The Potential for Machine Learning in Multi-Omics-based Microbiome Research

Student: Maxime van Sloten s3247171
Supervisor: dr. ir. H.J.M. Harmsen
12 February 2023

Abstract

In the last few decades extensive research has been done on the human microbiome, which is the sum of all microorganisms that live within and on our body. While the microbiome itself is influenced by many factors, it in turn also plays a vital role in influencing the health and disease of human beings. It is therefore no surprise that an increasing amount of effort has been put in gathering data on the human microbiome. The development of high-throughput technologies has allowed for the development of longitudinally personalized multi-omics profiling. Due to the complicated web of interactions between the human microbiome and the host, it is extremely complicated to determine significant health-related associations with certainty. A possible solution that has gotten considerable interest from scientists in recent years has been the rapidly developing field of machine learning (ML) due to its potential for integrating large datasets, creating models and predicting phenotypes. ML can generally be subdivided into two groups; unsupervised learning (like PCA and PCoA) and supervised learning (like SVM and RF) with deep learning as a technique that can be both unsupervised or supervised. ML applications have had positive results in recent years, as can be seen in microbiome studies on antimicrobial resistance and cancer where ML strategies rivaled and/or eclipsed traditional analyses. However, concepts such as the curse of dimensionality, high-quality data and interpretability are still problematic. Luckily, studies have provided solutions to these problems, such as autoencoders like VAE, data augmentation like MetaNN and deep forest algorithms. For this reason, a selection of these solutions should be more publicly used and perhaps even standardized in order to improve scientific quality of results in microbiome research.

Table of Contents **Page**

1. Introduction

In the last few decades extensive research has been done on the gut human microbiome, which is the sum of all microorganisms that live within our gastrointestinal tract, including the bacteria (mostly strict anaerobes), fungi, archaea, viruses and protozoans. It is highly specific to the individual, with a wide variety of factors influencing the composition of the various microbes living in the gastrointestinal tract such as genetics, diet, mode of delivery, and many more. While the microbiome itself is influenced by many factors, it in turn also plays a vital role in influencing the health and disease of human beings (Sekirov et al., 2010). For example, a recent study from 2019, showed that treatment with a broad spectrum of antibiotics results in dramatic reduction in gut bacterial load, lowered bacterial diversity, and enhanced inflammation (Hagan et al., 2019). Not only is the relationship between antibiotics and the microbiome visible with respect to the immune system, but also in the increasingly important emergence of antibiotic-resistance. Research has shown that both the commensal and pathogenic bacteria in the gut microbiome of humans can serve as a reservoir for antimicrobial-resistance genes (ARGs), which can then be transferred to other pathogenic bacteria that may pose serious health threats (Paul et al., 2022). A rather infamous example of an antibiotic-resistant bacteria is of course MRSA (methicillin-resistant *Streptococcus aureus*) which has earned its reputation from taking many lives, and notably also in hospitals. But while this is perhaps the best-known example of antibiotic-resistant bacteria, there are a myriad more. These developments underscore the importance of microbiome research. Not only is this to the benefit of combating antibiotic resistance, but also other fields since the gut microbiome acts on many of our bodily systems. It is therefore no surprise that an increasing amount of effort has been put in gathering data on the human gut microbiome. The development of high-throughput technologies has allowed for the development of longitudinally personalized multi-omics profiling. By combining various omics fields such as metabolomics, proteomics and genomics among others it is possible to compile a better picture of the human microbiome and its relation with human health (Lloyd-Price et al., 2019, Zhou et al., 2019, iHMP RNC, 2014). Due to the complicated web of interactions between the human microbiome and the host, it is extremely complicated to determine significant health-related associations with certainty. For this reason it is important to develop methods to efficiently extract information from multi-omics data in order to identify patterns that may otherwise be lost to us. A possible method that has gotten considerable interest from scientists in recent years has been the rapidly developing field of machine learning due to its potential for integrating large datasets, creating models and predicting phenotypes (Li et al., 2022, Angermueller et al., 2016, Beam et al., 2018, Ching et al., 2018). Promising advances have been made, but there are also significant challenges in the field that require solving. Here, we will first discuss the investigation of the human microbiome, followed by some of the various related omics fields, as well as an investigation into the current use, future potential applications and remaining challenges of machine learning techniques.

2. The Microbiome

Every person has his or her body colonized by commensal microorganisms in enormous quantities. A 70 kg “reference man” has an estimated number of bacterial cells of 3.8 x 10^{13}, which is roughly equal to our own human cells (Sender et al., 2016). The collectivity of all microorganisms by type in the human intestinal tract is also referred to as the gut microbiota and the collectivity of microorganisms and their genome is called the microbiome (Gill et al., 2006). The vast majority of these bacteria are located in the gastrointestinal tract, and of these bacteria, most reside in the distal gut. The human distal gut microbiome contains over 100 times as many genes as our human genome alone. Recent studies have shown that the gut microbiota has coevolved because of the mutually beneficial relationship. This microbiota and its microbiome have given us the capacity to extract nutrients that would otherwise not have been capable (Bäckhed et al., 2005). On top of this vast amount of bacteria in our gut, it has been found that bacterial diversity is significant between humans, and may contribute to variations in normal physiology (Eckburg et al., 2005). Not only is the microbiome important for the acquisition of nutrients, but it has important contributions to the development of the immune system (Mazmanian et al., 2005), protection against gut injury (Rakoff-Nahoum et al., 2004) and energy balance (Bäckhed et al., 2004) among others. Additionally, the human gut microbiota plays big roles in various diseases such as diabetes (Qin et al., 2012), obesity (Ley et al., 2006), inflammatory bowel disease (IBD) (Lloyd-Price et al., 2019), autism (Sharon et al., 2019) and other brain disorders (Kaur et al., 2021), liver disease (Chu et al., 2019) and cardiovascular diseases (Kazemian et al., 2020) to name some. Given the multitude of systems upon which the microbiota acts, the understanding of the factors contributing to the control of the microbiome itself has been of considerable interest. Studies have shown that there are an equally vast array of factors influencing the microbiota, including genetics (Kurilshikov et al., 2021), age (Wilmanski et al., 2021), diet (Zmora et al., 2019), drugs (Hagan et al., 2019, Wu et al., 2017) and exercise (Quiroga et al., 2020). A generalized summary of the interplay between the human microbiome and human health can be seen in Figure 1.

