BioMedInformatics

MDPI

Article

Bovine Milk Fat Intervention in Early Life and Its Impact on Microbiota, Metabolites and Clinical Phenotype: A Multi-Omics Stacked Regularization Approach

João Pereira [1,2,*,†], Lucas R. F. Bresser [1,2,*,†], Natal van Riel [1,3], Ellen Looijesteijn [4], Ruud Schoemaker [4], Laurien H. Ulfman [4], Prescilla Jeurink [4], Eva Karaglani [5], Yannis Manios [5], Rutger W. W. Brouwer [6,7], Wilfred F. J. van Ijcken [6,7] and Evgeni Levin [1,2,*]

[1] Department of Internal and Vascular Medicine, Amsterdam University Medical Center, 1105 AZ Amsterdam, The Netherlands; n.a.vanriel@amsterdamumc.nl
[2] HorAIzon B.V., 2645 LT Delfgauw, The Netherlands
[3] Department of Biomedical Engineering, Eindhoven University of Technology, 5600 MB Eindhoven, The Netherlands
[4] FrieslandCampina, 3818 LE Amersfoort, The Netherlands; ellen.looijesteijn@frieslandcampina.com (E.L.); ruud.schoemaker@frieslandcampina.com (R.S.); Laurien.Ulfman@frieslandcampina.com (L.H.U.); prescilla.jeurink@frieslandcampina.com (P.J.)
[5] Department of Nutrition and Dietetics, School of Health Science and Education, Harokopio University, GR-176 76 Athens, Greece; evkaraglani@gmail.com (E.K.); yannis.manios@gmail.com (Y.M.)
[6] Department of Cell Biology, Erasmus University Medical Center, 3015 GD Rotterdam, The Netherlands; r.w.w.brouwer@erasmusmc.nl (R.W.W.B.); w.vanijcken@erasmusmc.nl (W.F.J.v.I.)
[7] Center for Biomics, Erasmus University Medical Center, 3015 GD Rotterdam, The Netherlands
* Correspondence: joao.belo@horaizon.nl (J.P.); lucas.bresser@horaizon.nl (L.R.F.B.); evgeni.levin@horaizon.nl (E.L.)
† These authors contributed equally to this work.

check for updates

Citation: Pereira, J.; Bresser, L.R.F.; van Riel, N.; Looijesteijn, E.; Schoemaker, R.; Ulfman, L.H.; Jeurink, P.; Karaglani, E.; Manios, Y.; Brouwer, R.W.W.; et al. Bovine Milk Fat Intervention in Early Life and Its Impact on Microbiota, Metabolites and Clinical Phenotype: A Multi-Omics Stacked Regularization Approach. *Biomedinformatics* **2022**, *2*, 281–296. https://doi.org/10.3390/biomedinformatics2020018

Academic Editor: Jörn Lötsch

Received: 15 April 2022
Accepted: 18 May 2022
Published: 24 May 2022

Abstract: The integration and analysis of multi-omics modalities is an important challenge in bioinformatics and data science in general. A standard approach is to conduct a series of univariate tests to determine the significance for each parameter, but this underestimates the connected nature of biological data and thus increases the number of false-negative errors. To mitigate this issue and to understand how different omics' data domains are jointly affected, we used the Stacked Regularization model with Bayesian optimization over its full parameter space. We applied this approach to a multi-omics data set consisting of microbiota, metabolites and clinical data from two recent clinical studies aimed at detecting the impact of replacing part of the vegetable fat in infant formula with bovine milk fat on healthy term infants. We demonstrate how our model achieves a high discriminative performance, show the advantages of univariate testing and discuss the detected outcome in its biological context.

Keywords: Stacked Regularization; Bayesian optimization; infant formula; multi-omics; microbiota

1. Introduction

Integrating data from different data sources such as genomics, metabolomics, proteomics, images and/or clinical features is an active area of research. Traditionally, a paired *t*-test or a non-parametric counterpart is employed to determine significant differences and to detect relevant biomarkers. However, such approaches overlook the multivariate interaction between different biomarkers. This might lead to a suboptimal detection of biomarkers and an increased number of false-negative errors. Similarly, the use of multivariate but linear techniques such as the popular *mixOmics* [1] do not take into account the full complexity of biological data. Recently, several multi-omics methods have been developed, which can be categorized into three types: concatenation-based, where the data are concatenated for analysis; model-based, where separate models are built for each data set, and a final model is built from the bottom ones; and transformation-based models, where each data set is transformed into a particular format and then joined with the rest to build the model [2,3]. The model-based stacking generalization method is an attractive solution due its simplicity and high empirical performance [4,5]. However, it has two drawbacks: the models' hyperparameters are usually optimized separately, missing the global optimum, and each model only has access to its own domain and thus misses out on the information conveyed by other potentially correlated features/biomarkers. Here, we use a particular model-based strategy, Manifold Mixing for Stacked Regularization (MMSR) [6], that addresses both issues. This method is based on the manifold assumption, the idea that complex high-dimensional data lie on or close to a lower-dimensional manifold. Each data set manifold is then used to mix information across the various modalities, and the transformed data are then fed to a stacked model whose parameters are jointly optimized via Bayesian optimization. We extend the MMSR model for highly curved manifolds by partitioning the space into flat and curved regions while creating region-specific maps between the different manifolds. We then demonstrate its performance on the combined data of two intervention trials. The intervention studies examined the effect of infant formula (IF) containing bovine milk fat on microbiota, metabolomics and clinical parameters in healthy term infants. In both studies, we aim to understand the impact of different IFs with varying sn-2 palmitate content on various physiological and clinical parameters. Specifically, we compare an IF with a mixed-fat blend of 50% vegetable oils and 50% bovine milk fat (MF), containing 39% of sn-2-palmitate, with a standard formula with the same total fat content but originating from a 100% vegetable fat blend (VF), containing 10.1% of sn-2-palmitate. To our knowledge, this is the first IF intervention study analyzed using a multi-omics, multivariate model.

In order to understand the impact of different IFs on the overall biomarker profile, we investigate the model feature importance. Due to the model complexity, we take a model-agnostic approach related to *permutation importance* (PI). However, since PI is biased in the presence of highly correlated data, we use the *pairwise permutation algorithm* (PPA) [7] and discuss the multi-omics profile differences that the model exploits to discriminate between the two groups.

Contributions

- Novel application of the Manifold Mixing for Stacked Regularization framework;
- Extension of MMSR by using curvature-dependent domain partition and non-linear inter-manifold maps;
- Novel application of PPA;
- Exploration of the effect of milk-fat-containing formula on infant gut parameters using a multi-omics multivariate model.

2. Materials and Methods

2.1. Data set

The data set consisted of three sub-sets: microbiota, metabolites and clinical parameters. These data were obtained from 2 different clinical trials [8,9] including a total of 33 infants. In both studies, the infants consumed standard IF with 100% vegetable fat (VF) or IF with 50% milk fat (MF) for two weeks in a cross-over design. In this study, the intervention period was 2 weeks. At the end of each two-week intervention period, fecal samples were collected. The fecal samples were analyzed for microbiota composition and fecal metabolites specifically for the current model. Clinical parameters included anthropometric measurements, amount of formula consumed and data from questionnaires on stool characteristics and gastrointestinal symptoms as well as fecal biochemical profiles. This information was also collected at the end of each intervention period and has been reported before [8,9]. Both studies were approved by the Harokopio University's ethics committee (Athens, Greece). The first trial [8] was conducted in Athens and Larissa (Greece) between December 2017 and July 2018. The second study [9] was conducted between May

2019 and November 2019 in Athens, Greece. The first and second trials were conducted in agreement with the International Conference on Harmonisation guidelines on Good Clinical Practice and registered at the Dutch Trial Register (trialregister.nl) as Trial NL6702 and Trial NL7815, respectively.

2.2. Metabolomics

Lyophilized fecal samples were analyzed for metabolite composition by Biocrates (Innsbruck, Austria) for the first and by Fraunhofer Institute (Hannover, Germany) for the second trial. To extract the metabolites, samples were supplemented with 10 volumes of extraction buffer (85% ethanol in phosphate buffer) and vortexed thoroughly until dissolution. The homogenized samples were placed in a chilled ultrasonic bath for 5 min and centrifuged, and the resulting supernatants were used for the quantification of metabolites with an MxP Quant 500 kit (Biocrates, Innsbruck, Austria). Lipids and hexoses were measured by flow injection analysis-tandem mass spectrometry (FIA-MS/MS) using a Xevo® TQ-S instrument (Waters, Milford, MA, USA) with an electrospray ionization (ESI) source, and small molecules were measured by ultra-performance liquid chromatography-tandem mass spectrometry (UPLC-MS/MS) using the same Xevo® TQ-S instrument. Data were quantified using TargetLynx mass spectrometry software (Waters). In the model, only the data of the small molecules were included.

