A toolbox of machine learning software to support microbiome analysis

Marcos-Zambrano, Laura Judith

Frontiers Media SA
2023-11-22

Marcos-Zambrano, L J, Lopez-Molina, V M, 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, O, Berland, M, Calle, M L, Ceci, M, Duman, H, Guendogdu, A, Havulinna, A S, Kaka Bra, K H N, Kalluci, E, Karav, S, Lode, D, Lopes, M B, May, P, Nap, B, Nedyalkova, M, Paciencia, I, Pasic, L, Pujolassos, M, Shigdel, R, Susin, A, Thiele, I, Truica, C-O, Wilmes, P, Yilmaz, E, Yousef, M, Claesson, M J, Truu, J & Carrillo de Santa Pau, E 2023, 'A toolbox of machine learning software to support microbiome analysis', Frontiers in Microbiology, vol. 14. https://doi.org/10.3389/fmicb.2023.12508

http://hdl.handle.net/10138/569084
10.3389/fmicb.2023.1250806

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.

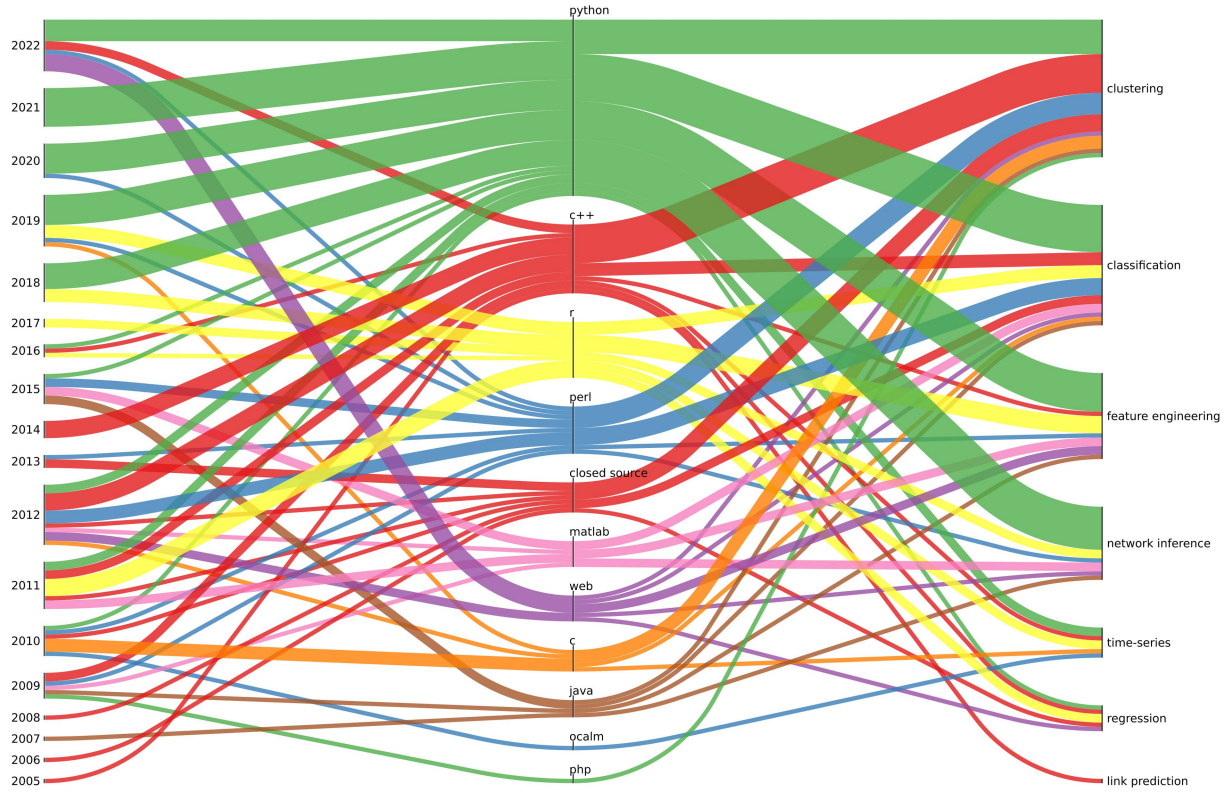

FIGURE 1
The relationship between the year of publication (left), programming language (centre), and ML task (right) is depicted for the most commonly used software in microbiome analysis. The thickness of the line represents the quantity of software projects associated with a particular relationship (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).

