Microbiome-based disease prediction with multimodal variational information bottlenecks

Filippo Grazioli*, Raman Siarheyeu, Giampaolo Pileggi, Andrea Meiser

NEC Laboratories Europe, Heidelberg, Germany

* filippo.grazioli@neclab.eu

Abstract

Scientific research is shedding more and more light on the interaction of the gut microbiome with the human body and its role in human health. Empirical evidence shows that the microbiome is strongly interconnected with the host immune system and can contribute to carcinogenesis. Furthermore, studies show that it can be used as a predictor for various diseases. In this work, we propose the Multimodal Variational Information Bottleneck (MVIB), a novel machine learning model capable of learning a joint representation of multiple heterogneous data modalities. Adopting an information theoretic interpretation of deep neural networks, MVIB allows to compute a joint stochastic encoding of the input data modalities, which acts as a minimal sufficient statistics of the inputs for the prediction of a target label. In a supervised learning setting, we use the MVIB to predict whether patients are affected by a certain disease by jointly analising the species-relative abundance and strain-level marker profiles extracted from shotgun metagenomic sequencing of microbial DNA. We propose various pre-processing methods for the marker profiles and validate our model on ten disease datasets. Additionaly, we perform transfer learning experiments by fine-tuning the model on the target disease dataset after a pre-training on all other source diseases. Furthermore, we adopt a saliency technique to interpret the outcome of the model classification and identify which microbial species and strain-level markers mostly contributed to a certain prediction. We show that MVIB achieves competitive results on the microbiome-based disease prediction task. The code for this work is publicly available at `https://github.com/nec-research/microbiome-mvib`.

Author summary

The gut microbiome is strongly correlated with human health and can be used to predict various diseases. Here, we propose the Multimodal Variational Information Bottleneck (MVIB), a novel machine learning method which allows to efficiently combine heterogeneous data modalities by computing a joint stochastic embedding. We use the MVIB to predict whether patients are affected by a certain disease and we shed light on which microbial species and gene markers are contributing significantly to the disease outcome. Specifically, we analyse species-relative abundance and strain-level marker profiles extracted from shotgun metagenomic sequencing of microbial DNA from ten different cohorts and obtained competitive classification performance on disease prediction. Furthermore, we adopt a method based on saliency to discover which microbial species and gene markers mostly contribute to the model prediction.

Introduction

The human microbiota consists in the totality of microorganisms which live in (and on) our body. Such communities of microorganisms, composed by an abundance of different species, many of them yet unknown, constitute highly complex and diverse ecosystems which interact with the host system (i.e. the human body). It is estimated, that the genes in the gut microbiome outnumber the human genes 100fold [1]. Recent studies have shown that the human microbiota plays a key role in human health and diseases [2]. Humans benefit from the presence of these microorganisms, as they enable important chemical processes. For example, the microbiota plays a crucial role in maintaining homeostasis, developing the host immune system, protecting the integrity of mucosal barriers and harvesting nutrients which would otherwise be inaccessible [2, 3]. There is also empirical evidence of how altered states of the microbiota can contribute to carcinogenesis and consistently affect therapeutic response [4]. [5] shows the potential use of the microbiota as a predictor for various diseases.

When we address the set of microbial species (microbiota) living in a host, together with information on their genome, we refer to the *microbiome*. High-throughput sequencing technologies has allowed scientists to capture a highly descriptive snapshot of microbial communities. 16S rRNA gene sequencing technology allows for a cost-effective profiling of the most common components of the human microbiome. Shotgun metagenomic sequencing technology can provide a higher resolution profile at the strain level. As the cost of shotgun metagenomic sequencing keeps decreasing and the resolution increasing, high-quality microbiome data is becoming increasingly more accessible and this will potentially lead to higher diffusion of microbiome-based applications in the human healthcare sector.

Various large-scale microbiome profiling studies, e.g. the Human Microbiome Project [6] and the MetaHIT project [7], aimed at establishing metagenomic protocols. However, despite various attempts to develop unified best practices, real standardised approaches in microbiome research have not yet been established [8–10]. The need for quantitative analysis of heterogeneous and multi-dimensional high-throughput omics data has allowed the flourishing of a broad spectrum of statistical and machine learning application for microbiome research [8]. An important machine learning challenge is disease prediction, used for sample classification in diagnostic or prognostic studies [11].

Although promising methods for microbiome-based disease prediction have been published, none addresses the entire data complexity provided by the shotgun metagenomic sequencing technology simultaneously in a single framework. Existing approaches either focus on species-relative abundance profiles, or on strain-level marker profiles. [12, 13] apply deep learning to abundance profiles of the human gut microbiome to solve the disease prediction task. MetAML [14] tackles the disease prediction problem by applying classical machine learning algorithms to either abundance or marker profiles. MicroPheno [15] sub-samples the 16s rRNA sequences via a bootstrapping process, computes a k-mer representation of the sub-sampled sequences and eventually uses the k-mer representations for disease prediction. DeepMicro [16] uses deep representation learning with autoencoders to compute embeddings of either abundance or marker profiles, which are then classified by classical downstream machine learning models. [17] targets the prediction of cardiovascular disease by using supervised learning on taxonomic features (microbial taxa). PopPhy-CNN [18] represents the microbial phylogenetic tree in a matrix together with the relative abundance of the microbial taxa and operates disease prediction with a convolutional neural network (CNN).

To the best of our knowledge, there is no existing approach which can effectively leverage on the combination of both species-relative abundance and strain-level marker profiles. This is therefore a fundamental open problem which needs to be addressed in order to fully exploit the shotgun metagenomics sequencing technology. In fact,

abundance and marker profiles provide complementary information: on the one hand, we can assume that some microbial species present in the human body have not yet been discovered, and their presence would therefore only be partially observable through gene markers. On the other hand, gene markers alone are not descriptive of the abundance distribution of the various (known) species.

Here, we present the *Multimodal Variational Information Bottleneck* (MVIB), a multimodal generalisation of the Deep Variational Information Bottleneck (Deep VIB) [19]. Leveraging on the theory of the Information Bottleneck (IB) [20], MVIB is capable to learn a joint encoding Z, that is maximally expressive about a target diagnosis Y, while being maximally compressive of the multimodal input sources $X^1, ..., X^M$. The input data modalities can be heterogeneous and in this work MVIB simultaneously combines the species-relative abundance profiles and the strain-level marker profiles to estimate the probability for a patient to be affected by a certain disease. Using deep neural networks to parametrise the relevant probability distributions allows to train MVIB in an end-to-end fashion and to optimise its objective via stochastic gradient descent. We validate MVIB with high-resolution shotgun metagenomics sequencing data for ten datasets and explore various data pre-processing techniques. Furthermore, we explain the model predictions by computing marker-specific and abundance-specific saliency maps [21]; this allows to discover which gene markers and microbial species mostly contributed to a positive disease prediction. Fig 1 depicts the whole pipeline. Experiments on transfer learning [22, 23] are performed in order to investigate the possibility to improve the accuracy of MVIB on the target disease prediction task by means of a pre-training step on source tasks with non-target disease cohorts.

MVIB is a powerful and general approach for combining multiple heterogeneous data modalities leveraging on variational inference and the IB theory. Our results show that MVIB considerably outperforms current methods for microbiome-based disease prediction.

Materials and methods

Datasets

For the experiments, publicly available human gut metagenomic samples of ten different disease cohorts were considered: inflammatory bowel disease (*IBD*), type 2 diabetes in European women (*EW-T2D*), type 2 diabetes in Chinese subjects (*C-T2D*), obesity (*Obesity*, $Obesity_{joint}$), liver cirrhosis (*Cirrhosis*), colorectal cancer (*Colorectal*, $Colorectal_{EMBL}$, $Early\text{-}Colorectal_{EMBL}$) and chronic high blood pressure (*Hypertension*). In brackets are the names we assigned to each dataset, used as identifiers in this work. In order to compare our results with DeepMicro [16] and MetAML [14], we considered the same datasets as in the original papers: IBD, EW-T2D, C-T2D, Obesity, Cirrhosis and Colorectal.

For each cohort dataset, affected patients and control subjects are available (Table 1). The IBD cohort present 25 sick subjects and 85 healthy controls [7]. The EW-T2D cohort presents 53 European women affected by type 2 diabetes and 43 healthy subjects [25]. The C-T2D cohort presents 170 Chinese subjects affected by type 2 diabetes and 174 healthy controls [26]. The Obesity cohort presents 164 obese subjects and 89 lean individuals [27]. The Cirrhosis cohort presents 118 sick subjects 114 healthy controls [28]. The Colorectal cohort presents 48 subjects affected by colorectal cancer and 73 healthy controls [29]. All 1,156 pre-processed human gut metagenomic samples that constitute the IBD, EW-T2D, C-T2D, Obesity, Cirrhosis and Colorectal datasets were obtained from the MetAML repository [14].