2.3. Microbiota Composition

DNA was isolated from the lyophilized fecal materials of both studies using a QIAmp fast FNA stool mini kit (Qiagen, Venlo, The Netherlands) and quantified using Quant-it assays (Thermo Fisher Scientific, Waltham, MA, USA). Per sample, 50 ng of DNA was used to generate dual-indexed sequencing libraries using the DNA Flex method (Illumina) with 5 cycles of PCR amplification. The resulting libraries were sequenced on an Illumina HiSeq2500 sequencer (Illumina). Paired-end reads of 300 base-pairs in length were generated using a modified sequencing protocol. Per sample, between 10 and 12 million clusters were generated. Basecalling was performed using BCL2FASTQ2 software (Illumina). Raw reads were checked and quality-filtered using fastp (v.0.20.0) [10]. Here, the adapter was detected and removed; a total of 5 bp in front for read1 was trimmed, and sliding-window quality trimming was applied (with a window width of 4 bp and threshold Q-score of 15). After trimming and adapter removal, reads shorter than 70 bp were removed. Paired-end reads that passed quality filtering were then mapped against the human genome (hg19) using Bowtie 2 (v.2.4.1) [11]. SAMtools (v.1.9) [12], sambamba (v.0.7.1) [13] and BEDtools (v.2.27.1) [14] were used to remove reads that were mapped to the human genome. The remaining high-quality, non-human reads were subsampled to 20 million paired-end reads per sample using seqtk (v.1.3r106) and fed to a humann3 pipeline (v.3.0.0.alpha.3) [15]. For each sample, the species-level microbial composition was inferred using MetaPhlAn3 (v.3.0.2) [15].

2.4. Software

All our models were implemented in Python 3.7, NumPy (version 1.21. 5), and Pandas (version 0.24.2) were used to prepare and manipulate the data set. For the base models, we used the tree-based XGBoost model from the XGBoost package version 1.6.1.

2.5. Stacked Regularization

We used a stacking framework [6], where each domain data set is fed to a different model, and the outputs of these models are subsequently fed into one or multiple layers of models that learn how to optimally combine the base-level predictions (see Figure 1). The goal is to compute $p\left(y|\mathbf{x}^1, \ldots, \mathbf{x}^M\right)$, where $\mathbf{x}^i$ are the coordinates of an instance from $\mathbf{X}$ in domain $\mathcal{X}^i$. The input is passed to a first layer of W_0 predictors $g_1^0(\mathbf{x}), \ldots, g_{W_0}^0(\mathbf{x})$, with:

$$g_i^0(\mathbf{x}) = p\left(y|\mathbf{x}^1, \ldots, \mathbf{x}^M, \boldsymbol{\theta}_i^0\right) , \tag{1}$$

where $\boldsymbol{\theta}_i^0$ are the hyperparameters of the ith model. We pass one data source per model, $g_i^0(\mathbf{x}) = p\left(y|\mathbf{x}^i, \boldsymbol{\theta}_i^0\right)$, so that the width of the first layer, W_0, is equal to the number of domains, M. The output from this layer is then passed to one or more layers of W_k models $g_1^k, \ldots, g_{W_k}^k$, which blend the outputs of the previous ones:

$$g_i^k(\mathbf{x}) = p\left(y|g_1^{k-1}(\cdot), \ldots, g_{W_k}^{k-1}(\cdot)), \boldsymbol{\theta}_i^k\right), k \in [1, L] \ , \tag{2}$$

where L is the total number of blending layers and $\boldsymbol{\theta}_i^k$ the hyperparameters of ith model from the kth layer. The last blending layer is then passed to a final model f that produces the output $f(\mathbf{x}) = p\left(y|g_1^L(\mathbf{x}), \ldots, g_{W_L}^L(\mathbf{x}), \boldsymbol{\theta}^{L+1}\right)$, where $\boldsymbol{\theta}^{L+1}$ are the hyperparameters of f.

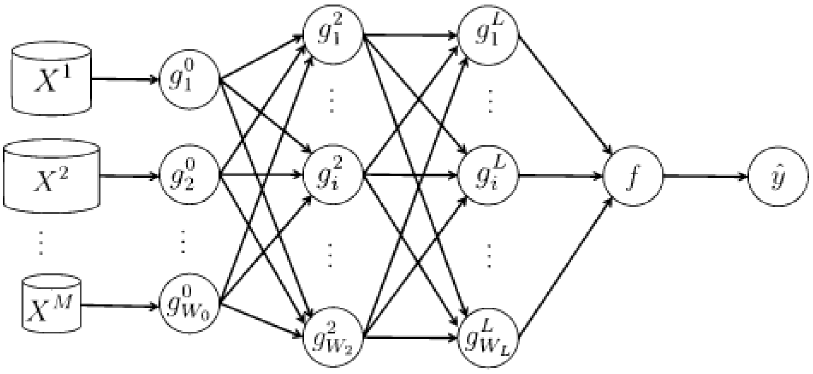


Figure 1. Illustration of the Stacked Regularization model. Each domain data set is fed to a different model, and the predictions are combined by models in higher level layers to produce a single prediction.

2.6. Manifold Mixing for Stacked Regularization

The stacking approach described in the previous section only shares information across different domains via the predictions produced by the base models. Instead, we would like the base models to have access to the other domains' information, since there are likely cross-domain interactions. To achieve this, we first assume that each domain's data lie in a lower-dimensional manifold and create a map between each pair of domains $\mathcal{X}^s \to \mathcal{X}^t$. Next, we use these maps to project the points from the original to the target manifolds to deform the local geometry so that the two become more similar. Consider a set of points S and two mappings taking the points in S to two coordinate systems of domains $\mathcal{X}^t$ and $\mathcal{X}^s$, $\varphi : S \to \mathbb{R}^{|t|}$, $\psi : S \to \mathbb{R}^{|s|}$; suppose subsets $\mathbf{X}^t$, $\mathbf{X}^s$ of data set $\mathbf{X}$ are measured in these coordinate systems. Let us introduce an approximation to the mapping, $\varphi \circ \psi^{-1} : \mathbb{R}^{|s|} \to \mathbb{R}^{|t|}$, from the coordinates of domain $\mathcal{X}^s$ to the coordinates of domain $\mathcal{X}^t$: $\mathbf{L}_s^t = \mathbf{X}^t\mathbf{X}^{s\intercal}(\mathbf{X}^s\mathbf{X}^{s\intercal})^{-1}$ defined as the solution to minimization problem $\min\limits_{\mathbf{L}_s^t} \sum_{i=1}^N ||\mathbf{x}_i^t - \mathbf{L}_s^t\mathbf{x}_i^s||^2$. Denote by $\mathbf{n}_i^t[j]$ the jth neighbor of instance $\mathbf{x}_i$ in domain $\mathcal{X}^t$. Let the array of the points in $\mathcal{X}^s$, which are the neighbors of instance $\mathbf{x}_i^t$ in domain $\mathcal{X}^t$, be $\mathbf{N}_i^{s\leftarrow t} = \left[\mathbf{x}^s_{\mathbf{n}_i^t[1]}, \mathbf{x}^s_{\mathbf{n}_i^t[2]}, \ldots, \mathbf{x}^s_{\mathbf{n}_i^t[k]}\right]$. Our goal is to 'mix' information from different manifolds. This is accomplished by projecting the neighbors of $\mathbf{x}_i^t$ from the source to the target domain and then finding the linear combination of the points that best reconstructs $\mathbf{x}_i^t$ in the original domain:

$$\min_{\mathbf{w}_i} \sum_i ||\mathbf{x}_i^t - \tilde{\mathbf{x}}_i^{t\leftarrow s}||^2 = \min_{\mathbf{w}_i} \sum_i ||\mathbf{x}_i^t - L_s^t\mathbf{N}_i^{s\leftarrow t}\mathbf{w}_i||^2, \tag{3}$$

where $\tilde{\mathbf{x}}_i^{t\leftarrow s}$ is the reconstruction of $\mathbf{x}_i^t$ using domain $\mathcal{X}^s$. We visualize how substituting $\mathbf{x}_i$ with $\tilde{\mathbf{x}}_i^{t\leftarrow s}$ might affect the target manifold in Figure 2.

