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 heterogeneous 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. We use the MVIB to predict whether patients are affected by a certain disease by jointly analysing the species-relative abundance and strain-level marker profiles extracted from shotgun metagenomic sequencing of microbial DNA. Additionally, we perform transfer learning experiments by fine-tuning the model on the target disease dataset after a pre-training on all other source diseases, as well as cross-study generalisation experiments. Furthermore, we adopt a saliency technique to interpret the outcome of the model classification and identify which microbial species and strain-level markers mostly contribute to a certain prediction. The combination of metabolomic data to the species-relative abundance and strain-level marker profiles is also investigated. We show that MVIB achieves competitive results on the microbiome-based disease prediction task and is very fast to train.

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 encoding. 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 contribute significantly to the prediction 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 obtain 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. Additionally, we perform experiments combining abundance and marker profiles with

metabolomic data extracted from mass spectrometry. We show that MVIB can operate accurate microbiome-based disease prediction and is also very fast to train.

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 by a factor of 100 [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 have 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 investigating the nature of the microbial components of the human genetic and metabolic landscape. 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 applications 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 encodings 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). [19] proposes the SIAMCAT R package, a toolbox for ML-based comparative metagenomics, whose input consists of a feature matrix (abundances of microbial taxa, genes, or pathways across all samples), plus optional meta-variables (e.g. demographics, lifestyle, clinical records of sample donors or technical parameters of data acquisition).

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) [20]. Leveraging on the theory of the Information Bottleneck (IB) [21], 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. Additionally, we show that MVIB can be used in a trimodal setting, enabling the joint analysis of metabolite, abundance and marker profiles for disease prediction. 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, plus one trimodal dataset which includes metabolomic data, in addition to abundance and marker profiles. Furthermore, we explain the model predictions by computing marker-specific and abundance-specific saliency maps [22]; 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 [23, 24] 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. Cross-study generalisation experiments are performed in order to investigate the generalisation capabilities of MVIB, when trained on samples coming from a given study and tested on samples which belong to a different one.

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*).

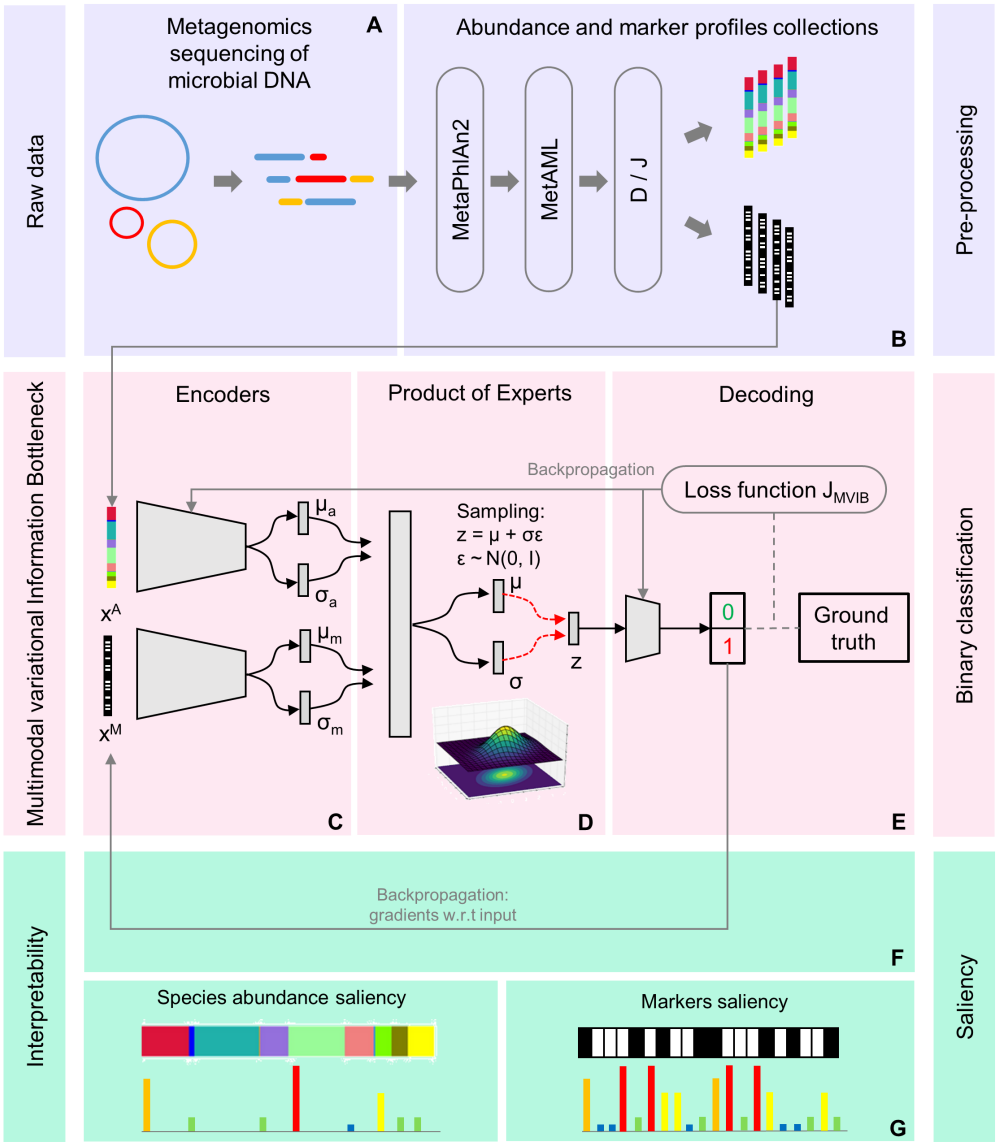

Fig 1. The MVIB applied for multimodal microbiome-based disease prediction with saliency. (A) Raw data: shotgun metagenomics sequencing data of the human gut microbiome. (B) Pre-processing steps. MetaPhlAn2 and MetAML are used for extracting the species-relative abundance profiles and strain-level marker profiles. Two pre-processing steps are proposed and investigated: *default* (D) and *joint* (J). They allow to constitute two datasets collections: D and J. (C) The probabilistic encoders of the MVIB model. (D) The Product of Experts computes a single joint posterior; an encoding $\boldsymbol{z}$ is sampled with the reparameterisation trick [25]. (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.

In brackets are the names we assigned to each dataset, used as identifiers in this work.

For each cohort dataset, affected patients and control subjects are available (Table 1). The IBD cohort presents 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 [26]. The C-T2D cohort presents 170 Chinese subjects affected by type 2 diabetes and 174 healthy controls [27]. The Obesity cohort presents 164 obese subjects and 89 lean individuals [28]. The Cirrhosis cohort presents 118 sick subjects and 114 healthy controls [29]. The Colorectal cohort presents 48 subjects affected by colorectal cancer and 73 healthy controls [30]. 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].

$\text{Obesity}_{\text{joint}}$, $\text{Colorectal}_{\text{EMBL}}$, $\text{Early-Colorectal}_{\text{EMBL}}$ and Hypertension are additional disease cohorts we used to test our disease prediction method. For these datasets, we performed all data pre-processing steps starting from high-throughput whole-genome sequencing reads. The $\text{Obesity}_{\text{joint}}$ cohort consists of the subjects from [28] used in MetAML [14], supplemented with 50 obese and 28 healthy additional subjects from the longitudinal study [31] (see also [32]). The $\text{Colorectal}_{\text{EMBL}}$ dataset consists of 96 affected individuals and 103 healthy controls extracted from [30]. In comparison with the Colorectal dataset from MetAML [14], $\text{Colorectal}_{\text{EMBL}}$ includes 78 extra subjects from [30] (cohorts F and G). $\text{Early-Colorectal}_{\text{EMBL}}$ presents 96 affected samples from dataset $\text{Colorectal}_{\text{EMBL}}$, but relabelled: 52 subjects are labelled as "early stage" (colorectal cancer stages 0, I and II), and 44 are labelled as "late stage" (stages III and IV). The Hypertension cohort presents shotgun metagenomic sequencing of faecal samples from 156 Chinese individuals with pre-hypertension (pHTN) or primary hypertension (HTN), and 40 healthy controls [33]. We considered both pHTN and HTN patients as affected subjects.