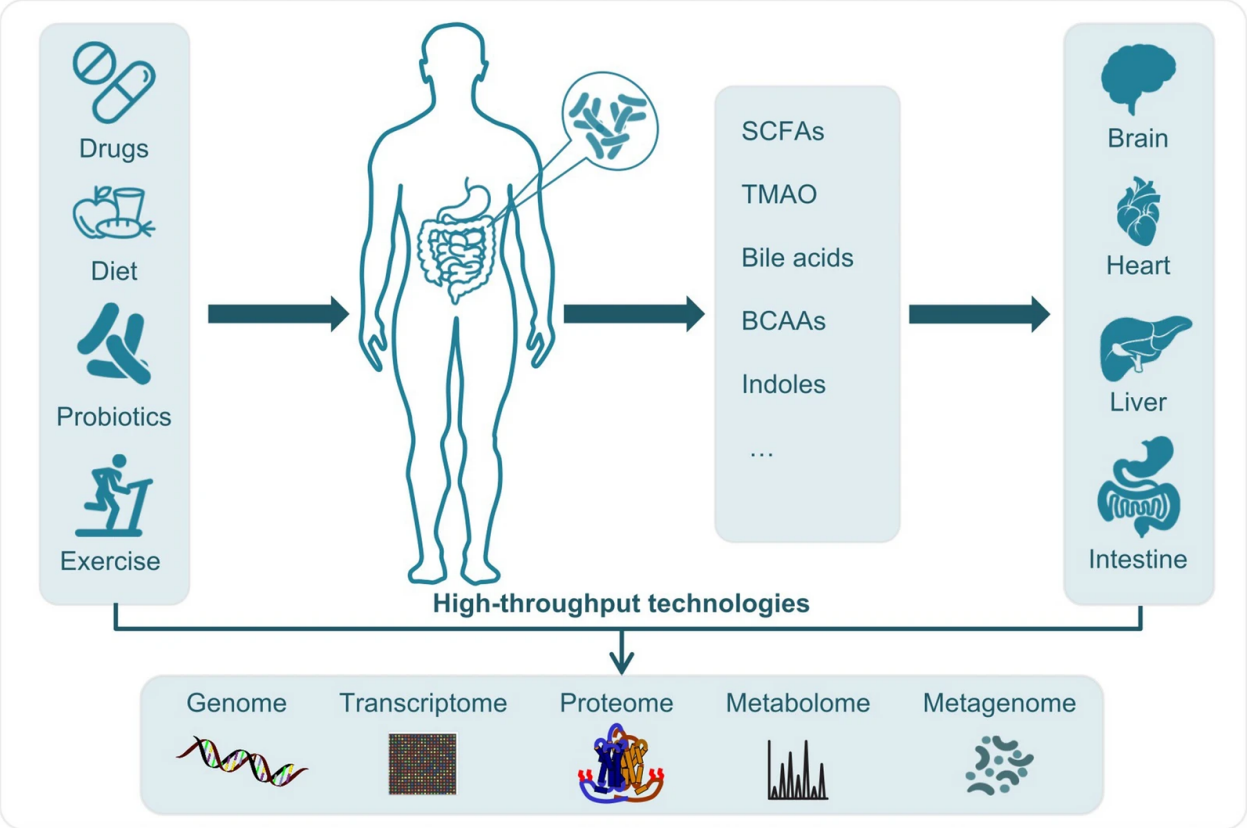

Figure 1. The complex interplay between the gut microbiome and human metabolism*, as depicted in a figure by Li et al., 2022. A collection of factors influence the microbiome, such as drugs, diet and exercise. The microbiome in turn influences the brain, heart and liver, among others, through its intermediaries like short-chain fatty acids (SCFAs), bile acids and indoles. The complexity of the interactions require high-throughput technologies to compile into multi-omics data and deepen our understanding of the associations between gut microbiota and human health.*

3. The Use of Multi-omics in Microbiome research

The intricacies of the interplay between microbiome and health have understandably led to the use of multiple '-omics' fields in order to better understand the relationship between the two and contribute to more personalized medicine. A first example is the field of genomics. Not only because of the classification of the various microorganisms within our gut microbiota community and their functions, but also because of the study of the impact of host genetics on the interactions with the microbiota (Li et al., 2022). Furthermore, the fact that our microbiome includes such an enormous amount of genetic information separate from our own genome highlights the importance of including these genes in our super-organismal view of our genetic landscape, especially given that the human gut microbiome has been implicated in the regulation of so many of our systems (Gill et al., 2006). On a similar note, a deeper understanding of our metabolome should include the metabolic networks of the microbial communities in our gut (Gill et al., 2006). Furthermore, the gut microbiome has been implicated in epigenetic regulation of brain disorders (Kaur et al., 2021), antimicrobial resistance (Huemer et al., 2020) and colorectal cancer development (Allen et al., 2019), showing the necessity of including epigenomics in the integrated model of microbiome interactions.

The National Institutes of Health Human Microbiome Project (NIH HMP) was an initiative to investigate the gut microbiome and its relationship with human health and lasted ten years, subdivided into two phases (HMP1 and HMP2). HMP1 focused on the identification and characterization of the various microbial communities of the body in a study on healthy adult subjects, and included a set of projects that included specific diseases. The huge in-depth investigation of HMP1 improved the identification of the taxonomic composition of the microbiome, but also led to the realization that that composition did not correlate well with the host phenotype, which was found to be influenced more by microbial function. HMP2 increased the amount of biological properties that were included in the study for both the host and the microbiome. It included studies of microbiome-associated conditions like pregnancy (for the vaginal microbiomes), inflammatory bowel diseases (for the gut microbiome), and prediabetes (for the gut and nasal microbiomes). The HMPs measured changes in microbial community composition, viromics, metabolomics, gene expression and protein profiles for the host and microbiome, and host-specific properties such as genomics, epigenomics, antibody and cytokine profiles, thereby expanding the resource base for microbiome research. HMP1 and HMP2 together have produced a total of 42 terabytes of multi-omic data, archived and curated by the Data Coordination Center (DCC) and available for use (iHMP RNC, 2019). There is growing evidence that the knowledge garnered by this research can provide us with ways to manipulate the gut microbiota as a potential strategy for disease treatment (Li et al., 2022). One avenue of treatment development is dietary intervention. A study by Ghosh et al., 2017, for example, shows that adherence to a Mediterranean diet alters the gut microbiome in a significant enough way that older people have reduced frailty,improved health status and better cognitive function (Ghosh et al., 2020). Another intervention that has seen increasing interest is fecal microbiota transplantation (FMT), where administration of fecal matter from a donor into the intestinal tract of a recipient has seen successes in treatment of, for example, recurrent *Clostridium difficile* infections and may be a potential treatment method for IBD, obesity, and brain disorders (Gupta et al., 2016, Xu et al., 2021).
But while an enormous amount of information was gathered from this study, it also raised more questions. Generally, it still remains a challenge to extract useful information from the huge multi-omics data for finding associations between the microbiome and host health. For this reason, it is becoming increasingly clear that advanced computational methods need to be developed to efficiently work with these big datasets.

4. Machine Learning

One of these potential methods is the use of machine learning (ML). ML is a form of artificial intelligence that is created to automatically learn and improve itself from input data without being explicitly programmed by a human. As mentioned above, high-throughput technologies have led to an enormous amount of data, which has previously been used in single-omics data analysis. However, this only tells part of the stories since the various omics fields all contribute to the phenotype. It has been argued that ML can contribute to the untangling of the various omics fields to get a comprehensive view from heterogeneous data in gut microbiota studies. For example, ML has already been used for phenotyping (both environmental and host

phenotypes), microbial classification (to determine abundance, diversity, and/or distribution), studying interactions between components of the microbiome, and monitoring changes in the microbiome (Hernández Medina et al., 2022). ML can generally be subdivided into two groups; unsupervised learning and supervised learning (See Figure 2) (Li et al., 2022).

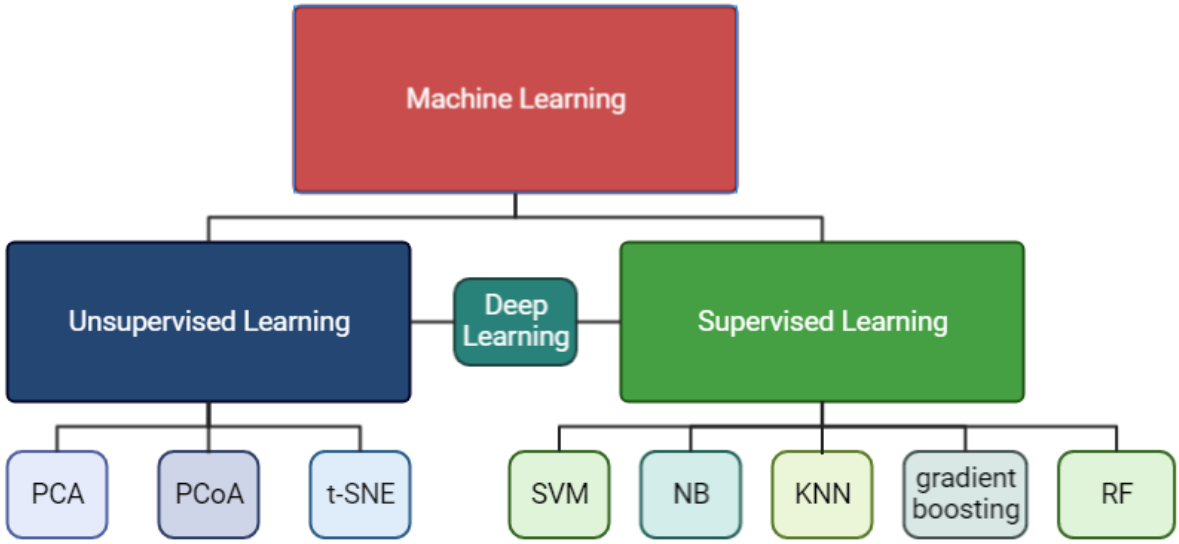

Figure 2. A summary of machine learning subdivisions and some of the most common techniques that are currently employed. *Machine learning is subdivided into two groups. 1) Unsupervised Learning techniques (such as principal component analysis (PCA), principal coordinate analysis (PCoA) and t-distributed stochastic neighbor embedding (t-SNE)). 2) Supervised Learning techniques (such as support vector machine (SVM), Naïve Bayes (NB), k-nearest neighbor (KNN), gradient boosting, random forest (RF) and others). There's also a technique called Deep Learning, which is a form of machine learning that can be unsupervised or supervised. Created with BioRender*