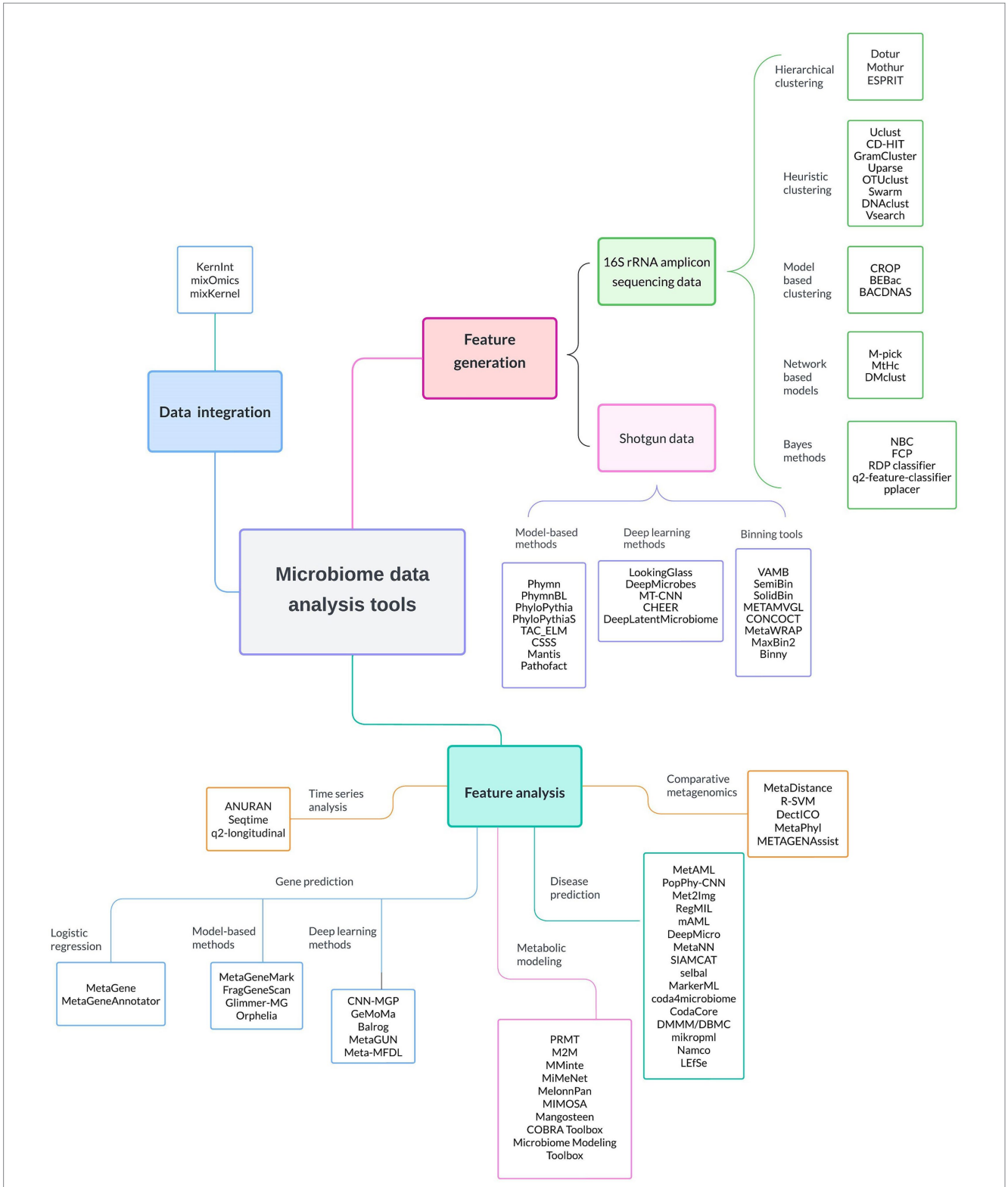

FIGURE 2
Comprehensive overview of the most commonly ML-based software applications employed in microbiome data analysis. These software tools are categorized based on their primary application into feature generation, feature analysis, and data integration. It is worth noting that numerous software 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