Universidad Politécnica
de Madrid

**Escuela Técnica Superior de
Ingenieros Informáticos**

Máster Universitario en Inteligencia Artificial

Trabajo Fin de Máster

# EXPLAINABLE MACHINE LEARNING FOR LONGITUDINAL MULTI-OMIC MICROBIOME

Autora: Paula Laccourreye Matamala

Tutores: Concha Bielza Lozoya y Pedro Larrañaga Múgica

Madrid, Julio 2021

*Trabajo Fin de Máster*
*Máster Universitario en Inteligencia Artificial*
*Título:* EXPLAINABLE MACHINE LEARNING FOR LONGITUDINAL MULTI-OMIC MICROBIOME ANALYSIS
Julio 2021

*Autora:* Paula Laccourreye Matamala

*Tutores:* Concha Bielza Lozoya y Pedro Larrañaga Múgica
Inteligencia Artificial
E.T.S. de Ingenieros Informáticos
Universidad Politécnica de Madrid

# Acknowledgements

I would like to express my sincere thanks to my professors and supervisors, Concha Bielza and Pedro Larrañaga, for giving me the opportunity to discover the fascinating world of human microbiome research from a machine learning approach and for their continuous patience, knowledge, and guidance through all the project.

# Resumen

La enfermedad inflamatoria intestinal (EII) engloba la enfermedad de Crohn y la colitis ulcerosa, dos patologías crónicas relacionadas con la inflamación del tracto gastrointestinal. Si bien se desconoce la causa exacta de la EII, existe una conexión clave entre la comunidad microbiana en el intestino, los genes y el sistema inmunológico. Lo mismo ocurre con otras enfermedades relacionadas con estos factores, como asma, diabetes, obesidad o cáncer colorrectal. Por esta razón, la investigación del microbioma humano se está convirtiendo en un foco de interés cada vez mayor en la salud humana y ha atraído la atención de grandes consorcios de investigación como el proyecto NIH Human Microbiome en Estados Unidos o MetaHIT en Europa.

Los conjuntos de datos de microbiomas son únicos en su caracterizada alta dimensionalidad y dispersión. Además, las comunidades microbianas son dinámicas y su composición cambia con el tiempo. Los avances recientes en la secuenciación de alto rendimiento han llevado a una explosión de datos multi-ómicos de diferentes fuentes que pueden proporcionar importantes conocimientos biológicos. Por lo tanto, los desafíos actuales en el estudio de los datos del microbioma humano implican el manejo de datos multidominio (heterogéneos), con alta dimensionalidad, así como el análisis de series temporales (datos irregularmente espaciados y escasos).

Sin embargo, las soluciones actuales para comprender mejor las relaciones entre el microbioma humano y las enfermedades no se han abordado en profundidad. En consecuencia, se deben explorar y desarrollar nuevos métodos explicables, dinámicos y multiómicos para el análisis del microbioma. Las redes bayesianas posibilitan un enfoque interesante ya que nos ayudan a comprender las formas básicas en que las diferentes entidades biológicas (taxones, genes, metabolitos) interactúan entre sí en un entorno determinado (por ejemplo el intestino humano).

En este trabajo se han desarrollado un conjunto de pasos de preprocesamiento para limpiar, filtrar, seleccionar e integrar con éxito datos longitudinales de metagenómica, metatranscriptómica y metabolómica procedentes del proyecto del microbioma humano. Proponemos un enfoque de red bayesiana dinámica que puede ayudar a construir modelos dinámicos que capturen el comportamiento del microbioma humano para comprender cómo la comunidad microbiana se comunica con el huésped y contribuye a la enfermedad. En concreto para este proyecto, estudiaremos enfermedades inflamatorias del intestino (EII): la enfermedad de Crohn y la colitis ulcerosa. Nuestra solución propuesta será muy valiosa para la actual y emergente medicina de precisión.

# Abstract

Inflammatory bowel disease (IBD) encompasses Crohn's disease and ulcerative colitis, two chronic disorders involving inflammation of the gastrointestinal tract. While the exact cause of IBD remains unknown, there is a key connection between the microbial community in the gut, genes, and the immune system. The same occurs with other disease conditions related to these factors such as asthma, diabetes, obesity, or colorectal cancer. For this reason, microbiome research is becoming an increasing major focus of interest in human health and has attracted the attention of large research consortiums such as the NIH Human Microbiome project in United States or MetaHIT in Europe.

Microbiome data sets are unique in their characterized high dimensionality and sparsity. Moreover, microbial communities are dynamic, and their compositions change with time. Recent advances in high-throughput sequencing have led to explosion of multi-omic data from different sources that are able to provide important biological insights. Thus, current challenges of studying microbiome data involve multidomain data (heterogeneity), high dimensionality and time series analysis (sparse and irregularly spaced data).

However, current solutions to better understand relationships between the human microbiome and disease have not been dealt with in depth. Consequently, new explainable, dynamic, and multi-omic methods to microbiome analysis must be explored and developed. Bayesian networks are an interesting approach as they help us understanding of the basic ways the different biological entities (taxa, genes, metabolites) interact with each other in a given environment (human gut).

We develop a set of preprocessing steps to successfully clean, filter, select and integrate informative metagenomics, metatranscriptomics and metabolomics longitudinal data from the Human Microbiome Project. We propose a dynamic Bayesian network approach that can assist in building dynamic models to capture the behavior of the human microbiome to understand how the microbial community communicates with the host and contributes toward disease. Specifically for this project, we will study inflammatory bowel diseases (IBD): Crohn's disease and ulcerative colitis. Our proposed solution will be very valuable for current and emerging predictive precision medicine.

# Table of Contents

# 1 Introduction

## 1.1 Introduction

Over the past decade the microbiome has been receiving increasing attention, especially with international initiatives like the Human Microbiome Project launched by the National Institute of Health in the United Sates or MetaHIT funded by the European Commission with over 14 institutions and more than 100 scientists collaborating. With the rise of high-throughput technologies and omic sciences, there has been increasing evidence that the human microbiome plays an important role in many disease status such as obesity, diabetes, *C.difficile* infection or colorectal cancer, among many others, generating significant attention in clinical applications for current and emerging diseases.

In this context, there is a lot of research activity in developing therapeutics and diagnostics. Even though important efforts have been put into the field, the functions, dynamics, and causation of dysbiosis state performed by the microbial community remains unclear. Machine learning models can help elucidate important connections, functions, and relationships between microbial community in the human host that can advance the discovery of novel therapeutic approaches.

## 1.2 Motivation

The community has raised some concerns with current studies of human microbiome research. Most studies only focus on describing static taxonomic composition of the human microbiome, overlooking temporal variability thus causing a major drawback in real world clinical applications as many diseases are characterized by periods of remission and exacerbation in symptoms. The present work aims to focus on longitudinal microbiome data in order to yield insights into the dynamic behavior of microbiota.

The project pursues to develop a machine learning methodology for microbiome research that is explicable and transparent in dynamic scenarios, i.e., with longitudinal data. Moreover, the combination of multiple genomic data types has been proven to result in increased performance (Zhu, et al., 2008). Thus, with this study we would like to establish identification of predictive and discriminatory omics features. This work will serve as an initial step towards the common goal of current scientists in the field to potentially uncover novel predictive and prognostic markers.

Bayesian networks are a suitable tool to model the interaction of many microbial communities in the human gut, as they are prepared for inferring complex networks from noisy data to predict clinical outcomes of relevance in a biologically interpretable manner (McGeachie, et al., 2016).

Therefore, the main objective of this work is to explore the state of the art of Bayesian networks applied to human microbiome data analysis identifying their current pitfalls and limitations as well as outlining solutions to be addressed in future research.

Our study was further motivated by the accessibility to one of largest studies of the human microbiome completed to date, a public database collected in an international collaborative research program, the Human Microbiome project. Although for our study we focus on one specific condition, our approach could potentially allow us to develop future models based on Bayesian networks for a wide spectrum of diseases.

Finally, we identified previous work failed to tackle a core issue in microbiome research which is usability and reproducibility. Currently, solutions are either solely focused on addressing the problem from a microbiology point of view, making it difficult to understand for machine learning experts, or centered on the technical aspect of machine learning models, leaving out the microbiology community. There is still a lack of a homogenous unification of both domains in studies, where high expertise is achieved and reflected in the design of machine learning pipelines applied to real world human microbiome data. For this reason, we will aim to develop a computational framework easy to use and interpret, that merge both the biotechnological aspect as well as the ML approach into building informative models of the dynamic human gut microbiome.

## 1.3 Objectives

The main goal is to develop new methodologies for analysis of human microbiome data through explainable models generated by machine learning techniques based on the concept of probabilistic graphical models.

This work will focus on the analysis of longitudinal microbiome data in order to identify patterns of variation, and link these to patterns of host status, such as, the presence or absence of a particular disease. This approach will hopefully allow us to capture the influence of individual microbial classes and functions on each other over time.

Moreover, this study will address both taxonomic composition and functional profile (both important and complementary) as most studies to date only investigate taxonomic composition. To do this, we will be integrating multi-omics with microbiome data and clinical information. Multi-omics data analysis can help us identify potential metabolic biomarkers from metabolite characterization thus contributing to the development of precision medicine, a field which is currently receiving much attention and promises to revolutionize healthcare and medical treatments within the next decade (McGrath & Ghersi, 2016). By integrating multi-omics in microbiome research, we aim to answer questions such as “What are microbes doing?” in order to find functions provided by the microbial community that are critical for human health, or “How do they interact?” characterizing host-microbiome interactions that can help us predict causal role of human microbes.

In particular we would like to perform inference of temporal interaction between the following three biological entities, as shown in Figure 1.

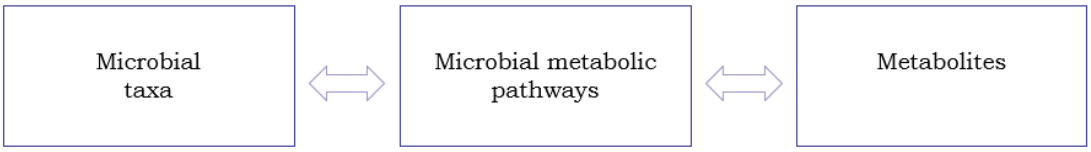


*Figure 1. Interactions among biological entities of the microbial communities in the human gut.*

Microbiome metagenomic, metatranscriptomics and metabolomic analysis permit gene-level and functional associations with disease.

Our study aims to extend current knowledge of associations between the human microbiome and health and disease through the application of data-driven machine learning models. Even though machine learning models and Bayesian networks have already been applied to the field, our approach is focused around model interpretability. With this in mind, dynamic Bayesian networks are presented as the technique that will help scientists obtain transparent and interpretable intelligent systems in benefit of human health. Our work aims to shed new light on the applications of dynamic Bayesian networks to describe temporal variation of the gut microbiota and dynamic relationships between taxonomic entities and clinical attributes.

An ambitious ultimate goal we considered was answering the question: “Are we able to move from correlation to causation?”. One of the objectives is to develop a methodology that can infer cause and not limited only to association. There have been many studies over the past years identifying and linking the composition of microbiota to health outcomes. However, an intriguing area in the field of microbiome research is being able to discriminate microbiome features that are causal for disease from those that are consequence of disease, being able to unveil if microbiota is actually causing or driving specific disease outcomes (Wang & Jia, 2016). Fecal transplants in germ-free mouse models have been an important starting point for research in this direction. Once we identify what or who is causing the disruption of the ecosystem, we will be able to shift research into exploring how to modify our microbiome to address disease status and establish microbial basis for secure and effective personalized treatments.

For all the above reasons, a key goal for us with this work is to build a general-purpose framework/protocol to study microbiome characteristics using machine learning that would be easy to use for either microbiology experts or computer scientists.

## 1.4 Structure of this document

The master thesis is organized as follows.

Chapter 2 starts with a review of the scientific literature in which first, the biological background required to understand the clinical problem, is introduced. Then we review general applications of machine learning methods to the problem summarizing their advantages and limitations. Finally, dynamic Bayesian networks are described. In Chapter 3 the proposed framework and analysis are explained. Chapter 4 presents our results. Finally, Chapter 5 describes the conclusions drawn from the project and discusses lines for future research.

# 2 Literature Review

The purpose of this chapter is to primarily introduce the clinical and biological background to the problem of human microbiome research (Section 2.1). Applications of machine learning for analysis of microbiome data will be reviewed (Section 2.2) with focus on two models: (1) Bayesian networks (BNs), presented as the proposed model, its theoretical background and previous work done with human microbiome data and (2) one of the most used models; random forests. Finally, we will end (Section 2.3) with current and upcoming limitations and challenges in the field.

## 2.1 Biological and clinical background

### 2.1.1 Human microbiome

The human body consists of about 100 trillion microbial cells vs 10 trillion human cells (Ursell, Metcalf, Wegener, & Knight, 2012) (Bull & Plummer, 2014). In genome terms, humans have around 20,000 human genes and around 2-20 million microbial genes (Gilbert, et al., 2018). Microbiome research is the discipline studying behavior and functions of the microbiota. Microbiota refers to the community of microorganisms such as bacteria, viruses, fungi and archaea residing within an environment while microbiome is the term used to describe the collection of all the genes which are contained in the human microbial community.

The field of human microbiome is a relatively new field of research which focuses on studying the microbial genes in the human or in other words the collection of all the genes which are contained in the human microbial community. Advances in the field have been driven mainly by: (i) cost reduction of high-throughput sequencing techniques which allowed for parallel analysis of DNA/RNA molecules; (ii) appearance of novel bioinformatics tools and computational pipelines that enable the analysis of microbiome sequencing data (Caporaso, et al., 2010) and (iii) availability of larger datasets (Turnbaugh, et al., 2007) (McDonald, et al., 2018).

Although microbiome research is currently being studied for many applications such as ecology, agriculture, biotechnology or plant health (Mueller & Sachs, 2015), (Moran, 2015) (Louca, Parfrey, & Doebeli, 2016) (Hou & Kolodkin-Gal, 2020), (Trivedi, Leach, Tringe, Sa, & Singh, 2020) there is a particular growing interest in Medicine in order to understand how the community of bacteria in the human body shape our health. Not only understanding “who is there”, but also “what are they doing”, “how are they doing it” and their interaction with the human host. This is, among others, due to the increasing published studies proving how the dysbiosis of microbes in different parts of the human body (oral, skin, gut, vaginal) are related to numerous health conditions and their risk and severity (Duvallet, Gibbons, Gurry, Irizarry, & Alm, 2017). Next, we detail the most relevant.

Colorectal carcinoma: (Zeller, et al., 2014) works identifying taxonomic makers for colorectal cancer (CRC) patient screening using a LASSO logistic regression classifier. (Wirbel, et al., 2019) carried out a meta-analysis of 768 subjects to define CRC functional taxonomic signatures for future early diagnosis. (Thomas, et al., 2019) and (Su, et al., 2020) developed a search-based strategy for disease detection and classification based on phylogeny-based composition, (Ai, et al., 2017) assessed different models for fecal microbiota-based CRC prediction whereas (Kharrat, et al., 2019) main goal was to identify microbiota that can have a causal role in developing CRC.

Autoimmune diseases: (Vatanen, et al., 2016) followed 222 infants from different regions with the aim of characterizing gut microbiome development in early life and identifying critical microbes that can contribute to immune modulation altering its normal course. Also, (Cornejo-Pareja, et al., 2020) results demonstrated altered gut microbiota in patients with common autoimmune thyroid diseases (Graves-Basedow's or Hashimoto's) and pointed out the relationship between them.

Metabolic disorders such as obesity or diabetes constitute a popular area of application of human microbiome research. Motivated by the goal of accelerating the development of new therapeutic strategies for the prevention of type II diabetes (Koivula, et al., 2014), or microbiota taxonomic and functional diversity association with type I diabetes (Leiva-Gea, et al., 2018), there has been intense research with promising outcomes (Qin J. , et al., 2012); (Zhou, et al., 2019); (Sikalidis & Maykish, 2020); (Doumatey, et al., 2020); (Kostic, et al., 2015).

Asthma/allergy: numerous studies have explored the association between gut microbiota and the risk of developing childhood asthma or allergies (Depner, et al., 2020). Findings from longitudinal studies in (Joseph, et al., 2016) and (Metwally, et al., 2019) show potential of using microbiome profiles for allergy prediction.

Thus, there are many reasons that are contributing to generate considerable interest in terms of explaining the role of the human microbiome in health.

Microbiota is essential for the correct functioning of our organism in many ways. Further, it enlarges the genetic and functional capability of its host. There has been intense study of the role and function of the human microbiome in maintaining a healthy state. Some of the tasks that the microbial community performs, that contribute to healthy state are: (i) protecting against infections and harmful pathogenic organisms by shaping the immune response and being partly responsible of maintaining and developing a healthy immune system (Hooper & Gordon, 2001), (Bengmark, 2013) and (Weng & Walker, 2013), (ii) providing essential nutritional compounds (chemical transformations), (iii) participating in the metabolism of xenobiotics (Clayton, Baker, Lindon, Everett, & Nicholson, 2009), (Carmody & Turnbaugh, 2014) and drugs (Li, He, & Jia, 2016) (Weersma, Zhernakova, & Fu, 2020), (iv) degradation of indigestible components of the host diet like polysaccharides (Turnbaugh, et al., 2009) and general aid in the digestion of food.

The importance of microbiome research has been presented, but to reveal valuable insights for clinical applications, realistic and accurate analysis of the microbiota must be done. Therefore, in this work we will focus on investigating the dynamics of human microbiome which is in fact its real nature. The microbiome of a human being matures and establishes in early life and keeps changing across lifetime. That is why many studies have focused on characterizing microbial communities following infant subjects from pre-term throughout their first months of life (La Rosa, et al., 2014), (Lugo-Martinez, Ruiz-Perez, Narasimhan, & Bar-Joseph, 2019), (Kostic, et al., 2015) and (Rouhani, et al., 2020). Some interesting areas of research related to the human microbiome dynamics are the temporal variability in healthy adults, the response (dynamic) to external and internal perturbations such as diet or environment and the associations of microbiome changes with host disease (Gerber, 2014). MDSINE (Bucci, et al., 2016) is a computational tool incorporating an algorithm for predicting dynamics of host-microbial interactions using time-series data. However, some limitations of this work need to be addressed. The framework presented is relatively general (captures only pairwise microbe-microbe interactions) and relies on a series of approximations to the underlying dynamical systems model (Lotka-Volterra dynamical systems). Moreover, results only cover simple systems (mice) instead of the complex human microbiota. (Ridenhour, et al., 2017) presented a study of the temporal variation of microbial communities. They used an ARIMA model with elastic net regularization to estimate ecological interactions and microbial dynamics from 16S sequencing data. In (Faust, Lahti, Gonze, de Vos, & Raes, 2015) authors studied the potential of time-varying networks and time series tools to capture temporal variation of microbial communities in response to perturbations. Their results pointed out how longitudinal analysis can reveal insights into microbial ecosystem dynamics and aid to explain what perturbations (external or internal factors) modulate microbe dynamics and stability.  These studies remark the importance of developing robust time-series analysis in order to uncover insights into microbial interactions and dynamics.

There are several different factors that have been clearly identified for driving the changes in the microbial community such as: (i) antibiotics (Buffie, et al., 2012), (Pérez-Cobas, et al., 2013), (Theriot, et al., 2014), (Ramirez, et al., 2020), (Pérez-Cobas, et al., 2013); (ii) human dietary lifestyle and habits (David, et al., 2014), (Asnicar, et al., 2021), (Berry, et al., 2020), (McNulty, et al., 2013); (iii) host internal process such as hormonal cycles; (iv) pregnancy (Romero, et al., 2014), (Ferrocino, et al., 2018), (Rothenberg, Wagner, Hamidi, Alekseyenko, & Azcarate-Peril, 2019); (v) host disease status (Silverman, 2019), (Baldini, et al., 2020), (Qin J. , et al., 2012); (vi) other microbe-microbe interactions and exchange with external environment or other hosts.

One of the most important features of microbiome is its diversity. We know that microbiome can differ within the same organism, for example, microbiome of the small intestine differs from the oral microbiome. Furthermore, the diversity of the  microbiome between two different individuals (only share 10-20%) is immense compared to the differences between their genomes (99.9% identical) (Ursell, Metcalf, Wegener, & Knight, 2012). The fact that individual species are not commonly shared across the human ecosystem makes the field of microbial research both intriguing and complex at the same time. Moreover, in this context of microbiome diversity, recent findings suggest that, rather than focusing on identifying the microbiome composition of each individual, the importance of microbiome research lies in its functionality given that different microbial species can carry out equivalent metabolic functions and the same species, different functions.

Findings have revealed there is a relationship between microbiome and disease status, but is this association causal or causative? Is it due to causation or correlation? Microbiome research studies have already demonstrated correlation and they are starting to transition towards causation. Although this remains a major challenge in current studies, recent publications have shown promising results for validating causal relationships. For example, (Sanna, et al., 2019) observed significant causal relationship between a specific microbial pathway (4-aminobutanoate degradation) and an increase in insulin secretion. (Koh, et al., 2020), identified in an animal model a microbe that produced a metabolite, Imidazol propionate, that promotes insulin resistance and impairs glucose metabolism. Therefore, this could act as a potential biomarker for type II diabetes. (Ridaura, et al., 2013), used the Missouri adolescent female twin study (MOAFTS) cohort to study fecal microbiota from twins discordance for obesity to demonstrate the influence of host obesity-associated phenotype. (Koeth, et al., 2013) results suggest that trimethylamine-N-oxide (TMAO) produced by gut microbiota contributes to atherosclerosis and cardiovascular disease risk. (Sampson, et al., 2016) reported that altered gut microbiota promotes α-synuclein-mediated motor deficits and brain pathology reducing microglia activation and enhancing Parkinson's disease neurological and motor dysfunction.

Experimental evidence through fecal microbiota transplantation (FMT) has been a major contributor in explaining the causal role of microbiota in disease. This procedure consists of transplanting microbial communities from mice, humans, and other organisms into germ-free (GF) animals (de Groot, Frissen, de Clercq, & Nieuwdorp, 2017). It is important to be aware that this approach may not be valid to every condition. Several studies have employed FMT to demonstrate causality and relationship between gut microbiota and disease (Turnbaugh, et al., 2009), (Arrieta, et al., 2015) and (Lopez & Grinspan, 2016). (Korpela, et al., 2020) showed the ability of maternal FMT during postnatal period for restoring the intestinal microbiota in cesarean section born infants. In 2013, the United States Food and Drug Administration approved FMT for clostridium difficile (*C. difficile)* infection in patients who have not responded to standard therapies (Wang, et al., 2019). Nevertheless, serious adverse events have been reported and have raised the alert of potential risks associated with this procedure limiting its use to experimental studies with human microbiota associated (HMA) murine models, rather than clinical use as a microbiome therapy. HMA murine models remain an interesting approach to establish causal relationships between altered (dysbiotic) gut microbiomes and human disease.

Latest studies are starting to move towards multi-omic approaches. The development of innovative systems biology techniques –such as genomics, metabolomics, transcriptomics, and proteomics– present an opportunity to understand the complex interactions and nature of pathologies. They allow direct analysis of genes, transcripts, metabolites, and proteins from biological samples of the microbiota (Segal, et al., 2019). Approaches that combine information from multiple data sources, such as metatranscriptomics, metaproteomics and metabolomics, have gained popularity due to their ability to provide deeper understandings into functional changes that occur in the microbiome over time (Gerber, 2014). Incorporating multi-omics in our study is key to answer the main questions we are interested in such as 'What are microbiomes doing' 'What functional chemistry is being carried out?' or 'What environmental products are consumed and excreted?'. Also, it is essential to answer questions like 'How are they doing it?', 'Which enzyme pathways are present?' or 'How is the pathway activity?' In this study, we will aim to give answer to four specific key questions outlined in Section 3.1. Figure 2 shows an overview of the different multi-omic approaches in microbiome research that we will present next.

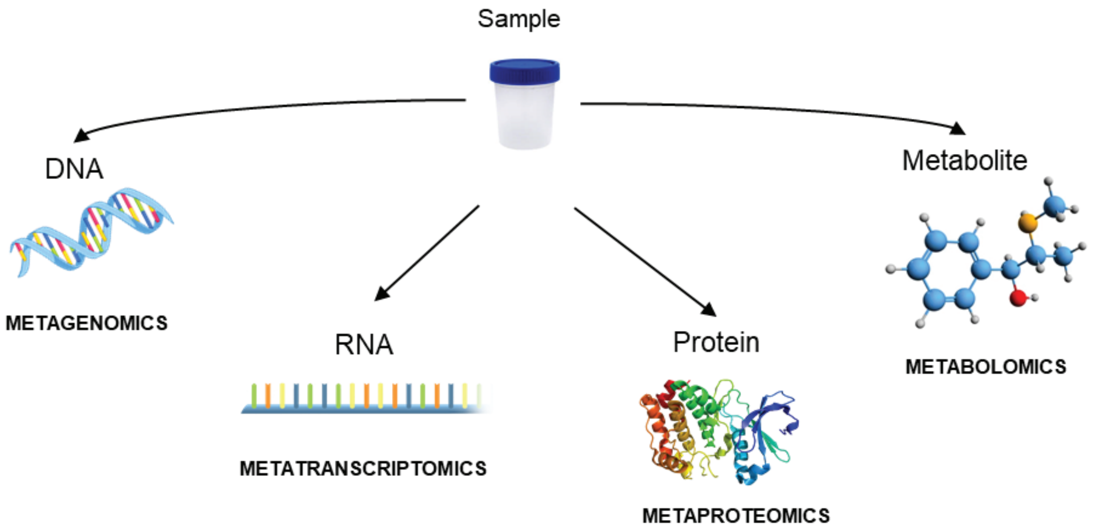


*Figure 2: Overview of multi-omic approaches.*

Promising results from (Yang, Karr, Watrous, & Dorrestein, 2011), (Lamendella, VerBerkmoes, & Jansson, 2012), (Mallick, Franzosa, Vatanen, Morgan, & Huttenhower, 2017) or (Cammarota, et al., 2020) have remarked the usefulness of integrating omic datasets to help unravel taxonomic and functional changes: what are microbes doing and how do they interact between them across time (Bodein, Chapleur, Droit, & Lê Cao, 2019).

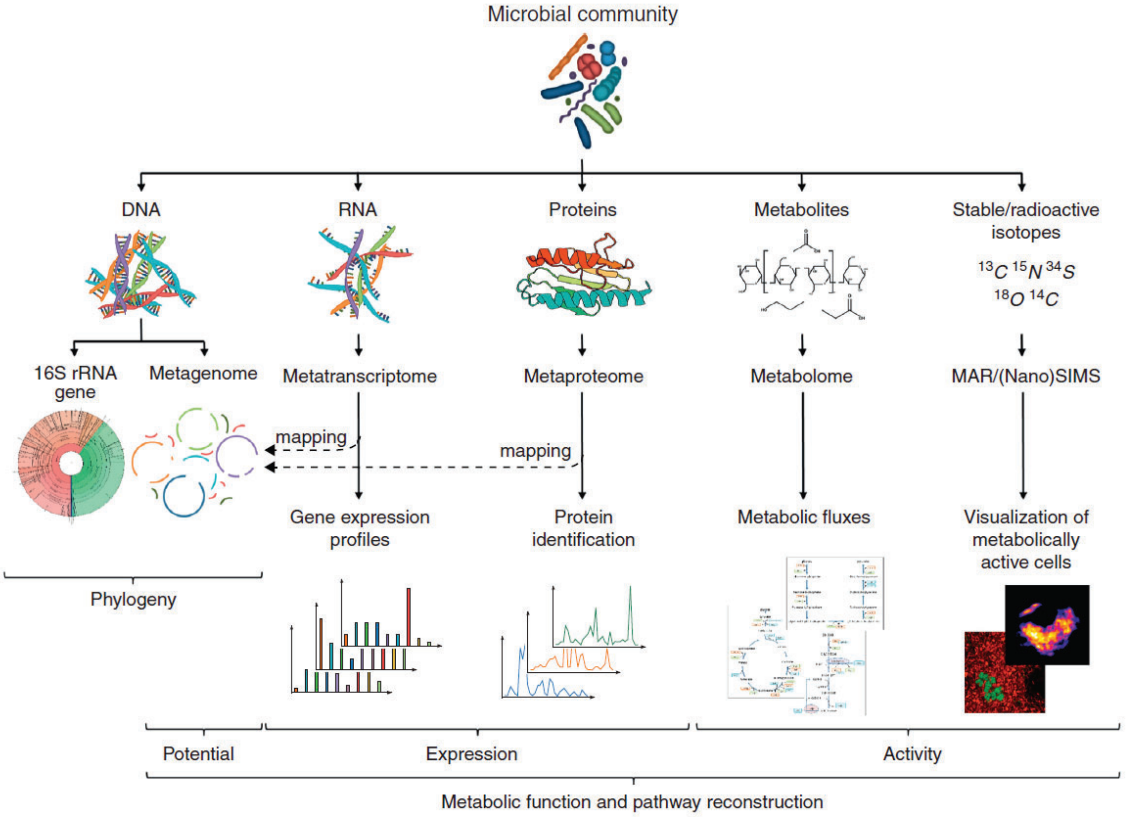


*Figure 3: Meta-omic approaches (Vanwonterghem, Jensen, Ho, Batstone, & Tyson, 2014).*

**Metagenomics** consists of the analysis of sequenced reads, namely genetic material, or metagenome, of microbial community or environment to determine the profile of microbial taxa (Figure 3). (Handelsman, Rondon, Brady, Clardy, & Goodman, 1998) described for the first-time metagenomics, a DNA sequencing approach to study the complex gut microbial community. This methodology allows us to analyze the bacteria in their natural state and define the whole structure of microbial communities.

**Metatranscriptomics** is the field that studies the genes that are expressed by the microbiota genes (or the community). Therefore, the analysis of metatranscriptomics data enables understanding how the microbiome responds to the environment (Figure 3). It helps answering questions of interest such as "What are the microorganisms doing?" and "Which functions are performed by the microorganism?". HUMAnN (Franzosa, et al., 2018) is the tool used to identify functions and pathways of expressed genes (gene family abundances) and how they contribute to microbiota community. It is crucial to analyze the different omics together in order to avoid wrong interpretation of results. When analyzing metatranscriptomics data isolated, the transcript abundance can be confounded with the underlying gene copy number. For example, transcript X may be more abundant in sample A relative to sample B because there are more copies of gene X in sample A relative to sample B (all of which are equally expressed) (Jagtap, et al., 2021).

**Metaproteomics** was originally defined (and still holds) as the characterization of the protein complement of environmental microbiota at a given time point (Wilmes & Bond, 2004), namely the study of proteins expressed by members of the microbiota (microbial proteome). The main advantage of metaproteomics is that it provides the phenotypes of microorganisms on the molecular level. This means it can help scientists study the structure, metabolism, and physiology of microbial community members.

