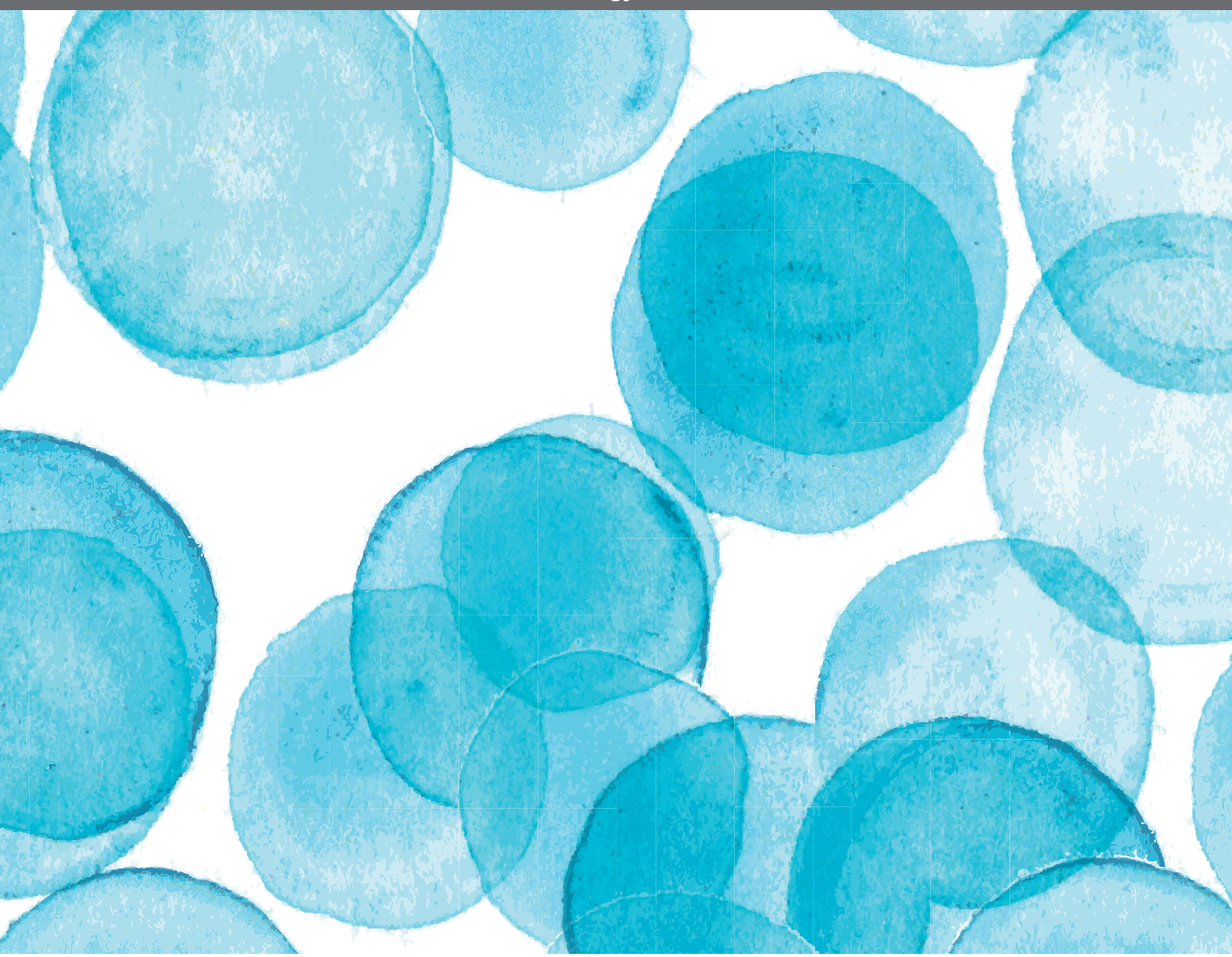
MICROBIOME AND MACHINE LEARNING

EDITED BY: Isabel Moreno Indias, Marcus Claesson, Aldert Zomer and David Gomez-Cabrero

PUBLISHED IN: Frontiers in Microbiology



frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88976-678-9
DOI 10.3389/978-2-88976-678-9

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

MICROBIOME AND MACHINE LEARNING

Topic Editors:
Isabel Moreno Indias, Universidad de Málaga, Spain
Marcus Claesson, University College Cork, Ireland
Aldert Zomer, Utrecht University, Netherlands
David Gomez-Cabrero, NavarraBiomed, Spain

Citation: Indias, I. M., Claesson, M., Zomer, A., Gomez-Cabrero, D., eds. (2022). Microbiome and Machine Learning. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-88976-678-9

Table of Contents

frontiers | Frontiers in Microbiology

EDITORIAL
published: 05 July 2022
doi: 10.3389/fmicb.2022.964921

Editorial: Microbiome and Machine Learning

Isabel Moreno-Indias[1,2]*, *Aldert L. Zomer*[3], *David Gómez-Cabrero*[4,5] and *Marcus J. Claesson*[6,7] *on behalf of ML4Microbiome*

[1] Unidad de Gestión Clínica de Endocrinología y Nutrición, Hospital Clínico Universitario Virgen de la Victoria, Instituto de Investigación Biomédica de Málaga y Plataforma en Nanomedicina-IBIMA Plataforma Bionand, Málaga, Spain, [2] Centro de Investigación Biomédica en Red de Fisiopatología de la Obesidad y la Nutrición (CIBEROBN), Instituto de Salud Carlos III, Madrid, Spain, [3] Department of Infectious Diseases and Immunology, Faculty of Veterinary Medicine, Utrecht University, Utrecht, Netherlands, [4] Navarrabiomed, Complejo Hospitalario de Navarra (CHN), IdiSNA, Universidad Pública de Navarra (UPNA), Pamplona, Spain, [5] Bioscience Program, Bioengineering Program, Biological and Environmental Science and Engineering Division, King Abdullah University of Science and Technology (KAUST), Jeddah, Saudi Arabia, [6] School of Microbiology and APC Microbiome Ireland, University College Cork, Cork, Ireland, [7] SeqBiome Ltd., Cork, Ireland

Keywords: microbiome, machine learning, personalized medicine, ML4microbiome, Artificial Intelligence

Editorial on the Research Topic

Microbiome and Machine Learning

The human microbiome has attracted more and more attention in the last decade. It has been recognized as a major player in the homeostasis of the host, and in this manner, in the pathophysiology of different diseases. One of the focus points of microbiome research has been advancing the development of personalized medicine approaches, which are potentially necessary to treat multifactorial diseases presenting with heterogenous phenotypes. While significant efforts have been made in terms of sampling, sequencing and analysis multiple well-described disease cohorts and controls, subsequent translation of this information into clinical use has unfortunately been slower than expected.

On the other hand, translation of microbiome insights to clinical practice faces different challenges. For instance, the many different analysis techniques specifically suited for the study of the microbiome have to be standardized, due to the otherwise over-shadowing methodological confounders. A second problem we are facing is that some particularities of microbiome data and its management makes the development of optimized and standardized methods that can deal with this kind of high dimensional data especially difficult. Machine learning (ML) offers great potential to be applied in analyzing these complex datasets. The main goal of the COST Action ML4Microbiome (https://www.cost.eu/actions/CA18131/) is to optimize, standardize and disseminate best practice of ML usage for human microbiome data. This Action has brought together Artificial Intelligence (AI)/ML experts and microbiome researchers to meet this aim, which will ultimately accelerate the advance in the translation of microbiome science.

This endeavor is, however, far from trivial due to several methodological challenges that must first be overcome. Microbiome data are inherently noisy and heterogeneous, there are several different data types, and in most cases many more features (taxa, genes etc.) than samples. In order to describe the current state-of-the-art of ML with microbiome data, Macros-Zambrano et al., reviewed ML use in terms of feature selection, biomarker identification, disease prediction and treatment. The review focused on real ML applications and outlined relevant software and repositories of microbiome data with associated research papers guiding the implementation of future ML efforts in this space. Indeed, ML4microbiome members also expressed their perspective on the past, present and future of the use of ML in microbiome in an accompanying review (Moreno-Indias et al.). Here, the main shortcomings identified were the small size of the datasets used so far, the necessity to combine statistical techniques that have been specifically tailored to fit the particular

OPEN ACCESS

Edited and reviewed by:
Ludmila Chistoserdova, University of Washington, United States

***Correspondence:**
Isabel Moreno-Indias
isabel.moreno@ibima.eu

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

Received: *09 June 2022*
Accepted: *10 June 2022*
Published: *05 July 2022*

Citation:
Moreno-Indias I, Zomer AL, Gómez-Cabrero D and Claesson MJ (2022) Editorial: Microbiome and Machine Learning. Front. Microbiol. 13:964921. doi: 10.3389/fmicb.2022.964921

characteristics of microbiome data, and the need for more user-friendly versions of these approaches to facilitate a wide range usage from different areas of expertise.

Original research manuscripts have also been submitted by the researchers in the field. The contributions published in this Research Topic have both improved current knowledge in particular fields, and contributed with new ML-based tools to be applied in the microbiome space. Some papers focus on the more technical part, including the comparison of Naive Bayes classifiers (NBC) vs. other 16S rRNA taxonomic classifiers based on Random Forest or Neural Networks (Ziemski et al.). The authors demonstrated and concluded that in practical scenarios NBC behave in a similar manner to the other classifiers. Although further improvements will arrive, at least for the moment, NBC use is still guaranteed.

In terms of development of new ML-based tools for enhancing microbiome data analysis have been part of this topic as well, Ramon et al. proposed the *kernInt* package to integrate metagenomic datasets with unsupervised and supervised microbiome analyses, including the recovery of microbial signatures through taxa importance. One important point is that *kernlnt* considers the compositionality of the microbiome data, and that this approach is adaptable enough to use with different applications.

Other applications presented in this Research Topic are two new tools developed based on two disciplines in continuous growth: virome and secretome (Fang and Zhou). Here, the authors used a deep learning approach in order to develop a prokaryote virus virion proteins (PVVPs) prediction tool called VirionFinder, to identify the complete and partial PVVPs from non-prokaryote virus virion proteins (non-PVVPs). The identification of this kind of proteins is a critical step for many viral analyses, such as species classification, phylogenetic analysis and the exploration of how prokaryote virus interact with their hosts. The researchers found that focusing only on a 20 amino acids sequence, instead of the whole or partial proteins VirionFinder, significantly improves sensitivity. Using real virome data further improved the recognition rate of PVVP-like sequences compared to previous tools.

Yu et al. presented their efforts on detecting secreted proteins by Gram-negative bacteria, which is particularly important due to their involvement in bacteria-host interactions. As it is currently challenging to distinguish between different types, especially between type III secreted effectors (T3SEs) and type IV secreted effectors (T4SEs), the authors proposed a deep learning solution for accurately distinguish T3SEs and T4SEs. The tool called DeepT3_4 is able to reach a recall of 80%, providing a promising tool for secretome analysis.

Several manuscripts submitted to this Research Topic have focused on a translational vision. Sudhakar et al. highlighted important computational applications to overcome some of the limitations encountered in microbiome lab-research to enhance our understanding of the microbe-host interactions, and how to fill the big gaps in terms of how the microbiome mechanistically influences host functions at both system and community levels (Sudhakar et al.). This comprehension allows us to progress the development of biomarkers uncovering mechanisms for therapeutic interventions and generating integrated signatures to stratify patients. Other authors have focused on particular diseases, such as Bakir-Gungor et al., who used different supervised and unsupervised ML models to investigate novel microbiota to find Type 2 diabetes (T2D) biomarkers. They increased the diagnostic accuracy and identified several species from *Bacteroides* and other genera that were relevant for the disease. These bacteria have been previously reported to play roles in T2D pathophysiology.

Finally, Vilne et al. presented a minireview on the use of ML in coronary artery disease and its risk prediction. The authors discussed the inclusion of diet-gut microbiome interactions in order to advance development of personalized medicine. Although microbiome data is of paramount importance for the development of a precision medicine approach, they argued that there are still several hurdles to take related to the homogenization of the data, both in terms of microbiome and diet. Once these have been addressed, the development of wearable biosensors for the patients' self-care may be possible.

In conclusion, the introduction of the use of ML in microbiome research is still in its infancy and much more research and methods development are necessary. These new approaches hold great potential for predicting individual health status, the Research Topic presented in this issue will hopefully aid in accelerating the transition. The ML4microbiome COST Action has made great strides in bringing the microbiome and ML community together which can lead to the necessary advancements in both research communities.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

FUNDING

This study was supported by the COST Action CA18131 "Statistical and machine learning techniques in human microbiome studies". IM-I was supported by the "MS type II" program (CPII21/00013) from the Instituto de Salud Carlos III and co-funded by Fondo Europeo de Desarrollo Regional-FEDER.

Conflict of Interest: MC was employed by company SeqBiome Ltd., Cork, County Cork, Ireland.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 21 January 2021
doi: 10.3389/fmicb.2021.605782

DeepT3_4: A Hybrid Deep Neural Network Model for the Distinction Between Bacterial Type III and IV Secreted Effectors

Lezheng Yu[1], Fengjuan Liu[2], Yizhou Li[3], Jiesi Luo[4] and Runyu Jing[3]**

[1] School of Chemistry and Materials Science, Guizhou Education University, Guiyang, China, [2] School of Geography and Resources, Guizhou Education University, Guiyang, China, [3] College of Cybersecurity, Sichuan University, Chengdu, China, [4] Department of Pharmacology, School of Pharmacy, Southwest Medical University, Luzhou, China

Gram-negative bacteria can deliver secreted proteins (also known as secreted effectors) directly into host cells through type III secretion system (T3SS), type IV secretion system (T4SS), and type VI secretion system (T6SS) and cause various diseases. These secreted effectors are heavily involved in the interactions between bacteria and host cells, so their identification is crucial for the discovery and development of novel anti-bacterial drugs. It is currently challenging to accurately distinguish type III secreted effectors (T3SEs) and type IV secreted effectors (T4SEs) because neither T3SEs nor T4SEs contain N-terminal signal peptides, and some of these effectors have similar evolutionary conserved profiles and sequence motifs. To address this challenge, we develop a deep learning (DL) approach called DeepT3_4 to correctly classify T3SEs and T4SEs. We generate amino-acid character dictionary and sequence-based features extracted from effector proteins and subsequently implement these features into a hybrid model that integrates recurrent neural networks (RNNs) and deep neural networks (DNNs). After training the model, the hybrid neural network classifies secreted effectors into two different classes with an *accuracy*, *F*-value, and *recall* of over 80.0%. Our approach stands for the first DL approach for the classification of T3SEs and T4SEs, providing a promising supplementary tool for further secretome studies.

Keywords: Gram-negative bacteria, secreted effector, deep learning-artificial neural network, recurrent neural networks, deep neural networks

OPEN ACCESS

Edited by:
Isabel Moreno Indias, University of Málaga, Spain

Reviewed by:
Jiangning Song, Monash University, Australia
Yejun Wang, Shenzhen University, China
Daniel Veltri, National Institutes of Health (NIH), United States

****Correspondence:***
Jiesi Luo
ljs@swmu.edu.cn
Runyu Jing
jingryedu@gmail.com

Specialty section:
This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology

Received: *13 September 2020*
Accepted: *04 January 2021*
Published: *21 January 2021*

Citation:
Yu L, Liu F, Li Y, Luo J and Jing R (2021) DeepT3_4: A Hybrid Deep Neural Network Model for the Distinction Between Bacterial Type III and IV Secreted Effectors. Front. Microbiol. 12:605782. doi: 10.3389/fmicb.2021.605782

INTRODUCTION

Protein secretion plays an important role in coordinating the interactions between bacteria and their surrounding environment. Through a variety of secretion systems, bacteria can release different types of proteins into the extracellular environment or even directly inject them into eukaryotic host cells (Galan and Waksman, 2018; McQuade and Stock, 2018). Since bacterial secreted proteins are commonly involved in important physiological activities of host cells, they have become a new research hotspot in recent years. To date, nine different types of secretion systems have been discovered from Gram-negative bacteria, which are named type I secretion system (T1SS) to type IX secretion system (T9SS), respectively (Lasica et al., 2017;

Lauber et al., 2018). Within these secretion systems, T1SS, T2SS, and T5SS can transport enzymes and other proteins into the surrounding environment, while type III secretion system (T3SS), type IV secretion system (T4SS), and type VI secretion system (T6SS) can transfer various effector proteins into host cells directly. These secreted effectors released through the latter three secretion systems are generally referred to as type III secreted effectors (T3SEs), type IV secreted effectors (T4SEs), and type VI secreted effectors (T6SEs) (An et al., 2018), and they can exert the virulence of Gram-negative bacteria in a number of ways, severely disrupting the normal function of host cells (Kim, 2018). Therefore, an in-depth study of secreted effectors is highly desirable for understanding the pathogenesis of bacteria and developing novel anti-microbial agents.

Over the past decade, dozens of machine learning-based computational approaches have been proposed to identify different types of secreted effectors (Zeng and Zou, 2019), including support vector machine (SVM) (Samudrala et al., 2009; Yang et al., 2010; Wang et al., 2011, 2014, 2017; Dong et al., 2013; Zou et al., 2013; Goldberg et al., 2016; Esna Ashari et al., 2019a,b), random forest (RF) (Yang et al., 2013), artificial neural network (ANN) (Löwer and Schneider, 2009), naive Bayes (NB) (Arnold et al., 2009), hidden Markov model (HMM) (Xu et al., 2010; Lifshitz et al., 2013; Wang et al., 2013), logistic regression (LR) (Esna Ashari et al., 2018), decision tree (DT) (Wang et al., 2019a), gradient boosting (Chen et al., 2020), deep learning (DL) (Xue et al., 2018, 2019; Açıcı et al., 2019; Fu and Yang, 2019; Hong et al., 2020; Li et al., 2020a), and their ensemble methods (Burstein et al., 2009; Hobbs et al., 2016; Wang et al., 2018, 2019b; Xiong et al., 2018; Li et al., 2020b). Some of these methods have achieved relatively high predictive accuracy, while they can recognize only one type of secreted effector, such as SIEVE (Samudrala et al., 2009), EffectiveT3 (Arnold et al., 2009), T3_MM (Wang et al., 2013), GenSET (Hobbs et al., 2016), Bastion3 (Wang et al., 2019a), DeepT3 (Xue et al., 2019), WEDeepT3 (Fu and Yang, 2019), ACNNT3 (Li et al., 2020a), and EP3 (Li et al., 2020b) for T3SEs; T4EffPred (Zou et al., 2013), T4SEpre (Wang et al., 2014), DeepT4 (Xue et al., 2018), PredT4SE-Stack (Xiong et al., 2018), Bastion4 (Wang et al., 2019b), T4SE-XGB (Chen et al., 2020), and CNN-T4SE (Hong et al., 2020) for T4SEs; and Bastion6 (Wang et al., 2018) for T6SEs. It is important to note that due to the small number of T6SEs for model construction, researchers usually pay more attention to identifying T3SEs and T4SEs rather than T6SEs. In addition, several multi-label classifiers have been developed to identify different types of Gram-negative bacterial secreted proteins simultaneously (Yu et al., 2013; Ding and Zhang, 2016; Liang et al., 2018; Yu et al., 2018; Kong and Zhang, 2019), but they are not good at distinguishing between T3SEs and T4SEs. Both T3SEs and T4SEs are non-classical secreted proteins (without classical N-terminal signal peptides) (Liang and Zhang, 2018; Zhang et al., 2020), and some of them have similar evolutionary conserved profiles and sequence motifs (Zou et al., 2013), so it is difficult to distinguish them accurately using current methods.

In this paper, we explore the use of various DL architectures and feature descriptors to identify and classify T3SEs and T4SEs. Four different DL architectures are used to build the classification models, including the convolutional neural networks (CNNs), recurrent neural networks (RNNs), convolutional-RNNs (CNN-RNNs), and deep neural networks (DNNs). For the CNN, RNN, and CNN-RNN architectures, we first characterize protein sequences using dictionary encoding and then generate amino-acid character embedding vectors to learn the features of two types of secreted effectors. The DNN architecture is designed as a multilayered neural network, whose input layer is fed traditional features or descriptors, including amino acid composition (AAC), dipeptide composition (DC), position-specific scoring matrix (PSSM), and their different combinations. We carry out extensive experiments for comparison and present a systematic analysis. Our results show that a hybrid neural network (architectures: RNN + DNN; features: dictionary encoding + AAC + DC) performs better than other models on the test and independent test datasets, enabling accurate classification of T3SEs and T4SEs. We also achieve interpretable DL for T3SEs and T4SEs classification via an advanced dimensionality reduction procedure and visualization, which unravels the predictions of models. Based on these results, we develop a DL approach, which is called DeepT3_4, by implementing both the raw sequence and sequence-derived features of effector proteins into the hybrid model. DeepT3_4 helps to understand the similar sequences and structures for some of T3SEs and T4SEs, facilitating the refined studies of different types of secreted effectors.

MATERIALS AND METHODS

Dataset Collection and Processing

Reliable data are the primary factor in establishing stable and effective predictors, and all experimental data used in this study were extracted from the Bacterial Secreted Effector Protein DataBase (SecretEPDB) (An et al., 2017). SecretEPDB provides a comprehensive annotation of the T3SEs, T4SEs, and T6SEs, including sequence, structure, and function annotations for these secreted effectors. A total of 1230 T3SEs, 731 T4SEs, and 181 T6SEs were collected in this database, and we selected all of the T3SE and T4SE samples as original data to construct the training and test datasets.

In order to avoid redundancy and homology bias, all effector proteins in the original data were aligned by CD-HIT (Huang et al., 2010) with a maximum sequence identity of 25%. After that, only 302 T3SEs and 375 T4SEs were kept. Subsequently, 70% of this dataset was randomly selected for building the benchmark dataset and the remaining 30% was used to establish the independent test set (Jiang et al., 2017). Finally, the benchmark dataset contained 211 T3SEs and 263 T4SEs, while the independent test set was consisted of 91 T3SEs and 112 T4SEs (**Supplementary Table S1**).

For further evaluating the performance of our method and comparing with other state-of-the-art approaches, other two independent test datasets were established by searching publicly available articles. The independent test dataset 2 contains 108 T3SEs and 30 T4SEs, which were extracted directly from Bastion3 (Wang et al., 2019a) and Bastion4 (Wang et al., 2019b),

respectively. The independent test dataset 3 is composed of 35 T3SEs and 75 T4SEs, which were collected from the studies of Yang et al. (2013) and Wang et al. (2017), respectively. In addition, other 1319 proteins were randomly selected to detect the performance of our method for identifying non-T3SEs and non-T4SEs.

Feature Extraction

Dictionary Encoding

Each amino acid in the protein sequence is represented by an ordinal number, in which each of the 20 basic amino acids is assigned a number from 1 to 20 (e.g., alanine is assigned a number of 1) (Veltri et al., 2018). Thus, each protein is represented by a one-letter code and transformed into an L-dimensional vector, where L is the length of the protein.

Amino Acid Composition (AAC) and Dipeptide Composition (DC)

For each protein sequence, a 20-dimensional vector $\{d_1, d_2, \ldots, d_{20}\}$ and a 400-dimensional vector $\{d_1, d_2, \ldots, d_{400}\}$ are used to represent the compositions of 20 common amino acids and all 400 possible amino acid pairs, respectively. The 20 elements in $\{d_1, d_2, \ldots, d_{20}\}$ represent the occurrence frequencies of each amino acid with a protein. The 400 elements in $\{d_1, d_2, \ldots, d_{400}\}$ represent the frequencies of dipeptides.

Position-Specific Scoring Matrix (PSSM)

The PSSM profiles contain the evolutionary information of a protein. Each element in PSSM indicates the substitution scores of the individual residue at that specific position in the multiple sequence alignment. To generate PSSM, each protein sequence in our training and test datasets was searched against the Swiss-Prot database using the PSI-BLAST (Altschul and Koonin, 1998) with three iterations and a cutoff E-value of 0.001. The generated PSSM from PSI-BLAST includes $L \times 20$ elements, where L is the length of a protein. This original profile is further used to calculate the PSSM feature by averaging the columns in PSSM profile and then is scaled to $[-1, 1]$. Finally, PSSM generates a 20-dimensional feature vector by characterizing a mutation of the corresponding amino acid type during the evolution process.

Deep Neural Networks

As the most popular machine learning algorithm, DL has been successfully applied to solve various problems, such as image recognition, speech recognition, language translation, and biological data analysis (Jurtz et al., 2017; Tang et al., 2019). There have been four common variations of DNNs, including the CNNs, the RNNs, the CNN-RNNs, and the DNNs. The CNNs have outstanding spatial information analysis capabilities and have been successfully applied in the prediction of secreted effectors (Xue et al., 2018, 2019; Açıcı et al., 2019), protein solubility (Khurana et al., 2018), and crystallization (Elbasir et al., 2019). Compared to CNNs, RNNs can handle sequential inputs effectively and recognize sequence motifs of varying length extraordinarily well, making them the preferred choice for machine translation, text generation, and image captioning (Esteva et al., 2019). In order to integrate the advantages of the CNNs and RNNs, the CNN-RNNs have been developed in recent years and applied to a variety of biological problems (Quang and Xie, 2016; Pan et al., 2018; Tayara and Chong, 2019). As a typical representative of feedforward neural network (FNN), DNNs are composed of multiple perceptrons of different layers and are therefore very suitable for solving non-linear problems and have been widely used in data classification and other fields (Kruse et al., 2013).

Deep Learning Architectures

To accurately classify the proteins of Gram-negative bacteria into separate secretion classes, we used DNNs with four different architectures. For the first three network architectures, including CNNs, RNNs, and CNN-RNNs, we encode the primary sequence using a dictionary amino acid representation as input and output one score between 0 and 1, corresponding to the probability of an effector protein of interest being a T3SE or a T4SE. The fourth architecture DNN is a standard multilayer neural network. The DNN model takes AAC, PSSM, DC, and their different combinations as inputs to predict the probability scores of two types of secreted effectors. We describe the overview of different DL architectures below.

The CNN consists of an embedding layer, a convolutional layer, a pooling layer, a fully connected layer, and an output layer. The first embedding layer transforms the input into a 256-dimensional vector representation. This transformation can best be thought of as a one-dimensional signal (over sequence position) spanning all amino acid signal channels. The input sequence is 1500 amino acids long, a number that was chosen to fit out dataset's longest sequence. If the length of the sequence exceeds 1500 amino acids, the excess will be ignored; otherwise, the "X" character (unknown residue) will be padded at the tail of the sequence to fit the 1500 length. The second convolutional layer has 250 filters, where the filter width is set to five. The convolutional layer is then followed by a max-pooling layer with a non-overlapping window of size 2 to halve the size of the input. Subsequently, a fully connected layer consisted of 650 neurons with a dropout ratio of 20% is chosen to receive the flattening results of the pooling layer (Bogard et al., 2019). All layers whose activation is not specified explicitly use rectified linear unit (ReLU) activations. Finally, the output layer employs the sigmoid activation function to provide the predicted probability score for the test sequence.

The RNN is made up of three types of layers: an embedding layer, a biLSTM layer, and an output layer. The bidirectional long short-term memory (biLSTM) is an enhanced version of general RNNs in which the scalar-valued hidden layer of RNNs is replaced by a biLSTM memory block. The biLSTM layer is a forward–backward structure along the input sequence consisting of two relatively separated RNN layers. We explored biLSTM layers with 32, 64, 128, and 256 neurons and from one to four layers deep. A biLSTM layer with 64 neurons and one layer of depth gave the best performance. Dropout of 20% is applied to biLSTM layer to prevent overfitting. The final output layer utilizes a sigmoid activation function to process the output of the biLSTM layer and gives a value (the probability score) for each protein sequence.

The CNN-RNN incorporates an embedding layer with embedding size 256 along with a 1D convolutional layer with filter size = 5. The max-pooling layer subsamples the 1D signal by a factor of two. The flattened pooling output is passed to a biLSTM layer of 64 hidden neurons, which finally connects to a sigmoid activation function that outputs the predicted probability score.

The DNN is constructed from three fully connected layers with decreasing sizes of features vectors (e.g., 400, 200, and 100 for the DC) to reduce feature dimensions toward convergence of model training. As an additional precaution, a dropout probability of 20% is used in each layer.

The same RNN and DNN architectures are used to construct the hybrid model. The RNN and DNN models are trained separately, and their last hidden layers are further concatenated and inputted into a sigmoid activation node. The RNN architecture consists of an embedding layer and a biLSTM layer. The biLSTM layer has 64 hidden units followed by a dropout rate of 20%. The DNN model has three fully connected layers with 420, 210, and 105 neurons, respectively.

Performance Evaluation

For evaluation, we used standard performance quantification metrics such as Recall (Sensitivity), Precision (PRE), Accuracy (ACC), F-value, and Matthew's correlation coefficient (MCC), which are defined as follows:

$$\mathrm{Re}call = \frac{TP}{TP + FN} \tag{1}$$

$$PRE = \frac{TP}{TP + FP} \tag{2}$$

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{3}$$

$$F - \mathrm{value} = 2 \times \frac{TP}{2TP + FP + FN} \tag{4}$$

$$MCC = \frac{TP \times TN - FN \times FP}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}} \tag{5}$$

where TP, FP, TN, and FN stand for true positive, false positive, true negative, and false negative, respectively.

Implementation

All DNNs were implemented by using autoBioSeqpy (Jing et al., 2020). The autoBioSeqpy is an easy-to-use DL tool for biological sequence classification. The main advantage of this tool is its simplicity. Users only need to prepare the input dataset. After that, data encoding, model development, training, evaluation, and figure generation workflows can be run through the command line interface, by which users can modify the parameters of the workflows easily. In addition, autoBioSeqpy is designed to separate the data encoding and model configuration into two relatively independent parts. The DL models can be built using python code (i.e., written in.py files) or json files (saved by Keras), so that the model can be flexibly adjusted according to user needs. Currently, the tool has been upgraded to version 2.0, which supports more complex network models and incorporates model visualization function. For example, layerUMAP is a portable command-line tool included in the autoBioSeqpy tool suite, written in python, that makes use of the uniform manifold approximation and projection (UMAP) for visual understanding of DL models (Melville, 2019).

We sampled a variety of hyperparameter sets for different DL models, including embedding dimension (32, 64, 128, and 256), dropout rate (10, 20, 30, and 40%), batch size (25, 50, 75, and 100), epoch number (20, 40, 60, and 80), learning rate (0.001, 0.005, 0.01, 0.05, and 0.1), convolution kernel size (3, 5, 7, 9, and 11), number of filters (50, 100, 150, 200, and 250), and number of neurons in BiLSTM (32, 64, 128, and 256). We took the sampled parameter set with best performance (mean *MCC* score) and varied each parameter individually while keeping the rest constant.

During the training process, we used binary_crossentropy as loss function of the network and it has been optimized using the Adam optimizer approach with a learning rate 0.001. We trained all models with 40 epochs and a batch size of 25. The weights of the parameters were updated within 40 epochs, and at the end of each epoch, the intermediate validation metric is calculated. After the training, the optimized parameters were evaluated by the predictions from the test dataset. All the training was conducted on a Windows 10 workstation with an NVIDIA GTX 1060 GPU with CUDA 10.2.95. To interpret the model, we visualized the decision map of model in two dimensions. We used the output of the last hidden layer of the model as the extracted output features, which were then projected into a 2D manifold via UMAP. Next, we used a two-color scheme to refer to T3SE and T4SE based on the extracted output features.

RESULTS

Overview of the Deep Learning Models

We first used a DL classification tool (autoBioSeqpy) (Jing et al., 2020) to design and evaluate all 11 DL classifiers. Classifiers are divided into three categories: (1) different model architectures but with the same model inputs (CNN, RNN, and their combination CNN-RNN); (2) the same model architecture but with different model inputs (DNN); and (3) a hybrid architecture combining the above two categories (RNN and DNN). **Figure 1** depicts all DL architectures in the effector classification. Details of the methods are reported in Section "Materials and Methods."

Effect of Model Architectures and Features on Performance

We analyzed the performance of 11 different models (CNN, RNN, CNN-RNN, seven DNN models with different input features, and a hybrid model) on our held-out test set. The benchmark dataset comprised 474 proteins, 70% of which was randomly extracted for establishing the training set, 20% for the test set, and the remaining 10% for the validation set. We performed an extensive random hyperparameter search for each model on the validation set, and then the top-performing tuned

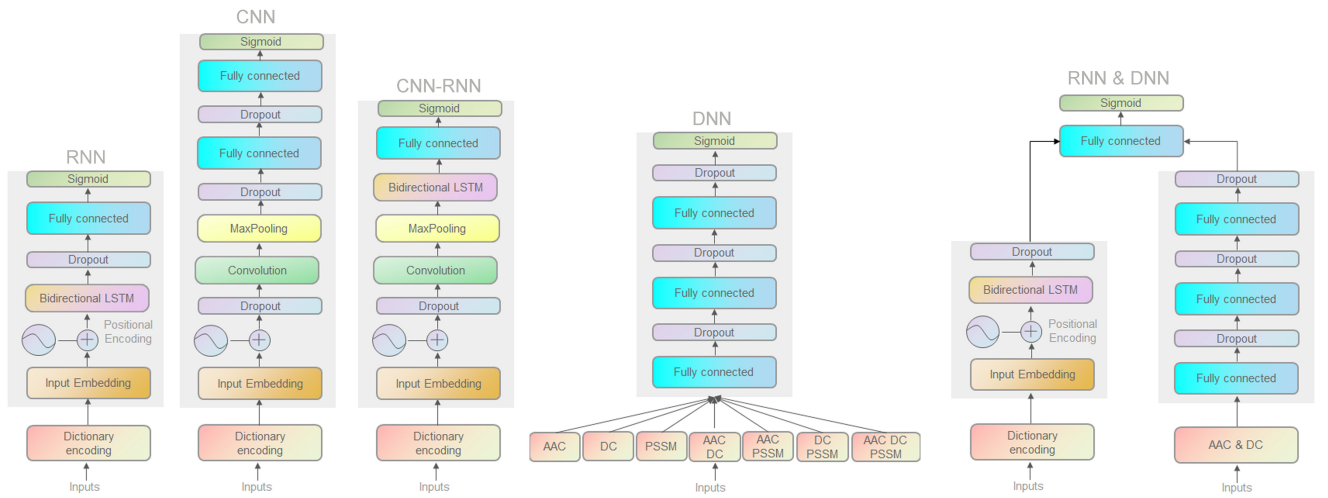


FIGURE 1 | Schematic illustration of the deep neural network architectures.

models were evaluated on the test set. A summary of our results is provided in **Figure 2**.

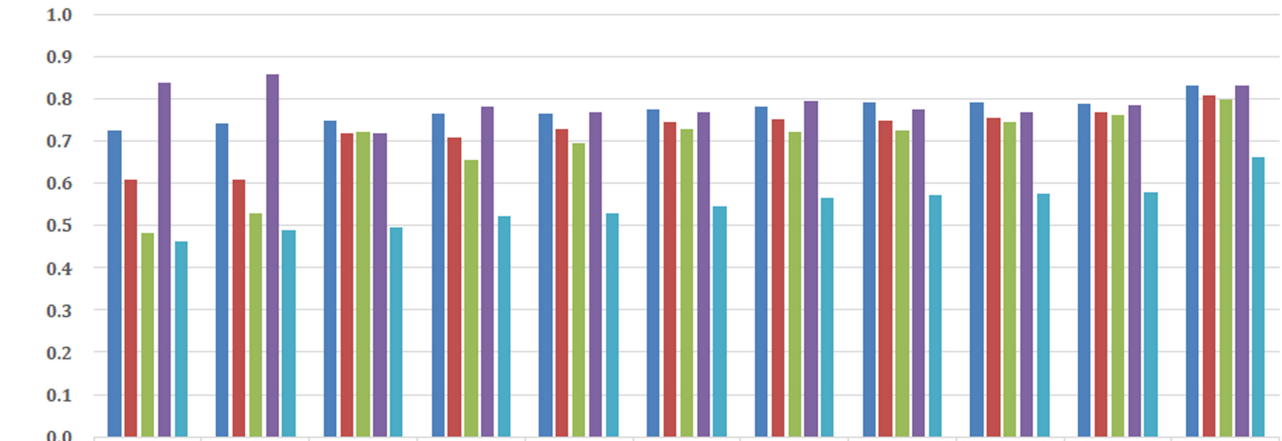
Since CNN, RNN, and CNN-RNN have the same model inputs, we first compared these three DL architectures. The RNN model afforded the best training performance with the highest scores of *Recall* (72.9%), *PRE* (77.0%), *ACC* (77.5%), *MCC* (0.546), and *F*-value (74.4%). The CNN model followed with an *ACC* of 76.6% and an *MCC* of 0.528. The CNN-RNN model showed the lowest performance (*ACC* = 74.9% and *MCC* = 0.496), which were lower than those of the RNN model as 2.6% and 0.050 on *ACC* and *MCC*, respectively.

Based on AAC, DC, PSSM, and their different combinations, seven different features were employed to build the DNN models. From **Figure 2**, it can be seen that DNN models trained by the single feature group (only AAC, PSSM, or DC) tended to obtain the relatively poor results, whereas those trained by a combination of features (AAC+PSSM, AAC+DC, PSSM+DC, and AAC+PSSM+DC) seemed to achieve better performance. For the seven DNN models, the model with AAC+DC yielded the best performance, and gave the highest scores of *Recall* (76.3%), *F*-value (77.0%), and *MCC* (0.578). The PSSM+DC model offered the highest *ACC* score (79.3%), but other four parameters were lower than the AAC+DC model. Although the model with AAC+PSSM+DC learns the most information, its overall predictive performance was also weaker than that of the model with AAC+DC. Surprisingly, the comprehensive performance of the model with single feature DC was almost comparable to those models trained with the combined features. This result indicates that DC is a very important feature for making a distinction between T3SEs and T4SEs.

After careful analysis of above results, we proposed a hybrid model to integrate the advantages of the RNN and best DNN models. This hybrid DNN model yielded the best overall prediction performance for the test dataset, and provided the highest scores of *Recall* (77.9%), *PRE* (83.3%), *ACC* (83.1%), *MCC* (0.663), and *F*-value (81.0%). Therefore, we chose this hybrid model as the final prediction model for this study. The receiver operating characteristic (ROC) curve, precision recall (P-R) curve, and accuracy-loss (acc-loss) curve were exploited to evaluate the performance of the hybrid model (**Figure 3**). Area values under the ROC curve (auROC) and P-R curve (auPRC) for the hybrid model were 0.877 and 0.832, respectively. We also trained the RNN and DNN (AAC+DC) models separately to evaluate the robustness of the models, showing auROCs of (0.804 and 0.847) and auPRCs of (0.795 and 0.794), respectively (**Supplementary Figures S1, S2**). The results suggest that the RNN and DNN (AAC+DC) models learned different sets of features that complement each other for the task of distinguishing between two types of secreted effectors.

Visualizing and Understanding Deep Learning Models

To investigate the ability of DL models to distinguish two types of secreted effectors, we analyzed the features extracted from the last hidden layer of three classification models [RNN, DNN (AAC+DC), and RNN and DNN (AAC+DC)]. **Figure 4** shows a UMAP (McInnes and Healy, 2018) for dimension reduction projection of these features. The points are color-coded based on the true class label. Therefore, T3SEs and T4SEs are characterized by purple and red points, respectively, with each point representing an effector. As shown in **Figure 4**, the features clearly distinguish the different secreted effectors. In the RNN architecture, some T3SEs are distributed across the T4SE cluster with no obvious pattern. The DNN and hybrid architectures have the advantage of very clearly clusters, which is consistent with the above classification results. Furthermore, studies have confirmed that T3SS could be divided into two subgroups, including the injectisome (non-flagellar) system and the fagellar system (Blocker et al., 2003; Puhar and Sansonetti, 2014). Therefore, as the secretory products of the T3SS, T3SEs could also be classified into two subtypes, which are shown



	DNN(PSSM)	DNN(AAC)	CNN-RNN	DNN(AAC+PSSM)	CNN	RNN	DNN(DC)	DNN(PSSM+DC)	DNN(AAC+PSSM+DC)	DNN(AAC+DC)	RNN&DNN(AAC+DC)
ACC	0.726	0.741	0.749	0.765	0.766	0.775	0.782	0.793	0.791	0.788	0.831
F-value	0.608	0.609	0.718	0.709	0.729	0.744	0.752	0.748	0.754	0.770	0.810
Recall	0.484	0.530	0.723	0.655	0.696	0.729	0.722	0.727	0.745	0.763	0.799
PRE	0.838	0.860	0.720	0.781	0.769	0.770	0.794	0.774	0.770	0.787	0.833
MCC	0.462	0.490	0.496	0.523	0.528	0.546	0.566	0.574	0.575	0.578	0.663

FIGURE 2 | Performance comparison of different model architectures and features on the test dataset.

by the two sub-populations in **Figure 4**. Thus, this result implies that T3SEs do have different sequences and conserved patterns as well.

Performance of Different Model Architectures and Features on the Independent Test Set

To test model performance on external data, an independent test set was obtained whose data were never used for training and testing. We used this dataset to further compare the predictive performance of models established by different architectures and features, and the results are shown in **Figure 5**. For the three DL architectures whose inputs are dictionary-encoded sequences, the RNN model also yielded the best overall prediction performance, achieving the highest scores of *Recall* (75.3%), *PRE* (79.4%), *ACC* (80.0%), *MCC* (0.596), and *F*-value (77.2%). The CNN-RNN model got the worst predictive performance, including the lowest scores of *Recall* (72.6%), *ACC* (75.6%), *MCC* (0.508), and *F*-value (72.7%). The AAC+DC model also afforded the best overall predictive performance among the seven DNN models, receiving the highest scores for *ACC* (80.0%), *MCC* (0.597), and *F*-value (77.5%). Finally, the hybrid RNN and DNN (AAC+DC) model obtained the best overall predictive performance on the independent test set, providing the highest scores for *Recall* (81.2%), *PRE* (80.0%), *ACC* (82.3%), *MCC* (0.645), and *F*-value (80.5%), respectively. We then evaluated the performance of RNN, DNN (AAC+DC), and their hybrid model using ROC, PR, and acc-loss curves (**Supplementary Figures S3, S4** and **Figure 6**). In terms of auROC and auPRC, the hybrid model also performed better than other two models. These results further suggest that combining learned features of RNN and DNN models can deliver a better model compared with individual models.

Development of DeepT3_4 and Comparison With Other Existing Methods

To further evaluate the performance of our hybrid DL model (named DeepT3_4), we used other two independent test datasets to compare the performance of DeepT3_4 with other three state-of-the-art approaches, including a typical T3SE predictor-Bastion3 (Wang et al., 2019a) and two representative T4SE classifiers-Bastion4 (Wang et al., 2019b) and CNN-T4SE (Hong et al., 2020). For the independent test dataset 2, all prediction results are listed in **Table 1**. As shown in the table, Bastion3 correctly identified all 108 T3SEs, but 12 T4SEs were incorrectly predicted as T3SEs; 29 T4SEs were correctly identified by Bastion4, but 25 T3SEs were incorrectly predicted as T4SEs. CNN-T4SE correctly identified the maximum number of T4SEs (29), but got the minimum number of T3SEs (59). When using DeepT3_4, 101 T3SEs and 26 T4SEs were correctly identified, and seven T3SEs and four T4SEs were misclassified. Although DeepT3_4 did not obtain the highest *Recall* for T3SEs and the highest *PRE* for T4SEs, it yielded the best overall prediction performance here. DeepT3_4 gave the highest scores of *ACC* (92.0%), *MCC* (0.775), and *F*-value (94.8%), which provided a 0.7%, 0.1%, and 0.040 improvement in *ACC*, *F*-value, and *MCC*, respectively. These results indicate that DeepT3_4 is stable and reliable in distinguishing T3SEs and T4SEs.

For the independent test dataset 3, all results are shown in **Table 2**. As we can see from this table, Bastion3 acquired the best overall prediction performance with the highest scores of *ACC*

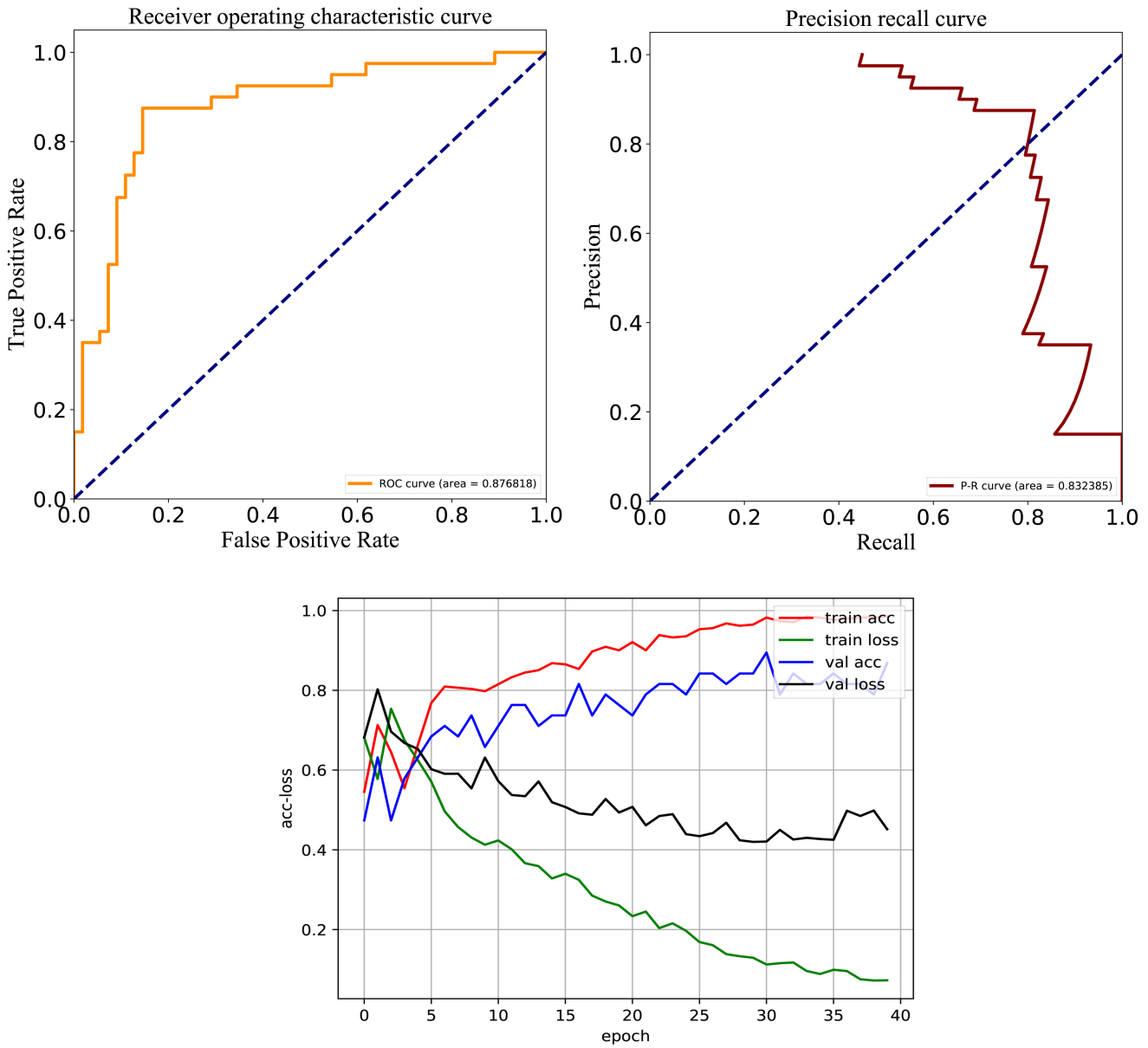


FIGURE 3 | ROC, PR, and acc-loss curves generated by autoBioSeqpy tool for the hybrid deep learning model on the test dataset.

(95.5%), *F*-value (93.0%), *Recall* (94.3%), and *MCC* (0.896). The performance of DeepT3_4 is slightly lower than that of Bastion3, and afforded the second highest scores of *ACC* (94.5%), *F*-value (91.4%), *Recall* (91.4%), and *MCC* (0.874). Though Bastion4 and CNN-T4SE got the highest score of *PRE* (100.0%), their overall prediction performances were worse than those of Bastion3 and DeepT3_4. It is noteworthy that for most of query sequences (known secreted effectors) in the independent test dataset 3, Bastion3 and Bastion4 did not provide the prediction results, but directly gave the search results of BastionDB and all results were marked as *Exp*. If both of Bastion3 and Bastion4 give the prediction results for all query sequences, we believe that DeepT3_4 will perform better than them.

Model Robustness Evaluation

To assess the effect of data scale on the predictive performance of DeepT3_4, we calculated and plotted learning curves to observe the relationship between the performance and data size. To generate learning curves, an external resampling mechanism with replacement was used to generate subsets with five different scales: 20, 40, 60, 80, and 100%. After resampling, the subset was split into five training-test groups for cross-validation. Each resampling was repeated 10 times to measure the robustness of the DL model. Thus, a total of 250 models (10 replicates * five scales * five folds) were built for predicting the generated test sets. **Supplementary Figures S5, S6** show the learning curves of the DeepT3_4 model using the ACC and MCC metrics. The

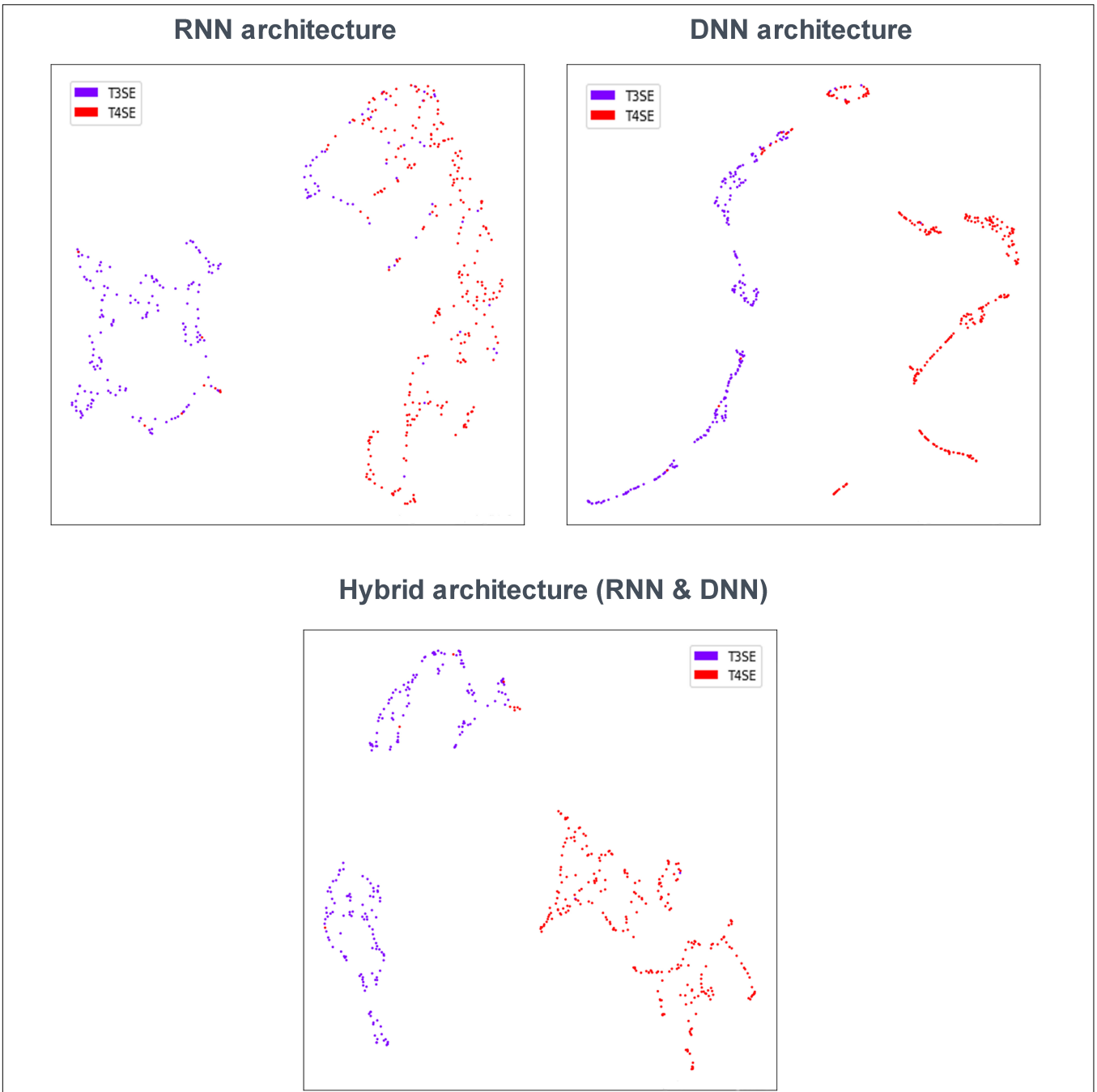
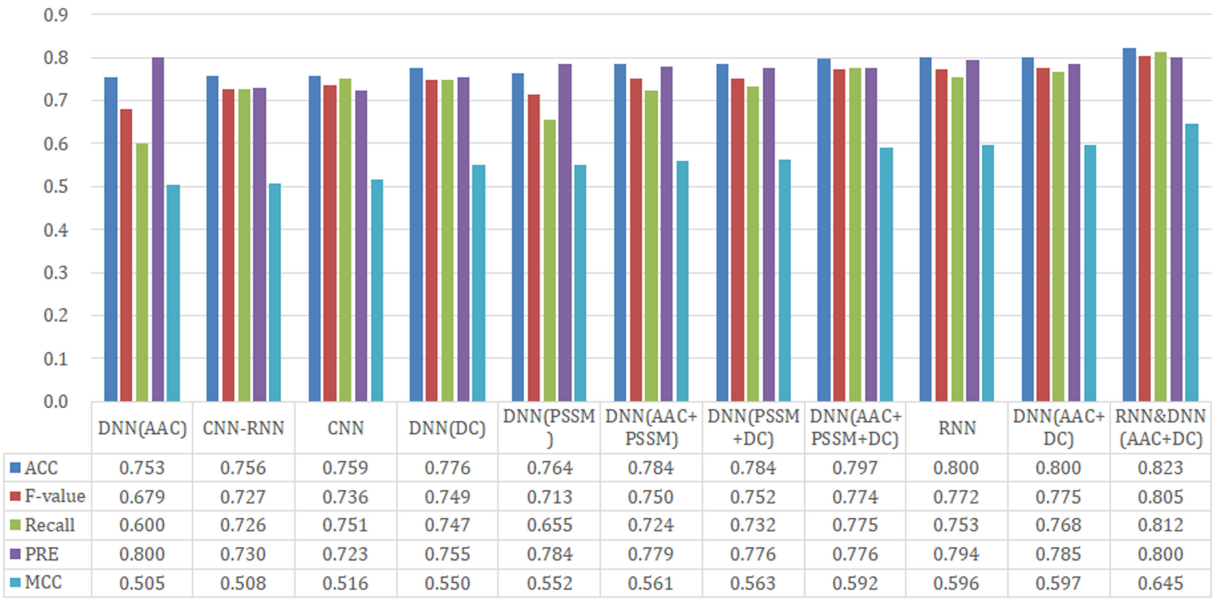


FIGURE 4 | UMAP visualization of learned features.

DeepT3_4 model becomes relatively stable when the scale of the dataset reaches 60% (about 325 samples). On this scale, the ACC and MCC scores in cross-validation are 81.0 ± 3.0% and 0.620 ± 0.060, respectively. Except for the learning curve, we also used fivefold cross-validation on a 100% scale dataset to further evaluate the generalizability of the model. The detailed results are shown in **Supplementary Table S2**. The DeepT3_4 achieves the average scores of 83.9 ± 2.6% for ACC and 0.677 ± 0.052 for MCC, which is consistent with the results of 10-time test (**Figure 2**). All together, these results illustrate the robustness of DeepT3_4, even on the small sample dataset.

DISCUSSION

In recent years, many excellent works have been done in the field of secreted effector prediction, such as Bastion3 (Wang et al., 2019a) and DeepT3 (Xue et al., 2019) for T3SEs and Bastion4



	DNN(AAC)	CNN-RNN	CNN	DNN(DC)	DNN(PSSM)	DNN(AAC+PSSM)	DNN(PSSM+DC)	DNN(AAC+PSSM+DC)	RNN	DNN(AAC+DC)	RNN&DNN(AAC+DC)
ACC	0.753	0.756	0.759	0.776	0.764	0.784	0.784	0.797	0.800	0.800	0.823
F-value	0.679	0.727	0.736	0.749	0.713	0.750	0.752	0.774	0.772	0.775	0.805
Recall	0.600	0.726	0.751	0.747	0.655	0.724	0.732	0.775	0.753	0.768	0.812
PRE	0.800	0.730	0.723	0.755	0.784	0.779	0.776	0.776	0.794	0.785	0.800
MCC	0.505	0.508	0.516	0.550	0.552	0.561	0.563	0.592	0.596	0.597	0.645

FIGURE 5 | Performance comparison of different model architectures and features on the independent test set.

(Wang et al., 2019b) and CNN-T4SE (Hong et al., 2020) for T4SEs. Different from these studies, we developed a hybrid DL approach by integrating RNN and DNN architectures to classify T3SEs and T4SEs in this work. We have carried out extensive experiments for comparison and have presented an in-depth analysis. For both the benchmark and independent test sets, the hybrid DNN model shows a consistently better performance than the others. The innovations of this study are as follows: (i) to the best of our knowledge, this is the first study to use DL to classify T3SEs and T4SEs; (ii) different DL architectures and features are employed to construct the predictors; (iii) clustering and visualization of model-extracted features using UMAP; (iv) the experimental results confirm that some of T3SEs and T4SEs may have similar evolutionary conservatism profiles and sequence motifs, which leads to limitations in the classification performance of computational methods.

The secretion signal of T3SEs is generally located at the N-terminal sequences (Yang et al., 2013), while the secretion signal of T4SEs is commonly found in the C-terminal sequences (Nagai et al., 2005). Therefore, some state-of-the-art methods choose only 50–100 N-terminal amino acid residues to identify T3SEs (Wang et al., 2013; Yang et al., 2013; Xue et al., 2019), or only 100 C-terminal amino acid residues to predict T4SEs (Zou et al., 2013; Wang et al., 2014; Xue et al., 2018). In order to assess the role of N-terminal or C-terminal sequence features in the classification of T3SEs and T4SEs, we calculated the sequence-based features within the first 100 N-terminal residues, the last 100 C-terminal residues, and the whole protein sequences using the best hybrid model, and further compared their performance using the independent test set consisting of 91 T3SEs and 112 T4SEs. All test results are listed in **Supplementary Table S3**. As can be seen from the table, the hybrid model trained by the full protein sequences achieved the best overall prediction performance and afforded the highest scores of *Recall* (81.2%), *PRE* (80.0%), *ACC* (82.3%), *MCC* (0.645), and *F*-value (80.5%). However, the performance of the hybrid models trained on the first 100 N-terminal and last 100 C-terminal residues is lower than that of the full-length sequence. These results suggest that the full sequences can better characterize the two types of secreted effectors.

We further developed DeepT3_4 to be able to predict non-T3SEs and non-T4SEs. To further estimate the performance of DeepT3_4, we employed a new dataset for a ternary classification, which is composed of 1319 other proteins. When tested on the new test set 10 times, DeepT3_4 obtained an overall average *ACC* of 88.2%, which is higher than that of the binary classification (*ACC* = 82.3%), suggesting that the addition of other types of protein sequences does not affect the predictive performance of our method.

In order to gain insight into the pathogenesis of bacteria and to effectively develop new drugs, an increasing number of studies have been conducted on various secreted effectors. Although DeepT3_4 can distinguish between T3SEs and T4SEs, there is still some room for further improvement. Moreover, there are still many issues to be solved in the study of secreted effectors. For example, T6SEs are widespread in various Gram-negative bacteria, but only a few computational methods are currently available to accurately identify them (Wang et al., 2018, 2020; Sen et al., 2019). The T3SS and T4SS can be divided into different subgroups (Costa et al., 2015), and thus their secretory products, T3SEs and T4SEs are also classified into different subfamilies (Bi et al., 2013; Zou et al., 2013). However, more detailed studies of the subfamilies of T3SEs and T4SEs are still rare. In addition, a new predictor has been built to recognize potential non-classical secreted proteins of Gram-positive bacteria recently (Zhang et al., 2020), which may

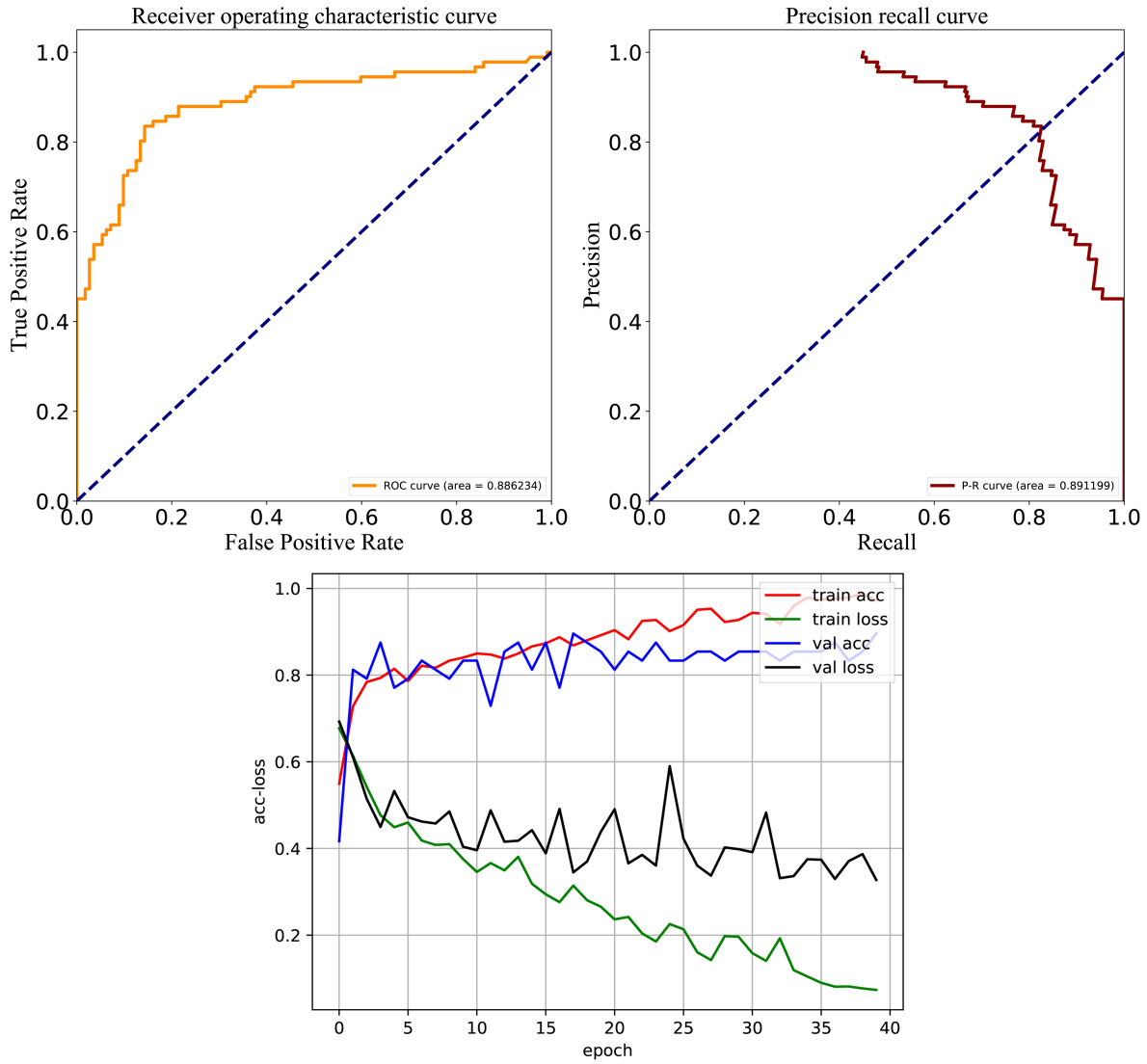


FIGURE 6 | ROC, PR, and acc-loss curves generated by autoBioSeqpy tool for the hybrid deep learning model on the independent test set.

TABLE 1 | Performance Comparisons of DeepT3_4 with other three methods on the independent test dataset 2.

Model	TP	TN	FN	FP	ACC (%)	*F*-value (%)	Recall (%)	PRE (%)	MCC
Bastion3	108	18	0	12	91.3	94.7	100.0	90.0	0.735
Bastion4	83	29	25	1	81.2	86.5	76.9	98.8	0.621
CNN-T4SE	59	29	49	1	63.8	70.2	54.6	98.3	0.427
DeepT3_4	101	26	7	4	92.0	94.8	93.5	96.2	0.775

TABLE 2 | Performance Comparisons of DeepT3_4 with other three methods on the independent test dataset 3.

Model	TP	TN	FN	FP	ACC (%)	*F*-value (%)	Recall (%)	PRE (%)	MCC
Bastion3	33	72	2	3	95.5	93.0	94.3	91.7	0.896
Bastion4	27	75	8	0	92.7	87.1	77.1	100.0	0.835
CNN-T4SE	28	75	7	0	93.6	88.9	80.0	100.0	0.855
DeepT3_4	32	72	3	3	94.5	91.4	91.4	91.4	0.874

spark a wave of researches on bacterial non-classical secreted proteins. Overall, we propose an effective computational method to accurately differentiate between T3SEs and T4SEs in this work, and hope it could facilitate more relevant researches on bacterial secreted effectors.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. These data can be found here: https://github.com/jingry/autoBioSeqpy/tree/2.0/examples/T3T4.

AUTHOR CONTRIBUTIONS

JL and LY conceived the study and wrote the manuscript. RJ contributed to the design, implementation, and testing of the model. LY and FL performed the data analysis. All authors read and agreed to the published version of the manuscript.

FUNDING

This research was funded by the National Natural Science Foundation of China (21803045), the Fund of Science and Technology Department of Guizhou Province [(2017)5790-07], and The Development Program for Youth Science and Technology Talents in Education Department of Guizhou Province [KY (2016)219].

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2021.605782/full#supplementary-material

REFERENCES

Açıcı, K., Aşuroğlu, T., Erdaş, ÇB., and Oğul, H. (2019). T4SS effector protein prediction with deep learning. *Data* 4:45. doi: 10.3390/data4010045

Altschul, S. F., and Koonin, E. V. (1998). Iterated Profile Searches with PSI-BLAST—A tool for discovery in protein databases. *Trends Biochem. Sci.* 23, 444–447. doi: 10.1016/S0968-0004(98)01298-5

An, Y., Wang, J., Li, C., Leier, A., Marquez-Lago, T., Wilksch, J., et al. (2018). Comprehensive assessment and performance improvement of effector protein predictors for bacterial secretion systems III, IV and VI. *Brief. Bioinform.* 19, 148–161. doi: 10.1093/bib/bbw100

An, Y., Wang, J., Li, C., Revote, J., Zhang, Y., Naderer, T., et al. (2017). SecretEPDB: a comprehensive web-based resource for secreted effector proteins of the bacterial types III, IV and VI secretion systems. *Sci. Rep.* 7:41031. doi: 10.1038/srep41031

Arnold, R., Brandmaier, S., Kleine, F., Tischler, P., Heinz, E., Behrens, S., et al. (2009). Sequence-based prediction of type III secreted proteins. *PLoS Pathog.* 5:e1000376. doi: 10.1371/journal.ppat.1000376

Bi, D., Liu, L., Tai, C., Deng, Z., Rajakumar, K., and Ou, H. Y. (2013). SecReT4: a web-based bacterial type IV secretion system resource. *Nucleic Acids Res.* 41, D660–D665. doi: 10.1093/nar/gks1248

Blocker, A., Komoriya, K., and Aizawa, S. (2003). Type III secretion systems and bacterial flagella: insights into their function from structural similarities. *Proc. Natl. Acad. Sci. U. S. A.* 100, 3027–3030. doi: 10.1073/pnas.0535335100

Bogard, N., Linder, J., Rosenberg, A. B., and Seelig, G. (2019). A deep neural network for predicting and engineering alternative polyadenylation. *Cell* 178, 91–106. doi: 10.1016/j.cell.2019.04.046

Burstein, D., Zusman, T., Degtyar, E., Viner, R., Segal, G., and Pupko, T. (2009). Genome-scale identification of *Legionella pneumophila* effectors using a machine learning approach. *PLoS Pathog.* 5:e1000508. doi: 10.1371/journal.ppat.1000508

Chen, T., Wang, X., Chu, Y., Wei, D., and Xiong, Y. (2020). T4SE-XGB: interpretable sequence-based prediction of type IV secreted effectors using eXtreme gradient boosting algorithm. *bioRxiv [Preprint]* doi: 10.1101/2020.06.18.158253

Costa, T. R., Felisberto-Rodrigues, C., Meir, A., Prevost, M. S., Redzej, A., Trokter, M., et al. (2015). Secretion systems in Gram-negative bacteria: structural and mechanistic insights. *Nat. Rev. Microbiol.* 13, 343–359. doi: 10.1038/nrmicro3456

Ding, S., and Zhang, S. (2016). A Gram-negative bacterial secreted protein types prediction method based on PSI-BLAST profile. *Biomed Res. Int*. 2016:3206741. doi: 10.1155/2016/3206741

Dong, X., Zhang, Y. J., and Zhang, Z. (2013). Using weakly conserved motifs hidden in secretion signals to identify type-III effectors from bacterial pathogen genomes. *PLoS One* 8:e56632. doi: 10.1371/journal.pone.0056632

Elbasir, A., Moovarkumudalvan, B., Kunji, K., Kolatkar, P. R., Mall, R., and Bensmail, H. (2019). DeepCrystal: a deep learning framework for sequence-based protein crystallization prediction. *Bioinformatics* 35, 2216–2225. doi: 10.1093/bioinformatics/bty953

Esna Ashari, Z., Brayton, K. A., and Broschat, S. L. (2019a). Prediction of T4SS effector proteins for Anaplasma phagocytophilum Using OPT4e, a new software tool. *Front. Microbiol.* 10:1391. doi: 10.3389/fmicb.2019.01391

Esna Ashari, Z., Brayton, K. A., and Broschat, S. L. (2019b). Using an optimal set of features with a machine learning-based approach to predict effector proteins for *Legionella pneumophila*. *PLoS One* 14:e0202312. doi: 10.1101/383570

Esna Ashari, Z., Dasgupta, N., Brayton, K. A., and Broschat, S. L. (2018). An optimal set of features for predicting type IV secretion system effector proteins for a subset of species based on a multi-level feature selection approach. *PLoS One* 13:e0197041. doi: 10.1371/journal.pone.0197041

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., et al. (2019). A guide to deep learning in healthcare. *Nat. Med.* 25, 24–29. doi: 10.1038/s41591-018-0316-z

Fu, X., and Yang, Y. (2019). WEDeepT3: predicting type III secreted effectors based on word embedding and deep learning. *Quant. Biol.* 7, 293–301. doi: 10.1007/s40484-019-0184-7

Galan, J. E., and Waksman, G. (2018). Protein-Injection machines in bacteria. *Cell* 172, 1306–1318. doi: 10.1016/j.cell.2018.01.034

Goldberg, T., Rost, B., and Bromberg, Y. (2016). Computational prediction shines light on type III secretion origins. *Sci. Rep.* 6:34516. doi: 10.1038/srep34516

Hobbs, C. K., Porter, V. L., Stow, M. L., Siame, B. A., Tsang, H. H., and Leung, K. Y. (2016). Computational approach to predict species-specific type III secretion system (T3SS) effectors using single and multiple genomes. *BMC Genomics* 17:1048. doi: 10.1186/s12864-016-3363-1

Hong, J., Luo, Y., Mou, M., Fu, J., Zhang, Y., Xue, W., et al. (2020). Convolutional neural network-based annotation of bacterial type IV secretion system effectors with enhanced accuracy and reduced false discovery. *Brief. Bioinform.* 21, 1825–1836. doi: 10.1093/bib/bbz120

Huang, Y., Niu, B., Gao, Y., Fu, L., and Li, W. (2010). CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* 26, 680–682. doi: 10.1093/bioinformatics/btq003

Jiang, J., Chen, Y., and Narayan, A. (2017). Offline-enhanced reduced basis method through adaptive construction of the surrogate training set. *J. Sci. Comput.* 73, 853–875. doi: 10.1007/s10915-017-0551-3

Jing, R., Li, Y., Xue, L., Liu, F., Li, M., and Luo, J. (2020). autoBioSeqpy: a deep learning tool for the classification of biological sequences. *J. Chem. Inf. Model.* 60, 3755–3764. doi: 10.1021/acs.jcim.0c00409

Jurtz, V. I., Johansen, A. R., Nielsen, M., Armenteros, J. J. A., Nielsen, H., Sønderby, C. K., et al. (2017). An introduction to deep learning on biological sequence data: examples and solutions. *Bioinformatics* 33, 3685–3690. doi: 10.1093/bioinformatics/btx531

Khurana, S., Rawi, R., Kunji, K., Chuang, G. Y., Bensmail, H., and Mall, R. (2018). DeepSol: a deep learning framework for sequence-based protein solubility prediction. *Bioinformatics* 34, 2605–2613. doi: 10.1093/bioinformatics/bty166

Kim, B. S. (2018). The modes of action of MARTX toxin effector domains. *Toxins* 10:507. doi: 10.3390/toxins10120507

Kong, L., and Zhang, L. (2019). An ensemble method for multi-type Gram negative bacterial secreted protein prediction by integrating different PSSM-based features. *SAR QSAR Environ. Res.* 30, 181–194. doi: 10.1080/1062936X.2019.1573438

Kruse, R., Borgelt, C., Klawonn, F., Moewes, C., Steinbrecher, M., and Held, P. (2013). "Multi-layer perceptrons," in *Computational Intelligence*, ed. D. Inkpen (Berlin: Springer), 47–81.

Lasica, A. M., Ksiazek, M., Madej, M., and Potempa, J. (2017). The type IX secretion system (T9SS): highlights and recent insights into its structure

and function. *Front. Cell. Infect. Microbiol.* 7:215. doi: 10.3389/fcimb.2017.00215

Lauber, F., Deme, J. C., Lea, S. M., and Berks, B. C. (2018). Type 9 secretion system structures reveal a new protein transport mechanism. *Nature* 564, 77–82. doi: 10.1038/s41586-018-0693-y

Li, J., Li, Z., Luo, J., and Yao, Y. (2020a). ACNNT3: Attention-CNN Framework for prediction of sequence-based bacterial type III secreted effectors. *Comput. Math. Methods Med.* 2020:3974598. doi: 10.1155/2020/3974598

Li, J., Wei, L., Guo, F., and Zou, Q. (2020b). EP3: an ensemble predictor that accurately identifies type III secreted effectors. *Brief. Bioinform* bbaa008. doi: 10.1093/bib/bbaa008

Liang, Y., and Zhang, S. (2018). Identify Gram-negative bacterial secreted protein types by incorporating different modes of PSSM into Chou's general PseAAC via Kullback-Leibler divergence. *J. Theor. Biol.* 454, 22–29. doi: 10.1016/j.jtbi.2018.05.035

Liang, Y., Zhang, S., and Ding, S. (2018). Accurate prediction of Gram-negative bacterial secreted protein types by fusing multiple statistical features from PSI-BLAST profile. *SAR QSAR Environ. Res.* 29, 469–481. doi: 10.1080/1062936X.2018.1459835

Lifshitz, Z., Burstein, D., Peeri, M., Zusman, T., Schwartz, K., Shuman, H. A., et al. (2013). Computational modeling and experimental validation of the *Legionella* and *Coxiella* virulence-related type-IVB secretion signal. *Proc. Natl. Acad. Sci. U.S.A.* 110, E707–E715. doi: 10.1073/pnas.1215278110

Löwer, M., and Schneider, G. (2009). Prediction of type III secretion signals in genomes of Gram-negative bacteria. *PLoS One* 4:e5917. doi: 10.1371/journal.pone.0005917

McInnes, L., and Healy, J. (2018). UMAP: uniform manifold approximation and projection for dimension reduction. *ArXiv [Preprint]* doi: 10.21105/joss.00861

McQuade, R., and Stock, S. P. (2018). Secretion systems and secreted proteins in Gram-negative entomopathogenic bacteria: their roles in insect virulence and beyond. *Insects* 9:68. doi: 10.3390/insects9020068

Melville, J. (2019). *uwot: The Uniform Manifold Approximation and Projection (UMAP) Method for Dimensionality Reduction*. Available online at: https://github.com/jlmelville/uwot (accessed October, 2020).

Nagai, H., Cambronne, E. D., Kagan, J. C., Amor, J. C., Kahn, R. A., and Roy, C. R. (2005). A C-terminal translocation signal required for Dot/Icm-dependent delivery of the *Legionella* RalF protein to host cells. *Proc. Natl. Acad. Sci. U. S. A.* 102, 826–831. doi: 10.1073/pnas.0406239101

Pan, X., Rijnbeek, P., Yan, J., and Shen, H. B. (2018). Prediction of RNA-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks. *BMC Genomics* 19:511. doi: 10.1186/s12864-018-4889-1

Puhar, A., and Sansonetti, P. J. (2014). Type III secretion system. *Curr. Biol.* 24, 784–791. doi: 10.1016/j.cub.2014.07.016

Quang, D., and Xie, X. (2016). DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res.* 44:e107. doi: 10.1093/nar/gkw226

Samudrala, R., Heffron, F., and McDermott, J. E. (2009). Accurate prediction of secreted substrates and identification of a conserved putative secretion signal for type III secretion systems. *PLoS Pathog.* 5:e1000375. doi: 10.1371/journal.ppat.1000375

Sen, R., Nayak, L., and De, R. K. (2019). PyPredT6: a python-based prediction tool for identification of Type VI effector proteins. *J. Bioinf. Comput. Biol.* 17:1950019. doi: 10.1142/S0219720019500197

Tang, B., Pan, Z., Yin, K., and Khateeb, A. (2019). Recent advances of deep learning in bioinformatics and computational biology. *Front. Genet.* 10:214. doi: 10.3389/fgene.2019.00214

Tayara, H., and Chong, K. T. (2019). Improving the quantification of DNA sequences using evolutionary information based on deep learning. *Cells* 8:1635. doi: 10.3390/cells8121635

Veltri, D., Kamath, U., and Shehu, A. (2018). Deep learning improves antimicrobial peptide recognition. *Bioinformatics* 34, 2740–2747. doi: 10.1093/bioinformatics/bty179

Wang, C., Li, J., Zhang, Y., and Guo, M. (2020). Identification of Type VI effector proteins using a novel ensemble classifier. *IEEE Access* 8, 75085–75093. doi: 10.1109/ACCESS.2020.2985111

Wang, J., Li, J., Yang, B., Xie, R., Marquez-Lago, T. T., Leier, A., et al. (2019a). Bastion3: a two-layer ensemble predictor of type III secreted effectors. *Bioinformatics* 35, 2017–2028. doi: 10.1093/bioinformatics/bty914

Wang, J., Yang, B., An, Y., Marquez-Lago, T. T., Leier, A., Wilksch, J., et al. (2019b). Systematic analysis and prediction of type IV secreted effector proteins by machine learning approaches. *Brief. Bioinform.* 20, 931–951. doi: 10.1093/bib/bbx164

Wang, J., Yang, B., Leier, A., Marquez-Lago, T. T., Hayashida, M., Rocker, A., et al. (2018). Bastion6: a bioinformatics approach for accurate prediction of type VI secreted effectors. *Bioinformatics* 34, 2546–2555. doi: 10.1093/bioinformatics/bty155

Wang, Y., Guo, Y., Pu, X., and Li, M. (2017). Effective prediction of bacterial type IV secreted effectors by combined features of both C-termini and N-termini. *J. Comput. Aided Mol. Des.* 31, 1029–1038. doi: 10.1007/s10822-017-0080-z

Wang, Y., Sun, M., Bao, H., and White, A. P. (2013). T3_MM: a Markov model effectively classifies bacterial type III secretion signals. *PLoS One* 8:e58173. doi: 10.1371/journal.pone.0058173

Wang, Y., Wei, X., Bao, H., and Liu, S. L. (2014). Prediction of bacterial type IV secreted effectors by C-terminal features. *BMC Genomics* 15:50. doi: 10.1186/1471-2164-15-50

Wang, Y., Zhang, Q., Sun, M. A., and Guo, D. (2011). High-accuracy prediction of bacterial type III secreted effectors based on position-specific amino acid composition profiles. *Bioinformatics* 27, 777–784. doi: 10.1093/bioinformatics/btr021

Xiong, Y., Wang, Q., Yang, J., Zhu, X., and Wei, D. Q. (2018). PredT4SE-Stack: prediction of bacterial type IV secreted effectors from protein sequences using a stacked ensemble method. *Front. Microbiol.* 9:2571. doi: 10.3389/fmicb.2018.02571

Xu, S., Zhang, C., Miao, Y., Gao, J., and Xu, D. (2010). Effector prediction in host-pathogen interaction based on a Markov model of a ubiquitous EPIYA motif. *BMC Genomics* 11:S1. doi: 10.1186/1471-2164-11-S3-S1

Xue, L., Tang, B., Chen, W., and Luo, J. (2018). A deep learning framework for sequence-based bacteria type IV secreted effectors prediction. *Chemom. Intell. Lab. Syst.* 183, 134–139. doi: 10.1016/j.chemolab.2018.11.002

Xue, L., Tang, B., Chen, W., and Luo, J. (2019). DeepT3: deep convolutional neural networks accurately identify Gram-negative bacterial type III secreted effectors using the N-terminal sequence. *Bioinformatics* 35, 2051–2057. doi: 10.1093/bioinformatics/bty931

Yang, X., Guo, Y., Luo, J., Pu, X., and Li, M. (2013). Effective identification of Gram-negative bacterial type III secreted effectors using position-specific residue conservation profiles. *PLoS One* 8:e84439. doi: 10.1371/journal.pone.0084439

Yang, Y., Zhao, J., Morgan, R. L., Ma, W., and Jiang, T. (2010). Computational prediction of type III secreted proteins from Gram-negative bacteria. *BMC Bioinformatics* 11(Suppl. 1):S47. doi: 10.1186/1471-2105-11-S1-S47

Yu, L., Liu, F., Du, L., and Li, Y. (2018). An improved approach for rapidly identifying different types of Gram-negative bacterial secreted proteins. *Nat. Sci.* 10, 168–177. doi: 10.4236/ns.2018.105018

Yu, L., Luo, J., Guo, Y., Li, Y., Pu, X., and Li, M. (2013). In silico identification of Gram-negative bacterial secreted proteins from primary sequence. *Comput. Biol. Med.* 43, 1177–1181. doi: 10.1016/j.compbiomed.2013.06.001

Zeng, C., and Zou, L. (2019). An account of *in silico* identification tools of secreted effector proteins in bacteria and future challenges. *Brief. Bioinform.* 20, 110–129. doi: 10.1093/bib/bbx078

Zhang, Y., Yu, S., Xie, R., Li, J., Leier, A., Marquez-Lago, T. T., et al. (2020). PeNGaRoo, a combined gradient boosting and ensemble learning framework for predicting non-classical secreted proteins. *Bioinformatics* 36, 704–712. doi: 10.1093/bioinformatics/btz629

Zou, L., Nan, C., and Hu, F. (2013). Accurate prediction of bacterial type IV secreted effectors using amino acid composition and PSSM profiles. *Bioinformatics* 29, 3135–3142. doi: 10.1093/bioinformatics/btt554

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

TECHNOLOGY AND CODE
published: 28 January 2021
doi: 10.3389/fmicb.2021.609048

kernInt: A Kernel Framework for Integrating Supervised and Unsupervised Analyses in Spatio-Temporal Metagenomic Datasets

Elies Ramon[1], Lluís Belanche-Muñoz[2], Francesc Molist[3], Raquel Quintanilla[4], Miguel Perez-Enciso[1,5] and Yuliaxis Ramayo-Caldas[4]*

[1] Plant and Animal Genomics, Statistical and Population Genomics Group, CSIC-IRTA-UAB-UB Consortium, Centre for Research in Agricultural Genomics (CRAG), Bellaterra, Spain, [2] Department of Computer Science, Polytechnic University of Catalonia, Barcelona, Spain, [3] Schothorst Feed Research B.V., Lelystad, Netherlands, [4] Animal Breeding and Genetics Program, IRTA, Caldes de Montbui, Spain, [5] Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain

OPEN ACCESS

Edited by:
Isabel Moreno Indias, University of Málaga, Spain

Reviewed by:
Xiaoquan Su, Qingdao University, China
Jun Chen, Mayo Clinic, United States
Yang Dai, University of Illinois at Chicago, United States

***Correspondence:**
Elies Ramon
eramongurrea@gmail.com

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

Received: *22 September 2020*
Accepted: *07 January 2021*
Published: *28 January 2021*

Citation:
Ramon E, Belanche-Muñoz L, Molist F, Quintanilla R, Perez-Enciso M and Ramayo-Caldas Y (2021) kernInt: A Kernel Framework for Integrating Supervised and Unsupervised Analyses in Spatio-Temporal Metagenomic Datasets. Front. Microbiol. 12:609048. doi: 10.3389/fmicb.2021.609048

The advent of next-generation sequencing technologies allowed relative quantification of microbiome communities and their spatial and temporal variation. In recent years, supervised learning (i.e., prediction of a phenotype of interest) from taxonomic abundances has become increasingly common in the microbiome field. However, a gap exists between supervised and classical unsupervised analyses, based on computing ecological dissimilarities for visualization or clustering. Despite this, both approaches face common challenges, like the compositional nature of next-generation sequencing data or the integration of the spatial and temporal dimensions. Here we propose a kernel framework to place on a common ground the unsupervised and supervised microbiome analyses, including the retrieval of microbial signatures (taxa importances). We define two compositional kernels (Aitchison-RBF and compositional linear) and discuss how to transform non-compositional beta-dissimilarity measures into kernels. Spatial data is integrated with multiple kernel learning, while longitudinal data is evaluated by specific kernels. We illustrate our framework through a single point soil dataset, a human dataset with a spatial component, and a previously unpublished longitudinal dataset concerning pig production. The proposed framework and the case studies are freely available in the *kernInt* package at https://github.com/elies-ramon/kernInt.

Keywords: microbiome, metagenomics, kernel, supervised, unsupervised, spatio-temporal, SVM, kPCA

Abbreviations: ANN, artificial neural network; ASV, amplicon sequence variant; JSK, Jensen-Shannon Kernel; kPCA, kernel principal components analysis; MDS, multidimensional scaling; MKL, multiple kernel learning; NGS, next-generation sequencing; NMSE, normalized mean squared error; OTU, operational taxonomic unit; PCA, principal components analysis; PCoA, principal coordinates analysis; RBF, radial basis function; RF, random forests; SVM, support vector machines

INTRODUCTION

The microbiome is defined as the ensemble of microorganisms and their genomes in a given environment. Microorganisms are present in ecological niches as diverse as soil, oceans, freshwater, plants, and animals, but a large fraction of these taxa cannot be cultivated with culture-dependent methods. The advent of next-generation sequencing (NGS) revolutionized this field by allowing the massive sequencing and quantification of microbial habitats.

Proper analysis of microbiome data is challenging for a variety of reasons. Abundance data obtained with NGS is multivariate, sparse and compositional in nature (Gloor et al., 2017). Also, microbial communities are very dynamic biological systems, thus justifying spatial or time-course studies (Bodein et al., 2019; Berg et al., 2020). The first approach on the field used statistical tools from standard ecological studies (Gloor et al., 2017). For example, one of the first steps in nearly all microbiome studies consists in computing alpha and beta-diversities. Beta-diversity measures, e.g., Bray-Curtis or Unifrac, quantify the difference in diversity between samples from different habitats. They are used for clustering analysis or, more commonly, for visualization techniques like principal coordinates analysis (PCoA) or multidimensional scaling (MDS). However, this approach has been challenged, as the abundance data obtained by NGS has a particular nature. The total number of reads delivered is bounded by an uninformative sum: the library size (i.e., the number of total reads per sample). Library size is uninformative because it does not contain information about the population. Instead, it is arbitrarily fixed by the sequencing process and may vary by orders of magnitude across samples (McMurdie and Holmes, 2014). This kind of data is called compositional and deserves a specific mathematical treatment (Gloor et al., 2017). In the case of metagenomics, extensive research is being done to translate current statistical techniques to this paradigm (Gloor et al., 2017; Silverman et al., 2017; Rivera-Pinto et al., 2018). One example is the proposal of using the compositional Aitchison distance instead of the classic beta-diversity measures (Quinn et al., 2018).

In machine learning, the aforementioned clustering, ordination and visualization techniques belong to the so-called unsupervised learning. Supervised learning, which is focused on prediction, is not so widespread in microbiome analysis yet, but the number of studies using this kind of approach is rapidly growing in the last years (Zhou and Gallins, 2019). Due to this rise in popularity, widely used libraries for microbiome analysis like QIIME2 (Bolyen et al., 2019) now include plugins for supervised learning in their toolbox. Typical available methods include random forests (RF), artificial neural networks (ANN), support vector machines (SVM), and ridge regression (Qu et al., 2019; Zhou and Gallins, 2019; Namkung, 2020). Among the aforementioned, RF are popular in the microbiome context and tend to outperform other methods (Zhou and Gallins, 2019; Namkung, 2020). ANN have shown excellent performance in some cases but are susceptible to overfitting, especially if sample size is greatly exceeded by the number of taxa, as is often the case in metagenomics and metataxonomics. A desirable feature for supervised methods is the identification of microbial signatures (i.e., taxa that are predictive of a certain phenotype), which may enable a biological interpretation of the results. RF are endowed with variable importance measures that can be used to this effect, while there is not such straightforward heuristic for ANN, although several possible strategies exist (Ibrahim, 2013). Another supervised method, *selbal* (Rivera-Pinto et al., 2018), is focused on the identification of microbial signatures based on balances (i.e., the geometric means of data from two groups of taxa), and has the particularity of being purely compositional.

