sci Eng* 9, 90–95. doi: 10.1109/MCSE.2007.55

Huxley, R. R., Ansary-Moghaddam, A., Clifton, P., Czernichow, S., Parr, C. L., and Woodward, M. (2009). The impact of dietary and lifestyle risk factors on risk of colorectal cancer: a quantitative overview of the epidemiological evidence. *Int. J. Cancer* 125, 171–180. doi: 10.1002/IJC.24343

Johnson, C. M., Wei, C., Ensor, J. E., Smolenski, D. J., Amos, C. I., Levin, B., et al. (2013). Meta-analyses of colorectal cancer risk factors. *Cancer Causes Control* 24, 1207–1222. doi: 10.1007/S10552-013-0201-5

Karsa, L. V., Lignini, T. A., Patnick, J., Lambert, R., and Sauvaget, C. (2010). The dimensions of the CRC problem. *Best Pract. Res. Clin. Gastroenterol.* 24, 381–396. doi: 10.1016/J.BPG.2010.06.004

Kostic, A. D., Chun, E., Robertson, L., Glickman, J. N., Gallini, C. A., Michaud, M., et al. (2013). *Fusobacterium nucleatum* potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment. *Cell Host Microbe* 14, 207–215. doi: 10.1016/J.CHOM.2013.07.007

Kostic, A. D., Gevers, D., Pedamallu, C. S., Michaud, M., Duke, F., Earl, A. M., et al. (2012). Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. *Genome Res.* 22, 292–298. doi: 10.1101/GR.126573.111

Liu, J., Huang, X., Chen, C., Wang, Z., Huang, Z., Qin, M., et al. (2023). Identification of colorectal cancer progression-associated intestinal microbiome and predictive signature construction. *J. Transl. Med.* 21:373. doi: 10.1186/S12967-023-04119-1

Liu, Y., Lau, H. C. H., Cheng, W. Y., and Yu, J. (2023). Gut microbiome in colorectal Cancer: clinical diagnosis and treatment. *Genomics Proteomics Bioinformatics* 21, 84–96. doi: 10.1016/J.GPB.2022.07.002

Lucas, C., Barnich, N., and Nguyen, H. T. T. (2017). Microbiota, inflammation and colorectal Cancer. *Int. J. Mol. Sci.* 18:310. doi: 10.3390/IJMS18061310

Mallick, H., Franzosa, E. A., Mclver, L. J., Banerjee, S., Sirota-Madi, A., Kostic, A. D., et al. (2019). Predictive metabolomic profiling of microbial communities using amplicon or metagenomic sequences. *Nat. Commun.* 10, 1–11. doi: 10.1038/s41467-019-10927-1

Marques, C., Oliveira, C. S. F., Alves, S., Chaves, S. R., Coutinho, O. P., Côrte-Real, M., et al. (2013). Acetate-induced apoptosis in colorectal carcinoma cells involves lysosomal membrane permeabilization and cathepsin D release. *Cell Death Dis.* 4:e507. doi: 10.1038/CDDIS.2013.29

Mizutani, S., Yamada, T., and Yachida, S. (2020). Significance of the gut microbiome in multistep colorectal carcinogenesis. *Cancer Sci.* 111, 766–773. doi: 10.1111/CAS.14298

Moore, W. E. C., and Moore, L. H. (1995). Intestinal floras of populations that have a high risk of colon cancer. *Appl. Environ. Microbiol.* 61, 3202–3207. doi: 10.1128/AEM.61.9.3202-3207.1995

Murovec, B., Deutsch, L., and Stres, B. (2021). General unified microbiome profiling pipeline (Gumpp) for large scale, streamlined and reproducible analysis of bacterial 16s rrna data to predicted microbial metagenomes, enzymatic reactions and metabolic pathways. *Meta* 11:336. doi: 10.3390/metabo11060336

Pandey, H., Tang, D. W. T., Wong, S. H., and Lal, D. (2023). Gut microbiota in colorectal Cancer: biological role and therapeutic opportunities. *Cancers (Basel)* 15:866. doi: 10.3390/CANCERS15030866

Pasolli, E., Truong, D. T., Malik, F., Waldron, L., and Segata, N. (2016). Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. *PLoS Comput. Biol.* 12:e1004977. doi: 10.1371/JOURNAL.PCBI.1004977

Pedregosa, F., Michel, V., Grisel Oliviergrisel, O., Blondel, M., Prettenhofer, P., Weiss, R., et al. (2011). Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.