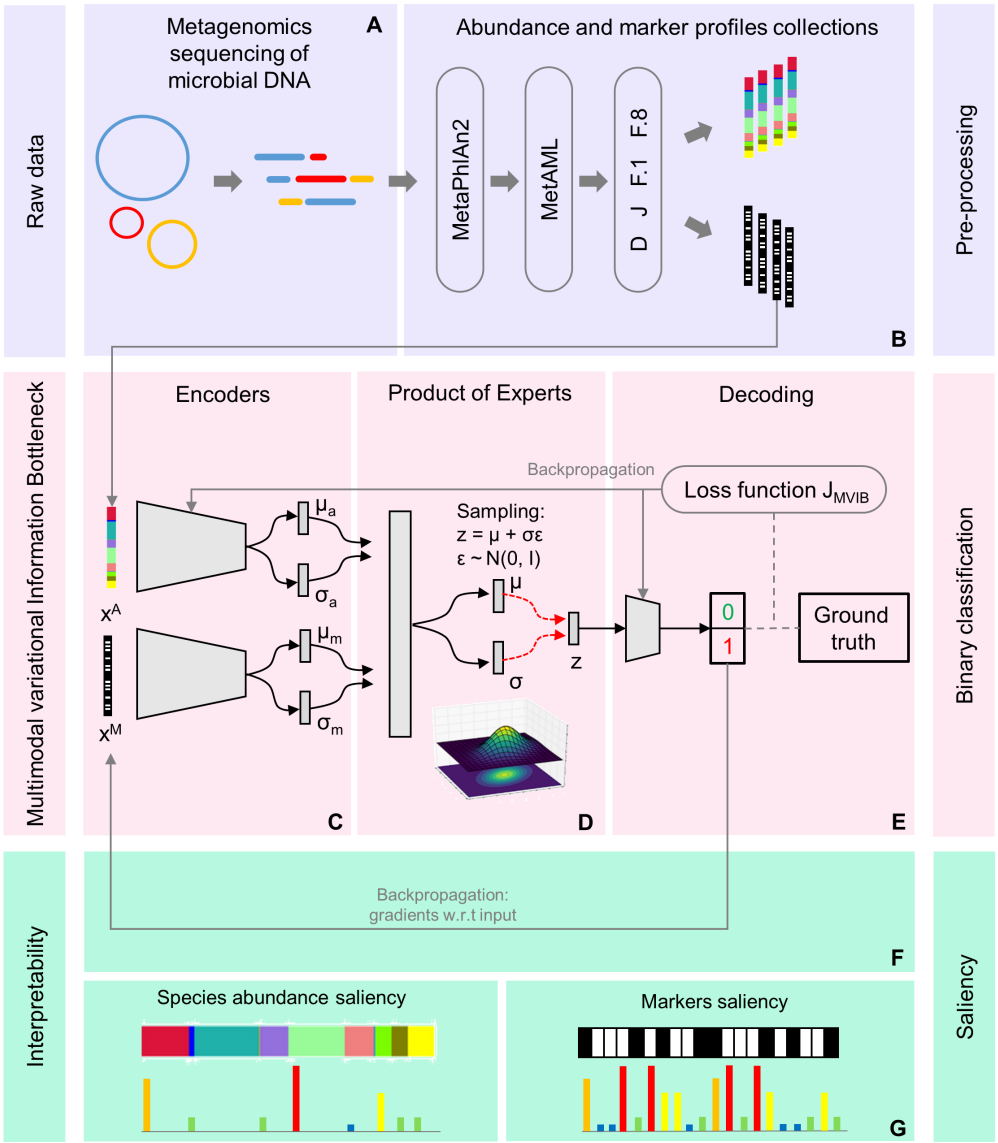

Fig 1. The MVIB applied for multimodal microbiome-based disease prediction with saliency. (A) Raw data: shotgun metagenomics sequencing data of the human gut microbiome. (B) Pre-processing steps. MetaPhlAn2 and MetAML are used for extracting the species-relative abundance profiles and strain-level marker profiles. Four pre-processing steps are proposed and investigated: *default* (D), *joint* (J), *filter.1* (F.1), *filter.8* (F.8). They allow to constitute four datasets collections: D, J, F.1 and F.8. (C) The probabilistic encoders of the MVIB model. (D) The Product of Experts computes a single joint posterior; an embedding $\boldsymbol{z}$ is sampled with the reparameterisation trick [24]. (E) A logistic regression decoder estimates the probability of the patient to be affected by a certain disease. (F, G) The gradients of the output class are computed w.r.t the input vectors to compute saliency maps.

$\text{Obesity}_{\text{joint}}$, $\text{Colorectal}_{\text{EMBL}}$, $\text{Early-Colorectal}_{\text{EMBL}}$ and Hypertension were additional disease cohorts we used to test our disease prediction method. We did all data per-processing for these subjects from high-throughput whole-genome sequencing

reads. The $\text{Obesity}_{\text{joint}}$ cohort consists of subjects from [27] used in MetAML repository [14] and supplemented with 50 obese and 28 healthy subjects mentioned in the longitudinal study [42]. The $\text{Colorectal}_{\text{EMBL}}$ consists of 96 affected and 103 healthy controls extracted from study [29]. Comparing to Colorectal dataset from MetAML repository [14], it includes 78 extra subjects from cohorts F and G [29]. The $\text{Early-Colorectal}_{\text{EMBL}}$ presents 96 affected samples from dataset $\text{Colorectal}_{\text{EMBL}}$ with 52 subjects labeled as "early stage" (colorectal cancer stages 0, I and II), and 44 labeled as "late stage" (stages III and IV). The Hypertension cohort presents shotgun metagenomic sequencing of fecal samples from 156 Chinese individuals with pre-hypertension (pHTN) or primary hypertension (HTN), and 40 healthy controls [30]. We considered both pHTN and HTN patients as affected subjects.

The entirety of the presented data derives from whole-genome shotgun metagenomic studies which adopted the Illumina paired-end sequencing technology (Fig 1A). For each sample that passed quality control, MetaPhlAn2 [31] was adopted to extract two microbiome profiles: (1) a species-level relative abundance profile, and (2) a strain-level marker profile. Default parameters were used for MetaPhlAn2. The strain-level marker profile consists in a vector of binary variables: 0 indicates the absence of a certain strain, while 1 indicates its presence. The species-relative abundance profile consists in a vector of continuous variables in $\mathbb{R}^{[0,1]}$, the sum of which is always 1. MetAML was used to pre-process the extracted abundance profile in order to exclude sub-species-level features. The dimensionality of the marker profiles ranges up to an order of magnitude of 10^5, while abundance profiles present an order of magnitude of 10^2. Table 1 presents an overview of the datasets described in this section.

Table 1. An overview of microbiome datasets used for disease prediction

Dataset name	Total samples	Controls	Affected	Data source references
IBD	110	85	25	[14], [7]
EW-T2D	96	43	53	[14], [25]
C-T2D	344	174	170	[14], [26]
Obesity	253	89	164	[14], [27]
$\text{Obesity}_{\text{joint}}$	331	117	214	[42], [43], [14], [27]
Cirrhosis	232	114	118	[14], [28]
Colorectal	121	73	48	[14], [29] (cohort F)
$\text{Colorectal}_{\text{EMBL}}$	199	103	96	[29] (cohorts F and G)
$\text{Early-Colorectal}_{\text{EMBL}}$	96	52	44	[29] (cohorts F and G)
Hypertension	196	40	156	[30]

Pre-processing and strain-level markers selection

First, all metagenomic samples from $\text{Colorectal}_{\text{EMBL}}$, $\text{Early-Colorectal}_{\text{EMBL}}$ and Hypertension datasets, as well as samples from $\text{Obesity}_{\text{joint}}$ that are not contained in [14], were run through quality control and MetaPhlAn2-based annotation pipeline to get species-level relative abundance and strain-level marker profiles compatible to the datasets from MetAML repository [14].

SRAs that match the runs of paired-end metagenome sequencing experiments were downloaded from the corresponding NCBI bioproject repositories and converted into fastq files with *fastq-dump v2.8.0*, part of NCBI SRA Toolkit [40]. The created run-level fastq files were aggregated by sample identifier. Afterwards, these aggregated fastq files passed the quality control of Kneaddata v0.7.4 [41] with default parameters. Reads trimmed to less than 60pb were removed. All processed files were then profiled with MetaPhlAn2 [31] to get species-relative abundances and strain-level marker profiles,

MetAML dataset selection script [14] was used to leave only species-level features in the abundance profiles.

We then created four collections of datasets, each of which derived from a different pre-processing methodology: *default*, *joint*, *filter.1* and *filter.8*. Each collection includes all disease datasets.

The *default* collection contains species relative adundances and gene markers as they were extracted by the MetPhlAn2 + MetAML pipeline. Only standardisation was applied to abundances.

The *joint* collection was created for transfer learning experiments. Abundances and markers were made homogeneous across all diseases by creating a union of the datasets to keep the same positional indexes of the microbial species and markers across all datasets.