As $\text{Colorectal}_{\text{EMBL}}$ is an extension of the Colorectal dataset, we define $\Delta Colorectal = Colorectal_{\text{EMBL}} - Colorectal$ as the additional set of samples. Analogously, as $\text{Obesity}_{\text{joint}}$ is an extension of Obesity, we define $\Delta Obesity = Obesity_{\text{joint}} - Obesity$. The subtractions are intended as the difference of two sets.

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 the quality control, MetaPhlAn2 [34] 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 for each sample. 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.

In addition to the datasets discussed so far, which present two data modalities, we performed experiments on a trimodal dataset, which includes metabolite profiles, in addition to the species-relative abundance and the strain-level marker profiles. We extracted 347 samples with metabolomic and metagenomic data from [35], which proposes a large cohort of participants who underwent colonoscopy to assess taxonomic and functional characteristics of the gut microbiota and metabolites. The metabolomic data was extracted by means of capillary electrophoresis time-of-flight mass spectrometry (CE-TOFMS). In this dataset, 220 samples belong to patients affected by colorectal cancer, polypoid adenomas or intramucosal carcinomas, in addition to more advanced lesions. The remaining 127 samples belong to healthy individuals. We refer to this trimodal dataset as $Colorectal_{Metabolic}$.

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], [26]
C-T2D	344	174	170	[14], [27]
Obesity	253	89	164	[14], [28]
$\text{Obesity}_{\text{joint}}$	331	117	214	[31], [32], [14], [28]
ΔObesity	78	28	50	[31], [32]
Cirrhosis	232	114	118	[14], [29]
Colorectal	121	73	48	[14], [30] (cohort F)
$\text{Colorectal}_{\text{EMBL}}$	199	103	96	[30] (cohorts F and G)
ΔColorectal	78	30	48	[30] (cohorts F and G)
$\text{Early-Colorectal}_{\text{EMBL}}$	96	52	44	[30] (cohorts F and G)
Hypertension	196	40	156	[33]
$\text{Colorectal}_{\text{Metabolic}}$	347	127	220	[35]

Pre-processing

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

SRAs that match the runs of paired-end metagenome sequencing experiments were downloaded from the corresponding NCBI bioproject repositories. They were then converted into fastq files with fastq-dump version 2.8.0, part of the NCBI SRA Toolkit [36]. The created run-level fastq files were aggregated by sample identifier. Afterwards, these aggregated fastq files passed the quality control of Kneaddata version 0.7.4 [37] with default parameters. Reads trimmed to less than 60 bp were removed. All processed files were then profiled with MetaPhlAn2 [34] to get species-relative abundances and strain-level marker profiles. The MetAML dataset selection script [14] was used to leave only species-level features in the abundance profiles.

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

The *default* collection contains species relative abundances and gene markers as they were extracted by the MetPhlAn2 + MetAML pipeline. Additionally, feature standardisation was applied to the abundance profiles.

The *joint* collection was created for transfer learning (see Section Transfer learning) and cross-study generalisation (see Section Cross-study generalisation) 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 strain markers across all datasets.

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 [21], 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 for maximising Eq 1 would be the identity $Z = X$. This would ensure a maximally informative representation, but (b) places a constraint on Z. In fact, due to (b), we want to "forget" as much information as possible about X. 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 Eq 2 causes Z to be predictive of Y, while the second term constraints Z to be a minimal sufficient statistics of X. β controls the trade-off.

As derived in [20] for the Deep Variational Information Bottleneck (Deep 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, Eq 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 [25], 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 Eq 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 Deep VIB objective of Eq 3 by considering that X is a collection of multimodal random input variables s.t. $X = \{X^i | i^{th}\ modality\ present\}$. In light of this, the posterior $p(Z|\boldsymbol{x})$ of Eq 3 consists actually in the joint posterior $p(Z|\boldsymbol{x}^1,...,\boldsymbol{x}^M)$, conditioned by the joint M available data modalities. Following the approach proposed for the Multimodal Variational Autoencoder [38], 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 Eq 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 [39]. We can now formulate the objective of the Multimodal Variational Information Bottleneck:

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

Implementation details

In this work, there are two main data modalities: the species-relative abundance profiles and the strain-level marker 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 $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$. 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 $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$. SiLU [40] activation functions are used. 0.4 drop-out is used at training time.

As described in Section Trimodal MVIB: combining metabolomics and metagenomics, experiments on the $\text{Colorectal}_{\text{Metabolic}}$ dataset demand a third additional stochastic encoder for the metabolite profiles. For this data modality, the same encoder architecture adopted for the abundance profiles is used.

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

The networks are 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 is set to 256. The training is performed for 200 epochs and, in order to avoid overfitting, the best model is selected by saving the weights corresponding to the epoch where the area under the curve (AUC) of the receiver operating characteristic (ROC) is 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 are implemented in Python 3.6 using PyTorch [41] version 1.7; code is publicly available at `https://github.com/nec-research/microbiome-mvib`.

Extending the MVIB objective with the triplet margin loss

The MVIB objective function proposed in Eq 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 Eq 5 with an additional triplet margin loss [42] term can lead to a more accurate predictor on various disease datasets.

Given a certain batch size B of samples on which the loss is meant to be computed, let B_+ be the number of sick samples and B_- the number of healthy control samples. Let $B_T = min\{B_+, B_-\}$. From a batch of size B, we randomly sample B_T sick samples

and B_T control samples without repetitions (this implies that the smaller set of samples, either sick or control, will be fully considered). These $2B_T$ samples constitute the *anchors* of the triplet margin loss. We represent the anchors with two sets: $\mathcal{A}_+$ (which contains B_T sick anchors) and $\mathcal{A}_-$ (which contains B_T control anchors). For each anchor sample in $\mathcal{A}_+$ and $\mathcal{A}_-$, we sample: (a) a subject of the same class (addressed in this context as *positive*) and (b) a sample of the opposite class (addressed as *negative*). In our implementation, this is obtained by shuffling the samples in the opposite-class anchor set. This allows to constitute a set of positive samples $\mathcal{P}_+$ (i.e. of the same class) and a set of negative samples $\mathcal{N}_+$ (i.e. of the opposite class) for the sick anchors $\mathcal{A}_+$. Analogously, a set of positive samples $\mathcal{P}_-$ and a set of negative samples $\mathcal{N}_-$ are obtained for the control anchors $\mathcal{A}_-$.

Patients contained in the $\mathcal{P}_-$ set are healthy control subjects, while subjects contained in $\mathcal{N}_-$ are sick patients. Analogously, samples contained in the $\mathcal{P}_+$ set belong to sick subjects, while samples contained in $\mathcal{N}_+$ are heathy control patients. This is because, in the context of the triplet margin loss, *positive* and *negative* actually mean of the *same* and of *opposite* ground truth label, respectively.

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 $\widetilde{q}(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(f_e^\mu(\boldsymbol{x}), f_e^\sigma(\boldsymbol{x}))$ (see Section Implementation details). 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 the right hand side of Eq 7 encourages the encodings of sick samples to be closer to each other in their K-dimensional Euclidean space and far away from the encodings of healthy control patients. In the same fashion, the second term of the equation encourages the encodings of healthy patients to be clustered in the same region of the latent space and to be distant from the encodings of sick patients.

Having defined the triplet loss objective of Eq 7 leads to an extension of the MVIB objective presented in Eq 5 which define as MVIB-T objective:

$$J_{MVIB-T} = J_{MVIB} + \lambda_T J_T \,. \quad (8)$$

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

Full multimodal objective

For the training of MVIB, we adopt the same training paradigm proposed for the Multimodal Variational Autoencoder [38]. The MVIB objective presented in Eq 5 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 certain data modalities are not available 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.