As microbial communities are highly dynamic systems, it is important to address their spatial and/or temporal variation (Berg et al., 2020). In spatial-structured studies, repeated samples of different sites (e.g., body sites, depth layers) are obtained from the same individuals or entities, thus raising the question of how to integrate them. A more general challenge is the integration of datasets coming from different sources (e.g., "omics"), which may have different data types. Several statistical methods have been proposed to solve this question in the microbiome field. Some examples are *Link-HD* (Zingaretti et al., 2020), *mixKernel* (Mariette et al., 2018), and *MOFA* (Argelaguet et al., 2018), all focused in the unsupervised learning setting. In most supervised methods, this integration is usually performed at the input data level (early integration), for example by concatenating the datasets; or after the model is built (late integration), combining their scores as in ensemble methods. However, early integration may be not possible if data nature differs across sources (Schölkopf et al., 2004). The case of the longitudinal studies (which follow the evolution over time of microbial communities) is more complex. Typically, longitudinal data is modeled by fitting a function (e.g., polynomial interpolation, splines) to the data points over time. To date, there exist few analytical tools for this kind of data in the microbiome field. Two examples can be found at Bodein et al. (2019) and Coenen et al. (2020), but they are restricted to unsupervised analysis.

Difficulties like the compositionality of data or how to accommodate the spatial and temporal dimensions affect supervised and unsupervised methods alike. However, there is a gap between the most widely used supervised learning methods and the unsupervised analyses typical of the microbiome field (**Figure 1A**). Libraries like QIIME2 juxtapose traditional analyses (e.g., PCoA) with many different and powerful prediction algorithms, but both branches remain independent from a mathematical point of view. It is true that some beta-diversity dissimilarity-based engines can be used as classifiers (Su et al., 2020; Shenhav et al., 2019). However, as these tools are strongly focused on distinguishing among a limited number of bacteria-related conditions, they are not aimed at regression problems, nor do they give any information about the microbial signatures. We consider that carrying out all aforementioned analyses in a common mathematical framework would provide a new, holistic view to microbiome studies. With all this in mind we propose a generic and flexible kernel framework (**Figure 1B**) as a way to handle unsupervised and supervised microbiome analyses (including the retrieval of microbial signatures), while paying special attention to data compositionality and spatial and temporal integration. Kernel methods are a family within

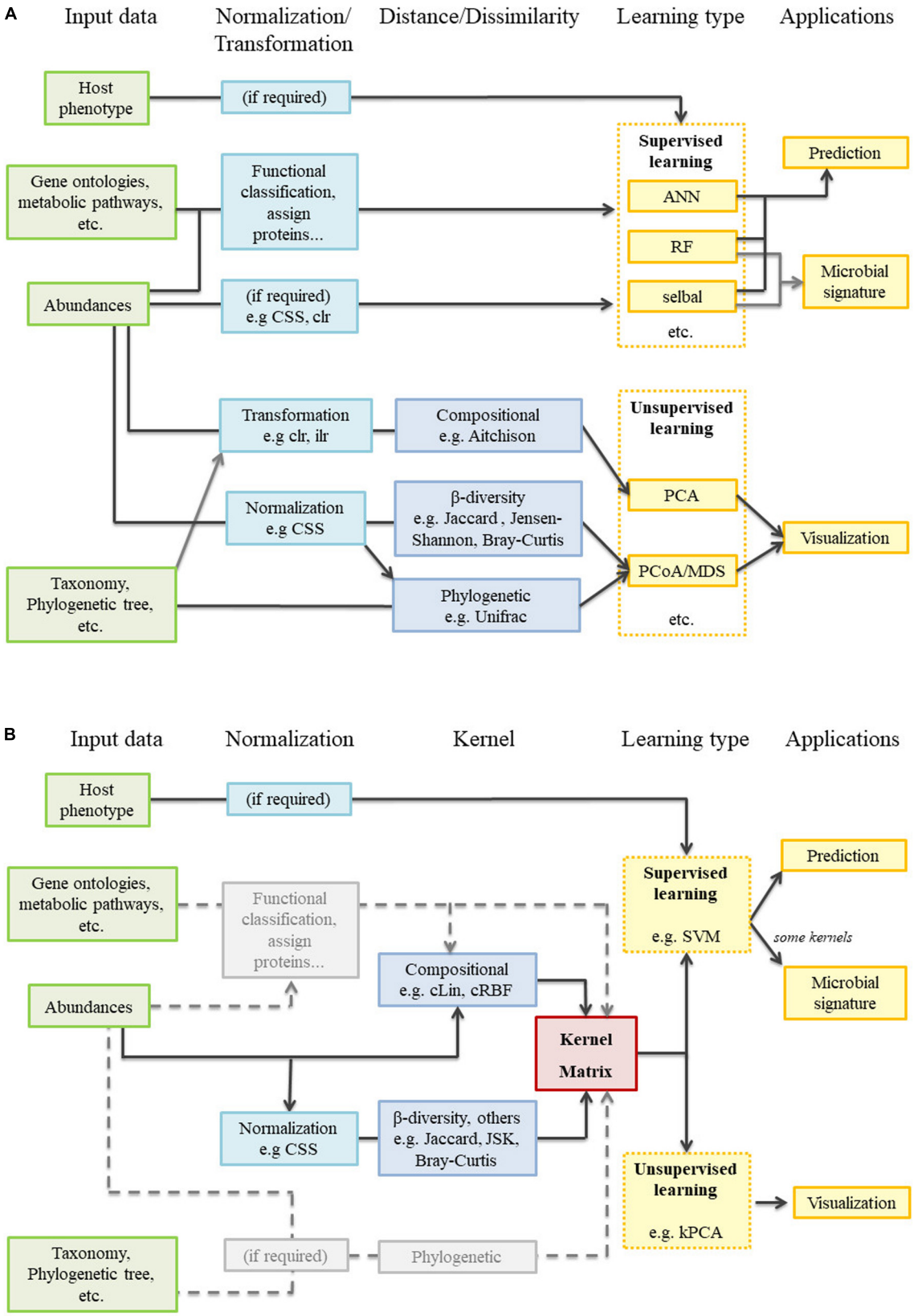


FIGURE 1 | Metagenomic analysis workflow. **(A)** Current state-of-the art: supervised and unsupervised learning are completely independent. **(B)** Kernel framework: the pivotal position of the kernel matrix is clearly observed. In gray, several tasks not performed during the present work but that merit future research.

machine learning methods that share the use of kernel functions or, simply, kernels. Some of these methods have been already applied to some specific problems or areas within microbiome analysis (Zhan et al., 2017; Mariette et al., 2018; Zhou and Gallins, 2019) but their potential has not been fully exploited. In this work, we propose two new compositional kernels and discuss how to translate non-compositional, but nonetheless widespread, beta-diversity matrices to the kernel framework. We perform supervised and unsupervised analyses from the same kernel matrix, and show how to extract microbial signatures. Spatial and longitudinal data are also treated with specific kernel tools. This kernel framework is illustrated with three case studies: a single point soil metagenomic dataset, a human dataset with a spatial component, and a previously unpublished longitudinal dataset concerning pig gut microbiota. An R package implementing the proposed methods, along with the analyzed datasets, is freely available at https://github.com/elies-ramon/kernInt.

MATERIALS AND METHODS

Kernels for Microbiome Data

A real symmetric two-place function is a kernel iff, for every finite set of objects $x_1, \ldots, x_N$, it generates a positive semi-definite matrix of dimension $N \times N$: the kernel matrix (Schölkopf et al., 2004; Shawe-Taylor and Cristianini, 2004). Probably the most widely known and used kernel functions are the linear and radial basis function (RBF) kernels, both defined for real vectors.

Intuitively, a kernel can be understood as a measure of the similarity between x_i and x_j. As objects $x_1, \ldots, x_N$ are never represented explicitly, kernels can be designed for non-standard data types if a notion of what is considered "similar" in that given context exists (Schölkopf et al., 2004). Each kernel provides a different grasp of the dataset. Furthermore, as similarity measures, kernels are related (but opposite) to the beta-diversities widely used in microbiome analyses. However, although every beta-diversity distance or dissimilarity is paired with a similarity counterpart, not all of them fulfill the aforementioned conditions and are, therefore, kernels.

We now present two compositional and two non-compositional kernels, all of them available in *kernInt*. In addition, users have the option of entering any kernel matrix, pre-computed with a kernel of their choice. In this work we are restricted only to kernels that can be obtained from taxonomic abundance tables, but further insights can be found in the Discussion.

Compositional Kernels

Here we define two kernels analogous to the linear and RBF kernels, but specific for compositional data. We introduce the Aitchison-RBF kernel as:

$$cRBF(\mathbf{x_i}, \mathbf{x_j}) = exp\left(-\gamma \sum_{k=1}^{D}\left(log\left(\frac{x_{ik}}{G(\mathrm{x_i})}\right) - log\left(\frac{x_{jk}}{G(\mathrm{x_j})}\right)\right)^2\right) \quad (1)$$

where x_i and x_j represent the taxonomic abundances in two different individuals, D is the number of different taxa, $G(.)$ is the geometric mean, and $\gamma > 0$ is a hyperparameter that has to be tuned. This non-linear kernel derives from the Aitchison distance, which is Euclidean and therefore (Eq. 1) is a valid kernel. The logarithm term can be identified as the compositional clr-transformation (Gloor et al., 2017) over the original data.

Analogously, we define the compositional linear kernel as:

$$cLin\ (\mathbf{x_i}, \mathbf{x_j}) = \sum_{k=1}^{D} log\left(\frac{x_{ik}}{G(\mathrm{x_i})}\right)\ log\left(\frac{x_{jk}}{G(\mathrm{x_j})}\right) \quad (2)$$

Although cRBF is related to Aitchison distance and has the advantage of non-linearity, cLin is easier to interpret and allows the retrieval of the microbial signatures.

Non-compositional Kernels

The most widely beta-diversity measures are Bray-Curtis, Unifrac and Jensen-Shannon (Gloor et al., 2017). Bray-Curtis and Jensen-Shannon are computed from taxonomic tables, while Unifrac additionally needs a phylogenetic tree. The Jensen-Shannon is metric and has a kernel counterpart that is already described in Bai and Hancock (2011) as the Jensen-Shannon Kernel (JSK):

$$JSK\left(\mathbf{x_i}, \mathbf{x_j}\right) = 1 - \frac{1}{2}\left[\sum_{k=1}^{D} x_{ik}\ \ln\left(\frac{2\ x_{ik}}{x_{ik}\ +\ x_{jk}}\right) + \sum_{k=1}^{D} x_{jk}\ ln\left(\frac{2\ x_{jk}}{x_{ik}\ +\ x_{jk}}\right)\right] \quad (3)$$

provided that x_i and x_j contain relative frequencies. The Bray-Curtis dissimilarity is semimetric, and so we propose using Jaccard, a similar distance (Gardener, 2014), instead. The Jaccard distance is paired with a well-known kernel (Bouchard et al., 2013) and has a variant suitable for quantitative data. The quantitative Jaccard (also known as Ružička) kernel is defined in Gardener (2014) as:

$$qJac(\mathbf{x_i}, \mathbf{x_j}) = \sum_{k=1}^{D} \frac{min\ \left(x_{ik}, x_{jk}\right)}{max\ \left(x_{ik}, x_{jk}\right)} \quad (4)$$

All aforementioned kernels have an asymptotic computational complexity of $O(N^2\ D)$.

Kernel Methods and Framework

Kernel methods share the use of symmetric and positive semi-definite matrices (i.e., kernel matrices), and not the original data, as input. That limits the potential similarity measures that one can use to only valid kernels, but also guarantees that every matrix generated can be processed by the kernel method. Furthermore, using kernels places all different analyses in a common mathematical ground (see **Figure 1B**), which we refer as the kernel framework. For phenotype prediction, we use SVM, a classical method that can perform regression and classification (both binary and multi-class). For the unsupervised analyses we use kernel principal components analysis (kPCA), a kernelized version of the standard algorithm. In both

cases, *kernInt* allows the user to choose the values of the hyperparameters and (in the case of SVM) to perform a complete cross-validation and performance evaluation using an independent test set.

Spatial Data

The kernel framework is particularly well suited for the integration of spatial or heterogeneous data types (Schölkopf et al., 2004; Mariette et al., 2018). This is because the integration can be done directly at the kernel matrices level. Let $\mathbf{K_1}, \ldots, \mathbf{K_M}$ be the kernel matrices computed from M different sources of data coming from the same individuals. Then, we can obtain a consensus kernel matrix $\mathbf{K}^*$:

$$\mathbf{K}^* = \sum_{z=1}^{M} \beta_z \, \mathbf{K_z} \tag{5}$$

with the restriction $\beta_z \geq 0$. The optimal β_z values can be obtained through an optimization process, which is known as multiple kernel learning (MKL) (Schölkopf et al., 2004). In unsupervised scenarios, a consensus matrix $\mathbf{K}^*$ can be obtained by choosing the β coefficients that maximize average similarity of $\mathbf{K}^*$ with all $\mathbf{K_z}$ matrices (Mariette et al., 2018).

Temporal Data

A time series is an ordered set of repeated samples indexed by time, in the form $\{x_i, t_i\}$. The natural way to summarize this type of data is through a function, which can be obtained using polynomial interpolation or splines. When data contains the time series of several individuals, it is commonly referred as longitudinal data.

The functional RBF kernel (Chen et al., 2013) translates the RBF kernel to accept real functions as input. Therefore, evolution over time among individuals is compared and used afterward for phenotype prediction or unsupervised tasks. Let $f(t)$ and $g(t)$ be univariate functions, so that they represent the variation of a single feature in two different individuals between the time interval $[t_a, t_b]$. Then, the kernel definition is:

$$fRBF\ (f, g) = exp\left(-\gamma \int_{t_a}^{t_b} \left|f\,(t) - g\,(t)\right|^2 dt\right) \tag{6}$$

In an analogous way, the functional linear kernel is defined as:

$$fLin\ (f, g)\ =\ \int_{t_a}^{t_b} f\,(t)\ g\,(t)\ dt \tag{7}$$

These kernels allow irregular sampling intervals and missing time points, but suffer of the cost of computing numerically the integral (e.g., if an algebraic solution is not possible). Computations can be simplified if fLin and fRBF are defined for discrete functions, so the modeling of time series as continuous functions is skipped. In this case, $f(t)$ and $g(t)$ may directly denote the original objects $\{x_i, t_i\}$, so each time value directly maps to a certain value of the feature variable x. If T is the total number of time points and Δt the time increment, then:

$$fRBF\ (f, g) =\ exp(-\gamma \sum_{i=1}^{T} (f(t_i) - g(t_i))^2) \tag{8}$$

$$fLin\ (f, g)\ =\ \Delta t \sum_{i=1}^{T} f\,(t_i)\ g\,(t_i) \tag{9}$$

The discrete approach is sound in cases with few data points, when the modeling is less reliable. However, contrarily to (Eqs 6, 7), these expressions cannot deal with irregular sampling times or missing data.

In multivariate scenarios, for instance microbiome data, many features are simultaneously sampled over time. Let f_k and g_k model taxon k in two individuals, being D the total number of taxa. The aforementioned kernels can be combined as in:

$$fRBF'(f, g) = \prod_{k=1}^{D} fRBF(f_k,\ g_k) \tag{10}$$

$$fLin'(f, g) = \sum_{k=1}^{D} fLin(f_k,\ g_k) \tag{11}$$

With a computational complexity of $O(N^2 TD)$ if (Eqs 8, 9) are used.

It should be noted that the kernel approach allows the integration of data that is both spatial and temporal-structured. *kernInt* first handles the temporal dimension using a kernel for longitudinal data (fLin or fRBF) over each space point, and then integrates the spatial dimension by performing MKL over the fLin or fRBF kernel matrices coming from the same individual.

Microbial Signature

In a broad sense, the "microbial signature" is the collection of taxa associated with a trait of interest that has a high predictive value in the context of a given model (Rivera-Pinto et al., 2018). It can be retrieved from a linear SVM using the orientation of the separating hyperplane (Guyon et al., 2002): if the plane is orthogonal to a particular feature dimension, then that feature is maximally informative. This method takes into account the correlation between taxa. As cLin is a translation of the linear kernel for compositional data, using (Eq. 2) we can retrieve the microbial signatures, which should be understood as the taxa importances after the clr-transformation. The same occurs when assessing the variable influence on the principal components in kPCA. A general permutation technique is proposed in Mariette et al. (2018), but using cLin permits obtaining the taxa influence in the same straightforward way than standard principal components analysis (PCA).

The linearity also permits extending the microbial signature retrieval, when using SVM, to the longitudinal and spatial cases. When performing MKL, as long as the cLin kernel is strictly applied to all sampled sites, the global importance of a given taxon among all sites can be computed as the weighted sum (using the optimal β coefficients) of its partial importance in

each site. In the longitudinal case, the global importance of each taxon k can be obtained from (Eq. 9) by addition of the partial importances over all T time points.

Case Studies and Data Pre-processing

We illustrate our framework with three case studies: a single point dataset, a dataset with a spatial component, and a longitudinal dataset. The latter is previously unpublished while the rest of the data is public.

Soil Dataset

Bacterial composition of soil varies significantly at a biogeographical scale, and is related to chemical and environmental factors. Here we reanalyzed a single point dataset by Lauber et al. (2009), who used 16S small-subunit ribosomal (16S rRNA) gene pyrosequencing to profile the bacterial communities of different soils across North and South America. Authors reported that soil pH was significantly correlated with beta-diversity distances between samples. They also found correlation with alpha diversity, which was highest in soils with near-neutral pHs. To perform our analysis, we retrieved the taxonomic abundances as well as the associated metadata from Qiita https://qiita.ucsd.edu/ (ID: 103). The number of operational taxonomic unit (OTUs) was 7,396, while the number of soil samples was 89. As a part of the pre-processing, we excluded sample number 89, with only 1 read, which was also not included in the original paper.

Smokers Dataset

Charlson et al. (2010) analyzed the impact of cigarette smoking on the global airway microbial population. Bacterial communities were profiled using 454 pyrosequencing of the 16S rRNA gene in four airway sites: the left and right sides of nasopharynx and oropharynx. Authors reported that composition was primarily determined by airway site, with individuals exhibiting minimal lateral or temporal variation. They used RF to predict the smoking status from the taxonomic abundances. We retrieved the dataset (metadata and taxonomic abundances) from Qiita (ID: 524) to perform our analysis. Of the original 70 individuals, we discarded those that reported airway illness or antibiotic usage in the 3 months prior to sampling. Thus, we analyzed the same 62 individuals of the original work (29 smokers and 33 non-smokers). Number of different OTUs was 2,817.

Pig Dataset

Here we present a previously unpublished dataset, which evaluates the relationship of pre-weaning diarrhea with the early gut microbiota colonization in piglets. Gut microbiota was profiled in 153 piglets during their first week of life. Between days 8 and 21 (weaning day), 79 out of the 153 piglets had diarrhea and were treated with antibiotics. Swab sampling was done within 5 min after farrowing (day 0) and at days 3 and 7 post-farrowing. DNA was extracted from fecal samples and profiled using Illumina sequencing of 16S rRNA gene in each of the three time points. The cleaned sequences were processed into amplicon sequence variants (ASVs). Further details are described in **Supplementary Method 0**. Analyses were carried out at the ASV (3,577 ASVs were obtained) and at the Genera taxonomic levels.

Experimental Set-Up

Analyses across the three datasets included a comparison with the original reports (for Soil and Smokers datasets), as well as contrast with results from RF. The cLin and cRBF kernels were applied directly to the raw counts, as they handle data in an inherently compositional manner. Before computing both kernels, a number under the detection limit was added to all dataset entries to handle zeroes (Quinn et al., 2018). An alternative normalization of data, the cumulative sum scaling (Paulson et al., 2013) was performed prior to applying the non-compositional Jensen-Shannon and Jaccard kernels. That way the compositional and non-compositional kernels could be compared. In the rest of cases (RF and longitudinal) we used the compositional clr-transform over data. RF were obtained with the R package *randomForest* (Liaw and Wiener, 2002), while the kernel approach was carried out using *kernInt* (which relies on the *kernlab* package for computing kPCA and SVM). A step-by-step guide with examples can be found at the *kernInt* package vignette: https://elies-ramon.github.io/kernInt/.

Unsupervised analyses were carried out using the whole datasets. Instead, for the supervised analyses, each dataset was split at random into the training set (80% of data) and the test set (20%). Optimal hyperparameters' values (number of trees in RF, cost in SVM, and γ for RBF-like kernels) and β coefficients for MKL were obtained by 5×5 cross-validation on the training set. Hyperparameters' ranges are in **Supplementary Table 1**. Once the best values were found, the final model was built using the whole training set. We repeated the whole process 40 times, each time with different 80/20 randomly split training/test partitions, to obtain an error distribution. Performance over the test set was computed using normalized mean squared error (NMSE) for regression and Accuracy for classification. We measured with the *microbenchmark* package the running time of computing the SVM models on a 64-bit Ubuntu 20.04 LTS workstation with Intel(R) Core(TM) i5-6300U CPU at 2.40 GHz and 12 GiB of RAM (see **Supplementary Figure 1**). For the sake of comparison, the running time of several RF implementations (including the *randomForest* package) can be found at Wright and Ziegler (2017).

For the Smokers and Pig case studies, additional considerations had to be taken into account. In the Smokers dataset, in addition to the kPCA analysis, we computed the similarity among kernel matrices of different body sites with the *mixKernel* package (Mariette et al., 2018). We compared the performance of data integration via MKL (the kernel approach) with that of RF when using early and late integration approaches. In the former case, the input of the RF was the concatenated data of the four sites. Instead, in the latter case we used the forests created for each site separately to vote for the final decision (Li et al., 2018).

In the Pig dataset, to make sure that the training and test sets were completely independent, piglets from the same litter (full sibs) were always placed either in one or other set. Performance of fLin and fRBF was contrasted to those of RF and their analogous non-longitudinal kernels (cLin and cRBF) using all available days

at once. For the non-longitudinal methods, 80% of the piglets were used to train the model, using their three time points data in separate rows, with time included as an additional variable. The remaining piglets were reserved to test the model, but using only one of their time points (either day 0, 3, or 7) chosen at random and discarding the rest. This way, both longitudinal and non-longitudinal approaches had the same test set size. Longitudinal kernels fLin and fRBF were computed using (Eq. 9) and (Eq. 8), as only three time points were available and we preferred not to interpolate the day's in-between. Also, using the expression for discrete functions we could obtain the microbial signatures. The information of all taxa was combined as in (Eqs 11, 10) and the training/test partitions were carried out as in the normal case. In a second step, the dataset was decomposed by sampling times and the analysis was carried out for days 0, 3, and 7 separately using RF, cLin and cRBF in the usual way.

Microbial signatures from SVMs were obtained from the hyperplane normal vector w. The importance of taxon k is computed by *kernInt* as $(w_k)^2$ (Guyon et al., 2002). When using RF, we used the mean decrease in node impurity (for regression tasks) and mean decrease in Gini index (for classification). Both RF and SVM give absolute values of taxa importance, so they were converted to relative values. We used the R package *MiRKAT* (Zhan et al., 2017) to test if the association of the target phenotype with the signatures we obtained was statistically significant.

RESULTS

Soil Data

The cLin kPCA over the bacterial abundances is shown in **Figure 2A**. The remaining kPCAs, which gave a similar profile, can be found at **Supplementary Figure 2**. Soil samples are clearly separated by their pH, in agreement with the original results. The U-shaped projection is typical of data structured by a gradual transition with few overlapping OTUs at the endpoints (**Supplementary Figure 3**). The peak diversity in near-neutral soils in contrast with extreme pHs may also have some effect (**Supplementary Figure 4**). In addition, we used SVMs with the four kernels described above to predict the pH of each soil site from the bacterial abundances. This was not done in the original work and so we used RF, a non-kernel, alternative method, as benchmark. Results are shown in **Figure 2B**. The best compositional kernel was cLin, having a median error of ∼0.09; and the best non-compositional one was JSK, with a median error of ∼0.10. In comparison, RF had a higher median error, almost the double of cLin, around 0.17.

To go further in the interpretation of the results, we analyzed the microbial signature retrieved from RF and cLin-SVM. The distribution of the importances was highly skewed. For subsequent analyses we kept only 5% of the taxa, which accounted for around the 90% (RF) and 95% (SVM) of total importance, with the two methods having 42% of OTUs in common. Top ten relevant taxa are shown in **Figure 2C** (RF) and **Figure 2D** (SVM). In agreement with the kPCA results, prediction is primarily driven by few OTUs of extreme pH ecosystems (e.g., genera *Rubrobacter* and *Balneimonas* on the basic side, orders *Solibacterales* and *RB41* on the acid side). We used *MiRKAT* to test the significance of the association of the pH with both the top ten and 5% most important taxa, according to the cLin kernel. In both cases, we obtained very low p-values ($<10^{-8}$).

Smokers Data

We predicted smoking status from the taxonomic abundances. At first models were built using the four sites separately, as in the original study. Authors used RF and reported a median accuracy of 64% on the right and 65% on the left oropharynx (i.e., throat), and 71% on the right and 68% on the left nasopharynx. We re-computed the RF accuracies with our data pre-processing, and obtained very similar results (**Figure 3A**), with the only exception of the right nasopharynx (new median accuracy: 66%). Regarding the kernels, the worst one was cLin (**Supplementary Figure 5**), which nonetheless gave similar accuracies to RF. The best kernel was the Jaccard kernel (**Figure 3A**), which improved substantially the RF accuracies, especially in the throat. Then, we combined the spatial-structured samples of the same individuals to test if accuracy increased when using an integrative approach (**Figure 3A**). For the kernels, we first used MKL to combine the kernel matrices at the airway level (nasopharynx on one hand and oropharynx on the other) and, finally, we integrated all sites. This decreased the error substantially and delivered the best classification result, with a median accuracy of 92%. As for the RF, we tested both the early integration approach and the late integration approach, and found that the latter granted better predictions. At best, integration of the four sites delivered a median accuracy of 83%. The results for the rest of kernels can be found in **Supplementary Figure 5**. In all cases, integration of the four datasets using our MKL proposal increased the accuracy in comparison to the individual models, and doing so gave better or equivalent results that those of RF integration approaches. The only exceptions to this trend were the nasopharynx and oropharynx models delivered by cLin (but not the model with the four sites combined).

Next, we recovered the overall microbial signature (i.e., across the four sampling sites). The importance distribution is not as skewed as in the Soil dataset: here the top 5% taxa accounted for the 62% of overall importance. The association of this subset of taxa with the target phenotype was highly significant (p-value $< 10^{-8}$). Top ten taxa are shown in **Figure 3B**. *Neisseria* sp. large impact in discriminating smokers from non-smokers was already reported in the original work, especially in oropharynx models. The rest of highlighted taxa in **Figure 3B** were also noted to have a role, either in models from nasopharynx alone or from both airways sites (Charlson et al., 2010). This mostly agrees with our results when the sampling sites are analyzed separately (**Supplementary Figure 6**).

Following the original work, differences in bacterial communities among the body sites were also analyzed. We present results for the Jaccard kernel in **Figures 3C,D**, while the rest are in **Supplementary Figure 7**. **Figure 3C** shows the similarity across kernel matrices derived from left and right nasopharynx and oropharynx. The highest similarity was achieved within matrices of the same airway site but

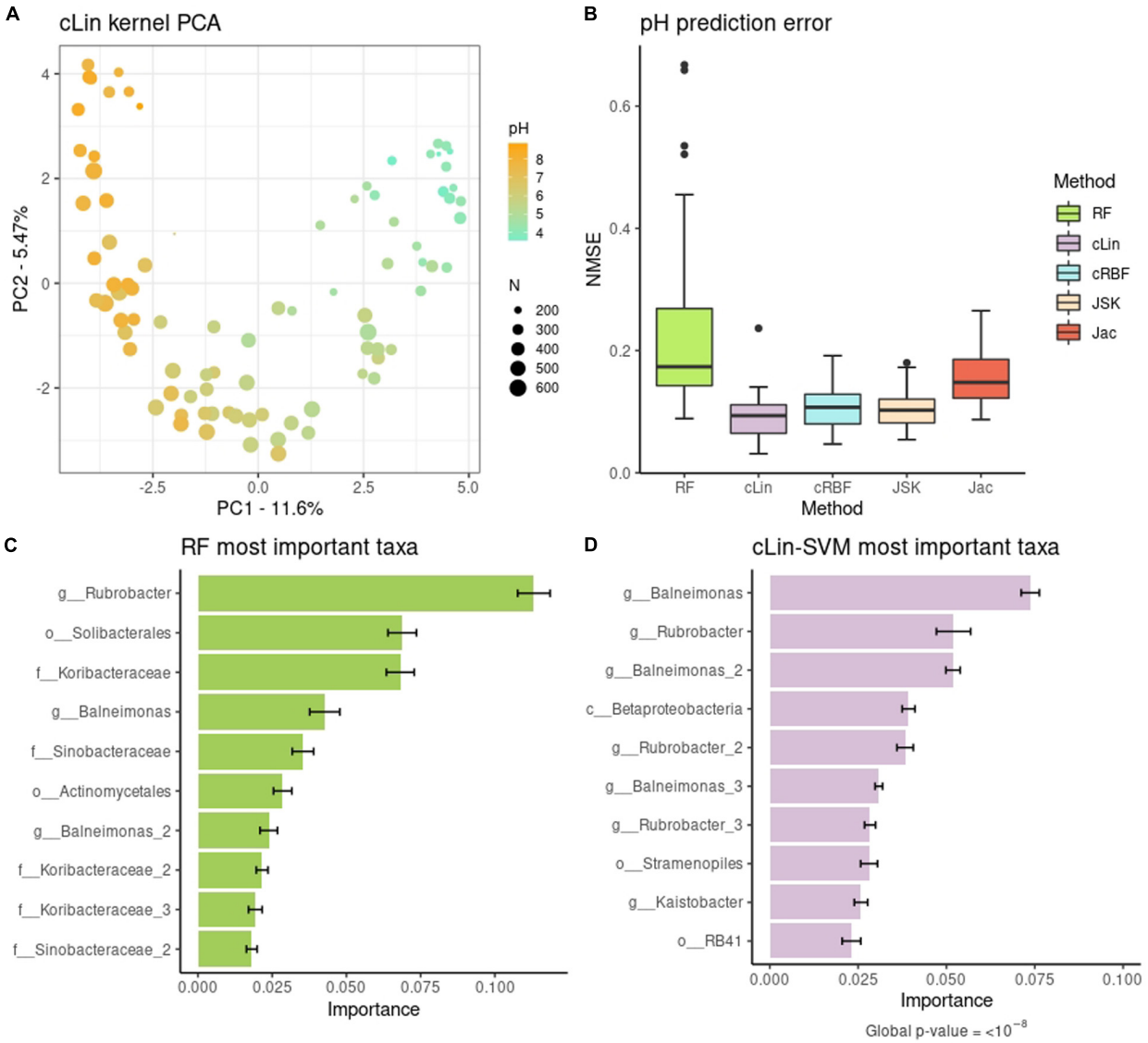


FIGURE 2 | (A) Compositional linear kPCA over the 88 soils. Color represents pH, while point size stands for the number of different observed taxa. **(B)** pH prediction error distribution over the 40 replicates. **(C,D)** Top relevant taxa for pH prediction according to RF and cLin-SVM. Standard error across the 40 replicates is marked with error bars.

different laterality. As in the original paper (**Supplementary Figure 8**), using a kPCA (**Figure 3D**) we could discriminate between nasopharynx and oropharynx sites (first PC) but not between left and right.

Pig Data

Evolution of gut microbiota from 153 healthy piglets over their first week of life was used to predict the occurrence of pre-weaning diarrhea. In **Figure 4A** we compared the performance of the longitudinal kernels (fLin and fRBF) vs. their analogous non-longitudinal kernels (cLin and cRBF) plus RF when using all available days at once. The longitudinal approach clearly outperformed the non-longitudinal approach at both Genera and ASVs levels. fRBF had a better performance than fLin, and worked best at the ASV level (with a median accuracy around 76%) than in Genera data (median accuracy ∼70%). Although aggregating taxa to the genus level is a relatively common practice –see e.g., Rivera-Pinto et al. (2018)–, in our case using a coarser taxonomic resolution decreased the accuracy. Within the non-longitudinal approach, we obtained similar accuracies using RF and kernels, and both were close to the median accuracy of the random model (50.1%). To further understand the results, the analysis was carried out in days 0, 3, and 7 separately using RF, cLin and cRBF kernels. **Figure 4B** reveals that all models from days 0 and 3 had no predictive power. Accuracy increased dramatically after day 7 to a maximum of 73% for cRBF (ASV level), only slightly worse than its analogous longitudinal kernel fRBF. We used the kernel machine test of *MiRKAT* to further

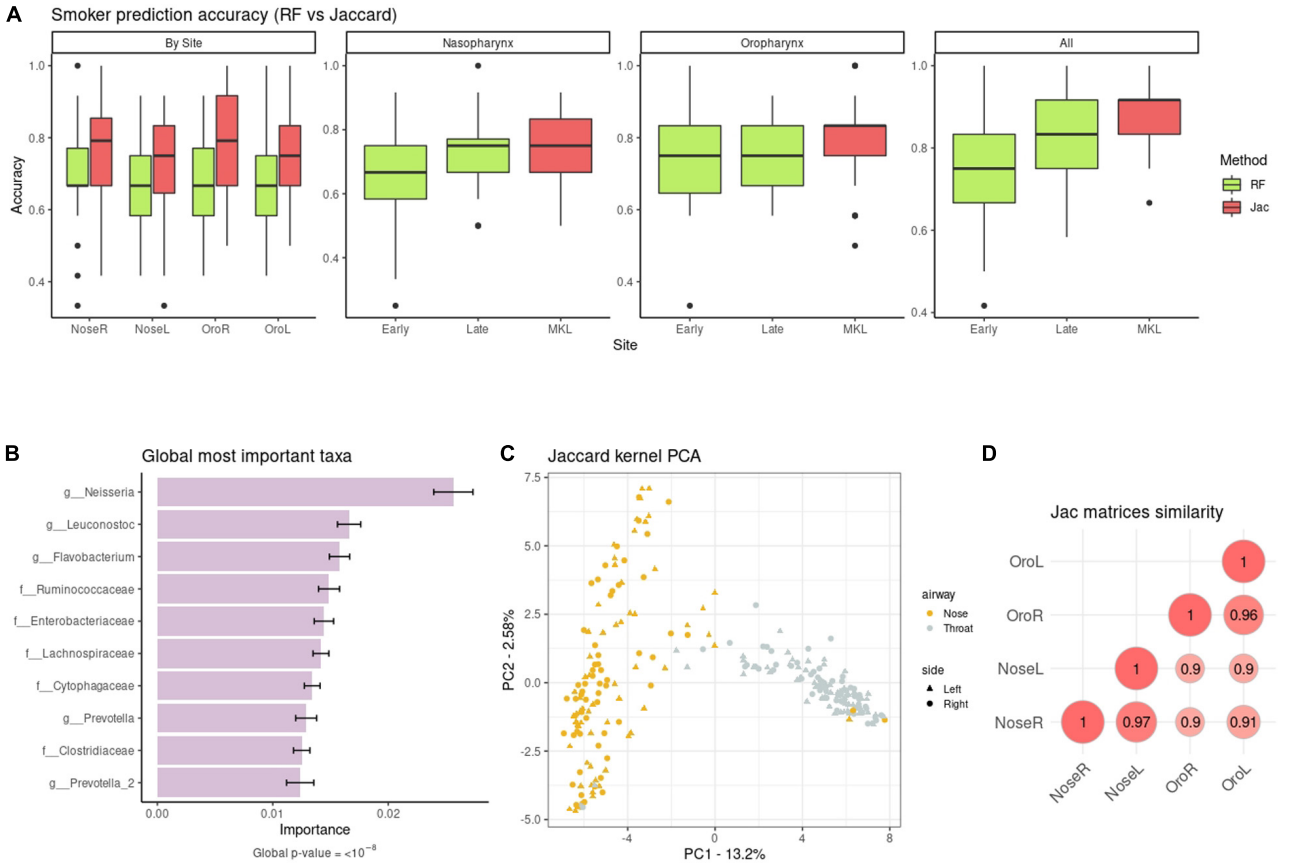


FIGURE 3 | (A) Non-smoker/smoker accuracies from taxonomic data: RF (green) vs. the quantitative Jaccard kernel (red). NoseL, NoseR, OroL, and OroR models are obtained from single datasets, the Nasopharynx and Oropharynx panels contain the information from the right and left sides, and All the combination of the four datasets. Early and Late refer to the early integration and late integration approaches for RF. **(B)** Top ten global cLin-SVM importances across all body sites. **(C)** Similarity across the Jaccard kernel matrices of the four sites (Nasopharynx Right and Left and Oropharynx Right and Left). **(D)** Jaccard kPCA of the taxonomic abundances. Color code represent airway site, whereas shape indicates the laterality of the samples.

confirm that days 0 and 3 were not significantly associated with phenotype, while day 7 was. As expected, only the kernel matrices of day 7 delivered significant p-values (Genera: cLin p-value $< 10^{-6}$, RBF p-value $< 10^{-7}$; ASV: cLin and cRBF p-values $< 10^{-8}$) after Bonferroni correction.

In a second step we analyzed the kPCA and microbial signatures, after discarding all models without predictive power. **Figures 5A,B** show the fLin and cLin (day 7) kPCA, while fRBF and cRBF are in **Supplementary Figure 9**. In all cases a partial separation between healthy and sick piglets, with a large area of overlap, is observed. Genera relevance on prediction of pre-weaning diarrhea is shown in **Figures 5C–E**. We discuss the microbial signature at the Genera level, as around 2/3 of the ASVs lack species assignation (**Supplementary Figure 10**). According to fLin, beneficial genera like *Lactobacillus* and *Bacteroides* had the higher overall importance during the first week. In day 7, it was striking the great importance given by cLin to *Desulfovibrio*, and secondarily to *Streptococcus*. RF also highlighted the butyrate-producing genus *Dorea*. Distribution of the microbial signature at the ASV level was skewed, but again, much less than in the Soil case study. The top 5% ASVs accounted for a 46% (fLin) and 58% (cLin) of the total importance, with an overlap between RF and cLin in day 7 of 2/3 of the ASVs. The association of the 5% most important taxa (global and day 7) with the phenotype was statistically significant according to MiRKAT (p-values $< 10^{-8}$).

DISCUSSION

The kernel framework allows performing a great diversity of analyses in a common ground, while allowing a great flexibility on how data is approached. However, within the microbiome field, previous application of kernel methods has been mostly restricted to specific areas. Zhan et al. (2017) proposed a kernel-based semi-parametric regression method for testing the association of the human microbiota communities with multiple phenotypes. Their method was implemented in the R package *MiRKAT*. In turn, Mariette et al. (2018) combined metagenomic data and environmental measures of the TARA ocean expedition using unsupervised MKL with the *mixKernel* package. In some reports that compare the performance of different supervised

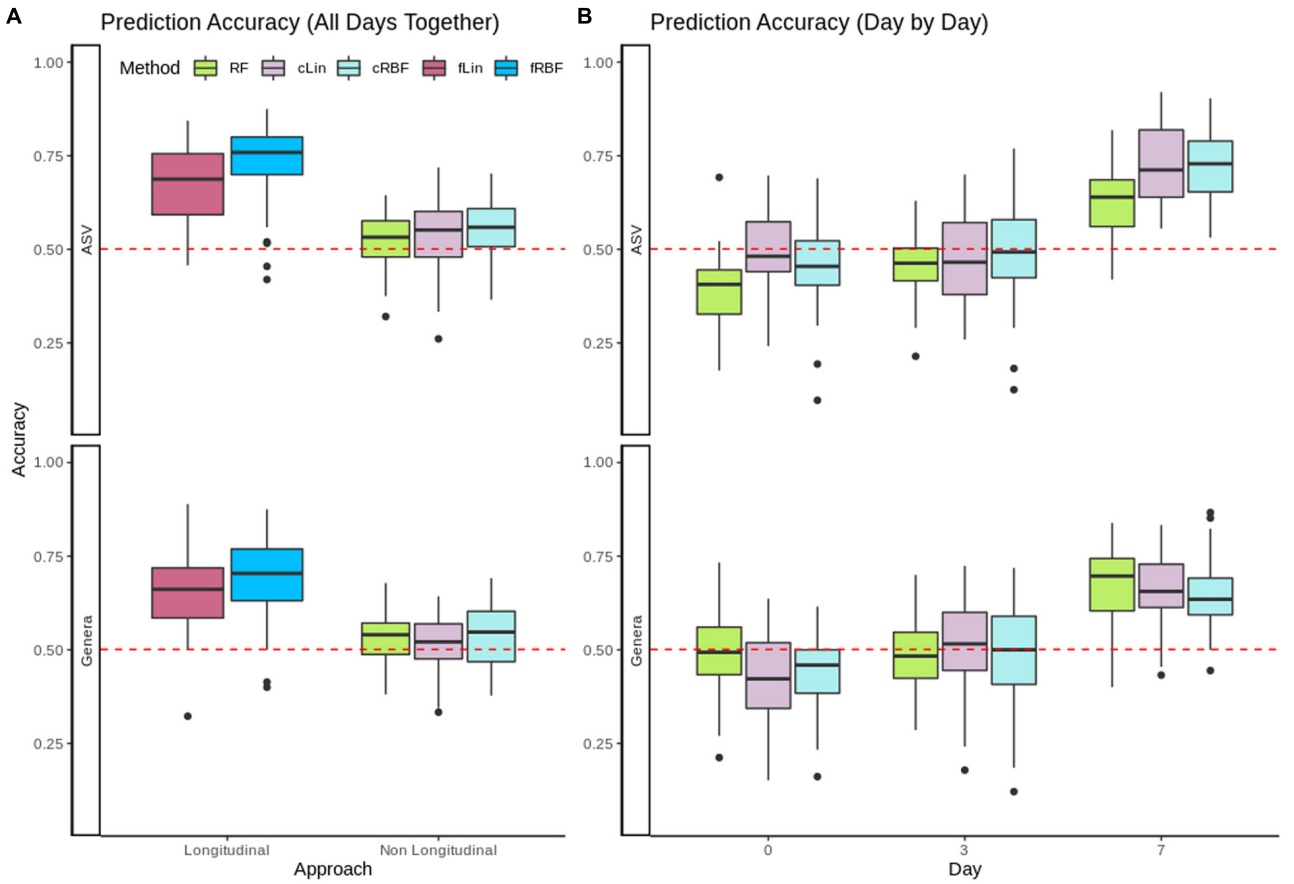


FIGURE 4 | (A) Accuracy for RF, non-longitudinal kernels (cRBF, cLin) and Longitudinal kernels (fLin, fRBF) in the prediction of neonatal diarrhea from all available days. **(B)** Accuracy for RF and non-longitudinal cLin and cRBF kernels from metagenomic data of days 0, 3, and 7 post-birth separately. In both panels, the red dashed line marks the accuracy of the random model.

methods in microbiome data, SVM often appear along RF or ANN (Qu et al., 2019; Zhou and Gallins, 2019; Namkung, 2020). Thus, kernel methods were mostly used in an isolated way, without exploiting the kernel framework ability to integrate a great range of analyses while giving a unitary view. Another advantage of this framework is that it can handle virtually any data type. However, to our best knowledge, it has not been previously applied to longitudinal microbiome studies. Finally, in previous works there was a lack of kernels that took into account the compositional nature of metagenomic datasets. Here we addressed all these questions, while also providing some examples of how previous kernel-based tools like *MiRKAT* and *mixKernel* can fit into our framework.

When comparing *kernInt* to a popular package for microbiome analysis like QIIME2, it becomes apparent that the former is more specific in its scope. *kernInt* is not concerned with sequence alignment, taxonomic assignation and quality control as QIIME2 is, but with the analysis once the abundance table is obtained. Both packages are aimed at community ecology analysis (in QIIME2: alpha and beta diversities, PCoA, etc.) and supervised learning areas. While *kernInt* does not have the great range of methods available in QIIME2, it improves the current state-of-the-art in the following points: (i) Proposal and implementation of specific kernels for microbiota, while QIIME2 currently provides default kernels for real vectors (the linear, RBF, polynomial and sigmoid kernels). (ii) As far as we know, SVM is available in QIIME2 but kPCA is not; therefore, it is not possible performing both supervised and unsupervised analysis under the same mathematical point of view (**Figure 1**). (iii) Integration of spatial and temporal samples: QIIME2 does not have a specific handling of spatial (and, potentially, multi-omic) data, while *kernInt* allows performing unsupervised, supervised and retrieval of microbial signatures in this kind of datasets. On the other hand, the QIIME2 "longitudinal" plugin implements several tools for longitudinal data, but the option of performing supervised learning from the variation of microbiota over time is absent.

Throughout this work, we summarized the microbiome analyses in three branches: unsupervised learning (represented by kPCA), supervised learning (SVM) and identification of phenotype-associated microbial signatures. The Soil case study clearly illustrated how all three types are intertwined and complementary. In agreement with the original publication, both SVM and kPCA results showed that taxonomic abundances and

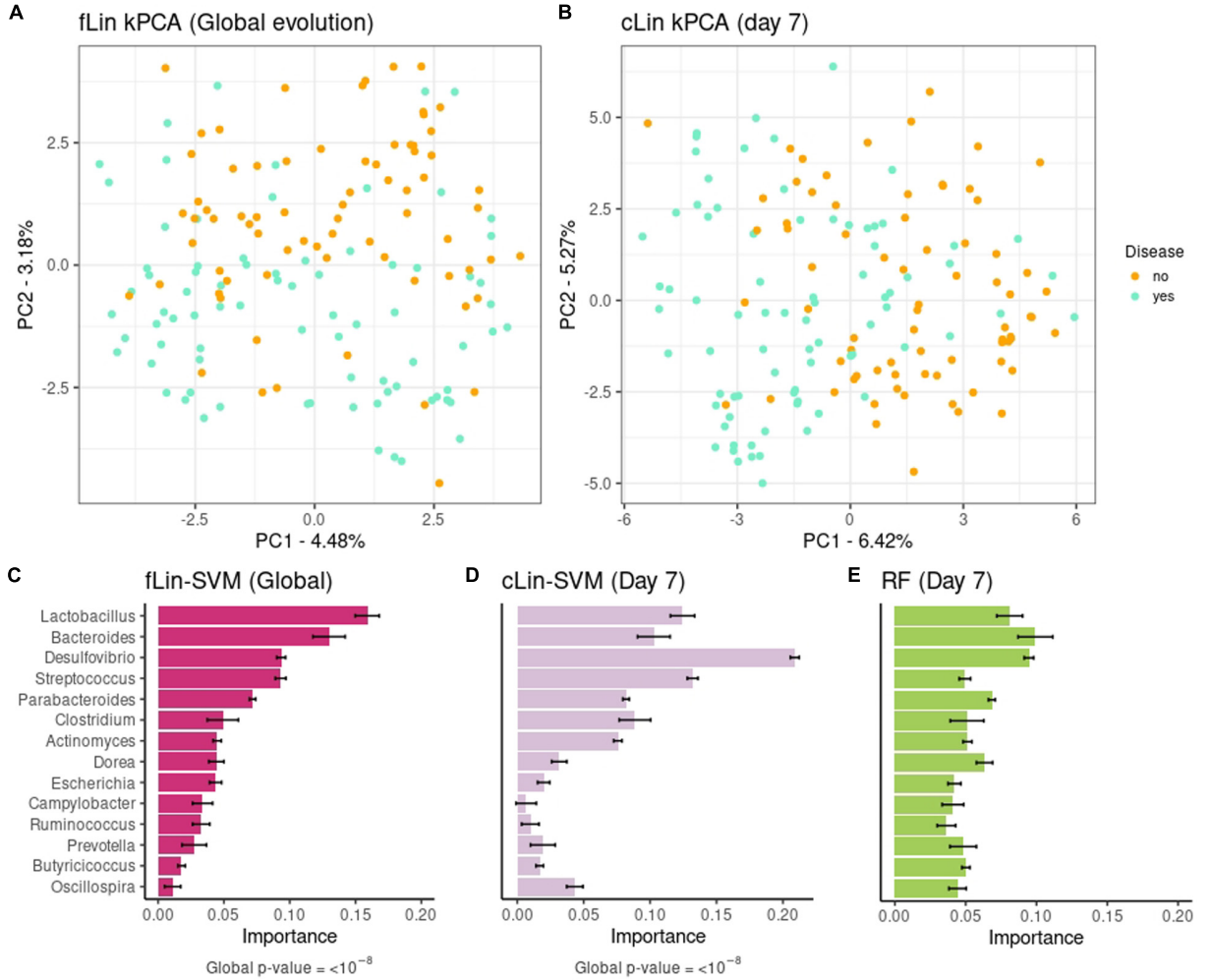


FIGURE 5 | Above: kPCA of fLin in panel **(A)** and of cLin (day 7) in panel **(B)**. Below: Microbial signatures at the Genera level. Global importance for the first week is in panel **(C)**. Importances for the day 7 according to the cLin kernel and RF are in panels **(D,E)**. Standard error across the 40 replicates is marked with error bars.

pH are strongly related. This granted a quite low prediction error (up to a median NMSE of 0.09) but, by itself, does not explain the underlying mechanism connecting microbial abundance and pH. Microbial signature revealed that the SVM learning is driven by few taxa of opposite pH ecosystems. For instance, *RB41* belong to the phylum *Acidobacteria*. The *Rubrobacter* genus contains well known extremophiles and, like the *Balneimonas* (renamed *Microvirga*) genus, has preference for clearly alkaline soils (Dahal and Kim, 2017; Chen et al., 2018). Furthermore, the arch in the kPCA projection indicated that communities from acid and basic habitats did not overlap (Morton et al., 2017). Taken together, these complementary views point that soil microbial structure is shaped by a gradual niche differentiation strongly modulated by the pH. This agrees with previous findings on this dataset (Lauber et al., 2009; Morton et al., 2017) but appears in a more concise and unified way using the kernel framework.

In comparison to other methods, the kernel framework did not only allow a holistic view of data, but also gave good results in each learning area. Concerning supervised learning, in general, the kernel methods tend to have an advantage over variable-oriented methods (e.g., in supervised learning: ridge regression, decision trees, RF, etc.) and over ANN (for the reasons stated in the "Introduction" section) when faced with $N << D$ data. This is a common scenario in metagenomics when working at the OTU or ASV level, but not necessarily in coarser taxonomic resolutions. This is illustrated with the different behavior of kernels with respect to RF in **Figure 4B** (see ASV vs. Genera results). In the other cases, SVM were consistently better (or at least equivalent) to RF in all the case studies that we analyzed. This disagrees with some previous reports in the microbiome area, e.g., Zhou and Gallins (2019). However, it should be noted that SVM performance depends on the kernel used, and these reports used generic linear and RBF kernels. Even when using kernels specific for metagenomic data, we observed differences among their mean NMSE or accuracies as large as fifteen percentage points. At the same time, our results suggest that there is not a single kernel that systematically achieves the best performance in every problem. We found that cLin was the best

one in the first case study, quantitative Jaccard in the second and fRBF in the third. In this scenario, we consider that the linear-like kernels like cLin are a safe starting point. They allow for the retrieval of the microbial signatures, are faster to compute and easier to interpret than non-linear kernels, and with high-dimensional data ($>10^3$–10^4 taxa) they tend to match the RBF kernel (usually considered the gold standard) in performance (Hsu et al., 2003; Keerthi and Lin, 2003). RBF may be useful if the number of different taxa is low, or when a strong non-linear relationship is suspected. The weakness of the compositional kernels that we proposed is that they cannot handle zeroes without pre-processing; instead, zeroes pose no problems to the quantitative Jaccard kernel. How to deal with zeroes is, currently, an open topic of research in compositional analysis (Weiss et al., 2017; Quinn et al., 2018). If there is not enough *a priori* information that permits selecting the kernel beforehand, visual assessment of the candidate kernel matrices via kPCA could be of some help. A more rigorous approach is to perform nested cross-validation (Cawley and Talbot, 2010) to avoid overfitting when selecting both the candidate hyperparameters' values and the best kernel for a given problem. Finally, phylogenetic kernels were beyond the scope of this work, nor the available datasets had the phylogenetic trees needed to compute them. However, they may be derived from (Eqs 1, 2) by replacing the clr term with other transformations, e.g., the PhILR transformation (Silverman et al., 2017). A phylogeny-based kernel was also proposed in Xiao and Chen (2017).

Concerning our unsupervised analyses, we observed that the main structure revealed by the original MDS/PCoA (ordination by pH in the Soil dataset, and by body site in the Smoker dataset) was conserved in our kPCAs. On the other hand, microbial signatures obtained with SVM had a biological interpretation. In general, the most important taxa retrieved from SVM coincided with those of RF (40–65% of overlap depending on the dataset), and could be recovered too when dealing with spatial and temporal-structured datasets. However, we acknowledge that a drawback of these signatures (though they handle well the cases of multicollinearity) is that they are based on linear kernels. In turn, RF can take into account both non-linearity and complex interactions among taxa. In any case, the informativeness of a microbial signature can be assessed by the prediction performance of the SVM model that generated it.

Apart of the aforementioned advantages of the kernel framework, we also showed how it can accommodate datasets with spatial and/or temporal components. We illustrated the integration of spatial-structured samples with the Smokers dataset. The analysis in the original work was carried out in each sampling site independently, with a maximum median accuracy of 71%. Here we showed how combining the body sites using MKL increased the median accuracy to 92%. Therefore, our results remark the relevance of using an integrative approach to improve the accuracy of phenotype prediction when spatial-structured samples of the same individuals are available.

In addition to the package and framework proposal, we analyzed a previously unpublished dataset profiling the microbiota evolution and pre-weaning diarrhea incidence in 153 piglets. Through this dataset we illustrated the kernel framework application to time-structured samples. Pre-weaning diarrhea is an important issue in pig production, as the antibiotic treatment increases both the emergence of resistances and the economic costs. It is already known that gut colonization starts immediately after birth, and it evolves from a highly variable to a more stable and homogeneous ecosystem over the first weeks. However, most of the current studies in pig production ignore early dynamics in gut microbiota (Mach et al., 2015; Han et al., 2018; Massacci et al., 2020). We wanted to test if pre-weaning diarrhea could be anticipated as soon as the first week of life. In this sense, our results suggest that the first stages of intestinal microbiota convey some valuable information indeed. kPCAs showed a partial separation between piglets affected of diarrhea vs. healthy piglets, and by using longitudinal kernels we achieved a moderate accuracy of 76%. However, it was unclear if this accuracy was to be attributed to a different taxa evolution in the two groups over the first week, or to a single time point with a great predictive value. The day-by-day prediction clarified this issue, and showed that day 7 achieved a median accuracy of 73% while the rest of points lacked predictive power. Even so, longitudinal kernels were able to slightly improve prediction (76% vs. 73% at the ASV level, and 69% vs. 64% using Genera), so global taxa evolution may also have a small role.

This is also seen in the underlying microbial signatures of the global first week (fLin) vs. day 7 (cLin). To be noted, in day 7 the most important genus was sulfate-reducing bacteria *Desulfovibrio*, which is known to have a relevant role during pig gut colonization (Mach et al., 2015). Instead, the global (longitudinal) model was mainly led by *Lactobacillus* and *Bacteroides*. Relationship of both genera to pre-weaning diarrhea is well sustained in literature. *Lactobacillus* spp. are well known probiotic bacteria, while members of *Bacteroides* genus are associated with increased infants gut microbial diversity (Stewart et al., 2018). Furthermore, both play an important role on mammals' gut microbial colonization (Sawicki et al., 2017; Wexler and Goodman, 2017) and are dominant in healthy pigs compared with diarrhea-affected piglets (Song et al., 2017), which gives confidence in the reliability of our findings.

In summary, our kernel framework successfully places the most important analyses in the microbiome field on a common ground, takes into account the compositionality of data, and is flexible enough to integrate spatial and temporal dimensions of the datasets.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: This manuscript utilizes proprietary data. Requests to access these datasets should be directed to YR-C/IRTA/yuliaxis.ramayo@irta.cat.

ETHICS STATEMENT

The animal study was reviewed and approved by the Central Authority for Scientific Procedures on Animals of Netherlands–Centrale Commissie Dierproeven (CCD).

AUTHOR CONTRIBUTIONS

YR-C, MP-E, RQ, and ER contributed to conception and design of the study. FM was in charge of the pig data sampling. YR-C and MP-E supervised the overall research, while LB-M supervised the machine learning part. ER performed the all analysis and wrote the first draft of the manuscript. YR-C, MP-E, and LB-M revised and wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This work was funded by projects PID2019-108829RB-I00, AGL2016–78709-R, and AGL2017–88849-R awarded by the Spanish Ministry of Economy and Competitiveness. ER has funding from a FI-AGAUR Ph.D. studentship grant, with the support of the Secretaria d'Universitats i Recerca de la Generalitat de Catalunya and the European Social Fund. YR-C is recipient of a Ramon y Cajal post-doctoral fellowship (RYC2019-027244-I) from the Spanish Ministry of Science and Innovation. We acknowledge further financial support from the Spanish Ministry of Economy and Competitiveness through the "Severo Ochoa Programme for Centres of Excellence in R&D" 2016–2019 (SEV-2015-0533)," and from the EU through the BFU2016–77236-P (MINECO/AEI/FEDER, EU).

ACKNOWLEDGMENTS

The authors warmly thank all technical staff from Schothorst Feed Research.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2021.609048/full#supplementary-material

REFERENCES

Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., et al. (2018). Multi-Omics Factor Analysis–a framework for unsupervised integration of multi-omics data sets. *Mole. Syst. Biol.* 14:e8124.

Bai, L., and Hancock, E. R. (2011). *Graph clustering using the jensen-shannon kernel. In International Conference on Computer Analysis of Images and Patterns*, Berlin: Springer, 2011, 394–401.

Berg, G., Rybakova, D., Fischer, D., Cernava, T., Vergčs, M. C. C., Charles, T., et al. (2020). Microbiome definition re-visited: old concepts and new challenges. *Microbiome* 8, 1–23.

Bodein, A., Chapleur, O., Droit, A., and Lê Cao, K. A. (2019). A generic multivariate framework for the integration of microbiome longitudinal studies with other data types. *Front. Genet.* 10:963. doi: 10.3389/fgene.2019.00963

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, et al. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* 37, 852–857.

Bouchard, M., Jousselme, A. L., and Doré, P. E. (2013). A proof for the positive definiteness of the Jaccard index matrix. *Int. J. Approx. Reas.* 54, 615–626. doi: 10.1016/j.ijar.2013.01.006

Cawley, G. C., and Talbot, N. L. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *J. Mach. Learn. Res.* 11, 2079–2107.

Charlson, E. S., Chen, J., Custers-Allen, R., Bittinger, K., Li, H., Sinha, R., et al. (2010). Disordered microbial communities in the upper respiratory tract of cigarette smokers. *PLoS one* 5:e15216. doi: 10.1371/journal.pone.0015216

Chen, R. W., Wang, K. X., Wang, F. Z., He, Y. Q., Long, L. J., and Tian, X. P. (2018). Rubrobacter indicoceani sp. nov., a new marine actinobacterium isolated from Indian Ocean sediment. *Int. J. Systemat. Evolut. Microbiol.* 68, 3487–3493. doi: 10.1099/ijsem.0.003018

Chen, H., Tang, F., Tino, P., and Yao, X. (2013). "Model-based kernel for efficient time series analysis," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, eds J. He, Y. Koren, R. Ghani, R. Parekh, I. Dhillon, and T. E. Senator (United States: Association for Computing Machinery), 392–400.

Coenen, A. R., Hu, S. K., Luo, E., Muratore, D., and Weitz, J. S. (2020). A Primer for Microbiome Time-Series Analysis. *Front. Genet.* 11:310. doi: 10.3389/fgene.2020.00310

Dahal, R. H., and Kim, J. (2017). Microvirga soli sp. nov., an alphaproteobacterium isolated from soil. *Int. J. Syst. Evolut. Microbiol.* 67, 127–133. doi: 10.1099/ijsem.0.001582

Gardener, M. (2014). *Community ecology: analytical methods using R and Excel*. United Kingdom: Pelagic Publishing Ltd.

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., and Egozcue, J. J. (2017). Microbiome Datasets Are Compositional: And This Is Not Optional. *Front. Microbiol.* 8:2224. doi: 10.3389/fmicb.2017.02224

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Mach. Learn.* 46, 389–423.

Han, G. G., Lee, J. Y., Jin, G. D., Park, J., Choi, Y. H., Kang, S. K., et al. (2018). Tracing of the fecal microbiota of commercial pigs at five growth stages from birth to shipment. *Scient. Rep.* 8, 1–9.

Hsu, C. W., Chang, C. C., and Lin, C. J. (2003). *A practical guide to support vector classification*. Available online at: https://www.csie.ntu.edu.tw/~{}cjlin/papers/guide/guide.pdf [accessed on June 12, 2020]

Ibrahim, O. M. (2013). A comparison of methods for assessing the relative importance of input variables in artificial neural networks. *J. Appl. Sci. Res.* 9, 5692–5700.

Keerthi, S. S., and Lin, C. J. (2003). Asymptotic behaviors of support vector machines with Gaussian kernel. *Neural. Comput.* 15, 1667–1689. doi: 10.1162/089976603321891855

Lauber, C. L., Hamady, M., Knight, R., and Fierer, N. (2009). Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. *Appl. Environ. Microbiol.* 75, 5111–5120. doi: 10.1128/AEM.00335-09

Li, Y., Wu, F. X., and Ngom, A. (2018). A review on machine learning principles for multi-view biological data integration. *Brief. Bioinform.* 19, 325–340.

Liaw, A., and Wiener, M. (2002). randomForest: Breiman and Cutler's random forests for classification and regression. *R Package Vers.* 4, 6–10.

Mach, N., Berri, M., Estellé, J., Levenez, F., Lemonnier, G., Denis, C., et al. (2015). Early-life establishment of the swine gut microbiome and impact on host phenotypes. *Environ. Microbiol. Rep.* 7, 554–569. doi: 10.1111/1758-2229.12285

Mariette, J., Villa-Vialaneix, N. (2018). Unsupervised multiple kernel learning for heterogeneous data integration. *Bioinformatics* 34, 1009–1015. doi: 10.1093/bioinformatics/btx682

Massacci, F. R., Berri, M., Lemonnier, G., Guettier, E., Blanc, F., Jardet, D., et al. (2020). Late weaning is associated with increased microbial diversity and Faecalibacterium prausnitzii abundance in the fecal microbiota of piglets. *Anim. Microb.* 2, 1–13.

McMurdie, P. J., and Holmes, S. (2014). Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput. Biol.* 10:e1003531. doi: 10.1371/journal.pcbi.1003531

Morton, J. T., Toran, L., Edlund, A., Metcalf, J. L., Lauber, C., and Knight, R. (2017). Uncovering the horseshoe effect in microbial analyses. *Msystems* 2, 166–e116.

Namkung, J. (2020). Machine learning methods for microbiome studies. *J. Microbiol.* 58, 206–216. doi: 10.1007/s12275-020-0066-8

Paulson, J. N., Stine, O. C., Bravo, H. C., and Pop, M. (2013). Differential abundance analysis for microbial marker-gene surveys. *Nat. Methods* 10, 1200–1203. doi: 10.1038/nmeth.2658

Qu, K., Guo, F., Liu, X., Lin, Y., and Zou, Q. (2019). Application of machine learning in microbiology. *Front. Microbiol.* 10:827. doi: 10.3389/fmicb.2019.00827

Quinn, T. P., Erb, I., Richardson, M. F., and Crowley, T. M. (2018). Understanding sequencing data as compositions: an outlook and review. *Bioinformatics* 34, 2870–2878. doi: 10.1093/bioinformatics/bty175

Rivera-Pinto, J., Egozcue, J. J., Pawlowsky-Glahn, V., Paredes, R., Noguera-Julian, M., and Calle, M. L. (2018). Balances: a New Perspective for Microbiome Analysis. *mSystems* 3, 53–e18.

Sawicki, C. M., Livingston, K. A., Obin, M., Roberts, S. B., Chung, M., and McKeown, N. M. (2017). Dietary fiber and the human gut microbiota: application of evidence mapping methodology. *Nutrients* 9:125. doi: 10.3390/nu9020125

Schölkopf, B., Tsuda, K., and Vert, J. P. (2004). *Kernel methods in computational biology*. New York: MIT press.

Shawe-Taylor, J., and Cristianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge: Cambridge university press.

Shenhav, L., Thompson, M., Joseph, T. A., Briscoe, L., Furman, O., Bogumil, D., et al. (2019). FEAST: fast expectation-maximization for microbial source tracking. *Nat. Methods* 16:627. doi: 10.1038/s41592-019-0431-x

Silverman, J. D., Washburne, A. D., Mukherjee, S., and David, L. A. (2017). A phylogenetic transform enhances analysis of compositional microbiota data. *Elife* 6:e21887.

Song, D., Peng, Q., Chen, Y., Zhou, X., Zhang, F., Li, A., et al. (2017). Altered gut microbiota profiles in sows and neonatal piglets associated with porcine epidemic diarrhea virus infection. *Scient. Rep.* 7, 1–10.

Stewart, C. J., Ajami, N. J., O'Brien, J. L., Hutchinson, D. S., Smith, D. P., Wong, M. C., et al. (2018). Temporal development of the gut microbiome in early childhood from the TEDDY study. *Nature* 562, 583–588. doi: 10.1038/s41586-018-0617-x

Su, X., Jing, G., Sun, Z., Liu, L., Xu, Z., McDonald, D., et al. (2020). Multiple-Disease Detection and Classification across Cohorts via Microbiome Search. *Msystems* 5, 150–e120.

Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., et al. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* 5:27.

Wexler, A. G., and Goodman, A. L. (2017). An insider's perspective: *Bacteroides* as a window into the microbiome. *Nat. Microbiol.* 2, 1–11.

Wright, M. N., and Ziegler, A. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *J. Statist. Softw.* 77, 1–17.

Xiao, J., and Chen, J. (2017). *Phylogeny-based kernels with application to microbiome association studies. In New Advances in Statistics and Data Science*. Cham: Springer, 217–237.

Zhan, X., Tong, X., Zhao, N., Maity, A., Wu, M. C., and Chen, J. (2017). A small-sample multivariate kernel machine test for microbiome association studies. *Genet. Epidemiol.* 41, 210–220. doi: 10.1002/gepi.22030

Zhou, Y. H., and Gallins, P. (2019). A review and tutorial of machine learning methods for microbiome host trait prediction. *Front. Genet* 10:579. doi: 10.3389/fgene.2019.00579

Zingaretti, L. M., Renand, G., Morgavi, D. P., and Ramayo-Caldas, Y. (2020). Link-HD: a versatile framework to explore and integrate heterogeneous microbial communities. *Bioinformatics* 36, 2298–2299. doi: 10.1093/bioinformatics/btz862

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 05 February 2021
doi: 10.3389/fmicb.2021.615711

VirionFinder: Identification of Complete and Partial Prokaryote Virus Virion Protein From Virome Data Using the Sequence and Biochemical Properties of Amino Acids

Zhencheng Fang[1,2] *and Hongwei Zhou*[1,3]*

[1] *Microbiome Medicine Center, Department of Laboratory Medicine, Zhujiang Hospital, Southern Medical University, Guangzhou, China,* [2] *Center for Quantitative Biology, Peking University, Beijing, China,* [3] *State Key Laboratory of Organ Failure Research, Southern Medical University, Guangzhou, China*

OPEN ACCESS

Edited by:
Isabel Moreno Indias, University of Málaga, Spain

Reviewed by:
Simon Roux, Joint Genome Institute, Lawrence Berkeley National Laboratory, United States
Felipe Hernandes Coutinho, Miguel Hernández University of Elche, Spain

***Correspondence:**
Hongwei Zhou
hzhou@smu.edu.cn; biodegradation@gmail.com

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

Received: *09 October 2020*
Accepted: *04 January 2021*
Published: *05 February 2021*

Citation:
Fang Z and Zhou H (2021) VirionFinder: Identification of Complete and Partial Prokaryote Virus Virion Protein From Virome Data Using the Sequence and Biochemical Properties of Amino Acids. Front. Microbiol. 12:615711. doi: 10.3389/fmicb.2021.615711

Viruses are some of the most abundant biological entities on Earth, and prokaryote virus are the dominant members of the viral community. Because of the diversity of prokaryote virus, functional annotation cannot be performed on a large number of genes from newly discovered prokaryote virus by searching the current database; therefore, the development of an alignment-free algorithm for functional annotation of prokaryote virus proteins is important to understand the viral community. The identification of prokaryote virus virion proteins (PVVPs) is a critical step for many viral analyses, such as species classification, phylogenetic analysis and the exploration of how prokaryote virus interact with their hosts. Although a series of PVVP prediction tools have been developed, the performance of these tools is still not satisfactory. Moreover, viral metagenomic data contains fragmented sequences, leading to the existence of some incomplete genes. Therefore, a tool that can identify partial PVVPs is also needed. In this work, we present a novel algorithm, called VirionFinder, to identify the complete and partial PVVPs from non-prokaryote virus virion proteins (non-PVVPs). VirionFinder uses the sequence and biochemical properties of 20 amino acids as the mathematical model to encode the protein sequences and uses a deep learning technique to identify whether a given protein is a PVVP. Compared with the state-of-the-art tools using artificial benchmark datasets, the results show that under the same specificity (*Sp*), the sensitivity (*Sn*) of VirionFinder is approximately 10–34% much higher than the *Sn* of these tools on both complete and partial proteins. When evaluating related tools using real virome data, the recognition rate of PVVP-like sequences of VirionFinder is also much higher than that of the other tools. We expect that VirionFinder will be a powerful tool for identifying novel virion proteins from both complete prokaryote virus genomes and viral metagenomic data. VirionFinder is freely available at https://github.com/zhenchengfang/VirionFinder.

Keywords: virome, metagenome, gene function annotation, deep learning, prokaryote virus virion protein

INTRODUCTION

Prokaryote virus are some of the most dominant biological entities in the viral community. Recently, a large number of experimental methods that enrich viral particles in the microbial community or computational methods that identify viral sequences in metagenomic data have been developed (Hayes et al., 2017; Khan Mirzaei et al., 2020; Martínez et al., 2020; Saak et al., 2020), leading to the discovery of a large number of novel prokaryote virus. The functional annotation of prokaryote virus genes is essential for understanding the composition and function of prokaryote virus in the microbial community. One of the most important tasks of functional annotation of prokaryote virus genes is the identification of prokaryote virus virion proteins (PVVPs) from non-prokaryote virus virion proteins (non-PVVPs). The PVVPs, which are also called structural proteins, are essential materials of the infectious viral particles, including shell proteins, envelope proteins, and viral particle enzymes (Feng et al., 2013). The identification of PVVPs plays an important role in understanding the interaction between a prokaryote virus and its host and can further help in developing antibacterial drugs (Lekunberri et al., 2017). Additionally, PVVPs are important for virus classification (Galiez et al., 2016), and it has been suggested that specific PVVPs can further serve as phylogenetic marker genes similar to 16S rDNA in bacteria (Seguritan et al., 2012) and therefore are important genes for viral phylogenetic analysis in the microbial community. Another important application of PVVPs is to identify prophages in bacterial chromosomes since the PVVP-enriched regions in bacterial chromosomes have a higher potential to be prophages (Roux et al., 2015). Although a series of experimental methods have been developed to identify PVVPs, such as protein array analysis, sodium dodecyl sulfate-polyacrylamide gel electrophoresis and mass spectrometry (Charoenkwan et al., 2020a), a fast and low-cost computational method is needed to accommodate the massive increase in sequencing data.

Computational methods based on similarity searches against known databases for PVVP identification are intuitive strategies, but such methods may not work well for viral metagenomic data. Because of its non-cultivable nature, the viral community contains a large number of novel prokaryote virus. It has been shown that many sequences in virome data are not present in the current database (Hayes et al., 2017). In addition, a large number of genes annotated on the prokaryote virus genomes of current database are predicted by related bioinformatics tools, such as GeneMark (Besemer and Borodovsky, 2005), and their function has not been subjected to experimental verification, indicating that the current knowledge of viral gene function is quite limited. Alignment-free algorithms, such as machine learning-based methods, bypass employing similarity search strategies and can identify novel PVVPs by universal features extracted from known data. Therefore, Alignment-free algorithms for PVVP identification may be better suited for virome studies. Recently, many alignment-free algorithms for such tasks have been developed, including iVIREONS (Seguritan et al., 2012), the algorithm developed by Feng et al. (2013), PVPred (Ding et al., 2014), the algorithm developed by Zhang et al. (2015), PVP-SVM (Manavalan et al., 2018), PhagePred (Pan et al., 2018), the algorithm developed by Tan et al. (2018), the algorithm developed by Ru et al. (2019), Pred-BVP-Unb (Arif et al., 2020), PVPred-SCM (Charoenkwan et al., 2020a) and Meta-iPVP (Charoenkwan et al., 2020b). To the best of our knowledge, among these algorithms, iVIREONS, PVPred, PVP-SVM, PVPred-SCM, and Meta-iPVP are currently available via web servers, while the other algorithms have not been released either via web servers (or the server was out of order) or one-click software packages. The biological support of these tools is that the amino acid composition between virion proteins and non-virion proteins is different. For example, it has been shown that the virion proteins contain more amino acids whose molecular weight is low (Ding et al., 2014). Based on this phenomenon, these tools constructed specific feature sets, such as the frequency of each amino acid on the protein, to characterize a given protein, and employed a shallow statistical model to distinguish the PVVP and non-PVVP according to the input feature sets. For example, the tool iVIREONS used the amino acid frequency as the feature sets and employed a shallow artificial neural network to classify the PVVP and non-PVVP (Seguritan et al., 2012); the tool PVPred used the g-gap dipeptide compositions as the feature sets and employed a support vector machine to classify the PVVP and non-PVVP (Ding et al., 2014); the tool PVP-SVM used the composition of amino acid, dipeptide and atom as well as the chain-transition-distribution and physicochemical properties as feature sets, and employed a support vector machine to classify the PVVP and non-PVVP (Manavalan et al., 2018); the tool PVPred-SCM used dipeptide composition as feature sets and employed a scoring card method to classify the PVVP and non-PVVP (Charoenkwan et al., 2020a); and the tool Meta-iPVP used the information of discriminative probabilistic features and employed a support vector machine to classify the PVVP and non-PVVP (Charoenkwan et al., 2020b). The performance of such methods relied heavily on the selected features (Ding et al., 2014). Since such features are constructed by the researcher empirically, the performance of these tools will be affected if inappropriate features are selected. In contrast, deep learning technique bypasses the process of artificial feature selection, and uses deep neural networks to extract useful features from the raw data automatically and therefore, deep learning may be more powerful in many bioinformatics tasks (Min et al., 2017). Thus, employing deep learning technique on the PVVP identification task may further improve the performance of the existing tools. Recently, a deep learning based method to identify specific virion proteins, namely capsid and tail, has been proposed (Abid and Zhang, 2018). Moreover, the existing tools are primarily designed for complete proteins while sequence assemblies of viral sequencing reads in metagenomic data are more difficult than chromosome-derived reads (Sutton et al., 2019; Martínez et al., 2020), indicating that virome data may contain fragmented sequences with some partial genes. Therefore, tools that can perform PVVP identification from partial genes are also needed.

In this work, we present VirionFinder. VirionFinder takes a sequence file containing all proteins from a single prokaryotic viral genome or viral metagenomic data in which viral sequences

are collected using experimental or computational method as input, and outputs a tabular file containing the judgment for each protein. Based on deep learning, VirionFinder can identify complete and partial PVVPs from virome data using the sequence and biochemical properties of amino acids. Evaluations showed that VirionFinder outperformed all the currently available tools.

MATERIALS AND METHODS

Dataset Construction

To create a benchmark dataset, we downloaded all the prokaryotic viruses from the RefSeq viral database ([1]downloaded in November 28, 2019). In addition to phage proteins, our dataset also contained proteins from archaeal viruses, which were also members of prokaryotic viruses. Dividing the data into training and testing sets according to the genome release day is a commonly used method to test an algorithm's ability to handle novel data (Zhou and Xu, 2010; Ren et al., 2017; Fang et al., 2019, 2020). To evaluate whether VirionFinder can identify a PVVP from a novel prokaryote virus, which is important for virology studies, we used the genomes released before 2018 to construct the training set, while the remaining genomes were used to construct the test set. According to the description from Seguritan et al. (2012), genes labeled one of the following key words "capsid," "tape measure," "portal," "tail," "fiber," "baseplate," "connector," "neck," and "collar" were extracted in the form of amino acid sequences to construct the PVVP set, while the remaining genes were used to construct the non-PVVP set. Genes labeled "hypothetical protein," "unnamed," "probable," "putative," or "similar to" were removed from the dataset as suggested by Seguritan et al. (2012). The accession lists of the PVVPs and non-PVVPs of the training and test sets are provided in **Supplementary Data Sheet 2**.

Mathematical Model of Amino Acid Sequences

Each protein sequence is represented by a "one-hot" matrix and a biochemical property matrix. We use a "one-hot" vector to represent a certain amino acid and use a "one-hot" matrix to represent a protein sequence. In the "one-hot" vector, each of the 20 amino acids is represented by a 20-dimensional vector with 19 bits are "0" and a certain bit is "1" (shown in **Supplementary Table 1**). In this way, a protein sequence of length L can be represented by a "one-hot" matrix with length L and width 20. It has been shown that deep learning techniques have a strong ability to extract complex features and specific motifs using sequence "one-hot" encoding (Jones et al., 2017), and this "one-hot" matrix will serve as the input of the deep neural network described below.

It has been shown that the biochemical properties of frequently occurring amino acids that make up PVVPs and non-PVVPs are significantly different. The study of Charoenkwan et al. (2020a) showed that there are 20 biochemical properties of amino acids in the AAindex database (Kawashima et al., 2008) that have a strong correlation with amino acids that make up PVVPs and non-PVVPs. The indexes of these 20 biochemical properties in the AAindex database are FUKS010107, FUKS010111, JACR890101, PRAM820102, QIAN880126, SNEP660102, KOEP990101, QIAN880124, RADA880105, WOLR790101, HUTJ700102, HUTJ700103, ZIMJ680103, FAUJ880104, LEVM760105, FAUJ880111, CHAM830104, LEVM760102, GEIM800101, and EISD860102. A detailed description of these 20 biochemical properties is provided in Supplementary Tables 2, 3 of the paper by Charoenkwan et al. (2020a). In addition to these 20 biochemical properties, Seguritan et al. (2012) suggested that the isoelectric point of amino acids (corresponding AAindex: ZIMJ680104) is an important property for classifying PVVPs and non-PVVPs. Moreover, Ding et al. (2014) found that amino acids that make up PVVPs are often small, and therefore, the molecular weight property (corresponding AAindex: FASG760101) may also be an important property for PVVP identification. In the biochemical property matrix, an amino acid is represented by a 22-dimensional vector in which each bit represents a corresponding AAindex value as mentioned above. Similar to the "one-hot" matrix, a protein sequence of length L can be represented by a biochemical property matrix with length L and width 22. Each AAindex value is normalized between 0 and 1 in the biochemical property matrix.

Design of the Deep Learning Neural Network

We designed a convolutional neural network with a "one-hot" path and a biochemical property path to extract the complex features from the input protein sequence and to further identify whether the given protein is a PVVP. The structure of the neural network is shown in **Figure 1**. In both the "one-hot" and biochemical property paths, we used a one-dimensional convolution operation to detect the sequence features from the "one-hot" matrix and the biochemical property matrix. The length of the convolution kernels is set to 8, the number of kernels of each path is set to 500, and we used the rectified linear unit (ReLU) function as the activation function to perform nonlinear transformations. After the convolution operation, 500 feature maps are generated for each of the "one-hot" matrix and the biochemical property matrix. We then used a one-dimensional global max pooling operation to handle each feature map, and then a 500-dimensional feature vector was generated for each of the "one-hot" matrix and the biochemical property matrix. The two 500-dimensional feature vectors are connected into a 1000-dimensional feature vector. After a batch normalization layer and a fully connected layer with the ReLU activation function, the sigmoid layer calculates a score between 0 and 1 reflecting the likelihood that the given protein is a PVVP. To prevent overfitting, in the training process, there is a dropout layer between the batch normalization layer and the fully connected layer, and a dropout layer between the fully connected layer and the sigmoid layer.

Unlike the existing tools, considering that there may be some incomplete genes in virome data, VirionFinder was trained using

[1]ftp://ftp.ncbi.nlm.nih.gov/refseq/release/viral/

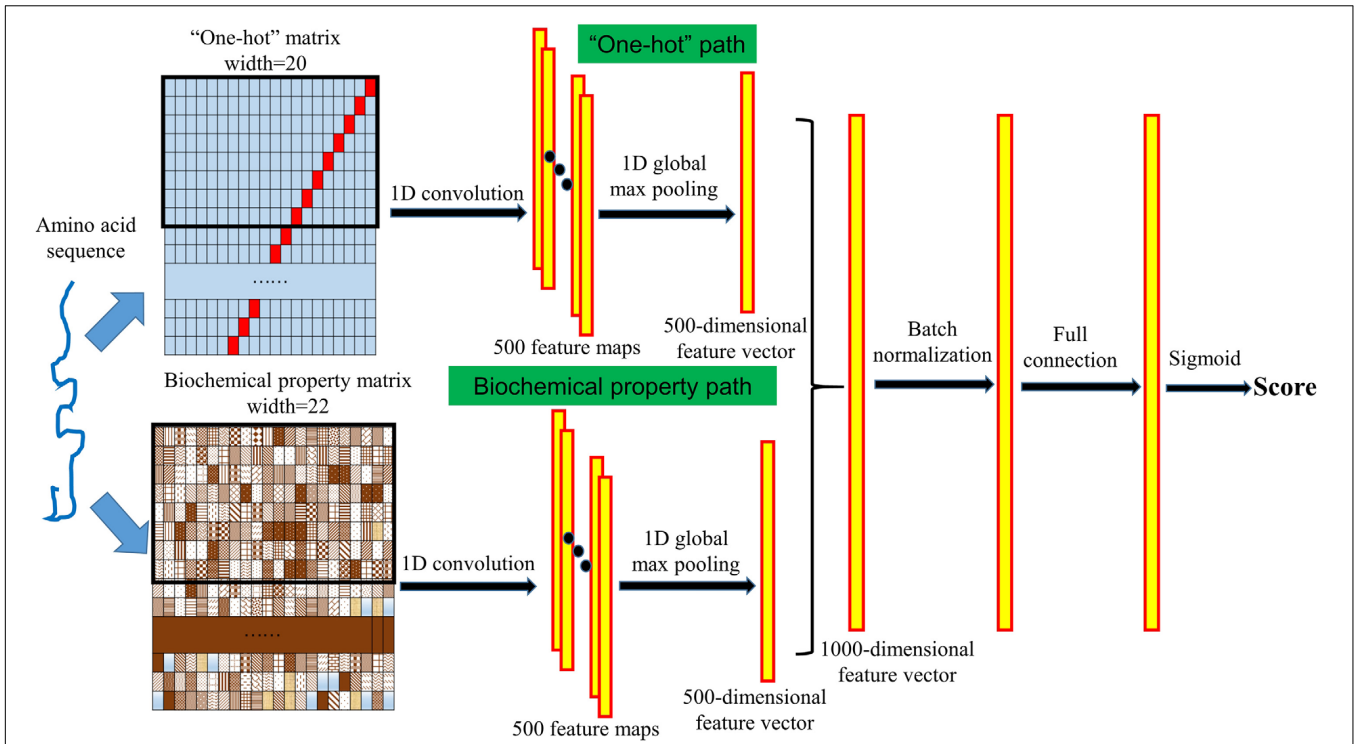


FIGURE 1 | Structure of VirionFinder. VirionFinder contains a "one-hot" path and a biochemical property path to extract complex features from the "one-hot" matrix and biochemical property matrix, respectively. For a given protein sequence, VirionFinder calculates a likelihood score reflecting whether the protein is a PVVP.

protein fragments rather than complete proteins, which helps VirionFinder extract the local features, specific motifs and local conserved functional domains more effectively than previous methods. Specifically, we randomly extracted protein fragments between 30 and 40 aa in the training set and test set, respectively. Finally, 200,000 fragments of both PVVPs and non-PVVPs were generated for the training set, respectively, while 5,000 fragments of both PVVPs and non-PVVPs were generated for the test set, respectively. In the training process, we used the Adam optimizer for the neural network, and the number of iteration epochs was set to 80. For the 10-fold cross validation performed on the training set fragments, VirionFinder achieved an average of area under the receiver operating characteristic curve (AUC) of 91.46% (±0.15%). For the amino acid fragments in the test set, we found that the neural network could achieve an AUC of 88.96%. Furthermore, we tried to remove the biochemical property path and "one-hot" path, respectively, and retrained VirionFinder. We found that these two single-path neural networks could achieve slightly lower AUCs of 87.60 and 85.46%, respectively, indicating that the neural network with both "one-hot" and biochemical property paths may be able to extract useful information from the input data more comprehensively than the neural networks with only one of these paths.

In the prediction process, for amino acid fragments longer than 40 aa, VirionFinder uses a scan window with a length of 40 aa to move across the protein sequence without overlapping, and a weighted average score is calculated for the whole sequence. For example, given a 90-aa sequence, VirionFinder will calculate three scores for the subsequences of 1–40, 41–80, and 81–90 aa. A weighted average score for these 3 scores will be calculated, and the weights for each score are 40/90, 40/90, and 10/90, respectively.

RESULTS

Performance Comparison Against the Benchmark Dataset

We first compared VirionFinder with the currently available tools, namely, iVIREONS, PVPred, PVP-SVM, PVPred-SCM, and Meta-iPVP. To evaluate each tool on both complete and partial genes more comprehensively, we performed the evaluation over four groups of test data with different sequence completeness levels. Group A contains all the complete proteins in the test set. In Group B, each protein in the test set was randomly cut to a subsequence of 75% of the full length. Similarly, Group C contained sequences of 50% of the full length, while Group D contained sequences of 25% of the full length. The evaluation criteria are the sensitivity and specificity, which are given by $Sn = TP/(TP+FN)$ and $Sp = TN/(TN+FP)$, respectively. For VirionFinder, the higher the score of a given protein, the more likely it is a PVVP. In general, a value of 0.5 can serve as the default threshold. To make our comparison more convincing, in the evaluation process, we let VirionFinder achieve the same Sp as the comparison tools by adjusting the threshold, and

under the same *Sp,* we compared the *Sn* of VirionFinder (denoted by *SnV*) with the *Sn* of the corresponding comparison tool (denoted by *SnC*). The results are shown in **Table 1**. In all cases, VirionFinder performed much better than the other tools. Among the comparison tools, Meta-iPVP, which is the newest tool released recently, and iVIREONS are the two best-performing tools, but VirionFinder not only achieves a higher performance but is also stabler for incomplete genes. We found that in the full-length sequences, under the same *Sp*, the *Sn* of VirionFinder is 12.62 and 13.59% higher than that of Meta-iPVP and iVIREONS, respectively, while in the 25% full length sequences, the *Sn* of VirionFinder is 16.18 and 17.15% higher than the *Sn* of these tools, indicating that the advantage of VirionFinder is more obvious in incomplete genes. Therefore, we conclude that VirionFinder can be used as a PVVP annotation tool not only for isolated complete prokaryote virus genomes but also for viral metagenomic data, in which some genes may be incomplete.

Evaluation Using Real Virome Data

We also evaluated VirionFinder and related tools using real viral metagenomic data. It is worth noting that real metagenomic data are hard to use as a benchmark dataset because real data contain a large number of sequences from unknown species that are not present in the current database, and therefore, such an evaluation must be qualitative. We collected lung virome data (Young et al., 2015) from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (accession: SRR5224158.1). We performed the quality control and assembly processes using SPAdes (Bankevich et al., 2012) pipeline by the command "spades.py –meta –1 file1.fastq –2 file2.fastq –o out_folder." The assembled contigs contain 24,230 sequences with a maximum length of 32,273 bp, an average length of 140.83 bp, and the minimum length of 55 bp, indicating that a large number of short reads are poorly assembled. We then used the MetaProdigal (Hyatt et al., 2012) to perform gene prediction. Among the predicted genes, only 7.02% were complete genes. To collect the potential PVVPs, we used position-specific iterated basic local alignment search tool BLAST (PSI-BLAST) to search all the predicted proteins in the PVVPs from the RefSeq viral database. PSI-BLAST was used here because such a homology search strategy is more sensitive for novel genes with low similarity to sequences in the current database. All potential PVVPs with *e*-values less than 1e-5 were collected. Among these potential PVVPs, VirionFinder identified 76.47% of them as PVVPs (using a default value of 0.5 as the threshold), while iVIREONS, PVPred, PVP-SVM, PVPred-SCM, and Meta-iPVP identified 52.94%, 17.65, 17.65, 52.94, and 70.59%, respectively (shown in **Figure 2**), indicating that VirionFinder can identify the highest proportion of PVVP-like sequences as PVVPs. Such results are also consistent with the quantitative comparison against the benchmark dataset in which VirionFinder is the best-performing tool, while the Meta-iPVP tool outperforms the other comparison tools. Additionally, we found that the PVPred and PVP-SVM tools can identify only a few potential PVVPs (<20%), indicating that these tools may not be able to adapt to the situation of virome data, in which a large number of genes are incomplete.

