mathematics

MDPI

Article

Explainable Machine Learning for Longitudinal Multi-Omic Microbiome

Paula Laccourreye [1,*], Concha Bielza [2] and Pedro Larrañaga [2]

[1] Digital Health & Biomedical Technologies, Vicomtech Foundation, Basque Research and Technology Alliance (BRTA), 20009 Donostia-San Sebastián, Spain

[2] Artificial Intelligence Department, Universidad Politécnica de Madrid, 28660 Madrid, Spain; mcbielza@fi.upm.es (C.B.); pedro.larranaga@fi.upm.es (P.L.)

* Correspondence: placcourreye@vicomtech.org

Abstract: Over the years, research studies have shown there is a key connection between the microbial community in the gut, genes, and immune system. Understanding this association may help discover the cause of complex chronic idiopathic disorders such as inflammatory bowel disease. 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 and relationships between microbes in the human host. Our study aims to extend the current knowledge of associations between the human microbiome and health and disease through the application of dynamic Bayesian networks to describe the temporal variation of the gut microbiota and dynamic relationships between taxonomic entities and clinical variables. We develop a set of preprocessing steps to clean, filter, select, integrate, and model informative metagenomics, metatranscriptomics, and metabolomics longitudinal data from the Human Microbiome Project. This study accomplishes novel network models with satisfactory predictive performance (accuracy = 0.648) for each inflammatory bowel disease state, validating Bayesian networks as a framework for developing interpretable models to help understand the basic ways the different biological entities (taxa, genes, metabolites) interact with each other in a given environment (human gut) over time. These findings can serve as a starting point to advance the discovery of novel therapeutic approaches and new biomarkers for precision medicine.

Keywords: computational methods; bioinformatics; Bayesian networks; human microbiome; omics; machine learning; interpretable artificial intelligence

MSC: 62H22

check for updates

Citation: Laccourreye, P.; Bielza, C.; Larrañaga, P. Explainable Machine Learning for Longitudinal Multi-Omic Microbiome. *Mathematics* **2022**, *10*, 1994. https://doi.org/10.3390/math10121994

Academic Editors: Manuel Franco, Juana María Vivo and Junseok Kim

Received: 30 April 2022
Accepted: 2 June 2022
Published: 9 June 2022

1. Introduction

Although microbiome research is currently being studied for many applications, such as in ecology, agriculture, biotechnology, and plant health [1–4], there is a particular growing interest in medicine to understand how the community of bacteria in the human body shapes 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. 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 States [5] or MetaHIT [6] funded by the European Commission. 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 statuses, such as obesity, autoimmune disorders, asthma/allergies, diabetes, *C. difficile* infection, and colorectal cancer among many others, generating significant attention in clinical applications for current and emerging diseases. This is, among others, due to the increasing published studies proving that the dysbiosis of microbes in different parts of the human body (oral, skin, gut, vaginal, etc.) is related to numerous health conditions and their risk and severity [7–15].

1.1. Dynamic Longitudinal Data

Nevertheless, to reveal valuable insights for clinical applications, realistic and accurate analyses of the microbiota must be done. The community has raised some concerns with current studies of human microbiome research. Most studies only focus on describing the static taxonomic composition of the human microbiome, overlooking temporal variability, thus causing major drawbacks in real-world clinical applications, as many diseases are characterised by periods of remission and exacerbation in symptoms. Therefore, this work will mainly focus on investigating the dynamics of the human microbiome (i.e., analysing longitudinal data), which is, in fact, its real nature. A series of studies have remarked the importance of developing robust time-series analysis to uncover insights into microbial interactions and dynamics [16,17]. In [18], the authors studied the potential of time-varying communities in response to perturbations and obtained results that pointed out how longitudinal analysis can reveal insights into microbial ecosystem dynamics and aid to explain why perturbations (external or internal factors) modulate microbiome dynamics and stability. Furthermore, several studies have pointed out the need for integrating omic datasets (e.g.,: metatranscriptomics, metabolomic, etc.) to help unravel taxonomic and functional changes [19–21].

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

1. Data size: Current datasets lack large-scale data, suffering 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 [22].
2. Comparability and reproducibility: The lack of validated clinical models and differences in methodologies is preventing the translation of valuable results into real-world clinical practice.
3. Inherent characteristics of microbiome data: Sparsity, compositionality, and high variability are the main statistical properties that describe microbiome data hence leading to several 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. Thus, the bottleneck has shifted from data generation to data analysis. Moreover, microbiome data is compositional, so instead of looking at the absolute abundances of cells, we are mapping reads, and there is a fixed sequencing depth, i.e., four reads/sample, given by the technology used to obtain the sequences.
4. Interpretability: Incorporating phylogenetic and functional relationships among organisms into unified dynamic models of the human microbiome is crucial. Studies need to integrate multi-omic datasets to fully understand microbes and their interactions instead of exploring unique taxonomic composition analysis.

1.2. State of the Research Field

Machine learning (ML) methods are a well-suited solution for handling microbiome analysis, unlocking its full biological and clinical potential. Traditional biostatistical analytical methods are sometimes ineffective and limited compared to ML techniques, given the inherently noisy and highly variable nature of microbiome data. It has not been until recent years that more studies have started to explore the power of ML methods to predict host traits from microbiome patterns [23–27]. Bayesian networks (BNs) are a powerful ML tool to model the interaction of many microbial communities in the human gut by inferring complex networks from noisy data to predict clinical outcomes of relevance in a biologically interpretable manner. Microbiome data exhibits strong temporal fluctuations that we are interested in modelling. The use of dynamic Bayesian networks (DBNs) can help us handle this characteristic behavior of the system by providing information about the ordering and dependencies between the time points or showing how one taxon/pathway/metabolome influences another over time. Despite the increasing interest in microbiome research, to the best of our knowledge, only a few studies have applied BNs to human microbiome data (Table 1) [28–35].

Table 1. State of the art of BNs models applied to human microbiome datasets. Studies appear in chronological order.

#	Study	Method	Dataset	Longitudinal Data	Meta-Omics	Goal
1	[28]	DBNs	Premature infant gut [36]	☒	□	Build a DBN model to identify important relationships between microbiome taxa and predict future changes in microbiome composition
2	[29]	BNs	Vaginal microbiome [37]	□	□	Demonstrate associations between women's sexual and menstrual habits, demographics, vaginal microbiome composition, and symptoms and diagnostics of bacterial vaginosis (BV)
3	[30]	DBNs	Infant gut [36]	☒	□	Obtaining inferences from time-series data
4	[31]	BNs	Twins UK [38]	□	□	Possible causal relationships between metabolites and body mass index (BMI)
5	[32]	BNs	Rectal cancer [32]	□	☒	Reveal differential microbial communities and functions in terms of therapeutic responses
6	[33]	BNs with the incremental dynamic analysis (IDA) method	Colorectal cancer [14,39]	□	□	Identify key species that are likely to be causal agents of colorectal cancer (CRC)
7	[34]	BNs with co-occurrence networks (CoNs)	Infant gut [36], vaginal [37], oral data [5]	□	□	Make an inference about colonisation order
8	[35]	DBNs	IBDMD (inflammatory bowel disease multi-omics database) [40]	☒	☒	Infer temporal relationships between entities in a microbial community and extend (Lugo-Martinez et al., 2019) to other omics

The first report on the use of DBNs for human microbiome data analysis, according to authors, was [28]. Their work is the pioneer study to build a DBN model to capture the influence of individual microbial classes on each other over time. The most important pitfalls of this study were 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 studies in the literature were either limited to using traditional BNs [29,31,32] (data analysed was static) or to the analyses of a single omic data set [33,34], thus, lacking a holistic view of the microbial community. Nevertheless, their preliminary work reported interesting results such as the confirmation of the importance of vaginal pH and *Gardnerella* as influencers on the Nugent score (bacterial vaginosis diagnostic) or the identification of key species likely to cause colorectal cancer (CRC). Finally, [35] extended the previous research group's activity [30] to account for multi-omic dataset integration. Their work employs four types of omic data: taxa, genes, host genes, and metabolites. From the studied literature, none of the existing studies cover all of the objectives set for this study, so we believe that, although similar work has been presented in recent years, the specific focus of our work is novel and will provide relevant insights to the community.

1.3. Interpretability

We could arguably state that, although not a novel trend, explainable artificial intelligence (XAI) is of broad and current interest. In recent years, innovative ML algorithms, such as deep learning, have become increasingly complex and sophisticated [41,42]. Consequently, there is an unprecedent need, requested by non-experts in the domain, of developing transparent and understandable models. Especially in the clinical field, explaining the reasoning behind the decisions and results is crucial for applicability in medicine [43]. To the best of our knowledge, this line of research had not been applied to microbiome research until the last two years and only in a few publications: [44,45] and [46].

1.4. Main Aim and Contributions of the Work

The present approach answers this need by generating ML prediction models based on probabilistic graphical models (BNs) and prior biological domain knowledge (input restrictions), which will help scientists obtain interpretable intelligent systems to benefit human health. This work pursues the development of a computational methodology for human microbiome research that is explainable and transparent in dynamic scenarios, i.e., with longitudinal data. More specifically, the aim of this work is to serve as a general-purpose framework/protocol to study microbiome characteristics using ML that would be easy to use for either microbiology experts or computer scientists. A set of preprocessing steps were developed to successfully clean, filter, select, integrate, and model informative metagenomics, metatranscriptomics, and metabolomics longitudinal data from the Human Microbiome Project. A summary of the main contributions of this work is presented in Table 2.

Table 2. Main contributions of this study.

#	Contributions
1	Statistical analysis of longitudinal, multi-omic human microbiome data
2	State-of-the-art review of interpretable artificial intelligence approaches (models and tools) for human microbiome data
3	Identification of temporal interactions and connections between the biological entities: microbial taxa, microbial metabolic pathways, metabolites
4	Address both taxonomic composition and functional profile
5	Network model for each specific disease state (UC, CD)
6	Novel proposed preprocessing framework for the IBD Human Microbiome Project data to serve as an analysis tool for non-ML experts

To sum up, this study highlights the potential of network-based approaches (such as probabilistic graphical models) applied to microbiome research, given the complexity and sparsity of the data [47], as a XAI solution.

In this first introductory section, we have reviewed the current state of the research field, presented the aim and contributions of the work, and highlighted the purpose of the study. The rest of the paper is organised as follows. In Section 2, we will briefly describe the main methods applied and present the proposed framework. Section 3 summarises our main findings. Conclusions drawn from the project and future research is discussed in Section 4, and Section 5 concludes the paper.

2. Materials and Methods

In this section, we explain the proposed framework of the study and describe the mathematical and computational models used. It starts with a description of the dataset used and continues with the pre-processing steps followed. The section ends up presenting the predictive ML model used.