In the *filter.1* and *filter.8* collections, standardisation was applied on abundances. Additionally, for each gene marker, an odds ratio (OR) score was calculated based on its abundance in affected and healthy groups (R package questionr) and gene markers were classified as associated or non-associated. P-values were corrected for multiple testing with the Benjamin & Hochberg method (R package stats). Only associated gene markers were included into the final collection according to P-values ($P < 0.1$ for *filter.1* and $P < 0.8$ for *filter.8*).

The Multimodal Variational Information Bottleneck

Let Y be a random variable representing a ground truth label associated with a set of multimodal input random variables $X^1, \ldots, X^M$. In order to provide a more compact notation, let us represent the collection of the available data modalities as a data point $X = \{X^i | i^{th}\ modality\ present\}$. Let Z be a stochastic encoding of X coming from an intermediate layer of a deep neural network and defined by a parametric encoder $p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta})$ representing the upstream part of such neural model. For the rest of this manuscript, we adopt the following notation: X, Y, Z are random variables; $\boldsymbol{x}$, $\boldsymbol{y}$, $\boldsymbol{z}$ are multidimensional instances of random variables; $f(\cdot;\boldsymbol{\theta})$ are functions parametrised by a vector of parameters $\boldsymbol{\theta}$; $\mathcal{S}$ represents a set.

Following the information bottleneck approach [20], our goal consists in learning an encoding Z which is (a) maximally informative about Y and (b) maximally compressive about X. Following an information theoretic approach, objective (a) implies maximising the mutual information $I(Z, Y; \boldsymbol{\theta})$ between the encoding Z and the target Y, where:

$$I(Z, Y; \boldsymbol{\theta}) = \int p(\boldsymbol{z}, \boldsymbol{y}|\boldsymbol{\theta}) \log \frac{p(\boldsymbol{z}, \boldsymbol{y}|\boldsymbol{\theta})}{p(\boldsymbol{z}|\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta})} d\boldsymbol{y}\, d\boldsymbol{z}\ . \tag{1}$$

A trivial solution of Equation 1 would be the identity $Z = X$. This would ensure a maximally informative representation, but (b) places a constraint on Z. This leads to the objective:

$$\max_{\boldsymbol{\theta}}\ R_{IB}(\boldsymbol{\theta}) = I(Z, Y; \boldsymbol{\theta}) - \beta I(Z, X; \boldsymbol{\theta})\ , \tag{2}$$

where $\beta \geq 0$ is a Lagrange multiplier. The first term on the right hand side of Equation 2 causes Z to be predictive of Y, while the second term constraints Z to be a a minimal sufficient statistics of X. β controls the trade-off.

As derived in [19] for the Deep Variational Information Bottleneck (VIB), assuming $q(\boldsymbol{y}|\boldsymbol{z})$ and $r(\boldsymbol{z})$ are variational approximations of the true $p(\boldsymbol{y}|\boldsymbol{z})$ and $p(\boldsymbol{z})$, respectively, Equation 2 can be rewritten as:

$$J_{VIB} = \frac{1}{N}\sum_{n=1}^{N} \mathop{\mathbb{E}}_{\boldsymbol{\epsilon}\sim p(\boldsymbol{\epsilon})}[-\log q(\boldsymbol{y}_n|f(\boldsymbol{x}_n,\boldsymbol{\epsilon}))] + \beta KL[p(Z|\boldsymbol{x}_n), r(Z)]\,, \tag{3}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ is an auxiliary Gaussian noise variable, KL is the Kullback-Leibler divergence and f is a vector-valued parametric deterministic encoding function. The introduction of $\boldsymbol{\epsilon}$ consists in the reparameterisation trick [24], which allows to write $p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta})d\boldsymbol{x} = p(\boldsymbol{\epsilon})d\boldsymbol{\epsilon}$, where $\boldsymbol{z} = f(\boldsymbol{x}, \boldsymbol{\epsilon})$ is now treated as a deterministic variable. This formulation allows the noise variable to be independent of the model parameters. This way, it is easy to compute gradients of the objective in Equation 3 and optimise via backpropagation. In this work, we let the variational approximate posteriors be multivariate Gaussians with a diagonal covariance structure $\boldsymbol{z} \sim p(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2\boldsymbol{I})$; a valid reparameterisation is $\boldsymbol{z} = \boldsymbol{\mu} + \boldsymbol{\sigma}\boldsymbol{\epsilon}$.

We generalise the formulation of the VIB objective of Equation 3 by considering that X is a collection of multimodal random input variables s.t. $X = \{X^i | i^{th}\ \textit{modality present}\}$. In light of this, the posterior $p(Z|\boldsymbol{x})$ of Equation 3 consists actually in the joint posterior $p(Z|\boldsymbol{x}^1, ..., \boldsymbol{x}^M)$, marginalised over the joint M available data modalities. Following the approach proposed for the Multimodal Variational Autoencoder [32], assuming conditional independence between the various modalities conditioned on Z and approximating $p(Z|\boldsymbol{x}^i)$ with $q(Z|\boldsymbol{x}^i) = \widetilde{q}(Z|\boldsymbol{x}^i)p(Z)$, where $\widetilde{q}(Z|\boldsymbol{x}^i)$ is the stochastic encoder of the ith mode and $p(Z)$ is a prior, the joint posterior can be expressed as a product of single-modality posteriors:

$$p(Z|\boldsymbol{x}^1, ..., \boldsymbol{x}^M) \propto \frac{\prod_{i=1}^{M} p(Z|\boldsymbol{x}^i)}{\prod_{i=1}^{M-1} p(Z)} \approx \frac{\prod_{i=1}^{M}[\widetilde{q}(Z|\boldsymbol{x}^i)p(Z)]}{\prod_{i=1}^{M-1} p(Z)} = p(Z)\prod_{i=1}^{M}\widetilde{q}(Z|\boldsymbol{x}^i)\,. \tag{4}$$

The formulation derived from Equation 4 is addressed as a product of experts (PoE). When the involved probability distributions are Gaussian, the PoE acquires a simple analytical solution, as the product of Gaussian experts is itself a Gaussian [33]. We can now formulate the objective of the Multimodal Variational Information Bottleneck:

$$J_{MVIB} = \frac{1}{N}\sum_{n=1}^{N} \mathop{\mathbb{E}}_{\boldsymbol{\epsilon}\sim p(\boldsymbol{\epsilon})}[-\log q(\boldsymbol{y}_n|f(\boldsymbol{x}_n^i, ..., \boldsymbol{x}_n^M, \boldsymbol{\epsilon}))] + \beta KL[p(Z)\prod_{i=1}^{M}\widetilde{q}(Z|\boldsymbol{x}_n^i), r(Z)]\,. \tag{5}$$

Implementation details

In this work, there are two relevant data modalities: the species-relative abundance profiles and the strain-level markers profiles. For each modality, the dedicated stochastic encoder has the form $\widetilde{q}(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(f_e^{\mu}(\boldsymbol{x}), f_e^{\sigma}(\boldsymbol{x}))$. f_e is a multi-layer perceptron (MLP). For abundance profiles, the MLP has 2 layers of the form *input dimension* - *input dimension*/2 - *input dimension*/2, followed by two parallel layers which output 2 vectors of size K for μ and σ. K is the size of the bottleneck. For a more stable computation, we let $f_e^{\sigma}(\boldsymbol{x})$ model the logarithm of the variance $\log \boldsymbol{\sigma}^2$. For marker profiles, the MLP has 2 layers of the form *input dimension*/2 - 1024 - 1024, followed by two parallel layers which output 2 vectors of size K for μ and σ. SiLU [34] activation functions are used. 0.4 drop-out is used at training time.

The decoder consists in a logistic regression model $q(\boldsymbol{y}|\boldsymbol{z}) = \sigma(f_d(\boldsymbol{z}))$, where $\sigma(\cdot)$ is the logistic sigmoid function and $f_d(\boldsymbol{z}) = \boldsymbol{w}^T\boldsymbol{z} + \boldsymbol{b}$. This implements the binary classification.

In Equation 5, $r(Z)$ and $p(Z)$ are treated as a K-dimensional spherical Gaussian distributions, $r(\boldsymbol{z}) = p(\boldsymbol{z}) = \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. The latent dimension K of the embedding was set to 256. β was set to 10^{-5}.

The networks were trained using the Adam optimiser, with a learning rate of 10^{-4} and a L_2 weight decay with $\lambda = 10^{-5}$. The batch size was set to 256. The training was performed for 200 epochs and, in order to avoid overfitting, the best model was selected by saving the weights corresponding to the epoch where the area under the curve (AUC) of the receiver operating characteristic (ROC) was maximum on the validation set.

All experiments were performed on a CentOS Linux 8 machine with NVIDIA GeForce RTX 2080 Ti GPUs and CUDA 10.2 installed, with the exception of the transfer learning experiments(see Section Transfer learning), which were performed with an NVIDIA TITAN RTX GPU. Algorithms were implemented in Python 3.6 using PyTorch 1.7; code is available at `https://github.com/nec-research/microbiome-mvib`.