Eq 5 can be reformulated as a full multimodal objective, which allows the model to be optimal in all multimodal and single-modality settings:

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

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 Eq 8. For all experiments performed in this work, the objective functions are computed in their full multimodal form, as described in this section.

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 our test samples are the same ones considered by DeepMicro and allows a fair benchmark of the models' performances. The same procedure was adopted for $\text{Obesity}_{\text{joint}}$, $\text{Colorectal}_{\text{EMBL}}$, Early-$\text{Colorectal}_{\text{EMBL}}$, Hypertension and $\text{Colorectal}_{\text{Metabolic}}$ 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.

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 [23, 24]. Iteratively, we first select a target disease from the set of 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. The target task shall therefore be interpreted as the prediction of the disease which we mostly care about.

Merging the 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 $(\mathcal{D}_T, \mathcal{T}_T)$.

For the transfer learning experiments, we adopted the *joint* datasets collection (see Section Pre-processing), as the microbial species and strain markers share the same positional indexes across the various datasets and present the same dimension.

As $\text{Colorectal}_{\text{EMBL}}$ is an extension of the Colorectal dataset, and $\text{Obesity}_{\text{joint}}$ is an extension of Obesity (see Section Datasets), applying the procedure described above

would allow the model to observe the samples shared by more than one dataset during both the source and the target task. This would lead to biased results. Therefore, in the transfer learning experiments, ΔObesity and ΔColorectal (see Section Datasets) are considered during the source task instead of $\text{Obesity}_{\text{joint}}$ and $\text{Colorectal}_{\text{EMBL}}$, respectively.

Cross-study generalisation

In order to further evaluate the generalisation capabilities of MVIB across different studies, and motivated by the fact of having various datasets for the same disease, we performed cross-study generalisation experiments. First, we identified the following ordered pairs of datasets: (EW-T2D, C-T2D), (C-T2D, EW-T2D), (ΔObesity, Obesity), (ΔColorectal, Colorectal), (Obesity, ΔObesity), (Colorectal, ΔColorectal) - see Section Datasets. Then, for each pair, we trained MVIB on the first source dataset, and tested it on the second target one. No fine-tuning was performed on the test dataset.

Explaining predictions with saliency

Following the same paradigm proposed by [22], 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 $\boldsymbol{x}^M$:

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

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

The magnitude of the derivatives presented in Eq 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 samples correctly classified as positive (i.e. sick) and compute their average. This allows to discover a ranking of the most influential gene markers and microbial species for each disease dataset (see Fig 1F,G). The computation of saliency maps is extremely quick and it only requires a single backpropagation pass.

Trimodal MVIB: combining metabolomics and metagenomics

In order to further investigate the multimodal learning capabilities of MVIB, we performed trimodal experiments on the $\text{Colorectal}_{\text{Metabolic}}$ dataset [35]. This dataset includes three data modalities for all samples: species-relative abundance, strain-level marker and metabolite profiles (see Section Datasets). For training the model in the trimodal setting, the same full multimodal training paradigm presented in Section Full multimodal objective was adopted. This allows to compute the MVIB objective for all possible data modalities combinations, as well as for the individual modalities.

In this trimodal setting, the objective function is computed in the following fashion:

$$\sum_{m \in \mathcal{M}} J_{MVIB}(\boldsymbol{x}^m) + \sum_{\{m,n\} \in \mathcal{MM}} J_{MVIB}(\boldsymbol{x}^m, \boldsymbol{x}^n) + J_{MVIB}(\{\boldsymbol{x}^m | m \in \mathcal{M}\}), \tag{11}$$

where $\mathcal{M} = \{A, M, Metabolic\}$ is the set of the three considered data modalities and $\mathcal{MM} = \{\{m,n\} | m,n \in \mathcal{M}\}$ is the set of all possible (unordered) modality pairs. S1 Fig

depicts the trimodal architecture of MVIB used for the experiments on the $\text{Colorectal}_{\text{Metabolic}}$ dataset.

For the trimodal experiments, the learning rate has been set to 10^{-5} and the bottleneck dimension K to 128, as we observed that this slightly improves learning.

Results

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

Table 2. MVIB performance overview.

Dataset	Random Forest			$\text{DeepMicro}_{\text{VAE}}$‡ [16]			PopPhy-CNN [18]	MVIB (A+M)	
	A	M	A+M	A	M	A+M	A	D	J
IBD	0.899 (0.037)	0.932 (0.025)	0.936 (0.017)	0.779 (0.032)	0.899 (0.039)	0.833 (0.034)	0.799 (0.052)	0.922 (0.020)	**0.936 (0.014)**
EW-T2D	0.825 (0.020)	0.812 (0.017)	0.789 (0.024)	0.640 (0.051)	0.853 (0.041)	0.741 (0.080)	0.591 (0.037)	**0.859 (0.023)**	0.853 (0.025)
C-T2D	0.717 (0.018)	0.736 (0.023)	0.734 (0.017)	0.715 (0.031)	0.719 (0.019)	0.714 (0.030)	0.641 (0.019)	0.750 (0.009)	**0.758 (0.012)**
Obesity	0.662 (0.022)	0.634 (0.019)	0.619 (0.017)	0.600 (0.030)	0.599 (0.014)	0.576 (0.040)	0.630 0.018)	0.662 (0.024)	**0.666 (0.027)**
Cirrhosis	0.897 (0.013)	0.896 (0.007)	0.905 (0.009)	0.781 (0.021)	0.891 (0.016)	0.867 (0.018)	0.893 (0.008)	**0.925 (0.005)**	0.924 (0.005)
Colorectal	0.803 (0.048)	0.771 (0.053)	0.764 (0.053)	0.739 (0.070)	0.737 (0.068)	0.548 (0.040)	**0.803 (0.023)**	0.780 (0.071)	0.777 (0.069)
$\text{Obesity}_{\text{joint}}$	0.806 (0.026)	0.795 (0.024)	0.799 (0.025)	0.662 (0.015)	0.637 (0.021)	0.600 (0.017)	0.697 (0.026)	0.815 (0.019)	**0.818 (0.018)**
$\text{Colorectal}_{\text{EMBL}}$	**0.890 (0.015)**	0.866 (0.024)	0.860 (0.016)	0.662 (0.027)	0.725 (0.049)	0.693 (0.034)	†	0.811 (0.010)	0.814 (0.013)
$\text{Early-Colorectal}_{\text{EMBL}}$	0.577 (0.046)	0.533 (0.039)	**0.582 (0.024)**	0.562 (0.070)	0.557 (0.048)	0.5616 (0.048)	†	0.535 (0.050)	0.543 (0.048)
Hypertension	0.662 (0.035)	**0.677 (0.023)**	0.631 (0.031)	0.523 (0.030)	0.561 (0.028)	0.658 (0.054)	†	0.603 (0.045)	0.591 (0.041)

All values are ROC AUC computed on the test sets. Values in brackets refer to the standard error over five repeated experiments (see Section Validation framework and performance evaluation). The first group of columns presents the results obtained with the Random Forest. The second group of columns presents the best results obtained with DeepMicro with VAE. The third column presents results obtained with PopPhy-CNN. The last 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) (see Section Pre-processing). 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 Eq 8).
†: for some datasets, we could not reproduce PopPhy-CNN results, as the original implementation appears to generate an endless loop at the step in which the phylogenetic tree is pruned.
‡: We selected the best test ROC AUC from the DeepMicro paper over multiple downstream classifier models (SVM, MLP, Random Forest).

To evaluate the performance of the MVIB, we performed a comparison with Random

Forest, DeepMicro [16] and PopPhy-CNN [18], which are state-of-the-art approaches for microbiome-based disease prediction at the time this paper is being written. In Table 2 we report the AUC of the ROC for benchmark methods, as well as for MVIB. MVIB results are reported for the dataset collections *default* and *joint* (see Section Pre-processing) and derive from the joint learning of abundance and marker profiles (A+M). A complete overview of the MVIB results for all bimodal dataset collections is available in S2 Table.