**Metabolomics** is the study of metabolites originated by the microbiota. Metabolomics, the latest of the 'omics' sciences, has progressively been gaining dominance and importance since it started emerging at the beginning of the 1990's. Metabolomics studies the set of metabolites present in a biological system, particularly in biofluids such as urine or blood. While genomics and proteomics give us information of what could have happened (in a living organism), metabolomics can reveal what is happening at present, and therefore help characterize phenotype of organisms. This can be very interesting in personalized medicine, where knowledge of metabolomic variables can serve to predict the reaction of a human to the administration of drugs, in such a way that the treatment could be individualized for each patient, choosing the best active ingredient and the most effective dose, reducing the risk to a harmful reaction. In metabolic pathways, a series of linked reactions take place, where an input (product of one reaction) is processed to produce an output product (substrate for next reaction), just as a fundamental software tool is programmed following an input-process-output model. Being able to identify, characterize, and simulate metabolic pathways constitutes a research area of broad and current interest. Scientists believe these small molecules are the means of communications between microbes and microbes and with their host (human cells). As seen earlier in this work, gut microbiota is involved in the creation of specific metabolites that affect important immune functions crucial for preventing health disorders.

As a whole, microbiome research has proven to have immense potential in the field of **personalized medicine** since each individual has different traits and a unique signature that can be represented by the microbiome. This will represent a relevant growing area of microbiome research in the upcoming years.

Biomarkers serve a crucial role in clinical research as they provide us with:

- Pharmacokinetics analysis
- Monitoring effectiveness
- Early detection of relapses
- Monitoring of safety/toxicity parameters
- Learning the drug mechanism of action
- Exploration of resistance mechanisms

For example, personalized medicine could be used to develop precision oncology for creating targeted therapy and immunotherapy. The microbiome signature we aim to characterize through our model could potentially serve as a biomarker for patient condition prediction and stratification. Increased use of metabolomics in the study of the microbiome will possibly allow broadening the search of biomarkers involving the presence of metabolites derived from microbial activity with certain pathologies and thus reducing interindividual variability. Some examples of potential biomarkers identified in the literature are *Fusobacterium nucleatum* (colon cancer), *Faecalibacterium praustnitzii* (Crohn’s disease).

In order to fully understand our problem and our data it is important to first introduce how microbiome data is generated. The sampling method, sequencing strategy and experimental setup you choose will affect the makeup of the studied community and the perceived abundance of bacteria in that ecosystem. Traditionally, scientist used to cultivate bacterial communities by isolating a specific strain, purifying it and sequencing its genomes, but the problem is the majority of bacteria in the environment is uncultivable. The field of computational biology and genomics experienced a major development due to high-throughput sequencing technologies. Over the years it became increasingly dependent on using DNA/RNA sequencing techniques. Currently data generation reaches the order of tens of thousands of genes sequenced in a single experiment (McGuire, et al., 2020). For metagenomics and microbiome analyses, we will be introducing 16S ribosomal RNA (rRNA) sequencing. A common strategy is to sequence a specific region of the genome, such as the 16S rRNA. This region is a ubiquitous conserved region used to identify, amplify, and sequence the genes. Most microbiome data are generated either by (a) targeted amplicon sequencing (usually region of 16S ribosomal RNA gene) or (b) by metagenomic shotgun sequencing.

A summary of the complete overview of the workflow for microbiome data analysis is presented in Figure 4. Once the data is generated, it is common for data relating to the microbiota to be organized into so-called operational taxonomic units (OTUs), that is, clusters of similar gene sequences. OTUs represent the abundance of particular bacteria.

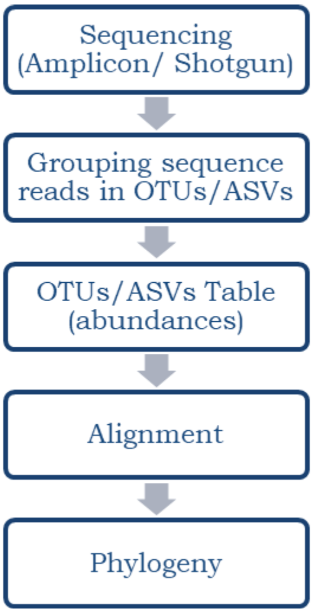

Sequencing (Amplicon/ Shotgun)
Grouping sequence reads in OTUs/ASVs
OTUs/ASVs Table (abundances)
Alignment
Phylogeny

*Figure 4: The general workflow for microbiome data analysis for both amplicon sequence variants (ASVs) and operational taxonomic units (OTUs).*

In our goal of making microbiome research straightforward and comprehensible to computer scientists or other researchers initiating in the field, in Table 1 we present a comparison between both approaches pointing out their corresponding advantages (highlighted in green) and disadvantages in order to aid in the decision of  what would be the best technique according to the aim of their microbiome study.

| **SEQUENCING METHOD** | |
|---|---|
| **Amplicon (e.g.: 16S rRNA marker gene)** | **Whole Genome Shotgun (WGS)** |
| ✓ Most widely used so more available data sets and analysis pipelines. | Expensive sequencing cost. |
| Technical variation from multiple sources and batches. Technical factors include DNA extraction, PCR primers, sequencing platforms or type of sequence reads **(Costea, 2017)**. | ✓ Less sensitive to technical differences in temporal dynamics of microbiome (Wirbel, et al., 2019), (Voigt, 2015). |
| ✓ Most common and cost-effective. | Higher computational cost (both data and computing intensive for analysis). Memory-RAM intensive. |
| Only genus level resolution. | ✓ Expanded taxonomic resolution to species-level. |
| Bacterial coverage only. | Able to identify species from all 3 taxonomic domains (bacteria, eukaryotes and archaea). |

| | |
|---|---|
| Does not directly quantify gene and functions. Limited applicable range for functional profiling. | ✓ Allows analysis of gene functions. |
| Reliability of bacterial classification decreases below genus level. It has poor specificity. | ✓ Offers the possibility to analyze strain or even SNP level dynamics of the microbiome. |
| ✓ Low risk of false positives. | High risk of false positives. |

*Table 1: Amplicon sequencing vs. shotgun metagenomic WGS: sequencing methods comparison.*

Apart from relative abundances of taxa, microbiome diversity can be analyzed by two methods: α-diversity analysis and β-diversity analysis. α diversity describes the diversity within sample/community, that is, species richness and evenness. For instance, the gut microbiota of lean individuals has been found to be significantly more diverse that those of obese individuals. βeta diversity on the other hand, describes diversity between communities/samples, namely differences in microbial composition between communities. It is a useful measure to understand how samples vary against each other. This rational is similar to the 'clustering' algorithms that show differences or similarities among samples. Popular metrics used to estimate distance between communities on 16S rRNA gene-based studies are those based on phylogenetic similarity. Once β diversity has been measured, the dataset can be visualized by principal coordinate analysis (PCoA). PCoA is an ordination technique widely described in the literature for analyzing the composition of different microbiomes. For example, explaining the differences in gut microbiome between non IBD and IBD patients. Some popular distance metrics used for comparing microbial communities are: UniFrac (Lozupone & Knight, 2005), Weighted UniFrac (Lozupone, Hamady, Kelley, & Knight, 2007), Bray-curtis (Bray & Curtis, 1957), Jaccard similarity coefficient.

Once data is generated and grouped into clusters, microbiome data analysis techniques to be performed can be placed into four general categories (Dhariwal, et al., 2017):

1. Taxonomic profiling: to characterize community compositions based on methods developed in ecology such as α-diversity (within-sample diversity) or β-diversity (between-sample diversity)

2. Functional profiling: to assign genes into different functional groups (i.e. metabolic pathways or biological processes) to understand their collective functional capacities. Usually when functional profiling is required, shotgun sequencing technique is preferably used.

3. Comparative analysis: to identify features that are significantly different among conditions under study.

4. Meta-analysis: to integrate user data with public data or knowledge accumulated for improved statistical power or biological understanding.

Resources as well as problems regarding downstream analysis of microbiome data are discussed in later sections. Nevertheless, a deeper analysis of the technologies and biological experimental setup for microbiome research is beyond the scope of this Master thesis.

### 2.1.2 Inflammatory Bowel Disease

Inflammatory bowel disease is a chronic inflammatory disorder of the gastrointestinal tract which comprises Crohn’s disease (CD) and ulcerative colitis (UC) (Gubatan, et al., 2021). IBD affects several million individuals worldwide and is one of the most-studied imbalances between microbes and the immune system. Crohn's disease and ulcerative colitis are complex disorders that are heterogeneous at the clinical, molecular, genetic, and microbial levels. IBD is a chronic disease characterized by periods of relapse and remission (symptom-free periods) (Liverani, Scaioli, Digby, Bellanova, & Belluzzi, 2016).

Still, after several years of research, scientists have not yet found a specific pathogen causing IBD, on the contrary to other diseases, and so it is believed that it is the overall microbial ecosystem dysbiosis causing IBD. Many studies have focused on demonstrating that disrupted composition of the gut microbiota is associated to patients with IBD (Franzosa, et al., 2019), (Lloyd-Price, et al., 2019), (Madgwick, et al., 2020), (Hacilar, Nalbantoglu, O, & Bakir-Gungor, 2020) and (Wang, et al., 2021). Even so, there is still a great need to develop a comprehensive map to understand the nature of microbial changes that will allow improvement of future diagnostic and therapeutic approaches in IBD.

Aside from the medical relevance of this chronic disease, we chose to focus on this disorder as the gut microbiome has been the most widely investigated area of the human microbiome providing us with considerable biological and clinical background we could use for validation and interpretation of our results. Moreover, the integrative Human Microbiome Project (Integrative HMP (iHMP) Research Network Consortium., 2014) offered us a dataset with all the inclusion criteria we needed: longitudinal, multi-omic, publicly available human microbiome dataset.

### 2.1.3 Databases, datasets and computational tools

One of the most important research programs in the field of microbiome analysis is the Human Microbiome Project (HMP), supported by the National Institutes of Health (NIH) common fund. The main goal of the study was to accelerate the classification of human microbiota and its impact in human health and diseases (Peterson, et al., 2009). The project was divided in two phases. Phase 1 (2008-2013) covered the characterization of the microbiomes of healthy human subjects at five major body sites: gastrointestinal tract, mouth, vagina, skin, and nasal cavity. The techniques used were 16S and metagenomic shotgun sequencing. Phase 2 (2013-2016), also known as the integrative human microbiome project (iHMP), generated resources to aid in the characterization of microbiome and human host from three different cohorts of microbiome-associated conditions:

- Pregnancy and preterm birth: MOMS-PI.
- Onset of inflammatory bowel disease: IBDMDB.
- Onset of type II diabetes: T2D.

The techniques used are multiple omics technologies. The iHMP (Integrative HMP (iHMP) Research Network Consortium., 2014) is one of the largest open data resources for studying longitudinal microbiome alterations and relation to disease. It is open source, but its download is not intuitive and requires Aspera client. The HMP Data Portal can be accessed by the following link:

https://portal.hmpdacc.org/. Currently, Bioconductor package “HMP2Data” is under review to access the three datasets in the R environment which we have found to be more convenient and straightforward.

In general terms, the field of microbiome research counts with a considerable amount of publicly available resources and databases designed to address a variety of problems. We aim to present the most common ones used in the literature that are readily encountered in the majority of studies.

- **ML Repo** (Vangay, Hillmann, & Knights, 2019) (https://knights-lab.github.io/MLRepo/). The Microbiome Learning (ML) Repo is a curated repository with the aim of explicitly defining a total of 33 classification and regression tasks. Metadata files are task specific. *ML Repo* includes both amplicon-based and shotgun metagenomics datasets. One of the advantages offered by this tool is, it is easily accessible web-based interface. In total it integrates data from 15 publicly available human microbiome datasets.

- **curatedMetagenomicData** (Pasolli, et al., 2017) is a microbiome-based curated repository offering a collection of shotgun-metagenomics datasets with varying human sample types with gene, pathway, and taxonomic abundance tables. Its data is only accessible via Bioconductor package (https://www.bioconductor.org/) and are stored as ExpressionSet objects which integrate metadata and abundance data. It is relevant to point out that this tool is directed specially to bioinformatic experts.

- **MicrobiomeHD** (Duvallet, Gibbons, Gurry, Irizarry, & Alm, 2017): is a microbiome-based curated repository presenting easily accessible taxonomic abundance tables with curated metadata. One drawback is its limitation only to amplicon-based sequencing data, human stool samples and case-control responses.

- **GMrepo** (Wu, et al., 2020): is a curated database of consistently annotated human gut metagenomes. GMrepo contains 66,133 human gut runs/samples from 295 projects; they are associated with 94 human phenotypes.

- **Mockrobiota** (Bokulich, et al., 2016): is a public resource for microbiome bioinformatics benchmarking using artificially constructed communities. Mockrobiota ensures data integrity and facilitates the microbiome research community manages replication and consistency across studies.

- **KEGG:** The Kyoto Encyclopedia of Genes and Genomes (Goto, et al., 1997) was developed by Kanehisa Laboratories in 1995 with the goal to store functional aspects of genomes based on the concept of binary relation between two molecules. Currently it can easily be considered one of the principal reference knowledge base sources of information of large-scale molecular data sets. Its specific database for pathway maps ’KEGG Pathway’ covers 537 pathway maps. Its latest new approach (Kanehisa, Sato, Furumichi, Morishima, & Tanabe, 2019) incorporates adequate knowledge representation for human genomes, specifically health-oriented information (human gene variants disease related).

- **MicrobiomeAnalyst** (Dhariwal, et al., 2017): is a web-based application to support microbiome data analysis providing comprehensive statistical data analysis, visualization, and meta-analysis for abundance tables and BIOM outputs. An important limitation of the tool is that it does not support multi-omic approaches and only handles taxonomic data as input (e.g.: OTU table, taxon list).

- **QIIME** (Caporaso, et al., 2010): is a widely extended analysis pipeline to deal with amplicon and metagenomic sequencing data. The tool helps scientist assign sequences to samples, cluster sequences of closely related organisms and perform statistical and visualization analysis. QIIME stands out for allowing the user to try many functionalities at different points of the whole analysis. For instance, when performing OTU picking the user can choose from diverse methods such as BLAST, UCLUST, CD-HIT or even *mothur* (Schloss, et al., 2009).

Typically, software tools presented above will ask the user to input a raw sequence data file in one of the following common data formats found in microbiome research:

- FASTA. A popular text format in the field of molecular biology for storing nucleotide and protein sequences.
- FASTQ. A format used for the output text file produced by biological sequencing techniques. The file contains the sequence reads and quality metrics or score for each sequence. If FASTQ file of 16S sequences contains primers and barcodes, a de-multiplexing step is required in order to split sequences by barcodes and one file per sample.
- BIOM, the biological observation matrix format (McDonald, et al., 2012) is a JSON-based file format for presenting sample and observation metadata.

Other open-access repositories and data-sharing platforms that are frequently consulted when conducting microbiome studies include but not limited to:

- Qiita microbial study management platform (https://qiita.ucsd.edu/).
- NCBI-SRA, the National Center for Biotechnology Information (NCBI) Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra).
- European Bioinformatics Institute (EBI) European Read Archive (EBI Metagenomics) (https://www.ebi.ac.uk/metagenomics/).

Problems with these sites is that most samples are organized by study and stored as raw or clean DNA sequences. Furthermore, metadata among studies is generally not unified, that is, every scientist uses their own protocol and structure.

## 2.2 Machine Learning in Microbiome Research

Machine learning (ML) methods are a well-suited solution for handling microbiome analysis, unlocking its full biological and clinical potential, as it is designed to be applied to large number of predictor variables, even sparse data and can learn predictive models indicating which predictor variables are important. Traditional biostatistical analytical methods are sometimes ineffective and limited compared to ML techniques given the inherently noisy and highly variable nature of microbiome data. The idea behind using ML and AI approaches is to aid in reducing the search of potential drivers of diseases coming from large scale human microbiota datasets. ML is also helpful when we have a hard classification task and there are no obvious keystone species identified by just using statistical analysis (Li, Zhang, Wu, Zhou, & Xu, 2018).

Furthermore, the gaining relevance of ML is demonstrated by initiatives like the European Cooperation in Science and Technology COST Action CA18131: Statistical and machine learning techniques in human microbiome studies. The main aim and objective of the Action is to create productive symbiosis between discovery-oriented microbiome researchers and data-driven ML experts, and to optimize and then to standardize the use of said techniques, following the creation of publicly available benchmark datasets.

There are many efforts into developing models and tools in the field. However it has not been until recent years, that more studies have been starting to explore the power of ML methods to predict host traits from microbiome patterns (Knights, Costello, & Knight, 2011), (Larsen & Dai, 2015), (Moitinho-Silva, et al., 2017). In (Fukui, et al., 2020) the authors used LASSO regularized multiple logistic regression to analyze fecal gut microbiota data from IBD patients with the aim of establishing an objective diagnostic tool. Although preliminary results suffer from a number of limitations (e.g. lack of multi-omic integration) it represents a good example of the usefulness and power of ML in disease-associated microbiome data. (Hacilar, Nalbantoglu, O, & Bakir-Gungor, 2020) used supervised and unsupervised ML algorithms to identify bacteria species as potential IBD biomarkers. Hacilar points out the importance of feature selection methodologies such as extreme gradient boosting (XGBoost) or conditional mutual information maximization (CMIM) to select features based on their relevance as metagenomic data is commonly characterized by having larger number of predictors (taxa) compared to the number of samples. The unsupervised learning approach described in (Shomorony, et al., 2020) also contributed to demonstrating the power of employing ML models on multimodal data for discovery of novel biomarkers and disease signatures. In addition, the study evidences the applicability of the approach on longitudinal data.

A recent review of the literature on this matter (Namkung, 2020) found that the random forest has proven to outperform other ML methods when analyzing microbiome profiles associated with disease status. (Cammarota, et al., 2020) also offered an overview of the role and limitations of ML driven approaches and how their flexibility can be exploited to address numerous potential applications in clinical settings. The author alerts the reader about how the quality and amount of input data are determining factors in the whole ML process.

Current research focus is shifting towards causality and complex modeling for clinical applications of diagnostics, prognostics, and therapeutics, where ML models have shown promising applications. Additionally, several studies have pointed out the need for integrating metatranscriptomics and metabolomics to measure microbial functions (Heshiki, et al., 2020). For instance, (Gupta, et al., 2020) used species-level abundances in their random forest-based classifier for prediction of disease. The solution presented, the Gut Microbiome Health Index (GMHI), is a robust index for evaluating health status based on the species-level taxonomic profile of a stool shotgun metagenome (gut microbiome) sample. However, (Gupta, et al., 2020) declared one main weakness of their study was precisely not incorporating metagenomic functional profiles.

With this in mind, we would like to enumerate a series of  potential input features in microbial studies to build a richer and more informative classifier or ML model: relative abundance of taxa vectors (from 16S rRNA gene sequence), sample-by-taxon abundance matrix, abundances of functional genes, OTUs, α-diversity and β-diversity, MetaPhlAn2 species-level relative abundances, MetaPhlAn2 strain-specific markers presence (or species) or relative abundance of each metabolic pathway (metatranscriptomics).

There is still a lack of good standardized analytical framework protocols and ML knowledge in the microbiome community. Thus, further intense work is needed to bridge the gap between microbiome researchers and computer scientists in the correct implementation and usage of these approaches. ML models will assist to provide systematic insights into possible causal or contributing roles of the microbiome. However, many experts in the community contend there is still considerably uncertainty in causal mechanisms and thus this remains an active area of research (Topçuoğlu, Lesniak, Ruffin, Wiens, & Schloss, 2020).

### 2.2.1 Bayesian networks

Among all the different ML approaches and models, BN-based analysis is certainly one of the most biologically interpretable (Wang, et al., 2019). The need for explainable artificial intelligence models is highly demanded by microbiome researchers nowadays. Therefore, in this work, we will focus on the application of BN to the microbiome research field.

A BN (Pearl, Probabilistic Reasoning in Intelligent Systems, 1988) can be defined as a graphical model used to describe the joint probability distribution over a set of random variables. By means of a directed acyclic graph (DAG), conditional (in)dependence relations (that can be causal under some circumstances) are represented by arcs, and random variables by nodes. This model offers an intuitive and solid approach to modelling uncertain knowledge.

In order to construct a BN, the structure S (DAG) which expresses the conditional (in)dependencies among triplets of variables and the parameters θ of the model that determine the conditional probability distributions need to be learned from observational data. Nevertheless, this task is nontrivial (Chickering D. , 1996), and has aroused considerable interest in the scientific community as many other NP-hard problems. Methods that address the challenge of learning causal structure from data can be classified in three main groups: constraint-based, score-based and hybrid methods.

*Constraint-based methods* (Spirtes, Glymour, & Scheines, Causation, prediction, and search, 1993) involve the use of statistical tests to discover conditional independencies between variables which are then employed to define the structure that represents these relationships. The three main steps that every method of this kind follow, given observational data are: (1) learn the skeleton of the network (undirected graph), (2) set all the directions of the v-structures, and (3) set all the directions of other arcs. The main method is the PC algorithm (Spirtes, Glymour, & Scheines, Causation, prediction, and search, 1993).

*Score-based methods* consists of giving a score to each DAG according to its ability to fit the given data based on a metric function and a heuristic method to search the space of solutions such as a greedy search (Cooper & Herskovits, 1992). For other examples of this method and a deeper understanding of its mechanism, the following articles can be reviewed: (Chickering D. , 2002); (Cano, Gomez-Olmedo, & Moral, 2008) and (Niinimaki & Parviainen, 2012).

*Hybrid methods* imply a combination of both above techniques in one algorithm such as the well-known technique Max-Min Hill Climbing (MMHC) (Tsamardinos, Brown, & Aliferis, 2006).

Inductive causation (IC) algorithm (Pearl & Verma, 1991) provided the first framework for learning the skeleton of a Bayesian network by using a backward strategy that starts with a complete graph that will be pruned following the results of statistical tests for conditional independencies. IC was closely followed by SGS algorithm (Spirtes, Glymour, & Scheines, 2000) and by the most popular method, the PC algorithm which constitutes both the first practical implementation and the improvement of the former algorithms. PC algorithm composed of two principal steps: (i) find the skeleton (detection phase) and (ii) make orientation of the edges (orientation phase). They showed the relevance of causal Markov and causal faithfulness assumptions for linear models. The Markov blanket of a random variable *X* in a BN, under the faithfulness assumption, consists of the union of the set of nodes (parents, children, and parents of children) (Pearl & Verma, 1990) of *X*. Therefore, the Markov blanket is the minimal set of nodes for which *X* is conditionally independent of all other nodes (Borchani, Bielza, Martínez-Martínez, & Larrañaga, 2012).

Other important local methods are: Grow Shrink, GS (Margaritis, 2003) and Incremental Association Markov blanket, IAMB (Tsamardinos & Aliferis, 2003) both of them follow a forward step-wise selection Markov blanket detection approach, so the learn in first place the Markov blanket of each node simplifying the identification of neighbors and hence reducing the number of conditional independence tests that need to be computed.

Several algorithms have been proposed that extend the PC algorithm. First, some of them focus on improving the efficiency in high-dimensional spaces and increasing data size such as the semi-interleaved HITON-PC (Aliferis, Statnikov, Tsamardinos, Mani, & Koutsoukos, 2010), the PC extension of (Kalisch & Bühlmann, 2007) and (Harris & Drton, 2013), parallel PC (Le, et al., 2015) and the reduced-PC algorithm (Sondhi & Shojaie, 2019).

A well-known extensions of IAMB algorithm is the Interleaved incremental association (Inter-IAMB) algorithm (Yaramakala & Margaritis, 2005) manages to reduce and avoid false positives in the Markov blanket detection phase.

Second, some important variants of the PC algorithm are: PC-stable, Conservative PC and Adjacency Conservative PC algorithm. The PC-stable algorithm (Colombo & Maathuis, 2014) manages to implement the orderindependency of input variables and obtain solid results. Order-dependence is one of the mayor drawbacks of PC-based methods as it limits its applicability to real-world problems and introduces errors. It has been targeted and solved in other work such as (Qi, Fan, Gao, & Liu, 2019). Conservative PC (CPC) (Ramsey, Zhang, & Spirtes, 2006), is an extended version of PC that aims at detecting violations of orientation-faithfulness in the orientation phase. The Adjacency Conservative PC algorithm (ACPC) (Lemeire, Meganck, Cartella, & Liu, 2012) extends the CPC algorithm but in this case, it detects violations of adjacency-faithfulness.

Once the structure of the network is known, the conditional probability distributions of each random variable (node) given its parents can be estimated. One approach to learning parameters for BN modeling is maximum likelihood estimation. The goal of this statistical method is to maximize the probability of obtaining $D$ for a specific value of $\theta$, where $D$ = {$\mathbf{x}^{(1)}$, $\mathbf{x}^{(2)}$,…,$\mathbf{x}^{(N)}$} represents the data set given the BN model $G$. This operation results in the likelihood function $p(D|G,\theta)$. An alternative approach is to use Bayesian estimation based on prior knowledge as a prior joint distribution over the parameters or structures.

When using BN, we would commonly be interested in capturing reasoning patterns under uncertainty. BNs allow us to do this by computing the distribution of some set of variables which we have not observed, a process known as probabilistic inference. In the simplest case, given an observation (evidence) $\boldsymbol{e}$ we can query the model to calculate the posterior probability of a target variable(s) or node $X_j$: $p(x_j|\boldsymbol{e})$. Multiple methods have been developed over the years to perform approximate inference (Henrion, 1990), (Shachter & Peot, 1990), (Golightly & Wilkinson, 2011) instead of exact inference as this latter case implies an intractable (NP-hard) problem for densely connected BNs. Nevertheless, (Dagum & Luby, 1993), demonstrated that even approximate inference is NP-hard.

An important consideration to take into account when working with BNs is the type of data being studied. Variables included in the network can be discrete or continuous and according to this, a different type of assumptions and parametric distributions will be estimated for the nodes. In the case of microbiome data, we will typically be dealing with continuous data. Most commonly used parametric distribution for this case would be *Gaussian* or *Gaussian mixture model* (Vatanen, et al., 2012) which models all conditional distributions as linear Gaussians and all continuous nodes follow a multivariate normal distribution $N(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$. However, we could still be presented with the case where we have both continuous and discrete variables in the same dataset such as clinical variables (continuous) and pathway abundances (discrete).

A conditional Gaussian Bayesian network (CGBN) models discrete nodes as conditionally independent probability distributions dependent on the values of their discrete parents and modeling continuous nodes as conditionally independent Gaussian distributions linearly dependent upon their Gaussian parents and with parameters conditioned the values of the discrete parents.

If the CGBN has a directed acyclic graph G over discrete variables Δ and continuous variables Ψ, where π(X) is the (possibly empty) set of parents of variable X according to G, and there is a set of conditional probability distributions P over Δ, and a set of conditional linear Gaussian density functions F over Ψ, then the model results in a multivariate normal mixture density over all variables (Madsen, 2008) is:

$$P(\Delta)f(\Psi|\Delta) = \prod_{x\in\Delta} P\big(x\big|\pi(x)\big) \prod_{y\in\Psi} f\big(y\big|\pi(y)\big)$$

When π(X) is empty, P(X) and f(Y) are just (unconditional) probability or density functions, respectively.

Therefore, as seen above, BN show great potential due to their ability to deal with uncertainties related to limited short and sparse data, and their power to detect informative patterns of the underlying system. Moreover, (Layeghifard, Hwang, & Guttman, 2017) discuss in their review the potential of network-based approaches applied to microbiome research. This work highlights that given the complexity and sparsity of microbiome data, network theory (e.g. probabilistic graphical models) provides a holistic approach for modeling biological systems and analyzing comprehensive interactions between microbial community members. Network biology constitutes a powerful tool for understanding of human microbiome, yet, due to its complexity in terms of implementation it still needs to be further developed.

As seen in Section 2.1.1, being able to yield insights into the dynamic behavior of microbiota, identify patterns of variation in longitudinal microbiome data and link these to patterns of host status are key in the advance of microbiome research. In this context, dynamic Bayesian networks (DBNs) (Dean & Kanazawa, 1989) represent an important approach for time-series human microbiome data analysis. DBNs, extend BNs to model time-series data (dynamic systems) (Murphy K. , 2002), where at each (discrete) time instance *t* (or slice), nodes correspond to random variables at time *t* and directed edges correspond to conditional dependencies in the DAG.

The edges of a DBN can be defined as (i) *inter-slice arcs*: the arcs that directly connect nodes from two consecutive time slices (ii) *intra-slice arcs*: the arcs that connect nodes from the same time slice. In DBNs certain assumptions are used: (i) *first-order Markov assumption,* i.e., *so* the probability of an observation at time *t* only depends on the observation at time *t*-1; (ii) *stationarity,* the data is generated by a distribution that does not change with time.

The resulting DBN can be modeled for $X^t = (X_1^t, \dots, X_n^t)$, $t = 1,\dots,T$ as shown in Figure 5 we have that,

$$P(X^1) \prod_{t=2}^{T} P(X^t \;|X^{t-1}) = P(X^1,\dots,X^T)$$

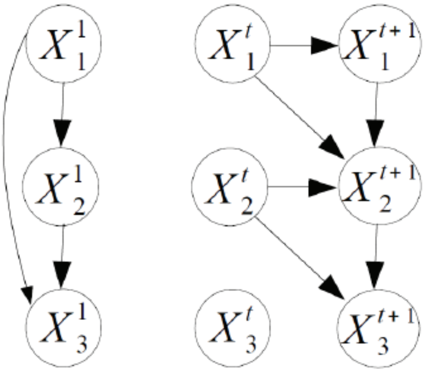


*Figure 5: Example of a DBN. Prior BN (left) and transition BN (right) for three variables.*

Due to the characteristics that microbiome data exhibit, with strong temporal fluctuations that we are interested in modeling, the use of DBNs can help us handle temporal behavior of the system, provide information about the ordering and dependencies between the time points or show how one taxon/pathway/metabolome influences another over time (the connections between nodes). Likewise, DBN are considered generative models which are an ideal tool for dealing with compositional data (microbiome data).

Van Gerven and colleagues presented in 2008 one of the first publications to introduce the use of DBNs in the complex domain of medicine for prognostic prediction models. DBNs have also been frequently applied in gene regulatory networks (Murphy & Mian, 1999), (Husmeier, 2003), (Zou, Denby, & Feng, 2009). Other studies like (Faust, Lahti, Gonze, de Vos, & Raes, 2015) did not cover the use of BNs, but they suggested that their method could be combined with DBNs to build a time varying DBN method.