Extending the MVIB objective with the triplet margin loss

The MVIB objective function proposed in Equation 5 consists of two terms: a supervised negative log-likelihood loss and a KL divergence which acts as a regulariser. As presented in Section Results, it was empirically observed that extending the J_{MVIB} of Equation 5 with an additional triplet margin loss [35] term can lead to a more accurate predictor on various disease datasets.

Given a certain batch size B of patient samples on which the loss and the gradient are computed, let B_P be the number of positive (sick) subjects and B_N the number of negative (control) patients. Let $B_T = min\{B_P, B_N\}$. From a batch of size B, we randomly sample B_T positive samples and B_T negative samples without repetitions (this implies that the smaller set of samples, either negative or positive, will be fully considered); these $2B_T$ samples constitute the *anchors* of the triplet margin loss and we represent them with two sets: $\mathcal{A}_+$ and $\mathcal{A}_-$. For each sampled anchor, (a) a positive sample of the same class and (b) a negative sample of the opposite class are sampled from the two anchor sets. In our implementation, this is obtained by shuffling the indexes of the anchors. This allows to constitute a set of positive samples $\mathcal{P}_+$ and a set of negative samples $\mathcal{N}_+$ for the positive anchors. A set of positive samples $\mathcal{P}_-$ and a set of negative samples $\mathcal{N}_-$ are obtained for the negative anchors too. Patients contained in the $\mathcal{P}_-$ set are sick subjects, while subjects contained in $\mathcal{N}_-$ are healthy control patients. This is because, in this context, *positive* and *negative* actually mean of the *same* and of *opposite* ground truth label.

We define the triplet margin loss as:

$$L_T(\mathcal{A}, \mathcal{P}, \mathcal{N}) = \sum_{\substack{\boldsymbol{x}_A \in \mathcal{A}, \\ \boldsymbol{x}_P \in \mathcal{P}, \\ \boldsymbol{x}_N \in \mathcal{N}}} max\{\|f_e^\mu(\boldsymbol{x}_A) - f_e^\mu(\boldsymbol{x}_P)\|_2^2 - \|f_e^\mu(\boldsymbol{x}_A) - f_e^\mu(\boldsymbol{x}_N)\|_2^2 + \alpha, 0\}\,, \quad (6)$$

where α is a tunable margin which we set to 1 and $f_e^\mu(\boldsymbol{x})$ is the MLP that computes the mean of $\widetilde{q}(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(f_e^\mu(\boldsymbol{x}), f_e^\sigma(\boldsymbol{x}))$. It follows that, for our $\mathcal{A}_+$ and $\mathcal{A}_-$ anchors sets, the triplet loss objective can be written as:

$$J_T = L_T(\mathcal{A}_+, \mathcal{P}_+, \mathcal{N}_+) + L_T(\mathcal{A}_-, \mathcal{P}_-, \mathcal{N}_-)\,. \quad (7)$$

Intuitively, the first term of of the right hand side of Equation 8 encourages the encodings of control patients to be closer to each other in their K-dimensional Euclidean space and far away from the encodings of sick patients. In the same fashion, the second

term of the equation encourages the encodings of sick patients to be clustered in the same region of the latent space and to be distant from the encodings of healthy patients. This leads to an extension of the MVIB objective presented in Equation 5 which define as MVIB-T objective:

$$J_{MVIB-T} = J_{MVIB} + \lambda_T J_T \ . \tag{8}$$

λ_T is a multiplying constant which we set to 1 through all experiments.

Sub-sampled training paradigm for missing data

For the training of MVIB, we adopt the same sub-sampled training paradigm proposed for the Multimodal Variational Autoencoder [32]. The MVIB objective presented in Equation 8 assumes that all M data modalities are present. This has the unfortunate consequence of not training the single-modality encoders $\widetilde{q}(\boldsymbol{z}|\boldsymbol{x}^i)$ individually. This implies that the model cannot be used if missing data modalities are presented at test time. In order to circumvent this limitation and allow the MVIB to work at test time with missing data modalities, we need to compute the MVIB objective for the combined data modalities, as well as for the individual modalities. Following the sub-sampled training paradigm, Equation 8 can be formulated as:

$$J_{MVIB}(\boldsymbol{x}^A) + J_{MVIB}(\boldsymbol{x}^M) + J_{MVIB}(\boldsymbol{x}^A, \boldsymbol{x}^M) \ , \tag{9}$$

where $\boldsymbol{x}^A$ represents the species-relative abundance profiles and $\boldsymbol{x}^M$ the strain-level marker profiles. This extension holds also for the MVIB-T objective of Equation 8.

Explaining predictions with saliency

Following the same paradigm proposed by [21], given a multimodal pair $(\boldsymbol{x}_0^A, \boldsymbol{x}_0^M)$, where $\boldsymbol{x}^A$ represents the species-relative abundance profile and $\boldsymbol{x}^M$ the strain-level marker profile, we would like to rank the gene markers of $\boldsymbol{x}^M$ and the microbial species of $\boldsymbol{x}^A$ based on their influence on the MVIB prediction. More formally, we compute the derivate of the MVIB class prediction with respect to both input vectors $\boldsymbol{x}^A$ and $\boldsymbol{x}^M$:

$$\boldsymbol{w}^A = \frac{\partial q(\boldsymbol{y}|f(\boldsymbol{x}^A, \boldsymbol{x}^M, \boldsymbol{\epsilon}))}{\partial \boldsymbol{x}^A}\Big|_{\boldsymbol{x}_0^A} \ , \ \boldsymbol{w}^M = \frac{\partial q(\boldsymbol{y}|f(\boldsymbol{x}^A, \boldsymbol{x}^M, \boldsymbol{\epsilon}))}{\partial \boldsymbol{x}^M}\Big|_{\boldsymbol{x}_0^M} \ , \tag{10}$$

where $q(\cdot)$ is the parametric probabilistic decoder of MVIB and $f(\cdot)$ represents its whole multimodal encoding block, following the same notation of Equation 5.

The magnitude of the derivatives presented in Equation 10 indicates which gene markers and species-relative abundances need to be changed the least to affect the class score the most. For each disease dataset, we compute the saliency maps for all test samples classified as positive (i.e. sick) and compute their average. This allows to discover a ranking of the most influential gene markers and species for each diseases datasets. The computation of saliency maps is extremely quick and it only requires a single backpropagation pass.

Transfer learning

Motivated by the hypothesis that altered microbiome states caused by two different diseases might in fact share some common patterns, we performed experiments following

a transfer learning paradigm [22, 23]. Iteratively, we first select a target disease from the set of ten considered datasets. This allows to define a target domain $\mathcal{D}_T = \{\mathcal{X}_T, p(\boldsymbol{X}_T)\}$, where $\mathcal{X}_T$ is the target feature space and $p(\boldsymbol{X}_T)$ is the marginal distribution of the set of samples of the target dataset $\boldsymbol{X}_T = \{\boldsymbol{x}_i | i \in \mathcal{X}_T, i = 1, ..., N\}$. A target task $\mathcal{T}_T = \{\mathcal{Y}_T, f\}$ is also defined, where $\mathcal{Y}_T$ is the label space of the target dataset and f the decision function which is expected to be learned. Merging the nine non-target datasets, we constitute a source domain and task $(\mathcal{D}_S, \mathcal{T}_S)$. For the source task, all sick patients are treated equally and they are assigned a positive ground truth label, independently on what pathology they actually have. We first train MVIB on the source domain and task $(\mathcal{D}_S, \mathcal{T}_S)$ and fine-tune it on the target domain and task $(\mathcal{D}_T, \mathcal{T}_T)$. We performed transfer learning experiments only for the *joint* datasets collection, as the microbial species and strain markers, as well as their positional indexes in the input feature vectors, are the same across the various diseases.

Validation framework and performance evaluation

With the goal of providing a non-biased estimate of the model's accuracy, we implemented a performance evaluation scheme inspired by DeepMicro [16]. We split each dataset into training and test sets with a 8:2 ratio. In particular, for the IBD, EW-T2D, C-T2D, Obesity, Cirrhosis and Colorectal datasets, we implemented the random train-test split using the same random partition seeds used by DeepMicro; this ensures that out test patients are the same ones considered by DeepMicro and allows a fair benchmark of the models' performances. The same procedure was adopted for $\text{Obesity}_{\text{joint}}$, $\text{Colorectal}_{\text{EMBL}}$, $\text{Early-Colorectal}_{\text{EMBL}}$ and Hypertension too. Considering only the training set, we performed a stratified 5-fold cross validation and used the validation sets to compute a validation ROC AUC score for selecting the epoch with the best model parameters. The five best models obtained via the 5-fold cross validation were then tested on the left-out test set and their predictions were ensembled via a majority vote. This procedure was repeated five times, each time with a different random partition seed, in order to ensure that the five experiments were conducted with independent random training-test splits. The resulting test performance metrics coming from the five independent experiments were then averaged and their mean was used for comparing model performance.

