Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Hierarchical Sparse Bayesian Multitask Learning for Disease Prediction in Pooled Microbiome Studies

Haonan Zhu
Lawrence Livermore National Laboratory
Andre R Goncalves
goncalves1@llnl.gov

Lawrence Livermore National Laboratory
Car Reen Kok
Lawrence Livermore National Laboratory
Camilo Valdes
Lawrence Livermore National Laboratory
Hiranmayi Ranganathan
Lawrence Livermore National Laboratory
Boya Zhang
Lawrence Livermore National Laboratory
Jose Manuel Martí
Lawrence Livermore National Laboratory
Monica K Borucki
Lawrence Livermore National Laboratory
Nisha J Mulakken
Lawrence Livermore National Laboratory
James B Thissen
Lawrence Livermore National Laboratory
Crystal Jaing
Lawrence Livermore National Laboratory
Alfred Hero
University of Michigan–Ann Arbor
Nicholas A Be
Lawrence Livermore National Laboratory

Method Article

Keywords: human microbiome, microbial biomarkers, disease prediction, multi-study analysis, Bayesian hierarchical models, uncertainty quantification

Posted Date: June 8th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9785007/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at BioData Mining on September 2nd, 2026. See the published version at https://doi.org/10.1186/s13040-026-00600-x.

Hierarchical Sparse Bayesian Multitask Learning for Disease Prediction in Pooled Microbiome Studies

Haonan Zhu[1,2], Andre R. Goncalves[1*], Car Reen Kok[1], Camilo Valdes[1], Hiranmayi Ranganathan[1], Boya Zhang[1], Jose Manuel Martí[1], Monica K. Borucki[1], Nisha J. Mulakken[1], James B. Thissen[1], Crystal Jaing[1], Alfred Hero[2], Nicholas A. Be[1*]

[1]Lawrence Livermore National Laboratory, Livermore, CA, USA.
[2]Electrical and Computer Engineering, University of Michigan, Ann Arbor, MI, USA.

*Corresponding author(s). E-mail(s): goncalves1@llnl.gov; be1@llnl.gov;

Abstract

Background: Microbiome-based disease prediction across pooled studies is challenging because the data are high dimensional, heterogeneous, and often too limited to support reliable study-specific models. We propose a hierarchical sparse Bayesian multitask logistic regression model that encourages shared sparsity across related studies while retaining interpretability and uncertainty quantification. To make posterior inference scalable, we derive a variational approximation for the model parameters.

Results: We evaluate the method on synthetic datasets and on a pooled metagenomic collection comprising 61 previously published microbiome studies spanning 19 disease conditions. In simulation, the proposed approach improves support recovery when regression coefficients share a common sparse structure across tasks. On the pooled microbiome application, the proposed method achieves competitive predictive performance while offering clear advantages in interpretability, consistently identifying sparse sets of informative taxa and providing calibrated probabilistic predictions with credible intervals, even under substantial cross-study heterogeneity.

Conclusion: These results suggest that the method is particularly useful as an interpretable, uncertainty-aware framework for extracting shared microbial signals from heterogeneous microbiome studies.

Keywords: human microbiome, microbial biomarkers, disease prediction, multi-study analysis, Bayesian hierarchical models, uncertainty quantification

1 Background

The human microbiome has an immense impact on human health, and a growing body of work demonstrates its promise as a basis for disease prediction, including in gastrointestinal (GI) health [1–5], cancer [6–8], neurological disease [3, 9], and immune dysfunction [4, 5]. This has motivated increasing interest in employing machine learning models to perform health prediction based on human gut microbiome data [10–13]. In practice, however, microbiome data poses particular difficulties for standard machine learning approaches. First, microbiome datasets are high dimensional and small in sample size, making overparameterized models prone to overfitting. Second, clinical applications require interpretable models that quantify prediction uncertainty, a property most deep neural networks lack [14]. One natural response to the sample-size problem is pooling data across studies, but this introduces its own complications.

While pooling microbiome studies offers an opportunity to improve disease prediction and feature discovery, it introduces substantial heterogeneity arising from differences in cohort composition, experimental design, laboratory protocols, and sequencing pipelines [15]. This heterogeneity makes it difficult to obtain models that are both stable and interpretable, a particularly important property in bioinformatics, where investigators are often interested not only in predictive accuracy but also in identifying microbial features that can be validated experimentally.

In this work, we propose a hierarchical sparse Bayesian multitask logistic regression model for binary classification across related microbiome studies. The model uses a shared sparsity structure to identify taxa that are informative across tasks while allowing task-specific effect sizes. Relative to optimization-based multitask methods [16–18], the Bayesian formulation provides posterior uncertainty estimates and a principled way to encode prior assumptions about relationships among studies.

Our approach also differs from prior multitask sparse models [16, 17, 19–22] in how the sparsity pattern is represented. Following [22], we use an overparameterized Bernoulli-Gaussian formulation rather than regularization penalties, which has been associated with improved support recovery [23]. Because the posterior distribution is not available in closed form, we use variational inference to obtain a scalable approximation for high-dimensional microbiome feature spaces.

We evaluate the model on synthetic datasets and on a pooled metagenomic resource containing 61 previously published studies across 19 disease conditions. The simulation experiments show that the approach performs well when the underlying tasks share sparse structure, particularly in support recovery. On the real microbiome data, the main benefit is not uniform improvement in predictive accuracy, but rather the ability to recover sparse, interpretable microbial signatures while providing calibrated uncertainty estimates, which are critical for clinical reliability and biological interpretability.

2 Methods

2.1 Data collection and preprocessing

To ensure comparability across studies, all raw DNA sequence data were processed using a unified pipeline. Sequence data were downloaded from the appropriate repository for each of the 61 studies referenced above. All datasets consisted of short-read data generated on Illumina sequencing platforms. Reads were quality filtered with

fastp [24]. Metagenomic classification and taxonomic assignment were performed with Centrifuge [25, 26], using a reference index built from the NCBI Nucleotide (nt) database [27] as of January 4, 2023. Post-processing and filtering with a minimum hit length of 40 were performed using Recentrifuge [28]. The output consisted of read counts assigned to each observed taxonomic identifier (taxID) in the reference index. Full details of the data collection and processing pipeline, including the complete list of studies, are provided in [29]. The list of all microbiome studies used in this work and their associated metadata is provided in Section 1 of the Supplementary Information.

Read count values for each taxonomic identifier were further processed by centered log ratio (CLR) transform [30] to provide features for the machine learning model. Seven total taxonomic rank levels were examined for purposes of this analysis, and we take the union set of all the taxIDs for each taxonomic rank to provide a consistent feature space. The taxonomic ranks are: Kingdom (11), Phylum (160), Class (313), Order (868), Family (2,338), Genus (7,519) and Species (31,679), where each number in the parentheses corresponds to the feature dimension (i.e., total number of taxIDs). The health condition studied for each patient is one-hot encoded into one of the 19 conditions (see Supplementary Information document).

2.2 Notations

We refer to the abundance of each taxonomic identifier (taxID) as a feature. Each study is referred to as a task, and the disease status of each individual is treated as a binary label. Bold upper-case letters denote matrices, bold lower case letters denote vectors and no bold lower case denote scalars. The Hadamard (element-wise) product of vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ is denoted by $\boldsymbol{a} \circ \boldsymbol{b}$, and diag denotes the function map a vector to a diagonal matrix with the vector as its diagonal entries. We denote the observed taxID count data after centered log ratio transform as $\boldsymbol{x}_t^i \in \mathbb{R}^d$ together with its label $y_t^i \in \{0, 1\}$, where d corresponds to the number of features (i.e., number of taxIDs),

$t = 1, \ldots, T$ denotes the different tasks (i.e., different studies), $i = 1, \ldots, n_t$ denotes the different subjects per study, n_t denotes the total number of subjects per task, and y_t^i reflect whether the subject is diseased (1) or control (0). For a given regression weight matrix $\boldsymbol{W} \in \mathbb{R}^{T \times d}$ for all the tasks, we denote $\boldsymbol{w}_t \in \mathbb{R}^d$ the row vector of $\boldsymbol{W}$ corresponds to t-th task across features, and $\boldsymbol{w}_{(j)} \in \mathbb{R}^T$ the column vector of $\boldsymbol{W}$ corresponds to j-th feature across tasks.