Despite the increasing interest microbiome research has arisen, to the best of our knowledge, very few studies have applied BN to human microbiome data. In Table 2 we resume the state of the art of BN applied to real human microbiome data. Studies appear in chronological order.

| # | Study | Method | Dataset | Longitudinal data | Meta-omics | Goal |
|---|---|---|---|---|---|---|
| 1 | (McGeachie, et al., 2016) | DBNs | Premature infant gut (La Rosa, et al., 2014) | Yes | No | Build a DBN model to identify important relationships between microbiome taxa and predict future changes in microbiome composition |
| 2 | (Noyes, Cho, Ravel, Forney, & Abdo, 2018) | BNs | Vaginal microbiome (Ravel, et al., 2011) | No | No | Demonstrate associations between women's sexual and menstrual habits, demographics, vaginal microbiome composition and symptoms and diagnostics of bacterial vaginosis (BV) |
| 3 | (Lugo-Martinez, Ruiz-Perez, Narasim | DBNs | Infant gut (La Rosa, et al., 2014) | Yes | No | Obtaining inferences from time-series data |

| | | | | | | |
|---|---|---|---|---|---|---|
| | han, & Bar-Joseph, 2019) | | | | | |
| 4 | (Howey, Shin, Relton, Davey Smith, & Cordell, 2020) | BNs | Twins UK (Moayyeri, Hammond, Hart, & Spector, 2013) | No | No | Possible causal relationships between metabolites and body mass index (BMI) |
| 5 | (Jang, et al., 2020) | BNs | Rectal cancer (Jang, et al., 2020) | No | Yes | Reveal differential microbial communities and functions in terms of therapeutic responses |
| 6 | (Kharrat N. A.-E., 2019) | BNs with the incremental dynamic analysis (IDA) method | Colorectal cancer (i) (Marchesi, 2011) (ii) (Zeller, et al., 2014) | No | No | Identify key species that are likely to be causal agents of colorectal cancer (CRC) |
| 7 | (Sazal, Mathee, Ruiz-Perez, Cickovski, & Narasimhan, 2020) | BNs with co-occurrence networks (CoNs) | Infant gut (La Rosa, et al., 2014)Vagi-nal: (Ravel, et al., 2011), oral data (HMP) | No | No | Make inference about colonization order |
| 8 | (Ruiz-Perez, et al., 2021) | DBNs | IBDMD (inflammatory bowel disease multi-omics database) (Lloyd-Price, et al., 2019) | Yes | Yes | Infer temporal relationships between entities in a microbial community and extend (Lugo-Martinez et al., 2019) to other omics. |

*Table 2: State of the art of BNs models applied to human microbiome datasets.*

The first report on the use of DBNs for human microbiome data analysis, according to authors, was (McGeachie, et al., 2016). Their work is the pioneer study to build a DBN model to capture the influence of individual microbial classes on each other over time. Most important pitfalls of this study are the simplification of data and models or vastly reducing the size of the data by aggregating the data at certain taxonomic levels. Moreover, the study was limited to taxonomic analysis only (non multi-omic) so the exact nature of the biological mechanisms underlying taxonomic relationships remain unknown.

Subsequent study found in the literature (Noyes, Cho, Ravel, Forney, & Abdo, 2018) was limited to the use of traditional BNs, thus data analyzed was static and the need, in some cases, to discretize the data could have probably resulted in loss of information. Nevertheless, their preliminary work reported interesting results and the learned BN model confirmed the importance of vaginal pH and *Gardnerella* as influencers on the Nugent score (bacterial vaginosis diagnostic).

In study number three of Table 2 (Lugo-Martinez, Ruiz-Perez, Narasimhan, & Bar-Joseph, 2019), the focus of the learned model was restricted to providing knowledge on how the abundance of taxa depended on the abundance of other taxa and clinical variables. The study presents a novel approach to analyze longitudinal microbiome data using temporal alignments before learning a DBN to account for the different paces of biological processes. When compared to prior published methods, the developed approach outperformed both the baseline and previous methods for the same dataset (infant gut). Unfortunately, this computational pipeline is only able to analyze a single omic data set. As many others have highlighted, Lugo-Martinez's approach needs to integrate additional molecular data (metabolomics/gene expression). Importantly, the author highlights critical challenges when dealing with time-series data such as sampling rates and missing values which could affect the accuracy of the network being modeled.

Other recent studies which successfully applied conventional BNs to human microbiome data are (Howey, Shin, Relton, Davey Smith, & Cordell, 2020) and (Jang, et al., 2020). As an innovative novel approach, Howey's work incorporates directed arcs representing genetic anchors to the BN analysis. Results of this study concluded that BNs outperform other recently-proposed methods and the model serves as a complementary approach to Mendelian randomization (Smith & Ebrahim, 2003) for analyzing causal relationships in complex biological scenarios. The dataset used in this case was comprised of fatty acid metabolites along with clinical metadata (body mass index). To learn the structure and parameters of the BN, authors used R package `bnlearn` with the integrated hill-climbing score-based algorithm and Bayesian information criterion (BIC) score. Jang's study used BNs to investigate species level of taxa in CR (complete response of rectal cancer treatment) vs non-CR patients by discretization of continuous variables (microbial taxa) into two states according to their relative abundance.

An alternative approach was presented by (Kharrat N. A.-E., 2019), where in order to identify key species likely to cause colorectal cancer (CRC), a BN was combined with the "Interventional-calculus when the DAG is absent" (IDA) method (Maathuis, Kalisch, & Bühlmann, Estimating high-dimensional intervention effects from observational data, 2009), (Maathuis, Colombo, Kalisch, & Bühlmann, 2010) to generate a model for inference of causal relationships.

(Sazal, Mathee, Ruiz-Perez, Cickovski, & Narasimhan, 2020) constructed co-occurrence networks (CoN) (Fernandez, Riveros, Campos, Mathee, & Narasimhan, 2015) using the Pearson correlation coefficient. A network was built for each cohort (medium BV and advanced BV). Next, the coefficients that were generated by the CoN were used to augment the BN. The combined output, referred to as the signed BN, a variant of BNs, was used to discover relationships present in the studied microbial community. The ultimate goal was to retrieve "*colonization order*" from bacterial abundance data.

Finally, (Ruiz-Perez, et al., 2021) extended previous research groups activity (Lugo-Martinez, Ruiz-Perez, Narasimhan, & Bar-Joseph, 2019) to account for multi-omic dataset integration. Their pre-processing pipeline includes temporal alignment of the data to correct for the different progression rates of each individual. The sub steps of the pipeline include: (1) filtering abnormal and noisy samples by computing the mean and standard deviation of the alignment error, and (2) removing all samples from an individual where alignment error exceeded a certain threshold as previously described in (Lugo-Martinez, Ruiz-Perez, Narasimhan, & Bar-Joseph, 2019). (Ruiz-Perez, et al., 2021) work employs prior knowledge (two sets of constraints: *skeleton* and *augmented)* to constraint the resulting model and reduce overfitting. Their model used four types of omic data: taxa, genes, host genes and metabolites.

As studied in this brief literature review, none of the existing studies cover all the objectives raised in Section 1.3. of this master thesis. Therefore, we believe that, even though similar work has been presented in very recent years in the literature, the specific focus of our work is novel and will provide relevant insights to the community.

### 2.2.2 Random forests

As (Moreno-Indias, et al., 2021) recent review of ML tools for microbiome research pointed out, the random forest (RF) remains the popular model of choice when applying ML models for disease-prediction tasks with microbiome data.

RFs (Breiman, 2001) are defined as a type of ML classifier consisting of a collection of tree-structured classifiers obtained from independent identically distributed random vectors of rows and columns and each tree casts a unit vote for the most popular class as the final output. RF is an approach that combines bagging and randomization by introducing randomness in a complementary and different form.

For our present study, using RFs was discarded for two main reasons: it was already broadly explored in the literature and moreover, even though their performance might be superior, it is not an intuitive model and acts as a "black box" which is far away from our goal of presenting a biological interpretable solution to the microbiology and clinical community. Nevertheless, in our work, we are interested in evaluating RF performance against the proposed DBNs as a benchmarking procedure in order to verify if RF also outperforms DBNs.

Table 3 serves as an example of published performance of RF (Pasolli, Truong, Malik, Waldron, & Segata, 2016). Prediction performance was evaluated by the area under the curve (AUC) metric. In each paper, RF was the method that outperformed the rest: support vector machines (SVM) (Cortes & Vapnik, 1995), Lasso (Tibshirani, 1996) and elastic net (Zou & Hastie, 2005).

| **Dataset** | **Num. Samples** | **Metric** | **Method** | **Value** |
|---|---|---|---|---|
| Liver cirrhosis (Qin, et al., 2014) | 232 | AUC | **RF** | **0.95** |
| | | | SVM | 0.92 |
| | | | Elastic net | 0.91 |
| | | | Lasso | 0.88 |
| Colorectal cancer<br>(Zeller, et al., 2014) | 121 | AUC | **RF** | **0.87** |
| | | | SVM | 0.81 |
| | | | Elastic net | 0.79 |
| | | | Lasso | 0.73 |
| IBD<br>(Qin J. , et al., 2010) | 110 | AUC | **RF** | **0.89** |
| | | | SVM | 0.86 |
| | | | Elastic Net | 0.83 |
| | | | Lasso | 0.81 |
| Obesity<br>(Le Chatelier, et al., 2013) | 253 | AUC | **RF** | **0.66** |
| | | | SVM | 0.65 |
| | | | Elastic net | 0.64 |
| | | | Lasso | 0.60 |

*Table 3. Benchmark of various machine learning techniques prove the superiority of random forest when applied to different metagenomic datasets (Pasolli, Truong, Malik, Waldron, & Segata, 2016).*

Other papers have also elucidated the competitive and accurate performance of RFs when compared to other methods (Zhou & Gallins, 2019). A newly published study (Sarrabayrouse, et al., 2021) by the prestigious Vall d'Hebron Research Institute in Spain also selected RF as the ML method of choice to classify unaffected relatives of IBD patients, discriminate CD from UC, and predict disease relapse with a good average performance.

### 2.2.3 Explainable AI

We could arguable state that, although not a novel trend, explainable Artificial Intelligence (XAI) is of broad and current interest. In recent years, innovative ML and AL algorithms such as deep learning have become increasingly complex and sophisticated (Castelvecchi, 2016), (Holzinger, Biemann, Pattichis, & Kell, 2017). Consequently, there is an unprecedent need, requested by non-experts in the domain, of developing transparent and understandable models. Specially in the clinical field, being able to explain the reasoning behind the decisions and results is of critical importance for applicability in medicine (Barredo Arrieta, et al., 2020).

To the best of our knowledge, this line of research hadn't been applied to microbiome research until this past year and in a minor number of publications (Prifti, et al., 2020), (Carrieri, et al., 2021), (Wong, et al., 2021), emphasizing the novelty of our approach.

## 2.3 Limitations and Challenges

The main challenges and opportunities encountered in the field of microbiome data analysis can be grouped into four areas:

I. Data Size

Current datasets lack large-scale data. Sample size keeps coming up as a protagonist limiting factor in the full potential of results. Rather than focusing on gathering volumes of low-quality data, studies need to concentrate on setting up experimental protocols that guarantee regular sampling and sufficient time-points for downstream analysis. Currently, studies suffer from economic and logistic constraints that limit and affect data collection standards. Further advantages could be taken once we define how to decode large-scale microbiome data in a precise and efficient manner (Su, Jing, Zhang, & Wu, 2020).

II. Comparability and reproducibility

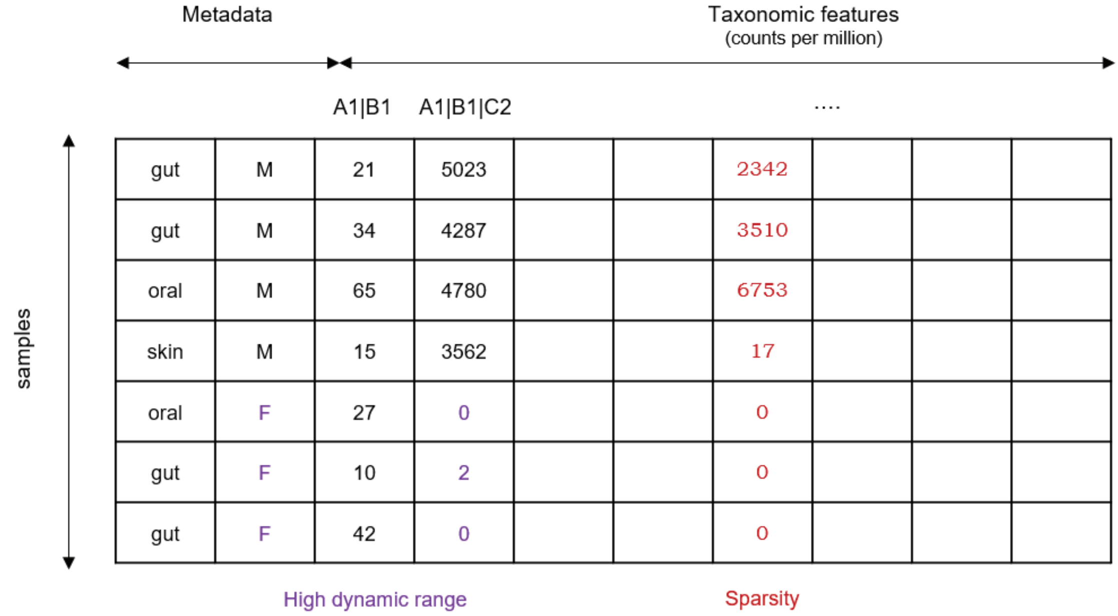
The lack of validated clinical models and differences in methodologies is preventing the translation of valuable results into the real-world clinical practice. Analysis of human microbiome data involves preprocessing of raw sequences and there is a latent need to develop manual curation and standardized protocols to prevent from derived variations in results. Linked to the following point, human microbiome presents huge host-to-host variability (heterogeneity) causing difficulties when trying to extend models to other cohorts (Duvallet, Gibbons, Gurry, Irizarry, & Alm, 2017).

III. Inherent characteristic of microbiome data

Sparsity, compositionality, high variability (Aitchison, 1982), (Gloor, Macklaim, Pawlowsky-Glahn, & Egozcue, 2017) are the main statistical properties that describe microbiome data and present computational challenges. High-throughput RNA-seq technologies used in the process of generating microbiome data from the sample often introduce technical artifacts that translate into errors and noise. For that reason, the bottleneck has shifted from data generation to data analysis, which is essentially influenced by the way the data is generated in a first instance. Moreover, microbial communities are highly complex, nonlinear, evolving systems that can be chaotic and therefore unpredictable (Faust & Raes, 2012).

Furthermore, microbiome data is compositional so instead of looking at absolute abundances of cells, we are mapping reads and there is a fixed sequencing depth, i.e., 4 reads/sample, given by the technology used to obtain the sequences. The number of reads is inferior to the number of cells in the experiment, so we are examining proportions of reads.

As seen in the below taxon-per-sample abundance matrix (Figure 6), most taxa are confined to a relatively small fraction of samples (sparsity).


Metadata
Taxonomic features
(counts per million)
samples
High dynamic range
Sparsity

| | | A1\|B1 | A1\|B1\|C2 | | | | .... | | |
|---|---|---|---|---|---|---|---|---|---|
| gut | M | 21 | 5023 | | | 2342 | | | |
| gut | M | 34 | 4287 | | | 3510 | | | |
| oral | M | 65 | 4780 | | | 6753 | | | |
| skin | M | 15 | 3562 | | | 17 | | | |
| oral | F | 27 | 0 | | | 0 | | | |
| gut | F | 10 | 2 | | | 0 | | | |
| gut | F | 42 | 0 | | | 0 | | | |

*Figure 6: Visual example of inherent characteristics of microbiome data. Metadata include microbiome area (gut, skin, mouth) and if subject is female (F) or male (M).*

## IV. Interpretability

With rapid advances in microbiome research, the community has raised some concerns related to the interpretability of models. After a detailed analysis of the state-of-the-art, we cannot emphasize enough the advantages of integrating multi-omic datasets with the goal of building a holistic view of microbial community and their interactions. Incorporating phylogenetic and functional relationships among organisms into unified dynamic models of human microbiome is crucial. Studies need to start moving away from unique taxonomic composition analysis which limits discovering the whole picture.

Other opportunities in the field, as pointed out in (Su, Jing, Zhang, & Wu, 2020), will be how microbial profiling methods will benefit from prospect improvements of species or strain-level resolution with full-length 16S and reduction in sequencing costs for WGS-based profiling through shallow WGS.

In general, even though BN and other ML models offer numerous advantages versus traditional statistical analysis they also suffer from a number of pitfalls. ML approaches find difficulties estimating the true real-life performance of the model. In our particular longitudinal data type, due to the complex dynamics of the microbiome, estimating confidence in predictions becomes more complicated.

Most studies involving time-series data (longitudinal studies) have mentioned having few unevenly spaced time points (Bodein, Chapleur, Droit, & Lê Cao, 2019) as a major limitation to accurate realistic results. However, we have detected an absence of studies addressing the sampling frequency problem associated with temporal data. Determining how frequently should a host-microbial ecosystem be sampled has been overlooked and remains unclear. This matter will probably depend on time-scale changes of interest. Nevertheless, careful consideration should be put into sampling frequency as undersampling could lead to temporal aliasing (Gerber, 2014). As in many other ML applications, identifying variables or attributes (e.g., OTUs) that will produce both good discrimination within the training data and good generalization to feature tests data is still a challenge.

# 3 Materials and Methods

## 3.1 Problem definition

Research questions that motivated this work and are:

- Q1: Which microorganisms are present in our sample?
- Q2: What functions are performed by the organisms?
- Q3: Can we identify a biomarker for IBD prediction and stratification?
- Q4: What is the microbiome profile (functions and communications) for each condition type?

The goal of this work is to use DBNs to find a biological interpretable model of gut microbial ecosystem in IBD patients that will give answer to the above questions.

## 3.2 Tools

Programming languages used for this project include:

- R: used for visualization and application of statistical analysis.
- Python: used for the preprocessing steps. The following specific packages were used: *Scikit-learn*, *numpy*, *pandas*.
- Matlab: used for DBN learning and inference.

We used open-source available software package *Cytoscape* (Shannon P. M., 2003) for interpretation and visualization of the resulting network.

## 3.3 Dataset

### 3.3.1 Description

We describe the data on which we test our dynamic Gaussian BN model and inference.

We used the dataset from the Inflammatory Bowel Disease Multiomics database (IBDMDB) iHMP study (Lloyd-Price, et al., 2019). As part of the Integrative Human Microbiome Project (HMP2 or iHMP), IBDMDB followed 132 subjects over the period of one year to generate integrated longitudinal molecular profiles of host and microbial activity during disease (up to 24 time points each). The IBDMDB dataset is a comprehensive multi-omics dataset that includes metatranscriptomics, metagenomics, proteomics, viromics, serology, host transcriptomics, 16S and metabolomics data. The results provide longitudinal profiling of the biological properties of the human gut microbiome in IBD and a comprehensive view of functional dysbiosis in the gut microbiome during IBD activity (Integrative HMP (iHMP) Research Network Consortium., 2014). Raw sequence data can be download from the BioProject NCBI site with accession code PRJNA398089.

The study's resources, results, and data, which are available through the IBDMDB webpage http://ibdmdb.org, provide the most comprehensive description to date of host and microbial activities in inflammatory bowel diseases.

As described in Figure 7 clinical metadata such as disease activity, treatment and interventions information and diet and environment details were collected. The data set contains a total of 2951 samples and participants were classified attending to their disease status into one of the three classes: non-IBD or controls with a total of 46 subjects, ulcerative colitis with a total of 46 subjects and Crohn's disease with 86 participants.

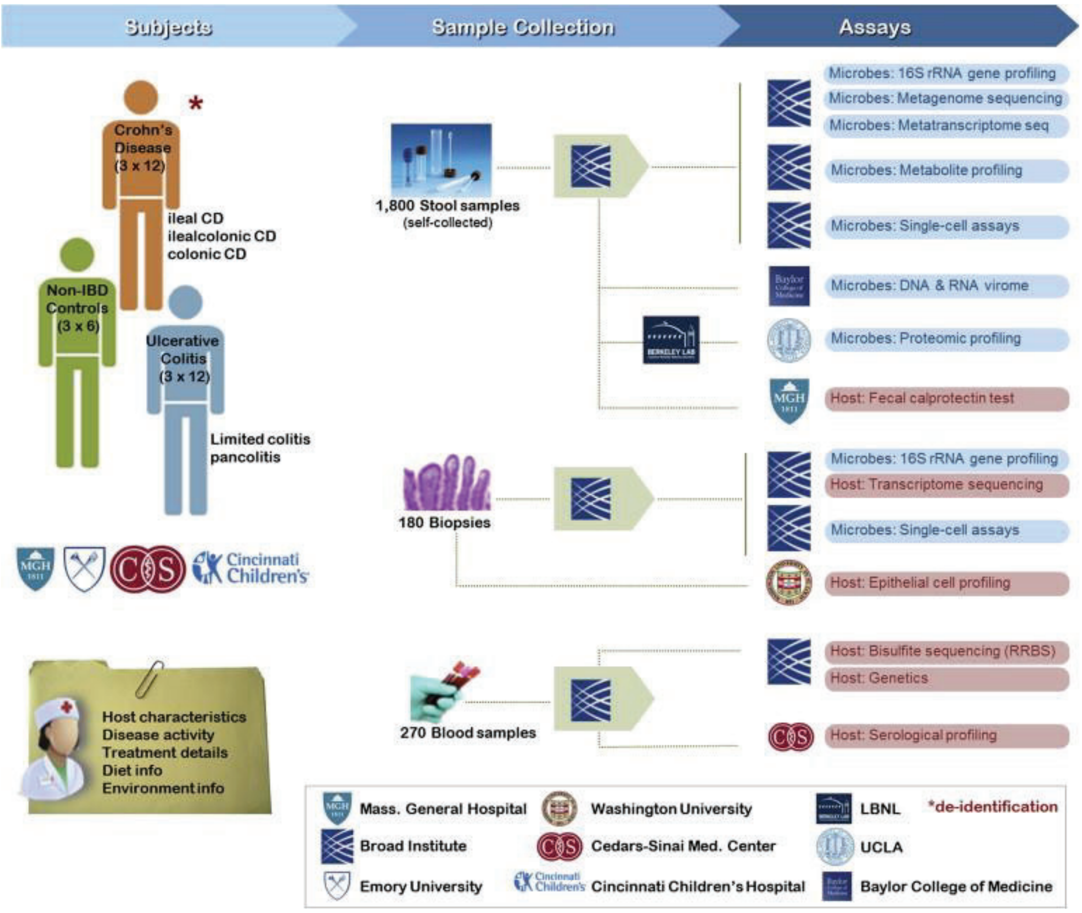

Subjects
Sample Collection
Assays
Crohn's Disease (3 x 12)
*
ileal CD
ilealcolonic CD
colonic CD
Non-IBD Controls (3 x 6)
Ulcerative Colitis (3 x 12)
Limited colitis
pancolitis
Cincinnati Children's
Host characteristics
Disease activity
Treatment details
Diet info
Environment info
1,800 Stool samples (self-collected)
180 Biopsies
270 Blood samples
Microbes: 16S rRNA gene profiling
Microbes: Metagenome sequencing
Microbes: Metatranscriptome seq
Microbes: Metabolite profiling
Microbes: Single-cell assays
Microbes: DNA & RNA virome
Microbes: Proteomic profiling
Host: Fecal calprotectin test
Microbes: 16S rRNA gene profiling
Host: Transcriptome sequencing
Microbes: Single-cell assays
Host: Epithelial cell profiling
Host: Bisulfite sequencing (RRBS)
Host: Genetics
Host: Serological profiling
Mass. General Hospital
Washington University
LBNL
*de-identification
Broad Institute
Cedars-Sinai Med. Center
UCLA
Emory University
Cincinnati Children's Hospital
Baylor College of Medicine

*Figure 7: Characterizing the gut microbial ecosystem for diagnosis and therapy in inflammatory bowel disease: sample collection, assay, and data generation workflow (Integrative HMP (iHMP) Research Network Consortium., 2014).*

Publicly available data repository shows high level results over all of the HMP2 pipelines. Thus, a series of preprocessing steps were applied by the research group leading the project before making the data available publicly. Raw data files are also accessible through the same repository for those interested in manipulating sequences with the complete process (e.g. QIIME), although it was not handled in the present study as bioinformatic analysis was not the focus and scope of this master thesis. For additional information on clinical, sample handling and data generation protocols, the following link can be consulted https://ibdmdb.org/cb/browser/.

As introduced in Section 2.1.1, the choice of sequencing technology will have certain influence in data composition, characteristics, and quality. The type, platform and source of high-throughput technologies used in HMP2 IBD study are described in Table 4.

| Type | Source | Platform | Number of Oligos/SNPs | SNP Batch Id |
|---|---|---|---|---|
| Whole exome sequencing | Illumina | HGSC VCRome 2.1 design (42Mb, NimbleGen, hg19) | N/A | N/A |
| Whole transcriptome shotgun sequencing (RNA-seq) | Illumina | HiSeq | N/A | N/A |
| Methylome sequencing | Illumina | HiSeq | N/A | N/A |
| Amplicon sequencing | Illumina | HiSeq | N/A | N/A |

*Table 4. Molecular data.*

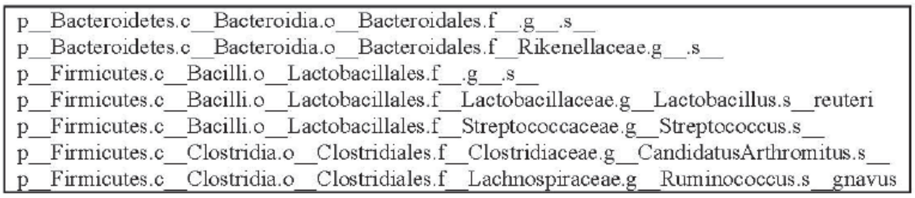
Selecting the dataset to be used for our implementation and proof of concept was fairly easy, as limited number of studies were eligible given our selection criteria: longitudinal, multi-omic, human and disease related data. However, considerable amount of time was dedicated to understanding and breaking down data files for downstream analysis. We elaborated a comprehensive summary table (Table 5) in aim of helping future researchers interested in using the same dataset accelerate their data inspection phase. The enormous dimensions of some files support the advantage of using AI algorithms versus traditional biological statistical analysis to efficiently extract full knowledge from data.

| | File name | File description | Dimension |
|---|---|---|---|
| **Metadata** | `hmp2_metadata.csv` | Full sample metadata table. Samples as rows and metadata as columns | 178×490 |
| **16S** | `taxonomic_profiles.tsv` | Biopsy 16S data. Contains OTUs IDs | 982×178 |
| **Metabolomics** | `iHMP_metabolomics.csv` | Metabolomics profiles | 81867×553 |
| **Metagenomics** | `ecs_relab.tsv` | MGX EC abundances with stratification | |
| | `taxonomic_profiles.tsv` | MetaPhlAn2 taxonomic profiles | 1479×1639 |
| | `pathabundance_relab.tsv` | MGX pathway abundances with stratification | 10884×1639 |
| | `species_counts_table.tsv` | Species count for each sample (total and after filter) | 1300×2 |

| | | | |
|---|---|---|---|
| | hmp2_mgx_taxonomy.tsv | Taxonomic profiles (pilot study) | 932×1639 |
| | hmp2_mgx_pathabundance.tsv | MGX pathway abundances without stratification | 22113×1639 |
| | ecs_3.tsv | MGX EC abundances without stratification | 108433×1639 |
| **Metatranscriptomics** | pathabundance_relab.tsv | MTX pathway abundances with stratification | 6061×736 |
| | hmp2_mtx_pathabundance.tsv | MTX pathway abundances (pilot study) | 8562×818 |
| | genefamilies.tsv | Gene families annotated using UniRef database | 2164739×736 |
| | ecs_3.tsv | MTX EC abundances without stratification | 83226×818 |
| | ecs_relab.tsv | MTX EC abundances with stratification | 70711×736 |

*Table 5: Structure of data files in the Human Microbiome Project II - IBD (Lloyd-Price, et al., 2019) study. Green rows highlight the files used in the present study for subsequent analysis.*

Another example of non-intuitive information that had to be carefully analyzed and deduced in order to fully understand was the choice of nomenclature (Figure 8).



```
p__Bacteroidetes.c__Bacteroidia.o__Bacteroidales.f__.g__.s__
p__Bacteroidetes.c__Bacteroidia.o__Bacteroidales.f__Rikenellaceae.g__.s__
p__Firmicutes.c__Bacilli.o__Lactobacillales.f__.g__.s__
p__Firmicutes.c__Bacilli.o__Lactobacillales.f__Lactobacillaceae.g__Lactobacillus.s__reuteri
p__Firmicutes.c__Bacilli.o__Lactobacillales.f__Streptococcaceae.g__Streptococcus.s__
p__Firmicutes.c__Clostridia.o__Clostridiales.f__Clostridiaceae.g__CandidatusArthromitus.s__
p__Firmicutes.c__Clostridia.o__Clostridiales.f__Lachnospiraceae.g__Ruminococcus.s__gnavus
```

*Figure 8: Example of nomenclature for taxa. P, phylum; c, class; o, order; f, family; g, genus; s, species. Taxonomic levels that lack information (e.g., f___.g____.s____) did not match named taxa present in the GreenGene database (http://greengenes.lbl.gov) a 16S rRNA gene database.*

Through the whole analysis we work with a column named "Participant ID" in order to identify subjects. The chosen nomenclature consisted of a capital letter in first place followed by four integers as seen in the following examples:

- **M**2014
- **P**6005
- **C**3001
- **H**4001
- **E**5009

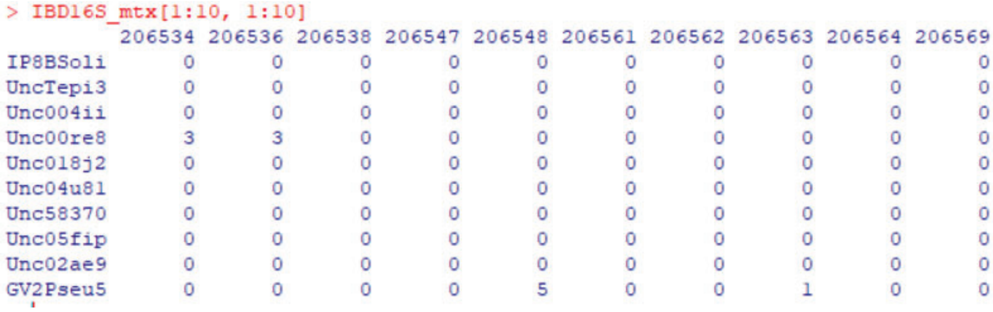
Capital letter correspond to the site where the subject was recruited: M for Massachusetts General Hospital (38 participants), P for MGH Pediatrics (17 participants), C for Cedars-Sinai Medical Center (33 participants), H for Cincinnati Children's Hospital (33 participants) and E for Emory University Hospital (11 participants).

We filtered the complete dataset in order for subjects to meet certain inclusion criteria. The inclusion criteria used for our study in particular was to have a minimum of four measured timepoints for all three omic types: metagenomics, metatranscriptomics and metabolomics. We filtered the subjects of interest with the information given in the metadata set. These restrictions yielded a total of 93 subjects: 47 with Crohn disease, 23 with ulcerative colitis and 23 controls or non-IBD. Further on, during preprocessing stage and data inspection we had to remove an additional group of two subjects as metatranscriptomics path abundance dataset did not satisfy the threshold of four time points, contrary of what metadata described. Therefore, the data used for downstream analysis contained a total of 91 subjects. Full details can be consulted in Appendix II.

### 3.3.2 Data inspection with bioinformatic tools

`HMP2Data` is a Bioconductor package (in R) providing the data of the integrative Human Microbiome Project (iHMP), the second phase of HMP project (HMP2). It contains 16S rRNA sequencing data from all three longitudinal studies: 1) MOMS-PI, pregnancy and preterm birth; 2) IBD, gut disease onset, inflammatory bowel disease; and 3) T2D, onset of type 2 diabetes and respiratory viral infection. In a preliminary stage, for data inspection purposes, raw data files were downloaded from the HMP Data Analysis and Coordination Center. Processed data is provided as matrices, `SummarizedExperiment`, `MultiAssayExperiment`, and `phyloseq` class objects. A straightforward preliminary inspection workflow was executed in order to understand the data before constructing the ad hoc preprocessing script.