4.1 Unsupervised Learning

Unsupervised learning entails the discovery of new hidden patterns from a given dataset without known dependent variables. These are also called data-driven predictions. The biggest part of these can be further categorized into two groups of techniques. The first is dimension reduction, which is a transformation from high-dimensional space into a low-dimensional space that simplifies the data and makes it more practical to work with by keeping the useful properties. Such data can, for example, be used to make visualizations by taking principal variables from high-dimensional feature space (Li et al., 2022, Hernández Medina et al., 2022). Examples include popular and classical methods such as principal component analysis (PCA) and principal coordinate analysis (PCoA), which are used to visualize and contrast microbial communities. Another method is t-distributed stochastic neighbor embedding (t-SNE) that can be used to identify non-linear relationships in complex microbiome datasets (Hernández Medina et al., 2022). The other group consists of clustering analyses. These include k-means clustering, hierarchical clustering and self-organizing map (SOM). These algorithms are used to cluster, meaning to form multiple groups, based on similarities or differences. This has been used to identify patterns in gut microbiota studies (Li et al., 2022).

4.2 Supervised learning

Supervised learning learns a function from input data that consists of independent and dependent variables across all samples. The dependent variables are used to train and develop the ML model. The created ML model can then be tasked to find patterns for the new samples, such as classification and regression (a method to identify the relationships between independent variables and a dependent variable) (Li et al., 2022). Supervised learning has some of the most classical of ML methods (Hernández Medina et al., 2022). Examples of supervised learning include support vector machine (SVM), Naïve Bayes (NB), k-nearest neighbor (KNN), gradient boosting and random forest (RF) (Li et al., 2022, Hernández Medina et al., 2022). A number of these supervised learning techniques will now be discussed.

4.2.1 Decision Tree

A decision tree is a commonly used method for predictive modeling. It uses a flowchart-like structure resembling a tree model to find a variable from input features. It makes decisions based on how to split up a dataset into groups that are similar. A positive aspect of this model is that it allows for easy interpretations of the trained model. Multiple other methods have developed from the decision tree method, such as random forest and gradient boosting (Li et al., 2022, Hernández Medina et al., 2022).

4.2.2 Random Forest

Random forest, also called bootstrap aggregation or bagging, is an aggregated collection of independently-trained decision trees that are trained on a randomly-sampled subset of the training dataset (Li et al., 2022, Hernández Medina et al., 2022). In essence, multiple decision trees are trained at the same time and the aggregation of the predicted structures is used to get the final predicted outcome (Li et al., 2022).

4.2.3 Gradient Boosting

Gradient boosting is another ensemble ML algorithm, similar to RF in that it starts with a weak learner, but differs in the sense that the weak learner is sequentially trained and improved based on the previous one until a model is made that fits the dataset best (Li et al., 2022). Two algorithms of this type have been developed called XGBoost and LightGBM, which differ in how the tree grows. XGBoost splits the tree level-wise, while LightGBM splits the tree leaf-wise. The LightGBM model is, for this reason, more accurate and faster since the leaf-wise method can cut down on more loss (Li et al., 2022).

4.2.4 K-nearest neighbor

K-nearest neighbors (k-NN) uses the principle of “Cicero pares cum paribus facillime congregantur” (meaning ‘birds of a feather flock together’). It uses known classifications of samples (i.e. from a training set) to classify an unknown sample within a group that is nearest to that point (Mucherino et al., 2009). It can be used for both classification and regression problems. What determines the “neighborhood” is a selected distance metric in a multidimensional feature space. Usually, these metrics are euclidean distances (the length of a line between two points) or correlation coefficients (Marcos-Zambrano et al., 2021). It has had

multiple applications in recent years, such as a study that developed a machine learning approach (KNN) to estimate the postmortem interval using skin microbiome samples. A KNN regressor was developed from a data set from nasal and ear samples which allowed for accurate prediction of postmortem interval to within 55 accumulated degree days, which is roughly equal to 2 days at 27.5 °C. The results were a successful proof-of-concept of the use of necrobiome data in forensics (Johnson et al., 2016). Another study by Hacilar et al. (2018) used KNN to classify fecal samples as belonging to a health or diseased (IBD) person using shotgun metagenomic data from 382 individuals (234 health and 148 IBD patients). They tested a variety of trained models and found that KNN + LogitBoost (a boosting classification algorithm) worked best (Mucherino et al., 2009, Hacilar et al., 2018).

4.2.5 Naïve Bayes Classifiers

Naïve Bayes (NB) Classifiers are a family of classifiers based on the Bayes' theorem which describes the probability of an event based on prior knowledge of conditions that are potentially related to the event. NB classifiers make use of this theorem with strong (naïve) assumptions of statistical independence between the features. One study by Werner et al. (2012) describes how it has been used for taxonomic classification in microbiome research using 16S rRNA gene sequences due to its automation, speed and accuracy. Werner et al. tested the influence of the training set on classification. They observed that it was most advantageous to use the largest, most diverse training set and even identified new phylogenetic clusters previously unclassified. Furthermore, they found that trimming the reference sequences to the primer region made the classification depth better, with higher confidence thresholds.

4.2.6 Support vector machines

Support vector machines (SVM) is a supervised ML algorithm mainly used for classification, although it can also be used for regression. Specifically, it is used for learning a decision boundary (which is a line where all (or most) samples of one class are on one side of that line) between the classes for two-group classification problems and regression tasks. It captures non-linear associations of microbiome and host information to maximize the distance (margin) between healthy and disease samples (Marcos-Zambrano et al., 2021, Wu et al., 2021). The only samples relevant for learning a decision boundary are those closest to it (the support vectors). SVM can be a convenient tool for when linear separation between classes is not possible in original feature space because it can make use of the kernel trick (which avoids the explicit mapping needed for linear learning algorithms) to estimate the decision boundary in higher-dimensional space (Marcos-Zambrano et al., 2021).

4.2.7 Deep Learning

Deep learning (DL) is a class of ML algorithms of both unsupervised and supervised techniques with various artificial neural network architectures. It is a deep neural network (DNN) that consists of nodes (neurons or units) of functions that extract information from input data and turn it into more abstract outputs that go to other nodes, thereby forming a connection of nodes in a network of multiple layers that can be organized in different architectures (Li et al., 2022, Hernández Medina et al., 2022, Ghannam et al., 2021). The simplest of these neural network architectures is the fully-connected neural network (FCNN), where all nodes of one layer are

connected to every node of the next, as can be seen in Figure 3. FCNN has been used to predict host phenotypes from metagenomic data (Hernández Medina et al., 2022). Deep learning has shown to handle multi-omics data successfully and is rather flexible in that it can adapt to new tasks. While deep learning techniques create more accurate models, it has been noted that it sacrifices interpretability behind the predictions, which can make applications difficult. Also, deep learning generates many hyperparameters that require larger datasets to learn from training (Li et al., 2022, Hernández Medina et al., 2022, Ghannam et al., 2021).