Results

MVIB achieves competitive results on the multimodal microbiome-based disease prediction task

To evaluate the performance of the MVIB, we performed a comparison with DeepMicro [16] and PopPhy-CNN [18], which are the state-of-the-art approaches for microbiome-based disease prediction at the time this paper is being written. In Table 2 we report the best AUC of the ROC for DeepMicro, PopPhy-CNN, and MVIB. MVIB results are reported for the dataset collections *default* and *joint*. For DeepMicro, both abundance (A) and markers (M) profiles results are reported. Furthermore, we also computed DeepMicro's performance on the multimodal disease prediction task by training the model on the concatenation of markers and abundance profiles (A+M). For PopPhy-CNN, only abundance (A) results are reported, as the method does not consider strain-level marker profiles.

The DeepMicro method presents a 2-steps approach: (1) train an autoencoder to embed the input; (2) train a downstream classifier on the embeddings computed at the autoencoder's bottleneck. Various autoencoder architectures and downstream models

are presented in [16] and there is no clear criteria concerning which combination works best. For the benchmark presented in Table 2, we consider the DeepMicro results obtained with the Variational Autoencoder [24]. We believe this is the most significant comparison, since MVIB is also based on variational inference. This allows to highlight the improvements derving from our end-to-end multi-modal architecture with respect to DeepMicro's 2-steps approach.

Considering the multimodal A+M results and the *default* and *joint* dataset collection (see Table 2), we observed that MVIB consistently outperforms DeepMicro on all datasets. Early-Colorectal$_{\text{EMBL}}$ and Hypertension, two of the newly introduced datasets, appear to be the hardest disease cohorts to detect by means of microbiome-based diagnosis and the MVIB only achieves 0.535 and 0.603 ROC AUC, respectively, on the *default* datasets collection. On the Obesity$_{\text{joint}}$ and Colorectal$_{\text{EMBL}}$ datasets, MVIB achieves 0.815 and 0.811 ROC AUC, respectively, on the *default* datasets collection

In comparison with PopPhy-CNN, MVIB achieves comparable results on Obesity. PopPhy-CNN achieves a 2% higher ROC AUC on Cirrhosis, but is outperformed by MVIB on C-T2D. In Table 2 the standard error for the PopPhy-CNN is omitted as not provided in [18].

A complete overview of the MVIB results for all diseases and all dataset collections is available in S1 Table.

Improving disease prediction with transfer learning

Driven by the hypothesis that different diseases might lead to some common altered patterns in the patients' microbiome, we performed transfer learning experiments. As described in Section Transfer learning, we first pre-trained MVIB on all non-target diseases datasets and then fine-tuned the model on the target diseases. For these experiments, only the *joint* datasets collection was adopted, as it presents feature vectors with the same dimensions across all disease cohorts for both the species-relative and strain-level marker profiles. Table 3 presents the MVIB classification performance achieved by performing a pre-training on the source task followed by a fine-tuning on the target diseases (see Section Transfer learning). Although the ROC AUC slightly decreases on the IBD, EW-T2D, Cirrhosis and Early-Colorectal$_{\text{EMBL}}$ datasets, consistent improvements can be observed on all other datasets, w.r.t randomly initialised models.

For the Colorectal and Colorectal$_{\text{EMBL}}$ datasets, the ROC AUC increases from 0.777 to 0.925 and from 0.814 to 0.850, respectively. For the Obesity and Obesity$_{\text{joint}}$ datasets, the ROC AUC increases from 0.666 to 0.976 and from 0.818 to 0.978, respectively. A minor improvement can also be observed for the C-T2D dataset.

The consistent improvement in classification achieved via transfer learning for the Colorectal, Colorectal$_{\text{EMBL}}$, Obesity and Obesity$_{\text{joint}}$ datasets can be motivated by the fact that, when these datasets are iteratively considered as the target, samples from other source datasets which share the same pathology are observed by the model during the solution of the source task. For instance, when the model is being fine-tuned on the Colorectal$_{\text{EMBL}}$ dataset, samples from Colorectal had already observed during the pre-training. For the Hypertension dataset, however, this argument does not apply and the model achieves 6% ROC AUC boost from knowledge acquired from different disease cohorts.

The triplet margin loss can improve classification

The first term of the original MVIB objective J_{MVIB} (see Equation 5) is a negative log-likelihood, which acquires the shape of a binary cross-entropy in the binary classification setting. We explored an extension of this objective by adding a triplet

Table 2. MVIB performance overview.

Dataset	DeepMicro-VAE [16]‡			PopPhy-CNN [18]	MVIB (A+M)	
	A	M	A+M	A	D	J
IBD	0.779 (0.032)	0.899 (0.039)	0.833 (0.034)	†	0.922 (0.020)	**0.936 (0.014)**
EW-T2D	0.640 (0.051)	0.853 (0.041)	0.741 (0.080)	†	**0.859 (0.023)**	0.853 (0.025)
C-T2D	0.715 (0.031)	0.719 (0.019)	0.714 (0.030)	0.690 (-)	0.750 (0.009)	**0.758 (0.012)**
Obesity	0.600 (0.030)	0.599 (0.014)	0.576 (0.040)	**0.666 (-)**	0.662 (0.024)	**0.666 (0.027)**
Cirrhosis	0.781 (0.021)	0.891 (0.016)	0.867 (0.018)	**0.946 (-)**	0.925 (0.005)	0.924 (0.005)
Colorectal	0.739 (0.070)	0.737 (0.068)	0.548 (0.040)	†	**0.780 (0.071)**	0.777 (0.069)
$\text{Obesity}_{\text{joint}}$	†	†	†	†	0.815 (0.019)	**0.818 (0.018)**
$\text{Colorectal}_{\text{EMBL}}$	†	†	†	†	0.811 (0.010)	**0.814 (0.013)**
$\text{Early-Colorectal}_{\text{EMBL}}$	†	†	†	†	0.535 (0.050)	**0.543 (0.048)**
Hypertension	†	†	†	†	**0.603 (0.045)**	0.591 (0.041)

All values are ROC AUCs computed on the test sets. Values in brackets refer to the standard error over five repeated experiments. The first group of column presents the best results obtained with DeepMicro with VAE. The second group of columns presents the MVIB results. A, M, A+M respectively express that results are obtained by processing abundance profiles only, marker profiles only or abundance and marker profiles simultaneously. D and J refer to the pre-processing techniques adopted and the collections of datasets obtained: *default* (D), *joint* (J). The MVIB implementation follows the design presented in Section Implementation details and the bottleneck dimension was set to 256. The J_{MVIB-T} objective is adopted (see Equation 8).
†: For these datasets, DeepMicro or PopPhy-CNN did not perform experiments.
‡: We selected the best test ROC AUC from the DeepMicro paper over multiple downstream classifier models (SVM, MLP, Random Forest).

margin loss term (see Equation 8). The triplet margin loss aims to encourage the latent distribution of samples which belong to the same class to cluster in a dedicate region of the latent Euclidean space. This aims at reducing variance. At the same time, the triplet margin loss encourages the distributions of samples of different classes to depart from each other in the latent space. This aims at increasing the separability of different classes and facilitating the classification task. Table 4 presents a comparison of the effects of the triplet margin loss on the MVIB classification performance. For the IBD, EW-T2D, Obesity, $\text{Obesity}_{\text{joint}}$ and $\text{Early-Colorectal}_{\text{EMBL}}$ datasets, the highest ROC AUC is achieved by optimising the J_{MVIB-T} objective. The J_{MVIB} leads to best results on the remaining datasets.

Fig 2A depicts the 95% confidence interval of the patients' stochastic embeddings $\boldsymbol{z} \sim p(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2\boldsymbol{I})$ deriving from the optimisation of the J_{MVIB-T} objective (see Equation 8). In comparison with the curves of Fig 2C, obtained with the original MVIB objective J_{MVIB} (see Equation 5), we observe that the triplet margin loss leads to Gaussian distributions which are clustered in the same region of the space. Furthermore,

Table 3. Comparison of pre-trained models against randomly initialised models.

Dataset	Random initilisation	Pre-trained model
	J	J
IBD	**0.936 (0.014)**	0.868 (0.033)
EW-T2D	**0.853 (0.025)**	0.826 (0.028)
C-T2D	0.758 (0.012)	**0.765 (0.010)**
Obesity	0.666 (0.027)	**0.976 (0.008)**
Cirrhosis	**0.924 (0.005)**	0.912 (0.008)
Colorectal	0.777 (0.069)	**0.925 (0.034)**
$\text{Obesity}_{\text{joint}}$	0.818 (0.018)	**0.978 (0.006)**
$\text{Colorectal}_{\text{EMBL}}$	0.814 (0.013)	**0.850 (0.031)**
$\text{Early-Colorectal}_{\text{EMBL}}$	**0.543 (0.048)**	0.523 (0.058)
Hypertension	0.591 (0.041)	**0.667 (0.033)**

The first column presents the results obtained with randomly initialised models. The second column displays classification results obtained by first pre-training the models on all source datasets, and the fine-tuning them on the target disease. The J_{MVIB-T} objective was adopted for both columns (see Equation 8). J refers to the adopted pre-processing technique: *joint* (see Section Pre-processing and strain-level markers selection). Reported values are ROC AUCs computed on the test sets. In brackets, the standard error over five repeated experiments is reported.

stochastic embeddings deriving from J_{MVIB} (see Equation 5) present higher variance in their $\boldsymbol{\sigma}^2$ (see Fig 2C). A better separability of the embedding means $\boldsymbol{\mu}$ can be achieved by the J_{MVIB-T} objective (see Fig 2B) in comparison with the J_{MVIB} objective (see Fig 2D).