TABLE 1 | Performance comparison between VirionFinder and related tools.

Group	Tool	*Sp* (%)	*SnC* (%)	*SnV* (%)	*SnV-SnC* (%)
Group A	VirionFinder vs. iVIREONS	71.37	78.32	91.91	13.59
Full length	VirionFinder vs. PVPred	90.07	44.01	71.52	27.51
	VirionFinder vs. PVP-SVM	84.31	48.22	82.20	33.98
	VirionFinder vs. PVPred-SCM	79.74	58.90	87.06	28.16
	VirionFinder vs. Meta-iPVP	66.67	81.88	94.50	12.62
Group B	VirionFinder vs. iVIREONS	73.20	74.76	88.67	13.92
75% of the full length	VirionFinder vs. PVPred	88.76	44.34	70.23	25.89
	VirionFinder vs. PVP-SVM	83.53	47.90	79.61	31.72
	VirionFinder vs. PVPred-SCM	75.95	59.22	86.41	27.18
	VirionFinder vs. Meta-iPVP	56.99	85.11	95.47	10.36
Group C	VirionFinder vs. iVIREONS	71.63	73.79	85.76	11.97
50% of the full length	VirionFinder vs. PVPred	85.88	50.49	66.99	16.50
	VirionFinder vs. PVP-SVM	82.22	46.93	73.79	26.86
	VirionFinder vs. PVPred-SCM	73.20	59.87	84.79	24.92
	VirionFinder vs. Meta-iPVP	56.21	81.55	95.47	13.92
Group D	VirionFinder vs. iVIREONS	72.29	59.87	77.02	17.15
25% of the full length	VirionFinder vs. PVPred	78.82	44.01	63.43	19.42
	VirionFinder vs. PVP-SVM	78.04	47.25	63.43	16.18
	VirionFinder vs. PVPred-SCM	67.45	56.63	84.79	28.16
	VirionFinder vs. Meta-iPVP	47.32	79.61	95.79	16.18

We let VirionFinder achieve the same Sp as the comparison tools by adjusting the threshold, and under the same Sp, we compared the Sn of VirionFinder (denoted by SnV) with the Sn of the corresponding comparison tool (denoted by SnC). The column of SnV-SnC presents the advantage of VirionFinder with the comparison tools.

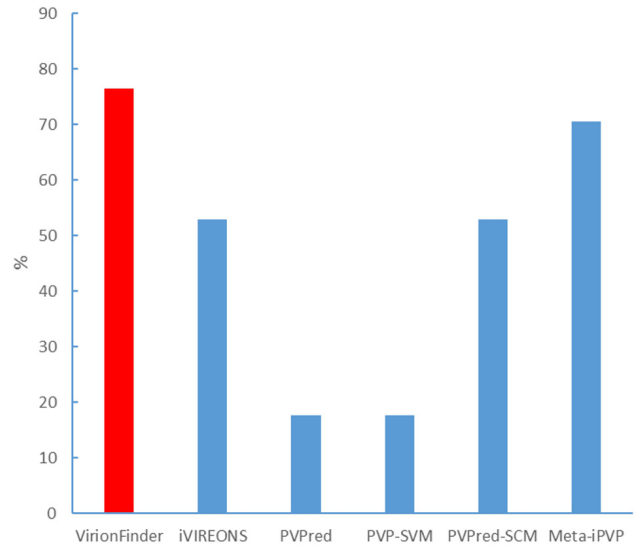


FIGURE 2 | Identification of potential PVVPs by VirionFinder and related tools.

Virion proteins are sometimes encoded next to each other on the genome. We analyzed the longest contig from the virome data. This contig contained 32,273 base pairs and 34 genes. The only gene which was identified as PVVP using PSI-BLAST was the 31st gene from the 5' end, which showed homology with the portal protein. We found that VirionFinder could continuously identify the 30th–33rd genes as PVVP. Correspondingly, iVIREONS and Meta-iPVP could continuously identify the 31st–32nd genes as PVVP; PVPred could not identify the 31st gene as PVVP but identify the 30th gene as PVVP; PVP-SVM continuously identify the 29th–34th genes as non-PVVP and PVPred-SCM continuously identify the 20th–32nd genes as non-PVVP. This showed that VirionFinder had the ability to identify more potential novel PVVPs around the known PVVPs.

We further observed the distribution of VirionFinder scores on all proteins. We found that the distribution showed obvious bimodal distribution (shown in **Supplementary Figure 1**). The bimodal distribution showed that VirionFinder judged most proteins as non-PVVPs with the scores very close to 0 and judged a small fraction of proteins as PVVPs with the scores very close to 1. This observation suggests that the rate of false-positive of VirionFinder is not insanely high and that VirionFinder is able to efficiently identify the subset of predicted CDS with a composition consistent with a PVVP, including likely a number of novel PVVPs.

We further collected 22 virome samples of healthy human gut from Norman et al. (2015). The accession list of the samples is provided in **Supplementary Table 2**. We assembled the short reads and performed gene prediction as we mentioned above, and a total of 278,150 genes were predicted. We used PSI-BLAST to find all PVVP-like sequences as we mentioned above. We found that VirionFinder can identify 83.37% of the PVVP-like sequences as PVVPs, indicating that VirionFinder can achieve robust performance in large scale viral metagenomic data.

It is worth noting that in the lung virome, only 17 out of 7,267 proteins were identified as PVVP with PSI-BLAST, and in the 22 samples of virome data from healthy human gut, only 8,563 out of 278,150 proteins were identified as PVVP with PST-BLAST. This relatively low frequency of PVVP detected suggests that there are some novel PVVPs not currently annotated in real virome data, and alignment-free tools like VirionFinder are needed to identify the most likely PVVPs from these large set of "hypothetical proteins." The related files, including the genes predicted by MetaProdigal, PSI-BLAST output files and VirionFinder result files, are stored in the VirionFinder GitHub website under the "virome" folder.

DISCUSSION AND CONCLUSION

In this work, we present VirionFinder to identify PVVPs using the sequence and biochemical properties of amino acids based on a deep learning technique. VirionFinder takes a complete or partial prokaryote virus protein as input and judges whether the given protein is a PVVP. Tests show that VirionFinder achieves a much better performance than the state-of-the-art tools.

Like other PVVP prediction tools, VirionFinder is designed primarily for prokaryotic viruses, which are dominant in the viral community. The protein sequences in the training set of VirionFinder are also derived from prokaryotic viruses. It is worth noting that the viral community also contains eukaryotic viruses, which are not included in our training set. To allow VirionFinder to better adapt to the real situation of the viral community, we will consider retraining VirionFinder regularly with eukaryotic viruses included in the future. On the other hand, many eukaryotic viruses, such as SARS-CoV-2, are RNA viruses that may not occur frequently in traditional metagenomic DNA sequencing data, and we therefore consider that the existence of eukaryotic viruses may not seriously affect the usage of VirionFinder. We will also consider developing another version of VirionFinder to handle RNA virus sequencing data.

Bacterial host contamination is another issue that need to be pay attention to when using VirionFinder. The training set of VirionFinder did not contain bacterial proteins and therefore, the existing of host contamination may lead to the false positive prediction of VirionFinder. We randomly collected 10,000 bacterial proteins from RefSeq database to test how VirionFinder judge these host proteins and we found that the scores of VirionFinder among these 10,000 bacterial proteins seemed to obey the normal distribution with the mean around 0.5 (shown in **Supplementary Figure 2**), indicating that VirionFinder cannot judge whether the host protein belongs to PVVP or non-PVVP. Therefore, we recommend that user can use related bioinformatics tools to filter out the sequences from host contamination as the preprocessing process before using VirionFinder. Some of the related tools which can distinguish viral sequences and bacterial sequences are listed in the review of Martínez et al. (2020).

In conclusion, VirionFinder achieves the highest performance on both the benchmark dataset and real virome data. It is expected that VirionFinder will be a powerful tool for virome studies.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

ZF and HZ proposed and designed the study, wrote and revised the manuscript. ZF constructed the datasets and wrote the code. Both authors contributed to the article and approved the submitted version.

FUNDING

This investigation was financially supported by the National Key R&D Program of China (2017YFC1310600) and National Natural Science Foundation of China (81925026).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2021.615711/full#supplementary-material

REFERENCES

Abid, D., and Zhang, L. (2018). DeepCapTail: A Deep Learning Framework to Predict Capsid and Tail Proteins of Phage Genomes. *bioRxiv* 23:477885.

Arif, M., Ali, F., Ahmad, S., Kabir, M., Ali, Z., and Hayat, M. (2020). Pred-BVP-Unb: Fast prediction of bacteriophage Virion proteins using un-biased multi-perspective properties with recursive feature elimination. *Genomics* 112, 1565–1574. doi: 10.1016/j.ygeno.2019.09.006

Bankevich, A., Nurk, S., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* 19, 455–477. doi: 10.1089/cmb.2012.0021

Besemer, J., and Borodovsky, M. (2005). GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Res.* 33, W451–W454.

Charoenkwan, P., Kanthawong, S., Schaduangrat, N., Yana, J., and Shoombuatong, W. (2020a). PVPred-SCM: Improved Prediction and Analysis of Phage Virion Proteins Using a Scoring Card Method. *Cells* 9:353. doi: 10.3390/cells9020353

Charoenkwan, P., Nantasenamat, C., Hasan, M. M., and Shoombuatong, W. (2020b). Meta-iPVP: a sequence-based meta-predictor for improving the prediction of phage virion proteins using effective feature representation. *J. Comput. Aided Mol. Des.* 34, 1105–1116. doi: 10.1007/s10822-020-00323-z

Ding, H., Feng, P. M., Chen, W., and Lin, H. (2014). Identification of bacteriophage virion proteins by the ANOVA feature selection and analysis. *Mol. Biosyst.* 10, 2229–2235. doi: 10.1039/c4mb00316k

Fang, Z., Tan, J., Wu, S., Li, M., Wang, C., Liu, Y., et al. (2020). PlasGUN: gene prediction in plasmid metagenomic short reads using deep learning. *Bioinformatics* 36, 3239–3241. doi: 10.1093/bioinformatics/btaa103

Fang, Z., Tan, J., Wu, S., Li, M., Xu, C., Xie, Z., et al. (2019). PPR-Meta: a tool for identifying phages and plasmids from metagenomic fragments using deep learning. *Gigascience* 8:giz066.

Feng, P. M., Ding, H., Chen, W., and Lin, H. (2013). Naïve Bayes classifier with feature selection to identify phage virion proteins. *Comput. Math. Methods Med.* 2013:530696.

Galiez, C., Magnan, C. N., Coste, F., and Baldi, P. (2016). VIRALpro: a tool to identify viral capsid and tail sequences. *Bioinformatics* 32, 1405–1407. doi: 10.1093/bioinformatics/btv727

Hayes, S., Mahony, J., Nauta, A., and van Sinderen, D. (2017). Metagenomic Approaches to Assess Bacteriophages in Various Environmental Niches. *Viruses* 9:127. doi: 10.3390/v9060127

Hyatt, D., LoCascio, P. F., Hauser, L. J., and Uberbacher, E. C. (2012). Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics* 28, 2223–2230. doi: 10.1093/bioinformatics/bts429

Jones, W., Alasoo, K., Fishman, D., and Parts, L. (2017). Computational biology: deep learning. *Emerg. Top Life Sci.* 1, 133–150.

Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., and Kanehisa, M. (2008). AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res.* 36, D202–D205.

Khan Mirzaei, M., Xue, J., and Costa, R. (2020). Challenges of Studying the Human Virome - Relevant Emerging Technologies. *Trends Microbiol.* 1:32622559.

Lekunberri, I., Subirats, J., Borrego, C. M., and Balcázar, J. L. (2017). Exploring the contribution of bacteriophages to antibiotic resistance. *Environ. Pollut.* 220, 981–984. doi: 10.1016/j.envpol.2016.11.059

Manavalan, B., Shin, T. H., and Lee, D. Y., (2018). Sequence-Based Prediction of Phage Virion Proteins Using a Support Vector Machine. *Front. Microbiol.* 9:476. doi: 10.3389/fmicb.2018.00476

Martínez, J. M., Martinez-Hernandez, F., and Martinez-Garcia, M. (2020). Single-virus genomics and beyond. *Nat. Rev. Microbiol.* 6, 1–12. doi: 10.1155/2008/893941

Min, S., Lee, B., and Yoon, S. (2017). Deep learning in bioinformatics. *Brief Bioinform.* 18, 851–869.

Norman, J. M., Handley, S. A., Baldridge, M. T., Droit, L., Liu, C. Y., Keller, B. C., et al. (2015). Disease-specific alterations in the enteric virome in inflammatory bowel disease. *Cell* 160, 447–460. doi: 10.1016/j.cell.2015.01.002

Pan, Y., Gao, H., Lin, H., Liu, Z., Tang, L., and Li, S. (2018). Identification of Bacteriophage Virion Proteins Using Multinomial Naïve Bayes with g-Gap Feature Tree. *Int. J. Mol. Sci.* 19:1779. doi: 10.3390/ijms19061779

Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A., and Sun, F. (2017). VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* 5:69.

Roux, S., Enault, F., Hurwitz, B. L., and Sullivan, M. B. (2015). VirSorter: mining viral signal from microbial genomic data. *PeerJ.* 3:e985. doi: 10.7717/peerj.985

Ru, X., Li, L., and Wang, C. (2019). Identification of Phage Viral Proteins With Hybrid Sequence Features. *Front. Microbiol.* 10:507. doi: 10.3389/fmicb.2019.00507

Saak, C. C., Dinh, C. B., and Dutton, R. J. (2020). Experimental approaches to tracking mobile genetic elements in microbial communities. *FEMS Microbiol. Rev.* 44, 606–630. doi: 10.1093/femsre/fuaa025

Seguritan, V., Alves, N. Jr., Arnoult, M., Raymond, A., Lorimer, D., Burgin, A. B. Jr., et al. (2012). Artificial neural networks trained to detect viral and phage structural proteins. *PLoS Comput. Biol.* 8:e1002657. doi: 10.1371/journal.pcbi.1002657

Sutton, T. D., Clooney, A. G., Ryan, F. J., Ross, R. P., and Hill, C. (2019). Choice of assembly software has a critical impact on virome characterisation. *Microbiome* 7:12.

Tan, J. X., Dao, F. Y., Lv, H., Feng, P. M., and Ding, H. (2018). Identifying Phage Virion Proteins by Using Two-Step Feature Selection Methods. *Molecules* 23:2000. doi: 10.3390/molecules23082000

Young, J. C., Chehoud, C., Bittinger, K., Bailey, A., Diamond, J. M., Cantu, E., et al. (2015). Viral metagenomics reveal blooms of anelloviruses in the respiratory

tract of lung transplant recipients. *Am. J. Transpl.* 15, 200–209. doi: 10.1111/ajt.13031

Zhang, L., Zhang, C., Gao, R., and Yang, R. (2015). An Ensemble Method to Distinguish Bacteriophage Virion from Non-Virion Proteins Based on Protein Sequence Characteristics. *Int. J. Mol. Sci.* 16, 21734–21758. doi: 10.3390/ijms160921734

Zhou, F., and Xu, Y. (2010). cBar: a computer program to distinguish plasmid-derived from chromosome-derived sequence fragments in metagenomics data. *Bioinformatics* 26, 2051–2052. doi: 10.1093/bioinformatics/btq299

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

REVIEW
published: 19 February 2021
doi: 10.3389/fmicb.2021.634511

Applications of Machine Learning in Human Microbiome Studies: A Review on Feature Selection, Biomarker Identification, Disease Prediction and Treatment

Laura Judith Marcos-Zambrano[1], Kanita Karaduzovic-Hadziabdic[2], Tatjana Loncar Turukalo[3], Piotr Przymus[4], Vladimir Trajkovik[5], Oliver Aasmets[6,7], Magali Berland[8], Aleksandra Gruca[9], Jasminka Hasic[10], Karel Hron[11], Thomas Klammsteiner[12], Mikhail Kolev[13], Leo Lahti[14], Marta B. Lopes[15,16], Victor Moreno[17,18,19,20], Irina Naskinova[13], Elin Org[6], Inês Paciência[21], Georgios Papoutsoglou[22], Rajesh Shigdel[23], Blaz Stres[24], Baiba Vilne[25], Malik Yousef[26,27], Eftim Zdravevski[5], Ioannis Tsamardinos[22], Enrique Carrillo de Santa Pau[1], Marcus J. Claesson[28], Isabel Moreno-Indias[29,30] and Jaak Truu[31]* on behalf of ML4Microbiome*

[1] Computational Biology Group, Precision Nutrition and Cancer Research Program, IMDEA Food Institute, Madrid, Spain, [2] Faculty of Engineering and Natural Sciences, International University of Sarajevo, Sarajevo, Bosnia and Herzegovina, [3] Faculty of Technical Sciences, University of Novi Sad, Novi Sad, Serbia, [4] Faculty of Mathematics and Computer Science, Nicolaus Copernicus University, Toruń, Poland, [5] Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Skopje, North Macedonia, [6] Institute of Genomics, Estonian Genome Centre, University of Tartu, Tartu, Estonia, [7] Department of Biotechnology, Institute of Molecular and Cell Biology, University of Tartu, Tartu, Estonia, [8] Université Paris-Saclay, INRAE, MGP, Jouy-en-Josas, France, [9] Department of Computer Networks and Systems, Silesian University of Technology, Gliwice, Poland, [10] University Sarajevo School of Science and Technology, Sarajevo, Bosnia and Herzegovina, [11] Department of Mathematical Analysis and Applications of Mathematics, Palacký University, Olomouc, Czechia, [12] Department of Microbiology, University of Innsbruck, Innsbruck, Austria, [13] South West University "Neofit Rilski", Blagoevgrad, Bulgaria, [14] Department of Computing, University of Turku, Turku, Finland, [15] NOVA Laboratory for Computer Science and Informatics (NOVA LINCS), FCT, UNL, Caparica, Portugal, [16] Centro de Matemática e Aplicações (CMA), FCT, UNL, Caparica, Portugal, [17] Oncology Data Analytics Program, Catalan Institute of Oncology (ICO) Barcelona, Spain, [18] Colorectal Cancer Group, Institut de Recerca Biomedica de Bellvitge (IDIBELL), Barcelona, Spain, [19] Consortium for Biomedical Research in Epidemiology and Public Health (CIBERESP), Barcelona, Spain, [20] Department of Clinical Sciences, Faculty of Medicine, University of Barcelona, Barcelona, Spain, [21] EPIUnit – Instituto de Saúde Pública da Universidade do Porto, Porto, Portugal, [22] Department of Computer Science, University of Crete, Heraklion, Greece, [23] Department of Clinical Science, University of Bergen, Bergen, Norway, [24] Group for Microbiology and Microbial Biotechnology, Department of Animal Science, University of Ljubljana, Ljubljana, Slovenia, [25] Bioinformatics Research Unit, Riga Stradins University, Riga, Latvia, [26] Department of Information Systems, Zefat Academic College, Zefat, Israel, [27] Galilee Digital Health Research Center (GDH), Zefat Academic College, Zefat, Israel, [28] School of Microbiology & APC Microbiome Ireland, University College Cork, Cork, Ireland, [29] Unidad de Gestión Clínica de Endocrinología y Nutrición, Instituto de Investigación Biomédica de Málaga (IBIMA), Hospital Clínico Universitario Virgen de la Victoria, Universidad de Málaga, Málaga, Spain, [30] Centro de Investigación Biomédica en Red de Fisiopatología de la Obesidad y la Nutrición (CIBEROBN), Instituto de Salud Carlos III, Madrid, Spain, [31] Institute of Molecular and Cell Biology, University of Tartu, Tartu, Estonia

The number of microbiome-related studies has notably increased the availability of data on human microbiome composition and function. These studies provide the essential material to deeply explore host-microbiome associations and their relation to the development and progression of various complex diseases. Improved data-analytical tools are needed to exploit all information from these biological datasets, taking into account the peculiarities of microbiome data, i.e., compositional, heterogeneous and sparse nature of these datasets. The possibility of predicting host-phenotypes based on taxonomy-informed feature selection to establish an association between microbiome

OPEN ACCESS

Edited by:
George Tsiamis,
University of Patras, Greece

Reviewed by:
Jonathan Badger,
National Cancer Institute (NCI), United States
Suleyman Yildirim,
Istanbul Medipol University, Turkey

***Correspondence:**
Laura Judith Marcos-Zambrano
judith.marcos@imdea.org
Jaak Truu
jaak.truu@ut.ee

Specialty section:
This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology

Received: 27 November 2020
Accepted: 01 February 2021
Published: 19 February 2021

Citation:
Marcos-Zambrano LJ, Karaduzovic-Hadziabdic K, Loncar Turukalo T, Przymus P, Trajkovik V, Aasmets O, Berland M, Gruca A, Hasic J, Hron K, Klammsteiner T, Kolev M, Lahti L, Lopes MB, Moreno V, Naskinova I, Org E, Paciência I, Papoutsoglou G, Shigdel R, Stres B, Vilne B, Yousef M, Zdravevski E, Tsamardinos I, Carrillo de Santa Pau E, Claesson MJ, Moreno-Indias I and Truu J (2021) Applications of Machine Learning in Human Microbiome Studies: A Review on Feature Selection, Biomarker Identification, Disease Prediction and Treatment. Front. Microbiol. 12:634511. doi: 10.3389/fmicb.2021.634511

and predict disease states is beneficial for personalized medicine. In this regard, machine learning (ML) provides new insights into the development of models that can be used to predict outputs, such as classification and prediction in microbiology, infer host phenotypes to predict diseases and use microbial communities to stratify patients by their characterization of state-specific microbial signatures. Here we review the state-of-the-art ML methods and respective software applied in human microbiome studies, performed as part of the COST Action ML4Microbiome activities. This scoping review focuses on the application of ML in microbiome studies related to association and clinical use for diagnostics, prognostics, and therapeutics. Although the data presented here is more related to the bacterial community, many algorithms could be applied in general, regardless of the feature type. This literature and software review covering this broad topic is aligned with the scoping review methodology. The manual identification of data sources has been complemented with: (1) automated publication search through digital libraries of the three major publishers using natural language processing (NLP) Toolkit, and (2) an automated identification of relevant software repositories on GitHub and ranking of the related research papers relying on learning to rank approach.

Keywords: microbiome, machine learning, disease prediction, biomarker identification, feature selection

INTRODUCTION

The human microbiome represents a complex community of trillions of microorganisms (bacteria, archaea, viruses, as well as microbial eukaryotes such as fungi, protozoa and helminths), well-known to affect general health and homeostasis, e.g., by actively participating in human metabolism and regulating the immune system. Several disease-related states have been linked with a disruption of the steady relationship between the gut microbiota and gut epithelial cells (dysbiosis) (Petersen and Round, 2014). In the last decade, the number of microbiome-related studies has increased notably, and big populational studies such the Human Microbiome Project (Human Microbiome Project Consortium, 2012), the metagenomics of the Human Intestinal Tract (Qin et al., 2010), and the American Gut Project (McDonald et al., 2018), among others, have considerably increased the available data on human microbiome composition and function. These studies provide the essential material to deeply explore host-microbiome associations and their relation to the development and progression of various complex diseases.

Most of the above-mentioned data were generated by amplicon sequencing, primarily by profiling the V3-V4 region of the 16S rRNA marker gene, which allows taxonomic identification of bacteria and archaea. A smaller number of studies have also used 18S rRNA marker gene sequencing to study the microbial eukaryotes such as fungi and protozoa (Elekwachi et al., 2017; Yarza et al., 2017). In both cases, amplicon sequences exhibiting a predefined level of sequence similarity (usually 97%) are commonly clustered into Operational Taxonomic Units (OTUs) that represent the abundance of a particular bacterial taxon (Blaxter et al., 2005). However, due to recent advances in high-throughput sequencing technologies, OTUs are increasingly being replaced by amplicon sequence variants (ASVs), which are un-clustered error-corrected reads (Callahan et al., 2017). After clustering (in case of OTUs) or denoising (in case of ASVs) and feature classification and annotation, the OTU/ASV table with the correspondent abundances is generated. Despite the cost-effective nature of this methodology, 16S rRNA gene sequencing has some drawbacks, e.g., (i) reliable bacterial classification is mostly possible down to the genus level (Winand et al., 2020); and (ii) limited information of the bacterial genes and their functions is obtained.

Another approach that is increasingly being used is the shotgun sequencing of microbial DNA without selecting a particular gene. This approach allows for more accurate classification of the microbial communities (even down to the strain level), and also permits the study of genes and their functions, e.g., by the construction of Gene Ontology (GO) (Ashburner et al., 2000) tables and Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000) pathways (Scholz et al., 2016).

Improved data-analytical tools are needed to exploit all the information from these biological datasets, considering the peculiarities of microbiome data, i.e., compositional data, heterogeneous and sparse nature of the datasets. The possibility of predicting host-phenotypes based on taxonomy-informed feature selection to establish an association between the microbiome, predict various disease states or improve human health is beneficial for personalized medicine. In fact, the gut microbiome has become an integral part of personalized medicine, as it not only significantly contributes to inter-individual variability in health and disease, but also represents a potentially modifiable factor that can be targeted by therapeutics in a personalized manner (Kashyap et al., 2017). In this regard, ML may provide new insights into biomedical analyses, by the development of models that can be used to predict outputs such as categorical labels, binary responses, or continuous values.

Recently, a number of studies have applied ML techniques to analyze human microbiome data, harvesting the hidden knowledge to uncover and understand diversity in taxonomy and function within microbial communities and their impacts on human health. Firstly, to support the taxonomic representation and differentiation in microbiology, models were developed to support the classification of microbial features (Cai and Sun, 2011; Bonder et al., 2012; Werner et al., 2012; Vervier et al., 2016). Secondly, ML was used for the inference of host phenotypes in disease prediction (Pasolli et al., 2016; Flemer et al., 2017; Asgari et al., 2019; LaPierre et al., 2019; Thomas et al., 2019), and finally, to support the use of microbial communities to stratify patients by the characterization of state-specific microbial signatures (Koohi-Moghadam et al., 2019; Wirbel et al., 2019; Yachida et al., 2019).

Here, we aim to review the application of the different ML techniques to human microbiome data analysis and the available ML-based software resources currently used in the analysis of human microbiome data. The review is mainly focused on the application of ML in microbiome studies related to causality and clinical use for diagnostics, prognostics, and therapeutics.

METHODS

Scoping Review Methodology – Identification, Selection, and Organization of Relevant Publications

This study follows the scoping review methodology for searching and assessment of the relevant studies (Arksey and O'Malley, 2005). The breadth of the ML methodology and data types in ML-based microbiome analysis hinder the thorough qualitative analyses of the selected papers, thus giving a scoping nature to this review which aims to search, select and synthesize the findings related to the application of ML in microbiome analysis and identify the available research evidence. The scientific methodology of all emerging review types is common as they rely on a formal and explicit methods for search, selection and evaluation of published studies (Moher et al., 2015). An example of such thorough review guidelines is Preferred Reporting Items for Systematic Review and Meta-Analysis (PRISMA) for systematic reviews in healthcare (Moher et al., 2010). The methodological framework for scoping reviews is established following the exact way how systematic reviews are conducted, providing sufficient details to reproduce the results (Moher et al., 2015). The workflow for a scoping review and adopted in this study, includes 5 stages (Arksey and O'Malley, 2005): (1) identification of a research question; (2) identification of relevant studies; (3) study selection; (4) charting the data; (5) collating, summarizing, and reporting the results.

As the motivation and relevance of the research question has already been extensively elaborated, we focus here on the methodology used to identify and select relevant studies. We have used both manual and automated search of literature corpus in the identification step, performing three independent processes:

- Manual search – crowdsourcing of the studies relevant for the review topic by all members of the COST Action CA18131 "Statistical and machine learning techniques in human microbiome studies". In this way, in total 54 papers were collected, and 35 papers are included in the final list.
- An automated search of digital libraries of three major publishers (PubMed, Springer and IEEE) using NLP Toolkit (Zdravevski et al., 2019) to automate the literature search, scanning, and eligibility assessment. This automated search was additionally constrained to the period from January 2008 to December 2019 (and including those). In total 5,935 papers were identified using this method, after removal of duplicates that appear as a result of multiple searches using the similar subsets of keywords. From that, 67 papers were selected for a manual check, and 37 papers are included in the final list.
- An automated search through the available GitHub resources using NLP algorithms to identify relevant software repositories and extract corresponding scientific papers. The papers were automatically ranked by relevance using the pointwise learning to rank approach (Fejzer et al. unpublished) trained using the manually collected and labeled papers. We found 357 repositories that matched human microbiome research (within 1339 matching microbiome research). In these locations, we found 410 papers, and based on model score, selected 29 papers. The final list includes 17 papers.

The study selection procedure comprised scanning and eligibility assessment steps. The scanning was used in NLP Toolkit thread and served to remove the duplicates and exclude the papers whose title and abstract could not be analyzed due to unavailability, parsing errors, or any other reason. The eligibility assessment step referred to all identified studies in order to select only those relevant for this review. For the studies identified by the NLP Toolkit, the relevance of the study was assessed based on the NLP augmented evaluation of title and abstract according to the prespecified criteria. The papers identified through an automated search of GitHub resources were scored for relevance using the trained model based on learning to rank approach. The detailed description of the methodology used in automated search and eligibility assessment for both NLP Toolkit and learning to rank approach are provided in **Supplementary Material**.

The scoping review workflow illustrating the number of identified, scanned, and articles included in this scoping review using all three data collection procedures is presented in **Figure 1**. The listing of all articles included in this study labeled with respect to different descriptors/keywords is available as Multimedia Appendix.

Medical Subject Headings Annotations

Medical Subject Headings (MeSH) is the NLM controlled vocabulary thesaurus used for the indexing of articles in PubMed. We have used this resource to catalog the 89 papers included in this review from a biomedical perspective to explore the areas that

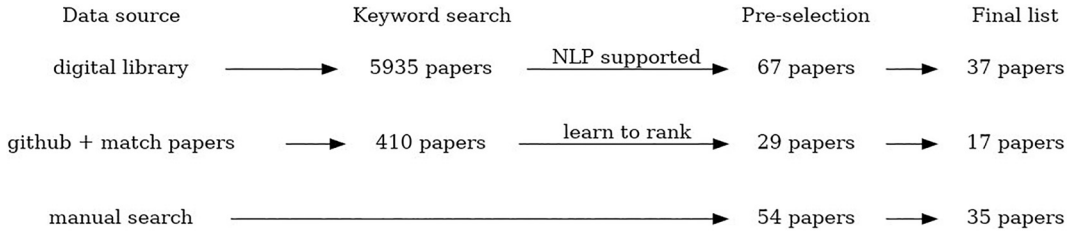


FIGURE 1 | The scheme summarizing the process of paper selection for this review.

are implementing ML techniques in human microbiome studies. The Wordclouds tool was used to summarize the information[1].

Data Acquisition From Different Resources

The human microbiome has been described as a fingerprint, unique and specific to each individual, set in early life and modeled by diet, lifestyle and environmental factors (Gilbert et al., 2018). Besides the high inter-variability of the microbiome, there are some shared functions between the different microbial strains, the so-called core human metagenome established by the analysis of large population studies. Moreover, the characterization of the microbial genes implied in human metabolic functions, the creation of a "gene catalog" of the human microbiome, and the description of differences between specific human conditions have been pointed out by assessing populational studies that have generated great amounts of metagenomics data. The list of main population studies, gene catalogs generated and database resources for analyzing microbiome data, respectively, are shown in **Table 1**.

Data Selection and Pre-processing for ML-Based Applications

Proper normalization of microbiome data is essential for obtaining relevant outcomes from their further processing (Weiss et al., 2017) including ML techniques, with the primary aim to ensure comparability of data across samples. The issue is the large variability in library sizes, constrained additionally by the maximum number of sequence reads of the instrument. This total count constraint induces strong dependencies among the abundances of the different taxa; an increase in the abundance of one taxon requires the decrease of the observed number of counts for some of the other taxa so that the total number of counts does not exceed the specified sequencing depth (Rivera-Pinto et al., 2018). Moreover, observed raw abundances and the total number of reads per sample are non-informative since extracted DNA was normalized during library preparation and also, they represent only a fraction or random sample of the original DNA content in the environment. While Weiss et al. (2017) proposed normalization strategies like cumulative sum scaling, variance stabilization, and trimmed-mean by M-values, none of them really captures the above property of scale invariance, known from the concept of compositional data as observations carrying relative information (Aitchison, 1986; Pawlowsky-Glahn et al., 2015; Filzmoser et al., 2018). A very simple approach of normalization to the total amount of extractable microbial DNA or the total number of targeted cells counted by either flow cytometry or qPCR represented a step in the right direction.

The main idea is to represent the original microbiome (compositional) data in new variables, formed by interpretable log-ratios or their aggregates (log-contrasts), and then to continue in standard statistical or ML processing. There is an increasing number of publications motivating and using the log-ratio methodology of compositional data for statistical processing of microbiome (e.g., Gloor et al., 2017; Silverman et al., 2017; Quinn et al., 2018; Randolph et al., 2018; Rivera-Pinto et al., 2018; Jiang et al., 2019; Quinn and Erb, 2020). However, it still cannot be considered as a mainstream concept in microbiome analysis, mostly due to the high dimensionality of samples and the necessity of dealing with (count) zeros. From the perspective of ML techniques, the outcome is not necessarily a better classification, this depends, as usual, on the capability of a specific method to extract information from (transformed) data, but the compositional approach should reveal relevant sources of differences (microbiome markers) among microbiome samples or groups of samples (e.g., diseased vs healthy).

Literature Review of ML Applications for Microbiome Studies

We finally selected 89 papers for review (35 manually selected, 37 using the automated NLP Toolkit search through PubMed, IEEE Xplore and Springer digital libraries, and 17 by searching in GitHub repositories). ML implies training and evaluation of models to identify, classify, and predict patterns from data. Unsupervised methods aim to identify plausible patterns in the data, without the use of ground truth/labels, while supervised approaches rely on the given labels to train the model and learn the mapping of input features to the labels at the output.

Here, we present the most frequently applied ML methods in microbiome studies, taking into account that ML applied on the large volumes of microbiome data can offer valuable insight into human-microbiome interactions We focused on those studies in which ML is used for: (i) the classification and prediction of microbial taxa, i.e., microbial classification and taxonomic assignment; (ii) the prediction of the host phenotype by linking microbial populations to phenotypes and ecological

[1]https://www.wordclouds.com

TABLE 1 | Different resources and databases for microbiome data acquisition.

Study name	Samples	Description	Data Availability/Ref.
Human Microbiome Project phase 1 and 2 (HMP1, iHMP or HMP2).	HMP1: Healthy adult population of 242 individuals. Samples from 35 body sites, retrieving 13,572 samples in total (i.e., feces = 2,151; buccal mucosa = 633; vagina = 551; other body sites = 10,237). HMP2: Data related to three main conditions: preterm birth, diabetes, and inflammatory bowel disease.	The project generated a huge number of nucleotide sequences of microorganisms, by simultaneously creating protocols to promote reproducible sampling and data generation in microbiome studies, essential for the establishment of computational methods for microbiome data analysis. The iHMP aimed to study host-microbiome interactions by joining the analysis of the immunity, metabolism, and molecular activity to untangle the complex interplay between the host and its microbiomes.	https://hmpdacc.org/.
Metagenomics of the Human Intestinal Tract (MetaHIT)	Overweight and obese adults, and patients with IBD from Spain and Denmark.	This project aimed to characterize the human gut metagenomes of healthy, overweight and obese adults, and patients with IBD from Spain and Denmark. The project has generated 576.7 Gb of sequence and predicted 3.3 million unique open reading frames (ORFs).	Li et al., 2014
American Gut (AGP).	"Wild-type" population. This project currently included microbial sequences from 15,096 samples of 11,336 human participants.	Initially, it was designed to study the North American population, but the initiative attracted also people from the United Kingdom, Australia and other countries. People volunteered to collect feces and fill a questionnaire of their general health status, disease history, and lifestyle to get their microbiome sequenced. The diversity of the data, and the high number of microbial sequences allowed to classify the microbiomes into four great categories, and differences according to country, sex, age, and race, were observed, moreover it adds up to ∼467 million of 16S rRNA V4 gene fragments.	http://americangut.org
The Integrated Gene Catalogs 1 and 2 (IGC, and IGC2).	Comprises more than nine million genes observed in gut microbes. Recently, an updated version of the catalog, denoted added 517,488 supplementary genes.	A catalog of microbial genes including important functions for host-bacterial interaction, and the determination of the so-called "minimal gut bacterial genome" that encompasses genes from bacteria found in most human guts. It has been applied successfully to study host-microbiome associations in the context of different diseases such as type 2 diabetes, obesity, and other pathologies. Genes with co-varying abundance levels can be clustered (Nielsen et al., 2014; Plaza Oñate et al., 2019) to allow taxonomic and functional profiling, and reveal potential disease markers in metagenome-wide association studies.	Qin et al., 2010; Wen et al., 2017
The Unified Human Gastrointestinal Genome (UHGG) and Protein (UHGP) catalogs.	286,997 microbial genomes from the available human gut microbiome datasets.	These catalogs were created by analyzing 286,997 microbial genomes and over 625 million protein sequences, including more than four thousand species. Up to 71% of the taxons analyzed are viable but non-culturable (VBNC), lacking viable culture indicating that most of the microbial diversity in the catalog remains to be characterized.	Almeida et al., 2021
MGnify (formerly EBI Metagenomics)	Diverse microbiome types, including ∼63.000 samples from human microbiome.	Free-access resource for browsing, analyzing, and archiving metagenomic and metatranscriptomic data. The platform contains an automated pipeline for the analysis of microbiome data to determine the taxonomic diversity along with functional and metabolic characteristics.	https://www.ebi.ac.uk/metagenomics/\penalty-\@M Mitchell et al., 2018, 2020
CuratedMetagenomeData	Includes taxonomic and metabolic functional profiles and curated metadata for the publicly available human microbiome samples generated by shotgun metagenomic sequencing.	Bioconductor (Gentleman et al., 2004) package that provides uniformly processed and manually annotated human microbiome data. All data is processed by using the same pipeline, i.e., MetaPhlAn2 (Truong et al., 2015) for taxonomic abundance, gene marker presence and absence, and HUMAnN2 (Franzosa et al., 2018) for coverage and abundance of metabolic pathways and gene families abundance.	Pasolli et al., 2017
Qiita	Sequencing, proteomics, taxonomic, transcriptomics, and metabolomics data.	Open-source management platform for microbial studies. It integrates different omics data, providing a database and compute resources for the analyses of microbiome data.	https://qiita.ucsd.edu/\penalty-\@M Gonzalez et al., 2018
ML Repo	15 published human microbiome datasets.	Public web-based repository of 33 curated classification and regression tasks from 15 published human microbiome datasets. Therefore, it is not only the data repository but it can also be used for benchmarking new machine learning approaches for microbiome data analyses.	https://knights-lab.github.io/MLRepo/\penalty-\@M Vangay et al., 2019

environments, i.e., disease prediction, and (iii) the usage of microbial communities for understanding disease mechanisms, and the further application in personalized medicine (companion test), i.e., biomarker-finding.

Finally, many of the reviewed ML methods have implemented within the Bioconductor packages, initially developed for the microchip/microarray-based data analyses (Gentleman et al., 2004). Consequently, the lessons learned enabled their integration into web portals, such as Microbiome Analyst[2] (Chong et al., 2020) for a comprehensive statistical, visual and meta-analysis of microbiome data.

Supervised Learning Methods

Supervised learning trains and evaluates the model based on the input data complemented with ground truth/labels indicating the outcomes for the given input samples. Common supervised learning approaches include regression analysis and statistical classification.

Logistic regression

Logistic regression (LR) is a statistical method that learns a model that predicts an outcome for a binary variable, Y, from one or more response variables categorical or continuous, X. (Hoffman, 2019).

Logistic regression has been used for establishing microbial signatures in bacterial vaginosis (Beck and Foster, 2015), a disease associated with the vagina microbiome, however, no single microbe has been found to cause it. The authors found that both classifiers identify largely similar microbial community features and that only a few features were necessary to generate models with high classification accuracy. Moreover, the authors investigated the importance of subsets of the microbial community features for the classification process. The taxa identified as more relevant were in line with those identified by previous studies, and classification performance was as well comparable.

In another study, a total of 300 biomarkers were selected from 13,990 features including clinical information and the matrix of relative gene abundance from 806 microbiomes of Chinese individuals (383 controls, 170 with type 2 diabetes, 130 with rheumatoid arthritis, and 123 with liver cirrhosis). Seven algorithms were used, and logistic regression achieved the highest accuracy. This study showed that gut microbiome biomarkers could distinguish abnormal cases from controls with a high level of specificity. The microbiome biomarkers found, present a promising predictive power for application in disease diagnostics, especially disease screening within a large-scale population (Wu et al., 2018).

Tap et al. (2017) set up a ML procedure to identify a microbial signature to predict the severity of Irritable Bowel Syndrome (IBS) using a LASSO-based logistic regression approach applied to 195 subjects. The performance was assessed using the AUROC, and a set of 90 robust OTUs was negatively associated with microbial richness, exhaled methane, presence of methanogens, and enterotypes enriched with the bacterial order *Clostridiales* or genus *Prevotella* (Tap et al., 2017). Fukui et al. (2020) used a similar LASSO logistic regression-based approach to extract a featured group of bacteria for identifying IBS patients. They then applied Random Forest models on the selected features to perform the classification between 85 IBS patients and from 26 healthy controls, obtaining a sensitivity of >80% and specificity of >90% (Fukui et al., 2020).

Linear discriminant analysis (LDA)

Linear Discriminant Analysis (LDA) is a generalization of Fisher's linear discriminant, a method used in statistics, pattern recognition and machine learning to find a linear combination of features that provides good separation between the classes of objects or events. When applied to microbiome data, this approach finds a linear combination of microbial features in the training data that models the multivariate mean differences between classes (Zhou and Gallins, 2019).

The linear discriminant analysis (LDA) effect size (LEfSe) method proposed by The Huttenhower Lab as part of bioBakery workflows for executing microbial community analyses[3] was specifically designed for biomarker discovery in metagenomic data (16S rRNA gene and whole-genome shotgun datasets). It performs high-dimensional class comparisons that determine the features: organisms, clades, operational taxonomic units, genes, or functions; most likely explaining differences between classes. It joins standard tests for statistical significance plus additional tests encoding biological consistency and effect relevance. The algorithm first uses the non-parametric factorial Kruskal-Wallis (KW) sum-rank test to detect features with significant differential abundance regarding the class of interest. Then, biological consistency is investigated using a set of pairwise tests among subclasses using the (unpaired) Wilcoxon rank-sum test, finally uses LDA to estimate the effect size of each differentially abundant feature and perform dimension reduction (Segata et al., 2011).

k-nearest neighbors (k-NN)

k-NN is based on simple classification rule, assigning the new sample to a class which is in the majority among the k training samples nearest to that point. The algorithm can be used both for classification and regression problems, depending on a type of the outcome variable (discrete or continuous). The neighborhood is defined using a selected distance metric in a multidimensional feature space. Euclidean distance or correlation coefficients are the most regularly used distance metrics. For continuous traits, a weighted average of the k nearest neighbor is used (Zhou and Gallins, 2019).

k-NN has been used to effectively determine the postmortem interval (PMI) using microbial samples from the skin microbiota found in the nasal and ear canals of cadavers. When the microbiota from both sites was considered jointly, the regression was successful, yielding a model that accurately predicts the postmortem interval to within 55 accumulated degree days (ADD), which represents about two days of decomposition at an average temperature of 27.5°C (Johnson et al., 2016).

Hacılar et al. (2018) compared several ML-based techniques to classify fecal samples as healthy or with disease [i.e., Inflammatory Bowel Disease (IBD)]. They used a dataset

[2]https://www.microbiomeanalyst.ca/

[3]http://huttenhower.sph.harvard.edu/lefse/

containing shotgun metagenomic data from 382 individuals (234 healthy and 148 IBD patients). The training set was a profile of gut microbial communities for each sample generated by MetaPhlAn2 (Segata et al., 2012). Several models were trained (RF, Adaboost, k-NN + LogitBoost, Decision tree, Neural network, LogitBoost and Furia) and 10-fold cross-validation was performed to evaluate the performance for each model. Finally, they added a feature selection (i.e., mRMR: minimum redundancy and maximal relevance) step before the training process. With and without feature selection k-NN + LogitBoost performed best with 0.87 and 0.86 accuracy scores, respectively (Hacılar et al., 2018).

Naïve Bayes classifiers

Naïve Bayes classifiers are a family of simple probabilistic classifiers based on the application of Bayes' theorem with strong (naïve) assumptions of statistical independence between the features. In one such study applying NB to microbiome data, Werner et al. (2012) investigated the influence of the training set on the results of the taxonomic classification of 16S rRNA gene sequences generated in microbiome studies. The classification using a naïve Bayes classifier indicated that taxonomic classification accuracy of 16S rRNA gene sequences improves when a Naive Bayes classifier is trained only on a selected region of the target sequences. This result was used for some other classifiers (e.g., in QIIME2) (Werner et al., 2012).

Support vector machines (SVM)

SVMs is a machine learning algorithm that aims to learn a decision boundary between the classes, so as to ensure the maximum achievable distance (margin) between the samples closest to the decision boundary. The samples relevant for learning a decision boundary are only those closest to it, called support vectors. When linear separation between classes is not possible in original feature space, the SVM uses the kernel trick to estimate the decision boundary in a higher-dimensional space (Cortes and Vapnik, 1995). SVM can as well be used for regression tasks.

A Sino-European team (Qin et al., 2010) led an early study using WGS data in order to identify dissociative genetic markers from fecal sample sequencing data for IBD and Type II diabetes (T2D). They used a variety of tools to process the raw reads: SOAPdenovo (Li et al., 2010) for assembly; MetaGene (Noguchi et al., 2006) for gene prediction; KEGG (Kanehisa et al., 2004) and eggNOG (Jensen et al., 2007) for functional annotation. They selected 50 marker genes for T2D (using mRMR: minimum redundancy and maximal relevance) out of a gene catalog containing roughly 300,000 genes. They also show that taxonomic abundance data segregates IBD and healthy individuals when performing PCoA.

Cui and Zhang (2013) described an alignment-free supervised classification procedure for the classification of metagenome samples into predefined classes with sequence signatures from shotgun metagenomics sequencing data by using recursive SVM, this approach integrates feature selection and classification steps in one method. They also applied the methodology on a real metagenome dataset to classify IBD and non-IBD samples. The accuracy obtained using the stringent leave-one-out cross-validation (LOOCV) was 88%, additionally permutation experiment were performed to evaluate statistical significance (Cui and Zhang, 2013).

Liu Y. et al. (2011) presented "MetaGUN"[4] a gene prediction method for identifying genes in metagenomic fragments based on SVM. Initially, input sequences were classified into phylogenetic groups, using a k-mer based sequence binning method. Afterward, for each group, the identification of protein-coding sequences was performed using SVM classifiers. MetaGUN applies universal prediction modules and a novel prediction module to identify protein-coding sequences. Entropy density profiles (EDP) of codon usage, Translation Initiation Side (TIS) scores and Open Reading Frame (ORF) length are employed as discriminative features and used as inputs into the classifiers to distinguish protein-coding sequences from non-coding sequences. In the last stage, TISs are relocated by employing a modified version of MetaTISA. The MetaGUN prediction method was compared with six existing metagenomic gene finders (Liu Y. et al., 2011). The results showed that the performance of MetaGUN is better for 3′ end of genes on longer fragments, and comparable results were obtained with Glimmer-MG on shorter fragments. For 5′ end of genes, with fragments of various lengths, MetaGUN outperformed other tested methods on the overall TISs. When applied on two healthy human gut microbiome samples, MetaGUN was able to find more novel genes than other methods (Liu Y. et al., 2011).

Ning and Beiko (2015) explored a phylogenetic approach in classification of oral microbiota using a ML approach focusing on classification using SVMs. The authors used phylogenetic information as the basis for the proposed custom kernels and as classifier features. Other than using the phylogenetic information (such as taxon and clade abundance), PICRUSt (Langille et al., 2013) that predicts molecular functions from 16S rRNA sequence data was used to generate additional input features. The proposed kernels based on UniFrac measure of community dissimilarity (Lozupone et al., 2011) did not result in improved performance. Even though the combinations of the selected input features were important predictors, they did not result in increased accuracy. The classification was performed on nine oral sites and resulted in a modest 81% prediction accuracy which indicates the challenges of classification of oral microbiota.

Another study, performed by Larsen and Dai (2015), demonstrated that the metabolome derived from the human gut microbiome might be predictive of host dysbiosis. Metagenomic enzyme profiles predicted from 16S rRNA microbiome community structures were used to generate metabolic models. The authors apply SVM to show that emergent property of the microbiome and its aggregate community metabolome of human gut are more predictive of dysbiosis than the microbiome community composition or predicted enzyme function profiles.

Artificial neural networks

Artificial neural networks refer to an interconnected feed-forward network of neural units each comprising multiple inputs

[4]http://bioinfo.ctb.pku.edu.cn/MetaGUN/

and a single output, organized in several layers to map a feature vector from the input layer, to the class label at the output layer. The inputs to each neuron are weighted outputs from the neurons from a previous layer, which are summed and non-linearly transformed at its output. The total number of hidden layers and the number of neurons within each hidden layer are specified by the user. All neurons from the input layer are connected to all neurons in the first hidden layer, with weights representing each connection. This process continues until the last hidden layer is connected. The backpropagation algorithm is used to modify the weights in a neural network optimizing for the classification accuracy. For microbiome data, OTUs/ASVs are commonly used at the input layer, with separate neurons for each OTU/ASV.

Lo and Marculescu (2019) describe a neural network platform for the classification of host phenotypes from metagenomic data, using a new data augmentation technique to mitigate the effects of data over-fitting. They tested the proposed framework on eight real datasets including data from HMP (Turnbaugh et al., 2007), and two diseases, i.e., IBD (Gevers et al., 2014), and esophagus diseases (esophagitis, Barrett's esophagus, esophagal adenocarcinoma; Yang et al., 2015), finding that the new proposed methodology outperforms other models previously used in the literature (Lo and Marculescu, 2019).

Deep Learning

Deep learning (DL) is a ML method that assumes using artificial neural networks (ANNs) with deep architectures, i.e., multiple hidden layers, yielding a higher level of abstraction and in general a significant improvement in performance given very large data sets. Another advantage to other ML methods is that DL architectures learn the feature representation given the raw data at its input, thus alleviating the feature engineering step. Currently, DL is thought to be the most advanced ML technique for a variety of applications (Chassagnon et al., 2020).

To classify human epithelial materials highly relevant for forensic investigations, Díez López et al. (2019) applied taxonomy-independent DL methods on skin, saliva, and vaginal microbiome data obtained from the Human Microbiome Project. A total of 1636 validated reference samples from these sites were used to identify most informative sequence positions via correspondence analysis. High-inertia positions were used as input matrix to train 50 DL networks based on a 4-layer ANN. Two sets of samples (110 test and 41 mock casework samples) were deployed to validate the output from the deep learning approach with most of the samples being classified correctly. This approach offers a more accurate and efficient tissue-classification approach compared to human biomarkers, as donor DNA-based methods often lead to cross-identification and low specificity due to overlaps in human cell composition. However, a successful application of DL methods in such a context ideally requires standardized biological and methodological conditions during the generation of training and test data (Díez López et al., 2019).

Another example of using DL approach for analyses of metagenomic data are DeepARG networks which are trained to predict antibiotic resistance genes (ARGs) from metagenomic data (Arango-Argoty et al., 2018). DeepARG consists of two models: DeepARG-LS, which was developed to classify ARGs based on full gene length sequences, and DeepARG-SS, which was developed to identify and classify ARGs from short sequence reads. The initial collection of ARGs was obtained from three major databases: CARD, ARDB, and UNIPROT and 30 ARG categories were used to train the models. To further evaluate and validate performance, the DeepARG-LS model was applied to all the ARG sequences in the MEGARes database (Lakin et al., 2017). Also, the ability of the DeepARG-LS model to predict novel ARGs was tested on a set of 76 metallo-beta-lactamase genes obtained from the study of Berglund et al. (2017). Based on the results the authors conclude that the DeepARG models can be used to get an overview or inference of the kinds of antibiotic resistance in a collection of sequences; however, still the downstream experimental validation is required to confirm whether the sequences truly confer resistance.

Asgari et al. (2019) used deep learning, Random Forest(RF) and SVM, for distinguishing among human body-sites, diagnosis of Crohn's disease, and predicting the environments from representative 16S gene sequences. Moreover, they also proposed a reference- and alignment-free approach for predicting environments and host phenotypes from 16S rRNA gene sequencing data based on k-mer representations. They described that for large datasets (10K samples per class) using DL provides more accurate predictions. However, when the number of samples is not large enough, RFs performed better on both OTUs and k-mer features. However, for classification over representative sequences as opposed to samples (pool of sequences), the SVM outperformed the RF classifier (Asgari et al., 2019).

Convolutional neural network CNNs are similar to traditional deep neural networks (DNNs), they are made up of layers of neurons that have learnable weights and biases. Each neuron receives some inputs, calculates a dot product, and optionally follows it with a non-linear function (Lopez Pinaya et al., 2020). In 2017, this team (Fioravanti et al., 2018) introduced a phylogenetic CNN that would enable the classification of gut microbiome metagenomic data into healthy or IBD phenotypes, summing up to a total of 6 classification tasks. Those phenotypes included the different subtypes of the disease: Crohn's disease (CD) and Ulcerative Colitis (UC), as well as the state of the pathology (flare or remission) and the part of the intestine that is affected for CD (ileum or colon). The dataset used for training (Sokol et al., 2017) contained bacterial and fungal community (16S rDNA and ITS) from 38 controls and 222 IBD patients. Pre-processing of the raw data was carried out using QIIME2 (Kuczynski et al., 2012), UCLUST (Edgar, 2010) and RAxML (Stamatakis, 2014), in order to get relative abundance, cluster the taxa and build a phylogenetic tree that will then be input to the CNN. A synthetic dataset was also constructed as deep learning performs better when trained on large datasets. To do so, they generated vectors in the Aitchison simplex that is spanned by the "real" dataset. This improved the performance of the CNN, which tends to overfit when trained only on the initial dataset. They compared the performance of their newly crafted CNN with more traditional learning models (LSVM, RF, Multi Layer Perceptron NN)

using the Matthews Correlation Coefficient (MCC) as a metric. Overall, for each of the six tasks, the CNN outperformed the other models.

Ensemble Methods

Ensemble methods combine multiple classifiers to obtain a better performance than a single classifier.

Random forests (RF)

RFs are an example of ensemble learning, in which a complex model is made by combining many simple models. In this case, simple models are decision trees. RFs use a bootstrap resampling on the given dataset to learn each decision tree using a single boostrap set. The final output of a RF is obtained using a majority voting of the individual decision trees. As these are well-studied methods, they are used as baselines for comparison in many studies (Breiman, 2001). The most widely used ML algorithm, RF classifiers have been frequently used along with Least Absolute Shrinkage and Selection Operator (LASSO) for feature selection, for stratification of patients (Flemer et al., 2017; Yachida et al., 2019) and biomarker finding (Koohi-Moghadam et al., 2019; Thomas et al., 2019; Wirbel et al., 2019) and finding of host-microbial signatures to detect fecal contamination in environmental samples (Roguet et al., 2018).

RF has been used for classification of pediatric patients of Crohn's disease (CD) according to disease state and treatment response by using the alpha diversity of the samples and the genetic risk score (GRS) of each patient (Douglas et al., 2018). They found higher classification accuracy with 16S rRNA datasets than shotgun metagenomics due to the higher contamination of human DNA in the shotgun metagenomes.

Ross et al. (2017) analyzed the impact of cohabitation on the individual composition of the skin microbiome. For the analysis, the authors used 16S rDNA amplicons of bacteria and archaea from 330 skin samples from 17 skin regions of 10 heterosexual cohabiting couples. Analysis was performed using both statistical and ML methods. Their results showed that the two most important factors that affect the skin microbiome are individuality and body region, which is in line with previous studies. The authors also showed that cohabitation strongly influences skin microbial community diversity. When RF method was applied for skin microbiome classification, accuracy greater than 86% was achieved (Ross et al., 2017).

Ai et al. (2019) took advantage of the continuously decreasing price of whole genome sequencing technology to diagnose colorectal cancer (CRC) based on fecal shotgun sequencing data. They used a dataset consisting of French and Austrian cohorts both containing 156 individuals (312 in total; 124 healthy and 188 CRC and adenoma patients). To preprocess the raw reads and produce the relative abundance of each taxon in the gut, they used the GRAMMy tool (Xia et al., 2011). In order to select taxa that best discriminate a healthy sample from a sample displaying tumor-related dysbiosis, ML techniques were implemented; feature (taxon) selection was carried out using information theory (mutual information) and a RF classifier was trained using a 6-fold cross-validation process. This resulted in the selection of a set of taxa whose abundance was a good indicator of the presence or not of CRC related dysbiosis in the gut (Ai et al., 2019).

Rahman et al. (2017) used metagenomes to identify antibiotic resistance genes in the infant gut microbiome. Their findings were in line with previous work showing that there is an increase of resistance gene levels after antibiotics intake, which is followed by the recovery of the microbial community. The authors also found that, over time, the formula feeding influences the gut resistome. A RF model was used to classify resistomes of formula-fed and breast-fed babies. Using feature importance, the trained model was then used in the selection of resistance genes. Furthermore, ML methods were used to select genes that can predict the change in relative abundance of an organism after the intake of vancomycin and cephalosporin antibiotics. The best results were obtained using the boosted decision trees (Rahman et al., 2017).

Yang et al. (2019) applied a RF classifier for forensic identification based on an individual's microbial sample using a combination of single-nucleotide polymorphisms (SNPs) in the 16S rRNA gene of *Cutibacterium acnes* and skin microbiome OTU table, achieving 93.3% accuracy. Their work also showed that the genotype of *C. acnes* 16S rRNA gene was more stable over time than that of the skin microbiome profile. The proposed method showed promising results for microbiome-based forensic identification (Yang et al., 2019).

Gupta et al. studied a cohort of patients with CRC from India by using shotgun metagenomics. They identified 20 potential microbial taxonomic markers based on their significant association with the health status, and 33 potential microbial gene markers using Weka and the Boruta R packages. They applied RF with the selected biomarkers and combined with two different cohorts from China and Austria successfully discriminated the Indian CRC from healthy microbiomes with high accuracy (Gupta et al., 2019).

Sze and Schloss (2016) conducted a meta-analysis to detect if specific microbiome-based markers can be associated with obesity. The authors selected ten previously published studies, re-calculated OTU tables with the available 16S rRNA sequencing data, applied RF models trained on each data set and tested them on the remaining data sets to predict the obesity status of the subjects. The authors found weak relationships between richness, evenness, and diversity and obesity status. Moreover, they also showed that most studies lack the power to detect small differences in alpha diversity metrics and phylum-level relative abundances. The analysis demonstrated that the ability to reliably classify individuals as obese only based on the composition of their microbiome was limited. The authors concluded that the involvement of the microbiome in obesity is not apparent based on the taxonomic information provided by 16S rRNA gene sequence data (Sze and Schloss, 2016).

Braun et al. (2019) studied patients with quiescent Celiac Disease (CD) and compared their microbiota with both CD and healthy patients. The RF model was used to prioritize taxa that best distinguish relapses from non-relapses. Top three taxa were used to construct the flare index that was significantly different for flare and no-flare samples. Flare index also significantly

correlated with microbial richness and microbial dysbiosis index (Braun et al., 2019).

Fabijanić and Vlahoviček (2016) utilized the translational optimization effect, a property of gene regulation, to distinguish subjects with liver cirrhosis from healthy controls using the RF classifier (Fabijanić and Vlahoviček, 2016). Another study that utilized the RF algorithm on gut microbiome data is described by Hasic and Music; the condition studied was Multiple Sclerosis (MS). The results demonstrate the best accuracy in distinguishing control samples from MS samples when genus-level taxa abundances were used as features. The model learned on one dataset was evaluated on another set of the MS samples coming from people living in another country. The classification accuracy on this test set was comparable to the error on the validation set (Telalovic and Azra, 2020).

Multiple decision tress

Travisany et al. (2015) proposed an ensemble method for microbial taxa prediction present in a specific environment as well as their abundances using multiple CARTs (classification and regression tree). The authors first constructed a dataset of genomic fragments by collecting genomes from publicly available databases. They built two predictors, one using a dataset with 98 genera of the gastrointestinal tract available from the Human Microbiome Project, and the other with 17 early studied genera of the gastrointestinal tract. They computed the statistics of k-mer frequencies, GC radio and GC skew for each read for a specific environment-associated dataset. The prediction was then performed by majority vote selection of multiple ($n = 558$) CART trees. The proposed method was evaluated using simulated and public human gut microbiome datasets. Using 17 representative genera, the authors achieved an accuracy of 77% in read assignments (Travisany et al., 2015).

Gradient boosting (GB)

A ML method that addresses regression and classification problems by generating a prediction model as an ensemble of weak predictors, mostly decision trees, and then averaging predictions over decision trees of fixed sizes. As with other forms of boosting, the process successively computes weights for the poorly predicted samples.

For the gut microbiome, GB has been applied by Zeevi et al. (2015). Their study included a cohort of 800 overweight or obese non-diabetic individuals, in which the gut microbiome was being profiled (relative abundances of 16S rRNA amplicon-based phyla, metagenome-based species and KEGG modules) along with their nutritional profiles, as well as several blood parameters and anthropometrics to successfully predict the post-meal glucose levels for each individual and each meal. Their ML model was based on a stochastic gradient boosting regression (Friedman, 2001). When using stochastic gradient boosting, at each iteration, a randomly selected subsample is drawn from the training data without replacement, which is then used to fit the model. Zeevi et al. used 80% of their samples and 40% of the features. They did not limit the depth of the three, however, it was required that the leaves have at least 60 instances (i.e., meals, in their case). In total, 4000 iterations were used with a learning rate of 0.002. The authors subsequently validated the output from the trained ML model in an independent cohort of 100 participants. Further, they conducted a blinded randomized controlled dietary intervention in another cohort based on the ML-based predictions, observing similar improvements in the post-meal glucose levels, accompanied by consistent alterations to the gut microbiota (Zeevi et al., 2015).

Faust et al. (2012) employed GB to investigate co-occurrence relationships in 16S rRNA data obtained from the Human Microbiome Project. Generalized boosted linear models were fitted using taxa abundance data from source sites to predict abundances of target taxa within targets sites. The analysis was augmented with the integration of a set of similarity and dissimilarity measures (Pearson and Spearman coefficients for correlation, Bray-Curtis and Kullback-Leibler as dissimilarity measures) to finally create a network of co-occurrence and co-exclusion relationships within the analyzed microbiomes. By putting these tools together, the authors were able to reveal that closer related taxa tend to co-occur in special vicinity or environmentally similar habitats whereas phylogenetically more distant microbes with similar functional aptitudes are more likely to compete. A major difficulty in developing this method was taking into account the compositional character of relative abundance data which could lead to spurious correlations. However, coupling permutations and repeated renormalization contributed to maintaining true correlations. While these observations were made on data from the Human Microbiome Project, the computational methodology can be transferred to other research questions involving marker gene sequencing (Faust et al., 2012).

GB has been applied to analyze a combination of 16S rRNA, host transcriptome, epigenome, genotype and dietary data from colonic biopsies of inflammatory bowel disease patients and healthy controls using XgBoost (Ryan et al., 2020). When microbiota information was combined with diet and host genotype, the disease classifications improved significantly, and even more so when host epigenome and microbiota data were combined.

Applications of Several Machine Learning Methods

Le Goallec et al. (2020) proposed a framework for building microbiome-derived indicators of host phenotypes of infant age, sex, breastfeeding status, historical antibiotic usage, country of origin, and delivery type. By leveraging five different types of data and their combinations (host demographics ("baseline" data) and the four microbiome data type: BioCyc pathway relative abundance, Co-Abundance Groups (CAGs) relative abundance, MetaPhlAn2 taxa relative abundance, and gene relative abundance, they compared the prediction performances of 8 machine learning methods: 2 different elastic net (Elastic Net Caret and Elastic Net 2) implementations, 2 random forest (RF Caret and RF2) implementations, 2 gradient boosted machine (GBM Caret and GBM2) implementations, support vector machines (SVM, kernels: linear, polynomial of degree 2 and radial), K-nearest neighbors (KNN) and naive Bayes (NB). In their investigation, they found that non-linear models and particularly the Gradient Boosted Machines (Caret) were

the most consistently effective at the classification of sex, breastfeeding status, country of origin. For other phenotypes such as age and prior antibiotic usage, the information encoded in the microbiome seems to be linear, as no significant difference was observed between the elastic nets and the tree-based methods. In these cases, linear methods were a better choice, because of the ease of interpretation. The authors concluded that significant pairwise relationships could be detected between phenotypes and biomarkers (Le Goallec et al., 2020).

A UK based team carried out a study aiming at building a hybrid classifier that would perform several classification tasks [IBD presence (1), subtype (2) and severity(3)] (Wingfield et al., 2016). A publicly available dataset of 16S rRNA containing fecal sequencing data from 37 healthy individuals and 122 IBD patients) was used in order to train the three aforementioned models. For each sample, the sequenced reads were pre-processed into taxonomic and functional profiles using QIIME2 (Kuczynski et al., 2012) and PICRUSt (Langille et al., 2013) respectively. Then, a pipeline of three consecutive classifiers (SVM for stages one and two, multilayer perceptron (MLP) for stage three) was developed and the classifiers were cross-validated. The outcomes of the different classification steps were disease-free, IBD remission and IBD active for stage one. Ulcerative colitis (UC), Crohn's disease and control for stage two and finally mild, moderate and severe for stage three. The average precision scores for the k-fold cross-validations were rather low, 0.71, 0.65 and 0.61 for stages one, two and three respectively, however the average area under the ROC curves were consistently better (ranging from 0.7 to 0.9).

In another study, a framework entitled Phy-PMRFI (Phylogeny-aware modeling for prediction of metagenomic functions using RF Feature Importance), the authors use ML for microbiome functional properties. They integrated quantitative profiles of taxa (abundance counts of OTUs) and biological information derived from the phylogeny of microbial taxa. This approach helped to select taxa at different taxonomic levels that reck in associating a metagenomic sample with the host environmental phenotypes. It implemented a phylogeny and abundance-aware matrix (PAAM) (Wassan et al., 2018b) that combines phylogeny with the abundance counts of microbial taxa. For Phy-PMRFI, the authors used RF to recognize microbial features that are useful for classifying phenotypic groups and improve metagenomic predictions. Afterward, the informative microbial taxa obtained acted as an input to three commonly used MLclassifiers: (1) SVM, (2) Logistic Regression, and (3) Naive Bayes, intending to identify if phylogenetic relatedness is a good predictor of functional similarity. For this, the authors used three microbiome datasets as cases to demonstrate the utility of the Phy-PMRFI framework in predicting functions of metagenomic data. They concluded that inclusion of the phylogenetic measure potentially maximizes the opportunity of classifying microbiome functions according to naturally inherent properties of taxa (Wassan et al., 2019).

Beck and Foster (2014) applied genetic programming, RF and logistic regression to classify microbial communities into bacterial vaginosis (BV) positive and negative categories. Using the mentioned classification models, most important features of the microbial community used to predict BV were also identified. The classification was applied to two different datasets. The authors obtained an accuracy above 90% for Nugent score and above 80% for the Amsel criteria. Even though different sets of most important features were identified by the tested classifiers, the shared features, in general, agree with the previous research (Beck and Foster, 2014).

In the context of the human gut microbiome, Zhu et al. (2020) proposed a DL ensemble feature selection model, Deep Forest, which is based on the RF method to perform microbiome-wide association studies (MWAS). When tested on three data sets using several classifiers, the proposed method achieved better classification performance than SVMs, k-NNs and convolutional neural networks (CNNs). Performance evaluation of Deep Forest was also evaluated in terms of feature selection. The method achieved better results with the selected reduced feature subset. When the selected features were compared to the existing literature, identified microbial biomarkers have found to have a relationship with the diseases (Zhu et al., 2020).

Statnikov et al. (2013) performed a comprehensive evaluation of 18 ML methods and five feature selection methods to perform body site and subject multicategory classification and diagnosis using microbiome data. The evaluation was performed on eight datasets using constructed OTU tables as input features for the ML methods. Performance of evaluated methods was measured using the proportion of correct classifications and relative classifier information metrics. From the evaluated methods, RF, SVM, kernel ridge regression, and Bayesian logistic regression with Laplace priors were among the best-performing methods with statistically similar levels of classification accuracy (Statnikov et al., 2013).

In work published by Eck et al. (2017) two datasets were analyzed. One distinguished skin from gut microbiome samples and the other IBD patients from healthy individuals. Several ML algorithms were applied: Linear SVM, RF, nearest shrunken centroids, logistic regression with l_2 regularization. The authors measured the most important taxa on species level (applying intergenic spacer profiling of 16S-23S rRNA) for the classification when applying different algorithms. The identification of such taxa facilitates biologically meaningful interpretation of the microbiota-based predictions (Eck et al., 2017).

Hollister et al. (2019) evaluated the relationships of pediatric IBS and abdominal pain with intestinal microbes and fecal metabolites. By leveraging both metagenomic and metabolomic information, and using LASSO feature selection, RF models, and SVM, the authors selected ten features including abundances and distributions of the metabolites, bacterial species, and functional pathways. Features selected were capable of distinguishing pediatric IBS cases from controls with an AUC of 0.93 and $\geq$ 80% accuracy. Moreover, the bacterial features and metabolites described appeared to be closely linked with abdominal pain and emphasized the importance of the microbiome-gut-brain axis to human health (Hollister et al., 2019).

Pasolli et al. (2016) used the SVM, RF classifiers, LASSO and elastic net regularized multiple logistic regression, Neural Networks and Bayesian logistic regression, and assessed the

prediction power of metagenomic data in linking the gut microbiome with disease states (Pasolli et al., 2016).

Liu Z. et al. (2011) developed a method called MetaDistance that integrates SVM and k-NN for multiclass classification and additionally performs feature selection. The proposed method showed good classification accuracy for classifying body sites and skin sites according to 16S rRNA gene data. Besides, the method was demonstrated to be robust for small sample sizes and unbalanced classes (Liu Z. et al., 2011).

Mohammed and Guda (2015) used a consensus-based ensemble of k-NN, SVM, RF, decision stump and Naive Bayes classifier to hierarchically predict enzymes encoded by the human gut microbiome. They further applied their method to analyze the enzyme profiles of lean vs obese and IBD vs non-IBD subjects (Mohammed and Guda, 2015).

Chen et al. (2016) explored the differences between the gut microbiome from three different races (Asian, European and American races), by analyzing the expression levels of their gut microbiome genes. They applied minimum redundancy maximum relevance incremental feature selection methods and four ML methods to determine the most relevant gut microbiome genes that are differentially expressed in individuals from different races. The approaches used were: RF, k-NN, sequential minimal optimization (a type of SVM method where training is performed using the sequential minimal optimization algorithm proposed by Platt (1998), and dagging (a type of meta classifier, where multiple models are built and integrated using majority voting). For performance evaluation, the authors used the overall prediction accuracy and Matthews's correlation coefficient (MCC). MCC was used since it is a suitable performance measure to evaluate model performance even in the case of imbalanced classes (Chicco and Jurman, 2020). Sequential minimal optimization method achieved the best performance results (overall prediction accuracy 99.6%, MCC 99.3%) in identifying 454 most important differentially expressed genes. The obtained results also show that the first 25 out of the 454 identified genes were observed to achieve accuracy greater than 96% and were analyzed in more detail. The identified genes reflected differences among analyzed races such as eating habits, living environments/geographic localization and metabolic levels, which are also known to influence the gut microbiome (Chen et al., 2016).

In more recent work, Zhou and Gallins (2019) evaluated the most commonly used supervised ML methods for microbiome host trait prediction: regression methods, linear discriminant analysis, SVM, similarity matrices and related kernel methods, k-NN, RFs, gradient boosting for decision trees, and neural networks. The authors first performed a comparative analysis based on the literature review of published work, focusing on 17 reported datasets generated from OTU tables. Additionally, the authors performed their own comparative analysis of the mentioned ML methods using three datasets available from MicrobiomeHD database[5] (Duvallet et al., 2017). For feature extraction, the authors applied a hierarchical feature engineering (HFE) (Oudah and Henschel, 2018). Among the compared methods, decision tree-based methods, in general, performed well, achieving similar results with the neural network models in the analyzed published literature. Furthermore, by applying HFE for OTU table feature reduction, better performance results were achieved for almost all of the evaluated methods (Zhou and Gallins, 2019).

Unsupervised Learning Methods

Unsupervised methods identify apparent patterns in the data, without the use of predefined labels. These are important exploratory tools to examine the data and to determine important data structures and correlation patterns (Zhou and Gallins, 2019).

Clustering

Hierarchical clustering is a classic unsupervised learning technique, which builds a hierarchy of nested clusters using a dendrogram, merging or splitting clusters based on different metrics (Zhou and Gallins, 2019). Cai and Sun (2011) used hierarchical clustering for classification of 16S rDNA sequences, they developed ESPRIT-Tree, a hierarchical clustering-based algorithm and demonstrated its utility by performing analysis of millions of 16S rRNA sequences, simultaneously addressing the space and computational issues. The novel algorithm exhibits a quasilinear time and space complexity comparable to greedy heuristic clustering algorithms while achieving a similar accuracy to the standard hierarchical clustering algorithm using 16S rRNA data (Cai and Sun, 2011). In another study, the authors applied hierarchical clustering for establishing possible relations between microbiota and disease-associated host changes, i.e., disease prediction. Here, the authors used as feature transcriptome (RNA-seq) signatures of the host cell (colonocytes), and the 16S rRNA data from gut microbiota. The authors treated colonic epithelial cells with live microbiota from five healthy individuals. Their results show an important role of gut microbiota in regulating host gene expression and suggest that manipulation of microbiome composition could be useful in future therapies (Richards et al., 2019).

Possible correlation between microbiota and disease-associated host changes is done through another microbiome communities clustering algorithm - a novel multivariate testing method called an adaptive Microbiome-based Sum of Powered score (aMiSPU) (Wu et al., 2016). The aMiSPU method is proposed to assess how the compositions of microbiotas are associated with human overall health. Since it is a data-driven approach based on a sum of powered score (SPU) tests and adaptive variable weighting, using a generalized taxon proportion combining microbial abundance information with phylogenetic tree information, it reduces the criticality of the choice of a phylogenetic distance which was a weak point in most previous methods. Most univariate tests depend on strong parametric assumptions on the distributions or mean-variance functional forms for microbiome data which results in a false positive (type I errors). So, some findings are considered significant when they have occurred by chance. As no assumption is imposed, the proposed method - a multivariate semi-parametric test - eliminates the chance of incorrectly

[5]https://github.com/cduvallet/microbiomeHD

rejecting a true null hypothesis that there is no association between any taxa and the outcome of interest. The evaluation of aMiSPU test on simulated and real data indicates that the aMiSPU test is better performing than several competing with well-controlled type I error rates. A by-product of the method is a ranking of the importance of the taxa and be used as a selection tool for the taxa which are likely to be associated with the outcome of interest. The MiSPU R package is public and accessible at https://github.com/ChongWu-Biostat/MiSPU. Its application for understanding the association between microbial communities (i.e., microbiotas) throughout the human body and disease can help in developing personalized medicine.

Biclustering is a powerful data mining technique that allows simultaneously clustering rows and columns of a data matrix to find submatrices that can overlap (Xie et al., 2019). In principle, there exist four categories of biclustering methods: (1) variance minimization methods, (2) two-way clustering methods, (3) motif and pattern recognition methods and (4) probabilistic and generative approaches (Madeira and Oliveira, 2004). For many years, biclustering algorithms have been widely used for the analysis of gene expression data, but new biclustering applications are emerging, such as detecting disease marker genera from gut microbiome as those methods are suitable to detect overlapping clusters on both microbes and hosts. Falony et al. (2016) used biclustering to identify sample subsets with specific taxonomic signatures detecting two stable clusters showing that partially overlapped with previously described enterotypes (Falony et al., 2016). Zhou et al. (2020) proposed an identifiable Bayesian multinomial matrix factorization model to infer overlapping clusters on both microbes and hosts. The authors demonstrate the utility of the proposed approach by comparing four alternative methods in simulations and then by applying it into Qin's IBD microbiome dataset revealing clusters which contain bacteria families that are known to be related to the inflammatory bowel disease and its subtypes according to biological literature (Zhou et al., 2020).

To cluster groups of communities with similar compositions into envirotypes or enterotypes and thus into "metacommunities" the Dirichlet multinomial mixture (DMM) generative modeling framework has been developed (Holmes et al., 2012). It assesses the community structure, including the sample density and size. Multinomial sampling coupled with Dirichlet prior was used before, but the extension of the prior to a mixture of Dirichlet components is a novelty in this work. The method describes each community by a vector, generated by one of finite possible Dirichlet mixture components with different hyperparameters, where each entry is the probability that a read is from given taxa. These vectors of the frequency of taxa occurrences in each sample are placed in a matrix, which is sparse as most species are observed with low abundance. This multinomial sampling is a discrete model that can be used for assessing the size and sparsity of a community. Moreover, it becomes a starting point for a generative modeling framework which explicitly describes a model for generating the studied data, and provides a means to cluster groups of communities with similar compositions. The product of the research is a software package for fitting DMM models which uses a Laplace approximation to integrate out the hyperparameters and estimate the evidence of the complete model. The authors leveraged the methodology to estimate the association of obesity with distinct microbiota by applying the DMM model to human gut microbe genera frequencies from Obese and Lean twins. They did not find a significant impact of body mass on community structure, but rather a possible relation to a disturbed enterotype. They conclude that disturbed states are associated with a more variable community, as this was observed apart from the obese twins, also in people suffering from inflammatory bowel disease (IBD) and ileal Crohn's disease (ICD).

Non-negative matrix factorization (NMF)

This method aims to extract hidden patterns from a series of high-dimensional vectors automatically and has been widely applied in many areas, such as image and natural language processing, and computational biology for dimensional reduction, unsupervised learning (clustering, semi-supervised clustering and co-clustering, etc.) and prediction (Zhang, 2012). The NMF analysis can provide a range of interpretable conclusions about the data sets. For metagenomic data, the features extracted can be mapped to metabolic pathways.

In the work by Cai et al. (2017), the authors use non-negative matrix factorization to identify key features of microbial communities, by analyzing 16S rDNA amplicon and functional data. Using three data sets: the difference in macrolide synthesis pathways for the non-ruminant herbivores; the change in gut and tongue microbial composition for person two in the moving picture data (Caporaso et al., 2011); and the differences in various pathways for the IBD microbiome dataset (Qin et al., 2010) the authors demonstrate how to interpret the features identified by NMF to draw meaningful biological conclusions and discover hitherto unidentified patterns in the data (Cai et al., 2017).

Other ML Methods

Causal inference methods

Causal inference methods provide exploratory data analysis of causal relationships between variables, e.g., relationship between microbial species and disease outcome.

Bayesian networks (BN). BN are probabilistic graphical models consisting of a directed acyclic graph (DAG). In this model, nodes correspond to random variables, and the directed edges correspond to potential conditional dependencies between them. In a recent study, authors constructed a BN model via Augmented Markov Blanket algorithm to identify microbial networks and species-related with the complete response after concurrent chemoradiation in rectal cancer. The BN analysis revealed a link between a specific taxon and an improved therapeutic response (Jang et al., 2020). BN has also been used in combination with other methods, in particular, the Intervention calculus when the DAG is absent (IDA) method (Kharrat et al., 2019), to identify microbial species that are likely to have a causal role in colorectal cancer (CRC) risk and onset.

Dynamic Bayesian networks (DBNs). Dynamic Bayesian Networks (DBNs) are BNs attested for modeling relationships over temporal data. In this regard, a DBN is a directed acyclic

graph where, at each time slice or instance, nodes correspond to random variables of interest and directed edges correspond to their conditional dependencies in the graph (Russell and Norvig, 2016). DNB has been used for analyzing longitudinal microbiome data sets to establish temporal relationships between different taxonomic ranks and other clinical factors that affect the microbiome (Lugo-Martinez et al., 2019). They studied longitudinal data sets from three human microbiome body sites: infant gut, vagina, and oral cavity, and use temporal alignments to normalize the differences in the progress of biological processes of each subject, they found that microbiome alignments improve the predictive performance of the methodology over previous studies of longitudinal datasets, and increase the ability to infer new and previously reported biological and environmental relationships between the components of the microbiome and other factors that influence it, this methodology allows to predict microbiome states and relationships based on longitudinal data applying DBN. Moreover, authors build up the CGBayesNets package that is freely available under the MIT Open Source license agreement.

In general, time series analyses represent a valuable approach to determine the resilience and variability of microbial communities. Perturbations and changing environmental conditions can drive communities into alternative stable states, while bi- and multi-stable states are mostly induced by member interactions within a microbial community. However, a detailed exploration of these temporal shifts is often restricted by either intensively sampled but small treatment groups or large studies, including only few sampling time points. Faust et al. (2015) compared twelve-time series analysis techniques used for high-throughput sequencing studies. These techniques mostly operate on cross-correlation, autocorrelation or network inference. Although the sampling scheme is highly dependent on the environment of interest, appropriate sampling frequency and regularity are crucial. These parameters define the resolution, completeness, sparsity, and noisiness of the data and potentially limit the explanatory power of the analysis output. By applying DBN techniques, incomplete data may be amended and used to model dependencies in time series. Apart from that, the identification of early warning signs indicating an upcoming change in microbiome-inherent networks could help to predict responses to environmental factors (Faust et al., 2015).

Mendelian randomization (MR). Mendelian randomization (MR) has been used to understand the causal role of gut microbiome in disease. MR uses human genetic variants, such as single nucleotide polymorphisms (SNP), as proxy measures for clinically relevant traits of interest (e.g., gut microbiome) to

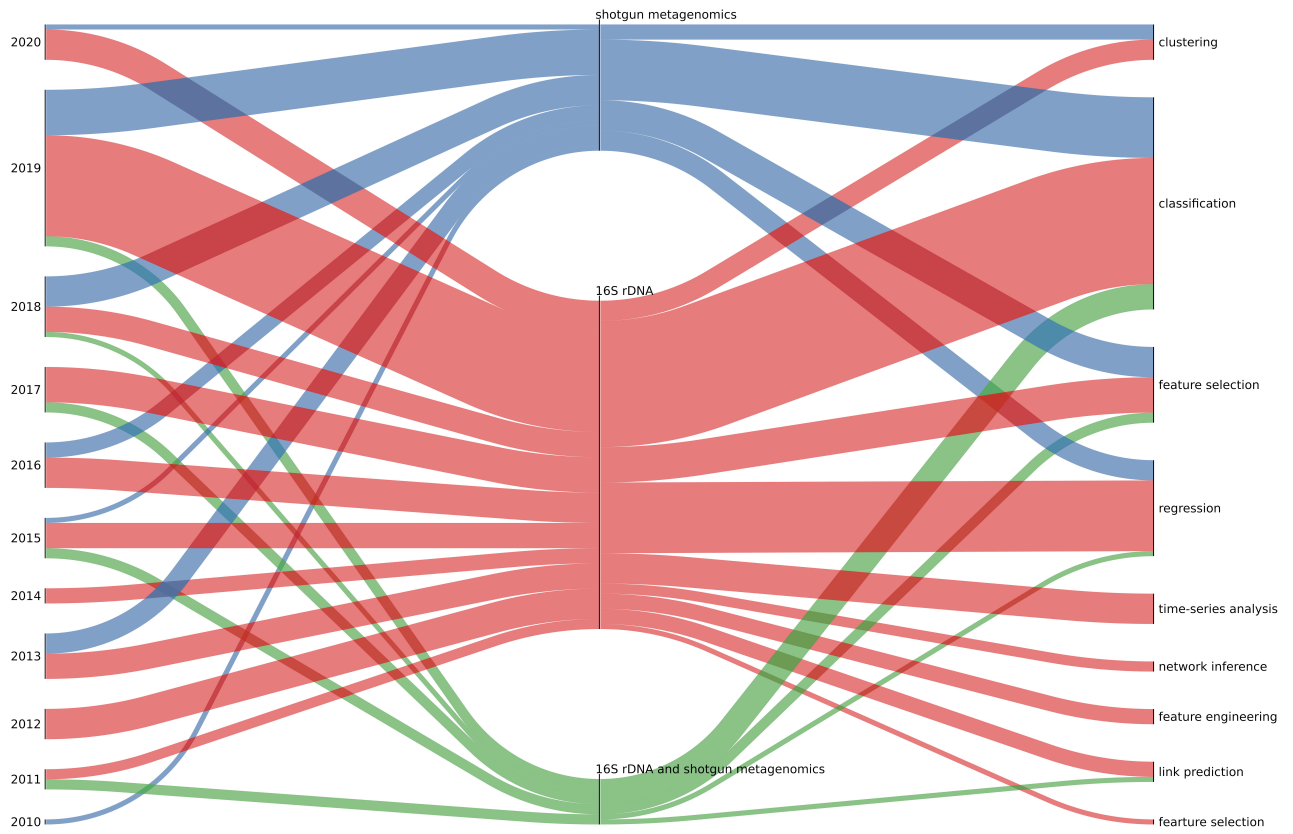


FIGURE 2 | Plot summarizing reviewed articles that apply machine learning in human microbiome data analysis. Articles are summarized based on microbiome input data type and broadly defined ML categories and constrained by year. Please note that in the case of the year 2020 the input does not cover all publications from this year.

estimate the causal relationship between a trait and a disease or health outcome, therefore eliminating confounding and reverse causation effects between the exposure of interest and outcome. In a bidirectional MR analysis on over 3800 individuals from the Flemish Gut Flora Project and two German cohorts, Hughes and co-workers (Hughes et al., 2020) were able to estimate relationships among five microbial traits and seven outcomes, namely waist circumference and body mass index.

Also, Sanna et al. (2019) used bidirectional MR to assess the causal role of the gut microbiome on metabolic traits, based on genome-wide genetic information, gut metagenomic sequence and fecal short-chain fatty acid (SCFA) levels from 952 normoglycemic individuals, combined with genome-wide-association summary statistics for 17 metabolic and anthropometric traits. The authors found a causal role of gut-produced fecal SCFA with respect to energy balance and glucose homeostasis. In particular, a genetically influenced shift in the gut microbiome toward increased production of butyrate with beneficial effects on beta-cell function, and host genetic variation resulting in increased fecal propionate levels affecting type 2 diabetes risk (Sanna et al., 2019).

Correlation-based network analysis. Seo et al. (2017) studied which of the gut microbes responded to probiotic intervention, and their association with gastrointestinal symptoms in healthy adult humans. The study consisted of 21 individuals after probiotics consumption for 60 days and evaluated the changes in microbiome composition through 16S rRNA amplicon

FIGURE 3 | Plot based on Wordcloud with MESH (Medical Subject Headings) terms annotated from the 89 articles.

TABLE 2 | Clinical Applications of Machine Learning for human microbiome studies.