The dataset used in this study is publicly available from the Inflammatory Bowel Disease Multiomics database (IBDMDB) iHMP study [40], which follows 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). Raw sequence data can be downloaded from the BioProject NCBI site with accession code PRJNA398089. This dataset provides the most comprehensive description to date of host and microbial activities in inflammatory bowel diseases. Participants were classified attending to their disease status: non-IBD controls, Crohn's disease (CD), and ulcerative colitis (UC). This was one of the few datasets that met our complete selection criteria: (i) longitudinal data, (ii) multi-omic, (iii) human-disease-related, and a (iv) minimum of four measured timepoints for all the omic types: metagenomics, metatranscriptomics, and metabolomics.

The subjects of interest were filtered by number of measured time points (min. 4) for each omic type and yielded a total of 93 subjects: n = 47 for CD, n = 23 UC, and n = 23 controls or non-IBD. During further preprocessing, we removed two additional subjects due to missing time points for metatranscriptomics path abundance. Therefore, the data used for downstream analysis contained a total of 91 subjects. A brief description of the datasets used in the present study can be seen in Table 3.

Table 3. Structure of data files used in the present study from the Human Microbiome Project II–IBD study [40].

Data Type	File Name	File Description	File Dimension
Metadata	hmp2_metadata.csv	Full sample metadata table; samples as rows and metadata as columns	178 × 490
Metabolomics	iHMP_metabolomics.csv	Metabolomics profiles	81,867 × 553
Metagenomics	taxonomic_profiles.tsv	MetaPhlAn2 taxonomic profiles	1479 × 1639
Metatranscriptomics	pathabundance_relab.tsv	MTX pathway abundances with stratification	6061 × 736

Data inspection, statistical analyses (e.g.,: normalisation or differential abundance analysis), and visualisation tasks were performed using the 'R' programming language and environment [RStudio version 4.2.0, including phyloseq and Bioconductor package among others]. Two of the most important metrics to explore the biological diversity of microbiome data are α diversity and β diversity. These two metrics allow us to study, on the one hand, diversity within a sample and observe how dysbiosis states (UC and CD) manifest an expected lower diversity measure compared to a healthy state [48] and, on the other hand, how samples vary against each other, considering the whole distribution of species in a community, in order to discern between clusters (non-IBD and IBD), e.g., to understand if sample A from patient A (UC) is more similar in composition to sample B from patient B (non IBD) or C from patient C (CD). 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 1) to study diversity within a sample, we could observe how dysbiosis states (UC and CD) manifested an expected lower diversity measure compared to a healthy state [40,48].

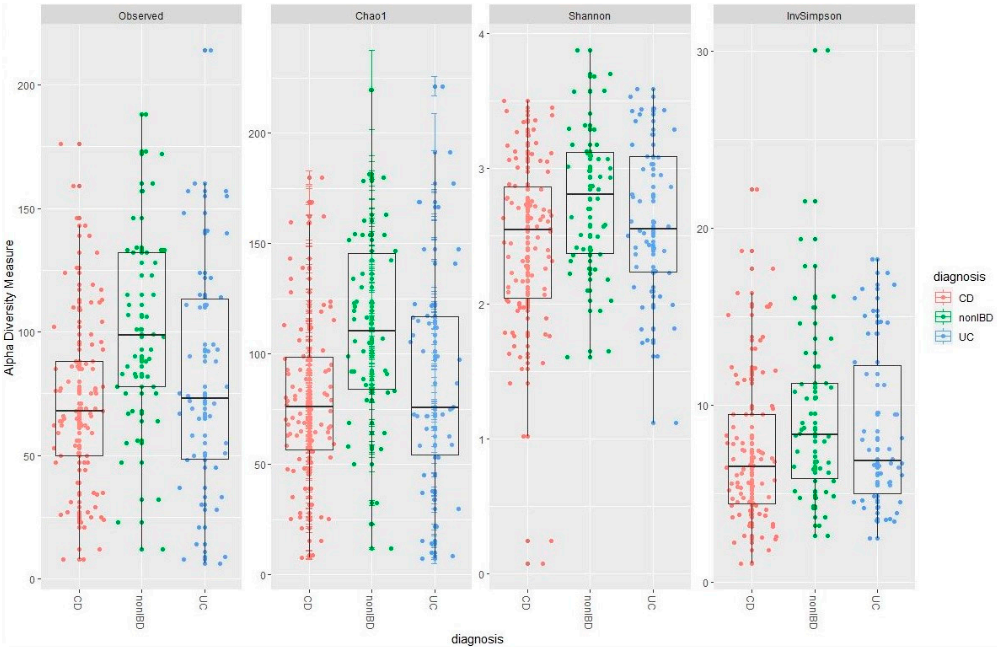

Figure 1. Comparison of α diversity measured by observed species, Chao1 index, Shannon diversity, and Simpson (from left to right). CD cluster is shown in red, UC in blue, and non-IBD subjects in green. Healthy control samples are significantly different from IBD samples. Shannon and Simpson indicate the uniformity of the abundance of different species in a sample. Figure generated using RStudio.

The results of comparing beta diversities from a qualitative and quantitative strategy are shown in Figure 2. Both Jaccard and Bray–Curtis indexes are reported, although similar results are obtained. 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 considering abundances.

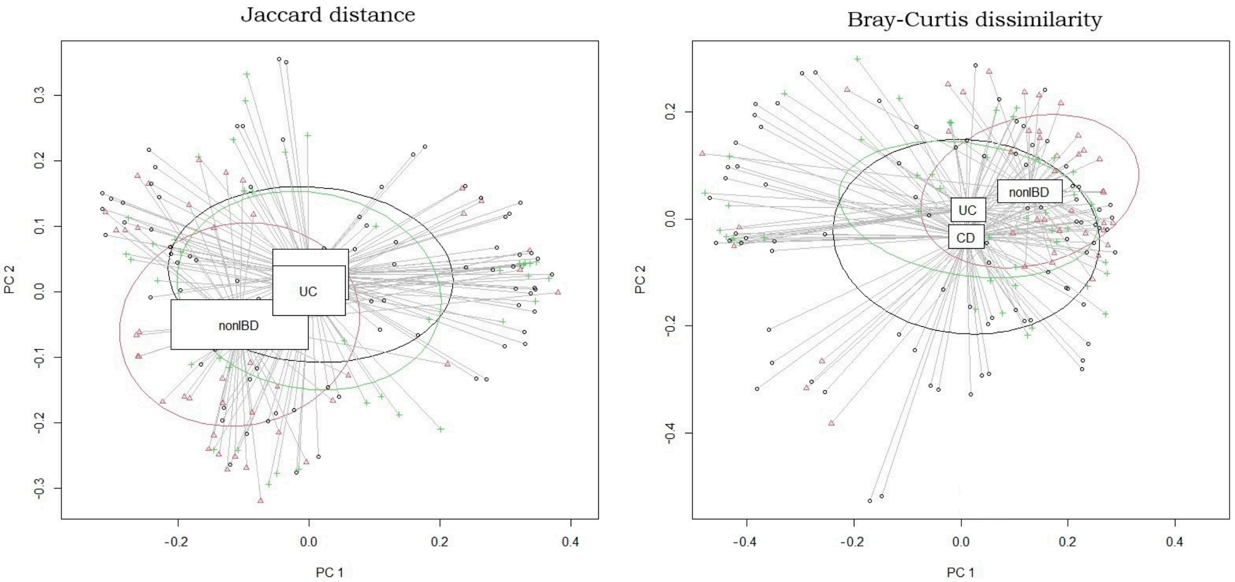

Figure 2. Beta diversity for the IBD dataset. Jaccard distance vs. Bray–Curtis dissimilarity. Once β diversity has been measured, the dataset can be visualised by principal coordinate analysis (PCoA). PCoA is an ordination technique widely described in the literature for analysing the composition of different microbiomes. β diversity analysis elucidate dissimilarities between samples (UC, CD and healthy). Figure generated using R.

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.

The preprocessing stage involved the interpretation of data format, the definition of data structures for their management, and correction tasks, including the elimination of noise and errors among others. An original preprocessing script was developed in Python to prepare the dataset that will be used for further analysis. The original datasets corresponding to the whole genome shotgun sequencing were loaded in CSV and TSV formats. The resulting ad hoc scripts are publicly available through the following link https://github.com/muia2021pl/Preprocessing_HumanMicrobiome (accessed on 29 April 2022). A total of five datasets are imported with the following feature sets, grouped by type:

- Clinical features (metadata): subject identification (e.g., "Subject ID"), time steps for sample time series (e.g., "week"), phenotype/cluster of each sample ("diagnosis"), external perturbations ("antibiotic")
- Metabolomic features: metabolic concentrations, mass-to-charge ratio (m/z) (continuous)
- Metagenomic features: taxonomic profile (continuous) corresponding to the relative abundance in percentage or counts per million
- Metatranscriptomic features: relative abundance of each metabolic pathway (continuous). The information is divided into two different datasets: HMP2 and HMP2 pilot. Data matrix (tables) will be preprocessed and expressed as abundances

Pre-processing tasks involved the removal of subjects with limited measured time points, rearranging indexes and columns, adding data-type identifiers to each variable and removing clinical variables (columns) with missing data. Additionally, log transformation and normalisation were applied to the data (for continuous variables). A normalisation method was used to remove technical bias in compositional data. Microbiome data are compositional for technical, biological, and computational reasons, thusly interpreted into relative counts. Moreover, each omic technology and data type has a variable number of columns and identified features. There are a number of standard normalisation methods used in the literature with this same final goal [49,50]. However, The use of log-ratios transformation is recommended for microbial taxa data normalisation [51,52]; therefore, the method of choice for our datasets, implemented in the ad hoc pre-processing Python script was centered log-ratios (CLR) transformation [53].

The CLR method is defined for a composition vector $\boldsymbol{x}_j$ as follows:

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

where x_{j} is the j-th sample, and $g(\boldsymbol{x}_j) = \left(\prod_{i=1}^{D} x_{i,j}\right)^{\frac{1}{D}}$ is the geometric mean (row-wise) of composition $\mathbf{x}$. The CLR transformation converts the relative abundance (or operational taxonomic unit counts) to ratios between all parts by calculating the geometric mean of all values (whole composition).