The effect of strain-level marker selection

Overall, we observed that MVIB achieved the highest performance on the disease prediction task when combined with the marker filtration techniques described in Section Pre-processing and strain-level markers selection (see Table 4 and S1 Table). For IBD, the F.1 filtration led to the highest ROC AUC. For all other datasets, the highest ROC AUC was achieved with the F.8 filtration. For the $\text{Early-Colorectal}_{\text{EMBL}}$ and Hypertension datasets, both F.1 and F.8 output empty markers sets due to the negligible difference in the markers distributions between sick and healthy control patients. These empirical results show that strain markers filtration can be beneficial and facilitate classification. Fig 3 depicts the means $\boldsymbol{\mu}$ of stochastic patient embeddings $\boldsymbol{z} \sim p(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2\boldsymbol{I})$. Depicted in the figure are the embeddings of test samples for the IBD dataset obtained by setting the latent dimension K to 2. It is possible to visualise the effect of the pre-processing method. In particular, F1 allows the highest separability. This is in line with the classification performance results presented in Table 4 for IBD.

Discovering the most influential microbial species and strain-level markers

In order to achieve interpretability, we implemented a method which allows to compute saliency, i.e. the areas of the input vectors which are most discriminative with respect to the predicted class. For each disease dataset in each dataset collection, we obtained saliency maps of both abundance and markers profiles by computing the derivative of

Table 4. Comparison of different objective functions and pre-processing techniques.

Dataset	J_{MVIB-T}				J_{MVIB}			
	D	F.1	F.8	J	D	F.1	F.8	J
IBD	0.922 (0.020)	**0.995 (0.003)**	0.984 (0.009)	0.936 (0.014)	0.915 (0.018)	0.988 (0.012)	0.981 (0.011)	0.915 (0.018)
EW-T2D	0.859 (0.023)	*	**0.996 (0.002)**	0.853 (0.025)	0.859 (0.024)	*	0.994 (0.004)	0.855 (0.027)
C-T2D	0.750 (0.009)	0.876 (0.009)	0.945 (0.003)	0.758 (0.012)	0.754 (0.014)	0.873 (0.010)	**0.949 (0.006)**	0.756 (0.015)
Obesity	0.662 (0.024)	*	**0.761 (0.020)**	0.666 (0.027)	0.673 (0.028)	*	0.751 (0.021)	0.672 (0.030)
Cirrhosis	0.925 (0.005)	0.935 (0.007)	0.934 (0.005)	0.924 (0.005)	0.930 (0.002)	0.936 (0.006)	**0.943 (0.005)**	0.928 (0.002)
Colorectal	0.780 (0.071)	0.955 (0.015)	0.980 (0.012)	0.777 (0.069)	0.788 (0.059)	0.944 (0.016)	**0.984 (0.009)**	0.796 (0.055)
$\text{Obesity}_{\text{joint}}$	0.815 (0.019)	0.894 (0.018)	**0.983 (0.005)**	0.818 (0.018)	0.825 (0.019)	0.898 (0.013)	0.980 (0.004)	0.825 (0.018)
$\text{Colorectal}_{\text{EMBL}}$	0.811 (0.010)	0.966 (0.012)	0.979 (0.009)	0.814 (0.013)	0.827 (0.012)	0.966 (0.011)	**0.989 (0.005)**	0.830 (0.011)
$\text{Early-Colorectal}_{\text{EMBL}}$	0.535 (0.050)	*	*	**0.543 (0.048)**	0.525 (0.051)	*	*	0.533 (0.053)
Hypertension	0.603 (0.045)	*	*	0.591 (0.041)	0.622 (0.048)	*	*	**0.624 (0.049)**

All values are ROC AUCs computed on the test sets. Values in brackets refer to the standard error over five repeated experiments. The first group of columns presents the results obtained optimising the J_{MVIB-T} objective (see Equation 8), which includes the triplet margin loss. The second group of columns presents the results obtained optimising the original objective function J_{MVIB} (see Equation 5). D, F.1, F.8 and J refer to the four pre-processing techniques adopted and the four collections of datasets obtained: *default* (D), *joint* (J), *filter.1* (F.1), *filter.8* (F.8).
*: For these datasets, the F.1 or F.8 filtrations did not preserve any markers and experiments could not be preformed.

the model prediction with respect to the inputs (see Equation 10). Saliency maps were computed for all positive predictions obtained on the test sets. For each disease, a single saliency map was then obtained by averaging the saliency maps of the positive predictions. For each disease column, abundance and markers saliency have been standardised using Equation 11:

$$\boldsymbol{z} = \frac{\boldsymbol{w} - \boldsymbol{\mu}}{\boldsymbol{\sigma}} \,, \tag{11}$$

where $\boldsymbol{w}$ is the saliency map computed with Equation 10, $\boldsymbol{\mu}$ is its mean and $\boldsymbol{\sigma}$ its standard deviation. After computing the standardised saliency maps, features that presented a value higher than 0.5 across all diseases have been set to 0. This allows to ignore gradients which are high across all datasets and to better highlight the microbial species and gene markers which have different saliency for a specific disease.

Fig 4 depicts the saliency maps for the abundance and markers profiles obtained from the *joint* datasets collection. All diseases appear to share some common patterns in the saliency gradients. In particular, there are certain microbial species and gene markers which are highlighted across all disease datasets. There is, however, clear differentiation between the saliency maps across the various datasets and this highlights

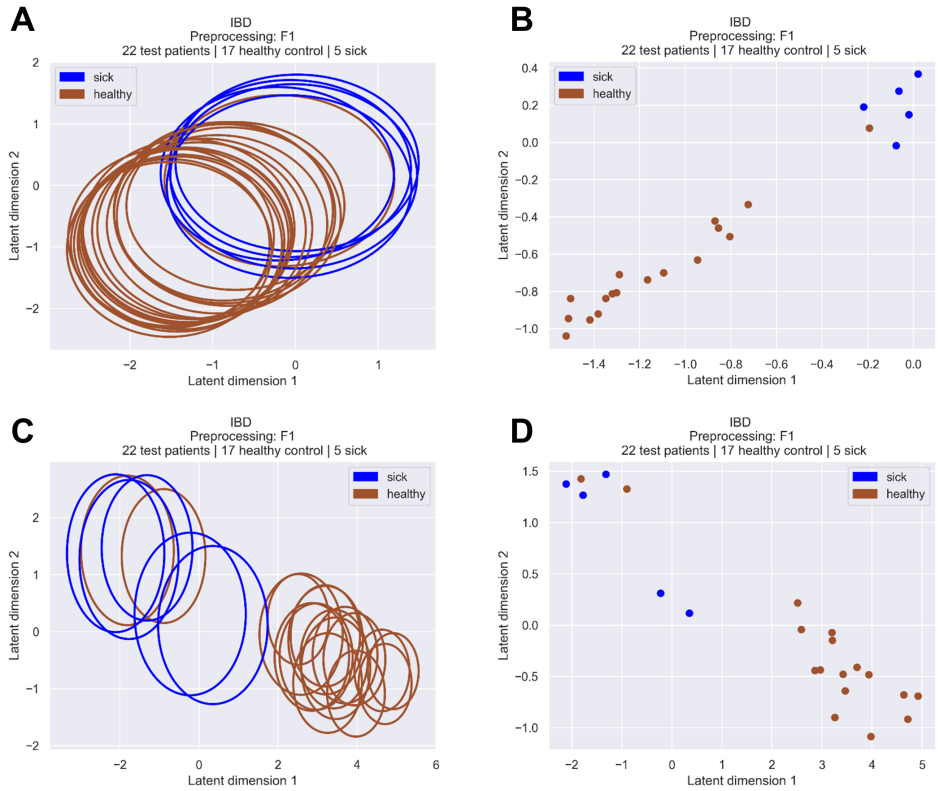

Fig 2. Effect of the triplet margin loss on the stochastic embeddings of the patients. The depicted curves are the 95% confidence intervals of the patients' stochastic embeddings $\boldsymbol{z} \sim p(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2\boldsymbol{I})$; the points are their means $\boldsymbol{\mu}$. The displayed embeddings consist in a set of 22 test samples obtained from a random training-test split of the IBD dataset. The K dimension of the latent space has been set to 2 in order to allow a 2D visualisation. (A, B) optimisation of the J_{MVIB-T} objective (see Equation 8). (C, D) optimisation of the J_{MVIB} objective (see Equation 5).

how specific microbial species and gene markers are most discriminative with respect to one specific disease. This technique allowed us to discover a ranking of the most influential gene markers and microbial species for each diseases datasets.

