frontiers | Frontiers in Microbiology

TYPE Perspective
PUBLISHED 25 September 2023
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

Check for updates

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

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/

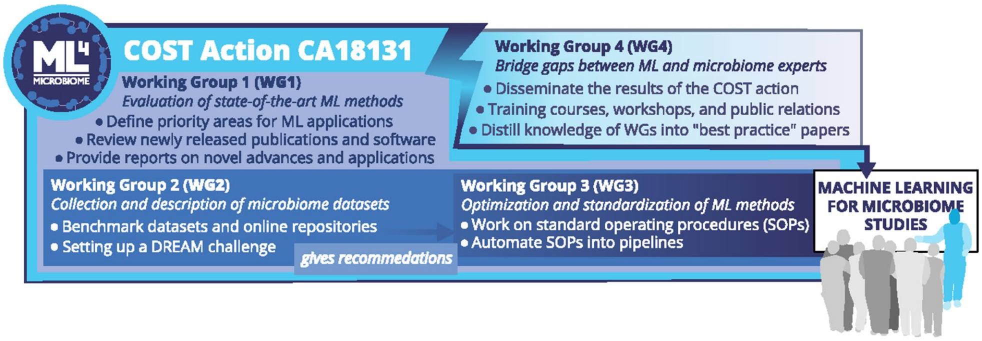
ML4 MICROBIOME
COST Action CA18131
Working Group 1 (WG1)
Evaluation of state-of-the-art ML methods
• Define priority areas for ML applications
• Review newly released publications and software
• Provide reports on novel advances and applications
Working Group 4 (WG4)
Bridge gaps between ML and microbiome experts
• Disseminate the results of the COST action
• Training courses, workshops, and public relations
• Distill knowledge of WGs into "best practice" papers
Working Group 2 (WG2)
Collection and description of microbiome datasets
• Benchmark datasets and online repositories
• Setting up a DREAM challenge
gives recomendations
Working Group 3 (WG3)
Optimization and standardization of ML methods
• Work on standard operating procedures (SOPs)
• Automate SOPs into pipelines
MACHINE LEARNING FOR MICROBIOME STUDIES

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

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.

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