2.3 Hierarchical sparse Bayesian multitask logistic regression model

Although the human microbiome comprises a vast number of microbial species, only a small subset is typically associated with disease [31], motivating the assumption that only a sparse subset of taxa is informative for disease prediction. We encode this assumption with a Bernoulli-Gaussian prior [23, 32] in which a binary inclusion vector $\boldsymbol{z}$ determines whether a feature is active and the continuous weights $\boldsymbol{w}_t$ determine the effect size for each task. Because $\boldsymbol{z}$ is shared across tasks, the model favors a common sparse support, while still allowing the magnitude of each effect to vary from study to study. Excluded taxa are assigned exactly zero weight rather than merely shrunk, and a beta hyperprior on the inclusion probability allows the overall sparsity level to be inferred from the data.

Further, to enforce the information sharing across different tasks, the row of $\boldsymbol{W}$, denoted by $\boldsymbol{w}_{(j)}$ are assumed to be $i.i.d$ draws from a multivariate Gaussian distribution with mean $\mathbf{0}$ and covariance $\boldsymbol{\Sigma}_0$. A Wishart prior is proposed for this shared covariance matrix to provide a method for utilizing expert knowledge about the underling relationships among the studies. This enables us to exploit the Wishart-normal conjugacy to obtain efficient inference later. The proposed conditional model can be summarized:

$$
\begin{aligned}
y_t^i | \boldsymbol{w}_t, \boldsymbol{z}; \boldsymbol{x}_t^i &\overset{\text{independent}}{\sim} \text{Bernoulli}\left(\text{sigmoid}\left(\left\langle (\boldsymbol{w}_t \circ \boldsymbol{z}), \boldsymbol{x}_t^i \right\rangle\right)\right) \\
&\forall i = 1, \ldots, n_t, \forall t = 1, \ldots, T, \\
z_j | \theta &\overset{\text{i.i.d}}{\sim} \text{Bernoulli}(\theta) \quad \forall j = 1, \ldots, d, \\
\theta &\sim \text{Beta}(\alpha_0, \beta_0), \\
\boldsymbol{w}_{(j)} | \boldsymbol{\Sigma}_0 &\overset{\text{i.i.d}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_0) \quad \forall j = 1, \ldots, d, \\
\boldsymbol{\Sigma}_0^{-1} &\sim \text{Wishart}(v_0, \boldsymbol{V}_0).
\end{aligned}
$$

where $\alpha_0, \beta_0, v_0, \boldsymbol{V}_0$ are hyperparameters selected by the experimenter. Smaller value of the ratio $\frac{\alpha_0}{\alpha_0+\beta_0}$ corresponds to a prior belief of fewer informative features while the magnitude α_0 controls the confidence of prior belief, and $v_0, \boldsymbol{V}_0$ reflects the prior knowledge about the covariance structure of the regression coefficient $\boldsymbol{W}$ across different tasks. See Fig. 1 for a visualization of the proposed model and its cluster extension used in subsection 3.2.

The model induces the following conditional log-probability, up to an additive constant, where Γ denotes the gamma function.

$$
\begin{aligned}
&\log\left(p\left(\boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\Sigma}_0, \theta \mid \boldsymbol{X}; v_0, \boldsymbol{V}_0, \alpha_0, \beta_0\right)\right) \\
&= -\frac{1}{2}\operatorname{tr}\left(\boldsymbol{V}_0^{-1}\boldsymbol{\Sigma}_0^{-1}\right) \\
&\quad + \frac{v_0 + d - T - 1}{2}\log\det\left(\boldsymbol{\Sigma}_0^{-1}\right) \\
&\quad - \frac{v_0}{2}\log\det\left(\boldsymbol{V}_0\right) \\
&\quad + (\alpha_0 - 1)\log(\theta) + (\beta_0 - 1)\log(1-\theta) \\
&\quad + \log\Gamma(\alpha_0 + \beta_0) - \log\Gamma(\alpha_0) - \log\Gamma(\beta_0) \\
&\quad - \sum_j \frac{1}{2}\left\langle \boldsymbol{w}_{(j)}, \boldsymbol{\Sigma}_0^{-1}\boldsymbol{w}_{(j)} \right\rangle \\
&\quad + \sum_t \sum_i y_t^i \left(\left\langle (\boldsymbol{w}_t \circ \boldsymbol{z}), \boldsymbol{x}_t^i \right\rangle\right)
\end{aligned}
$$

$$
\begin{aligned}
&- \sum_t \sum_i \log \left(\exp \left(\left\langle (\boldsymbol{w}_t \circ \boldsymbol{z}), \boldsymbol{x}_t^i \right\rangle \right) + 1 \right) \\
&+ \left(\sum_j z_j \right) \log \theta + \left(d - \sum_j z_j \right) \log (1 - \theta) . \qquad (1)
\end{aligned}
$$

2.4 Variational inference

With the combination of the logistic function and the hierarchical structure, inference from the exact posterior distribution of the conditional model is intractable since the posterior distribution is not available in closed form. We resort to a variational approach [33, 34] where we approximate the posterior distribution with a simpler distribution, and the approximation is iteratively refined. We refer interested readers to [33] for a comprehensive review on variational inference (VI) as a general approach for Bayesian inference.

Section 2.4.1 introduces the mean-field approximation used to approximate the posterior distribution along with the variational objective function, and Section 2.4.2 summarizes the optimization algorithm based on coordinate ascent [35].

2.4.1 Mean-field approximation and variational lower bound

We adopt a mean-field variational family because it balances flexibility and computational tractability. The approximation factorizes the posterior over the global sparsity parameter, the shared covariance matrix, the binary inclusion variables, and the task-specific regression coefficients. This factorization makes it possible to derive closed-form updates for most terms while preserving the key latent structure of the model.

In this work, we propose to approximate the true posterior of the proposed model by the function:

$$
q(\theta, \boldsymbol{W}, \boldsymbol{\Sigma}_0, \boldsymbol{z}) = q(\theta; \alpha, \beta)\, q\left(\boldsymbol{\Sigma}_0^{-1}; v, \boldsymbol{V}\right)
$$

$$\prod_j q\left(z_j;\phi_j\right) q(\boldsymbol{w}_{(j)};\boldsymbol{m}_{(j)},\boldsymbol{\Sigma}_j). \tag{2}$$

Here, $q\left(\theta;\alpha,\beta\right)$ is a beta distribution with parameters (α,β), $q\left(z_j;\phi_j\right)$ is a Bernoulli distribution with parameters ϕ_i, $q\left(\boldsymbol{\Sigma}_0^{-1};v,\boldsymbol{V}\right)$ is a Wishart distribution with parameters $(v,\boldsymbol{V})$, and $q\left(\boldsymbol{w}_{(j)};\boldsymbol{m}_{(j)},\boldsymbol{\Sigma}_j\right)$ is a multivariate Gaussian distribution with mean $\boldsymbol{m}_{(j)}$ and covariance $\boldsymbol{\Sigma}_j$ for $j=1,\ldots,d$. Together, these factors define the variational family used to approximate the true posterior.

The optimization objective of variational inference (VI) is to minimize the KL-divergence between the approximation Eqn. 2 and the true posterior by maximizing the evidence lower bound (ELBO):

$$\begin{aligned}
\mathrm{ELBO}(q) &= \mathbb{E}_q\left[\log f(\boldsymbol{X},\boldsymbol{Y},\theta,\boldsymbol{z},\boldsymbol{W},\boldsymbol{\Sigma}_0)\right] + \mathrm{Entropy}(q) \\
&= -\frac{1}{2}\operatorname{tr}(v\boldsymbol{V}_0^{-1}\boldsymbol{V}) + \frac{v_0+d}{2}\log\det(\boldsymbol{V}) \\
&\quad + \frac{v_0+d-v}{2}\psi_T\left(\frac{v}{2}\right) + \frac{vT}{2} + \ln\Gamma_T\left(\frac{v}{2}\right) \\
&\quad + \left(\alpha_0 + \sum_j \phi_j - \alpha\right)\psi(\alpha) + \ln B(\alpha,\beta) \\
&\quad + \left(\beta_0 + d - \sum_j \phi_j - \beta\right)\psi(\beta) \\
&\quad + (\alpha+\beta-d-\alpha_0-\beta_0)\,\psi(\alpha+\beta) \\
&\quad - \frac{1}{2}\sum_j \operatorname{tr}(v\boldsymbol{V}(\boldsymbol{m}_{(j)}\boldsymbol{m}_{(j)}^\top + \boldsymbol{\Sigma}_j)) \\
&\quad + \sum_t\sum_i y_t^i\left(\left\langle(\boldsymbol{m}_t\circ\boldsymbol{\phi}),\boldsymbol{x}_t^i\right\rangle\right) \\
&\quad - \sum_t\sum_i \mathbb{E}_{\boldsymbol{W},\boldsymbol{z}\sim q}\left[\log\left(\exp\left(\left\langle(\boldsymbol{w}_t\circ\boldsymbol{z}),\boldsymbol{x}_t^i\right\rangle\right)+1\right)\right] \\
&\quad + \frac{1}{2}\sum_j \log\det(\boldsymbol{\Sigma}_j) - \sum_j \phi_j\log(\phi_j) \\
&\quad - \sum_j (1-\phi_j)\log(1-\phi_j).
\end{aligned} \tag{3}$$