For Random Forest and DeepMicro, both abundance-only (A) and markers-only (M) results are reported. Furthermore, we also computed the performance on the multimodal disease prediction task by training these two models on the concatenation of abundance and marker profiles (A+M). We computed Random Forest results adopting Scikit-learn [43] version 0.23.2. The Random Forest results reported in Table 2 derive from a fine-tuning through cross-validated grid-search over the hyperparameter space reported in S3 File. 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 encodings 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 [25]. We believe this is the most significant comparison, since MVIB is also based on variational inference. This allows to highlight the improvements deriving from our end-to-end multimodal architecture with respect to DeepMicro's 2-steps approach.

For PopPhy-CNN, we edited the source code of the validation and testing steps in order to make them equivalent to the approach described in Section Validation framework and performance evaluation. The original PopPhy-CNN implementation, in fact, presented a slightly different validation procedure, and we wanted to ensure that all results presented in Table 2 derive from the same validation approach.

Random Forest performs overall very well, but outperforms MVIB only on the $\text{Colorectal}_{\text{EMBL}}$, $\text{Early-Colorectal}_{\text{EMBL}}$ and Hypertension datasets.

We observed that MVIB consistently outperforms DeepMicro on all datasets, both in the single-modality (A or B) and in the multimodal settings (A+M).

In comparison with PopPhy-CNN, MVIB achieves better results on all datasets, except for Colorectal, where PopPhy-CNN achieves 2% higher ROC AUC. For the $\text{Colorectal}_{\text{EMBL}}$, $\text{Early-Colorectal}_{\text{EMBL}}$ and Hypertension datasets, we could not reproduce PopPhy-CNN results, as the original implementation appears to generate an endless loop at the step in which the phylogenetic tree is pruned. In Table 2 we indicated this by means of the † symbol.

$\text{Early-Colorectal}_{\text{EMBL}}$ and Hypertension appear to be the hardest diseases to predict by means of microbiome-based diagnosis and the MVIB only achieves 0.535 and 0.603 ROC AUC, respectively, on the *default* datasets collection.

Multimodal ablation study

In order to further evaluate the multimodal learning capabilities of MVIB, we performed an ablation study aimed at comparing classification performances obtained in various different multimodal and single-modality settings. Fig 2 depicts the results of the ablation study.

For the ablation study, MVIB has been trained optimising the full multimodal objective (see Section Full multimodal objective), which ensures that all encoders are trained individually, as well as jointly. At inference time, we tested the model by either

considering only abundance or only marker profiles as inputs, for the single-modality settings, or by passing both the abundance and marker profiles simultaneously.

As presented in Fig 2, the markers-only results (MVIB-M) are consistently better w.r.t. the abundance-only ones (MVIB-A). Furthermore, in the bimodal setting (MVIB-A+M), the model performance does not degrade but remains comparable to the markers-only setting. On the IBD, Cirrhosis, Colorectal and Hypertension datasets, the bimodal MVIB (MVIB-A+M) achieves slightly better results w.r.t both single-modality settings.

Based on the results obtained in this ablation study, it is possible to conclude that MVIB can efficiently combine heterogeneous data modalities. In particular, even though abundance-only test results present consistently lower ROC AUC w.r.t the markers-only ones, the combination of the two modalities does not lead to a performance trade-off. In fact, in the bimodal setting, MVIB achieves performances which are either as good as the markers-only ones, or slightly better. It is possible to conclude that MVIB is a suitable multimodal learning approach for situations in which one modality is less informative (or harder to learn) than another. MVIB allows in fact to consider both modalities simultaneously without incurring in performance drops due to the increased feature space and ensures classification performances which are as good as those deriving from the best data modality - or even slightly better.

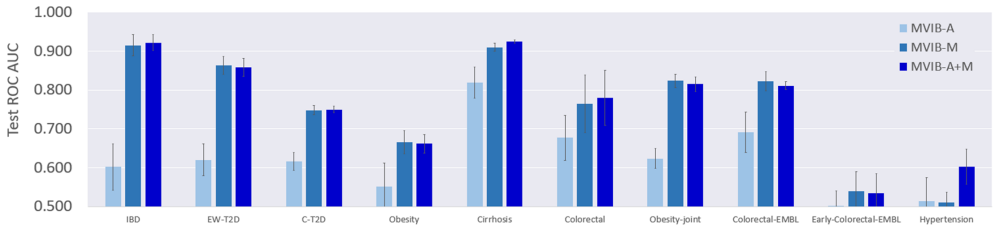

Fig 2. Ablation study - comparison of single-modality and bimodal MVIB performances. Depicted values are test ROC AUC, error bars are standard error obtained by repeating the experiment five times on different random train/test splits (see Section Validation framework and performance evaluation). The J_{MVIB} objective (see Eq 5) has been adopted for optimisation. MVIB-A: results obtained by passing only species-relative abundance profiles at test time. MVIB-M: results obtained by passing only strain-level marker profiles at test time. MVIB-A+M: results obtained by passing both species-relative abundance and strain-level marker profiles at test time.

The triplet margin loss can improve classification

The first term of the original MVIB objective J_{MVIB} (see Eq 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, namely J_{MVIB-T}, by adding a triplet margin loss term (see Eq 8).

The triplet margin loss aims to encourage the latent distribution of samples which belong to the same class to cluster in a dedicated region of the latent Euclidean space. 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 3 presents a comparison of the effects of the triplet margin loss on the MVIB classification performance. For the IBD, EW-T2D, C-T2D and Early-Colorectal$_{\text{EMBL}}$ datasets, the highest ROC AUC is achieved by optimising the J_{MVIB-T} objective, which includes the triplet margin loss. The J_{MVIB} objective leads to best results on the

remaining datasets.

Table 3. Comparison of different objective functions and pre-processing techniques.

Dataset	J_{MVIB-T}		J_{MVIB}	
	D	J	D	J
IBD	0.922 (0.020)	**0.936 (0.014)**	0.915 (0.018)	0.915 (0.018)
EW-T2D	**0.859 (0.023)**	0.853 (0.025)	0.859 (0.024)	0.855 (0.027)
C-T2D	0.750 (0.009)	**0.758 (0.012)**	0.754 (0.014)	0.756 (0.015)
Obesity	0.662 (0.024)	0.666 (0.027)	**0.673 (0.028)**	0.672 (0.030)
Cirrhosis	0.925 (0.005)	0.924 (0.005)	**0.930 (0.002)**	0.928 (0.002)
Colorectal	0.780 (0.071)	0.777 (0.069)	0.788 (0.059)	**0.796 (0.055)**
$\text{Obesity}_{\text{joint}}$	0.815 (0.019)	0.818 (0.018)	0.825 (0.019)	**0.825 (0.018)**
$\text{Colorectal}_{\text{EMBL}}$	0.811 (0.010)	0.814 (0.013)	0.827 (0.012)	**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	0.602 (0.045)	0.608 (0.043)	0.622 (0.048)	**0.624 (0.049)**

All values are ROC AUC computed on the test sets. Values in brackets refer to the standard error over five repeated experiments (see Section Validation framework and performance evaluation). The first group of columns presents the results obtained optimising the J_{MVIB-T} objective (see Eq 8), which includes the triplet margin loss. The second group of columns presents the results obtained optimising the original objective function J_{MVIB} (see Eq 5). D and J refer to the two pre-processing techniques adopted and the two collections of datasets obtained: *default* (D) and *joint* (J) (see Section Pre-processing).

Fig 3A depicts the 95% confidence interval of the patients' stochastic encodings $\boldsymbol{z} \sim p(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2\boldsymbol{I})$ deriving from the optimisation of the J_{MVIB} (see Eq 5). Such Gaussian distributions are the output of the PoE (see Eq 4) and consist in fact in the joint posterior distribution of the latent encoding, conditioned on all input data modalities. In comparison with the curves of Fig 3C, obtained with the J_{MVIB-T} objective (see Eq 8), we observe that the triplet margin loss leads to Gaussian distributions which present a higher variance. Conversely, stochastic encodings deriving from J_{MVIB} (see Eq 5) present smaller variance (see Fig 3A).