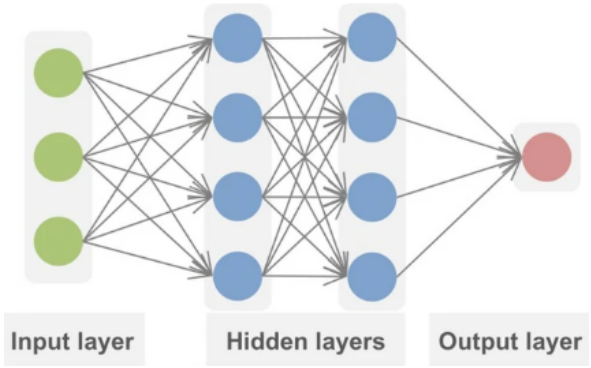

Figure 3. An example of a Deep Learning Fully-Connected Neural Network (FCNN), *where all nodes from one layer are connected to the nodes of the next. (taken from Li et al., 2022)*

4.3 Workflow of machine learning modeling

As described above, multiple supervised machine learning algorithms have been made. The various steps in the development of a correct model using these algorithms can generally be summed to four distinct steps and can be seen in Figure 4 (Li et al., 2022).

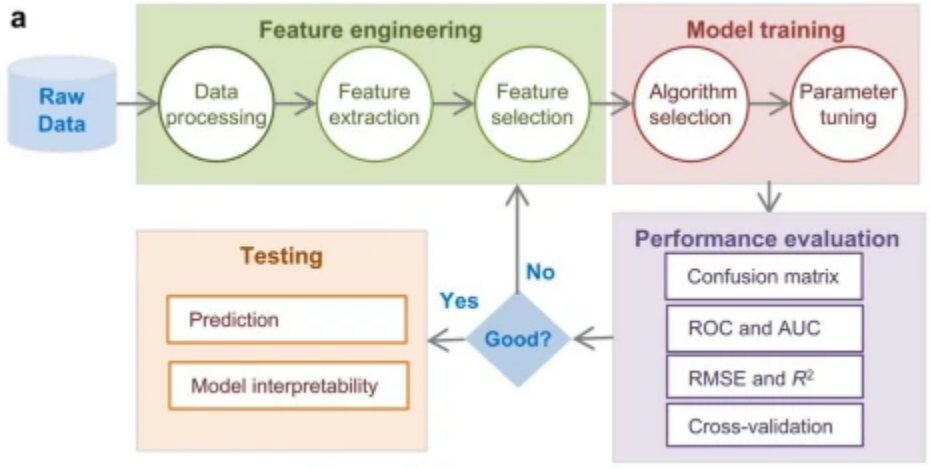

Figure 4. The general workflow of machine learning modeling. *In this figure the four steps of modeling can be seen: feature engineering, model training, performance evaluation and model application/testing (taken from Li et al., 2022).*

Feature engineering involves data pre-processing (which includes cleaning, normalization and transformation of the data), feature extraction (to build a feature vector that represents a decreased number of variables from raw data in order to select relevant information from the raw data) and feature selection processes. Selection of the relevant information to form a

feature vector can still be difficult, especially if the dataset consists of many variables, such as genes, proteins and metabolites, which require much computing power and memory. Inadequate training of the model can lead to overfitting. Overfitting is an undesirable behavior where the algorithm gives predictions for training data but not for new data, leading to inaccurate predictions when given a new dataset (Li et al., 2022).
The training process typically consists of parameter tuning and feature engineering until the model can no longer be improved. Multiple different models and approaches can be developed and benchmarked, after which the one with the best results is selected for application and evaluation (Li et al., 2022).
As for the performance evaluation, this step typically applies methods such as a confusion matrix (a table to describe the performance of a classification model on a set of test data), a receiver operating characteristics (ROC) (a plot of test sensitivity as the y coordinate versus its 1-specificity or false positive rate as the x coordinate) and assessment metrics like area under ROC curve (AUC) (a measure of the overall performance of a diagnostic test, interpreted as the average value of sensitivity for all possible values of specificity) to test the performance of classification and regression tasks. The AUC summarizes the model performance into a single measure from zero to one, which facilitates comparison of different classifiers (Li et al., 2022, Park et al., 2004). When evaluation is finished, the model can be tested with new data.

4.4 Integrative strategies for multi-omics data

Machine Learning strategies can be employed to better integrate the vast amount of multi-omics data such as metabolomics, transcriptomics, metagenomics and others. The individual omics data analyses have been very popular and extensively used in previous years. However, these only provide a partial perspective of the true complexity of a biological system. Therefore, in order to gain a more thorough and complete understanding of these systems, ML has been proposed to be used because of the significant potential to integrate heterogeneous data in, for example, gut microbiota studies (Li et al., 2022). Li et al. (2022) posits that there are three types of methods for the integration of multi-omics data in ML. The first of these integration strategies is data-driven modeling, which makes use of directly combining each omics data into a larger matrix before the training step for the machine learning model (Li et al., 2022). In other words, the data integration step happens at the early stage of modeling. Data-driven modeling has already been applied in multiple microbiota studies. For example, a study on hypertension made use of metagenomic and metabolomic data was to train random forest classifiers evaluated with ROC and AUC for statistical analyses. Through this method, they were able to reveal that gut microbiota dysbiosis contributes to the development of hypertension (Li et al., 2017). Another study on personalized nutrition created a machine-learning algorithm (a stochastic gradient boosting regression algorithm) that integrated blood parameters, dietary habits, physical activity and gut microbiota. This algorithm was capable of accurately predicting personalized postprandial glycemic response to real-life meals (Zeevi et al., 2015). Another type of integration strategies first transform the omics data in an intermediate form, such as a graph, a kernel matrix and a deep neural network. These can then be combined for the training and analysis of the model (Li et al., 2022). An example of this can be seen in a study by Hira et al. (2021) on the integration of multi-omics analysis of ovarian cancer by using variational autoencoders (VAE), which is a deep learning-based dimensionality reduction technique. The

algorithm was used for mono-omics, integration of di-omics and tri-omics data analysis of ovarian cancer through cancer samples identification, molecular subtypes clustering and classification and survival analysis. This method was found to successfully classify transcriptional subtypes with an accuracy range of 87.1% - 95,7% and proved that VAE-based methods can be used in cancer prognosis. Further conclusions Hira et al. made were that VAE outperformed existing dimensionality reduction techniques and integrated multi-omics analyses performed better or similar compared to mono-omics analyses. Finally, another type of integration first lets the machine learning algorithm train the model using each omics data, and then the predictive outcomes of the trained individual models are pulled together to create a combined model (Li et al., 2022).

5. Recent Applications and Potential Uses

As can be gathered from some of the examples of papers that have already been given, the utility of machine learning has borne its fruits in recent years. Applications range from phenotypic prediction and biomarker discovery to precision medicine for recommended therapeutics and nutrition and patient stratification and classification of disease subtypes (Li et al., 2022). Having already discussed some of the possible applications of machine learning for a wide variety of medical fields in some capacity in above examples, we shall discuss a few applications that show the variety and utility of machine learning in two specific fields: Antimicrobial Resistance and Cancer.

5.1 Antimicrobial Resistance

As described above, antimicrobial resistance (AMR) is an ever increasing problem and the microbiome can act as a reservoir for bacteria to develop and spread AMR genes. The development of new methods to improve treatment of bacterial infections and the discovery of potential AMR genes is therefore warranted. A few studies in this field will now be investigated in more detail. The first of these studies, by Madrigal et al. (2022), sought to identify such genes by looking at the surface microbiome of the International Space station across three flights in eight different locations during 12 months. Whole genomes of 226 strains, 21 shotgun metagenome sequences, and 24 metagenome-assembled genomes (MAGs) were retrieved and used to this end. The data was analyzed using a deep learning model specifically made for the identification of antibiotic resistance genes: DeepARG. They were able to identify hundreds of AMR genes from many isolates. For example, they identified AMR dominance for *Kalamiella piersonii*, which is a bacterium also found in urinary tract infections in humans, as well as strains related to *Enterobacter bugandensis* and *Bacillus cereus* (Madrigal et al., 2022). Another study by Ren et al. (2022) evaluated four machine learning methods for the prediction of AMR for the antibiotics ciprofloxacin, cefotaxime, ceftazidime and gentamicin: logistic regression, support vector machine, random forest and convolutional neural network. For training for the models they made use of whole genome sequencing data and corresponding phenotype information for antibiotics for 987 *E.coli* strains. The models were then evaluated using ROC and AUC. In this study, Ren at al. were able to demonstrate that these models effectively predicted AMR on whole-genome sequencing data, with RF and CNNs performing better than LR and SVM. They were also able to identify novel secondary mutations associated with AMR for each antibiotic

(Ren et al., 2021). In a third study, an approach called *INferring Drug Interactions using chemo-Genomics and Orthology* (INDIGO) was used that was capable of predicting antibiotic combinations that interact synergistically or antagonistically in inhibiting bacterial growth on the chemogenomic profiles of the individual antibiotics. INDIGO is an algorithm that makes use of random forests to build a model that links the interaction outcome of drug combinations to the joint chemogenomic profile of the drug pair. Chemogenomic profiling gives insights into the mechanism of action of drugs by measuring the fitness of gene-knockout strains treated with (in this case antibiotic) compounds. A large database of publicly available chemogenomic data in *E.coli* was used to identify predictive genetic features of antibiotic synergy and antagonism in order to identify new drug interactions. Identification of orthologs of genes of *E.coli* led to the prediction of drug combinations for *Mycobacterium tuberculosis* and *Staphylococcus aureus*, two types of bacteria frequently implicated in mortally dangerous infections (Chandrasekaran et al., 2016). In the context of the microbiome, it could be worthwhile to apply this method with INDIGO to identify combinations of antibiotics that should be avoided so as to limit the damage to the fragile ecosystem in the gut.