where ψ is the digamma function, ψ_T is the multivariate digamma function and γ_T is the multivariate gamma function. Since the expectations of the sigmoid functions do

not admit closed form solutions, we approximate the sigmoid functions by quadratic lower bounds:

$$
\begin{aligned}
& -\log\left(\exp\left(\left\langle (\boldsymbol{w}_t \circ \boldsymbol{z}), \boldsymbol{x}_t^i \right\rangle\right) + 1\right) \\
& \geq -\log\left(\exp\left(\left\langle (\boldsymbol{w}_t' \circ \boldsymbol{z}'), \boldsymbol{x}_t^i \right\rangle\right) + 1\right) \\
& \quad - \frac{\left\langle \boldsymbol{x}_t^i, \boldsymbol{w}_t \circ \boldsymbol{z} - \boldsymbol{w}_t' \circ \boldsymbol{z}' \right\rangle}{\exp\left(\left\langle (\boldsymbol{w}_t' \circ \boldsymbol{z}'), \boldsymbol{x}_t^i \right\rangle\right) + 1} \\
& \quad - \frac{1}{8} \left(\boldsymbol{w}_t \circ \boldsymbol{z} - \boldsymbol{w}_t' \circ \boldsymbol{z}'\right)^\top \boldsymbol{x}_t^i \left(\boldsymbol{x}_t^i\right)^\top \left(\boldsymbol{w}_t \circ \boldsymbol{z} - \boldsymbol{w}_t' \circ \boldsymbol{z}'\right). \qquad (4)
\end{aligned}
$$

for $i = 1, \ldots, n_t, t = 1, \ldots, T$, and $\boldsymbol{w}_t', \boldsymbol{z}'$ are deterministic reference points of choice. This type of approximations has been used to design majorize-minimization (MM) algorithms for the logistic regression problem [36, 37]. The resulting approximation is, up to a constant:

$$
\begin{aligned}
& -\sum_t \sum_i \mathbb{E}_{\boldsymbol{W}, \boldsymbol{z} \sim q}\left[\log\left(\exp\left(\left\langle (\boldsymbol{w}_t \circ \boldsymbol{z}), \boldsymbol{x}_t^i \right\rangle\right) + 1\right)\right] \\
& \geq -\sum_t \sum_i \log\left(\exp\left(\left\langle (\boldsymbol{w}_t' \circ \boldsymbol{z}'), \boldsymbol{x}_t^i \right\rangle\right) + 1\right) \\
& \quad + \sum_t \sum_i \frac{\left\langle \boldsymbol{w}_t' \circ \boldsymbol{z}', \boldsymbol{x}_t^i \right\rangle}{\exp\left(-\left\langle (\boldsymbol{w}_t' \circ \boldsymbol{z}'), \boldsymbol{x}_t^i \right\rangle\right) + 1} \\
& \quad - \frac{1}{8} \sum_t \sum_i \left\langle \boldsymbol{w}_t' \circ \boldsymbol{z}', \boldsymbol{x}_t^i \right\rangle^2 \\
& \quad - \sum_t \sum_i \frac{1}{\exp\left(-\left\langle (\boldsymbol{w}_t' \circ \boldsymbol{z}'), \boldsymbol{x}_t^i \right\rangle\right) + 1} \left\langle \boldsymbol{x}_t^i, (\boldsymbol{m}_t \circ \boldsymbol{\phi}) \right\rangle \\
& \quad + \frac{1}{4} \sum_t \sum_i \left\langle \boldsymbol{w}_t' \circ \boldsymbol{z}', \boldsymbol{x}_t^i \right\rangle \left\langle \boldsymbol{m}_t \circ \boldsymbol{\phi}, \boldsymbol{x}_t^i \right\rangle \\
& \quad - \frac{1}{8} \sum_t \sum_i \left\langle \boldsymbol{m}_t \circ \boldsymbol{\phi}, \boldsymbol{x}_t^i \right\rangle^2 \\
& \quad + \frac{1}{8} \sum_t \sum_j \left(m_{t,j}^2 (\phi_j - 1) - (\boldsymbol{\Sigma}_j)_{t,t}\right) \phi_j \sum_i (x_t^{ij})^2. \qquad (5)
\end{aligned}
$$

Alternative lower bounds are available [35, 38, 39], but they introduce additional variational parameters that scale with the feature dimension d. We avoid those formulations here in order to keep inference manageable for high-dimensional microbiome datasets.

2.4.2 Coordinate Ascent Variational Inference (CAVI)

Coordinate ascent variational inference (CAVI) [33, 35] optimizes one block of variational parameters at a time while holding the remaining blocks fixed. In this setting, that strategy yields an iterative algorithm that alternates among updates for the coefficient means and covariances, the feature-inclusion probabilities, and the global hyperparameters. The resulting procedure is substantially more scalable than sampling-based alternatives for the high-dimensional microbiome setting considered here.

One useful result (Equation 18 of [33]) states that if we are to approximate a general posterior distribution $p(\boldsymbol{\xi} \mid \text{data})$ with a mean-field approximation $q(\boldsymbol{\xi}) := \prod_j q_j(\xi_j)$, the CAVI update for j-th latent variable ξ_j (i.e., the optimal solution $q_j^\star(\xi_j)$) is proportional to the exponentiated conditional expected log of the joint:

$$q_j^\star \propto \exp\left(\mathbb{E}_{\boldsymbol{\xi}_{-j} \sim q_{-j}}\left[\log\left(p\left(\xi_j, \boldsymbol{\xi}_{-j} \mid \text{data}\right)\right)\right]\right). \tag{6}$$

where $\boldsymbol{\xi}_{-j}$ corresponds to all but the j-th latent variable.

A standard result in variational inference states that, under a mean-field factorization, each variational factor is updated by exponentiating the conditional expectation of the joint log-density with respect to the remaining factors. We use this identity to derive the updates summarized in Algorithm 1; full derivations are provided in Section 2 of the Supplementary Information document.

Algorithm 1 CAVI for Bayesian Multitask Sparse Logistic Regression

procedure CAVI(Inputs: $\left(\boldsymbol{x}_t^i, y_t^i\right)_{t=1,\ldots T, i=1,\ldots n_t}$)
  for all itr $= 1, \ldots,$ Niter **do**
    for all $j = 1, \ldots d$ **do**

$$\tilde{\boldsymbol{X}}_j \leftarrow \operatorname{diag}\left(\begin{bmatrix} \sum_i \phi_j (x_1^{ij})^2 \\ \vdots \\ \sum_i \phi_j (x_T^{ij})^2 \end{bmatrix}\right)$$

$$\boldsymbol{\Sigma}_j \leftarrow \left(v\boldsymbol{V} + \tfrac{1}{4}\tilde{\boldsymbol{X}}_j\right)^{-1}$$

    end for
    for all $j = 1, \ldots, d$ **do**
      for all $t = 1, \ldots, T, i = 1, \ldots, n_t$ **do**

$$\tilde{y_t^i} \leftarrow \text{sigmoid}\left(\textstyle\sum_j m_{tj} x_t^{ij} \phi_j\right)$$

      end for

$$\boldsymbol{m}_{(j)} \leftarrow \boldsymbol{\Sigma}_j \begin{bmatrix} \phi_j \sum_i \left(y_1^i - \tilde{y_1^i}\right) x_1^{ij} \\ \vdots \\ \phi_j \sum_i \left(y_T^i - \tilde{y_T^i}\right) x_T^{ij} \end{bmatrix} + \tfrac{1}{4}\boldsymbol{\Sigma}_j \begin{bmatrix} \phi_j^2 \sum_i \left(x_1^{ij}\right)^2 m_{1j} \\ \vdots \\ \phi_j^2 \sum_i \left(x_T^{ij}\right)^2 m_{Tj} \end{bmatrix}$$