Disease	Datasets	Features	Aim	Method	Citation
Crohn's Disease (CD)	BISCUIT cohort (Hansen et al., 2012; Pascal et al., 2017), CD *n* = 20, Controls *n* = 20, Validation Cohort RISK cohort (Gevers et al., 2014).	Shotgun metagenomics data and 16S rRNA gene data.	Classify pediatric CD patients by disease state and treatment response.	Random forest.	Douglas et al., 2018
Colorectal Cancer (CRC)	Patients with CRC S0 *n* = 27, Patients with CRC SIII/IV *n* = 54. Healthycontrols *n* = 127.	Shotgun metagenomics data (Species, KO genes, Metabolite profiles).	Classification of CRC patients according to cancer stage.	Feature selection by LASSO. Random forest.	Yachida et al., 2019
Colorectal Cancer (CRC)	Fecal CRC metagenomes: *n* = 38 previously published, *n* = 22 new. Control *n* = 60.	Feature selection by LASSO. Features: IGC gene abundances.	Predict taxonomic and functional microbiome CRC signatures.	Feature selection by LASSO.: Random forest.	Wirbel et al., 2019
Colorectal cancer (CRC)	Stool: Controls *n* = 62, CRC *n* = 69, Polyps *n* = 23. Swabs: Controls *n* = 25, CRC *n* = 45, Polyps *n* = 21.	Log-ratio transformed values of OTUs present in at least 5% of individuals.	Development of an oral and fecal microbiota classifier that distinguish individuals with CRC and adenomas from controls.	Feature selection by LASSO. Random forest.	Flemer et al., 2017
Colorectal cancer (CRC)	Cohort 1: CRC *n* = 29, adenomas *n* = 27, controls = 24. Cohort 2: CRC *n* = 32, Control = 28. Validation Datasets: CRC *n* = 313, Adenomas *n* = 143, Controls = 308.	Taxonomic species-level abundances, gene-family and pathways related abundances.	Finding of reproducible microbiome markers and disease-predictive models for CRC.	Supervised Learning Methods: Random forest.	Thomas et al., 2019
Colorectal cancer (CRC)	Previously published data from France, Hong Kong and Austria. France (Zeller et al., 2014).	Shotgun metagenomics, FASTA.	Discovery of biomarkers from WGS that could be used to build a machine learning classifier for CRC prediction.	Supervised Learning Methods: Random forest. Neural network. Support vector machine.	Koohi-Moghadam et al., 2019
Abnormal cases vs. Controls	Controls *n* = 383. Abnormal Cases: Type 2 diabetes *n* = 170, Rheumatoid arthritis *n* = 130, Liver cirrhosis *n* = 123.	Shotgun metagenomics.	Develop a pipeline to address the challenging characterization of multilabel samples from type 2 diabetes, rheumatoid arthritis, and liver cirrhosis.	Logistic Regression.	Wu et al., 2018)
Bacterial Vaginosis (BV)	Dataset 1: Asymptomatic BV-:299. Asymptomatic BV + :97 Dataset 2: Asymptomatic BV-:6. Asymptomatic BV + :214.	OTU tables from 16S rRNA gene data.	Establishing microbial signatures in bacterial vaginosis (BV).	Logistic Regression, Genetic Programming, and Random Forest.	Beck and Foster, 2015
Colorectal cancer (CRC)	*n* = 30 Controls *n* = 30 CRC patients from Previously published datasets from Austria (*n* = 57 health ycontrols, *n* = 46 CRC patients) and China (*n* = 53 healthy controls and 75 CRC patients) (Feng et al., 2015; Purcell et al., 2017).	Shotgun Metagenomics data (mOTU, MGS, Methaphlan species) Gene counts.	Identify cohort-specific non-invasive biomarkers to be used in diagnosis of CRC.	Weka "CfsSubsetEval" + Boruta algorithm for feature selection. RF with 33 genes and 20 taxonomic markers.	Gupta et al., 2019
Obesity	Data from 10 previously published studies (*n* = 2.786 subjects) (Turnbaugh et al., 2006; Wu et al., 2011; Human Microbiome Project Consortium, 2012; Zupancic et al., 2012; Escobar et al., 2014; Goodrich et al., 2014; Schubert et al., 2014; Ross et al., 2015; Zeevi et al., 2015; Baxter et al., 2016).	OTU tables from 16S rRNA gene data.	Predict obesity status on the basis of the microbial composition of the microbiome.	Random Forest.	Sze and Schloss, 2016
Pediatric irritable bowel syndrome (IBS)	*n* = 23 IBS patients *n* = 22 Healthy Controls.	Shotgun metagenomics, Gene Counts and pathways, Metabolomics.	Evaluate the relationship between pediatric IBS and abdominal pain with intestinal microbes and fecal metabolites.	RF LASSO feature selection SVM naïve Bayes.	Hollister et al., 2019

(Continued)

TABLE 2 | Continued

Disease	Datasets	Features	Aim	Method	Citation
Gastrointestinal symptoms in healthy humans	*n* = 21 volunteers after probiotics consumption for 60 days.	16S rRNA gene data.	Establish which of the gut microbes respond to probiotics interventions.	Correlation-based network analysis. Dimensionality reduction.	Seo et al., 2017
Chron's Disease	Chron's Disease dataset: *n* = 731 Pediatric patients with CD *n* = 628 Non-CD. *n* = 300 healthy controls from HMP (Turnbaugh et al., 2007).	16S rRNA gene data.	Use of deep learning methods and classic machine learning approaches for distinguishing among human body sites, diagnosis of Crohn's disease, and predicting the environments from representative 16S gene sequences.	RF, SVM, Deep Learning.	Asgari et al., 2019
Inflammatory Bowel Disease (IBD) and esophagus diseases	*n* = 3501 samples from different datasets (Costello et al., 2009; Knights et al., 2011).	16S rRNA gene data.	Classification of metagenomic data using Neural Networks approaches.	Neural Networks. Comparison with supervised ML methods (Linear regression, Boosting gradients, SVM, RF).	Lo and Marculescu, 2019
Islet autoimmunity (IA) and Type 1 Diabetes (T1D).	*n* = 10,913 metagenomes in stool samples from persistent confirmed IA or T1D vs controls. (TEDDY cohort) (Hagopian et al., 2011).	Shotgun metagenomics. Gene count.	Describe the functional profile of the developing gut microbiome in relation to islet autoimmunity, T1D and other early childhood events.	RF to separate between case-controls.	Vatanen et al., 2018
Irritable Bowel Syndrome	71 samples from 22 children with IBS (pediatric Rome III criteria) and 22 healthy children.	16S rRNA gene data.	Finding microbial signatures for Irritable Bowel Syndrome.	Random Forest.	Saulnier et al., 2011
Sclerosing cholangitis	46 controls and 80 patients with PSC during ERC (37 with early disease, 32 with advanced disease, and 11 with biliary dysplasia).	16S rRNA gene data.	Explore the microbial involvement in the etiopathogenesis and risk for development of biliary neoplasia in primary sclerosing cholangitis.	Generalized linear models.	Pereira et al., 2017
Allergy	Skin microbiota samples from 118 individuals.	16S rRNA gene data.	Analyzing atopic sensitization (i.e., allergic disposition) in a random sample of adolescents.	Linear and logistic regression, and PCA.	Hanski et al., 2012
Liver disease	FINRISK population cohort (Borodulin et al., 2018).	Shallow shotgun metagenome sequencing.	Study the link between the Fatty Liver Index (FLI) and gut microbiome composition in a population sample in Finland.	Gradient boosting.	Ruuskanen et al., 2020
Liver disease	A large population-based cohort ($N \geq 7{,}115$) and ∼15 years of electronic health register follow-up of the FINRISK population cohort (Borodulin et al., 2018).	Shallow shotgun metagenome sequencing.	Investigate the predictive ability of gut microbial markers in conjunction with conventional risk factors, for incident liver disease and alcoholic liver disease.	Gradient boosting.	Liu et al., 2020
Serum lipids	Healthy Finnish adults (*n* = 25, 18 females, 7 males).	16S rRNA gene data.	Evaluate the association between the gut microbiome and lipid profile.	Linear models, unsupervised hierarchical clustering.	Lahti et al., 2013
IBD (Crohn's disease, Ulcerative Colitis, collagenous colitis) vs healthy	Three publicly available human metagenomics data sets as Use Cases (Turnbaugh et al., 2009; Koren et al., 2013; Halfvarson et al., 2017).	OTU tables.	Predicting gut microbiome functional role.	Supervised Learning method comparison.	Wassan et al., 2018a
Obesity	267 children aged 7–18 years from the American Gut Project (McDonald et al.).	16S rRNA gene data.	Composition of gut microbiota and its associations with BMI level, weight change and lifestyle.	Linear decomposition model.	Bai et al., 2019
Postmortem Changes	144 sample swabs were from 21 cadavers.	16S rRNA gene data.	Use of necrobiome data in the prediction of the Postmortem interval.	Regression.	Johnson et al., 2016

sequencing. They used correlation-based network analysis and dimensionality reduction to assess the effect of probiotics consumption and found that probiotic intervention reduced the abundance of potential bacteria such as *Citrobacter* and *Klebsiella* spp. in the human gut microbial community. Moreover, they found that probiotic intervention may reduce the flatulence through downregulation of *Methanobrevibacter* spp. abundance (Seo et al., 2017).

Biomedical Applications of ML Techniques in Human Microbiome Analyses

Figure 2 summarizes reviewed papers based on the input data type and ML method type. The most dominant input data type in the case application of ML methods for human microbiome analysis has been 16S rRNA amplicon-based sequencing data either in the form of OTU or ASV tables while usage of shotgun metagenomes has increased during recent years. There are a small number of studies that have tested ML methods on both amplicon-based and shotgun datasets. Most often applied ML methods have been feature classification, selection and regression. Most often different ensemble learning methods have been applied while deep learning has been used in few cases. The number of yearly published papers using ML for microbiome data analysis has been slightly growing during years 2011–2018 and increased more than twice in 2019 compared to the previous year (**Supplementary Figure 2**).

The application of DP to human microbiome analysis is not well captured by our dataset as its application for human microbiome analysis is an emerging field. Recent example includes disease state prediction (inflammatory bowel disease, type 2 diabetes, liver cirrhosis, obesity) using deep representation learning framework that deploys various autoencoders to learn robust low-dimensional representations from high-dimensional microbiome profiles and trains classification models based on the learned representation (Oh and Zhang, 2020) or that relate key microbial biomarkers with metabolite biomarkers in gut microbiome (Le et al., 2020).

Our results indicate that the biomedical application of ML for analyses of human microbiome datasets has been mainly focused on the characterization of differently abundant microbial groups between different body sites and the effect of diet on microbiome composition and dynamics. The gut microbiome datasets have been extensively used to stratify and classify patients according to symptoms or characteristics to assist in the diagnosis

TABLE 3 | Available Resources for applying ML to human microbiome studies.

Tool Name	Description	References
Feature Selection with the R Package MXM	Includes several feature selection algorithms. In particular, the Statistically Equivalent Signatures (SES) algorithm that is very suitable for microbiome data because it scales up to high dimensions and requires few samples. It also reports "multiple biosignatures" meaning multiple, minimal-size subsets of features that lead to an equally predictive model. A more recent feature selection algorithm that scales up well to high dimensional data called Forward-Backward Selection with Early Dropping (FBED) also implemented in the MXM R package; It is preferable to SES when the sample size is higher.	Lagani et al., 2017; Borboudakis and Tsamardinos, 2019
Automated Machine Learning (AutoML) with JADBio.	End-to-end AutoML tool designed to deliver predictive and diagnostic models to non-experts while drastically increasing the productivity of expert analysts. Several qualifications make JADbio (www.jadbio.com) very suitable for microbiome data analysis. First, it accepts numerical measurements (e.g., abundance tables), as well as discrete predictors (e.g. experimental factors and curated metadata), and incomplete datasets with missing values. Second, it facilitates a novel out-of-sample bootstrapping protocol able to provide accurate, non-optimistic estimates of predictive performance even in cases of low sample sizes (e.g., 40) and hundreds of thousands of features Finally. It uses SES and FBED to return the corresponding *biosignatures.* This allows the creation of predictive models that are equally good up to statistical equivalence, thus, providing the researcher with choices when designing new cost-benefit diagnostic assays.	Tsamardinos et al., 2018, 2020
Microbiome network inference with SCENERY.	SCENERY is a free online application that allows users to perform several network learning tasks (scenery.csd.uoc.gr). It is the first of its kind to facilitate advanced algorithms for the inference of association networks, probabilistic causal networks and Bayesian networks. The qualifications of SCENERY have been successfully shown on the single-cell cytometry domain. At the moment, SCENERY does not treat missing values or compositionality, yet, it is readily applicable to the microbiome data domain for inferring causal or non-causal networks of microbiome molecules and species.	Papoutsoglou et al., 2017
The Microbiome Modeling Toolbox	Comprehensive toolbox to model (i) microbe-microbe and host-microbe metabolic interactions, and (ii) microbial communities using microbial genome-scale metabolic reconstructions and metagenomic data.	Baldini et al., 2019
Constraint-based reconstruction and analysis (COBRA) Toolbox v.3.0.	Software suite for quantitative prediction of cellular and multicellular biochemical networks with constraint-based modeling.	Heirendt et al., 2019
Reconstruction, Analysis and Visualization of Metabolic Networks (RAVEN).	RAVEN is a commonly used MATLAB toolbox for genome-scale metabolic model reconstruction, curation and constraint-based modeling and simulation.	Wang et al., 2018
Fizzy: feature subset selection for metagenomics	Python command line tool compatible with BIOM format, for microbial ecologists that implements information-theoretic subset selection methods for biological data formats.	Ditzler et al., 2015; http://github.com/EESI/Fizzy.

TABLE 4 | Common problems in machine-learning analyses.

Problem type	Problem description
Not cross-validating the feature selection step	Perhaps the most common pitfall of performance estimation is that of performing feature selection on the complete, labeled dataset (e.g. by differential expression) and subsequently cross-validating only the modeling algorithm on the same data (Hastie et al., 2009). The same account for any other step on the pipeline that peeks at the labels or the outcome to predict. In the case of large sample and balanced datasets the overestimation should be unnoticeable. On small sample or imbalanced datasets, however, overestimation can become quite significant (Tsamardinos et al., 2020). An analyst should cross-validate all steps of the analysis as atoms, including the preprocessing, imputation, feature selection, and modeling to obtain accurate estimates of performance.
Not correcting for winner's curse	A second common error is reporting the cross-validation predictive performance of the winning algorithm or ML pipeline as the final performance estimate. For example, an analyst may try 1000 combinations of different algorithms for each step of the analysis with various values for their hyper-parameters and find that the winning combination has a cross-validated accuracy of 80%. This estimate is on average, overestimated because of the "winner's curse" (Ioannidis, 2008). The overestimation due to the winner's curse is again large in small or imbalanced datasets. It is not uncommon to find 0.7 AUC when the true one equals random guessing (0.5 AUC) due to the winner's curse. Other estimation protocols need to be applied in these cases. The simplest solution is to withhold a separate test set to estimate the performance of the winning model; unfortunately, this technique loses samples to estimation and cannot be applied when samples are scarce. Techniques that remove the winner's curse in small samples are the nested cross-validation and the bootstrap bias-corrected CV (Tsamardinos et al., 2018).
Not stratifying the split to folds	Another typical error occurs when randomly splitting the available samples, either for creating an external validation dataset, or to perform cross-validation, without accounting the class imbalance and sample dependency. The partitioning should be stratified, i.e., the class distribution should be maintained in the folds. When the classes are imbalanced, sample stratification leads to improved performance estimations (Tsamardinos et al., 2015).
Not handling repeated measurements	When sampling is correlated, e.g., the same subject is measured repeatedly, care needs to be exercised. Treating samples as identically and independently distributed (i.i.d.) as cross-validation assumes, provides overestimated performance estimations. When samples are grouped in repeated measurements, one should take care to assign all samples in the group in the same fold. This way, they all belong in the train set or the test set during cross-validation and never in both.
Splitting data inappropriately	When building ML models, typically data is broken into training and test sets. The training set is used to teach the model, and the model's performance is evaluated by how well it describes the test set. Researchers typically split the data at random that may not be the correct approach always. The "right" way to split data might not be obvious, but careful consideration and trying several approaches may give more insight (Riley, 2019).

and management of diseases with a preference on those related with gut microbiome, due to easy accessibility for obtaining fecal samples, such as inflammatory bowel diseases, obesity and colorectal neoplasms (see **Figure 3**). A list of selected studies on the application of machine learning to human microbiome data in biomedical research is presented in **Table 2**.

However, it should be noted that many of the reviewed papers are focused on the comparison of the performance of different ML methods, developing workflows or creating new ML approaches considering the technical aspects of ML related to the nature and complexity of the microbiome data, but without a clear biological or clinical question behind to solve. A detailed analysis of the dataset obtained showed that 20 of 89 papers used their own unique datasets, while the rest of publications made repetitive and intensive use of a limited number of datasets to develop ML solutions, like the Human Microbiome Project widely used for microbiome body composition studies. Besides, we identified 9 papers related to the development of ML methods for microbiome longitudinal analysis that are mainly based on the reuse of five datasets (Caporaso et al., 2011; Gajer et al., 2012; David et al., 2014; La Rosa et al., 2014; DiGiulio et al., 2015) with Gajer et al. being reused in four of them. In addition, we need to highlight the limited sample size in many of the studies what compromises the applicability and the conclusions of the ML methods reviewed.

Table 3 summarizes the main available resources for applying different ML methods to human microbiome studies. Most of the reviewed studies have applied ML methods incorporated in general data analysis packages. As stated by Moreno-Indias et al. (2021). it is important to foster the development of user-friendly ML-based tools for translational and clinical personnel. This process is strongly dependent on open-source software ecosystems as application of ML in microbiome data analysis is rapidly evolving field and involves high degree of multidisciplinary.

Building prediction models for the analysis of microbiome or similar biological data often requires the design of an ML pipeline in which different algorithms for data preprocessing, imputation, feature selection, and modeling are combined along with their hyper-parameter values. The implementation of such a complex modeling strategy could be tedious and requires substantial human resources to optimize. Most importantly, however, this process is prone to serious methodological errors that lead to models whose training performance estimates are inflated (overestimated) and, thus, fail to generalize on external validation datasets. Some common pitfalls of ML application are listed in **Table 4**.

CONCLUSION

Human microbiome research has received increasing interest during recent years, mainly due to the large potential applicability of metagenomics data from human microbiome studies in personalized medicine. International and interdisciplinary efforts have made possible to collect large volumes of microbiome data, facilitating the development and implementation of different ML methods. Here we reviewed the different ML methods developed and applied to human microbiome data analysis for an insight of the development in the field with their

achievements and pitfalls. Although the data presented here is mostly centered on the analysis of bacterial community, many principles reviewed could be applied in general, regardless of the microbiome feature type. The advantages of ML techniques over classical statistical models are to infer relationships between variables for automatic pattern discovery and handling with multi-dimensional data. Therefore, these methods have been widely used for classification, biomarker identification, gene prediction or association studies in human microbiome research. Based on the performed review, most common machine learning algorithms that were used for microbiome analysis were Random Forest, Support Vector Machines, Logistic Regression and k-NN. Since there are several factors that need to be considered during the selection of the ML algorithm (i.e., number of features, number of observations, data quality, data type etc.), it is recommended to apply and evaluate more than one method and select the one with the best performance. However, other ML applications that will be of high interest in the near future are underrepresented like deep learning, spatiotemporal and dynamic modeling, methods for longitudinal and mechanistic analyses or integrative methods for data from different sources to understand microbiome-host interaction and diseases. Nevertheless, the full deployment of ML techniques in human microbiome studies for a complete application and integration in the personalized medicine field requires further efforts. Personalized medicine requires a deep understanding of features characterizing individual particularities and responses a frequent lack of ML methods. ML models with high complexity often come with a loss of interpretability running as black boxes. In many cases, ML methods fail to provide easily, understandable and interpretable predictions essential to identify mistakes or biases in the input data when the model is trained. Moreover, ML methods introduced in this review require fine-tuning of many hyper-parameters to achieve optimal results being a time-consuming task given the high number of possible alternatives. In addition, for training powerful ML methods with reliable results a large amount of data and a lot of computing resources are required. In general, ML methods introduced in this review are based on datasets with a limited number of cases and without other independent datasets what conditions their results and applicability. Therefore, from our review perspective future efforts in the field should be focused in (1) create standards (incl data pre-processing) for the development and deployment of ML techniques with an easy, transparent, and trustable interpretability for non-experts taking in account the peculiarities of microbiome data; (2) increase the number and quality of human microbiome studies; (3) create efficient data structures and ML repositories following Findable, Accessible, Interoperable and Reusable (FAIR) principles and (4) build bridges between different disciplines, microbiology, biology, statistics, bioinformatics, engineering and others to increase interdisciplinary for innovative solutions. COST Action CA18131 on *Statistical and Machine Learning Techniques in Human Microbiome Studies* (ML4Microbiome) is highly committed to pursuit these objectives in collaboration with the international community and extended discussions on contemporary challenges and proposed solutions are addressed by the ML4Microbiome consortium in Moreno-Indias et al. (2021).

AUTHOR CONTRIBUTIONS

MC, EC, IM-I, and JT conceived the review. LM-Z and JT coordinated, supervised and wrote the draft, the **Supplementary Information** and the final manuscript. KK-H, TL, PP, VT, and EC performed the analysis, prepared the figures and **Supplementary Information**, wrote the draft and the final manuscript. OA, MB, MC, AG, JH, KH, TK, MK, LL, ML, VM, IM-I, IN, EO, IP, GP, RS, BS, BV, EZ, IT, and MY revised draft manuscript, provided comments, included manual references and wrote parts of the final manuscript. All the authors discussed and approved the final version of the manuscript.

FUNDING

This study was supported by COST Action CA18131 "Statistical and machine learning techniques in human microbiome studies". Estonian Research Council grant PRG548 (JT). Spanish State Research Agency Juan de la Cierva Grant IJC2019-042188-I (LM-Z). EO was founded and OA was supported by Estonian Research Council grant PUT 1371 and EMBO Installation grant 3573. AG was supported by Statutory Research project of the Department of Computer Networks and Systems.

ACKNOWLEDGMENTS

The authors are grateful to all COST Action CA18131 "Statistical and machine learning techniques in human microbiome studies" members for their contribution in discussion about evaluation process of ML methods currently used in microbiome research during action workshops.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2021.634511/full#supplementary-material

REFERENCES

Ai, D., Pan, H., Han, R., Li, X., Liu, G., and Xia, L. C. (2019). Using decision tree aggregation with random forest model to identify gut microbes associated with colorectal cancer. *Genes* 10:112. doi: 10.3390/genes10020112

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. London: Chapman and Hall.

Almeida, A., Nayfach, S., Boland, M., and Strozzi, F. (2021). A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.* 39, 105–114 doi: 10.1038/s41587-020-0603-3

Arango-Argoty, G., Garner, E., Pruden, A., Heath, L. S., Vikesland, P., and Zhang, L. (2018). DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome* 6:23.

Arksey, H., and O'Malley, L. (2005). Scoping studies: towards a methodological framework. *Int. J. Soc. Res. Methodol.* 8, 19–32. doi: 10.1080/1364557032000119616

Asgari, E., Garakani, K., McHardy, A. C., and Mofrad, M. R. K. (2019). MicroPheno: predicting environments and host phenotypes from 16S rRNA gene sequencing using a k-mer based representation of shallow sub-samples. *Bioinformatics* 35:1082. doi: 10.1093/bioinformatics/bty652

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat. Genet.* 25, 25–29. doi: 10.1038/75556

Bai, J., Hu, Y., and Bruner, D. W. (2019). Composition of gut microbiota and its association with body mass index and lifestyle factors in a cohort of 7-18 years old children from the American Gut Project. *Pediatr. Obes.* 14:e12480. doi: 10.1111/ijpo.12480

Baldini, F., Heinken, A., Heirendt, L., Magnusdottir, S., Fleming, R. M. T., and Thiele, I. (2019). The Microbiome Modeling Toolbox: from microbial interactions to personalized microbial communities. *Bioinformatics* 35, 2332–2334. doi: 10.1093/bioinformatics/bty941

Baxter, N. T., Ruffin, M. T., Rogers, M. A., and Schloss, P. D. (2016). Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. *Genome Med.* 8:37. doi: 10.1186/s13073-016-0290-3

Beck, D., and Foster, J. A. (2014). Machine learning techniques accurately classify microbial communities by bacterial vaginosis characteristics. *PLoS One* 9:e87830. doi: 10.1371/journal.pone.0087830

Beck, D., and Foster, J. A. (2015). Machine learning classifiers provide insight into the relationship between microbial communities and bacterial vaginosis. *Biodata Min.* 8:23. doi: 10.1186/s13040-015-0055-3

Berglund, F., Marathe, N. P., Österlund, T., Bengtsson-Palme, J., Kotsakis, S., Flach, C.-F., et al. (2017). Identification of 76 novel B1 metallo-β-lactamases through large-scale screening of genomic and metagenomic data. *Microbiome* 5:134. doi: 10.1186/s40168-017-0353-8

Blaxter, M., Mann, J., Chapman, T., Thomas, F., Whitton, C., Floyd, R., et al. (2005). Defining operational taxonomic units using DNA barcode data. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 360, 1935–1943. doi: 10.1098/rstb.2005.1725

Bonder, M. J., Abeln, S., Zaura, E., and Brandt, B. W. (2012). Comparing clustering and pre-processing in taxonomy analysis. *Bioinformatics* 28, 2891–2897. doi: 10.1093/bioinformatics/bts552

Borboudakis, G., and Tsamardinos, I. (2019). Forward-backward selection with early dropping. *J. Mach. Learn. Res.* 20, 276–314.

Borodulin, K., Tolonen, H., Jousilahti, P., Jula, A., Juolevi, A., Koskinen, S., et al. (2018). Cohort profile: the national FINRISK STUDY. *Int. J. Epidemiol.* 47, 696i–696i. doi: 10.1093/ije/dyx239

Braun, T., Di Segni, A., BenShoshan, M., Neuman, S., Levhar, N., Bubis, M., et al. (2019). Individualized dynamics in the gut microbiota precede Crohn's disease flares. *Am. J. Gastroenterol.* 114, 1142–1151. doi: 10.14309/ajg.0000000000000136

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Cai, Y., Gu, H., and Kenney, T. (2017). Learning microbial community structures with supervised and unsupervised non-negative matrix factorization. *Microbiome* 5:110. doi: 10.1186/s40168-017-0323-1

Cai, Y., and Sun, Y. (2011). ESPRIT-Tree: hierarchical clustering analysis of millions of 16S rRNA pyrosequences in quasilinear computational time. *Nucleic Acids Res.* 39:e95. doi: 10.1093/nar/gkr349

Callahan, B. J., McMurdie, P. J., and Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J.* 11, 2639–2643. doi: 10.1038/ismej.2017.119

Caporaso, J. G., Lauber, C. L., Costello, E. K., Berg-Lyons, D., Gonzalez, A., Stombaugh, J., et al. (2011). Moving pictures of the human microbiome. *Genome Biol.* 12:R50. doi: 10.1186/gb-2011-12-5-r50

Chassagnon, G., Vakalopolou, M., Paragios, N., and Revel, M.-P. (2020). Deep learning: definition and perspectives for thoracic imaging. *Eur. Radiol.* 30, 2021–2030. doi: 10.1007/s00330-019-06564-3

Chen, L., Zhang, Y. H., Huang, T., and Cai, Y. D. (2016). Gene expression profiling gut microbiota in different races of humans. *Sci. Rep.* 6:23075. doi: 10.1038/srep23075

Chicco, D., and Jurman, G. (2020). The advantages of the matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genom.* 21:6. doi: 10.1186/s12864-019-6413-7

Chong, J., Liu, P., Zhou, G., and Xia, J. (2020). Using microbiomeanalyst for comprehensive statistical, functional, and meta-analysis of microbiome data. *Nat. Protoc.* 15, 799–821. doi: 10.1038/s41596-019-0264-1

Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.* 20, 273–297. doi: 10.1007/BF00994018

Costello, E. K., Lauber, C. L., Hamady, M., Fierer, N., Gordon, J. I., and Knight, R. (2009). Bacterial community variation in human body habitats across space and time. *Science* 326, 1694–1697. doi: 10.1126/science.1177486

Cui, H., and Zhang, X. (2013). Alignment-free supervised classification of metagenomes by recursive SVM. *BMC Genom.* 14:641. doi: 10.1186/1471-2164-14-641

David, L. A., Materna, A. C., Friedman, J., Campos-Baptista, M. I., Blackburn, M. C., Perrotta, A., et al. (2014). Host lifestyle affects human microbiota on daily timescales. *Genome Biol.* 15:R89. doi: 10.1186/gb-2014-15-7-r89

Díez López, C., Vidaki, A., Ralf, A., Montiel González, D., Radjabzadeh, D., Kraaij, R., et al. (2019). Novel taxonomy-independent deep learning microbiome approach allows for accurate classification of different forensically relevant human epithelial materials. *Forensic Sci. Int. Genet.* 41, 72–82. doi: 10.1016/j.fsigen.2019.03.015

DiGiulio, D. B., Callahan, B. J., McMurdie, P. J., Costello, E. K., Lyell, D. J., Robaczewska, A., et al. (2015). Temporal and spatial variation of the human microbiota during pregnancy. *Proc. Natl. Acad. Sci. U.S.A.* 112, 11060–11065. doi: 10.1073/pnas.1502875112

Ditzler, G., Morrison, J. C., Lan, Y., and Rosen, G. L. (2015). Fizzy: feature subset selection for metagenomics. *BMC Bioinform.* 16:358. doi: 10.1186/s12859-015-0793-8

Douglas, G. M., Hansen, R., Jones, C. M. A., Dunn, K. A., Comeau, A. M., Bielawski, J. P., et al. (2018). Multi-omics differentially classify disease state and treatment outcome in pediatric Crohn's disease. *Microbiome* 6:13. doi: 10.1186/s40168-018-0398-3

Duvallet, C., Gibbons, S. M., Gurry, T., Irizarry, R. A., and Alm, E. J. (2017). Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nat. Commun.* 8:1784. doi: 10.1038/s41467-017-01973-8

Eck, A., Zintgraf, L. M., de Groot, E. F. J., de Meij, T. G. J., Cohen, T. S., Savelkoul, P. H. M., et al. (2017). Interpretation of microbiota-based diagnostics by explaining individual classifier decisions. *BMC Bioinform.* 18:441. doi: 10.1186/s12859-017-1843-1

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26, 2460–2461. doi: 10.1093/bioinformatics/btq461

Elekwachi, C. O., Wang, Z., Wu, X., Rabee, A., and Forster, R. J. (2017). Total rRNA-Seq analysis gives insight into bacterial, fungal, protozoal and archaeal communities in the rumen using an optimized rna isolation method. *Front. Microbiol.* 8:1814. doi: 10.3389/fmicb.2017.01814

Escobar, J. S., Klotz, B., Valdes, B. E., and Agudelo, G. M. (2014). The gut microbiota of colombians differs from that of Americans, Europeans and Asians. *BMC Microbiol.* 14:311. doi: 10.1186/s12866-014-0311-6

Fabijanić, M., and Vlahoviček, K. (2016). Big data, evolution, and metagenomes: predicting disease from gut microbiota codon usage profiles. *Methods Mol. Biol.* 1415, 509–531. doi: 10.1007/978-1-4939-3572-7_26

Falony, G., Joossens, M., Vieira-Silva, S., Wang, J., Darzi, Y., Faust, K., et al. (2016). Population-level analysis of gut microbiome variation. *Science* 352, 560–564. doi: 10.1126/science.aad3503

Faust, K., Lahti, L., Gonze, D., de Vos, W. M., and Raes, J. (2015). Metagenomics meets time series analysis: unraveling microbial community dynamics. *Curr. Opin. Microbiol.* 25, 56–66. doi: 10.1016/j.mib.2015.04.004

Faust, K., Sathirapongsasuti, J. F., Izard, J., Segata, N., Gevers, D., Raes, J., et al. (2012). Microbial co-occurrence relationships in the human microbiome. *PLoS Comput. Biol.* 8:e1002606. doi: 10.1371/journal.pcbi.1002606

Feng, Q., Liang, S., Jia, H., Stadlmayr, A., Tang, L., Lan, Z., et al. (2015). Gut microbiome development along the colorectal adenoma-carcinoma sequence. *Nat. Commun.* 6:6528. doi: 10.1038/ncomms7528

Filzmoser, P., Hron, K., and Templ, M. (2018). *Applied Compositional Data Analysis*. Cham: Springer.

Fioravanti, D., Giarratano, Y., Maggio, V., Agostinelli, C., Chierici, M., Jurman, G., et al. (2018). Phylogenetic convolutional neural networks in metagenomics. *BMC Bioinform.* 19:49. doi: 10.1186/s12859-018-2033-5

Flemer, B., Warren, R. D., Barrett, M. P., Cisek, K., Das, A., Jeffery, I. B., et al. (2017). The oral microbiota in colorectal cancer is distinctive and predictive. *Gut* 67, 1454–1463. doi: 10.1136/gutjnl-2017-314814

Franzosa, E. A., McIver, L. J., Rahnavard, G., Thompson, L. R., Schirmer, M., Weingart, G., et al. (2018). Species-level functional profiling of metagenomes and metatranscriptomes. *Nat. Methods* 15, 962–968. doi: 10.1038/s41592-018-0176-y

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Ann. Stat.* 29, 1189–1232. doi: 10.1214/aos/1013203451

Fukui, H., Nishida, A., Matsuda, S., Kira, F., Watanabe, S., Kuriyama, M., et al. (2020). Usefulness of machine learning-based gut microbiome analysis for identifying patients with irritable bowels syndrome. *J. Clin. Med. Res.* 9:403. doi: 10.3390/jcm9082403

Gajer, P., Brotman, R. M., Bai, G., Sakamoto, J., Schütte, U. M. E., Zhong, X., et al. (2012). Temporal dynamics of the human vaginal microbiota. *Sci. Transl. Med.* 4:132ra52. doi: 10.1126/scitranslmed.3003605

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* 5:R80. doi: 10.1186/gb-2004-5-10-r80

Gevers, D., Kugathasan, S., Denson, L. A., Vázquez-Baeza, Y., Van Treuren, W., Ren, B., et al. (2014). The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host Microb.* 15, 382–392. doi: 10.1016/j.chom.2014.02.005

Gilbert, J. A., Blaser, M. J., Caporaso, J. G., Jansson, J. K., Lynch, S. V., and Knight, R. (2018). Current understanding of the human microbiome. *Nat. Med.* 24, 392–400. doi: 10.1038/nm.4517

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., and Egozcue, J. J. (2017). Microbiome datasets are compositional: and this is not optional. *Front. Microbiol.* 8:2224. doi: 10.3389/fmicb.2017.02224

Gonzalez, A., Navas-Molina, J. A., Kosciolek, T., McDonald, D., Vázquez-Baeza, Y., Ackermann, G., et al. (2018). Qiita: rapid, web-enabled microbiome meta-analysis. *Nat. Methods* 15, 796–798. doi: 10.1038/s41592-018-0141-9

Goodrich, J. K., Waters, J. L., Poole, A. C., Sutter, J. L., Koren, O., Blekhman, R., et al. (2014). Human genetics shape the gut microbiome. *Cell* 159, 789–799. doi: 10.1016/j.cell.2014.09.053

Gupta, A., Dhakan, D. B., Maji, A., Saxena, R., Vishnu Prasoodanan, P. K., Mahajan, S., et al. (2019). Association of Flavonifractor plautii, a flavonoid-degrading bacterium, with the gut microbiome of colorectal cancer patients in India. *mSystems* 4:e00438-19. doi: 10.1128/msystems.00438-19

Hacılar, H., Nalbantoğlu, O. U., and Bakir-Güngör, B. (2018). "Machine learning analysis of inflammatory bowel disease-associated metagenomics dataset," in *Proceedings of the 2018 3rd International Conference on Computer Science and Engineering (UBMK)*, Sarajevo.

Hagopian, W. A., Erlich, H., Lernmark, A., Rewers, M., Ziegler, A. G., Simell, O., et al. (2011). The environmental determinants of diabetes in the young (TEDDY): genetic criteria and international diabetes risk screening of 421 000 infants. *Pediatr. Diabetes* 12, 733–743. doi: 10.1111/j.1399-5448.2011.00774.x

Halfvarson, J., Brislawn, C. J., Lamendella, R., Vázquez-Baeza, Y., Walters, W. A., Bramer, L. M., et al. (2017). Dynamics of the human gut microbiome in inflammatory bowel disease. *Nat. Microbiol.* 2:17004. doi: 10.1038/nmicrobiol.2017.4

Hansen, R., Russell, R. K., Reiff, C., Louis, P., McIntosh, F., Berry, S. H., et al. (2012). Microbiota of de-novo pediatric IBD: increased *Faecalibacterium prausnitzii* and reduced bacterial diversity in Crohn's but not in ulcerative colitis. *Am. J. Gastroenterol.* 107, 1913–1922. doi: 10.1038/ajg.2012.335

Hanski, I., von Hertzen, L., Fyhrquist, N., Koskinen, K., Torppa, K., Laatikainen, T., et al. (2012). Environmental biodiversity, human microbiota, and allergy are interrelated. *Proc. Natl. Acad. Sci. U.S.A.* 109, 8334–8339. doi: 10.1073/pnas.1205624109

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. New York, NY: Springer, doi: 10.1007/978-0-387-84858-7

Heirendt, L., Arreckx, S., Pfau, T., Mendoza, S. N., Richelle, A., Heinken, A., et al. (2019). Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0. *Nat. Protoc.* 14, 639–702. doi: 10.1038/s41596-018-0098-2

Hoffman, J. I. E. (2019). "Logistic regression," in *Basic Biostatistics for Medical and Biomedical Practitioners*, ed. J. I. E. Hoffman (Amsterdam: Elsevier), 581–589. doi: 10.1016/b978-0-12-817084-7.00033-4

Hollister, E. B., Oezguen, N., Chumpitazi, B. P., Luna, R. A., Weidler, E. M., Rubio-Gonzales, M., et al. (2019). Leveraging human microbiome features to diagnose and stratify children with irritable bowel syndrome. *J. Mol. Diagn.* 21, 449–461. doi: 10.1016/j.jmoldx.2019.01.006

Holmes, I., Harris, K., and Quince, C. (2012). Dirichlet multinomial mixtures: generative models for microbial metagenomics. *PLoS One* 7:e30126. doi: 10.1371/journal.pone.0030126

Hughes, D. A., Bacigalupe, R., Wang, J., Rühlemann, M. C., Tito, R. Y., Falony, G., et al. (2020). Genome-wide associations of human gut microbiome variation and implications for causal inference analyses. *Nat. Microbiol.* 5, 1079–1087. doi: 10.1038/s41564-020-0743-8

Human Microbiome Project Consortium (2012). Structure, function and diversity of the healthy human microbiome. *Nature* 486, 207–214. doi: 10.1038/nature11234

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology* 19, 640–648. doi: 10.1097/EDE.0b013e31818131e7

Jang, B.-S., Chang, J. H., Chie, E. K., Kim, K., Park, J. W., Kim, M. J., et al. (2020). Gut microbiome composition is associated with a pathologic response after preoperative chemoradiation in patients with rectal cancer. *Int. J. Radiat. Oncol. Biol. Phys.* 107, 736–746. doi: 10.1016/j.ijrobp.2020.04.015

Jensen, L. J., Julien, P., Kuhn, M., von Mering, C., Muller, J., Doerks, T., et al. (2007). eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res.* 36, D250–D254. doi: 10.1093/nar/gkm796

Jiang, P., Green, S. J., Chlipala, G. E., Turek, F. W., and Vitaterna, M. H. (2019). Reproducible changes in the gut microbiome suggest a shift in microbial and host metabolism during spaceflight. *Microbiome* 7:113. doi: 10.1186/s40168-019-0724-4

Johnson, H. R., Trinidad, D. D., Guzman, S., Khan, Z., Parziale, J. V., DeBruyn, J. M., et al. (2016). A machine learning approach for using the postmortem skin microbiome to estimate the postmortem interval. *PLoS One* 11:e0167370. doi: 10.1371/journal.pone.0167370

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi: 10.1093/nar/28.1.27

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. (2004). The KEGG resource for deciphering the genome. *Nucleic Acids Res.* 32, D277–D280. doi: 10.1093/nar/gkh063

Kashyap, P. C., Chia, N., Nelson, H., Segal, E., and Elinav, E. (2017). Microbiome at the frontier of personalized medicine. *Mayo Clin. Proc.* 92, 1855–1864. doi: 10.1016/j.mayocp.2017.10.004

Kharrat, N., Assidi, M., Abu-Elmagd, M., Pushparaj, P. N., Alkhaldy, A., Arfaoui, L., et al. (2019). Data mining analysis of human gut microbiota links *Fusobacterium* spp. with colorectal cancer onset. *Bioinformation* 15, 372–379. doi: 10.6026/97320630015372

Knights, D., Costello, E. K., and Knight, R. (2011). Supervised classification of human microbiota. *FEMS Microbiol. Rev.* 35, 343–359. doi: 10.1111/j.1574-6976.2010.00251.x

Koohi-Moghadam, M., Borad, M. J., Tran, N. L., Swanson, K. R., Boardman, L. A., Sun, H., et al. (2019). MetaMarker: a pipeline for de novo discovery of novel metagenomic biomarkers. *Bioinformatics* 35, 3812–3814. doi: 10.1093/bioinformatics/btz123

Koren, O., Knights, D., Gonzalez, A., Waldron, L., Segata, N., Knight, R., et al. (2013). A guide to enterotypes across the human body: meta-analysis of microbial community structures in human microbiome datasets. *PLoS Comput. Biol.* 9:e1002863. doi: 10.1371/journal.pcbi.1002863

Kuczynski, J., Stombaugh, J., Walters, W. A., González, A., Gregory Caporaso, J., and Knight, R. (2012). Using QIIME to Analyze 16S rRNA gene sequences from microbial communities. *Curr. Protoc. Microbiol.* 27, 1E.5.1–1E.5.20. doi: 10.1002/9780471729259.mc01e05s27

La Rosa, P. S., Warner, B. B., Zhou, Y., Weinstock, G. M., Sodergren, E., Hall-Moore, C. M., et al. (2014). Patterned progression of bacterial populations in the premature infant gut. *Proc. Natl. Acad. Sci. U.S.A.* 111, 12522–12527. doi: 10.1073/pnas.1409497111

Lagani, V., Athineou, G., Farcomeni, A., Tsagris, M., and Tsamardinos, I. (2017). Feature selection with the R Package MXM: discovering statistically equivalent feature subsets. *J. Statist. Softw.* 80, 1–25. doi: 10.18637/jss.v080.i07

Lahti, L., Salonen, A., Kekkonen, R. A., Salojarvi, J., Jalanka-Tuovinen, J., Palva, A., et al. (2013). Associations between the human intestinal microbiota,

Lactobacillus rhamnosus GG and serum lipids indicated by integrated analysis of high-throughput profiling data. *PeerJ* 1:e32. doi: 10.7717/peerj.32

Lakin, S. M., Dean, C., Noyes, N. R., Dettenwanger, A., Ross, A. S., Doster, E., et al. (2017). MEGARes: an antimicrobial resistance database for high throughput sequencing. *Nucleic Acids Res.* 45, D574–D580. doi: 10.1093/nar/gkw1009

Langille, M. G. I., Zaneveld, J., Gregory Caporaso, J., McDonald, D., Knights, D., Reyes, J. A., et al. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat. Biotechnol.* 31, 814–821. doi: 10.1038/nbt.2676

LaPierre, N., Ju, C. J.-T., Zhou, G., and Wang, W. (2019). MetaPheno: a critical evaluation of deep learning and machine learning in metagenome-based disease prediction. *Methods* 166, 74–82. doi: 10.1016/j.ymeth.2019.03.003

Larsen, P. E., and Dai, Y. (2015). Metabolome of human gut microbiome is predictive of host dysbiosis. *Gigascience* 4:42. doi: 10.1186/s13742-015-0084-3

Le, V., Quinn, T. P., Tran, T., and Venkatesh, S. (2020). Deep in the bowel: highly interpretable neural encoder-decoder networks predict gut metabolites from gut microbiome. *BMC Genom.* 21:256. doi: 10.1186/s12864-020-6652-7

Le Goallec, A., Tierney, B. T., Luber, J. M., Cofer, E. M., Kostic, A. D., and Patel, C. J. (2020). A systematic machine learning and data type comparison yields metagenomic predictors of infant age, sex, breastfeeding, antibiotic usage, country of origin, and delivery type. *PLoS Comput. Biol.* 16:e1007895. doi: 10.1371/journal.pcbi.1007895

Li, J., Jia, H., Cai, X., Zhong, H., Feng, Q., Sunagawa, S., et al. (2014). An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* 32, 834–841. doi: 10.1038/nbt.2942

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., et al. (2010). De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* 20, 265–272. doi: 10.1101/gr.097261.109

Liu, Y., Guo, J., and Zhu, H. (2011). "Gene prediction in metagenomic fragments based on the SVM algorithm," in *Proceedings of the 2011 4th International Conference on Biomedical Engineering and Informatics (BMEI)*, Shanghai, doi: 10.1109/bmei.2011.6098588

Liu, Z., Hsiao, W., Cantarel, B. L., Drábek, E. F., and Fraser-Liggett, C. (2011). Sparse distance-based learning for simultaneous multiclass classification and feature selection of metagenomic data. *Bioinformatics* 27, 3242–3249. doi: 10.1093/bioinformatics/btr547

Liu, Y., Meric, G., Havulinna, A. S., Teo, S. M., Ruuskanen, M., Sanders, J., et al. (2020). Early prediction of liver disease using conventional risk factors and gut microbiome-augmented gradient boosting. *medRxiv* [Preprint], doi: 10.1101/2020.06.24.20138933

Lo, C., and Marculescu, R. (2019). MetaNN: accurate classification of host phenotypes from metagenomic data using neural networks. *BMC Bioinform.* 20:314. doi: 10.1186/s12859-019-2833-2

Lopez Pinaya, W. H., Vieira, S., Garcia-Dias, R., and Mechelli, A. (2020). "Convolutional neural networks," in *Machine Learning*, eds A. Mechelli and S. Vieira (Amsterdam: Elsevier), 173–191. doi: 10.1016/b978-0-12-815739-8.00010-9

Lozupone, C., Lladser, M. E., Knights, D., Stombaugh, J., and Knight, R. (2011). UniFrac: an effective distance metric for microbial community comparison. *ISME J.* 5, 169–172. doi: 10.1038/ismej.2010.133

Lugo-Martinez, J., Ruiz-Perez, D., Narasimhan, G., and Bar-Joseph, Z. (2019). Dynamic interaction network inference from longitudinal microbiome data. *Microbiome* 7:54. doi: 10.1186/s40168-019-0660-3

Madeira, S. C., and Oliveira, A. L. (2004). Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 1, 24–45. doi: 10.1109/TCBB.2004.2

McDonald, D., Hyde, E., Debelius, J. W., Morton, J. T., Gonzalez, A., Ackermann, G., et al. (2018). American gut: an open platform for citizen science microbiome research. *mSystems* 3:e0031-18. doi: 10.1128/mSystems.00031-18

Mitchell, A. L., Almeida, A., Beracochea, M., Boland, M., Burgin, J., Cochrane, G., et al. (2020). MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Res.* 48, D570–D578. doi: 10.1093/nar/gkz1035

Mitchell, A. L., Scheremetjew, M., Denise, H., Potter, S., Tarkowska, A., Qureshi, M., et al. (2018). EBI Metagenomics in 2017: enriching the analysis of microbial communities, from sequence reads to assemblies. *Nucleic Acids Res.* 46, D726–D735. doi: 10.1093/nar/gkx967

Mohammed, A., and Guda, C. (2015). Application of a hierarchical enzyme classification method reveals the role of gut microbiome in human metabolism. *BMC Genomics* 16(Suppl. 7):S16. doi: 10.1186/1471-2164-16-S7-S16

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., and PRISMA Group (2010). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Int. J. Surg.* 8, 336–341. doi: 10.1016/j.ijsu.2010.02.007

Moher, D., Stewart, L., and Shekelle, P. (2015). All in the family: systematic reviews, rapid reviews, scoping reviews, realist reviews, and more. *Syst. Rev.* 4:183. doi: 10.1186/s13643-015-0163-7

Moreno-Indias, I., Lahti, L., Nedyalkova, M., Elbere, I., Roshchupkin, G., Adilovic, M., et al. (2021). Statistical and machine learning techniques in human microbiome studies: contemporary challenges and solutions. *Front. Microbiol.* 12:635781. doi: 10.3389/fmicb.2021.635781

Nielsen, H. B., Almeida, M., Juncker, A. S., Rasmussen, S., Li, J., Sunagawa, S., et al. (2014). Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat. Biotechnol.* 32, 822–828. doi: 10.1038/nbt.2939

Ning, J., and Beiko, R. G. (2015). Phylogenetic approaches to microbial community classification. *Microbiome* 3:47. doi: 10.1186/s40168-015-0114-5

Noguchi, H., Park, J., and Takagi, T. (2006). MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Res.* 34, 5623–5630. doi: 10.1093/nar/gkl723

Oh, M., and Zhang, L. (2020). DeepMicro: deep representation learning for disease prediction based on microbiome data. *Sci. Rep.* 10:6026. doi: 10.1038/s41598-020-63159-5

Oudah, M., and Henschel, A. (2018). Taxonomy-aware feature engineering for microbiome classification. *BMC Bioinform.* 19:227. doi: 10.1186/s12859-018-2205-3

Papoutsoglou, G., Athineou, G., Lagani, V., Xanthopoulos, I., Schmidt, A., Éliás, S., et al. (2017). SCENERY: a web application for (causal) network reconstruction from cytometry data. *Nucleic Acids Res.* 45, W270–W275. doi: 10.1093/nar/gkx448

Pascal, V., Pozuelo, M., Borruel, N., Casellas, F., Campos, D., Santiago, A., et al. (2017). A microbial signature for Crohn's disease. *Gut* 66, 813–822. doi: 10.1136/gutjnl-2016-313235

Pasolli, E., Schiffer, L., Manghi, P., Renson, A., Obenchain, V., Truong, D. T., et al. (2017). Accessible, curated metagenomic data through ExperimentHub. *Nat. Methods* 14, 1023–1024. doi: 10.1038/nmeth.4468

Pasolli, E., Truong, D. T., Malik, F., Waldron, L., and Segata, N. (2016). Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. *PLoS Comput. Biol.* 12:e1004977. doi: 10.1371/journal.pcbi.1004977

Pawlowsky-Glahn, V., Egozcue, J. J., and Tolosana-Delgado, R. (2015). *Modeling and Analysis of Compositional Data*. Chichester: John Wiley & Sons.

Pereira, P., Aho, V., Arola, J., Boyd, S., Jokelainen, K., Paulin, L., et al. (2017). Bile microbiota in primary sclerosing cholangitis: impact on disease progression and development of biliary dysplasia. *PLoS One* 12:e0182924. doi: 10.1371/journal.pone.0182924

Petersen, C., and Round, J. L. (2014). Defining dysbiosis and its influence on host immunity and disease. *Cell. Microbiol.* 16, 1024–1033. doi: 10.1111/cmi.12308

Platt, J. C. (1998). Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. Technical Report MSR-TR-98-14, Microsoft Research.

Plaza Oñate, F., Le Chatelier, E., Almeida, M., Cervino, A. C. L., Gauthier, F., Magoulès, F., et al. (2019). MSPminer: abundance-based reconstitution of microbial pan-genomes from shotgun metagenomic data. *Bioinformatics* 35, 1544–1552. doi: 10.1093/bioinformatics/bty830

Purcell, R. V., Visnovska, M., Biggs, P. J., Schmeier, S., and Frizelle, F. A. (2017). Distinct gut microbiome patterns associate with consensus molecular subtypes of colorectal cancer. *Sci. Rep.* 7:11590. doi: 10.1038/s41598-017-11237-6

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464, 59–65. doi: 10.1038/nature08821

Quinn, T. P., and Erb, I. (2020). Interpretable log contrasts for the classification of health biomarkers: a new approach to balance selection. *mSystems* 5:e00230-19. doi: 10.1128/mSystems.00230-19

Quinn, T. P., Erb, I., Richardson, M. F., and Crowley, T. M. (2018). Understanding sequencing data as compositions: an outlook and review. *Bioinformatics* 34, 2870–2878. doi: 10.1093/bioinformatics/bty175

Rahman, S. F., Olm, M. R., Morowitz, M. J., and Banfield, J. F. (2017). Machine learning leveraging genomes from metagenomes identifies influential antibiotic resistance genes in the infant gut microbiome. *bioRxiv* [Preprint], doi: 10.1101/185348

Randolph, T. W., Zhao, S., Copeland, W., Hullar, M., and Shojaie, A. (2018). Kernel-penalized regression for analysis of microbiome data. *Ann. Appl. Stat.* 12, 540–566. doi: 10.1214/17-AOAS1102

Richards, A. L., Muehlbauer, A. L., Alazizi, A., Burns, M. B., Findley, A., Messina, F., et al. (2019). Gut microbiota has a widespread and modifiable effect on host gene regulation. *mSystems* 4:e00323-18. doi: 10.1128/mSystems.00323-18

Riley, P. (2019). Three pitfalls to avoid in machine learning. *Nature* 572, 27–29. doi: 10.1038/d41586-019-02307-y

Rivera-Pinto, J., Egozcue, J. J., Pawlowsky-Glahn, V., Paredes, R., Noguera-Julian, M., and Calle, M. L. (2018). Balances: a new perspective for microbiome analysis. *mSystems* 3:e0053-18. doi: 10.1128/mSystems.00053-18

Roguet, A., Eren, A. M., Newton, R. J., and McLellan, S. L. (2018). Fecal source identification using random forest. *Microbiome* 6:185. doi: 10.1186/s40168-018-0568-3

Ross, A. A., Doxey, A. C., and Neufeld, J. D. (2017). The skin microbiome of cohabiting couples. *mSystems* 2:e0043-17. doi: 10.1128/mSystems.00043-17

Ross, M. C., Muzny, D. M., McCormick, J. B., Gibbs, R. A., Fisher-Hoch, S. P., and Petrosino, J. F. (2015). 16S gut community of the cameron county hispanic cohort. *Microbiome* 3:7. doi: 10.1186/s40168-015-0072-y

Russell, S. J., and Norvig, P. (2016). *Artificial Intelligence: a Modern Approach*. Malaysia: Pearson Education Limited.

Ruuskanen, M. O., Åberg, F., Männistö, V., Havulinna, A. S., Méric, G., Liu, Y., et al. (2020). Links between gut microbiome composition and fatty liver disease in a large population sample. *medRxiv* [Preprint], doi: 10.1101/2020.07.30.20164962

Ryan, F. J., Ahern, A. M., Fitzgerald, R. S., Laserna-Mendieta, E. J., Power, E. M., Clooney, A. G., et al. (2020). Colonic microbiota is associated with inflammation and host epigenomic alterations in inflammatory bowel disease. *Nat. Commun.* 11:1512. doi: 10.1038/s41467-020-15342-5

Sanna, S., van Zuydam, N. R., Mahajan, A., Kurilshikov, A., Vich Vila, A., Võsa, U., et al. (2019). Causal relationships among the gut microbiome, short-chain fatty acids and metabolic diseases. *Nat. Genet.* 51, 600–605. doi: 10.1038/s41588-019-0350-x

Saulnier, D. M., Riehle, K., Mistretta, T.-A., Diaz, M.-A., Mandal, D., Raza, S., et al. (2011). Gastrointestinal microbiome signatures of pediatric patients with irritable bowel syndrome. *Gastroenterology* 141, 1782–1791. doi: 10.1053/j.gastro.2011.06.072

Scholz, M., Ward, D. V., Pasolli, E., Tolio, T., Zolfo, M., Asnicar, F., et al. (2016). Strain-level microbial epidemiology and population genomics from shotgun metagenomics. *Nat. Methods* 13, 435–438. doi: 10.1038/nmeth.3802

Schubert, A. M., Rogers, M. A. M., Ring, C., Mogle, J., Petrosino, J. P., Young, V. B., et al. (2014). Microbiome data distinguish patients with *Clostridium difficile* infection and non-*C. difficile*-associated diarrhea from healthy controls. *mBio* 5:e001021-14. doi: 10.1128/mBio.01021-14

Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W. S., et al. (2011). Metagenomic biomarker discovery and explanation. *Genome Biol.* 12:R60. doi: 10.1186/gb-2011-12-6-r60

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., and Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat Methods* 9, 811–814. doi: 10.1038/nmeth.2066

Seo, M., Heo, J., Yoon, J., Kim, S.-Y., Kang, Y.-M., Yu, J., et al. (2017). *Methanobrevibacter* attenuation via probiotic intervention reduces flatulence in adult human: a non-randomised paired-design clinical trial of efficacy. *PLoS One* 12:e0184547. doi: 10.1371/journal.pone.0184547

Silverman, J. D., Washburne, A. D., Mukherjee, S., and David, L. A. (2017). A phylogenetic transform enhances analysis of compositional microbiota data. *eLife* 6:e21887. doi: 10.7554/eLife.21887

Sokol, H., Leducq, V., Aschard, H., Pham, H.-P., Jegou, S., Landman, C., et al. (2017). Fungal microbiota dysbiosis in IBD. *Gut* 66, 1039–1048. doi: 10.1136/gutjnl-2015-310746

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi: 10.1093/bioinformatics/btu033

Statnikov, A., Henaff, M., Narendra, V., Konganti, K., Li, Z., Yang, L., et al. (2013). A comprehensive evaluation of multicategory classification methods for microbiomic data. *Microbiome* 1:11. doi: 10.1186/2049-2618-1-11

Sze, M. A., and Schloss, P. D. (2016). Looking for a signal in the noise: revisiting obesity and the microbiome. *mBio* 7:e01018-16. doi: 10.1128/mBio.01018-16

Tap, J., Derrien, M., Tornblom, H., Brazeilles, R., Cools-Portier, S., Dore, J., et al. (2017). Identification of an intestinal microbiota signature associated with severity of irritable bowel syndrome. *Gastroenterology* 152, 111–123.e8. doi: 10.1053/j.gastro.2016.09.049

Telalovic, H. J., and Azra, M. (2020). Using data science for medical decision making case: role of gut microbiome in multiple sclerosis. *BMC Med. Inform. Decis. Mak.* 20:262. doi: 10.1186/s12911-020-01263-2

Thomas, A. M., Manghi, P., Asnicar, F., Pasolli, E., Armanini, F., Zolfo, M., et al. (2019). Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nat. Med.* 25, 667–678. doi: 10.1038/s41591-019-0405-7

Travisany, D., Galarce, D., Maass, A., and Assar, R. (2015). "predicting the metagenomics content with multiple CART trees," in *Mathematical Models in Biology: Bringing Mathematics to Life*, eds V. Zazzu, M. B. Ferraro, and M. R. Guarracino (Cham: Springer International Publishing), 145–160. doi: 10.1007/978-3-319-23497-7_11

Truong, D. T., Franzosa, E. A., Tickle, T. L., Scholz, M., Weingart, G., Pasolli, E., et al. (2015). MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* 12, 902–903. doi: 10.1038/nmeth.3589

Tsamardinos, I., Charonyktakis, P., Lakiotaki, K., Borboudakis, G., Zenklusen, J. C., Juhl, H., et al. (2020). Just add data: automated predictive modeling and biosignature discovery. *bioRxiv* [Preprint], doi: 10.1101/2020.05.04.075747

Tsamardinos, I., Greasidou, E., and Borboudakis, G. (2018). Bootstrapping the out-of-sample predictions for efficient and accurate cross-validation. *Mach. Learn.* 107, 1895–1922. doi: 10.1007/s10994-018-5714-4

Tsamardinos, I., Rakhshani, A., and Lagani, V. (2015). Performance-estimation properties of cross-validationbased protocols with simultaneous hyper-parameter optimization. *Int. J. Artif. Intell. Tools* 24, 1–14. doi: 10.1007/978-3-319-07064-3_1

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., and Gordon, J. I. (2007). The human microbiome project. *Nature* 449, 804–810. doi: 10.1038/nature06244

Turnbaugh, P. J., Ley, R. E., Mahowald, M. A., Magrini, V., Mardis, E. R., and Gordon, J. I. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* 444, 1027–1031. doi: 10.1038/nature05414

Turnbaugh, P. J., Ridaura, V. K., Faith, J. J., Rey, F. E., Knight, R., and Gordon, J. I. (2009). The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice. *Sci. Transl. Med.* 1:6ra14. doi: 10.1126/scitranslmed.3000322

Vangay, P., Hillmann, B. M., and Knights, D. (2019). Microbiome Learning Repo (ML Repo): a public repository of microbiome regression and classification tasks. *Gigascience* 8:giz042. doi: 10.1093/gigascience/giz042

Vatanen, T., Franzosa, E. A., Schwager, R., Tripathi, S., Arthur, T. D., Vehik, K., et al. (2018). The human gut microbiome in early-onset type 1 diabetes from the TEDDY study. *Nature* 562, 589–594. doi: 10.1038/s41586-018-0620-2

Vervier, K., Mahé, P., Tournoud, M., Veyrieras, J.-B., and Vert, J.-P. (2016). Large-scale machine learning for metagenomics sequence classification. *Bioinformatics* 32, 1023–1032. doi: 10.1093/bioinformatics/btv683

Wang, H., Marcišauskas, S., Sánchez, B. J., Domenzain, I., Hermansson, D., Agren, R., et al. (2018). RAVEN 2.0: a versatile toolbox for metabolic network reconstruction and a case study on *Streptomyces coelicolor*. *PLoS Comput. Biol.* 14:e1006541. doi: 10.1371/journal.pcbi.1006541

Wassan, J. T., Wang, H., Browne, F., and Zheng, H. (2018a). A comprehensive study on predicting functional role of metagenomes using machine learning methods. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 16, 751–763. doi: 10.1109/TCBB.2018.2858808

Wassan, J. T., Wang, H., Browne, F., and Zheng, H. (2018b). "PAAM-ML: a novel phylogeny and abundance aware machine learning modelling approach for microbiome classification," in *Proceedings of the 2018 IEEE International*

Conference on Bioinformatics and Biomedicine (BIBM), Madrid, doi: 10.1109/BIBM.2018.8621382

Wassan, J. T., Wang, H., Browne, F., and Zheng, H. (2019). Phy-PMRFI: phylogeny-aware prediction of metagenomic functions using random forest feature importance. *IEEE Trans. Nanobiosci.* 18, 273–282. doi: 10.1109/tnb.2019.2912824

Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., et al. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* 5:27. doi: 10.1186/s40168-017-0237-y

Wen, C., Zheng, Z., Shao, T., Liu, L., Xie, Z., Le Chatelier, E., et al. (2017). Quantitative metagenomics reveals unique gut microbiome biomarkers in ankylosing spondylitis. *Genome Biol.* 18:142. doi: 10.1186/s13059-017-1271-6

Werner, J. J., Koren, O., Hugenholtz, P., DeSantis, T. Z., Walters, W. A., Caporaso, J. G., et al. (2012). Impact of training sets on classification of high-throughput bacterial 16s rRNA gene surveys. *ISME J.* 6, 94–103. doi: 10.1038/ismej.2011.82

Winand, R., Bogaerts, B., Hoffman, S., Lefevre, L., Delvoye, M., Van Braekel, J., et al. (2020). Targeting the 16s rRNA gene for bacterial identification in complex mixed samples: Comparative evaluation of second (Illumina) and third (oxford nanopore technologies) generation sequencing technologies. *Int. J. Mol. Sci.* 21:298.

Wingfield, B., Coleman, S., McGinnity, T. M., and Bjourson, A. J. (2016). "A metagenomic hybrid classifier for paediatric inflammatory bowel disease," in *Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN)*, Vancouver, BC, doi: 10.1109/ijcnn.2016.7727318

Wirbel, J., Pyl, P. T., Kartal, E., Zych, K., Kashani, A., Milanese, A., et al. (2019). Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat. Med.* 25, 679–689. doi: 10.1038/s41591-019-0406-6

Wu, C., Chen, J., Kim, J., and Pan, W. (2016). An adaptive association test for microbiome data. *Genome Med.* 8:56. doi: 10.1186/s13073-016-0302-3

Wu, G. D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y.-Y., Keilbaugh, S. A., et al. (2011). Linking long-term dietary patterns with gut microbial enterotypes. *Science* 334, 105–108. doi: 10.1126/science.1208344

Wu, H., Cai, L., Li, D., Wang, X., Zhao, S., Zou, F., et al. (2018). Metagenomics biomarkers selected for prediction of three different diseases in chinese population. *Biomed Res. Int.* 2018:2936257. doi: 10.1155/2018/2936257

Xia, L. C., Cram, J. A., Chen, T., Fuhrman, J. A., and Sun, F. (2011). Accurate genome relative abundance estimation based on shotgun metagenomic reads. *PLoS One* 6:e27992. doi: 10.1371/journal.pone.0027992

Xie, J., Ma, A., Fennell, A., Ma, Q., and Zhao, J. (2019). It is time to apply biclustering: a comprehensive review of biclustering applications in biological and biomedical data. *Brief. Bioinform.* 20, 1449–1464. doi: 10.1093/bib/bby014

Yachida, S., Mizutani, S., Shiroma, H., Shiba, S., Nakajima, T., Sakamoto, T., et al. (2019). Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. *Nat. Med.* 25, 968–976. doi: 10.1038/s41591-019-0458-7

Yang, J., Tsukimi, T., Yoshikawa, M., Suzuki, K., Takeda, T., Tomita, M., et al. (2019). *Cutibacterium acnes* (*Propionibacterium acnes*) 16S rRNA genotyping of microbial samples from possessions contributes to owner identification. *mSystems* 4:e001672-17. doi: 10.1128/mSystems.00594-19

Yang, L., Yachimski, P. S., Brodie, E., Nelson, K. E., and Pei, Z. (2015). "Foregut microbiome, development of esophageal adenocarcinoma, project," in *Encyclopedia of Metagenomics*, eds. S. K. Highlander, F. Rodriguez-Valera and B A. White (Cham: Springer), 186–189. doi: 10.1007/978-1-4899-7475-4_709

Yarza, P., Yilmaz, P., Panzer, K., Glöckner, F. O., and Reich, M. (2017). A phylogenetic framework for the kingdom Fungi based on 18S rRNA gene sequences. *Mar. Genom.* 36, 33–39. doi: 10.1016/j.margen.2017.05.009

Zdravevski, E., Lameski, P., Trajkovik, V., Chorbev, I., Goleva, R., Pombo, N., et al. (2019). "Automation in systematic, scoping and rapid reviews by an NLP toolkit: a case study in enhanced living environments," in *Enhanced Living Environments. Lecture Notes in Computer Science*, Vol. 11369, eds I. Ganchev, N. Garcia, C. Dobre, C. Mavromoustakis, and R. Goleva (Cham: Springer), 1–18. doi: 10.1007/978-3-030-10752-9_1

Zeevi, D., Korem, T., Zmora, N., Israeli, D., Rothschild, D., Weinberger, A., et al. (2015). Personalized nutrition by prediction of glycemic responses. *Cell* 163, 1079–1094. doi: 10.1016/j.cell.2015.11.001

Zeller, G., Tap, J., Voigt, A. Y., Sunagawa, S., Kultima, J. R., Costea, P. I., et al. (2014). Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.* 10:766. doi: 10.15252/msb.20145645

Zhang, Z.-Y. (2012). "Nonnegative matrix factorization: models, algorithms and applications," in *Data Mining: Foundations and Intelligent Paradigms: Volume 2: Statistical, Bayesian, Time Series and other Theoretical Aspects*, eds D. E. Holmes and L. C. Jain (Berlin: Springer), 99–134. doi: 10.1007/978-3-642-23241-1_6

Zhou, F., He, K., Li, Q., Chapkin, R. S., and Ni, Y. (2020). Bayesian biclustering for microbial metagenomic sequencing data via multinomial matrix factorization. *arXiv* [Preprint], Available online at: http://arxiv.org/abs/2005.08361 (accessed 08 February, 2021).

Zhou, Y.-H., and Gallins, P. (2019). A review and tutorial of machine learning methods for microbiome host trait prediction. *Front. Genet.* 10:579. doi: 10.3389/fgene.2019.00579

Zhu, Q., Li, B., He, T., Li, G., and Jiang, X. (2020). Robust biomarker discovery for microbiome-wide association studies. *Methods* 173, 44–51. doi: 10.1016/j.ymeth.2019.06.012

Zupancic, M. L., Cantarel, B. L., Liu, Z., Drabek, E. F., Ryan, K. A., Cirimotich, S., et al. (2012). Analysis of the gut microbiota in the old order Amish and its relation to the metabolic syndrome. *PLoS One* 7:e43052. doi: 10.1371/journal.pone.0043052

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

PERSPECTIVE
published: 22 February 2021
doi: 10.3389/fmicb.2021.635781

Statistical and Machine Learning Techniques in Human Microbiome Studies: Contemporary Challenges and Solutions

Isabel Moreno-Indias[1,2], Leo Lahti[3], Miroslava Nedyalkova[4], Ilze Elbere[5], Gennady Roshchupkin[6], Muhamed Adilovic[7], Onder Aydemir[8], Burcu Bakir-Gungor[9], Enrique Carrillo-de Santa Pau[10], Domenica D'Elia[11], Mahesh S. Desai[12,13], Laurent Falquet[14,15], Aycan Gundogdu[16,17], Karel Hron[18], Thomas Klammsteiner[19], Marta B. Lopes[20,21], Laura Judith Marcos-Zambrano[10], Cláudia Marques[22], Michael Mason[23], Patrick May[24], Lejla Pašić[25], Gianvito Pio[26], Sándor Pongor[27], Vasilis J. Promponas[28], Piotr Przymus[29], Julio Saez-Rodriguez[30], Alexia Sampri[31], Rajesh Shigdel[32], Blaz Stres[33,34,35], Ramona Suharoschi[36], Jaak Truu[37], Ciprian-Octavian Truică[38], Baiba Vilne[39], Dimitrios Vlachakis[40], Ercument Yilmaz[41], Georg Zeller[42], Aldert L. Zomer[43], David Gómez-Cabrero[44] and Marcus J. Claesson[45] on Behalf of ML4Microbiome*

[1] Instituto de Investigación Biomédica de Málaga (IBIMA), Unidad de Gestión Clìnica de Endocrinologìa y Nutrición, Hospital Clìnico Universitario Virgen de la Victoria, Universidad de Málaga, Málaga, Spain, [2] Centro de Investigación Biomeìdica en Red de Fisiopatologtìa de la Obesidad y la Nutrición (CIBEROBN), Instituto de Salud Carlos III, Madrid, Spain, [3] Department of Computing, University of Turku, Turku, Finland, [4] Human Genetics and Disease Mechanisms, Latvian Biomedical Research and Study Centre, Riga, Latvia, [5] Latvian Biomedical Research and Study Centre, Riga, Latvia, [6] Department of Epidemiology, Erasmus Medical Center, Rotterdam, Netherlands, [7] Department of Genetics and Bioengineering, International University of Sarajevo, Sarajevo, Bosnia and Herzegovina, [8] Department of Electrical and Electronics Engineering, Karadeniz Technical University, Trabzon, Turkey, [9] Department of Computer Engineering, Abdullah Gul University, Kayseri, Turkey, [10] Computational Biology Group, Precision Nutrition and Cancer Research Program, IMDEA Food Institute, Madrid, Spain, [11] Department for Biomedical Sciences, Institute for Biomedical Technologies, National Research Council, Bari, Italy, [12] Department of Infection and Immunity, Luxembourg Institute of Health, Esch-sur-Alzette, Luxembourg, [13] Odense Research Center for Anaphylaxis, Department of Dermatology and Allergy Center, Odense University Hospital, University of Southern Denmark, Odense, Denmark, [14] Department of Biology, University of Fribourg, Fribourg, Switzerland, [15] Swiss Institute of Bioinformatics, Lausanne, Switzerland, [16] Department of Microbiology and Clinical Microbiology, Faculty of Medicine, Erciyes University, Kayseri, Turkey, [17] Metagenomics Laboratory, Genome and Stem Cell Center (GenKök), Erciyes University, Kayseri, Turkey, [18] Department of Mathematical Analysis and Applications of Mathematics, Palacký University, Olomouc, Czechia, [19] Department of Microbiology, University of Innsbruck, Innsbruck, Austria, [20] NOVA Laboratory for Computer Science and Informatics (NOVA LINCS), FCT, UNL, Caparica, Portugal, [21] Centro de Matemática e Aplicações (CMA), FCT, UNL, Caparica, Portugal, [22] CINTESIS, NOVA Medical School, NMS, Universidade Nova de Lisboa, Lisbon, Portugal, [23] Computational Oncology, Sage Bionetworks, Seattle, WA, United States, [24] Bioinformatics Core, Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg, [25] Sarajevo Medical School, University Sarajevo School of Science and Technology, Sarajevo, Bosnia and Herzegovina, [26] Department of Computer Science, University of Bari Aldo Moro, Bari, Italy, [27] Faculty of Information Tehnology and Bionics, Pázmány University, Budapest, Hungary, [28] Bioinformatics Research Laboratory, Department of Biological Sciences, University of Cyprus, Nicosia, Cyprus, [29] Faculty of Mathematics and Computer Science, Nicolaus Copernicus University, Toruñ, Poland, [30] Institute of Computational Biomedicine, Heidelberg University, Faculty of Medicine and Heidelberg University Hospital, Heidelberg, Germany, [31] Division of Informatics, Imaging and Data Sciences, School of Health Sciences, University of Manchester, Manchester, United Kingdom, [32] Department of Clinical Science, University of Bergen, Bergen, Norway, [33] Jozef Stefan Institute, Ljubljana, Slovenia, [34] Biotechnical Faculty, University of Ljubljana, Ljubljana, Slovenia, [35] Faculty of Civil and Geodetic Engineering, University of Ljubljana, Ljubljana, Slovenia, [36] Molecular Nutrition and Proteomics Lab, Faculty of the Food Science and Technology, Institute of Life Sciences, University of Agricultural Sciences and Veterinary Medicine of Cluj-Napoca, Cluj-Napoca, Romania, [37] Institute of Molecular and Cell Biology, University of Tartu, Tartu, Estonia, [38] Department of Computer Science and Engineering, Faculty of Automatic Control and Computers, University Politehnica of Bucharest, Bucharest, Romania, [39] Bioinformatics Research Unit, Riga Stradins University, Riga, Latvia, [40] Laboratory

OPEN ACCESS

Edited by:
Nikos Kyrpides, Lawrence Berkeley National Laboratory, United States

Reviewed by:
Stephen Nayfach, Lawrence Berkeley National Laboratory, United States
Jonathan Badger, National Cancer Institute (NCI), United States

***Correspondence:**
Isabel Moreno-Indias
isabel.moreno@ibima.eu

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

Received: 30 November 2020
Accepted: 28 January 2021
Published: 22 February 2021

Citation:
Moreno-Indias I, Lahti L, Nedyalkova M, Elbere I, Roshchupkin G, Adilovic M, Aydemir O, Bakir-Gungor B, Santa Pau EC-d, D'Elia D, Desai MS, Falquet L, Gundogdu A, Hron K, Klammsteiner T, Lopes MB, Marcos-Zambrano LJ, Marques C, Mason M, May P, Pašić L, Pio G, Pongor S, Promponas VJ, Przymus P, Saez-Rodriguez J, Sampri A, Shigdel R, Stres B, Suharoschi R, Truu J, Truică C-O, Vilne B, Vlachakis D, Yilmaz E, Zeller G, Zomer AL, Gómez-Cabrero D and Claesson MJ (2021) Statistical and Machine Learning Techniques in Human Microbiome Studies: Contemporary Challenges and Solutions. Front. Microbiol. 12:635781. doi: 10.3389/fmicb.2021.635781

of Genetics, Department of Biotechnology, School of Applied Biology and Biotechnology, Agricultural University of Athens, Athens, Greece, [41] Department of Computer Technologies, Karadeniz Technical University, Trabzon, Turkey, [42] European Molecular Biology Laboratory, Structural and Computational Biology Unit, Heidelberg, Germany, [43] Department of Infectious Diseases and Immunology, Faculty of Veterinary Medicine, Utrecht University, Utrecht, Netherlands, [44] Navarrabiomed, Complejo Hospitalario de Navarra (CHN), IdiSNA, Universidad Pública de Navarra (UPNA), Pamplona, Spain, [45] School of Microbiology and APC Microbiome Ireland, University College Cork, Cork, Ireland

The human microbiome has emerged as a central research topic in human biology and biomedicine. Current microbiome studies generate high-throughput omics data across different body sites, populations, and life stages. Many of the challenges in microbiome research are similar to other high-throughput studies, the quantitative analyses need to address the heterogeneity of data, specific statistical properties, and the remarkable variation in microbiome composition across individuals and body sites. This has led to a broad spectrum of statistical and machine learning challenges that range from study design, data processing, and standardization to analysis, modeling, cross-study comparison, prediction, data science ecosystems, and reproducible reporting. Nevertheless, although many statistics and machine learning approaches and tools have been developed, new techniques are needed to deal with emerging applications and the vast heterogeneity of microbiome data. We review and discuss emerging applications of statistical and machine learning techniques in human microbiome studies and introduce the COST Action CA18131 "ML4Microbiome" that brings together microbiome researchers and machine learning experts to address current challenges such as standardization of analysis pipelines for reproducibility of data analysis results, benchmarking, improvement, or development of existing and new tools and ontologies.

Keywords: machine learning, microbiome, ML4Microbiome, personalized medicine, biomarker identification

INTRODUCTION

The microbiome has long been defined as a community of commensal, symbiotic, or pathogenic microorganisms that inhabit a particular body site or environment (Lederberg and McCray, 2001). The current apprehension of the microbiome encompasses the totality of microorganisms and their interactions, interplay with the host and the surrounding environment, and is further influenced by constant co-evolution (Berg et al., 2020). Understanding the composition, balance, and role of the microbiome in human health and disease has become a field of extensive research over the past decade (Wang and Kasper, 2014; Gagnière et al., 2016; Sampson et al., 2016; Barratt et al., 2017). The potential for applications in biomedicine and biotechnology has been especially evident from gut microbiome studies. Furthermore, microbiome research has become an important subject of popular science and led to the acceleration of development in different biotechnology industry sectors.

Some of the key topics in this field cover early life (Tamburini et al., 2016), mechanisms of colonization resistance against pathogens (Buffie and Pamer, 2013; Kim et al., 2017), and stability and individuality of adult microbiota (Mehta et al., 2018), and its associations with diseases, diet, medication, and lifestyle in various populations across the globe (Segata et al., 2011; Schmidt et al., 2018; Cullen et al., 2020). Moreover, the research focus is shifting toward considering the role of genetics and environment (Org et al., 2015; Roslund et al., 2020), as well as of diet (Singh et al., 2017), and to translate this knowledge into microbiota-based clinical solutions (Lynch et al., 2019).

Compared to many other fields of multi-omic studies, microbiomes are dynamic ecosystems with active host regulation. This adds interesting new dimensions and complexity to the analyses and interpretation of data. Thus, the field also requires additional ecological perspectives. The advances in high-throughput sequencing technologies have accelerated microbiome research (Malla et al., 2019), but the volume of data and their complexity sets challenges for analysis. Adaptive statistical and machine learning (ML) methodologies can help us to overcome many of these barriers, but these methodologies need to be adjusted to the specific properties of microbiome data.

Microbiome Data Properties and Analysis Challenges

Two commonly used strategies for microbiome profiling include the sequencing of a highly conserved region, such as the bacterial 16S ribosomal RNA (16S rRNA), and the untargeted sequencing of genetic material present in the sample, as in shotgun metagenomics (see **Box 1** for more information) (Nayfach et al., 2019). The quality of microbiome data and profiling is influenced by experimental, biological, and environmental factors (Poussin et al., 2018). Further variation arises from differences in sequence

BOX 1 | Common data types in microbiome research.

Amplicon data. Amplicon based approaches are the most widely used high-throughput method for microbiome studies. Amplicon studies comprise data from specific regions of various types of marker genes used for taxonomic profile determination of microbiome: 16S ribosomal RNA (16S rRNA) gene for prokaryotes; 18S ribosomal RNA (18S rRNA) gene for eukaryotes; internal transcribed spacers (ITS) for fungi. These data are characterized by variability in the selected regions, amplification primers and amplification protocols. Due to the sequence similarity, the data are often organized into operational taxonomic units (OTUs) (Schmitt et al., 2012). The two most popular approaches for obtaining groups of related OTUs are based on (i) aligning sequences to a reference database or (ii) clustering sequences based on sequence identity (*de novo* approach). Once OTU clusters are defined, taxonomic information is given for the representative sequences of each OTU to deduce the phylogeny. However, probabilistic techniques such as DADA2 (Callahan et al., 2016) have recently gained more attention, and are now increasingly used to replace the standard OTU clustering approaches by ASVs, which are un-clustered error-corrected reads. Although amplicon sequencing is cost-effective, the reliability of bacterial classification decreases below genus level, and this methodology does not directly quantify bacterial genes and functions.

Shotgun metagenomics data. A growing number of studies use shotgun metagenomics and offer untargeted sequence data from the analyzed samples. These data typically include contamination from host or environmental reads as well. The non-host DNA can be used for taxonomic analysis or functional profiling of all types of microorganisms present in the microbiome–it allows the analyses of bacteria, viruses, fungi and parasites at the same time. Sequences from metagenomic data can be classified using existing databases or assembled *de novo*. This type of analysis offers the possibility to analyze strain or even SNP level dynamics of the microbiome (Quince et al., 2017; Zeevi et al., 2019) as well as reconstruction of draft genomes, which enables the identification of novel organisms and provides a way to link functions with taxa. Depending on the aims of the study, shotgun metagenomics can provide a variable amount of data as shallow, deep, or even ultradeep sequencing (Hillmann et al., 2018).

Metatranscriptome data. Metatranscriptomics characterize the expressed transcripts of the analyzed community at a given time point/conditions transcripts of the analyzed community by RNA sequencing data. Depending on the sequencing depth, with this method it is possible to obtain information on gene expression levels both for the microbiome communities and for the host. This requires the highest sequencing depth, most stringent standards for sample storage and processing, and data analysis workflows and benchmarking for these data are only in the developmental stage. Despite these advantages, metatranscriptomes will need to be supported by additional shotgun metagenomics measurements for accurate interpretation.

Other-omics data such as metabolomics data, metaproteomics data. These data represent directly measured metabolites or expressed proteins, therefore providing additional functional information. Similarly, these data can contain information both from the microbiome and the host.

filtering, clustering, taxonomic assignment and binning, as different bioinformatic tools and pipelines are in use. This lack of standardization introduces statistical biases, and subsequent challenges for reproducibility and cross-study comparisons (Lozupone et al., 2013; Falony et al., 2016; Zhernakova et al., 2016). Some of the first large microbiome profiling studies, as the Human Microbiome Project (Turnbaugh et al., 2007) and the MetaHIT project (Qin et al., 2010), were established as a population-scale framework to develop metagenomic protocols (for a more comprehensive list of large-scale microbiome studies, see Marcos-Zambrano et al., 2021). Despite various attempts to standardize methods, a gold standard of microbiome research is yet to be established (Quince et al., 2017; Knight et al., 2018).

The special characteristics of metagenomic sequencing data are posing additional challenges for statistical analysis. For instance, the large inter-individual variability, heteroscedastic variation (i.e., variance increasing with mean abundance) and large biological and technical variations are often not properly approximated by classical Gaussian or log-normal models, requiring customized analytical approaches. Microbiome data sets tend to be sparse and skewed, and typically include many more microbial features compared to the number of samples or observations collected in most microbiome studies to date (**Supplementary Table S1**). Moreover, microbiome features often exhibit complex and hierarchical dependency structures in terms of taxonomies or co-variation in abundance and function. Moreover, unaligned and misaligned sequence reads, and challenges to distinguish technical and biological variation especially at the level of low-abundant organisms add additional challenges to the microbiome analyses. The demand to represent microbiome data with an arbitrary, but fixed sum of components without loss of information are known from the concept of compositional data (Aitchison, 1986; Gloor et al., 2017). Furthermore, complementary multi-omic and other data types (**Box 1**) may require different modeling approaches. The integration of different types of data often lacks rigorous model selection procedures, correction for multiple testing, handling of missing data features/labels, or data harmonization and integration (Namkung, 2020).

Finally, the reliability and integration of relevant metadata such as demographics, health, diet, age, medication, lifestyle, and other factors are critical for drawing informative insights from microbiome studies. However, these crucial pieces of information are most often missing or insufficiently machine-readable in publicly available data resources, thus forming bottlenecks on data reuse.

Statistics and Machine Learning Aspects

Microbiome research has set fresh challenges for statistical analysis. Instead of a thorough literature review of this rapidly expanding and heterogeneous field, we provide hereby a topical perspective on the application of ML techniques in microbiome research (for an extensive review, please see Marcos-Zambrano et al., 2021).

One of the most common applications of ML is dimensionality reduction, which facilitates the exploration and visualization of community similarity and distribution across the population of study samples. Non-linear approaches have become a common choice due to the inherent complexity of microbial communities, including methods such as PCoA, UMAP, and other techniques (Legendre and Legendre, 2012; Becht et al., 2019; Kobak and Berens, 2019), as well as autoencoders (Oh and Zhang, 2020) have been taken into use. Many automated analysis pipelines readily include these methods (Buza et al., 2019; Liao et al., 2019).

Clustering has found many applications in microbiome research, ranging from data preprocessing to downstream community analyses. A popular method is the denoiser DADA2

(Callahan et al., 2016), designed to identify unique 16S rRNA amplicon sequence variants (ASVs) (Davis et al., 2018). In metagenome sequencing studies, probabilistic methods have been used to assemble contigs into genome bins based on information of abundance and sequence information; CONCOCT (Alneberg et al., 2014) implements non-parametric clustering based on a variational Gaussian mixture model. The advantage of the non-parametric approach is the automated determination of the cluster number based on the model, rather than *post hoc* evaluation indices such as the Kalinski-Harabasz or Silhouette index. In the downstream analysis of microbiome data, a notable application of clustering algorithms has been the identification of microbiome *community types,* used to stratify individuals into specific subgroups based on microbiome composition (Holmes et al., 2012; Costea et al., 2018). Recently, more detailed assemblage models have been developed to identify latent factors and sub-communities that can complement ecosystem-wide stratification that focuses on overarching community types. Examples include phylofactor (Washburne et al., 2019), tipping elements (Lahti et al., 2014), non-negative matrix factorization, latent Dirichlet allocation, and other latent mixture models (Sankaran and Holmes, 2019).

Classification methods are commonly used in taxonomic assignment of metagenomic reads to annotate genome sequences (Treangen et al., 2013; Tamames et al., 2019) or in the production of metagenome-assembled genomes (Murovec et al., 2020). Another application is sample classification in diagnostic or prognostic studies (Pasolli et al., 2016; Aryal et al., 2020). Common ML algorithms such as random forest, support vector machines (SVM), elastic net, and LASSO have all been used for disease-prediction tasks (Pasolli et al., 2016), and automated feature selection schemes have been reported to perform will in comparison with standard tests in disease prediction (Ai et al., 2017). Instead of hard classification, some applications focus on detecting estimated percentage contribution, or soft classification, of each potential source environment related to the sample (Knights et al., 2011; Shenhav et al., 2019; McGhee et al., 2020).

Deep learning (DL) is increasingly applied in microbiome research Convolutional Neural Networks (CNNs) (Armour et al., 2019) have recently been augmented with phylogenetic tree information (Reiman et al., 2018), or combined neural networks with random forests (Rahman and Rangwala, 2020). Variable evaluation metrics including accuracy, precision, recall, F1-score and area under curve (AUC), have been used, highlighting the need for standardized benchmarks regarding well-defined modeling tasks; systematic evaluations have been carried out for instance for metagenome-based disease prediction and differentiation of body sites based on microbiome composition (Asgari et al., 2018; Reiman et al., 2018; Díez López et al., 2019; LaPierre et al., 2019). DL has been also applied to classify antibiotic resistance genes (ARGs) derived from metagenomic data (Arango-Argoty et al., 2018) and to overcome the lack of well-curated taxonomic trees for newly discovered species in metagenome assembled genomes (Murovec et al., 2020). DL has also been used to predict how gut microbiome responds to perturbations by antibiotics (Rahman et al., 2018). Whereas DL methods are notoriously data-hungry, recent applications have shown promising performance with moderate training sample sizes.

A vast number of microbiome studies quantify associations between the abundances of specific metagenomic and functional features, and key covariates such as health and disease, and other factors including diet, medication, geography, or stool consistency (Turnbaugh et al., 2007; Qin et al., 2010; Falony et al., 2016; Zhernakova et al., 2016). The analysis covers a vast spectrum of standard ML methods with additional adaptations to microbiome data. Popular approaches include adaptations of linear discriminant analysis (Segata et al., 2011), negative binomials (Love et al., 2014), and Dirichlet distributions (Fernandes et al., 2014), and non-parametric methods (Weiss et al., 2017; Lin and Peddada, 2020). Non-parametric regression models, such as Gaussian processes, have been also used to study associations between microbiome diversity and external conditions (Arbel et al., 2016). Common techniques for community comparisons include regularized discriminant analysis (RDA) (Legendre and Legendre, 2012), random forest (Sze and Schloss, 2018; Topçuoğlu et al., 2020), and gradient boosting (Qin et al., 2020; Topçuoğlu et al., 2020). Further strategies have been developed in order to consider hierarchical dependencies between taxonomic groups to control for multiple testing and to identify the appropriate taxonomic levels for associations (Sankaran and Holmes, 2014; Washburne et al., 2017).

Other emerging applications include spatio-temporal modeling of microbiome variation both at the individual and population levels as well as the biogeographical variation within and across body sites; agent-based models provide interesting opportunities in this area (Juhász et al., 2014; Lin et al., 2018). Probabilistic joint species distribution models have also been recently applied in the microbiome context (Björk et al., 2018). Bayesian ML techniques can help to deal with uncertainties related to the limited information in short and sparse time series or spatial sampling. The uncertainty, the limited sampling density, or the limited amount of labeled examples when training a model can also be addressed through semi-supervised methods. Prospective analyses predicting long-term incident of health and disease risk based on microbiome composition have remained scarce due to the lack of large-scale cohorts with long-term follow-ups, but the need for prospective analysis methods is now emerging (Liu et al., 2020; Salosensaari et al., 2020). Mendelian randomization and related techniques are finding applications to understand the causal role of gut microbiome in disease (Sanna et al., 2019; Hughes et al., 2020).

DISCUSSION

Statistics and ML provide tools to extract useful information from scarce, noisy, and limited data. In particular, within microbiome data, this has to be balanced with the complexity

and limited understanding of the host-regulated ecological processes and the high levels of individual variation. ML has great potential to improve disease diagnosis and identify personalized biomarkers, due to its ability to detect informative patterns in the data with limited prior knowledge of the underlying system.

One of the main shortcomings is, however, the use of inappropriately small datasets, as apparent from the example studies (and their corresponding datasets) listed in **Supplementary Table S1**. Data accumulation will further enhance the use of more advanced ML technologies. Efficient data structures and making microbiome data Findable, Accessible, Interoperable, and Reusable (FAIR)[1] can provide invaluable support for the open development of statistical and ML tools to help to advance the field (Shetty and Lahti, 2019). Consequently, data repositories maintained by large consortia could serve as a central resource for the research community (Meyer et al., 2008; Mitchell et al., 2020). However, to this aim, the submission of the metadata must follow controlled vocabulary and minimal standards (ten Hoopen et al., 2017).

Some of the main challenges in detecting associations between specific microbiome features and key covariates are related to choosing appropriate distributional assumptions including sparsity and compositionality, appropriate feature selection, controlling for technical biases such as read count variations, the potential confounding effects, and multiple testing. Successful solutions often present combinations of statistical techniques that have been specifically tailored to fit the particular characteristics of microbiome data. Besides, over-fitting, incomplete model selection or performance assessment can lead to poor generalizability of the results in previously unseen data sets and lack of reproducibility. It is essential to understand the principles underlying each method and follow the recommended guidelines in order to ensure compliance with the modeling assumptions (Rule et al., 2019) and avoid overfitting (Eetemadi et al., 2020). Another important driver for the field is the development of suitable data structures in statistical programming languages, such as the R/Bioconductor ecosystem as curatedMetagenomicData (Pasolli et al., 2016) and the *phyloseq* (McMurdie and Holmes, 2013) or *TreeSummarizedExperiment* classes (Huang et al., 2020), that permit standardization and efficient collaborative development of methods.