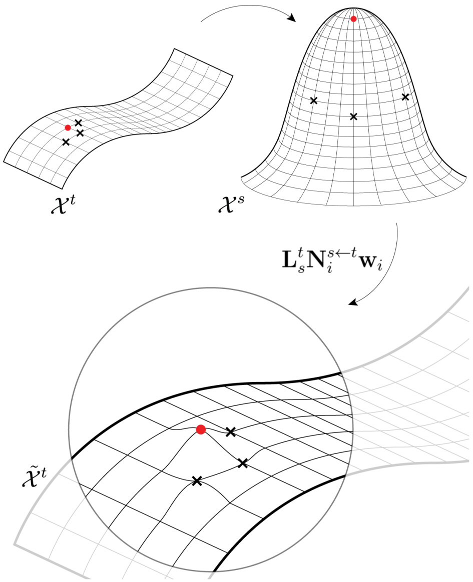


Figure 2. Target manifold being deformed by the source manifold using the Manifold Mixing method. The crosses are the neighbors of point $\mathbf{x}_i$ (point in red) in the target domain. These neighbors are mapped from the source to the target domain and then used to locate $\mathbf{x}_i$. This causes the target manifold to be locally deformed by the source manifold.

After setting the derivative with respect to $\mathbf{w}_i$ to zero, the optimal solution corresponds to:

$$\mathbf{w}_i = \left(\left(\tilde{\mathbf{N}}_i^{t\leftarrow s} \right)^{\intercal} \tilde{\mathbf{N}}_i^{t\leftarrow s} \right)^{-1} \left(\tilde{\mathbf{N}}_i^{t\leftarrow s} \right)^{\intercal} \mathbf{x}_i^t \ , \tag{4}$$

where $\tilde{\mathbf{N}}_i^{t\leftarrow s} = \mathbf{L}_s^t \mathbf{N}^{s\leftarrow t}$, the neighbors of $\mathbf{x}_i$ in $\mathcal{X}_t$ projected from their coordinates in $\mathcal{X}_s$ back to the coordinates in $\mathcal{X}_t$. We can now transform original space $\mathcal{X}^t$ into the space reconstructed from the other domains, $\tilde{\mathcal{X}}^t$, by computing for each instance the weighted mean of its reconstructions:

$$\tilde{\mathbf{x}}_i^t = \beta_t \mathbf{x}_i^t + \sum_{s \neq d} \beta_s \tilde{\mathbf{x}}_i^{t\leftarrow s} \ , \tag{5}$$

where β_j can be seen as the prior of domain $\mathcal{X}^j$'s relevance, and $\sum_j \beta_j = 1$. When evaluating a new point $\mathbf{x}_{new}$, first, the nearest neighbors from the training set are found; then, the reconstruction is given by $\tilde{\mathbf{N}}_i^{s\leftarrow t} \mathbf{w}_{new}$.

Linear map $\mathbf{L}_s^t$ might incur a large error when dealing with highly curved manifolds. In this case, a more accurate solution can be found by partitioning the source domain into disjoint curved subspaces and then constructing a map for each subspace plus one for the flat regions' union. We extend the original method by using a recursive algorithm to partition the space in this way, which is presented in Algorithm 1. The curvature is determined by computing the nearest neighbor matrix first and then finding the difference between the direct second neighbor distance and the sum of the first plus the second neighbor distances. Additionally, the linear map can be substituted for a non-linear map $\phi : \mathbb{R}^{|s|} \rightarrow \mathbb{R}^{|t|}$; thus, $\tilde{\mathbf{N}}_i^{t\leftarrow s} = \phi\left(\mathbf{N}^{s\leftarrow t}\right)$ would be used in Equation (4) instead.

To train the stacked model, we perform Bayesian optimization over the whole model parameter space to search for the best global set of hyperparameters while keeping the computation time relatively low. Decision tree-based XGBoost models are used in each layer; the hyperparameter space is defined by: N_estimator (200,400,800), max_depth (3,5,10), min_samples_split (3,5,10) and max_features (None, sqrt, log2).

Algorithm 1 partition_curved_regions.

```
Input: $\mathcal{NN}$, $q$(queue), $\mathcal{C}$(curvature array), $\mathcal{S}_p$(list of curved subspaces), $visited$, $parent\_partition\_queue$
Return: $\mathcal{S}_p$, $q$
$i \leftarrow q.pop()$
if $i$ not in $visited$ then
    $partitioning \leftarrow$ IsHighCurvature($i$, $\mathcal{C}$)
    $visited.add(i)$
    $neighbors \leftarrow \mathcal{NN}[i]$
    if $partitioning$ then
        $parent\_partitioning \leftarrow parent\_partition\_queue.pop()$
        if $parent\_partitioning$ then
            $\mathcal{S}_p$[-1]$.append(i)$
        else
            $\mathcal{S}_p.append($list$(i))$
        end if
        $\mathcal{S}_p$[-1]$.extend(neighbors)$
    end if
    for $n$ in $neighbors$ do
        if $n$ not in $visited$ and not in $q$ then
            $q.push(n)$, $parent\_partition\_queue.push(partitioning)$
        end if
    end for
    while length($q$)>0 do
        $\mathcal{S}_p$, $q \leftarrow partition\_curved\_regions(\mathcal{NN}, q, \mathcal{C}, \mathcal{S}_p, visited, parent\_partition\_queue)$
    end while
end if
```

2.7. Pairwise Permutation Algorithm

Permutation importance (PI) is a popular algorithm that measures feature importance by comparing model performance before and after randomly permuting the feature values. Whenever there are highly correlated features, the values given by PI are biased, since the model can still rely on the other correlated features. In this work, we use the *pairwise permutation algorithm* (PPA) [7] to account for the correlation bias. The PPA starts by computing the correlation matrix between all the features and then permutes pairs of features whose hierarchically clustered correlation exceeds a pre-defined distance threshold α = 1. The importance value is then given by:

$$PPI_i = \frac{1}{\sum\limits_{j=1}^{M} |\mathbf{R}_{i,j}|} \left(PI_{i,i} + \sum_{\substack{j=1 \\ j \neq i}}^{M} \mathbb{1}\big(|\mathbf{R}_{i,j}| > \alpha\big) |\mathbf{R}_{i,j}| \cdot PI_{i,j} \right), \tag{6}$$

where **R** is the feature hierarchical clustered correlation distance matrix; $\mathbb{1}$ is the indicator function; and $PI_{i,j}$ is the permutation importance value computed by permuting the ith and jth features together.

3. Results

3.1. IF Data Set Model Performance

For each feature in the metabolomics, microbiota and clinical parameters, we computed the value differences between the first and second measurements (cross-over design) and normalized them using zero-mean unit variance standardization. A binary variable indicating whether MF was fed first or second was used as the outcome variable.

In order to evaluate the stacked model performance, we used Leave-One-Out cross-validation combined with subsampling to prevent overfitting and computed the area under the receiver operating characteristic curve (ROC AUC). We tested the model on 10 subsamples with 75% of the data size. Moreover, we performed Bayesian optimization over the whole model's combined parameter space using 20% of the data as validation set. This routine is depicted in Figure 3. To account for errors due to the relatively low sample size, we used a binary tuning parameter indicating whether the mixing described in Section 2.6 coupled with a non-linear map was turned on. In order to control for potential confounders such as age, gender and length/weight, we included them in the clinical feature data set.

The model obtained an ROC AUC of 0.93 ± 0.03, thus achieving high discrimination between both interventions (Figure 4). To determine the statistical significance of the results, we performed a permutation test to evaluate the model performance on bootstrapped data with randomly permuted output values. The model was significant at the $p \leq 0.01$ level. The permutation test score distribution is included in Appendix A (Figure A1).