Methods such as simple autoencoders, or the PCA, which are commonly used for dimensionality reduction, provide deterministic encodings, i.e. compressed representations of the input without a rigorous estimate of the uncertainty. Our method allows to compute *stochastic* encodings, i.e. to represent the input not only as a point in a latent space, but as a probability distribution. Computing probability distributions allows to estimate the confidence of the stochastic encodings.

Intuitively, it is preferable to obtain stochastic encodings which present a small variance, as this allows for a better separability and classification. Fig 3 allows to visually interpret the effect of including the triplet margin loss (see Eq 8) in the

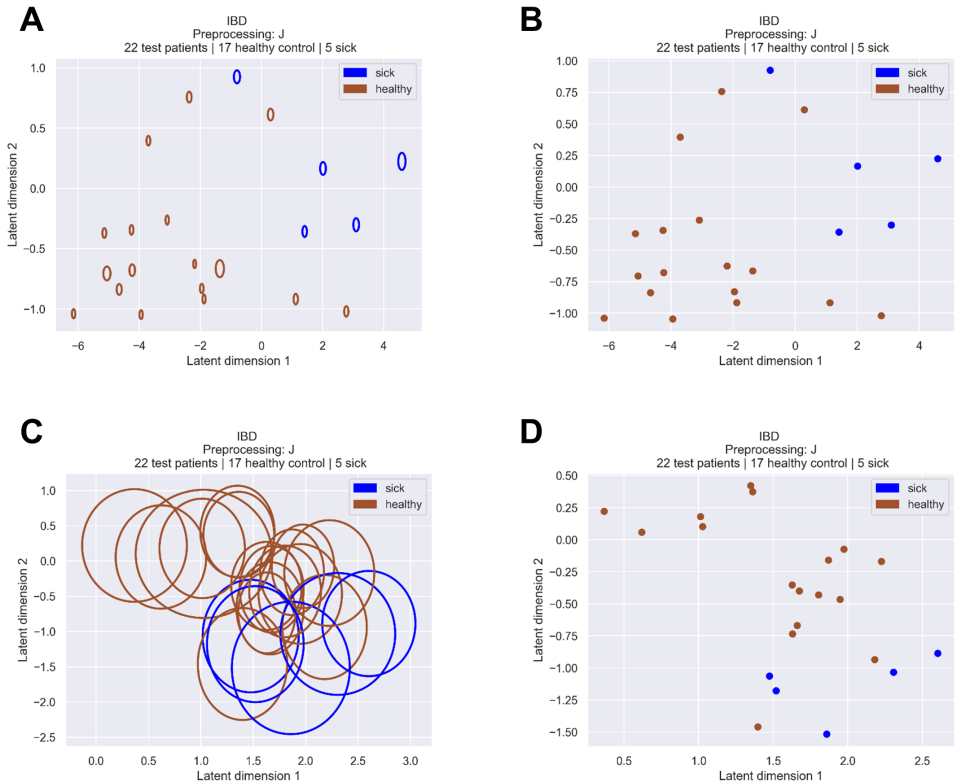

Fig 3. Effect of the triplet margin loss on the stochastic encodings of the patients. The depicted curves are the 95% confidence intervals of the patients' stochastic encodings $\boldsymbol{z} \sim p(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2\boldsymbol{I})$; the points are their means $\boldsymbol{\mu}$. The displayed encodings consist in a set of 22 test samples obtained from a random training-test split of the IBD dataset (i.e. the 20% of the dataset not used for training). 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} objective (see Eq 5). (C, D) optimisation of the J_{MVIB-T} objective (see Eq 8).

objective function and to visualise the results of the different losses on the patients' stochastic encodings. We conclude that, although adding the triplet margin loss can lead to better classification results on some datasets (see Table 3), the stochastic encodings derived from the optimisation of the J_{MVIB-T} objective (see Eq 8) present higher variance w.r.t. those obtained from the optimisation of the J_{MVIB} objective (see Eq 5).

In addition to the plots of Fig 3, which only refer to the IBD dataset, S4 File contains the plots of the MVIB stochastic encodings for all the datasets considered in this work. The stochastic encodings depicted in S4 File are obtained from the *joint* datasets collection. They are available for models trained by optimising the J_{MVIB-T} objective, as well as models trained by optimising the J_{MVIB} objective, in order to allow comparison.

S5 File presents for each dataset a comparison between the PCA 2D projections and the mean of the MVIB 2D stochastic encodings. For both the PCA and the MVIB plots presented in S5 File, the *default* datasets collection (see Section Pre-processing) has been adopted. It is possible to observe that the PCA projections completely fail at clustering the two samples classes (healthy and sick) in dedicated areas of the 2D plane. Conversely, MVIB encodes sick and healthy samples in two clearly separated areas of

the 2D plane, creating two distinct clusters.

Transfer learning across different diseases can improve disease prediction

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 abundance and strain-level marker profiles (see Section Pre-processing). Table 4 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). The ROC AUC slightly decreases on the Cirrhosis and Colorectal target datasets. A consistent performance reduction is observed on the IBD and EW-T2D target datasets. Improvements w.r.t randomly initialised models can be observed on all other datasets: C-T2D, Obesity and Hypertension.

Motivated by the fact that $\text{Colorectal}_{\text{EMBL}}$ is an extension of the Colorectal dataset, and $\text{Obesity}_{\text{joint}}$ is an extension of Obesity (see Section Datasets), the $\text{Colorectal}_{\text{EMBL}}$ and $\text{Obesity}_{\text{joint}}$ datasets were not considered during the transfer learning experiments. This ensures that the samples which are present in more than one dataset are not observed during both the source and the target tasks. ΔObesity and ΔColorectal are adopted during the source task only (pre-training).

Table 4. Comparison of pre-trained models against randomly initialised models.

Dataset	Random initialisation	Pre-trained model
	J	J
IBD	**0.936 (0.014)**	0.882 (0.021)
EW-T2D	**0.853 (0.025)**	0.780 (0.024)
C-T2D	0.758 (0.012)	**0.774 (0.013)**
Obesity	0.666 (0.027)	**0.679 (0.024)**
Cirrhosis	**0.924 (0.005)**	0.918 (0.007)
Colorectal	**0.777 (0.069)**	0.763 (0.068)
Hypertension	0.591 (0.041)	**0.645 (0.047)**

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 then fine-tuning them on the target disease (see Section Transfer learning). The J_{MVIB-T} objective was adopted for both columns (see Eq 8). J refers to the adopted pre-processing technique: *joint* (see Section Pre-processing). Reported values are ROC AUC computed on the test sets. In brackets, the standard error over five repeated experiments is reported see Section Validation framework and performance evaluation).

Cross-study generalisation results and benchmark

Generalisation is a fundamental requirement for machine learning models. It consists in the capacity to make correct predictions on samples which were not observed at training time. In order to further investigate how well MVIB can generalise, we performed cross-study experiments as described in Section Cross-study generalisation. First, six

ordered pairs of datasets were identified: (EW-T2D, C-T2D), (C-T2D, EW-T2D), (ΔObesity, Obesity), (ΔColorectal, Colorectal), (Obesity, ΔObesity), (Colorectal, ΔColorectal). The first dataset of each pair was exclusively used for training, while the second one for testing without fine-tuning. In order to guarantee that abundance and marker features share the same positional indexes across the source and target datasets, the *joint* dataset collection was employed for all cross-study experiments (see Section Pre-processing).

In order to compare the generalisation capabilities of MVIB with a state-of-the-art machine learning model used in microbiome research, we performed benchmark experiments with Random Forest. The Random Forest was trained on the concatenation of the abundance and marker profiles (multimodal setting). For the Random Forest, a cross-validated grid-search over a defined hyperparameter space was adopted for fine tuning (see S3 File). MVIB did not undergo hyperparameter tuning, but employed the default implementation described in Section Implementation details, with the exception of a lower learning rate, which was set to 10^{-5}, as we observed this improves convergence.