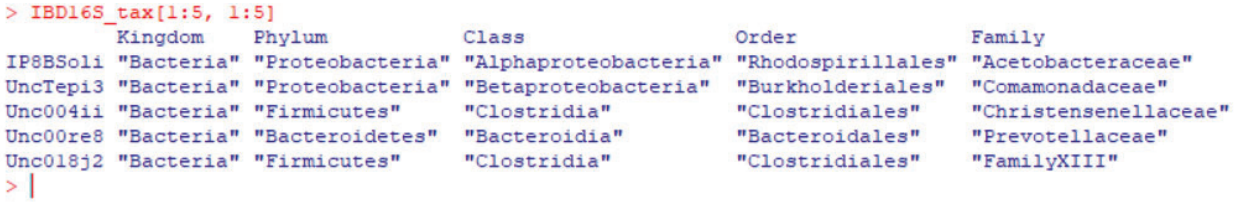
Step 1: Load 16S data as a matrix, rows are SILVA IDs, columns are sample names as can be seen in Figure 9.



```
> IBD16S_mtx[1:10, 1:10]
         206534 206536 206538 206547 206548 206561 206562 206563 206564 206569
IP8BSoli      0      0      0      0      0      0      0      0      0      0
UncTepi3      0      0      0      0      0      0      0      0      0      0
Unc004ii      0      0      0      0      0      0      0      0      0      0
Unc00re8      3      3      0      0      0      0      0      0      0      0
Unc018j2      0      0      0      0      0      0      0      0      0      0
Unc04u81      0      0      0      0      0      0      0      0      0      0
Unc58370      0      0      0      0      0      0      0      0      0      0
Unc05fip      0      0      0      0      0      0      0      0      0      0
Unc02ae9      0      0      0      0      0      0      0      0      0      0
GV2Pseu5      0      0      0      0      5      0      0      1      0      0
```

*Figure 9: 16S data matrix.*

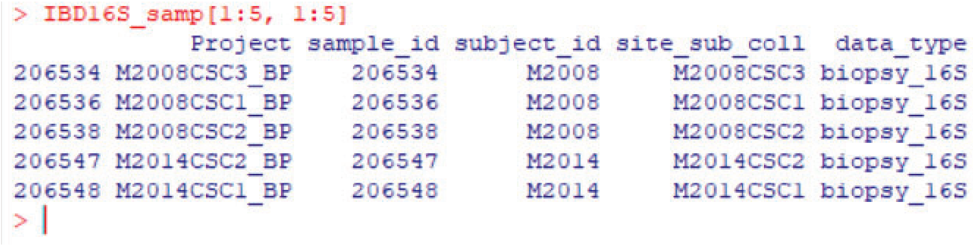
Step 2: Load the SILVA taxonomy table as a matrix, rows are SILVA IDs, columns are taxonomic ranks as seen in Figure 10.



```
> IBD16S_tax[1:5, 1:5]
         Kingdom    Phylum           Class                  Order              Family
IP8BSoli "Bacteria" "Proteobacteria" "Alphaproteobacteria" "Rhodospirillales" "Acetobacteraceae"
UncTepi3 "Bacteria" "Proteobacteria" "Betaproteobacteria"  "Burkholderiales"  "Comamonadaceae"
Unc004ii "Bacteria" "Firmicutes"     "Clostridia"          "Clostridiales"    "Christensenellaceae"
Unc00re8 "Bacteria" "Bacteroidetes"  "Bacteroidia"         "Bacteroidales"    "Prevotellaceae"
Unc018j2 "Bacteria" "Firmicutes"     "Clostridia"          "Clostridiales"    "FamilyXIII"
> |
```

*Figure 10: Taxonomy table.*

Step 3: Load the 16S sample annotation data as a matrix, rows are samples, columns are annotations as seen in Figure 11.



```
> IBD16S_samp[1:5, 1:5]
                 Project sample_id subject_id site_sub_coll  data_type
206534 M2008CSC3_BP    206534      M2008      M2008CSC3 biopsy_16S
206536 M2008CSC1_BP    206536      M2008      M2008CSC1 biopsy_16S
206538 M2008CSC2_BP    206538      M2008      M2008CSC2 biopsy_16S
206547 M2014CSC2_BP    206547      M2014      M2014CSC2 biopsy_16S
206548 M2014CSC1_BP    206548      M2014      M2014CSC1 biopsy_16S
> |
```

*Figure 11: Metadata file with clinical information.*

Table 6 shows the summarized information of dimensions of the files that were analyzed.

| | Rows | Columns |
|---|---:|---:|
| OTU table | 982 taxa | 178 samples |
| Sample data | 178 samples | 490 sample variables |
| Taxonomy table | 982 taxa | 6 taxonomic ranks |

*Table 6: Summary table for data visual inspection analysis output.*

As seen in Table 6, only three type of data files are integrated in the current version of the `HMP2Data` package, presenting a major drawback for our desired application, thus we decided to discard its use for this project. However, exploring these files gave us a broad initial idea of what our data looked like, and we highly recommend researches unfamiliar with microbiome data to look into it as an initial step due to its accessibility and friendly use.

We computed the corresponding histogram for each data attribute in order to visualize the distribution of sample depths (16S taxonomic profiles) (Figure 12) and OTU frequencies (Figure 13).

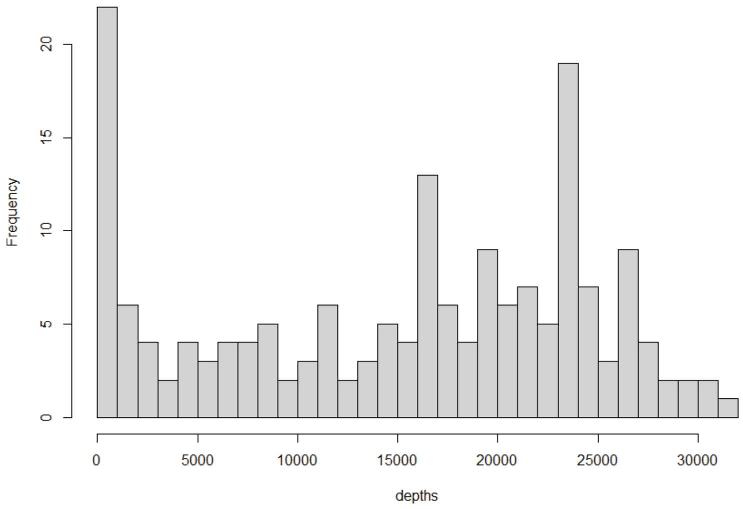


*Figure 12: Histogram of depths.*

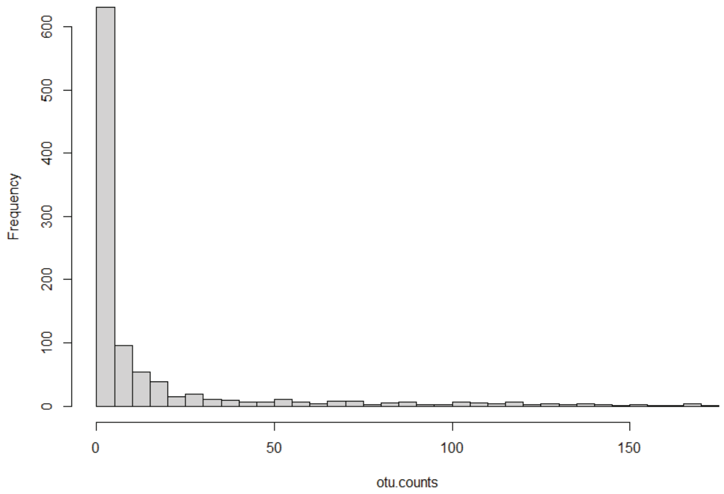


*Figure 13: Histogram of OTU counts.*

Species distribution typically follow a negative binomial: many subjects have small relative abundance of that species, and a very few subjects exist with a high relative abundance of a specific species. For visualization purposes as well as downstream analysis we know remove OTUs present in <10% of samples. Figure 14 shows the resulting histogram.

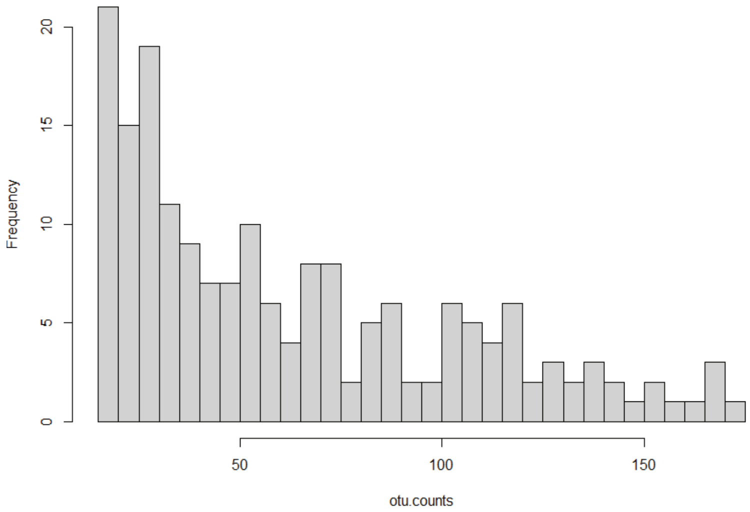


*Figure 14: Histogram of OTU counts after filtering OTUs present in <10% of samples.*

After these first steps of data inspection one of the metrics we are interested in exploring when analyzing microbiome data or any other genomic related dataset is diversity. Two of the most important metrics for biological diversity are α diversity and β diversity (see Section 2.1.1).

α diversity measurement is constrained by the sequencing depth (total number of reads per sample). Rarefying (e.g., through vegan R package: rarefy), selecting the appropriate sample depth, is necessary before calculating α diversity.

By computing α diversity (Figure 15) to study diversity within a sample, we could observe how dysbiosis states (UC and CD) manifested an expected lower diversity measure compared to healthy state (Wright, et al., 2015) (Lloyd-Price, et al., 2019).

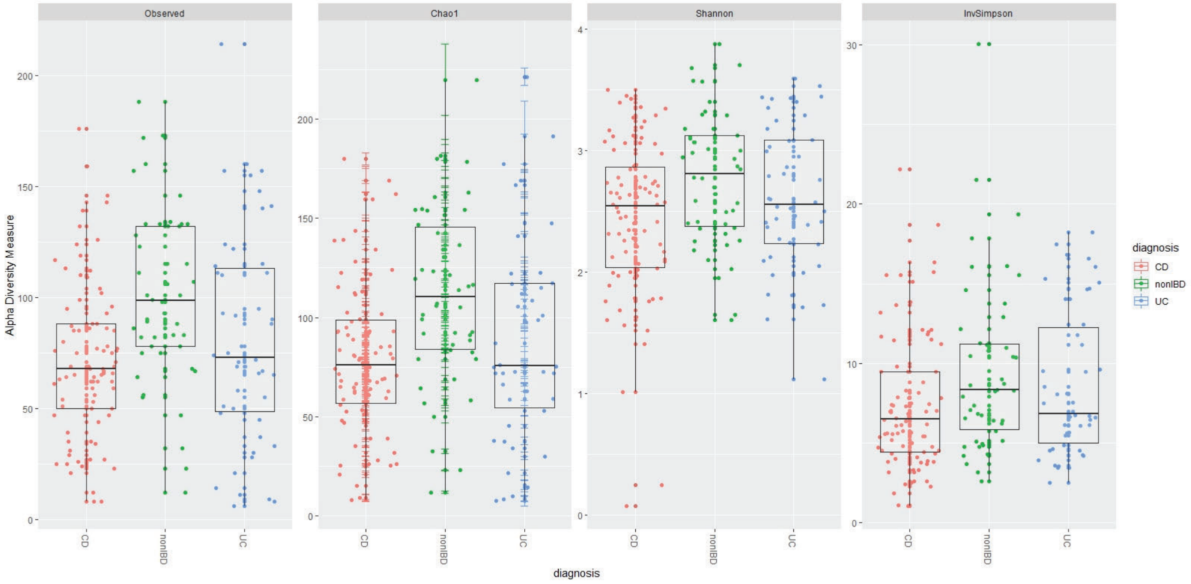


*Figure 15: Comparison of a diversity measured by observed species, Chao1 index, Shannon diversity and Simpson. CD cluster is shown in red, UC in blue and nonIBD subjects in green. Healthy control samples are significantly different from IBD samples. Shannon and Simpson indicate the uniformity of the abundance of different species in a sample.*

β diversity describes how samples vary against each other taking into account the whole distribution of species in a community. β diversity can help us discern between clusters (non IBD and IBD patients), for instance in our case, understand if sample A is more similar in composition to sample B (non IBD) or C (IBD). Results of comparing beta diversities from a qualitative and quantitative strategy are shown in Figure 16. We used both Jaccard and Bray-Curtis indexes, although similar results where reported, we wanted to analyze if there was an advantage in using one of them for our particular data. Commonly, Jaccard index is recommended when dealing with large spatial scales and datasets with presence/absence of data. Bray-Curtis on the other hand is preferred when taking into account abundances. β diversity analysis elucidate dissimilarities between samples (UC, CD and healthy).

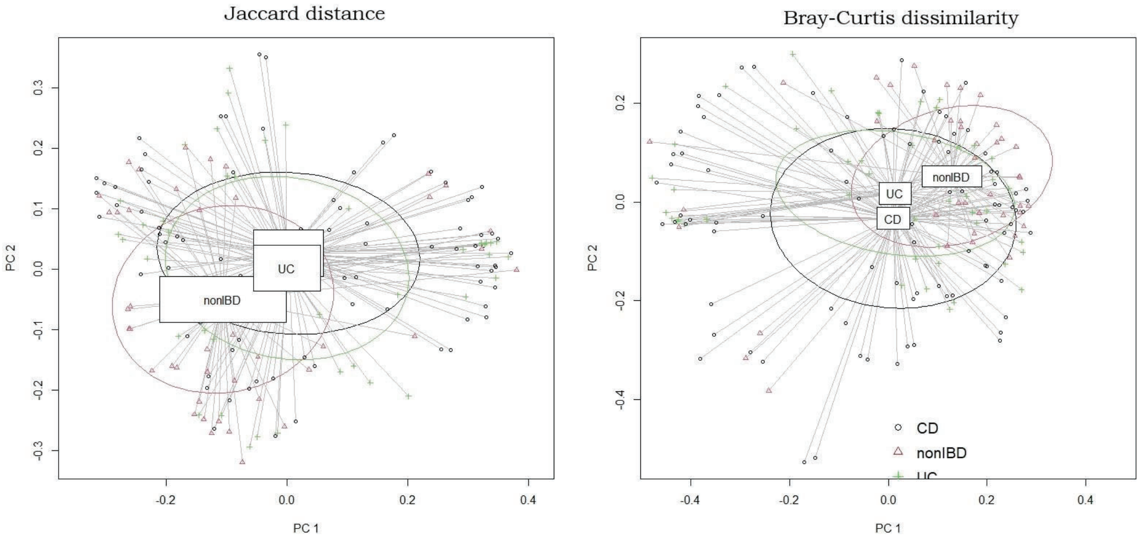


*Figure 16: Beta diversity for IBD dataset. Jaccard distance vs. Bray-Curtis dissimilarity.*

These kinds of metrics are useful to build a preliminary idea of our data. However, given the complexity and diversity of microbiome data, further computational tools and analysis need to be applied to fully understand our data.

### 3.3.3 Preprocessing

This stage involves interpretation of the data format, the definition of the data structures for their management and correction tasks including elimination of noise and errors among others.

Figure 17 sums up the necessary previous steps (preprocessing workflow) followed to obtain the dataset prior to the analysis in the preprocessing stage. For more information of each step, the following links can be consulted:

- Center for Metagenomics and Microbiome Research (CMMR): link.
- UPARSE (Edgar, 2013): https://drive5.com/uparse/

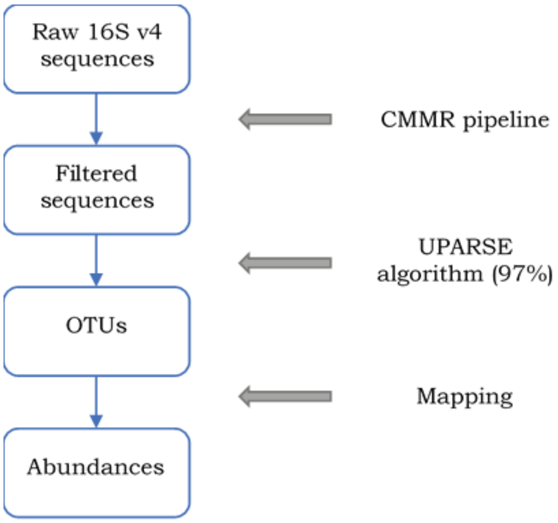


*Figure 17: IBDMDB protocol followed for data preprocessing.*

a) Raw sequences data file

The starting point for microbiome data analysis is the “raw data file” quality check and processing. Every sequencing apparatus and experiment protocol will provide different characteristics and measurements in the raw data file, and it is vital to understand the type of raw data to be analyzed before any downstream analysis is performed.

In our case, for quality control processing, metagenomic and metatranscriptomic data samples were run through KneadData which first trimmed reads to preserve high quality sequence and removed any adapter contaminants. Surviving sequences were filtered to remove any human contaminants using the human genome (hg38). Data are organized by paired and orphan reads. When one read in a pair passes a filtering step and the other does not the surviving read is an orphan.

b) Taxonomic profiling

One common adopted approach is to collapse OTUs to genus or higher taxonomic levels. In general, two approaches can be applied depending on the data units we have: (i) *clustering* if we deal with OTUs; (ii) *denoising* if we deal with amplicon sequence variants (ASVs).

Taxonomic profiles of metagenomic data were generated using MetaPhlAn2. MetaPhlAn2 is a taxonomic classification computational tool for metagenomic phylogenetic analysis. It allows profiling the composition of a microbial community from metagenomic sequencing data by assigning DNA sequence to its microbial species of origin (taxa). MetaPhlAn2 generates:

- Species-level taxonomic profiles expressed as relative abundance from kingdom to strain level.
- Presence of unique, clade-specific markers.
- Abundance of unique, clade-specific markers.

A common step in microbiome data preprocessing is the low count removal. Species abundances are passed through a basic filter requiring each species to have at least 0.01% abundance. This step aims to counteract sequencing errors. A total of 540 species were identified. After basic filtering 109 species remained.

Figure 18 shows a hierarchical clustering of samples and species, using top 25 species with highest mean relative abundance among samples. Abundances were log10 transformed prior to clustering, and the “average linkage” clustering on the Euclidean distance metric was used to cluster samples. The species dendrogram is based on pairwise (Spearman) correlation between species. Samples are columns and species are rows. The color bar represents relative abundances on a log10 scale. The use of double dendrogram colored heatmaps is ubiquitous in the microbiome literature.

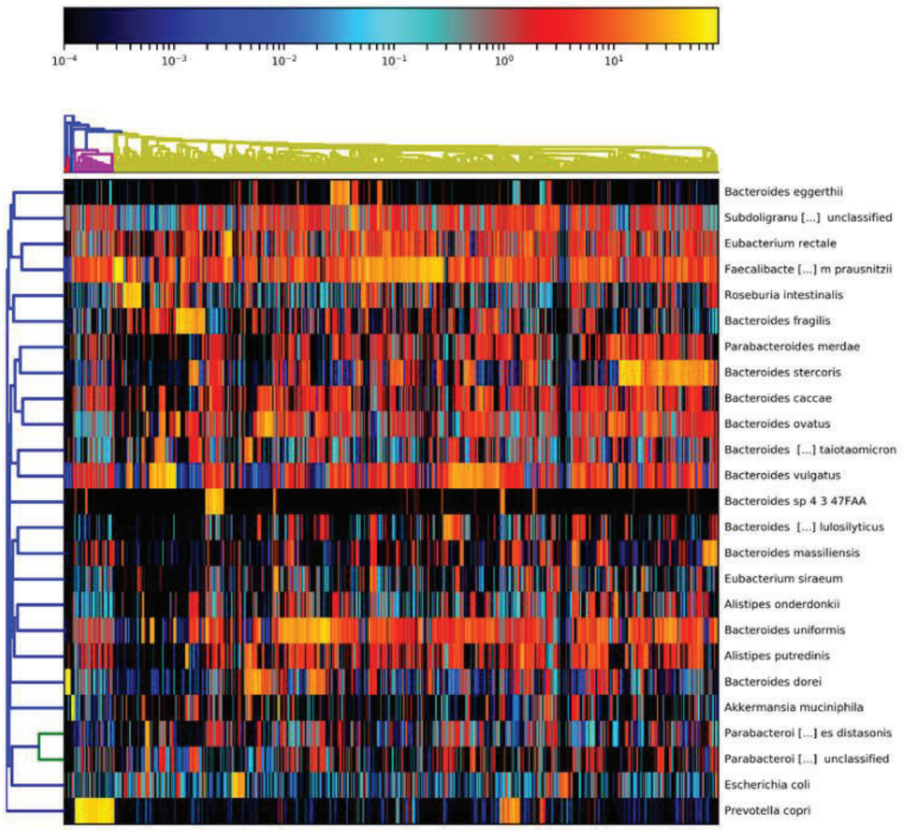


*Figure 18: Top 25 species based on average relative abundances. Heatmap was generated using Hclust2 (https://ibdmdb.org/) .*

Figure 19 presents a stacked barpot of the 15 most abundant species among samples. This information can be interesting for validation of our final results in order to understand if our model is capturing true interactions of the studied microbial community.

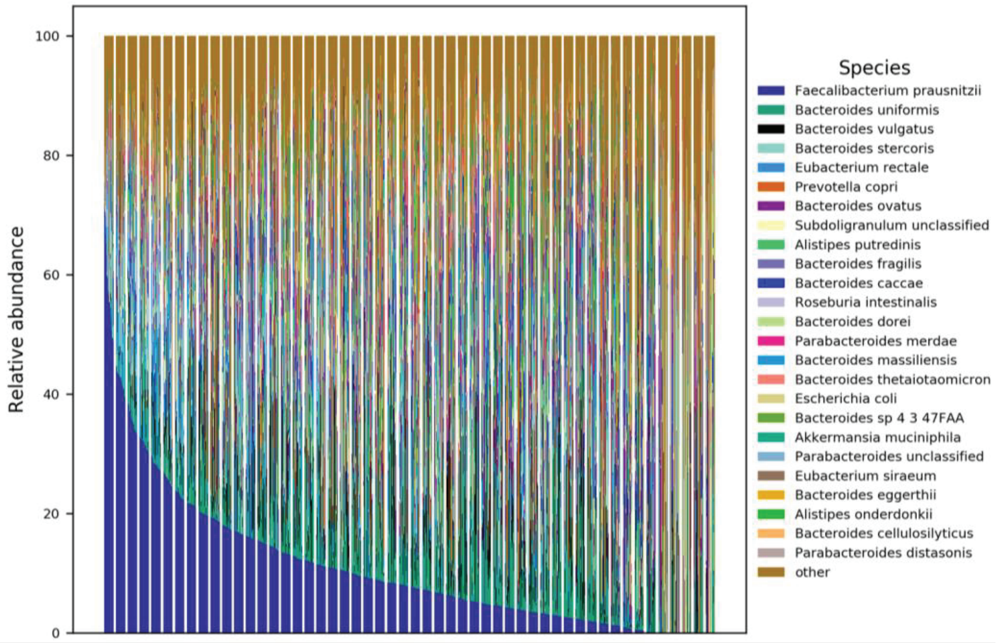


*Figure 19: Top 15 species by average abundances. Stacked barplot of 15 most abundant species among samples. (https://ibdmdb.org/).*

c) Functional profiling

HUMAnN2 is a pipeline for functional profiling which comprises profiling the presence or absence and abundance of microbial genes and pathways in a community from metagenomic or metatranscriptomic sequencing data using the UniRef and MetaCyc databases. With this analysis we are able to describe the metabolic potential of a microbial community. Thanks to function profiling we are able to answer to the question “What are the microbes in the community of interest doing or capable of doing?”

HUMAnN2 generates:

- Abundance of gene families
- Metabolic pathway coverage
- Metabolic pathway abundance
- Enzyme commission (EC) enzyme modules

For metatranscriptomics pathway abundances in IBDMDB, hierarchical clustering was executed using top 50 pathways with highest mean relative abundance among samples. The “average linkage” clustering on the Euclidean distance metric was used to cluster samples. The pathway dendrogram is based on pairwise (Spearman) correlation between pathways. Samples are columns and pathway are rows. The heatmaps were generated with Hclust2. The most abundant DNA features are not necessarily those with the highest transcription (RNA) levels. From the dendrogram of the left side of Figure 20 we can see which RNA pathways are more correlated to each other.

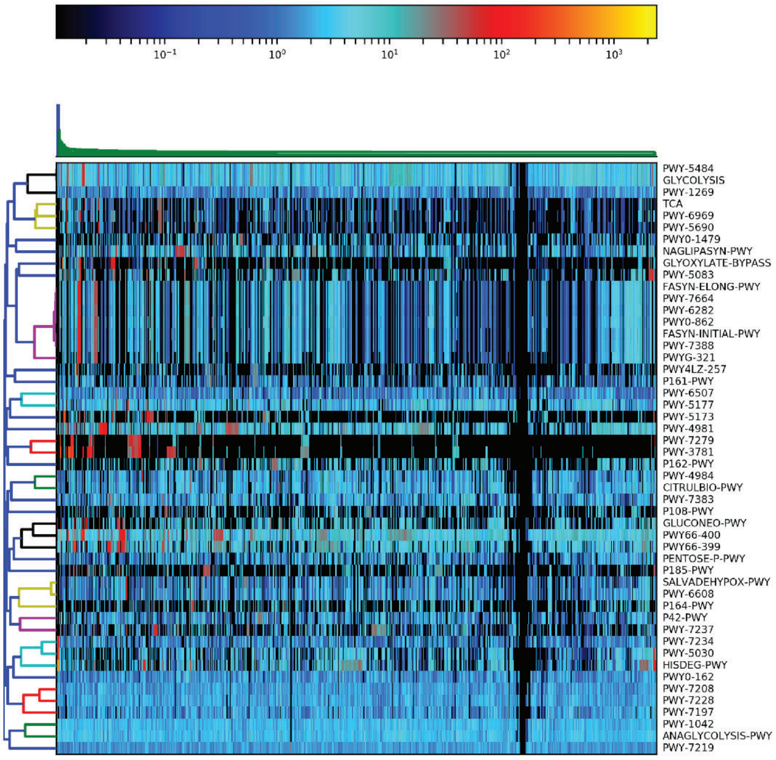


*Figure 20: Top 50 pathways by average abundance (log10). Color bar represents relative abundances on a log10 scale. Abundances were log10 transformed prior to clustering (https://ibdmdb.org/). A dendrogram is added on the side that is created with hierarchical clustering.*

d) Dimensional reduction

Principal coordinate analysis (PCoA) using distance measures of Bray-Curtis dissimilarity is useful for microbiome data. PCoA can be applied to visualize high-dimensional microbiome data patterns. PCoA performs a rotation of the inter-sample distance matrix (after centering) to represent those distances as accurately as possible in a small number of dimensions. For the PCoA plot, relative abundances are passed through a basic filter requiring each terminal taxa to have at least 0.01 % abundance in at least 10 % of all samples.

Figure 21 show PCoA of variance among samples, based on Bray-Curtis dissimilarities between species profiles of samples. Numbers in parenthesis on each axis represent the amount of variance explained by that axis. Distances quantify the similarity in terms of taxonomic species abundances.

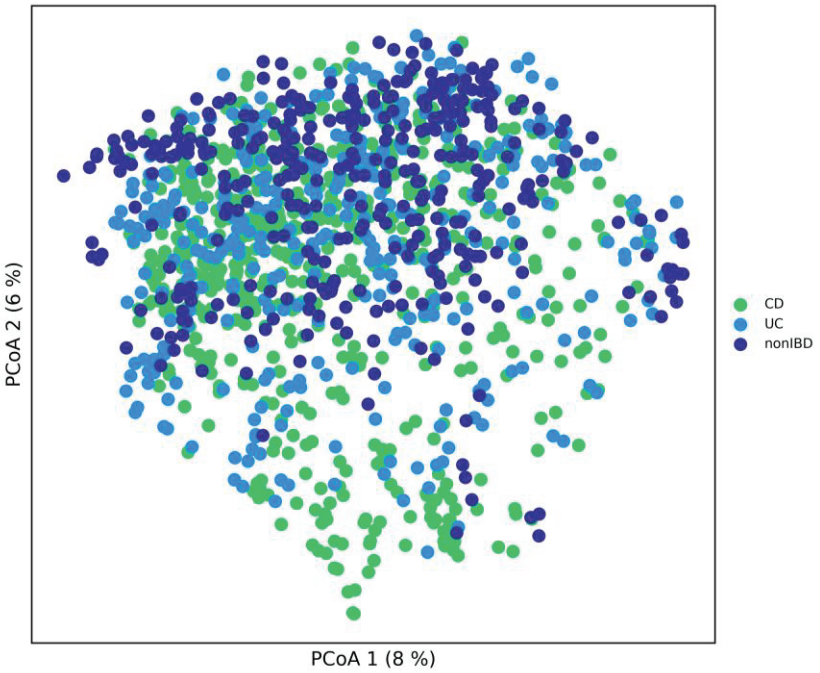


*Figure 21: Ordination of species abundances. Principal coordinate analysis of variance among samples, based on Bray-Curtis dissimilarities between species profiles of samples. Numbers in parentheses on each axis represent the amount of variance explained by that axis (https://ibdmdb.org/).*

After the first tasks of data exploration through specific bioinformatics tools and popular statistical methods used in microbiome research, we developed our own Python preprocessing script to prepare the dataset that will be used in further analysis as the input to our model. The resulting ad hoc scripts can be consulted through the following link https://github.com/muia2021pl/TFM_microbiome.

We first loaded the corresponding original datasets in .csv and .tsv formats with the information corresponding to the whole genome shotgun sequencing. A total of five datasets are imported: metadata with clinical variables, metabolomics, metagenomics with taxonomic profiles abundances and metatranscriptomics with path abundances (see Table 5). In the latter case, the information is divided in two different datasets: HMP2 and HMP2 pilot. Data matrix (tables) we will be preprocessing are expressed as abundances.

Next, a series of preprocessing steps are performed on each dataset separately. Tasks involved removal of subjects with limited measured time points (threshold set at minimum of four time-points across three omics), rearrange indexes and columns, adding data type identifiers to each variable (column) or removing unnecessary columns for the problem in scope. In this preprocessing stage, log transformation and normalization were applied to the data (for continuous variables). To be able to apply ML techniques, we might need to transpose the data set. This is since biological data has the predictor variables as rows (instead of columns).

### Interpolation

At this point we considered applying some approach to overcome irregular sampling. Spline modelization (Interpolation) is a popular approach that uses continuous curves to interpolate time points that might be missing or are inconsistent between different types of data (Bodein, Chapleur, Droit, & Lê Cao, 2019). Some proposed approaches include smoothing spline ANOVA (Paulson, Talukder, & Bravo, Longitudinal differential abundance analysis of microbial marker-gene surveys using smoothing splines, 2017), negative binomial smoothing splines (Metwally, et al., 2018) or Gaussian cubic splines (Luo, Ziebell, & An, 2017). One limitation that these approaches share is the fact that they are univariate and cannot infer ecological interactions. Although we explored the possibility of including these technique in our preprocessing framework, we finally decided not to implement it as further research must be done in order to appropriately exploit its potential. More will be discussed in Section 5.2.