5.2. Cancer

The study by Hira et al. (2021) already shows the potential for integrative machine learning applications in research, but did not include microbiome data in its investigations. However, other studies have found links between the microbiome and cancer development/prognosis. There are three such studies, among many in recent years, that have used these integrative machine learning applications and gathered positive results. These will now be discussed in more detail. The first study, by Yang et al. (2022), used a multi-omics machine learning framework in predicting the survival of colorectal cancer (CRC) patients. CRC is the third most universal cancer globally, and so identification of biomarkers is critical for personalized therapies. This identification was achieved by looking at mRNA, miRNA and tissue microbiome levels and training models to evaluate the accuracy of potential biomarkers in predicting CRC survival. Yang et al. concluded that the microbiome of CRC tissue had the best predictive power on three-year survival of CRC patients. 26 differential microbial communities and 13 differentially expressed genes were screened out in the process, with *Thermoanaerobacterium, Parabacteroides, Oceanicaulis* and *Acetonema* being more abundantly present in short-term survival CRC patients, while *Methylotenera, Candidatus_Riesia* and *Aquamicrobium* were enriched in long-term survival CRC patients (Yang et al., 2022). Another study by Uyar et al. (2021) used multi-omics data integration through the use of advanced deep learning methods to uncover multi-omic ‘fingerprints’ associated with clinical and molecular features for multiple cancer types. Uyar et al. made use of MAUI, a stacked beta-variational auto-encoder, that is capable of reducing the high dimensional multi-omic feature datasets (Mutations, gene expression, DNA methylation, copy number variations) into low dimensional factors. MAUI was further successfully used for the modeling of clinical parameters, predicting and characterisation of molecular cancer subtypes, prognostic stratification of patients based on survival outcomes, and response or resistance to cancer treatment (Uyar et al., 2021). A third study By Zhang et al. published by Elsevier in Pharmacological Research in February 2023 employed another integrative multi-omics machine learning strategy to identify determinants of gut microbiota and tumor immunological status in CRC. Zhang et al. used the gut microbiome, gut metabolome,

host tumor transcriptome and host tumor immune profile of different CRC patient populations to analyze the data individually and integratedly to identify gut microbial markers that were capable of distinguishing CRC samples from healthy controls. A Least Absolute Shrinkage and Selector Operation (LASSO)-penalized logistic regression model (a form of supervised machine learning) was developed to select the most likely genera. Additionally, Support Vector Machine-Recursive Feature Elimination (SVM-RFE) was applied to screen important features. For integration of microbiome and metabolomics, DIABLO (Data Integration analysis for Biomarker discovery using a Latent component method for Omics), a dimension reduction method, was used to identify metagenomic and metabolomic signatures. Zhang et al. provided evidence for the direct effect of the CRC microbiota in the tumor progression, tumor immune status and immunotherapy response. For example, *Fusobacterium* and *Clostridium* were considerably increased in CRC compared to healthy controls. *Lactobacillus, Faecalibacterium,* and *Bifidobacterium* were depleted in CR patients. Additionally, the Lasso and SVM-FRE models successfully identified microbial markers for early detection of CRC. DIABLO identified strong interplay across gut microbes, metabolites, and well-defined functional genes in CRC samples, notably in immunity signaling pathways such as T cell proliferation, chemotaxis and defense response to viruses (Zhang et al., 2023).

6. Remaining Challenges and Future Perspectives

While ML has shown to be a promising tool to analyze the gut microbiome, there are still a number of hurdles left to overcome in order to perfect these techniques and make it even more competitive against other analysis techniques. The high-dimensional and heterogeneous data used in microbiome studies has large amounts of molecular features (such as genes, species, metabolites among others) but often has relatively small sample sizes. This is also referred to as the ‘curse of dimensionality’ and can make it difficult to develop accurate prediction models, which can lead to overfitting problems by ML algorithms. Possible solutions to this problem include cross-validation, reduction of the model complexity, and more robust training with more data (Li et al., 2022). Another solution is to employ an algorithm that has been used in some of the strategies mentioned above: autoencoder-based deep learning methods such as VAE. These transform high-dimensional features into low-dimensional representations, making high-dimensional datasets more easy to handle (Hira et al., 2021).That being said, deep learning contains many hyperparameters and requires much data. With lowering costs to acquiring data per sample and the already extensive data collection by the Human Microbiome Project, ML methods such as deep learning are likely to increase in importance in the near future (Li et al., 2022, iHMP RNC, 2014). Having said that, there can be multiple confounding factors (factors that are not included in the analysis model but are significantly associated with response variables) like drugs, age and diet among others still make it challenging to build robust and accurate ML models and could therefore be integrated more in ML models to account for these factors and improve disease associations by machine learning algorithms (Li et al., 2022). Another issue is the matter of reliable, correctly-labeled and high quality data. Firstly, an important consideration is that microbiome datasets may have deficiencies and biases, which can adversely affect the training process of a machine learning algorithm. It is therefore paramount that extra care must be taken in the use of data, and preferably use larger

datasets in ML training (Hernández Medina et al., 2022). Imbalanced datasets can have a negative effect on the accuracy of training classifiers. This means that classes have to be balanced, either by adding data to the smaller classes in the dataset or by discarding data from the larger set (Li et al., 2022). A possible solution to this could be the generation of new data using data augmentation. A study by Lo et al. (2019) already showed the value of this by using MetaNN, a neural network framework that uses a data augmentation technique, which outperformed existing models in classification accuracy for metagenomic data (Lo et al., 2019). Finally, and perhaps most importantly, interpretability of machine learning results can still be difficult, especially with deep learning models that use many hidden layers. ML makes links between an input and a response but does not figure out the mechanism behind the relationship. This is why machine learning models are sometimes seen as ‘black boxes’. Elucidating underlying mechanisms and factors of pathogenesis, which can stand in the way of developing treatment methods in the clinic. This is why deep learning, and perhaps ML in general, is likely to be adopted in the research field first, and only later in the clinic (Ching et al., 2018, Hernández Medina et al., 2022). An example of facilitating interpretation can be seen in deep forest algorithms, which are decision tree algorithms that can assign and rank importance of critical features. It has already shown results in microbiome-wide association studies (Li et al., 2022, Hernández Medina et al., 2022). We have now seen the current applications of machine learning techniques, as well as a few of the remaining challenges that prevent these techniques from being employed even more than they are today. In order to increase the utility of machine learning within the microbiome field, and ultimately increase the scientific output in the search for more personalized and advanced therapies related to the microbiome, these machine learning techniques will have to become more sophisticated. For this reason, we propose a collection of specific improvements that future researchers can focus on to better utilize machine learning and hopefully increase the quality of machine learning output. These are based on some of the proposed solutions given in literature to rectify the remaining problems (see Figure 5). These improvements are aimed mainly to reduce or prevent the curse of dimensionality, data quality and problems with interpretability, as described in the Remaining Challenges section.