Fig 4 shows the cross-study results. MVIB performed better than the Random Forest on the (EW-T2D, C-T2D) and (ΔColorectal, Colorectal) experiments. Conversely, the Random Forest achieved better generalisation on (C-T2D, EW-T2D), (ΔObesity, Obesity) and (Obesity, ΔObesity). The two methods obtained the same results on (Colorectal, ΔColorectal).

Based on the empirical results which we obtained, it is possible to conclude that MVIB offers competitive cross-study generalisation when benchmarked with a fine-tuned Random Forest.

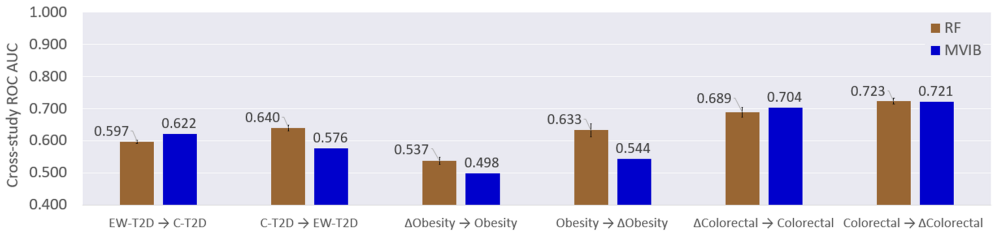

Fig 4. MVIB cross-study generalisation and benchmark. Values are test ROC AUC computed by first training the models on a source dataset and then testing it on a target dataset (see Section Cross-study generalisation). RF: Random Forest. All datasets used for cross-study experiments belong to the J collection (see Section Pre-processing). For MVIB, the J_{MVIB-T} objective has been adopted for the optimisation (see Eq 8). The datasets reported on the x-axis shall be interpreted as: *train* → *test*. For the Random Forest, the error bars represent the standard error over five repeated experiments and account for the stochasticity of the Scikit-learn implementation. The standard error is missing for the MVIB results, as our PyTorch implementation has been made deterministic.

Discovering the most salient microbial species and strain-level markers

In order to achieve interpretability, we implemented a method which allows to compute saliency, i.e. detecting the areas of the input vectors which are most discriminative with respect to the predicted class (see Section Explaining predictions with saliency). For each disease dataset in the *default* dataset collection (see Section Pre-processing), we obtained saliency maps of both abundance and marker profiles. Such saliency maps

were obtained by computing the derivative of the model's positive predictions with respect to the inputs (see Eq 10). Since we were mostly interested in the magnitude of such gradients, rather than the sign, the absolute value was considered.

The aim of this analysis is the discovery of the most salient microbial species and strain-level markers for each considered dataset, i.e. the features which mostly affect the outcome of positive disease predictions. Aiming at discovering true biological insights about the relationships between the microbiome and the analysed diseases, we considered saliency maps derived from true positive predictions. After having trained MVIB as described in Section Materials and methods, each dataset was passed through the model to compute predictions for all samples. Saliency maps were then computed (see Eq 10). Saliency vectors coming from true positive predictions where then extracted from the dataset batch. As described in Section Validation framework and performance evaluation, experiments were repeated five times with different independent train-test splits. Furthermore, for each train-test split, five different best models derived from a 5-fold cross-validation were ensembled. This leads to 25 different models, and therefore 25 sets of saliency maps, which were then averaged.

Fig 5 depicts the top 25 microbial species and strain markers sorted by saliency for the $\text{Colorectal}_{\text{EMBL}}$ dataset (see Section Datasets).

For the microbial species (Fig 5A), the prevalence in healthy and affected individuals was calculated by means of a Wilcoxon test. For each microbial species, two samples were constructed from the abundance values: (1) healthy individuals, (2) affected individuals. FDR correction was applied to adjust for multiple hypotheses testing. Species prevalence supported by adjusted p-values < 0.1 was reported as significant. In Fig 5A, species which significantly prevail in healthy individuals are marked in blue, while prevalent specie in positive individual are marked in red.

S6 File contains the plots with the top 25 microbial species and strain markers sorted by mean saliency for all datasets considered in this work, analogous to what Fig 5 depicts for the $\text{Colorectal}_{\text{EMBL}}$ dataset.

Metabolomics can improve colorectal cancer prediction when combined with metagenomics

With the aim of further investigating the multimodal learning capabilities of MVIB, we performed experiments on the $\text{Colorectal}_{\text{Metabolic}}$ dataset extracted from [35] (see Section Datasets). For each sample, this dataset presents three modalities: species-relative abundance, strain-level marker and metabolite profiles.

First, we trained MVIB following the full multimodal training paradigm presented in Section Trimodal MVIB: combining metabolomics and metagenomics. This ensures that all encoders are individually and jointly trained and that MVIB can perform single-modality and multimodal predictions at test time. We then tested the model in various multimodal and single-modality fashions: abundance profiles only (A), marker profiles only (M), abundance + marker profiles (A+M), abundance + marker + metabolite profiles (A+M+Metabolic).

We compared MVIB with a fine-tuned Random Forest. For fine-tuning the Random Forest, a cross-validated grid-search over a defined hyperparameter space was used (see S3 File). In the A+M and A+M+Metabolic multimodal settings, the various input modalities have been concatenated and then fed into the Random Forest.

Fig 6 shows the obtained experimental results. It is possible to observe that the multimodal settings A+M and A+M+Metabolic allow MVIB to reach higher test ROC AUC w.r.t the Random Forest. Furthermore, adding metabolomic data (A+M+Metabolic) allows to achieve the best classification results. Results obtained in the single-modality M setting are comparable to those obtained in the trimodal setting.