    end for
    for all $j = 1, \ldots, d$ **do**
      for all $t = 1, \ldots, T, i = 1, \ldots, n_t$ **do**

$$\tilde{y_t^i} \leftarrow \text{sigmoid}\left(\textstyle\sum_j m_{tj} x_t^{ij} \phi_j\right)$$

      end for

$$\phi_j \leftarrow \text{sigmoid}\Big(\psi(\alpha) - \psi(\beta) + \textstyle\sum_t \sum_i (y_t^i - \tilde{y}_t^i) m_{tj} x_t^{ij} + \tfrac{1}{4}\sum_t (m_{tj}^2 \phi_j) \sum_i (x_t^{ij})^2 - \tfrac{1}{8}\sum_t m_{tj}^2 \sum_i (x_t^{ij})^2 - \tfrac{1}{8}\sum_t (\Sigma_j)_{tt} \sum_i (x_t^{ij})^2\Big)$$

    end for

$$\alpha \leftarrow \alpha_0 + \textstyle\sum_j \phi_j,$$

$$\beta \leftarrow \beta_0 + d - \textstyle\sum_j \phi_j,$$

$$\theta \leftarrow \tfrac{\alpha}{\alpha+\beta},$$

$$v \leftarrow v_0 + d,$$

$$\boldsymbol{V} \leftarrow \left(\boldsymbol{V}_0^{-1} + \textstyle\sum_j \boldsymbol{m}_{(j)} \boldsymbol{m}_{(j)}^\top + \boldsymbol{\Sigma}_j\right)^{-1}$$

  end for
end procedure

2.5 Experimental setup

We evaluate the proposed method on both synthetic and real microbiome datasets. The synthetic experiments are designed to test whether the model can recover shared sparse structure under controlled conditions, whereas the real-data experiments assess predictive behavior, calibration, and feature interpretability on pooled studies. We

compare BayesMTL with single-task, pooled, and multitask baselines that represent common alternatives for sparse classification, which are described below.

1. Single-task Logistic Classifier with sparsity penalty (STL-LC) [40, 41]: this is a single-task model, where we fit independent logistic models to each task separately. This is an extension of standard LASSO to the binary classification setting.
2. Pooled Logistic Classifier with sparsity penalty (Pooled-LC): we train a single logistic regression model on the combined data from all datasets.
3. MTFL (Multitask Feature Learning) [16]: this is an optimization based approach to multitask learning based on $\ell_{2,1}$-norm regularization. The proposed method can be seen as a Bayesian Hierarchical extension, and the difference in modeling is manifested in how the sparsity pattern is encouraged: we use an overparameterized Bernoulli-Gaussian model, which has better support recovery properties [23].
4. MSSL (Multitask Sparse Structure Learning) [18]: this is an optimization based multitask learning approach, where the imposed sparsity structure is on the precision (inverse covariance matrix) of the regression coefficients across tasks. The optimization problem of this formulation is equivalent to the graphical lasso problem [42, 43] for covariance estimation.

Hyperparameters for all methods are selected using validation cross-entropy. Because the available datasets are relatively small and class imbalance is common, we use repeated stratified cross-validation to reduce the sensitivity of results to a particular train-test split. We report balanced accuracy as the primary predictive metric because it is more informative than raw accuracy under class imbalance.

3 Results and Discussion

3.1 Synthetic Datasets

We generate synthetic datasets to examine whether our algorithm is able to: 1) recover the support of the regression coefficients that correspond to informative features for the prediction, which are evaluated by balanced accuracy; 2) recover the ground truth regression coefficients (up to normalization), which are evaluated by cosine distance (i.e., one minus the normalized inner product).

The synthetic experiments are designed to isolate the conditions under which the proposed multitask prior should help. Specifically, they test whether BayesMTL can recover the correct sparse support and the underlying coefficient structure when tasks truly share informative features. By varying both sparsity level and sample-size imbalance, these simulations probe performance across easier and harder multitask settings. Additional details of the synthetic datasets designed for this experiment is presented in Table 1.

Support recovery: We convert support recovery into a binary prediction problem in which each feature is classified as either relevant or irrelevant. Figure 2 shows that the proposed Bayesian method performs best in most settings, with particularly strong performance when the true coefficients share a sparse support.

Weight recovery: We also assess how accurately each method recovers the underlying regression coefficients. Because logistic predictions are invariant to global scaling, we use cosine distance as the evaluation metric. Table 2 shows that the proposed Bayesian method performs best in all but the dense setting.

3.2 Microbiome Data

In this subsection, we apply the proposed model to previously published microbiome datasets described in Section 1. We aim to assess whether leveraging shared structure

	Balanced	Imbalanced
Dense ($\theta = 0.8$)	Dataset-1	Dataset-4
Sparse ($\theta = 0.2$)	Dataset-2	Dataset-5
Ultra Sparse ($\theta = 0.05$)	Dataset-3	Dataset-6

Table 1: Summary of the simulated dataset, where the sample sizes of balanced data are generated by a Poisson distribution with rate 24, and sample sizes of the imbalanced data are generated by a negative binomial distribution with stopping-time parameter $r = 1$ and probability of success $p = 0.04$. For all the simulations, the number of features is 100 and number of tasks is 10. For the imbalanced datasets, we amplify all the sample sizes by 6 to ensure that both positive samples and negative samples are present across all tasks. Both settings have an expected sample size 30, where the imbalanced case has more variations in sample sizes among different tasks. The θ parameter corresponds to the expected percentage of the predictive features.

across multiple diseases improves classification performance, and whether the model can highlight microbial taxa associated with each disease.

Figure 3 summarizes balanced accuracy across taxonomic ranks. We observe that jointly modeling diseases tends to be more beneficial at coarser taxonomic levels, where

	BayesMTL	MTFL	MSSL	Pooled-LC	STL-LC
Dataset-1	0.74 (0.06)	0.56 (0.02)	**0.55 (0.02)**	0.98 (0.04)	0.73 (0.02)
Dataset-2	**0.34 (0.09)**	0.56 (0.01)	0.56 (0.01)	1.02 (0.05)	0.56 (0.05)
Dataset-3	**0.10 (0.04)**	0.59 (0.03)	0.59 (0.02)	0.99 (0.12)	0.26 (0.04)
Dataset-4	0.67 (0.10)	0.58 (0.08)	**0.58 (0.05)**	1.02 (0.03)	0.73 (0.07)
Dataset-5	**0.46 (0.11)**	0.59 (0.08)	0.57 (0.04)	0.57 (0.09)	1.04 (0.06)
Dataset-6	**0.22 (0.12)**	0.65 (0.05)	0.61 (0.04)	1.07 (0.10)	0.35 (0.11)

Table 2: Summary of the weights recovery measured in cosine distance. The cosine distance is bounded between 0 and 2 with 0 meaning perfect recovery. The proposed Bayesian (BayesMTL) approach outperforms the benchmark methods when there is a shared sparsity structure across regression coefficients of different tasks.

shared patterns across studies are more likely to emerge, although these improvements are not uniform across all datasets. This variability is consistent with the substantial heterogeneity in study design, patient populations, and sequencing protocols. Beyond predictive performance, a key advantage of the proposed model is its ability to produce sparse and interpretable solutions. As shown in Table 3, the model selects a substantially smaller set of taxa than competing approaches while maintaining comparable accuracy. In the context of microbiome analysis, this is particularly valuable, as it narrows the set of candidate taxa for downstream biological investigation.

Taxon Ranks	BayesMTL	MTFL	MSSL	Pooled-LC	STL-LC
Kingdom	0.29 (0.01)	0.54 (0.03)	0.47 (0.03)	0.88 (0.04)	**0.26 (0.02)**
Phylum	0.18 (0.02)	0.30 (0.002)	0.31 (0.003)	0.78 (0.03)	**0.086 (0.02)**
Class	0.12 (0.01)	0.24 (0.006)	0.30 (0.05)	0.69 (0.02)	**0.070** (0.005)
Order	0.11 (0.01)	0.20 (0.01)	0.22 (0.01)	0.54 (0.009)	**0.046** (0.008)
Family	0.043 (0.02)	0.41 (0.3)	0.17 (0.002)	0.41 (0.003)	**0.041 (0.007)**
Genus	**0.018 (0.007)**	0.40 (0.3)	0.15 (0.0001)	0.33 (0.003)	0.035 (0.006)
Species	**0.029 (0.005)**	0.15 (0.03)	0.15 (0.0008)	0.28 (0.004)	0.032 (0.005)

Table 3: Summary of the results for sparsity ratios (percentage of regression coefficients that are zero). The bold number indicates that the corresponding method gives the sparsest result for the given taxonomic rank, and the values in parentheses represent standard deviations computed over 5 different runs. The proposed approach is the only multitask method that provides a sparse solution, identifying common microbes across studies of the same disease category that are informative for predicting health status.