### Normalization

As presented in Section 2.3, we do not have equally spaced time points in our data. Moreover, each omics technology produces count or abundance tables with samples in rows and features in columns (genes, proteins, species, ...) and each data type has a variable number of columns depending on the technology and number of identified features.

Consequently in every analysis in microbiome research (as in many other fields) one of the first things to be done in the preprocessing stage is data normalization and transformation. Microbiome data are compositional, because of technical, biological, and computational reasons, thus interpreted into relative counts. Taxa abundance needs to be adjusted for compositionality. In case normalization needs to be applied to the data, the use of log-ratios transformation is recommended for microbial taxa data normalization (Gloor, Macklaim, Pawlowsky-Glahn, & Egozcue, 2017), (Mars, et al., 2020). There are a number of standard normalization methods used in the literature with the same final goal: removing technical bias in compositional data (Paulson, Stine, Bravo, & Pop, 2013), (Badri, Kurtz, Müller, & Bonneau, 2018).

*(1) Centered log-ratio (CLR) Transformation* introduced by (Aitchison, 1982) converts the relative abundances (or OUT counts) to ratios between all parts by calculating the geometric mean of all values (whole composition). This was the method of choice for our dataset, implemented in the ad-hoc preprocessing Python script. $x_1$

It is defined for a composition $x$ as follows:

$$clr\,(x_j) = \left[\ \ln \frac{x_{1,j}}{g(x_j)}, \ldots, \ln \frac{x_{D,j}}{g(x_j)}\ \right]$$

where $x_j$ is the j-th sample and $g_m(x_j) = (\prod_{i=1}^{D} x_i)^{1/D}$ is the geometric mean (row-wise) of vector $x$.

*(2) Total sum scaling (TSS)* divides each individual feature count with the total library size (OUT counts in a sample) to yield the relative proportion of counts for that feature. The total sum sums up to 1.

*(3) Cumulative Sum Scaling (CSS)* calculates a scaling factor as the quantile of the count distribution of samples assuming that at this range, counts are derived from a common distribution. It was developed by (Paulson, Stine, Bravo, & Pop, 2013) in addition to `metagenomeSeq` Bioconductor package for differential analysis implementation.

It is worth mentioning other methods that can yield correct performance such as relative log expression (RLE) proposed by (Anders & Huber, 2010), trimmed mean of M-values (TMM) proposed by (Robinson & Oshlack, 2010) or upper quantile (Bullard, Purdom, Hansen, & Dudoit, 2010).

**Other preprocessing steps**

Other useful preprocessing steps could be dealing with missing values. Technical problems during experiments commonly occur and result in missing values in the data set. Important to note that removing samples (rows) or variables (columns) can have limitations such as reducing knowledge (data size) and prediction power. Several longitudinal studies have explored additional complementary preprocessing steps such as interpolation and detrending with the aim of making time points equidistant and comparable. These last approaches were not tackled in our study but will be discussed in Section 5 as future lines of research.

### 3.3.4 Features

The feature sets, grouped by type, included in this work are:

Clinical features (metadata)
- Variables to identify subjects (e.g., "Subject.ID")
- Variables to identify time steps for sample time series ("week")

- Variables to indicate the phenotype/cluster of each sample (named as "diagnosis").
- Variables to indicate external perturbations ("antibiotic")

Metagenomic features

- Taxonomic profiles. Type: continuous. Taxonomic features correspond to the relative abundance in percentage or counts per million.

Metatranscriptomic features

- Relative abundance of each metabolic pathway. Type: continuous.

Metabolomics features

- Metabolic concentrations. Type: continuous.

One major challenge with integrating multi-omics is that combining different types of biological information increases the number of analyzed features while keeping the number of observations/samples (subjects) constant (Figure 22). Feature selection can therefore improve prediction accuracy of our model.

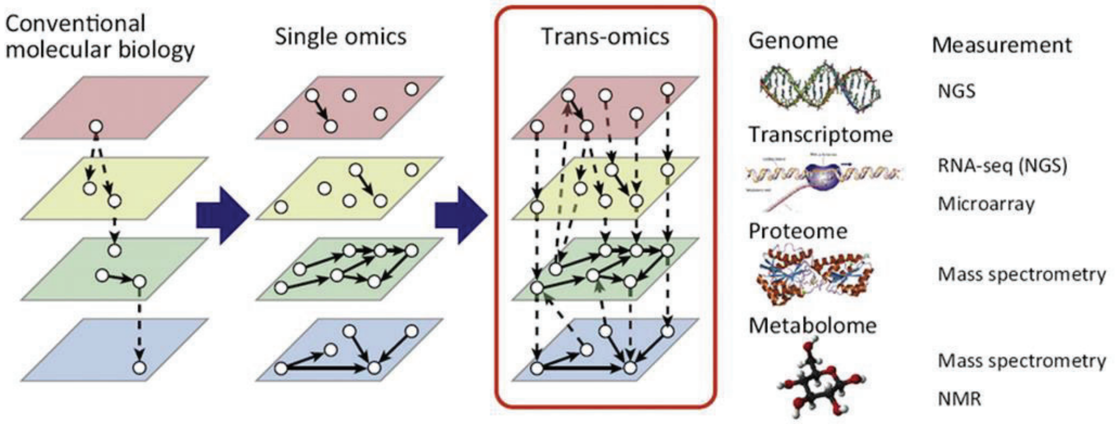
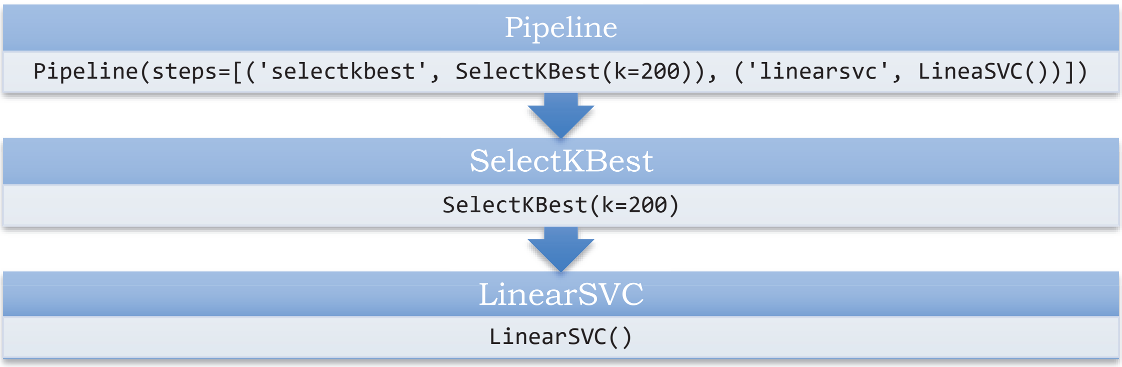


*Figure 22: Multi-omics integration (Yugi, Kubota, Hatano, & Kuroda, 2016).*

Feature subset selection techniques are an essential step in every ML analysis pipeline. Some of the advantages it offers involve reducing overfitting, making interpretability easier, improving accuracy by reducing misleading or noisy data and reducing training time, as less input data will accelerate training times.
As a preprocessing dimensionality reduction step, we first filtered predictor variables with near zero variance using `VarianceThreshold()` class from the `sklearn.feature_selection` module.

In our case, features with a training-set variance equal to the zero threshold will be removed. The default is to keep all features with non-zero variance, i.e. remove the features that have the same value in all samples.

For simplicity in our case, we focus on univariate feature selection filter algorithms. Common univariate feature selection works by selecting the best features based on univariate statistical tests. These tests are easy and simple to use but they do not account for intercorrelation among features. We applied Select-K-Best to select the k most important features with the highest scores based in Chi-squared statistics. As we are using sparse data, chi2 and mutual information `scikit-learn` Python implementation offer a good solution as they will deal with the data without making it dense (processing sparse matrices without casting them internally to dense numpy arrays). Next step, we used 10-fold cross-validation linear support vector classifier as the external estimator implemented with Python `scikit-learn` package. We chose this particular variation of support vector machines as it is recommended for optimal performance with sparse data and handles multiclass. At last, we further discard features using the Bayes factor as explained in subsequent Section 3.4.1.



```
Pipeline(steps=[('selectkbest', SelectKBest(k=200)), ('linearsvc', LineaSVC())])
SelectKBest(k=200)
LinearSVC()
```

For performance assessment purposes, univariate Chi2, ANOVA and mutual information will be tested and compared to perform feature selection on the full dataset for each omic type. The result is a set of informative features that can be utilized for downstream ML analysis, Table 7.

| **Data** | **FS technique** | **Eval. Metric**<br>(Accuracy) |
|---|---|---|
| Metagenomics | Univariate Chi2 | 0.82 |
| | Univariate ANOVA | 0.78 |
| | Univariate MI | 0.73 |
| Metabolomics | Univariate Chi2 | 0.67 |
| | Univariate ANOVA | 0.69 |
| | Univariate MI | 0.69 |
| Metatranscriptomics | Univariate Chi2 | 0.55 |
| | Univariate ANOVA | 0.56 |
| | Univariate MI | 0.50 |

*Table 7: Classification report of univariate feature selection techniques. Results shown correspond to K=200 best features. When K=100 the accuracy decreased.*

## 3.4 Implementation

Using the preprocessed multi-omics dataset our next step was to learn a graphical structure from temporal data collected from a dynamic system.

### 3.4.1 Dynamic Bayesian network

In order to select a software package to implement the DBN, we first evaluated and compared several options, see Table 8.

| Name | Language | Data type | Learning | Inference |
|---|---|---|---|---|
| *CGBayesNets*<br>(McGeachie, Chang, & Weiss, 2014) | Matlab | Discrete and continuous | Yes | Yes |
| *dbnR*<br>https://github.com/dkesada/dbnR | R | Continuous data (only) | Yes | Yes |
| *Bnlearn*<br>(Scutari, 2010) | R | Continuous and discrete | Yes | No |
| *Bnfinder*<br>(Wilczyński & Dojer, 2009) | Python | Discrete and continuous* | Yes | Yes |

*Table 8: Software packages benchmarking. (*only if 1 discrete parent and no children).*

The requirements we needed the software tool to meet, and support consisted of learning and inference of DBNs in the presence of both discrete and continuous data. As seen from Table 8 the options are quite limited and further implementations and capabilities extensions should be explored. However, these developments were out of the scope for the present work.

CGBayesNets (McGeachie, Chang, & Weiss, 2014) builds a two-stage DBN of the microbiome population dynamics. It considers current time samples and the immediate previous ones. It performs inference with mixed continuous and discrete networks as a CGBN; while other packages do not. CGBayesNets uses Bayesian marginal likelihood to guide network search for inference.  It also provides functions for employing cross-validation (CV) and bootstrapping for model performance and verification. CGBayesNets could be used with the ultimate goal of finding a network predictive of the phenotype (case/control status). Still, one limitation of this package is its inability to support the use of intra-edges. For this reason, we used the modified version of CGBayesNets implemented by (Lugo-Martinez, Ruiz-Perez, Narasimhan, & Bar-Joseph, 2019) where intra-edges are allowed and BIC and AIC networks scoring functions are included.

dbnR package is an alternative good option. It covers learning and doing inference (forecast in the future) over Gaussian DBNs of arbitrary Markovian order. It extends some of the functionality offered by the 'bnlearn' package to learn the networks from data and perform exact inference. It offers two structure learning algorithms for DBNs and the possibility to perform forecasts of arbitrary length. A tool for visualizing the structure of the net is also provided via the 'visNetwork' package. The only drawback with this package is the fact that it does not support discrete variables. In our case and in many other microbiome studies, clinical variables (discrete and continuous) bring valuable information to the model and have a key role in the analysis. A solution the author provides in order to deal with discrete data is to perform clustering on our data based on our discrete variables (clinical metadata in our case) and next train a continuous network model for each cluster.

bnlearn package in R performs Bayesian network structure learning and parameter learning. This package implements constraint-based (e.g., PC, GS, IAMB, Inter-IAMB…etc.), pairwise (ARACNE (Margolin, et al., 2006)), score-based (e.g., Hill-Climbing) and hybrid (MMHC (Tsamardinos, Brown, & Aliferis, 2006), hybrid HPC (Gasse, Aussem, & Haytham, 2012)) structure learning algorithms for discrete, Gaussian, and conditional Gaussian networks, along with many score functions and conditional independence tests. In order to implement simulated dynamic functionality (not supported by the package) we could create a blacklist with restricted edges, in order to prohibit backward edges in time. Unfortunately, this solution will notably increase running time and computational resources. Lastly, bnlearn has the additional limitation of not implementing inference.

BNfinder can also be used to infer DBN from time series data. It performs structure learning using two scoring criteria: Bayesian-Dirichlet equivalence (BDe) (Heckerman, Geiger, & Chickering, 1995) and minimal description length (MDL) (Rissanen, 1978), (Grünwald, 2007). These scores, although designed for discrete variables, are used in this implementation to handle continuous variables under the assumption that conditional distributions belong to a family of Gaussian mixtures (one discrete parent and zero children) (McGeachie, Chang, & Weiss, 2014).

## I. Learning the structure of a DBN

First, the network structure is learned from the dataset. We need to set some parameters for the learning algorithm. Prior assumed distributions for each node are needed to determine the posterior probability of the data.

- Prior equivalent sample size $\nu = 10$.
- Prior assumed standard deviation: $\sigma^2 = 1$
- Maximum number of parents =3.

As filtering strategy, to prune dataset and reduce number of variables, we implement the Bayes factor of association with the phenotype (i.e. disease). Bayes factors can be computed for the dependence of each variable with the phenotype variable. It will help us determine the strength of association a variable has with the phenotype of interest. The Bayes factor is a Bayesian likelihood ratio test that computes the ratio of posterior probabilities of two quantities: 1) the probability of the variable being statistically dependent upon the phenotype, and 2) the probability of that variable being independent of the phenotype, both given in log scale. For values > Bayes factor, the variable is more likely to be associated with the phenotype than not. This is suitable for filtering for domains with too many variables to be considered by usual Bayes network methods. Bayes factor reduces the dataset down to a manageable number of informative variables by limiting further investigation to variables with log Bayes factor surpassing a predetermined threshold (in our case 5, 10 and 15).

CGBayesNets provides four types of network learning algorithms: (i) K2-style search (Cooper & Herskovits, 1992); (ii) greedy, exhaustive, hill-climbing search (every step adds arc that increases likelihood the most); (iii) pheno-centric search (Chang & McGeachie, 2011); (iv) simulated annealing (Kirkpatrick, Gelatt Jr, & Vecchi, 1983). Theoretical foundation of CGBayesNet can be consulted in Appendix I.

The main framework for learning DBNs consists of the following steps 1) combine time-series data into a larger column matrix with each time point matrix below the prior time point matrix, 2) learn the BN using StateTrackerSearch() function with dynamic Bayes net option enabled to allow cycles, allow self-loops, 3) unroll BN into a 2TBN: a 2-timepoint BN; all arcs are from time point one to time point two, 4) unroll dataset from timeseries matrix, 5) use normal techniques to predict with unrolled 2TBN.

MakeTSBNData(), assembles a 2-stage DBN dataset from times series data. It takes input data and arranges it, so the first time a subject id is encountered, its slotted into the $T_0$ data. The second time it is encountered, its slotted into both the $T_1$ and the $T_0$ data. The last time it is encountered, it is only slotted into the $T_T$ data.

Bootstrapping functionality is also implemented in the software that can be used to compare the performance of networks formed by starting with the phenotype node ('diagnosis') and then adding, in sequence, the most frequent edge occurring in the bootstrap networks and measuring the performance of that network on the dataset in cross-validation. Among models with equal or similar performance, we should opt for the most parsimonious model.

FullBNLearn() performs an 'exhaustive' search through possible arcs using a hill-climbing algorithm to learn a CGBN on the data. Though the author refers to it as an exhaustive search, it is important to note, it does not consider all possible networks, but rather all possible legal arcs between any two nodes. Bayesian Dirichlet equivalent sample-size uniform (BDeu) measure of marginal likelihood of the data (Heckerman, Geiger, & Chickering, 1995), is used as the network scoring metric.

It is important to note that we adapted original implementation scripts (code) (McGeachie, Chang, & Weiss, 2014) to serve our particular purposes as out data and final goal was different from similar studies that also used CGBayesNet (McGeachie, et al., 2016), (Lugo-Martinez, Ruiz-Perez, Narasimhan, & Bar-Joseph, 2019), (Ruiz-Perez, et al., 2021).

Furthermore, we performed DBN structure constraining by using an adjacency matrix as an input to the model. This matrix is configured in such a way that will only allow edges between specific nodes therefore reducing complexity and avoiding overfitting. The selected configuration was based on biological basic knowledge following (Ruiz-Perez, et al., 2021) model for reproducibility and comparison: clinical variables are independent, taxa is responsible for the expression of genes, and these genes are involved in metabolic pathways. In the same way, metabolites produced in $t_i$ will impact taxa abundance and growth in the next time slice $t_{i+1}$.

### II. Parameter learning

Once we have the structure of the DBN, we have to fold our dataset and fit the parameters of the DBN. This can be done in by calling the `LearnParams()` function in CGBayesNet to learn the marginal distributions of each node in the BN based on the data and the Bayesian priors.

As in (Lugo-Martinez, Ruiz-Perez, Narasimhan, & Bar-Joseph, 2019) and (Ruiz-Perez, et al., 2021), we maximized the likelihood of the data for a given structure using *maximum log-likelihood estimation* (MLE).

### III. Visualization

The model outputs both trivial Graph Format (*.tgf*) and GraphML (*.graphml*). For our implementation, we will use an output file GraphML version of the network, *output_file.graphml*, that can then be loaded into network visualization software such as Cytoscape (Shannon, et al., 2003). Additionally, we prepared an ad-hoc script in R to generate a custom style XML file for our output networks, encoding several properties of the underlying graph, such as node shape, arc line type and transparency of abundances to visualize in Cytoscape.

### IV. Inference and forecasting

Once we have fitted the model, we can perform inference over the learned model. When using BN, any variable can be used as the target node of the inference. Furthermore, in our particular case, DBNs, variables in the next time slices are predicted from the values in the previous slices.

CGBayesNet, implements the Cowell algorithm (Cowell, 2005) to perform inference in conditional linear Gaussian network nodes, as it is a numerically stable approach, combined with a simple variable elimination algorithm for inference between discrete nodes in the network (Koller & Friedman, 2009).

## 3.4.2 Random forest

To see whether BN approach generated a better performance in terms of predictive power of disease state, in comparison to the current gold standard for ML applications to microbiome data, a RF classification model was trained to classify disease outcome.  This part of the analysis was done with scikit-learn library in Python.

The input features consisted of taxonomic profiles (metagenomes), relative abundance of each metabolic pathway (metatranscriptomics) and metabolic intensities, which were previously normalized in data preprocessing step. The RF model was learned from the training data, using hyperparameter tuning (hyperparameter grid search) and 5-fold cross-validation. Training, test, and validation sets were randomly chosen for learning and subsequent performance evaluation. The performance of the model was evaluated on the training and previously unseen test datasets (Figure 23).

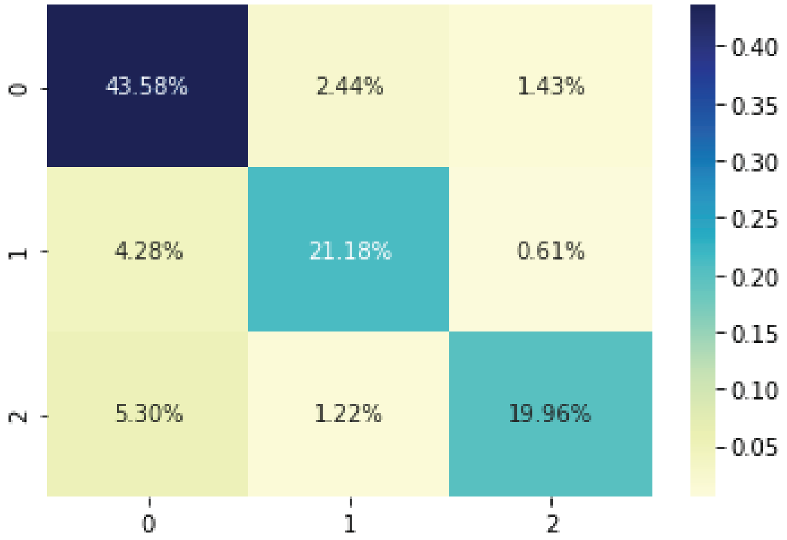
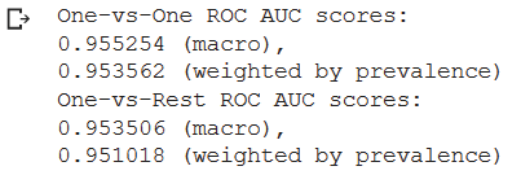


*Figure 23: RF classifier confusion matrix (heatmap). n_estimators = 500. Y-axis represents true labels and X-axis the predicted labels. CD =0, UC= 1 and nonIBD = 2.*

As the implementation of the RF was done in Python, we compute the Area Under the Receiver Operating Characteristic Curve (ROC AUC) from prediction scores from `sklearn.metrics` as it is the implementation that can be used for multiclass problems.

Next, we present the AUC score using the One-vs-Rest (OvR) and One-vs-One (OvO) schemes for multi-class classification. OvR splits a multi-class classification into one binary classification problem while OvO does the same split but for each pair of classes. We report both macro average, and a prevalence-weighted average, however, no substantial difference is presented:



```
One-vs-One ROC AUC scores:
0.955254 (macro),
0.953562 (weighted by prevalence)
One-vs-Rest ROC AUC scores:
0.953506 (macro),
0.951018 (weighted by prevalence)
```

## 3.5 Performance assessment

We can perform analysis of predicting performance for different learning algorithms by two approaches: *k-fold cross-validation* and *bootstrapping* (Friedman, 2000). Cross validation will commonly be applied to estimate the performance of the learned model on an unseen replication dataset. Bootstrapping will aid in obtaining estimates of the frequency of individual arcs within a given BN.

In summary, the algorithms iteratively add the most frequent arc from the frequency matrix until a network of n nodes in the Markov blanket of the phenotype is achieved.

# 4 Results

In this study we constructed a DBN model of the gut microbial ecosystem from the Inflammatory Bowel Disease Multi-omics dataset of the Human Microbiome Project. We used a two-stage DBN model, where two slices are modeled and learned at time.

Our ultimate purpose was to identify a bacterial signature that describes the dynamics of adult microbial gut as well as compare differences in signatures between subjects with UC, CD and healthy. In order to do this, we (i) prepared a framework that covered main microbiome analysis preprocessing steps, (ii) modeled interactions between different omics, (iii) construct and learned a dynamic structure for each disease state (UC & DC) to infer which is the most probable dynamics, i.e. identify a *maximum a posteriori* (MAP) (iv) construct a model adding a 'diagnosis' node to the network and study its outgoing arcs.

The preprocessing steps, implemented through our ad-hoc script, involved filtering subjects with limited time points, integrating three omic types in one matrix, normalization, and feature selection. Based on these preprocessing steps, the resulting dataset used for modeling consisted of 91 subjects, 200 microbial taxa, 200 expressed metabolic pathways and 200 metabolites. As clinical variables, the week in which the sample was obtained, and the use (binary) of antibiotics were included. In addition, Bayes factor score was used to further reduce the dataset. We used prior knowledge as input to the learning DBN algorithm in order to constraint the resulting output model and prevent overfitting.

The full network learned by the model comprised of 182 nodes per time slice: 37 microbial taxa; 19 gene pathways; 29 metabolites and three clinical variables. We constructed a model (i) with and without bootstrap (10 repetitions due to restricted computational tools), (ii) with restriction matrix (prior knowledge) as shown in Figure 24, and (iii) different Bayes factor score thresholds (threshold = 5, 10, 15) were explored as part of a hyperparameter tuning phase. Connections with largest Bayes factors are more likely to represent a true causal association.

| clinical<br>▲ | MGX<br>● | MTX<br>◆ | metabolite<br>■ |
|---|---|---|---|
| 2 | 3 | 2 | 2 |
| 0 | 2 | 3 | 2 |
| 0 | 2 | 1 | 3 |
| 0 | 2 | 2 | 1 |

*Figure 24: Restriction matrix. Self-loops identified by 1, inter-edges identified as 2 and intra-edges identified as 3.*

For illustrative purposes, we trained a DBN model on a subset of the 50 best entities of each omic type and maximum number of parents of 3. Results for the combined diagnosis model (all health conditions in same model), are shown in Figure 25. Nodes represent bacterial taxon, metabolites, clinical data, or metabolic pathways.

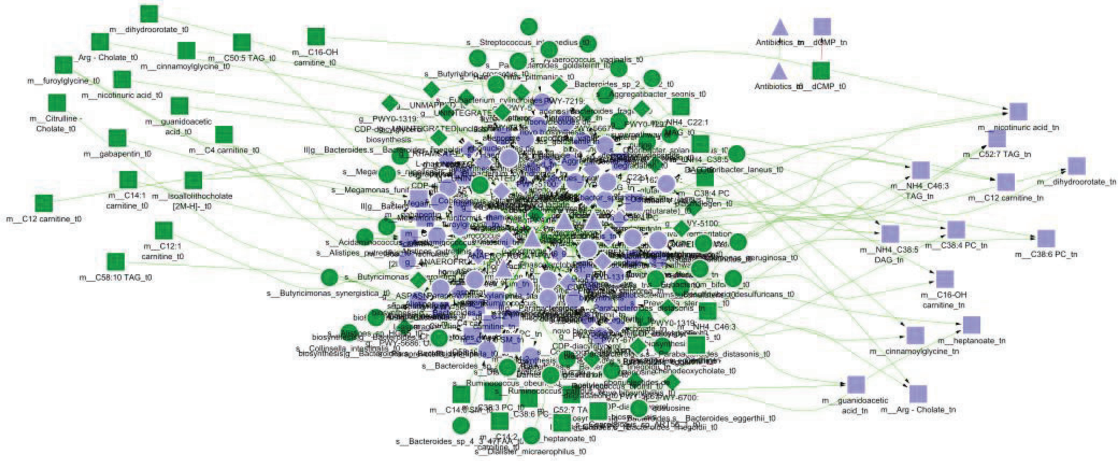
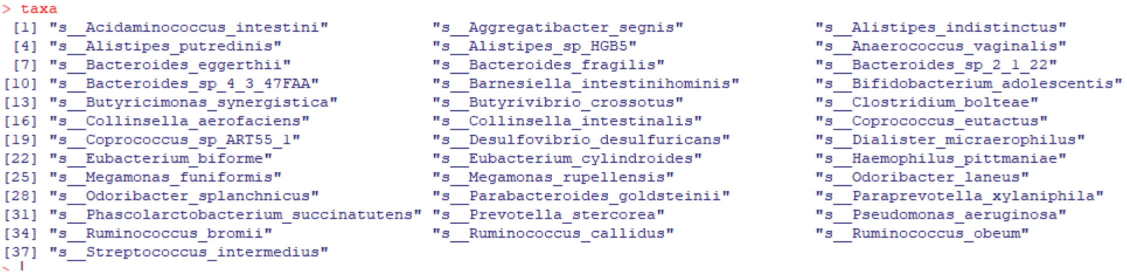


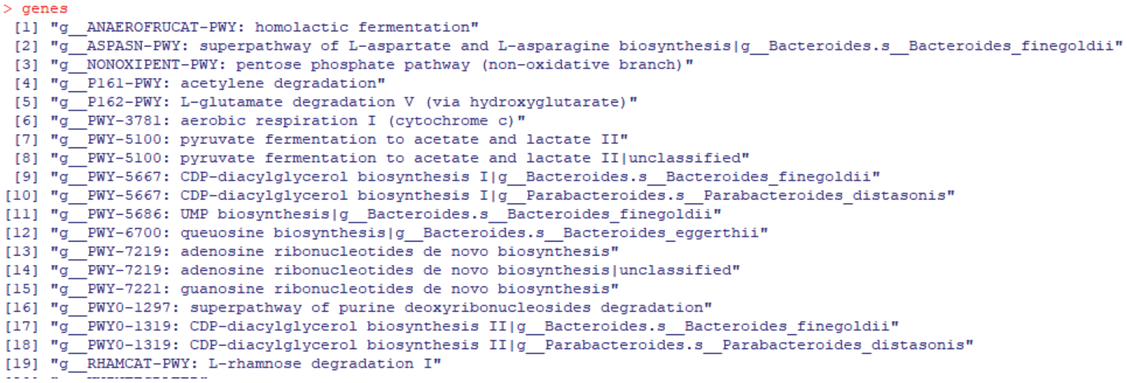
*Figure 25: Learned DBN with 'restriction matrix' constraints on the top 50 best features per omic type. Green nodes represent time slice $t_i$ and purple nodes the consecutive $t_{i+1}$. Metabolite nodes are represented by squares, species (taxa) by circles, clinical variables by triangles and metabolic pathways by diamonds. Total number of nodes is 182, and total number of edges 231.*

Our network is composed of the nodes grouped by typology, presented in Figure 26, Figure 27 and Figure 28.