The microbiome field is moving from associations to causality, mechanisms, and prediction, and ML will aid in this transition. Data obtained from ML methods can help to propose new hypotheses to be tested in experimental models, as well as to accelerate the translation of the microbiome data into clinical practice. Its optimal use will presumably trigger the improvement of the searching of biomarker candidates for disease diagnostics, prognostics, and the use of statistical inference for causal insights (Pearl, 2009; Walhout et al., 2013), as with the increasing need to model temporal and dynamical variation. But these advances will appear through validation of the results obtained by sequencing (e.g., using an independent approach such as qPCR), followed by combinations with other omics, especially with metabolomics and metatranscriptomics.

Interpretability by non-experts is an essential consideration when ML models are put in practice by translational researchers.

[1]https://microbiomedata.org/fair/

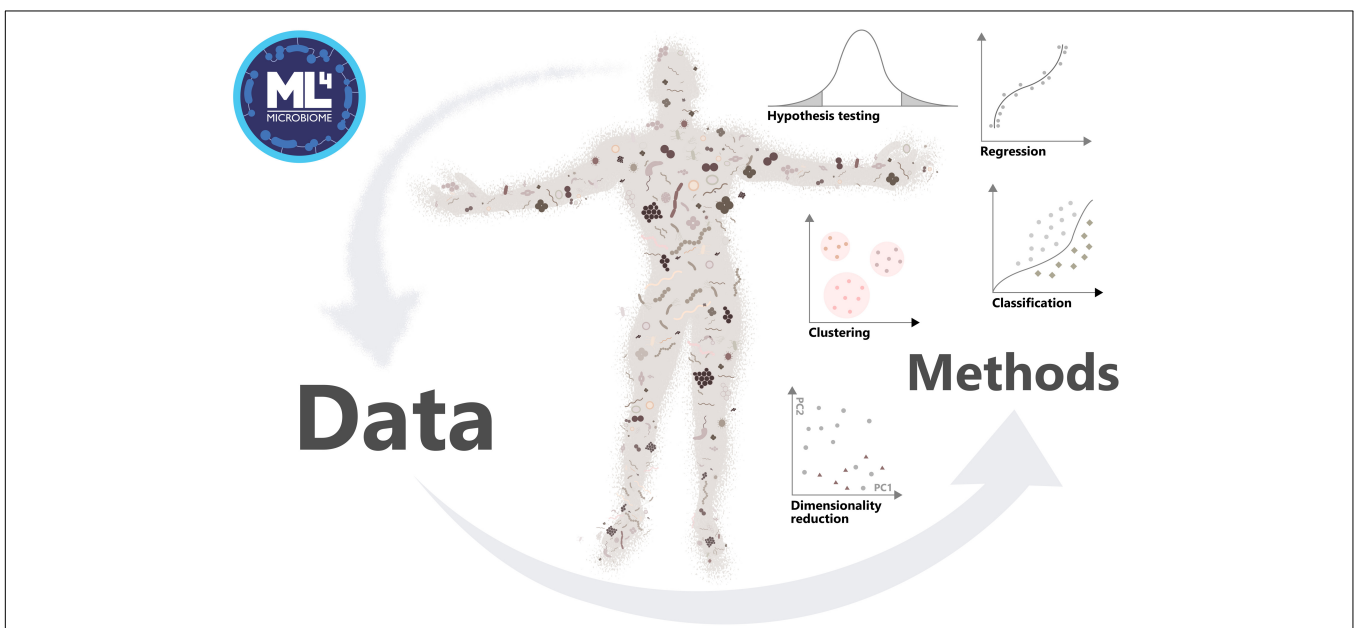


FIGURE 1 | Approach of the Action ML4Microbiome to the implementation of machine learning methods in microbiome research, driving change of the field in personalized medicine.

To overcome existing trade-offs between model interpretability and performance (Topçuoğlu et al., 2020) an active collaboration and joint education/training of researchers from statistical, biomedical and clinical fields is essential. Therefore, one main priority is the development of user-friendly tools for translational and clinical personnel, who may have limited experience with bioinformatics methods. In this line, popular software like mothur (Schloss et al., 2009, QIIME2 (Bolyen et al., 2019), and MicrobiomeAnalyst (Chong et al., 2020), the R/Bioconductor ecosystem (Qin et al., 2010), Anvi'o (Eren et al., 2015), and Biobakery (McIver et al., 2018) have incorporated ML methods into their applications in a readily usable format. Hence, the role of open source software ecosystems is critical for the overall development of the whole field. This can support and advance open collaboration networks and co-creation models that have been further complemented with open benchmark data sets (Olson et al., 2017) and reproducible notebooks (Rule et al., 2019). None of the above, however, can be achieved without multidisciplinary training of "next-generation" experts that could be integrated in clinical environments, ultimately facilitating clinical decision-making based on microbiome data as part of personalized medicine strategies (Gómez-López et al., 2019).

In order to accelerate this transition, the COST (European Cooperation in Science and Technology) Action "ML4Microbiome" (Machine Learning for Microbiome) started in 2019 with the aim to coordinate a synergistic network of the use of ML in Microbiome research at the European level. This COST Action CA18131 on *Statistical and Machine Learning Techniques in Human Microbiome Studies* is a step toward tackling the challenges by strengthening the network of European researchers in this emerging research area (**Figure 1**). A space of discussion to break down barriers of communication between fields, as well as their engagement, is being constructed through its four working groups (WG) and several networking and training events http://www.ml4microbiome.eu. It is also planned to launch a DREAM challenge[2]. DREAM challenges are crowdsourced benchmark efforts. By decoupling the method development (open to any scientist) to their evalution (by the organizers based on hold-back data, these challenges provide an unbiased and transparent assessment of methods (Saez-Rodriguez et al., 2016). Furthermore, the action ML4Microbiome identified multiple shortcomings in the current research that need to be taken into consideration. The field will benefit from increasing sample sizes, and the availability of spatial and longitudinal profiling that can be used to train more detailed and accurate models of microbiome variation. The development of interpretable and transparent ML methods will help to bridge the gap between methodological and applied fields. ML4Microbiome is open for new multi-disciplinary collaborations and collaborative ML methods development, and is welcoming researchers to participate in workshops, courses, source code/tool development aiming to promote the use of appropriate statistical and machine learning methods in metagenomics.

[2] www.dreamchallenges.org

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

IM-I, AZ, DG-C, and MC conceived the manuscript. IM-I, LL, MN, IE, and GR coordinated, supervised, and wrote the draft, the **Supplementary Information**, and the final manuscript. MA, OA, BB-G, ES, DD'E, MD, LF, AG, KH, TK, ML, LM-Z, CM, MM, PM, LP, GP, SP, VP, PP, AS, RSh, BS, RSu, JT, C-OT, BV, DV, EY, GZ, JS-R, AZ, DG-C, and MC revised draft manuscript, provided comments, included manual references, and wrote parts of the final manuscript. All the authors discussed and approved the final version of the manuscript.

FUNDING

This study was supported by the COST Action CA18131 "Statistical and machine learning techniques in human microbiome studies." IM-I was supported by the "MS type I" program (CP16/00163) from the Instituto de Salud Carlos III and co-funded by Fondo Europeo de Desarrollo Regional-FEDER. MN was grateful for the additional support by the project "Information and Communication Technologies for a Single Digital Market in Science, Education and Security" of the Scientific Research Center, NIS-3317 and National roadmaps for research infrastructures (RIs) grant number NIS-3318. LL was supported by Academy of Finland (decision 295741). IE was supported by H2020-EU.4.b. project "Integration of knowledge and biobank resources in comprehensive translational approach for personalized prevention and treatment of metabolic disorders (INTEGROMED)" (grant agreement ID 857572). MD was supported by the Luxembourg National Research Fund (FNR) CORE grant (C18/BM/12585940).

ACKNOWLEDGMENTS

We are grateful to all COST Action CA18131 "Statistical and machine learning techniques in human microbiome studies" members for their contributions to the discussion about the topics in this perspective, and especially to the WG4 and WG1.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2021.635781/full#supplementary-material

Supplementary Table 1 | Summary and main characteristics of human microbiome studies employing ML approaches.

REFERENCES

Ai, L., Tian, H., Chen, Z., Chen, H., Xu, J., and Fang, J.-Y. (2017). Systematic evaluation of supervised classifiers for fecal microbiota-based prediction of colorectal cancer. *Oncotarget* 8, 9546–9556. doi: 10.18632/oncotarget.14488

Aitchison, J. (1986). *THE statistical Analysis of Compositional Data*. New York, NY: Chapman and Hall.

Alneberg, J., Bjarnason, B. S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., et al. (2014). Binning metagenomic contigs by coverage and composition. *Nat. Methods* 11, 1144–1146. doi: 10.1038/nmeth.3103

Arango-Argoty, G., Garner, E., Pruden, A., Heath, L. S., Vikesland, P., and Zhang, L. (2018). DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome* 6:23. doi: 10.1186/s40168-018-0401-z

Arbel, J., Mengersen, K., and Rousseau, J. (2016). Bayesian nonparametric dependent model for partially replicated data: the influence of fuel spills on species diversity. *Ann. Appl. Stat.* 10, 1496–1516. doi: 10.1214/16-AOAS944

Armour, C. R., Nayfach, S., Pollard, K. S., and Sharpton, T. J. (2019). A metagenomic meta-analysis reveals functional signatures of health and disease in the human gut microbiome. *mSystems* 4:e00332-18. doi: 10.1128/mSystems.00332-18

Aryal, S., Alimadadi, A., Manandhar, I., Joe, B., and Cheng, X. (2020). Machine learning strategy for gut microbiome-based diagnostic screening of cardiovascular disease. *Hypertens. Dallas Tex* 1979, 1555–1562. doi: 10.1161/HYPERTENSIONAHA.120.15885

Asgari, E., Garakani, K., McHardy, A. C., and Mofrad, M. R. K. (2018). MicroPheno: predicting environments and host phenotypes from 16S rRNA gene sequencing using a k-mer based representation of shallow sub-samples. *Bioinform. Oxf. Engl.* 34, i32–i42. doi: 10.1093/bioinformatics/bty296

Barratt, M. J., Lebrilla, C., Shapiro, H.-Y., and Gordon, J. I. (2017). The gut microbiota, food science, and human nutrition: a timely marriage. *Cell Host Microbe* 22, 134–141. doi: 10.1016/j.chom.2017.07.006

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W. H., Ng, L. G., et al. (2019). Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* 37, 38–44. doi: 10.1038/nbt.4314

Berg, G., Rybakova, D., Fischer, D., Cernava, T., Vergès, M.-C. C., Charles, T., et al. (2020). Microbiome definition re-visited: old concepts and new challenges. *Microbiome* 8:103. doi: 10.1186/s40168-020-00875-0

Björk, J. R., Hui, F. K. C., O'Hara, R. B., and Montoya, J. M. (2018). Uncovering the drivers of host-associated microbiota with joint species distribution modelling. *Mol. Ecol.* 27, 2714–2724. doi: 10.1111/mec.14718

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., et al. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* 37, 852–857. doi: 10.1038/s41587-019-0209-9

Buffie, C. G., and Pamer, E. G. (2013). Microbiota-mediated colonization resistance against intestinal pathogens. *Nat. Rev. Immunol.* 13, 790–801. doi: 10.1038/nri3535

Buza, T. M., Tonui, T., Stomeo, F., Tiambo, C., Katani, R., Schilling, M., et al. (2019). iMAP: an integrated bioinformatics and visualization pipeline for microbiome data analysis. *BMC Bioinformatics* 20:374. doi: 10.1186/s12859-019-2965-4

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., and Holmes, S. P. (2016). DADA2: high-resolution sample inference from Illumina amplicon data. *Nat. Methods* 13, 581–583. doi: 10.1038/nmeth.3869

Chong, J., Liu, P., Zhou, G., and Xia, J. (2020). Using MicrobiomeAnalyst for comprehensive statistical, functional, and meta-analysis of microbiome data. *Nat. Protoc.* 15, 799–821. doi: 10.1038/s41596-019-0264-1

Costea, P. I., Hildebrand, F., Arumugam, M., Bäckhed, F., Blaser, M. J., Bushman, F. D., et al. (2018). Enterotypes in the landscape of gut microbial community composition. *Nat. Microbiol.* 3, 8–16. doi: 10.1038/s41564-017-0072-8

Cullen, C. M., Aneja, K. K., Beyhan, S., Cho, C. E., Woloszynek, S., Convertino, M., et al. (2020). Emerging priorities for microbiome research. *Front. Microbiol.* 11:136. doi: 10.3389/fmicb.2020.00136

Davis, N. M., Proctor, D. M., Holmes, S. P., Relman, D. A., and Callahan, B. J. (2018). Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome* 6:226. doi: 10.1186/s40168-018-0605-2

Díez López, C., Vidaki, A., Ralf, A., Montiel González, D., Radjabzadeh, D., Kraaij, R., et al. (2019). Novel taxonomy-independent deep learning microbiome approach allows for accurate classification of different forensically relevant human epithelial materials. *Forensic Sci. Int. Genet.* 41, 72–82. doi: 10.1016/j.fsigen.2019.03.015

Eetemadi, A., Rai, N., Pereira, B. M. P., Kim, M., Schmitz, H., and Tagkopoulos, I. (2020). The computational diet: a review of computational methods across diet, microbiome, and health. *Front. Microbiol.* 11:393. doi: 10.3389/fmicb.2020.00393

Eren, A. M., Esen, ÖC., Quince, C., Vineis, J. H., Morrison, H. G., Sogin, M. L., et al. (2015). Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* 3:e1319. doi: 10.7717/peerj.1319

Falony, G., Joossens, M., Vieira-Silva, S., Wang, J., Darzi, Y., Faust, K., et al. (2016). Population-level analysis of gut microbiome variation. *Science* 352, 560–564. doi: 10.1126/science.aad3503

Fernandes, A. D., Reid, J. N., Macklaim, J. M., McMurrough, T. A., Edgell, D. R., and Gloor, G. B. (2014). Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome* 2:15. doi: 10.1186/2049-2618-2-15

Gagnière, J., Raisch, J., Veziant, J., Barnich, N., Bonnet, R., Buc, E., et al. (2016). Gut microbiota imbalance and colorectal cancer. *World J. Gastroenterol.* 22, 501–518. doi: 10.3748/wjg.v22.i2.501

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., and Egozcue, J. J. (2017). Microbiome datasets are compositional: and this is not optional. *Front. Microbiol.* 8:2224. doi: 10.3389/fmicb.2017.02224

Gómez-López, G., Dopazo, J., Cigudosa, J. C., Valencia, A., and Al-Shahrour, F. (2019). Precision medicine needs pioneering clinical bioinformaticians. *Brief. Bioinform.* 20, 752–766. doi: 10.1093/bib/bbx144

Hillmann, B., Al-Ghalith, G. A., Shields-Cutler, R. R., Zhu, Q., Gohl, D. M., Beckman, K. B., et al. (2018). Evaluating the information content of shallow shotgun metagenomics. *mSystems* 3, e69–e18. doi: 10.1128/mSystems.00069-18

Holmes, I., Harris, K., and Quince, C. (2012). Dirichlet multinomial mixtures: generative models for microbial metagenomics. *PLoS One* 7:e30126. doi: 10.1371/journal.pone.0030126

Huang, R., Soneson, C., Ernst, F. G. M., Rue-Albrecht, K. C., Yu, G., Hicks, S. C., et al. (2020). TreeSummarizedExperiment: a S4 class for data with hierarchical structure. *F1000Research* 9:1246. doi: 10.12688/f1000research.26669.1

Hughes, D. A., Bacigalupe, R., Wang, J., Rühlemann, M. C., Tito, R. Y., Falony, G., et al. (2020). Genome-wide associations of human gut microbiome variation and implications for causal inference analyses. *Nat. Microbiol.* 5, 1079–1087. doi: 10.1038/s41564-020-0743-8

Juhász, J., Kertész-Farkas, A., Szabó, D., and Pongor, S. (2014). Emergence of collective territorial defense in bacterial communities: horizontal gene transfer can stabilize microbiomes. *PLoS One* 9:e0095511. doi: 10.1371/journal.pone.0095511

Kim, S., Covington, A., and Pamer, E. G. (2017). The intestinal microbiota: antibiotics, colonization resistance, and enteric pathogens. *Immunol. Rev.* 279, 90–105. doi: 10.1111/imr.12563

Knight, R., Vrbanac, A., Taylor, B. C., Aksenov, A., Callewaert, C., Debelius, J., et al. (2018). Best practices for analysing microbiomes. *Nat. Rev. Microbiol.* 16, 410–422. doi: 10.1038/s41579-018-0029-9

Knights, D., Kuczynski, J., Charlson, E. S., Zaneveld, J., Mozer, M. C., Collman, R. G., et al. (2011). Bayesian community-wide culture-independent microbial source tracking. *Nat. Methods* 8:761. doi: 10.1038/nmeth.1650

Kobak, D., and Berens, P. (2019). The art of using t-SNE for single-cell transcriptomics. *Nat. Commun.* 10:5416. doi: 10.1038/s41467-019-13056-x

Lahti, L., Salojärvi, J., Salonen, A., Scheffer, M., and de Vos, W. M. (2014). Tipping elements in the human intestinal ecosystem. *Nat. Commun.* 5:4344. doi: 10.1038/ncomms5344

LaPierre, N., Ju, C. J.-T., Zhou, G., and Wang, W. (2019). MetaPheno: a critical evaluation of deep learning and machine learning in metagenome-based disease prediction. *Methods San Diego Calif.* 166, 74–82. doi: 10.1016/j.ymeth.2019.03.003

Lederberg, J., and McCray, A. T. (2001). 'Ome sweet 'omics– a genealogical treasury of words. *Scientist* 15:8. doi: 10.1089/clinomi.03.09.05

Legendre, P., and Legendre, L. (2012). *Numerical Ecology*. Amsterdam: Elsevier.

Liao, T., Wei, Y., Luo, M., Zhao, G.-P., and Zhou, H. (2019). tmap: an integrative framework based on topological data analysis for population-scale microbiome stratification and association studies. *Genome Biol.* 20:293. doi: 10.1186/s13059-019-1871-4

Lin, C., Culver, J., Weston, B., Underhill, E., Gorky, J., and Dhurjati, P. (2018). GutLogo: agent-based modeling framework to investigate spatial and temporal dynamics in the gut microbiome. *PLoS One* 13:e0207072. doi: 10.1371/journal.pone.0207072

Lin, H., and Peddada, S. D. (2020). Analysis of compositions of microbiomes with bias correction. *Nat. Commun.* 11:3514. doi: 10.1038/s41467-020-17041-7

Liu, Y., Meric, G., Havulinna, A. S., Teo, S. M., Ruuskanen, M., Sanders, J., et al. (2020). Early prediction of liver disease using conventional risk factors and gut microbiome-augmented gradient boosting. *medRxiv* [Preprint]. doi: 10.1101/2020.06.24.20138933

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15:550. doi: 10.1186/s13059-014-0550-8

Lozupone, C. A., Stombaugh, J., Gonzalez, A., Ackermann, G., Wendel, D., Vázquez-Baeza, Y., et al. (2013). Meta-analyses of studies of the human microbiota. *Genome Res.* 23, 1704–1714. doi: 10.1101/gr.151803.112

Lynch, S. V., Ng, S. C., Shanahan, F., and Tilg, H. (2019). Translating the gut microbiome: ready for the clinic? *Nat. Rev. Gastroenterol. Hepatol.* 16, 656–661. doi: 10.1038/s41575-019-0204-0

Malla, M. A., Dubey, A., Kumar, A., Yadav, S., Hashem, A., and Abd_Allah, E. F. (2019). Exploring the human microbiome: the potential future role of next-generation sequencing in disease diagnosis and treatment. *Front. Immunol.* 9:2968. doi: 10.3389/fimmu.2018.02868

Marcos-Zambrano, L. J., Karaduzovic-Hadziabdic, K., Przymus, P., Trajkovik, V., Aasmets, O., Berland, M., et al. (2021). Applications of machine learning in human microbiome studies: a review on feature selection, biomarker identification, disease prediction and treatment. *Front. Microbiol.* doi: 10.3389/fmicb.2021.634511

McGhee, J. J., Rawson, N., Bailey, B. A., Fernandez-Guerra, A., Sisk-Hackworth, L., and Kelley, S. T. (2020). Meta-SourceTracker: application of Bayesian source tracking to shotgun metagenomics. *PeerJ* 8:e8783. doi: 10.7717/peerj.8783

McIver, L. J., Abu-Ali, G., Franzosa, E. A., Schwager, R., Morgan, X. C., Waldron, L., et al. (2018). bioBakery: a meta'omic analysis environment. *Bioinformatics* 34, 1235–1237. doi: 10.1093/bioinformatics/btx754

McMurdie, P. J., and Holmes, S. (2013). phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* 8:e61217. doi: 10.1371/journal.pone.0061217

Mehta, R. S., Abu-Ali, G. S., Drew, D. A., Lloyd-Price, J., Subramanian, A., Lochhead, P., et al. (2018). Stability of the human faecal microbiome in a cohort of adult men. *Nat. Microbiol.* 3, 347–355. doi: 10.1038/s41564-017-0096-0

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E., Kubal, M., et al. (2008). The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 9:386. doi: 10.1186/1471-2105-9-386

Mitchell, A. L., Almeida, A., Beracochea, M., Boland, M., Burgin, J., Cochrane, G., et al. (2020). MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Res.* 48, D570–D578. doi: 10.1093/nar/gkz1035

Murovec, B., Deutsch, L., and Stres, B. (2020). Computational framework for high-quality production and large-scale evolutionary analysis of metagenome assembled genomes. *Mol. Biol. Evol.* 37, 593–598. doi: 10.1093/molbev/msz237

Namkung, J. (2020). Machine learning methods for microbiome studies. *J. Microbiol.* 58, 206–216. doi: 10.1007/s12275-020-0066-8

Nayfach, S., Shi, Z. J., Seshadri, R., Pollard, K. S., and Kyrpides, N. C. (2019). New insights from uncultivated genomes of the global human gut microbiome. *Nature* 568, 505–510. doi: 10.1038/s41586-019-1058-x

Oh, M., and Zhang, L. (2020). DeepMicro: deep representation learning for disease prediction based on microbiome data. *Sci. Rep.* 10:6026. doi: 10.1038/s41598-020-63159-5

Olson, R. S., La Cava, W., Orzechowski, P., Urbanowicz, R. J., and Moore, J. H. (2017). PMLB: a large benchmark suite for machine learning evaluation and comparison. *BioData Min.* 10:36. doi: 10.1186/s13040-017-0154-4

Org, E., Parks, B. W., Joo, J. W. J., Emert, B., Schwartzman, W., Kang, E. Y., et al. (2015). Genetic and environmental control of host-gut microbiota interactions. *Genome Res.* 25, 1558–1569. doi: 10.1101/gr.194118.115

Pasolli, E., Truong, D. T., Malik, F., Waldron, L., and Segata, N. (2016). Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights. *PLoS Comput. Biol.* 12:e1004977. doi: 10.1371/journal.pcbi.1004977

Pearl, J. (2009). Causal inference in statistics: an overview. *Stat. Surv.* 3, 96–146. doi: 10.1214/09-SS057

Poussin, C., Sierro, N., Boué, S., Battey, J., Scotti, E., Belcastro, V., et al. (2018). Interrogating the microbiome: experimental and computational considerations in support of study reproducibility. *Drug Discov. Today* 23, 1644–1657. doi: 10.1016/j.drudis.2018.06.005

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., et al. (2010). A human gut microbial gene catalog established by metagenomic sequencing. *Nature* 464, 59–65. doi: 10.1038/nature08821

Qin, Y., Meric, G., Long, T., Watrous, J., Burgess, S., Havulinna, A., et al. (2020). Genome-wide association and Mendelian randomization analysis prioritizes bioactive metabolites with putative causal effects on common diseases. *medRxiv* [Preprint]. doi: 10.1101/2020.08.01.20166413

Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., and Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* 35, 833–844. doi: 10.1038/nbt.3935

Rahman, M. A., and Rangwala, H. (2020). IDMIL: an alignment-free interpretable deep multiple instance learning (MIL) for predicting disease from whole-metagenomic data. *Bioinformatics* 36, i39–i47. doi: 10.1093/bioinformatics/btaa477

Rahman, S. F., Olm, M. R., Morowitz, M. J., and Banfield, J. F. (2018). Machine learning leveraging genomes from metagenomes identifies influential antibiotic resistance genes in the infant gut microbiome. *mSystems* 3:e00123-17. doi: 10.1128/mSystems.00123-17

Reiman, D., Metwally, A. A., and Dai, Y. (2018). PopPhy-CNN: a phylogenetic tree embedded architecture for convolution neural networks for metagenomic data. *bioRxiv* [Preprint]. doi: 10.1101/257931

Roslund, M. I. Puhakka R., Grönroos, N., Nurminen, N., Oikarinen, N., Gazal, A. M., (2020). Biodiversity intervention enhances immune regulation and health-associated commensal microbiota among daycare children. *Sci. Adv.* 6:eaba2578. doi: 10.1126/sciadv.aba2578

Rule, A., Birmingham, A., Zuniga, C., Altintas, I., Huang, S.-C., Knight, R., et al. (2019). Ten simple rules for writing and sharing computational analyses in Jupyter Notebooks. *PLoS Comput. Biol.* 15:e1007007. doi: 10.1371/journal.pcbi.1007007

Saez-Rodriguez, J., Costello, J. C., Friend, S. H., Kellen, M. R., Mangravite, L., Meyer, P., et al. (2016). Crowdsourcing biomedical research: leveraging communities as innovation engines. *Nat. Rev. Genet.* 17, 470–486. doi: 10.1038/nrg.2016.69

Salosensaari, A., Laitinen, V., Havulinna, A. S., Meric, G., Cheng, S., Perola, M., et al. (2020). Taxonomic signatures of long-term mortality risk in human gut microbiota. *medRxiv* [Preprint]. doi: 10.1101/2019.12.30.19015842

Sampson, T. R., Debelius, J. W., Thron, T., Janssen, S., Shastri, G. G., Ilhan, Z. E., et al. (2016). Gut microbiota regulate motor deficits and neuroinflammation in a model of Parkinson's disease. *Cell* 167, 1469.e12–1480.e12. doi: 10.1016/j.cell.2016.11.018

Sankaran, K., and Holmes, S. (2014). structSSI: simultaneous and selective inference for grouped or hierarchically structured data. *J. Stat. Softw.* 59, 1–21. doi: 10.18637/jss.v059.i13

Sankaran, K., and Holmes, S. P. (2019). Latent variable modeling for the microbiome. *Biostat. Oxf. Engl.* 20, 599–614. doi: 10.1093/biostatistics/kxy018

Sanna, S., van Zuydam, N. R., Mahajan, A., Kurilshikov, A., Vich Vila, A., Võsa, U., et al. (2019). Causal relationships among the gut microbiome, short-chain fatty acids and metabolic diseases. *Nat. Genet.* 51, 600–605. doi: 10.1038/s41588-019-0350-x

Schmidt, T. S. B., Raes, J., and Bork, P. (2018). The human gut microbiome: from association to modulation. *Cell* 172, 1198–1215. doi: 10.1016/j.cell.2018.02.044

Schmitt, S., Tsai, P., Bell, J., Fromont, J., Ilan, M., Lindquist, N., et al. (2012). Assessing the complex sponge microbiota: core, variable and species-specific bacterial communities in marine sponges. *ISME J.* 6, 564–576. doi: 10.1038/ismej.2011.116

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., et al. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* 75, 7537–7541. doi: 10.1128/AEM.01541-09

Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W. S., et al. (2011). Metagenomic biomarker discovery and explanation. *Genome Biol.* 12:R60. doi: 10.1186/gb-2011-12-6-r60

Shenhav, L., Thompson, M., Joseph, T. A., Briscoe, L., Furman, O., Bogumil, D., et al. (2019). FEAST: fast expectation-maximization for microbial source tracking. *Nat. Methods* 16, 627–632. doi: 10.1038/s41592-019-0431-x

Shetty, S. A., and Lahti, L. (2019). Microbiome data science. *J. Biosci.* 44:115.

Singh, R. K., Chang, H.-W., Yan, D., Lee, K. M., Ucmak, D., Wong, K., et al. (2017). Influence of diet on the gut microbiome and implications for human health. *J. Transl. Med.* 15:73. doi: 10.1186/s12967-017-1175-y

Sze, M. A., and Schloss, P. D. (2018). Leveraging existing 16S rRNA gene surveys to identify reproducible biomarkers in individuals with colorectal tumors. *mBio* 9:e00630-18. doi: 10.1128/mBio.00630-18

Tamames, J., Cobo-Simón, M., and Puente-Sánchez, F. (2019). Assessing the performance of different approaches for functional and taxonomic annotation of metagenomes. *BMC Genomics* 20:960. doi: 10.1186/s12864-019-6289-6

Tamburini, S., Shen, N., Wu, H. C., and Clemente, J. C. (2016). The microbiome in early life: implications for health outcomes. *Nat. Med.* 22, 713–722. doi: 10.1038/nm.4142

ten Hoopen, P., Finn, R. D., Bongo, L. A., Corre, E., Fosso, B., Meyer, F., et al. (2017). The metagenomic data life-cycle: standards and best practices. *GigaScience* 6:gix047. doi: 10.1093/gigascience/gix047

Topçuoğlu, B. D., Lesniak, N. A., Ruffin, M. T., Wiens, J., and Schloss, P. D. (2020). A framework for effective application of machine learning to microbiome-based classification problems. *mBio* 11:e00434-20. doi: 10.1128/mBio.00434-20

Treangen, T. J., Koren, S., Sommer, D. D., Liu, B., Astrovskaya, I., Ondov, B., et al. (2013). MetAMOS: a modular and open source metagenomic assembly and analysis pipeline. *Genome Biol.* 14:R2. doi: 10.1186/gb-2013-14-1-r2

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., and Gordon, J. I. (2007). The human microbiome project. *Nature* 449, 804–810. doi: 10.1038/nature06244

Walhout, M., Vidal, M., and Dekker, J. (2013). *Handbook of Systems Biology*. Amsterdam: Elsevier.

Wang, Y., and Kasper, L. H. (2014). The role of microbiome in central nervous system disorders. *Brain. Behav. Immun.* 38, 1–12. doi: 10.1016/j.bbi.2013.12.015

Washburne, A. D., Silverman, J. D., Leff, J. W., Bennett, D. J., Darcy, J. L., Mukherjee, S., et al. (2017). Phylogenetic factorization of compositional data yields lineage-level associations in microbiome datasets. *PeerJ* 5:e2969. doi: 10.7717/peerj.2969

Washburne, A. D., Silverman, J. D., Morton, J. T., Becker, D. J., Crowley, D., Mukherjee, S., et al. (2019). Phylofactorization: a graph partitioning algorithm to identify phylogenetic scales of ecological data. *Ecol. Monogr.* 89:e01353. doi: 10.1002/ecm.1353

Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., et al. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* 5:27. doi: 10.1186/s40168-017-0237-y

Zeevi, D., Korem, T., Godneva, A., Bar, N., Kurilshikov, A., Lotan-Pompan, M., et al. (2019). Structural variation in the gut microbiome associates with host health. *Nature* 568, 43–48. doi: 10.1038/s41586-019-1065-y

Zhernakova, A., Kurilshikov, A., Bonder, M. J., Tigchelaar, E. F., Schirmer, M., Vatanen, T., et al. (2016). Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science* 352, 565–569. doi: 10.1126/science.aad3369

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

REVIEW
published: 11 May 2021
doi: 10.3389/fmicb.2021.618856

Computational Biology and Machine Learning Approaches to Understand Mechanistic Microbiome-Host Interactions

Padhmanand Sudhakar[1,2,3], Kathleen Machiels[1], Bram Verstockt[1,4], Tamas Korcsmaros[2,3] and Séverine Vermeire[1,4]*

[1] Department of Chronic Diseases, Metabolism and Ageing, Translational Research Center for Gastrointestinal Disorders (TARGID), KU Leuven, Leuven, Belgium, [2] Earlham Institute, Norwich, United Kingdom, [3] Quadram Institute Bioscience, Norwich, United Kingdom, [4] Department of Gastroenterology and Hepatology, University Hospitals Leuven, KU Leuven, Leuven, Belgium

OPEN ACCESS

Edited by:
Isabel Moreno Indias, University of Málaga, Spain

Reviewed by:
Zhili He, University of Oklahoma, United States
Christopher L. Hemme, University of Rhode Island, United States
Swagatika Sahoo, Indian Institute of Technology Madras, India

***Correspondence:**
Padhmanand Sudhakar
padhmanand.sudhakar@kuleuven.be
orcid.org/0000-0003-1907-4491

Specialty section:
This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology

Received: *18 October 2020*
Accepted: *19 March 2021*
Published: *11 May 2021*

Citation:
Sudhakar P, Machiels K, Verstockt B, Korcsmaros T and Vermeire S (2021) Computational Biology and Machine Learning Approaches to Understand Mechanistic Microbiome-Host Interactions. Front. Microbiol. 12:618856. doi: 10.3389/fmicb.2021.618856

The microbiome, by virtue of its interactions with the host, is implicated in various host functions including its influence on nutrition and homeostasis. Many chronic diseases such as diabetes, cancer, inflammatory bowel diseases are characterized by a disruption of microbial communities in at least one biological niche/organ system. Various molecular mechanisms between microbial and host components such as proteins, RNAs, metabolites have recently been identified, thus filling many gaps in our understanding of how the microbiome modulates host processes. Concurrently, high-throughput technologies have enabled the profiling of heterogeneous datasets capturing community level changes in the microbiome as well as the host responses. However, due to limitations in parallel sampling and analytical procedures, big gaps still exist in terms of how the microbiome mechanistically influences host functions at a system and community level. In the past decade, computational biology and machine learning methodologies have been developed with the aim of filling the existing gaps. Due to the agnostic nature of the tools, they have been applied in diverse disease contexts to analyze and infer the interactions between the microbiome and host molecular components. Some of these approaches allow the identification and analysis of affected downstream host processes. Most of the tools statistically or mechanistically integrate different types of -omic and meta -omic datasets followed by functional/biological interpretation. In this review, we provide an overview of the landscape of computational approaches for investigating mechanistic interactions between individual microbes/microbiome and the host and the opportunities for basic and clinical research. These could include but are not limited to the development of activity- and mechanism-based biomarkers, uncovering mechanisms for therapeutic interventions and generating integrated signatures to stratify patients.

Keywords: health, disease, microbiome-host interactions, molecular mechanisms, computational approaches, machine learning, basic and clinical research

INTRODUCTION: MICROBIOME-HOST INTERACTIONS

Across different niches and ecosystems, micro-organisms including bacteria, viruses, archaea inhabit a wide range of hosts (Braga et al., 2016). This community of microbes imparts various functions such as making nutrients accessible to the host (Martin et al., 2019), modulating the host immune system (Mendes et al., 2019), warding off pathogens (Pickard et al., 2017), maintaining homeostasis (Ohland and Jobin, 2015; Penny et al., 2018) among others. These functions are in turn driven primarily by molecular interactions between microbial and host molecules such as proteins, RNA and metabolites (Hughes and Sperandio, 2008; Braga et al., 2016). Deciphering these interactions could not only reveal the microbe-host cross-talk but also provide us with insights into formulating therapeutic strategies aimed at maintaining health and/or ameliorating disease states. The past decades have witnessed a surge in research interest to study microbial communities (and their interactions) which inhabit various niches – from the gut to the soil ecosystem. This was made possible by technological advancements leading to plummeting costs of 16S and metagenomic sequencing, higher sequencing depth and resolution (Levy and Myers, 2016; Jacob et al., 2019; Valli et al., 2020), novel *in vitro* systems (Shah et al., 2016; Eain et al., 2017; May et al., 2017), and new methodologies for high-throughput profiling of multiple -omic data types such as metaproteomics, metabolomics, lipidomics (Muller et al., 2013; Roume et al., 2015). However, due to many other limitations related to scale, scope, feasibility and sample availability for parallel omic read -outs, experimentally determining the inter-species microbe-host interactions is a challenging task (Fritz et al., 2013). Computational methods can overcome some of these limitations thereby enhancing our understanding of microbe-host interactions (Dix et al., 2016). In this review, we outline some key concepts, tools, and methods involved in computationally inferring the molecular mechanisms mediating microbe-host interactions.

BIOLOGICAL NETWORKS: CONCEPTS AND APPLICATIONS

Biological networks represent relationships (termed edges) between any two biological entities (species, organisms, and molecules, etc.) which are usually called as nodes. At the level of molecules (genes, proteins, metabolites, RNAs, and small molecules, etc.), biological networks could either denote the physical interactions (e.g., protein–protein, protein-DNA, and RNA-protein, etc.) between molecules or any measure of association (e.g., co-expression and co-occurrence) between molecules (Gosak et al., 2018). In this paper, we will refer only to physical interactions. Physical interactions can be classified based on various criteria such as molecular types (protein–protein, protein-DNA, and RNA-protein, etc.), experimental scale (high-throughput or low-throughput), source (experimentally determined or computationally predicted), directionality (directed or undirected), relational signs (positive or negative relationships) and coverage (genome-wide or targeted). Since biological networks provide the larger context in which genes or proteins tend to exert their action, researchers can thereby fine-tune their hypotheses. Networks have largely been used in the domain of biological sciences (a) as a scaffold to integrate either singular or multiple contextual -omic datasets such as gene expression, proteomics, etc., measured in response to intrinsic or extrinsic stimuli (Charitou et al., 2016), (b) as a graph to trace potential signaling and regulatory pathways connecting any two nodes (Azeloglu and Iyengar, 2015), (c) to perform functional analysis at a local or global level (Emmert-Streib and Glazko, 2011), (d) to reconstruct the networks of non-model organisms from those of model organisms (Thompson et al., 2015), (e) to discover drug and disease targets (Huang et al., 2018), and (f) to infer globally or locally conserved signatures such as modules, motifs, etc (Wong et al., 2012). Various resources of molecular interactions and tools for integrative network analysis have been compiled and developed by the research community of network biologists. Since a very detailed description of the resources and tools is out of scope of the current review, readers are hereby referred to Pedamallu and Ozdamar (2014), Miryala et al. (2018), Romano et al. (2019).

Due to their utility in capturing contextual backgrounds and communication between molecular entities, biological networks have been used to not only study intra-species interactions but also inter-species cross-talks. Molecular ecological networks (Deng et al., 2012; Heleno et al., 2014) are a case in point by which the concept of networks are used to study the interactions between molecules (derived from different species or even kingdoms) in a larger ecological context (Yang et al., 2017; Meyer et al., 2020; Yu et al., 2020; Zheng et al., 2020). At the very core of it, a typical molecular ecological network inference workflow (Zhou et al., 2010; Deng et al., 2012; Chen et al., 2017) starts with the generation of meta -omic datasets (such as metagenomics, metatranscriptomics, and metaproteomics, etc.) followed by differential abundance testing between samples from contrasting conditions. Various measures of correlations and associations can then be applied to determine the distance between samples based on the differences and similarities in terms of the molecular features measured in the -omic datasets across the sample classes. Such correlations or associations can be used as a primary point of reference to investigate the possibility of mechanistic interactions which could in turn be driving the associative relationships. Furthermore, a network based representation of the feature-space can be used to compare samples with each other or to associate network properties such as the presence of motifs and modules to higher-level ecological traits/phenotypes. However, since molecular ecological networks do not directly infer molecular mechanisms which is the topic of this review, a detailed discussion on the topic is not undertaken.

COMPUTATIONAL METHODS IN MICROBIOME-HOST INTERACTIONS: FILLING THE GAPS

Computational methods bring in various advantages to the analysis of interactions between the host and individual microbes

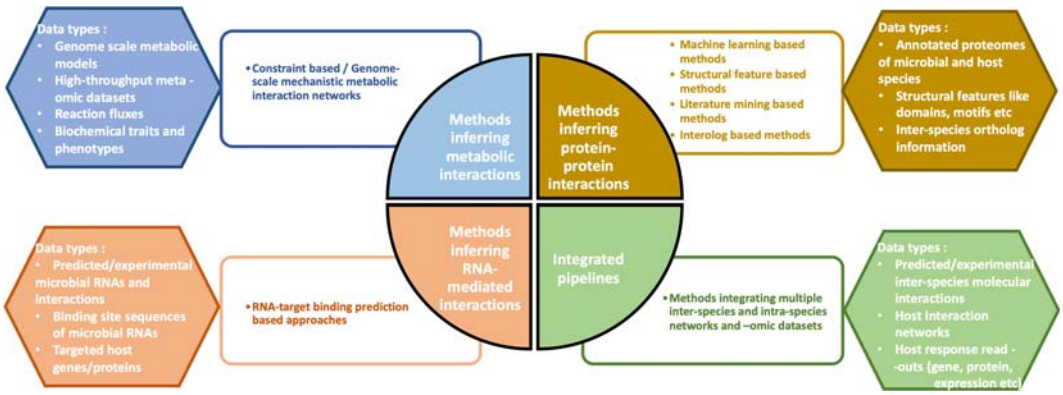


FIGURE 1 | Overview of the four different categories of computational methods which help infer the molecular mechanisms of microbe-host interactions. Some examples of data types corresponding to each of the four methods are depicted.

and/or the microbial community. These include their attributes of (a) enhancing scalability, i.e., perform the computational inferences for a large number of variables and samples, (b) improving reproducibility (if complemented by inter-operability, automation, proper version control and sufficient documentation), (c) assessing performance by using a series of metrics, (d) shortlisting and prioritizing interactions, (e) and thereby (f) enabling the fine-tuning of hypothesis for experimental and/or epidemiological studies. Although most of the methods hitherto have focused on inferring the interactions between individual microbial species (mostly well studied pathogens) and the host, a few methods have been developed to predict the interactions at a community level. In principle, many of the methods which have been used to infer interactions of single species can be scaled up (with appropriate modifications) to infer community level interactions.

CLASSIFICATION OF COMPUTATIONAL METHODS IN MICROBIOME-HOST INTERACTIONS

From a mechanistic view-point, the most widely studied interaction types in interspecies cross-talks include (a) microbial metabolite-mediated networks, (b) protein–protein interactions (PPIs), and (c) RNA-mediated interactions. Accordingly, many of the computational methods developed to investigate microbe-host interactions have focused on the three above-mentioned interaction types (**Figure 1**). As a fourth method approach, integrated pipelines combine multiple microbial and host -omic data types and networks to infer the cumulative functional effects of inter-species interactions/communication on the host.

Approaches Inferring Mechanistic Metabolic Interactions

The metabolomic layer (which comprises the enzymes, metabolites, and the reactional interactions between them) has a prominent influence on both health and disease states associated with alterations in microbiota composition (Wong et al., 2016; Martinez et al., 2017). Metabolic networks can thus represent and capture the underlying mechanisms driving various phenotypes (Pey et al., 2013; Samal et al., 2017; Zampieri et al., 2019). Computational approaches aimed at inferring the microbe-host co-metabolic networks can be classified into three prominent categories namely (a) Community-wide metabolic network modeling using metagenomic datasets: this approach is based on the assumption that the metagenomic read-outs represent the gene-distribution structure of the entire microbial community. The autonomy of species – i.e., information about which gene is derived from which species, are disregarded. Thus, the metabolic network reconstructed using this approach consists of relationships (reactions) catalyzed by enzymes (encoded by the measured genes) between molecular entities (metabolites) at a community level. (b) High throughput data driven approaches using metabolic datasets – this data-driven methodology uses targeted or untargeted profiling of metabolites from different groups of samples. Subsequently, multi-variate modeling methods and various statistical methods including simple PCAs are applied to identify biomarkers which distinguish different sample groups from each other. (c) Genome scale reconstruction applying constraint-based modeling approaches which are described below. The first two methods do not provide direct mechanistic insights and hence are not covered further in this review.

Genome-scale reconstruction models provide mechanistic information by integrating multiple inputs. These inputs include the curated genome scale metabolic models of both the host and microbial species, high-throughput meta -omic datasets including metabolites, reaction fluxes, biochemical traits and accessory phenotypic data. However, due to the strenuous nature of various steps involved in constructing the models and in scaling it up to multiple species or multiple hosts, only a handful of studies have applied this concept to infer microbe-host co-metabolic interactions (**Table 1**). The AGORA (assembly of gut organisms through reconstruction and analysis) collection is a resource of genome-scale metabolic models for 773 human gut bacterial species using a combination of metagenomics and

experimental data from literature. Furthermore, the framework employed by AGORA is amenable to scale-up given its easy adaptability to novel species of interest. AGORA also serves as a source of genome scale metabolic models reconstructed in a standardized manner. Thus, various studies have in turn used the genome scale models from the AGORA resource to construct context-specific models (Bauer et al., 2017; Bunesova et al., 2018; Tramontano et al., 2018; Pryor et al., 2019; Yilmaz et al., 2019). Recently, the authors of AGORA and their collaborators extended the framework to 7206 strains by incorporating information on the drug-metabolizing potential of the bacterial strains (Heinken et al., 2020).

The reported studies on genome-scale reconstruction models have been distributed across many different ecological contexts such as the human and rumen gut ecosystems (Islam et al., 2019), microbe-plant interactions, human alveolar macrophages, the effect of viral demands on the metabolism of human macrophages, microbe-host interactions in Parkinson's Disease to name a few. Due to the mechanistic nature of such models, they can be used as a template for further integrating other -omic datasets. This not only refines the models thereby increasing their predictive power but also assigns contextuality.

By incorporating the individual reconstructed metabolic models of tomato (*Solanum lycopersicum*) and the tomato late blight pathogen *Phytophthora infestans*, Rodenburg et al. (2019) pointed out specific pathways which mediate the dependencies of the pathogen on the metabolism of *S. lycopersicum*. The individual metabolic models for *S. lycopersicum* and *P. infestans* were derived by manually adding reactions and sub-cellular localization of metabolites and reactions (based on curation of literature) to the corresponding genome-scale models. Furthermore, by over-laying dual RNA-seq transcriptomic datasets from the host-pathogen duo into the co-metabolic network, various metabolic changes characterizing the scavenging nature of *P. infestans* were revealed. A similar study was performed in a mammalian setting wherein co-metabolic interactions and metabolic exchanges were inferred between the respiratory pathogen *Mycobacterium tuberculosis* and human alveolar macrophages (Bordbar et al., 2010). The metabolic model for the alveolar macrophages was derived from Recon1, the global human metabolic model (Thiele et al., 2013b). Briefly, a curated version of Recon1 was overlaid with gene expression data for healthy, inactivated alveolar macrophages and combined with information on flux limits for major pathways of central metabolism and a host of heterogeneous datasets such as immunohistological staining, transporter proteins, etc (Bordbar et al., 2010). The macrophage model was then combined with that of *Francisella tularensis* and corrected for compartment-specific reactions and metabolites. Unsurprisingly, given the advancement in terms of data generated and metabolic models made available, most of the genome-scale metabolic reconstruction studies (**Table 1**) were carried out for the gut ecosystem (Heinken et al., 2013; Heinken and Thiele, 2015; Ding et al., 2016; Islam et al., 2019).

Other microbe-host co-metabolic studies have been performed using publicly available tools based on constraint-based modeling approaches. The Constraint-based reconstruction and analysis (COBRA) toolbox (Heirendt et al., 2019) is one such compendium of methods containing various user-guided steps to reconstruct genome-scale metabolic models. It is characterized by properties such as interoperability, customized reconstruction, modeling, visualization, modeling, simulation, and integration of -omic datasets in various contexts (compartments, cell-types, etc.). By harnessing these properties, researchers have used the COBRA toolbox to model and investigate microbe-host metabolic interactions (Heinken et al., 2013; Thiele et al., 2013a) in the context of mammalian health with implications on human health. A representative study of the gut ecosystem using the COBRA toolbox integrated two previously published constraint-based models of mouse and a gut commensal *Bacteroides thetaiotaomicron* (Heinken et al., 2013). The *B. thetaiotaomicron* model was generated by the manual curation of a seed model produced by Model Seed (Henry et al., 2010) from the genome sequence annotated using RAST (Aziz et al., 2008) (which is a prokaryotic genome annotation tool). The mouse metabolic model was compiled by integrating a previously annotated and reconstructed model with gene essentiality data from experiments followed by corrections for duplicate reactions. The two models were then brought together by setting rules based on the subcellular localization of metabolites and reactions. The integrated metabolic model could capture many of the phenotypes exhibited *in vivo* namely the dependence of *B. thetaiotaomicron* on glycans derived from the metabolism of the host as well as the host diet itself (Heinken et al., 2013). It is noteworthy to mention that the authors also introduced novel methodologies such as Pareto analysis to complement the power of the COBRA toolbox. Pareto analysis is a bi-objective linear programming-based methodology which enables the analysis and identification of growth dependencies and trade-offs between the microbe and the host as captured by their metabolic networks.

TABLE 1 | Studies using genome-scale metabolic models and constraint based approaches to infer mechanistic co-metabolic interactions between microbial and host species.

Study	Context
Rodenburg et al. (2019)	Integrated metabolic model of *P. infestans* infecting tomato (*S. lycopersicum*)
Islam et al. (2019)	Genome-scale metabolic model between key members in the rumen microbiome and the viral phages
Hertel et al. (2019)	Integrated constraint-based model revealing microbe-host interactions in Parkinson's Disease
Aller et al. (2018)	Genome-scale model integrating biochemical demands arising from virus production and human macrophage cell metabolism
Ding et al. (2016)	Simulation of co-metabolic model of different enteropathogens in response to various host environments
Heinken and Thiele (2015)	*In silico* microbe-host gut co-metabolic model to predict effects of different host dietary schemes
Heinken et al. (2013)	Experimentally validated gut co-metabolic model between commensal bacterium *B. thetaiotaomicron* and mouse
Bordbar et al. (2010)	*Francisella tularensis* infecting human alveolar macrophage supported by high-throughput data from infected conditions

A similar study (Hertel et al., 2019) was performed using the COBRA toolbox in conjunction with other supplementary tools such as the Microbiome Modeling Toolbox (Baldini et al., 2019) which can integrate the individual reconstructed models together into one reconstructed model in addition to other useful properties (such as inferring interactions by taxa, reconstruction of pairwise/community co-metabolic networks, compartment-based modeling, pareto analysis, and various downstream operations) to extend the constraint-based modeling framework. The study integrated the microbiome and longitudinal metabolomic datasets from patients with Parkinson's disease (Hertel et al., 2019). This microbiome-host -omic integration study provided clues as to how alterations in particular co-metabolized pathways (by both the host and microbiome) such as sulfur metabolism could contribute to the varying severity of the disease. In particular, the authors were able to identify that changes in the co-metabolized pathways could be driven by particular members of the gut microbiota. This opens up possibilities to design gut microbiome-based therapies to treat or even prevent Parkinson's disease.

Approaches Inferring Protein–Protein Interactions (PPIs)

Protein–protein interactions are one of the most well-studied interaction types mediating inter-species communication (Schweppe et al., 2015). Accordingly, a large number of computational microbe-host interaction studies have focused on PPIs. Congruently, PPI-based approaches have also been propelled by the adoption of concepts from other domains of computational biology and computational sciences in general. Hence, PPI-based approaches can be sub-classified into four predominant methods (**Table 2**) depending on the concepts used (1) Machine learning based PPI methods, (2) Structural feature based PPI methods, (3) Data/Literature mining based PPI methods, and (4) Interolog based PPI methods. In this section, we provide a brief overview of the concepts involved in each of these methods (**Table 2**) and provide a few representative examples.

Structural Feature Based PPI Methods

Interactions between proteins are usually a by-product of physical interactions between structural features of the proteins and/or could be characterized indirectly by co-occurring functional features of the proteins (Ding and Kihara, 2018). Structural features of the proteins include their domain and motif architectures/compositions, amino acid composition and frequencies, post-translational modification signatures, amino acid k-mers, mimicry motifs and 3D structural properties (Ding and Kihara, 2018). Structural feature-based PPI prediction, applied initially for intra-species PPIs, was subsequently extended to inter-species studies. Essentially, the fundamental principle on which structural feature-based PPI prediction methods work involves the use of mechanistic evidence between structural features to identify potentially interacting proteins. These could include for example interactions between domains, between domains and motifs, post-translational modifications and pairwise structural similarity (Ding and Kihara, 2018). Such structural studies have been confined to considerably well studied

TABLE 2 | Computational approaches and methods inferring protein–protein interactions mediating inter-kingdom cross-talk between microbial and host organisms.

Method and corresponding studies	Reported use-case (host-microbe)
Machine learning based methods	
Leite et al. (2018)	Bacteria–phage
Tastan et al. (2009); Qi et al. (2010), Dyer et al. (2011); Nouretdinov et al. (2012), Shoombuatong et al. (2012); Mei (2013), Hongjaisee et al. (2019)	Human–HIV
Kshirsagar et al. (2013)	Human–*F. tularensis,* Human–*Y. pestis,* Human–*B. anthracis,* Human-*S. typhi*
Wuchty (2011)	Human–*Plasmodium falciparum*
Kösesoy et al. (2019)	Human–*Y. pestis,* Human–*B. anthracis*
Cui et al. (2012); Emamjomeh et al. (2014), Kim et al. (2017)	Human–Hepatitis C virus
HOPITOR (Basit et al., 2018)	Generic (Human–virus PPIs)
Liao et al. (2011)	Human–*Schistosoma japonicum*
Mei et al. (2018); Sun et al. (2018)	Human–*Francisella tularensis*
Kargarfard et al. (2016)	3 hosts and 674 influenza strains
Cui et al. (2012); Dong et al. (2015), Kim et al. (2017)	Human–Human papillomavirus
Lai et al. (2012)	Human–Influenza A virus
Mei and Zhu (2014a)	Human–HTLV retroviruses
Mei and Zhu (2014b)	Human–*Salmonella*
Lian et al. (2019)	Human–*Y. pestis*
Structural feature based methods (features used)	
Dyer at al. (2007) (DDI)	Human–*Plasmodium falciparum*
Nourani et al. (2016) (DDI)	Human–multiple viruses
Sudhakar et al. (2019) (DDI and DMI)	Human–multiple bacterial pathogens
Doolittle and Gomez (2011) (PSS)	Human–Dengue virus, *Aedes aegypti*–Dengue virus
Cui et al. (2016) (PSS)	Human–HIV, Human–*Francisella tularensis*
P-HIPSTer (Lasso et al., 2019) (PSS)	Human–multiple viruses
Chen at al. (2019) (PSS)	Human–Dengue virus 2, Human–West Nile virus
Guven-Maiorov et al. (2017) (Mimicry)	Human–*Helicobacter pylori*
Mahajan and Mande (2017) (DDI)	Human–*Francisella tularensis*
Zhang et al. (2017a) (DMI)	Grass carp–Grass carp reovirus
Mehrotra et al. (2017) (PSS, DDI, and localization)	Human–*Leptospira interrogans*, Human–*Leptospira biflexa*
Halehalli and Nagarajaram (2015) (DDI, DMI)	Human–multiple viruses
SugarBindDB (Mariethoz et al., 2016) (glycan mediated PPIs)	Generic
Rajasekharan et al. (2013) (PSS)	Human–Chandipura virus
Carducci et al. (2010) (DDI)	Human–papillomavirus type 16
Franzosa and Xia (2011) (PSS and sequence identity)	Human–multiple viruses
Sahu et al. (2014) (DDI)	Arabidopsis-*Pseudomonas syringae*
Zhou et al. (2018) (DDI)	Human–Dengue virus, *Aedes aegypti*–Dengue virus
Kim et al. (2017) (DDI)	Human–multiple viruses
Kerr et al. (2015) (Computational docking)	Human–Dengue virus 2, Human–West Nile virus
Evans et al. (2009) (DMI)	Human–HIV
Doxey and McConkey (2013) (Mimicry)	Human–*Francisella tularensis*

(Continued)

TABLE 2 | Continued

Method and corresponding studies	Reported use-case (host-microbe)
Mei and Zhang (2020) (Mimicry)	Human-*S. typhimurium* and Human-Human respiratory syncytial virus
Data/Literature mining based methods	
Thieu et al. (2012)	Generic
Viruses.STRING (Cook et al., 2018)	319 hosts and 239 viruses
Li et al. (2018)	Human–Epstein-Barr virus
Saik et al. (2016)	Human–Hepatitis C virus
García-Pérez et al. (2018)	Human–Influenza A virus
"Interolog" based methods	
Krishnadev and Srinivasan (2008); Lee et al. (2008)	Human–*Plasmodium falciparum*
Krishnadev and Srinivasan (2011)	Human–*E. coli,* Human–*S. typhimurium,* Human–*Y. pestis*
Tyagi et al. (2009)	Human–*Helicobacter pylori*
Cui et al. (2016)	Human–HIV, Human–*Francisella tularensis*
Schleker et al. (2012)	Human–*Salmonella, Salmonella–A. thaliana*
Li et al. (2012)	*A. thaliana–Ralstonia solanacearum*
Wallqvist et al. (2017)	Human–*Coxiella burnetii*
Cuesta-Astroz et al. (2019)	Human and 15 eukaryotic parasites
Zhou et al. (2014); Cui et al. (2016)	Human–*Francisella tularensis*
Barh et al. (2013)	Human–*Corynebacterium pseudotuberculosis*, Human–*Corynebacterium diphtheriae*, Human–*Francisella tularensis*, Human–*Corynebacterium ulcerans*, Human–*Y. pestis*, and Human–*E. coli*

DDI, domain–domain interaction; DMI, domain-motif interaction; PSS, pairwise structural similarity. ***Supplementary Table 1*** *provides further details into the novelty of the methods and results.*

species pairs involving H. sapiens and prominent viral and bacterial pathogens (**Table 2**). Along with pairwise structural similarity-based methods using 3D protein complexes, domain–domain interaction (DDI) and domain-motif interaction (DMI) based methods are one of the most commonly used methods within the structural feature based methodological framework for predicting inter-species PPIs. Due to the ease of annotating domains and motifs, DDI- and DMI-based methods have been harnessed widely (**Table 2**). While DDI based methods have been applied to infer PPIs for a large number of species-pairs including Human–*Plasmodium falciparum* (Dyer et al., 2007), Human–*Francisella tularensis* (Zhou et al., 2013; Mahajan and Mande, 2017), Human–*Leptospira interrogans* (Mehrotra et al., 2017), Human–*Leptospira biflexa* (Mehrotra et al., 2017), Human–papillomavirus type 16 (Carducci et al., 2010), Arabidopsis–*Pseudomonas syringae* (Sahu et al., 2014), Rice–*Xanthomonas oryzae* (Kim et al., 2008), they have the inherent disadvantage of not being able to explicitly discern directionality.

On the other hand, DMIs provide directionality for PPIs, thus indicating the flow of signal transduction (Akiva et al., 2012; Gibson et al., 2015). For example, if a microbial protein A contains a domain known to be interacting with a motif on the host protein B, it is graphically represented as A > B, translating into "microbial protein A modulates host protein B." Due to their specificity, DMI-based methods are preferred over DDI based methods for research questions seeking to answer the role of post-translational modifications elicited on host proteins by microbial proteins or vice versa. However, due to the short sequence length of protein sequence motifs, even the most stringent search strategies have the tendency to result in thousands of false-positive hits while performing motif searches on a proteome-wide basis (Perkins et al., 2010; Idrees et al., 2018). Therefore, proper quality controls need to be applied to filter out false-positives based on structural properties such as the occurrence of truly interacting motifs within disordered regions and outside globular domains (Perkins et al., 2010; Idrees et al., 2018; **Figure 2**).

Several studies (**Table 2**) have been conducted to apply the principles of DMIs to predict PPIs for multiple microbe-host species-combinations including grass carp-grass carp reovirus (Zhang et al., 2017a), human-multiple bacterial pathogens (Sudhakar et al., 2019) and human-multiple viruses (Evans et al., 2009; Halehalli and Nagarajaram, 2015). By integrating DMI predictions between grass carp and grass carp reovirus (GCRV) proteins with differential gene expression and tissue-specific gene expression followed by functional enrichment, Zhang et al. (2017a) were able to pinpoint several signaling pathways modulated by GCRV. The authors also highlight an enrichment of host genes expressed in the intestinal niche suggesting that GCRV might have a higher influence on the gut. Recently, we conducted a study (Sudhakar et al., 2019) using DDI and DMI based methods to identify cross-talks between several bacterial pathogens including *Salmonella* and autophagy – a prominent biological process involved in host cellular homeostasis. Firstly, to identify microbial proteins targeted by selective autophagy, we scanned the bacterial proteins for the presence of the recognition motifs corresponding to the selective autophagy receptors p62 and NDP52 and the autophagy adapter protein LC3. Conversely, to infer the modulation of host autophagy by the bacterial pathogens, DMI and DDI based methods were used to identify the bacterial proteins which are able to bind to/modulate the 37 core autophagy host proteins. By overlapping the two above-mentioned sets of predictions, bacterial proteins involved in interplays were identified. Such bacterial proteins are also targeted by the host autophagy machinery for clearance and degradation. This was followed by experimentally verifying the effect on autophagy of a *Salmonella* protease involved in human-*Salmonella* interplay.

A variation of the motif-based methodologies is the use of motifs to characterize pathogen mimicry. This essentially involves the identification of eukaryotic linear motifs on microbial proteins which in turn can hijack host proteins and thereby promote antagonistic binding (Hurford and Day, 2013; Via et al., 2015). Motif-mediated molecular mimicry therefore rewires the host signaling and regulatory networks by titrating essential host proteins and enabling the microbe to create favorable micro-environments in the host cell by altering immune responses for example (Cusick et al., 2012). In addition to motifs, molecular mimicry can also be mediated at the level of protein, structural and interface levels. At the

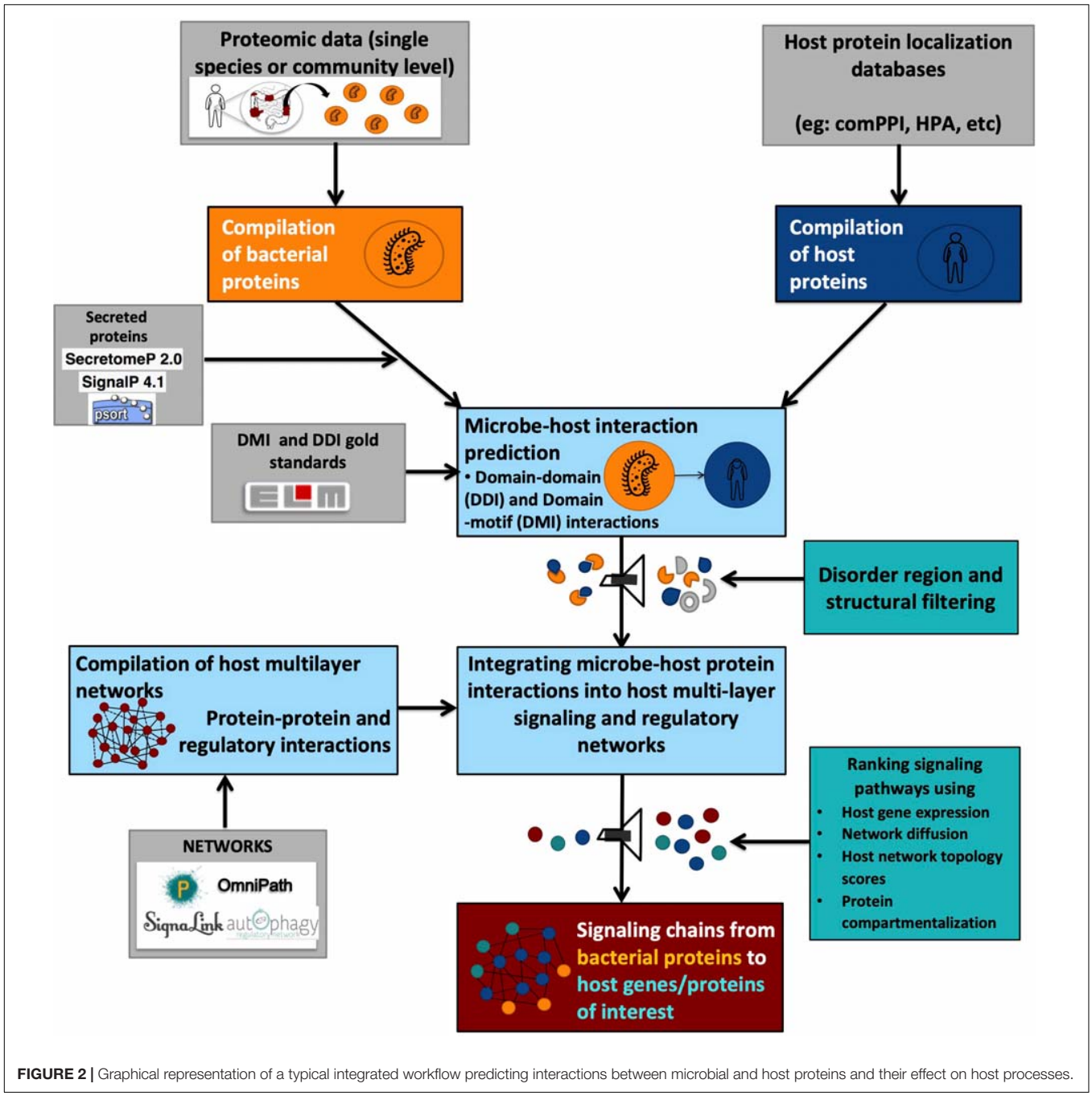


FIGURE 2 | Graphical representation of a typical integrated workflow predicting interactions between microbial and host proteins and their effect on host processes.

protein level, specific studies investigating the role of molecular mimicry in the pathogenesis of prominent bacterial pathogens (Doxey and McConkey, 2013) including *Salmonella typhimurium* and *Human respiratory syncytial virus* (Mei and Zhang, 2020) have been carried out (**Table 2**). At the interface level, Guven-Maiorov et al. (2017) devised a computational method to infer mimicry induced by a prominent gastric cancer causing pathogen *Helicobacter pylori*. Besides DDI and DMI based methods, researchers have also used other structure-based methodologies such as pairwise structural similarity (PSS) to predict inter-species PPIs. PSS methods at their very core are based on the premise that proteins possessing similar structures have a greater probability of interacting with the same set of protein partners (Ding and Kihara, 2018). This has been applied to infer the interactions with the host of various pathogens such as Dengue virus (Doolittle and Gomez, 2011), HIV (Cui et al., 2016), *Francisella tularensis* (Cui et al., 2016), West Nile virus (Chen et al., 2019), Chandipura virus (Rajasekharan et al., 2013), and other viral pathogens (Franzosa and Xia, 2011; Lasso et al., 2019).

As a means of ensuring proper quantitative evaluation of *de novo* PPI predictions, emerging computational methods such as

machine learning have been used in conjunction with structural-feature based PPI prediction methods. In order to avoid repetitions, methods using ML for evaluating the performance of structural feature dependent PPI predictions are discussed in the next subsection.

Machine Learning Based PPI Methods

Due to their ability to discern complex patterns among a large number of features in big datasets, machine learning (ML) methods have found favor in various applications of computational biology and bioinformatics (Shastry and Sanjay, 2020) including the prediction of microbe-host molecular interactions. A variety of supervised and unsupervised methods have been used to predict the interactions between microbial and host proteins (**Table 2**). In general, supervised machine learning methods utilize features from "gold-standard" interaction datasets to identify potential protein–protein interaction pairs from the user provided list of microbial and host proteins (Zhang et al., 2017b). In supervised methods, the "gold-standard" datasets are either compiled from high-throughput experimental methodologies or from curated lists of interactions from the literature (Zhang et al., 2017b). In the case of ML being used in combination with "interolog" based methods (explained in section 5.2.4), "gold-standard" PPI datasets can also be retrieved from other related or unrelated microbe-host species pairs depending on the scope of the study. Some of the features used to infer *de novo* PPI predictions include protein properties such as post-translational modifications, chemical composition, tissue distribution, molecular weight, domain/motif compositions, ontologies, gene expression, amino-acid frequencies, homology to human binding partners, and relevance of proteins in host network. By using these features, supervised methods are able to discern truly interacting protein pairs from all possible pairs of microbial and host proteins (Zhang et al., 2017b).

Supervised methods can also be differentiated by the kind of ML methodology/model used for the task of rightly classifying truly interacting protein pairs. Several supervised studies employing individual ML models [such as I2-regularized logistic regression (Mei et al., 2018), random forests (RF) (Kösesoy et al., 2019), etc], support vector machine (SVM) (Cui et al., 2012; Shoombuatong et al., 2012; Kim et al., 2017) have been applied to infer PPIs between microbial and host species. SVMs use a framework of searching and finding the best hyperplane (aka decision boundary represented by a mathematical equation) to separate sample with different labels corresponding to a class. Several variations of the SVM exist to handle data with underlying linear or non-linear relationships (Byvatov and Schneider, 2003).

Using four different ML models namely RF, SVM, Artificial Neural Networks (ANN) and K-Nearest Neighbors (K-NN), and multiple lines of -omic evidence including experimental PPIs as predictive features, Leite et al. (2018) devised a model based on a supervised protocol to accurately predict bacterium-phage interactions. The model, a type of ensemble learning, due to its generic nature, can also be used to predict interactions between any two given species, given the availability of informative feature sets. Ensemble learning (Che et al., 2011), combines multiple individual classifiers to achieve a final classification and has been used to predict PPI based HIV-human and hepatitis C virus-human networks (Mei, 2013; Emamjomeh et al., 2014). Ensemble classification methods outperform individual classifiers based on several use-cases (Krawczyk, 2015; Haque et al., 2016; Yijing et al., 2016; Lin et al., 2019) and can be generalized into three distinct categories namely bagging, boosting and stacked generalization. The last of the three approaches, stacked generalization, was used by Emamjomeh et al. (2014) to predict PPIs between human and the hepatitis C virus. While bagging assigns training sets to individual classifiers based on a random selection of the initial training dataset with replacement for subsequent sampling runs, boosting involves the creation and evaluation of classifiers in a sequential manner, with the succeeding classifier assigning more weights to the misclassification errors committed by the preceding classifier. The "boosted" weights are then normalized for all the instances in the entire dataset which is then used as the training dataset for the next classifier after which the final classification step is carried out based on the weighted individual classifiers. The stacked generalization methodology is designed to overcome some of the errors committed by the individual classifiers even if they are used in the ensemble framework. The stacked approach achieves this by using a "stacks" of base learners so that its output is the input for a meta-learner which knows how best to combine the base learners' outputs. The training data may or may not overlap between the two stacks and can be specified accordingly.

Various auxiliary algorithms have been used in conjunction with machine learning methods to predict inter-species PPIs. An example of such a study includes the use of a novel protein sequence based feature extraction method called Location Based Encoding (LBE) with different classifier models including RFs. Such integrated methodologies have been used to predict protein interactions with the human host of two important pathogens – *Bacillus anthracis* and *Yersinia pestis* (Kösesoy et al., 2019). LBE is a methodology which complements the ML approaches for PPIs by differentiating proteins only based on the locations of the amino acids in the sequence (Li et al., 2009).

Supervised methods are sometimes constrained due to the small size of "gold-standard" datasets that restricts the inference and prediction of proteome-wide PPIs between the full list of proteins of any two given species. Mei and Zhu (2014a) harness the power of multi-instance AdaBoost, a type of boosting-based ensemble learning protocol, which is a multi-instance learning based ML method, to reconstruct proteome-wide Human T-cell leukemia virus-human PPI networks using homology knowledge derived protein features. AdaBoost improves classification performance by combining multiple weak classifiers into one strong classifier. It works in part by assigning more weight to instances which can only be classified with greater difficulty than to instances which can be easily classified (Kim et al., 2012). The dearth of true interacting protein-pairs has also prompted researchers to use unsupervised or semi-supervised approaches to infer microbe-host PPIs. Qi et al. (2010) complement the list of true interactions with a list of protein-pairs wherein association evidence exists with no interaction evidence between the proteins of a pair. Supervised learning is performed thereafter with a

multilayer perceptron network and by using the true interaction list. Subsequently, the semi-supervised approach uses the same network layers of the supervised classifier but instead trains on the protein-pairs with association evidence only. By using this hybrid approach, the authors report improved performance for predicting interactions between HIV and human proteins (Qi et al., 2010).

Data/Literature Mining Based PPI Methods

Even though many databases have been compiled to collect, curate and store microbe-host PPIs (Kumar and Nanduri, 2010; Durmus Tekir et al., 2013; Cook et al., 2018; Gao et al., 2018; Singh et al., 2019), these are mostly confined to well-studied pathogens and are predominantly comprised of interactions from high-throughput experiments. Contrastingly, in the literature, there exist inter-species PPIs from low-throughput experiments with some of them from non-model organisms, and commensal microbes, but mostly distributed over several individual studies. Very often, the inter-species PPI databases and repositories do not capture these sparse interactions. Hence, researchers have adapted and modified data- and text-mining tools to search for and extract microbe-host PPIs from existing literature. Retrieving such PPIs not only helps in increasing the number of true positive and true negative interactions (which helps aid the predictive performance of algorithms) but also extends our knowledge of existing microbe-host interactions. Motivated by the above explained need to mine-out microbe-host PPIs, Thieu et al. (2012) combine and compare the performance of a language based method based on a link grammar parser to a supervised ML methodology (SVM) and report that the combined approach results in a higher classification accuracy when compared to existing literature mining methods. As part of a bigger analytical framework aimed at uncovering the cellular mechanisms involved in human B lymphocytes during *Epstein-Barr virus* infection, Li et al. (2018) use a big-data mining methodology to identify a diverse range of inter-species molecular interactions including PPIs. Similar text/data mining approaches were also executed to extract PPI-mediated interactions of the human host with multiple viruses such as Hepatitis C virus (Saik et al., 2016) and Influenza A virus (García-Pérez et al., 2018; **Table 2**).

Interolog Based PPI Methods

For most species-pairs of interest, especially those belonging to the category of non-model organisms, there is a scarcity of experimentally verified PPIs. This has necessitated the development of novel bioinformatic methods, one of which is the inference of interactions from existing experimentally determined inter-species PPIs (Kshirsagar et al., 2015). These types of methodologies are usually based on the principle of homology (hence the term "interolog": meaning interacting orthologs) – either at the level of proteins or protein structural features or both. Protein features used for homology based extrapolation include but are not limited to domains, motifs, amino-acid k-mers, and 3D structural properties (Kshirsagar et al., 2015). Interolog based approaches have been applied to harness the large volume of experimentally verified PPIs for model organisms including prominent bacterial/viral pathogens. Despite the potentially large coverage that can be achieved by such approaches, there exist several disadvantages of using interolog approaches as a silver bullet for inferring inter-species PPIs especially for novel species-pairs. These disadvantages are attributed to different pathogenic mechanisms between the microbes in the context of infecting different host species, different cellular localizations, and varying activity levels (expression, post-translational modifications, etc.) of the orthologous microbial proteins. Such differences lead to accessibility bottlenecks i.e., the ability of the proteins to physically access host proteins and thereby interact. Hence, interolog based approaches need to be complemented with additional filtering and quality control steps such as selecting proteins from infection-relevant cellular compartments, expression/activity measurements, etc.

Interolog based methods have been used to infer inter-species PPIs for many prominent pathogens and parasites (**Table 2**). Different versions of the interolog approach have been used to extrapolate PPIs corresponding to interactions between the human host and various pathogens such as *Plasmodium falciparum* (Krishnadev and Srinivasan, 2008; Lee et al., 2008), *Escherichia coli* (Krishnadev and Srinivasan, 2011), *S. typhimurium* (Krishnadev and Srinivasan, 2011; Schleker et al., 2012), *Y. pestis* (Krishnadev and Srinivasan, 2011), *Helicobacter pylori* (Tyagi et al., 2009), HIV (Cui et al., 2016), *Francisella tularensis* (Zhou et al., 2014; Cui et al., 2016), *Coxiella burnetii* (Wallqvist et al., 2017), *Corynebacterium pseudotuberculosis* (Barh et al., 2013), *Corynebacterium diphtheriae* (Barh et al., 2013), and *Corynebacterium ulcerans* (Barh et al., 2013). Using PPIs from the STRING database as the starting interaction set, Cuesta-Astroz et al. (2019) used the interolog methodology to predict PPIs between 15 different eukaryotic pathogens and the human host. To assign species-specific and lifecycle- specific contextuality, the authors confined the analysis to proteins from particular cellular compartments which are relevant to the infection process. From the analysis of the ensuing PPI networks, various invasion and evasion mechanisms adopted commonly and specifically by particular parasites were inferred (Cuesta-Astroz et al., 2019). Schleker et al. (2012) present another version of the interolog approach to predict human-*Salmonella* and *A. thaliana-Salmonella* PPI networks. As a source of template PPIs, publicly available interaction databases are used along with databases containing 3D structures between Pfam domains. As an add-on to the sequence based orthology of proteins, domain based orthology is also performed in order to reduce the false positive rates. Several additional filtering strategies such as restriction to predicted transmembrane proteins, relevance in host network and functional attributes such as gene ontology are used to make the PPIs more specific.

Approaches Inferring RNA Mediated Interactions

The role of RNAs, especially non-coding RNAs such as long non-coding RNAs (lncRNAs) and microRNAs (miRNAs) in mediating molecular microbe-host interactions have been

reported in the literature (Li et al., 2015b; Agliano et al., 2019). RNA molecules are either secreted by the microbial cell into the host cell or are packaged into vesicles along with other molecules which are then taken up by the host cell by endocytosis (Weiberg et al., 2014; Huang et al., 2019; Ahmadi Badi et al., 2020). Such microbial RNAs then modulate host cell activity by either binding to DNA, messenger RNAs or proteins. Thus, by salvaging and titrating host components, microbial RNAs modulate regulatory and signaling networks and subsequently host cell activity (Duval et al., 2017; Agliano et al., 2019; Shirahama et al., 2020). However, in contrast to PPI based methods, even though RNA-mediated microbe-host interactions are well studied from an experimental point of view, very few methods or studies exist that have systemically and systematically applied computational analysis (**Table 3**). As such, the resources which exist in the domain of RNA-mediated microbe-host interactions comprise of databases such as ViRBase (Li et al., 2015b) which is predominantly a source of experimentally verified virus–host non-coding RNA-associated interactions. In addition, it also contains predicted binding sites of virus non-coding RNAs on host proteins and RNAs. A prominent study which comprehensively examines and evaluates the role of RNAs in microbe-host interactions is that of Saçar Demirci and Adan (2020) who investigated the roles in infection of miRNA-like sequences encoded within the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) genome. They used a modified version of izMiR (Allmer et al., 2016), a SVM based ML method to predict pre-miRNAs which are homologous to the human precursor miRNAs from miRbase. The SVM based ML method identified several viral hairpin sequences which were smaller in length compared to the human miRNA precursors while many of the human and viral miRNA precursors were similar in length and shared identical minimum free energy, a feature used by the izMiR workflow (Allmer et al., 2016). Based on this observation, a revised classifier trained using only the known human miRNAs was used on the entire SARS-CoV-2 hairpin dataset which resulted in the identification of potential hairpins from which mature miRNA candidates were extracted. As a next step, the psRNATarget tool (Dai et al., 2018) was used to predict *de novo* the human genes targeted by the inferred viral miRNAs. Functional analysis of the human genes targeted revealed that the SARS-CoV-2miRNAs can affect various host processes including transcription, defense systems, Wnt and EGFR signaling pathways.

TABLE 3 | Examples of studies utilizing computational approaches to infer RNA-mediated interactions between microbes and hosts.

Study	Context
Saçar Demirci and Adan (2020)	Analysis revealing the potential interactions between mature micro-RNA like viral RNA sequences and host genes
ViRBase (Li et al., 2015b)	Source of experimentally verified virus–host non-coding RNA-associated interactions; also contains predicted binding sites of virus non-coding RNAs on host proteins and RNAs

Approaches Utilizing Pipelines Integrating Multiple-Omic Datasets

Besides the computational methods based on particular types of molecular interactions, some integrated pipelines (**Table 4**) have been compiled to infer mechanistic microbe-host interactions. In general, such pipelines (**Figure 2**) incorporate the prediction of at least one molecular interaction type between microbial and host molecular components followed by various other functionalities such as integration of host responses. **Table 5** provides a non-exhaustive overview of the different tools, databases and resources which are available in the public domain to compile integrated workflows based on PPIs for example.

KBase (Arkin et al., 2018) is an integrated bioinformatics platform enabling users to share datasets with the research community as well as facilitating the integration, and analysis of -omic datasets from microbes and plants by creating computational workflows. Recently, we developed MicrobioLink (Andrighetti et al., 2020), an integrated pipeline which carries out *de novo* DDI and DMI based microbe-host PPI prediction followed by quality control using information from disordered region predictions from built-in tools such as IUPred (Mészáros et al., 2018). The pipeline then utilizes network diffusion principles and tools (Paull et al., 2013) to infer the molecular mechanisms and signaling pathways which mediate the effect of microbial proteins on host responses as measured by transcriptomic or proteomic read-outs. Flexibility is provided for users to feed in the desired datasets at any given step of the pipeline. Given the advent of new computational tools in inter-species interactions and pipeline management platforms, it is expected that an increasing number of dedicated bioinformatic workflows for microbe-host interactions will be developed in the near future.

DISCUSSION: OPPORTUNITIES AND CHALLENGES

Opportunities

Clinical and Translational Research

Since the aforementioned computational tools help researchers narrow down on both microbial and host components involved in mechanistic cross-talks, the tools may discover molecules which can delineate different clinical phenotypes. In addition,

TABLE 4 | Integrated pipelines used to infer microbe-host interactions by combining heterogeneous -omic datasets.

Methodology	Functionalities
MicrobioLink (Andrighetti et al., 2020)	Integrating microbe-host protein interaction networks with host responses and host regulatory/signaling networks using network diffusion principles
KBase (Arkin et al., 2018)	Integrated platform enabling data sharing, integration, and analysis of -omic datasets from microbes, plants, and their communities by creating computational workflows
Li et al. (2015a)	Identifying critical effectors involved in host-pathogen interactions by integrating multiple lines of -omic evidence

TABLE 5 | A non-exhaustive catalog of resources, tools and databases to compile protein–protein interaction based workflows for inferring microbe (microbiome)-host interactions.

Step in workflow	Resource/Tool/Database
Source of proteomes (sequence information)	UniProt (The UniProt Consortium, 2018), HumanPSD (Hodges et al., 2002), YPD (Payne and Garrels, 1997), PombePD (Costanzo et al., 2001), WormPD (Costanzo et al., 2001), and SWISS-PROT (Bairoch and Apweiler, 1996)
Source of proteomic datasets (expression information)	ProteomicsDB (Schmidt et al., 2018), Human Protein Atlas (HPA) (Thul and Lindskog, 2018), PRIDE (Perez-Riverol et al., 2019), PeptideAtlas (Desiere et al., 2006), MassIVE.quant (Choi et al., 2020), jPOSTrepo (Okuda et al., 2017), iProX (Ma et al., 2019), and Panorama Public (Sharma et al., 2018)
Proteomic annotations (structural features)	InterPro (Mitchell et al., 2019), Pfam (El-Gebali et al., 2019), ELM (Gouw et al., 2018), and PDB (Burley et al., 2017)
Protein sub-cellular localization (databases and prediction tools)	ComPPI (Veres et al., 2015), HPA (Thul and Lindskog, 2018), LocDB (Rastogi and Rost, 2011), LocSigDB (Negi et al., 2015), COMPARTMENTS (Binder et al., 2014), eSLDB (Pierleoni et al., 2007), SCLpred-EMS (Kaleel et al., 2020), DeepLoc (Almagro Armenteros et al., 2017), PSORTdb (Peabody et al., 2016), SecretomeP (Bendtsen et al., 2004), and Signal P (Armenteros et al., 2019)
Base information for prediction of PPIs	Domain-domain predictions – DOMINE (Raghavachari et al., 2008) and Domain-motif predictions – ELM (Gouw et al., 2018)
Quality control of inferred PPIs (using disordered region prediction)	IUPred (Mészáros et al., 2018), PrDOS (Ishida and Kinoshita, 2007), D2P2 (Oates et al., 2013), PONDR-FIT (Xue et al., 2010), DISOPRED (Ward et al., 2004), MFDp2 (Mizianty et al., 2013), and Meta-Disorder (Kozlowski and Bujnicki, 2012)
Network resources	OmniPath (Türei et al., 2016), IntAct (Orchard et al., 2014), Reactome (Fabregat et al., 2018), STRING (Szklarczyk et al., 2017), HTRI (Bovolenta et al., 2012), and DoRothEA (Garcia-Alonso et al., 2018)
Network diffusion approaches	NBS (Hofree et al., 2013), HotNet (Vandin et al., 2011), TieDie (Basha et al., 2013; Paull et al., 2013), RegMod (Qiu et al., 2010), and stSVM21 (Cun and Fröhlich, 2013)
Databases for host gene expression	GEO (Clough and Barrett, 2016) and ArrayExpress (Parkinson et al., 2007)

they can also be possible targets for therapeutic interventions. In other words, mechanistic predictions combined with clinical meta-data have a dual-purpose – they provide information on molecular components which could both represent and drive clinical phenotypes (Younesi, 2015) and thereby could potentially minimize our reliance on association-based biomarkers alone which need not explain causality (Levenson and Mori, 2014). The discovery of such mechanistic knowledge warrants the combinatorial use of different methodologies including machine learning and molecular interaction analysis. While many community level studies have been conducted on meta -omic datasets for the clinical classification of patients and the discovery of associative biomarkers (Wen et al., 2017; Yu et al., 2020; Clos-Garcia et al., 2019; Conteville et al., 2019), they have not incorporated mechanistic inferences. On the other hand, most mechanistic studies (**Tables 2**, **3**) have been carried out on particular pathogens/microbial species without including clinical meta-data and/or clinical classifications.

Multi-omic approaches integrating heterogeneous -omic datasets from patients have been implemented for several diseases including IBD (Lloyd-Price et al., 2019) which are associated with microbial dysbiosis. However, these studies do not provide the required mechanistic insights for formulating therapeutic interventions. Beltran and Brito (2019) devised an integrated methodology to unravel the molecular mechanisms underlying the microbe-host interactions associated with various diseases such as colorectal cancer, IBD, obesity and type-2 diabetes. The aforementioned study represents one of the first and few initiatives to use community-wide microbe-host interaction predictions using meta -omic datasets from patients to discover mechanistic interactions driving the clinical phenotypes. By combining orthology based approaches to extrapolate interactions from experimental PPIs, machine learning and patient derived -omic datasets, the authors identified a subset of inter-species PPIs which are associated with disease phenotypes (Beltran and Brito, 2019). Thiele et al. (2020) published a novel study by integrating different levels of information (dietary information, physiological parameters, organ weights, and organ connectivities, etc.) and datasets such as molecular -omics (proteomics, metabolomics, metabolites produced by the gut microbiota) in an organ specific manner to arrive at a whole-body-model of human metabolism. Although not fully mechanistic, with this model, the authors were able to predict biomarkers of inherited metabolic diseases and host-microbiome co-metabolism. Such integrated studies and workflows combining statistical and mechanistic inference of multi -omic datasets awaits further adoption and application in the research on various diseases associated with microbial dysbiosis.

Research on Comparative Ecological Networks

The tools and resources listed in this review can be used to infer and predict molecular interactions between species in several contexts [microbe/microbiota in host, microbe/microbiota in several hosts, microbe (vs) microbe, and microbiota (vs) microbe, etc]. In almost all of the above-mentioned cases, molecular interactions between the autonomous entities (be it species or communities) could be driving the emergent phenotypes. Since the tools discussed in this manuscript also concern themselves with extrapolating interactions based on homology between species-pairs, it could be a right fit to predict *de novo* interaction relationships for species with very little experimental interaction information.

For example, Crohn's disease, a sub-type of IBD, is characterized by the dysbiosis of the gut microbiome (Joossens et al., 2011; Schaubeck et al., 2016; Shaw et al., 2016). This results in persistent inflammation of the gut mucosal barrier as a result of the unbalanced host responses (co-influenced by host genetic factors as well) to the dysbiosed microbiome and its various components such as proteins, metabolites, etc (Li et al., 2014; Lavelle and Sokol, 2020). Some of the CD patients also display lesions of the skin during or after therapeutic regimens (Huang et al., 2012; Gravina et al., 2016). It is known that the

skin also houses a complex microbial community which plays a role in maintaining homeostasis (Schommer and Gallo, 2013; Chen et al., 2018). Understanding the mechanisms by which CD medications impact the microbe-host interactions in the gut as well as the skin could help in avoiding the unintended side-effects of therapy in CD.

Yet another relevant context to apply the tools discussed herein is the inference of underlying molecular mechanisms which mediate the evasion of immune responses by bacterial pathogens in various hosts and their importance in transmission between hosts. We recently showed that bacterial pathogens and autophagy, a primary intracellular line of defense in the host, are engaged in an evolutionary tug of war, as evidenced by the presence of various interplays and cross-talks (Sudhakar et al., 2019). Given the exposure of host animals such as poultry and cattle to xenobiotic compounds such as antibiotics, many zoonotic pathogens are under constant selection pressure to evolve survival strategies to modulate/evade/survive within the host animal (Harada and Asai, 2010). This opens the door for impending risks of transmission (from animal hosts to human hosts or between various animal hosts) via the food chain of zoonotic species which have been selected for survival over many generations of persistence in the host (Farrell and Davies, 2019; Mollentze and Streicker, 2020). Microbe-host interaction mechanisms are at the evolutionary cross-roads of such transmission events between hosts. In this context, studying such interactions is expected to provide deeper insights into designing strategies to prevent and/or minimize spill-over transmission events.