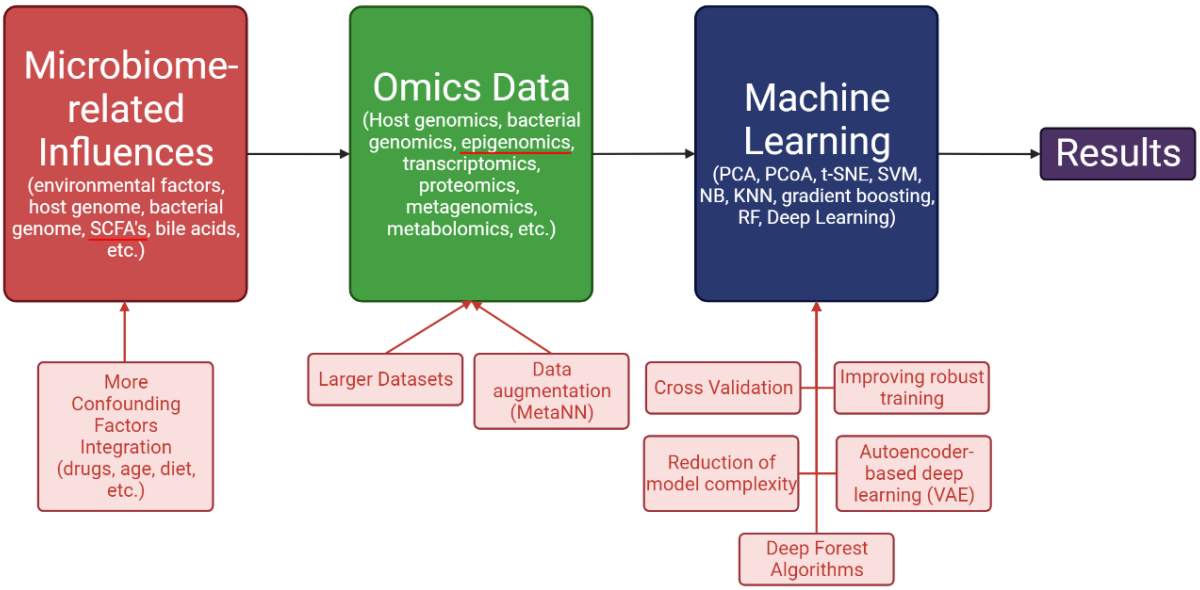

Figure 5. The steps of using machine learning in multi-omics data in microbiome research with additions to improve the machine learning applications. *The additions are in the bottom half of the figure and can have influence on various steps of machine learning utilization, such as the sample collection (of factors contributing to microbiome-related influences such as the host genome), the various omics types (and the size of the datasets) and machine learning itself (dimensional reduction, training, complexity, etc.). Created with BioRender.*

7. Conclusion and Discussion

As we have seen, the microbiome is massively influential in determining the health of people. For this reason it is no wonder that much effort has been put in acquiring as much data as possible in recent years. While this has certainly guided the field in a positive direction, new challenges have come to light. For example, the data has been put to good use in single-omics studies, but single-omics can only give us a very limited view of microbiome interactions, especially given the fact that the microbiome influences so many different systems in our body. It is therefore imperative that the various omics fields are integrated into a single model to better capture the interactions at play. However, this is easier said than done since conventional analysis methods have proven to handle multi-omics data with some difficulty. A possible solution to this problem is machine learning, which has gained increased popularity in recent years. As we have seen, there are a myriad of different techniques that employ machine learning. Their applications have proven their effectiveness in handling single-omics and multi-omics data. Machine learning results often rival or eclipse those of conventional analyses. However, a few challenges still stand in the way of more widespread use of machine learning techniques in multi-omics studies. Concepts such as the curse of dimensionality, high-quality data and interpretability are still problematic. Luckily, studies have provided solutions to these problems. For this reason, a selection of these solutions should be more publicly used and perhaps even standardized in order to improve scientific quality of results in the microbiome field. The use of autoencoders like VAE, data augmentation like MetaNN and deep forest algorithms have proven to be very effective in this regard. Taken together with the fact that A.I. is expanding in other fields like the recently popularized ChatGPT, it is not far-fetched to

conclude that we only stand at the precipice of an increased rate of machine learning uses, especially in the complicated scientific fields like microbiome research.

8. Bibliography

1) Allen, J., Sears, C.L. Impact of the gut microbiome on the genome and epigenome of colon epithelial cells: contributions to colorectal cancer development. *Genome Med* 11, 11 (2019). https://doi.org/10.1186/s13073-019-0621-2
2) Angermueller C, Parnamaa T, Parts L, Stegle O. Deep learning for computational biology. Mol Syst Biol. 2016;12:878.
3) Bäckhed F, Ding H, Wang T, Hooper LV, Koh GY, Nagy A, Semenkovich CF, Gordon JI. The gut microbiota as an environmental factor that regulates fat storage. Proc Natl Acad Sci U S A. 2004 Nov 2;101(44):15718-23. doi: 10.1073/pnas.0407076101. Epub 2004 Oct 25. PMID: 15505215; PMCID: PMC524219.
4) Bäckhed F, Ley RE, Sonnenburg JL, Peterson DA, Gordon JI. Host-bacterial mutualism in the human intestine. Science. 2005 Mar 25;307(5717):1915-20. doi: 10.1126/science.1104816. PMID: 15790844.
5) Beam AL, Kohane IS. Big Data and Machine Learning in Health Care. JAMA. 2018;319:1317–8.
6) Chandrasekaran S, Cokol-Cakmak M, Sahin N, Yilancioglu K, Kazan H, Collins JJ, Cokol M. Chemogenomics and orthology-based design of antibiotic combination therapies. Mol Syst Biol. 2016 May 24;12(5):872. doi: 10.15252/msb.20156777. PMID: 27222539; PMCID: PMC5289223.
7) Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, Ferrero E, Agapow PM, Zietz M, Hoffman MM, et al. Opportunities and obstacles for deep learning in biology and medicine. J R Soc Interface. 2018.
8) Chu H, Duan Y, Yang L, Schnabl B. Small metabolites, possible big changes: a microbiota-centered view of non-alcoholic fatty liver disease. Gut. 2019 Feb;68(2):359-370. doi: 10.1136/gutjnl-2018-316307. Epub 2018 Aug 31. PMID: 30171065.
9) Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, Sargent M, Gill SR, Nelson KE, Relman DA. Diversity of the human intestinal microbial flora. Science. 2005 Jun 10;308(5728):1635-8. doi: 10.1126/science.1110591. Epub 2005 Apr 14. PMID: 15831718; PMCID: PMC1395357.
10) Ghannam, R. B., & Techtmann, S. M. (2021). Machine learning applications in microbial ecology, human microbiome studies, and environmental monitoring. *Computational and Structural Biotechnology Journal*, *19*, 1092–1107. https://doi.org/10.1016/J.CSBJ.2021.01.028
11) Ghosh TS, Rampelli S, Jeffery IB, Santoro A, Neto M, Capri M, Giampieri E, Jennings A, Candela M, Turroni S, Zoetendal EG, Hermes GDA, Elodie C, Meunier N, Brugere CM, Pujos-Guillot E, Berendsen AM, De Groot LCPGM, Feskins EJM, Kaluza J, Pietruszka B, Bielak MJ, Comte B, Maijo-Ferre M, Nicoletti C, De Vos WM, Fairweather-Tait S, Cassidy A, Brigidi P, Franceschi C, O'Toole PW. Mediterranean diet intervention alters the gut microbiome in older people reducing frailty and improving health status: the