S2 File contains all saliency maps computed for all disease and all datasets collections; microbial species and gene markers are ranked according to the saliency score.

Discussion

The problem of microbiome-based disease prediction is challenging, for several reasons. The first is the fact that species-relative abundance and strain-level marker profiles are mutually informative and only provide a full observation of the microbiome when considered together. The second is the dimensionality of the data, which consistently exceeds the number of available samples. The third is that the microbiome rapidly changes, not only when a person is affected by a disease, but also as a function of lifestyle elements, e.g. diet [1], stress [36] and sleep [37]. This work tries to tackle all three challenges.

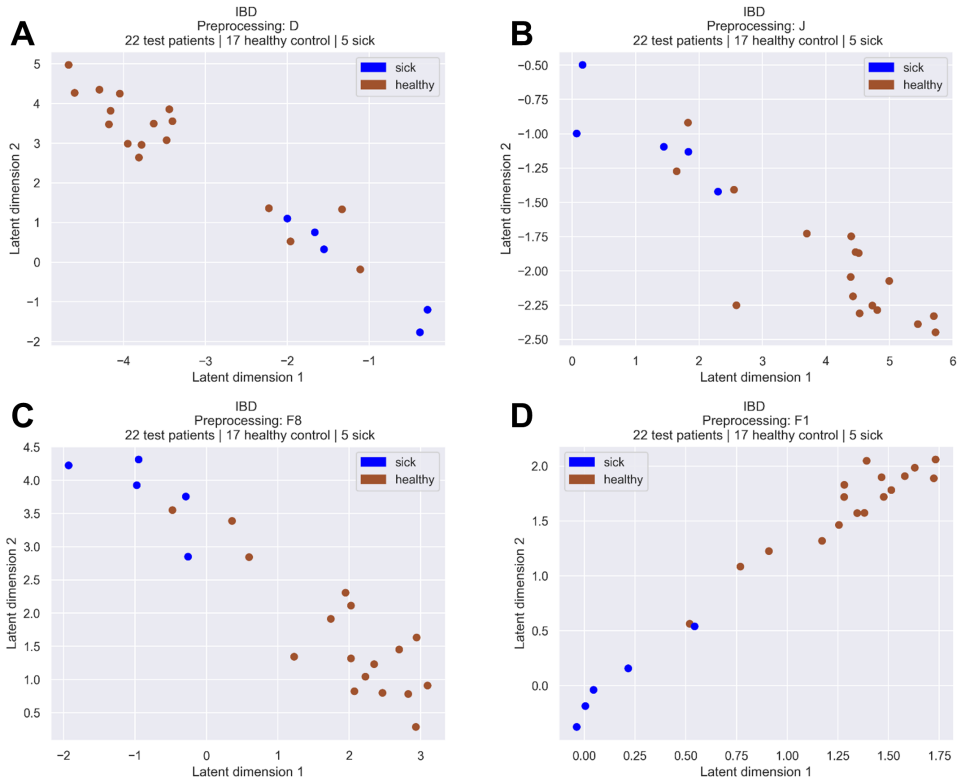

Fig 3. Effect of different pre-processing methods on the stochastic embeddings of the patients. The depicted points consist in the means $\boldsymbol{\mu}$ of stochastic patient embeddings $\boldsymbol{z} \sim p(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2\boldsymbol{I})$. The displayed embeddings consist in a set of 22 test samples obtained from a random training-test split of the IBD dataset. The K dimension of the latent space has been set to 2 in order to allow a 2D visualisation. The various plots highlight the effect of the four pre-processing techniques D, J, F1, F8 on the embeddings and the achieved separability of the samples in the latent space (see Section Pre-processing and strain-level markers selection). (A) *default* pre-processing. (B) *joint* pre-processing. (C) *filter.8* pre-processing. (D) *filter.1* pre-processing.

MVIB consists in a novel multi-modal learning approach which allows to compute a joint stochastic encoding of the species-relative abundance and strain-level marker profiles extracted from shotgun metagenomic sequencing of the microbial DNA. The stochastic encoding is maximally informative about the target label and maximally compressive about the input data modalities. This allows to combine information from both the species-relative abundance and strain-level marker profiles, which allows to compute a complete representation of the host microbiome only when considered together.

MVIB obtains competitive results when compared to the state-of-the-art methods DeepMicro and PopPhy-CNN for microbiome-based disease prediction. Furthermore, our method has the advantage of being end-to-end, while DeepMicro requires a 2-step training approach. A much more complex fine-tuning is also needed by DeepMicro, deriving from the fact that there are no well-defined criteria regarding which autoencoder architecture and downstream classifier combination works best, given a certain dataset. Conversely, we did not perform any kind of fine-tuning in this work and obtained competitive results on all examined datasets.

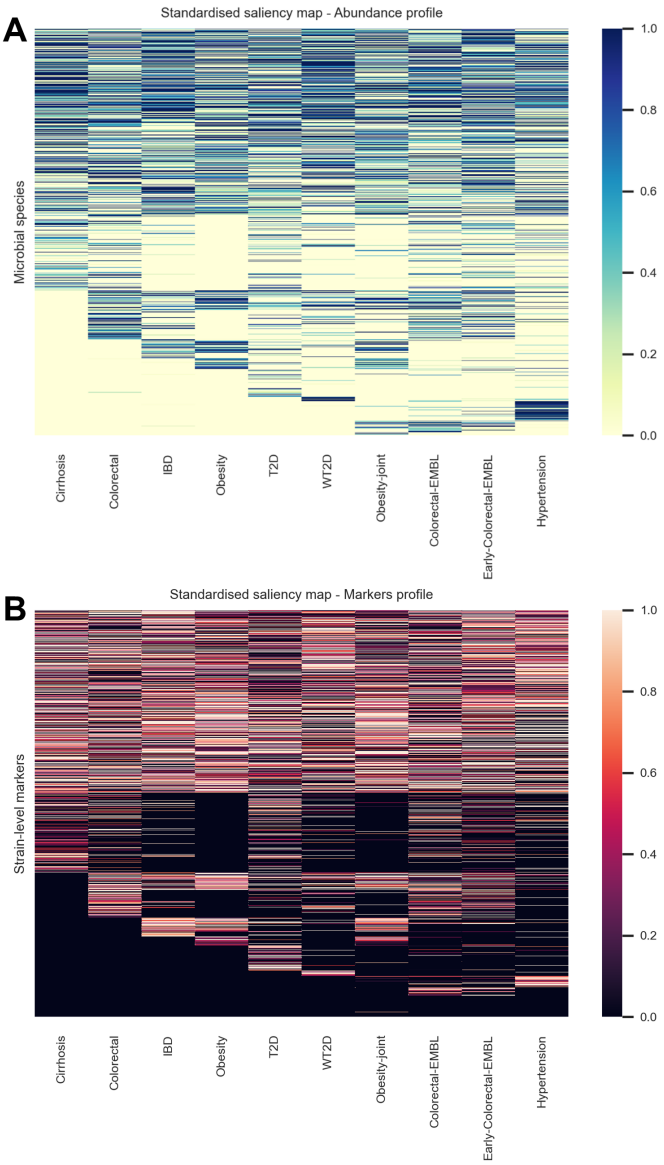

Fig 4. Saliency maps for species-relative abundance and strain-level marker profiles. Saliency maps computed for the J datasets collection using Equation 10 and Equation 11. Standardisation has been applied for each disease. Features that presented a value higher than 0.5 across all diseases have been set to 0. (A) Standardised saliency for microbial species. (B) Standardised saliency for microbial gene markers.

One of the long-term goals for MVIB is to timely identify microbiome patterns which can highlight the presence of a disease in its early stages. In the context of disease prevention, the analysis of the gut microbiome could in fact allow to timely intervene and prevent the evolution of a disease from an early and curable stage to more advance and severe ones. Compelling is also possibility to extend MVIB in order to enable the analysis of temporal longitudinal microbiome data with the aim of detecting anomalous temporal patterns.

Recent studies [38, 39] shed light on the role of the microbiome in cancer immunotherapy and on its influence on the response to cancer treatment. Evidence

suggests that modulating the gut microbiome could augment therapeutic responses. Leveraging on the multi-modal learning capabilities of MVIB, a possible direction for future research consists in extending MVIB to further host-specific and cancer-specific data modality for the estimation of the probability of success of a given immunotherapy, or for the optimisation of a given cancer treatment.

Supporting information

S1 Table. Complete experimental results for the multimodal microbiome-based disease prediction task with MVIB. Results obtained optimising the J_{MVIB-T} objective (see Equation 8). Experiments are executed five times with random independent training-test splits. Values in brackets refer to the standard error over the repeated experiments. All values in the table refer to metrics computed on the test sets. auROC: area under the under the receiver operating characteristic curve. AC: classification accuracy. F1: F1 score. P: precision. R: recall. D, F.1, F.8 and J refer to the four pre-processing techniques adopted and the four collections of datasets obtained: *default* (D), *joint* (J), *filter.1* (F.1), *filter.8* (F.8).

*: For these datasets, the F.1 or F.8 filtrations did not preserve any markers and experiments could not be preformed.

S2 File. Saliency rankings for abundance and marker profiles. These files present the saliency maps computed for both abundance and marker profiles for the D, F.1, F.8 and J datasets collections. Results are obtained by optimising the J_{MVIB-T} objective (see Equation 8). Microbial species and strain-level markers are ranked based on the computed saliency score.

Acknowledgments

We would like to acknowledge Dr. Carolin Lawrence and Dr. Francesco Alesiani for valuable discussion.

References

1. Gilbert JA, Blaser MJ, Caporaso JG, Jansson JK, Lynch SV, Knight R. Current understanding of the human microbiome. Nature medicine. 2018. 24(4):392-400.

2. Cho I, Blaser MJ. The human microbiome: at the interface of health and disease. Nature Reviews Genetics. 2012. 13(4):260-270.

3. Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, Chinwalla AT, et al. Structure, function and diversity of the healthy human microbiome. Nature. 2012. 486(7402):207.

4. McQuade JL, Daniel CR, Helmink BA, Wargo JA. Modulating the microbiome to improve therapeutic response in cancer. The Lancet Oncology. 2019. 20(2):e77-e91.

5. Eloe-Fadrosh EA, Rasko DA. The human microbiome: from symbiosis to pathogenesis. Annual review of medicine. 2013. 64:145-163.

6. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. The human microbiome project. Nature. 2007. 449(7164):804-810.

7. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature. 2010. 464(7285):59-65.

8. Moreno-Indias I, Lahti L, Nedyalkova M, Elbere I, Roshchupkin G, Adilovic M, et al. Statistical and machine learning techniques in human microbiome studies: contemporary challenges and solutions. Frontiers in Microbiology. 2021. 12:277.

9. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. Shotgun metagenomics, from sampling to analysis. Nature biotechnology. 2017. 35(9):833-844.

10. Knight R, Vrbanac A, Taylor BC, Aksenov A, Callewaert C, Debelius J, et al. Best practices for analysing microbiomes. Nature Reviews Microbiology. 2018. 16(7):410-422.

11. LaPierre N, Ju CJT, Zhou G, Wang W. MetaPheno: A critical evaluation of deep learning and machine learning in metagenome-based disease prediction. Methods. 2019. 166:74-82.

12. Nguyen TH, Chevaleyre Y, Prifti E, Sokolovska N, Zucker JD. Deep learning for metagenomic data: using 2d embeddings and convolutional neural networks. arXiv:171200244 [Preprint]. 2017 [cited 2021 May 21]. Available from: https://arxiv.org/abs/1712.00244

13. Nguyen TH, Prifti E, Chevaleyre Y, Sokolovska N, Zucker JD. Disease classification in metagenomics with 2d embeddings and deep learning. arXiv:180609046 [Preprint]. 2018 [cited 2021 May 21]. Available from: https://arxiv.org/abs/1806.09046

14. Pasolli E, Truong DT, Malik F, Waldron L, Segata N. Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. PLoS computational biology. 2016. 12(7):e1004977.

15. Asgari E, Garakani K, McHardy AC, Mofrad MR. MicroPheno: predicting environments and host phenotypes from 16S rRNA gene sequencing using a k-mer based representation of shallow sub-samples. Bioinformatics. 2018. 34(13):i32-i42.

16. Oh M, Zhang L. DeepMicro: deep representation learning for disease prediction based on microbiome data. Scientific reports. 2020. 10(1):1-9.

17. Aryal S, Alimadadi A, Manandhar I, Joe B, Cheng X. Machine Learning Strategy for Gut Microbiome-Based Diagnostic Screening of Cardiovascular Disease. Hypertension. 2020. 76(5):1555-1562.

18. Reiman D, Metwally AA, Sun J, Dai Y. PopPhy-CNN: a phylogenetic tree embedded architecture for convolutional neural networks to predict host phenotype from metagenomic data. IEEE journal of biomedical and health informatics. 2020. 24(10):2993-3001.

19. Alemi AA, Fischer I, Dillon JV, Murphy K. Deep variational information bottleneck. arXiv:161200410 [Preprint]. 2016 [cited 2021 May 21]. Available from: https://arxiv.org/abs/1612.00410

20. Tishby N, Pereira FC, Bialek W. The information bottleneck method. arXiv:0004057 [Preprint]. 2000 [cited 2021 May 21]. Available from: https://arxiv.org/abs/physics/0004057

21. Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv:13126034 [Preprint]. 2013 [cited 2021 May 21]. Available from: https://arxiv.org/abs/1312.6034

22. Pan SJ, Yang Q. A survey on transfer learning. IEEE Transactions on knowledge and data engineering. 2009. 22(10):1345-1359.

23. Zhuang F, Qi Z, Duan K, Xi D, Zhu Y, Zhu H, et al. A comprehensive survey on transfer learning. Proceedings of the IEEE. 2020. 109(1):43-76.

24. Kingma DP, Welling M. Auto-encoding variational bayes. arXiv:13126114 [Preprint]. 2013 [cited 2021 May 21]. Available from: https://arxiv.org/abs/1312.6114

25. Karlsson FH, Tremaroli V, Nookaew I, Bergström G, Behre CJ, Fagerberg B, et al. Gut metagenome in European women with normal, impaired and diabetic glucose control. Nature. 2013. 498(7452):99-103.

26. Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature. 2012. 490(7418):55-60.

27. Le Chatelier E, Nielsen T, Qin J, Prifti E, Hildebrand F, Falony G, et al. Richness of human gut microbiome correlates with metabolic markers. Nature. 2013. 500(7464):541-546.

28. Qin N, Yang F, Li A, Prifti E, Chen Y, Shao L, et al. Alterations of the human gut microbiome in liver cirrhosis. Nature. 2014. 513(7516):59-64.

29. Zeller G, Tap J, Voigt AY, Sunagawa S, Kultima JR, Costea PI, et al. Potential of fecal microbiota for early-stage detection of colorectal cancer. Molecular systems biology. 2014. 10(11):766.

30. Li J, Zhao F, Wang Y, et al. Gut microbiota dysbiosis contributes to the development of hypertension. Microbiome. 2017. 5(14).

31. Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nature methods. 2015. 12(10):902-903.

32. Wu M, Goodman N. Multimodal generative models for scalable weakly-supervised learning. arXiv:180205335 [Preprint]. 2018 [cited 2021 May 21]. Available from: https://arxiv.org/abs/1802.05335

33. Cao Y, Fleet DJ. Generalized product of experts for automatic and principled fusion of Gaussian process predictions. arXiv:14107827 [Preprint]. 2014 [cited 2021 May 21]. Available from: https://arxiv.org/abs/1410.7827

34. Hendrycks D, Gimpel K. Gaussian error linear units (gelus). arXiv:160608415 [Preprint]. 2016 [cited 2021 May 21]. Available from: https://arxiv.org/abs/1606.08415

35. Balntas V, Riba E, Ponsa D, Mikolajczyk K. Learning local feature descriptors with triplets and shallow convolutional neural networks. BMVC. 2016.

36. Stothart MR, Palme R, Newman AE. It's what's on the inside that counts: stress physiology and the bacterial microbiome of a wild urban mammal. Proceedings of the Royal Society B. 2019. 286(1913):20192111.

37. Smith RP, Easson C, Lyle SM, Kapoor R, Donnelly CP, Davidson EJ, et al. Gut microbiome diversity is associated with sleep physiology in humans. PLoS One. 2019. 14(10):e0222394.

38. Gopalakrishnan V, Spencer CN, Nezi L, Reuben A, Andrews M, Karpinets T, et al. Gut microbiome modulates response to anti–PD-1 immunotherapy in melanoma patients. Science. 2018. 359(6371):97-103.

39. Gopalakrishnan V, Helmink BA, Spencer CN, Reuben A, Wargo JA. The influence of the gut microbiome on cancer, immunity, and cancer immunotherapy. Cancer cell. 2018. 33(4):570-580.

40. NCBI. SRA-Tools. Last visited 2021 May 21. Available from: http://ncbi.github.io/sra-tools/

41. Biobakery. KneadData. Last visited 2021 May 21. Available from: https://github.com/biobakery/kneaddata

42. Li S, Zhu A,Benes V et al. Durable coexistence of donor and recipient strains after fecal microbiota transplantation. Science. 2016. 352(6285):586-589.

43. Voigt A, Costea P, Kultima J, et al. Temporal and technical variability of human gut metagenomes. Genome Biology. 2015. 16(73).