Challenges

Over the past decade, various advances in the domain of computational analysis of microbe-host interactions have been made. However, despite this progress, there remain many challenges as described below. These challenges also present opportunities and the need to come up with innovative approaches and solutions.

Catching Up With Complex Infection Processes

Infection biology has taken new strides over the past years with new molecule classes (Katiyar-Agarwal and Jin, 2010; Rana et al., 2015; Duval et al., 2017; Long et al., 2017; Peters et al., 2019; Acuña et al., 2020) and cell-types (Chattopadhyay et al., 2018) being discovered as having a role in the infection process. With that, novel interaction types between various molecular classes are also unearthed (Silmon de Monerri and Kim, 2014). In some cases, computational methods have not caught up with molecular mechanisms. For example, hepadnaviruses utilize host DNA ligases to generate covalently closed circular DNAs which play a major role in mediating viral infection and persistence (Long et al., 2017). Similarly long non-coding RNAs are known to be involved in host-pathogen interactions (Duval et al., 2017; Agliano et al., 2019). However, till date, computational methods do not exist to predict or infer the mechanisms by which the viruses recruit the host DNA ligases or directly modulate the biogenesis, conformation and activity of long non-coding RNAs. Hence, computational method developments are always a step behind the complexity associated with infection biology. This gap is all the more prevalent for commensal organisms in contrast to pathogens due to the constant and historically prevalent study bias.

Lack of Experimental Datasets

Non-model organisms and non-pathogenic organisms such as probiotics and commensals also suffer from a considerable knowledge gap in terms of known/experimentally verified molecular interactions. This affects the performance of computational methods considerably due to the need for large sets of true positives for the satisfactory performance and assessment of predictive algorithms (Jiao and Du, 2016). In addition, this also influences the coverage and accuracy of interolog approaches since they harness already existing true positive datasets for extrapolating to the species-pairs of interest based on orthology.

False-Positives

As with any computational algorithm, microbe-host interaction prediction methods also face the curse of false positives. This issue could be exacerbated by the availability of relatively small true positive (truly interacting) and true negative (non-interacting sets) datasets (Jiao and Du, 2016). Furthermore, the evolutionary distance and difference in infection process between the template species-pairs and the species-pair of interest as well as the absence of orthologous molecular components involved in the interactions could also contribute to the inflated false positive rates, reduced performance and coverage.

Community-Wide Interaction Prediction

Most of the microbe-host interaction computational tools have been directed at uncovering interactions corresponding to individual microbe-host pairs. This is a major drawback of existing methodologies, especially given the fact that phenotypes related to health and disease are associated with changes in community wide alterations (Clemente et al., 2012; Koboziev et al., 2014; Wang et al., 2017; Bailey and Holscher, 2018; Dominguez-Bello et al., 2019).

Modeling Dynamics of Microbe-Host Interactions

Last but not the least, current methods involved in microbe-host interaction analysis are not equipped to handle the dynamic nature of natural ecosystems and ecological niches in which the interactions are embedded. Although it is a generic drawback of many bioinformatic approaches, this challenge will need coordinated efforts between modelers, experimental biologists and bioinformaticians.

CONCLUSION

Since the advent and expansion of high-throughput sequencing technologies, various observational studies of microbial communities inhabiting various ecological niches (inside host

organisms for example) have been carried out. This has mostly resulted in associations with health- or disease-associated phenotypes. However, there is a huge gap in terms of the mechanisms mediated by these microbial communities and how these mechanisms contribute to the observed phenotypes. Despite the availability of experimental datasets which capture some of these mechanisms such as PPIs, these are either confined to model organisms or well-studied pathogens. Computational approaches provide researchers with the tools to upscale microbe-host interaction research by enabling them to make *de novo* inter-species molecular interactions and to extrapolate existing microbe-host interaction datasets to the species-pairs of interest. Computational methods may aid the study of microbe-host interaction by reducing the variable space, prioritizing interactions, and eventually building hypothesis for further experimental verification.

AUTHOR CONTRIBUTIONS

PS performed the literature review and wrote the manuscript. KM provided critical feedbacks and contributed to the text. BV contributed to relevant discussion about the clinical implications. TK and SV supervised the work and provided valuable discussions, feedbacks, and comments. All authors contributed to the article and approved the submitted version.

FUNDING

PS was supported by the ERC Advanced Grant (ERC-2015-AdG, 694679, CrUCCial). TK was supported by a fellowship in computational biology at the Earlham Institute (Norwich, United Kingdom) in partnership with the Quadram Institute (Norwich, United Kingdom) and strategically supported by the BBSRC (BB/J004529/1, BB/P016774/1, and BB/CSP17270/1). SV is a senior clinical investigator of the Research Foundation Flanders (FWO), Belgium.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2021.618856/full#supplementary-material

Supplementary Table 1 | Studies using genome-scale metabolic models and constraint based approaches to infer mechanistic co-metabolic interactions between microbial and host species.

REFERENCES

Acuña, S. M., Floeter-Winter, L. M., and Muxel, S. M. (2020). MicroRNAs: biological regulators in pathogen-host interactions. *Cells* 9:113. doi: 10.3390/cells9010113

Agliano, F., Rathinam, V. A., Medvedev, A. E., Vanaja, S. K., and Vella, A. T. (2019). Long noncoding RNAs in host-pathogen interactions. *Trends Immunol.* 40, 492–510. doi: 10.1016/j.it.2019.04.001

Ahmadi Badi, S., Bruno, S. P., Moshiri, A., Tarashi, S., Siadat, S. D., and Masotti, A. (2020). Small RNAs in outer membrane vesicles and their function in host-microbe interactions. *Front. Microbiol.* 11, 1209. doi: 10.3389/fmicb.2020.01209

Akiva, E., Friedlander, G., Itzhaki, Z., and Margalit, H. (2012). A dynamic view of domain-motif interactions. *PLoS Comput. Biol.* 8, e1002341. doi: 10.1371/journal.pcbi.1002341

Aller, S., Scott, A., Sarkar-Tyson, M., and Soyer, O. S. (2018). Integrated human-virus metabolic stoichiometric modelling predicts host-based antiviral targets against Chikungunya. Dengue and Zika viruses. *J. R. Soc. Interface* 15:20180125. doi: 10.1098/rsif.2018.0125

Allmer, J., Allmer, J., and Saçar Demirci, M. D. (2016). izMiR: computational ab initio microRNA detection. *Protoc. Exch. [Preprint]*. doi: 10.1038/protex.2016.047

Almagro Armenteros, J. J., Sønderby, C. K., Sønderby, S. K., Nielsen, H., and Winther, O. (2017). DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics* 33, 3387–3395. doi: 10.1093/bioinformatics/btx431

Andrighetti, T., Bohar, B., Lemke, N., Sudhakar, P., and Korcsmaros, T. (2020). MicrobioLink: an integrated computational pipeline to infer functional effects of microbiome-host interactions. *Cells* 9:1278. doi: 10.3390/cells9051278

Arkin, A. P., Cottingham, R. W., Henry, C. S., Harris, N. L., Stevens, R. L., Maslov, S., et al. (2018). Kbase: the united states department of energy systems biology knowledgebase. *Nat. Biotechnol.* 36, 566–569. doi: 10.1038/nbt.4163

Armenteros, J. J. A., Tsirigos, K. D., Sønderby, C. K., Petersen, T. N., Winther, O., Brunak, S., et al. (2019). SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.* 37, 420–423. doi: 10.1038/s41587-019-0036-z

Azeloglu, E. U., and Iyengar, R. (2015). Signaling networks: information flow, computation, and decision making. *Cold Spring Harb. Perspect. Biol.* 7:a005934. doi: 10.1101/cshperspect.a005934

Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., et al. (2008). The RAST server: rapid annotations using subsystems technology. *BMC Genomics* 9:75. doi: 10.1186/1471-2164-9-75

Bailey, M. A., and Holscher, H. D. (2018). Microbiome-mediated effects of the mediterranean diet on inflammation. *Adv. Nutr.* 9, 193–206. doi: 10.1093/advances/nmy013

Bairoch, A., and Apweiler, R. (1996). The SWISS-PROT protein sequence data bank and its new supplement TREMBL. *Nucleic Acids Res.* 24, 21–25. doi: 10.1093/nar/24.1.21

Baldini, F., Heinken, A., Heirendt, L., Magnusdottir, S., Fleming, R. M. T., and Thiele, I. (2019). The microbiome modeling toolbox: from microbial interactions to personalized microbial communities. *Bioinformatics* 35, 2332–2334. doi: 10.1093/bioinformatics/bty941

Barh, D., Gupta, K., Jain, N., Khatri, G., León-Sicairos, N., Canizalez-Roman, A., et al. (2013). Conserved host-pathogen PPIs. globally conserved inter-species bacterial PPIs based conserved host-pathogen interactome derived novel target in *Corynebacterium pseudotuberculosis*, *Corynebacterium diphtheriae*, *Francisella tularensis*, *Corynebacterium ulcerans*, *Y. pestis*, and *E. coli* targeted by Piper betel compounds. *Integr. Biol. (Camb)* 5, 495–509. doi: 10.1039/c2ib20206a

Basha, O., Tirman, S., Eluk, A., and Yeger-Lotem, E. (2013). ResponseNet2.0: revealing signaling and regulatory pathways connecting your proteins and genes–now with human data. *Nucleic Acids Res.* 41, W198–W203. doi: 10.1093/nar/gkt532

Basit, A. H., Abbasi, W. A., Asif, A., Gull, S., and Minhas, F. U. A. A. (2018). Training host-pathogen protein-protein interaction predictors. *J. Bioinform. Comput. Biol.* 16:1850014. doi: 10.1142/S0219720018500142

Bauer, E., Zimmermann, J., Baldini, F., Thiele, I., and Kaleta, C. (2017). BacArena: individual-based metabolic modeling of heterogeneous microbes in complex communities. *PLoS Comput. Biol.* 13:e1005544. doi: 10.1371/journal.pcbi.1005544

Bendtsen, J. D., Jensen, L. J., Blom, N., Von Heijne, G., and Brunak, S. (2004). Feature-based prediction of non-classical and leaderless protein secretion. *Protein Eng. Des. Sel.* 17, 349–356. doi: 10.1093/protein/gzh037

Beltran, J. F., and Brito, I. (2019). Host-microbiome protein-protein interactions capture mechanisms in human disease. *BioRxiv*. doi: 10.1101/821926

Binder, J. X., Pletscher-Frankild, S., Tsafou, K., Stolte, C., O'Donoghue, S. I., Schneider, R., et al. (2014). COMPARTMENTS: unification and visualization of protein subcellular localization evidence. *Database (Oxford)* 2014:bau012. doi: 10.1093/database/bau012

Bordbar, A., Lewis, N. E., Schellenberger, J., Palsson, B. Ø, and Jamshidi, N. (2010). Insight into human alveolar macrophage and *Francisella tularensis* interactions via metabolic reconstructions. *Mol. Syst. Biol.* 6:422. doi: 10.1038/msb.2010.68

Bovolenta, L. A., Acencio, M. L., and Lemke, N. (2012). HTRIdb: an open-access database for experimentally verified human transcriptional regulation interactions. *BMC Genomics* 13:405. doi: 10.1186/1471-2164-13-405

Braga, R. M., Dourado, M. N., and Araújo, W. L. (2016). Microbial interactions: ecology in a molecular perspective. *Braz. J. Microbiol.* 47(Suppl. 1), 86–98. doi: 10.1016/j.bjm.2016.10.005

Bunesova, V., Lacroix, C., and Schwab, C. (2018). Mucin cross-feeding of infant bifidobacteria and *Eubacterium hallii*. *Microb. Ecol.* 75, 228–238. doi: 10.1007/s00248-017-1037-1034

Burley, S. K., Berman, H. M., Kleywegt, G. J., Markley, J. L., Nakamura, H., and Velankar, S. (2017). Protein data bank (PDB): the single global macromolecular structure archive. *Methods Mol. Biol.* 1607, 627–641. doi: 10.1007/978-1-4939-7000-1_26

Byvatov, E., and Schneider, G. (2003). Support vector machine applications in bioinformatics. *Appl. Bioinform.* 2, 67–77.

Carducci, M., Licata, L., Peluso, D., Castagnoli, L., and Cesareni, G. (2010). Enriching the viral-host interactomes with interactions mediated by SH3 domains. *Amino Acids* 38, 1541–1547. doi: 10.1007/s00726-009-0375-z

Charitou, T., Bryan, K., and Lynn, D. J. (2016). Using biological networks to integrate, visualize and analyze genomics data. *Genet. Sel. Evol.* 48:27. doi: 10.1186/s12711-016-0205-201

Chattopadhyay, P. K., Roederer, M., and Bolton, D. L. (2018). A deadly dance: the choreography of host-pathogen interactions, as revealed by single-cell technologies. *Nat. Commun.* 9, 4638. doi: 10.1038/s41467-018-06214-0

Che, D., Liu, Q., Rasheed, K., and Tao, X. (2011). Decision tree and ensemble learning algorithms with their applications in bioinformatics. *Adv. Exp. Med. Biol.* 696, 191–199. doi: 10.1007/978-1-4419-7046-6_19

Chen, J., Sun, J., Liu, X., Liu, F., Liu, R., and Wang, J. (2019). Structure-based prediction of West Nile virus-human protein-protein interactions. *J. Biomol. Struct. Dyn.* 37, 2310–2321. doi: 10.1080/07391102.2018.1479659

Chen, Y. E., Fischbach, M. A., and Belkaid, Y. (2018). Skin microbiota-host interactions. *Nature* 553, 427–436. doi: 10.1038/nature25177

Chen, Z., Zheng, Y., Ding, C., Ren, X., Yuan, J., Sun, F., et al. (2017). Integrated metagenomics and molecular ecological network analysis of bacterial community composition during the phytoremediation of cadmium-contaminated soils by bioenergy crops. *Ecotoxicol. Environ. Saf.* 145, 111–118. doi: 10.1016/j.ecoenv.2017.07.019

Choi, M., Carver, J., Chiva, C., Tzouros, M., Huang, T., Tsai, T.-H., et al. (2020). MassIVE.quant: a community resource of quantitative mass spectrometry-based proteomics datasets. *Nat. Methods* 17, 981–984. doi: 10.1038/s41592-020-0955-950

Clemente, J. C., Ursell, L. K., Parfrey, L. W., and Knight, R. (2012). The impact of the gut microbiota on human health: an integrative view. *Cell* 148, 1258–1270. doi: 10.1016/j.cell.2012.01.035

Clos-Garcia, M., Andrés-Marin, N., Fernández-Eulate, G., Abecia, L., Lavín, J. L., van Liempd, S., et al. (2019). Gut microbiome and serum metabolome analyses identify molecular biomarkers and altered glutamate metabolism in fibromyalgia. *EBioMedicine* 46, 499–511. doi: 10.1016/j.ebiom.2019.07.031

Clough, E., and Barrett, T. (2016). The gene expression omnibus database. *Methods Mol. Biol.* 1418, 93–110. doi: 10.1007/978-1-4939-3578-9_5

Conteville, L. C., Oliveira-Ferreira, J., and Vicente, A. C. P. (2019). Gut microbiome biomarkers and functional diversity within an amazonian semi-nomadic hunter-gatherer group. *Front. Microbiol.* 10, 1743. doi: 10.3389/fmicb.2019.01743

Cook, H. V., Doncheva, N. T., Szklarczyk, D., von Mering, C., and Jensen, L. J. (2018). Viruses.STRING: a virus-host protein-protein interaction database. *Viruses* 10:519. doi: 10.3390/v10100519

Costanzo, M. C., Crawford, M. E., Hirschman, J. E., Kranz, J. E., Robertson, L. S., et al. (2001). YPD, PombePD and WormPD: model organism volumes of the BioKnowledge library, an integrated resource for protein information. *Nucleic Acids Res.* 29, 75–79. doi: 10.1093/nar/29.1.75

Cuesta-Astroz, Y., Santos, A., Oliveira, G., and Jensen, L. J. (2019). Analysis of predicted host-parasite interactomes reveals commonalities and specificities related to parasitic lifestyle and tissues tropism. *Front. Immunol.* 10:212. doi: 10.3389/fimmu.2019.00212

Cui, G., Fang, C., and Han, K. (2012). Prediction of protein-protein interactions between viruses and human by an SVM model. *BMC Bioinformatics* 13(Suppl. 7):S5. doi: 10.1186/1471-2105-13-S7-S5

Cui, T., Li, W., Liu, L., Huang, Q., and He, Z.-G. (2016). Uncovering new pathogen-host protein-protein interactions by pairwise structure similarity. *PLoS One* 11:e0147612. doi: 10.1371/journal.pone.0147612

Cun, Y., and Fröhlich, H. (2013). Network and data integration for biomarker signature discovery via network smoothed T-statistics. *PLoS One* 8:e73074. doi: 10.1371/journal.pone.0073074

Cusick, M. F., Libbey, J. E., and Fujinami, R. S. (2012). Molecular mimicry as a mechanism of autoimmune disease. *Clin. Rev. Allergy Immunol.* 42, 102–111. doi: 10.1007/s12016-011-8294-7

Dai, X., Zhuang, Z., and Zhao, P. X. (2018). psRNATarget: a plant small RNA target analysis server (2017 release). *Nucleic Acids Res.* 46, W49–W54. doi: 10.1093/nar/gky316

Deng, Y., Jiang, Y.-H., Yang, Y., He, Z., Luo, F., and Zhou, J. (2012). Molecular ecological network analyses. *BMC Bioinformatics* 13:113. doi: 10.1186/1471-2105-13-113

Desiere, F., Deutsch, E. W., King, N. L., Nesvizhskii, A. I., Mallick, P., Eng, J., et al. (2006). The PeptideAtlas project. *Nucleic Acids Res.* 34, D655–D658. doi: 10.1093/nar/gkj040

Ding, T., Case, K. A., Omolo, M. A., Reiland, H. A., Metz, Z. P., Diao, X., et al. (2016). Predicting essential metabolic genome content of niche-specific enterobacterial human pathogens during simulation of host environments. *PLoS One* 11:e0149423. doi: 10.1371/journal.pone.0149423

Ding, Z., and Kihara, D. (2018). Computational methods for predicting protein-protein interactions using various protein features. *Curr. Protoc. Protein Sci.* 93, e62. doi: 10.1002/cpps.62

Dix, A., Vlaic, S., Guthke, R., and Linde, J. (2016). Use of systems biology to decipher host-pathogen interaction networks and predict biomarkers. *Clin. Microbiol. Infect.* 22, 600–606. doi: 10.1016/j.cmi.2016.04.014

Dominguez-Bello, M. G., Godoy-Vitorino, F., Knight, R., and Blaser, M. J. (2019). Role of the microbiome in human development. *Gut* 68, 1108–1114. doi: 10.1136/gutjnl-2018-317503

Dong, Y., Kuang, Q., Dai, X., Li, R., Wu, Y., Leng, W., et al. (2015). Improving the understanding of pathogenesis of human papillomavirus 16 via mapping protein-protein interaction network. *Biomed Res. Int.* 2015:890381. doi: 10.1155/2015/890381

Doolittle, J. M., and Gomez, S. M. (2011). Mapping protein interactions between dengue virus and its human and insect hosts. *PLoS Negl. Trop. Dis.* 5:e954. doi: 10.1371/journal.pntd.0000954

Doxey, A. C., and McConkey, B. J. (2013). Prediction of molecular mimicry candidates in human pathogenic bacteria. *Virulence* 4, 453–466. doi: 10.4161/viru.25180

Durmus Tekir, S., Çakir, T., Ardiç, E., Sayilirbas, A. S., Konuk, G., Konuk, M., et al. (2013). PHISTO: pathogen-host interaction search tool. *Bioinformatics* 29, 1357–1358. doi: 10.1093/bioinformatics/btt137

Duval, M., Cossart, P., and Lebreton, A. (2017). Mammalian microRNAs and long noncoding RNAs in the host-bacterial pathogen crosstalk. *Semin. Cell Dev. Biol.* 65, 11–19. doi: 10.1016/j.semcdb.2016.06.016

Dyer, M. D., Murali, T. M., and Sobral, B. W. (2007). Computational prediction of host-pathogen protein-protein interactions. *Bioinformatics* 23, i159–i166. doi: 10.1093/bioinformatics/btm208

Dyer, M. D., Murali, T. M., and Sobral, B. W. (2011). Supervised learning and prediction of physical interactions between human and HIV proteins. *Infect. Genet. Evol.* 11, 917–923. doi: 10.1016/j.meegid.2011.02.022

Eain, M. M. G., Baginska, J., Greenhalgh, K., Fritz, J. V., Zenhausern, F., and Wilmes, P. (2017). Engineering solutions for representative models of the gastrointestinal human-microbe interface. *Engineering* 3, 60–65. doi: 10.1016/J.ENG.2017.01.011

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., et al. (2019). The Pfam protein families database in 2019. *Nucleic Acids Res.* 47, D427–D432. doi: 10.1093/nar/gky995

Emamjomeh, A., Goliaei, B., Zahiri, J., and Ebrahimpour, R. (2014). Predicting protein-protein interactions between human and hepatitis C virus via an ensemble learning method. *Mol. Biosyst.* 10, 3147–3154. doi: 10.1039/c4mb00410h

Emmert-Streib, F., and Glazko, G. V. (2011). Network biology: a direct approach to study biological function. *Wiley Interdiscip. Rev. Syst. Biol. Med.* 3, 379–391. doi: 10.1002/wsbm.134

Evans, P., Dampier, W., Ungar, L., and Tozeren, A. (2009). Prediction of HIV-1 virus-host protein interactions using virus and host sequence motifs. *BMC Med. Genomics* 2:27. doi: 10.1186/1755-8794-2-27

Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., et al. (2018). The reactome pathway knowledgebase. *Nucleic Acids Res.* 46, D649–D655. doi: 10.1093/nar/gkx1132

Farrell, M. J., and Davies, T. J. (2019). Disease mortality in domesticated animals is predicted by host evolutionary relationships. *Proc. Natl. Acad. Sci. U S A.* 116, 7911–7915. doi: 10.1073/pnas.1817323116

Franzosa, E. A., and Xia, Y. (2011). Structural principles within the human-virus protein-protein interaction network. *Proc. Natl. Acad. Sci. U S A.* 108, 10538–10543. doi: 10.1073/pnas.1101440108

Fritz, J. V., Desai, M. S., Shah, P., Schneider, J. G., and Wilmes, P. (2013). From meta-omics to causality: experimental models for human microbiome research. *Microbiome* 1:14. doi: 10.1186/2049-2618-1-14

Gao, N. L., Zhang, C., Zhang, Z., Hu, S., Lercher, M. J., Zhao, X.-M., et al. (2018). MVP: a microbe-phage interaction database. *Nucleic Acids Res.* 46, D700–D707. doi: 10.1093/nar/gkx1124

Garcia-Alonso, L., Iorio, F., Matchan, A., Fonseca, N., Jaaks, P., Peat, G., et al. (2018). Transcription factor activities enhance markers of drug sensitivity in cancer. *Cancer Res.* 78, 769–780. doi: 10.1158/0008-5472.CAN-17-1679

García-Pérez, C. A., Guo, X., Navarro, J. G., Aguilar, D. A. G., and Lara-Ramírez, E. E. (2018). Proteome-wide analysis of human motif-domain interactions mapped on influenza a virus. *BMC Bioinformatics* 19:238. doi: 10.1186/s12859-018-2237-2238

Gibson, T. J., Dinkel, H., Van Roey, K., and Diella, F. (2015). Experimental detection of short regulatory motifs in eukaryotic proteins: tips for good practice as well as for bad. *Cell Commun. Signal.* 13:42. doi: 10.1186/s12964-015-0121-y

Gosak, M., Markoviè, R., Dolenšek, J., Slak Rupnik, M., Marhl, M., Stožer, A., et al. (2018). Network science of biological systems at different scales: a review. *Phys. Life Rev.* 24, 118–135. doi: 10.1016/j.plrev.2017.11.003

Gouw, M., Michael, S., Sámano-Sánchez, H., Kumar, M., Zeke, A., Lang, B., et al. (2018). The eukaryotic linear motif resource - 2018 update. *Nucleic Acids Res.* 46, D428–D434. doi: 10.1093/nar/gkx1077

Gravina, A. G., Federico, A., Ruocco, E., Lo Schiavo, A., Romano, F., Miranda, A., et al. (2016). Crohn's disease and skin. *United Eur. Gastroenterol. J.* 4, 165–171. doi: 10.1177/2050640615597835

Guven-Maiorov, E., Tsai, C.-J., Ma, B., and Nussinov, R. (2017). Prediction of host pathogen interactions for *Helicobacter pylori* by interface mimicry and implications to gastric Cancer. *J. Mol. Biol.* 429, 3925–3941. doi: 10.1016/j.jmb.2017.10.023

Halehalli, R. R., and Nagarajaram, H. A. (2015). Molecular principles of human virus protein-protein interactions. *Bioinformatics* 31, 1025–1033. doi: 10.1093/bioinformatics/btu763

Haque, M. N., Noman, N., Berretta, R., and Moscato, P. (2016). Heterogeneous ensemble combination search using genetic algorithm for class imbalanced data classification. *PLoS One* 11:e0146116. doi: 10.1371/journal.pone.0146116

Harada, K., and Asai, T. (2010). Role of antimicrobial selective pressure and secondary factors on antimicrobial resistance prevalence in *Escherichia coli* from food-producing animals in Japan. *J. Biomed. Biotechnol.* 2010:180682. doi: 10.1155/2010/180682

Heinken, A., Acharya, G., Ravcheev, D. A., Hertel, J., Nyga, M., Okpala, O. E., et al. (2020). AGORA2: large scale reconstruction of the microbiome highlights wide-spread drug-metabolising capacities. *BioRxiv [preprint]* doi: 10.1101/2020.11.09.375451

Heinken, A., Sahoo, S., Fleming, R. M. T., and Thiele, I. (2013). Systems-level characterization of a host-microbe metabolic symbiosis in the mammalian gut. *Gut Microbes* 4, 28–40. doi: 10.4161/gmic.22370

Heinken, A., and Thiele, I. (2015). Systematic prediction of health-relevant human-microbial co-metabolism through a computational framework. *Gut Microbes* 6, 120–130. doi: 10.1080/19490976.2015.1023494

Heirendt, L., Arreckx, S., Pfau, T., Mendoza, S. N., Richelle, A., Heinken, A., et al. (2019). Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0. *Nat. Protoc.* 14, 639–702. doi: 10.1038/s41596-018-0098-92

Heleno, R., Garcia, C., Jordano, P., Traveset, A., Gómez, J. M., Blüthgen, N., et al. (2014). Ecological networks: delving into the architecture of biodiversity. *Biol. Lett.* 10:20131000. doi: 10.1098/rsbl.2013.1000

Henry, C. S., DeJongh, M., Best, A. A., Frybarger, P. M., Linsay, B., and Stevens, R. L. (2010). High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat. Biotechnol.* 28, 977–982. doi: 10.1038/nbt.1672

Hertel, J., Harms, A. C., Heinken, A., Baldini, F., Thinnes, C. C., Glaab, E., et al. (2019). Integrated analyses of microbiome and longitudinal metabolome data reveal microbial-host interactions on sulfur metabolism in Parkinson's disease. *Cell Rep.* 29, 1767–1777.e8. doi: 10.1016/j.celrep.2019.10.035

Hodges, P. E., Carrico, P. M., Hogan, J. D., O'Neill, K. E., Owen, J. J., Mangan, M., et al. (2002). Annotating the human proteome: the human proteome survey database (HumanPSD) and an in-depth target database for G protein-coupled receptors (GPCR-PD) from incyte genomics. *Nucleic Acids Res.* 30, 137–141. doi: 10.1093/nar/30.1.137

Hofree, M., Shen, J. P., Carter, H., Gross, A., and Ideker, T. (2013). Network-based stratification of tumor mutations. *Nat. Methods* 10, 1108–1115. doi: 10.1038/nmeth.2651

Hongjaisee, S., Nantasenamat, C., Carraway, T. S., and Shoombuatong, W. (2019). HIVCoR: a sequence-based tool for predicting HIV-1 CRF01_AE coreceptor usage. *Comput. Biol. Chem.* 80, 419–432. doi: 10.1016/j.compbiolchem.2019.05.006

Huang, B. L., Chandra, S., and Shih, D. Q. (2012). Skin manifestations of inflammatory bowel disease. *Front. Physiol.* 3:13. doi: 10.3389/fphys.2012.00013

Huang, C.-Y., Wang, H., Hu, P., Hamby, R., and Jin, H. (2019). Small RNAs – Big players in plant-microbe interactions. *Cell Host Microbe* 26, 173–182. doi: 10.1016/j.chom.2019.07.021

Huang, J. K., Carlin, D. E., Yu, M. K., Zhang, W., Kreisberg, J. F., Tamayo, P., et al. (2018). Systematic evaluation of molecular networks for discovery of disease genes. *Cell Syst.* 6, 484–495.e5. doi: 10.1016/j.cels.2018.03.001

Hughes, D. T., and Sperandio, V. (2008). Inter-kingdom signalling: communication between bacteria and their hosts. *Nat. Rev. Microbiol.* 6, 111–120. doi: 10.1038/nrmicro1836

Hurford, A., and Day, T. (2013). Immune evasion and the evolution of molecular mimicry in parasites. *Evolution* 67, 2889–2904. doi: 10.1111/evo.12171

Idrees, S., Pérez-Bercoff, Å., and Edwards, R. J. (2018). SLiM-Enrich: computational assessment of protein-protein interaction data as a source of domain-motif interactions. *PeerJ* 6, e5858. doi: 10.7717/peerj.5858

Ishida, T., and Kinoshita, K. (2007). PrDOS: prediction of disordered protein regions from amino acid sequence. *Nucleic Acids Res.* 35, W460–W464. doi: 10.1093/nar/gkm363

Islam, M. M., Fernando, S. C., and Saha, R. (2019). Metabolic modeling elucidates the transactions in the rumen microbiome and the shifts upon virome interactions. *Front. Microbiol.* 10:2412. doi: 10.3389/fmicb.2019.02412

Jacob, J. J., Veeraraghavan, B., and Vasudevan, K. (2019). Metagenomic next-generation sequencing in clinical microbiology. *Ind. J. Med. Microbiol.* 37, 133–140. doi: 10.4103/ijmm.IJMM_19_401

Jiao, Y., and Du, P. (2016). Performance measures in evaluating machine learning based bioinformatics predictors for classifications. *Quant. Biol.* 4, 320–330. doi: 10.1007/s40484-016-0081-2

Joossens, M., Huys, G., Cnockaert, M., De Preter, V., Verbeke, K., Rutgeerts, P., et al. (2011). Dysbiosis of the faecal microbiota in patients with Crohn's disease and their unaffected relatives. *Gut* 60, 631–637. doi: 10.1136/gut.2010.223263

Kaleel, M., Zheng, Y., Chen, J., Feng, X., Simpson, J. C., Pollastri, G., et al. (2020). SCLpred-EMS: subcellular localization prediction of endomembrane system

and secretory pathway proteins by Deep N-to-1 convolutional neural networks. *Bioinformatics* 36, 3343–3349. doi: 10.1093/bioinformatics/btaa156

Kargarfard, F., Sami, A., Mohammadi-Dehcheshmeh, M., and Ebrahimie, E. (2016). Novel approach for identification of influenza virus host range and zoonotic transmissible sequences by determination of host-related associative positions in viral genome segments. *BMC Genomics* 17:925. doi: 10.1186/s12864-016-3250-3259

Katiyar-Agarwal, S., and Jin, H. (2010). Role of small RNAs in host-microbe interactions. *Annu. Rev. Phytopathol.* 48, 225–246. doi: 10.1146/annurev-phyto-073009-114457

Kerr, S. A., Jackson, E. L., Lungu, O. I., Meyer, A. G., Demogines, A., Ellington, A. D., et al. (2015). Computational and functional analysis of the virus-receptor interface reveals host range trade-offs in new world arenaviruses. *J. Virol.* 89, 11643–11653. doi: 10.1128/JVI.01408-1415

Kim, B., Alguwaizani, S., Zhou, X., Huang, D.-S., Park, B., and Han, K. (2017). An improved method for predicting interactions between virus and human proteins. *J. Bioinform. Comput. Biol.* 15:1650024. doi: 10.1142/S0219720016500244

Kim, J.-G., Park, D., Kim, B.-C., Cho, S.-W., Kim, Y. T., Park, Y.-J., et al. (2008). Predicting the interactome of Xanthomonas oryzae pathovar oryzae for target selection and DB service. *BMC Bioinformatics* 9:41. doi: 10.1186/1471-2105-9-41

Kim, T.-H., Park, D.-C., Woo, D.-M., Jeong, T., and Min, S.-Y. (2012). "Multi-class classifier-based adaboost algorithm," in *Intelligent Science and Intelligent Data Engineering Lecture Notes in Computer Science*, eds Y. Zhang, Z.-H. Zhou, C. Zhang, and Y. Li (Berlin: Springer), 122–127. doi: 10.1007/978-3-642-31919-8_16

Koboziev, I., Reinoso Webb, C., Furr, K. L., and Grisham, M. B. (2014). Role of the enteric microbiota in intestinal homeostasis and inflammation. *Free Radic. Biol. Med.* 68, 122–133. doi: 10.1016/j.freeradbiomed.2013.11.008

Kösesoy, Ý, Gök, M., and Öz, C. (2019). A new sequence based encoding for prediction of host-pathogen protein interactions. *Comput. Biol. Chem.* 78, 170–177. doi: 10.1016/j.compbiolchem.2018.12.001

Kozlowski, L. P., and Bujnicki, J. M. (2012). MetaDisorder: a meta-server for the prediction of intrinsic disorder in proteins. *BMC Bioinformatics* 13:111. doi: 10.1186/1471-2105-13-111

Krawczyk, B. (2015). Forming ensembles of soft one-class classifiers with weighted bagging. *New Gener. Comput.* 33, 449–466. doi: 10.1007/s00354-015-0406-400

Krishnadev, O., and Srinivasan, N. (2008). A data integration approach to predict host-pathogen protein-protein interactions: application to recognize protein interactions between human and a malarial parasite. *In Silico Biol. (Gedrukt)* 8, 235–250.

Krishnadev, O., and Srinivasan, N. (2011). Prediction of protein-protein interactions between human host and a pathogen and its application to three pathogenic bacteria. *Int. J. Biol. Macromol.* 48, 613–619. doi: 10.1016/j.ijbiomac.2011.01.030

Kshirsagar, M., Carbonell, J., and Klein-Seetharaman, J. (2013). Multitask learning for host-pathogen protein interactions. *Bioinformatics* 29, i217–i226. doi: 10.1093/bioinformatics/btt245

Kshirsagar, M., Schleker, S., Carbonell, J., and Klein-Seetharaman, J. (2015). Techniques for transferring host-pathogen protein interactions knowledge to new tasks. *Front. Microbiol.* 6, 36. doi: 10.3389/fmicb.2015.00036

Kumar, R., and Nanduri, B. (2010). HPIDB–a unified resource for host-pathogen interactions. *BMC Bioinformatics* 11, S16. doi: 10.1186/1471-2105-11-S6-S16

Lai, Y.-H., Li, Z.-C., Chen, L.-L., Dai, Z., and Zou, X.-Y. (2012). Identification of potential host proteins for influenza a virus based on topological and biological characteristics by proteome-wide network approach. *J. Proteomics* 75, 2500–2513. doi: 10.1016/j.jprot.2012.02.034

Lasso, G., Mayer, S. V., Winkelmann, E. R., Chu, T., Elliot, O., Patino-Galindo, J. A., et al. (2019). A structure-informed atlas of human-virus interactions. *Cell* 178, 1526–1541.e16. doi: 10.1016/j.cell.2019.08.005

Lavelle, A., and Sokol, H. (2020). Gut microbiota-derived metabolites as key actors in inflammatory bowel disease. *Nat. Rev. Gastroenterol. Hepatol.* 17, 223–237. doi: 10.1038/s41575-019-0258-z

Levenson, V., and Mori, Y. (2014). The era of personalized medicine: mechanistic or correlative biomarkers? *Per. Med.* 11, 361–364. doi: 10.2217/pme.14.10

Lee, S.-A., Chan, C., Tsai, C.-H., Lai, J.-M., Wang, F.-S., Kao, C.-Y., et al. (2008). Ortholog-based protein-protein interaction prediction and its application to inter-species interactions. *BMC Bioinformatics* 9(Suppl. 12):S11. doi: 10.1186/1471-2105-9-S12-S11

Leite, D. M. C., Brochet, X., Resch, G., Que, Y.-A., Neves, A., and Peña-Reyes, C. (2018). Computational prediction of inter-species relationships through omics data analysis and machine learning. *BMC Bioinformatics* 19:420. doi: 10.1186/s12859-018-2388-2387

Levy, S. E., and Myers, R. M. (2016). Advancements in next-generation sequencing. *Annu. Rev. Genomics. Hum. Genet.* 17, 95–115. doi: 10.1146/annurev-genom-083115-22413

Li, C.-W., Jheng, B.-R., and Chen, B.-S. (2018). Investigating genetic-and-epigenetic networks, and the cellular mechanisms occurring in Epstein-Barr virus-infected human B lymphocytes via big data mining and genome-wide two-sided NGS data identification. *PLoS One* 13:e0202537. doi: 10.1371/journal.pone.0202537

Li, Q., Wang, C., Tang, C., He, Q., Li, N., and Li, J. (2014). Dysbiosis of gut fungal microbiota is associated with mucosal inflammation in Crohn's disease. *J. Clin. Gastroenterol.* 48, 513–523. doi: 10.1097/MCG.0000000000000035

Li, W., Fan, X., Long, Q., Xie, L., and Xie, J. (2015a). Mycobacterium tuberculosis effectors involved in host-pathogen interaction revealed by a multiple scales integrative pipeline. *Infect. Genet. Evol.* 32, 1–11. doi: 10.1016/j.meegid.2015.02.014

Li, Y., Wang, C., Miao, Z., Bi, X., Wu, D., Jin, N., et al. (2015b). ViRBase: a resource for virus-host ncRNA-associated interactions. *Nucleic Acids Res.* 43, D578–D582. doi: 10.1093/nar/gku903

Li, X., Liao, B., Shu, Y., Zeng, Q., and Luo, J. (2009). Protein functional class prediction using global encoding of amino acid sequence. *J. Theor. Biol.* 261, 290–293. doi: 10.1016/j.jtbi.2009.07.017

Li, Z.-G., He, F., Zhang, Z., and Peng, Y.-L. (2012). Prediction of protein-protein interactions between Ralstonia solanacearum and *Arabidopsis thaliana*. *Amino Acids* 42, 2363–2371. doi: 10.1007/s00726-011-0978-z

Lian, X., Yang, S., Li, H., Fu, C., and Zhang, Z. (2019). Machine-Learning-Based predictor of human-bacteria protein-protein interactions by incorporating comprehensive host- network properties. *J. Proteome Res.* 18, 2195–2205. doi: 10.1021/acs.jproteome.9b00074

Liao, Q., Yuan, X., Xiao, H., Liu, C., Lv, Z., Zhao, Y., et al. (2011). Identifying *Schistosoma japonicum* excretory/secretory proteins and their interactions with host immune system. *PLoS One* 6:e23786. doi: 10.1371/journal.pone.0023786

Lin, W.-C., Lu, Y.-H., and Tsai, C.-F. (2019). Feature selection in single and ensemble learning-based bankruptcy prediction models. *Expert Systems* 36:e12335. doi: 10.1111/exsy.12335

Lloyd-Price, J., Arze, C., Ananthakrishnan, A. N., Schirmer, M., Avila-Pacheco, J., Poon, T. W., et al. (2019). Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* 569, 655–662. doi: 10.1038/s41586-019-1237-9

Long, Q., Yan, R., Hu, J., Cai, D., Mitra, B., Kim, E. S., et al. (2017). The role of host DNA ligases in hepadnavirus covalently closed circular DNA formation. *PLoS Pathog.* 13:e1006784. doi: 10.1371/journal.ppat.1006784

Ma, J., Chen, T., Wu, S., Yang, C., Bai, M., Shu, K., et al. (2019). iProX: an integrated proteome resource. *Nucleic Acids Res.* 47, D1211–D1217. doi: 10.1093/nar/gky869

Mahajan, G., and Mande, S. C. (2017). Using structural knowledge in the protein data bank to inform the search for potential host-microbe protein interactions in sequence space: application to *Mycobacterium tuberculosis*. *BMC Bioinformatics* 18:201. doi: 10.1186/s12859-017-1550-y

Mariethoz, J., Khatib, K., Alocci, D., Campbell, M. P., Karlsson, N. G., Packer, N. H., et al. (2016). SugarBindDB, a resource of glycan-mediated host-pathogen interactions. *Nucleic Acids Res.* 44, D1243–D1250. doi: 10.1093/nar/gkv1247

Martin, A. M., Yabut, J. M., Choo, J. M., Page, A. J., Sun, E. W., Jessup, C. F., et al. (2019). The gut microbiome regulates host glucose homeostasis via peripheral serotonin. *Proc. Natl. Acad. Sci. U S A.* 116, 19802–19804. doi: 10.1073/pnas.1909311116

Martinez, K. B., Leone, V., and Chang, E. B. (2017). Microbial metabolites in health and disease: navigating the unknown in search of function. *J. Biol. Chem.* 292, 8553–8559. doi: 10.1074/jbc.R116.752899

May, S., Evans, S., and Parry, L. (2017). Organoids, organs-on-chips and other systems, and microbiota. *Emerg. Top. Life Sci.* 1, 385–400. doi: 10.1042/ETLS20170047

Mehrotra, P., Ramakrishnan, G., Dhandapani, G., Srinivasan, N., and Madanan, M. G. (2017). Comparison of *Leptospira interrogans* and *Leptospira biflexa* genomes: analysis of potential leptospiral-host interactions. *Mol. Biosyst.* 13, 883–891. doi: 10.1039/c6mb00856a

Mei, S. (2013). Probability weighted ensemble transfer learning for predicting interactions between HIV-1 and human proteins. *PLoS One* 8:e79606. doi: 10.1371/journal.pone.0079606

Mei, S., Flemington, E. K., and Zhang, K. (2018). Transferring knowledge of bacterial protein interaction networks to predict pathogen targeted human genes and immune signaling pathways: a case study on *Francisella tularensis*. *BMC Genomics* 19:505. doi: 10.1186/s12864-018-4873-4879

Mei, S., and Zhang, K. (2020). In silico unravelling pathogen-host signaling cross-talks via pathogen mimicry and human protein-protein interaction networks. *Comput. Struct. Biotechnol. J.* 18, 100–113. doi: 10.1016/j.csbj.2019.12.008

Mei, S., and Zhu, H. (2014a). AdaBoost based multi-instance transfer learning for predicting proteome-wide interactions between *Salmonella* and human proteins. *PLoS One* 9:e110488. doi: 10.1371/journal.pone.0110488

Mei, S., and Zhu, H. (2014b). Computational reconstruction of proteome-wide protein interaction networks between HTLV retroviruses and Homo sapiens. *BMC Bioinformatics* 15:245. doi: 10.1186/1471-2105-15-245

Mendes, V., Galvão, I., and Vieira, A. T. (2019). Mechanisms by which the gut microbiota influences cytokine production and modulates host inflammatory responses. *J. Interferon Cytokine Res.* 39, 393–409. doi: 10.1089/jir.2019.0011

Mészáros, B., Erdos, G., and Dosztányi, Z. (2018). IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res.* 46, W329–W337. doi: 10.1093/nar/gky384

Meyer, J. M., Leempoel, K., Losapio, G., and Hadly, E. A. (2020). Molecular ecological network analyses: an effective conservation tool for the assessment of biodiversity, trophic interactions, and community structure. *Front. Ecol. Evol.* 8:588430. doi: 10.3389/fevo.2020.588430

Miryala, S. K., Anbarasu, A., and Ramaiah, S. (2018). Discerning molecular interactions: a comprehensive review on biomolecular interaction databases and network analysis tools. *Gene* 642, 84–94. doi: 10.1016/j.gene.2017.11.028

Mitchell, A. L., Attwood, T. K., Babbitt, P. C., Blum, M., Bork, P., Bridge, A., et al. (2019). InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res.* 47, D351–D360. doi: 10.1093/nar/gky1100

Mizianty, M. J., Peng, Z., and Kurgan, L. (2013). MFDp2: accurate predictor of disorder in proteins by fusion of disorder probabilities, content and profiles. *Intrinsically Disord Proteins* 1:e24428. doi: 10.4161/idp.24428

Mollentze, N., and Streicker, D. G. (2020). Viral zoonotic risk is homogenous among taxonomic orders of mammalian and avian reservoir hosts. *Proc. Natl. Acad. Sci. U S A.* 117, 9423–9430. doi: 10.1073/pnas.1919176117

Muller, E. E. L., Glaab, E., May, P., Vlassis, N., and Wilmes, P. (2013). Condensing the omics fog of microbial communities. *Trends Microbiol.* 21, 325–333. doi: 10.1016/j.tim.2013.04.009

Negi, S., Pandey, S., Srinivasan, S. M., Mohammed, A., and Guda, C. (2015). LocSigDB: a database of protein localization signals. *Database (Oxford)* 2015:bav003. doi: 10.1093/database/bav003

Nourani, E., Khunjush, F., and Durmuñ, S. (2016). Computational prediction of virus-human protein-protein interactions using embedding kernelized heterogeneous data. *Mol. Biosyst.* 12, 1976–1986. doi: 10.1039/c6mb00065g

Nouretdinov, I., Gammerman, A., Qi, Y., and Klein-Seetharaman, J. (2012). Determining confidence of predicted interactions between HIV-1 and human proteins using conformal method. *Pac. Symp. Biocomput.* 2012, 311–322.

Oates, M. E., Romero, P., Ishida, T., Ghalwash, M., Mizianty, M. J., Xue, B., et al. (2013). D2P2: database of disordered protein predictions. *Nucleic Acids Res.* 41, D508–D516. doi: 10.1093/nar/gks1226

Ohland, C. L., and Jobin, C. (2015). Microbial activities and intestinal homeostasis: a delicate balance between health and disease. *Cell. Mol. Gastroenterol. Hepatol.* 1, 28–40. doi: 10.1016/j.jcmgh.2014.11.004

Okuda, S., Watanabe, Y., Moriya, Y., Kawano, S., Yamamoto, T., Matsumoto, M., et al. (2017). jPOSTrepo: an international standard data repository for proteomes. *Nucleic Acids Res.* 45, D1107–D1111. doi: 10.1093/nar/gkw1080

Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., et al. (2014). The MIntAct project –IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* 42, D358–D363. doi: 10.1093/nar/gkt1115

Parkinson, H., Kapushesky, M., Shojatalab, M., Abeygunawardena, N., Coulson, R., Farne, A., et al. (2007). ArrayExpress–a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res.* 35, D747–D750. doi: 10.1093/nar/gkl995

Paull, E. O., Carlin, D. E., Niepel, M., Sorger, P. K., Haussler, D., and Stuart, J. M. (2013). Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion Through Interacting Events (TieDIE). *Bioinformatics* 29, 2757–2764. doi: 10.1093/bioinformatics/btt471

Payne, W. E., and Garrels, J. I. (1997). Yeast Protein Database (YPD): a database for the complete proteome of *Saccharomyces cerevisiae*. *Nucleic Acids Res.* 25, 57–62. doi: 10.1093/nar/25.1.57

Peabody, M. A., Laird, M. R., Vlasschaert, C., Lo, R., and Brinkman, F. S. L. (2016). PSORTdb: expanding the bacteria and archaea protein subcellular localization database to better reflect diversity in cell envelope structures. *Nucleic Acids Res.* 44, D663–D668. doi: 10.1093/nar/gkv1271

Pedamallu, C. S., and Ozdamar, L. (2014). "A review on protein-protein interaction network databases," in *Modeling, Dynamics, Optimization and Bioeconomics I Springer Proceedings in Mathematics & Statistics*, eds A. A. Pinto and D. Zilberman (Cham: Springer International Publishing), 511–519. doi: 10.1007/978-3-319-04849-9_30

Penny, H. A., Hodge, S. H., and Hepworth, M. R. (2018). Orchestration of intestinal homeostasis and tolerance by group 3 innate lymphoid cells. *Semin. Immunopathol.* 40, 357–370. doi: 10.1007/s00281-018-0687-688

Perez-Riverol, Y., Csordas, A., Bai, J., Bernal-Llinares, M., Hewapathirana, S., Kundu, D. J., et al. (2019). The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.* 47, D442–D450. doi: 10.1093/nar/gky1106

Perkins, J. R., Diboun, I., Dessailly, B. H., Lees, J. G., and Orengo, C. (2010). Transient protein-protein interactions: structural, functional, and network properties. *Structure* 18, 1233–1243. doi: 10.1016/j.str.2010.08.007

Peters, J. M., Solomon, S. L., Itoh, C. Y., and Bryson, B. D. (2019). Uncovering complex molecular networks in hostŰpathogen interactions using systems biology. *Emerg. Top. Life Sci.* 3, 371–378. doi: 10.1042/ETLS20180174

Pey, J., Tobalina, L., de Cisneros, J. P. J., and Planes, F. J. (2013). A network-based approach for predicting key enzymes explaining metabolite abundance alterations in a disease phenotype. *BMC Syst. Biol.* 7:2. doi: 10.1186/1752-0509-7-62

Pickard, J. M., Zeng, M. Y., Caruso, R., and Núñez, G. (2017). Gut microbiota: role in pathogen colonization, immune responses, and inflammatory disease. *Immunol. Rev.* 279, 70–89. doi: 10.1111/imr.12567

Pierleoni, A., Martelli, P. L., Fariselli, P., and Casadio, R. (2007). eSLDB: eukaryotic subcellular localization database. *Nucleic Acids Res.* 35, D208–D212. doi: 10.1093/nar/gkl775

Pryor, R., Norvaisas, P., Marinos, G., Best, L., Thingholm, L. B., Quintaneiro, L. M., et al. (2019). Host-Microbe-Drug-Nutrient screen identifies bacterial effectors of metformin therapy. *Cell* 178, 1299–1312.e29. doi: 10.1016/j.cell.2019.08.003

Qi, Y., Tastan, O., Carbonell, J. G., Klein-Seetharaman, J., and Weston, J. (2010). Semi-supervised multi-task learning for predicting interactions between HIV-1 and human proteins. *Bioinformatics* 26, i645–i652. doi: 10.1093/bioinformatics/btq394

Qiu, Y.-Q., Zhang, S., Zhang, X.-S., and Chen, L. (2010). Detecting disease associated modules and prioritizing active genes based on high throughput data. *BMC Bioinformatics* 11:26. doi: 10.1186/1471-2105-11-26

Raghavachari, B., Tasneem, A., Przytycka, T. M., and Jothi, R. (2008). DOMINE: a database of protein domain interactions. *Nucleic Acids Res.* 36, D656–D661. doi: 10.1093/nar/gkm761

Rajasekharan, S., Rana, J., Gulati, S., Sharma, S. K., Gupta, V., and Gupta, S. (2013). Predicting the host protein interactors of *Chandipura* virus using a structural similarity-based approach. *Pathog Dis.* 69, 29–35. doi: 10.1111/2049-632X.12064

Rana, A., Ahmed, M., Rub, A., and Akhter, Y. (2015). A tug-of-war between the host and the pathogen generates strategic hotspots for the development of novel therapeutic interventions against infectious diseases. *Virulence* 6, 566–580. doi: 10.1080/21505594.2015.1062211

Rastogi, S., and Rost, B. (2011). LocDB: experimental annotations of localization for homo sapiens and *Arabidopsis thaliana*. *Nucleic Acids Res.* 39, D230–D234. doi: 10.1093/nar/gkq927

Rodenburg, S. Y. A., Seidl, M. F., Judelson, H. S., Vu, A. L., Govers, F., and de Ridder, D. (2019). Metabolic model of the phytophthora infestans-tomato interaction reveals metabolic switches during host colonization. *mBio* 10:e00454-19. doi: 10.1128/mBio.00454-419

Romano, P., Dräger, A., Fiannaca, A., Giugno, R., La Rosa, M., et al. (2019). The 2017 Network Tools and Applications in Biology (NETTAB) workshop: aims, topics and outcomes. *BMC Bioinformatics* 20:125. doi: 10.1186/s12859-019-2681-2680

Roume, H., Heintz-Buschart, A., Muller, E. E. L., May, P., Satagopam, V. P., Laczny, C. C., et al. (2015). Comparative integrated omics: identification of key functionalities in microbial community-wide metabolic networks. *NPJ Biofilms Microb.* 1:15007. doi: 10.1038/npjbiofilms.2015.7

Saçar Demirci, M. D., and Adan, A. (2020). Computational analysis of microRNA-mediated interactions in SARS-CoV-2 infection. *PeerJ* 8:e9369. doi: 10.7717/peerj.9369

Sahu, S. S., Weirick, T., and Kaundal, R. (2014). Predicting genome-scale *Arabidopsis-Pseudomonas* syringae interactome using domain and interolog-based approaches. *BMC Bioinformatics* 15(Suppl. 11):S13. doi: 10.1186/1471-2105-15-S11-S13

Saik, O. V., Ivanisenko, T. V., Demenkov, P. S., and Ivanisenko, V. A. (2016). Interactome of the hepatitis C virus: literature mining with ANDSystem. *Virus Res.* 218, 40–48. doi: 10.1016/j.virusres.2015.12.003

Samal, S. S., Radulescu, O., Weber, A., and Fröhlich, H. (2017). Linking metabolic network features to phenotypes using sparse group lasso. Bioinformatics 33, 3445–3453. doi: 10.1093/bioinformatics/btx427

Schaubeck, M., Clavel, T., Calasan, J., Lagkouvardos, I., Haange, S. B., Jehmlich, N., et al. (2016). Dysbiotic gut microbiota causes transmissible Crohn's disease-like ileitis independent of failure in antimicrobial defence. *Gut* 65, 225–237. doi: 10.1136/gutjnl-2015-309333

Schleker, S., Garcia-Garcia, J., Klein-Seetharaman, J., and Oliva, B. (2012). Prediction and comparison of *Salmonella-human* and *Salmonella-Arabidopsis* interactomes. *Chem. Biodivers.* 9, 991–1018. doi: 10.1002/cbdv.201100392

Schmidt, T., Samaras, P., Frejno, M., Gessulat, S., Barnert, M., Kienegger, H., et al. (2018). ProteomicsDB. *Nucleic Acids Res.* 46, D1271–D1281. doi: 10.1093/nar/gkx1029

Schommer, N. N., and Gallo, R. L. (2013). Structure and function of the human skin microbiome. *Trends Microbiol.* 21, 660–668. doi: 10.1016/j.tim.2013.10.001

Schweppe, D. K., Harding, C., Chavez, J. D., Wu, X., Ramage, E., Singh, P. K., et al. (2015). Host-Microbe Protein Interactions during Bacterial Infection. *Chem. Biol.* 22, 1521–1530. doi: 10.1016/j.chembiol.2015.09.015

Shah, P., Fritz, J. V., Glaab, E., Desai, M. S., Greenhalgh, K., Frachet, A., et al. (2016). A microfluidics-based in vitro model of the gastrointestinal human-microbe interface. *Nat. Commun.* 7:11535. doi: 10.1038/ncomms11535

Sharma, V., Eckels, J., Schilling, B., Ludwig, C., Jaffe, J. D., MacCoss, M. J., et al. (2018). Panorama public: a public repository for quantitative data sets processed in skyline. *Mol. Cell Proteomics* 17, 1239–1244. doi: 10.1074/mcp.RA117.000543

Shastry, K. A., and Sanjay, H. A. (2020). "Machine learning for bioinformatics," in *Statistical Modelling and Machine Learning Principles for Bioinformatics Techniques, Tools, and Applications Algorithms for Intelligent Systems* eds. K. G. Srinivasa, G. M. Siddesh, and S. R. Manisekhar (Singapore: Springer Singapore), 25–39. doi: 10.1007/978-981-15-2445-5_3

Shaw, K. A., Bertha, M., Hofmekler, T., Chopra, P., Vatanen, T., Srivatsa, A., et al. (2016). Dysbiosis, inflammation, and response to treatment: a longitudinal study of pediatric subjects with newly diagnosed inflammatory bowel disease. *Genome Med.* 8:75. doi: 10.1186/s13073-016-0331-y

Shirahama, S., Miki, A., Kaburaki, T., and Akimitsu, N. (2020). Long non-coding RNAs involved in pathogenic infection. *Front. Genet.* 11:454. doi: 10.3389/fgene.2020.00454

Shoombuatong, W., Hongjaisee, S., Barin, F., Chaijaruwanich, J., and Samleerat, T. (2012). HIV-1 CRF01_AE coreceptor usage prediction using kernel methods based logistic model trees. *Comput. Biol. Med.* 42, 885–889. doi: 10.1016/j.compbiomed.2012.06.011

Silmon de Monerri, N. C., and Kim, K. (2014). Pathogens hijack the epigenome: a new twist on host-pathogen interactions. *Am. J. Pathol.* 184, 897–911. doi: 10.1016/j.ajpath.2013.12.022

Singh, N., Bhatia, V., Singh, S., and Bhatnagar, S. (2019). MorCVD: a unified database for host-pathogen protein-protein interactions of cardiovascular diseases related to microbes. *Sci. Rep.* 9:4039. doi: 10.1038/s41598-019-40704-5

Sudhakar, P., Jacomin, A.-C., Hautefort, I., Samavedam, S., Fatemian, K., Ari, E., et al. (2019). Targeted interplay between bacterial pathogens and host autophagy. *Autophagy* 15, 1620–1633. doi: 10.1080/15548627.2019.1590519

Sun, J., Yang, L.-L., Chen, X., Kong, D.-X., and Liu, R. (2018). Integrating multifaceted information to predict *Mycobacterium tuberculosis*-human protein-protein interactions. *J. Proteome Res.* 17, 3810–3823. doi: 10.1021/acs.jproteome.8b00497

Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., et al. (2017). The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* 45, D362–D368. doi: 10.1093/nar/gkw937

Tastan, O., Qi, Y., Carbonell, J. G., and Klein-Seetharaman, J. (2009). Prediction of interactions between HIV-1 and human proteins by information integration. *Pac. Symp. Biocomput.* 2009, 516–527. doi: 10.1142/9789812836939_0049

The UniProt Consortium (2018). UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* 46:2699. doi: 10.1093/nar/gky092

Thiele, I., Heinken, A., and Fleming, R. M. T. (2013a). A systems biology approach to studying the role of microbes in human health. *Curr. Opin. Biotechnol.* 24, 4–12. doi: 10.1016/j.copbio.2012.10.001

Thiele, I., Swainston, N., Fleming, R. M. T., Hoppe, A., Sahoo, S., Aurich, M. K., et al. (2013b). A community-driven global reconstruction of human metabolism. *Nat. Biotechnol.* 31, 419–425. doi: 10.1038/nbt.2488

Thiele, I., Sahoo, S., Heinken, A., Hertel, J., Heirendt, L., Aurich, M. K., et al. (2020). Personalized whole-body models integrate metabolism, physiology, and the gut microbiome. *Mol. Syst. Biol.* 16:e8982. doi: 10.15252/msb.20198982

Thieu, T., Joshi, S., Warren, S., and Korkin, D. (2012). Literature mining of host-pathogen interactions: comparing feature-based supervised learning and language-based approaches. *Bioinformatics* 28, 867–875. doi: 10.1093/bioinformatics/bts042

Thompson, D., Regev, A., and Roy, S. (2015). Comparative analysis of gene regulatory networks: from network reconstruction to evolution. *Annu. Rev. Cell Dev. Biol.* 31, 399–428. doi: 10.1146/annurev-cellbio-100913-112908

Thul, P. J., and Lindskog, C. (2018). The human protein atlas: a spatial map of the human proteome. *Protein Sci.* 27, 233–244. doi: 10.1002/pro.3307

Tramontano, M., Andrejev, S., Pruteanu, M., Klünemann, M., Kuhn, M., Galardini, M., et al. (2018). Nutritional preferences of human gut bacteria reveal their metabolic idiosyncrasies. *Nat. Microbiol.* 3, 514–522. doi: 10.1038/s41564-018-0123-129

Türei, D., Korcsmáros, T., and Saez-Rodriguez, J. (2016). OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nat. Methods* 13, 966–967. doi: 10.1038/nmeth.4077

Tyagi, N., Krishnadev, O., and Srinivasan, N. (2009). Prediction of protein-protein interactions between *Helicobacter pylori* and a human host. *Mol. Biosyst.* 5, 1630–1635. doi: 10.1039/b906543c

Valli, R. X. E., Lyng, M., and Kirkpatrick, C. L. (2020). There is no hiding if you Seq: recent breakthroughs in *Pseudomonas aeruginosa* research revealed by genomic and transcriptomic next-generation sequencing. *J. Med. Microbiol.* 69, 162–175. doi: 10.1099/jmm.0.001135

Vandin, F., Upfal, E., and Raphael, B. J. (2011). Algorithms for detecting significantly mutated pathways in cancer. *J. Comput. Biol.* 18, 507–522. doi: 10.1089/cmb.2010.0265

Veres, D. V., Gyurkó, D. M., Thaler, B., Szalay, K. Z., Fazekas, D., Korcsmáros, T., et al. (2015). ComPPI: a cellular compartment-specific database for protein-protein interaction network analysis. *Nucleic Acids Res.* 43, D485–D493. doi: 10.1093/nar/gku1007

Via, A., Uyar, B., Brun, C., and Zanzoni, A. (2015). How pathogens use linear motifs to perturb host cell networks. *Trends Biochem. Sci.* 40, 36–48. doi: 10.1016/j.tibs.2014.11.001

Wallqvist, A., Wang, H., Zavaljevski, N., Memiševiæ, V., Kwon, K., Pieper, R., et al. (2017). Mechanisms of action of *Coxiella burnetii* effectors inferred from host-pathogen protein interactions. *PLoS One* 12:e0188071. doi: 10.1371/journal.pone.0188071

Wang, B., Yao, M., Lv, L., Ling, Z., and Li, L. (2017). The human microbiota in health and disease. *Engineering* 3, 71–82. doi: 10.1016/J.ENG.2017.01.008

Ward, J. J., McGuffin, L. J., Bryson, K., Buxton, B. F., and Jones, D. T. (2004). The DISOPRED server for the prediction of protein disorder. *Bioinformatics* 20, 2138–2139. doi: 10.1093/bioinformatics/bth195

Weiberg, A., Wang, M., Bellinger, M., and Jin, H. (2014). Small RNAs: a new paradigm in plant-microbe interactions. *Annu. Rev. Phytopathol.* 52, 495–516. doi: 10.1146/annurev-phyto-102313-045933

Wen, C., Zheng, Z., Shao, T., Liu, L., Xie, Z., Le Chatelier, E., et al. (2017). Quantitative metagenomics reveals unique gut microbiome biomarkers in ankylosing spondylitis. *Genome Biol.* 18:42. doi: 10.1186/s13059-017-1271-6

Wong, A. C. N., Vanhove, A. S., and Watnick, P. I. (2016). The interplay between intestinal bacteria and host metabolism in health and disease: lessons from *Drosophila melanogaster*. *Dis. Model. Mech.* 9, 271–281. doi: 10.1242/dmm.023408

Wong, E., Baur, B., Quader, S., and Huang, C.-H. (2012). Biological network motif detection: principles and practice. *Brief. Bioinformatics* 13, 202–215. doi: 10.1093/bib/bbr033

Wuchty, S. (2011). Computational prediction of host-parasite protein interactions between *Plasmodium falciparum* and *H. sapiens*. *PLoS One* 6:e26960. doi: 10.1371/journal.pone.0026960

Xue, B., Dunbrack, R. L., Williams, R. W., Dunker, A. K., and Uversky, V. N. (2010). PONDR-FIT: a meta-predictor of intrinsically disordered amino acids. *Biochim. Biophys. Acta* 1804, 996–1010. doi: 10.1016/j.bbapap.2010.01.011

Yang, G., Peng, M., Tian, X., and Dong, S. (2017). Molecular ecological network analysis reveals the effects of probiotics and florfenicol on intestinal microbiota homeostasis: an example of sea cucumber. *Sci. Rep.* 7:4778. doi: 10.1038/s41598-017-05312-5311

Yijing, L., Haixiang, G., Xiao, L., Yanan, L., and Jinling, L. (2016). Adapted ensemble classification algorithm based on multiple classifier system and feature selection for classifying multi-class imbalanced data. *Knowledge-Based Systems* 94, 88–104. doi: 10.1016/j.knosys.2015.11.013

Yilmaz, B., Juillerat, P., Øyås, O., Ramon, C., Bravo, F. D., Franc, Y., et al. (2019). Microbial network disturbances in relapsing refractory Crohn's disease. *Nat. Med.* 25, 323–336. doi: 10.1038/s41591-018-0308-z

Younesi, E. (2015). Disease systems modeling for discovery of mechanistic biomarkers. *Eur. J. Mol. Clin. Med.* 2:61. doi: 10.1016/j.nhtm.2014.11.023

Yu, H., Xue, D., Wang, Y., Zheng, W., Zhang, G., and Wang, Z.-L. (2020). Molecular ecological network analysis of the response of soil microbial communities to depth gradients in farmland soils. *Microbiologyopen* 9:e983. doi: 10.1002/mbo3.983

Zampieri, G., Vijayakumar, S., Yaneske, E., and Angione, C. (2019). Machine and deep learning meet genome-scale metabolic modeling. *PLoS Comput. Biol.* 15:e1007084. doi: 10.1371/journal.pcbi.1007084

Zhang, A., He, L., and Wang, Y. (2017a). Prediction of GCRV virus-host protein interactome based on structural motif-domain interactions. *BMC Bioinformatics* 18:145. doi: 10.1186/s12859-017-1500-1508

Zhang, M., Su, Q., Lu, Y., Zhao, M., and Niu, B. (2017b). Application of machine learning approaches for protein-protein interactions prediction. *Med. Chem.* 13, 506–514. doi: 10.2174/1573406413666170522150940

Zheng, Q., Zhang, M., Zhang, T., Li, X., Zhu, M., and Wang, X. (2020). Insights from metagenomic, metatranscriptomic, and molecular ecological network analyses into the effects of chromium nanoparticles on activated sludge system. *Front. Environ. Sci. Eng.* 14:60. doi: 10.1007/s11783-020-1239-1238

Zhou, H., Gao, S., Nguyen, N. N., Fan, M., Jin, J., Liu, B., et al. (2014). Stringent homology-based prediction of *H. sapiens-M. tuberculosis* H37Rv protein-protein interactions. *Biol. Direct* 9:5. doi: 10.1186/1745-6150-9-5

Zhou, H., Rezaei, J., Hugo, W., Gao, S., Jin, J., Fan, M., et al. (2013). Stringent DDI-based prediction of *H. sapiens-M. tuberculosis* H37Rv protein-protein interactions. *BMC Syst. Biol.* 7:S6. doi: 10.1186/1752-0509-7-S6-S6

Zhou, J., Deng, Y., Luo, F., He, Z., Tu, Q., and Zhi, X. (2010). Functional molecular ecological networks. *mBio* 1:e00169-10. doi: 10.1128/mBio.00169-110

Zhou, X., Park, B., Choi, D., and Han, K. (2018). A generalized approach to predicting protein-protein interactions between virus and host. *BMC Genomics* 19:568. doi: 10.1186/s12864-018-4924-4922

Conflict of Interest: BV received lecture fees from AbbVie, Ferring Pharmaceuticals, Janssen, R-Biopharm, and Takeda; consultancy fees from Janssen and Sandoz. SV: research grant: MSD, AbbVie, Takeda, Pfizer, and J&J; lecture fee: MSD, AbbVie, Takeda, Ferring, Centocor, Hospira, Pfizer, J&J, and Genentech/Roche; consultancy: MSD, AbbVie, Takeda, Ferring, Centocor, Hospira, Pfizer, J&J, Genentech/Roche, Celgene, Mundipharma, Celltrion, SecondGenome, Prometheus, Shire, ProDigest, Gilead, and Galapagos. SV is a senior clinical investigator of the Research Foundation–Flanders (FWO). The work of TK was supported by BenevolentAI and Unilever.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 18 June 2021
doi: 10.3389/fmicb.2021.644487

Beating Naive Bayes at Taxonomic Classification of 16S rRNA Gene Sequences

Michal Ziemski[1†], Treepop Wisanwanichthan[2†], Nicholas A. Bokulich[1‡] and Benjamin D. Kaehler[2*‡]*

[1] Laboratory of Food Systems Biotechnology, Institute of Food, Nutrition, and Health, ETH Zürich, Zurich, Switzerland,
[2] School of Science, University of New South Wales, Canberra, ACT, Australia

Naive Bayes classifiers (NBC) have dominated the field of taxonomic classification of amplicon sequences for over a decade. Apart from having runtime requirements that allow them to be trained and used on modest laptops, they have persistently provided class-topping classification accuracy. In this work we compare NBC with random forest classifiers, neural network classifiers, and a perfect classifier that can only fail when different species have identical sequences, and find that in some practical scenarios there is little scope for improving on NBC for taxonomic classification of 16S rRNA gene sequences. Further improvements in taxonomy classification are unlikely to come from novel algorithms alone, and will need to leverage other technological innovations, such as ecological frequency information.

Keywords: microbiome, metagenomics, marker-gene sequencing, taxonomic classification, machine learning, neural networks

OPEN ACCESS

Edited by:
Isabel Moreno Indias, Universidad de Málaga, Spain

Reviewed by:
Hao Lin, University of Electronic Science and Technology of China, China
Sotiris Kotsiantis, University of Patras, Greece

****Correspondence:***
Nicholas A. Bokulich
nicholas.bokulich@hest.ethz.ch
Benjamin D. Kaehler
b.kaehler@adfa.edu.au

†These authors share first authorship

‡These authors share senior authorship

Specialty section:
This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology

Received: *21 December 2020*
Accepted: *31 May 2021*
Published: *18 June 2021*

Citation:
Ziemski M, Wisanwanichthan T, Bokulich NA and Kaehler BD (2021) Beating Naive Bayes at Taxonomic Classification of 16S rRNA Gene Sequences. Front. Microbiol. 12:644487. doi: 10.3389/fmicb.2021.644487

INTRODUCTION

Microbial communities are integral components of diverse ecosystems on planet Earth, supporting both environmental and human health (The Human Microbiome Project Consortium, 2012; Thompson et al., 2017). Investigating the role of microorganisms in these environments often involves characterizing the composition of these communities using high-throughput DNA sequencing methods, most commonly of universal marker genes, such as small subunit rRNA genes (Thompson et al., 2017). Even short sequences (e.g., as obtained from "second-generation" sequencing instruments) of 16S rRNA gene hypervariable domains can differentiate bacterial families and genera (Liu et al., 2008), making these marker genes popular targets for microbial census studies.

A critical step in any microbial census study is the taxonomic classification of observed DNA sequences, to infer the relative abundance of different taxonomic groups. This is performed by comparison of observed sequences to a reference database of sequences from known taxa, using an appropriate taxonomic classifier (Robeson et al., 2020). A large number of taxonomic classification methods have been developed and benchmarked for classification of marker gene sequences (Bokulich et al., 2018b; Gardner et al., 2019), but among the most successful and ubiquitous in microbiome studies have been naive Bayes classifiers (NBC). The primacy of NBC was established by the Ribosomal Database Project (RDP) classifier (Wang et al., 2007), which utilized an NBC and demonstrated that genus-level accuracy could be achieved from short 16S rRNA gene sequences. The superiority of NBC for marker-gene sequence classification has proven robust over time,

as we have shown more recently with the various classifiers implemented in q2-feature-classifier (Bokulich et al., 2018b), a taxonomic classification plugin for the popular QIIME 2 microbiomics software platform (Bolyen et al., 2019). Furthermore, we demonstrated that the accuracy of NBC could be significantly enhanced by providing ecological information about the expected frequency of different taxonomic groups in specific natural environments (Kaehler et al., 2019) to enable more reliable species-level classification of 16S rRNA gene sequences. This improves upon the assumptions of earlier NBC for marker-gene sequences (e.g., RDP classifier), which assume uniform class weights, i.e., that microbial species are equally likely to be observed.

Newer methods for taxonomic classification have been developed and tested, but have failed to reliably exceed the accuracy of NBC for marker-gene taxonomic classification, both in individual benchmarks (Lu and Salzberg, 2020) and in independent benchmarks (Almeida et al., 2018; Gardner et al., 2019). Notably, all benchmarks to date (except those in Kaehler et al., 2019) have tested NBC with uniform class weights, underlining that naive Bayes remains most accurate even without full optimization for specific sample types. This led us to consider three questions in the current study:

1. Could taxonomic frequency information benefit other taxonomic classifiers?
2. Could newer supervised learning algorithms exceed the accuracy of NBC?
3. Do decreasing performance advances in the microbiome taxonomy classification literature indicate that we are reaching an upper limit of performance for classification of short marker-gene sequences?

Class weight information can be utilized by a variety of supervised classification methods, so we hypothesized that using class weights could provide these methods with a much-needed performance boost. We chose two newer machine learning classification algorithms that have been successfully applied to other problems in bioinformatics (e.g., sample classification, e.g., Bokulich et al., 2016, 2018a; Roguet et al., 2018), but little-explored for DNA sequence annotation: Random Forests (RF) (Breiman, 2001) and convolutional neural networks (CNN) (Lecun et al., 1998). These algorithms have shown favorable performance against the RDP classifier in isolated tests (Chaudhary et al., 2015; Fiannaca et al., 2018; Busia et al., 2019; Zhao et al., 2020) but have not been independently benchmarked, nor compared against NBC with ecologically informed class weights.

We demonstrate that RF and CNN come close to but fail to exceed the accuracy of NBC when utilizing class weight information. Additionally, we use a "perfect" classifier to establish an upper bound for classification accuracy. We discover that, at least for short reads of 150 nt, there can be almost no improvement over an NBC if class weights are used. If longer reads are used (all of the V4 region) then there is limited scope for improvement, but again only if class weights are used. Finally, NBCs remain easier and faster to train than RF and CNN classifiers with fewer hardware requirements.

RESULTS

We selected RF and CNN classifiers as promising methods for DNA sequence taxonomy classification, due to promising performance of various implementations in recent isolated reports (Chaudhary et al., 2015; Fiannaca et al., 2018; Busia et al., 2019; Zhao et al., 2020). In particular, the use of ensemble classification by RF is a potentially attractive means of efficiently calculating class probabilities via random selection of sequence data in each decision tree. The ability of CNNs to learn complex patterns, and in particular to model spatial organization in sequence data (Busia et al., 2019), make CNNs promising for DNA sequence annotation tasks. We utilized a kmer bagging approach for feature extraction prior to both RF and NBC classification, as has been commonly implemented in NBC including the RDP classifier (Wang et al., 2007; Bokulich et al., 2018b). However, kmer bagging fails to leverage the mid- to long-range spatial organization of DNA sequences. Hence, we used a Word2Vec (Mikolov et al., 2013) encoding for feature extraction prior to CNN classification, similar to the spatial encoding schemes implemented for other CNN classifiers (Busia et al., 2019; Zhao et al., 2020).

Random Forest Classifiers

We performed hyperparameter tuning of the RF classifiers following a two-tiered approach. Cross validation was performed on sequences in the Greengenes reference data set (McDonald et al., 2012) and on sample compositions derived from real samples downloaded from the Qiita database (Gonzalez et al., 2018). All of the tests of the NBC and RF classifiers that we performed used taxonomic weighting information (Kaehler et al., 2019).

First, a grid search was performed on a comparatively smaller data set to select hyperparameters with primary performance effects (all samples of 150 nt length labeled as sediment (non-saline) in Qiita on 20 March 2019 (Thompson et al., 2017), 188 samples, downloaded using q2-clawback (Kaehler et al., 2019) see **Supplementary Material** and **Supplementary Table 1** for details). A second grid search was performed on a much larger animal distal gut data set (downloaded from Qiita on 23 May, 2019 with the same parameters, 22,454 samples). The results of the initial tests were that max_depth and max_features were the only classifier parameters that had a meaningful impact on classification accuracy. Except for confidence, all parameters are those of the scikit-learn classifier (see section "Materials and Methods"). Additionally, a confidence parameter of 0.7 was found to give greater accuracy than a confidence parameter of 0.9.

In all cases, classification accuracy was measured using F-measure for species-level classification. We also tested final results using the Matthews correlation coefficient (MCC) (Chicco and Jurman, 2020). MCC was chosen to reduce the reported bias in F-measure in the presence of imbalanced classes. In all cases results were qualitatively the same. For the same confidence

level, parameter choice also gave the same rankings for MCC as F-measure (see **Supplementary Figure 1**).

The parameters selected for the second tier of parameter tuning are shown in **Table 1**. A full grid search testing all combinations of parameters was not performed because of operational difficulties balancing requests for walltime, number of CPUs, and memory usage on a shared computational resource (see **Supplementary Table 2** and **Supplementary Figure 4**) and negligible impacts on performance (**Figure 1**). If it is not possible to train a classifier on a machine with 14 CPUs and 3TB of memory in under 24 h, it is not useful to the wider community regardless of accuracy (Bokulich et al., 2020), and hence these configurations were not pursued further. A max_depth of None implies that nodes were expanded until all nodes were pure. max_features of None implies that the maximum number of features was the number of features.

Results indicate that maximum tree depth (max_depth) exerted the greatest influence on classification accuracy (**Figure 1** and **Supplementary Figure 1**). Regardless of the confidence level, increasing the maximum depth leads to an increase in F-measure. The most significant change can be observed between 16 and 64 nodes (average F-measures of 0.636 ∓ 0.006 and 0.768 ∓ 0.004 at confidence 0.7, respectively, standard error measured over folds). Increasing max_depth beyond 64, however, does not lead to an appreciable increase in accuracy, and using unlimited tree depth (i.e., max_depth = None; tree nodes are expanded until leaf purity is achieved) yields marginally higher F-measures at all confidence levels ($F = 0.779 \mp 0.005$ at confidence 0.7) (**Figure 1**).

Decreasing the number of features (max_features) to be taken into account while deciding on node splitting resulted in a modest decrease in classification accuracy [0.636 ∓ 0.006 and 0.608 ∓ 0.003 average F-measure for using all of the features as compared to sqrt (number of features)]. This performance decrease was least pronounced at lower confidence levels. Similarly, increasing the number of estimators (n_estimator, i.e., trees in the forest) had no or very low influence on classification accuracy, regardless of the confidence level. Increasing the number of estimators from 100 to 1,000 reliably caused memory issues, however, particularly with a maximum tree depth of 64 (see **Supplementary Table 2**).

None of the parameter sets tested in our study outperformed the NBC at any of the confidence levels we tested (Wilcoxon signed-rank $p < 0.05$). To test whether reducing the classification confidence threshold further beyond the level of 0.6 could help increase RF's performance, we trained and evaluated an additional set of classifiers with fixed parameters (max. number of features, 100 estimators, max. tree depth) while varying confidence in the range 0.3–0.5. Decreasing the confidence marginally increased the test set's recall and F-measures at the cost of precision (**Figure 1** and **Supplementary Figure 1**), however the accuracy achieved by the NBC could still not be obtained (**Figure 1**).

Interestingly, precision of the RF classifier tested with most of the parameter sets could in many cases outperform the NB model and it was rather insensitive to parameter changes given a confidence level (**Supplementary Figure 1**, top panel). It was the recall, however, that not only varied greatly between parameter sets, but also could never come close to that of the NB (**Supplementary Figure 1**, bottom panel).

TABLE 1 | Parameter values used for computationally intensive grid search on animal-distal-gut samples.

Parameter	Values		
n_estimators	100	1,000	–
max_depth	16	64	None
max_features	sqrt	None	–
Confidence	0.6	0.7	0.8

Convolutional Neural Networks

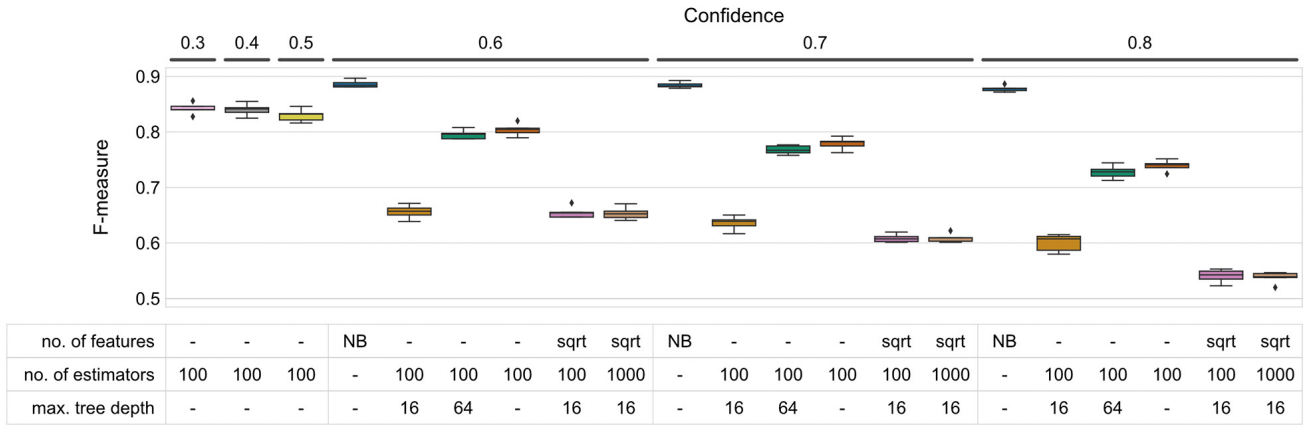
Following our tests of RF classifiers, we were interested in evaluating whether we could leverage recent advances in neural network-based models for superior taxonomic classification. Cross validation was performed as described for RF. To reduce run time, we used a relatively small data set that consisted of all 5,632 of the animal distal gut 150 nt samples that were available from Qiita on 1 June, 2018 (downloaded using q2-clawback; Kaehler et al., 2019).

More specifically, we focused on CNNs as their performance is favorable in the literature (Chaudhary et al., 2015; Fiannaca et al., 2018; Busia et al., 2019; Zhao et al., 2020) and for their relatively parsimonious parameterization and insensitivity to insertion and deletion events. Before feeding the DNA sequences to the network, we applied the Word2Vec model in its Continuous-bag-of-words (CBOW) implementation to convert genetic information into a series of 300-element vectors. That not only allowed us to convert the k-mers into numerical values but also carried additional information about relatedness/similarity between any two k-mers within a given sequence.

For most of our tests we used a simple neural network with a single (one-dimensional) convolutional layer followed by a global max pooling layer and a classification layer (**Figure 2A**, architecture I). We varied the number of filters and kernel size of each filter (see **Table 2**) to test which of those parameters would have the greatest influence on the model performance (measured as Precision, Recall, F-measure, and MCC at the species level, similarly as was done for the Random Forest models).

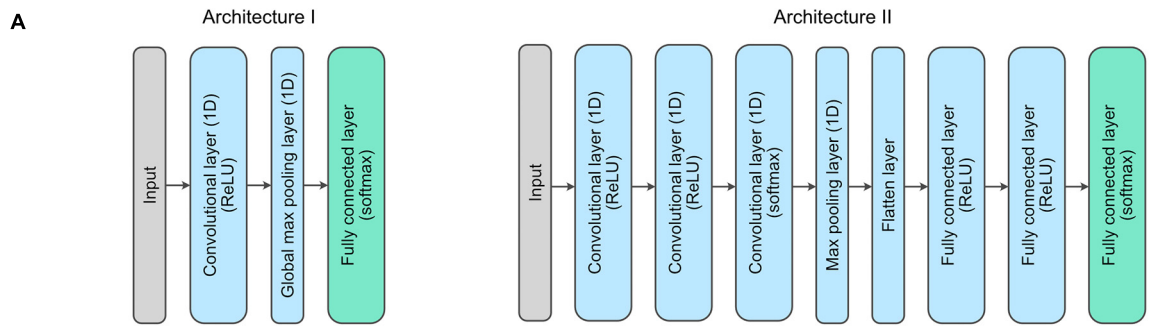
Increasing either the number of filters or kernel size resulted in an increase of classification accuracy with average F-measures between 0.710 ∓ 0.009 and 0.801 ∓ 0.006 for models with filters = 64/kernel_size = 3 and filters = 512/kernel_size = 7, respectively (evaluated at 0.7 confidence level, **Figure 2B** and **Supplementary Figure 2**). Based on these initial findings, we selected a subset of parameter configurations to test the effect of confidence settings on CNN classification performance. Reducing the confidence parameter to 0.5 improved performance (average F-measure of 0.849 ∓ 0.005, filters = 512/kernel_size = 7), but further improvements were not observed when we further reduced confidence to 0.25 (see alternate model specifications in **Supplementary Material**).

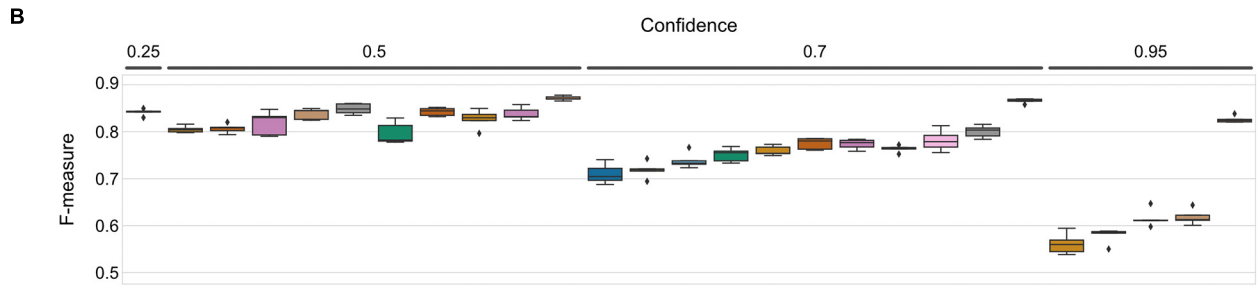
We also attempted to improve performance by then extending the test range of number of filters and kernel size to 4,096 and 14 at a confidence level of 0.5 and found that doubling the number of filters or kernel size had little to



no. of features	-	-	-	NB	-	-	-	sqrt	sqrt	NB	-	-	-	sqrt	sqrt	NB	-	-	-	sqrt	sqrt
no. of estimators	100	100	100	-	100	100	100	100	1000	-	100	100	100	100	1000	-	100	100	100	100	1000
max. tree depth	-	-	-	-	16	64	-	16	16	-	16	64	-	16	16	-	16	64	-	16	16

FIGURE 1 | F-measure accuracy performance of RF and NB classifiers. Box-and-whisker plots indicate the median and quartile distributions of F-measures for each classifier and configuration, across 5-fold of CV. RF classifier configurations were tested via a grid search across the hyperparameters listed in the subset table. NB classifiers do not have equivalent parameters, and hence only NB is listed in the table beneath bars representing NB classifiers. Both RF and NB classifiers were tested at multiple confidence levels. None of the tested parameter sets outperformed the NBC at any of the confidence levels (Wilcoxon signed-rank test $p < 0.05$).





architecture	II	I	I	I	I	I	I	I	I	I	NB	I	I	I	I	I	I	I	I	I	I	NB	I	I	I	I	NB
no. of filters	1024	128	128	256	256	512	1024	2048	2048	4096	-	64	64	64	128	128	128	256	256	512	512	-	128	128	256	256	-
kernel size	7	5	7	5	7	7	7	7	14	7	-	3	5	7	3	5	7	5	7	5	7	-	5	7	5	7	-

FIGURE 2 | F-measure accuracy performance of CNN and NB classifiers. **(A)** CNN architectures implemented in this benchmark. **(B)** Box-and-whisker plots indicate the median and quartile distributions of F-measures for each classifier and configuration, across 5-fold of CV. CNN classifier configurations were tested via a grid search across the hyperparameters listed in the subset table. NB classifiers do not have equivalent parameters, and hence only NB is listed in the table beneath bars representing NB classifiers. Both CNN and NB classifiers were tested at multiple confidence levels. None of the tested networks outperformed the NBC at a given confidence level (Wilcoxon test $p < 0.05$).

no effect on the classifier accuracy (average F-measure of 0.843 ∓ 0.004, filters = 2,048/kernel_size = 7, 0.827 ∓ 0.009, filters = 2,048/kernel_size = 14). Finally, we tested a variety of different network architectures and two feature-extraction methods other than Word2Vec (see **Supplementary Materials** for details). One of the better results is represented by

TABLE 2 | Parameter values used for grid search using the convolutional neural network.

Parameter	Values			
Filters	64	128	256	512
Kernel size	3	5	7	–
Confidence	0.5	0.7	0.95	–

Architecture II in **Figure 2B**, which also used one-hot-encoding of individual nucleotides to build a sequence of vectors for input to the neural network.

While exhaustively testing all of the possible neural network architectures is not practical, a pattern emerged in our testing. That is that with tuning, it was possible to approach an average F-measure of around 0.85, but none of the models that we tested outperformed the NBC, which on the same data set with reads of the same length had an average F-measure of 0.866 $\mp$ 0.002 (all differences between CNN and NB results were statistically significant at $p < 0.05$ when evaluated at the same confidence level).

Comparing accuracy reported between F-measure and MCC, again the differences were qualitatively the same and different configurations were ranked almost identically within confidence levels.

Moreover, also in the case of CNN classifiers it is recall that plays a major role in differentiating between different parameter configurations (**Supplementary Figure 2**). While classification precision remained on an approximately similar level for most of the configurations tested, the recall increased as model complexity increased (in terms of model parameters). Regardless of the parameters used, however, CNN recall was always lower than that of NBC at a given confidence level (**Supplementary Figure 2**).