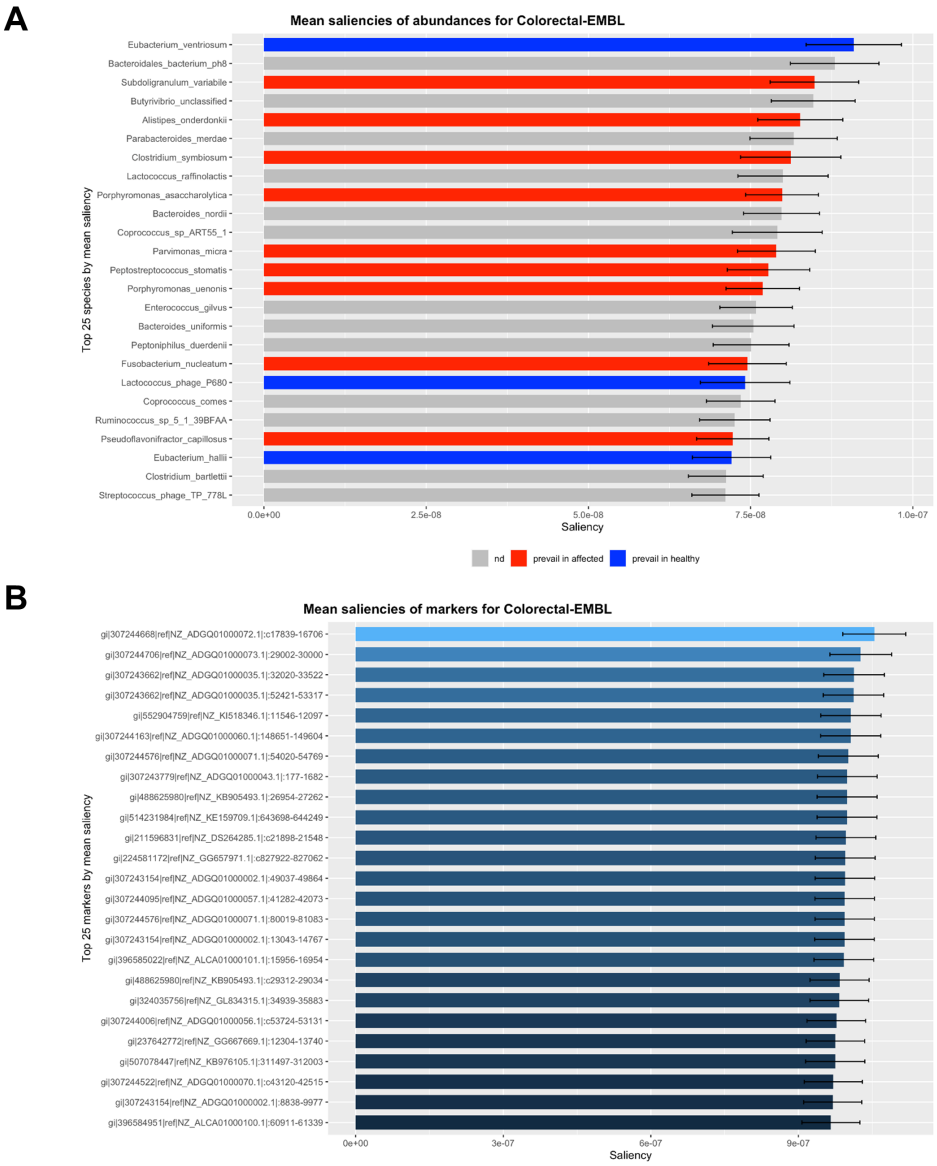

Fig 5. Top microbial species and strain markers sorted by saliency for the Colorectal$_{\text{EMBL}}$ dataset. Saliency maps computed for the Colorectal$_{\text{EMBL}}$ dataset (see Section Datasets) using Eq 10. Error bars represent the standard error over the five repeated experiments and the five ensembled MVIB models for each experiment (see Section Validation framework and performance evaluation). (A) Top 25 microbial species sorted by mean saliency. Species prevalence in healthy (blue) and affected (red) individuals were calculated using a Wilcoxon test for each microbial species for two unpaired samples: healthy and affected individuals. (B) Top 25 microbial strain markers sorted by mean saliency. The colour gradient matches the sorting for a better visualisation.

The Random Forest performs better than MVIB only in the abundance-only single-modality setting (A). Most notably, the Random Forest performance tends to degrade when further data modalities are added. This shows the superiority of MVIB in combining multiple heterogeneous data modalities.

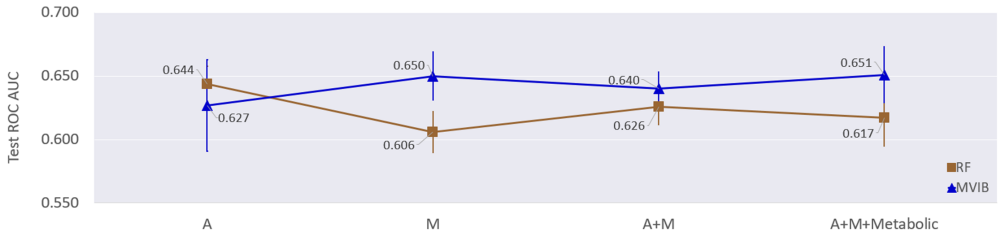

Fig 6. Comparison of different multimodal and single-modality models for colorectal cancer prediction. Values are test ROC AUC, error bars are standard error obtained by repeating the experiment five times on different random train/test splits (see Section Validation framework and performance evaluation). RF: Random Forest. A: species-relative abundance profiles. M: strain-level marker profiles. Metabolic: metabolite profiles. For MVIB, the J_{MVIB} objective has been adopted for the optimisation (see Eq 5). Experiments are executed on the $\text{Colorectal}_{\text{Metabolic}}$ dataset extracted from [35] (see Section Datasets).

Empirical analysis of the training time

In this section, we report the analysis of the training time of various machine learning models trained on the trimodal $\text{Colorectal}_{\text{Metabolic}}$ dataset extracted from [35] (see Section Datasets). We opted to operate this analysis with this dataset because it presents three data modalities, as well as the highest number of samples out of the datasets considered in this work.

For benchmark, we compared the training time of MVIB with the training time of a Random Forest (RF) and a Support Vector Machine (SVM). For both the RF and the SVM, we considered the default implementation of Scikit-learn [43] version 0.23.2, which runs in CPU. This means that the training time does not include hyperparameter optimisation.

For MVIB, we set the learning rate to 10^{-4} and the latent dimension K to 256. We trained MVIB with a batch size of 256 for a maximum of 50 epochs, as convergence (i.e. best validation ROC AUC) is normally observed within the first 20 epochs. We considered the training complete at the epoch in which the model achieves the highest ROC AUC on the validation set. No ensembling was performed, i.e. a single model was trained in this set of experiments. MVIB experiments were executed on a NVIDIA GeForce RTX 2080 Ti GPU.

We performed experiments in various multimodal and single-modality fashions: abundance profiles only (A), marker profiles only (M), abundance + marker profiles (A+M), abundance + marker + metabolite profiles (A+M+Metabolic). For RF and SVM, in the multimodal settings, the input feature vectors of the various modalities were concatenated. Fig 7 depicts the empirical results of the training time analysis.

Although the SVM is the fastest model to complete training in the abundance-only setting (A), its training time dramatically increases in the markers-only setting (M), as well as in the the two multimodal settings (A+M and A+M+Metabolic), reaching an average training time of more than 7 seconds. The feature space dimensions in the various settings are reported in Fig 7. The RF exhibits the same trend: it achieves the fastest training in the A setting, but requires a slightly longer time in all other settings. Compared with the SVM, the RF is always much faster and never requires more than 1 second to complete training. For MVIB we observe a different behaviour. The training time does not appear to correlate with the feature space dimension. In fact the shortest training time is achieved in the M setting, which presents a much higher feature space dimension w.r.t. A. In all other settings, MVIB converges in around 1 second.

In the A setting, despite the smaller feature space dimension, MVIB requires a

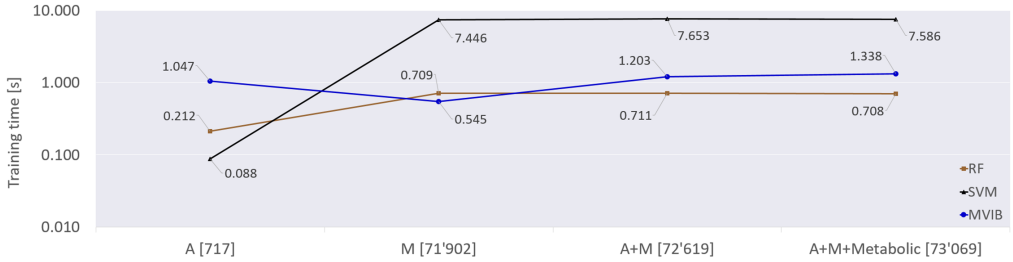

Fig 7. Training time comparison across different machine learning models and various feature space dimensions. The values reported in this graph consist in the training time measured in seconds. The $\text{Colorectal}_{\text{Metabolic}}$ dataset extracted from [35] (see Section Datasets) has been used for training all models. The depicted training times are obtained by averaging the run times of five experiments with different random train/test splits (see Section Validation framework and performance evaluation). A: species-relative abundance profiles. M: strain-level marker profiles. Metabolic: metabolite profiles. RF: Random Forest. SVM: Support Vector Machine. For MVIB, the J_{MVIB} objective has been adopted for the optimisation (see Eq 5). On the x-axis, next to the modality name, the feature space dimension is reported in square brackets.

longer training time w.r.t. the M setting. This is because the training duration of MVIB is mainly determined by the number of epochs needed to reach the highest validation ROC AUC. The number of required epochs is mainly a function of the learning rate and the batch size. The single forward pass and backpropagation steps are extremely quickly and highly parallelized, hence the feature space dimension plays a less relevant role in the training time of MVIB w.r.t. RF and SVM.