Calibration analysis complements the balanced-accuracy results by assessing how well the model's predicted probabilities reflect observed outcomes. Figure 4 shows that the model is reasonably well calibrated on the training data and retains useful calibration on the test data, although performance degrades near the probability extremes. This pattern suggests overconfident predictions at the boundaries, which may arise from both the choice of the logit link and the domain shift introduced by pooling heterogeneous microbiome studies. Even when predictive accuracy is mixed, the Bayesian formulation remains valuable because it exposes uncertainty rather than returning

only point predictions. Calibration curves for all taxonomic ranks are shown in the Supplementary Information document.

To interpret the fitted model, we combine posterior inclusion behavior and coefficient magnitude into a sparsity coefficient for each taxID. Figure 5 highlights four representative disease categories at the Class taxonomic rank. These examples were selected because they each combine multiple studies with non-trivial total sample sizes and, taken together, span sparse, intermediate, and dense feature-selection regimes. Diabetes and diarrhea are associated with relatively small sets of repeatedly selected taxa, autoimmune disease shows an intermediate pattern, and cancer yields a much broader set of informative features. Rather than suggesting a universal microbiome signature, these results indicate that the degree of sparsity is itself disease dependent. This reinforces the value of a model that can represent both shared structure and uncertainty in feature selection. The remaining disease categories for the Class taxonomic rank, along with selection-stability curves across all taxonomic ranks, are provided in the Supplementary Information document. These allow comparison of sparsity levels across disease groups and taxonomic ranks.

3.2.1 Diarrhea case study: feature importance

To further examine the feature selection capability of the model, we focus on the diarrhea datasets [44–46] since they have shown a relatively consistent sparse patterns (Fig. 5) and are directly relevant to gut and fecal microbiome studies. CLR transform is a normalization technique applied to each sample across features, therefore the magnitude of the transformed data varies across features. We propose to assign feature importance of j-th feature (taxID) for task t and sample i by combining the regression coefficient ($\{\boldsymbol{w}_t\}$), the sparsity parameters ($\boldsymbol{z}$) and the CLR transformed abundance

data ($\boldsymbol{x}_t^i$):

$$\left| \frac{w_{t,j} z_j x_t^{ij}}{\sqrt{\sum_j \left(w_{t,j} z_j x_t^{ij} \right)^2}} \right|. \tag{7}$$

The proposed importance weight can be interpreted as the relative contribution to the predicted log-odds ratio. From the estimated posterior distribution, we draw samples to explore the full distribution of the feature importance. The most important features will correspond to the features with consistently high importance weights across draws, samples, and tasks. After obtaining important features for each taxonomic rank, we visualize the result using taxonomic lineage information, where lineages with at least 5 taxIDs selected are shown in Fig. 6.

3.2.2 Diarrhea case study: biological interpretation

The majority of lineages identified by the model are relevant in the context of the gut microbiome and diet, and highlights the association of *Escherichia coli*, gut bacteriophages (order *Crassvirales*) and a gut fungi, *Debaryomycetaceae* with diarrhea. *Enteroaggregative E. coli*, in particular has been routinely discussed to play a causal role in diarrhea [47–49]. Although our analysis does link *Petitvirales* to healthy controls, these results associate *Crassvirales* with diarrhea (Fig. 6), while previous reports link the abundance of *Crassvirales* with gut health. A number of factors may influence the gut microbiome composition including age, diet, and geographic location [50–52] thus the datasets used to establish this model may have been influenced by population demographics. Specifically, the diarrhea datasets used here are derived from a mixed population of subjects from Bangladesh and the United States [44–46]. Additionally, this analysis does not capture other global variables that can contribute to a disease state such as differences in virome richness and diversity [53]. Finally, differences in bioinformatics pipelines and reference databases used to classify sequences likely contributes to reported differences. For example, up to 99% of phage sequences

do not map to reference viral genomes [51, 52] thus increasing the likelihood of variability in identification of bacteriophage genomes, which are known for their complex and mosaic genome composition. Furthermore, the identification of plant-related features within the Magnoliopsida class suggest the capability of our approach to identify potentially disease-associated dietary components for hypothesis generation. Bacteriophages belonging to orders *Crassvirales* and *Petitvirales* are highly abundant in the human gut microbiome and have been associated with health [50, 52, 54, 55].

4 Conclusions

In this work, we proposed a hierarchical sparse Bayesian multitask logistic regression model for predicting human health status from microbiome abundance data. The model leverages a shared sparsity structure to identify informative taxa across related tasks, while providing uncertainty quantification through a fully probabilistic framework. We developed an efficient variational inference procedure to enable scalable learning in high-dimensional settings.

Simulation studies demonstrate that the proposed approach performs particularly well when tasks share an underlying sparse structure, leading to improved feature recovery and predictive performance. On a real-world microbiome dataset pooled from multiple studies, the model achieves competitive predictive accuracy while consistently identifying sparse and interpretable sets of taxa. These results highlight its ability to extract biologically meaningful signals from heterogeneous data while maintaining well-calibrated probabilistic predictions.

There are several directions for future work. One natural extension is to explore alternative link functions, such as the probit link, which have flatter tails and may reduce the overconfident predictions observed with the logit formulation. Another direction is to incorporate additional hierarchical structure into the model to account for relevant metadata, such as study-specific factors or patient demographics.

It is also of interest to extend the model to more general prediction settings. In particular, a multi-label formulation would allow each task to include multiple binary outcomes, reflecting the fact that diseases are not mutually exclusive in many clinical settings. A related extension is to consider multiclass classification, where each task involves more than two outcome categories, such as different stages of disease progression.

Supplementary information. A supplementary information document accompanies this manuscript, containing references and metadata for all microbiome studies used in this work, as well as additional results.

Acknowledgements. Not applicable.

Declarations

Funding

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344 (LLNL-MI-853606). It was supported by the LLNL Laboratory Directed Research and Development Program, the ETI Consortium under US Department of Energy grant DE-NA0003921, and the US Army Research Office under grant number W911NF1910269.

Competing interests

The authors declare that they have no competing interests.

Ethics approval and consent to participate

This study used previously published, de-identified metagenomic datasets obtained from public repositories. No new human participants were recruited and no directly identifiable human data were collected by the authors for this study.

Consent for publication

All authors have read and approved the manuscript for publication.

Availability of data and materials

The datasets utilized in this study can be found in https://zenodo.org/records/17315984.

Code availability

The code that implements the Bayesian MTL model proposed in this paper can be found here: https://github.com/llnl/BayesMTL

Authors' contributions

HZ: Conceptualization, Investigation, Methodology, Computational Experiments, Formal analysis, Software, Writing – original draft. ARG: Conceptualization, Methodology, Software, Writing – review & editing. CRK: Data curation, Formal analysis, Writing – review & editing. CV: Methodology, Writing – review & editing. HR: Methodology, Writing – review & editing. BZ: Methodology, Writing – review & editing. JMM: Data curation, Writing – review & editing. JBT: Data curation, Writing – review & editing. NJM: Data curation, Writing – review & editing. MB: Methodology, Writing – review & editing. CJ: Project administration, Funding acquisition, Writing – review & editing. AH: Methodology, Supervision, Discussion, Writing – review & editing. NAB: Project administration, Funding acquisition, Supervision, Writing – review & editing.

References

[1] Lee M, Chang EB. Inflammatory bowel diseases (IBD) and the microbiome—searching the crime scene for clues. Gastroenterology. 2021;160(2):524–537.

[2] Nishida A, Inoue R, Inatomi O, Bamba S, Naito Y, Andoh A. Gut microbiota in the pathogenesis of inflammatory bowel disease. Clinical journal of gastroenterology. 2018;11:1–10.

[3] McGrattan AM, McGuinness B, McKinley MC, Kee F, Passmore P, Woodside JV, et al. Diet and inflammation in cognitive ageing and Alzheimer's disease. Current nutrition reports. 2019;8:53–65.

[4] De Luca F, Shoenfeld Y. The microbiome in autoimmune diseases. Clinical & Experimental Immunology. 2019;195(1):74–85.