NU-AGE 1-year dietary intervention across five European countries. Gut. 2020 Jul;69(7):1218-1228. doi: 10.1136/gutjnl-2019-319654. Epub 2020 Feb 17. PMID: 32066625; PMCID: PMC7306987.
12) Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, Gordon JI, Relman DA, Fraser-Liggett CM, Nelson KE. Metagenomic analysis of the human distal gut microbiome. Science. 2006;312:1355–9
13) Gupta S, Allen-Vercoe E, Petrof EO. Fecal microbiota transplantation: in perspective. Therap Adv Gastroenterol. 2016 Mar;9(2):229-39. doi: 10.1177/1756283X15607414. PMID: 26929784; PMCID: PMC4749851.
14) Hacılar H., O. U. Nalbantoğlu and B. Bakir-Güngör, "Machine Learning Analysis of Inflammatory Bowel Disease-Associated Metagenomics Dataset," 2018 3rd International Conference on Computer Science and Engineering (UBMK), Sarajevo, Bosnia and Herzegovina, 2018, pp. 434-438, doi: 10.1109/UBMK.2018.8566487.
15) Hagan T, Cortese M, Rouphael N, Boudreau C, Linde C, Maddur MS, Das J, Wang H, Guthmiller J, Zheng NY, Huang M, Uphadhyay AA, Gardinassi L, Petitdemange C, McCullough MP, Johnson SJ, Gill K, Cervasi B, Zou J, Bretin A, Hahn M, Gewirtz AT, Bosinger SE, Wilson PC, Li S, Alter G, Khurana S, Golding H, Pulendran B. Antibiotics-Driven Gut Microbiome Perturbation Alters Immunity to Vaccines in Humans. Cell. 2019 Sep 5;178(6):1313-1328.e13. doi: 10.1016/j.cell.2019.08.010. PMID: 31491384; PMCID: PMC6750738.
16) Hernández Medina, R., Kutuzova, S., Nielsen, K.N. *et al.* Machine learning and deep learning applications in microbiome research. *ISME COMMUN.* 2, 98 (2022). https://doi.org/10.1038/s43705-022-00182-9
17) Hira MT, Razzaque MA, Angione C, Scrivens J, Sawan S, Sarker M. Integrated multi-omics analysis of ovarian cancer using variational autoencoders. Sci Rep. 2021 Mar 18;11(1):6265. doi: 10.1038/s41598-021-85285-4. Erratum in: Sci Rep. 2021 Aug 11;11(1):16671. PMID: 33737557; PMCID: PMC7973750.
18) Huemer M, Mairpady Shambat S, Brugger SD, Zinkernagel AS. Antibiotic resistance and persistence-Implications for human health and treatment perspectives. EMBO Rep. 2020 Dec 3;21(12):e51034. doi: 10.15252/embr.202051034. Epub 2020 Dec 8. PMID: 33400359; PMCID: PMC7726816.
19) Integrative HMPRNC. The Integrative Human Microbiome Project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. Cell Host Microbe. 2014;16:276–89
20) Johnson HR, Trinidad DD, Guzman S, Khan Z, Parziale JV, DeBruyn JM, Lents NH. A Machine Learning Approach for Using the Postmortem Skin Microbiome to Estimate the Postmortem Interval. PLoS One. 2016 Dec 22;11(12):e0167370. doi: 10.1371/journal.pone.0167370. PMID: 28005908; PMCID: PMC5179130.
21) Kaur H, Singh Y, Singh S, Singh RB. Gut microbiome-mediated epigenetic regulation of brain disorder and application of machine learning for multi-omics data analysis. Genome. 2021 Apr;64(4):355-371. doi: 10.1139/gen-2020-0136. Epub 2020 Oct 8. PMID: 33031715.

22) Kazemian N, Mahmoudi M, Halperin F, Wu JC, Pakpour S. Gut microbiota and cardiovascular disease: opportunities and challenges. Microbiome. 2020 Mar 14;8(1):36. doi: 10.1186/s40168-020-00821-0. PMID: 32169105; PMCID: PMC7071638.
23) Kurilshikov A, Medina-Gomez C, Bacigalupe R, Radjabzadeh D, Wang J, Demirkan A, Le Roy CI, Raygoza Garay JA, Finnicum CT, Liu X, Zhernakova DV, Bonder MJ, Hansen TH, Frost F, Rühlemann MC, Turpin W, Moon JY, Kim HN, Lüll K, Barkan E, Shah SA, Fornage M, Szopinska-Tokov J, Wallen ZD, Borisevich D, Agreus L, Andreasson A, Bang C, Bedrani L, Bell JT, Bisgaard H, Boehnke M, Boomsma DI, Burk RD, Claringbould A, Croitoru K, Davies GE, van Duijn CM, Duijts L, Falony G, Fu J, van der Graaf A, Hansen T, Homuth G, Hughes DA, Ijzerman RG, Jackson MA, Jaddoe VWV, Joossens M, Jørgensen T, Keszthelyi D, Knight R, Laakso M, Laudes M, Launer LJ, Lieb W, Lusis AJ, Masclee AAM, Moll HA, Mujagic Z, Qibin Q, Rothschild D, Shin H, Sørensen SJ, Steves CJ, Thorsen J, Timpson NJ, Tito RY, Vieira-Silva S, Völker U, Völzke H, Võsa U, Wade KH, Walter S, Watanabe K, Weiss S, Weiss FU, Weissbrod O, Westra HJ, Willemsen G, Payami H, Jonkers DMAE, Arias Vasquez A, de Geus EJC, Meyer KA, Stokholm J, Segal E, Org E, Wijmenga C, Kim HL, Kaplan RC, Spector TD, Uitterlinden AG, Rivadeneira F, Franke A, Lerch MM, Franke L, Sanna S, D'Amato M, Pedersen O, Paterson AD, Kraaij R, Raes J, Zhernakova A. Large-scale association analyses identify host factors influencing human gut microbiome composition. Nat Genet. 2021 Feb;53(2):156-165. doi: 10.1038/s41588-020-00763-1. Epub 2021 Jan 18. PMID: 33462485; PMCID: PMC8515199.
24) Ley RE, Turnbaugh PJ, Klein S, Gordon JI. Microbial ecology: human gut microbes associated with obesity. Nature. 2006 Dec 21;444(7122):1022-3. doi: 10.1038/4441022a. PMID: 17183309.
25) Li, J., Zhao, F., Wang, Y. *et al.* Gut microbiota dysbiosis contributes to the development of hypertension. *Microbiome* 5, 14 (2017). https://doi.org/10.1186/s40168-016-0222-x
26) Li, P., Luo, H., Ji, B. *et al.* Machine learning for data integration in human gut microbiome. *Microb Cell Fact* 21, 241 (2022). https://doi.org/10.1186/s12934-022-01973-4
27) Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, Poon TW, Andrews E, Ajami NJ, Bonham KS, Brislawn CJ, et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature. 2019;569:655–62
28) Lo C, Marculescu R. MetaNN: accurate classification of host phenotypes from metagenomic data using neural networks. BMC Bioinformatics. 2019 Jun 20;20(Suppl 12):314. doi: 10.1186/s12859-019-2833-2. PMID: 31216991; PMCID: PMC6584521.
29) Madrigal, P., Singh, N.K., Wood, J.M. *et al.* Machine learning algorithm to characterize antimicrobial resistance associated with the International Space Station surface microbiome. *Microbiome* 10, 134 (2022). https://doi.org/10.1186/s40168-022-01332-w
30) Marcos-Zambrano, L. J., Karaduzovic-Hadziabdic, K., Loncar Turukalo, T., Przymus, P., Trajkovik, V., Aasmets, O., Berland, M., Gruca, A., Hasic, J., Hron, K., Klammsteiner, T., Kolev, M., Lahti, L., Lopes, M. B., Moreno, V., Naskinova, I., Org, E., Paciência, I., Papoutsoglou, G., … Truu, J. (2021). Applications of Machine Learning in Human Microbiome Studies: A Review on Feature Selection, Biomarker Identification, Disease