One major challenge with integrating multi-omics is that combining different types of biological information increases the number of analysed features while keeping the number of observations/samples (subjects) constant. Feature subset selection [54] can therefore improve the prediction accuracy of our model. As a preprocessing dimensionality reduction step, predictor variables with zero variance were filtered out using the VarianceThreshold() class from the sklearn.feature_selection module (scikit-learn 1.1.1.). We removed features with a training set variance equal to zero. Firstly, univariate feature selection filter algorithms were applied, which work by selecting features based on univariate statistical tests. SelectKBest() (scikit-learn 1.0.2 in Python) was used to select the k most important features with the highest scores based on Chi-squared statistics, ANOVA, and mutual information (MI). Chi2 ANOVA and MI scikit-learn Python implementations offer a good solution to deal with sparsity without making the data dense (without casting it internally to dense numpy arrays). Feature importance using random forest (RF) was also implemented. This embedded method combines filter and wrapper methods offering higher accuracy and a more generalisable approach. Additionally, 10-fold cross-validated linear support vector classifier (LinearSVC()) to create a linear support vector machine (SVM) model in scikit-learn was chosen as the external estimator implemented with Python scikit-learn package (scikit-learn 1.1.1). For performance assessment purposes, the four methods will be compared to perform feature subset selection on the full dataset for each omic type (Table 4). The set of informative features that were utilised for downstream ML analysis was obtained from the RF feature importance results.

Table 4. Classification accuracy of univariate feature subset selection techniques. Results shown for univariate feature selection correspond to k = 200 best features. Feature importance using RF is also reported.

Data Type	FSS Technique	Evaluation Metric (Accuracy)
Metagenomics	Univariate Chi2	0.82
	Univariate ANOVA	0.78
	Univariate MI	0.73
	Random forest	0.92
Metabolomics	Univariate Chi2	0.67
	Univariate ANOVA	0.69
	Univariate MI	0.69
	Random forest	0.78
Metatranscriptomics	Univariate Chi2	0.55
	Univariate ANOVA	0.56
	Univariate MI	0.50
	Random forest	0.68

Disease-State Prediction Model

Once the multi-omic data preprocessing step was completed, the next step was to learn a graphical structure for temporal data.

Among all the different ML approaches and models, BN-based analysis is certainly one of the most biologically interpretable, as its resulting networks can be easily understood [55–57]. The need for interpretable artificial intelligence models is highly demanded by microbiome researchers nowadays. Therefore, in this work, we will focus on the application of BN to the human microbiome research field.

A BN can be defined as a probabilistic graphical model used to encode the joint probability distribution over a set of random variables [58]. By means of a directed acyclic graph (DAG), probabilistic conditional (in)dependence relations (that can be causal under some circumstances) are represented by arcs and random variables (X_i) by nodes. This model offers an intuitive and solid approach to modelling uncertain knowledge. In order to construct a BN, the structure G 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 [59] and has aroused considerable interest in the scientific community, as have many other NP-hard problems. Methods that address the challenge of learning causal structure from data can be classified into three main groups: constraint-based, score-based, and hybrid methods. An inductive causation (IC) algorithm [60] provided the first framework for learning the skeleton for 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 the SGS algorithm [61] and by the most popular method, the PC algorithm that constitutes both the first practical implementation and an improvement on the former algorithms. The PC algorithm was composed of two principal steps: (i) finding the skeleton (detection phase) and (ii) making the 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) of X [60]. Therefore, the Markov blanket is the minimal set of nodes for which X is conditionally independent of all other nodes [62].

Other important local methods are: Grow Shrink, GS [63], and Incremental Association Markov blanket, IAMB [64] both of them follow a forward step-wise selection Markov blanket detection approach, so they learn in the first place the Markov blanket of each node, simplifying the identification of neighbours and hence, reducing the number of conditional independence tests that need to be computed.

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 modelling is maximum likelihood estimation. The goal of this statistical method is to maximise the probability of obtaining D for a specific value of θ, where $D = \{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(\mathrm{N})}\}$ 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 a 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 that we have not observed, a process known as probabilistic inference. In the simplest case, given an observation (evidence) **e**, we can query the model to calculate the posterior probability of a target variable(s) or node X_{j}: $\mathrm{p}(x_{\mathrm{j}} \mid \mathbf{e})$. Multiple methods have been developed over the years to perform approximate inference [65–67], instead of exact inference as this latter case implies an intractable problem for densely connected BNs. Nevertheless, [68] demonstrated that even approximate inference is NP-hard.

An important matter to consider 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 [69] which models all conditional distributions as linear Gaussians and all continuous nodes jointly follow a multivariate normal distribution $\mathrm{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\sum})$. 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 probability distributions, conditional on the values of their discrete parents and models continuous nodes as Gaussian distributions linearly dependent upon their Gaussian parents and with parameters conditional on the values of the discrete parents. If the CGBN has a DAG, G, over discrete variables Δ and continuous variables Ψ, where π(X) is the (possible 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 [70]:

$$\mathrm{P}(\Delta) f(\psi|\Delta) = \prod_{x\in\Delta} P\left(x \middle| \pi(x)\right) \prod_{y\in\psi} f\left(y \middle| \pi(y)\right) \tag{2}$$

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

Therefore, as seen above, BNs show great potential due to their ability to deal with uncertainties related to limited short (in time) and sparse data and their power to detect informative patterns of the underlying system. Moreover, BNs can be self-explanatory, i.e., the explanation of the model and reasoning process can be inferred graphically [56].

Dynamic Bayesian Network

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, DBNs [71] represent an important approach for time-series human microbiome data analysis. DBNs extend BNs to model time-series data (dynamic systems), 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 or more consecutive times slices (always directed forward in time), or (ii) intra-slice arcs: the arcs that connect nodes from the same time slice. In our DBN-based approach, certain assumptions are used: (i) a first-order Markov assumption, i.e., the probability of an observation at time *t* only depends on the observation at time *t*-1; (ii) stationarity, i.e., the data is generated by a distribution that does not change with time. The resulting DBN can be modelled for $\mathbf{X}^{t} = (X_1^t, \ldots, X_n^t)$ for n variables at each time slice $t = 1, \ldots, T$, as shown in Figure 3.

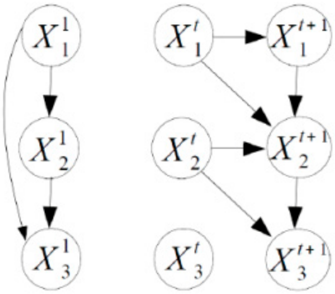

Figure 3. Example of a DBN structure. Prior BN (**left**) and transition BN (**right**) for three variables.

Assuming a first-order Markovian transition model, we have that

$$P\left(\mathbf{X}^1,\ldots,\mathbf{X}^T\right) = \mathrm{P}(\mathrm{X}^1)\prod_{t=2}^{\mathrm{T}} P\left(\mathbf{X}^t \,\middle|\, \mathbf{X}^{t-1}\right) = P\left(\mathbf{X}^1,\ldots,\mathbf{X}^T\right) \tag{3}$$

In order to select the right software package to implement the DBN, several existing options were evaluated and compared first. Table 5 presents the results from the software package benchmarking. The requirements we needed the selected software to meet and support consisted of learning and the inference of DBNs in the presence of both discrete and continuous data.

Table 5. Software packages benchmarking (* only if one discrete parent and no children).

Name	Language	Data Type	Learning	Inference
dbnR [72]	R	Continuous data	Yes	Yes
Bnlearn [73]	R	Discrete and continuous	Yes	No
Bnfinder [74]	Python	Discrete and continuous *	Yes	Yes
CGBayesNets [75]	Matlab	Discrete and continuous	Yes	Yes

dbnR package is a good alternative 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 visualising 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 then train a continuous network model for each cluster.

The 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 [76]), score-based (e.g., hill climbing), and hybrid (MMHC [77], hybrid HPC [78]) 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) [79] and minimal description length (MDL) [80,81]. 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) [75].

CGBayesNets [75] 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 a 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 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, this was the selected software tool to use in our work. A modified version of CGBayesNets implemented by [30] was used, wherein intra-edges are allowed, and BIC and AIC networks scoring functions are included.

In the first place, the network structure is learned from the dataset. A number of parameters need to be set for the learning algorithm. Prior assumed distributions for each node are needed to determine the posterior probability of the data. For this study:

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

As a filtering strategy, to prune the dataset and reduce the 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 to 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 the usual Bayes network methods. The Bayes factor reduces the dataset 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) a K2-style search [82]; (ii) greedy, exhaustive, a hill climbing search (every step adds arc that increases the likelihood the most); (iii) a pheno-centric search [83]; (iv) simulated annealing [84]. The theoretical foundation of CGBayesNet can be seen in Appendix A.

The main framework for learning DBNs consists of the following steps (1) combining time-series data into a larger column matrix with each time-point matrix below the prior time-point matrix, (2) learning the BN using StateTrackerSearch() function with the dynamic Bayes net option enabled to allow cycles and self-loops, (3) unrolling BN into a 2TBN: a 2-timepoint BN (all arcs are from time-point one to time-point two), (4) unrolling the dataset from the timeseries matrix, and (5) using normal techniques to make predictions 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, it is slotted into the T0 data. The second time it is encountered, it is slotted into both the T1 and the T0 data. The last time it is encountered, it is only slotted into the TT 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 that it does not consider all possible networks but rather all possible legal arcs between any two nodes. The Bayesian–Dirichlet-equivalent sample-size uniform (BDeu) measure of the marginal likelihood of the data [79] is used as the network scoring metric.

It is important to note that we adapted original implementation scripts (code) [75] to serve our particular purposes, as our data and final goal were different from similar studies that also used CGBayesNet [28,30,35].

Furthermore, we performed DBN structure, constraining it by using an adjacency matrix as an input to the model. This matrix is configured in such a way that only allows edges between specific nodes, therefore reducing complexity and avoiding overfitting. The selected configuration was based on basic biological knowledge, following [35]'s 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 ti will impact taxa abundance and growth in the next time-slice, $t_i + 1$.

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 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 [30] and [35], we maximised the likelihood of the data for a given structure using maximum log-likelihood estimation (MLE).

The model outputs both trivial Graph Format (.tgf) and GraphML (.graphml). For this implementation, an output file GraphML version of the network will be used, output_file.graphml, which can then be loaded into network visualisation software, such as Cytoscape [85]. Additionally, an ad hoc script was developed in R to generate a custom style XML file for the output networks, encoding several properties of the underlying graph, such as node shape, arc line type, and transparency of abundances to visualise them in Cytoscape.

Once the model has been fitted, inference over the learned model can be performed. When using BN, any variable can be used as the target node of the inference. Furthermore, in this particular case (DBNs) variables in the next time-slices are predicted from the values in the previous slices.

CGBayesNet implements the Cowell algorithm [86] 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 [87].

3. Results

3.1. DBN Model

In this study a DBN model of the gut microbial ecosystem was built from the Inflammatory Bowel Disease Multi-omics dataset of the Human Microbiome Project.