[5] Thaiss CA, Zmora N, Levy M, Elinav E. The microbiome and innate immunity. Nature. 2016;535(7610):65–74.

[6] Banerjee S, Tian T, Wei Z, Peck KN, Shih N, Chalian AA, et al. Microbial signatures associated with oropharyngeal and oral squamous cell carcinomas. Scientific reports. 2017;7(1):4036.

[7] Sepich-Poore GD, Zitvogel L, Straussman R, Hasty J, Wargo JA, Knight R. The microbiome and human cancer. Science. 2021;371(6536):eabc4552.

[8] 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.

[9] Boddy SL, Giovannelli I, Sassani M, Cooper-Knock J, Snyder MP, Segal E, et al. The gut microbiome: a key player in the complexity of amyotrophic lateral sclerosis (ALS). BMC medicine. 2021;19(1):1–14.

[10] Zhou YH, Gallins P. A review and tutorial of machine learning methods for microbiome host trait prediction. Frontiers in genetics. 2019;10:579.

[11] 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.

[12] Cammarota G, Ianiro G, Ahern A, Carbone C, Temko A, Claesson MJ, et al. Gut microbiome, big data and machine learning to promote precision medicine for cancer. Nature reviews gastroenterology & hepatology. 2020;17(10):635–648.

[13] Manandhar I, Alimadadi A, Aryal S, Munroe PB, Joe B, Cheng X. Gut microbiome-based supervised machine learning for clinical diagnosis of inflammatory bowel diseases. American Journal of Physiology-Gastrointestinal and Liver Physiology. 2021;320(3):G328–G337.

[14] Molnar C. Interpretable machine learning. Lulu. com; 2020.

[15] Goncalves AR, Ranganathan H, Valdes C, Zhu H, Zhang B, Kok CR, et al. Beyond microbial abundance: metadata integration enhances disease prediction in human microbiome studies. Frontiers in Microbiology. 2026;Volume 16 - 2025. https://doi.org/10.3389/fmicb.2025.1695501.

[16] Liu J, Ji S, Ye J. Multi-task feature learning via efficient l2, 1-norm minimization. In: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence; 2009. p. 339–348.

[17] Jalali A, Sanghavi S, Ruan C, Ravikumar P. A dirty model for multi-task learning. Advances in neural information processing systems. 2010;23.

[18] Gonçalves AR, Von Zuben FJ, Banerjee A. Multi-task sparse structure learning with gaussian copula models. The Journal of Machine Learning Research. 2016;17(1):1205–1234.

[19] Xue Y, Liao X, Carin L, Krishnapuram B. Multi-Task Learning for Classification with Dirichlet Process Priors. Journal of Machine Learning Research. 2007;8(1).

[20] Zhang Y, Schneider J. Learning multiple tasks with a sparse matrix-normal penalty. Advances in neural information processing systems. 2010;23.

[21] Guo S, Zoeter O, Archambeau C. Sparse Bayesian multi-task learning. Advances in Neural Information Processing Systems. 2011;24.

[22] Goncalves A, Ray P, Soper B, Widemann D, Nygård M, Nygård JF, et al. Bayesian multitask learning regression for heterogeneous patient cohorts. Journal of Biomedical Informatics. 2019;100:100059.

[23] Bazot C, Dobigeon N, Tourneret JY, Hero AO. A Bernoulli-Gaussian model for gene factor analysis. In: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE; 2011. p. 5996–5999.

[24] Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34(17):i884–i890.

[25] Kim D, Song L, Breitwieser FP, Salzberg SL. Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome research. 2016;26(12):1721–1729.

[26] Marti JM, Kok CR, Thissen JB, Mulakken NJ, Avila-Herrera A, Jaing CJ, et al. Addressing the dynamic nature of reference data: a new nt database for robust metagenomic classification. bioRxiv. 2024;p. 2024–06.

[27] Coordinators NR. Database resources of the national center for biotechnology information. Nucleic acids research. 2012;41(D1):D8–D20.

[28] Martí JM. Recentrifuge: Robust comparative analysis and contamination removal for metagenomics. PLoS computational biology. 2019;15(4):e1006967.

[29] Kok CR, Mulakken NJ, Thissen JB, Marti JM, Lee R, Trainer JB, et al. Meta2DB: Curated Shotgun Metagenomic Feature Sets and Metadata for Health State Prediction. bioRxiv. 2024;p. 2024–10.

[30] Aitchison J. The statistical analysis of compositional data. Journal of the Royal Statistical Society: Series B (Methodological). 1982;44(2):139–160.

[31] Wang B, Yao M, Lv L, Ling Z, Li L. The human microbiota in health and disease. Engineering. 2017;3(1):71–82.

[32] Soussen C, Idier J, Brie D, Duan J. From Bernoulli–Gaussian deconvolution to sparse signal restoration. IEEE Transactions on Signal Processing. 2011;59(10):4572–4584.

[33] Blei DM, Kucukelbir A, McAuliffe JD. Variational inference: A review for statisticians. Journal of the American statistical Association. 2017;112(518):859–877.

[34] Zhang C, Bütepage J, Kjellström H, Mandt S. Advances in variational inference. IEEE transactions on pattern analysis and machine intelligence. 2018;41(8):2008–2026.

[35] Bishop CM, Nasrabadi NM. Pattern recognition and machine learning. vol. 4 of Information Science and Statistics. Springer; 2006. Available from: https://books.google.com/books?id=qWPwnQEACAAJ.

[36] Böhning D, Lindsay BG. Monotonicity of quadratic-approximation algorithms. Annals of the Institute of Statistical Mathematics. 1988;40(4):641–663.

[37] Hunter DR, Lange K. A tutorial on MM algorithms. The American Statistician. 2004;58(1):30–37.

[38] Jaakkola TS, Jordan MI. A variational approach to Bayesian logistic regression models and their extensions. In: Sixth International Workshop on Artificial Intelligence and Statistics. PMLR; 1997. p. 283–294.

[39] Drugowitsch J. Variational Bayesian inference for linear and logistic regression. arXiv preprint arXiv:13105438. 2013;.

[40] Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological). 1996;58(1):267–288.

[41] Banerjee O, El Ghaoui L, d'Aspremont A. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. The Journal of Machine Learning Research. 2008;9:485–516.

[42] Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. Biostatistics. 2008;9(3):432–441.

[43] Mazumder R, Hastie T. The graphical lasso: New insights and alternatives. Electronic journal of statistics. 2012;6:2125.

[44] David LA, Weil A, Ryan ET, Calderwood SB, Harris JB, Chowdhury F, et al. Gut microbial succession follows acute secretory diarrhea in humans. MBio. 2015;6(3):10–1128.

[45] Kieser S, Sarker SA, Sakwinska O, Foata F, Sultana S, Khan Z, et al. Bangladeshi children with acute diarrhoea show faecal microbiomes with increased Streptococcus abundance, irrespective of diarrhoea aetiology. Environmental microbiology. 2018;20(6):2256–2269.

[46] McDonald D, Hyde E, Debelius JW, Morton JT, Gonzalez A, Ackermann G, et al. American gut: an open platform for citizen science microbiome research.

Msystems. 2018;3(3):10–1128.

[47] Robins-Browne RM, Hartland EL. Escherichia coli as a cause of diarrhea. Journal of gastroenterology and hepatology. 2002;17(4):467–475.

[48] Okhuysen PC, DuPont HL. Enteroaggregative Escherichia coli (EAEC): a cause of acute and persistent diarrhea of worldwide importance. The Journal of infectious diseases. 2010;202(4):503–505.

[49] Petro CD, Duncan JK, Seldina YI, Allué-Guardia A, Eppinger M, Riddle MS, et al. Genetic and virulence profiles of enteroaggregative Escherichia coli (EAEC) isolated from deployed military personnel (DMP) with travelers' diarrhea. Frontiers in Cellular and Infection Microbiology. 2020;10:200.

[50] Smith L, Goldobina E, Govi B, Shkoporov AN. Bacteriophages of the order Crassvirales: what do we currently know about this keystone component of the human gut virome? Biomolecules. 2023;13(4):584.