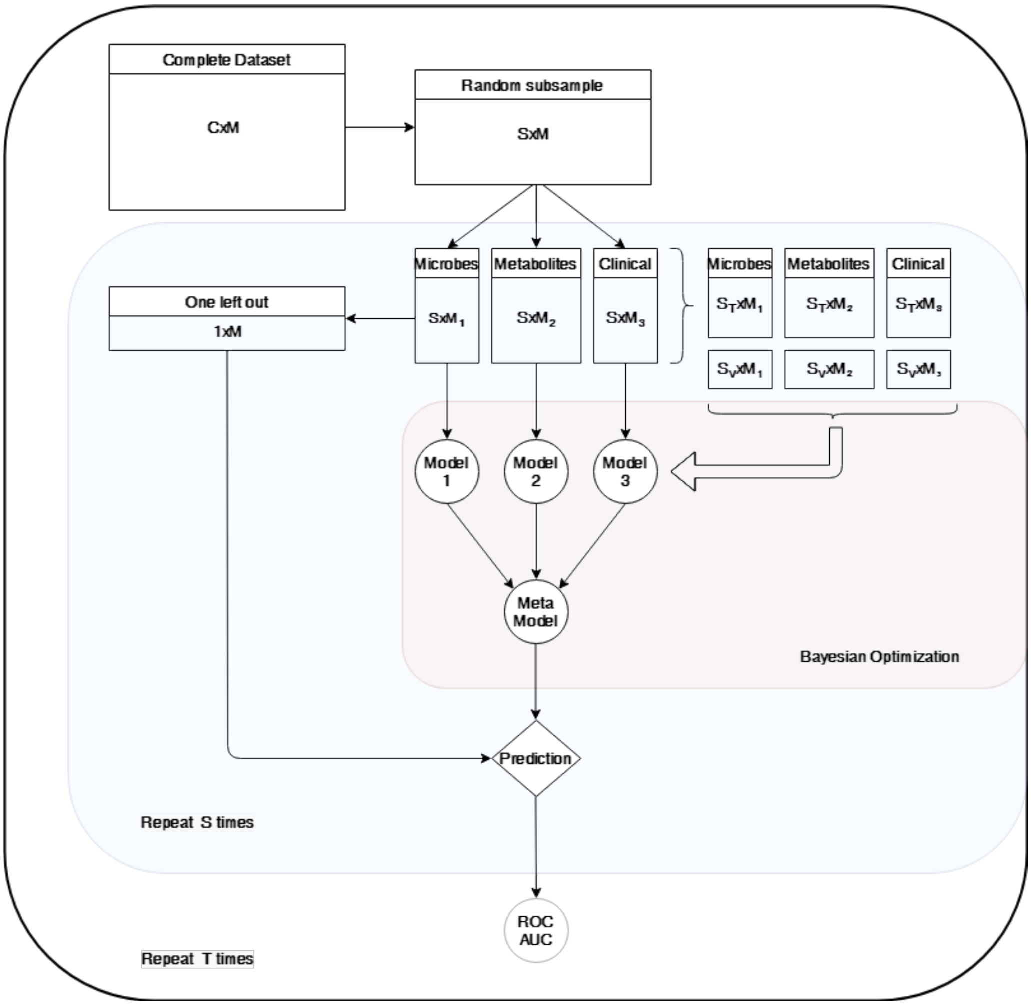


Figure 3. Schematic of the classification routine. The complete data set was subsampled T times with a sample size of S. For each sample, the data set was divided into its domains (microbes, metabolites and clinical), and each was split into the training/validation set over which we performed Bayesian optimization to tune the model. In parallel, one data instance whose label was predicted by the model was left out. This was performed for each of the S data instances in the sample. The ROC AUC curve was then computed for the aggregated predictions, and the score was averaged over T samples.

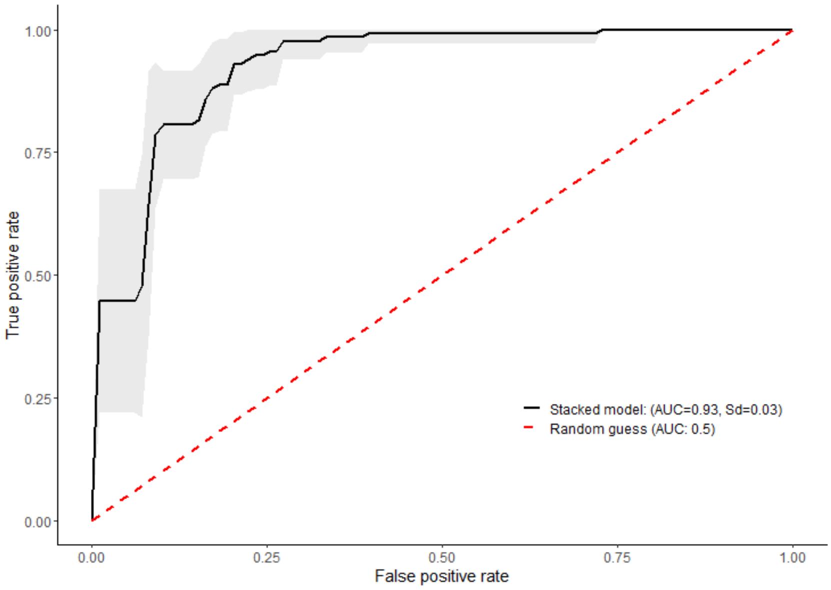


Figure 4. ROC curve for the IF data set treatment prediction. The black line represents the average model performance and the gray area the standard deviation.

3.2. Feature Importance

The goal of our approach is to determine (combinations of) biomarkers that change due to fat sources in IF. Since we are confident in the model pattern recognition ability due to its high performance, we now investigated the most discriminating features between the two groups. We performed a permutation importance test to determine the relative contribution of each feature to the model, first for each modality and then for the overall data set. We then plotted the box plots to check the univariate differences for the top 10 features of each data set. The top 25 feature importance values and the top 12 box plots for each data set are depicted in Figures A2 and A3. As can be seen from Figure A2b,d, the model relies on lauric acid as the most important discriminant feature, with a higher fecal excretion in VF intervention. The biochemical parameters account for most of the discrimination performance, with a higher fecal excretion of total soaps and palmitic acid soap in the VF intervention. We further investigated the feature profile by creating a normalized spider-plot using the top 10 features of each data set (Figure 5A,B). Since the overwhelming difference in lauric acid (C12) between the two groups likely overshadowed the difference in other markers in the combined profile, we recomputed feature importance in the absence of this fatty acid. Excluding C12 highlighted the importance of a.o glutamate and its descendant GABA (Figure 5A,B). We computed the standard permutation importance results in Appendix A (Figure A4). The PPA ranked total and palmitic acid soaps on top, while it ranked stearic acid considerably lower. This suggests that stearic acid was likely highly ranked in the standard permutation importance list because the degree of correlation among the other fatty acid measurements masked its true importance.

Since all measures of weight and length, as well as age and gender, were attributed approximately zero importance, it is safe to assume these were not relevant confounders in this study.

We computed the difference in means p-values for each of the overall top 25 biomarkers using the *t*-test or Wilcoxon test when appropriate. From Figure 6, all of the bacteria except *Veilonella parvula*, as well as most metabolites in the list, had non-significant concentration differences. Using the traditional univariate approach would therefore cause one to miss these biomarkers.

In order to investigate feature interaction, we measured the Spearman correlation coefficient among the top features (Figure 7). The correlation between the microbial species and the top markers in the metabolites/clinical markers might explain why the model was able to pick up some bacteria with no significant differences between the groups, such as

Veilonella parvula and its positive association with the FA soaps. Overall, there was a strong correlation between the different soap measurements, and GABA was also correlated with most of these. Interestingly, *Veilonella parvula* exhibited a very similar pattern of correlation with soaps compared to GABA.