A two-stage DBN model was implemented, wherein two slices were modelled 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 a health status. In order to do this, we (i) prepared a framework that covered main microbiome analysis pre-processing steps, (ii) modelled interactions between different omics, (iii) constructed and learned a dynamic structure for each disease state (UC and DC) to infer which is the most probable dynamics, i.e., to identify a maximum a posteriori (MAP), and (iv) constructed a model adding a 'diagnosis' node to the network and studied its outgoing arcs.

3.2. Pre-Processing

The pre-processing steps, implemented through our ad-hoc script, involved filtering subjects with limited time points, integrating three omic types in one matrix, normalisation, and feature selection (see Section 2). Feature importance was computed for each omic data type (Figures 4–6). Based on these pre-processing steps, the resulting dataset used for modelling 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 or not (binary) of antibiotics was included. In addition, the Bayes factor score was used to further reduce the dataset. Prior biological knowledge was used as input to the DBN learning algorithm in order to constrain (arcs between nodes, arc directions, and strength) the resulting output model and prevent overfitting.

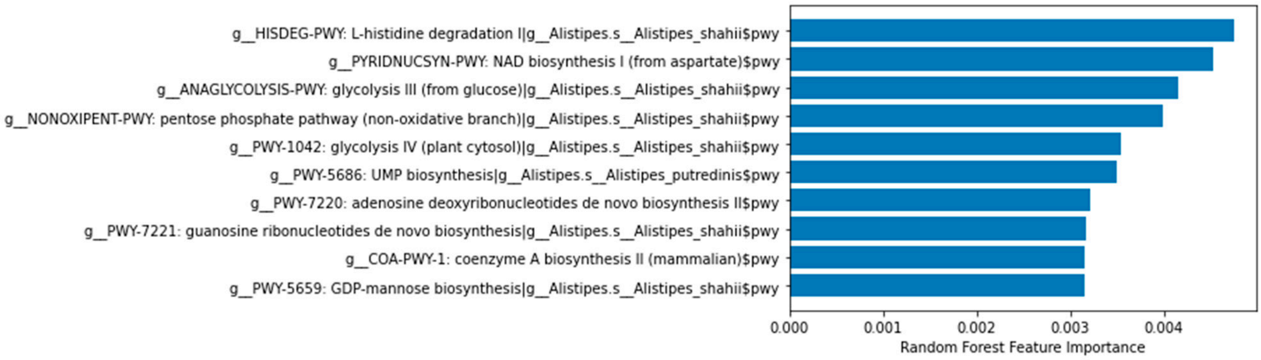

Figure 4. Feature importance (metatranscriptomics data) with a RF. Only the first ten features are shown for illustrative purposes. Generated with a Python script.

3.3. The Resulting Network

The full network comprised 91 nodes per time slice: 37 microbial taxa, 22 gene pathways, 29 metabolites, and three clinical variables. We constructed a network model (i) with and without bootstrapping (10 repetitions due to restricted computational tools), (ii) with a restriction matrix, and (iii) with different Bayes factor score thresholds (thresholds = 5, 10, 15) explored as part of a hyperparameter tuning phase. Connections with the largest Bayes factors are more likely to represent a true causal association.

For illustrative purposes, we trained a DBN model on a subset of the 50 best entities of each omic type and set a maximum number of parents of three. The results for the combined diagnosis model (the three health conditions in the same model) are shown in Figure 7. Nodes represent bacterial taxon, metabolites, clinical data, or metabolic pathways. Results for the independent models for each condition (CD, UC) are also reported in Figures 8 and 9, respectively. The full names of nodes can be seen in Supplementary Material Table S1.

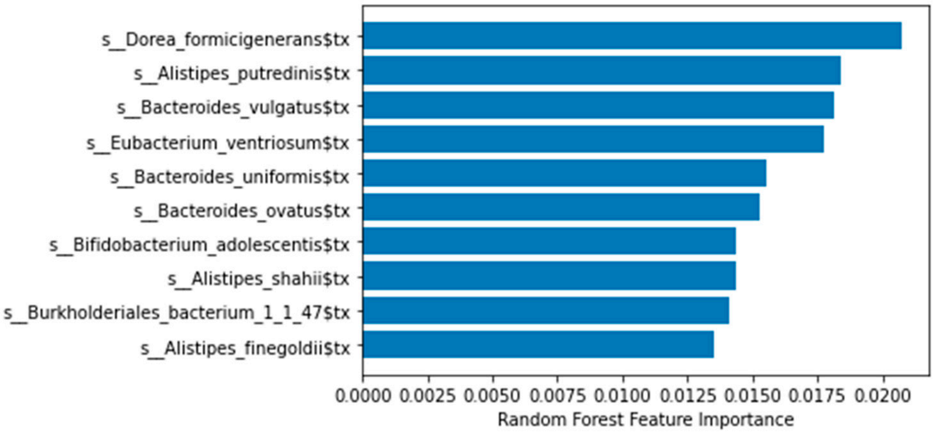

Figure 5. Feature importance (metagenomic data) with a RF. Only the first ten features are shown for illustrative purposes. Generated with a Python script.

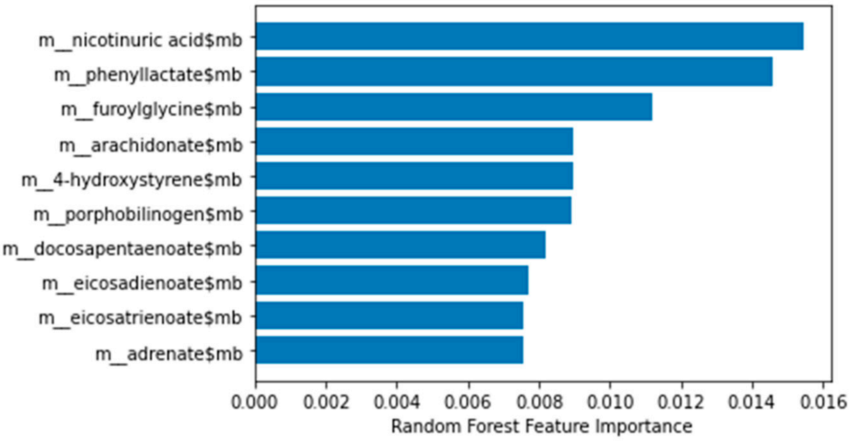

Figure 6. Feature importance (metabolomics data) with a RF. Only the first ten features are shown for illustrative purposes. Generated with a Python script.

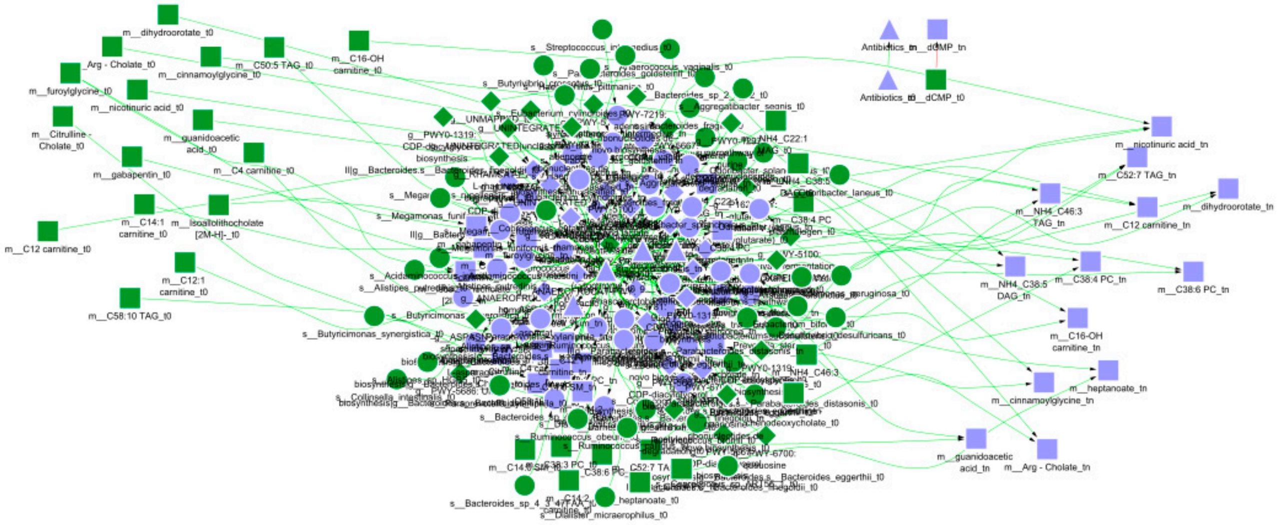

Figure 7. Learned DBN with 'restriction matrix' constraints on the top-50 best features per omic type. Green nodes represent time slice t_i and purple nodes the consecutive t_i+1. Metabolite nodes are represented by squares, species (taxa) by circles, clinical variables by triangles, and metabolic pathways by diamonds. The total number of nodes is 182, and the total number of edges is 231.