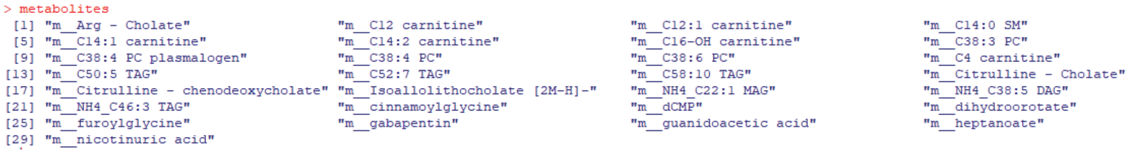
```
> taxa
 [1] "s__Acidaminococcus_intestini"           "s__Aggregatibacter_segnis"            "s__Alistipes_indistinctus"
 [4] "s__Alistipes_putredinis"                "s__Alistipes_sp_HGB5"                 "s__Anaerococcus_vaginalis"
 [7] "s__Bacteroides_eggerthii"               "s__Bacteroides_fragilis"              "s__Bacteroides_sp_2_1_22"
[10] "s__Bacteroides_sp_4_3_47FAA"            "s__Barnesiella_intestinihominis"      "s__Bifidobacterium_adolescentis"
[13] "s__Butyricimonas_synergistica"          "s__Butyrivibrio_crossotus"            "s__Clostridium_bolteae"
[16] "s__Collinsella_aerofaciens"             "s__Collinsella_intestinalis"          "s__Coprococcus_eutactus"
[19] "s__Coprococcus_sp_ART55_1"              "s__Desulfovibrio_desulfuricans"       "s__Dialister_micraerophilus"
[22] "s__Eubacterium_biforme"                 "s__Eubacterium_cylindroides"          "s__Haemophilus_pittmaniae"
[25] "s__Megamonas_funiformis"                "s__Megamonas_rupellensis"             "s__Odoribacter_laneus"
[28] "s__Odoribacter_splanchnicus"            "s__Parabacteroides_goldsteinii"       "s__Paraprevotella_xylaniphila"
[31] "s__Phascolarctobacterium_succinatutens" "s__Prevotella_stercorea"              "s__Pseudomonas_aeruginosa"
[34] "s__Ruminococcus_bromii"                 "s__Ruminococcus_callidus"             "s__Ruminococcus_obeum"
[37] "s__Streptococcus_intermedius"
> |
```

*Figure 26: Metagenomics (taxa) type nodes.*



```
> genes
 [1] "g__ANAEROFRUCAT-PWY: homolactic fermentation"
 [2] "g__ASPASN-PWY: superpathway of L-aspartate and L-asparagine biosynthesis|g__Bacteroides.s__Bacteroides_finegoldii"
 [3] "g__NONOXIPENT-PWY: pentose phosphate pathway (non-oxidative branch)"
 [4] "g__P161-PWY: acetylene degradation"
 [5] "g__P162-PWY: L-glutamate degradation V (via hydroxyglutarate)"
 [6] "g__PWY-3781: aerobic respiration I (cytochrome c)"
 [7] "g__PWY-5100: pyruvate fermentation to acetate and lactate II"
 [8] "g__PWY-5100: pyruvate fermentation to acetate and lactate II|unclassified"
 [9] "g__PWY-5667: CDP-diacylglycerol biosynthesis I|g__Bacteroides.s__Bacteroides_finegoldii"
[10] "g__PWY-5667: CDP-diacylglycerol biosynthesis I|g__Parabacteroides.s__Parabacteroides_distasonis"
[11] "g__PWY-5686: UMP biosynthesis|g__Bacteroides.s__Bacteroides_finegoldii"
[12] "g__PWY-6700: queuosine biosynthesis|g__Bacteroides.s__Bacteroides_eggerthii"
[13] "g__PWY-7219: adenosine ribonucleotides de novo biosynthesis"
[14] "g__PWY-7219: adenosine ribonucleotides de novo biosynthesis|unclassified"
[15] "g__PWY-7221: guanosine ribonucleotides de novo biosynthesis"
[16] "g__PWY0-1297: superpathway of purine deoxyribonucleosides degradation"
[17] "g__PWY0-1319: CDP-diacylglycerol biosynthesis II|g__Bacteroides.s__Bacteroides_finegoldii"
[18] "g__PWY0-1319: CDP-diacylglycerol biosynthesis II|g__Parabacteroides.s__Parabacteroides_distasonis"
[19] "g__RHAMCAT-PWY: L-rhamnose degradation I"
```

*Figure 27: Metatranscriptomics (RNA pathways) type nodes.*



```
> metabolites
 [1] "m__Arg - Cholate"                  "m__C12 carnitine"                  "m__C12:1 carnitine"        "m__C14:0 SM"
 [5] "m__C14:1 carnitine"                "m__C14:2 carnitine"                "m__C16-OH carnitine"       "m__C38:3 PC"
 [9] "m__C38:4 PC plasmalogen"           "m__C38:4 PC"                       "m__C38:6 PC"               "m__C4 carnitine"
[13] "m__C50:5 TAG"                      "m__C52:7 TAG"                      "m__C58:10 TAG"             "m__Citrulline - Cholate"
[17] "m__Citrulline - chenodeoxycholate" "m__Isoallolithocholate [2M-H]-"    "m__NH4_C22:1 MAG"          "m__NH4_C38:5 DAG"
[21] "m__NH4_C46:3 TAG"                  "m__cinnamoylglycine"               "m__dCMP"                   "m__dihydroorotate"
[25] "m__furoylglycine"                  "m__gabapentin"                     "m__guanidoacetic acid"     "m__heptanoate"
[29] "m__nicotinuric acid"
```

*Figure 28: Metabolites type nodes.*

Additionally, we present Table 9 with comparison of different network search algorithms: (i) K2-style, (ii) pheno-centric search, and (iii) simulated annealing. We found CGBayesNet had one major disadvantage for our study when reporting performance (AUC) because the implementation is only available for binary phenotype and not multiclass as our case (UC, CD, healthy). For this reason, for predictive performance analysis we learned a DBN model for CD and non IBD subjects, i.e., transforming our problem to a binary one.

| Network | Bootstrap realizations | Total Continuous Nodes | Total Discrete Nodes | AUC (%) |
|---|---|---|---|---|
| A | 20 | 3 | 153 | 60.83 |
| B | 20 | 3 | 153 | 54.04 |
| C | 20 | 3 | 40 | 54.85 |

*Table 9: Results of four Bayesian networks from bootstrapping on microbiome multi-omic time-series dataset. AUC is reported as a measure of predictive accuracy of the network. Number of continuous nodes = 150 and number of discrete nodes = 3.*

Table 9 shows the predictive performance of thre different networks on predicting IBD condition (diagnosis = CD) in the iHMP2 IBD dataset. Total nodes (continuous and discrete) reports the size of the network. AUC reports the convex hull of the area under receiver operator characteristic curve, which measures prediction at various sensitivity and specificity combinations. We used 20 bootstrap realizations of the dataset and computed networks at various edge frequencies. The consensus Markov-blanket had 71 nodes.

The AUC values and general performance (Table 9 and Figure 29) was lower than we expected and there is certainly room for improvement.

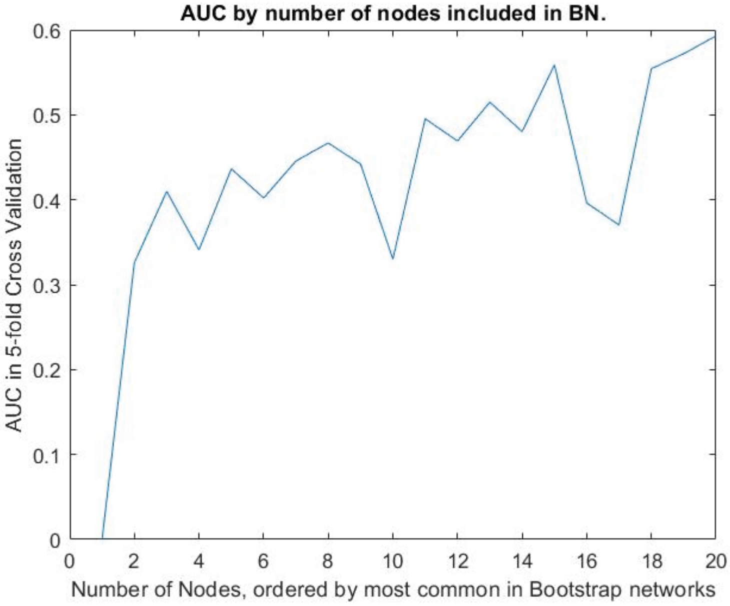


*Figure 29: Results of performance comparison of networks. A bootstrapping-produced continuous adjacency matrix is used to compare the performance of networks formed by starting with the phenotype node, and then adding, in sequence, the most frequent edge occurring in the bootstrap networks, and measuring the performance of that network on the dataset in cross-validation.*

Results for the independent models for each condition (CD, UC, healthy) are also reported in Figures 30, 31 and 32. Full name of nodes can be consulted in Appendix III.

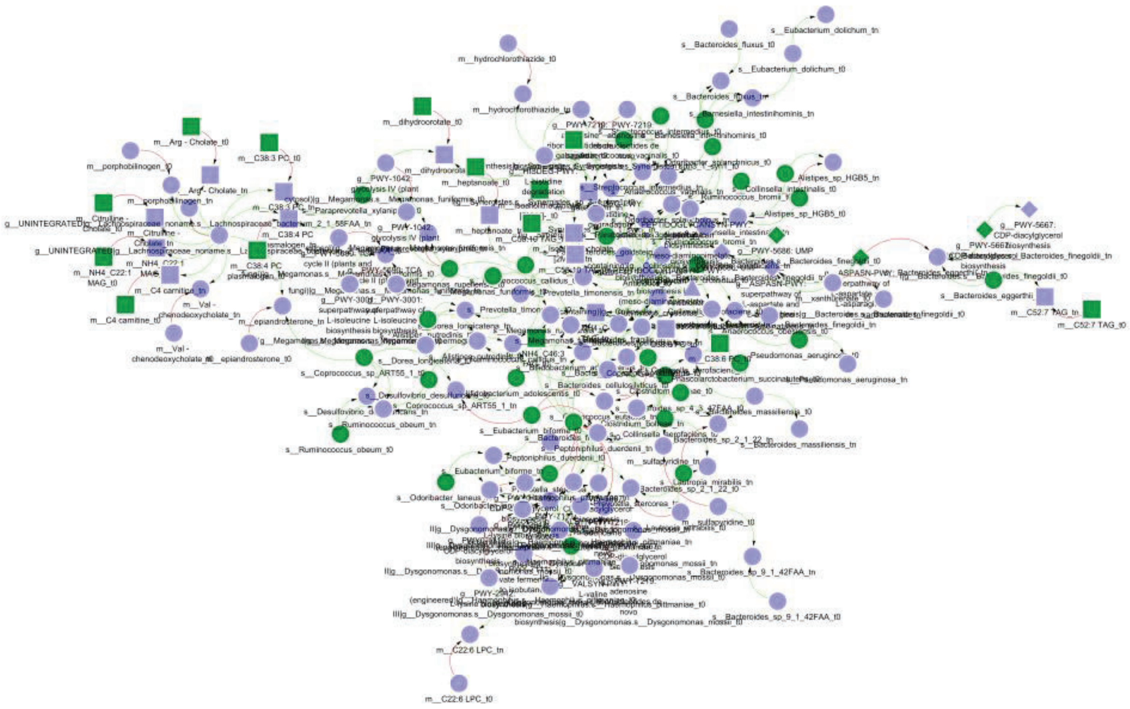

s__Eubacterium_dolichum_tn
s__Bacteroides_fluxus_t0
s__Eubacterium_dolichum_t0
m__hydrochlorothiazide_t0
m__hydrochlorothiazide_tn
s__Bacteroides_fluxus_tn
s__Barnesiella_intestinihominis_tn
m__dihydroorotate_t0
m__Arg - Cholate_t0
m__C38:3 PC_t0
m__porphobilinogen_t0
m__Arg - Cholate_tn
m__porphobilinogen_tn
Alistipes_sp_HGB5_tn
Alistipes_sp_HGB5_t0
Collinsella_intestinalis_t0
g__PWY-5667:
CDP-diacylglycerol
biosynthesis
m__C52:7 TAG_tn
m__C52:7 TAG_t0
s__Bacteroides_eggerthii
m__NH4_C22:1 MAG_t0
m__C4 carnitine_tn
m__C4 carnitine_t0
m__Val - chenodeoxycholate_tn
m__Val - chenodeoxycholate_t0
m__epiandrosterone_tn
m__epiandrosterone_t0
s__Coprococcus_sp_ART55_1_t0
s__Coprococcus_sp_ART55_1_tn
s__Desulfovibrio_desulfuricans_tn
s__Ruminococcus_obeum_tn
s__Ruminococcus_obeum_t0
s__Pseudomonas_aeruginosa_tn
s__Bacteroides_massiliensis_t0
s__Bacteroides_massiliensis_tn
s__Bacteroides_sp_2_1_22_t0
s__Eubacterium_biforme_t0
s__Eubacterium_biforme_tn
s__Peptoniphilus_duerdenii_t0
m__sulfapyridine_tn
m__sulfapyridine_t0
s__Lautropia_mirabilis_tn
s__Odoribacter_laneus_tn
s__Odoribacter_laneus_t0
s__Bacteroides_sp_9_1_42FAA_tn
s__Bacteroides_sp_9_1_42FAA_t0
s__Dysgonomonas_mossii_tn
s__Dysgonomonas_mossii_t0
m__C22:6 LPC_tn
m__C22:6 LPC_t0

*Figure 30: DBN for patients with UC with 'restriction matrix' constraints on the top 50 best features per omic type. Green nodes represent time slice tᵢ and purple nodes the consecutive ti+1. Metabolite nodes are represented by squares, species (taxa) by circles, clinical variables by triangles and metabolic pathways by diamonds.*

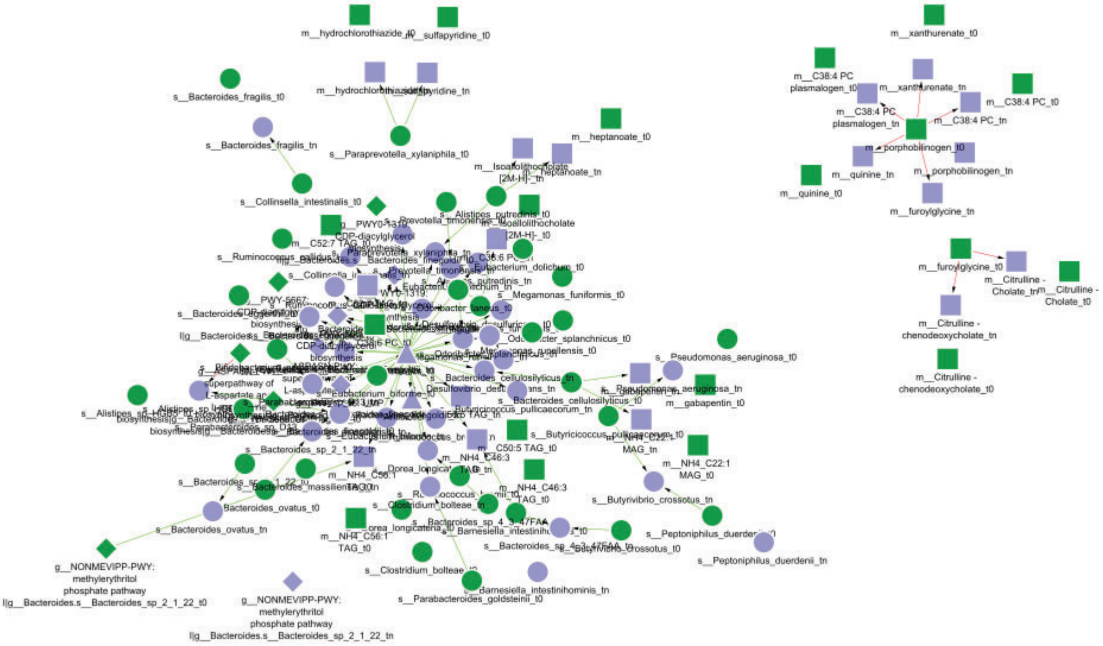

m__hydrochlorothiazide_t0
m__sulfapyridine_t0
m__hydrochlorothiazide_tn
m__sulfapyridine_tn
s__Bacteroides_fragilis_t0
s__Bacteroides_fragilis_tn
s__Paraprevotella_xylaniphila_t0
m__heptanoate_t0
m__heptanoate_tn
s__Collinsella_intestinalis_t0
s__Alistipes_putredinis_t0
s__Ruminococcus_gnavus_t0
s__Eubacterium_dolichum_t0
s__Megamonas_funiformis_t0
s__Bacteroides_cellulosilyticus_tn
s__Bacteroides_cellulosilyticus_t0
s__Pseudomonas_aeruginosa_t0
s__Pseudomonas_aeruginosa_tn
m__gabapentin_t0
s__Butyricicoccus_pullicaecorum_tn
s__Butyricicoccus_pullicaecorum_t0
m__C50:5 TAG_t0
m__NH4_C22:1 MAG_tn
m__NH4_C22:1 MAG_t0
s__Butyrivibrio_crossotus_tn
s__Butyrivibrio_crossotus_t0
s__Peptoniphilus_duerdenii_t0
s__Peptoniphilus_duerdenii_tn
s__Bacteroides_sp_2_1_22_tn
s__Bacteroides_ovatus_t0
s__Bacteroides_ovatus_tn
s__Clostridium_bolteae_t0
s__Clostridium_bolteae_tn
s__Barnesiella_intestinihominis_tn
s__Parabacteroides_goldsteinii_t0
m__NH4_C56:1 TAG_t0
m__NH4_C46:3 TAG_t0
g__NONMEVIPP-PWY: methylerythritol phosphate pathway I|g__Bacteroides.s__Bacteroides_sp_2_1_22_t0
g__NONMEVIPP-PWY: methylerythritol phosphate pathway I|g__Bacteroides.s__Bacteroides_sp_2_1_22_tn
m__xanthurenate_t0
m__xanthurenate_tn
m__C38:4 PC plasmalogen_t0
m__C38:4 PC plasmalogen_tn
m__C38:4 PC_t0
m__C38:4 PC_tn
m__porphobilinogen_t0
m__porphobilinogen_tn
m__quinine_tn
m__quinine_t0
m__furoylglycine_tn
m__furoylglycine_t0
m__Citrulline - Cholate_tn
m__Citrulline - Cholate_t0
m__Citrulline - chenodeoxycholate_tn
m__Citrulline - chenodeoxycholate_t0

*Figure 31: DBN for CD patients with ‘restriction matrix’ constraints on the top 50 best features per omic type. Green nodes represent time slice $t_i$ and purple nodes the consecutive $t_{i+1}$.* Metabolite nodes are represented by squares, species (taxa) by circles, clinical variables by triangles and metabolic pathways by diamonds.

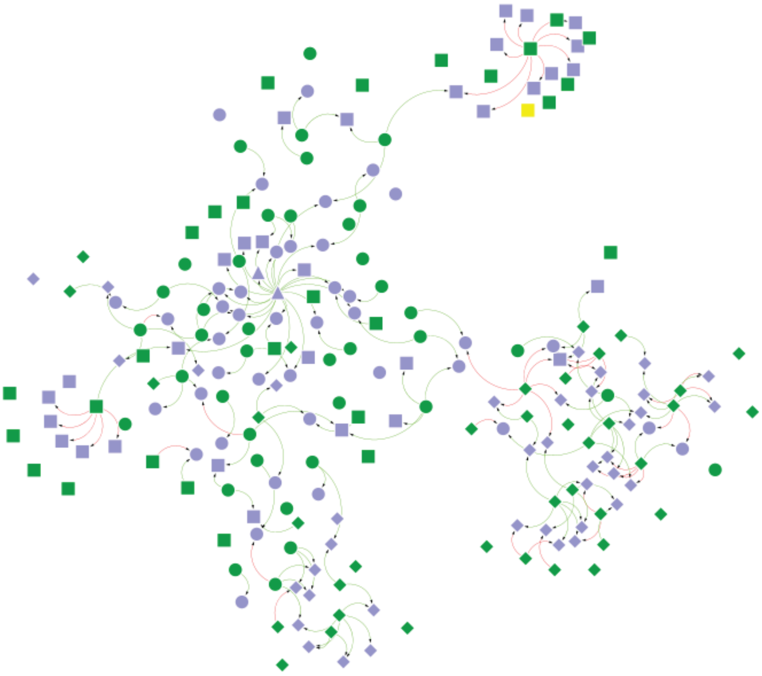


*Figure 32: DBN for non-IBD (control) subjects with ‘restriction matrix’ constraints on the top 50 best features per omic type. Green nodes represent time slice $t_i$ and purple nodes the consecutive $t_{i+1}$.* Metabolite nodes are represented by squares, species (taxa) by circles, clinical variables by triangles and metabolic pathways by diamonds *(due to network size, zoom in was required in order to show node labels, however, full network could not be displayed if zoomed in).*

In preliminary results we did observe, as expected from results of previous studies in the literature, IBD is associated with overall community dysbiosis rather than a specific bacterial species. For instance, a combination of increase in *Actinobacteria* and *Proteobacteria* with decrease in *Clostridium* and *Faecalibacterium* is observed in subjects with this condition. Although taxa abundance analysis was not the goal of our work, it does indeed show the model points in the right direction.

This sequence of analysis demonstrates the utility of DBNs ability to generate and test predictive models in human multi omic microbiome datasets.  The final model may be suggestive of a set of taxa, gene metabolic pathways and metabolites whose expression is dysregulated in patients with IBD. Also some of these attributes can be used for further biological inquiry as predictors of other attributes, thus used as a predictive model. For instance, metabolites in $t_i$ connected to taxa in $t_{i+1}$ may be used as predictors. This relationship is found for our UC model, where we identified the following chain:

NH4_C46:3 TAG (metabolite) in $t_0$ → *Alistipes_putredinis* in $t_n$

Also found in the literature to play a critical role in inflammation and disease (Parker, Wearsch, Veloo, & Rodriguez-Palacios, 2020).

Important to note that we are aware that a possible source of error could be the XML style file created for visualization purposes (Cytoscape) which could have influenced the results obtained. Further studying of the visualization software should be conducted in order to obtain full potential and functionality.

One of the objectives we set at the beginning of the project was to perform benchmark test of our model (DBN) compared to RF for the same dataset. However, in the end we did not produce enough performance data to do a correct assessment and establish conclusions.

# 5 Conclusions and Future Research

## 5.1 Conclusions

In this study we have examined the state of the art of different artificial intelligence techniques and methods to solve a current scientific problem of interest: analyzing the human microbiome temporal changes associated to disease state. We developed a tailored preprocessing script for the Inflammatory Bowel Disease Multiomics Dataset from the iHMP project, that covers an unmet need, as, to the best of our knowledge, data access, preparation, and integration of these datasets for machine learning models have not been developed yet. Furthermore, we have applied a powerful artificial intelligence approach, DBNs, to solve the problem with an innovative configuration and approach by integrating longitudinal multi-omic data for characterization of a model for each disorder (UC, CD) and the healthy state (non IBD). Also, our work has presented analysis of different software packages to construct the solution and selected one of them (`CGBayesNet`) for the implementation process.

Our implementation consisted of the following steps:

1. Preprocessing of data set.
2. Fitting the DBN model in two step structure and parameter learning. The output of this step was a 2-stage dynamic Bayes net class object (DBN).
3. Inference and test the DBN on a subset of variables given the evidence on the other variables. The output of this step was the predicted values and log probabilities of observing a less likely outcome for each variable than the value assigned to that variable by the input data.
4. Network visualization and analysis for biological interpretation of results.

We have shown DBNs can serve as a valuable approach to capture temporal variability in microbiome data and our results have found the 50 most important taxa, metabolic pathways, and metabolites for each condition.

The usefulness of BNs for microbiome analysis has been presented in our study. The use of prior biological domain knowledge as input restrictions matrix allowed us to prove the value of this approach versus other popular ML models (e.g. RF) to build explainable interpretable models. However, BN have a series of limitations. First, heavy assumptions that can be easily violated are required for valid inference. Second, model search in presence of large number of variables (as in real human microbiome data) requires massive computational power and its performance is affected by overall sample size. Third, BNs cannot explain a cyclic or feedback relationship among variables.

Our work clearly has some limitations that should be addressed in future research. Furthermore, a gold standard of human microbiome analysis needs to be established. Despite this, we are confident that our research will serve as a basis for future studies on dynamic Bayesian networks modeling of longitudinal human microbiome data.

## 5.2 Future research

Even though intense research has been done in causal discovery from dynamic human microbiome data there are still some necessities and goals that need to be covered in order for this field to keep active. This research has given rise to many questions in need of further investigation.

In terms of data preprocessing which we believe is determining for posterior quality of results and performance of model, some intriguing lines of research can be further explored. Firstly, spline interpolation method that combines time-course modeling with multivariate approaches to capture ecological interactions can result in increase of accuracy and performance of our model. Approaches such as multivariate analysis (Bodein, Chapleur, Droit, & Lê Cao, 2019) and linear mixed model splines (*LMMS)*, (Straube, 2015) implemented in R package `lmms` should be explored in future research. Also interesting along these same lines, being able to determine the ideal sampling frequency. It is known that different sampling frequencies change the associations inferred (Fuhrman, 2015) and that ideal sampling frequency depends on the system, so, can we determine a value for human microbiome studies? Secondly, feature selection. A different approach we could have followed would have been to obtain two different subsets by two different approaches and then obtain selected features that are the intersections of both subsets. Our focus has not been set on the literature review and state of the art of feature selection techniques for human microbiome analysis, therefore, this is a recommended matter to explore in future work. As expected, the choice of variables due to the used feature selection techniques, have influenced the results obtained. Moreover, although we further applied preprocessing steps to reduce sparsity of our data, the input dataset used for learning and inference still contained sparse features with missing data.

Concerning dataset selection, it would be convenient to test the model in multiple cohorts and potentially different class balances. For instance, exploring datasets that include geographically diverse populations. It has been seen that identical twins can become key in human microbiome studies because these subjects are not constrained by confounding factors (genetic confounding). The concept of impact of confounding factors on human microbiome analysis is still poorly understood, yet of critical importance (Bajaj, 2015). For studies focused on causal inference, as (Pasolli, Truong, Malik, Waldron, & Segata, 2016) advices, results will be biased in presence of confounding factors due to violation of the causal sufficiency assumption. This concept should be considered when designing a clinical study and selecting patient cohort, for instance, randomized controlled trial, a gold standard for causal inference, could be carried out in order to eliminate selection bias or confounding. The choice of dataset is one of the main limiting factor on the resulting performance of our model. For our study, we explored publicly available datasets, but we would like to point out for future studies an interesting cohort composed of identical twins, thus not subject to genetic confounding, namely *TwinsUK* (Verdi, 2019).

Furthermore, we selected four specific types of data as our input features (clinical, taxa, metatranscriptomic pathways and metabolites) but alternatives such as host gene expression or other environmental features (diet, patient a smoker) could be used in future work to learn the structure of the DBN.

Finally, with regards to evaluation and performance assessment, further work should be done to improve and support accountability of conclusions and results. Our work did not include a detailed performance report in results section. Based on this, our goal in next stages of this study will be to: report multiple evaluation metrics with confidence intervals and perform model hypothesis interpretation based on feature importance. Moreover, it would be useful to use an additional dataset from a research institution or medical center to further validate the model's performance (not used for training nor testing). For the validation of output, we suggest that the assistance of microbiology experts to interpret resulting models could greatly enrich conclusions and help point out new directions for future investigations.

The gut microbiome has been extensively studied but due to its high complexity and inter-individual heterogeneity it is not yet fully understood. Although machine learning methods, and in particular dynamic Bayesian networks, are promising techniques to infer useful insights, there is still considerable work to be done in some areas.  Other suggestions of topics that have not been explored in this Master thesis, but evidence substantial interest could be the study of other microbes such as the skin microbiome, oral cavity microbiome or the respiratory system microbiome that can produce equally interesting results and that have not been widely explored and characterized yet.

# 6 Bibliography

Ai, L., Tian, H., Chen, Z., Chen, H., Xu, J., & Fang, J. Y. (2017). Systematic evaluation of supervised classifiers for fecal microbiota-based prediction of colorectal cancer. *Oncotarget*, 9546–9556.

Aitchison, J. (1982). The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society*, 139-177.

Aliferis, C., Statnikov, A., Tsamardinos, I., Mani, S., & Koutsoukos, X. (2010). Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification Part I: Algorithms and Empirical Evaluation. *Journal of Machine Learning Research*, 171−234.

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome biology*, R106.

Arrieta, M. C., Stiemsma, L. T., Dimitriu, P. A., Thorson, L., Russell, S., Yurist-Doutsch, S., . . . Kollmann, T. (2015). Early infancy microbial and metabolic alterations affect risk of childhood asthma. *Science translational medicine*, 307ra152.

Asnicar, F., Berry, S. E., Valdes, A. M., Nguyen, L. H., Piccinno, G., Drew, D. A., . . . Thomas, A. M. (2021). Microbiome connections with host metabolism and habitual diet from 1,098 deeply phenotyped individuals. *Nature*, 321–332.

Badri, M., Kurtz, Z., Müller, C., & Bonneau, R. (2018). Normalization methods for microbial abundance data strongly affect correlation estimates. *bioRxiv*, 406264.

Bajaj, J. S. (2015). Salivary microbiota reflects changes in gut microbiota in cirrhosis with hepatic encephalopathy. *Hepatology (Baltimore, Md.)*, 1260–1271.

Baldini, F., Hertel, J., Sandt, E., Thinnes, C. C., Neuberger-Castillo, L., Pavelka, L., . . . Consortium, N.-P. (2020). Parkinson's disease-associated alterations of the gut microbiome predict disease-relevant changes in metabolic functions. *BMC biology*, 62.

Barredo Arrieta, A., Díaz-Rodríguez, N. D., Bennetot, A., Tabik, S., Barbado, A., García, S., . . . Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Information Fusion*, 82–115.

Bengmark, S. (2013). Gut microbiota, immune development and function. *Pharmacological research*, 87–113.

Berry, S. E., Valdes, A. M., Drew, D. A., Asnicar, F., Mazidi, M., Wolf, J., . . . Ordovas, J. M. (2020). Human postprandial responses to food and potential for precision nutrition. *Nature medicine*, 964–973.

Bodein, A., Chapleur, O., Droit, A., & Lê Cao, K. A. (2019). A Generic Multivariate Framework for the Integration of Microbiome Longitudinal Studies With Other Data Types. *Frontiers in genetics*, 963.

Bokulich, N. A., Rideout, J. R., Mercurio, W. G., Shiffer, A., Wolfe, B., Maurice, C. F., . . . Caporaso, J. G. (2016). mockrobiota: a Public Resource for Microbiome Bioinformatics Benchmarking. *mSystems*, e00062-16.

Borchani, H., Bielza, C., Martínez-Martínez, P., & Larrañaga, P. (2012). Markov blanket-based approach for learning multi-dimensional Bayesian network classifiers: an application to predict the European Quality of Life-5 Dimensions (EQ-5D) from the 39-item Parkinson's Disease Questionnaire (PDQ-39). *Journal of biomedical informatics, 45*, 1175–1184.

Bray, J., & Curtis, J. (1957). An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecological monographs*, 325-349.

Breiman, L. (2001). Random Forests. *Machine Learning*, 5–32.

Bucci, V., Tzen, B., Li, N., Simmons, M., Tanoue, T., Bogart, E., . . . Gerber, G. K. (2016). MDSINE: Microbial Dynamical Systems INference Engine for microbiome time-series analyses. *Genome biology*, 121.

Buffie, C. G., Jarchum, I., Equinda, M., Lipuma, L. G., Viale, A., Ubeda, C., . . . Pamer, E. G. (2012). Profound alterations of intestinal microbiota following a single dose of clindamycin results in sustained susceptibility to Clostridium difficile-induced colitis. *Infection and immunity*, 62–73.

Bull, M. J., & Plummer, N. T. (2014). Part 1: The Human Gut Microbiome in Health and Disease. *Integrative medicine (Encinitas, Calif.) vol. 13,6*, 17-22.

Bullard, J. H., Purdom, E., Hansen, K. D., & Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *MC bioinformatics*, 94.

Cammarota, G., Ianiro, G., Ahern, A., Carbone, C., Temko, A., Claesson, M. J., . . . Tortora, G. (2020). Gut microbiome, big data and machine learning to promote precision medicine for cancer. *Nature reviews. Gastroenterology & hepatology.*, 635–648.

Cano, A., Gomez-Olmedo, M., & Moral, S. (2008). A score based ranking of the edges for the PC algorithm. *Proceedings of the 4th European Workshop on Probabilistic Graphical Models*.

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., . . . McDonald, D. M. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature methods*, 335–336.

Carmody, R. N., & Turnbaugh, P. J. (2014). Host-microbial interactions in the metabolism of therapeutic and diet-derived xenobiotics. *The Journal of clinical investigation*, 4173–4181.

Carrieri, A. P., Haiminen, N., Maudsley-Barton, S., Gardiner, L. J., Murphy, B., Mayes, A. E., . . . Hoptroff. (2021). Explainable AI reveals changes in skin microbiome composition linked to phenotypic differences. *Scientific reports*, 4565.