Finally, we were interested in checking whether the networks described above were prone to overfitting. Large model capacity (expressed as number of model parameters) with respect to the amount of training data can lead to the network learning features of the training set that are not universally relevant, thus reducing the accuracy when evaluating the model on the test set. We compared training histories of the architecture I with 512 filters and kernel size of 7 and architecture II with 1,024 filters and kernel size of 7 (**Supplementary Figure 5**). For all of the results that we report we trained for either 5 or 10 epochs (**Supplementary Table 3**), and overfitting was not evident at that stage in either of these examples.

The Perfect Classifier

The underwhelming performance exhibited by RF and CNN classifiers led us to hypothesize that NBCs may already be approaching the upper limit of classification accuracy for this problem and hence alternative algorithms alone cannot exceed this performance. To test this hypothesis, we constructed a *perfect* classifier to measure the upper bound of classification accuracy for a given classification task. This classifier performs in-sample testing where the classifier can only fail if two or more species share exactly the same sequence. Where they do share the same sequence, one matching classification is chosen at random as the label for that sequence. The performance of such a classifier represents the upper limit of possible classification accuracy (Busia et al., 2019; Robeson et al., 2020).

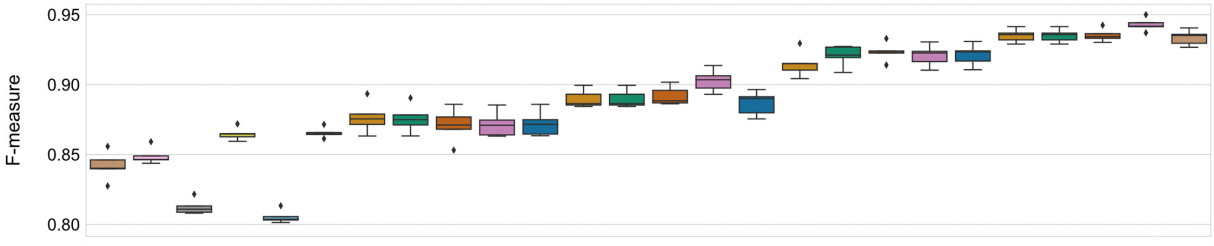
We trained perfect classifiers with and without taxonomic class weighting to assess the upper bound of accuracy when using sequence information alone (uniform weights) or when leveraging ecological information. We also tested a range of confidences and for 150 nt amplicons or sequences that captured all of V4. See section "Materials and Methods" for implementation details of how confidence affected the perfect classifiers. We used the same data set that we had used for testing CNNs for this purpose.

It is an implementation detail of the CNN classifiers that variable length amplicons are difficult to handle, so our above tests truncated sequences at 150 nt. So comparing the 150 nt perfect classifiers first, average F-measures for perfect classifiers that used class weight information varied between 0.887 and 0.903. The best NBC with the same constraints achieved an average F-measure of 0.865 $\mp$ 0.002 (significantly different from the perfect classifier; Wilcoxon rank test $p < 0.05$). Note that the NBCs were performing out-of-sample cross validation whereas the perfect classifier is in-sample and therefore naturally inflated. While there is a small gap between these two figures, it certainly strongly limits scope for improvement over NBCs.

The story is slightly different if the perfect classifier was allowed to use all of the V4. In that case, where class weight information is utilized, the perfect classifier scored average F-measure between 0.934 and 0.943 whereas the similarly constrained NBC achieved 0.866 $\mp$ 0.002 (differences statistically significant at $p < 0.05$; Wilcoxon rank test). In other words, it performed roughly identically to where V4 was constrained to its first 150 nt. This is consistent with other empirical results (Liu et al., 2008; Bokulich et al., 2018b).

Interestingly, MCC gave slightly different results to F-measure for the perfect classifier for classifiers with high confidence levels (0.75 and 0.95). At these confidence levels, MCC was penalized with respect to the lower confidence levels. At lower confidence levels all differences were statistically significant when comparing the top-performing classifiers to "perfect" classifiers evaluated with similar parameters (Wilcoxon rank sum test, $p < 0.05$). As the purpose of the perfect classifier is to provide an upper bound on classification accuracy, however, the significant result indicates that additional optimization can only yield diminishingly small performance improvements.

Finally, it is interesting to note the effect of incorporating class weight information on the NBC and perfect classifiers. For the NBC, the uniform classifiers (that did not use that information) performed almost the same for truncated and untruncated V4 sequences but were around 0.06 worse than when class weight information was used (average F-measure 0.812 $\mp$ 0.002 and 0.805 $\mp$ 0.002 respectively). For the perfect classifiers there was a clear progression where using all of V4 always improved accuracy and incorporating class weight information also increased accuracy. Class weight information did not improve classification accuracy for 150 nt sequences to



classifier	RF	CNN	NB	NB	NB	NB	PC	PC	PC	PC	PC	PC	PC	PC	PC	PC	PC	PC	PC	PC	PC	PC	PC	PC	PC	PC
weights	w	w	u	w	u	w	u	u	u	u	u	w	w	w	w	w	u	u	u	u	u	w	w	w	w	w
16S V4	150	150	150	150	FL	FL	150	150	150	150	150	150	150	150	150	150	FL	FL	FL	FL	FL	FL	FL	FL	FL	FL
confidence	0.3	0.5	0.7	0.7	0.7	0.7	0.05	0.25	0.5	0.75	0.95	0.05	0.25	0.5	0.75	0.95	0.05	0.25	0.5	0.75	0.95	0.05	0.25	0.5	0.75	0.95

FIGURE 3 | "Perfect" classifiers demonstrate the upper bound of classifier performance for V4 and full 16S rRNA gene sequences. These classifiers only fail when two species share an identical sequence, assuming uniform weights. Taxonomic weighting slightly increases classification accuracy both for V4 and full-length sequences. Box-and-whisker plots indicate the median and quartile distributions of F-measures for each classifier, across 5-fold of CV. The top-performing NB, RF, and CNN classifiers (trained and tested on V4 sequences) are compared to the "perfect" classifiers to demonstrate that the upper bound of performance is already being approached. All differences were statistically significant when comparing the top-performing classifiers to "perfect" classifiers evaluated with similar parameters (Wilcoxon rank sum test, $p < 0.05$). (Weights: w, weighted; u, uniform; 16S V4, 150–150 nt fragment of 16S rRNA V4 region; FL, all of the V4 region).

match that of uniform classification on full V4 for the perfect classifier (**Figure 3**).

DISCUSSION

The goal of this study was to evaluate the utility of newer supervised learning techniques for taxonomic classification of 16S rRNA gene sequences, and in particular whether models based on convolutional neural networks could leverage ecological distribution information to match classification performance as shown previously for NBC (Kaehler et al., 2019). The implementations tested here managed to approach the classification accuracy of NBC, but even optimized CNNs could not match or exceed the performance of NBC, corroborating the recent findings of others (Zhao et al., 2020). Importantly, the goal of this study was not to test the exact implementations of RF or CNN classifiers developed by others (which have shown promising results but to our knowledge were not designed to leverage taxonomic weight information), but rather to evaluate the potential promise of advances in extracting taxonomic weight information (Kaehler et al., 2019) combined with spatial embedding of sequence information (Mikolov et al., 2013) for exceeding the taxonomic classification performance of NBC. Further independent benchmarks by others, and evaluation in more diverse test scenarios (e.g., non-16S rRNA gene targets) are warranted to further assess and optimize the performance of deep learning algorithms for taxonomic classification (Bokulich et al., 2020).

Following optimization of the hyperparameters evaluated in this study, RF was able to approach the accuracy performance delivered by NBC. Computational resources required to train this classifier, however, are substantially greater and could prohibit its practical application. Particularly large memory and computation time (<20 CPU hours vs. hundreds of CPU hours for NBC and RF, respectively) needed for training seemed to be a problem for some of the better-performing parameter sets due to a requirement to train many large trees (i.e., comprising many split nodes).

Given the best set of parameters and an optimized model architecture, CNN classifiers could approach, but not match, NBC accuracy performance. Moreover, training CNNs required a significant amount of computational resources and specific hardware, particularly in the case of more complex networks with many parameters. Our testing was only made feasible by employing modern graphics processing units (GPUs). GPUs are widely used to train neural networks, and this capacity has been suggested as an attractive feature of CNNs for taxonomic classification versus conventional methods (Busia et al., 2019). Even though training networks presented in this study required only a couple of hours (∼3 h) on a single GPU (NVIDIA GeForce RTX 2080) compared to ∼20 CPU hours for a typical NBC, GPUs can be considerably more expensive and difficult to configure and maintain, and hence are out of reach or less attractive to many researchers.

Our "perfect" classifier tests underline the fact that evolutionary conservation in most genetic targets for microbiome profiling limits the degree of taxonomic resolution that is possible, particularly when sequencing short marker-gene reads. Hence, mature, existing methods for classification (NBC and some alignment-based classifiers) have already neared the upper limits of classification accuracy. The relationship between read length, primer selection, marker-gene target, sequence entropy, and taxonomic resolution has been well documented for 16S rRNA genes and other common targets, and even with long sequence reads (e.g., full-length 16S rRNA genes) species-level resolution can be challenging for many clades (Wang et al., 2007; Liu et al., 2008; Bokulich et al., 2018b; Johnson et al., 2019; Robeson et al., 2020). This is in part complicated by muddled microbial taxonomies (Oren and Garrity, 2014; Yarza et al., 2014) and misannotations

and other issues with reference databases used for taxonomic classification (Kozlov et al., 2016). Further improvements in taxonomy classification are unlikely to come from novel algorithms alone, and will require some combination of the following:

1. Use of spatial dependency in DNA sequences or other latent information. In spite of the current disappointing results, others have demonstrated the promise of spatially aware feature extraction prior to CNN classification for taxonomy or sample predictions (Busia et al., 2019; Zhao et al., 2020).
2. Use of ecological information from prior studies to hone classification accuracy (Kaehler et al., 2019).
3. Improvement of reference sequence and taxonomy databases (Parks et al., 2018; Robeson et al., 2020).
4. Longer read lengths and/or marker-gene targets (Johnson et al., 2019; Milani et al., 2020)
5. Improvements are not limited to accuracy, and could include more efficient classifiers with less runtime, memory, or other resource requirements (Bokulich et al., 2020).

We note that none of the methods compared in this work incorporated a feature selection step. It is possible that a feature selection step might increase performance of these methods (except for the perfect classifier), and warrants future investigation.

CONCLUSION

Naive Bayes classifiers have demonstrated robust performance for taxonomic classification of DNA sequences for more than a decade (Wang et al., 2007), and recent improvements have further increased their accuracy (Bokulich et al., 2018b; Kaehler et al., 2019). Newer supervised learning methods such as neural networks offer exciting features with potential to further improve pattern recognition in microbiome data but so far have only demonstrated small or no improvements for taxonomic classification specifically (Zhao et al., 2020). In the current study, we find further evidence that NBCs remain supreme for taxonomic classification, even when applying taxonomic weighting and spatial encoding of sequence information, as well as hyperparameter tuning to optimize RF and CNN classifiers for 16S rRNA gene classification. It is worth noting that both RF and CNN classifiers comfortably outperform NBCs when they use taxonomic weighting information but the NBC does not. We demonstrate that NBCs are already nearing the performance limit of taxonomic classification of short 16S rRNA gene reads, indicating that further improvements will require technological and biological improvements or by leveraging other information (e.g., ecological observations) beyond sequence information alone. CNNs and other methods remain promising, however, and further optimization and benchmarking is warranted to fully assess the opportunities of deep learning techniques for microbial classification.

MATERIALS AND METHODS

Random Forests

Cross validation of RF classifiers was performed using the methodology described in Kaehler et al. (2019). The RF classifier was tested using the standard q2-feature-classifier (Bokulich et al., 2018b) using a custom scikit-learn classifier specification to implement the scikit-learn random forest classifier (Pedregosa et al., 2011). Feature extraction was performed using the standard bag of overlapping 7-mers approach, also using scikit-learn. The code for the q2-feature classifier is available at https://github.com/qiime2/q2-feature-classifier. Cross validation code and classifier specifications are available at https://github.com/BenKaehler/paycheck.

Greengenes release 13_8 (McDonald et al., 2012) was used for the reference database and sample data was downloaded from Qiita (Gonzalez et al., 2018) using q2-clawback (Kaehler et al., 2019). 188 samples labeled as sediment (non-saline) were downloaded on 20 March 2019 and 22,454 samples labeled as animal distal gut were downloaded on 23 March 2019. These samples have been uploaded to Zenodo[1].

Convolutional Neural Networks

Cross validation of CNN classifiers was again performed using the methodology described in Kaehler et al. (2019). Neural networks were implemented using the Tensorflow library[2] via the Keras interface[3]. Feature extraction was performed using the Word2Vec algorithm (Rehurek and Sojka, 2010). A fork of the standard q2-feature-classifier was necessary to accommodate Keras models and is available at https://github.com/BenKaehler/q2-feature-classifier. Cross validation code and classifier specifications are available at https://github.com/BenKaehler/paycheck.

Greengenes release 13.8 (McDonald et al., 2012) was used for the reference database and sample data was downloaded from Zenodo[4]. That data was the Qiita animal distal gut data originally used in Kaehler et al. (2019).

In the embedding step, sequences were trimmed to 150 nt. Each sequence was converted into a "sentence" of overlapping 7-mers, which were then used as input to the Word2Vec algorithm as implemented in Gensim (Rehurek and Sojka, 2010) to transform each sequence into a sequence of 144 length-300 vectors. A window of 5 words was used for training and the Common Bag of Words (CBOW) algorithm (Rehurek and Sojka, 2010) was selected. Those images were then presented to the various neural network models as described in "Results" section.

Perfect Classifier

The perfect classifier used the same data set as the CNN experiments and tests were entirely in-sample. The perfect classifier tests were "cross validation" tests only in the sense that they used the same frozen, randomized test sets as the

[1]https://zenodo.org/record/4361424#.X90hI1MzaV4
[2]https://www.tensorflow.org/
[3]https://keras.io/
[4]https://zenodo.org/record/2548899#.X9rPG1MzaV4

CNN experiments to reduce random variation between the two. Code for the perfect classifier is available at https://github.com/BenKaehler/paycheck.

"Perfect" classification was made possible by in-sample testing because every sequence that was used for testing had already been seen by the classifier. Sample weight cross validation was also in-sample, in that the same aggregate weights were used for every test, although we found that performing weight-wise out-of-sample testing did not make a qualitative difference to the results.

For each sequence, a list of taxa that matched that exact sequence was compiled. Weights for each taxon in the list were calculated using the taxonomic weighting information or by equally weighting taxa for uniform weights. In both cases the weights were normalized for each sequence. If one of the taxa's weight was greater than the chosen confidence level, the taxon with the maximum weight was chosen. If two or more taxa had equal maximum weight (as most often happened in the uniform case), one was chosen at random. If the confidence level was not exceeded by any weights, weights were aggregated at the second lowest taxonomic level and the procedure was repeated until a potentially truncated taxon was assigned.

Statistical Analysis

To assess whether the classification performance (expressed as F-measure) differs significantly between various models (i.e., random forest and convolutional neural network variations) and the Naive Bayes classifier, we employed a two-tailed Wilcoxon rank sum test (when comparing CNN to NB results where sample sets differed) and a two-tailed Wilcoxon signed rank test (for all other comparisons). The analysis was performed at $\alpha = 0.05$ using all of the test samples available for a given model (combined across all the folds) followed by Hommel correction for multiple testing (Hommel, 1988).

Additionally, to account for the potential bias resulting from highly imbalanced classes we evaluated all the models using the Matthews correlation coefficient (MCC) (Chicco and Jurman, 2020). MCC metric was shown to be a more reliable metric as it assesses the entire confusion matrix (i.e., true positive and negative, false positives and negatives), proportionally to class size.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

BK and NB conceived and designed experiments. TW and BK performed data retrieval and initial experiments. MZ, TW, BK, and NB performed the analysis and interpretation and wrote the manuscript. All authors reviewed and approved the final manuscript.

FUNDING

This study was not supported by external funding sources. Open-access publishing costs were covered by ETH Zürich library.

ACKNOWLEDGMENTS

We thank Anja Adamov (ETH Zürich) for insightful discussions on training and evaluating results of the neural networks.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2021.644487/full#supplementary-material

REFERENCES

Almeida, A., Mitchell, A. L., Tarkowska, A., and Finn, R. D. (2018). Benchmarking taxonomic assignments based on 16S rRNA gene profiling of the microbiota from commonly sampled environments. *Gigascience* 7:giy054. doi: 10.1093/gigascience/giy054

Bokulich, N. A., Collins, T. S., Masarweh, C., Allen, G., Heymann, H., Ebeler, S. E., et al. (2016). Associations among wine grape microbiome, metabolome, and fermentation behavior suggest microbial contribution to regional wine characteristics. *MBio* 7:e00631-16. doi: 10.1128/mBio.00631-16

Bokulich, N. A., Dillon, M. R., Bolyen, E., Kaehler, B. D., Huttley, G. A., and Caporaso, J. G. (2018a). q2-sample-classifier: machine-learning tools for microbiome classification and regression. *J. Open Res. Softw.* 3:934. doi: 10.21105/joss.00934

Bokulich, N. A., Kaehler, B. D., Rideout, J. R., Dillon, M., Bolyen, E., Knight, R., et al. (2018b). Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome* 6:90.

Bokulich, N. A., Ziemski, M., Robeson, M., and Kaehler, B. (2020). Measuring the microbiome: best practices for developing and benchmarking microbiomics methods. *Comput. Struct. Biotechnol. J.* 18, 4048–4062. doi: 10.1016/j.csbj.2020.11.049

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., et al. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* 37, 852–857.

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32.

Busia, A., Dahl, G. E., Fannjiang, C., Alexander, D. H., Dorfman, E., Poplin, R., et al. (2019). A deep learning approach to pattern recognition for short DNA sequences. *bioRxiv* [Preprint] doi: 10.1101/353474

Chaudhary, N., Sharma, A. K., Agarwal, P., Gupta, A., and Sharma, V. K. (2015). 16S classifier: a tool for fast and accurate taxonomic classification of 16S rRNA hypervariable regions in metagenomic datasets. *PLoS One* 10:e0116106. doi: 10.1371/journal.pone.0116106

Chicco, D., and Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* 21:6. doi: 10.1186/s12864-019-6413-7

Fiannaca, A., La Paglia, L., La Rosa, M., Lo Bosco, G., Renda, G., Rizzo, R., et al. (2018). Deep learning models for bacteria taxonomic classification of metagenomic data. *BMC Bioinformatics* 19:198. doi: 10.1186/s12859-018-2182-6

Gardner, P. P., Watson, R. J., Morgan, X. C., Draper, J. L., Finn, R. D., Morales, S. E., et al. (2019). Identifying accurate metagenome and amplicon software via

a meta-analysis of sequence to taxonomy benchmarking studies. *PeerJ* 7:e6160. doi: 10.7717/peerj.6160

Gonzalez, A., Navas-Molina, J. A., Kosciolek, T., McDonald, D., Vázquez-Baeza, Y., Ackermann, G., et al. (2018). Qiita: rapid, web-enabled microbiome meta-analysis. *Nat. Methods* 15, 796–798. doi: 10.1038/s41592-018-0141-9

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika* 75, 383–386. doi: 10.1093/biomet/75.2.383

Johnson, J. S., Spakowicz, D. J., Hong, B.-Y., Petersen, L. M., Demkowicz, P., Chen, L., et al. (2019). Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nat. Commun.* 10:5029.

Kaehler, B. D., Bokulich, N. A., McDonald, D., Knight, R., Caporaso, J. G., and Huttley, G. A. (2019). Species abundance information improves sequence taxonomy classification accuracy. *Nat. Commun.* 10:4643.

Kozlov, A. M., Zhang, J., Yilmaz, P., Glöckner, F. O., and Stamatakis, A. (2016). Phylogeny-aware identification and correction of taxonomically mislabeled sequences. *Nucleic Acids Res.* 44, 5022–5033. doi: 10.1093/nar/gkw396

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 2278–2324. doi: 10.1109/5.726791

Liu, Z., DeSantis, T. Z., Andersen, G. L., and Knight, R. (2008). Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *Nucleic Acids Res.* 36:e120. doi: 10.1093/nar/gkn491

Lu, J., and Salzberg, S. L. (2020). Ultrafast and accurate 16S rRNA microbial community analysis using Kraken 2. *Microbiome* 8:124. doi: 10.1186/s40168-020-00900-2

McDonald, D., Price, M. N., Goodrich, J., Nawrocki, E. P., DeSantis, T. Z., Probst, A., et al. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J.* 6, 610–618. doi: 10.1038/ismej.2011.139

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*. Available online at: http://arxiv.org/abs/1301.3781 (Accessed December 9, 2020)

Milani, C., Alessandri, G., Mangifesta, M., Mancabelli, L., Lugli, G. A., Fontana, F., et al. (2020). Untangling species-level composition of complex bacterial communities through a novel metagenomic approach. *mSystems* 5, e404–e420. doi: 10.1128/mSystems.00404-20

Oren, A., and Garrity, G. M. (2014). Then and now: a systematic review of the systematics of prokaryotes in the last 80 years. *Antonie Van Leeuwenhoek* 106, 43–56. doi: 10.1007/s10482-013-0084-1

Parks, D. H., Chuvochina, M., Waite, D. W., Rinke, C., Skarshewski, A., Chaumeil, P.-A., et al. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* 36, 996–1004. doi: 10.1038/nbt.4229

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.

Rehurek, R., and Sojka, P. (2010). "Software framework for topic modelling with large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, (Valletta: ELRA), 45–50.

Robeson, M. S., O'Rourke, D. R., Kaehler, B. D., Ziemski, M., Dillon, M. R., Foster, J. T., et al. (2020). RESCRIPt: reproducible sequence taxonomy reference database management for the masses. Cold Spring Harbor Laboratory. *bioRxiv* [Preprint] doi: 10.1101/2020.10.05.326504

Roguet, A., Eren, A. M., Newton, R. J., and McLellan, S. L. (2018). Fecal source identification using random forest. *Microbiome* 6:185.

The Human Microbiome Project Consortium (2012). A framework for human microbiome research. *Nature* 486, 215–221. doi: 10.1038/nature11209

Thompson, L. R., Sanders, J. G., McDonald, D., Amir, A., Ladau, J., Locey, K. J., et al. (2017). A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* 551, 457–463.

Wang, Q., Garrity, G. M., Tiedje, J. M., and Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* 73, 5261–5267. doi: 10.1128/aem.00062-07

Yarza, P., Yilmaz, P., Pruesse, E., Glöckner, F. O., Ludwig, W., Schleifer, K.-H., et al. (2014). Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat. Rev. Microbiol.* 12, 635–645. doi: 10.1038/nrmicro3330

Zhao, Z., Woloszynek, S., Agbavor, F., Mell, J. C., Sokhansanj, B. A., and Rosen, G. (2020). Learning, visualizing and exploring 16S rRNA structure using an attention-based deep neural network. *bioRxiv* [Preprint] doi: 10.1101/2020.10.12.336271

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 25 August 2021
doi: 10.3389/fmicb.2021.628426

Discovering Potential Taxonomic Biomarkers of Type 2 Diabetes From Human Gut Microbiota *via* Different Feature Selection Methods

Burcu Bakir-Gungor[1], Osman Bulut[1], Amhar Jabeer[1], O. Ufuk Nalbantoglu[2] and Malik Yousef[3,4]*

[1] Department of Computer Engineering, Faculty of Engineering, Abdullah Gül University, Kayseri, Turkey, [2] Department of Computer Engineering, Genome and Stem Cell Center, Erciyes University, Kayseri, Turkey, [3] Department of Information Systems, Zefat Academic College, Zefat, Israel, [4] Galilee Digital Health Research Center, Zefat Academic College, Zefat, Israel

OPEN ACCESS

Edited by:
David Gomez-Cabrero, NavarraBiomed, Spain

Reviewed by:
Ren-You Gan, Institute of Urban Agriculture, Chinese Academy of Agricultural Sciences, China
Vincenzo Lagani, Ilia State University, Georgia

***Correspondence:**
Burcu Bakir-Gungor
burcu.gungor@agu.edu.tr

Specialty section:
This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology

Received: *11 November 2020*
Accepted: *03 May 2021*
Published: *25 August 2021*

Citation:
Bakir-Gungor B, Bulut O, Jabeer A, Nalbantoglu OU and Yousef M (2021) Discovering Potential Taxonomic Biomarkers of Type 2 Diabetes From Human Gut Microbiota via Different Feature Selection Methods. Front. Microbiol. 12:628426. doi: 10.3389/fmicb.2021.628426

Human gut microbiota is a complex community of organisms including trillions of bacteria. While these microorganisms are considered as essential regulators of our immune system, some of them can cause several diseases. In recent years, next-generation sequencing technologies accelerated the discovery of human gut microbiota. In this respect, the use of machine learning techniques became popular to analyze disease-associated metagenomics datasets. Type 2 diabetes (T2D) is a chronic disease and affects millions of people around the world. Since the early diagnosis in T2D is important for effective treatment, there is an utmost need to develop a classification technique that can accelerate T2D diagnosis. In this study, using T2D-associated metagenomics data, we aim to develop a classification model to facilitate T2D diagnosis and to discover T2D-associated biomarkers. The sequencing data of T2D patients and healthy individuals were taken from a metagenome-wide association study and categorized into disease states. The sequencing reads were assigned to taxa, and the identified species are used to train and test our model. To deal with the high dimensionality of features, we applied robust feature selection algorithms such as Conditional Mutual Information Maximization, Maximum Relevance and Minimum Redundancy, Correlation Based Feature Selection, and select K best approach. To test the performance of the classification based on the features that are selected by different methods, we used random forest classifier with 100-fold Monte Carlo cross-validation. In our experiments, we observed that 15 commonly selected features have a considerable effect in terms of minimizing the microbiota used for the diagnosis of T2D and thus reducing the time and cost. When we perform biological validation of these identified species, we found that some of them are known as related to T2D development mechanisms and we identified additional species as potential biomarkers. Additionally, we attempted to find the subgroups of T2D patients using *k*-means

clustering. In summary, this study utilizes several supervised and unsupervised machine learning algorithms to increase the diagnostic accuracy of T2D, investigates potential biomarkers of T2D, and finds out which subset of microbiota is more informative than other taxa by applying state-of-the art feature selection methods.

Keywords: feature selection, metagenomic analysis, classification, machine learning, type 2 diabetes, human gut microbiome

INTRODUCTION

Trillions of living creatures live in our bodies, especially in our gut. These organisms are important to regulate our immune system. They provide energy, break down foreign matters, produce some hormones, etc., which are extremely important for our health. The gut microbiome including different types and amounts of microorganisms is crucial for human health and human disorders (Valdes et al., 2018). With the help of new technologies and methods, we can get gut microbiome data. In other words, we can measure their amount in our gut more easily than ever before. Hence, we can try to go after some correlation signs between these creatures and human diseases. Type 2 diabetes (T2D) is one of such diseases, which affects millions of people around the world. Approximately 9–11% of people in the United States and China have T2D. Four hundred sixty-three million people in the world, who are older than 20, have diabetes. One of three people in the United States, who are older than 20, has prediabetes. Seventy percent of these prediabetic individuals will also have diabetes (James et al., 2003; National Diabetes Clearinghouse, 2011; Tabak et al., 2012; Diabetes.co.uk, 2019; International Diabetes Federation, 2019; Centers for Disease Control and Prevention, 2020).

Several studies have been conducted on human microbiota and its relations with type 1 diabetes, T2D, or obesity (Turnbaugh et al., 2009; Vrieze et al., 2012; Trøseid et al., 2013; Boulangé et al., 2016; Chobot et al., 2018; Peters et al., 2018). Brunetti (2007) defined T2D as a worldwide epidemic in 2010 and claimed that obesity was one of the most important driving forces for the development of T2D. This is varied by ethnicity though. For North America, the relationship between T2D and obesity is 90%. Whereas it is smaller than 40% in South Asia (International Diabetes Federation, 2003; James et al., 2003). The microbiota studies for obesity is also important for T2D studies. Not all obese individuals have also T2D, but 86% of T2D individuals are obese or overweight (Daousi et al., 2006; Narayan et al., 2007). The diet is one of the important factors that affect the gut microbiota (Falony et al., 2016; Zhernakova et al., 2016). found that while the dietary changes have a 57% role for the gut microbiota variations, the genetic mutations only have 12% role. Despite that there are some contrary arguments, it is reported in Zhang et al. (2010) that we can slow down the increase of obesity, and so the T2D, by regulating the variations of our gut microbiota by doing dietary changes. After the meal, even the glycemic action type of a body can be affected by its gut microbiota composition (Zeevi et al., 2015; Mendes-Soares et al., 2019). Some studies show that biotin deficiency may be associated with T2D (Maebashi et al., 1993; Wu et al., 2020) and biotin supplementation may help glucose regulation (Fernandez-Mejia, 2005; Albarracin et al., 2008; Lazo de la Vega-Monroy et al., 2013).

Conducting different studies to discover the associations and the relationships between variations of the gut microbiota and T2D is essential. For example, Karlsson et al. (2013) emphasize the importance of gender, age, and family history in these kinds of studies. Therefore, in order to minimize the source of variation, they worked on such data that consist of 145 women who are 70 years old. Interestingly, they found that some *Lactobacillus* species are increased and some *Clostridium* species are decreased in the microbiomes of the T2D patients. They got 0.83 AUC with a metagenomics cluster level. Increased *Clostridium clostridioforme* and decreased *Roseburia* in T2D patients are common findings of Karlsson et al. (2013) and Qin et al. (2012). Larsen et al. (2010) and Lê et al. (2013) also found that *Lactobacillus* species are increased in T2D patients.

Forslund et al. (2015) presented a different perspective such that the possible effects of the T2D drugs on the human gut microbiome also need to be taken into account. They also addressed the need to disentangle microbiota signs of the disease from the medications that patients use. Forslund et al. (2015), Wu et al. (2017), and Sun et al. (2018) show the effects of the most commonly used anti-T2D drug metformin. But they also found that metformin-untreated T2D is still associated with the butyrate producer species deficiency. The importance of butyrate-producing species for glucose health is also emphasized by Karlsson et al. (2013), Qin et al. (2012), Allin et al. (2018), and Sanna et al. (2019). Wu et al. (2020) also showed that butyrate producers' deficiency and the loss of genes for butyrate synthesis from both proteins and carbohydrates start to occur even from the prediabetic level. Diet is also important at this point, as mentioned before. The function of butyrate producers is also regulated by diet, especially fiber intake, which positively affects glucose control (Makki et al., 2018; Zhou et al., 2019).

Wu et al. (2020) also considered the potential effects of drugs on gut microbiota, and they studied the diabetes treatment-naive T2D cohort. Their findings were also in agreement with earlier studies (Qin et al., 2012; Karlsson et al., 2013; Forslund et al., 2015; Allin et al., 2018). They showed that their finding was independent of metformin, other confounding factors affecting gut microbiota, and also other confounders like age, BMI, and sex. Their microbiome-based machine learning model to detect T2D samples and healthy samples generated a 0.78 AUC score.

Zhong et al. (2019) worked on 254 samples of Chinese cohort. They found that *Dialister nvisus* (MLG-3376) and *Roseburia hominis* (MLG-14865 and 14920) are lower in the T2D patients who were also reported before by Forslund et al. (2015). They also found that *Streptococcus salivarius* (MLG-6991) is high in the

pre-sick people, which is in agreement with the previous findings of Allin et al. (2018) in the Danish prediabetic cohort. Zhong et al. showed that *Megasphaera elsdenii* (MLG-1568) was found in higher amounts in T2D patients compared to the pre-DM and healthy individuals. A similar finding was previously presented by He et al. (2018) by conducting a study on 7,000 individuals from South China.

On the other hand, Thingholm et al. (2019) claim that we need to differentiate the gut microbiota of obese individuals with T2D and obese individuals without T2D. This is proposed because they show different functional capacities and composition. Obesity is more associated with alterations in microbiome composition than T2D. They also concluded that only nominal increases in *Escherichia/Shigella* happen in the microbiomes of T2D patients. Also, medications and dietary supplements are highly related to gut microbiome variations (Thingholm et al., 2019).

Another important point to consider is the daily changes of the microbiota. There are some studies about gut microbiota's diurnal oscillations in composition (Thaiss et al., 2014; Liang et al., 2015; Kuang et al., 2019). More specifically to diabetes, Reitmeier et al. (2020) found that T2D patients exhibit disrupted circadian rhythms in their microbiome. They show that arrhythmic bacterial signatures have an additional value for the classification of T2D, and they found that 13 arrhythmic bacterial species contribute to risk profiling of T2D. On the other hand, they found that daily dietary habits (like mealtime or number of meals per day) are independent of gut microbiota composition (Reitmeier et al., 2020).

A recent survey paper by Marcos-Zambrano et al. (2021) summarized the applications of machine learning in the human microbiome studies and reviewed popular feature selection, biomarker identification, disease prediction, and treatment strategies. In this review, the most widely used machine learning algorithms that were used for microbiome analysis were reported as Random Forest, support vector machines (SVM), Logistic Regression, and k-NN. However, no clear recommendation is given and they have suggested to perform comparison study to choose the one with the optimal performance. All of those algorithms require a parameter tuning step to achieve its optimal model.

In this study, we analyzed T2D-associated metagenomic dataset *via* some feature selection algorithms such as Fleuret's Conditional Mutual Information Maximization (CMIM), Peng's Maximum Relevance and Minimum Redundancy (mRMR), Fast Correlation Based Filter (FCBF), and select K best (SKB). To assess the performance of different classifiers, in our preliminary analysis, we used Random Forest (RF), Decision Tree, Logitboost, Adaboost, SVM, and K-NN as classification methods. In our further experiments, we focused on RF classifier. In summary, this study utilizes both supervised and unsupervised machine learning algorithms (i) to generate a classification model that aids T2D diagnosis, (ii) to investigate potential pathobionts of T2D, and (iii) to find out subgroups of T2D patients.

The rest of this paper is organized as follows. In section "Materials and Methods", we present the dataset that we have used in this study and we describe our methodology. In section "Experiments", we present our findings when we apply feature selection algorithms, classification methods, and clustering algorithms to T2D-associated metagenomic data. In section "Discussions", we discuss the identified species in our study as candidate taxonomic biomarkers of T2D and compare them with the gold standard features that are known to be associated with T2D in literature. In section "Conclusion", we conclude the manuscript.

MATERIALS AND METHODS

In this study, we used the raw microbiome DNA sequencing data of 290 human samples. The raw sequencing data of samples were obtained from the repository provided by Qin et al. (2012), deposited in the NCBI Sequence Read Archive under accession numbers SRA045646 and SRA050230, and categorized into disease states based on the associated metadata. The raw sequences were subject to quality filtering steps, which were described in the SOP of the Human Microbiome Project Consortium (2012). After preprocessing, using MetaPhlAn2 taxonomic classification tool, metagenome samples were assigned to its microbial species of origin (taxa) and the relative abundance composition of each taxon of a sample was inferred accordingly. These taxa and their relative abundances formed the features to be employed in the machine learning algorithms. As illustrated in **Table 1**, the data consist of 290 samples and 1,455 microbial species. One hundred thirty-five of the samples are T2D patients, and 155 are healthy. **Table 1** presents some lines of the metagenomics dataset for T2D, following the initial preprocessing of the original data. The relative abundance values of each species for each sample are shown in this dataset. The features correspond to different species including bacteria, viruses, and archea. The samples have one of the two class labels, i.e., healthy (shown with 0) and T2D patient (shown with 1).

Figure 1 shows the workflow of our methodology. As shown in **Figure 1**, the following flowchart is applied: (i) the application of feature selection to detect the most important species for the development of T2D (T2D-associated microorganisms), (ii) model construction and classification, and (iii) application of clustering algorithms to specify subgroups of patients and control samples.

Feature Selection

The dimension of the data is 1,455 (1,455 microbial species) that might influence the performance of the classification algorithms. Thus, a feature selection process is necessary to reduce the dimension of the model and make it also easier for classification and for interpretation. In order to select informative features, in other words to reduce the number of taxa (species), min Redundancy Max Relevance (mRMR) (Brown et al., 2012), Lasso (Tibshirani, 1996), Elastic Net (Zou and Hastie, 2005), and iterative sure select algorithm (Duvallet et al., 2017) have been applied in literature.

We suggest that using some feature selection algorithms such as Peng's mRMR (Brown et al., 2012), Fleuret's CMIM

TABLE 1 | The metagenomics dataset of T2D, after the initial preprocessing of the original metagenomics data.

	Methanobrevibacter smithii	*Methanosphaera*	*Acidobacteriaceae*	...	*Megasphaera* sp. BV3C16	Class label (healthy/T2D patient)
Sample 1	0.334	0	0		0	0 (Healthy)
Sample 2	0.141	0	0	0.632	0.03	1 (T2D patient)
...						
Sample 290						

The relative abundance values of each species for each sample are shown in this dataset. The features correspond to different species including bacteria, viruses, and archea. The samples have one of the two class labels, i.e., healthy (shown with 0) and T2D patient (shown with 1).

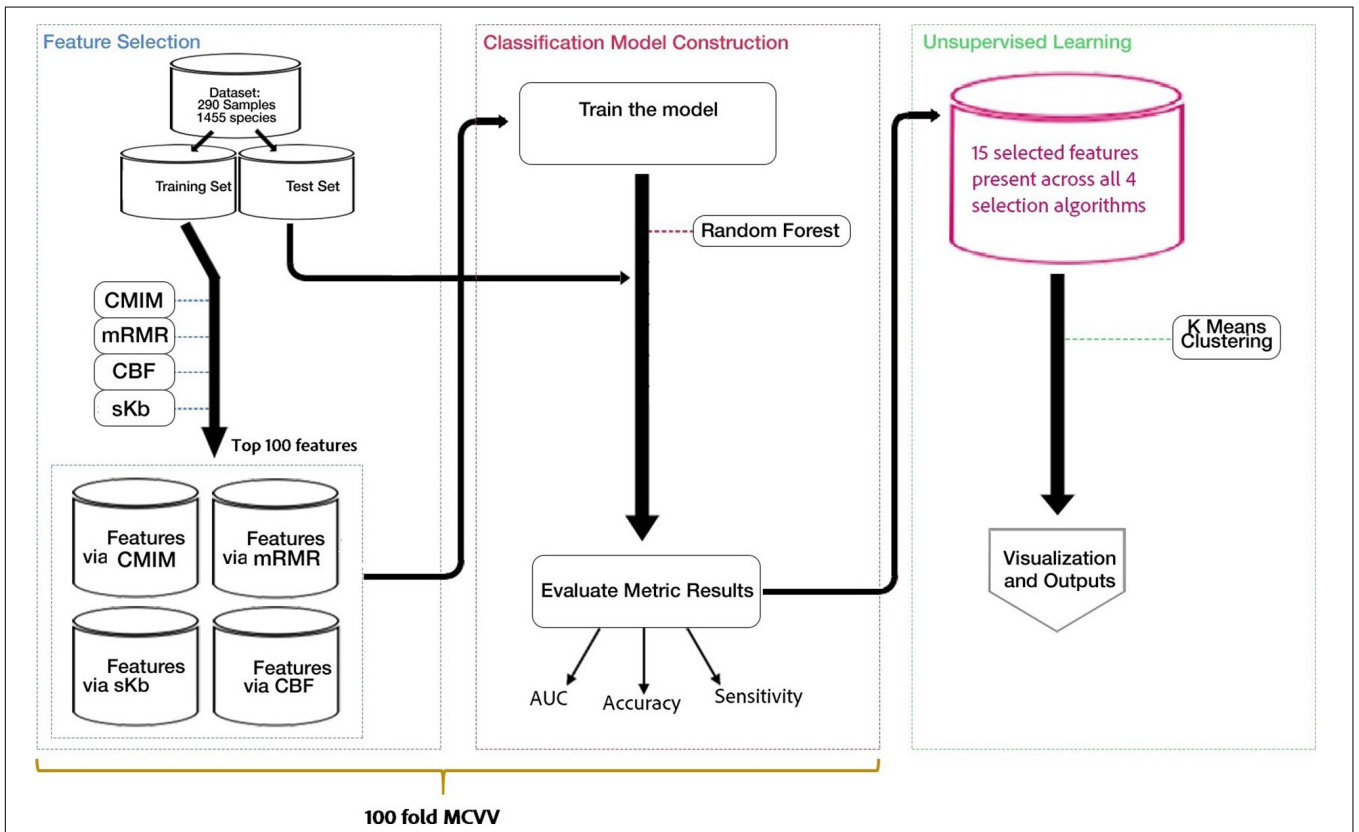


FIGURE 1 | Flowchart of our method, including three main parts. (i) Feature selection methods are applied to detect the most important species for the development of T2D (T2D-associated microorganisms). (ii) Using the selected features, models are constructed and used for classification. (iii) *K*-means clustering algorithm is applied on data to specify subgroups of patients and control samples.

(Fleuret, 2004), FCBF (Senliol et al., 2008), and SKB (Pedregosa et al., 2011) could improve classification performance, and by reducing the number of features, we can detect candidate taxonomic biomarkers.

Basically, the mRMR (Brown et al., 2012) method aims to select the features that have the least correlation between themselves (min redundancy) and the highest correlation with a class to predict (max relevance). In order to find the best subset of features, this method starts with an empty set and uses mutual information to weight features and forward selection technique with sequential search strategy. It is a multivariate feature selection method, which calculates the dependency between each feature pair, in addition to class relevance.

Conditional Mutual Information Maximization (Fleuret, 2004) determines the importance of features based on their conditional entropy and mutual information with the class. If the feature carries additional information, it selects that feature. Similarly, FCBF (Senliol et al., 2008) ranks features based on their mutual information with the class to predict, and then removes the features whose mutual information is less than a predefined threshold. It uses the idea of "predominant correlation". It selects features in a classifier-independent manner, selecting features with high correlation with the target variable, but little correlation with other variables. Notably, the correlation used here is not the classical Pearson or Spearman correlations, but Symmetrical Uncertainty (SU). SU is based on information theory, drawing from the concepts of

Shannon entropy and information gain. In other words, FCBF aims at reducing redundancy among selected features. FCBF provides an interpretable and robust option, with results that are generally good. The application of filter-based feature selections for big data analysis in the biomedical sciences not only can have a direct effect in classification efficiency but also might lead to interesting biological interpretations and possible quick identification of biomarkers.

Select K best scores the features against the class label using a function and selecting features according to the *k* highest score (Pedregosa et al., 2011). CMIM, mRMR, FCBF, and SKB feature selection methods are applied using the skfeature and sklearn libraries in Python 3[1].

Hacilar et al. (2019) applied some of these feature selection methods on inflammatory bowel disease-associated metagenomics dataset and reported to obtain good performance metrics. Most of those feature selection approaches are well studied and well known to achieve good results in human microbiome studies, as reported in a recent review (Marcos-Zambrano et al., 2021).

Classification Model Construction

In order to evaluate the effects of different classification methods, in our preliminary analysis, we have used Decision Tree, RF, LogitBoost, AdaBoost, an ensemble of SVM with kNN (*k* nearest neighbor), and an ensemble of the Logitboost with kNN. Since the tree model is easy for interpretation and since one can easily convert the model into rule set, in our further experiments, we continued with RF. Additionally, RF is one of the most used algorithms in the human microbiome studies as reported by Marcos-Zambrano et al. (2021).

We designed our actual experiments as follows. We used 100-fold Monte Carlo cross-validation (MCCV), which is the process of randomly selecting (without replacement) some fraction of the data to generate the training set and then assigning the rest to the test set (Xu and Liang, 2001). This process is repeated multiple times, and new training and test partitions are randomly generated each time. We have chosen 90% for training and 10% for testing. As shown in **Figure 1**, the feature selection methods are applied on the training set.

The Konstanz Information Miner (KNIME) platform (Berthold et al., 2008) is used for the implementation of our methodology. We used the RF predictor node from H20 library in KNIME.

Model Performance Evaluation

In order to evaluate model efficiency, we measured a range of statistical measures such as sensitivity, specificity, accuracy, and F1 measure for each created model. In this respect, we used the following formulations:

$$Sensitivity\,(Recall) = True\,Positive \Big/ \left(True\,Positive + False\,Negative\right) \quad (1)$$

$$Precision = True\,Positive \Big/ \left(True\,Positive + False\,Positive\right) \quad (2)$$

$$Specificity = True\,Negative \Big/ \left(True\,Negative + False\,Positive\right) \quad (3)$$

$$F1 - measure = \left(2^{*}Precision^{*}Recall\right) \Big/ \left(Precision + Recall\right) \quad (4)$$

$$Accuracy = (True\,Positive + True\,Negative)/(True\,Positive + True\,Negative + False\,Positive + False\,Negative). \quad (5)$$

Additionally, the area under the receiver operating characteristic (ROC) curve (AUC) is used to approximate the probability of the classifier that would score a randomly selected positive instance higher than a randomly selected negative instance.

The average of 100-fold MCCV (Xu and Liang, 2001) results is reported for all performance measures.

Unsupervised Learning

In order to find subgroups of patients and subgroups of healthy people, we have applied the *k*-means algorithm. *k*-means (Steinley and Brusco, 2007) is an unsupervised clustering algorithm that groups the data into clusters based on similarity or distance metric. *k*-means algorithm minimizes the error inside groups and maximizes the distance between the clusters. We have considered the Euclidean distance metric in our analysis. We used the Elbow method[2] to determine the optimum number of clusters. In this method, the slow down point denotes the optimum number of clusters.

EXPERIMENTS

Feature Selection and Classification

We have 1,455 features in our data, and we investigated for irrelevant and uninformative features. For this purpose, we applied four most well-studied feature selection algorithms, which are CMIM, mRMR, FCBF, and SKB. In our preliminary analysis, in order to evaluate the effects of different classification methods, Decision Tree, RF, LogitBoost, AdaBoost, an ensemble of SVM with kNN (*k* nearest neighbor), and an ensemble of the Logitboost with kNN are applied. As shown in **Supplementary Table 1** and **Supplementary Figure 1**, RF classifier generated the best performance results and we decided to continue with this classifier in our further experiments.

At the end of our experiments with 100-fold MCCV and RF classifier (as shown in **Figure 1**), we have listed the top 100 and top 500 identified features for each feature selection method in **Supplementary Tables 2**, **3**, respectively. The commonalities between those top 100 and top 500 identified feature sets are investigated, and the commonly detected 15 and 199 features within top 100 and top 500 identified features are shown in **Supplementary Tables 2**, **3**, respectively. The commonalities between top 100 identified feature sets, and the details of the 15 features, which are selected by all of the feature selection methods, are shown in

[1]https://www.python.org/about/

[2]https://predictivehacks.com/k-means-elbow-method-code-for-python/

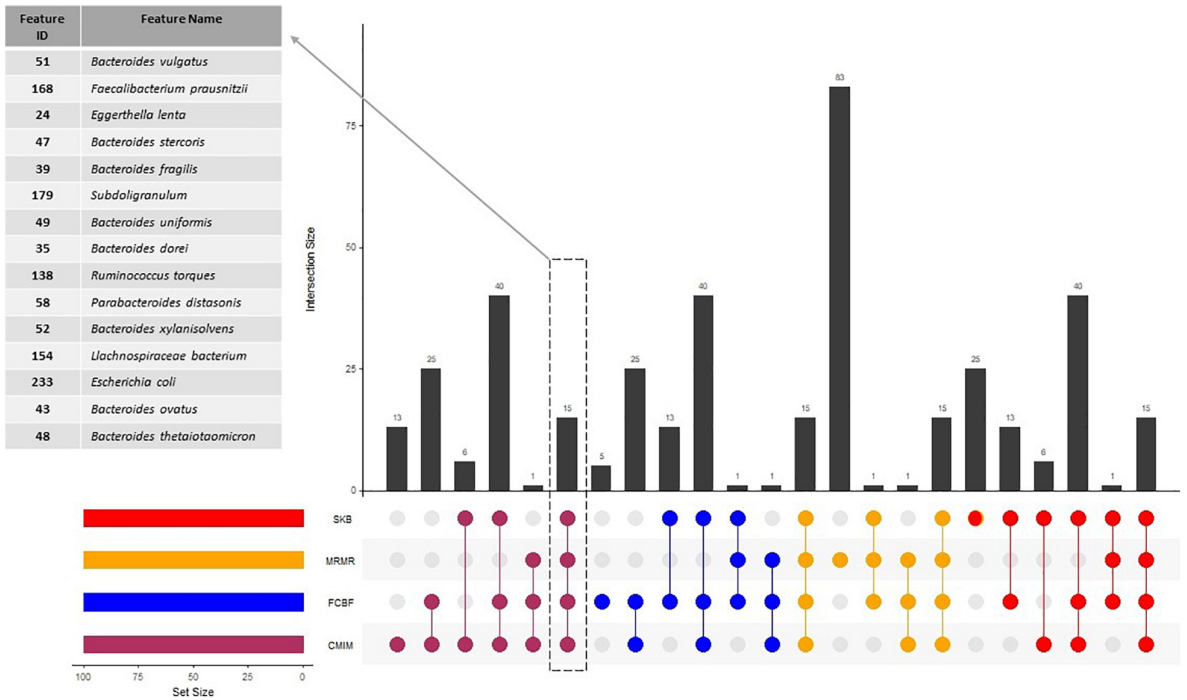


FIGURE 2 | Numbers of features, which are selected by different feature selection algorithms. The commonalities between the selected features by different methods are also illustrated.

Figure 2. In addition to the commonalities in species level, we investigated the commonalities in genus level. Nineteen genera are selected by all of the feature selection methods, as shown in **Supplementary Figure 2**.

By using several metrics as described in section "Model Performance Evaluation", we have compared the performances of (i) all features (without feature selection); (ii) top 100 and top 500 features selected using CMIM, mRMR, FCBF, and SKB; (iii) 15 and 199 features that are common among top 100 and top 500 features of all four tested feature selection methods; (iv) 329 identified features of 19 commonly detected genera in all four tested feature selection methods (**Supplementary Table 4**); and (v) 162 features of the gold standard genera that are reported to be associated with T2D in Gurung et al. (2020), as shown in **Supplementary Table 5**. A detailed comparative evaluation of our findings is presented in **Table 2** and **Figure 3**. As shown in **Figure 3**, the generated RF model resulted in 0.79 F1-score, 0.74 AUC, and 73% accuracy when all 1,455 features are used (without applying feature selection methods). On the other hand, when 199 features that are commonly selected in the top 500 features of all feature selection methods are used, the generated RF model resulted in 0.79 F1-score, 0.75 AUC, and 73% accuracy. Those selected 199 features performed as good as all features, even 1% higher in terms of AUC metric. Those selected 199 features also performed better compared to the performance (0.78 F1-score, 0.71 AUC, and 71% accuracy) of the 162 features (species) that belong to the gold standard genera, which are reported to be associated with T2D in a recent review paper (Gurung et al., 2020). By only using the 15 features that are commonly selected in the top 100 features list of all four tested feature selection methods, 0.75 F1-score, 0.62 AUC, and 64% accuracy metrics were obtained. In other words, T2D diagnosis could be possible with 64% accuracy by checking only the amounts of 15 specific species among 1,455 different species. As shown in **Figure 3**, the model using only those 15 species resulted in almost the same F1-score (0.75), with the F1-score obtained using all features (0.79). Checking the amounts of fewer features means less time and cost. In this respect, only using 15 features yielded comparable evaluation metrics.

Feature Correlations

The pairwise correlations of 15 features, which are commonly selected by all four tested feature selection methods, may be important for the further studies of T2D in terms of developing probiotics. For this reason, we have calculated the pairwise correlations of those 15 selected features using the tool in[3], and we have generated a heat map, as presented in **Figure 4**. It can be concluded from **Figure 4** that there are no important positive correlations between any two species among any two pairs of 15 selected species. This result indicates that each one of the selected 15 features has its own information and each feature (species) has an independent contribution to T2D development.

Clustering

We attempt to answer whether there could be any direct relationship between specific species and T2D subgroups. In order to answer this question, we used *k*-means clustering

[3]https://github.com/bhattbhavesh91/GA_Sessions/blob/master/ga_dsmp_5jan2019/16_feature_selection.ipynb

TABLE 2 | Comparative evaluation of the different feature selection methods, based on different performance metrics.

Methods		Accuracy	Recall	Specificity	Precision	AUC	F1	Number of features
CMIM	Score	0.71	0.90	0.48	0.72	0.72	0.78	100
	Std. dev.	0.10	0.11	0.34	0.15	0.11	0.05	
	Score	0.73	0.89	0.53	0.72	0.74	0.78	500
	Std. dev.	0.08	0.11	0.25	0.12	0.07	0.04	
FCBF	Score	0.68	0.91	0.41	0.68	0.70	0.76	100
	Std. dev.	0.08	0.10	0.27	0.10	0.09	0.04	
	Score	0.72	0.91	0.48	0.71	0.74	0.78	500
	Std. dev.	0.09	0.10	0.28	0.12	0.09	0.05	
MRMR	Score	0.63	0.95	0.23	0.62	0.59	0.74	100
	Std. dev.	0.06	0.12	0.27	0.10	0.12	0.02	
	Score	0.73	0.86	0.57	0.74	0.74	0.78	500
	Std. dev.	0.07	0.11	0.28	0.14	0.08	0.03	
SKB	Score	0.69	0.91	0.41	0.68	0.71	0.77	100
	Std. dev.	0.08	0.10	0.27	0.10	0.09	0.04	
	Score	0.71	0.92	0.46	0.69	0.74	0.78	500
	Std. dev.	0.08	0.08	0.25	0.10	0.09	0.04	
Commonly identified species (using top 100 features of each feature selection method)	Score	0.64	0.96	0.25	0.62	0.62	0.75	15
	Std. dev.	0.06	0.06	0.19	0.06	0.1	0.03	
Commonly identified species (using top 500 features of each feature selection method)	Score	0.73	0.89	0.54	0.73	0.75	0.79	199
	Std. dev.	0.08	0.09	0.25	0.11	0.09	0.05	
Identified species of commonly detected genus names	Score	0.71	0.91	0.46	0.70	0.73	0.78	329
	Std. dev.	0.09	0.09	0.28	0.11	0.09	0.05	
Species of gold standard genera of T2D	Score	0.71	0.91	0.46	0.70	0.71	0.78	162
	Std. dev.	0.09	0.11	0.28	0.11	0.10	0.05	
All features	Score	0.73	0.89	0.52	0.72	0.74	0.79	1,455
	Std. dev.	0.08	0.09	0.26	0.11	0.09	0.05	

algorithm and subgrouped the healthy samples and sick samples separately. As shown in **Supplementary Figure 3**, we decided to generate four subgroups for healthy samples and four subgroups for sick samples. **Figure 5** illustrates the identified healthy and T2D subgroups and the presence of the species in each of these subgroups. In **Figure 6**, we displayed more in detail the presence of four selected species in each of the healthy subgroups and one T2D subgroup, which covers 86% of the T2D patients from our dataset. It can be concluded from **Figures 5**, **6** that even though the samples were divided into subgroups, a single species may not have a direct effect on the development of T2D for a specific group. Nevertheless, there are a few observations that we can make: (i) *Bacteroides vulgatus* (shown in green in **Figures 5A**, **6C**) is mainly observed in healthy subgroups (healthy 0, 2, and 3) and found in reduced amounts in T2D patients. (ii) *Eggerthella lenta* is observed in reduced amounts in all healthy subgroups compared to the biggest subgroup of T2D patients (sick0), which includes 86% of the T2D patients from our dataset (shown in **Figure 6A**). (iii) *Bacteroides stercoris* (shown in red in **Figure 5A**) is present in reduced amounts in three of the healthy groups (healthy 0, 1, 2), compared to the biggest subgroup of T2D patients (sick0 in **Figure 6B**). (iv) Similarly, *Subdoligranulum* (shown in light green in **Figure 5B**) is present in reduced amounts in three of the healthy groups (healthy 0, 1, and 2), compared to the biggest subgroup of T2D patients (sick0 in **Figure 6D**).

DISCUSSION

The human gut microbiome contains trillions of living species. T2D is a disease that affects approximately 500 million people in the world. Like many other diseases, T2D might have a special association with gut microbiota (Manor et al., 2020). In the last decade, the identification of gut microbiota related to T2D has served as a stimulus for exponential advances in scientific production (Gurung et al., 2020). Multiple factors are reported to be involved in the changes of gut microbiota and hence its relationship to T2D (Sharma and Tripathi, 2019). The contribution of various molecular mechanisms of gut microbiota to T2D has been recently reviewed in Aw and Fukuda (2018). In order to change the gut microbiota to our benefit, several possibilities are currently available, and these

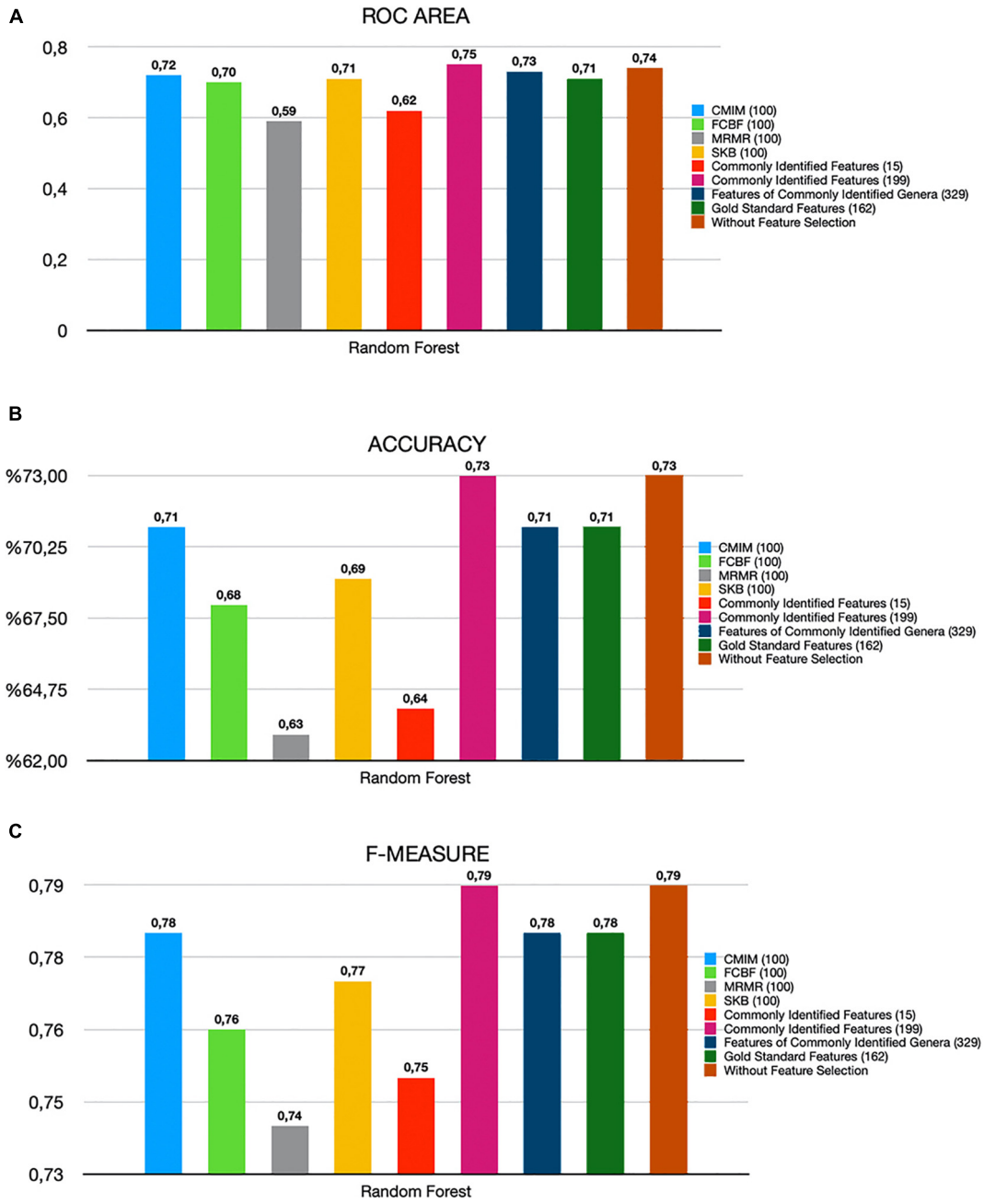


FIGURE 3 | Comparative evaluation of different feature selection methods based on **(A)** ROC area, **(B)** accuracy, and **(C)** F-measure metrics.

possibilities are providing encouraging results. In this respect, in this study, by analyzing the T2D-associated metagenomics data using several supervised and unsupervised machine learning algorithms, we attempt to discover potential taxonomic biomarkers of T2D. Our metagenomics dataset includes the amounts of 1,455 species, measured on the gut microbiota of 290 humans. We used different feature selection algorithms including CMIM, mRMR, FCBF, and SelectKBest. In our preliminary study, we used different classification algorithms including RF, Decision Tree, LogitBoost, AdaBoost, SVM + k means, and Logitboost + k means. In these preliminary experiments, as shown in **Supplementary Table 1** and **Supplementary Figure 1**, we observed that RF resulted in best performance metrics and we decided to continue with our experiments using RF classifier.

All tested feature selection methods commonly identified 15 specific features (as shown in **Figure 2**). Using the amounts of these 15 features, our generated model with RF could predict the T2D status of a sample with 64% accuracy. Compared to the 73% accuracy level using all 1,455 features, 73% accuracy level using 199 selected features, and 71% accuracy level using 162 gold

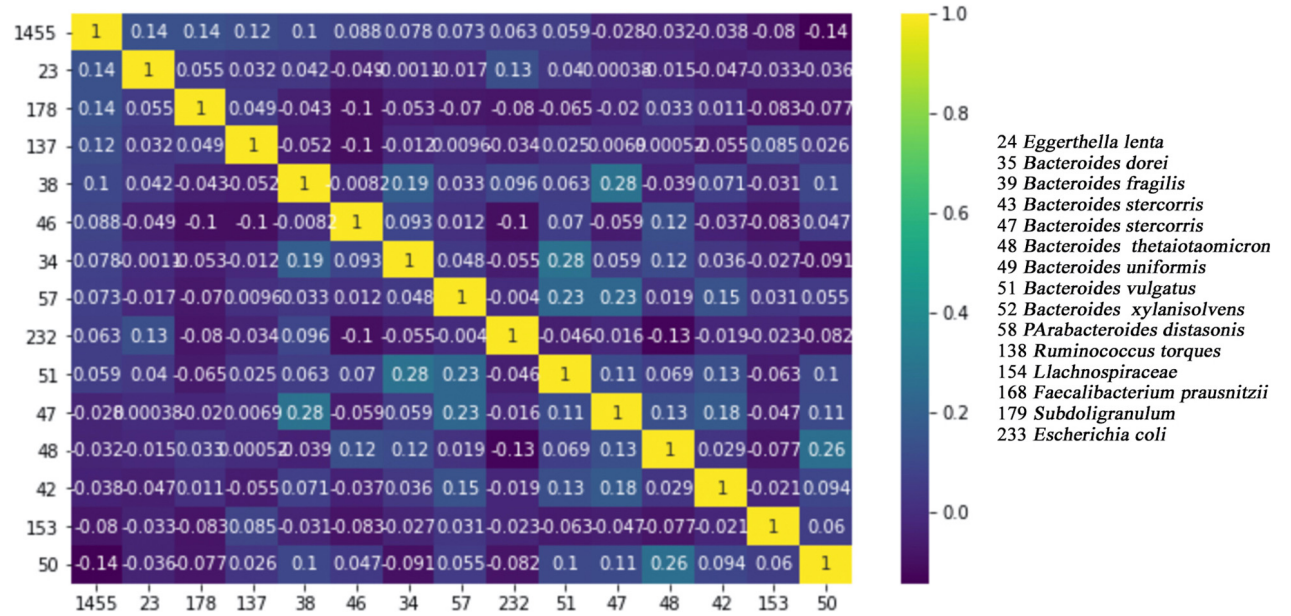


FIGURE 4 | Pairwise correlation heat map of 15 commonly identified features. While number 1 (shown in yellow) indicates full correlation, number 0 (shown in dark blue) indicates no correlation.

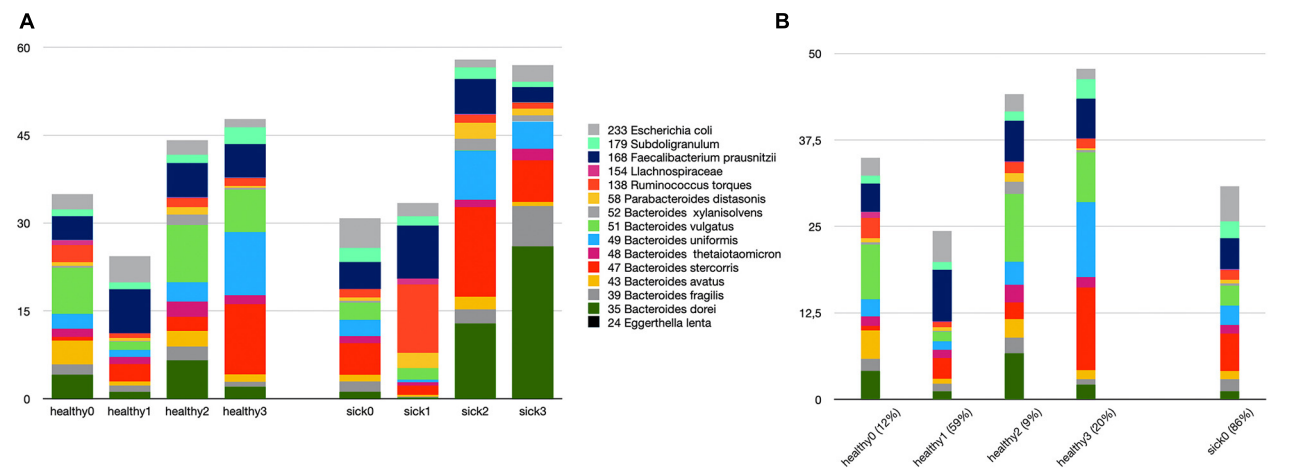


FIGURE 5 | The relative amounts of 15 species **(A)** in all healthy and T2D subgroups. **(B)** Zoomed-in view of all healthy subgroups and one T2D subgroup, which covers more than 86% of all samples.

standard features, these 15 selected features yielded reasonable accuracy results with much lower features. Also, the model using only those 15 species resulted in almost the same F1-score (0.75), with the F1-score obtained using all features (0, 79), as shown in **Figure 3**. Hence, these features could be further evaluated as potential taxonomic biomarkers of T2D. The identified features are *Bacteroides dorei*, *Bacteroides fragilis*, *Bacteroides ovatus*, *Bacteroides stercoris*, *Bacteroides thetaiotaomicron*, *Bacteroides uniformis*, *Bacteroides vulgatus*, *Bacteroides xylanisolvens*, *E. lenta*, *Escherichia coli*, *Faecalibacterium prausnitzii*, *Lachnospiraceae bacterium*, *Parabacteroides distasonis*, *Ruminococcus torques*, and *Subdoligranulum*. The associations of most of these features with T2D is also reported in literature as follows.

A recent review paper (Gurung et al., 2020) summarized the potential mechanisms of microbiota and its effects on the metabolism of T2D patients. Briefly, microbiota modulates inflammation, interacts with dietary constituents, and affects gut permeability, glucose and lipid metabolism, insulin sensitivity, and overall energy homeostasis in the mammalian host. In that study, Gurung et al. highlighted specific taxa that can affect T2D and presented the possible roles of these species in terms of T2D development. They surveyed 42 human observational studies on T2D and the bacterial microbiome, and they reported

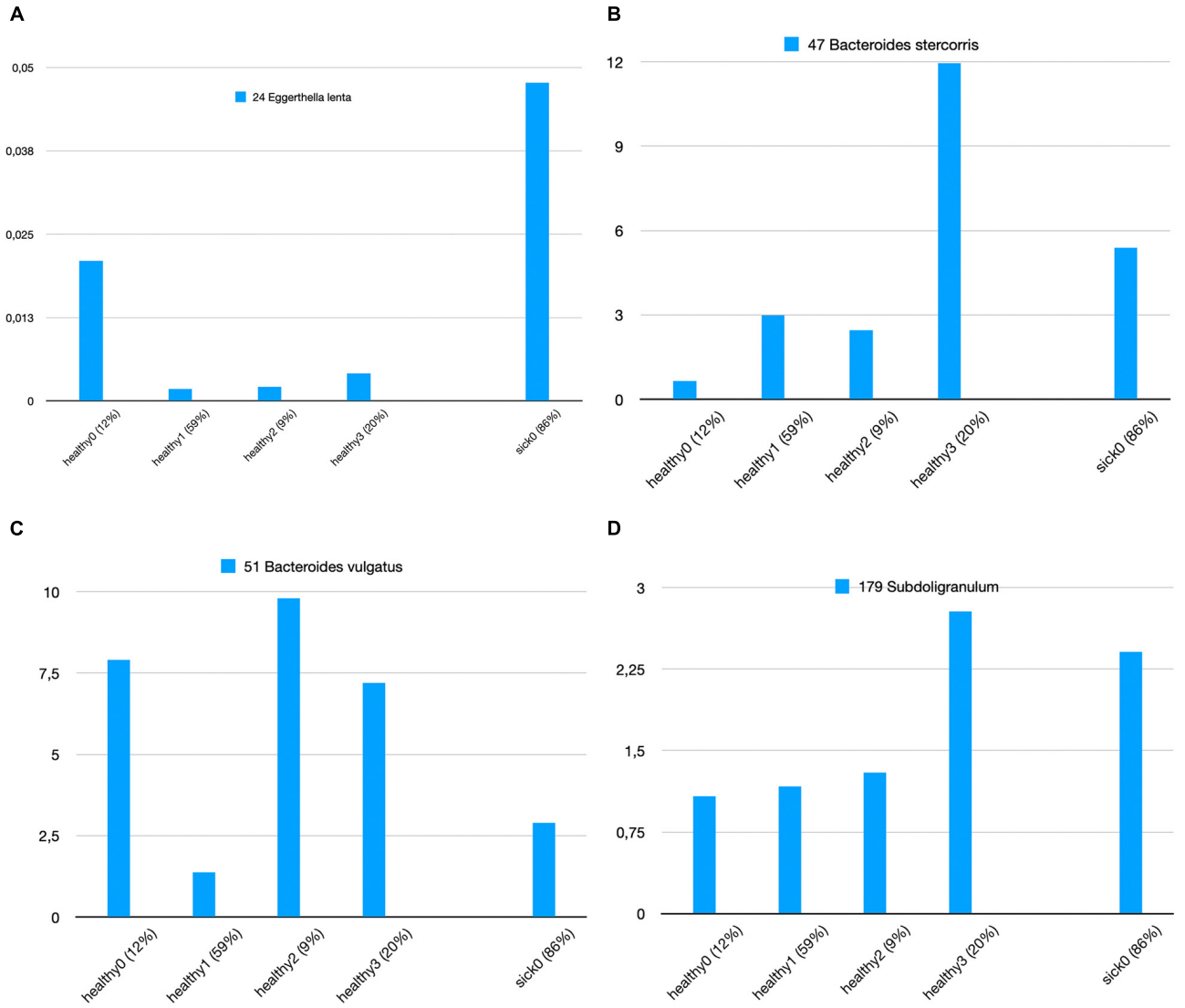


FIGURE 6 | Zoomed-in view of all healthy subgroups and the biggest T2D subgroup for **(A)** *Eggerthella lenta*, **(B)** *Bacteroides stercoris*, **(C)** *Bacteroides vulgatus*, and **(D)** *Subdoligranulum*.

Bacteroides as the second most commonly reported genus (Gurung et al., 2020). The studies that investigated this genus on the species level indicated that the levels of *Bacteroides intestinalis*, *Bacteroides 20-3*, and *Bacteroides vulgatus* were dropped in T2D patients, and the levels of *Bacteroides stercoris* were increased after sleeve gastrectomy surgery in T2D patients with diabetes remission (Wu, 2010; Karlsson et al., 2013; Zhang et al., 2013; Murphy et al., 2017). Additionally, two experimental animal studies tested the ability of *Bacteroides* in order to treat diet-induced metabolic disease (Cano, 2012; Yang, 2017). These studies indicated that the administration of *Bacteroides acidifaciens* (Yang, 2017) and *Bacteroides uniformis* (Cano, 2012) improved glucose intolerance and insulin resistance in diabetic mice. In another study, using a mouse model, Yoshida et al. (2018) found that *B. vulgatus* and *B. dorei* upregulates the expression of tight junction genes in the colon, which leads to reduction in gut permeability, reduction of lipopolysaccharides production, and amelioration of endotoxemia. T2D is known to be associated with increased levels of pro-inflammatory cytokines, chemokines, and inflammatory proteins (Gurung et al., 2020). Along this line, using mono-associated mice, Hoffman et al. (2016) reported that *Bacteroides thetaiotaomicron* reduces Th1, Th2, and Th17 cytokines. Chang et al. (2017) demonstrated that the induction of IL-10 by *Bacteroides fragilis* may contribute to the improvement of glucose metabolism because the overexpression of this cytokine in muscle protects from aging-related insulin resistance (Dagdeviren, 2017; Gurung et al., 2020). Taken together, these studies indicate that *Bacteroides* species play a beneficial role on glucose metabolism in humans and experimental animals. Among these *Bacteroides*

species, *B. dorei*, *B. fragilis*, *B. stercoris*, *B. thetaiotaomicron*, *B. uniformis*, and *B. vulgatus* are identified among the top 15 features list in our study. In addition to these species as potential taxonomic biomarkers of T2D, in this study, we suggest *B. ovatus* and *B. xylanisolvens* as two potential taxonomic biomarkers of T2D. Among the abovementioned *Bacteroides* species, *B. intestinalis*, *B. 20-3*, and *B. acidifaciens* did not exist in our top 15 species list.

In addition to the genera of *Bacteroides*, the effect of *Faecalibacterium* genus with respect to T2D development is discussed in the same review paper by Gurung et al. (2020). Gao et al. (2018) and Salamon et al. (2018) reported the lower frequencies of *Faecalibacterium* in the disease group of case–control study on T2D. While this genus was mostly reported to be decreased after different types of antidiabetic treatments ranging from metformin and herbal medicine (Tong et al., 2018) to bariatric surgery (Murphy et al., 2017), one study reported an opposite effect (Patrone et al., 2016). The studies that investigate this genus at species level usually detected *Faecalibacterium prausnitzii*. *F. prausnitzii* and the peptides secreted by this bacterium are shown to perform anti-inflammatory effects in different chemically induced colitis models in mice (Sokol et al., 2008; Quévrain et al., 2016; Breyner et al., 2017). In different human case–control studies, *F. prausnitzii* was found to be negatively associated with T2D (Furet et al., 2010; Graessler et al., 2013; Karlsson et al., 2013; Zhang et al., 2013; Remely et al., 2014). Although *F. prausnitzii* is commonly used as a probiotic for colitis (Rossi et al., 2015), only a few studies suggested using *F. prausnitzii* as a probiotic for metabolic disease. As shown in **Figure 2**, our top 15 features list includes *F. prausnitzii* and we suggest it as a potential taxonomic biomarker of T2D.

The genera of *Ruminococcus* has also been reported to have a positive association with T2D in the recent review paper by Gurung et al. (2020). Gurung et al. added that the studies reporting species levels of these bacteria reported conflicting results (Graessler et al., 2013; Murphy et al., 2017; Wu et al., 2017). For example, while Wu et al. (2017) found that *Ruminococcus* sp. *SR1/5* is enriched by metformin treatment, Murphy et al. (2017) demonstrated that *Ruminococcus bromii* is enriched and *Ruminococcus torques* is decreased after bariatric surgery and diabetes remission. Among these *Ruminococcus* species, *Ruminococcus torques* is identified among the top 15 features list in our study.

A recent study by Wang et al. (2019) demonstrated that *P. distasonis* prevents obesity and metabolic dysfunctions by producing succinate and secondary bile acids. Using ob/ob and high-fat diet-fed mice, they showed the metabolic benefits of *P. distasonis* in terms of decreasing weight gain, hyperglycemia, and hepatic steatosis. As shown in **Figure 2**, we detected *P. distasonis* in the top 15 features list in our study and we suggested it as a potential taxonomic biomarker of T2D.

Recently, the metformin treatment, which is the most prescribed antidiabetic drug, is shown to disturb the intestinal microbes. Hence, the compositional shifts were detected in the representative gut microbiomes of T2D patients undergoing metformin treatment. *Subdoligranulum variabile* is one of those microbes that is found to display increased abundance in those T2D patients undergoing metformin treatment (Forslund et al., 2015; Mardinoglu et al., 2016; Wu et al., 2017). As shown in **Figure 2**, we identified *S. variabile* in the top 15 features list.

Qin et al. (2012) demonstrated that the opportunistic pathogens (e.g., *Clostridium hatheway*, *Bacteroides caccae*, *E. coli*, and *E. lenta*) are increased in diabetes. On the other hand, Doumatey et al. (2020) reported that they did not find any evidence of such enrichment in their study, where they analyzed the gut microbiome profiles of T2D patients in Urban Africans. As shown in **Figure 2**, our top 15 features list includes *E. coli* and *E. lenta*. Although our top 15 features list did not include *C. hatheway*, different strains of this species are identified by all four tested feature selection methods, as shown in **Supplementary Tables 2**, **4**. We realized that different strains of this species such as *C. hathewayi_GCF_000160095*, *Clostridium hathewayi_GCF_000235505*, and *C. hathewayi unclassified* are detected in the top 100 lists of all four tested feature selection methods, as shown in **Supplementary Table 2**. Also, increased levels of *C. clostridioforme* in T2D patients are reported by Karlsson et al. (2013) and Qin et al. (2012). In our study, *C. clostridioforme* is included within the 199 commonly identified features of top 500 selected features, as shown in **Supplementary Table 3**, and the genera of *Clostridium* is identified by all tested feature selection methods, as shown in **Supplementary Figure 2**.

Lachnospiraceae species constitute the core of gut microbiota. They colonize the intestinal lumen from the birth, and during the host's life, they increase both in terms of the diversity of their species and their relative abundances. Although they are commonly found in the gut microbiota and their members are among the main producers of short-chain fatty acids, different *Lachnospiraceae* species are also associated with different intra- and extraintestinal diseases (Vacca et al., 2020). Kostic et al. (2015) reported that *Lachnospiraceae* genus negatively affects glucose metabolism, which leads to inflammation and promotes the onset of T1D. Along this line, using both human and mouse models, some other metagenomics studies demonstrated that *Lachnospiraceae* may also be specifically associated with T2D (Qin et al., 2012; Kameyama and Itoh, 2014). As shown in **Figure 2**, we detected *Lachnospiraceae* in the top 15 features list in our study.

The recent review paper by Gurung et al. (2020) pointed out that in addition to the genera of *Bacteroides*, the genera of *Bifidobacterium* is another beneficial genera and it is most frequently reported in the studies of T2D. They reported that the genera of *Bifidobacterium* is most consistently supported by the literature in terms of containing the microbes potentially protective against T2D (Gurung et al., 2020). For example, Wu et al. (2017) and Murphy et al. (2017) found a negative association between *Bifidobacterium adolescentis*, *Bifidobacterium bifidum*, *Bifidobacterium pseudocatenulatum*, *Bifidobacterium longum*, *Bifidobacterium dentium*, and disease in patients treated with metformin or after undergoing gastric bypass surgery. Although *Bifidobacterium* has not been used alone as probiotics for T2D, most of the animal studies that

tested different species from this genus (*B. bifidum*, *B. longum*, *B. infantis*, *B. animalis*, *B. pseudocatenulatum*, and *B. breve*) showed improvement of glucose tolerance (Le, 2015; Moya-Perez et al., 2015; Wang, 2015; Aoki, 2017; Kikuchi et al., 2018). These studies strengthen the idea that *Bifidobacterium* naturally habituating the human gut or introduced as probiotics play a protective role in T2D. In our study, several *Bifidobacterium* species (including *B. bifidum*, *B. longum*, *B. pseudocatenulatum*, *B. breve*, *B. animalis*, *B. adolescentis*, and *B. dentium*) are found as important features in the top 100 features lists of each one of four tested feature selection methods (as can be seen in **Supplementary Table 2**). However, each feature selection method selected a different *Bifidobacterium* species. When we get the intersection of the selected features from four different methods, these *Bifidobacterium* species did not show up in the top 15 selected features list. But on the genus level, *Bifidobacterium* is identified by all feature selection methods (as can be seen in **Supplementary Table 2** and **Supplementary Figure 2**). Once we focus on commonly detected genera instead of commonly detected species in all four tested feature selection methods, these *Bifidobacterium* species showed up among those 329 features, and using these features, 0.78 F1-score, 0.73 AUC, and 71% accuracy performance metrics are obtained, as shown in **Figure 3**. On the other hand, when we generate the list of top 500 selected features from each feature selection method and check for the commonly identified features among these four lists (as shown in **Supplementary Table 3**), we end up with 199 commonly selected features. *Bifidobacterium longum*, *B. pseudocatenulatum*, and *B. breve* existed in this list. Classification using these 199 commonly selected features resulted in 73% accuracy, 0.75 ROC, and 0.79 F1-measure, as shown in **Figure 3**. Those selected 199 features also performed better compared to the performance (0.78 F1-score, 0.71 AUC, and 71% accuracy) of the 162 features (species) that belong to the gold standard genera, which are reported to be associated with T2D in a recent review paper (Gurung et al., 2020). **Figure 3** illustrates the comparative evaluation of all the feature selection methods.

Similarly, in our analyses, several *Ruminococcus* species (including *R. gnavus*, *R. obeum*, *R. torques*, *R. albus*, *R. callidus*, *R. sp*, *R. lactaris*, *R. champanellensis*, and *R. flavefaciens*) and several *Blautia* species including *B. hansenii*, *B. producta*, and *B. sp_KLE_1732* are detected as important features in the top 100 features lists of each one of four tested feature selection methods (as can be seen in **Supplementary Table 2**). Accordingly, these species are included in the identified features list of commonly detected genera in all four tested feature selection methods, shown in **Supplementary Table 4**. In Gurung et al. (2020), *Ruminococcus*, *Blautia*, and *Fusobacterium* were reported to be positively associated with T2D. The genera of *Fusobacterium* is identified only by SKB feature selection method, as shown in **Supplementary Table 4**.

On the other hand, two genera (*Akkermansia* and *Roseburia*) that were found to be negatively associated with T2D in Gurung et al. (2020) did not show up in the commonly identified genera list (**Supplementary Figure 2**). However, these two genera were detected in the top 100 lists of different feature selection methods, as shown in **Supplementary Tables 2**, **4**. As shown in **Supplementary Table 4**, while the genera of *Akkermansia* is identified by FCBF and SKB feature selection methods, the genera of *Roseburia* is identified by all tested feature selection methods except mRmR.

Pasolli et al. (2016) attempted to classify the T2D patients and healthy samples using the metagenomic-associated dataset of T2D, downloaded from Qin et al. (2012). They followed the same preprocessing as we performed. Before applying MetaPhlAn2, the samples were subject to standard pre-processing as described in the SOP of the Human Microbiome Project. Similar to our study, they used species abundance as input data and tested the performances of the SVM and RF classifiers and also evaluated Lasso and elastic net regularized multiple logistic regression. On T2D-associated metagenomics dataset, without applying any feature selection, they obtained 0.75 F1-score, 0.62 AUC, and 64% accuracy using RF classifier, as shown in **Figure 2** of their study. Our RF model without applying feature selection methods resulted in 0.79 F1-score, 0.74 AUC, and 73% accuracy, as shown in **Figure 3** and **Table 2**.

Pasolli et al. (2016) also investigated the effect of different feature selection algorithms. On the T2D-associated metagenomics dataset, by only using 40 species (features) that are selected using Lasso feature selection, Pasolli et al. (2016) obtained 0.70 AUC using RF classifier, as shown in **Supplementary Figures 2**, **3**. In our study, by only using 15 species, 0.74 AUC is obtained using RF classifier, as shown in **Figure 3** and **Table 2**. We can conclude that there is added value in studying T2D through metagenomics and machine learning.

Lastly, we clustered the healthy samples and cases according to these 15 features (the amounts of 15 selected species) using *k*-means clustering. Hence, we attempt to distinguish the subgroups of healthy samples and sick samples. While the relative amounts of 15 selected species are shown in **Figure 5** for all healthy and T2D subgroups, in **Figure 6**, the relative amounts of some specific species are shown for all four healthy subgroups vs. sick0 subgroup, which covers 86% of all the patient samples. Once we evaluated **Figures 5**, **6**, we had some important observations. For example, it can be deduced from **Figure 6A** that the amount of *E. lenta* in healthy samples is at least 10–11 times less than its amount in patients. Therefore, the abundance of *E. lenta* can be evaluated as a candidate taxonomic biomarker for T2D disorder. Qin et al. (2012) also demonstrated that the levels of opportunistic pathogens such as *E. lenta* are increased in diabetes. **Figures 6B–D** indicate that *Bacteroides stercoris* (which is numbered as 47), *Bacteroides vulgatus* (which is numbered as 51), and *Subdoligranulum* (which is numbered as 179) can be considered as candidate taxonomic biomarkers of T2D. In literature, the levels of *Bacteroides vulgatus* were reported to be dropped in T2D patients and the levels of *Bacteroides stercoris* were reported to be increased after sleeve gastrectomy surgery in T2D patients with diabetes remission (Wu, 2010; Karlsson et al., 2013; Zhang et al., 2013; Murphy et al., 2017). In another study, using a mouse model, Yoshida et al. found that

B. vulgatus upregulates the expression of tight junction genes in the colon, which leads to reduction in gut permeability, reduction of lipopolysaccharides production, and amelioration of endotoxemia (57). *Subdoligranulum variabile* is one of those microbes that is found to display increased abundance in those T2D patients undergoing metformin treatment (Forslund et al., 2015; Mardinoglu et al., 2016; Wu et al., 2017).

CONCLUSION

Human gut microbiota, which consists of nearly 200 prevalent bacterial species and approximately 1,000 uncommon species, is considered as a multicellular organ. Gut microbiota can affect the host immune system, which is central to program several host activities (Qin et al., 2010). Hence, the metagenomic analysis of the human gut microbiome provides novel insights for several diseases, including T2D. Although several studies reported the significance of the gut microbiota in pathophysiology of T2D, this field is still in its infancy. The existing studies concluded that some microbial taxa and related molecular mechanisms may be involved in glucose metabolism related to T2D. Nevertheless, such simple interpretations are not enough to explain the heterogeneity and complexity of T2D, and the redundancy of gut microbiota further complicates these analyses. Along this line, in this study, we used the T2D-associated metagenomics data and developed a machine learning model to increase the diagnostic accuracy of T2D. We discovered potential taxonomic biomarkers of T2D and investigated which subset of microbiota is more informative than other taxa applying some of the state-of-the art feature selection methods. In our experiments, especially 15 species came into prominence. We present support from literature regarding the association of these species with T2D. Hence, we propose these species as candidate taxonomic biomarkers of T2D, where wet lab scientists can design validation experiments.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: The data is taken from the following paper: Qin et al. (2012). Raw sequencing data is obtained from NCBI Sequence Read Archive with SRA045646 accession number.

AUTHOR CONTRIBUTIONS

BB-G conceived the ideas and designed the study. AJ, ON, and MY conducted the experiments. BB-G, OB, AJ, and MY analyzed the results. BB-G, OB, AJ, ON, and MY participated in the discussion of the results and writing of the article. All authors read and approved the final version of the manuscript.

FUNDING

The work of BB-G has been supported by the Abdullah Gul University Support Foundation (AGUV). The work of MY has been supported by the Zefat Academic College. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2021.628426/full#supplementary-material

Supplementary Figure 1 | Preliminary performance evaluation results for T2D-associated metagenomics dataset using 10 fold cross-validation. Comparative evaluation of different classifiers using different feature selection methods based on **(A)** ROC area, **(B)** accuracy, and **(C)** F-measure.

Supplementary Figure 2 | Numbers of identified genus, which are selected by different feature selection algorithms. The commonalities between the selected genus in different methods are also illustrated.

Supplementary Figure 3 | Selection of the optimum number of the clusters for the **(A)** controls and **(B)** T2D patients. Using the elbow method, four clusters are found as the optimum number of clusters for both the controls [as shown in panel **(A)**] and T2D patients [as shown in panel **(B)**].

Supplementary Table 1 | Preliminary analysis results for T2D-associated metagenomics dataset using 10 fold cross-validation. Evaluation of different classification methods based on different performance measures (a) without feature selection and after applying (b) CMIM, (c) mRMR, (d) FCBF, and (e) SKB feature selection algorithms.

Supplementary Table 2 | Top 100 selected features for each feature selection method and 15 commonly identified features among these four lists.

Supplementary Table 3 | Top 500 selected features for each feature selection method and 199 commonly identified features among these four lists.

Supplementary Table 4 | Identified features of commonly detected genera in top 100 lists of all four tested feature selection methods.

Supplementary Table 5 | Features of gold standard genera that are reported to be associated with T2D in Gurung et al. (2020).

REFERENCES

Albarracin, C. A., Fuqua, B. C., Evans, J. L., and Goldfine, I. D. (2008). Chromium picolinate and biotin combination improves glucose metabolism in treated, uncontrolled overweight to obese patients with type 2 diabetes. *Diabetes Metab. Res. Rev.* 24, 41–51. doi: 10.1002/dmrr.755

Allin, K. H., Tremaroli, V., Caesar, R., Jensen, B. A. H., Damgaard, M. T. F., Bahl, M. I., et al. (2018). Aberrant intestinal microbiota in individuals with prediabetes. *Diabetologia* 61, 810–820. doi: 10.1007/s00125-018-4550-1

Aoki, R. (2017). A proliferative probiotic bifidobacterium strain in the gut ameliorates progression of metabolic disorders via microbiota modulation and acetate elevation. *Sci. Rep.* 7:43522.