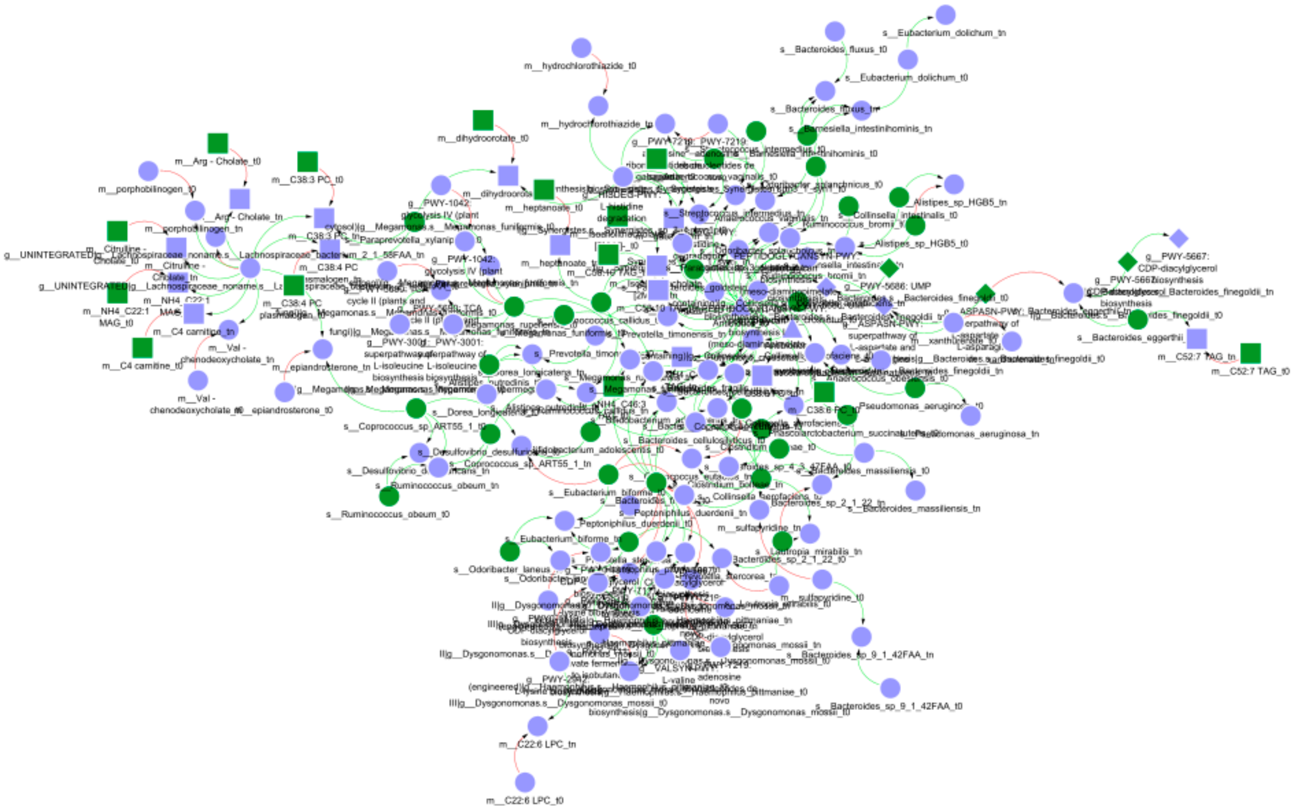
s__Eubacterium_dolichum_tn
s__Bacteroides_fluxus_t0
s__Eubacterium_dolichum_t0
m__hydrochlorothiazide_t0
s__Bacteroides_fluxus_tn
m__hydrochlorothiazide_tn
s__Barnesiella_intestinihominis_tn
m__dihydroorotate_t0
m__Arg - Cholate_t0
m__C38:3 PC_t0
m__porphobilinogen_t0
m__Arg - Cholate_tn
m__porphobilinogen_tn
g__UNINTEGRATED
m__Citrulline - Cholate_t0
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__Alistipes_sp_HGB5_tn
s__Alistipes_sp_HGB5_t0
s__Collinsella_intestinalis_t0
g__PWY-5667: CDP-diacylglycerol biosynthesis
s__Bacteroides_finegoldii_tn
s__Bacteroides_finegoldii_t0
s__Bacteroides_eggerthii
m__C52:7 TAG_tn
m__C52:7 TAG_t0
s__Pseudomonas_aeruginosa_t0
s__Pseudomonas_aeruginosa_tn
s__Bacteroides_massiliensis_t0
s__Bacteroides_massiliensis_tn
s__Eubacterium_biforme_t0
s__Peptoniphilus_duerdenii_t0
s__Peptoniphilus_duerdenii_tn
s__Eubacterium_biforme_tn
m__sulfapyridine_tn
m__sulfapyridine_t0
s__Lautropia_mirabilis_tn
s__Bacteroides_sp_2_1_22_t0
s__Odoribacter_laneus
s__Bacteroides_sp_9_1_42FAA_tn
s__Bacteroides_sp_9_1_42FAA_t0
Illg__Dysgonomonas.s__Dysgonomonas_mossii_t0
m__C22:6 LPC_tn
m__C22:6 LPC_t0

Figure 8. DBN for patients with UC with 'restriction matrix' constraints in the top-50 best features per omic type. Colours and shapes are as in Figure 7.

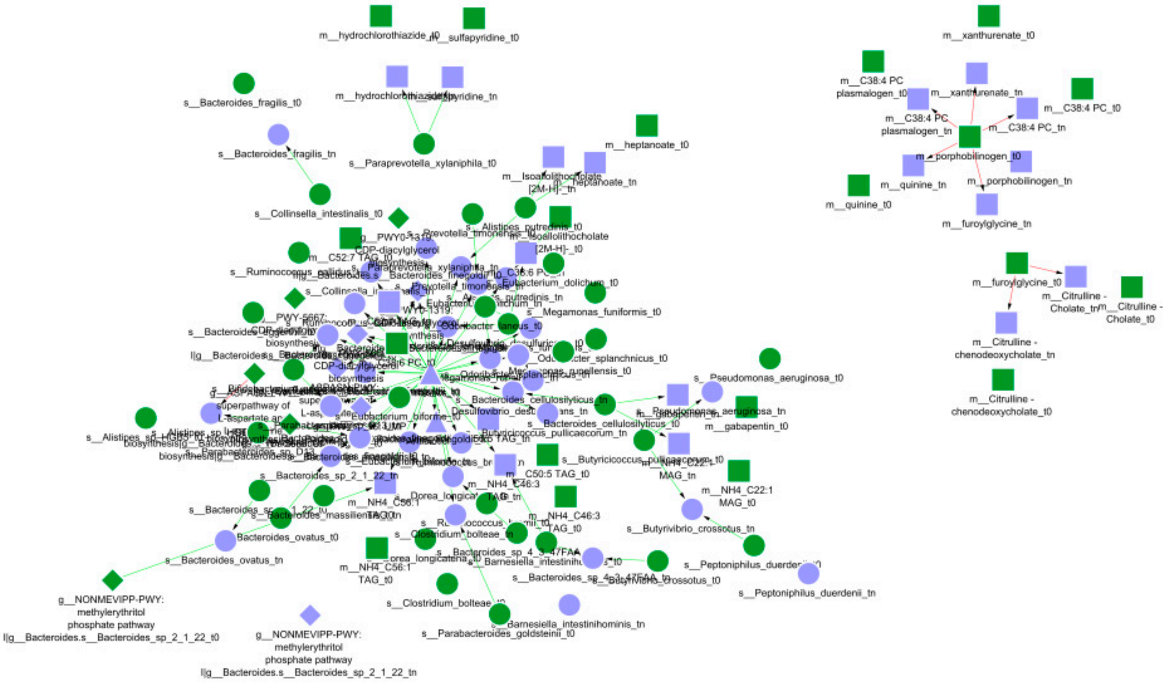
m__hydrochlorothiazide_t0
m__sulfapyridine_t0
s__Bacteroides_fragilis_t0
m__hydrochlorothiazide_tn
m__sulfapyridine_tn
s__Bacteroides_fragilis_tn
s__Paraprevotella_xylaniphila_t0
m__heptanoate_t0
m__Isoallolithocholate [2M-H]-_tn
m__heptanoate_tn
s__Collinsella_intestinalis_t0
s__Alistipes_putredinis_t0
m__C52:7 TAG_t0
s__Eubacterium_dolichum_t0
s__Megamonas_funiformis_t0
s__Odoribacter_splanchnicus_t0
s__Pseudomonas_aeruginosa_t0
s__Pseudomonas_aeruginosa_tn
m__gabapentin_t0
s__Bacteroides_cellulosilyticus_t0
s__Butyricicoccus_pullicaecorum_tn
s__Butyricicoccus_pullicaecorum_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_4_3_47FAA_t0
s__Barnesiella_intestinihominis_t0
s__Barnesiella_intestinihominis_tn
s__Parabacteroides_goldsteinii_t0
s__Clostridium_bolteae_t0
s__Clostridium_bolteae_tn
m__NH4_C56:1 TAG_t0
m__NH4_C46:3 TAG_t0
m__C50:5 TAG_t0
s__Dorea_longicatena_t0
s__Bacteroides_sp_2_1_22_tn
s__Bacteroides_massiliensis_tn
s__Bacteroides_ovatus_t0
s__Bacteroides_ovatus_tn
s__Alistipes_sp_HGB5
g__NONMEVIPP-PWY: methylerythritol phosphate pathway Ilg__Bacteroides.s__Bacteroides_sp_2_1_22_t0
g__NONMEVIPP-PWY: methylerythritol phosphate pathway Ilg__Bacteroides.s__Bacteroides_sp_2_1_22_tn
m__xanthurenate_t0
m__C38:4 PC plasmalogen_t0
m__xanthurenate_tn
m__C38:4 PC_t0
m__C38:4 PC plasmalogen_tn
m__C38:4 PC_tn
m__porphobilinogen_t0
m__quinine_tn
m__porphobilinogen_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 9. DBN for CD patients with 'restriction matrix' constraints on the top-50 best features per omic type. Colours and shapes are as in Figure 7.

3.4. Analysis and Interpretation of Experimental Results

In preliminary results, we did observe that, as expected from results of similar 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 a decrease in Clostridium and Faecalibacterium, is observed in subjects with this condition.

The final network model (Figure 7) may be suggestive of a set of taxa, gene metabolic pathways, and metabolites whose expression is dysregulated in patients with IBD. Furthermore, 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 (Figure 10), wherein the following chain was identified:

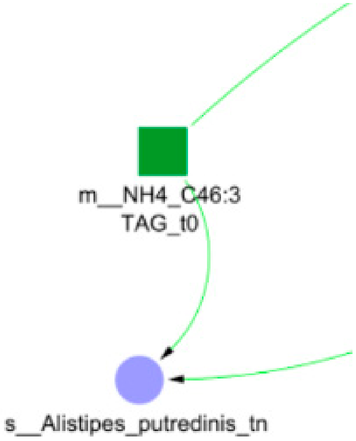

Figure 10. Identified connections between metabolite (green square) and taxa (purple circle) through the inspection of the dynamic Bayesian network model for UC.

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

Another stimulating finding (which supports the idea that the proposed model points in the right direction) is the connection between the methylerythritol phosphate (MEP) pathway (in t_i) and the NH4_C56:1 TAG (metabolite) in t_i+1, The end-product of this pathway, isopentenyl diphosphate (IPP), is known to play a crucial role in inflammation and disease [88]. The resulting analysis demonstrates the utility of DBNs ability to generate and test predictive models in human multi-omic microbiome datasets.

We performed comparisons of three different network search algorithms: (i) a K2-style, (ii) a pheno-centric search, and (iii) simulated annealing. A total of 20 bootstrap realisations of the dataset were performed and networks at various edge frequencies were computed. The consensus Markov-blanket had 71 nodes. The resulting best BN (with a total of 3 discrete nodes and 150 continuous nodes) reported a predictive performance of 64.8%, which is considered satisfactory and admissible in the present context [89].

4. Discussion

Our open implementation consisted of the following steps:

1. The pre-processing of the data set.
2. The fitting of the DBN model in two steps: structure and parameter learning. The output of this step was a 2-stage dynamic Bayes net class object (DBN).
3. The inference and test of 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 that the value assigned to that variable by the input data.