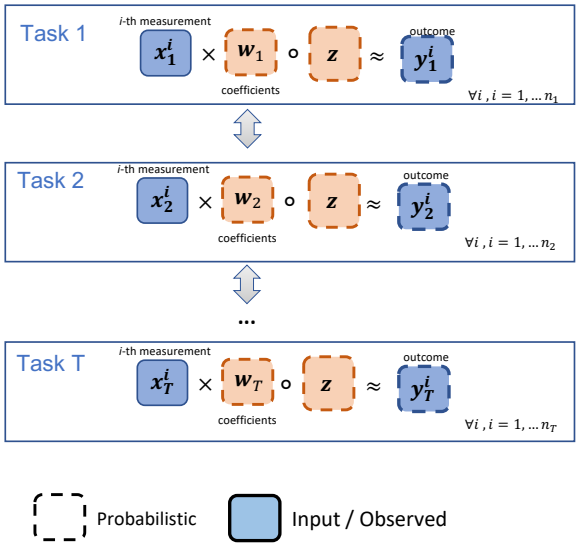
[51] Gregory AC, Zablocki O, Zayed AA, Howell A, Bolduc B, Sullivan MB. The gut virome database reveals age-dependent patterns of virome diversity in the human gut. Cell host & microbe. 2020;28(5):724–740.

[52] Ezzatpour S, Mondragon Portocarrero AdC, Cardelle-Cobas A, Lamas A, López-Santamarina A, Miranda JM, et al. The human gut Virome and its relationship with nontransmissible chronic diseases. Nutrients. 2023;15(4):977.

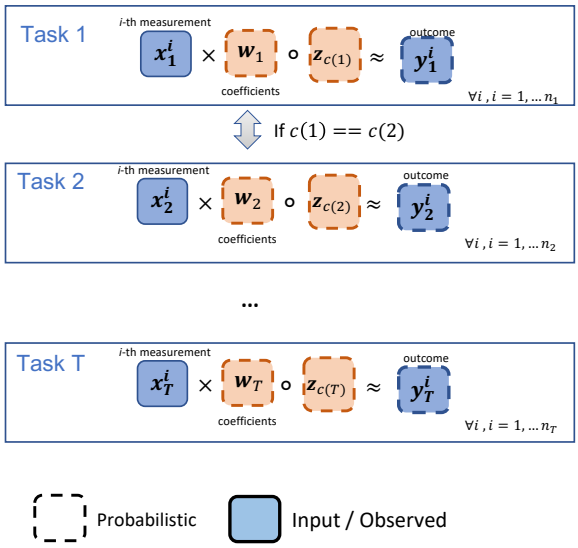
[53] Zhu Q, Dupont CL, Jones MB, Pham KM, Jiang ZD, DuPont HL, et al. Visualization-assisted binning of metagenome assemblies reveals potential new pathogenic profiles in idiopathic travelers' diarrhea. Microbiome. 2018;6:1–20.

[54] Shkoporov AN, Clooney AG, Sutton TD, Ryan FJ, Daly KM, Nolan JA, et al. The human gut virome is highly diverse, stable, and individual specific. Cell host & microbe. 2019;26(4):527–541.

[55] Shkoporov AN, O'Regan O, Smith L, Khokhlova EV, Draper LA, Ross RP, et al. Dynamic nature of viral and bacterial communities in human faeces. Iscience. 2024;27(2).

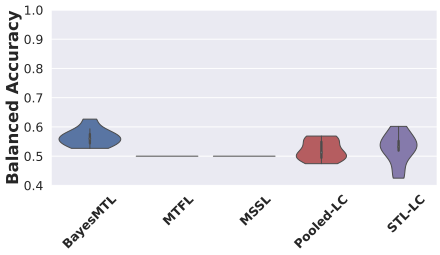


(a) The based model

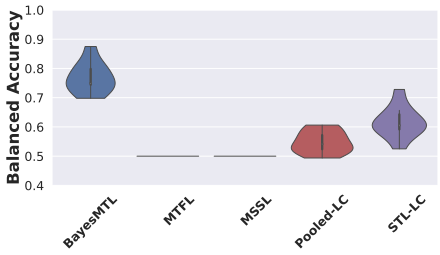


(b) The clustering extension

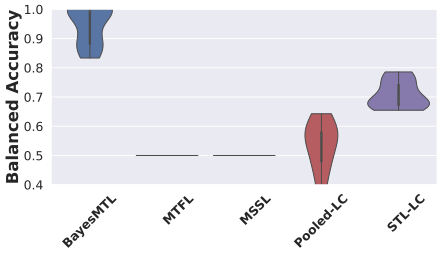
Fig. 1: Graphical visualization of the proposed probabilistic model. For the base model in (a), the information sharing cross different tasks is enforced by common priors between the regression weights and sparsity parameters. For the clustering extension (b), the information sharing is limited to tasks from the same cluster (i.e., disease category).



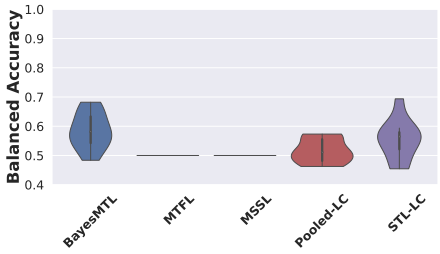
(a) Dataset1: balanced and dense



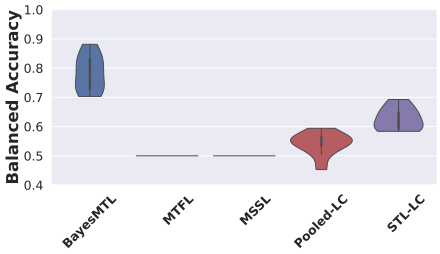
(b) Dataset2: balanced and sparse



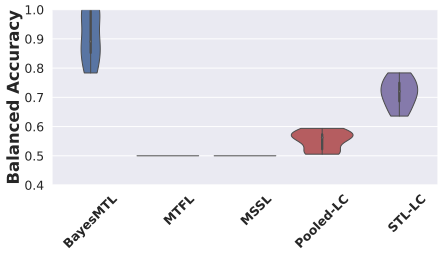
(c) Dataset3: balanced and ultra sparse



(d) Dataset4: imbalanced and dense

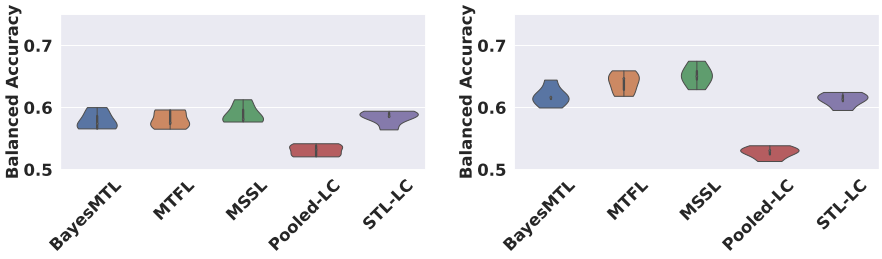


(e) Dataset5: imbalanced and sparse



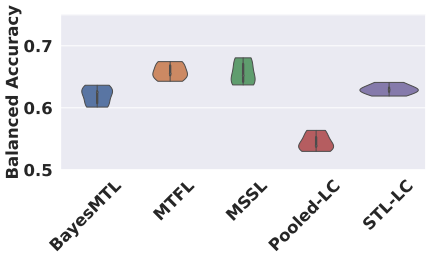
(f) Dataset6: imbalanced and ultra sparse

Fig. 2: Summary of the support recovery results for the simulated data evaluated using balanced accuracy across 10 runs. BayesMTL consistently outperforms baseline methods, particularly when shared sparsity exists across regression coefficients.