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature*, 20–23.

Chang, H. H., & McGeachie, M. (2011). Phenotype prediction by integrative network analysis of SNP and gene expression microarrays. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, (págs. 6849–6852).

Chickering, D. (1996). Learning Bayesian networks is NP-Complete. In *Learning from Data* (pp. 121-130). New york: Springer.

Chickering, D. (2002). Optimal Structure Identification With Greedy Search. *Journal of Machine Learning Research*, 507-554.

Clayton, T. A., Baker, D., Lindon, J. C., Everett, J. R., & Nicholson, J. K. (2009). Pharmacometabonomic identification of a significant host-microbiome metabolic interaction affecting human drug metabolism. *Proceedings of the National Academy of Sciences of the United States of America*, 14728–14733.

Colombo, D., & Maathuis, M. (2014). Order-Independent Constraint-Based Causal StructureLearning. *Journal of Machine Learning Researc*, ) 3921-39.

Cooper, G., & Herskovits, E. (1992). A Bayesian Method for the Induction of Probabilistic Networks from Data. *Machine Learning*, 309–347.

Cornejo-Pareja, I., Ruiz-Limón, P., Gómez-Pérez, A. M., Molina-Vega, M., Moreno-Indias, I., & Tinahones, F. J. (2020). Differential Microbial Pattern Description in Subjects with Autoimmune-Based Thyroid Diseases: A Pilot Study. *Journal of personalized medicine*, 192.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 273-297.

Costea, P. I.-V. (2017). Towards standards for human fecal sample processing in metagenomic studies. *Nature biotechnology*, 1069–1076.

Cowell, R. (2005). Local Propagation in Conditional Gaussian Bayesian Networks. *Journal of Machine Learning Research*, 1517–1550.

Dagum, P., & Luby, M. (1993). Approximating probabilistic inference in Bayesian belief networks is NP-hard. *Artificial Intelligence*, 141-153.

David, L. A., Maurice, C. F., Carmody, R. N., Gootenberg, D. B., Button, J. E., Wolfe, B. E., . . . Turnbaugh, P. J. (2014). Diet rapidly and reproducibly alters the human gut microbiome. *Nature*, 559–563.

de Groot, P. F., Frissen, M. N., de Clercq, N. C., & Nieuwdorp, M. (2017). Fecal microbiota transplantation in metabolic syndrome: History, present and future. *Gut microbes*, 253–267.

Dean, T. a. (1989). A model for reasoning about persistence and causation. *Computational Intelligence*, 142-150.

Dean, T., & Kanazawa, K. (1989). A model for reasoning about persistence and causation. *Computational Intelligence*, 142-150.

Depner, M., Taft, D. H., Kirjavainen, P. V., Kalanetra, K. M., Karvonen, A. M., Peschel, S., . . . Pekkanen, J. (2020). Maturation of the gut microbiome during the first year of life contributes to the protective farm effect on childhood asthma. *Nature*, 1766–1775.

Dhariwal, A., Chong, J., Habib, S., King, I. L., Agellon, L. B., & Xia, J. (2017). MicrobiomeAnalyst: a web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data. *Nucleic acids research*, W180–W188.

Doumatey, A. P., Adeyemo, A., Zhou, J., Lei, L., Adebamowo, S. N., Adebamowo, C., & Rotimi, C. N. (2020). Gut Microbiome Profiles Are Associated With Type 2 Diabetes in Urban Africans. *Frontiers in cellular and infection microbiology*, 63.

Duvallet, C., Gibbons, S. M., Gurry, T., Irizarry, R. A., & Alm, E. J. (2017). Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nature communications*, 1784.

Edgar, R. C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature methods*, 996–998.

Faust, K., & Raes, J. (2012). Microbial interactions: from networks to models. *Nature reviews. Microbiology*, 538–550.

Faust, K., Lahti, L., Gonze, D., de Vos, W. M., & Raes, J. (2015). Metagenomics meets time series analysis: unraveling microbial community dynamics. *Current opinion in microbiology*, 56–66.

Fernandez, M., Riveros, J. D., Campos, M., Mathee, K., & Narasimhan, G. (2015). Microbial "social networks". *BMC genomics*, S6.

Ferrocino, I., Ponzo, V., Gambino, R., Zarovska, A., Leone, F., Monzeglio, C., . . . Bo, S. (2018). Changes in the gut microbiota composition during pregnancy in patients with gestational diabetes mellitus (GDM). *Scientific reports*, 12216.

Franzosa, E. A., McIver, L. J., Thompson, L. R., Schirmer, M., Weingart, G., Lipson, K. S., . . . Huttenhower, C. (2018). Species-level functional profiling of metagenomes and metatranscriptomes. *Nature methods*, 962–968.

Franzosa, E. A., Sirota-Madi, A., Avila-Pacheco, J., Fornelos, N., Haiser, H. J., Reinker, S., . . . Imhann, F. (2019). Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nature microbiology*, Nature microbiology.

Friedman, N. L. (2000). Using Bayesian networks to analyze expression data. *Journal of computational biology : a journal of computational molecular cell biology*, 601–620.

Fuhrman, J. A. (2015). Marine microbial community dynamics and their ecological interpretation. *Nature reviews. Microbiology*, 133–146.

Fukui, H., Nishida, A., Matsuda, S., Kira, F., Watanabe, S., Kuriyama, M., . . . Miwa, H. (2020). Usefulness of Machine Learning-Based Gut Microbiome Analysis for Identifying Patients with Irritable Bowels Syndrome. *Journal of clinical medicine*, 2403.

Gasse, M., Aussem, A., & Haytham, E. (2012). An experimental comparison of hybrid algorithms for Bayesian network structure learning. *Machine Learning and Knowledge Discovery in Databases*, 58-73.

Gerber, K. G. (2014). The dynamic microbiome. *FEBS letters*, 4131–4139.

Gilbert, J. A., Blaser, M. J., Caporaso, J. G., Jansson, J. K., Lynch, S. V., & Knight, R. (2018). Current understanding of the human microbiome. *Nature medicine*, 392-400.

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., & Egozcue, J. J. (2017). Microbiome Datasets Are Compositional: And This Is Not Optional. *Frontiers in microbiology*, 2224.

Golightly, A., & Wilkinson, D. J. (2011). Bayesian parameter inference for stochastic biochemical network models using particle Markov chain Monte Carlo. *Interface focus*, 807–820.

Goto, S., Bono, H., Ogata, H., Fujibuchi, W., Nishioka, T., Sato, K., & Kanehisa, M. (1997). Organizing and computing metabolic pathway data in terms of binary relations. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 175–186.

Grünwald, P. (2007). *The minimum description length principle.* Cambridge, MA: MIT Press.

Gubatan, J., Levitte, S., Patel, A., Balabanis, T., Wei, M. T., & Sinha, S. R. (2021). Artificial intelligence applications in inflammatory bowel disease: Emerging technologies and future directions. *World journal of gastroenterology*, 1920–1935.

Gupta, V. K., Kim, M., Bakshi, U., Cunningham, K. Y., Davis, J. M., Lazaridis, K. N., . . . Sung, J. (2020). A predictive index for health status using species-level gut microbiome profiling. *Nature communications*, 4635.

Hacilar, H., Nalbantoglu, O., O, A., & Bakir-Gungor, B. (2020). Inflammatory Bowel Disease Biomarkers of Human Gut Microbiota Selected via Ensemble Feature Selection Methods. *arXiv*.

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., & Goodman, R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & biology*, R245–R249.

Harris, N., & Drton, M. (2013). PC Algorithm for Nonparanormal Graphical Models. *Journal of Machine Learning Research 14*, 3365-3383.

Heckerman, D., Geiger, D., & Chickering, D. (1995). Learning Bayesian networks: the combination of knowledge and statistical data. *Machine Learning*, 197-243.

Henrion, M. (1990). An Introduction to Algorithms for Inference in Belief Nets. En *Machine Intelligence and Pattern Recognition* (págs. 129-138). North-Holland.

Heshiki, Y., Vazquez-Uribe, R., Li, J., Ni, Y., Quainoo, S., Imamovic, L., . . . Panagiotou, G. (2020). Predictable modulation of cancer treatment outcomes by the gut microbiota. *Microbiome*, 28.

Holzinger, A., Biemann, C., Pattichis, C., & Kell, D. (2017). What do we need to build explainable AI systems for the medical domain? *arXiv*.

Hooper, L. V., & Gordon, J. I. (2001). Commensal host-bacterial relationships in the gut. *Science*, 1115–1118.

Hou, Q., & Kolodkin-Gal, I. (2020). Harvesting the complex pathways of antibiotic production and resistance of soil bacilli for optimizing plant microbiome. *FEMS microbiology ecology*, 96.

Howey, R., Shin, S. Y., Relton, C., Davey Smith, G., & Cordell, H. J. (2020). Bayesian network analysis incorporating genetic anchors complements conventional Mendelian randomization approaches for exploratory analysis of causal relationships in complex data. *PLoS genetics*, e1008198.

Husmeier, D. (2003). Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics*, 2271–2282.

Integrative HMP (iHMP) Research Network Consortium. (2014). The Integrative Human Microbiome Project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. *Cell host & microbe*, 276–289.

Jagtap, P., Mehta, S., Sajulga, R., Batut, B., Leith, E., Kumar, P., & Hiltemann, S. (2021, June 15). *Metatranscriptomics analysis using microbiome RNA-seq data (Galaxy Training Materials).* Retrieved from Galaxy Training: https://training.galaxyproject.org/training-material/topics/metagenomics/tutorials/metatranscriptomics/tutorial.html

Jang, B. S., Chang, J. H., Chie, E. K., Kim, K., Park, J. W., Kim, M. J., . . . Kim, H. J. (2020). Gut Microbiome Composition Is Associated with a Pathologic Response After Preoperative Chemoradiation in Patients with Rectal Cancer. *International journal of radiation oncology, biology, physics,*, 736–746.

Joseph, C. L., Zoratti, E. M., Ownby, D. R., Havstad, S., Nicholas, C., Nageotte, C., . . . Johnson, C. C. (2016). Exploring racial differences in IgE-mediated food allergy in the WHEALS birth cohort. *Annals of allergy, asthma & immunology : official publication of the American College of Allergy, Asthma, & Immunology*, 219–224.

Kalisch, M., & Bühlmann, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research*.

Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K., & Tanabe, M. (2019). New approach for understanding genome variations in KEGG. *Nucleic acids research*, D590–D595.

Kharrat, N. A.-E. (2019). Data mining analysis of human gut microbiota links Fusobacterium spp. with colorectal cancer onset. *Bioinformation*, 372–379.

Kharrat, N., Assidi, M., Abu-Elmagd, M., Pushparaj, P. N., Alkhaldy, A., Arfaoui, L., . . . Rebai, A. (2019). Data mining analysis of human gut microbiota links Fusobacterium spp. with colorectal cancer onset. *Bioinformation*, 372–379.

Kirkpatrick, S., Gelatt Jr, C., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, 671-680.

Knights, D., Costello, E. K., & Knight, R. (2011). Supervised classification of human microbiota. *FEMS microbiology reviews*, 343–359.

Koeth, R. A., Wang, Z., Levison, B. S., Buffa, J. A., Org, E., Sheehy, B. T., . . . Hazen, S. (2013). Intestinal microbiota metabolism of L-carnitine, a nutrient in red meat, promotes atherosclerosis. *Nature*, 576–585.

Koh, A., Mannerås-Holm, L., Yunn, N. O., Nilsson, P. M., Ryu, S. H., Molinaro, A., . . . Bäckhed, F. (2020). Microbial Imidazole Propionate Affects Responses to Metformin through p38γ-Dependent Inhibitory AMPK Phosphorylation. *Cell metabolism*, 643–653.

Koivula, R. W., Heggie, A., Barnett, A., Cederberg, H., Hansen, T. H., Koopman, A. D., . . . Bruna. (2014). Discovery of biomarkers for glycaemic deterioration before and after the onset of type 2 diabetes: rationale and design of the epidemiological studies within the IMI DIRECT Consortium. *Diabetologia*, 1132–1142.

Koller, D., & Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques.* MIT Press.

Korpela, K., Helve, O., Kolho, K. L., Saisto, T., Skogberg, K., Dikareva, E., . . . de Vos, W. M. (2020). Maternal Fecal Microbiota Transplantation in Cesarean-Born Infants Rapidly Restores Normal Gut Microbial Development: A Proof-of-Concept Study. *Cell*, 324–334.

Kostic, A. D., Gevers, D., Siljander, H., Vatanen, T., Hyötyläinen, T., Hämäläinen, A. M., . . . Clish, C. B. (2015). The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes. *Cell host & microbe*, 260–273.

La Rosa, P. S., Warner, B. B., Zhou, Y., Weinstock, G. M., Sodergren, E., Hall-Moore, C. M., . . . Tarr, P. I. (2014). Patterned progression of bacterial populations in the premature infant gut. *Proceedings of the National Academy of Sciences of the United States of America*, 12522–12527.

La Rosa, P. S.-M. (2014). Patterned progression of bacterial populations in the premature infant gut. *Proceedings of the National Academy of Sciences of the United States of America*, 12522–12527.

Lamendella, R., VerBerkmoes, N., & Jansson, J. K. (2012). 'Omics' of the mammalian gut--new insights into function. *Current opinion in biotechnology*, 491–500.

Larsen, P. E., & Dai, Y. (2015). Metabolome of human gut microbiome is predictive of host dysbiosis. *GigaScience*, 42.

Layeghifard, M., Hwang, D. M., & Guttman, D. S. (2017). Disentangling Interactions in the Microbiome: A Network Perspective. *Trends in microbiology*, 217–228.

Le Chatelier, E., Nielsen, T., Qin, J., Prifti, E., Hildebrand, F., Falony, G., . . . Bertalan, M. (2013). Richness of human gut microbiome correlates with metabolic markers. *Nature*, 541–546.

Le, T., Hoang, T., Li, J., Liu, L., Liu, H., & Hu, S. (2015). A Fast PC Algorithm for High Dimensional Causal Discovery with Multi-Core PCs. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.

Leiva-Gea, I., Sánchez-Alcoholado, L., Martín-Tejedor, B., Castellano-Castillo, D., Moreno-Indias, I., Urda-Cardona, A., . . . Queipo-Ortuño, M. I. (2018). Gut Microbiota Differs in Composition and Functionality Between Children With Type 1 Diabetes and MODY2 and Healthy Control Subjects: A Case-Control Study. *Diabetes care*, 2385–2395.

Lemeire, J., Meganck, S., Cartella, F., & Liu, T. (2012). Conservative independence-based causal structure learning in absence of adjacency faithfulness. *International Journal of Approximate Reasoning*, 1305-1325.

Levan, S. R. (2019). Author Correction: Elevated faecal 12,13-diHOME concentration in neonates at high risk for asthma is produced by gut bacteria and impedes immune tolerance. *Nature microbiology*.

Li, H., He, J., & Jia, W. (2016). The influence of gut microbiota on drug metabolism and toxicity. *Expert opinion on drug metabolism & toxicology*, 31-40.

Li, M., Zhang, J., Wu, B., Zhou, Z., & Xu, Y. (2018). Identifying Keystone Species in the Microbial Community Based on Cross- Sectional Data. *Current gene therapy*, 296-306.

Liverani, E., Scaioli, E., Digby, R. J., Bellanova, M., & Belluzzi, A. (2016). How to predict clinical relapse in inflammatory bowel disease patients. *World journal of gastroenterology*, 1017–1033.

Lloyd-Price, J., Arze, C., Ananthakrishnan, A. N., Schirmer, M., Avila-Pacheco, J., Poon, T. W., . . . Mall. (2019). Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature*, 655–662.

Lopez, J., & Grinspan, A. (2016). Fecal Microbiota Transplantation for Inflammatory Bowel Disease. *Gastroenterology & hepatology*, 374–379.

Louca, S., Parfrey, L. W., & Doebeli, M. (2016). Decoupling function and taxonomy in the global ocean microbiome. *Science*, 1272–1277.

Lozupone, C. A., Hamady, M., Kelley, S. T., & Knight, R. (2007). Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. *Applied and environmental microbiology*, 1576–1585.

Lozupone, C., & Knight, R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities. *Applied and environmental microbiology*, 8228–8235.

Lugo-Martinez, J., Ruiz-Perez, D., Narasimhan, G., & Bar-Joseph, Z. (2019). Dynamic interaction network inference from longitudinal microbiome data. *Microbiome*, 54.

Luo, D., Ziebell, S., & An, L. (2017). An informative approach on differential abundance analysis for time-course metagenomic sequencing data. *Bioinformatics*, 1286–1292.

Maathuis, M. H., Colombo, D., Kalisch, M., & Bühlmann, P. (2010). Predicting causal effects in large-scale systems from observational data. *Nature methods*, 247–248.

Maathuis, M. H., Kalisch, M., & Bühlmann, P. (2009). Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*.

Madgwick, M., Sudhakar, P., Tabib, N. S., Norvaisas, P., Creed, P., Verstockt, B., . . . Korcsmáros, T. (2020). P070 Machine learning approaches to identify IBD biomarkers from longitudinal microbiome data. *Journal of Crohn's and Colitis*, S170–S171.

Madsen, A. (2008). Belief Update in CLG Bayesian Networks With Lazy Propagation. *International Journal of Approximate Reasoning*, 503-521.

Mallick, H. M., Franzosa, E. A., Vatanen, T., Morgan, X. C., & Huttenhower, C. (2017). Experimental design and quantitative analysis of microbial community multiomics. *Genome biology*, 228.

Marchesi, J. R. (2011). Towards the human colorectal cancer microbiome. *PloS one*, e20447.

Margaritis, D. (2003). Learning Bayesian network model structure from data. *Doctoral thesis*. Pittsburgh, PA.

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., & Califano, A. (2006). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics*, 7.

Mars, R., Yang, Y., Ward, T., Houtti, M., Priya, S., Lekatz, H. R., . . . ... Kashyap, P. C. (2020). Longitudinal Multi-omics Reveals Subset-Specific Mechanisms Underlying Irritable Bowel Syndrome. *Cell*, 1460–1473.e17.

McDonald, D., Clemente, J. C., Kuczynski, J., Rideout, J. R., Stombaugh, J., Wendel, D., . . . Caporaso, J. G. (2012). The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. *GigaScience*, 7.

McDonald, D., Hyde, E., Debelius, J. W., Morton, J. T., Gonzalez, A., Ackermann, G., . . . Gogul, G. (2018). American Gut: an Open Platform for Citizen Science Microbiome Research. *mSystems*, e00031-18.

McGeachie, M. J., Chang, H. H., & Weiss, S. T. (2014). CGBayesNets: conditional Gaussian Bayesian network learning and inference with mixed discrete and continuous data. *PLoS computational biology*, e1003676.

McGeachie, M., Sordillo, J., Gibson, T., Weinstock, G., Liu, Y., Gold, D., . . . Litonjua, A. (2016). Longitudinal Prediction of the Infant Gut Microbiome with Dynamic Bayesian Networks. *Scientific reports*, 20359.

McGrath, S., & Ghersi, D. (2016). Building towards precision medicine: empowering medical professionals for the next revolution. *BMC medical genomics*, 23.

McGuire, A. L., Gabriel, S., Tishkoff, S. A., Wonkam, A., Chakravarti, A., Furlong, E., . . . Kim, J. S. (2020). The road ahead in genetics and genomics. *The road ahead in genetics and genomics.*, 581–596.

McNulty, N. P., Wu, M., Erickson, A. R., Pan, C., Erickson, B. K., Martens, E. C., . . . Gordon, J. I. (2013). Effects of diet on resource utilization by a model human gut microbiota containing Bacteroides cellulosilyticus WH2, a symbiont with an extensive glycobiome. *PLoS biology*, e1001637.

Metwally, A. A., Yang, J., Ascoli, C., Dai, Y., Finn, P. W., & Perkins, D. L. (2018). MetaLonDA: a flexible R package for identifying time intervals of differentially abundant features in metagenomic longitudinal studies. *Microbiome*, 32.

Metwally, A. A., Yu, P. S., Reiman, D., Dai, Y., Finn, P. W., & Perkins, D. L. (2019). Utilizing longitudinal microbiome taxonomic profiles to predict food allergy via Long Short-Term Memory networks. *PLoS computational biology*.

Moayyeri, A., Hammond, C. J., Hart, D. J., & Spector, T. D. (2013). The UK Adult Twin Registry (TwinsUK Resource). *Twin research and human genetics : the official journal of the International Society for Twin Studies,*, 144–149.

Moitinho-Silva, L., Steinert, G., Nielsen, S., Hardoim, C., Wu, Y. C., McCormack, G. P., . . . Hentschel, U. (2017). Predicting the HMA-LMA Status in Marine Sponges by Machine Learning. *Frontiers in microbiology*, 752.

Moran, M. A. (2015). The global ocean microbiome. *Science*, aac8455.

Moreno-Indias, I., Lahti, L., Nedyalkova, M., Elbere, I., Roshchupkin, G., Adilovic, M., . . . Marcos-Zambrano, L. J. (2021). Statistical and Machine Learning Techniques in Human Microbiome Studies: Contemporary Challenges and Solutions. *Frontiers in microbiology*, 635781.

Mueller, U. G., & Sachs, J. L. (2015). Engineering Microbiomes to Improve Plant and Animal Health. *Trends in microbiology*, 606–617.

Murphy, K. (2002). Dynamic Bayesian Networks:Representation, Inference and Learning. University of California, Berkeley.

Murphy, K., & Mian, S. (1999). *Modelling Gene Expression Data using Dynamic Bayesian Networks.*

Namkung, J. (2020). Machine learning methods for microbiome studies. *Journal of microbiology (Seoul, Korea)*, 206–216.

Niinimaki, T., & Parviainen, P. (2012). Local Structure Discovery in Bayesian Networks. *Proceedings of Uncertainty in Artificial Intelligence, Workshop on Causal Structure Learning*, 634-643.

Novakovic, J., & Veljovic, A. (2011). C-Support Vector Classification: Selection of kernel and parameters in medical diagnosis. *IEEE 9th International Symposium on Intelligent Systems and Informatics*, (págs. 465-470).

Noyes, N., Cho, K. C., Ravel, J., Forney, L. J., & Abdo, Z. (2018). Associations between sexual habits, menstrual hygiene practices, demographics and the vaginal microbiome as revealed by Bayesian network analysis. *PloS one*, e0191625.

Parker, B. J., Wearsch, P. A., Veloo, A., & Rodriguez-Palacios, A. (2020). The Genus Alistipes: Gut Bacteria With Emerging Implications to Inflammation, Cancer, and Mental Health. *Frontiers in immunology*, 906.

Pasolli, E., Schiffer, L., Manghi, P., Renson, A., Obenchain, V., Truong, D. T., . . . Waldron, L. (2017). Accessible, curated metagenomic data through ExperimentHub. *Nature methods*, 1023-1024.

Pasolli, E., Truong, D. T., Malik, F., Waldron, L., & Segata, N. (2016). Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights. *PLoS computational biology*, e1004977.

Paulson, J. N. (2013). Differential abundance analysis for microbial marker-gene surveys. *Nature methods*, 1200–1202.

Paulson, J. N., Stine, O. C., Bravo, H. C., & Pop, M. (2013). Differential abundance analysis for microbial marker-gene surveys. *Nature methods*, 1200–1202.

Paulson, J. N., Talukder, H., & Bravo, H. C. (2017, January 10). *Longitudinal differential abundance analysis of microbial marker-gene surveys using smoothing splines.* Retrieved from bioRxiv: https://www.biorxiv.org/content/10.1101/099457v1

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems.* San Francisco, CA.: Morgan Kaufmann Publishers Inc.

Pearl, J., & Verma, T. (1990). Equivalence and synthesis of causal models. *Proceedings of the sixth conference on uncertainty in artificial intelligence*, (pp. 220-7).

Pearl, J., & Verma, T. (1991). A theory of inferred causation. *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning*, (págs. 441-452). Cambridge, MA, USA.

Pérez-Cobas, A. E., Gosalbes, M. J., Friedrichs, A., Knecht, H., Artacho, A., Eismann, K., . . . Ott, S. J. (2013). Gut microbiota disturbance during antibiotic therapy: a multi-omic approach. *Gut*, 1591–1601.

Peterson, J., Garges, S., Giovanni, M., McInnes, P., Wang, L., Schloss, J. A., . . . Group, N. H. (2009). The NIH Human Microbiome Project. *Genome research*, 2317–2323.

Prifti, E., Chevaleyre, Y., Hanczar, B., Belda, E., Danchin, A., Clément, K., & Zucker, J. D. (2020). Interpretable and accurate prediction models for metagenomics data. *GigaScience*.

Qi, X., Fan, X., Gao, Y., & Liu, Y. (2019). Learning Bayesian network structures using weakest mutual-information-first strategy. *International Journal of Approximate Reasoning*, 84-98.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., . . . ... Wang, J. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 59–65.

Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., . . . Wang, J. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, 55–60.

Qin, N., Yang, F., Li, A., Prifti, E., Chen, Y., Shao, L., . . . ... Li, L. (2014). Alterations of the human gut microbiome in liver cirrhosis. *Nature*, 59–64.

Ramirez, J., Guarner, F., Bustos Fernandez, L., Maruy, A., Sdepanian, V. L., & Cohen, H. (2020). Antibiotics as Major Disruptors of Gut Microbiota. *Frontiers in cellular and infection microbiology*, 572912.

Ramsey, J., Zhang, J., & Spirtes, P. (2006). Adjacency-Faithfulness and Conservative Causal Inference. *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*.

Ravel, J., Gajer, P., Abdo, Z., Schneider, G. M., Koenig, S. S., McCulle, S. L., . . . Forney, L. J. (2011). Vaginal microbiome of reproductive-age women. *Proceedings of the National Academy of Sciences of the United States of America*, 4680–4687.

Ridaura, V. K., Faith, J. J., Rey, F. E., Cheng, J., Duncan, A. E., Kau, A. L., . . . Ursell, L. (2013). Gut microbiota from twins discordant for obesity modulate metabolism in mice. *Science*.

Ridenhour, B. J., Brooker, S. L., Williams, J. E., Van Leuven, J. T., Miller, A. W., Dearing, M. D., & Remien, C. H. (2017). Modeling time-series data from microbial communities. *The ISME journal, 11*, 2526–2537.

Rissanen, J. (1978). Modeling by the shortest data description. *Automatica*, 14:465-471.

Robinson, M. D., & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome biology*, R25.

Romero, R., Hassan, S. S., Gajer, P., Tarca, A. L., Fadrosh, D. W., Nikita, L., . . . Ravel, J. (2014). The composition and stability of the vaginal microbiota of normal pregnant women is different from that of non-pregnant women. *Microbiome*, 4.

Rothenberg, S. E., Wagner, C. L., Hamidi, B., Alekseyenko, A. V., & Azcarate-Peril, M. (2019). Longitudinal changes during pregnancy in gut microbiota and methylmercury biomarkers, and reversal of microbe-exposure correlations. *Environmental research*, 700–712.

Rouhani, S., Griffin, N. W., Yori, P. P., Gehrig, J. L., Olortegui, M. P., Salas, M. S., . . . Gordon, J. I. (2020). Diarrhea as a Potential Cause and Consequence of Reduced Gut Microbial Diversity Among Undernourished Children in Peru. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 989–999.

Ruiz-Perez, D., Lugo-Martinez, J., Bourguignon, N., Mathee, K., Lerner, B., Bar-Joseph, Z., & Narasimhan, G. (2021). Dynamic Bayesian Networks for Integrating Multi-omics Time Series Microbiome Data. *mSystems*, e01105-20.

Sampson, T. R., Debelius, J. W., Thron, T., Janssen, S., Shastri, G. G., Ilhan, Z. E., . . . Knight, R. (2016). Gut Microbiota Regulate Motor Deficits and Neuroinflammation in a Model of Parkinson's Disease. *Cell*, 1469–1480.

Sanna, S., van Zuydam, N. R., Mahajan, A., Kurilshikov, A., Vich Vila, A., Võsa, U., . . . McCarthy, M. I. (2019). Causal relationships among the gut microbiome, short-chain fatty acids and metabolic diseases. *Nature*, 600–605.

Sarrabayrouse, G., Elias, A., Yáñez, F., Mayorga, L., Varela, E., Bartoli, C., . . . Manichanh, C. (2021). Fungal and Bacterial Loads: Noninvasive Inflammatory Bowel Disease Biomarkers for the Clinical Setting. *mSystems*, e01277-20.

Sazal, M., Mathee, K., Ruiz-Perez, D., Cickovski, T., & Narasimhan, G. (2020). Inferring directional relationships in microbial communities using signed Bayesian networks. *BMC genomics*, 663.

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., . . . Weber, C. F. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and environmental microbiology*, 7537–7541.

Scutari, M. (2010). Learning Bayesian Networks with the bnlearn R package. *Journal of Statistical Software*, 1-22.

Sebastiani, P., Abad, M., & Ramoni, M. (2005). Bayesian Networks for Genomic Analysis. *Genomic Signal Processing and Statistics*, 281-320.

Segal, J. P., Mullish, B. H., Quraishi, M. N., Acharjee, A., Williams, H., Iqbal, T., . . . Marchesi, J. R. (2019). The application of omics techniques to understand the role of the gut microbiota in inflammatory bowel disease. *Therapeutic advances in gastroenterology*.

Shachter, R., & Peot, M. (1990). Simulation Approaches to General Probabilistic Inference on Belief Networks. In *Uncertainty in Artificial Intelligence* (pp. 221-231).

Shannon, P. M. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. . *Genome research*, 2498–2504.

Shannon, P., Markiel, A., Ozier, O., Baliga, N., Wang, J., Ramage, D., . . . T, I. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 498-504.

Shomorony, I., Cirulli, E. T., Huang, L., Napier, L. A., Heister, R. R., Hicks, M., . . . Karow, D. S. (2020). An unsupervised learning approach to identify novel signatures of health and disease from multimodal data. *Genome medicine*, 7.

Sikalidis, A. K., & Maykish, A. (2020). The Gut Microbiome and Type 2 Diabetes Mellitus: Discussing a Complex Relationship. *Biomedicines*, 8.

Silverman, G. (2019). The microbiome in SLE pathogenesis. *Nature reviews. Rheumatology*, 72–74.

Smith, G. D., & Ebrahim, S. (2003). 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International journal of epidemiology*, 1-22.

Sondhi, A., & Shojaie, A. (2019). The Reduced PC-Algorithm: Improved Causal Structure Learning in Large Random Networks. *Journal of Machine Learning Research*.

Spirtes, P., Glymour, C., & Scheines, R. (1993). Causation, prediction, and search. En *Lecture Notes in Statistics* (pág. volume 81). New York: Springer-Verlag.

Spirtes, P., Glymour, C., & Scheines, R. (2000). Causation, Prediction, and Search. Cambridge: MIT Press.

Straube, J. G. (2015). A Linear Mixed Model Spline Framework for Analysing Time Course 'Omics' Data. . *PloS one*, e0134540.

Su, X., Jing, G., Sun, Z., Liu, L., Xu, Z., McDonald, D., . . . Xu, J. (2020). Multiple-Disease Detection and Classification across Cohorts via Microbiome Search. *mSystems*, e00150-20.

Su, X., Jing, G., Zhang, Y., & Wu, S. (2020). Method development for cross-study microbiome data mining: Challenges and opportunities. *Computational and structural biotechnology journal*, 2075–2080.