4. Dynamic Bayesian network visualisation and analysis for the biological interpolation of results.

The study examines, in an unprecedented exhaustively manner, the current state of the art of DBNs to solve a current scientific problem of interest: analysing the human microbiome temporal changes associated with disease states. A tailored analytical framework for data pre-processing was developed for the Inflammatory Bowel Disease Multiomics Dataset from the iHMP project, which covers an unmet need, as, to the best of our knowledge, the data access, preparation, and integration of these datasets for machine learning models have not been developed yet. Furthermore, a powerful artificial intelligence approach, DBNs, was applied to solve the problem with an innovative configuration and approach by integrating longitudinal multi-omic data for the characterisation of a model for each disorder (UC, CD) and the healthy state (non-IBD) and also an overall model. An additional original contribution presented in this manuscript is the practical analysis of the different software packages currently available to construct the solution. Furthermore, the usefulness of BNs for microbiome analysis has been presented. The use of prior biological domain knowledge as an input restrictions matrix allowed us to prove the value of this approach versus other popular ML models (e.g., RF, DL) in building explainable and interpretable models. Nevertheless, BNs 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 the overall sample size. Third, BNs cannot explain a cyclic or feedback relationship among variables.

The results obtained in terms of performance (accuracy = 0.65) and biological associations are in line with previous studies found in the literature. However, the focus of this work was not to build a model that improves state-of-the-art predictive performance, as numerous ML models, such as fandom forest, already exist, and they are more appropriate for this objective. The goal was to fill an existing gap by examining the reliability and power of interpretable ML models for use by non-experts in the domain (AI) in future clinical investigations as an alternative tool for novel knowledge discovery.

5. Conclusions

In this work, we have successfully achieved a dynamic Bayesian network model that has implicitly collected temporal relationships that can help clinicians and researchers in the domain (gut microbiota) explore and discover new biomarkers. A model was obtained for each state of disease (UC, CD, healthy). Although self-explanatory for clinicians once visually computed, they could not have been obtained easily by researchers with no prior knowledge of both the analysis of omic data (bioinformatics) and artificial intelligence (computer science). As far as the authors are aware, no comprehensive work was dedicated to this same research objective. Moreover, this is the first study to use DBNs to capture temporal variability in microbiome data, identifying the 50 most important taxa, metabolic pathways, and metabolites for each condition.

The proposed methodology, BNs, and, more specifically, DBNs provide valuable insights and explanation about the predictions and probabilistic dependencies among the variables. The gut microbiome has been extensively studied, but, due to its high complexity and inter-individual heterogeneity, it is not yet fully understood. Although ML methods, and in particular, DBNs, are a promising technique to infer useful insights, there is still considerable work to be done in some areas.

Limitations and Future Research

In terms of ground truth assessment, the proposed ML model, DBN (CGBayesNet) was not compared with other DBN algorithms on a gold-standard dataset, as there are currently no computational tools or benchmark datasets available in the literature for this evaluation. Moreover, for the selected dataset (IBDMDB), no other available tool reviewed met the complete criteria to replicate the same analysis performed. Even so, future work should aim to test the model in multiple cohorts and potentially different class balances. For our study, we explored publicly available datasets for reproducibility and replicability purposes, but, for future studies, an interesting cohort composed of identical twins, thus not subject to genetic confounding, could be TwinsUK (https://twinsuk.ac.uk/, accessed on 17 February 2022). Moreover, other microbes such as the skin microbiome, oral cavity microbiome, or the respiratory system microbiome can produce equally interesting results and have not been widely explored and characterised yet.

Regarding evaluation and performance assessment, further work needs to be done by reporting multiple evaluation metrics and performing experimental validation on an independent cohort to translate research into clinical practice.

Further analysis and interpretation of results by domain experts (e.g., clinicians or microbiologists) is required to extract new knowledge applicable in clinical settings. Nonetheless, we hope this study encourages future collaborations between scientists from different fields to open interdisciplinary research lines contributing to this promising area of study.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/math10121994/s1, Table S1: Nodes of the complete DBN.

Author Contributions: Conceptualisation, P.L. (Pedro Larrañaga) and C.B.; methodology, P.L. (Paula Laccourreye), C.B., P.L. (Pedro Larrañaga); software, P.L. (Paula Laccourreye); validation, P.L. (Paula Laccourreye); formal analysis, P.L. (Paula Laccourreye); investigation, P.L. (Paula Laccourreye); data curation, P.L. (Paula Laccourreye); writing—original draft preparation, P.L. (Paula Laccourreye); writing—review and editing, P.L. (Pedro Larrañaga), C.B.; visualisation, P.L. (Paula Laccourreye); supervision, P.L. (Pedro Larrañaga), C.B. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been partially supported by the Spanish Ministry of Science and Innovation through the PID2019-109247GB-I00 project, and by the BBVA Foundation (2019 Call) through the "Score-based nonstationary temporal Bayesian networks. Applications in climate and neuroscience" (BAYES-CLIMA-NEURO) project.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study is openly available in the IBDMDB Downloads page: https://ibdmdb.org/tunnel/public/summary.html. IBDMDB data is associated with a NCBI BioProject. This data can be found here: https://www.ncbi.nlm.nih.gov/bioproject/398089 (PRJNA398089). For additional information on clinical, sample handling and data-generation protocols, the following link can be consulted at https://ibdmdb.org/cb/browser/, accessed on 29 April 2022. Both trivial Graph Format (.tgf) and GraphML (.graphml) outputs are available from the corresponding author, P.L. (Paula Laccourreye), upon request.

Acknowledgments: The author would like to thank Ivan Macía Oliver and Alba Garin-Muga (Digital Health and Biomedical Technologies, Vicomtech) for the support given towards the preparation and completion of this manuscript. We thank the reviewers for their valuable contributions and suggestions for improvement.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

The theoretical foundation of CGBayesNet is presented in [75]. To determine the best network model of the data, CGBayesNet computes the marginal likelihood of candidate network structures, conditioned upon the data, and chooses the network model that maximises 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) realising 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),θ_{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 modelled with Dirichlet priors, and 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:

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

where n_{ijk} is the number of data points satisfying y_i = k for π(y_i) in configuration j, and α_{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 that depends on its discrete parents, with a conditional variance $\sigma^2{}_{ij} = 1/\tau_{ij}$. The joint likelihood of the continuous nodes is then

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

with x_{ij} the values of continuous parents of y_i in case k, and β_{ij} is the vector of regression parameters given discrete parents of y_i = j. CGBayesNet follows [90] and uses a Gamma prior distribution for τ and a conditional multivariate Gaussian prior density on regression parameters β. Thus,

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

Furthermore, β is described by

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

For the identity matrix I, and β_{ij0} = E(β_{ij} | τ_{ij}). The above equations represent the main semantics of CGBayesNet.

References

1. Moran, M.A. The Global Ocean Microbiome. *Science* **2015**, *350*, aac8455. [CrossRef] [PubMed]
2. Mueller, U.G.; Sachs, J.L. Engineering Microbiomes to Improve Plant and Animal Health. *Trends Microbiol.* **2015**, *23*, 606–617. [CrossRef] [PubMed]