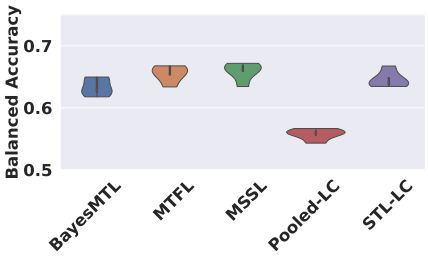


(a) Kingdom

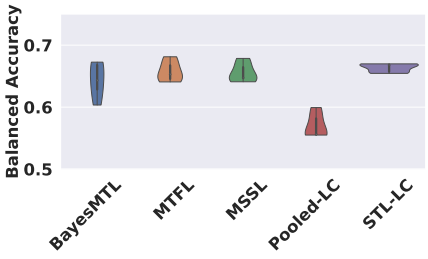
(b) Phylum



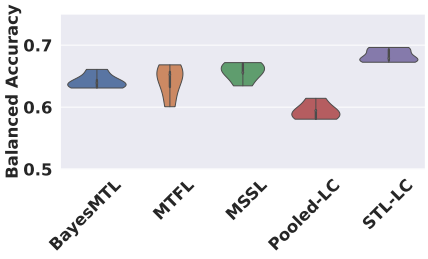
(c) Class



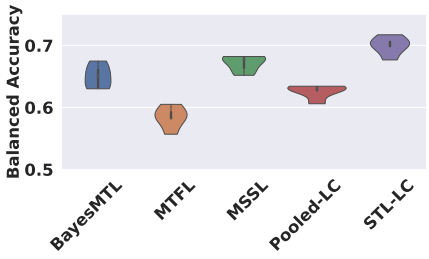
(d) Order



(e) Family

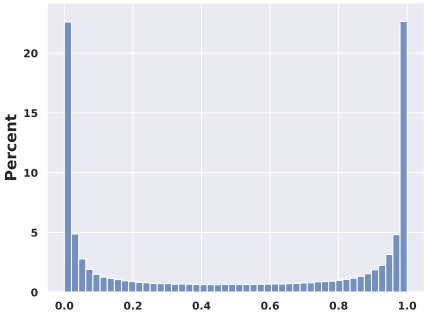


(f) Genus

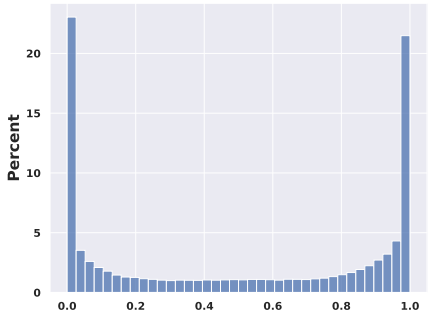


(g) Species

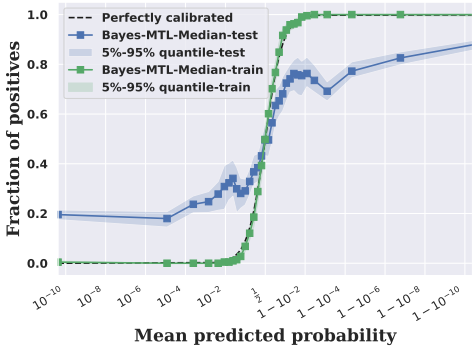
Fig. 3: Summary of the prediction performance evaluated by balanced accuracy across 7 taxonomic ranks. Due to the heterogeneous nature of the data, we do not see an improvement of the proposed approach over single-task models. However, the proposed approach is the only multitask method that provides a sparse solution, i.e., identifying common microbes across studies of the same disease category that are informative for prediction along with uncertainty quantification through the posterior distribution.



(a) Predicted probabilities on training data.

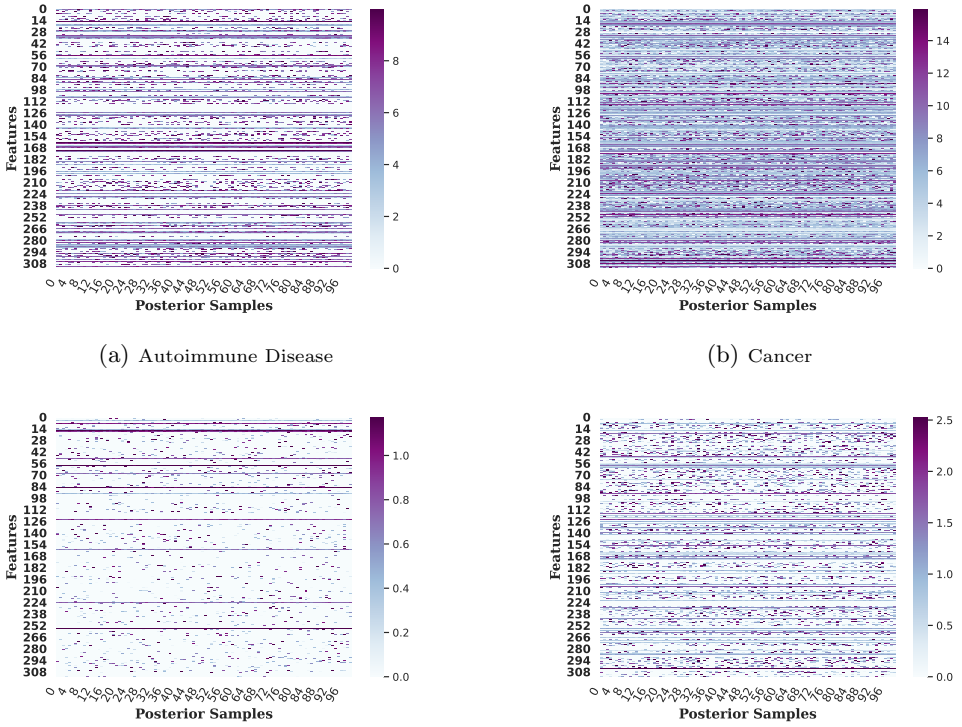


(b) Predicted probabilities on test data.



(c) Calibration curves for the training data and the test data with 90% marginal credible intervals.

Fig. 4: Calibration analysis for the proposed model on the Class taxonomic rank. Fig. (a) and Fig. (b) show the histograms of the predicted probabilities and training and test data respectively. Due to the choice of logit as link function, the predicted probabilities are concentrated around the boundaries. Fig. (c) shows the calibration curves of the predictions from training and test data. The model achieves near perfect calibration on the training data, and the degradation of performance on the test data at the boundary values indicates that the logit function as a link function is resulting in over-confident predictions.



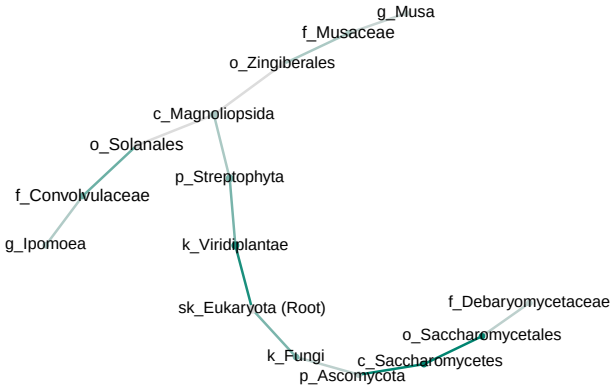
(a) Autoimmune Disease

(b) Cancer

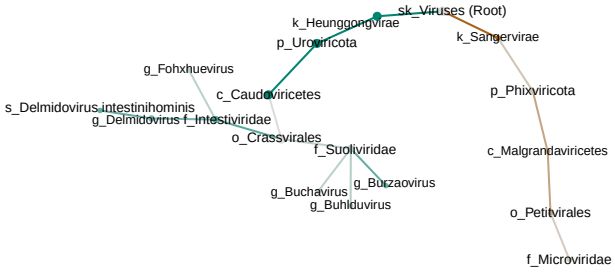
(c) Inflamatory Bowel Disease

(d) Diarrhea

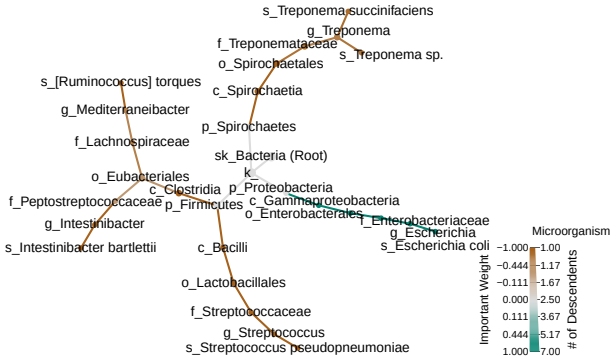
Fig. 5: Representative feature-sparsity patterns at the Class taxonomic rank. These four disease categories were selected because each pools multiple studies with non-trivial total sample sizes and together they illustrate sparse, intermediate, and dense feature-selection regimes. Diabetes and diarrhea show relatively small sets of repeatedly selected taxa, autoimmune disease shows an intermediate pattern, and cancer shows broader support across taxa. The remaining disease categories for the Class taxonomic rank are provided in the Supplementary Information document.



(a) Eukaryota



(b) Viruses



(c) Bacteria

Fig. 6: Lineage analysis of the important features in diarrhea datasets for three distinct taxonomic categories: Eukaryota, Viruses and Bacteria. Each node in the tree corresponds to a taxID. A positive sign (red colored) implies the given taxID is disease associated while a negative sign (green colored) implies the taxID is control associated. The identified lineages include *Escherichia coli* and *Crassvirales* bacteriophages, which have been studied in the context of diarrhea [47–50].

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supp.pdf