Prediction and Treatment. *Frontiers in Microbiology*, *12*, 313. https://doi.org/10.3389/FMICB.2021.634511/BIBTEX
31) Mazmanian SK, Liu CH, Tzianabos AO, Kasper DL. An immunomodulatory molecule of symbiotic bacteria directs maturation of the host immune system. Cell. 2005 Jul 15;122(1):107-18. doi: 10.1016/j.cell.2005.05.007. PMID: 16009137.
32) Mucherino, A., Papajorgji, P.J., Pardalos, P.M. (2009). *k*-Nearest Neighbor Classification. In: Data Mining in Agriculture. Springer Optimization and Its Applications, vol 34. Springer, New York, NY. https://doi.org/10.1007/978-0-387-88615-2_4
33) Park SH, Goo JM, Jo CH. Receiver operating characteristic (ROC) curve: practical review for radiologists. Korean J Radiol. 2004 Jan-Mar;5(1):11-8. doi: 10.3348/kjr.2004.5.1.11. PMID: 15064554; PMCID: PMC2698108.
34) Paul D, Das B. Gut microbiome in the emergence of antibiotic-resistant bacterial pathogens. Prog Mol Biol Transl Sci. 2022;192(1):1-31. doi: 10.1016/bs.pmbts.2022.07.009. Epub 2022 Sep 3. PMID: 36280316.
35) Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, Liang S, Zhang W, Guan Y, Shen D, Peng Y, Zhang D, Jie Z, Wu W, Qin Y, Xue W, Li J, Han L, Lu D, Wu P, Dai Y, Sun X, Li Z, Tang A, Zhong S, Li X, Chen W, Xu R, Wang M, Feng Q, Gong M, Yu J, Zhang Y, Zhang M, Hansen T, Sanchez G, Raes J, Falony G, Okuda S, Almeida M, LeChatelier E, Renault P, Pons N, Batto JM, Zhang Z, Chen H, Yang R, Zheng W, Li S, Yang H, Wang J, Ehrlich SD, Nielsen R, Pedersen O, Kristiansen K, Wang J. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature. 2012 Oct 4;490(7418):55-60. doi: 10.1038/nature11450. Epub 2012 Sep 26. PMID: 23023125.
36) Quiroga R, Nistal E, Estébanez B, Porras D, Juárez-Fernández M, Martínez-Flórez S, García-Mediavilla MV, de Paz JA, González-Gallego J, Sánchez-Campos S, Cuevas MJ. Exercise training modulates the gut microbiota profile and impairs inflammatory signaling pathways in obese children. Exp Mol Med. 2020 Jul;52(7):1048-1061. doi: 10.1038/s12276-020-0459-0. Epub 2020 Jul 6. PMID: 32624568; PMCID: PMC8080668.
37) Rakoff-Nahoum S, Paglino J, Eslami-Varzaneh F, Edberg S, Medzhitov R. Recognition of commensal microflora by toll-like receptors is required for intestinal homeostasis. Cell. 2004 Jul 23;118(2):229-41. doi: 10.1016/j.cell.2004.07.002. PMID: 15260992.
38) Ren Y, Chakraborty T, Doijad S, Falgenhauer L, Falgenhauer J, Goesmann A, Hauschild AC, Schwengers O, Heider D. Prediction of antimicrobial resistance based on whole-genome sequencing and machine learning. Bioinformatics. 2021 Oct 6;38(2):325–34. doi: 10.1093/bioinformatics/btab681. Epub ahead of print. PMID: 34613360; PMCID: PMC8722762.
39) Sekirov, I., Russell, S. L., Caetano M Antunes, L., & Finlay, B. B. (2010). Gut microbiota in health and disease. *Physiological Reviews*, *90*(3), 859–904. https://doi.org/10.1152/PHYSREV.00045.2009/ASSET/IMAGES/LARGE/Z9J003102543 0009.JPEG
40) Sender R, Fuchs S, Milo R. Revised Estimates for the Number of Human and Bacteria Cells in the Body. PLoS Biol. 2016 Aug 19;14(8):e1002533. doi: 10.1371/journal.pbio.1002533. PMID: 27541692; PMCID: PMC4991899.
41) Sharon G, Cruz NJ, Kang DW, Gandal MJ, Wang B, Kim YM, Zink EM, Casey CP, Taylor BC, Lane CJ, Bramer LM, Isern NG, Hoyt DW, Noecker C, Sweredoski MJ, Moradian A,

Borenstein E, Jansson JK, Knight R, Metz TO, Lois C, Geschwind DH, Krajmalnik-Brown R, Mazmanian SK. Human Gut Microbiota from Autism Spectrum Disorder Promote Behavioral Symptoms in Mice. Cell. 2019 May 30;177(6):1600-1618.e17. doi: 10.1016/j.cell.2019.05.004. PMID: 31150625; PMCID: PMC6993574.

42) Uyar, B., Ronen, J., Franke, V., Gargiulo, G., & Akalin, A. (2021). Multi-omics and deep learning provide a multifaceted view of cancer. *BioRxiv*, 2021.09.29.462364. https://doi.org/10.1101/2021.09.29.462364

43) Werner JJ, Koren O, Hugenholtz P, DeSantis TZ, Walters WA, Caporaso JG, Angenent LT, Knight R, Ley RE. Impact of training sets on classification of high-throughput bacterial 16s rRNA gene surveys. ISME J. 2012 Jan;6(1):94-103. doi: 10.1038/ismej.2011.82. Epub 2011 Jun 30. PMID: 21716311; PMCID: PMC3217155.

44) Wilmanski T, Diener C, Rappaport N, Patwardhan S, Wiedrick J, Lapidus J, Earls JC, Zimmer A, Glusman G, Robinson M, Yurkovich JT, Kado DM, Cauley JA, Zmuda J, Lane NE, Magis AT, Lovejoy JC, Hood L, Gibbons SM, Orwoll ES, Price ND. Gut microbiome pattern reflects healthy ageing and predicts survival in humans. Nat Metab. 2021 Feb;3(2):274-286. doi: 10.1038/s42255-021-00348-0. Epub 2021 Feb 18. Erratum in: Nat Metab. 2021 Apr;3(4):586. PMID: 33619379; PMCID: PMC8169080.

45) Wu H, Esteve E, Tremaroli V, Khan MT, Caesar R, Mannerås-Holm L, Ståhlman M, Olsson LM, Serino M, Planas-Fèlix M, Xifra G, Mercader JM, Torrents D, Burcelin R, Ricart W, Perkins R, Fernàndez-Real JM, Bäckhed F. Metformin alters the gut microbiome of individuals with treatment-naive type 2 diabetes, contributing to the therapeutic effects of the drug. Nat Med. 2017 Jul;23(7):850-858. doi: 10.1038/nm.4345. Epub 2017 May 22. PMID: 28530702.

46) Wu, S., Chen, Y., Li, Z., Li, J., Zhao, F., & Su, X. (2021). Towards multi-label classification: Next step of machine learning for microbiome research. *Computational and Structural Biotechnology Journal*, *19*, 2742–2749. https://doi.org/10.1016/J.CSBJ.2021.04.054

47) Xu, H. M., Huang, H. L., Zhou, Y. L., Zhao, H. L., Xu, J., Shou, D. W., Liu, Y. di, Zhou, Y. J., & Nie, Y. Q. (2021). Fecal Microbiota Transplantation: A New Therapeutic Attempt from the Gut to the Brain. *Gastroenterology Research and Practice*, *2021*. https://doi.org/10.1155/2021/6699268

48) Yang M, Yang H, Ji L, Hu X, Tian G, Wang B, Yang J. A multi-omics machine learning framework in predicting the survival of colorectal cancer patients. Comput Biol Med. 2022 Jul;146:105516. doi: 10.1016/j.compbiomed.2022.105516. Epub 2022 Apr 18. PMID: 35468406.

49) Zeevi D, Korem T, Zmora N, Israeli D, Rothschild D, Weinberger A, Ben-Yacov O, Lador D, Avnit-Sagi T, Lotan-Pompan M, et al. Personalized Nutrition by Prediction of glycemic responses. Cell. 2015;163:1079–94.

50) Zhang, S. L., Cheng, L. S., Zhang, Z. Y., Sun, H. T., & Li, J. J. (2023). Untangling determinants of gut microbiota and tumor immunologic status through a multi-omics approach in colorectal cancer. *Pharmacological Research*, *188*, 106633. https://doi.org/10.1016/J.PHRS.2022.106633

51) Zhou WY, Sailani MR, Contrepois K, Zhou YJ, Ahadi S, Leopold SR, Zhang MJ, Rao V, Avina M, Mishra T, et al. Longitudinal multi-omics of host-microbe dynamics in prediabetes. Nature. 2019;569:663.
52) Zmora N, Suez J, Elinav E. You are what you eat: diet, health and the gut microbiota. Nat Rev Gastroenterol Hepatol. 2019 Jan;16(1):35-56. doi: 10.1038/s41575-018-0061-2. PMID: 30262901.