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 [44] and sleep [45]. This work tries to tackle all three challenges.

MVIB consists in a novel multimodal 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. Furthermore, as presented in Section Trimodal MVIB: combining metabolomics and metagenomics, we show that MVIB can also handle the trimodal setting, which additionally involves metabolomic data extracted from mass spectrometry.

The stochastic encoding computed by MVIB is maximally informative about the target label and maximally compressive about the input data modalities. This allows to combine information from the species-relative abundance, the strain-level marker and the metabolite profiles, which allow to compute a complete representation of the host microbiome only when considered together.

MVIB obtains competitive results when compared with state-of-the-art methods for microbiome-based disease prediction, such as Random Forest, DeepMicro and PopPhy-CNN (see Section MVIB achieves competitive results on the multimodal microbiome-based disease prediction task). 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.

We showed that by computing saliency maps (see Section Explaining predictions with saliency) we can identify the input features which mostly contributed to a given prediction. Additionally, we showed that MVIB offers competitive generalisation capabilities (see Section Cross-study generalisation results and benchmark), and that its training time scales well with the size of the feature space (see Section Empirical analysis of the training time).

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. Although this temporal capability has not been investigated in this work, this represents one of our objectives for future research. 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 possibility to extend the architecture MVIB in order to enable the analysis of temporal longitudinal multimodal microbiome data with the aim of detecting anomalous temporal patterns. Although the current architecture of MVIB does not allow to include temporal information, using stochastic encoders which are capable to capture temporal dynamics (e.g. LSTMs [46]) might allow the model to operate predictions on longitudinal data.

Recent studies [47–53] shed light on the role of the microbiome in cancer immunotherapy and on its influence on the response to cancer treatments. Evidence suggests that modulating the gut microbiome could augment therapeutic responses. Leveraging on the multimodal learning capabilities of MVIB, a possible direction for future research consists in using MVIB to combine microbiome-specific, host-specific, treatment-specific and cancer-specific data modalities for the estimation of the probability of success of a given immunotherapy, or for the optimisation of a given cancer treatment. [54, 55] are recent microbiome-based machine learning methods which aim at predicting whether melanoma patients treated with immune checkpoint inhibitors will respond to the therapy. These methods represent a first effort in predicting how the microbiome affects the patients' response to cancer treatments. Nevertheless, they only analyse microbiome-specific data, and do not follow a holistic multimodal approach which would requite further host-specific, treatment-specific and cancer-specific data modalities.

Supporting information

S1 Fig. Trimodal MVIB architecture. This figure depicts the trimodal architecture of MVIB used for combining species-relative abundance, strain-level marker and metabolite profiles for colorectal cancer prediction (see Section Trimodal MVIB: combining metabolomics and metagenomics).

S2 Table. Complete experimental results for the multimodal microbiome-based disease prediction task with MVIB. Results obtained optimising the J_{MVIB-T} objective (see Eq 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. ROC AUC: area under the receiver operating characteristic curve. AC: classification accuracy. F1: F1 score. P: precision. R: recall. D and J refer to the two

pre-processing techniques adopted and the two collections of datasets obtained: *default* (D) and *joint* (J).

S3 File. Random Forest hyperparameter space for cross-validated grid-search. The file contains the hyperparameter space considered for Random Forest in cross-validated grid-search.

S4 File. Effect of the triplet margin loss on the stochastic encodings of the patients. For all datasets considered in this work, this file presents plots of the 2D MVIB stochastic encodings analogous to Fig 3. The depicted curves are the 95% confidence intervals of the patients' stochastic encodings $\boldsymbol{z} \sim p(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2\boldsymbol{I})$; the points are their means $\boldsymbol{\mu}$. The displayed encodings consist only in the test samples obtained from random training-test splits (i.e. the 20% of the dataset not used for training). The K dimension of the latent space has been set to 2 in order to allow a 2D visualisation. Plots derived from both the optimisation of the J_{MVIB-T} objective (see Eq 8) and the optimisation of the J_{MVIB} objective (see Eq 5) are included. Five copies of all plots are available, as they are obtained by training the model with five different independent train-test random splits (see Section Validation framework and performance evaluation).

S5 File. PCA 2D projections and MVIB 2D stochastic encoding This file presents, for each dataset, the plots of the PCA 2D projections, as well as the plots of the mean of the MVIB 2D stochastic encodings. For the MVIB stochastic encodings $\boldsymbol{z} \sim p(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2\boldsymbol{I})$, the depicted points represent the mean $\boldsymbol{\mu}$. The K dimension of the latent space has been set to 2 in order to allow a 2D visualisation of the encodings. For training MVIB, the J_{MVIB-T} objective (see Eq 8) has been optimised. For MVIB, five copies of the means plots are available, as they are obtained by training the model with five different independent train-test random splits (see Section Validation framework and performance evaluation). Both the PCA and the MVIB plots have been created starting from the *default* datasets collection (see Section Pre-processing).

S6 File. Top microbial species and strain markers sorted by saliency for all datasets These files present the plots of the top 25 microbial species and strain markers for all datasets considered in this work, analogous to what Fig 5 depicts for the $\text{Colorectal}_{\text{EMBL}}$ dataset. Additionally, the scripts used to create the plots are included.

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. Wirbel J, Zych K, Essex M, Karcher N, Kartal E, et al. Microbiome meta-analysis and cross-disease comparison enabled by the SIAMCAT machine learning toolbox Genome Biology volume. 2021. 22(1):1-27.

20. 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

21. 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

22. 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

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

24. 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.

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

26. 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.

27. 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.

28. 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.

29. 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.

30. 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.

31. 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.

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

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

34. 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.

35. Yachida S, Mizutani S, Shiroma H, Shiba S, Nakajima T, et al Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer Nature Medicine. 2019. 25(6):968-976

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

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

38. 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

39. 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

40. 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

41. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library Advances in Neural Information Processing Systems. 2019. 8024–8035

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

43. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, et al. Scikit-learn: Machine Learning in Python Journal of Machine Learning Research. 2011. 12:2825-2830

44. 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):2019-2111.

45. 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.

46. Hochreiter S, Schmidhuber J. Long short-term memory Neural computation. 1997. 9(8):1735-1780.

47. 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.

48. 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.

49. Davar D, Dzutsev AK, McCulloch JA, Rodrigues RR, Chauvin JM, et al. Fecal microbiota transplant overcomes resistance to anti–PD-1 therapy in melanoma patients Science. 2021. 371(6529):595-602.

50. Hakozaki T, Richard C, Elkrief A, Hosomi Y, Benlaïfaoui M, et al. The Gut Microbiome Associates with Immune Checkpoint Inhibition Outcomes in Patients with Advanced Non–Small Cell Lung Cancer Cancer Immunology Research. 2020. 8(10):1243-1250.

51. Mager L, Burkhard R, Pett N, Cooke NCA, Brown K, et al. Microbiome-derived inosine modulates response to checkpoint inhibitor immunotherapy Science. 2020. 369(6510):1481-1489.

52. Gong J, Chehrazi-Raffle A, Placencio-Hickok V, Guan M, Hendifar A, et al. The gut microbiome and response to immune checkpoint inhibitors: preclinical and clinical strategies Clinical and translational medicine. 2019. 8(1):1-14.

53. Peters BA, Wilson M, Moran U, Pavlick A, Izsak A, et al. Relating the gut metagenome and metatranscriptome to immunotherapy responses in melanoma patients Genome medicine. 2019, 11(1):1-14.

54. Limeta A, Ji B, Levin M, Gatto F, Nielsen J. Meta-analysis of the gut microbiota in predicting response to cancer immunotherapy in metastatic melanoma. JCI Insight. 2020, 5(23):e140940.

55. Oh M, Zhang L. DeepGeni: Deep generalized interpretable autoencoder elucidates gut microbiota for better cancer immunotherapy. bioRxiv [Preprint]. 2021 [cited 2021 August 2021]. Available from: https://doi.org/10.1101/2021.05.06.443032