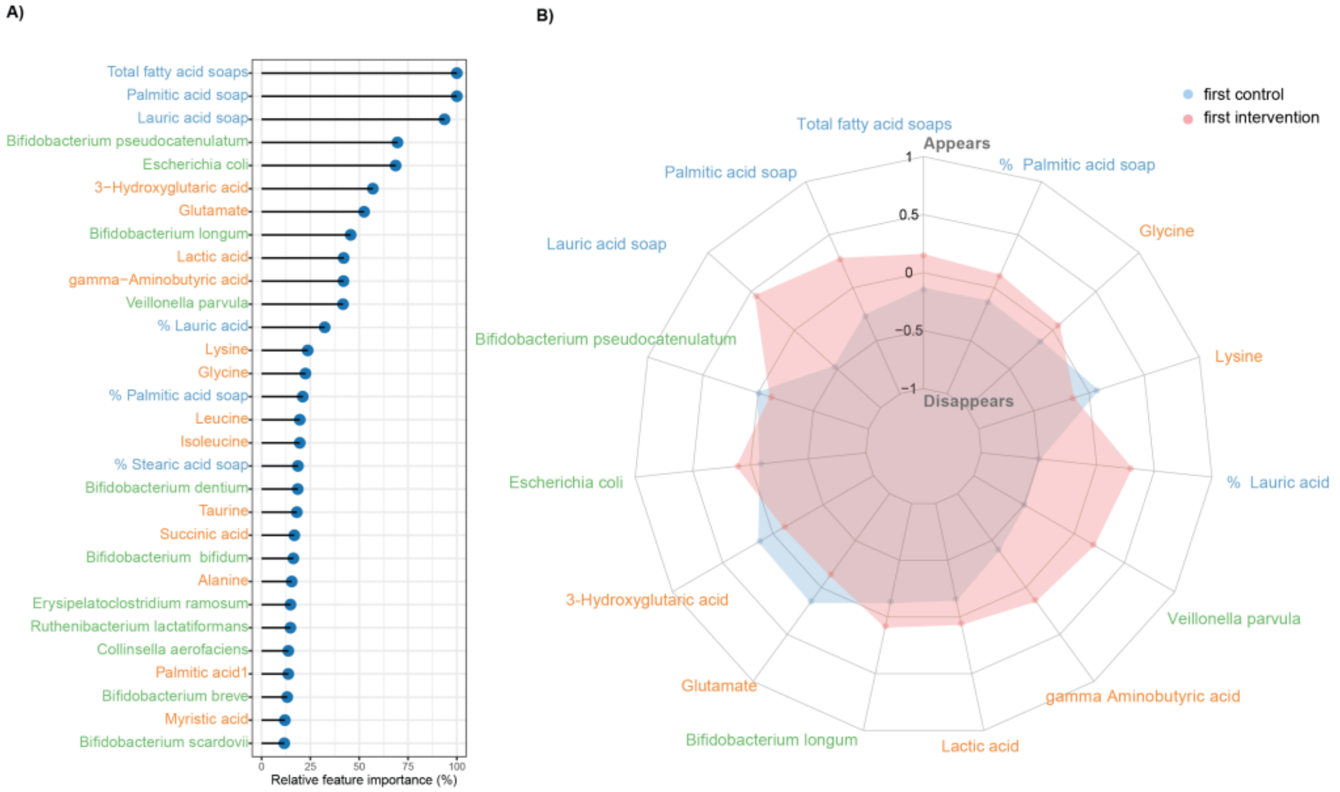


Figure 5. Top feature importance values (**A**) and relative change profile (**B**) of the overall markers after lauric acid (C12) removal.

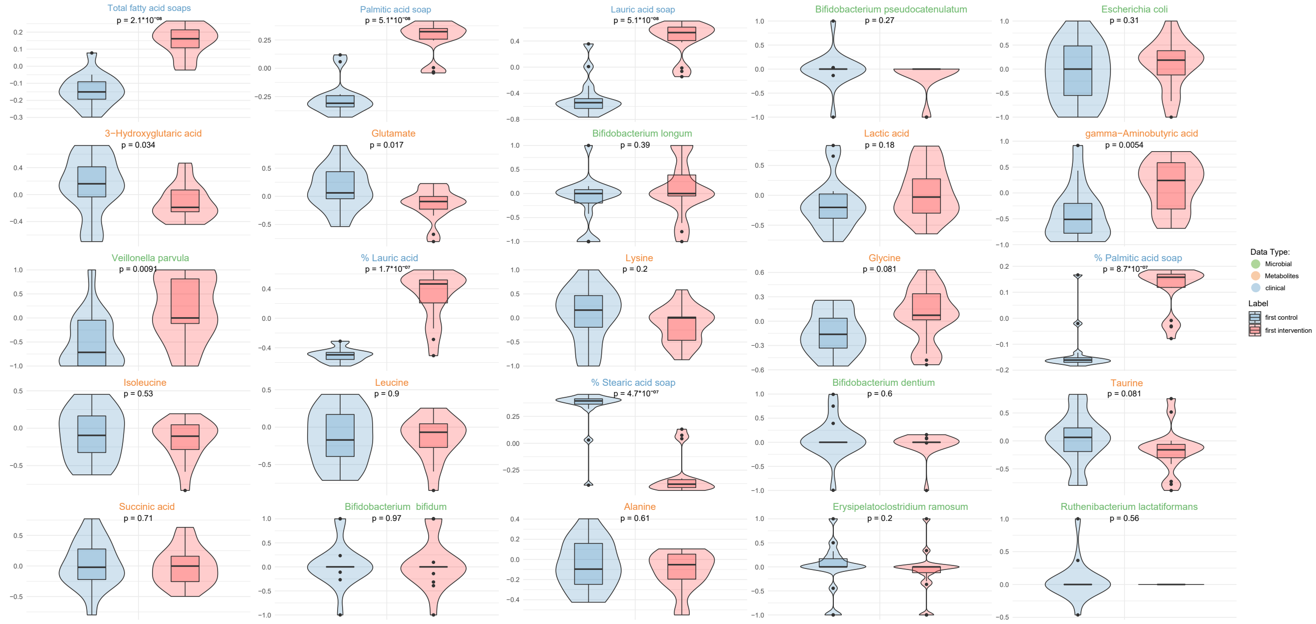


Figure 6. Box plots of the models' top 25 overall biomarkers and their respective univariate difference in means test *p*-value.

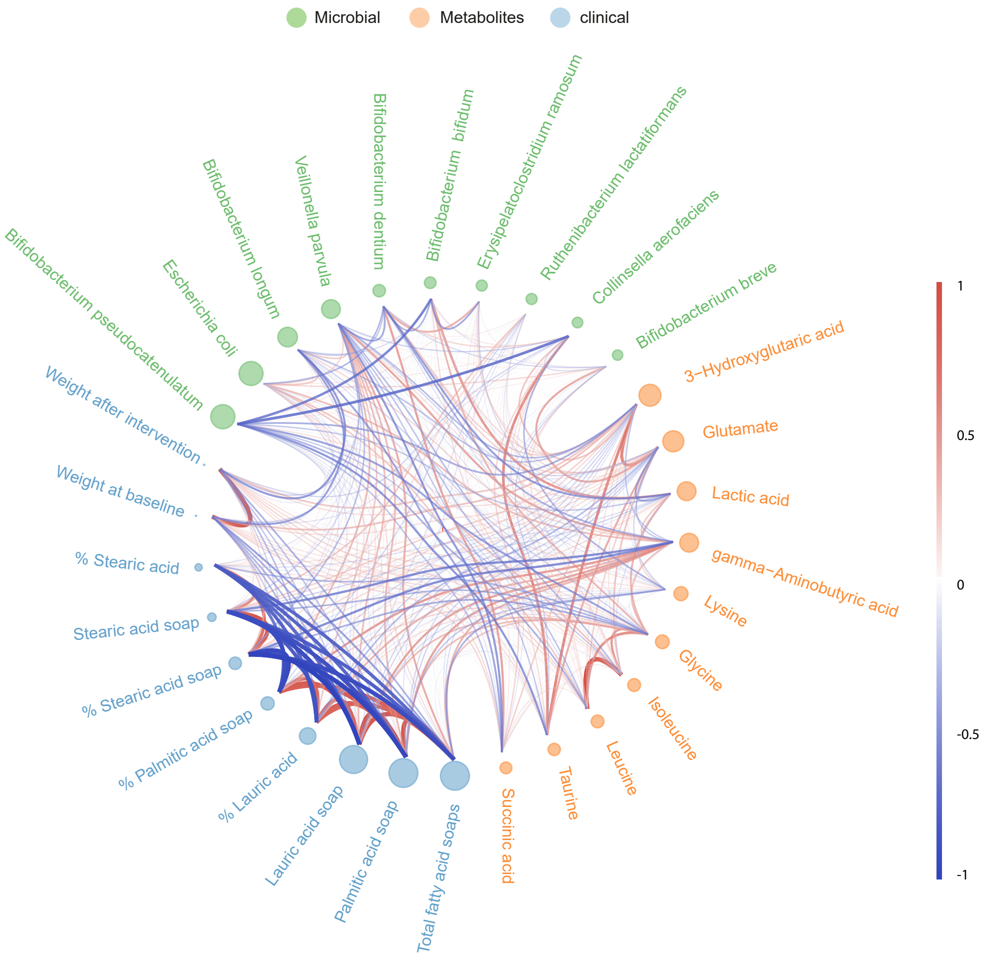


Figure 7. Circular correlation plot between the different domains' top features. Node size is proportional to feature importance, while line thickness/opacity is proportional to correlation between two markers.

4. Discussion

In this study, we demonstrate how a multivariate 'multi-omics' model is capable of identifying relevant biomarkers that do not meet the usual univariate statistical significance. This is evidenced by the high performance of the model and non-zero permutation importance for those non-statistically significant features. Although some of the attributed importance can be explained by the correlation with the markers exhibiting a more pronounced difference, the fact that the model picked up features with little mean differences and no correlations with markers with large differences is evidence that the model is able to capture highly non-linear interactions between the features and output that cannot be accounted by linear nor rank-based correlation alone. This is a major advantage over linear methods such as the popular mixOmics [1].

Regarding biomarker discovery, a major advantage of using a multi-omics approach is the possibility to compare the relative importance of features across different modalities. Although *permutation importance* is one of the most frequently used feature-ranking methods, it is known to be biased in the presence of correlated features. This is especially relevant in the medical/biology domains where the features form an intertwined network of interacting components. The contrast between the highest-ranked features in the PPA and that of standard PI makes it clear that PI underestimates the relevance of members in the fatty acid group due to their high correlation.

The top feature with the most significant difference between the VF and MF interventions was lauric acid (C12). The large difference in fecal lauric acid likely results from differences in its content between the two IFs; lauric acid concentrations were 6 and 10.4

mol% of triglycerides in, respectively, MF and VF [8]. Despite the reported antimicrobial activity of this FA against dominant gut bacteria, including *Bifidobacterium* spp., in the work by Matsue et al. [16], the relative abundances were not significantly lower in the VF intervention for any of the bacteria included in our model. The model is over-reliant on lauric acid due to its large difference, which overshadows patterns arising from other markers. Interestingly, removing it resulted in higher feature importance for the metabolites GABA, glutamate, glycine and taurine. GABA is produced by several microbial species, such as *Bifidobacterium*, *Bacteroides* and *Escherichia* species [17]. The correlation between GABA/Glutamate and several microbes in the top 10 list could mean that changing milk fat content has an impact on the synthesis of GABA by specific microorganisms. GABA and glutamate are inhibitory/excitatory neurotransmitters; thus, their balance in the central nervous system plays an important role in maintaining a stable nervous condition [18]. It is, however, still unclear how this is related to GABA and glutamate concentrations in the gut. Interestingly, *Veillonella parvula* and GABA correlated positively with each other and both correlated similarly with the most important fatty acids and fatty acid soaps. This positive correlation between GABA and *Veillonella parvula* is consistent with results of a recent clinical trial with preterm infants (Russell et al., 2021). Our data further showed a negative correlation between glutamate and GABA, which is likely to be explained by the fact that the decarboxylation of glutamate results in the synthesis of GABA and CO_2. Although associations of *Veillonella* and intestinal GABA levels with brain-related diseases have been suggested [19,20], further studies are required to establish causal relationships and to determine the relevance of small differences in healthy populations.

The contribution of taurine and glycine to discriminate between the interventions might be indicative of differences in (secondary) bile salt metabolism due to the different fat sources used. This is further corroborated by the strong positive correlation between taurine and *Collinsella aerofaciens* that might be related to the known bile salt hydrolase activity of this organism. Thus, some of the discriminatory factors presented in this work suggest a relationship between the nervous and microbial digestive systems. Further studies are required to establish causal relationships among diet, changes in the microbiome and its metabolites, and clinical endpoint. One pitfall of the current methodology is the lack of detail in explaining the relationship between the covariates. Specifically, the impact on the microbiome is likely highly non-linear based on its distribution and model reliance. Unfortunately, the current importance method only provides a holistic explanation of which features play a role in distinguishing the two groups and does not discriminate the output distribution as a function of different covariate values. To achieve this ambitious goal, methods that study the impact on individual infants' biomarker profiles should be used in future studies.

5. Conclusions

Determining the optimal IF can impact infants' overall health and growth at their critical early stage in human development. Modern technology allows a large amount of biomarkers to be collected, providing a more complete picture of the impact different IFs have on infants' metabolism. The usual univariate analysis underscores the importance of the biological context and feature interactions; thus, a multivariate approach is preferred to reduce type II errors. However, this presents a challenge since data from different domains likely have very different distributions which can be difficult to model. We demonstrated how a stacked model optimized over the whole space of parameter combinations achieved very high performance when discriminating infant treatment interventions. This resulted in a general overview of the IF impact, and as expected, fatty acid soaps were the leading factor of discrimination, as was also found using the univariate analysis. Although the difference in microbe concentrations was the smallest of all the feature groups, the combined model selected several bacteria, highlighting a possible impact on the infant microbiome. Presumably, the model was able to pick up the microbial signal due to the correlation with the top markers in the clinical and metabolomics groups, something that would be missed using univariate models.

Author Contributions: Conceptualization, E.L., J.P., L.R.F.B., N.v.R., W.F.J.v.I., L.H.U., R.W.W.B. and E.L. (Ellen Looijesteijn); methodology, L.R.F.B., J.P. and E.L. (Evgeni Levin); investigation, E.L. (Evgeni Levin), W.F.J.v.I., Y.M.; formal analysis, L.R.F.B., J.P. and R.S.; visualization, L.R.F.B., J.P., E.L. (Ellen Looijesteijn), R.S., R.W.W.B. and E.K. writing—original draft, L.R.F.B., J.P., E.L. (Evgeni Levin), E.L. (Ellen Looijesteijn), W.F.J.v.I., R.W.W.B., Y.M. and L.H.U.; writing—review and editing, J.P., L.R.F.B., N.v.R., E.L. (Ellen Looijesteijn), R.S., L.H.U., P.J., E.K., Y.M., R.W.W.B., W.F.J.v.I. and E.L. (Evgeni Levin). All authors have read and agreed to the published version of the manuscript.

Funding: The present study was funded by FrieslandCampina Nederland B.V., HORAIZON BV and in part by an Agri and Food Public-Private Partnership, project TKI-AF18142 entitled 'Combining deep sequencing omics approaches to translate an altered microbial composition towards health benefits'. This project received financial support from Top sector Agri and Food (AF18142). Within this top sector, businesses, knowledge institutes and government work together on innovations for safe and healthy food for 9 billion people in a resilient world.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data are available upon request via lucas.bresser@horaizon.nl.

Acknowledgments: We thank the IT team of HORAIZON BV for support with computational servers and data storage. In addition, we thank Bas Voermans for the collaboration.

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

Appendix A

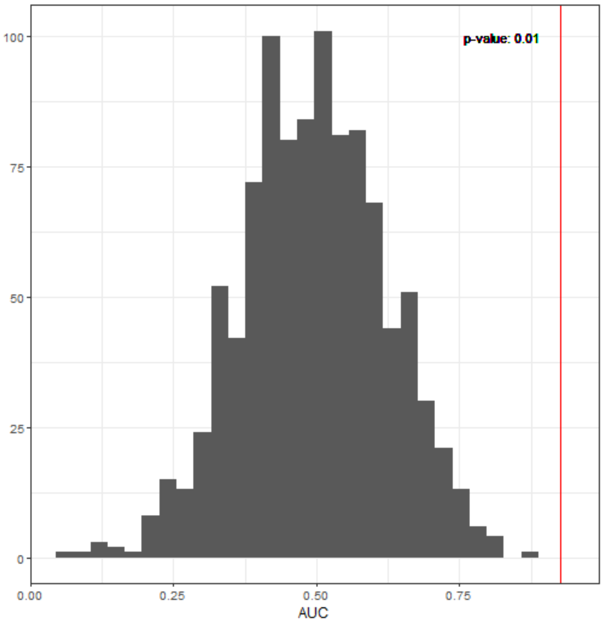
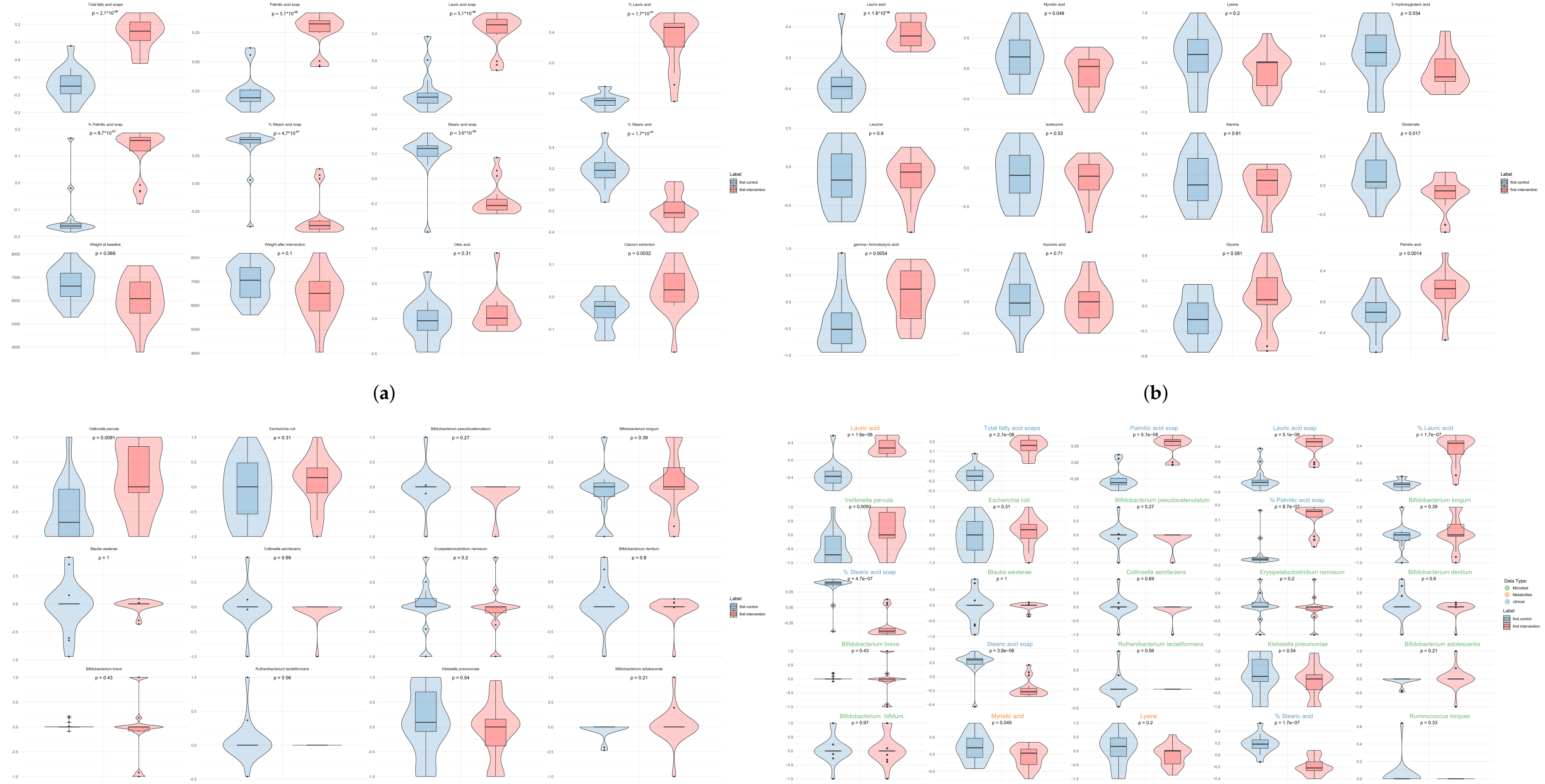


Figure A1. Distribution of the model ROC AUC values when the output values were permuted. The model performance was significant at the 0.05 level.



(**a**) (**b**)

(**c**) (**d**)

Figure A2. Box plots for each domain and the combined. (**a**) Box plots of the top 12 clinical features. (**b**) Box plots of the top 12 metabolite features. (**c**) Box plots of the top 12 microbial features. (**d**) Box plots of the top 12 combined features.

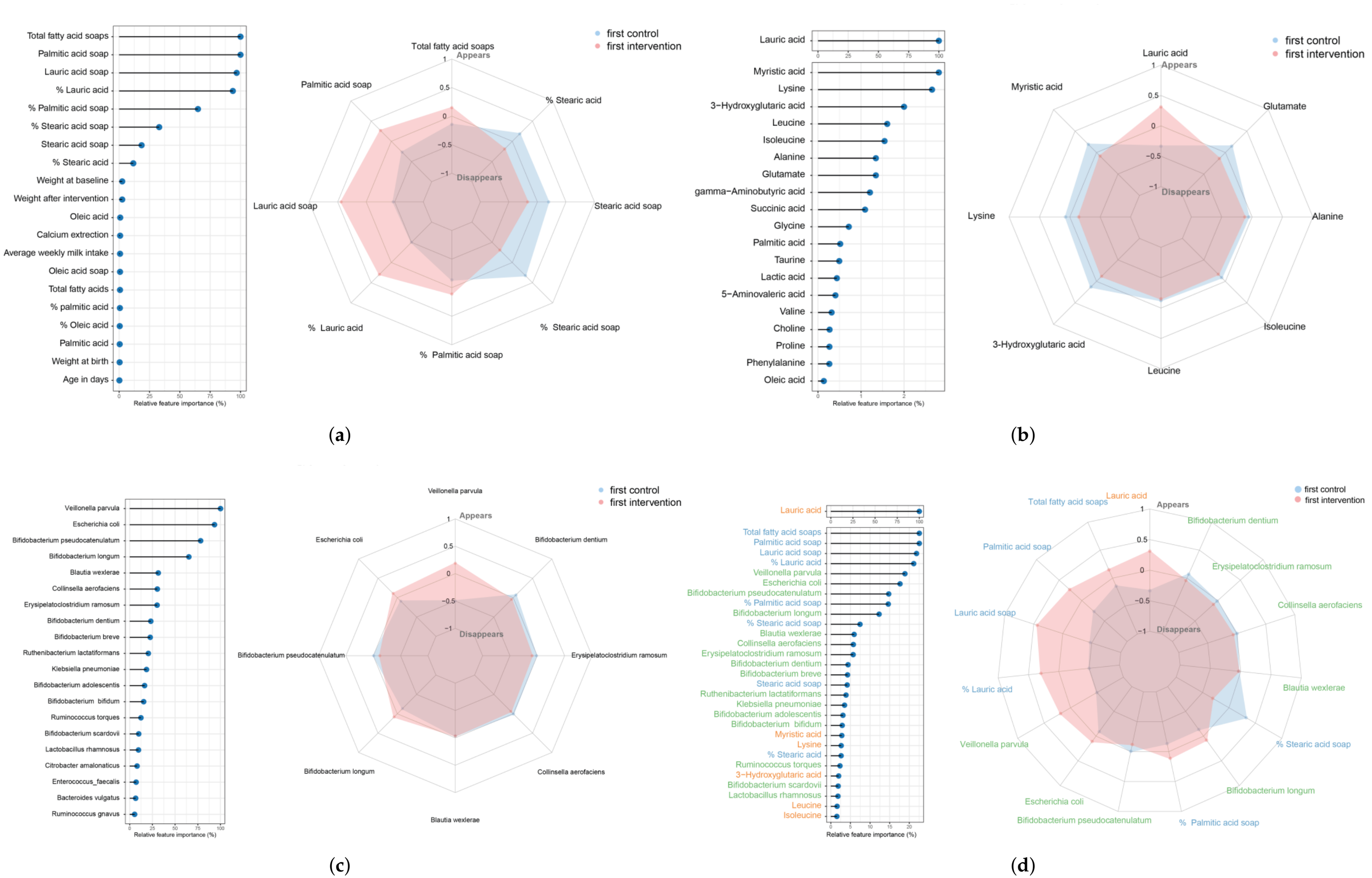


Figure A3. Spider plots of the top 10 combined features. (**a**) Top 25 important clinical features and the respective top 8 radar plot. (**b**) Top 25 important metabolite features and the respective top 8 radar plot. (**c**) Top 25 important microbial features and the respective top 8 radar plot. (**d**) Top 25 important combined features and the respective top 10 radar plot.

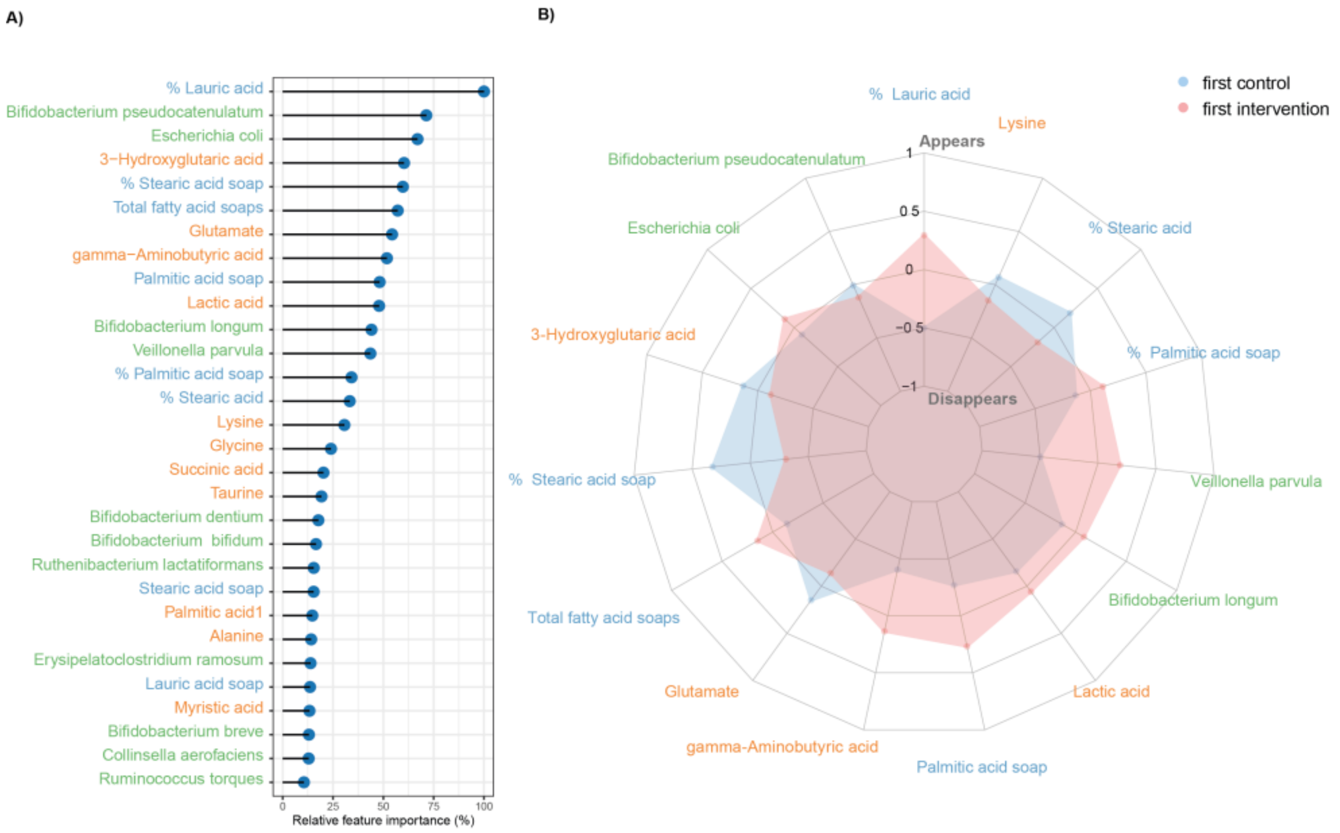


Figure A4. Top importance values (**A**) and relative change profile (**B**) of the overall markers after lauric acid (C12) removal computed using standard permutation importance.

References

1. Rohart, F.; Gautier, B.; Singh, A.; Lê Cao, K.A. mixOmics: An R package for ’omics feature selection and multiple data integration. *PLoS Comput. Biol.* **2017**, *13*, e1005752. [CrossRef] [PubMed]
2. Reel, P.S.; Reel, S.; Pearson, E.; Trucco, E.; Jefferson, E. Using machine learning approaches for multi-omics data analysis: A review. *Biotechnol. Adv.* **2021**, *49*, 107739. [CrossRef] [PubMed]
3. Hasin, Y.; Seldin, M.; Lusis, A. Multi-omics approaches to disease. *Genome Biol.* **2017**, *18*, 83. [CrossRef] [PubMed]
4. Wolpert, D.H. Stacked generalization. *Neural Netw.* **1992**, *5*, 241–259. [CrossRef]
5. Ma, Z.; Wang, P.; Gao, Z.; Wang, R.; Khalighi, K. Ensemble of machine learning algorithms using the stacked generalization approach to estimate the warfarin dose. *PLoS ONE* **2018**. [CrossRef] [PubMed]
6. Pereira, J.; Stroes, E.S.G.; Groen, A.K.; Zwinderman, A.H.; Levin, E. Manifold Mixing for Stacked Regularization. In *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2019*; Communications in Computer and Information Science; Springer: Cham, Switzerland, 2019; Volume 1167, pp. 444–452. [CrossRef]
7. Maaslandand, T.; Pereira, J.; Bastos, D.; de Goffau, M.; Nieuwdorp, M.; Zwinderman, A.H.; Levin, E. Interpretable Models via Pairwise permutations algorithm. *arXiv* **2021**, arXiv:2111.09145v1.
8. Manios, Y.; Karaglani, E.; Thijs-Verhoeven, I.; Vlachopapadopoulou, E.; Papazoglou, A.; Maragoudaki, E.; Manikas, Z.; Kampani, T.M.; Christaki, I.; Vonk, M.M.; et al. Effect of milk fat-based infant formulae on stool fatty acid soaps and calcium excretion in healthy term infants: Two double-blind randomised cross-over trials. *BMC Nutr.* **2020**, *6*, 46. [CrossRef] [PubMed]
9. Looijesteijn, E.; Brouwer, R.W.W.; Schoemaker, R.J.W.; Ulfman, L.H.; Ham, S.L.; Jeurink, P.; Karaglani, E.; IJcken, W.F.J.; Manios, Y. Effect of Bovine Milk Fat-based Infant Formulae on Microbiota, Metabolites and Stool Parameters in Healthy Term Infants in A Randomized, Crossover, Placebo-controlled Trial. *BMC Nutr.* **2022**. [CrossRef]
10. Chen, S.; Zhou, Y.; Chen, Y.; Gu, J. fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **2018**, *34*, i884–i890. [CrossRef] [PubMed]
11. Langmead, B.; Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *BMC Nutr.* **2012**, *9*, 357–359. [CrossRef] [PubMed]
12. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R. 1000 Genome Project Data Processing Subgroup The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079. [CrossRef] [PubMed]
13. Tarasov, A.; Vilella, A.J.; Cuppen, E.; Nijman, I.J.; Prins, P. Sambamba: Fast processing of NGS alignment formats. *Bioinformatics* **2015**, *31*, 2032–2034. [CrossRef] [PubMed]

14. Quinlan, A.R.; Hall, I.M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **2010**, *26*, 841–842. [CrossRef] [PubMed]
15. Beghini, F.; McIver, L.J.; Blanco-Míguez, A.; Dubois, L.; Asnicar, F.; Maharjan, S.; Mailyan, A.; Manghi, P.; Scholz, M.; Thomas, A.M.; et al. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *eLife* **2021**, *10*, e65088. [CrossRef] [PubMed]
16. Matsue, M.; Mori, Y.; Nagase, S.; Sugiyama, Y.; Hirano, R.; Ogai, K.; Ogura, K.; Kurihara, S.; Okamoto, S. Measuring the Antimicrobial Activity of Lauric Acid against Various Bacteria in Human Gut Microbiota Using a New Method. *Cell Transplant.* **2019**, *28*, 1528–1541. [CrossRef] [PubMed]
17. Strandwitz, P.; Kim, K.H.; Terekhova, D.; Liu, J.K.; Sharma, A.; Levering, J.; McDonald, D.; Dietrich, D.; Ramadhar, T.R.; Lekbua, A.; et al. GABA-modulating bacteria of the human gut microbiota. *Nat. Microbiol.* **2019**, *4*, 396–403. [CrossRef] [PubMed]
18. Altaib, H.; Nakamura, K.; Abe, M.; Badr, Y.; Yanase, E.; Nomura, I.; Suzuki, T. Differences in the Concentration of the Fecal Neurotransmitters GABA and Glutamate Are Associated with Microbial Composition among Healthy Human Subjects. *Microorganisms* **2021**, *9, 378*. [CrossRef] [PubMed]
19. Wan, L.; Ge, W.R.; Zhang, S.; Sun, Y.L.; Wang, B.; Yang, G. Case-Control Study of the Effects of Gut Microbiota Composition on Neurotransmitter Metabolic Pathways in Children With Attention Deficit Hyperactivity Disorder. *Front. Neurosci.* **2020**, *14*, 127. [CrossRef] [PubMed]
20. Zheng, P.; Zeng, B.; Liu, M.; Chen, J.; Pan, J.; Han, Y.; Liu, Y.; Cheng, K.; Zhou, C.; Wang, H.; et al. The gut microbiome from patients with schizophrenia modulates the glutamate-glutamine-GABA cycle and schizophrenia-relevant behaviors in mice. *Sci. Adv.* **2019**, *5*, eaau8317. [CrossRef] [PubMed]