Qi, Z., Zhibo, Z., Jing, Z., Zhanbo, Q., Shugao, H., Weili, J., et al. (2022). Prediction model of poorly differentiated colorectal cancer (CRC) based on gut bacteria. *BMC Microbiol.* 22:312. doi: 10.1186/S12866-022-02712-W

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464, 59–65. doi: 10.1038/NATURE08821

Rubinstein, M. R., Wang, X., Liu, W., Hao, Y., Cai, G., and Han, Y. W. (2013). *Fusobacterium nucleatum* promotes colorectal carcinogenesis by modulating E-cadherin/β-catenin signaling via its FadA adhesin. *Cell Host Microbe* 14, 195–206. doi: 10.1016/J.CHOM.2013.07.012

Sahuri-Arisoylu, M., Mould, R. R., Shinjyo, N., Bligh, S. W. A., Nunn, A. V. W., Guy, G. W., et al. (2021). Acetate induces growth arrest in Colon Cancer cells through modulation of mitochondrial function. *Front. Nutr.* 8:588466. doi: 10.3389/FNUT.2021.588466

Sánchez-Alcoholado, L., Laborda-Illanes, A., Otero, A., Ordóñez, R., González-González, A., Plaza-Andrades, I., et al. (2021). Relationships of gut microbiota composition, short-chain fatty acids and polyamines with the pathological response to neoadjuvant radiochemotherapy in colorectal cancer patients. *Int. J. Mol. Sci.* 22:549. doi: 10.3390/ijms22179549

Schloss, P. D. (2020). Reintroducing mothur: 10 years later. *Appl. Environ. Microbiol.* 86:2343. doi: 10.1128/AEM.02343-19

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., et al. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* 75, 7537–7541. doi: 10.1128/AEM.01541-09

Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W. S., et al. (2011). Metagenomic biomarker discovery and explanation. *Genome Biol.* 12:R60. doi: 10.1186/GB-2011-12-6-R60

Siegel, R., DeSantis, C., and Jemal, A. (2014). Colorectal cancer statistics, 2014. *CA Cancer J. Clin.* 64, 104–117. doi: 10.3322/CAAC.21220

Šket, R., Deutsch, L., Prevoršek, Z., Mekjavić, I. B., Plavec, J., Rittweger, J., et al. (2020). Systems view of deconditioning during spaceflight simulation in the PlanHab project: the departure of urine 1 H-NMR metabolomes from healthy state in young males subjected to bedrest inactivity and hypoxia. *Front. Physiol.* 11:1550. doi: 10.3389/fphys.2020.532271

Su, Q., Liu, Q., Lau, R. I., Zhang, J., Xu, Z., Yeoh, Y. K., et al. (2022). Faecal microbiome-based machine learning for multi-class disease diagnosis. *Nat. Commun.* 13:6818. doi: 10.1038/s41467-022-34405-3

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., et al. (2021). Global Cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* 71, 209–249. doi: 10.3322/CAAC.21660

Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R., and Wu, C. H. (2007). UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* 23, 1282–1288. doi: 10.1093/BIOINFORMATICS/BTM098

Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., and Wu, C. H. (2015). UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* 31, 926–932. doi: 10.1093/BIOINFORMATICS/BTU739

Thomas, A. M., Manghi, P., Asnicar, F., Pasolli, E., Armanini, F., Zolfo, M., et al. (2019). Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nat. Med.* 25, 667–678. doi: 10.1038/S41591-019-0405-7