Theriot, C. M., Koenigsknecht, M. J., Carlson, P. E., Jr, H. G., Nelson, A. M., Li, B., . . . Young, V. B. (2014). Antibiotic-induced shifts in the mouse gut microbiome and metabolome increase susceptibility to Clostridium difficile infection. *Nature communications*, 3114.

Thomas, A. M., Manghi, P., Asnicar, F., Pasolli, E., Armanini, F., Zolfo, M., . . . Shiba. (2019). Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nature medicine*, 667–678.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society.*, 267-288.

Topçuoğlu, B. D., Lesniak, N. A., Ruffin, M. T., Wiens, J., & Schloss, P. D. (2020). A Framework for Effective Application of Machine Learning to Microbiome-Based Classification Problems. *mBio*, e00434-20.

Trivedi, P., Leach, J. E., Tringe, S. G., Sa, T., & Singh, B. K. (2020). Plant-microbiome interactions: from community assembly to plant health. *Nature reviews. Microbiology*, 607–621.

Tsamardinos, I., & Aliferis, A. (2003). Algorithms for large scale markov blanket discovery. *Proceedings of the Sixteenth International Florida Artificial Intelligence Research Society Conference.* Florida, USA.

Tsamardinos, I., Brown, L., & Aliferis, C. (2006). The max–min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 31-78.

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., & Gordon, J. I. (2007). The human microbiome project. *Nature*, 804–810.

Turnbaugh, P. J., Ridaura, V. K., Faith, J. J., Rey, F. E., Knight, R., & Gordon, J. I. (2009). The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice. *Science translational medicine*, 6ra14.

Ursell, L. K., Metcalf, J. L., Wegener, L., & Knight, R. (2012). Defining the human microbiome. *Nutrition Reviews*, S38-S44.

van Gerven, M. A. (2008). Dynamic Bayesian networks as prognostic models for clinical patient management. *Journal of biomedical informatics*, 515–529.

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

Vanwonterghem, I., Jensen, P. D., Ho, D. P., Batstone, D. J., & Tyson, G. W. (2014). Linking microbial community structure, interactions and function in anaerobic digesters using new molecular techniques. *Current opinion in biotechnology*, 55–64.

Vatanen, T., Kostic, A. D., d'Hennezel, E., Siljander, H., Franzosa, E. A., Yassour, M., . . . Szabo, S. J. (2016). Variation in Microbiome LPS Immunogenicity Contributes to Autoimmunity in Humans. *Cell*, 842–853.

Vatanen, T., Kuusela, M., Malmi, E., Raiko, T., Aaltonen, T., & Nagai, Y. (2012). Semi-supervised detection of collective anomalies with an application in high energy particle physics. *The 2012 International Joint Conference on Neural Networks*, 1-8.

Verdi, S. A. (2019). TwinsUK: The UK Adult Twin Registry Update. *Twin research and human genetics : the official journal of the International Society for Twin Studies*, 523–529.

Voigt, A. Y. (2015). Temporal and technical variability of human gut metagenomes. *Genome biology*, 73.

Wang, J. W., Kuo, C. H., Kuo, F. C., Wang, Y. K., Hsu, W. H., Yu, F. J., . . . Wu, D. C. (2019). Fecal microbiota transplantation: Review and update. *Journal of the Formosan Medical Association = Taiwan yi zhi*, S23–S31.

Wang, J., & Jia, H. (2016). Metagenome-wide association studies: fine-mining the microbiome. *Nature reviews. Microbiology*, 508–522.

Wang, Q., Wang, K., Wu, W., Giannoulatou, E., Ho, J., & Li, L. (2019). Host and microbiome multi-omics integration: applications and methodologies. *Biophysical reviews*, 55–65.

Wang, Z. G., Gao, P., Pu, Q., Lin, P., Qin, S., Xie, N., . . . Wu, M. (2021). Microbial and genetic-based framework identifies drug targets in inflammatory bowel disease. *Theranostics*, 7491–7506.

Weersma, R. K., Zhernakova, A., & Fu, J. (2020). Interaction between drugs and the gut microbiome. *Gut*, 1510–1519.

Weng, M., & Walker, W. A. (2013). The role of gut microbiota in programming the immune phenotype. *Journal of developmental origins of health and disease*, 203–214.

Wilczyński, B., & Dojer, N. (2009). BNFinder: exact and efficient method for learning Bayesian networks. *Bioinformatics (Oxford, England)*, 286–287.

Wilmes, P., & Bond, P. L. (2004). The application of two-dimensional polyacrylamide gel electrophoresis and downstream analyses to a mixed community of prokaryotic microorganisms. *Environmental microbiology*, 911–920.

Wirbel, J., Pyl, P. T., Kartal, E., Zych, K., Kashani, A., Milanese, A., . . . Gandini, S. (2019). Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nature medicine*, 679-689.

Wong, C. W., Yost, S. E., Lee, J. S., Gillece, J. D., Folkerts, M., Reining, L., . . . Yuan, Y. (2021). Analysis of Gut Microbiome Using Explainable Machine Learning Predicts Risk of Diarrhea Associated With Tyrosine Kinase Inhibitor Neratinib: A Pilot Study. *Frontiers in oncology*, 604584.

Wright, E. K., Kamm, M. A., Teo, S. M., Inouye, M., Wagner, J., & Kirkwood, C. D. (2015). Recent advances in characterizing the gastrointestinal microbiome in Crohn's disease: a systematic review. *Inflammatory bowel diseases*, 1219-1228.

Wu, S., Sun, C., Li, Y., Wang, T., Jia, L., Lai, S., . . . Chen, W. H. (2020). GMrepo: a database of curated and consistently annotated human gut metagenomes. *Nucleic acids research*, D545–D553.

Y, X., J, S., D, C., Y, X., J, S., & D., C. (2018). Multivariate community analysis. *Statistical Analysis of Microbiome Data with R*, 285-330.

Yang, J. Y., Karr, J. R., Watrous, J. D., & Dorrestein, P. C. (2011). Integrating '-omics' and natural product discovery platforms to investigate metabolic exchange in microbiomes. *Current opinion in chemical biology*, 79–87.

Yaramakala, S., & Margaritis, D. (2005). Speculative MarkovBlanket DiscoveryforOptimalFeature Selection. *Proceedings of the Fifth IEEE International Conference on Data Mining.*

Yugi, K., Kubota, H., Hatano, A., & Kuroda, S. (2016). Trans-Omics: How To Reconstruct Biochemical Networks Across Multiple 'Omic' Layers. *Trends in biotechnology*, 276–290.

Zeller, G., Tap, J., Voigt, A. Y., Sunagawa, S., Kultima, J. R., Costea, P. I., . . . Yama. (2014). Potential of fecal microbiota for early-stage detection of colorectal cancer. *Molecular systems biology*, 766.

Zhou, W., Sailani, M. R., Contrepois, K., Zhou, Y., Ahadi, S., Leopold, S. R., . . . Petersen, L. (2019). Longitudinal multi-omics of host-microbe dynamics in prediabetes. *Nature*, 663–671.

Zhou, Y. H., & Gallins, P. (2019). A Review and Tutorial of Machine Learning Methods for Microbiome Host Trait Prediction. *Frontiers in genetics*, 579.

Zhu, J., Zhang, B., Smith, E. N., Drees, B., Brem, R. B., Kruglyak, L., . . . Schadt, E. E. (2008). Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nature genetics*, 854–861.

Zou, C., Denby, K. J., & Feng, J. (2009). Granger causality vs. dynamic Bayesian network inference: a comparative study. *BMC bioinformatics*, 122.

Zou, H., & Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society.*, 301-320.

# 7 Appendix

## I. Theoretical Foundation of CGBayesNet (McGeachie, Chang, & Weiss, 2014)

To determine the best network model of the data, CGBayesNet computes the marginal likelihood of candidate network structures, conditioned upon the data, and choose the network model that maximizes the marginal likelihood. The posterior probability of the Bayesian network model G, given the data D, is p(G|D)p(D), and it uses Bayes' theorem to equate p(G|D)p(D) = p(D|G)p(G), or :

$$p(G|D) \propto p(D|G)p(D),$$

where p(G) is the prior probability of a network model and p(D) is the prior probability of the data, and p(D|G) is the marginal likelihood:

$$p(D|G) = \int p(D|\theta, G)p(\theta|G)d\theta.$$

Here p(D| θ,G) is the likelihood of the data given the network G and distribution parameters θ, and p(θ|G) is the prior density of the parameters θ. The marginal likelihood p(D|G) is computed by averaging out the distribution parameters θ from the likelihood function, p(D|G, θ).

The Bayesian network semantics provides a decomposition of the likelihood as follows: for a given set of distribution parameters θ, a dataset D of size |D| = d, variables $y_i$ in I = (Δ union Ψ) realizing values $y_{ik}$ in {$y_{i1}$, $y_{i2}$, … $y_{id}$} in D, given parents π($y_i$) taking values $u_{ik}$ when $y_i$ takes value $y_{ik}$:

$$p(D|G,\theta) \propto \left[\prod_{i\in\mathrm{I}} \prod_{k=1}^{d} p(y_{ik}|\pi(y_{ik}) = u_{ik}, \theta_{ik})\right]$$

where p($y_{ik}$| π($y_i$),$\theta_{ik}$) is the probability of $y_i$ having value $y_{ik}$ in D with parent values $u_{ik}$ and distribution parameters θ. Distribution parameters for discrete nodes are modeled with Dirichlet priors, priors for Gaussian nodes are described below. In the discrete case, we denote by |$y_i$| and |π($y_i$)| the number of different values that $y_i$ and π($y_i$) can assume, respectively; then the discrete nodes have (joint) likelihood:

$$P(\Delta|\theta) = \prod_{i\in\Delta} \prod_{j}^{|\pi(y_i)|} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + n_{ij})} \prod_{k}^{|y_i|} \frac{\Gamma(\alpha_{ijk} + n_{ijk})}{\Gamma(\alpha_{ijk})}$$

where $n_{ijk}$ is the number of data points satisfying $y_i$ = k for π($y_i$) in configuration j, and $\alpha_{ijk}$ is the hyper parameter of the Dirichlet distribution indicating a prior assumed sample size. Γ(.) denotes the gamma function. Continuous nodes $y_i$ have Gaussian distributions with a mean that is a linear function of its continuous parents and depending on its discrete parents, with a conditional variance $\sigma^2_{ij}$ = 1/$\tau_{ij}$. The joint likelihood of the continuous nodes is then

$$p(\Psi|\theta) = \prod_{i\in\Psi} \left(\frac{\tau_{ij}}{2\pi}\right)^{n/2} e^{\left[\left(-{}^{\tau_{ij}}/_{2}\right)(y_{ik}-X_i\beta_{ij})^T (y_{ik}-X_i\beta_{ij})\right]}$$

with $x_{ij}$ the values of continuous parents of $y_i$ in case k, and $\beta_{ij}$ the vector of regression parameters given discrete parents of $y_i$ = j. CGBayesNet follows (Sebastiani, Abad, & Ramoni, 2005) and use a Gamma prior distribution for τ and a conditional multivariate Gaussian prior density on regression parameters β. Thus,

$$\tau_{ij} \sim \Gamma(\alpha_{i1}, \alpha_{i2}), \qquad p(\tau) = \frac{\tau_{ij}^{\alpha_{i1}-1} e^{-\tau_{ij}/\alpha_{i2}}}{\alpha_{i2}{}^{\alpha_{i1}}\Gamma(\alpha_{i1})}$$

And β is described by

$$\beta_{ij} | \tau_{ij} \sim N\left(\beta_{ij0}, \left(\tau_{ij} I\right)^{-1}\right)$$

For the identity matrix I, and $\beta_{ij0} = E(\beta_{ij} \mid \tau_{ij})$. The above equations represent the main semantics of CGBayesNets.

## II. Number of samples per Omic type for each participant

| | **Number of samples per omic type** | | | |
|---|---|---|---|---|
| **Participant ID** | **Metabolites** | **Metagenomics** | **Metatranscriptomics** | **Metatranscriptomics trimmed** |
| C3001 | 6 | 16 | 6 | 6 |
| C3002 | 6 | 15 | 7 | 7 |
| C3003 | 5 | 10 | 7 | 7 |
| C3004 | 5 | 24 | 20 | 20 |
| C3005 | 5 | 12 | 9 | 9 |
| C3006 | 5 | 11 | 7 | 6 |
| C3008 | 4 | 13 | 5 | 5 |
| C3009 | 6 | 12 | 8 | 6 |
| C3010 | 6 | 14 | 7 | 7 |
| C3011 | 5 | 22 | 5 | 5 |
| C3012 | 6 | 14 | 8 | 8 |
| C3013 | 5 | 22 | 5 | 5 |
| C3015 | 5 | 23 | 15 | 15 |
| C3016 | 7 | 20 | 15 | 14 |
| C3017 | 6 | 23 | 5 | 5 |
| C3021 | 4 | 9 | 7 | 7 |
| C3022 | 6 | 22 | 7 | 7 |
| C3023 | 7 | 12 | 7 | 7 |
| C3027 | 6 | 24 | 8 | 7 |
| C3028 | 5 | 12 | 5 | 5 |
| C3031 | 4 | 11 | 9 | 9 |
| C3034 | 5 | 10 | 6 | 6 |
| C3035 | 5 | 12 | 8 | 8 |

| | | | | |
|---|---|---|---|---|
| C3037 | 8 | 11 | 7 | 6 |
| E5001 | 5 | 13 | 9 | 9 |
| E5004 | 5 | 12 | 5 | 4 |
| E5009 | 0 | 11 | 4 | 0 |
| E5013 | 5 | 13 | 7 | 7 |
| H4001 | 8 | 12 | 5 | 5 |
| H4004 | 7 | 11 | 8 | 8 |
| H4006 | 6 | 23 | 19 | 19 |
| H4007 | 0 | 12 | 4 | 0 |
| H4008 | 6 | 24 | 8 | 8 |
| H4009 | 6 | 23 | 8 | 7 |
| H4010 | 7 | 13 | 7 | 6 |
| H4013 | 6 | 14 | 7 | 7 |
| H4014 | 6 | 12 | 8 | 8 |
| H4015 | 6 | 22 | 7 | 7 |
| H4016 | 6 | 14 | 5 | 4 |
| H4017 | 4 | 18 | 12 | 10 |
| H4018 | 6 | 13 | 7 | 7 |
| H4019 | 6 | 24 | 13 | 13 |
| H4020 | 5 | 23 | 18 | 18 |
| H4022 | 5 | 14 | 5 | 5 |
| H4023 | 5 | 22 | 8 | 8 |
| H4024 | 4 | 23 | 7 | 7 |
| H4027 | 5 | 11 | 5 | 5 |
| H4030 | 5 | 13 | 8 | 7 |
| H4031 | 5 | 12 | 7 | 7 |
| H4032 | 5 | 13 | 4 | 4 |
| H4035 | 7 | 24 | 18 | 16 |
| H4038 | 6 | 13 | 6 | 6 |
| H4039 | 4 | 13 | 9 | 9 |
| H4040 | 4 | 11 | 7 | 7 |
| H4043 | 5 | 10 | 4 | 4 |
| H4045 | 6 | 13 | 7 | 7 |
| M2008 | 6 | 17 | 12 | 12 |
| M2014 | 6 | 14 | 7 | 7 |
| M2021 | 4 | 8 | 5 | 5 |
| M2025 | 5 | 11 | 9 | 9 |
| M2026 | 5 | 17 | 7 | 7 |
| M2027 | 5 | 13 | 12 | 12 |
| M2028 | 5 | 20 | 5 | 5 |
| M2034 | 5 | 20 | 6 | 6 |
| M2039 | 5 | 15 | 10 | 10 |
| M2041 | 5 | 14 | 9 | 9 |
| M2042 | 5 | 23 | 8 | 8 |

| | | | | |
|---|---|---|---|---|
| M2047 | 6 | 14 | 10 | 9 |
| M2060 | 4 | 12 | 5 | 4 |
| M2061 | 7 | 13 | 10 | 10 |
| M2064 | 5 | 22 | 10 | 10 |
| M2068 | 4 | 26 | 8 | 8 |
| M2069 | 6 | 25 | 21 | 21 |
| M2071 | 5 | 10 | 9 | 9 |
| M2072 | 5 | 24 | 13 | 13 |
| M2075 | 6 | 11 | 5 | 5 |
| M2077 | 6 | 14 | 9 | 8 |
| M2079 | 5 | 14 | 8 | 8 |
| M2083 | 5 | 18 | 7 | 7 |
| M2084 | 6 | 23 | 8 | 8 |
| M2085 | 7 | 14 | 8 | 8 |
| M2097 | 5 | 11 | 7 | 7 |
| P6005 | 6 | 16 | 9 | 9 |
| P6009 | 6 | 22 | 7 | 7 |
| P6010 | 6 | 23 | 19 | 19 |
| P6013 | 4 | 23 | 6 | 6 |
| P6016 | 6 | 14 | 8 | 8 |
| P6018 | 5 | 25 | 17 | 17 |
| P6024 | 4 | 11 | 4 | 4 |
| P6028 | 6 | 11 | 8 | 6 |
| P6035 | 4 | 11 | 7 | 7 |
| P6037 | 4 | 8 | 7 | 7 |
| P6038 | 5 | 12 | 8 | 8 |
| **Total** | **492** | **1448** | **777** | **749** |

## III List of DBN nodes

| diagnosis: CD | | |
|---|---|---|
| **Metagenomics** | **Metatranscriptomics** | **metabolites** |
| [1] "s__Alistipes_putredinis" | [1] "g__ASPASN-PWY: superpathway of L-aspartate and L-asparagine biosynthesis\|g__Bacteroides.s__Bacteroides_finegoldii" | [1] "m__Arg - Cholate" |
| [2] "s__Alistipes_sp_HGB5" | [2] "g__COA-PWY-1: coenzyme A biosynthesis II (mammalian)\|g__Bacteroides.s__Bacteroides_salyersiae" | [2] "m__C10 carnitine" |
| [3] "s__Anaerococcus_obesiensis" | [3] "g__COA-PWY-1: coenzyme A biosynthesis II (mammalian)\|g__Dysgonomonas.s__Dysgonomonas_mossii" | [3] "m__C12 carnitine" |
| [4] "s__Anaerococcus_vaginalis" | [4] "g__COA-PWY-1: coenzyme A biosynthesis II (mammalian)\|g__Ruminococcus.s__Ruminococcus_callidus" | [4] "m__C12:1 carnitine" |
| [5] "s__Bacteroides_cellulosilyticus" | [5] "g__HISDEG-PWY: L-histidine degradation I\|g__Synergistes.s__Synergistes_sp_3_1_syn1" | [5] "m__C14:0 SM" |
| [6] "s__Bacteroides_eggerthii" | [6] "g__NONMEVIPP-PWY: methylerythritol phosphate pathway I\|g__Bacteroides.s__Bacteroides_sp_2_1_22" | [6] "m__C14:1 carnitine" |
| [7] "s__Bacteroides_fluxus" | [7] "g__NONMEVIPP-PWY: methylerythritol phosphate pathway I\|g__Dysgonomonas.s__Dysgonomonas_mossii" | [7] "m__C14:2 carnitine" |
| [8] "s__Bacteroides_fragilis" | [8] "g__PANTO-PWY: phosphopantothenate biosynthesis I\|g__Alistipes.s__Alistipes_sp_HGB5" | [8] "m__C16-OH carnitine" |
| [9] "s__Bacteroides_massiliensis" | [9] "g__PANTO-PWY: phosphopantothenate biosynthesis I\|g__Dysgonomonas.s__Dysgonomonas_gadei" | [9] "m__C18:1 LPC plasmalogen" |
| [10] "s__Bacteroides_ovatus" | [10] "g__PANTO-PWY: phosphopantothenate biosynthesis I\|g__Dysgonomonas.s__Dysgonomonas_mossii" | [10] "m__C18:1-OH carnitine" |
| [11] "s__Bacteroides_sp_1_1_30" | [11] "g__PEPTIDOGLYCANSYN-PWY: peptidoglycan biosynthesis I (meso-diaminopimelate containing)\|g__Collinsella.s__Collinsella_aerofaciens" | [11] "m__C20:1 LPC" |
| [12] "s__Bacteroides_sp_2_1_22" | [12] "g__PWY-1042: glycolysis IV (plant cytosol)\|g__Megamonas.s__Megamonas_funiformis" | [12] "m__C22:6 LPC" |
| [13] "s__Bacteroides_sp_4_3_47FAA" | [13] "g__PWY-2942: L-lysine biosynthesis III\|g__Dysgonomonas.s__Dysgonomonas_gadei" | [13] "m__C38:3 PC" |
| [14] "s__Bacteroides_sp_9_1_42FAA" | [14] "g__PWY-2942: L-lysine biosynthesis III\|g__Dysgonomonas.s__Dysgonomonas_mossii" | [14] "m__C38:4 PC plasmalogen" |
| [15] "s__Barnesiella_intestinihominis" | [15] "g__PWY-3001: superpathway of L-isoleucine biosynthesis I\|g__Megamonas.s__Megamonas_hypermegale" | [15] "m__C38:4 PC" |

| | | |
|---|---|---|
| [16] "s__Bifidobacterium_adolescentis" | [16] "g__PWY-5097: L-lysine biosynthesis VI\|g__Dysgonomonas.s__Dysgonomonas_gadei" | [16] "m__C38:6 PC" |
| [17] "s__Butyricicoccus_pullicaecorum" | [17] "g__PWY-5097: L-lysine biosynthesis VI\|g__Dysgonomonas.s__Dysgonomonas_mossii" | [17] "m__C4 carnitine" |
| [18] "s__Butyrivibrio_crossotus" | [18] "g__PWY-5188: tetrapyrrole biosynthesis I (from glutamate)\|g__Ruminococcus.s__Ruminococcus_callidus" | [18] "m__C46:1 TAG" |
| [19] "s__Clostridium_bolteae" | [19] "g__PWY-5659: GDP-mannose biosynthesis\|g__Alistipes.s__Alistipes_sp_HGB5" | [19] "m__C46:2 TAG" |
| [20] "s__Collinsella_aerofaciens" | [20] "g__PWY-5667: CDP-diacylglycerol biosynthesis I\|g__Alistipes.s__Alistipes_sp_HGB5" | [20] "m__C48:3 TAG" |
| [21] "s__Collinsella_intestinalis" | [21] "g__PWY-5667: CDP-diacylglycerol biosynthesis I\|g__Bacteroides.s__Bacteroides_finegoldii" | [21] "m__C50:5 TAG" |
| [22] "s__Coprococcus_eutactus" | [22] "g__PWY-5667: CDP-diacylglycerol biosynthesis I\|g__Dysgonomonas.s__Dysgonomonas_gadei" | [22] "m__C52:6 TAG" |
| [23] "s__Coprococcus_sp_ART55_1" | [23] "g__PWY-5667: CDP-diacylglycerol biosynthesis I\|g__Dysgonomonas.s__Dysgonomonas_mossii" | [23] "m__C52:7 TAG" |
| [24] "s__Desulfovibrio_desulfuricans" | [24] "g__PWY-5686: UMP biosynthesis\|g__Bacteroides.s__Bacteroides_finegoldii" | [24] "m__C58:10 TAG" |
| [25] "s__Dorea_longicatena" | [25] "g__PWY-5690: TCA cycle II (plants and fungi)\|g__Megamonas.s__Megamonas_funiformis" | [25] "m__Citrulline - Cholate" |
| [26] "s__Dysgonomonas_gadei" | [26] "g__PWY-5695: urate biosynthesis/inosine 5'-phosphate degradation\|g__Dysgonomonas.s__Dysgonomonas_gadei" | [26] "m__Citrulline - chenodeoxycholate" |
| [27] "s__Dysgonomonas_mossii" | [27] "g__PWY-5695: urate biosynthesis/inosine 5'-phosphate degradation\|g__Dysgonomonas.s__Dysgonomonas_mossii" | [27] "m__Isoallolithocholate [2M-H]-" |
| [28] "s__Erysipelotrichaceae_bacterium_5_2_54FAA" | [28] "g__PWY-6121: 5-aminoimidazole ribonucleotide biosynthesis I\|g__Alistipes.s__Alistipes_sp_HGB5" | [28] "m__N1-acetylspermine" |
| [29] "s__Eubacterium_biforme" | [29] "g__PWY-6147: 6-hydroxymethyl-dihydropterin diphosphate biosynthesis I\|g__Bacteroides.s__Bacteroides_sp_2_1_22" | [29] "m__NH4_C22:1 MAG" |
| [30] "s__Eubacterium_dolichum" | [30] "g__PWY-6386: UDP-N-acetylmuramoyl-pentapeptide biosynthesis II (lysine-containing)\|g__Dysgonomonas.s__Dysgonomonas_mossii" | [30] "m__NH4_C46:3 TAG" |
| [31] "s__Haemophilus_pittmaniae" | [31] "g__PWY-6387: UDP-N-acetylmuramoyl-pentapeptide biosynthesis I (meso-diaminopimelate containing)\|g__Dysgonomonas.s__Dysgonomonas_mossii" | [31] "m__NH4_C48:3 TAG" |
| [32] "s__Lautropia_mirabilis" | [32] "g__PWY-6700: queuosine biosynthesis\|g__Dysgonomonas.s__Dysgonomonas_mossii" | [32] "m__NH4_C56:1 TAG" |
| [33] "s__Megamonas_funiformis" | [33] "g__PWY-6703: preQ0 biosynthesis\|g__Dysgonomonas.s__Dysgonomonas_mossii" | [33] "m__Val - chenodeoxycholate" |

<table>
<tr><td>[34] "s__Megamonas_rupellensis"</td><td>[34] "g__PWY-7111: pyruvate fermentation to isobutanol (engineered)|g__Haemophilus.s__Haemophilus_pittmaniae"</td><td>[34] "m__alpha-hydroxymetoprolol"</td></tr>
<tr><td>[35] "s__Odoribacter_laneus"</td><td>[35] "g__PWY-7199: pyrimidine deoxyribonucleosides salvage|g__Alistipes.s__Alistipes_sp_HGB5"</td><td>[35] "m__biotin"</td></tr>
<tr><td>[36] "s__Odoribacter_splanchnicus"</td><td>[36] "g__PWY-7208: superpathway of pyrimidine nucleobases salvage|g__Desulfovibrio.s__Desulfovibrio_desulfuricans"</td><td>[36] "m__cinnamoylglycine"</td></tr>
<tr><td>[37] "s__Parabacteroides_goldsteinii"</td><td>[37] "g__PWY-7219: adenosine ribonucleotides de novo biosynthesis|g__Alistipes.s__Alistipes_sp_HGB5"</td><td>[37] "m__dTMP"</td></tr>
<tr><td>[38] "s__Parabacteroides_sp_D13"</td><td>[38] "g__PWY-7219: adenosine ribonucleotides de novo biosynthesis|g__Dysgonomonas.s__Dysgonomonas_gadei"</td><td>[38] "m__dihydroorotate"</td></tr>
<tr><td>[39] "s__Paraprevotella_xylaniphila"</td><td>[39] "g__PWY-7219: adenosine ribonucleotides de novo biosynthesis|g__Dysgonomonas.s__Dysgonomonas_mossii"</td><td>[39] "m__epiandrosterone"</td></tr>
<tr><td>[40] "s__Peptoniphilus_duerdenii"</td><td>[40] "g__PWY-7219: adenosine ribonucleotides de novo biosynthesis|g__Synergistes.s__Synergistes_sp_3_1_syn1"</td><td>[40] "m__furosemide"</td></tr>
<tr><td>[41] "s__Phascolarctobacterium_succinatutens"</td><td>[41] "g__PWY-7221: guanosine ribonucleotides de novo biosynthesis|g__Dysgonomonas.s__Dysgonomonas_mossii"</td><td>[41] "m__furoylglycine"</td></tr>
<tr><td>[42] "s__Prevotella_stercorea"</td><td>[42] "g__PWY-7228: superpathway of guanosine nucleotides de novo biosynthesis I|g__Dysgonomonas.s__Dysgonomonas_mossii"</td><td>[42] "m__gabapentin"</td></tr>
<tr><td>[43] "s__Prevotella_timonensis"</td><td>[43] "g__PWY0-1319: CDP-diacylglycerol biosynthesis II|g__Alistipes.s__Alistipes_sp_HGB5"</td><td>[43] "m__guanidoacetic acid"</td></tr>
<tr><td>[44] "s__Pseudomonas_aeruginosa"</td><td>[44] "g__PWY0-1319: CDP-diacylglycerol biosynthesis II|g__Bacteroides.s__Bacteroides_finegoldii"</td><td>[44] "m__heptanoate"</td></tr>
<tr><td>[45] "s__Ruminococcus_bromii"</td><td>[45] "g__PWY0-1319: CDP-diacylglycerol biosynthesis II|g__Dysgonomonas.s__Dysgonomonas_gadei"</td><td>[45] "m__hydrochlorothiazide"</td></tr>
<tr><td>[46] "s__Ruminococcus_callidus"</td><td>[46] "g__PWY0-1319: CDP-diacylglycerol biosynthesis II|g__Dysgonomonas.s__Dysgonomonas_mossii"</td><td>[46] "m__nicotinuric acid"</td></tr>
<tr><td>[47] "s__Ruminococcus_champanellensis"</td><td>[47] "g__PYRIDOXSYN-PWY: pyridoxal 5'-phosphate biosynthesis I|g__Dysgonomonas.s__Dysgonomonas_mossii"</td><td>[47] "m__porphobilinogen"</td></tr>
<tr><td>[48] "s__Ruminococcus_obeum"</td><td>[48] "g__UNINTEGRATED|g__Clostridium.s__Clostridium_symbiosum"</td><td>[48] "m__quinine"</td></tr>
<tr><td>[49] "s__Streptococcus_intermedius"</td><td>[49] "g__UNINTEGRATED|g__Lachnospiraceae_noname.s__Lachnospiraceae_bacterium_2_1_58FAA"</td><td>[49] "m__sulfapyridine"</td></tr>
<tr><td>[50] "s__Synergistes_sp_3_1_syn1"</td><td>[50] "g__VALSYN-PWY: L-valine biosynthesis|g__Haemophilus.s__Haemophilus_pittmaniae"</td><td>[50] "m__xanthurenate"</td></tr>
</table>

# 8 Glossary

- **Microbe**: microorganisms. For example: bacteria, eucarya and archea.

- **Microbiota:** a collection or community of microbes.

- **Human microbiota**: the particular community of microbes residing in and on the human body. Organism-level.

- **Human microbiome**: the term is how scientists refer to microbial genes living in the human body. The full collection of all the genes which are contained in the human microbial community. Gene-level. Consists of about 100 trillion microbial cells (vs 10 trillion human cells).

- **Dysbiosis**: disrupted microbial ecosystem. Can lead to a variety of human disease states.

- **Transcriptomics pathway abundance:** The abundance of a pathway in the sample is computed as a function of the abundances of the pathway's component reactions, with each reaction's abundance computed as the sum over abundances of genes catalyzing the reaction. The abundance is proportional to the number of complete "copies" of the pathway in the community. Unlike gene abundance, a pathway's abundance at community-level is not necessarily the sum of the abundance values of each species. Gene family and pathway abundances are in RPKs (reads per kilobase), accounting for gene length but not sample sequencing depth.