3. Louca, S.; Parfrey, L.W.; Doebeli, M. Decoupling Function and Taxonomy in the Global Ocean Microbiome. *Science* **2016**, *353*, 1272–1277. [CrossRef] [PubMed]
4. Hou, Q.; Kolodkin-Gal, I. Harvesting the Complex Pathways of Antibiotic Production and Resistance of Soil Bacilli for Optimizing Plant Microbiome. *FEMS Microbiol. Ecol.* **2020**, *96*, fiaa142. [CrossRef]
5. Turnbaugh, P.J.; Ley, R.E.; Hamady, M.; Fraser-Liggett, C.M.; Knight, R.; Gordon, J.I. The Human Microbiome Project. *Nature* **2007**, *449*, 804–810. [CrossRef]
6. Ehrlich, S.D. MetaHIT: The European Union Project on Metagenomics of the Human Intestinal Tract. In *Metagenomics of the Human Body*; Nelson, K.E., Ed.; Springer: New York, NY, USA, 2011; pp. 307–316. ISBN 978-1-4419-7089-3.
7. Vatanen, T.; Kostic, A.D.; d'Hennezel, E.; Siljander, H.; Franzosa, E.A.; Yassour, M.; Kolde, R.; Vlamakis, H.; Arthur, T.D.; Hämäläinen, A.-M.; et al. Variation in Microbiome LPS Immunogenicity Contributes to Autoimmunity in Humans. *Cell* **2016**, *165*, 842–853. [CrossRef]
8. Cornejo-Pareja, I.; Ruiz-Limón, P.; Gómez-Pérez, A.M.; Molina-Vega, M.; Moreno-Indias, I.; Tinahones, F.J. Differential Microbial Pattern Description in Subjects with Autoimmune-Based Thyroid Diseases: A Pilot Study. *J. Pers. Med.* **2020**, *10*, 192. [CrossRef]
9. Depner, M.; Taft, D.H.; Kirjavainen, P.V.; Kalanetra, K.M.; Karvonen, A.M.; Peschel, S.; Schmausser-Hechfellner, E.; Roduit, C.; Frei, R.; Lauener, R.; et al. Maturation of the Gut Microbiome during the First Year of Life Contributes to the Protective Farm Effect on Childhood Asthma. *Nat. Med.* **2020**, *26*, 1766–1775. [CrossRef]
10. Joseph, C.L.M.; Zoratti, E.M.; Ownby, D.R.; Havstad, S.; Nicholas, C.; Nageotte, C.; Misiak, R.; Enberg, R.; Ezell, J.; Johnson, C.C. Exploring Racial Differences in IgE-Mediated Food Allergy in the WHEALS Birth Cohort. *Ann. Allergy Asthma Immunol.* **2016**, *116*, 219–224.e1. [CrossRef]
11. Metwally, A.A.; Yu, P.S.; Reiman, D.; Dai, Y.; Finn, P.W.; Perkins, D.L. Utilizing Longitudinal Microbiome Taxonomic Profiles to Predict Food Allergy via Long Short-Term Memory Networks. *PLoS Comput. Biol.* **2019**, *15*, e1006693. [CrossRef]
12. Leiva-Gea, I.; Sánchez-Alcoholado, L.; Martín-Tejedor, B.; Castellano-Castillo, D.; Moreno-Indias, I.; Urda-Cardona, A.; Tinahones, F.J.; Fernández-García, J.C.; Queipo-Ortuño, M.I. 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* **2018**, *41*, 2385–2395. [CrossRef] [PubMed]
13. Qin, J.; Li, Y.; Cai, Z.; Li, S.; Zhu, J.; Zhang, F.; Liang, S.; Zhang, W.; Guan, Y.; Shen, D.; et al. A Metagenome-Wide Association Study of Gut Microbiota in Type 2 Diabetes. *Nature* **2012**, *490*, 55–60. [CrossRef] [PubMed]
14. Zeller, G.; Tap, J.; Voigt, A.Y.; Sunagawa, S.; Kultima, J.R.; Costea, P.I.; Amiot, A.; Böhm, J.; Brunetti, F.; Habermann, N.; et al. Potential of Fecal Microbiota for Early-Stage Detection of Colorectal Cancer. *Mol. Syst. Biol.* **2014**, *10*, 766. [CrossRef]
15. Wirbel, J.; Pyl, P.T.; Kartal, E.; Zych, K.; Kashani, A.; Milanese, A.; Fleck, J.S.; Voigt, A.Y.; Palleja, A.; Ponnudurai, R.; et al. Meta-Analysis of Fecal Metagenomes Reveals Global Microbial Signatures That Are Specific for Colorectal Cancer. *Nat. Med.* **2019**, *25*, 679–689. [CrossRef] [PubMed]
16. Ridenhour, B.J.; Brooker, S.L.; Williams, J.E.; Van Leuven, J.T.; Miller, A.W.; Dearing, M.D.; Remien, C.H. Modeling Time-Series Data from Microbial Communities. *ISME J.* **2017**, *11*, 2526–2537. [CrossRef]
17. Bucci, V.; Tzen, B.; Li, N.; Simmons, M.; Tanoue, T.; Bogart, E.; Deng, L.; Yeliseyev, V.; Delaney, M.L.; Liu, Q.; et al. MDSINE: Microbial Dynamical Systems Inference Engine for Microbiome Time-Series Analyses. *Genome Biol.* **2016**, *17*, 121. [CrossRef]
18. Faust, K.; Lahti, L.; Gonze, D.; de Vos, W.M.; Raes, J. Metagenomics Meets Time Series Analysis: Unraveling Microbial Community Dynamics. *Curr. Opin. Microbiol.* **2015**, *25*, 56–66. [CrossRef]
19. Heshiki, Y.; Vazquez-Uribe, R.; Li, J.; Ni, Y.; Quainoo, S.; Imamovic, L.; Li, J.; Sørensen, M.; Chow, B.K.C.; Weiss, G.J.; et al. Predictable Modulation of Cancer Treatment Outcomes by the Gut Microbiota. *Microbiome* **2020**, *8*, 28. [CrossRef]
20. Cammarota, G.; Ianiro, G.; Ahern, A.; Carbone, C.; Temko, A.; Claesson, M.J.; Gasbarrini, A.; Tortora, G. Gut Microbiome, Big Data and Machine Learning to Promote Precision Medicine for Cancer. *Nat. Rev. Gastroenterol. Hepatol.* **2020**, *17*, 635–648. [CrossRef]
21. Bodein, A.; Chapleur, O.; Droit, A.; Lê Cao, K.-A. A Generic Multivariate Framework for the Integration of Microbiome Longitudinal Studies with Other Data Types. *Front. Genet.* **2019**, *10*, 963. [CrossRef]
22. Su, X.; Jing, G.; Zhang, Y.; Wu, S. Method Development for Cross-Study Microbiome Data Mining: Challenges and Opportunities. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 2075–2080. [CrossRef] [PubMed]
23. Knights, D.; Costello, E.K.; Knight, R. Supervised Classification of Human Microbiota. *FEMS Microbiol. Rev.* **2011**, *35*, 343–359. [CrossRef] [PubMed]
24. Larsen, P.E.; Dai, Y. Metabolome of Human Gut Microbiome Is Predictive of Host Dysbiosis. *Gigascience* **2015**, *4*, 42. [CrossRef] [PubMed]
25. Moitinho-Silva, L.; Steinert, G.; Nielsen, S.; Hardoim, C.C.P.; Wu, Y.-C.; McCormack, G.P.; López-Legentil, S.; Marchant, R.; Webster, N.; Thomas, T.; et al. Predicting the HMA-LMA Status in Marine Sponges by Machine Learning. *Front. Microbiol.* **2017**, *8*, 752. [CrossRef]
26. Fukui, H.; Nishida, A.; Matsuda, S.; Kira, F.; Watanabe, S.; Kuriyama, M.; Kawakami, K.; Aikawa, Y.; Oda, N.; Arai, K.; et al. Usefulness of Machine Learning-Based Gut Microbiome Analysis for Identifying Patients with Irritable Bowels Syndrome. *J. Clin. Med.* **2020**, *9*, 2403. [CrossRef]
27. Hacilar, H.; Nalbantoglu, O.U.; Aran, O.; Bakir-Gungor, B. Inflammatory Bowel Disease Biomarkers of Human Gut Microbiota Selected via Ensemble Feature Selection Methods. *arXiv* **2020**, arXiv:2001.03019.

28. McGeachie, M.J.; Sordillo, J.E.; Gibson, T.; Weinstock, G.M.; Liu, Y.-Y.; Gold, D.R.; Weiss, S.T.; Litonjua, A. Longitudinal Prediction of the Infant Gut Microbiome with Dynamic Bayesian Networks. *Sci. Rep.* **2016**, *6*, 20359. [CrossRef]
29. Noyes, N.; Cho, K.-C.; Ravel, J.; Forney, L.J.; Abdo, Z. Associations between Sexual Habits, Menstrual Hygiene Practices, Demographics and the Vaginal Microbiome as Revealed by Bayesian Network Analysis. *PLoS ONE* **2018**, *13*, e0191625. [CrossRef]
30. Lugo-Martinez, J.; Ruiz-Perez, D.; Narasimhan, G.; Bar-Joseph, Z. Dynamic Interaction Network Inference from Longitudinal Microbiome Data. *Microbiome* **2019**, *7*, 54. [CrossRef]
31. Howey, R.; Shin, S.-Y.; Relton, C.; Smith, G.D.; Cordell, H.J. Bayesian Network Analysis Incorporating Genetic Anchors Complements Conventional Mendelian Randomization Approaches for Exploratory Analysis of Causal Relationships in Complex Data. *PLoS Genet.* **2020**, *16*, e1008198. [CrossRef]
32. Jang, B.-S.; Chang, J.H.; Chie, E.K.; Kim, K.; Park, J.W.; Kim, M.J.; Song, E.-J.; Nam, Y.-D.; Kang, S.W.; Jeong, S.-Y.; et al. Gut Microbiome Composition Is Associated with a Pathologic Response After Preoperative Chemoradiation in Patients with Rectal Cancer. *Int. J. Radiat. Oncol. Biol. Phys.* **2020**, *107*, 736–746. [CrossRef] [PubMed]
33. Kharrat, N.; Assidi, M.; Abu-Elmagd, M.; Pushparaj, P.N.; Alkhaldy, A.; Arfaoui, L.; Naseer, M.I.; El Omri, A.; Messaoudi, S.; Buhmeida, A.; et al. Data Mining Analysis of Human Gut Microbiota Links *Fusobacterium* spp. with Colorectal Cancer Onset. *Bioinformation* **2019**, *15*, 372–379. [CrossRef] [PubMed]
34. Sazal, M.; Mathee, K.; Ruiz-Perez, D.; Cickovski, T.; Narasimhan, G. Inferring Directional Relationships in Microbial Communities Using Signed Bayesian Networks. *BMC Genom.* **2020**, *21*, 663. [CrossRef] [PubMed]
35. Ruiz-Perez, D.; Lugo-Martinez, J.; Bourguignon, N.; Mathee, K.; Lerner, B.; Bar-Joseph, Z.; Narasimhan, G. Dynamic Bayesian Networks for Integrating Multi-Omics Time Series Microbiome Data. *Msystems* **2021**, *6*, e01105-20. [CrossRef]
36. La Rosa, P.S.; Warner, B.B.; Zhou, Y.; Weinstock, G.M.; Sodergren, E.; Hall-Moore, C.M.; Stevens, H.J.; Bennett, W.E.; Shaikh, N.; Linneman, L.A.; et al. Patterned Progression of Bacterial Populations in the Premature Infant Gut. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 12522–12527. [CrossRef]
37. Ravel, J.; Gajer, P.; Abdo, Z.; Schneider, G.M.; Koenig, S.S.K.; McCulle, S.L.; Karlebach, S.; Gorle, R.; Russell, J.; Tacket, C.O.; et al. Vaginal Microbiome of Reproductive-Age Women. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 4680–4687. [CrossRef]
38. Moayyeri, A.; Hammond, C.J.; Hart, D.J.; Spector, T.D. The UK Adult Twin Registry (TwinsUK Resource). *Twin Res. Hum. Genet.* **2013**, *16*, 144–149. [CrossRef]
39. Marchesi, J.R.; Dutilh, B.E.; Hall, N.; Peters, W.H.M.; Roelofs, R.; Boleij, A.; Tjalsma, H. Towards the Human Colorectal Cancer Microbiome. *PLoS ONE* **2011**, *6*, e20447. [CrossRef]
40. Lloyd-Price, J.; Arze, C.; Ananthakrishnan, A.N.; Schirmer, M.; Avila-Pacheco, J.; Poon, T.W.; Andrews, E.; Ajami, N.J.; Bonham, K.S.; Brislawn, C.J.; et al. Multi-Omics of the Gut Microbial Ecosystem in Inflammatory Bowel Diseases. *Nature* **2019**, *569*, 655–662. [CrossRef]
41. Castelvecchi, D. Can We Open the Black Box of AI? *Nat. News* **2016**, *538*, 20. [CrossRef]
42. Holzinger, A.; Biemann, C.; Pattichis, C.S.; Kell, D.B. What Do We Need to Build Explainable AI Systems for the Medical Domain? *arXiv* **2017**, arXiv:1712.09923.
43. Barredo Arrieta, A.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]
44. Prifti, E.; Chevaleyre, Y.; Hanczar, B.; Belda, E.; Danchin, A.; Clément, K.; Zucker, J.-D. Interpretable and Accurate Prediction Models for Metagenomics Data. *GigaScience* **2020**, *9*, giaa010. [CrossRef] [PubMed]
45. Carrieri, A.P.; Haiminen, N.; Maudsley-Barton, S.; Gardiner, L.-J.; Murphy, B.; Mayes, A.E.; Paterson, S.; Grimshaw, S.; Winn, M.; Shand, C.; et al. Explainable AI Reveals Changes in Skin Microbiome Composition Linked to Phenotypic Differences. *Sci. Rep.* **2021**, *11*, 4565. [CrossRef]
46. Wong, C.W.; Yost, S.E.; Lee, J.S.; Gillece, J.D.; Folkerts, M.; Reining, L.; Highlander, S.K.; Eftekhari, Z.; Mortimer, J.; Yuan, Y. Analysis of Gut Microbiome Using Explainable Machine Learning Predicts Risk of Diarrhea Associated with Tyrosine Kinase Inhibitor Neratinib: A Pilot Study. *Front. Oncol.* **2021**, *11*, 283. [CrossRef]
47. Pan, A.Y. Statistical Analysis of Microbiome Data: The Challenge of Sparsity. *Curr. Opin. Endocr. Metab. Res.* **2021**, *19*, 35–40. [CrossRef]
48. Wright, E.K.; Kamm, M.A.; Teo, S.M.; Inouye, M.; Wagner, J.; Kirkwood, C.D. Recent Advances in Characterizing the Gastrointestinal Microbiome in Crohn's Disease: A Systematic Review. *Inflamm. Bowel Dis.* **2015**, *21*, 1219–1228. [CrossRef]
49. Paulson, J.N.; Stine, O.C.; Bravo, H.C.; Pop, M. Robust Methods for Differential Abundance Analysis in Marker Gene Surveys. *Nat. Methods* **2013**, *10*, 1200–1202. [CrossRef]
50. Badri, M.; Kurtz, Z.D.; Müller, C.L.; Bonneau, R. Normalization Methods for Microbial Abundance Data Strongly Affect Correlation Estimates. *BioRxiv* **2018**, 406264. [CrossRef]
51. Gloor, G.B.; Macklaim, J.M.; Pawlowsky-Glahn, V.; Egozcue, J.J. Microbiome Datasets Are Compositional: And This Is Not Optional. *Front. Microbiol.* **2017**, *8*, 2224. [CrossRef]
52. Mars, R.A.T.; Yang, Y.; Ward, T.; Houtti, M.; Priya, S.; Lekatz, H.R.; Tang, X.; Sun, Z.; Kalari, K.R.; Korem, T.; et al. Longitudinal Multi-Omics Reveals Subset-Specific Mechanisms Underlying Irritable Bowel Syndrome. *Cell* **2020**, *182*, 1460–1473.e17. [CrossRef] [PubMed]
53. Aitchison, J. The Statistical Analysis of Compositional Data. *J. R. Stat. Soc. Ser. B* **1982**, *44*, 139–177. [CrossRef]

54. Saeys, Y.; Inza, I.; Larrañaga, P. A Review of Feature Selection Techniques in Bioinformatics. *Bioinformatics* **2007**, *23*, 2507–2517. [CrossRef] [PubMed]
55. Wang, J.-W.; Kuo, C.-H.; Kuo, F.-C.; Wang, Y.-K.; Hsu, W.-H.; Yu, F.-J.; Hu, H.-M.; Hsu, P.-I.; Wang, J.-Y.; Wu, D.-C. Fecal Microbiota Transplantation: Review and Update. *J. Formos Med. Assoc.* **2019**, *118* (Suppl. S1), S23–S31. [CrossRef]
56. Mihaljević, B.; Bielza, C.; Larrañaga, P. Bayesian Networks for Interpretable Machine Learning and Optimization. *Neurocomputing* **2021**, *456*, 648–665. [CrossRef]
57. Needham, C.J.; Bradford, J.R.; Bulpitt, A.J.; Westhead, D.R. A Primer on Learning in Bayesian Networks for Computational Biology. *PLoS Comput. Biol.* **2007**, *3*, e129. [CrossRef]
58. Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*; Morgan Kaufmann: Burlington, MA, USA, 1988.
59. Chickering, D.M. Learning Bayesian Networks Is NP-Complete. In *Learning from Data: Artificial Intelligence and Statistics V.*; Fisher, D., Lenz, H.-J., Eds.; Lecture Notes in Statistics; Springer: New York, NY, USA, 1996; pp. 121–130. ISBN 978-1-4612-2404-4.
60. Verma, T.; Pearl, J. Equivalence and Synthesis of Causal Models. In Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence, Virtual Event, 27–29 July 1990; Elsevier Science Inc.: New York, NY, USA, 1990; pp. 255–270.
61. Spirtes, P.; Glymour, C.; Scheines, R. *Causation, Prediction, and Search*, 2nd ed.; Adaptive Computation and Machine Learning Series; A Bradford Book: Cambridge, MA, USA, 2001; ISBN 978-0-262-19440-2.
62. Borchani, H.; Bielza, C.; Martı´nez-Martı´n, P.; Larrañaga, P. 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). *J. Biomed. Inform.* **2012**, *45*, 1175–1184. [CrossRef]
63. Margaritis, D. *Learning Bayesian Network Model Structure from Data*; Carnegie-Mellon Univ Pittsburgh Pa School of Computer Science: Pittsburgh, PA, USA, 2003.
64. Tsamardinos, I.; Aliferis, C.F.; Statnikov, A. Algorithms for Large Scale Markov Blanket Discovery. *FLAIRS Conf.* **2003**, *2*, 376–380.
65. Henrion, M. An Introduction to Algorithms for Inference in Belief Nets. In *Machine Intelligence and Pattern Recognition*; Henrion, M., Shachter, R.D., Kanal, L.N., Lemmer, J.F., Eds.; Uncertainty in Artificial Intelligence; Elsevier: Amsterdam, The Netherlands, 1990; Volume 10, pp. 129–138.
66. Shachter, R.D.; Peot, M.A. Simulation Approaches to General Probabilistic Inference on Belief Networks. In *Machine Intelligence and Pattern Recognition*; Henrion, M., Shachter, R.D., Kanal, L.N., Lemmer, J.F., Eds.; Uncertainty in Artificial Intelligence; Elsevier: Amsterdam, The Netherlands, 1990; Volume 10, pp. 221–231.
67. Golightly, A.; Wilkinson, D.J. Bayesian Parameter Inference for Stochastic Biochemical Network Models Using Particle Markov Chain Monte Carlo. *Interface Focus* **2011**, *1*, 807–820. [CrossRef]
68. Dagum, P.; Luby, M. Approximating Probabilistic Inference in Bayesian Belief Networks Is NP-Hard. *Artif. Intell.* **1993**, *60*, 141–153. [CrossRef]
69. Reynolds, D. Gaussian Mixture Models. In *Encyclopedia of Biometrics*; Li, S.Z., Jain, A., Eds.; Springer US: Boston, MA, USA, 2009; pp. 659–663. ISBN 978-0-387-73003-5.
70. Madsen, A.L. Belief Update in CLG Bayesian Networks with Lazy Propagation. *Int. J. Approx. Reason.* **2008**, *49*, 503–521. [CrossRef]
71. Dean, T.; Kanazawa, K. A Model for Reasoning about Persistence and Causation. *Comput. Intell.* **1989**, *5*, 142–150. [CrossRef]
72. Quesada, D. DbnR: Dynamic Bayesian Network Learning and Inference. Available online: https://github.com/dkesada/dbnR (accessed on 10 January 2022).
73. Scutari, M. Learning Bayesian Networks with the Bnlearn R Package. *J. Stat. Softw.* **2010**, *35*, 1–22. [CrossRef]
74. Wilczyński, B.; Dojer, N. BNFinder: Exact and Efficient Method for Learning Bayesian Networks. *Bioinformatics* **2009**, *25*, 286–287. [CrossRef]
75. McGeachie, M.J.; Chang, H.-H.; Weiss, S.T. CGBayesNets: Conditional Gaussian Bayesian Network Learning and Inference with Mixed Discrete and Continuous Data. *PLoS Comput. Biol.* **2014**, *10*, e1003676. [CrossRef]
76. Margolin, A.A.; Nemenman, I.; Basso, K.; Wiggins, C.; Stolovitzky, G.; Favera, R.D.; Califano, A. ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC Bioinform.* **2006**, *7*, S7. [CrossRef]
77. Tsamardinos, I.; Brown, L.E.; Aliferis, C.F. The Max-Min Hill-Climbing Bayesian Network Structure Learning Algorithm. *Mach. Learn.* **2006**, *65*, 31–78. [CrossRef]
78. Gasse, M.; Aussem, A.; Elghazel, H. An Experimental Comparison of Hybrid Algorithms for Bayesian Network Structure Learning. In *Machine Learning and Knowledge Discovery in Databases*; Flach, P.A., De Bie, T., Cristianini, N., Eds.; Springer: Berlin/Heidelberg, Germany, 2012; pp. 58–73.
79. Heckerman, D.; Geiger, D.; Chickering, D.M. Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. *Mach. Learn.* **1995**, *20*, 197–243. [CrossRef]
80. Rissanen, J. Modeling by Shortest Data Description. *Automatica* **1978**, *14*, 465–471. [CrossRef]
81. Grünwald, P.D. *The Minimum Description Length Principle*; Adaptive Computation and Machine Learning Series; MIT Press: Cambridge, MA, USA, 2007; ISBN 978-0-262-07281-6.
82. Cooper, G.F.; Herskovits, E. A Bayesian Method for the Induction of Probabilistic Networks from Data. *Mach. Learn.* **1992**, *9*, 309–347. [CrossRef]

83. Chang, H.-H.; McGeachie, M. Phenotype Prediction by Integrative Network Analysis of SNP and Gene Expression Microarrays. In Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Boston, MA, USA, 30 August—3 September 2011; pp. 6849–6852. [CrossRef]
84. Kirkpatrick, S.; Gelatt, C.D.; Vecchi, M.P. Optimization by Simulated Annealing. *Science* **1983**, *220*, 671–680. [CrossRef] [PubMed]
85. Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N.S.; Wang, J.T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **2003**, *13*, 2498–2504. [CrossRef] [PubMed]
86. Cowell, R.G. Local Propagation in Conditional Gaussian Bayesian Networks. *J. Mach. Learn. Res.* **2005**, *6*, 1517–1550.
87. Koller, D.; Friedman, N. *Probabilistic Graphical Models: Principles and Techniques*; Adaptive Computation and Machine Learning Series; MIT Press: Cambridge, MA, USA, 2009; ISBN 978-0-262-01319-2.
88. Parker, B.J.; Wearsch, P.A.; Veloo, A.C.M.; Rodriguez-Palacios, A. The Genus Alistipes: Gut Bacteria With Emerging Implications to Inflammation, Cancer, and Mental Health. *Front. Immunol.* **2020**, *11*, 906. [CrossRef] [PubMed]
89. Huang, Q.; Zhang, X.; Hu, Z. Application of Artificial Intelligence Modeling Technology Based on Multi-Omics in Noninvasive Diagnosis of Inflammatory Bowel Disease. *J. Inflamm. Res.* **2021**, *14*, 1933–1943. [CrossRef] [PubMed]
90. Sebastiani, P.; Abad, M.; Ramoni, M.F. Bayesian Networks for Genomic Analysis. *Genom. Signal Process. Stat.* **2005**, *2*, 281–320.