Tsamardinos, I., Charonyktakis, P., Papoutsoglou, G., Borboudakis, G., Lakiotaki, K., Zenklusen, J. C., et al. (2022). Just add data: automated predictive modeling for knowledge discovery and feature selection. *NPJ Precision Oncol.* 6, 38–17. doi: 10.1038/s41698-022-00274-8

Van Rossum, G., and Drake, F. L. (2009). Python 3 reference manual. Scotts Valley, CA: CreateSpace.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* 17, 261–272. doi: 10.1038/s41592-019-0686-2

Vu, H., Muto, Y., Hayashi, M., Noguchi, H., Tanaka, K., Yamamoto, Y., et al. (2022). Complete genome sequences of three Phocaeicola vulgatus strains isolated from a healthy Japanese individual. *Microbiol. Resour. Announ.* 11, e0112421–e0111145. doi: 10.1128/MRA.01124-21

Wang, W. L., Xu, S. Y., Ren, Z. G., Tao, L., Jiang, J. W., and Zheng, S. S. (2015). Application of metagenomics in the human gut microbiome. *World J. Gastroenterol.* 21, 803–814. doi: 10.3748/WJG.V21.I3.803

Waskom, M. (2021). seaborn: statistical data visualization. *J. Open Source Softw.* 6:3021. doi: 10.21105/joss.03021

Wong, C. C., and Yu, J. (2023). Gut microbiota in colorectal cancer development and therapy. *Nat. Rev. Clin. Oncol.* 20, 429–452. doi: 10.1038/S41571-023-00766-X

Yi, Y., Wang, J., Liang, C., Ren, C., Lian, X., Han, C., et al. (2023). LC-MS-based serum metabolomics analysis for the screening and monitoring of colorectal cancer. *Front. Oncol.* 13:1173424. doi: 10.3389/FONC.2023.1173424

Yu, J., Feng, Q., Wong, S. H., Zhang, D., Yi Liang, Q., Qin, Y., et al. (2017). Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. *Gut* 66, 70–78. doi: 10.1136/GUTJNL-2015-309800

Zackular, J. P., Rogers, M. A. M., Ruffin, M. T., and Schloss, P. D. (2014). The human gut microbiome as a screening tool for colorectal cancer. *Cancer Prev. Res. (Phila.)* 7, 1112–1121. doi: 10.1158/1940-6207.CAPR-14-0129

Zagato, E., Pozzi, C., Bertocchi, A., Schioppa, T., Saccheri, F., Guglietta, S., et al. (2020). Endogenous murine microbiota member *Faecalibaculum rodentium* and its human homologue protect from intestinal tumour growth. *Nat. Microbiol.* 5, 511–524. doi: 10.1038/S41564-019-0649-5

Zeller, G., Tap, J., Voigt, A. Y., Sunagawa, S., Kultima, J. R., Costea, P. I., et al. (2014). Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.* 10:766. doi: 10.15252/MSB.20145645

Zhang, W., An, Y., Qin, X., Wu, X., Wang, X., Hou, H., et al. (2021). Gut microbiota-derived metabolites in colorectal Cancer: the bad and the challenges. *Front. Oncol.* 11:739648. doi: 10.3389/FONC.2021.739648

Zhang, H. L., Zhang, A. H., Miao, J. H., Sun, H., Yan, G. L., Wu, F. F., et al. (2019). Targeting regulation of tryptophan metabolism for colorectal cancer therapy: a systematic review. *RSC Adv.* 9, 3072–3080. doi: 10.1039/C8RA08520J

Zhou, D., Chen, Y., Wang, Z., Zhu, S., Zhang, L., Song, J., et al. (2024). Integrating clinical and cross-cohort metagenomic features: a stable and non-invasive colorectal cancer and adenoma diagnostic model. *Front. Mol. Biosci.* 10:1298679. doi: 10.3389/FMOLB.2023.1298679

Frontiers in Microbiology

Explores the habitable world and the potential of microbial life

The largest and most cited microbiology journal which advances our understanding of the role microbes play in addressing global challenges such as healthcare, food security, and climate change.

Discover the latest Research Topics

See more →

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact