Multi-View Integrative Approach For Imputing Short-Chain Fatty Acids and Identifying Key factors predicting Blood SCFA

Anqi Liu[1#], Bo Tian[2#], Chuan Qiu[1], Kuan-Jui Su[1], Lindong Jiang[1], Chen Zhao[3], Meng Song[4], Yong Liu[2], Gang Qu[5], Ziyu Zhou[9], Xiao Zhang[1], Shashank Sajjan Mungasavalli Gnanesh[1], Vivek Thumbigere-Math[7], Zhe Luo[1], Qing Tian[1], Li-Shu Zhang[6], Chong Wu[8], Zhengming Ding[9*], Hui Shen[1*], Hong-Wen Deng[1*]

[#] Co-first authors

[*] Co-corresponding authors

[1]Tulane Center for Biomedical Informatics and Genomics, Deming Department of Medicine, School of Medicine, Tulane University, New Orleans, LA, USA

[2]Center for System Biology, Data Sciences, and Reproductive Health, School of Basic Medical Science, Central South University, Yuelu, Changsha, P.R. China

[3]College of Computing and Software Engineering, Kennesaw State University, GA, USA

[4]College of Science, Xi'an Shiyou University, Xi'an, P.R. China

[5]Department of Biomedical Engineering, Tulane University, New Orleans, Louisiana, USA

[6]School of Physical Science and Engineering, College of Life Sciences and Bioengineering, Beijing Jiaotong University, Beijing, China

[7]Division of Periodontics, University of Maryland Baltimore School of Dentistry, Baltimore, USA

[8]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, USA

[9]School of Science and Engineering, Tulane University, New Orleans, LA, USA

Abstract

Short-chain fatty acids (SCFAs) are the main metabolites produced by bacterial fermentation of dietary fiber within gastrointestinal tract. SCFAs produced by gut microbiotas (GMs) are absorbed by host, reach bloodstream, and are distributed to different organs, thus influencing host physiology. However, due to the limited budget or the poor sensitivity of instruments, most studies on GMs have incomplete blood SCFA data, limiting our understanding of the metabolic processes within the host. To address this gap, we developed an innovative multi-task multi-view integrative approach (M^2AE, Multi-task Multi-View Attentive Encoders), to impute blood SCFA levels using gut metagenomic sequencing (MGS) data, while taking into account the intricate interplay among the gut microbiome, dietary features, and host characteristics, as well as the nuanced nature of SCFA dynamics within the body. Here, each view represents a distinct type of data input (i.e., gut microbiome compositions, dietary features, or host characteristics). Our method jointly explores both view-specific representations and cross-view correlations for effective predictions of SCFAs. We applied M^2AE to two in-house datasets, which both include MGS and blood SCFAs profiles, host characteristics, and dietary features from 964 subjects and 171 subjects, respectively. Results from both of two datasets demonstrated that M^2AE outperforms traditional regression-based and neural-network based approaches in imputing blood SCFAs. Furthermore, a series of gut bacterial species (e.g., *Bacteroides thetaiotaomicron* and *Clostridium asparagiforme*), host characteristics (e.g., race, gender), as well as dietary features (e.g., intake of fruits, pickles) were shown to contribute greatly to imputation of blood SCFAs. These findings demonstrated that GMs, dietary features and host characteristics might contribute to the complex biological processes involved in blood SCFA productions. These might pave the way for a deeper and more nuanced comprehension of how these factors impact human health.

47 **Keywords:** Short-Chain Fatty Acids, Gut Microbiotas, Metagenome, Imputation, Deep learning

Introduction

Short-chain fatty acids (SCFAs) are vital metabolites produced by the bacterial fermentation of dietary fiber in the gastrointestinal tract [1]. In a healthy gut, common bacteria such as Prevotella, Bacteroides, Ruminococcaceae, and Lachnospiraceae [2, 3] generate principal SCFAs, including acetate, propionate, and butyrate. These SCFAs are absorbed and distributed to various organs, influencing host physiology by maintaining an intestinal anaerobic environment and regulating energy metabolism [4-7]. Given the effects of blood SCFAs on host health, understanding the factors that regulate their production is important for the development of strategies to modulate blood SCFA levels to promote health and prevent diseases [8].

Recent studies have shown that fermentative bacteria species such as *Faecalibacterium prausnitzii* and *Eubacterium rectale*, which are abundant in the human gut, could efficiently metabolize complex carbohydrates, particularly resistant starches, into butyrate [9]. Wang *et al.* found that an imbalance of "Good" and "Bad" gut microbiota led to the attenuation of the bacterial metabolite SCFAs [10, 11]. These findings demonstrated that gut microbiotas (GMs) are crucial determinants for blood SCFAs. Dietary features, specifically fibers and macronutrients (i.e., fat, protein, and carbohydrate) intake, are pivotal, as they determine the substrate availability for microbial fermentation in the gut, which subsequently impacts the synthesis of SCFAs [12]. Meanwhile, host characteristics (e.g., age, race) can significantly modulate blood levels of SCFAs by influencing the synthesis, uptake, and utilization of blood SCFAs within the body, possibly through the direct or indirect regulation of metabolic processes and immune responses [13]. For instance, the increase in blood SCFAs may be due to increased uptake of SCFAs in the colon, in part due to increased nutrient intake, a complete bypass of SCFA transporters and increased passive uptake of SCFAs [14]. Despite the growing interest in SCFAs research, the majority of

existing research does not fully account for the complex interplay among host characteristics, dietary features, and GMs, which are crucial for generating accurate results across a wide range of applications [15].

Due to the limited budget or the poor sensitivity of instruments, most of current studies focusing on the gut microbiome are lacking in complete measurements of blood SCFAs [16], which can limit subsequent analyses and conceivably results in the neglect of pivotal insights into metabolic processes within the host. Gut microbiome as measured by metagenomic sequencing (MGS) data can determine SCFA concentrations, influencing host phenotypes by affecting metabolism, immune responses, and energy homeostasis [17, 18]. Integrating SCFA data with MGS data enables multi-omic analyses that reveal broader metabolic impacts and potential links between gut microbiota composition and physiological or disease-related outcomes [19]. Hence, developing a model to impute blood SCFA levels using the MGS data is beneficial and essential for advancing our understanding in this field. To the best of our knowledge, the imputation of blood SCFAs in the host using the MGS data remains largely untapped, marking a significant gap in our comprehension of the dynamics and implications of blood SCFAs production. By developing predictive models that integrate gut microbial compositions with host characteristics and dietary features, we can better understand the complex interplay between these variables and their impacts on blood SCFA production, potentially paving the way for personalized interventions to optimize blood SCFA levels and promote overall well-being [20].

In this study, leveraging metagenomic sequencing technology and deep learning methods, we unveiled an innovative approach that captures the intricate interplays among GMs, dietary features, and host characteristics to impute the absolute abundances of human blood SCFAs. Our method addresses the challenge of incomplete SCFA data by imputing it using MGS data, which will

94 further facilitate the integration of SCFAs and MGS. This integration will enable more
95 comprehensive multi-omic analyses, providing deeper insights into the influence of gut microbial
96 composition on SCFA levels and their subsequent impact on host phenotypes in future research.
97 By applying our approach to two in-house generated datasets, we demonstrated our model
98 outperforms traditional regression-based and neural-network based approaches in imputing blood
99 SCFAs. Accurate imputation of incomplete blood SCFA data will enable researchers to conduct
100 more comprehensive studies exploring metabolic processes and their potential implications for
101 health.

Methods

Subject recruitment and sample collection

A total of 964 unrelated males, aged 20-51 years, were recruited for this study as the first dataset (Dataset 1). An additional 171 unrelated subjects, aged 20-85 years, were recruited for this study, forming the second dataset (Dataset 2). All the subjects were living in New Orleans, Louisiana and its surrounding areas. We excluded subjects who had chronic or recent temporary conditions (e.g., gastroenteritis or inter-continental travel in the past 3 months) that may have significantly disturbed gut microbiota compositions, as described previously [21-27]. Each subject provided stool and blood samples for metagenomic and SCFA profiling, respectively. We used the OMNIgene•GUT (OMR-200) all-in-one system (DNA GenoTEK, Ottawa, CA) for stool sample collection. Stool samples were frozen at -80°C after sample procurement until DNA extraction. Serum (for 964 subjects) or plasma (for 171 subjects) was extracted from 10 ml of blood samples from each subject according to the protein precipitation protocol [28] developed for metabolomics analysis, aliquoted, and stored at -80°C until used for further analysis. The 964 subjects in the first dataset (Dataset 1) also completed three questionnaires—the Louisiana Osteoporosis Questionnaire, the Metagenomic Study Supplementary Questionnaire, and the Food Frequency Questionnaire—to provide relevant covariate information (e.g., demographic factors, lifestyle factors and dietary features). The 171 subjects in the second dataset (Dataset 2) completed only two questionnaires—the Louisiana Osteoporosis Questionnaire and the Metagenomic Study Supplementary Questionnaire—to provide similar covariate information (e.g., demographic factors, but part of lifestyle factors and dietary features). Each subject signed an informed consent, and the study protocols were approved by the Institutional Review Boards (IRBs) of Tulane University. All data were treated with confidentiality, ensuring the anonymity of the participants.

Metagenome profiling

Metagenomic DNA was extracted from stool samples using the Nucleospin Soil kit (MACHEREY-NAGEL) according to manufacturer's instructions, as previously described [19, 29-33]. After a few washes, DNA was eluted with 50 μl elution buffer and stored at -80°C until used for further sequencing. For Dataset 1, 530 samples were sequenced at LC Sciences (Houston, TX), and 434 samples were sequenced at BGI Americas (Cambridge, MA). For Dataset 2, all the 171 subjects were sequenced at LC Sciences.

For the samples sequenced in LC Sciences (Houston, TX), the DNA library was constructed by TruSeq Nano DNA LT Library Preparation Kit (Illumina Inc.). And then we performed the paired-end 2×150 bp sequencing on an Illumina Hiseq 4000 platform at the LC Sciences following the vendor's recommended protocol.

Raw sequencing reads were processed to obtain valid reads for further analysis. First, sequencing adapters were removed from sequencing reads using cutadapt v1.9 [34]. Secondly, low quality reads were trimmed by fqtrim v0.94 [35] using a sliding-window algorithm. Thirdly, reads were aligned to the host genome using bowtie2 v2.2.0 [36] to remove host contamination. Once quality-filtered reads were obtained, they were *de novo* assembled to construct the metagenome for each sample by IDBA-UD v1.1.1 [37]. All coding regions (CDS) of metagenomic contigs were predicted by MetaGeneMark v3.26 [38]. CDS sequences of all samples were clustered by CD-HIT v4.6.1 [39] to obtain unigenes. Unigene abundance for a certain sample were estimated by TPM based on the number of aligned reads by bowtie2 v2.2.0 [36]. The lowest common ancestor taxonomy of unigenes were obtained by aligning them against the NCBI NR database by DIAMOND v 0.9.[40].

For samples sequenced in BGI Americas, the sequencing library was generated using MGI

Easy Universal DNA Library Prep Set Kit (MGI Inc.). The established library was sequenced on BGI DNBSEQ platform using the 100 bp pair-end sample preparation protocol. Quality control (QC) of raw reads was performed using Fastp [41] to filter low-quality reads. The high-quality reads were aligned to the host genome using bowtie2 [36] to remove human reads. The gene profiles were generated by aligning high-quality sequencing reads to the 9.9M integrated gene catalog (IGC) by using the Human Microbiome Project Unified Metabolic Analysis Network (HUMAnN2) [42] .

Serum/Plasma SCFA profiling

Eight SCFAs (acetic acid, propionic acid, isobutyric acid, butyric acid, 2-methylbutyric acid, isovaleric acid, valeric acid and hexanoic acid) in serum/plasma samples were analyzed by Metabolon Inc. (Durham, NC) using LC-MS/MS, as previously described [29-33]. The serum/plasma samples were spiked with stable labelled internal standards and were homogenized and subjected to protein precipitation with an organic solvent. After centrifugation, an aliquot of the supernatant was derivatized. The reaction mixture was injected onto an Agilent 1290/AB Sciex QTrap 5500 LC MS/MS system equipped with a C18 reversed phase UHPLC column. The mass spectrometer was operated in negative mode using electrospray ionization (ESI). The peak area of the individual analyte product ions was measured against the peak area of the product ions of the corresponding internal standards. Quantitation was performed using a weighted linear least squares regression analysis generated from fortified calibration standards prepared immediately prior to each run. LC-MS/MS raw data were collected and processed using SCIEX OS-MQ software v1.7. Three levels of QC samples were prepared by diluting with phosphate-buffered saline (PBS) and/or spiking with stock solutions to obtain the appropriate concentrations for each level (low-concentration QC, medium-concentration QC, and high-concentration QC). Sample analysis was carried out in a 96-

well plate format containing two calibration curves to determine SCFA concentrations and six QC samples (per plate) to monitor assay performance. Accuracy was evaluated using the corresponding QC replicates in the sample runs. QCs met acceptance criteria at all levels for all analytes. QC acceptance criteria are at least 50% of QC samples at each concentration level per analyte should be within ±20.0% of the corresponding historical mean, and at least 2/3 of all QC samples per analyte should fall within ±20.0% of the corresponding historical mean.

While SCFA levels were measured in Datasets 1 and 2 using serum and plasma samples, respectively, SCFA concentrations tend to be highly consistent between serum and plasma samples [43, 44]. This consistency arises because both serum and plasma SCFAs reflect similar metabolic states and distributions in the bloodstream [44]. Thus, validating the model on these two datasets is appropriate, as both serum and plasma SCFA data provide reliable and comparable insights into SCFA dynamics.

Data preprocessing

To remove noise and experimental artifacts in the data and better interpret the results, proper preprocessing for each view data is essential. For gut metagenomic sequencing data, we kept only the GM species that exist in all the subjects for data harmonization across the two data sets. For dietary and clinical data, we filtered out variables with missing rate > 20% and kept all the variables that, to our knowledge [45-48], could be pertinent to the study. Missing data for dietary and host characteristics data were imputed using multiple imputation through R package 'mice' [49]. Missing values in SCFAs were imputed with the minimum of values of all subjects for each SCFA (missing rate <1%). We randomly selected 75% of the samples as the training set and the remaining 25% of the samples in the dataset as the test set. To avoid potential bias, the training data and testing

data have been processed for data normalization separately. We applied log normalization to each type of data, transforming the features to reduce skewness and bring the values into a comparable scale.

Overview of Multi-task Multi-View Attentive Encoders (M^2AE) model

M^2AE is a framework for prediction tasks with multi-view data as input. Each view corresponds to a distinct category of data input, i.e., gut microbiome compositions, dietary features, or host characteristics. The workflow of M^2AE is shown in Fig. 1 and can be summarized into two components. (1) View-specific representation learning via attentive encoders. For each view, an attentive encoder is designed in a symmetric auto-encoder fashion, where the encoder part is composited with one graph convolutional module and two fully-connected layers for view-specific representation learning. (2) Multi-view integration via the View Interactive Network (VIN). A cross-view interactive tensor is calculated using the latent representations from all the view-specific networks. A VIN is then trained with the cross-view discovery tensor to produce the final predictions. VIN can effectively learn the intra-view and inter-view correlations in the higher-level space for better prediction with multi-view data. M^2AE is an end-to-end model, where both view-specific attentive encoders and VIN module are trained jointly. We describe each component in detail in the following sections.

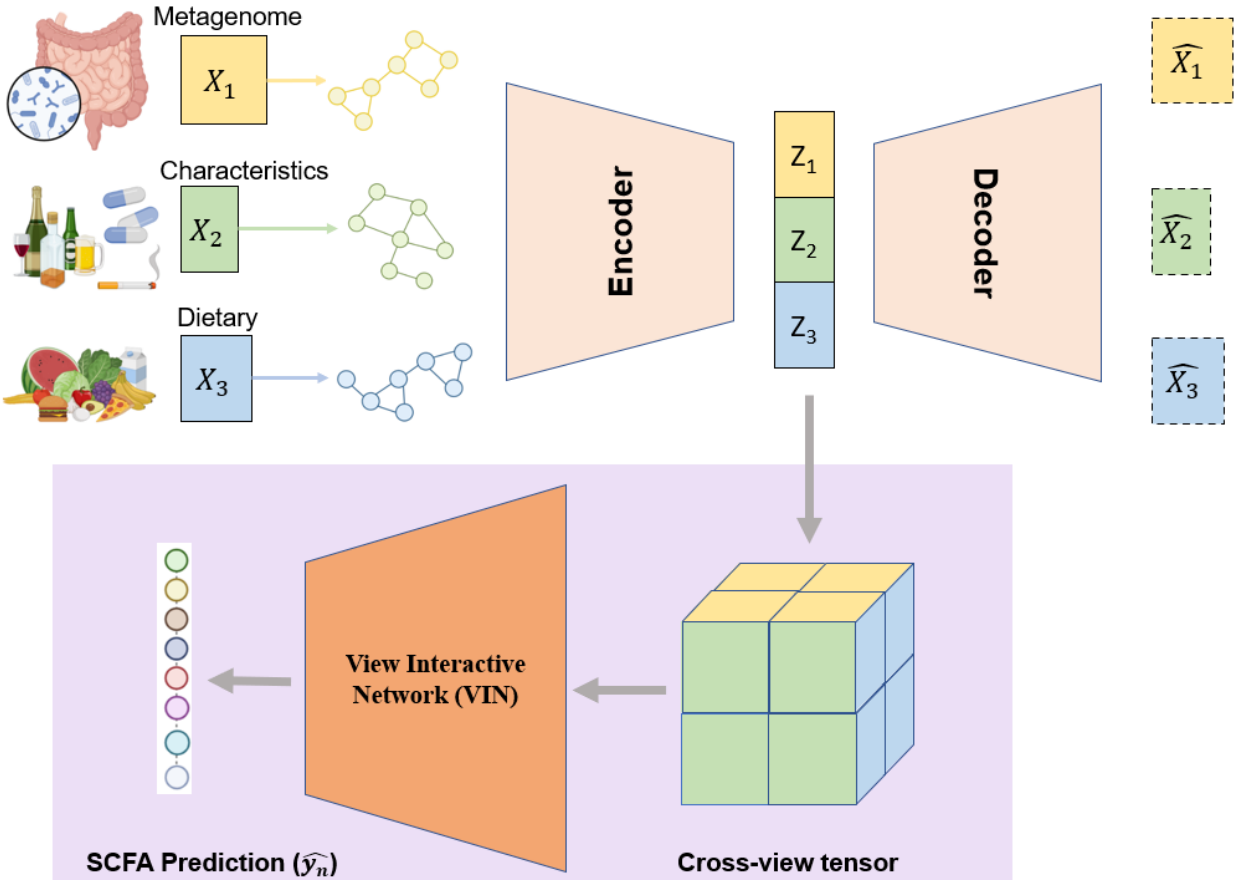


Fig. 1. Overview of M^{2}AE.

M^{2}AE is a framework for prediction tasks with multi-view data as input. The workflow of M^{2}AE can be summarized into two components. (1) View-specific representation learning via attentive encoders. For each view, an attentive encoder was designed in a symmetric auto-encoder fashion, where the encoder part is composed with one graph convolutional module and two fully-connected layers for view-specific representation learning. (2) Multi-view integration via the View Interactive Network (VIN). A cross-view interactive tensor was calculated using the latent representations from all the view-specific networks. A VIN was then trained with the cross-view discovery tensor to produce the final predictions. VIN can effectively learn the intra-view and inter-view correlations in the higher-level space for better prediction with multi-view data. M^{2}AE is an end-to-end model, where both view-specific attentive encoders and VIN module are trained jointly.

Attentive encoders (AEs) for view-specific representation learning

We design each attentive encoder (AE) in an autoencoder manner with one encoder and one symmetric decoder. The encoder contains one graph convolutional module and two fully-connected layers for view-specific representation learning of each type of data input.

The graph convolutional module is implemented to map node features to low-dimensional space and utilizes a simple inner product layer to aggregate the features for feature embedding [50]. By viewing each sample as a node, a view-specific graph can be constructed for each type of view by utilizing both the features (relative microbial abundance/dietary features/host characteristics) of each node and the relationships between nodes. Specifically, in each view, the input sample feature matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ contains the features of all samples, where n is the number of samples and d is the number of features. The input adjacency matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ characterizes the relationships between samples by computing the cosine similarity among pairs of nodes and edges. Thus, the graph convolutional module can be built by stacking multiple convolutional layers with each layer defined as:

$$\mathbf{H}^{(l+1)} = f\left(\mathbf{H}^{(l)}, \mathbf{A}\right) = \sigma(\mathbf{A}\mathbf{H}^{(l)}\mathbf{W}^{(l)}) \quad (1)$$

where $\mathbf{H}^{(l)}$ is the input of the lth layer and $\mathbf{W}^{(l)}$is the weight matrix of the l-th layer and $\mathbf{H}^{(0)} = \mathbf{X}$. $\sigma(\cdot)$ denotes a non-linear activation function. For $\mathbf{A}_{ij}$, it states the adjacency between node i and node j in the graph and is calculated as:

$$\mathbf{A}_{ij} = \begin{cases} s\left(\mathbf{x}_i, \mathbf{x}_j\right), \text{if } i \neq j \text{ and } s\left(\mathbf{x}_i, \mathbf{x}_j\right) \geq \epsilon \\ 0, otherwise \end{cases} \quad (2)$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ are the feature vectors of node i and node j, respectively. $s\left(\mathbf{x}_i, \mathbf{x}_j\right) = \frac{\mathbf{x}_i \cdot \mathbf{x}_j}{\|\mathbf{x}_i\|_2 \|\mathbf{x}_j\|_2}$ is the cosine similarity between node i and node j. The threshold ϵ is determined given a parameter k, which represents the average number of edges per node that are retained including self-connections:

$$k = \sum_{i,j} I(s(\mathbf{x}_i, \mathbf{x}_j) \geq \epsilon)/n \quad (3)$$

where $I(\cdot)$ is the indicator function. The parameter k in Eq. (3) is tuned over [51] $\{2, 5, 10\}$ with the training data, and the same k value is adopted across all experiments on the same dataset. Note that for $k = 1$, $\mathbf{A}$ will turn out to be an identity matrix.

Our graph convolutional module will output a latent feature $\mathbf{F}$ per view, which is then fed into the subsequent two fully-connected layers generating a further latent feature $\mathbf{Z}$ for each view. Thus, the adjacency matrix in the decoder is calculated as $\widetilde{\mathbf{A}} = \text{sigmoid}(\mathbf{Z}\mathbf{Z}^{\mathbf{T}})$ [52], which is sent to the decoder to reconstruct the original input.

For the model training, we aim to minimize the mean absolute error (MAE) between the input feature matrix $\mathbf{X}$ and the reconstructed matrix $\widehat{\mathbf{X}}$ for all views:

$$L_{AE} = \text{MAE}(\mathbf{X}, \widehat{\mathbf{X}}) = \frac{1}{N \times D} \sum_{n=1}^{N} \sum_{d=1}^{D} |x_{nd} - \hat{x}_{nd}| \quad (4)$$

where $\text{MAE}(\cdot)$ represents the mean absolute error function. x_{nd} is the input feature of the nth sample and dth feature, $\hat{x}_{nd}$ is the predicted feature of the nth sample and dth feature.

So far, we have learned the view-specific representation $\mathbf{Z}$ and we will introduce to fuse each view for the final prediction task in the following section.

VIN for multi-view integration

Current approaches leveraging multi-view data for biomedical prediction tasks traditionally either concatenate features from disparate views directly or fuse these features within a low-level feature space [53-56]. However, properly aligning multiple views remains a consistent challenge, as improper alignment can have detrimental effects. On the other hand, view correlation discovery network (VCDN) [57] can exploit the higher-level cross-view correlations in class label level, as different types of data can provide unique distinctiveness for the production of SCFAs. Inspired by this, we develop VIN, which consolidates three latent features from gut microbiome, dietary

features, and host characteristics to learn higher-level intra-view and cross-view correlations, thereby improving SCFA predictions.

For the latent representations of the nth sample from three types of views $\hat{z}_n^{(m)}$, $m = 1, 2, 3$, we construct a cross-view interactive tensor $\mathbf{C}_n$, where each entry of $\mathbf{C}_n$ is calculated as:

$$C_{n,a1a2a3} = \hat{z}_{n,a1}^{(1)} \hat{z}_{n,a2}^{(2)} \hat{z}_{n,a3}^{(3)} \tag{5}$$

where $\hat{z}_{n,a}^{(m)}$ denotes the ath entry of $\hat{z}_n^{(m)}$. Then, the obtained tensor $\mathbf{C}_n$ is reshaped to a c^3 dimensional vector $\mathbf{c}_n$ and is forwarded to the final prediction. VIN$(\cdot)$ is designed as a network with one graph convolutional layer and one fully-connected layer with the output dimension of c (In this case, we have eight SCFAs as outputs, so we set c = 8). We aim to minimize the mean absolute error between the predicted and ground-truth SCFAs as:

$$L_{\text{VIN}} = \sum_{n=1}^{N} \text{MAE}(\text{VIN}(\mathbf{c}_n), \mathbf{y}_n) \tag{6}$$

where $\mathbf{y}_n$ represents the absolute abundances of eight SCFAs in the nth sample, N represents the sample size.

To this end, VIN$(\cdot)$ could reveal the latent intra-view and cross-view correlations and help to improve the learning performance. By utilizing VIN$(\cdot)$ to integrate latent representations from different types of views, the final prediction made by M^{2}AE is based on both the latent representation from each view and the learned cross-view correlation knowledge.

Overall, we optimize our M^{2}AE by minimizing the attentive encoder loss and view interactive network losses in an iterative manner. During one epoch of the training process, we first fix VIN$(\cdot)$ and update $AE_m(\cdot)$, $m = 1, 2, 3$, for each type of view to minimize the loss function L_{AE}^m. Then we fix the view-specific AEs and update VIN$(\cdot)$ to minimize L_{VIN}. View-specific AEs and VIN are updated alternately until convergence.

Model performance evaluation

To evaluate the model's performance in imputing blood SCFAs, we computed the mean absolute errors (MAE) and root mean squared errors (RMSE) for each subject. The average MAE and RMSE were then calculated by averaging these metrics across all subjects. We evaluated the models on five different randomly generated training and test splits, and the mean and standard deviation of the evaluation metrics across these five experiments were computed.

Identification of influential factors for blood SCFAs

To identify significant factors for SCFAs, we defined a feature contribution score for each SCFA p across three different views as:

$$f_{d\to p}^{(m)} = \frac{1}{N} \cdot \sum_{n=1}^{N} \left| g_{d\to p}^{(m)} \right| \quad (7)$$

where $f_{d\to p}^{(m)}$ denotes the contribution score of the feature d to the SCFA p in view m and $g_{d\to p}^{(m)}$ denotes the gradient of the SCFA p with respect to the input feature d in view m. Using this approach, we analyzed the contribution of each feature in different types of views on the test set. Features with the largest contribution scores in each view were considered to be the most important ones. Considering the inherent variability during training, we executed five repeated experiments in one dataset and reported the results by summing up the feature contribution scores across these five repeated experiments.

KEGG pathway analyses were conducted to identify significant biological pathways enriched in prominent bacterial species associated with SCFAs, by searching on the website of Kyoto Encyclopedia of Genes and Genomes (https://www.genome.jp/kegg/).

Results

Datasets

To validate the proposed M^{2}AE model, we applied it to two different in-house datasets. We adopted the same data preprocessing pipeline described in the Methods section. For a fair comparison with existing approaches, we used the same methodology to construct the training and testing sets for evaluation.

Dataset 1 consists of data from 964 unrelated males who provided both stool and blood samples for metagenomic and serum SCFAs profiling, along with dietary features, and host characteristics data. The basic characteristics of the samples are shown in Table 1. Features used to predict serum SCFAs include 194 gut bacterial species (relative abundance), 33 dietary features (e.g., intake of fruits, vegetables) and 17 host characteristics (e.g., age, race). The host characteristics and dietary habits used in Dataset 1 are listed in Supplementary Table 1.

Table 1. Sample Characteristics for Dataset 1 (964 subjects).

	Caucasian	African American	Total
N (%)	577 (59.85%)	387 (40.15%)	964 (100%)
Age: Mean (range)	35.84 (20-51)	39.20 (20-51)	37.19 (20-51)
Height (cm): Mean (SD)	175.27 (6.90)	174.94 (6.97)	175.14 (6.93)
Weight (kg): Mean (SD)	82.97 (15.83)	82.85 (17.46)	82.92 (16.52)
BMI (kg/m^2): Mean (SD)	27.02 (5.07)	27.04 (5.31)	27.03 (5.18)
Regular exercise: n (%)	459 (79.55%)	256 (66.15%)	715 (74.17%)
Smoking: n (%)	391 (67.76%)	292 (75.45%)	683 (70.85%)
Alcohol drinking: n (%)	427 (74.00%)	210 (54.26%)	637 (66.08%)

Dataset 2 includes data from 171 unrelated subjects who provided both stool and blood

samples for metagenomic profiling and plasma SCFAs profiling, along with dietary features and host characteristics data. The basic characteristics of these samples are presented in Table 2. Features used to predict plasma SCFAs include 646 gut bacterial species (relative abundance), 3 dietary features (e.g., intake of milk and yogurt), and 11 host characteristics (e.g., age, gender, and race). The host characteristics and dietary habits used in Dataset 2 are listed in Supplementary Table 2.

Table 2. Sample Characteristics for Dataset 2 (171 subjects).

	Male (58 (33.92%))		**Female** (113 (66.08%))		
	Caucasian	**African American**	**Caucasian**	**African American**	**Total**
N (%)	19 (77.19%)	39 (22.81%)	77 (78.95%)	36 (21.05%)	171 (100%)
Age: Mean (range)	58.16 (44-76)	58.13 (51-72)	41.66 (20-85)	40.00 (21-69)	46.90 (20-85)
Height (cm): Mean (SD)	173.32 (8.07)	172.93 (7.57)	161.80 (6.96)	165.52 (6.33)	166.40 (8.63)
Weight (kg): Mean (SD)	85.41 (14.12)	83.72 (18.85)	66.10 (14.53)	77.34 (17.30)	74.63 (17.97)
BMI (kg/m^2): Mean (SD)	28.53 (4.93)	27.81 (4.90)	25.24 (5.27)	28.20 (6.02)	26.82 (5.47)
Regular exercise: n (%)	29 (83.04%)	29 (16.96%)	88 (85.38%)	25 (14.62%)	124 (72.51%)
Smoking: n (%)	35 (86.55%)	23 (13.45%)	103 (94.15%)	10 (5.85%)	75 (43.86%)
Alcohol drinking: n (%)	41 (90.06%)	17 (9.94%)	98 (91.23%)	15 (8.77%)	110 (64.33%)

M^2AE outperformed existing multi-view integration prediction methods

As shown in Table 3, we compared the prediction performance of M^2AE with the following existing regression algorithms for our data: (1) K-nearest neighbor regression (KNN), (2) Random forest regression (RF), (3) Gradient boosting-based regression (XGBoost), (4) Fully-connected neural network (NN) regression and (5) Linear regression. Deep fully-connected NN were also trained with MAE loss. Among the compared methods, KNN, RF, XGBoost, and NN were trained with the direct concatenation of the processed multi-view data as input. All methods were trained with the same processed data. The average MAE and average RMSE across all subjects were computed to compare the performance of different models. The choices for each hyper-parameter are relegated to Supplementary Table 3.

Table 3. Performance Comparison of different models for Dataset 1 and Dataset 2.

	Dataset 1		Dataset 2	
Methods	**Mean MAE**	**Mean RMSE**	**Total MAE**	**Total RMSE**
Linear Regression	0.641 ± 0.016	0.813 ± 0.020	0.495 ± 0.016	0.620 ± 0.002
RF	0.481 ± 0.012	0.619 ± 0.014	0.393 ± 0.018	0.494 ± 0.004
NN	0.951 ± 0.100	1.170 ± 0.121	2.232 ± 0.160	2.673 ± 0.062
KNN	0.518 ± 0.011	0.662 ± 0.014	0.423 ± 0.021	0.530 ± 0.005
XGBoost	0.529 ± 0.013	0.679 ± 0.013	0.447 ± 0.021	0.569 ± 0.005
AE(MLP)	0.458 ± 0.007	0.589 ± 0.010	0.392 ± 0.021	0.502 ± 0.004
AE(GCN)	0.467 ± 0.006	0.601 ± 0.008	0.385 ± 0.018	0.494 ± 0.004
AE(2GCN+1MLP)	0.470 ± 0.015	0.602 ± 0.015	0.388 ± 0.021	0.498 ± 0.003
VIN(MLP)	0.458 ± 0.011	0.590 ± 0.015	0.384 ± 0.019	0.493 ± 0.003
VIN(GCN)	0.597 ± 0.029	0.731 ± 0.032	0.384 ± 0.019	0.493 ± 0.003
M^{2}AE	**0.449 ± 0.010**	**0.581 ± 0.012**	**0.382 ± 0.017**	**0.489 ± 0.030**

Note: Bold represents the best performance in different criteria.

We observed that M^{2}AE outperformed the other methods in the prediction tasks by showing the smallest mean MAE and mean RMSE in both Dataset 1 and Dataset 2 (Table 3), indicating the superior learning capability of M^{2}AE. Interestingly, although deep learning-based methods have shown great promises in regression applications, the deep learning-based method NN did not show clear improvements over other approaches. This observation suggested that proper design of deep learning algorithms specific to multi-view integration applications was to some degree required to achieve superior prediction performance.

M^2AE outperformed its variations and other methods in SCFA prediction tasks

M^2AE integrates view-specific learning via AEs with cross-view interactive fusion via VIN for effective SCFA predictions. To examine the necessity of AEs and VIN for effective SCFA predictions, we performed extensive ablation studies of our proposed method where two additional variations of M^2AE were compared. (1) AutoEncoder_VIN: a) Fully-connected NNs with the same number of layers and the same dimensions of hidden layers as the encoder part in M^2AE were used for view-specific representation learning; b) GCNs with the same number of layers and the same dimensions of hidden layers as the encoder part in M^2AE were used for view-specific representation learning; c) An AE containing two graph convolutional layers and one fully-connected layer with the same dimensions of hidden layers as the encoder part in M^2AE were used for view-specific representation learning. The multi-view integration component utilized VIN, which was the same as M^2AE. (2) AE_NN/GCN: the view-specific representation component utilized one graph convolutional layer and two fully-connected layers, which was the same as M^2AE. a) A fully-connected NN with the same number of layers and the same dimensions of hidden layers as VIN was used for multi-view integration; b) A GCN with the same number of layers and the same dimensions of hidden layers as VIN was used for multi-view integration. Note that Autoencoder_VIN itself is also a novel approach. To the best of our knowledge, there is no existing method that applies AEs to multi-view data integration and imputation problems.

As shown in Table 3, we observed that M^2AE outperformed Autoencoder_VIN and AE_NN/GCN in all prediction tasks across both Datasets 1 and 2. The better performance of average MAE and average RMSE in M^2AE than AE_NN/GCN indicates that our usage of VIN combines one graph convolutional layer and one fully-connected layer for multi-view integration and prediction tasks makes important contributions to the performance boost of M^2AE comparing

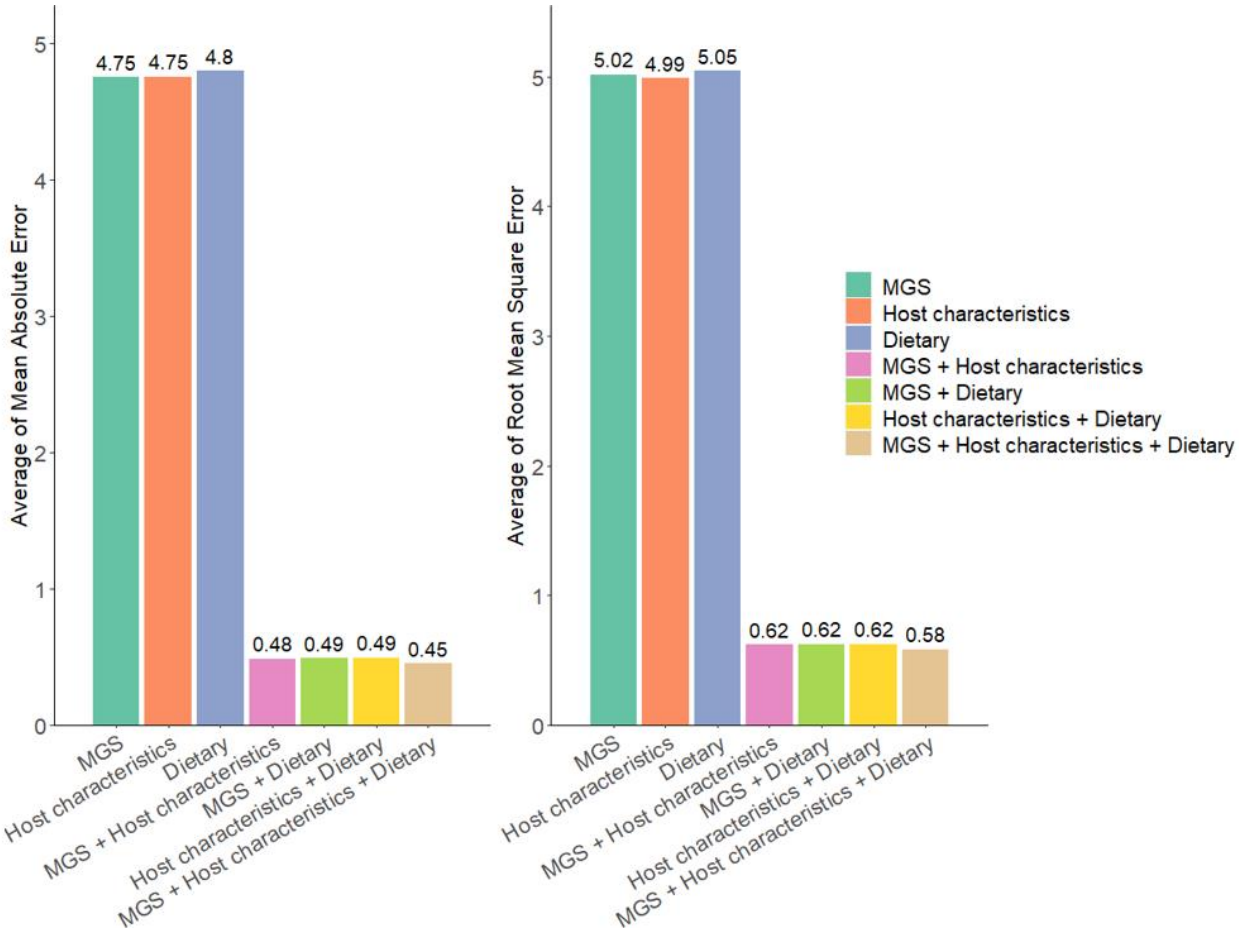
with existing methods. Compared with traditional NN and GCN that only learn from view information from one pathway, the AE further exploits the graph structural information within the data using a more flexible way. This can be essential to a more comprehensive understanding of the type of view as it captures the connections and correlations among samples. Therefore, AEs were needed for effective view-specific representation learning to fully exploit the advantages of VIN, and these two components could be trained jointly to achieve superior results for multi-view prediction tasks across both datasets.

Performance of M^{2}AE under different types of views

To further demonstrate the necessity of integrating multiple types of data to boost the prediction performance, we compared the prediction performance of M^{2}AE with three types of views (MGS + host characteristics + dietary features), M^{2}AE with two types of views (MGS + host characteristics, MGS + dietary features, and host characteristics + dietary features), and the view-specific AEs trained with single-view data before integration (MGS only, host characteristics only, and dietary features only). Figs. 2 and 3 show that by exploring the cross-view interactive fusion through VIN, the prediction performance was consistently improved by integrating prediction results from multiple views. Specifically, in all the prediction tasks, M^{2}AE models trained with three types of views achieved the best performance compared with M^{2}AE models trained with two types of views. Moreover, the M^{2}AE models trained with two types of views both outperformed the single-view AE models.

It is well known that gut microbiome is closely influenced by host characteristics and dietary habits, as noted in previous studies [12, 13]. This interdependence suggests that these factors, when considered together, may collectively enhance the predictive power of models. The results,

therefore, strongly support the necessity of integrating multiple types of data in predictive models. By leveraging cross-view interactions and fusing diverse data types, we captured a more comprehensive understanding of the complex interplay between microbiome, host, and environmental factors, leading to significantly improved prediction accuracy. This multi-view approach is crucial for advancing personalized medicine and for a more profound understanding of complex biological phenomena.



74

75 **Fig. 2.** Performance comparison of multi-view data prediction via M^{2}AE and single-view data

76 prediction via Attentive Encoders in Dataset 1 (n = 5 experiments for each model).

77 Means of evaluation metrics from different experiments are shown in the figure. MGS + Host
78 characteristics + Dietary refers to M^{2}AE with three types of views combining MGS, host
79 characteristics, and dietary features data. MGS + Host characteristics refers to M^{2}AE with two
80 types of views combining gut microbiomes and host characteristics. MGS + Dietary refers to
81 M^{2}AE with two types of views combining gut microbiomes and dietary features. Host
82 characteristics + Dietary refers to M^{2}AE with two types of views combining host characteristics
83 and dietary features. MGS, host characteristics and dietary features refer to the view-specific
84 attentive encoders trained with single-view gut microbiomes, host characteristics, and dietary
85 features.

86

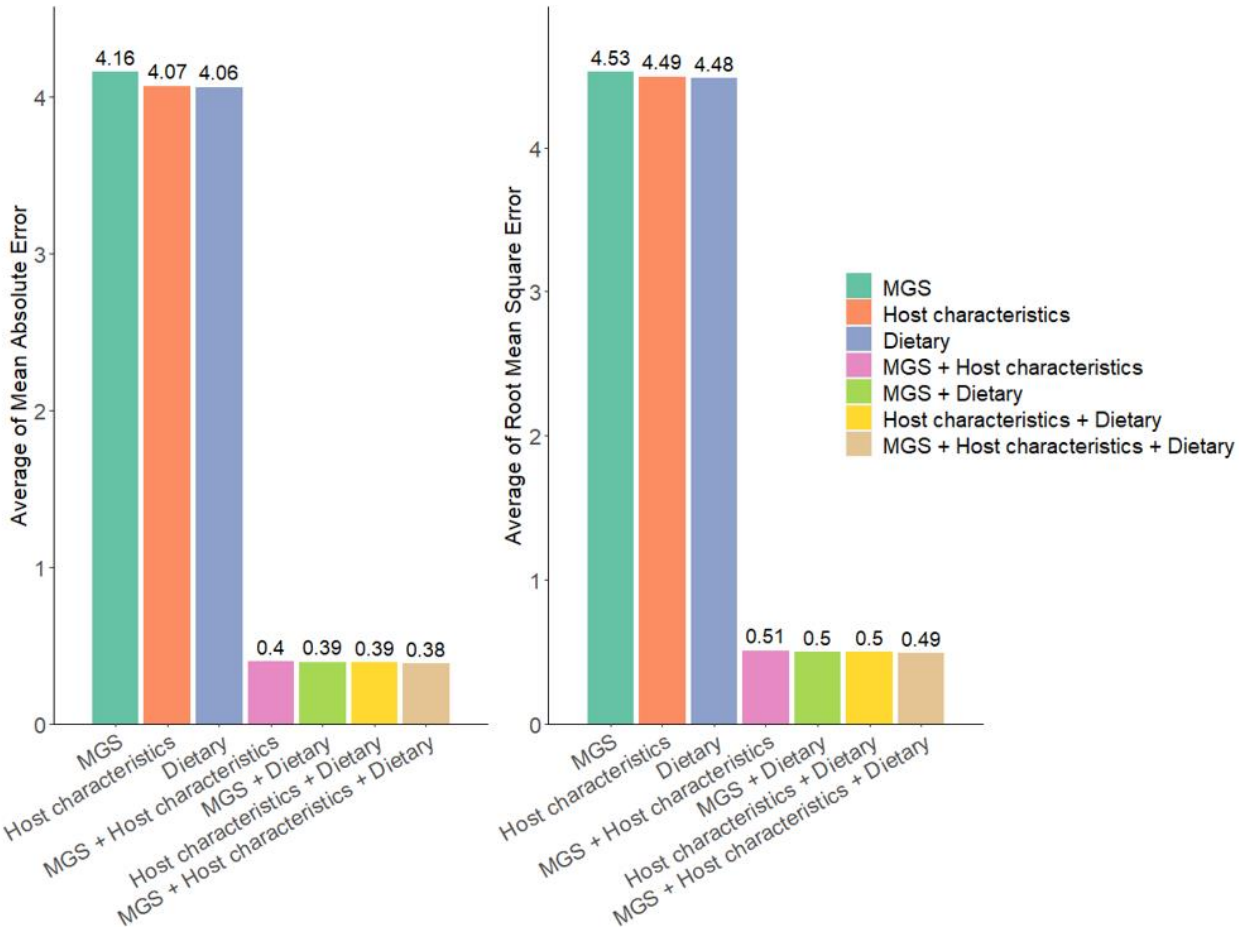


Fig. 3. Performance comparison of multi-view data prediction via M^2AE and single-view data prediction via Attentive Encoders in Dataset 2 (n = 5 experiments for each model).

Means of evaluation metrics from different experiments are shown in the figure. MGS + Host characteristics + Dietary refers to M^2AE with three types of views combining MGS, host characteristics, and dietary features data. MGS + Host characteristics refers to M^2AE with two types of views combining gut microbiomes and host characteristics. MGS + Dietary refers to M^2AE with two types of views combining gut microbiomes and dietary features. Host characteristics + Dietary refers to M^2AE with two types of views combining host characteristics and dietary features. MGS, host characteristics and dietary features refer to the view-specific attentive encoders trained with single-view gut microbiomes, host characteristics, and dietary features.

M^2AE identified important factors associated with blood SCFAs

In our analysis, we identified key features influencing SCFA production by selecting top-ranked features from two datasets, as detailed in Supplementary Tables 4-11. Among these, *Faecalibacterium prausnitzii* and *Rothia mucilaginosa* emerged as significant contributors to various SCFAs production. Several species from the *Bacteroides* genus were also highlighted for their role in SCFA biosynthesis. For example, *Bacteroides thetaiotaomicron* and *Bacteroides fragilis* were major producers of acetic acid, while *Bacteroides vulgatus* was linked to valeric acid. *Bacteroides fragilis* was also associated with 2-methylbutyric acid and isobutyric acid, and *Bacteroides eggerthii* was found to be important for isovaleric acid production. Numerous species within the *Clostridium* genus were identified as key contributors to specific SCFAs. *Clostridium bolteae* was particularly relevant for isobutyric acid production, while *Clostridium asparagiforme* played a significant role in butyric acid synthesis. *Streptococcus salivarius* was associated with butyric acid production. Additionally, species like *Leuconostoc gelidum* was noted for its relevance to valeric acid biosynthesis pathways. KEGG pathway analysis further revealed that several of these bacterial species, including *Faecalibacterium prausnitzii*, *Rothia mucilaginosa*, *Bacteroides thetaiotaomicron*, *Bacteroides fragilis*, *Bacteroides vulgatus*, *Clostridium asparagiforme*, and *Leuconostoc gelidum*, are enriched in SCFA-related biological processes, such as fatty acid biosynthesis, degradation, and metabolism.

Moreover, host characteristics, such as gender, race, age, height, weight, physical activity levels, and the use of probiotics, antibiotics, and gastric acid-lowering medications, were found to correlate with SCFA levels (Supplementary Tables 4-11). Dietary habits, including the consumption of pickles, fruits, cereals, eggs, meat, fats, coffee, and chocolate, also significantly influenced SCFA production. Overall, the factors identified by the M^2AE model showed

122 substantial diversities between different SCFAs.

123

Discussion

Recent development in high-throughput profiling technologies and integrative analysis of multi-view data offered advanced and powerful approaches to dissect complex biological problems. In this study, we pioneered an innovative approach, M^2AE, for imputing the abundances of blood SCFAs, and performed multi-view prediction for blood SCFAs data by synthesizing the information of gut microbiome, dietary features and host characteristics. This method jointly explores view-specific representation and cross-view correlation for effective prediction, and demonstrated superior performance compared with other methods. M^2AE also effectively identified prominent factors that showed strong associations with blood SCFAs.

Through literature mining, we found interesting evidence supporting the biological connections between these prominent factors and blood SCFAs and interesting relationships among some of these prominent factors.

Gut microbiotas

Our analysis identified several GM species, such as *Bacteroides thetaiotaomicron, Bacteroides fragilis*, *Bacteroides vulgatus*, *Bacteroides eggerthii*, *Clostridium asparagiforme, Clostridium bolteae, Faecalibacterium prausnitzii, Rothia mucilaginosa, Streptococcus salivarius* and *Leuconostoc gelidum*, as significant contributors to the production of blood SCFAs. Specifically, *Bacteroides thetaiotaomicron* and *Bacteroides fragilis* are prominent contributors to acetic acid production due to their ability to ferment complex carbohydrates into intermediate metabolites, such as lactate and succinate, which can be further converted into SCFAs by other gut bacteria [30, 58, 59]. *Bacteroides vulgatus* exhibited a negative association with blood valeric acid levels [60], possibly due to its negative interactions with other gut bacteria, which affect substrate availability

for valeric acid production [61]. *Bacteroides eggerthii* has been identified as a significant contributor to the production of isovaleric acid, primarily via leucine fermentation [62]. Moreover, consistent with prior study, *Clostridium asparagiforme* plays a major role in butyrate production by fermenting glucose into lactate, which is then converted to butyrate by other bacteria [63]. *Clostridium bolteae* can utilize valine through fermentation pathways, leading to the production of isobutyric acid as a metabolic byproduct [64]. *Faecalibacterium prausnitzii* is positively correlated with butyric and valeric acid [59, 65, 66]. *Rothia mucilaginosa* ferments glucose to produce acetate [67].

As above, we identified a series of gut bacterial species that have been proved to play a role in SCFA production. In addition, we identified some novel putative factors that might affect blood SCFAs. For example, *Bacteroides fragilis* was associated with 2-methylbutyric acid and isobutyric acid, *Streptococcus salivarius* was associated with butyric acid production, and *Leuconostoc gelidum* was related to valeric acid. *Bacteroides* species contribute to amino acid metabolism [68], which might lead to the production of SCFAs, such as isobutyric acid and 2-methylbutyric acid. *Streptococcus salivarius*, primarily known for its presence in the oral cavity, can also inhabit the gut and metabolize carbohydrates via fermentation [69], potentially contributing to the production of butyric acid. Similarly, *Leuconostoc gelidum* is a lactic acid bacterium known for fermenting carbohydrates to produce lactic acid [70], which might serve as a substrate for other gut microbes, potentially leading to the production of valeric acid. These findings enhance our understanding of the complex interactions between GM and blood SCFA levels. Meanwhile, these insights could help in validating our model by supporting the observed associations between specific bacterial species and SCFA production, as well as their potential influence on systemic SCFA levels. However, more in-depth studies might be needed to further unravel the underlying mechanisms.

Dietary features

Incorporating dietary features into our study enhanced our understanding in regulating the complex biological regulation of blood SCFAs production. Our findings demonstrated that the intake of various dietary components, such as pickles, fruits, cereals, eggs, meat, fat oil, coffee, and chocolate, influences blood SCFA levels. Diets could shape the microbiome by promoting the growth of bacteria that preferentially use the ingested nutrients [71]. For instance, fermented foods like pickles and fiber-rich foods like fruits and cereals promote the growth of beneficial bacteria that ferment sugars and fibers into SCFAs, such as acetic, propionic, and butyric acids [72, 73]. High-protein and high-fat diets, including eggs, steak, and fat oil, can promote the growth of *Bacteroides* species, which are adept at protein degradation and fat metabolism, thereby affecting SCFA production [10, 74-76]. Additionally, coffee and dark chocolate were linked to SCFA production due to their bioactive compounds, like caffeine, chlorogenic acid, and polyphenols, which modulate the gut microbiota and fermentation activities [77, 78]. These findings emphasize the critical role of diet in regulating SCFA levels and support our model's effectiveness in identifying dietary determinants of SCFA production.

Host characteristics

Our study identified several factors in host characteristics—such as gender, race, age, height, weight, BMI, physical activity, and the use of probiotics, antibiotics, and gastric acid-lowering medications—were correlated with blood SCFA levels. Race was associated with SCFA production, aligning with previous findings that African Americans have lower fecal acetate levels compared to white participants [45], potentially reflecting similar trends in blood SCFAs [79]. Gender and age also affect SCFA production due to differences in gut microbiota diversity and composition across

groups [80, 81]. Body composition indicators, such as BMI, weight, and height, correlate with SCFA levels, as reduced gut microbiota diversity in overweight or obese individuals often results in increased SCFA production [82-84], which is linked to energy storage and lipid metabolism [85-87]. Physical activities such as biking and swimming affect muscle lactate metabolism [88-90], and *Veillonella* species in the gut can convert lactate into SCFAs like acetic acid [91]. Additionally, probiotics increase SCFA production by boosting SCFA-producing bacteria [92], whereas antibiotics and gastric acid-lowering medications reduce microbial diversity and alter gut environments, impacting SCFA levels [84, 93, 94]. These identified host characteristics, in turn, demonstrate the effectiveness of our model in imputing SCFA levels by integrating gut microbiome compositions, dietary features, and host characteristics, providing a comprehensive understanding of the determinants influencing SCFA production.

Our current results indicated that the regulation of blood SCFAs could be a complex procedure, the GMs can be important factors. Besides, dietary habits and host characteristics might also influence blood SCFAs directly or through interactions with GMs. However, there are a few limitations in this study. First, all the subjects in our study were Caucasians and African Americans, making it necessary yet to generalize the results to other racial populations. The validation of our model in diverse populations would enhance its applicability. This necessitates further model validation in different cohorts that incorporate metagenomic and blood SCFAs profiling, dietary features, and host characteristics of a similar scope. Second, our study utilized two different datasets for model validation: one with serum SCFA measurements and the other with plasma SCFA measurements. While previous studies suggest that SCFA levels are generally consistent between serum and plasma samples, minor variations might still exist due to the different

biological matrices. These variations were not directly analyzed in our study, as the primary goal was to validate the model's performance across different datasets rather than compare serum and plasma SCFA levels. To strengthen the model's accuracy and reliability, future studies should validate our models using additional datasets with more serum and plasma SCFA measurements. Third, another limitation of our study is the sex distribution across the two datasets used for model validation. Dataset 1 included males, while Dataset 2 included both males and females. This imbalance in sex representation between the two datasets could affect the model's ability to generalize across different sexes. However, because Dataset 2 has a relatively small number of male participants, splitting this dataset by sex for separate analyses would result in low statistical power, leading to potentially unreliable results. To maintain robust model validation, we combined the male and female samples in Dataset 2, which helps to preserve adequate sample size for analysis. We adjusted for sex as a covariate in our model to account for potential sex-specific differences in SCFA production. However, future studies with more balanced sex distributions or larger sample sizes for both sexes would provide a more comprehensive understanding and enhance the robustness of the model. Fourth, the integration of data from two different sequencing platforms in Dataset 1 could be considered a limitation. However, deep learning models are particularly capable of finding common patterns across heterogeneous data by learning generalizable features that are not specific to any single sequencing platform. We applied uniform normalization within the model to standardize the data and employed regularization techniques like dropout and mean absolute error loss to prevent overfitting to sequencing platform-specific characteristics. While these strategies allow the model to generalize effectively, future studies could include more data generated from either consistent or varied sequencing platforms to further validate the model's robustness across different technical conditions. Fifth, another limitation of

our study is that the features in each view of the two datasets differ, which could result in variability in the important features identified by the model. Although many of the identified important features are well-known factors related to SCFA production, demonstrating the model's effectiveness to some extent, this variability highlights the need for caution in interpreting results. However, it is worth noting that the imputation performance was strong across both datasets, highlighting the model's generalizability despite these differences. It also opens up new opportunities to explore and identify additional relevant features that might be specific to certain datasets or conditions. Future studies should aim to include datasets with consistent feature sets across all views to enhance comparability and validate the model's ability to generalize findings across different contexts.

To summarize, we have blazed a trail with our innovative method that synthesizes information from the gut microbiome, dietary features, and host characteristics to perform multi-view imputation for blood SCFAs data. This can also help us identify key factors or pathways that regulate blood SCFAs in the future study. Our research highlights the utility of integrating information on gut microbiome, dietary features, and host characteristics, providing fresh perspectives on the potential regulatory mechanisms affecting blood SCFAs.

Data Availability

The raw data presented in this study can be found in online repositories. The names of the repositories and accession numbers can be found below: NCBI BioProjects PRJNA1015234 and PRJNA1015228.

Code Availability

The source code of this work can be downloaded from GitHub (https://github.com/Wonderangela123/M2AE).

Acknowledgement

This work is made possible with partial support by grants from the NIH (U19AG055373, R01AG061917, R01AG068232, P20GM109036 and P20GM103629).

Reference

1. Dalile B, Van Oudenhove L, Vervliet B, Verbeke K. The role of short-chain fatty acids in microbiota-gut-brain communication. Nat Rev Gastroenterol Hepatol. 2019;16(8):461-78. doi: 10.1038/s41575-019-0157-3. PubMed PMID: 31123355.

2. Sakamoto M, Takagaki A, Matsumoto K, Kato Y, Goto K, Benno Y. Butyricimonas synergistica gen. nov., sp. nov. and Butyricimonas virosa sp. nov., butyric acid-producing bacteria in the family 'Porphyromonadaceae' isolated from rat faeces. Int J Syst Evol Microbiol. 2009;59(Pt 7):1748-53. Epub 20090619. doi: 10.1099/ijs.0.007674-0. PubMed PMID: 19542124.

3. Huo W, Feng Z, Hu S, Cui L, Qiao T, Dai L, Qi P, Zhang L, Liu Y, Li J. Effects of polysaccharides from wild morels on immune response and gut microbiota composition in non-treated and cyclophosphamide-treated mice. Food Funct. 2020;11(5):4291-303. Epub 20200501. doi: 10.1039/d0fo00597e. PubMed PMID: 32356856.

4. Berni Canani R, Di Costanzo M, Leone L. The epigenetic effects of butyrate: potential therapeutic implications for clinical practice. Clin Epigenetics. 2012;4(1):4. Epub 20120227. doi: 10.1186/1868-7083-4-4. PubMed PMID: 22414433; PMCID: PMC3312834.

5. Wang Y, Yao W, Li B, Qian S, Wei B, Gong S, Wang J, Liu M, Wei M. Nuciferine modulates the gut microbiota and prevents obesity in high-fat diet-fed rats. Exp Mol Med. 2020;52(12):1959-75. Epub 20201201. doi: 10.1038/s12276-020-00534-2. PubMed PMID: 33262480; PMCID: PMC8080667.

6. Overby HB, Ferguson JF. Gut Microbiota-Derived Short-Chain Fatty Acids Facilitate Microbiota:Host Cross talk and Modulate Obesity and Hypertension. Curr Hypertens Rep. 2021;23(2):8. Epub 20210203. doi: 10.1007/s11906-020-01125-2. PubMed PMID: 33537923; PMCID: PMC7992370.

7. den Besten G, van Eunen K, Groen AK, Venema K, Reijngoud DJ, Bakker BM. The role of short-chain fatty acids in the interplay between diet, gut microbiota, and host energy metabolism. J Lipid Res. 2013;54(9):2325-40. Epub 20130702. doi: 10.1194/jlr.R036012. PubMed PMID: 23821742; PMCID: PMC3735932.

8. Rios-Covian D, Ruas-Madiedo P, Margolles A, Gueimonde M, de Los Reyes-Gavilan CG, Salazar N. Intestinal Short Chain Fatty Acids and their Link with Diet and Human Health. Front Microbiol. 2016;7:185. Epub 20160217. doi: 10.3389/fmicb.2016.00185. PubMed PMID: 26925050; PMCID: PMC4756104.

9. Louis P, Flint HJ. Diversity, metabolism and microbial ecology of butyrate-producing bacteria from the human large intestine. FEMS Microbiol Lett. 2009;294(1):1-8. Epub 20090213. doi: 10.1111/j.1574-6968.2009.01514.x. PubMed PMID: 19222573.

10. Chen J, Xiao Y, Li D, Zhang S, Wu Y, Zhang Q, Bai W. New insights into the mechanisms of high-fat diet mediated gut microbiota in chronic diseases. iMeta. 2023;2(1):e69. doi: https://doi.org/10.1002/imt2.69.

11. Wang G, Li X, Zhao J, Zhang H, Chen W. Lactobacillus casei CCFM419 attenuates type 2 diabetes via a gut microbiota dependent mechanism. Food Funct. 2017;8(9):3155-64. doi: 10.1039/c7fo00593h. PubMed PMID: 28782784.

12. Conlon MA, Bird AR. The Impact of Diet and Lifestyle on Gut Microbiota and Human Health. Nutrients. 2015;7(1):17-44.

13. Besten G, Eunen K, Groen AK, Venema K, Reijngoud K, Bakker BM. The role of short-chain fatty acids in the interplay between diet, gut microbiota, and host energy metabolism. J Lipid Res. 2013;54(9):2325–40.

14. You H, Tan Y, Yu D, Qiu S, Bai Y, He J, Cao H, Che Q, Guo J, Su Z. The Therapeutic Effect of SCFA-Mediated Regulation of the Intestinal Environment on Obesity. Front Nutr. 2022;9:886902. Epub 20220517. doi: 10.3389/fnut.2022.886902. PubMed PMID: 35662937; PMCID: PMC9157426.

15. Zeevi D, Korem T, Zmora N, Israeli D, Rothschild D, Weinberger A, Ben-Yacov O, Lador D, Avnit-Sagi T, Lotan-Pompan M, Suez J, Mahdi JA, Matot E, Malka G, Kosower N, Rein M, Zilberman-Schapira G, Dohnalova L, Pevsner-Fischer M, Bikovsky R, Halpern Z, Elinav E, Segal E. Personalized Nutrition by Prediction of Glycemic Responses. Cell. 2015;163(5):1079-94. doi: 10.1016/j.cell.2015.11.001. PubMed PMID: 26590418.

16. Wang Y, Gao X, Lv J, Zeng Y, Li Q, Wang L, Zhang Y, Gao W, Wang J. Gut Microbiome Signature Are Correlated With Bone Mineral Density Alterations in the Chinese Elders. Front Cell Infect Microbiol. 2022;12:827575. Epub 20220331. doi: 10.3389/fcimb.2022.827575. PubMed PMID: 35433497; PMCID: PMC9008261.

17. Ramos Meyers G, Samouda H, Bohn T. Short Chain Fatty Acid Metabolism in Relation to Gut Microbiota and Genetic Variability. Nutrients. 2022;14(24). Epub 20221216. doi: 10.3390/nu14245361. PubMed PMID: 36558520; PMCID: PMC9788597.

18. den Besten G, Lange K, Havinga R, van Dijk TH, Gerding A, van Eunen K, Muller M, Groen AK, Hooiveld GJ, Bakker BM, Reijngoud DJ. Gut-derived short-chain fatty acids are vividly assimilated into host carbohydrates and lipids. Am J Physiol Gastrointest Liver Physiol. 2013;305(12):G900-10. Epub 20131017. doi: 10.1152/ajpgi.00265.2013. PubMed PMID: 24136789.

19. Lin X, Xiao HM, Liu HM, Lv WQ, Greenbaum J, Gong R, Zhang Q, Chen YC, Peng C, Xu XJ, Pan DY, Chen Z, Li ZF, Zhou R, Wang XF, Lu JM, Ao ZX, Song YQ, Zhang YH, Su KJ, Meng XH, Ge CL, Lv FY, Luo Z, Shi XM, Zhao Q, Guo BY, Yi NJ, Shen H, Papasian CJ, Shen J, Deng HW. Gut microbiota impacts bone via Bacteroides vulgatus-valeric acid-related pathways. Nat Commun. 2023;14(1):6853. Epub 20231027. doi: 10.1038/s41467-023-42005-y. PubMed PMID: 37891329; PMCID: PMC10611739.

20. Zmora N, Suez J, Elinav E. You are what you eat: diet, health and the gut microbiota. Nat Rev Gastroenterol Hepatol. 2019;16(1):35-56. doi: 10.1038/s41575-018-0061-2. PubMed PMID: 30262901.

21. Greenbaum J, Su KJ, Zhang X, Liu Y, Liu A, Zhao LJ, Luo Z, Tian Q, Shen H, Deng HW. A multiethnic whole genome sequencing study to identify novel loci for bone mineral density. Hum Mol Genet. 2022;31(7):1067-81. doi: 10.1093/hmg/ddab305. PubMed PMID: 34673960; PMCID: PMC8976433.

22. He H, Liu Y, Tian Q, Papasian CJ, Hu T, Deng HW. Relationship of sarcopenia and body composition with osteoporosis. Osteoporos Int. 2016;27(2):473-82. Epub 2015/08/06. doi: 10.1007/s00198-015-3241-8. PubMed PMID: 26243357.

23. Song M, Greenbaum J, Luttrell Jt, Zhou W, Wu C, Luo Z, Qiu C, Zhao LJ, Su KJ, Tian Q, Shen H, Hong H, Gong P, Shi X, Deng HW, Zhang C. An autoencoder-based deep learning method for genotype imputation. Front Artif Intell. 2022;5:1028978. Epub 20221103. doi: 10.3389/frai.2022.1028978. PubMed PMID: 36406474; PMCID: PMC9671213.

24. Du Y, Xu T, Yin Z, Espinoza S, Xie Y, Gentry C, Tian Q, Zhao LJ, Shen H, Luo Z, Deng HW. Associations of physical activity with sarcopenia and sarcopenic obesity in middle-aged and older adults: the Louisiana osteoporosis study. BMC Public Health. 2022;22(1):896. Epub 2022/05/06. doi: 10.1186/s12889-022-13288-5. PubMed PMID: 35513868; PMCID: PMC9074188.

25. Du Y, Zhao LJ, Xu Q, Wu KH, Deng HW. Socioeconomic status and bone mineral density in adults by race/ethnicity and gender: the Louisiana osteoporosis study. Osteoporos Int. 2017;28(5):1699-709. Epub 2017/02/27. doi: 10.1007/s00198-017-3951-1. PubMed PMID: 28236128.

26. Jeng C, Zhao LJ, Wu K, Zhou Y, Chen T, Deng HW. Race and socioeconomic effect on sarcopenia and sarcopenic obesity in the Louisiana Osteoporosis Study (LOS). JCSM Clin Rep. 2018;3(2). Epub 2019/08/30. PubMed PMID: 31463425; PMCID: PMC6712972.

27. Ning H, Du Y, Zhao LJ, Tian Q, Feng H, Deng HW. The mediating effect of skeletal muscle index on the relationship between menarcheal age and bone mineral density in premenopausal women by race/ethnicity. Menopause. 2021;28(10):1143-9. Epub 2021/07/28. doi: 10.1097/GME.0000000000001814. PubMed PMID: 34313616; PMCID: PMC8514157.

28. Bruce SJ, Tavazzi I, Parisod V, Rezzi S, Kochhar S, Guy PA. Investigation of human blood plasma sample preparation for performing metabolomics using ultrahigh performance liquid chromatography/mass spectrometry. Anal Chem. 2009;81(9):3285-96. doi: 10.1021/ac8024569. PubMed PMID: 19323527.

29. Lv WQ, Lin X, Shen H, Liu HM, Qiu X, Li BY, Shen WD, Ge CL, Lv FY, Shen J, Xiao HM, Deng HW. Human gut microbiome impacts skeletal muscle mass via gut microbial synthesis of the short-chain fatty acid butyrate among healthy menopausal women. J

Cachexia Sarcopenia Muscle. 2021;12(6):1860-70. Epub 20210901. doi: 10.1002/jcsm.12788. PubMed PMID: 34472211; PMCID: PMC8718076.

30. Shen WD, Lin X, Liu HM, Li BY, Qiu X, Lv WQ, Zhu XZ, Greenbaum J, Liu RK, Shen J, Xiao HM, Deng HW. Gut microbiota accelerates obesity in peri-/post-menopausal women via Bacteroides fragilis and acetic acid. Int J Obes (Lond). 2022;46(10):1918-24. Epub 20220817. doi: 10.1038/s41366-022-01137-9. PubMed PMID: 35978102.

31. Liu HM, Lin X, Meng XH, Zhao Q, Shen J, Xiao HM, Deng HW. Integrated metagenome and metabolome analyses of blood pressure studies in early postmenopausal Chinese women. J Hypertens. 2021;39(9):1800-9. doi: 10.1097/HJH.0000000000002832. PubMed PMID: 33758158.

32. Tian B, Xu LL, Jiang LD, Lin X, Shen J, Shen H, Su KJ, Gong R, Qiu C, Luo Z, Yao JH, Wang ZQ, Xiao HM, Zhang LS, Deng HW. Identification of the serum metabolites associated with cow milk consumption in Chinese Peri-/Postmenopausal women. Int J Food Sci Nutr. 2024;75(6):537-49. Epub 20240625. doi: 10.1080/09637486.2024.2366223. PubMed PMID: 38918932.

33. Tian B, Yao JH, Lin X, Lv WQ, Jiang LD, Wang ZQ, Shen J, Xiao HM, Xu H, Xu LL, Cheng X, Shen H, Qiu C, Luo Z, Zhao LJ, Yan Q, Deng HW, Zhang LS. Metagenomic study of the gut microbiota associated with cow milk consumption in Chinese peri-/postmenopausal women. Front Microbiol. 2022;13:957885. Epub 20220816. doi: 10.3389/fmicb.2022.957885. PubMed PMID: 36051762; PMCID: PMC9425034.

34. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. Embnet Journal. 2011;17.

35. Pertea G. fqtrim: v0.9.4 (Version 0.9.4) 2015. Available from: http://ccb.jhu.edu/software/fqtrim/index.shtml.

36. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357-9. Epub 20120304. doi: 10.1038/nmeth.1923. PubMed PMID: 22388286; PMCID: PMC3322381.

37. Peng Y, Leung HC, Yiu SM, Chin FY. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics. 2012;28(11):1420-8. Epub 20120411. doi: 10.1093/bioinformatics/bts174. PubMed PMID: 22495754.

38. Zhu W, Lomsadze A, Borodovsky M. Ab initio gene identification in metagenomic sequences. Nucleic Acids Res. 2010;38(12):e132. Epub 20100419. doi: 10.1093/nar/gkq275. PubMed PMID: 20403810; PMCID: PMC2896542.

39. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22(13):1658-9. Epub 20060526. doi: 10.1093/bioinformatics/btl158. PubMed PMID: 16731699.

40. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12(1):59-60. Epub 20141117. doi: 10.1038/nmeth.3176. PubMed PMID: 25402007.

41. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34(17):i884-i90. doi: 10.1093/bioinformatics/bty560. PubMed PMID: 30423086; PMCID: PMC6129281.

42. Franzosa EA, McIver LJ, Rahnavard G, Thompson LR, Schirmer M, Weingart G, Lipson KS, Knight R, Caporaso JG, Segata N, Huttenhower C. Species-level functional profiling of metagenomes and metatranscriptomes. Nat Methods. 2018;15(11):962-8. Epub 20181030. doi: 10.1038/s41592-018-0176-y. PubMed PMID: 30377376; PMCID: PMC6235447.

43. Bowen RA, Remaley AT. Interferences from blood collection tube components on clinical chemistry assays. Biochem Med (Zagreb). 2014;24(1):31-44. Epub 20140215. doi: 10.11613/BM.2014.006. PubMed PMID: 24627713; PMCID: PMC3936985.

44. Zhao HJ, Chen Y, Liu T, McArthur K, Mueller NT. Short-Chain Fatty Acids and Preeclampsia: A Scoping Review. Nutr Rev. 2024. Epub 20240525. doi: 10.1093/nutrit/nuae057. PubMed PMID: 38796843.

45. Hester CM, Jala VR, Langille MG, Umar S, Greiner KA, Haribabu B. Fecal microbes, short chain fatty acids, and colorectal cancer across racial/ethnic groups. World J Gastroenterol. 2015;21(9):2759-69. doi: 10.3748/wjg.v21.i9.2759. PubMed PMID: 25759547; PMCID: PMC4351229.

46. Jenkins DJ, Kendall CW, Popovich DG, Vidgen E, Mehling CC, Vuksan V, Ransom TP, Rao AV, Rosenberg-Zand R, Tariq N, Corey P, Jones PJ, Raeini M, Story JA, Furumoto EJ, Illingworth DR, Pappu AS, Connelly PW. Effect of a very-high-fiber vegetable, fruit, and nut diet on serum lipids and colonic function. Metabolism. 2001;50(4):494-503. doi: 10.1053/meta.2001.21037. PubMed PMID: 11288049.

47. Sowah SA, Riedl L, Damms-Machado A, Johnson TS, Schubel R, Graf M, Kartal E, Zeller G, Schwingshackl L, Stangl GI, Kaaks R, Kuhn T. Effects of Weight-Loss Interventions on Short-Chain Fatty Acid Concentrations in Blood and Feces of Adults: A Systematic Review. Adv Nutr. 2019;10(4):673-84. doi: 10.1093/advances/nmy125. PubMed PMID: 31075175; PMCID: PMC6628843.

48. Guinan J, Wang S, Hazbun TR, Yadav H, Thangamani S. Antibiotic-induced decreases in the levels of microbial-derived short-chain fatty acids correlate with increased gastrointestinal colonization of Candida albicans. Sci Rep. 2019;9(1):8872. Epub 20190620. doi: 10.1038/s41598-019-45467-7. PubMed PMID: 31222159; PMCID: PMC6586901.

49. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software. 2011;45(3):1-67. doi: 10.18637/jss.v045.i03.

50. Li J, Lu G, Wu Z, Ling F. Multi-view representation model based on graph autoencoder. Information Sciences. 2023;632:439-53.

51. 2022 Alzheimer's disease facts and figures. Alzheimers Dement. 2022;18(4):700-89. Epub 2022/03/16. doi: 10.1002/alz.12638. PubMed PMID: 35289055.

52. Zhang X, Wang X, Shivashankar GV, Uhler C. Graph-based autoencoder integrates spatial transcriptomics with chromatin images and identifies joint biomarkers for Alzheimer's disease. Nat Commun. 2022;13(1):7480. Epub 20221203. doi: 10.1038/s41467-022-35233-1. PubMed PMID: 36463283; PMCID: PMC9719477.

53. Singh A, Shannon CP, Gautier B, Rohart F, Vacher M, Tebbutt SJ, Le Cao KA. DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. Bioinformatics. 2019;35(17):3055-62. doi: 10.1093/bioinformatics/bty1054. PubMed PMID: 30657866; PMCID: PMC6735831.

54. Serra A, Fratello M, Fortino V, Raiconi G, Tagliaferri R, Greco D. MVDA: a multi-view genomic data integration methodology. BMC Bioinformatics. 2015;16:261. Epub 20150819. doi: 10.1186/s12859-015-0680-3. PubMed PMID: 26283178; PMCID: PMC4539887.

55. Zhu X, Suk HI, Zhu Y, Thung KH, Wu G, Shen D. Multi-view Classification for Identification of Alzheimer's Disease. Mach Learn Med Imaging. 2015;9352:255-62. Epub 20151002. doi: 10.1007/978-3-319-24888-2_31. PubMed PMID: 26900608; PMCID: PMC4758364.

56. Li Y, Wu FX, Ngom A. A review on machine learning principles for multi-view biological data integration. Brief Bioinform. 2018;19(2):325-40. doi: 10.1093/bib/bbw113. PubMed PMID: 28011753.

57. Wang L. Correlation Discovery for Multi-view and Multi-label Learning: Northeastern University; 2021.

58. Samuel BS, Gordon JI. A humanized gnotobiotic mouse model of host-archaeal-bacterial mutualism. Proc Natl Acad Sci U S A. 2006;103(26):10011-6. Epub 20060616. doi: 10.1073/pnas.0602187103. PubMed PMID: 16782812; PMCID: PMC1479766.

59. Wrzosek L, Miquel S, Noordine ML, Bouet S, Joncquel Chevalier-Curt M, Robert V, Philippe C, Bridonneau C, Cherbuy C, Robbe-Masselot C, Langella P, Thomas M. Bacteroides thetaiotaomicron and Faecalibacterium prausnitzii influence the production of mucus glycans and the development of goblet cells in the colonic epithelium of a gnotobiotic model rodent. BMC Biol. 2013;11:61. Epub 20130521. doi: 10.1186/1741-7007-11-61. PubMed PMID: 23692866; PMCID: PMC3673873.

60. Lin X, et al. Gut microbiota impacts bone via B.vulgatus-valeric acid-related pathways2020. doi: chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.medrxiv.org/content/10.1101/2020.03.16.20037077v2.full.pdf.

61. Coyte KZ, Rakoff-Nahoum S. Understanding Competition and Cooperation within the Mammalian Gut Microbiome. Curr Biol. 2019;29(11):R538-R44. doi: 10.1016/j.cub.2019.04.017. PubMed PMID: 31163167; PMCID: PMC6935513.

62. Wang H, Xie L, Liu S, Dai A, Chi X, Zhang D. Non-targeted metabolomics and microbial analyses of the impact of oat antimicrobial peptides on rats with dextran sulfate sodium-induced enteritis. Front Nutr. 2022;9:1095483. Epub 20230111. doi: 10.3389/fnut.2022.1095483. PubMed PMID: 36712538; PMCID: PMC9875082.

63. Thomson P, Medina DA, Ortuzar V, Gotteland M, Garrido D. Anti-inflammatory effect of microbial consortia during the utilization of dietary polysaccharides. Food Res Int. 2018;109:14-23. Epub 20180411. doi: 10.1016/j.foodres.2018.04.008. PubMed PMID: 29803436.

64. Watanabe M, Kaku N, Ueki K, Ueki A. Falcatimonas natans gen. nov., sp. nov., a strictly anaerobic, amino-acid-decomposing bacterium isolated from a methanogenic reactor of cattle waste. Int J Syst Evol Microbiol. 2016;66(11):4639-44. Epub 20160808. doi: 10.1099/ijsem.0.001403. PubMed PMID: 27506535.

65. Xu D, Feng M, Chu Y, Wang S, Shete V, Tuohy KM, Liu F, Zhou X, Kamil A, Pan D, Liu H, Yang X, Yang C, Zhu B, Lv N, Xiong Q, Wang X, Sun J, Sun G, Yang Y. The Prebiotic Effects of Oats on Blood Lipids, Gut Microbiota, and Short-Chain Fatty Acids in Mildly Hypercholesterolemic Subjects Compared With Rice: A Randomized, Controlled Trial. Front Immunol. 2021;12:787797. Epub 20211209. doi: 10.3389/fimmu.2021.787797. PubMed PMID: 34956218; PMCID: PMC8697019.

66. Song CH, Kim N, Nam RH, Choi SI, Jang JY, Kim EH, Choi J, Choi Y, Yoon H, Lee SM, Seok YJ. The Possible Preventative Role of Lactate- and Butyrate-Producing Bacteria in Colorectal Carcinogenesis. Gut Liver. 2024;18(4):654-66. Epub 20231130. doi: 10.5009/gnl230385. PubMed PMID: 38030382; PMCID: PMC11249946.

67. Gao B, Gallagher T, Zhang Y, Elbadawi-Sidhu M, Lai Z, Fiehn O, Whiteson KL. Tracking Polymicrobial Metabolism in Cystic Fibrosis Airways: Pseudomonas aeruginosa Metabolism and Physiology Are Influenced by Rothia mucilaginosa-Derived Metabolites. mSphere. 2018;3(2). Epub 20180425. doi: 10.1128/mSphere.00151-18. PubMed PMID: 29695623; PMCID: PMC5917424.

68. Allison C, Macfarlane GT. Influence of pH, nutrient availability, and growth rate on amine production by Bacteroides fragilis and Clostridium perfringens. Appl Environ Microbiol. 1989;55(11):2894-8. doi: 10.1128/aem.55.11.2894-2898.1989. PubMed PMID: 2560361; PMCID: PMC203187.

69. Abranches J, Zeng L, Kajfasz JK, Palmer SR, Chakraborty B, Wen ZT, Richards VP, Brady LJ, Lemos JA. Biology of Oral Streptococci. Microbiol Spectr. 2018;6(5). doi: 10.1128/microbiolspec.GPP3-0042-2018. PubMed PMID: 30338752; PMCID: PMC6287261.

70. Johansson P, Sade E, Hultman J, Auvinen P, Bjorkroth J. Pangenome and genomic taxonomy analyses of Leuconostoc gelidum and Leuconostoc gasicomitatum. BMC Genomics. 2022;23(1):818. Epub 20221209. doi: 10.1186/s12864-022-09032-3. PubMed PMID: 36494615; PMCID: PMC9733070.

71. Zeng X, Xing X, Gupta M, Keber FC, Lopez JG, Lee YJ, Roichman A, Wang L, Neinast MD, Donia MS, Wuhr M, Jang C, Rabinowitz JD. Gut bacterial nutrient preferences quantified in vivo. Cell. 2022;185(18):3441-56 e19. doi: 10.1016/j.cell.2022.07.020. PubMed PMID: 36055202; PMCID: PMC9450212.

72. Swain MR, Anandharaj M, Ray RC, Parveen Rani R. Fermented fruits and vegetables of Asia: a potential source of probiotics. Biotechnol Res Int. 2014;2014:250424. Epub 20140528. doi: 10.1155/2014/250424. PubMed PMID: 25343046; PMCID: PMC4058509.

73. Chung WS, Walker AW, Louis P, Parkhill J, Vermeiren J, Bosscher D, Duncan SH, Flint HJ. Modulation of the human gut microbiota by dietary fibres occurs at the species level. BMC Biol. 2016;14:3. Epub 20160111. doi: 10.1186/s12915-015-0224-3. PubMed PMID: 26754945; PMCID: PMC4709873.

74. Mendez-Salazar EO, Ortiz-Lopez MG, Granados-Silvestre MLA, Palacios-Gonzalez B, Menjivar M. Altered Gut Microbiota and Compositional Changes in Firmicutes and Proteobacteria in Mexican Undernourished and Obese Children. Front Microbiol. 2018;9:2494. Epub 20181016. doi: 10.3389/fmicb.2018.02494. PubMed PMID: 30386323; PMCID: PMC6198253.

75. Luo J, Zhang C, Liu R, Gao L, Ou S, Liu L, Peng X. Ganoderma lucidum polysaccharide alleviating colorectal cancer by alteration of special gut bacteria and regulation of gene expression of colonic epithelial cells. Journal of Functional Foods. 2018;47:127-35.

76. Bisht A, Goh KKT, Matia-Merino L. The fate of mamaku gum in the gut: effect on in vitro gastrointestinal function and colon fermentation by human faecal microbiota. Food Funct. 2023;14(15):7024-39. Epub 20230731. doi: 10.1039/d3fo01665j. PubMed PMID: 37439088.

77. Nishitsuji K, Watanabe S, Xiao J, Nagatomo R, Ogawa H, Tsunematsu T, Umemoto H, Morimoto Y, Akatsu H, Inoue K, Tsuneyama K. Effect of coffee or coffee components on gut microbiome and short-chain fatty acids in a mouse model of metabolic syndrome. Sci Rep. 2018;8(1):16173. Epub 20181101. doi: 10.1038/s41598-018-34571-9. PubMed PMID: 30385796; PMCID: PMC6212590.

78. Garcia-Cordero J, Martinez A, Blanco-Valverde C, Pino A, Puertas-Martin V, San Roman R, de Pascual-Teresa S. Regular Consumption of Cocoa and Red Berries as a Strategy to Improve Cardiovascular Biomarkers via Modulation of Microbiota Metabolism in Healthy Aging Adults. Nutrients. 2023;15(10). Epub 20230513. doi: 10.3390/nu15102299. PubMed PMID: 37242181; PMCID: PMC10223313.

79. Yamamura R, Nakamura K, Kitada N, Aizawa T, Shimizu Y, Nakamura K, Ayabe T, Kimura T, Tamakoshi A. Associations of gut microbiota, dietary intake, and serum short-

chain fatty acids with fecal short-chain fatty acids. Biosci Microbiota Food Health. 2020;39(1):11-7. Epub 20191005. doi: 10.12938/bmfh.19-010. PubMed PMID: 32010539; PMCID: PMC6971417.

80. Mueller S, Saunier K, Hanisch C, Norin E, Alm L, Midtvedt T, Cresci A, Silvi S, Orpianesi C, Verdenelli MC, Clavel T, Koebnick C, Zunft HJ, Dore J, Blaut M. Differences in fecal microbiota in different European study populations in relation to age, gender, and country: a cross-sectional study. Appl Environ Microbiol. 2006;72(2):1027-33. doi: 10.1128/AEM.72.2.1027-1033.2006. PubMed PMID: 16461645; PMCID: PMC1392899.

81. Fransen F, van Beek AA, Borghuis T, Meijer B, Hugenholtz F, van der Gaast-de Jongh C, Savelkoul HF, de Jonge MI, Faas MM, Boekschoten MV, Smidt H, El Aidy S, de Vos P. The Impact of Gut Microbiota on Gender-Specific Differences in Immunity. Front Immunol. 2017;8:754. Epub 20170630. doi: 10.3389/fimmu.2017.00754. PubMed PMID: 28713378; PMCID: PMC5491612.

82. Liu BN, Liu XT, Liang ZH, Wang JH. Gut microbiota in obesity. World J Gastroenterol. 2021;27(25):3837-50. doi: 10.3748/wjg.v27.i25.3837. PubMed PMID: 34321848; PMCID: PMC8291023.

83. Salazar N, Ponce-Alonso M, Garriga M, Sanchez-Carrillo S, Hernandez-Barranco AM, Redruello B, Fernandez M, Botella-Carretero JI, Vega-Pinero B, Galeano J, Zamora J, Ferrer M, de Los Reyes-Gavilan CG, Del Campo R. Fecal Metabolome and Bacterial Composition in Severe Obesity: Impact of Diet and Bariatric Surgery. Gut Microbes. 2022;14(1):2106102. doi: 10.1080/19490976.2022.2106102. PubMed PMID: 35903014; PMCID: PMC9341356.

84. Lange O, Proczko-Stepaniak M, Mika A. Short-Chain Fatty Acids-A Product of the Microbiome and Its Participation in Two-Way Communication on the Microbiome-Host Mammal Line. Curr Obes Rep. 2023;12(2):108-26. Epub 20230519. doi: 10.1007/s13679-023-00503-6. PubMed PMID: 37208544; PMCID: PMC10250490.

85. Fernandes J, Su W, Rahat-Rozenbloom S, Wolever TM, Comelli EM. Adiposity, gut microbiota and faecal short chain fatty acids are linked in adult humans. Nutr Diabetes. 2014;4(6):e121. Epub 20140630. doi: 10.1038/nutd.2014.23. PubMed PMID: 24979150; PMCID: PMC4079931.

86. Heiss CN, Olofsson LE. Gut Microbiota-Dependent Modulation of Energy Metabolism. J Innate Immun. 2018;10(3):163-71. Epub 20171108. doi: 10.1159/000481519. PubMed PMID: 29131106; PMCID: PMC6757175.

87. Chambers ES, Preston T, Frost G, Morrison DJ. Role of Gut Microbiota-Generated Short-Chain Fatty Acids in Metabolic and Cardiovascular Health. Curr Nutr Rep. 2018;7(4):198-206. doi: 10.1007/s13668-018-0248-8. PubMed PMID: 30264354; PMCID: PMC6244749.

88. Richard NA, Koehle MS. Optimizing recovery to support multi-evening cycling competition performance. Eur J Sport Sci. 2019;19(6):811-23. Epub 20181227. doi: 10.1080/17461391.2018.1560506. PubMed PMID: 30589619.

609 89. Cieminski K, Flis DJ, Dzik KP, Kaczor JJ, Wieckowski MR, Antosiewicz J, Ziolkowski W.
610 Swim Training Affects on Muscle Lactate Metabolism, Nicotinamide Adenine
611 Dinucleotides Concentration, and the Activity of NADH Shuttle Enzymes in a Mouse
612 Model of Amyotrophic Lateral Sclerosis. Int J Mol Sci. 2022;23(19). Epub 20220929. doi:
613 10.3390/ijms231911504. PubMed PMID: 36232801; PMCID: PMC9569676.

614 90. Fasching P, Rinnerhofer S, Wultsch G, Birnbaumer P, Hofmann P. The First Lactate
615 Threshold Is a Limit for Heavy Occupational Work. J Funct Morphol Kinesiol. 2020;5(3).
616 Epub 20200825. doi: 10.3390/jfmk5030066. PubMed PMID: 33467281; PMCID:
617 PMC7739369.

618 91. Scheiman J, Luber JM, Chavkin TA, MacDonald T, Tung A, Pham LD, Wibowo MC,
619 Wurth RC, Punthambaker S, Tierney BT, Yang Z, Hattab MW, Avila-Pacheco J, Clish CB,
620 Lessard S, Church GM, Kostic AD. Meta-omics analysis of elite athletes identifies a
621 performance-enhancing microbe that functions via lactate metabolism. Nat Med.
622 2019;25(7):1104-9. Epub 20190624. doi: 10.1038/s41591-019-0485-4. PubMed PMID:
623 31235964; PMCID: PMC7368972.

624 92. Markowiak-Kopec P, Slizewska K. The Effect of Probiotics on the Production of Short-
625 Chain Fatty Acids by Human Intestinal Microbiome. Nutrients. 2020;12(4). Epub
626 20200416. doi: 10.3390/nu12041107. PubMed PMID: 32316181; PMCID: PMC7230973.

627 93. Modi SR, Collins JJ, Relman DA. Antibiotics and the gut microbiota. J Clin Invest.
628 2014;124(10):4212-8. Epub 20141001. doi: 10.1172/JCI72333. PubMed PMID: 25271726;
629 PMCID: PMC4191029.

630 94. Dethlefsen L, Huse S, Sogin ML, Relman DA. The pervasive effects of an antibiotic on the
631 human gut microbiota, as revealed by deep 16S rRNA sequencing. PLoS Biol.
632 2008;6(11):e280. doi: 10.1371/journal.pbio.0060280. PubMed PMID: 19018661; PMCID:
633 PMC2586385.

Supplementary Tables

Supplementary Table 1. Features (Host characteristics and dietary habits) used in Dataset 1.

Host Characteristics	Dietary Habits
Age	Whether you drink milk (Yes/No)
Race	Whether you eat yogurt (Yes/No)
Height (cm)	Whether you eat cheese (Yes/No)
Weight (kg)	Quantity of eating vegetables
Whether currently exercise on a fairly regular basis (Yes/No)	Quantity of eating fruit
Whether you smoke(d) cigarettes (Yes/No)	Frequency of eating eggs
Whether you drink (drank) alcohol (Yes/No)	Quantity of eating eggs in a day
Whether you are currently taking any probiotics supplements (Yes/No)	Frequency of eating steak
Whether you have you taken any probiotics in the past 12 months (Yes/No)	Quantity of eating steak in a day
BMI	Frequency of eating pork chops
Whether you have used antibiotics in the past 12 months (Yes/No)	Quantity of eating pork chops in a day
Whether you are currently taking any gastric acid lowering medications (Yes/No)	Frequency of eating cookies
Whether you have you taken any gastric acid lowering medications in the past 12 months (Yes/No)	Quantity of eating cookies in a day
Frequency of bicycling or swimming	Frequency of eating chocolate candy
Time in bicycling or swimming for each time	Quantity of eating chocolate candy in a day
Frequency of heavy work	Frequency of eating cold cereals
Time in heavy work for each time	Quantity of eating cold cereals in a day
	Quantity of drinking water in a day
	Frequency of drinking milky coffee
	Quantity of drinking milky coffee in a day
	Frequency of drinking non-milky coffee
	Quantity of drinking non-milky coffee in a day
	Frequency of drinking hot tea
	Quantity of drinking hot tea in a day
	Frequency of eating pickles
	Quantity of eating pickles in a day
	Frequency of taking folic acid
	Frequency of taking fiber supplements
	Frequency of taking calcium
	Quantity of taking calcium in a day
	Frequency of taking fat oil
	Number of meals in a day
	How much eating fried fish, fish sticks, fish sandwich, breaded fillets (g)

639 **Supplementary Table 2.** Features (Host characteristics and dietary habits) used in Dataset 2.

Host Characteristics	**Dietary Habits**
Age	Whether you drink milk (Yes/No)
Gender	Whether you eat yogurt (Yes/No)
Race	Whether you eat cheese (Yes/No)
Height (cm)	
Weight (kg)	
Whether currently exercise on a fairly regular basis (Yes/No)	
Whether you smoke(d) cigarettes (Yes/No)	
Whether you drink (drank) alcohol (Yes/No)	
Whether you are currently taking any probiotics supplements (Yes/No)	
Whether you have you taken any probiotics in the past 12 months (Yes/No)	
BMI	

640

641

642

643

644 **Supplementary Table 3.** Hyperparameters Tuning.

Hyperparameters	**Values**
Dropout	0.1, 0.3, 0.5, 0.7
Number of epochs	1000, 1500, 2000, 2500, 3000
Average number of edges per node that are retained including self-connections (k)	2, 5, 10

645

646

Supplementary Table 4. Important factors associated with blood acetic acids.

Gut Microbiotas		**Host Characteristics**		**Dietary Features**	
Dataset 1	**Dataset 2**	**Dataset 1**	**Dataset 2**	**Dataset 1**	**Dataset 2**
Subdoligranulum_unclassified	Bacteroides_intestinalis	Weight (kg)	Gender	Frequency of eating pickles	Whether you eat yogurt
Ruminococcus_torques	candidate_division_TM7_single_cell_isolate_TM7b	Time in bicycling or swimming for each time	Race	Frequency of eating steak	Whether you drink milk
Oscillibacter_unclassified	Atopobium_vaginae	Whether you have you taken any probiotics in the past 12 months (Yes/No)	Height (cm)	Frequency of taking fat oil	Whether you eat cheese
Bacteroides_uniformis	Weissella_confusa	Height (cm)	Weight (kg)	Quantity of eating eggs in a day	
Clostridium_asparagiforme	Lachnospiraceae_bacterium_2_1_46FAA	Whether you have used antibiotics in the past 12 months (Yes/No)	Whether you have you taken any probiotics in the past 12 months (Yes/No)	Quantity of eating cold cereals in a day	
Ruminococcus_sp_5_1_39BFAA	Bombyx_mori_nucleopolyhedrovirus				
Rothia_mucilaginosa	Streptococcus_mutans				
Prevotella_copri	Faecalibacterium_prausnitzii				
Roseburia_intestinalis	Clostridium_sp_HGF2				
Bacteroides_thetaiotaomicron	Ruminococcus_obeum				
Desulfovibrio_piger	Bordetella_unclassified				
Alistipes_unclassified	Anaerococcus_vaginalis				
Bifidobacterium_catenulatum	Alistipes_sp_HGB5				
Lactobacillus_salivarius	Avian_carcinoma_virus				
Clostridium_leptum	New_World_begomovirus_associated_satellite_DNA				
Acidaminococcus_sp_BV3L6	Sapporo_virus				

Erysipelotrichaceae_bacterium_6_1_45	Bifidobacterium_breve
Bacteroides_fragilis	Gemella_sanguinis
Parabacteroides_merdae	Fusobacterium_necrophorum
Alistipes_senegalensis	Culex_flavivirus
Mitsuokella_unclassified	Anaerotruncus_colihominis
Bacteroides_caccae	Lachnospiraceae_bacterium_5_1_63FAA
Turicibacter_unclassified	Bacteroides_cellulosilyticus
Bacteroides_sp_4_3_47FAA	Porphyromonas_asaccharolytica
Blautia_hansenii	Alistipes_finegoldii
Clostridium_clostridioforme	Facklamia_unclassified
Rhodococcus_erythropolis	Staphylococcus_phage_phiETA3
Clostridium_citroniae	Cupriavidus_unclassified
Ruminococcus_albus	Chicory_yellow_mottle_virus_large_satellite_RNA
Bacteroides_eggerthii	Erysipelotrichaceae_bacterium_21_3
	Gemella_unclassified
	Porphyromonas_uenonis
	Classical_swine_fever_virus
	Eubacterium_infirmum
	Enterovirus_B
	Escherichia_unclassified
	Eubacterium_sp_3_1_31
	Coprobacillus_unclassifi

ed

Bovine_viral_diarrhea_virus_2

Fusobacterium_nucleatum

Coprobacter_fastidiosus

Coprobacillus_sp_D6

Clostridiales_bacterium_1_7_47FAA

Ruminococcus_sp_JC304

Bacteroides_dorei

Mycoplasma_hominis

Actinomyces_urogenitalis

Bovine_viral_diarrhea_virus_3

Royal_Farm_virus

Prevotella_timonensis

Pseudomonas_thermotolerans

Coprococcus_sp_ART55_1

Bacteroides_pectinophilus

Blautia_hydrogenotrophica

Dorea_unclassified

Collinsella_aerofaciens

Sutterella_wadsworthensis

Eubacterium_biforme

Clostridiales_bacterium_BV3C26

Clostridium_sp_KLE_1755

Melon_aphid_borne_yell

ows_virus

Veillonella_dispar

Bacteroides_salyersiae

Clostridium_bartlettii

Lachnospiraceae_bacterium_9_1_43BFAA

Campylobacter_ureolyticus

Lactobacillus_antri

Montana_myotis_leukoencephalitis_virus

Entamoeba_dispar

Leuconostoc_gelidum

Caulobacter_unclassified

Fusobacterium_periodonticum

Coriobacteriaceae_bacterium_BV3Ac1

Cardiobacterium_valvarum

Bacteroides_fragilis

Bifidobacterium_pseudolongum

Porphyromonas_bennonis

Orthohepadnavirus_unclassified

Ruminococcus_gnavus

Alistipes_sp_AP11

Enterovirus_D

Lactobacillus_mucosae

Streptococcus_macedonicus

Bacteroides_thetaiotaomicron

Coprobacillus_sp_29_1

									Chicken_anemia_virus
									Lachnospiraceae_bacterium_1_4_56FAA
									Eggerthella_lenta
									Cetobacterium_somerae
									Circovirus_like_genome_SAR_A
									Granulicatella_elegans
									Ruminococcus_flavefaciens
									Parascardovia_denticolens
									Pseudomonas_phage_MP38
									Bilophila_unclassified
									Actinomyces_viscosus
									Actinomyces_naeslundii
									Streptococcus_agalactiae
									Salmon_pancreas_disease_virus
									C2likevirus_unclassified

Note: Top 30 gut microbiota species in dataset 1; top 100 gut microbiota species in dataset 2. Top 5 host characteristics features and top 5 dietary features in datasets 1 and 2.

Supplementary Table 5. Important factors associated with blood butyric acids.

Gut Microbiotas		Host Characteristics		Dietary Features	
Dataset 1	**Dataset 2**	**Dataset 1**	**Dataset 2**	**Dataset 1**	**Dataset 2**
Ruminococcus_torques	Ruminococcaceae_bacterium_D16	Weight (kg)	Gender	Frequency of eating pickles	Whether you eat yogurt
Subdoligranulum_unclassified	Nilaparvata_lugens_honeydew_virus_2	Whether you have you taken any probiotics in the past 12 months (Yes/No)	Race	Quantity of eating cold cereals in a day	Whether you drink milk
Roseburia_intestinalis	Bordetella_unclassified	Whether you are currently taking any gastric acid lowering medications (Yes/No)	Weight (kg)	Quantity of eating chocolate candy in a day	Whether you eat cheese
Oscillibacter_unclassified	Providencia_unclassified	Height (cm)	BMI	Frequency of eating steak	
Bifidobacterium_catenulatum	Weissella_confusa	Whether you have used antibiotics in the past 12 months (Yes/No)	Whether you have you taken any probiotics in the past 12 months (Yes/No)	Quantity of eating steak in a day	
Bacteroides_thetaiotaomicron	Coprococcus_sp_ART55_1				
Parabacteroides_merdae	Coriobacteriaceae_bacterium_BV3Ac1				
Turicibacter_unclassified	Eubacterium_infirmum				
Bacteroides_caccae	Corynebacterium_durum				
Bacteroides_uniformis	Apoi_virus				
Prevotella_copri	Alistipes_finegoldii				
Rothia_mucilaginosa	Plectrovirus_unclassified				
Erysipelotrichaceae_bacterium_6_1_45	Actinomyces_turicensis				
Erysipelotrichaceae_bacterium_6_1_45	Anaerostipes_caccae				
Blautia_hansenii	Alistipes_indistinctus				
Ruminococcus_sp_5_1_39BFAA	Parabacteroides_johnsonii				

Alistipes_senegalensis	Erysipelotrichaceae_bacterium_21_3
Desulfovibrio_piger	Eggerthella_unclassified
Bacteroides_vulgatus	Lactobacillus_animalis
Veillonella_unclassified	Lactobacillus_vaginalis
Streptococcus_salivarius	Campylobacter_ureolyticus
Lactobacillus_plantarum	Atopobium_parvulum
Rhodococcus_erythropolis	Enterococcus_casseliflavus
Eubacterium_limosum	Leuconostoc_pseudomesenteroides
Faecalibacterium_prausnitzii	Megasphaera_micronuciformis
Ruminococcus_albus	Desulfovibrio_termitidis
Oscillibacter_sp_KLE_1745	Enterococcus_avium
Propionibacterium_freudenreichii	Peptostreptococcus_unclassified
Clostridium_asparagiforme	Lachnospiraceae_bacterium_5_1_63FAA
Weissella_unclassified	Clostridium_sp_HGF2
	Quail_picornavirus_QPV1_HUN_2010
	Leuconostoc_carnosum
	Dialister_succinatiphilus
	Prevotella_bergensis
	Actinobacillus_unclassified
	Ruminococcus_obeum
	Bifidobacterium_pseudocatenulatum
	Alistipes_sp_AP11

Collinsella_stercoris

Porphyromonas_uenonis

Streptococcus_tigurinus

Turicibacter_sanguinis

Actinomyces_massiliensis

Bacteroides_intestinalis

Anaerostipes_sp_3_2_56FAA

Cupriavidus_unclassified

Weissella_paramesenteroides

Holdemania_unclassified

Clostridium_hathewayi

Streptococcus_salivarius

Chicory_yellow_mottle_virus_large_satellite_RNA

Leuconostoc_mesenteroides

Odoribacter_splanchnicus

Culex_flavivirus

Clostridiales_bacterium_BV3C26

Blautia_hydrogenotrophica

Clostridiaceae_bacterium_JC118

Clostridium_glycolicum

Clostridium_sp_KLE_1755

Bifidobacterium_pseudolongum

Collinsella_aerofaciens

Veillonella_unclassified

Leuconostoc_inhae

Dysgonomonas_unclassifi

ed

Okra_yellow_crinkle_Cameroon_alphasatellite

Streptococcus_australis

Lactobacillus_helveticus

Bovine_viral_diarrhea_virus_2

Rhodopseudomonas_palustris

Bacteroides_pectinophilus

Bacteroides_coprocola

Passion_fruit_woodiness_virus

Slackia_unclassified

Prevotella_timonensis

Bombyx_mori_nucleopolyhedrovirus

Hepatitis_C_virus

Fusobacterium_necrophorum

Anaerococcus_vaginalis

Avian_carcinoma_virus

Clostridium_bifermentans

Gastropod_associated_circular_ssDNA_virus

Actinomyces_urogenitalis

Olsenella_unclassified

Bilophila_wadsworthia

Hafnia_alvei

Oat_blue_dwarf_virus

Dorea_formicigenerans

Akkermansia_muciniphila

Hepatitis_A_virus

Streptococcus_mutans

	Lactobacillus_salivarius
	Lactobacillus_casei_paracasei
	Pepper_vein_yellows_virus
	Oscillibacter_sp_KLE_1745
	Bacteroidales_bacterium_ph8
	Mycoplasma_hominis
	Ruminococcus_bromii
	Pseudomonas_thermotolerans
	Cardiobacterium_valvarum
	Lindernia_anagallis_yellow_vein_virus_satellite_DNA_beta

Note: Top 30 gut microbiota species in dataset 1 ; top 100 gut microbiota species in dataset 2. Top 5 host characteristics features and top 5 dietary features in datasets 1 and 2.

649
650

Supplementary Table 6. Important factors associated with blood hexanoic acids.

Gut Microbiotas		**Host Characteristics**		**Dietary Features**	
Dataset 1	**Dataset 2**	**Dataset 1**	**Dataset 2**	**Dataset 1**	**Dataset 2**
Subdoligranulum_unclassified	Ruminococcus_bromii	Weight (kg)	Consent_Age	Quantity of eating steak in a day	Whether you eat yogurt
Ruminococcus_torques	Erysipelotrichaceae_bacterium_21_3	Frequency of heavy work	Gender	Quantity of eating eggs in a day	Whether you drink milk
Oscillibacter_unclassified	**Faecalibacterium_prausnitzii**	Height (cm)	Height (cm)	Frequency of eating steak	Whether you eat cheese
Alistipes_unclassified	Lachnospiraceae_bacterium_5_1_63FAA	Exercise_Regular	Race	Frequency of eating pickles	
Clostridium_asparagiforme	Ruminococcaceae_bacterium_D16	Time in heavy work for each time	Whether you have you taken any probiotics in the past 12 months (Yes/No)	Frequency of taking fat oil	
Bacteroides_uniformis	Desulfovibrio_termitidis				
Desulfovibrio_piger	Collinsella_stercoris				
Rothia_mucilaginosa	Coriobacteriaceae_bacterium_BV3Ac1				
Ruminococcus_sp_5_1_39BFAA	candidate_division_TM7_single_cell_isolate_TM7b				
Clostridium_leptum	Lactobacillus_phage_PL_1				
Bacteroides_thetaiotaomicron	Clostridium_sp_HGF2				
Roseburia_intestinalis	Dorea_formicigenerans				
Lactobacillus_salivarius	Classical_swine_fever_virus				
Prevotella_copri	Solobacterium_moorei				
Clostridium_citroniae	Lactobacillus_casei_paracasei				
Bacteroides_nordii	Gemella_unclassified				
Faecalibacterium_prausnitzii	Actinomyces_naeslundii				
Parabacteroides_unclassified	Turicibacter_sanguinis				
Bifidobacterium_cate	Lactobacillus_saerimneri				

nulatum	
Bacteroides_sp_4_3_47FAA	Ruminococcus_obeum
Clostridium_bolteae	Bordetella_unclassified
Clostridium_clostridioforme	Lindernia_anagallis_yellow_vein_virus_satellite_DNA_beta
Acidaminococcus_sp_BV3L6	Enterovirus_B
Erysipelotrichaceae_bacterium_6_1_45	Lachnospiraceae_bacterium_4_1_37FAA
Paraprevotella_clara	Clostridiaceae_bacterium_JC118
Bacteroides_fragilis	Roseburia_inulinivorans
Bacteroides_dorei	Weissella_confusa
Leuconostoc_gelidum	**Acidaminococcus_sp_BV3L6**
Bacteroides_eggerthii	Ruminococcus_sp_JC304
Blautia_hansenii	Clostridium_bifermentans
	Chicory_yellow_mottle_virus_large_satellite_RNA
	Campylobacter_ureolyticus
	Oat_blue_dwarf_virus
	Streptococcus_vestibularis
	Coriobacteriaceae_bacterium_phI
	Enterobacteria_phage_SfV
	Facklamia_unclassified
	Actinomyces_viscosus
	Ruminococcus_champanellensis
	Actinomyces_massiliensis
	Lactobacillus_animalis
	Anaerostipes_caccae
	Quail_picornavirus_QPV1_HUN_2010
	West_Nile_virus

Atopobium_vaginae

Lactococcus_lactis

Bacteroides_intestinalis

Sida_yellow_vein_Vietnam_virus_satellite_DNA_beta

Streptococcus_parasanguinis

Enterococcus_avium

Rothia_dentocariosa

Bovine_rhinitis_B_virus

Weissella_paramesenteroides

Bovine_viral_diarrhea_virus_2

Barbel_circovirus

Parvimonas_unclassified

Dysgonomonas_unclassified

Ruminococcus_sp_5_1_39BFAA

Fusobacterium_mortiferum

Gastropod_associated_circular_ssDNA_virus

Parascardovia_denticolens

Erysipelotrichaceae_bacterium_5_2_54FAA

Lactobacillus_vaginalis

Neisseria_flavescens

Phascolarctobacterium_succinatutens

Entamoeba_dispar

Circovirus_like_genome_SAR_A

Sutterella_wadsworthensis

Human_cosavirus_B

Providencia_unclassified

Leuconostoc_carnosum

Peptostreptococcus_anaerobius

Kelp_fly_virus
Fusobacterium_necrophorum
Hepatitis_GB_virus_B
Pennisetum_mosaic_virus
Gemella_haemolysans
Blautia_hydrogenotrophica
Bacteroides_cellulosilyticus
Eubacterium_infirmum
Bilophila_wadsworthia
Petunia_vein_clearing_virus
Porphyromonas_asaccharolytica
Clostridiales_bacterium_BV3C26
Holdemania_unclassified
Lactococcus_phage_jm2
Klebsiella_phage_KP36
Plectrovirus_unclassified
Bifidobacterium_breve
Parabacteroides_sp_20_3
Okra_yellow_crinkle_Cameroon_alphasatellite
Border_disease_virus
Melon_aphid_borne_yellows_virus
Oligella_urethralis
Lactobacillus_antri
Candidatus_Zinderia_insecticola
Haemophilus_sputorum
Malvastrum_leaf_curl_Philippines_betasatellite
Weissella_cibaria
Actinomyces_urogenitalis

651 Note: Top 30 gut microbiota species in dataset 1 ; top 100 gut microbiota species in dataset 2. Top 5 host characteristics features and top 5 dietary features in
652 datasets 1 and 2.
653

Supplementary Table 7. Important factors associated with blood 2-methylbutyric acids.

Gut Microbiotas		**Host Characteristics**		**Dietary Features**	
Dataset 1	**Dataset 2**	**Dataset 1**	**Dataset 2**	**Dataset 1**	**Dataset 2**
Subdoligranulum_unclassified	candidate_division_TM7_single_cell_isolate_TM7b	Weight (kg)	Race	Frequency of eating steak	Whether you drink milk
Ruminococcus_torques	Ruminococcus_obeum	Time in bicycling or swimming for each time	Gender	Frequency of eating pickles	Whether you eat yogurt
Bacteroides_uniformis	Weissella_confusa	Height (cm)	BMI	Quantity of eating eggs in a day	Whether you eat cheese
Bifidobacterium_catenulatum	Lactobacillus_phage_PL_1	Age	Whether you have you taken any probiotics in the past 12 months (Yes/No)	Frequency of taking fat oil	
Oscillibacter_unclassified	**Faecalibacterium_prausnitzii**	Race	Weight (kg)	Quantity of eating steak in a day	
Clostridium_asparagiforme	Bovine_viral_diarrhea_virus_2				
Ruminococcus_albus	Anaerostipes_unclassified				
Alistipes_senegalensis	Coriobacteriaceae_bacterium_BV3Ac1				
Alistipes_unclassified	Okra_yellow_crinkle_Cameroon_alphasatellite				
Roseburia_intestinalis	Klebsiella_phage_KP36				
Veillonella_unclassified	Candidatus_Zinderia_insecticola				
Faecalibacterium_prausnitzii	Bordetella_unclassified				
Bacteroides_thetaiotaomicron	Cardiobacterium_valvarum				
Prevotella_copri	Eubacterium_siraeum				
Rothia_mucilaginosa	Zinnia_leaf_curl_virus_associated_DNA_beta				
Acidaminococcus_sp_BV3L6	Granulicella_unclassified				
Clostridium_leptum	Bifidobacterium_pseudolongum				
Desulfovibrio_piger	Clostridium_sp_KLE_1755				

Erysipelotrichaceae_bacterium_6_1_45	Ruminococcus_champanellensis
Ruminococcus_sp_5_1_39BFAA	Orthohepadnavirus_unclassified
Lactobacillus_salivarius	Campylobacter_ureolyticus
Clostridium_scindens	Parabacteroides_sp_20_3
Weissella_unclassified	Streptococcus_agalactiae
Paraprevotella_clara	Dysgonomonas_mossii
Bacteroides_eggerthii	Dorea_formicigenerans
Bacteroides_fragilis	Alloprevotella_tannerae
Bacteroides_sp_4_3_47FAA	Aggregatibacter_segnis
Roseburia_hominis	Borrelia_duttonii
Bacteroides_nordii	Granulicatella_unclassified
Bacteroides_caccae	Holdemania_sp_AP2
	Gemella_haemolysans
	Enterobacteria_phage_SfV
	Streptococcus_vestibularis
	Sapporo_virus
	Streptococcus_mutans
	Anaerotruncus_colihominis
	Erysipelotrichaceae_bacterium_21_3
	Coprobacter_fastidiosus
	Eubacterium_ramulus
	Haemophilus_pittmaniae
	Olsenella_unclassified
	Alistipes_sp_HGB5
	Actinomyces_urogenitalis
	Lachnospiraceae_bacterium_5_1_63FAA
	Alloprevotella_unclassified

Pseudomonas_phage_MP38
Lactobacillus_crispatus
Prevotella_amnii
Bacteroides_coprocola
Anaerococcus_vaginalis
Enterococcus_hirae
Eubacterium_biforme
Barbel_circovirus
Cetobacterium_somerae
Butyrivibrio_crossotus
Peptostreptococcus_anaerobius
Bovine_viral_diarrhea_virus_3
Pediococcus_lolii
Helicobasidium_mompa_endornavirus_1
Culex_flavivirus
Leuconostoc_gasicomitatum
Oligella_urethralis
Bacteroides_dorei
Bovine_rhinitis_B_virus
Brachyspira_unclassified
Streptococcus_gordonii
Abiotrophia_defectiva
Lachnospiraceae_bacterium_4_1_37FAA
Alistipes_finegoldii
Desulfovibrio_termitidis
Lachnospiraceae_bacterium_9_1_43BFAA
Clostridium_bifermentans

Entamoeba_dispar

Lachnospiraceae_bacterium_8_1_57FAA

Infectious_flacherie_virus

Bifidobacterium_breve

Lactobacillus_fermentum

Lactobacillus_gasseri

Rothia_mucilaginosa

Pseudomonas_aeruginosa

Bacteroides_fragilis

Clostridium_sp_HGF2

Apoi_virus

Staphylococcus_phage_phiETA3

Turicibacter_sanguinis

Lactobacillus_saerimneri

Erysipelotrichaceae_bacterium_6_1_45

Sepik_virus

Porphyromonas_asaccharolytica

Clostridium_difficile

Streptococcus_macedonicus

Gastropod_associated_circular_ssDNA_virus

Subdoligranulum_variabile

Hippeastrum_mosaic_virus

Prevotella_bergensis

Colombian_datura_virus

Dorea_unclassified

Royal_Farm_virus

Enterovirus_B

	Blautia_hydrogenotrophica

Note: Top 30 gut microbiota species in dataset 1 ; top 100 gut microbiota species in dataset 2. Top 5 host characteristics features and top 5 dietary features in datasets 1 and 2.

Supplementary Table 8. Important factors associated with blood valeric acids.

Gut Microbiotas		**Host Characteristics**		**Dietary Features**	
Dataset 1	**Dataset 2**	**Dataset 1**	**Dataset 2**	**Dataset 1**	**Dataset 2**
Subdoligranulum_unclassified	Ruminococcus_obeum	Weight (kg)	Gender	Frequency of eating pickles	Whether you drink milk
Ruminococcus_torques	Weissella_confusa	Whether you have you taken any gastric acid lowering medications in the past 12 months (Yes/No)	Race	Quantity of eating chocolate candy in a day	Whether you eat yogurt
Oscillibacter_unclassified	Epsilon15likevirus_unclassified	Whether you have you taken any probiotics in the past 12 months (Yes/No)	Height (cm)	Quantity of eating fruit	Whether you eat cheese
Rothia_mucilaginosa	**Faecalibacterium_prausnitzii**	Time in heavy work for each time	Weight (kg)	Quantity of eating cold cereals in a day	
Alistipes_unclassified	Classical_swine_fever_virus	Frequency of bicycling or swimming	BMI	Frequency of eating steak	
Desulfovibrio_piger	Campylobacter_ureolyticus				
Clostridium_asparagiforme	Desulfovibrio_termitidis				
Bacteroides_thetaiotaomicron	Clostridium_sp_HGF2				
Bacteroides_uniformis	Bovine_rhinitis_B_virus				
Bacteroides_nordii	C2likevirus_unclassified				
Clostridium_bolteae	candidate_division_TM7_single_cell_isolate_TM7b				
Parabacteroides_unclassified	Eubacterium_sp_3_1_31				
Bacteroides_dorei	Corynebacterium_propinquum				
Clostridium_citroniae	Gemella_unclassified				
Collinsella_unclassified	Prevotella_bergensis				
Clostridium_clostridioforme	Neisseria_flavescens				
Bacteroides_fragilis	Cyclovirus_NGchicken15_NGA_2009				

Faecalibacterium_prausnitzii	Okra_yellow_crinkle_Cameroon_alphasatellite
Clostridium_leptum	Alloprevotella_unclassified
Eubacterium_ramulus	Lachnospiraceae_bacterium_5_1_63FAA
Bacteroides_sp_4_3_47FAA	Peptostreptococcus_anaerobius
Actinomyces_viscosus	Bacteroides_vulgatus
Ruminococcus_sp_5_1_39BFAA	Actinomyces_urogenitalis
Parabacteroides_merdae	Ruminococcus_champanellensis
Leuconostoc_gelidum	Lactobacillus_phage_PL_1
Prevotella_copri	Gastropod_associated_circular_ssDNA_virus
Lactobacillus_salivarius	Lactobacillus_antri
Escherichia_unclassified	Corynebacterium_jeikeium
Bacteroides_eggerthii	Coprococcus_comes
Erysipelotrichaceae_bacterium_6_1_45	Bordetella_unclassified
	Rhodospirillum_unclassified
	Neisseria_unclassified
	Streptococcus_parasanguinis
	JC_polyomavirus
	Prevotella_bivia
	Lachnospiraceae_bacterium_4_1_37FAA
	Pseudomonas_alcaligenes
	Bilophila_wadsworthia
	Pseudomonas_phage_MP38

Parabacteroides_sp_20_3
Stomatobaculum_longum
Lactobacillus_saerimneri
Slow_bee_paralysis_virus
Parvimonas_unclassified
Providencia_unclassified
Enterovirus_B
Megasphaera_genomosp_type_1
Alistipes_putredinis
Cetobacterium_somerae
Pepper_vein_yellows_virus
Actinobacillus_unclassified
Clostridiales_bacterium_1_7_47FAA
Alistipes_sp_HGB5
Porphyromonas_uenonis
Eubacterium_cylindroides
Bifidobacterium_longum
Enterobacter_cloacae
Enterobacteria_phage_HK225
Bifidobacterium_breve
Bacteroides_coprocola
Streptococcus_gordonii
Atopobium_parvulum
Blautia_producta
Atopobium_vaginae
Bean_common_mosaic_necrosis_virus
Melon_aphid_borne_yellows_virus

Morganella_morganii

Streptococcus_mutans

Pestivirus_strain_Aydin_04_TR

Streptococcus_tigurinus

Lachnospiraceae_bacterium_8_1_57FAA

Rothia_dentocariosa

Bacteroides_cellulosilyticus

Clostridium_sp_KLE_1755

Solobacterium_moorei

Porphyromonas_asaccharolytica

Leuconostoc_gelidum

Veillonella_unclassified

Clostridium_difficile

Orthohepadnavirus_unclassified

Clostridium_hathewayi

Hippeastrum_mosaic_virus

Streptococcus_gallolyticus

Eubacterium_dolichum

Collinsella_stercoris

Alloprevotella_tannerae

Acidaminococcus_sp_D21

Dorea_unclassified

Fusobacterium_nucleatum

Candida_glabrata

Anaerotruncus_colihominis

Anaerococcus_prevotii

					Lactobacillus_sakei
					Coprobacillus_unclassified
					Hepatitis_GB_virus_B
					Helicobasidium_mompa_endornavirus_1
					Fig_fleck_associated_virus
					Prevotella_timonensis
					Cardiobacterium_valvarum
					Collinsella_aerofaciens

Note: Top 30 gut microbiota species in dataset 1 ; top 100 gut microbiota species in dataset 2. Top 5 host characteristics features and top 5 dietary features in datasets 1 and 2.

Supplementary Table 9. Important factors associated with blood propionic acids.

Gut Microbiotas		**Host Characteristics**		**Dietary Features**	
Dataset 1	**Dataset 2**	**Dataset 1**	**Dataset 2**	**Dataset 1**	**Dataset 2**
Subdoligranulum_unclassified	Weissella_confusa	Whether you have you taken any probiotics in the past 12 months (Yes/No)	Gender	Frequency of eating pickles	Whether you eat yogurt
Ruminococcus_torques	Ruminococcaceae_bacterium_D16	Weight (kg)	Race	Frequency of eating steak	Whether you drink milk
Oscillibacter_unclassified	Desulfovibrio_termitidis	Whether you have used antibiotics in the past 12 months (Yes/No)	BMI	Quantity of eating steak in a day	Whether you eat cheese
Bacteroides_uniformis	Ruminococcus_obeum	Age	Consent_Age	Frequency of drinking non-milky coffee	
Roseburia_intestinalis	Lachnospiraceae_bacterium_5_1_63FAA	Height (cm)	Weight (kg)	Quantity of eating chocolate candy in a day	
Bacteroides_thetaiotaomicron	Providencia_unclassified				
Bifidobacterium_catenulatum	Cyclovirus_NGchicken15_NGA_2009				
Rothia_mucilaginosa	Enterovirus_B				
Alistipes_unclassified	Porphyromonas_uenonis				
Desulfovibrio_piger	Turicibacter_sanguinis				
Prevotella_copri	Lactococcus_garvieae				
Ruminococcus_sp_5_1_39BFAA	Ruminococcus_champanellensis				
Clostridium_asparagiforme	Classical_swine_fever_virus				
Blautia_hansenii	Prevotella_nanceiensis				
Faecalibacterium_prausnitzii	Culex_flavivirus				
Acidaminococcus_sp_BV3L6	Epsilon15likevirus_unclassified				
Turicibacter_unclassified	Enterobacteria_phage_HK225				
Parabacteroides_merdae	Bordetella_unclassified				
Erysipelotrichaceae_bacterium_6_1_45	Actinobacillus_unclassified				

Bacteroides_caccae	Prevotella_bergensis
Lactobacillus_salivarius	Olsenella_unclassified
Ruminococcus_albus	**Faecalibacterium_prausnitzii**
Clostridium_citroniae	Melon_aphid_borne_yellows_virus
Alistipes_senegalensis	Peptostreptococcus_anaerobius
Bacteroides_nordii	Streptococcus_gallolyticus
Bacteroides_eggerthii	Collinsella_aerofaciens
Bacteroides_sp_4_3_47FAA	Hafnia_alvei
Clostridium_leptum	Streptococcus_mutans
Veillonella_unclassified	candidate_division_TM7_single_cell_isolate_TM7b
Bacteroides_vulgatus	Dorea_formicigenerans
	Eubacterium_sp_3_1_31
	Parabacteroides_johnsonii
	Odoribacter_splanchnicus
	Lactobacillus_helveticus
	Bacteroides_phage_B124_14
	Clostridium_hathewayi
	Clostridium_bifermentans
	Corynebacterium_jeikeium
	Bacteroides_pectinophilus
	Blautia_hydrogenotrophica
	Eubacterium_ramulus
	Enterobacteriaceae_bacterium_9_2_54FAA
	Gastropod_associated_circular_ssDNA_virus
	Lactobacillus_phage_PL_1
	Fig_fleck_associated_virus
	Hippeastrum_mosaic_virus
	Streptococcus_gordonii
	Anaerotruncus_colihominis

Clostridium_sp_HGF2

Colombian_datura_virus

Streptococcus_salivarius

C2likevirus_unclassified

Hepatitis_GB_virus_B

Streptococcus_tigurinus

Plectrovirus_unclassified

Campylobacter_ureolyticus

Avian_carcinoma_virus

Chlorobium_phaeobacteroides

Coriobacteriaceae_bacterium_BV3Ac1

Clostridium_nexile

Clostridium_difficile

Gemella_unclassified

Weissella_paramesenteroides

Actinomyces_turicensis

Stomatobaculum_longum

Chicory_yellow_mottle_virus_large_satellite_RNA

Okra_yellow_crinkle_Cameroon_alphasatellite

Prevotella_buccalis

Lachnospiraceae_bacterium_8_1_57FAA

Bilophila_wadsworthia

Mobiluncus_curtisii

Corynebacterium_amycolatum

Veillonella_unclassified

Lachnospiraceae_bacterium_7_1_58FAA

Collinsella_stercoris

Pestivirus_strain_Aydin_04_TR

	Ruminococcus_lactaris
	Anaerococcus_prevotii
	Bacteroides_ovatus
	Clostridium_perfringens
	Alistipes_indistinctus
	Actinomyces_urogenitalis
	Bifidobacterium_pseudolongum
	Bifidobacterium_longum
	Neisseria_flavescens
	Parabacteroides_sp_20_3
	Facklamia_unclassified
	Erysipelotrichaceae_bacterium_21_3
	Streptococcus_phage_ALQ13_2
	Human_cosavirus_B
	Parsnip_yellow_fleck_virus
	Bovine_rhinitis_B_virus
	Sida_yellow_vein_Vietnam_virus_satellite_DNA_beta
	Leuconostoc_mesenteroides
	Ruminococcus_bromii
	Alistipes_sp_HGB5
	Roseburia_inulinivorans
	Clostridiales_bacterium_BV3C26
	Candida_glabrata
	Slackia_unclassified

Note: Top 30 gut microbiota species in dataset 1 ; top 100 gut microbiota species in dataset 2. Top 5 host characteristics features and top 5 dietary features in datasets 1 and 2.

Supplementary Table 10. Important factors associated with blood isobutyric acids.

Gut Microbiotas		**Host Characteristics**		**Dietary Features**	
Dataset 1	**Dataset 2**	**Dataset 1**	**Dataset 2**	**Dataset 1**	**Dataset 2**
Subdoligranulum_unclassified	candidate_division_TM7_single_cell_isolate_TM7b	Height (cm)	Race	Frequency of eating pickles	Whether you drink milk
Ruminococcus_torques	Weissella_confusa	Weight (kg)	Gender	Quantity of eating steak in a day	Whether you eat yogurt
Oscillibacter_unclassified	Ruminococcus_obeum	Time in bicycling or swimming for each time	Height (cm)	Frequency of eating steak	Whether you eat cheese
Bacteroides_uniformis	Streptococcus_mutans	Age	Whether you have you taken any probiotics in the past 12 months (Yes/No)	Quantity of eating eggs in a day	
Clostridium_asparagiforme	Campylobacter_ureolyticus	Time in heavy work for each time	BMI	Frequency of drinking non-milky coffee	
Alistipes_unclassified	Clostridium_sp_HGF2				
Rothia_mucilaginosa	Corynebacterium_jeikeium				
Bacteroides_thetaiotaomicron	Eubacterium_sp_3_1_31				
Faecalibacterium_prausnitzii	Desulfovibrio_termitidis				
Bifidobacterium_catenulatum	Bifidobacterium_longum				
Roseburia_intestinalis	Bacteroides_vulgatus				
Desulfovibrio_piger	**Faecalibacterium_prausnitzii**				
Clostridium_leptum	Epsilon15likevirus_unclassified				
Ruminococcus_albus	Ruminococcus_champanellensis				
Ruminococcus_sp_5_1_39BFAA	Bacteroides_coprocola				
Prevotella_copri	Bordetella_unclassified				
Alistipes_senegalen	Providencia_unclassified				

sis	
Lactobacillus_salivarius	Classical_swine_fever_virus
Bacteroides_eggerthii	Bovine_rhinitis_B_virus
Erysipelotrichaceae_bacterium_6_1_45	Cetobacterium_somerae
Acidaminococcus_sp_BV3L6	Okra_yellow_crinkle_Cameroon_alphasatellite
Bacteroides_nordii	Lactobacillus_phage_PL_1
Veillonella_unclassified	Enterobacteria_phage_SfV
Bacteroides_sp_4_3_47FAA	Parabacteroides_sp_20_3
Bacteroides_caccae	Cyclovirus_NGchicken15_NGA_2009
Clostridium_citroniae	Prevotella_bergensis
Blautia_hansenii	Rhodospirillum_unclassified
Clostridium_bolteae	Enterobacteria_phage_HK225
Bacteroides_fragilis	C2likevirus_unclassified
Roseburia_inuliniv orans	Pseudomonas_alcaligenes
	Neisseria_flavescens
	Corynebacterium_propinquum
	Coriobacteriaceae_bacterium_BV3Ac1
	Peptostreptococcus_anaerobius
	Eubacterium_siraeum
	Bilophila_wadsworthia
	Circovirus_like_genome_SAR_A
	Bifidobacterium_pseudolongum
	Clostridium_hathewayi

Alistipes_sp_HGB5

Leuconostoc_gasicomitatum

Enterovirus_B

Melon_aphid_borne_yellows_virus

Anaerotruncus_colihominis

Eubacterium_cylindroides

Bifidobacterium_breve

Dysgonomonas_unclassified

Fusobacterium_nucleatum

JC_polyomavirus

Prevotella_buccalis

Cardiobacterium_valvarum

Actinobacillus_unclassified

Erysipelotrichaceae_bacterium_21_3

Lactobacillus_antri

Facklamia_unclassified

Lachnospiraceae_bacterium_5_1_63FAA

Prevotella_bivia

Gemella_haemolysans

Clostridium_difficile

Solobacterium_moorei

Pepper_vein_yellows_virus

Veillonella_unclassified

Eubacterium_ramulus

Turicibacter_sanguinis

Prevotella_timonensis

Lindernia_anagallis_yellow_vein_virus_satellite_DNA_beta

Enterococcus_casseliflavus

Royal_Farm_virus

Collinsella_stercoris

Human_cosavirus_B

Megasphaera_unclassified

Atopobium_parvulum

Clostridium_ramosum

Atopobium_vaginae

Fig_fleck_associated_virus

Marvinbryantia_formatexigen
s

Collinsella_aerofaciens

Acidaminococcus_sp_D21

Eubacterium_dolichum

Leuconostoc_gelidum

Haemophilus_pittmaniae

Prevotella_nanceiensis

Dorea_unclassified

Lactococcus_raffinolactis

Clostridium_bolteae

Helicobasidium_mompa_endo
rnavirus_1

Lactobacillus_sakei

Lactobacillus_vaginalis

Porphyromonas_uenonis

Pestivirus_strain_Aydin_04_T
R

Anaerostipes_unclassified

Eggerthella_sp_1_3_56FAA

Oligella_urethralis

Lachnospiraceae_bacterium_8
_1_57FAA

Halomonas_unclassified

Actinomyces_turicensis

Gastropod_associated_circular

	_ssDNA_virus
	Streptococcus_vestibularis
	Hippeastrum_mosaic_virus
	Eubacterium_brachy

Note: Top 30 gut microbiota species in dataset 1 ; top 100 gut microbiota species in dataset 2. Top 5 host characteristics features and top 5 dietary features in datasets 1 and 2.

Supplementary Table 11. Important factors associated with blood isovaleric acids.

Gut Microbiotas		**Host Characteristics**		**Dietary Features**	
Dataset 1	**Dataset 2**	**Dataset 1**	**Dataset 2**	**Dataset 1**	**Dataset 2**
Subdoligranulum_unclassified	Desulfovibrio_termitidis	Height (cm)	Gender	Frequency of eating pickles	Whether you eat yogurt
Ruminococcus_torques	Ruminococcaceae_bacterium_D16	Weight (kg)	Race	Frequency of eating steak	Whether you drink milk
Oscillibacter_unclassified	Lactobacillus_phage_PL_1	Time in bicycling or swimming for each time	Height (cm)	Frequency of drinking non-milky coffee	Whether you eat cheese
Bacteroides_uniformis	Bovine_viral_diarrhea_virus_2	Time in heavy work for each time	Weight (kg)	Quantity of eating steak in a day	
Rothia_mucilaginosa	Lachnospiraceae_bacterium_5_1_63FAA	Whether you have you taken any probiotics in the past 12 months (Yes/No)	BMI	Quantity of eating chocolate candy in a day	
Clostridium_asparagiforme	Clostridium_sp_KLE_1755				
Bifidobacterium_catenulatum	Lactobacillus_saerimneri				
Bacteroides_thetaiotaomicron	Alloprevotella_tannerae				
Roseburia_intestinalis	Okra_yellow_crinkle_Cameroon_alphasatellite				
Alistipes_unclassified	Dorea_formicigenerans				
Desulfovibrio_piger	Ruminococcus_obeum				
Faecalibacterium_prausnitzii	Porphyromonas_asaccharolytica				
Bacteroides_nordii	Turicibacter_sanguinis				
Prevotella_copri	Peptostreptococcus_unclassified				
Acidaminococcus_sp_BV3L6	Alloprevotella_unclassified				
Ruminococcus_albus	Eubacterium_ramulus				
Alistipes_senegalensis	Bacteroides_ovatus				
Clostridium_leptum	Coriobacteriaceae_bacterium_BV3Ac1				
Ruminococcus_sp_5_1_39BFAA	Avian_carcinoma_virus				
Bacteroides_eggerth	Weissella_confusa				

Veillonella_unclassified	Fig_fleck_associated_virus
Parabacteroides_merdae	Streptococcus_salivarius
Bacteroides_caccae	Ruminococcus_champanellensis
Erysipelotrichaceae_bacterium_6_1_45	**Faecalibacterium_prausnitzii**
Clostridium_citroniae	Slow_bee_paralysis_virus
Blautia_hansenii	Olsenella_unclassified
Bacteroides_sp_4_3_47FAA	Klebsiella_pneumoniae
Lactobacillus_salivarius	Pestivirus_strain_Aydin_04_TR
Bacteroides_fragilis	Enterococcus_avium
Weissella_unclassified	Culex_flavivirus
	Campylobacter_ureolyticus
	candidate_division_TM7_single_cell_isolate_TM7b
	Apoi_virus
	Odoribacter_unclassified
	Barbel_circovirus
	Enterobacteria_phage_SfV
	Lactococcus_lactis
	Weissella_unclassified
	Lactobacillus_delbrueckii
	Holdemania_unclassified
	Actinomyces_urogenitalis
	Clostridium_sp_HGF2
	Bilophila_wadsworthia
	Aggregatibacter_segnis
	Actinobacillus_unclassified

Valsa_ceratosperma_hypovirus_1

Bacteroides_eggerthii

Campylobacter_concisus

Proteus_mirabilis

Nilaparvata_lugens_honeydew_virus_2

Zinnia_leaf_curl_virus_associated_DNA_beta

Orthohepadnavirus_unclassified

Klebsiella_phage_KP36

Collinsella_stercoris

Hippeastrum_mosaic_virus

Pseudomonas_phage_MP38

Anaerostipes_unclassified

Oligella_urethralis

Bifidobacterium_pseudolongum

Coprobacillus_unclassified

Candidatus_Zinderia_insecticola

Porphyromonas_uenonis

Collinsella_aerofaciens

Phascolarctobacterium_succinatutens

Parabacteroides_johnsonii

Clostridium_bifermentans

Bacteroides_sp_1_1_30

Holdemania_sp_AP2

Bovine_viral_diarrhea_virus_3

Granulicatella_unclassified

Cardiobacterium_valvarum

Lachnospiraceae_bacterium_8_1_57FAA

Blautia_producta

Escherichia_sp_TW09276	
Akkermansia_muciniphila	
Streptococcus_peroris	
Lindernia_anagallis_yellow_vein _virus_satellite_DNA_beta	
Corynebacterium_amycolatum	
Butyrivibrio_crossotus	
Hepatitis_C_virus	
Stomatobaculum_longum	
Alistipes_finegoldii	
Sapporo_virus	
Human_adenovirus_B	
Peptostreptococcus_anaerobius	
Actinomyces_naeslundii	
Collinsella_intestinalis	
Parabacteroides_sp_20_3	
Corynebacterium_durum	
Colombian_datura_virus	
Bacteroides_fragilis	
Bordetella_unclassified	
Anaerostipes_caccae	
Lactobacillus_gasseri	
Eubacterium_siraeum	
Gastropod_associated_circular_s sDNA_virus	
Anaerotruncus_colihominis	
Barnesiella_intestinihominis	
Mobiluncus_curtisii	
Leuconostoc_inhae	

Note: Top 30 gut microbiota species in dataset 1 ; top 100 gut microbiota species in dataset 2. Top 5 host characteristics features and top 5 dietary features in datasets 1 and 2.