Aw, W., and Fukuda, S. (2018). Understanding the role of the gut ecosystem in diabetes mellitus. *J. Diabetes Investig.* 9, 5–12. doi: 10.1111/jdi.12673

Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., et al. (2008). "KNIME: The Konstanz Information Miner," in *Data Analysis, Machine Learning and Applications. Studies in Classification, Data Analysis,*

and Knowledge Organization, eds C. Preisach, H. Burkhardt, L. Schmidt-Thieme, and R. Decker (Berlin: Springer), 319–326. doi: 10.1007/978-3-540-78246-9_38

Boulangeé, C. L., Neves, A. L., Chilloux, J., Nicholson, J. K., and Dumas, M. E. (2016). Impact of the gut microbiota on inflammation, obesity, and metabolic disease. *Genome Med.* 8:42.

Breyner, N. M., Michon, C., de Sousa, C. S., Vilas Boas, P. B., Chain, F., Azevedo, V. A., et al. (2017). Microbial anti-inflammatory molecule (MAM) from *Faecalibacterium prausnitzii* shows a protective effect on DNBS and DSS-induced colitis model in mice through inhibition of NF- kappa B pathway. *Front. Microbiol.* 8:114. doi: 10.3389/fmicb.2017.00114

Brown, G., Pocock, A., Zhao, M., and Luján, M. (2012). Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *J. Mach. Learn. Res.* 13, 27–66.

Brunetti, P. (2007). The lean patient with type 2 diabetes: characteristics and therapy challenge. *Int. J. Clin. Pract. Suppl.* 153, 3–9. doi: 10.1111/j.1742-1241.2007.01359.x

Cano, G. (2012). *Bacteroides* uniformis CECT 7771 ameliorates metabolic and immunological dysfunction in mice with high-fat-diet induced obesity. *PLoS One* 7:e41079. doi: 10.1371/journal.pone.0041079

Centers for Disease Control and Prevention (2020). *National Diabetes Statistics Report, 2020: Estimates of Diabetes and its Burden in the United States*. Atlanta, GA: Centers for Disease Control and Prevention.

Chang, Y. C., Ching, Y. H., Chiu, C. C., Liu, J. Y., Hung, S. W., and Huang, W. C. (2017). TLR2 and interleukin-10 are involved in *Bacteroides* fragilis-mediated prevention of DSS-induced colitis in gnotobiotic mice. *PLoS One* 12:e0180025. doi: 10.1371/journal.pone.0180025

Chobot, A., Górowska-Kowolik, K., Sokołowska, M., and Jarosz-Chobot, P. (2018). Obesity and diabetes-not only a simple link between two epidemics. *Diabetes Metab. Res. Rev* 34:e3042. doi: 10.1002/dmrr.3042

Dagdeviren, S. (2017). IL-10 prevents aging-associated inflammation and insulin resistance in skeletal muscle. *FASEB J.* 31, 701–710. doi: 10.1096/fj.201600832r

Daousi, C., Casson, I. F., Gill, G. V., MacFarlane, I. A., Wilding, J. P. H. H., and Pinkney, J. H. (2006). Prevalence of obesity in type 2 diabetes in secondary care: association with cardiovascular risk factors. *Postgrad. Med. J.* 82, 280–284. doi: 10.1136/pmj.2005.039032

Diabetes.co.uk (2019). *The Global Diabetes Community–Diabetes In Chı na–2019*. Available online at: https://www.diabetes.co.uk/global-diabetes/diabetes-in-china.html#:~{}:text=The%20number%20of%20people%20with,diabetes%20are%20occurring%20each%20year (accessed October, 2020).

Doumatey, A. P., Adeyemo, A., Zhou, J., Lei, L., Adebamowo, S. N., Adebamowo, C., et al. (2020). Gut microbiome profiles are associated with type 2 diabetes in Urban Africans. *Front. Cell Infect. Microbiol.* 10:63. doi: 10.3389/fcimb.2020.00063

Duvallet, C., Gibbons, S. M., Gurry, T., Irizarry, R. A., and Alm, E. J. (2017). Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nat. Commun.* 8:1784.

Falony, G., Joossens, M., Vieira-Silva, S., Wang, J., Darzi, Y., Faust, K., et al. (2016). Population-level analysis of gut microbiome variation. *Science* 352, 560–564.

Fernandez-Mejia, C. (2005). Pharmacological effects of biotin. *J. Nutr. Biochem.* 16, 424–427. doi: 10.1016/j.jnutbio.2005.03.018

Fleuret, F. (2004). Fast binary feature selection with conditional mutu-al information. *J. Mach. Learn. Res.* 13, 1531–1555.

Forslund, K., Hildebrand, F., Nielsen, T., Falony, G., Le Chatelier, E., Sunagawa, S., et al. (2015). Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. *Nature* 528, 262–266. doi: 10.1038/nature15766

Furet, J. P., Kong, L. C., Tap, J., Poitou, C., Basdevant, A., Bouillot, J. L., et al. (2010). Differential adaptation of human gut microbiota to bariatric sur-gery-induced weight loss: links with metabolic and low-grade inflammation markers. *Diabetes* 59, 3049–3057. doi: 10.2337/db10-0253

Gao, R., Zhu, C., Li, H., Yin, M., Pan, C., Huang, L., et al. (2018). Dysbiosis signatures of gut microbiota along the sequence from healthy, young patients to those with overweight and obesity. *Obesity (Silver Spring)* 26, 351–361. doi: 10.1002/oby.22088

Graessler, J., Qin, Y., Zhong, H., Zhang, J., Licinio, J., Wong, M. L., et al. (2013). Metagenomic sequencing of the human gut microbiome beforeand after bariatric surgery in obese patients with type 2 diabetes: correlation with inflammatory and metabolic parameters. *Pharmacogenomics J.* 13, 514–522. doi: 10.1038/tpj.2012.43

Gurung, M., Li, Z., You, H., Rodrigues, R., Jump, D. B., Morgun, A., et al. (2020). Role of gut microbiota in type 2 diabetes pathophysiology. *EBioMedicine* 51:102590. doi: 10.1016/j.ebiom.2019.11.051

Hacilar, H., Nalbantoglu, O. U., Aran, O., and Bakir-Gungor, B. (2019). Inflammatory bowel disease biomarkers of human gut microbiota selected via ensemble feature selection methods. *arXiv* [Preprint] arXiv:2001.03019,

He, Y., Wu, W., Zheng, H. M., Li, P., McDonald, D., Sheng, H. F., et al. (2018). Regional variation limits applications of healthy gut microbiome reference ranges and disease models. *Nat. Med.* 24, 1532–1535. doi: 10.1038/s41591-018-0164-x

Hoffmann, T. W., Pham, H. P., Bridonneau, C., Aubry, C., Lamas, B., Martin-Gallausiaux, C., et al. (2016). Microorganisms linked to inflammatory bowel disease- associated dysbiosis differentially impact host physiology in gnotobiotic mice. *ISME J.* 10, 460–477. doi: 10.1038/ismej.2015.127

Human Microbiome Project Consortium (2012). Structure, function and diversity of the healthy human microbiome. *Nature* 486, 207–214. doi: 10.1038/nature11234

International Diabetes Federation (2003). *Diabetes Atlas*, 2nd Edn. Brussels: International Diabetes Federation.

International Diabetes Federation (2019). *IDF Diabetes Atlas*, 9th Edn. Brussels: International Diabetes Federation.

James, W. P. T., Jackson-Leach, R., Mhurdu, C. N., Kalamara, E., Shayeghi, M., Rigby, N. J., et al. (2003). "Overweight and obesity," in *Comparative Quantification of Health Risks: Global and Regional Burden of Disease Attributable to Selected Major Risk Factors*, eds M. Ezzati, A. D. Lopez, A. Rodgers, and C. J. L. Murray (Geneva: WHO).

Kameyama, K., and Itoh, K. (2014). Intestinal colonization by a Lachnospiraceaebacterium contributes to the development of diabetes in obese mice. *Microbes Environ.* 29, 427–430. doi: 10.1264/jsme2.me14054

Karlsson, F. H., Tremaroli, V., Nookaew, I., Bergström, G., Behre, C. J., Fagerberg, B., et al. (2013). Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature* 498, 99–103. doi: 10.1038/nature12198

Kikuchi, K., Ben Othman, M., and Sakamoto, K. (2018). Sterilized bifidobacteria suppressed fat accumulation and blood glucose level. *Biochem. Biophys. Res. Commun.* 501, 1041–1047. doi: 10.1016/j.bbrc.2018.05.105

Kostic, A. D., Gevers, D., Siljander, H., Vatanen, T., Hyötylöinen, T., Hömölöinen, A. M., et al. (2015). The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes. *Cell Host Microbe* 17, 260–273. doi: 10.1016/j.chom.2015.01.001

Kuang, Z., Wang, Y., Li, Y., Ye, C., Ruhn, K. A., Behrendt, C. L., et al. (2019). The intestinal microbiota programs diurnal rhythms in host metabolism through histone deacetylase 3. *Science* 365, 1428–1434. doi: 10.1126/science.aaw3134

Larsen, N., Vogensen, F. K., van den Berg, F. W. J., Nielsen, D. S., Andreasen, A. S., Pedersen, B. K., et al. (2010). Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults. *PLoS One* 5:e9085. doi: 10.1371/journal.pone.0009085

Lazo de la Vega-Monroy, M. L., Larrieta, E., German, M. S., Baez-Saldana, A., and Fernandez-Mejia, C. (2013). Effects of biotin supplementation in the diet on insulin secretion, islet gene expression, glucose homeostasis and beta-cell proportion. *J. Nutr. Biochem.* 24, 169–177. doi: 10.1016/j.jnutbio.2012.03.020

Lê, K. A., Li, Y., Xu, X., Yang, W., Liu, T., Zhao, X., et al. (2013). Alterations in fecal *Lactobacillus* and *Bifidobacterium* species in type 2 diabetic patients in Southern China population. *Front. Physiol.* 3:496. doi: 10.3389/fphys.2012.00496

Le, T. K. (2015). *Bifidobacterium* species lower serum glucose, increase expressions of insulin signaling proteins, and improve adipokine profile in diabetic mice. *Biomed. Res.* 36, 63–70. doi: 10.2220/biomedres.36.63

Liang, X., Bushman, F. D., and FitzGerald, G. A. (2015). Rhythmicity of the intestinal microbiota is regulated by gender and the host circadian clock. *Proc. Natl. Acad. Sci. U.S.A.* 112, 10479–10484. doi: 10.1073/pnas.1501305112

Maebashi, M., Makino, Y., Furukawa, Y., Ohinata, K., Kimura, S., and Sato, T. (1993). Therapeutic evaluation of the effect of biotin on hyperglycemia in patients with non-insulin dependent diabetes mellitus. *J. Clin. Biochem. Nutr.* 14, 211–218. doi: 10.3164/jcbn.14.211

Makki, K., Deehan, E. C., Walter, J., and Bäckhed, F. (2018). The impact of dietary fiber on gut microbiota in host health and disease. *Cell Host Microbe* 23, 705–715. doi: 10.1016/j.chom.2018.05.012

Manor, O., Dai, C. L., Kornilov, S. A., Smith, B., Price, N. D., Lovejoy, J. C., et al. (2020). Health and disease markers correlate with gut microbiome omposition across thousands of people. *Nat. Commun.* 11:5206.

Marcos-Zambrano, L. J., Karaduzovic-Hadziabdic, K., Przymus, P., Trajkovik, V., Aasmets, O., Berland, M., et al. (2021). Applications of machine learning in human microbiome studies: a review on feature selection, biomarker identification, disease prediction and treatment. *Front. Microbiol.* 12:634511. doi: 10.3389/fmicb.2021.634511

Mardinoglu, A., Boren, J., and Smith, U. (2016). Confounding effects of metformin on the human gut microbiome in type 2 diabetes. *Cell Metab.* 23, 10–12. doi: 10.1016/j.cmet.2015.12.012

Mendes-Soares, H., Raveh-Sadka, T., Azulay, S., Edens, K., Ben-Shlomo, Y., Cohen, Y., et al. (2019). Assessment of a personalized approach to predicting postprandial glycemic responses to food among individuals without diabetes. *JAMA Netw. Open* 2:e188102. doi: 10.1001/jamanetworkopen.2018.8102

Moya-Perez, A., Neef, A., and Sanz, Y. (2015). *Bifidobacterium pseudocatenulatum* CECT 7765 reduces obesity-associated inflammation by restoring the lymphocyte-macrophage balance and gut microbiota structure in high-fat diet-fed mice. *PLoS One* 10:e0126976. doi: 10.1371/journal.pone.0126976

Murphy, R., Tsai, P., Jüllig, M., Liu, A., Plank, L., and Booth, M. (2017). Differential changes in gut microbiota after gastric bypass and sleeve gastrectomy bariatric surgery vary according to diabetes remission. *Obes. Surg.* 27, 917–925. doi: 10.1007/s11695-016-2399-2

Narayan, K. M. V., Boyle, J. P., Thompson, T. J., Gregg, E. W., and Williamson, D. F. (2007). Effect of BMI on lifetime risk for diabetes in the U.S. *Diabetes Care* 30, 1562–1566. doi: 10.2337/dc06-2544

National Diabetes Clearinghouse (2011). *National Diabetes Statistics 2011*. Bethesda, MD: National Institute of Health.

Pasolli, E., Truong, D. T., Malik, F., Waldron, L., and Segata, N. (2016). Machine learning meta- analysis of large metagenomic datasets: tools and biological insights. *PLoS Comput. Biol.* 12:e1004977. doi: 10.1371/journal.pcbi.1004977

Patrone, V., Vajana, E., Minuti, A., Callegari, M. L., Federico, A., Loguercio, C., et al. (2016). Postoperative changes in fecal bacterial communities and fer-mentation products in obese patients undergoing bilio-intestinal bypass. *Front. Microbiol.* 7:200. doi: 10.3389/fmicb.2016.00200

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., and Thirion, B. (2011). Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* 12, 2825–2830.

Peters, B. A., Shapiro, J. A., Church, T. R., Miller, G., Trinh-Shevrin, C., Yuen, E., et al. (2018). A taxonomic signature of obesity in a large study of American adults. *Sci. Rep.* 8:9749.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464, 59–65.

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

Quévrain, E., Maubert, M. A., Michon, C., Chain, F., Marquant, R., Tailhades, J., et al. (2016). Identification of an anti-inflammatory protein from *Faecalibacterium prausnitzii*, a commensal bacterium deficient in Crohn's disease. *Gut* 65, 415–425.

Reitmeier, S., Kiessling, S., Clavel, T., List, M., Almeida, E. L., Ghosh, T. S., et al. (2020). Arrhythmic gut microbiome signatures predict risk of type 2 diabetes. *Cell Host Microbe* 28, 258–272.e6.

Remely, M., Aumueller, E., Merold, C., Dworzak, S., Hippe, B., Zanner, J., et al. (2014). Effects of short chain fatty acid producing bacteria on epi-genetic regulation of FFAR3 in type 2 diabetes and obesity. *Gene* 537, 85–92. doi: 10.1016/j.gene.2013.11.081

Rossi, O., Khan, M. T., Schwarzer, M., Hudcovic, T., Srutkova, D., Duncan, S. H., et al. (2015). *Faecalibacterium prausnitzii* strain HTF-F and its extracellular polymeric matrix attenuate clinical parameters in DSS-Induced colitis. *PLoS One* 10:e0123013. doi: 10.1371/journal.pone.0123013

Salamon, D., Sroka-Oleksiak, A., Kapusta, P., Szopa, M., Mrozińska, S., Ludwig-Słomczyńska, A. H., et al. (2018). Characteristics of gut microbiota in adult patients with type 1and type 2 diabetes based on nextgeneration sequencing of the 16S rRNA genefragment. *Pol. Arch. Intern. Med.* 128, 336–343.

Sanna, S., van Zuydam, N. R., Mahajan, A., Kurilshikov, A., Vich Vila, A., Vo þ sa, U., et al. (2019). Causal relationships among the gut microbiome, short-chain fatty acids and metabolic diseases. *Nat. Genet.* 51, 600–605. doi: 10.1038/s41588-019-0350-x

Senliol, B., Gulgezen, G., Yu, L., and Cataltepe, Z. (2008). "Fast Correlation Based Filter (FCBF) with a Different Search Strategy," in *Proceedings of the 2008 23rd International Symposium on Computer and Information Sciences*, Istanbul.

Sharma, S., and Tripathi, P. (2019). Gut microbiome and type 2 diabetes: where we are and where to go? *J. Nutr. Biochem.* 63, 101–108. doi: 10.1016/j.jnutbio.2018.10.003

Sokol, H., Pigneur, B., Watterlot, L., Lakhdari, O., Bermúdez-Humarán, L. G., Gratadoux, J. J., et al. (2008). *Faecalibacterium prausnitzii* is an anti-inflammatory commensal bacte- rium identified by gut microbiota analysis of Crohn disease patients. *Proc. Natl. Acad. Sci. U.S.A.* 105, 16731–16736. doi: 10.1073/pnas.0804812105

Steinley, D., and Brusco, M. J. (2007). Initializing k-means batch clustering: a critical evaluation of several techniques. *J. Classif.* 24, 99–121. doi: 10.1007/s00357-007-0003-0

Sun, L., Xie, C., Wang, G., Wu, Y., Wu, Q., Wang, X., et al. (2018). Gut microbiota and intestinal FXR mediate the clinical benefits of metformin. *Nat. Med.* 24, 1919–1929. doi: 10.1038/s41591-018-0222-4

Tabak, A. G., Herder, C., Rathmann, W., Brunner, E. J., and Kivimeaki, M. (2012). Prediabetes: a high-risk state for diabetes development. *Lancet* 379, 2279–2290. doi: 10.1016/s0140-6736(12)60283-9

Thaiss, C. A., Zeevi, D., Levy, M., Zilberman-Schapira, G., Suez, J., Tengeler, A. C., et al. (2014). Transkingdom control of microbiota diurnal oscillations promotes metabolic homeostasis. *Cell* 159, 514–529. doi: 10.1016/j.cell.2014.09.048

Thingholm, L. B., Rühlemann, M. C., Koch, M., Fuqua, B., Laucke, G., Boehm, R., et al. (2019). Obese individuals with and without type 2 diabetes show different gut microbial functional capacity and composition. *Cell Host Microbe* 26, 252–264.e10.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B (Methodol.)* 58, 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x

Tong, X., Xu, J., Lian, F., Yu, X., Zhao, Y., Xu, L., et al. (2018). Structural alteration of gut microbiota during the amelioration of human type 2 diabetes with hyperlipidemia by metformin and a traditional Chi- nese herbal formula: a multicenter, randomized, open label clinical trial. *mBio* 9, e2392–e2317.

Trøseid, M., Nestvold, T. K., Rudi, K., Thoresen, H., Nielsen, E. W., and Lappegård, K. T. (2013). Plasma lipopolysaccharide is closely associated with glycemic control and abdominal obesity: evidence from bariatric surgery. *Diabetes Care* 36, 3627–3632. doi: 10.2337/dc13-0451

Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., et al. (2009). A core gut microbiome in obese and lean twins. *Nature* 457, 480–484. doi: 10.1038/nature07540

Vacca, M., Celano, G., Calabrese, F. M., Portincasa, P., Gobbetti, M., and De Angelis, M. (2020). The controversial role of human gut *Lachnospiraceae*. *Microorganisms* 8:573. doi: 10.3390/microorganisms8040573

Valdes, A. M., Walter, J., Segal, E., and Spector, T. D. (2018). Role of the gut microbiota in nutrition and health. *BMJ* 361:k2179. doi: 10.1136/bmj.k2179

Vrieze, A., Van Nood, E., Holleman, F., Salojaervi, J., Kootte, R. S., Bartelsman, J. F. W. M., et al. (2012). Transfer of intestinal microbiota from lean donors increases insulin sensitivity in individuals with metabolic syndrome. *Gastroenterology* 143, 913–916.e7.

Wang, J. (2015). Modulation of gut microbiota during probiotic-mediated attenuation of metabolic syndrome in high fat diet-fed mice. *ISME J.* 9, 1–15. doi: 10.1038/ismej.2014.99

Wang, K., Liao, M., Zhou, N., Bao, L., Ma, K., Zheng, Z., et al. (2019). Para*Bacteroides* distasonis alleviates obesity and metabolic dysfunctions via production of succinate and secondary bile acids. *Cell Rep.* 26, 222–235.e5.

Wu, H., Esteve, E., Tremaroli, V., Khan, M. T., Caesar, R., Manneras-Holm, L., et al. (2017). Metformin alters the gut microbiome of individuals with treatment-naive type 2 diabetes, contributing to the therapeutic effects of the drug. *Nat. Med.* 23, 850–858.

Wu, H., Tremaroli, V., Schmidt, C., Lundqvist, A., Olsson, L. M., Krämer, M., et al. (2020). The gut microbiota in prediabetes and diabetes: a population-based cross-sectional study. *Cell Metab.* 32, 379–390.e3. doi: 10.1016/j.cmet.2020.06.011

Wu, X. (2010). Molecular characterisation of the faecal microbiota in patients with type II diabetes. *Curr. Microbiol.* 61, 69–78. doi: 10.1007/s00284-010-9582-9

Xu, Q.-S., and Liang, Y.-Z. (2001). Monte Carlo cross validation. *Chemom. Intell. Lab. Syst.* 56, 1–11. doi: 10.1016/s0169-7439(00)00122-2

Yang, J. Y. (2017). Gut commensal *Bacteroides* acidifaciens prevents obesity and improves insulin sensitivity in mice. *Mucosal Immunol.* 10, 104–116. doi: 10.1038/mi.2016.42

Yoshida, N., Emoto, T., Yamashita, T., Watanabe, H., Hayashi, T., Tabata, T., et al. (2018). *Bacteroides* vulgatus and *Bacteroides* dorei reduce gut microbial lipopolysaccharide production and inhibit atherosclerosis. *Circulation* 138, 2486–2498. doi: 10.1161/circulationaha.118.033714

Zeevi, D., Korem, T., Zmora, N., Israeli, D., Rothschild, D., Weinberger, A., et al. (2015). Personalized nutrition by prediction of glycemic responses. *Cell* 163, 1079–1094.

Zhang, C., Zhang, M., Wang, S., Han, R., Cao, Y., Hua, W., et al. (2010). Interactions between gut microbiota, host genetics and diet relevant to development of metabolic syndromes in mice. *ISME J.* 4, 232–241. doi: 10.1038/ismej.2009.112

Zhang, X., Shen, D., Fang, Z., Jie, Z., Qiu, X., Zhang, C., et al. (2013). Human gut microbiota changes reveal the progression of glucose intolerance. *PLoS One* 8:e71108. doi: 10.1371/journal.pone.0071108

Zhernakova, A., Kurilshikov, A., Bonder, M. J., Tigchelaar, E. F., Schirmer, M., Vatanen, T., et al. (2016). Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science* 352, 565–569. doi: 10.1126/science.aad3369

Zhong, H., Ren, H., Lu, Y., Fang, C., Hou, G., Yang, Z., et al. (2019). Distinct gut metagenomics and metaproteomics signatures in prediabetics and treatment-naïve type 2 diabetics. *EBioMedicine* 47, 373–383. doi: 10.1016/j.ebiom.2019.08.048

Zhou, W., Sailani, M. R., Contrepois, K., Zhou, Y., Ahadi, S., Leopold, S. R., et al. (2019). Longitudinal multi- omics of host-microbe dynamics in prediabetes. *Nature* 569, 663–671.

Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* 67(Pt 2), 301–320. doi: 10.1111/j.1467-9868.2005.00503.x

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Microbiology

MINI REVIEW
published: 11 April 2022
doi: 10.3389/fmicb.2022.627892

Could Artificial Intelligence/Machine Learning and Inclusion of Diet-Gut Microbiome Interactions Improve Disease Risk Prediction? Case Study: Coronary Artery Disease

Baiba Vilne[1,2]*, *Juris Ķibilds*[3], *Inese Siksna*[3], *Ilva Lazda*[3], *Olga Valciņa*[3] and *Angelika Krūmiņa*[3,4]

[1] Bioinformatics Lab, Riga Stradins University, Riga, Latvia, [2] COST Action CA18131 - Statistical and Machine Learning Techniques in Human Microbiome Studies, Brussels, Belgium, [3] Institute of Food Safety, Animal Health and Environment BIOR, Riga, Latvia, [4] Department of Infectology and Dermatology, Riga Stradins University, Riga, Latvia

OPEN ACCESS

Edited by:
Isabel Moreno Indias, Universidad de Málaga, Spain

Reviewed by:
Alinne Castro, Dom Bosco Catholic University, Brazil

***Correspondence:**
Baiba Vilne
baiba.vilne@rsu.lv

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

Received: *10 November 2020*
Accepted: *24 February 2022*
Published: *11 April 2022*

Citation:
Vilne B, Ķibilds J, Siksna I, Lazda I, Valciņa O and Krūmiņa A (2022) Could Artificial Intelligence/Machine Learning and Inclusion of Diet-Gut Microbiome Interactions Improve Disease Risk Prediction? Case Study: Coronary Artery Disease. Front. Microbiol. 13:627892. doi: 10.3389/fmicb.2022.627892

Coronary artery disease (CAD) is the most common cardiovascular disease (CVD) and the main leading cause of morbidity and mortality worldwide, posing a huge socio-economic burden to the society and health systems. Therefore, timely and precise identification of people at high risk of CAD is urgently required. Most current CAD risk prediction approaches are based on a small number of traditional risk factors (age, sex, diabetes, LDL and HDL cholesterol, smoking, systolic blood pressure) and are incompletely predictive across all patient groups, as CAD is a multi-factorial disease with complex etiology, considered to be driven by both genetic, as well as numerous environmental/lifestyle factors. Diet is one of the modifiable factors for improving lifestyle and disease prevention. However, the current rise in obesity, type 2 diabetes (T2D) and CVD/CAD indicates that the "one-size-fits-all" approach may not be efficient, due to significant variation in inter-individual responses. Recently, the gut microbiome has emerged as a potential and previously under-explored contributor to these variations. Hence, efficient integration of dietary and gut microbiome information alongside with genetic variations and clinical data holds a great promise to improve CAD risk prediction. Nevertheless, the highly complex nature of meals combined with the huge inter-individual variability of the gut microbiome poses several Big Data analytics challenges in modeling diet-gut microbiota interactions and integrating these within CAD risk prediction approaches for the development of personalized decision support systems (DSS). In this regard, the recent re-emergence of Artificial Intelligence (AI) / Machine Learning (ML) is opening intriguing perspectives, as these approaches are able to capture large and complex matrices of data, incorporating their interactions and identifying both linear and non-linear relationships. In this Mini-Review, we consider (1) the most used AI/ML approaches and their different use cases for CAD risk prediction (2) modeling of the

content, choice and impact of dietary factors on CAD risk; (3) classification of individuals by their gut microbiome composition into CAD cases vs. controls and (4) modeling of the diet-gut microbiome interactions and their impact on CAD risk. Finally, we provide an outlook for putting it all together for improved CAD risk predictions.

Keywords: machine learning, diet, gut microbiome, personalized nutrition, coronary artery disease, artificial intelligence, risk prediction

1. INTRODUCTION

Coronary artery disease (CAD) is the most common cardiovascular disease (CVD) and the main leading cause of morbidity and mortality worldwide, posing a huge socio-economic burden to the society and health systems (Lopez et al., 2006). Currently, our health care system is facing a paradigm shift from a "one size fits all" approach to a more optimized model to identify prevention strategies and treatments tailored to each individual, the so called personalized medicine. Moreover, the vision of prevention has also transformed toward a concept of "positive health" and primordial prevention—the prevention of disease risk factors before they actually occur, i.e., through targeted modifications of person's environment/lifestyle (Movsisyan et al., 2020). Therefore, timely and precise identification of people at high risk of CAD is of utmost importance for the personalized cardiology (Alaa et al., 2019), as such persons may need more aggressive health promotion strategies, especially the modifiable CAD risk factors could be effectively reduced or even eliminated in this way (Movsisyan et al., 2020).

Over the past two decades, numerous approaches for CAD risk prediction have been developed and several have also entered the clinical routine such as the Framingham Risk Score (FRS) (Wilson et al., 1998) or the Systematic Coronary Risk Evaluation (SCORE) metrics (Conroy et al., 2003), extensively reviewed elsewhere (Damen et al., 2016; Westerlund et al., 2021). However, these approaches are mostly based on a limited number of predictors—the traditional CAD risk factors (age, sex, diabetes, systolic blood pressure, LDL/HDL cholesterol, smoking). Hence, incompletely predictive of disease onset, progression and clinical outcome across all patient groups (Alaa et al., 2019), overestimating the 10-year CAD/CVD risk, especially for high-risk individuals and European populations (Damen et al., 2016). These models typically do not take into account the fact that the treatment options have improved and that, by modifying the person's environment/lifestyle, the disease risk can be reduced over time (Westerlund et al., 2021).

CAD is a multi-factorial disease with complex etiology, considered to be driven by both environment/lifestyle and genetic factors (Davey Smith et al., 2005; Erdmann et al., 2018; Vilne and Schunkert, 2018). Over the last 14 years, several large-scale genome-wide association studies have aimed to identify the genetic factors associated with CAD risk (Samani et al., 2007; Erdmann et al., 2009; Tregouet et al., 2009; Schunkert et al., 2011; Deloukas et al., 2012; Nikpay et al., 2015; Howson et al., 2017; Nelson et al., 2017; Webb et al., 2017; van der Harst and Verweij, 2018) and their functional consequences (Brænne et al., 2015; Kessler et al., 2015, 2016, 2017; Zhao et al., 2016; Aherrahrou et al., 2017; Vilne et al., 2017; Lempiäinen et al., 2018; Schunkert et al., 2018; Neiburga et al., 2021). It is currently a matter of intense debate, whether it might be time to implement genetic variations in the clinical routine CAD risk predictions (Inouye et al., 2018; Khera et al., 2018; Cecile et al., 2019; Gola et al., 2020; Lieb and Vasan, 2020).

At the same time, the contribution of environmental/lifestyle factors, in particular, dietary factors have remained less investigated (Khera et al., 2017; Dimovski et al., 2019). Diet is one of the modifiable factors for disease prevention and dietary recommendations have been formulated for decades to guide us toward changing our eating habits in favor of healthy choices. For example, the consumption of foods abundant in cholesterol and fats, such as (processed) red meats, have been associated with increased CAD risk and mortality (Bernstein et al., 2010; Micha et al., 2010). First evidence suggests that even for individuals at high genetic CAD risk and with pre-existing non-modifiable risk factors (age, sex, positive family history) adherence to a healthy lifestyle could be associated with an almost 50% lower relative risk of CAD (Khera et al., 2017; Dimovski et al., 2019), indicating that the inclusion of dietary factors can substantially improve CAD risk prediction, as compared to standard Cox models without these additional variables (Rigdon and Basu, 2019; Ho et al., 2020). With the advent of biosensors and wearable health technology connected to mobile apps, large-scale longitudinal food diaries and images of meals consumed are increasingly becoming available and are even being integrated within electronic health records (Verma et al., 2018; Dinh-Le et al., 2019; Moraes Lopes et al., 2020), whereas further advances in and rapidly decreasing costs of next generation sequencing generate increasing data volumes describing the human gut microbiome qualitative and quantitative composition and function (Eetemadi et al., 2020), thus providing valuable sources of data for integration in the context of personalized diet recommendation systems (Eetemadi et al., 2020), which could be further integrated into clinical decision support systems for improved CAD risk predictions. However, the current rise in obesity, type 2 diabetes (T2D) and CVD/CAD (Pallazola et al., 2019), indicates that the "one-size-fits-all" approach may not be efficient, due to significant variation in inter-individual responses to diet (Hughes et al., 2019), and that interactions between diet and other factors need to be considered (Qi, 2012).

Recently, the human gut microbiome has emerged as a potential and previously under-explored contributor to these variations (Bashiardes et al., 2018), as the composition

and function of this complex community of trillions of microorganisms (including bacteria, archaea, viruses, and microbial eukaryotes) (Garud and Pollard, 2020) is modulated by dietary components, e.g., the well-known beneficial impact of the so called Mediterranean diet (De Filippis et al., 2016). This impact is partly mediated through the metabolization and transformation of different nutrients by the gut microbiome, generating secondary metabolites, with changed retention time, bioactivity and different impact on health outcomes: being either protective, such as the short-chain fatty acids (SCFA) or promoting the disease development such as hydrogen sulfite or bile acids (Ni et al., 2015; Hughes et al., 2019; Eetemadi et al., 2020). Changes in the qualitative and quantitative composition of the gut microbiome have been increasingly linked to a number of diseases, including obesity (Turnbaugh et al., 2009; Maruvada et al., 2017; Miyamoto et al., 2019) and CVD/CAD (Koeth et al., 2013; Miele et al., 2015; Tang et al., 2017; Ascher and Reinhardt, 2018). Hence, efficient integration of dietary factors with the gut microbiome holds a great promise to revolutionize the way diseases are treated, through dietary recommendations and lifestyle changes or even the optimization of our gut microbiome, personalized to each individual and the desired health outcomes (Bashiardes et al., 2018; Eetemadi et al., 2020).

Taken together, the multifactorial and complex etiology of CAD (driven by both genetic and environmental/lifestyle factors), combined with the highly complex nature of meals (containing multiple ingredients and spices) and with the additional complexity and huge inter-individual variability of the gut microbiome (Marcos-Zambrano et al., 2021; Moreno-Indias et al., 2021), resulting in completely different responses to identical meals (Zeevi et al., 2015), urgently calling for more advanced problem-solving approaches. Moreover, with the development of high-throughput omic measurement platforms and digitalization of health records, the field is rapidly entering the Big Data era, as the volumes of these data are increasing exponentially (Stephens et al., 2015) and need to be transformed into valuable knowledge. In this regard, the recent re-emergence of advanced computational data-driven technologies such as Artificial Intelligence (AI)/Machine Learning (ML) approaches are opening intriguing perspectivesfor the integration of omics data (genetic variations, gut microbiome) with additional clinical (Reel et al., 2021) and environmental/lifestyle and the development personalized CAD diagnostics tools (Alizadehsani et al., 2019). AI/ML represent automated approaches that are adaptive and able to capture large and heterogeneous matrices of data extracting meaningful patterns and identifying both linear and non-linear relationships between these high-dimensional input variables and the outcomes (Alaa et al., 2019; Rigdon and Basu, 2019; Bodnar et al., 2020; Moraes Lopes et al., 2020). Especially, Deep Learning (DL) approaches, hold a great promise for future progress due to its capabilities to learn from input raw data, instead of using hand-crafted features that require domain expertise (Ching et al., 2018; Solares et al., 2020).

In this Mini Review, we explore, whether the inclusion of dietary factors and/or gut microbiome data in combination with the power of AI/ML could potentially improve CAD risk prediction. In particular, we consider: (1) the most used AI/ML approaches for CAD risk prediction; (2) the use cases of AI/ML approaches to model the content, choice and impact of dietary factors and how this could be used to predict CAD risk; (3) the use cases of AI/ML approaches to classify individuals by their gut microbiome composition into CAD cases vs. controls and to (4) model the diet-gut microbiome interactions and their impact on CAD risk (as illustrated in **Figure 1** and summarized in **Table 1**). (5) Finally, we provide an outlook for putting it all together into a smart clinical decision support system (DSS), considering the traditional risk factors in combination with individual's genetic variations, as well as dietary factors and gut microbiome and discuss the potential of AI/ML based methods vs. conventional approaches for risk predictions.

2. ARTIFICIAL INTELLIGENCE / MACHINE LEARNING APPROACHES FOR CAD RISK PREDICTION

Artificial Intelligence (AI) / Machine Learning (ML) has recently caught the interest of both academia and industry, and the different approaches have been explicitly reviewed elsewhere, e.g., Cao et al., 2018; Goecks et al., 2020. Hence, we only give a very brief overview, highlighting some of the most popular and widely used approaches and common terminology in the field, to prepare the reader for the sections to follow.

In general, AI/ML-based approaches can be considered as a set of methods that can effectively use large and complex data sets to extract meaningful patterns (i.e., "learn") in order to use this "knowledge" to make predictions on other data (Vilne et al., 2019) and improve with experience (Libbrecht and Noble, 2015). Moreover AI/ML can be performed either (1) in a unsupervised manner by exploring and detecting what types of labels best explain the data i.e., using unlabeled data; (2) in a supervised manner by classifying, predicting and explaining the data, requiring labels (Vilne et al., 2019; Eetemadi et al., 2020), or (3) in a semi-supervised manner, taking advantage of both unlabeled and labeled data, where only a subset of data is labeled (Libbrecht and Noble, 2015).

In particular, supervised learning has gained much attention recently (Reel et al., 2021), as it allows to define certain outputs that can be used for classification of patients, and will be the main focus of this Mini-Review. In this class, one of the most popular is the Random Forest (RF) approach (Breiman, 2001), which randomly selects a subset from the training data to construct an ensemble of Decision Tree (DT) predictors to aggregate the predictions, by this attempting to lower the variance and deal with the issue of overfitting. Decision Tree (DT) approach is also a commonly used classifier, splitting the input data into branch-like segments, according to a certain parameter (Goecks et al., 2020).

Another popular method in the field is the Support Vector Machine (SVM) classifier, representing a pattern classification technique, based on the idea of transforming the original data that is not linearly separable to a higher dimensional space and finding a hyperplane separating the data into classes, based on a priori defined criteria, with the aim to

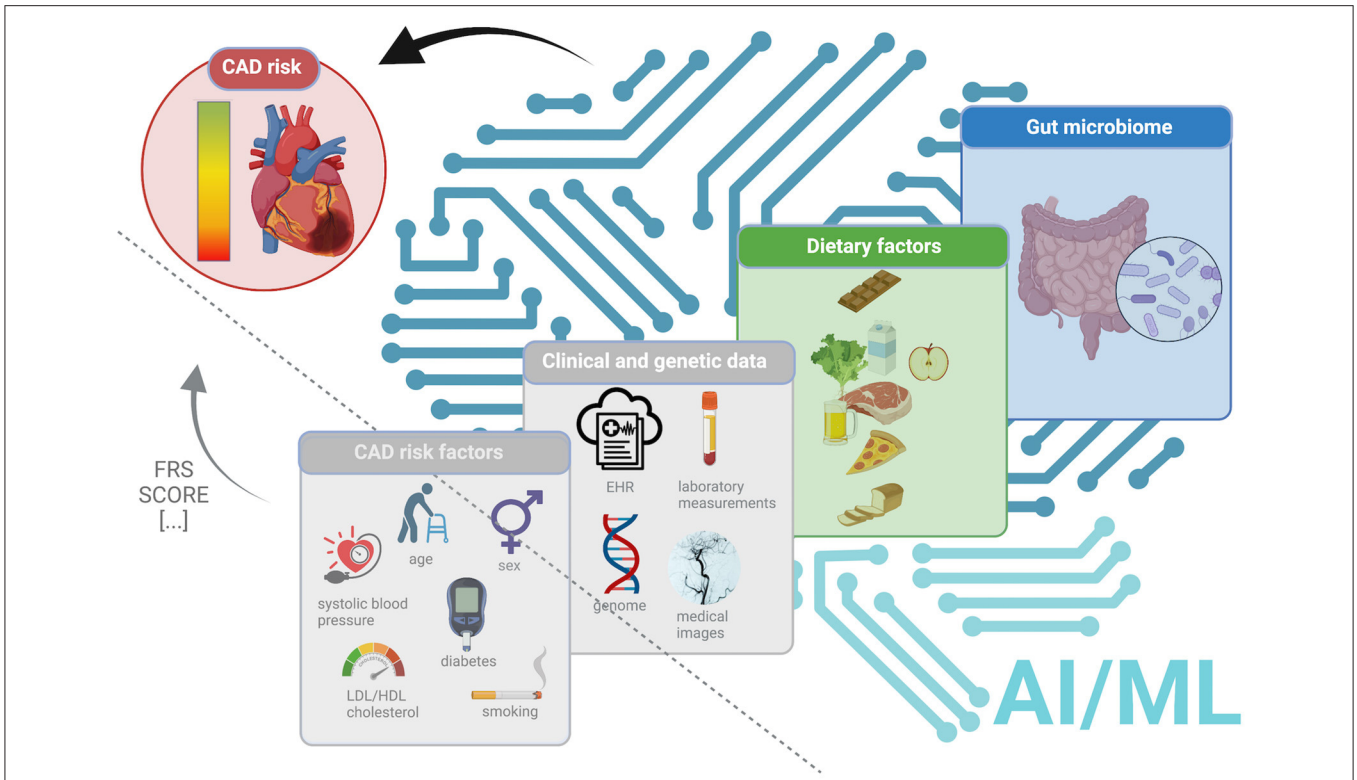


FIGURE 1 | An overview of the current status and future directions to improve CAD risk prediction. The left panel (triangle separated by a dashed line) demonstrates the current status (in gray, as not explicitly considered in this Mini-Review), demonstrates the current metrics such as FRS (Wilson et al., 1998) or SCORE (Conroy et al., 2003) using the traditional CAD risk factors (age, sex, diabetes, systolic blood pressure, LDL/HDL cholesterol, smoking). The right panel (separated by a dashed line) highlights the possible future directions to improve CAD risk prediction using AI/ML approaches and, alongside with clinical data and genetic variations (in gray, as not explicitly considered in this Mini-Review) dietary factors (in green) and/or gut microbiome (in blue).

overcome overfitting (Suykens et al., 2001). However, further improvements may be necessary when dealing with omics data (Han and Jiang, 2014).

Gradient Boosting (GB), such as Stochastic Gradient Boosting Regression (Friedman, 2001) is a technique that, similar to RF, constructs multiple decision trees by drawing a random samples from the data set (termed bagging). However, instead of constructing many parallel deep trees, it constructs multiple shallow trees (weak learners) and in a sequential manner (i.e., one after the other) so that the next tree improves upon the classification of previous trees in an additive manner. GB is known to perform best with fewer input variables of low dimensionality, whereas RF performs better with many input variables or high dimensionality (Hughes et al., 2019).

Finally, Artificial Neural Networks (ANN), and their extension, Deep Learning (DL), are graph computing models, which, at least to some extent, should mimic the functioning of the human brain, hence their computing units are called neurons and are interconnected for passing information to each other. Moreover, networks of neurons are additionally organized in layers. The first one is an input layer, receiving the training data. This is followed by several hidden layers. The last one is an output layer, which performs the actual prediction of the class (McCulloch and Pitts, 1990). ANN have been demonstrated to outperform other AI/ML approaches in many areas, especially (medical) image analyses (Eetemadi et al., 2020).

Performance of an AI/ML classifier is often expressed as the area under the curve (AUC), where a value of 0.5 indicates poor performance (equal to a random guess) while higher values (approaching 1) indicate better classification performance, allowing an easy comparison of the success of various implementations of AI/ML approaches (Bradley, 1997). However, considering that, in most cases, the users are more interested in positive outputs (i.e., people at high CAD risk), some other performance measures would need to be considered as well, such as the Jaccard index (J) or the F1-score, focusing on the fraction of true positives (Jiao and Du, 2016). Moreover, if the input data sets are imbalanced (i.e., many more controls than CAD patients in the training set), precision-recall (PR) curve should be considered along the ROC curve and additional performance measures, such as the balanced accuracy (BAcc) and the Matthew's Correlation Coefficient (MCC) considered (Jiao and Du, 2016). We refer the interested reader to Jiao and Du (2016) for more details. Moreover, if the input data is not normally distributed, maximum likelihood estimation (MLE) should be used to model this data and determine the model parameters for the evaluation metrics (Maximum-likelihood method, 2001).

TABLE 1 | A list of the case studies related to improved CAD risk prediction considered in this Mini-Review.

Category	Study purpose	AI/ML approaches(s) used	References
Dietary factors	To create an automated mobile vision food diary (Im2Calories), which can recognize the nutritional contents and calories of an individual's meal from its image.	Deep Learning (DL)/Convolutional Neural Network (CNN), adjusted for a mobile phone and images taken "in the wild"	Myers et al., 2015
	Use public food diaries of MyFitnessPal app users to study the food components of a successful ("below" the user defined "daily calories goal") or un-successful ("above") diet.	Support Vector Machine (SVM)	Weber and Achananuparp, 2016
	Use the data from the ThinkSlim app, to assess and predict individual's eating behavior in relation to their individual states (location, activity, emotions).	Decision Tree (DT), tailored to longitudinal real-time data	Spanakis et al., 2017
	Evaluate, how healthy Brazilian children and teens respond inter-individually to nutritional intervention of multivitamins and minerals, to develop recommendations for optimizing the levels of these supplements.	Elastic Net (EN) penalized regression model	Mathias et al., 2018
	Investigate whether the consideration of additional variables (in total 473 available variables, including dietary and nutritional information) could increase the accuracy of CVD risk prediction in 423,604 UK Biobank participants.	AutoPrognosis	Alaa et al., 2019
	Investigate whether the consideration of dietary information can improve CVD risk prediction.	Gradient Boosted Machines (GBMs) and Random Forests (RF), tailored to the analyses of survival data	Rigdon and Basu, 2019
Gut microbiome	Assess the potential of the (mainly gut) microbiome species-level abundances to be used for the classification of healthy vs. unhealthy (including obese and T2D patients) individuals.	Random Forests (RF), Support Vector Machine (SVM)	Pasolli et al., 2016
	Predict different traits, including cholesterol levels and BMI using the gut microbiome data in healthy participants.	Regularization of Learning Networks (RLN), Deep Neural Networks (DNNs), Gradient Boosting Trees (GBTs), Linear Models (LM)	Ira Shavitt, 2018
	Compare the composition of the gut microbiome in CAD patients vs. healthy controls.	Random Forests (RF)	Zhu et al., 2018
	Test, whether gut microbiome could be potentially used for diagnostic screening of CVD.	Random Forests (RF), Support Vector Machine (SVM), Decision Trees (DT), Elastic-Net (EN) and Neural Networks (NN)	Aryal et al., 2020
Dietary factors and gut microbiome	Identify associations between the gut microbiome composition and the concentration of butyrate, in response to dietary supplementation with resistant starch.	Random Forests (RF)	Venkataraman et al., 2016
	Investigate, the post-meal glucose levels in response to 46,898 standardized and real-life meals, in conjunction with the gut microbiome composition.	Stochastic Gradient Boosting Regression (SGBR)	Zeevi et al., 2015
	To validate the predictions by Zeevi et al. (2015) in an independent 327 cohort of individuals.	Stochastic Gradient Boosting Regression (SGBR)	Mendes-Soares et al., 2019
	Develop standardized protocols for the analyses of the diet-induced gut microbiome changes.		Spector et al., 2019
	Compare the post-meal glucose levels in response to the traditionally made sourdough-leavened whole-grain bread vs. industrially made white bread, in conjunction with the gut microbiome composition.	Stochastic Gradient Boosting Regression (SGBR)	Korem et al., 2017
	Use the gut microbiome data to predict changes of TMAO in healthy individuals after choline intake or screening population at high risks of CVD.	Random Forests (RF)	Lu et al., 2017

Considering that only a few studies so far have used dietary factors (Alaa et al., 2019; Rigdon and Basu, 2019) or gut microbiome (Zhu et al., 2018; Aryal et al., 2020), and no studies using both (the closest being Zeevi et al., 2015 related to blood glucose levels), in combination with AI/ML for CAD risk prediction, we also consider closely related research on dietary factors (in green), gut microbiome (in blue) and combinations of both (in turquoise) in other disease settings vs. healthy individuals.

The added value of AI/ML models in CAD diagnostics has been thoroughly reviewed before, examining 149 relevant studies between 1992 and 2019 (Alizadehsani et al., 2019). Most of this research focused on the usage of clinical data (symptom, examination and echo features), laboratory measurements and medical images (e.g., coronary computed tomography angiography, myocardial perfusion imaging or intravascular ultrasound) (Alizadehsani et al., 2019). The Authors observed that there were three approaches applied to almost all the datasets—ANN/DL, DT, and SVM—most probably due to their ease of use, low computational burden and encouraging performance (Alizadehsani et al., 2019). In particular, studies

with best performances (i.e., with a reported accuracy of >98%) used ANN and SVM as their classifiers, which may be due to the use of non-linear kernel functions (Alizadehsani et al., 2019). However, it was concluded that further investigation are needed to determine which approaches are most appropriate for a particular feature category (e.g., ejection fraction, regional wall motion abnormality, and valvular heart disease extracted from echo). Of note, however, neither of the data types highlighted above provide any information on the molecular bases of a disease, which could possibly yield a more timely and precise diagnosis, or even risk prediction, allowing for individually tailored treatments (Westerlund et al., 2021) or even prevention strategies, toward the goal of "positive health", resulting in a significantly improved life-span and quality (Movsisyan et al., 2020).

Genomic data have been used in combination with AI/ML for CAD risk prediction. In particular, (penalized) logistic regression, Naïve Bayes (NB), RF, SVM, and GB were compared vs. polygenic risk scores (PRS) on a data set of 7,736 CAD cases vs. 6,774 controls, testing the final models on an independent data set (527 CAD cases vs. 473 controls) (Gola et al., 2020). Interestingly, they found that PRS actually outperformed AI/ML-based approaches in predicting CAD status (AUC∼0.92 vs. ∼0.81 for NB and SVM and AUC∼0.75 for RF and GB). The Authors conclude that "there is no need to use a sledge-hammer to crack the nut", i.e., the assumption of linear additive effects influencing the risk of CAD seems sufficient. On the other hand, PRS may not be a suitable option, if the goal would be to predict the changes in CAD risk over time or the particular molecular basis driving the development and progression CAD (Westerlund et al., 2021).

This is were additional data layers such dietary factors and gut microbiome, as an integrator of this information (Bashiardes et al., 2018; Eetemadi et al., 2020), come in. However, only a few studies so far have used dietary factors (Alaa et al., 2019; Rigdon and Basu, 2019) or gut microbiome (Zhu et al., 2018; Aryal et al., 2020), and no studies using both (the closest being Zeevi et al., 2015 related to blood glucose levels), in combination with AI/ML for CAD risk prediction. We further discuss these few studies and also consider closely related research on dietary factors, gut microbiome and combinations of both in other disease settings vs. healthy individuals.

3. PERFORMING DIET-BASED CAD RISK PREDICTION USING AI/ML

Dietary information is mainly collected *via* questionnaires, either through self-reporting or by a trained interviewer. For self-reporting, a food frequency questionnaire and dietary recall can be used, where participants report their meal intake either every 24 h or over a longer period through a checklist of food items (Eetemadi et al., 2020). At the same time, fitness apps are gaining increased popularity, as food logging can be performed during its consumption or even by capturing an image of the meal, thus the bias related to individual's memory can be reduced (Weber and Achananuparp, 2016; Verma et al., 2018; Eetemadi et al., 2020). Clearly, such food tracking would be of utmost importance for a more efficient management of patients with obesity, T2D and CVD/CAD (Bernstein et al., 2010; Pallazola et al., 2019), when successfully coupled with an effective coaching to modulate it toward healthy food choices (Spanakis et al., 2017). The AI/ML approaches can be leveraged for such purposes (Verma et al., 2018).

In this regard, Myers et al. (2015) created a Google app, Im2Calories, to predict the nutritional contents and calories of individual's meal from its image, using a Convolutional Neural Network/DL-based classifier, which was modified to run on a mobile phone analyzing images taken by users, demonstrating promising first results in this direction.

Weber and Achananuparp (2016) used public food diaries of >4,000 MyFitnessPal users to train a SVM classifier to distinguish between a successful ("below" a user specified "daily calories goal") vs. un-successful ("above" the goal) diet and analyzed the different dietary factors influencing these two outcomes. It was observed that "oil", "butter", "mcdonalds", "dessert" or "pork" vs. "poultry" were related to being "above" the calories goal. Moreover, there were less food logging on the weekend and the users were most likely to be "above" the calories goal (Weber and Achananuparp, 2016).

Spanakis et al. (2017) made use of data collected from the fitness app ThinkSlim, to link the individual states (like location, activity, emotions- cheerful, relaxed vs. sad, bored, stressed, angry, worried) of healthy-weight vs. overweight individuals to their dietary choices or wishes, using a Decision Tree (DT)-based classifier, modified to use longitudinal real-time data. They derived several groups of individuals with similar eating behavior and used this information to warn the participants before the individual states that may lead to unhealthy eating behavior (Spanakis et al., 2017).

Mathias et al. (2018) conducted a six-week study to evaluate, how 136 healthy Brazilian children and teens (9–13 years old) responded to multivitamins and minerals, to develop recommendations for optimizing their levels, based on several clinical, anthropometric and food intake parameters. These data were then used to predict each individual's response to the intervention, based on these measures using an Elastic Net penalized regression model.

However, none of the above mentioned studies were directly related to CAD risk prediction. There have been only a few studies considering the dietary factors for CAD risk prediction, so far. Alaa et al. (2019) analyzed 423,604 UK Biobank participants without CVD at baseline with the aim to predict their future disease risk. They investigated, whether AI/ML-based approaches could possibly improve disease risk prediction, as compared to conventional approaches (such as FRS) and whether considering additional information (i.e., 473 variables, including dietary information) could increase the accuracy of their predictions. They used AutoPrognosis, which allows to automatically select and tune the best possible AI/ML approaches, by comprising different imputation strategies, feature selection and processing, as well as classification and calibration approaches. They observed and improvement in comparison to (AUC∼0.77 vs. ∼0.72 for FRS) conventional approaches (Alaa et al., 2019).

Rigdon and Basu (2019) performed a retrospective study using AI/ML exploring whether considering randomly sampled sparse nutrition data could possibly improve CVD mortality risk

prediction. They made use of NHANES interview data collected from 1999 to 2011 linked to the National Death Index (NDI) in the US, selecting 29,390 participants as their training set and further 12,600 participants as their test set. Similarly to Alaa et al. (2019), they aimed at testing whether AI/ML-based approaches vs. standard (Cox) models and considering additional predictor variables (dietary information) could possibly improve CVD mortality risk prediction. They applied two DT-based AI/ML approaches the Gradient Boosted Machines (GBM) (Chen et al., 2013) and RF (Ishwaran et al., 2008), tailored to the analyses of survival data to demonstrated that the inclusion of dietary information significantly improved risk prediction, as compared to the standard models and when including only the traditional risk factors. In particular, they found that a standard Cox model without dietary factors overestimated the CVD mortality risk nearly two-fold, whereas AI/ML models in combination with these additional data substantially improved their predictions (AUC∼0.87 vs. ∼0.93).

4. PERFORMING GUT MICROBIOME-BASED CAD RISK PREDICTION USING AI/ML

In addition to genetic and environmental/life-style factors, gut microbiota has emerged as a additional factor influencing the CAD risk (Aryal et al., 2020). Clearly, researchers have asked, whether gut microbiome profiling combined with AI/ML approaches could be used for improved CAD/CVD risk prediction. In the last 10 years, a number of studies have demonstrated that there is a possible relationship between the gut microbiome composition, such as changes in the abundance of *Bacteroidetes, Firmicutes, Lactobacillus, Streptococcus, Bifidobacterium, Roseburia*, or *Escherichia* spp. and the development of several diseases, including obesity (Turnbaugh et al., 2009; Maruvada et al., 2017; Miyamoto et al., 2019) hypertension (Karbach et al., 2016), and CVD (Karlsson et al., 2012; Koeth et al., 2013; Miele et al., 2015; Kelly et al., 2016; Tang et al., 2017; Ascher and Reinhardt, 2018).

Several studies have used AI/ML approaches to classify test subjects into groups (such as healthy vs. disease) based on microbiome data. In most studies, relative abundances of microbiome taxa are used as features, obtained either by amplicon sequencing of the 16S rRNA phylogenetic marker gene or by shotgun metagenomic sequencing (Hughes et al., 2019). As the costs of shotgun metagenomic sequencing decrease, the functional profiles derived from metagenome sequences can be expected to increasingly be used as input features with AI/ML approaches (Eetemadi et al., 2020; Sanchez-Rodriguez et al., 2020). Pasolli et al. (2016) utilized 2,424 shotgun metagenomic samples from eight studies to assess the potential of the (mainly gut) microbiome species-level abundances to be used in order to differentiate healthy vs. unhealthy (including obese and T2D) individuals and compare the prediction accuracy of RF vs. SVM approaches. Interestingly, for T2D and obesity, the models demonstrated lower discrimination ability as compared, for example, to liver cirrhosis (AUC of 0.74 and 0.65 vs. 0.94, respectively), suggesting less significant changes in microbiome composition related to T2D and obesity. Comparing the accuracy of RF vs. SVM, in all cases, RF demonstrated similar or even better results than SVM (for T2D: AUC∼0.74 vs. ∼0.66, respectively).

However, although the approach of using gut microbiome in combination with AI/ML approaches for disease risk prediction is not novel, it has not been widely applied for CAD, yet (Aryal et al., 2020). Zhu et al. (2018) compared the composition of the gut microbiome between 70 CAD patients vs. 98 healthy controls and used RF to potentially differentiate these two groups of individuals, achieving and AUC of 0.67. In addition, the gut microbiome of CAD patients displayed decreased diversity and richness, with decreased abundances of *Faecalibacterium, Roseburia*, and *Eubacterium rectale* (the butyrate producers) and increased abundances of *Escherichia-Shigella* and *Enterococcus*. More recently, in order to test whether gut microbiome could be potentially used for diagnostic screening of CVD, Aryal et al. (2020) applied five different AI/ML approaches (RF, SVM, DT, Elastic-Net and Neural Networks) to the gut microbiome relative abundances of 478 CVD patients vs. 473 healthy controls, collected as part of the American Gut Project [https://microsetta.ucsd.edu/american-gut-project/] and profiled using fecal 16S ribosomal RNA sequencing. However, when using 39 differential bacterial taxa as features, the best AUC this study could achieve was AUC∼0.58 (with Neural Networks), followed by Elastic-Net (AUC∼0.57), SVM (AUC∼0.55) and DT (AUC∼0.51). Interestingly, the performance of RF significantly improved (AUC∼0.65) when trained with the top 500 high-variance OTU features, instead of taxonomic features, whereas the AUC of Neural Networks dropped (AUC∼0.48). Furthermore, highly contributing OTU features (HCOFs) were selected based on their variable importance (0–100, where 0: no contribution to the model and 100: max contribution to the model) to further reduce the dimensionality of the OTU feature space. The top 100 HCOFs with the highest scores were selected for training the RF model. As a result, the RF models trained with the top 20 and top 25 HCOFs achieved further improved performance (AUC∼0.70).

5. CONSIDERING DIET-GUT MICROBIOME INTERACTIONS FOR CAD RISK PREDICTION USING AI/ML

It can be assumed that particular diets, such as those high in fats and/or sugars might lead to variations in the gut microbiome composition and changes in its functional capacity that potentially might facilitate the development of diseases, including metabolic disorders such as obesity, insulin resistance and atherosclerosis/CVD (Sanchez-Rodriguez et al., 2020). Despite the close link between our diet and gut microbiome, the number of studies collecting and analyzing both types of data is sparse and either not considering the full spectrum of dietary factors (Lu et al., 2017) or not directly addressing the prediction of CVD/CAD risk (Zeevi et al., 2015; Venkataraman et al., 2016; Spector et al., 2019).

Zeevi et al. (2015) used a the GB approach to investigate whether individuals' gut microbiome profiles in combination with several other sources of information (blood parameters, anthropometrics, self-reported lifestyle behaviors and physical activity) could predict glucose levels in response to standardized and real-life meals in a cohort of 800 overweight or obese non-diabetic individuals, observing high inter-individual variability, even in response to identical meals, suggesting that dietary recommendations need to be personalized. Later, these predictions were validated by Mendes-Soares et al. (2019) in an independent cohort of 327 individuals and by Korem et al. (2017), when focusing on the consumption of sourdough-leavened whole-grain bread vs. industrially made white bread, also using the GB approach. In the latter case, the relative abundances of *Coprobacter fastidiosus* and *Lachnospiraceae bacterium* were among the most informative features.

Venkataraman et al. (2016) used RF to predict whether the gut microbiome composition of individuals can predict their response to dietary supplementation with resistant starch, as measured using fecal butyrate concentrations. This study could identify three different response groups—enhanced, high and low, and could attribute these differences to the increase of starch-degrading bacteria *Bifidobacterium adolescentis* and *Ruminococcus bromii* in the enhanced and high, but not in the low fecal butyrate concentration group.

Spector et al. (2019), as part of the PREDICT study [http://www.tim-spector.co.uk/predict/], is actively working toward personalized nutrition tools by systematically analyzing the diet-induced gut microbiome changes using AI/ML approaches in order to stratify individual responses to dietary interventions based on the individual's gut microbiome and develop standardized protocols for the purpose. Among others, this study has demonstrated that shotgun metagenomic sequencing may be more accurate than 16S rRNA amplicon sequencing, as it allows also capturing individual-specific strain-level features, thus improving the stratification.

In the context of CVD/CAD risk prediction, most studies have focused on the circulating levels of the diet- and gut microbiota-dependent metabolite trimethylamine-N-oxide (TMAO) (Trøseid et al., 2020). Lu et al. (2017) used RF and the gut microbiome data to predict changes of TMAO levels after choline intake, as a potential approach for screening population at high risk of CVD and identified the beta (inter-individual) diversity of the gut microbiome as a significant predictor (AUC of 0.86) of increased vs. decreased plasma TMAO level.

6. DISCUSSION

Timely and precise identification of people at high risk of CAD is of utmost importance for the development of personalized treatment strategies (Alaa et al., 2019; Westerlund et al., 2021), as such persons may need more aggressive health promotion strategies, especially the modifiable CAD risk factors could be effectively reduced or even eliminated in this way (Movsisyan et al., 2020). Although, numerous algorithms for CAD risk prediction have been developed over the years and several have also entered the clinical routine (FRS Wilson et al., 1998, SCORE Conroy et al., 2003), these are typically based on a limited number of traditional CAD risk factors (age, sex, diabetes, LDL and HDL cholesterol, smoking, systolic blood pressure) and are not suitable across all patient groups (Alaa et al., 2019) and do not take into account the fact that by modifying the person's environment/lifestyle the disease risk could be reduced over time (Westerlund et al., 2021).

The added value of AI/ML approaches in CAD diagnostics has been explored before, however, so far, most of this research has focused on the usage of clinical data and medical images (Alizadehsani et al., 2019), thus providing no information on the molecular bases of a disease (Westerlund et al., 2021). A small number of studies has used genomic data have been used in combination with AI/ML for CAD risk prediction (Gola et al., 2020). However, AI/ML approaches underperformed in comparison to a simple PRS, assuming linear additive effects (Gola et al., 2020). This is were additional data layers such dietary factors and gut microbiome, as an integrator of this information (Bashiardes et al., 2018; Eetemadi et al., 2020), come in. However, only a few studies so far have used dietary factors (Alaa et al., 2019; Rigdon and Basu, 2019) or gut microbiome (Zhu et al., 2018; Aryal et al., 2020), and no studies using both [the closest being (Zeevi et al., 2015) related to blood glucose levels], in combination with AI/ML for CAD risk prediction. We further discuss these few studies and also consider closely related research on dietary factors, gut micorbiome and combinations of both in other disease settings vs. healthy individuals.

With the advent of wearable biosensors connected to mobile applications, large-scale longitudinal food diaries and images of meals consumed will become increasingly available providing a valuable data source for such investigations (Munos et al., 2016). The future vision for personalized nutrition has led to great interest for advancements in the diagnostics and decision support systems (DSS) that would allow continuous assessment of individual's dietary features, in conjunction with gut microbiome composition and additional information, such as access to the electronic health record (EHR) and lifestyle and environment information, physical activity from the biosensors and wearable health technology. All of it would aid in forming tailored recommendations such as choosing an optimal meal for lowering post-meal glucose levels (as shown by Zeevi et al., 2015) in patients with T2D. Although, the recent re-emergence of AI/ML approaches is opening intriguing perspectives in this direction, it must be remembered that these data-driven technologies and their predictions strongly depend on the quantity and quality of the input data. In this regard, several limitations to the current food intake and composition databases have been observed. Apparently, these databases currently contain only 0.5% of the known nutritional compounds (Eetemadi et al., 2020). Another issue is the data standardization, which is challenging as complex dietary patterns need to be captured in an organized manner, translating chemicals constituents of the food into the intake of energy and nutrients (Verma et al., 2018). Currently, the most widely applied methods of food intake monitoring include the food diaries, which make it difficult to convert the food descriptions into the

energy. Additional challenges arise when the food is collected from different sources, i.e., individual and/or hospital-based sources.The missing data problem could be partly addressed through improved data imputation techniques, which should be complemented by improved food intake monitoring and data collection methods, creating integrated databases with defined standard formats for annotation and classification, considering the FAIR (Findability, Accessibility, Interoperability, and Reuse) data principles [https://www.go-fair.org/fair-principles/]. Initiatives such as the EuroDISH project [https://www.eurofir.org/our-resources/past-projects/eurodish/] are already working in this direction.

There are a number of challenges and limitations related to the application of AI/ML approaches for microbiome studies, as thoroughly and systematically summarized in several literature reviews (Marcos-Zambrano et al., 2021; Moreno-Indias et al., 2021) by the members of the COST Action CA18131 "ML4Microbiome" (https://www.cost.eu/actions/CA18131/), bringing together AI/ML experts and microbiome researchers. Overall, similar to other high-throughput studies, one of the main limitations in current research has been the usage of inappropriate study design, including small datasets and lack of additional data to estimate confounding effects, especially considering the well-known huge variations in microbiome composition across individuals and body sites and their strong dependence on the environment/lifestyle factors such as geographic location, diet and medications (Marcos-Zambrano et al., 2021; Moreno-Indias et al., 2021). In order to identify generalized responses, a much larger number of individuals spanning a range of microbiome types and a careful adjustment for potential confounding effects would be required (Johnson et al., 2020). In addition, a number of data processing/statistical and AI/ML challenges have been observed, such as the selection of appropriate normalization methods to address the variability in raw read counts, inappropriate distributional assumptions considering the data sparsity, compositional nature and complex and hierarchical dependency structures, the choice of suitable feature selection approaches, i.e., requiring customized analytical approaches (Eetemadi et al., 2020; Marcos-Zambrano et al., 2021; Moreno-Indias et al., 2021). In fact, successful examples often present a combination of different statistical approaches, specifically tailored to the characteristics of different data types (Marcos-Zambrano et al., 2021; Moreno-Indias et al., 2021). On top of that the dependence on the reference databases is a well-known major limitation of the sequence alignment-based approaches, used to assign taxa in sequencing studies (Chaudhary et al., 2015; Vilne et al., 2019), resulting in large numbers of uncharacterized microbes (the "microbial dark matter") (Marcos-Zambrano et al., 2021). Finally, the field of high-throughput sequencing overall needs a rigorous assessment, benchmarking and standardization of approaches and tools (Vilne et al., 2019), to allow cross-study comparisons and modeling (Marcos-Zambrano et al., 2021). Currently, the integration of microbiome data across several studies is difficult due to the above mentioned factors, as well as the differences in sample collection, storage and processing protocols in the wet-lab, which may introduce biases (Eetemadi et al., 2020). Hence, all findings should be validated, e.g., using quantitative PCR (Jian et al., 2020). Finally, also for these data, the above mentioned FAIR data principles should be widely incorporated [https://www.go-fair.org/fair-principles/] to facilitate such efforts. For more details we refer the reader to Marcos-Zambrano et al. (2021), Moreno-Indias et al. (2021). However, we note that these current limitations related to microbiome studies are posing additional challenges for CAD risk prediction.

Moreover, several studies have shown that the inter-individual responses to dietary factors may differ, mostly due to the differences in the gut microbiome composition (Zeevi et al., 2015; Korem et al., 2017; Mendes-Soares et al., 2019). However, especially in the context of multi-factorial diseases, such as CAD, the differences in individual genetic predisposition (Nikpay et al., 2015; Nelson et al., 2017) and its down-stream implications (Brænne et al., 2015; Vilne et al., 2017; Lempiäinen et al., 2018; Vilne and Schunkert, 2018) in addition to variations in other (besides diet) environmental and lifestyle factors such as physical activity, stress and sleep may play and important role in these responses (Khera et al., 2017). Hence, emphasizing the need to collect a wide variety of measures in large populations that would allow for stratification of patients in sub-groups and perform longitudinal sampling to also capture the dynamics of these responses. Endeavors to standardize the study protocols have already started (Spector et al., 2019).

On the other hand, benchmark investigations have demonstrated that, whether a particular AI/ML approach would actually improve the predictions compared to conventional approaches, may depended on the specific dataset at hand (Westerlund et al., 2021). For example, in microbiome studies, DL approaches have been demonstrated to underperform in comparison to GB, possibly due to the potentially large variability in the relative importance of different input features. To overcome this limitation, Ira Shavitt (2018) have proposed an approach called Regularization of Learning Networks (RLN). They used it to predict a number of traits related to disease risk, such as cholesterol levels and body mass index (BMI) from the gut microbiome data of 2,574 healthy individuals. They evaluated four different AI/ML approaches [RLN, GB, DL, and Linear Models (LM)] and, although, GB still outperformed the other three, RLN performed significantly better than DL (15% vs. ca. 2% less explained variance than GB on average).

Currently, the number of studies investigating the potential of gut-microbiome in combination with AI/ML to predict CVD risk is limited (Aryal et al., 2020) an so is the prediction power of these models, with a max AUC of 0.70, when training a RF model with the top 25 highest contributing OTU features (Aryal et al., 2020). However, it must be noted that the authors did not normalize the OTU data across all the samples to test the option of classifying new samples without the need for repeated processing (Aryal et al., 2020). In addition, this study addressed the prediction of CVD, which, as the authors themselves recognize (Aryal et al., 2020) includes a range of conditions (from hypertension and atherosclerosis to CAD). Hence, these predictions may improve when stratifying CVD patients into specific disease sub-types. Moreover, another interesting observation from this study is the fact that bacterial taxonomic features achieved a lower (AUC∼0.58) AUC, in comparison to high-variance OTU features (AUC∼0.65), and especially when further reducing the

dimensionality of the feature space by pre-selecting the top 25 highest contributing OTU features (AUC∼0.70) (Aryal et al., 2020). From the usage in clinical routine, focusing on a small number of highly contributing OTUs may be indeed more practical, analogous to the handful of traditional CAD risk factors, however, we will need further studies to arrive replicate these findings and arrive at these gut microbiome biomarkers. Furthermore, their mechanistic implications in CVD need to be further investigated (Aryal et al., 2020). Gut microbiota as the only type of data used for diagnostic classification of non-CVD vs. CVD may not be sufficient Especially, considering that gut microbiome can be influenced by other features such as diet and medications, hence these data should be always collected in parallel.

Clearly, AI/ML (especially DL approaches due to their capabilities to learn from input raw data, instead of using hand-crafted features that require domain expertise, Ching et al., 2018) in combination with timely access to numerous, potentially relevant, data sources [e.g., gut microbiome and genetic data, in addition to the current 7 CVD metrics smoking, physical activity, body mass index, blood pressure, cholesterol, glucose and dietary factors (Angell et al., 2020), combined with longitudinal clinical data from electronic health records (Matlock et al., 2013; Reynolds et al., 2017)] also holds a great promise for the improvements of public health surveillance systems, formulation of policies by forecasting the impact of a factor or intervention on the burden of disease and the cost of care, and to propose recommendations to stakeholders (medical institutions, public health authorities, scientific communities) enabling public health action and measure progress with the aim to reduce the huge socio-economic burden of CVD/CAD and increase healthy life expectancy in future (Angell et al., 2020; Roger et al., 2020). The same is true for the implementation of personalized decision support system (DSS) for CAD risk prediction and patient management that would be a great support for clinicians in health care.

However, despite the rapid development of several technologies and advancements in Big Data analytics, the implementation of such systems that would integrate comprehensive health and related data (such as genetic variations, dietary factors, gut microbiome) to provide either generalized recommendations for public health surveillance and policy makers or individual recommendations for the routine clinical practice, still poses a number of challenges that will need to be overcame first, in order to move toward their implementation and usability in practice. Overall, such systems will need to deal with heterogeneous datasets and we will require a rigorous assessment, benchmarking and standardization of AI/ML-based CVD/CAD risk prediction models, ensuring model availability and extensive multiple external validations and calibration across different disease outcomes, populations (in men and women separately) and geographical regions *via* head-to-head comparisons across different studies and model impact and performance generalizability assessment and to identify potential sources of heterogeneity (Damen et al., 2016; Marcos-Zambrano et al., 2021; Westerlund et al., 2021).

Moreover, In April 2016, the European Union adopted new rules regarding the use of personal information, the General Data Protection Regulation, which imposes additional legal and privacy constraints when analyzing sensitive health data, hence model training will need to be accomplished within a differential privacy framework without sharing the raw data (e.g., federated learning) and considering other rules regarding the use of personal information as input for decision-making approaches, such as the 'right to an explanation', meaning that when using AI/ML, we must be able to explain how a decision was reached, especially if the ground-truth is unknown (Ching et al., 2018). This calls for the AI/ML models to be human-interpretable, reliable and explainable to aid the formulation guidelines or personalized advice on treatment strategy, or even prevention, plans (Ching et al., 2018; Westerlund et al., 2021).

In any case, the AI/ML tools will not be a replacement for the human experts, who are still an integral part of the knowledge discovery process, hence, managing huge amounts of health data will need to become an integral part of future medical, policy making and research activity, across sub-disciplines (Moreira et al., 2019).

AUTHOR CONTRIBUTIONS

BV wrote the manuscript. JĶ, IS, IL, AK, and OV participated in revising and editing the manuscript. All authors have read and approved the final version of the manuscript.

FUNDING

This research was funded by the Latvian Council of Science within the project Gut microbiome composition and diversity among health and lifestyle induced dietary regimen, project No. lzp-2018/2-0266.

REFERENCES

Aherrahrou, R., Aherrahrou, Z., Schunkert, H., and Erdmann, J. (2017). Coronary artery disease associated gene phactr1 modulates severity of vascular calcification *in vitro*. *Biochem. Biophys. Res. Commun*. 491, 396–402. doi: 10.1016/j.bbrc.2017.07.090

Alaa, A. M., Bolton, T., Di Angelantonio, E., Rudd, J. H. F., and van der Schaar, M. (2019). Cardiovascular disease risk prediction using automated machine learning: a prospective study of 423,604 UK biobank participants. *PLoS ONE* 14, e0213653. doi: 10.1371/journal.pone.0213653

Alizadehsani, R., Abdar, M., Roshanzamir, M., Khosravi, A., Kebria, P. M., Khozeimeh, F., et al. (2019). Machine learning-based coronary artery disease diagnosis: a comprehensive review. *Comput. Biol. Med*. 111, 103346. doi: 10.1016/j.compbiomed.2019.103346

Angell, S. Y., McConnell, M. V., Anderson, C. A., Bibbins-Domingo, K., Boyle, D. S., Capewell, S., et al. (2020). The American heart association 2030 impact goal:

a presidential advisory from the american heart association. *Circulation* 141, e120–e138. doi: 10.1161/CIR.0000000000000758

Aryal, S., Alimadadi, A., Manandhar, I., Joe, B., and Cheng, X. (2020). Machine learning strategy for gut microbiome-based diagnostic screening of cardiovascular disease. *Hypertension* 76, 1555–1562. doi: 10.1161/HYPERTENSIONAHA.120.15885

Ascher, S., and Reinhardt, C. (2018). The gut microbiota: an emerging risk factor for cardiovascular and cerebrovascular disease. *Eur. J. Immunol.* 48, 564–575. doi: 10.1002/eji.201646879

Bashiardes, S., Godneva, A., Elinav, E., and Segal, E. (2018). Towards utilization of the human genome and microbiome for personalized nutrition. *Curr. Opin. Biotechnol.* 51, 57–63. doi: 10.1016/j.copbio.2017.11.013

Bernstein, A. M., Sun, Q., Hu, F. B., Stampfer, M. J., Manson, J. E., and Willett, W. C. (2010). Major dietary protein sources and risk of coronary heart disease in women. *Circulation* 122, 876–883. doi: 10.1161/CIRCULATIONAHA.109.915165

Bodnar, L. M., Cartus, A. R., Kirkpatrick, S. I., Himes, K. P., Kennedy, E. H., Simhan, H. N., et al. (2020). Machine learning as a strategy to account for dietary synergy: an illustration based on dietary intake and adverse pregnancy outcomes. *Am. J. Clin. Nutr.* 111, 1235–1243. doi: 10.1093/ajcn/nqaa027

Brænne, I., Civelek, M., Vilne, B., Di Narzo, A., Johnson, A. D., Zhao, Y., et al. (2015). Prediction of causal candidate genes in coronary artery disease loci. *Arterioscler. Thromb. Vasc. Biol.* 35, 2207–2217. doi: 10.1161/ATVBAHA.115.306108

Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recogn.* 30, 1145–1159. doi: 10.1016/S0031-3203(96)00142-2

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Cao, C., Liu, F., Tan, H., Song, D., Shu, W., Li, W., et al. (2018). Deep learning and its applications in biomedicine. *Genomics Proteomics Bioinform.* 16, 17–32. doi: 10.1016/j.gpb.2017.07.003

Cecile, A., Janssens, J. W., and Joyner, M. J. (2019). Polygenic risk scores that predict common diseases using millions of single nucleotide polymorphisms: is more, better? *Clin. Chem.* 65, 609–611. doi: 10.1373/clinchem.2018.296103

Chaudhary, N., Sharma, A. K., Agarwal, P., Gupta, A., and Sharma, V. K. (2015). 16s classifier: a tool for fast and accurate taxonomic classification of 16s rrna hypervariable regions in metagenomic datasets. *PLoS ONE* 10, e0116106. doi: 10.1371/journal.pone.0116106

Chen, Y., Jia, Z., Mercola, D., and Xie, X. (2013). A gradient boosting algorithm for survival analysis via direct optimization of concordance index. *Comput. Math. Methods Med.* 2013, 873595. doi: 10.1155/2013/873595

Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., et al. (2018). Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface* 15, 20170387. doi: 10.1098/rsif.2017.0387

Conroy, R. M., Pyorala, K., Fitzgerald, A. P., Sans, S., Menotti, A., De Backer, G., et al. (2003). Estimation of ten-year risk of fatal cardiovascular disease in Europe: the score project. *Eur. Heart J.* 24, 987–1003. doi: 10.1016/s0195-668x(03)00114-3

Damen, J. A. A. G., Hooft, L., Schuit, E., Debray, T. P. A., Collins, G. S., Tzoulaki, I., et al. (2016). Prediction models for cardiovascular disease risk in the general population: systematic review. *BMJ* 353, i2416. doi: 10.1136/bmj.i2416

Davey Smith, G., Ebrahim, S., Lewis, S., Hansell, A. L., Palmer, L. J., and Burton, P. R. (2005). Genetic epidemiology and public health: hope, hype, and future prospects. *Lancet* 366, 1484–1498. doi: 10.1016/S0140-6736(05)67601-5

De Filippis, F., Pellegrini, N., Vannini, L., Jeffery, I. B., La Storia, A., Laghi, L., et al. (2016). High-level adherence to a mediterranean diet beneficially impacts the gut microbiota and associated metabolome. *Gut* 65, 1812–1821. doi: 10.1136/gutjnl-2015-309957

Deloukas, P., Kanoni, S., Willenborg, C., Farrall, M., Assimes, T. L., Thompson, J. R., et al. (2012). Large-scale association analysis identifies new risk loci for coronary artery disease. *Nat. Genet.* 45, 25–33. doi: 10.1038/ng.2480

Dimovski, K., Orho-Melander, M., and Drake, I. (2019). A favorable lifestyle lowers the risk of coronary artery disease consistently across strata of non-modifiable risk factors in a population-based cohort. *BMC Public Health* 19, 1575. doi: 10.1186/s12889-019-7948-x

Dinh-Le, C., Chuang, R., Chokshi, S., and Mann, D. (2019). Wearable health technology and electronic health record integration: scoping review and future directions. *JMIR mHealth uHealth* 7, e12861. doi: 10.2196/12861

Eetemadi, A., Rai, N., Pereira, B. M. P., Kim, M., Schmitz, H., and Tagkopoulos, I. (2020). The computational diet: a review of computational methods across diet, microbiome, and health. *Front. Microbiol.* 11, 393. doi: 10.3389/fmicb.2020.00393

Erdmann, J., Grosshennig, A., Braund, P. S., Konig, I. R., Hengstenberg, C., Hall, A. S., et al. (2009). New susceptibility locus for coronary artery disease on chromosome 3q22.3. *Nat. Genet.* 41, 280–282. doi: 10.1038/ng.307

Erdmann, J., Kessler, T., Munoz Venegas, L., and Schunkert, H. (2018). A decade of genome-wide association studies for coronary artery disease: the challenges ahead. *Cardiovasc. Res.* 114, 1241–1257. doi: 10.1093/cvr/cvy084

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Ann. Statist.* 29, 1189–1232. doi: 10.1214/aos/1013203451

Garud, N. R., and Pollard, K. S. (2020). Population genetics in the human microbiome. *Trends Genet.* 36, 53–67. doi: 10.1016/j.tig.2019.10.010

Goecks, J., Jalili, V., Heiser, L. M., and Gray, J. W. (2020). How machine learning will transform biomedicine. *Cell* 181, 92–101. doi: 10.1016/j.cell.2020.03.022

Gola, D., Erdmann, J., Muller-Myhsok, B., Schunkert, H., and Konig, I. R. (2020). Polygenic risk scores outperform machine learning methods in predicting coronary artery disease status. *Genet. Epidemiol.* 44, 125–138. doi: 10.1002/gepi.22279

Han, H., and Jiang, X. (2014). Overcome support vector machine diagnosis overfitting. *Cancer Inform.* 13(Suppl 1), 145–158. doi: 10.4137/CIN.S13875

Ho, F. K., Gray, S. R., Welsh, P., Petermann-Rocha, F., Foster, H., Waddell, H., et al. (2020). Associations of fat and carbohydrate intake with cardiovascular disease and mortality: prospective cohort study of uk biobank participants. *BMJ* 368, m688. doi: 10.1136/bmj.m688

Howson, J. M. M., Zhao, W., Barnes, D. R., Ho, W.-K., Young, R., Paul, D. S., et al. (2017). Fifteen new risk loci for coronary artery disease highlight arterial-wall-specific mechanisms. *Nat. Genet.* 49, 1113–1119. doi: 10.1038/ng.3874

Hughes, R. L., Marco, M. L., Hughes, J. P., Keim, N. L., and Kable, M. E. (2019). The role of the gut microbiome in predicting response to diet and the development of precision nutrition models-part I: overview of current methods. *Adv. Nutr.* 10, 953–978. doi: 10.1093/advances/nmz022

Inouye, M., Abraham, G., Nelson, C. P., Wood, A. M., Sweeting, M. J., Dudbridge, F., et al. (2018). Genomic risk prediction of coronary artery disease in 480,000 adults: Implications for primary prevention. *J. Am. Coll. Cardiol.* 72, 1883–1893. doi: 10.1016/j.jacc.2018.07.079

Ira Shavitt, E. S. (2018). "Regularization learning networks: deep learning for tabular datasets," in *32nd Conference on Neural Information Processing Systems (NeurIPS 2018)* (Montreal, QC).

Ishwaran, Udaya, B., Kogalur, E. H. B., and Lauer, M. S. (2008). Random survival forests. *Ann. Appl. Stat.* 2, 841–860. doi: 10.1214/08-AOAS169

Jian, C., Luukkonen, P., Yki-Jarvinen, H., Salonen, A., and Korpela, K. (2020). Quantitative PCR provides a simple and accessible method for quantitative microbiota profiling. *PLoS ONE* 15, e0227285. doi: 10.1371/journal.pone.0227285

Jiao, Y., and Du, P. (2016). Performance measures in evaluating machine learning based bioinformatics predictors for classifications. *Quant. Biol.* 4, 320–330. doi: 10.1007/s40484-016-0081-2

Johnson, A. J., Zheng, J. J., Kang, J. W., Saboe, A., Knights, D., and Zivkovic, A. M. (2020). A guide to diet-microbiome study design. *Front. Nutr.* 7, 79. doi: 10.3389/fnut.2020.00079

Karbach, S. H., Schonfelder, T., Brandao, I., Wilms, E., Hormann, N., Jackel, S., et al. (2016). Gut microbiota promote angiotensin ii-induced arterial hypertension and vascular dysfunction. *J. Am. Heart Assoc.* 5, e003698. doi: 10.1161/JAHA.116.003698

Karlsson, F. H., Fak, F., Nookaew, I., Tremaroli, V., Fagerberg, B., Petranovic, D., et al. (2012). Symptomatic atherosclerosis is associated with an altered gut metagenome. *Nat. Commun.* 3, 1245. doi: 10.1038/ncomms2266

Kelly, T. N., Bazzano, L. A., Ajami, N. J., He, H., Zhao, J., Petrosino, J. F., et al. (2016). Gut microbiome associates with lifetime cardiovascular disease risk profile among bogalusa heart study participants. *Circ. Res.* 119, 956–964. doi: 10.1161/CIRCRESAHA.116.309219

Kessler, T., Vilne, B., and Schunkert, H. (2016). The impact of genome-wide association studies on the pathophysiology and therapy of cardiovascular disease. *EMBO Mol. Med.* 8, 688–701. doi: 10.15252/emmm.201506174

Kessler, T., Wobst, J., Wolf, B., Eckhold, J., Vilne, B., Hollstein, R., et al. (2017). Functional characterization of the, javax.xml.bind.jaxbelement@3a826464, coronary artery disease risk locus. *Circulation* 136, 476–489. doi: 10.1161/CIRCULATIONAHA.116.024152

Kessler, T., Zhang, L., Liu, Z., Yin, X., Huang, Y., Wang, Y., et al. (2015). Adamts-7 inhibits re-endothelialization of injured arteries and promotes vascular remodeling through cleavage of thrombospondin-1. *Circulation* 131, 1191–1201. doi: 10.1161/CIRCULATIONAHA.114.014072

Khera, A. V., Chaffin, M., Aragam, K. G., Haas, M. E., Roselli, C., Choi, S. H., et al. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet*. 50, 1219–1224. doi: 10.1038/s41588-018-0183-z

Khera, A. V., Emdin, C. A., and Kathiresan, S. (2017). Genetic risk, lifestyle, and coronary artery disease. *N. Engl. J. Med*. 376, 1194–1195. doi: 10.1056/NEJMc1700362

Koeth, R. A., Wang, Z., Levison, B. S., Buffa, J. A., Org, E., Sheehy, B. T., et al. (2013). Intestinal microbiota metabolism of l-carnitine, a nutrient in red meat, promotes atherosclerosis. *Nat. Med*. 19, 576–585. doi: 10.1038/nm.3145

Korem, T., Zeevi, D., Zmora, N., Weissbrod, O., Bar, N., Lotan-Pompan, M., et al. (2017). Bread affects clinical parameters and induces gut microbiome-associated personal glycemic responses. *Cell Metab*. 25:1243.e5–1253.e5. doi: 10.1016/j.cmet.2017.05.002

Lempiäinen, H., Brænne, I., Michoel, T., Tragante, V., Vilne, B., Webb, T. R., et al. (2018). Network analysis of coronary artery disease risk genes elucidates disease mechanisms and druggable targets. *Sci. Rep*. 8, 3434. doi: 10.1038/s41598-018-20721-6

Libbrecht, M. W., and Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nat. Rev. Genet*. 16, 321–332. doi: 10.1038/nrg3920

Lieb, W., and Vasan, R. S. (2020). An update on genetic risk scores for coronary artery disease: are they useful for predicting disease risk and guiding clinical decisions? *Expert Rev. Cardiovasc. Ther*. 18, 443–447. doi: 10.1080/14779072.2020.1797489

Lopez, A. D., Mathers, C. D., Ezzati, M., Jamison, D. T., and Murray, C. J. L. (2006). Global and regional burden of disease and risk factors, 2001: systematic analysis of population health data. *Lancet* 367, 1747–1757. doi: 10.1596/978-0-8213-6262-4

Lu, J.-Q., Wang, S., Yin, J., Wu, S., He, Y., Zheng, H.-M., et al. (2017). [A machine learning model using gut microbiome data for predicting changes of trimethylamine-n-oxide in healthy volunteers after choline consumption]. *J. Southern Med. Univ*. 37, 290–295. doi: 10.3969/j.issn.1673-4254.2017.03.02

Marcos-Zambrano, L. J., Karaduzovic-Hadziabdic, K., Loncar Turukalo, T., Przymus, P., Trajkovik, V., Aasmets, O., et al. (2021). Applications of machine learning in human microbiome studies: a review on feature selection, biomarker identification, disease prediction and treatment. *Front. Microbiol*. 12, 634511. doi: 10.3389/fmicb.2021.634511

Maruvada, P., Leone, V., Kaplan, L. M., and Chang, E. B. (2017). The human microbiome and obesity: moving beyond associations. *Cell Host Microbe* 22, 589–599. doi: 10.1016/j.chom.2017.10.005

Mathias, M. G., Coelho-Landell, C. A., Scott-Boyer, M.-P., Lacroix, S., Morine, M. J., Salomao, R. C., et al. (2018). Clinical and vitamin response to a short-term multi-micronutrient intervention in Brazilian children and teens: from population data to interindividual responses. *Mol. Nutr. Food Res*. 62, e1700613. doi: 10.1002/mnfr.201700613

Matlock, D. D., Groeneveld, P. W., Sidney, S., Shetterly, S., Goodrich, G., Glenn, K., et al. (2013). Geographic variation in cardiovascular procedure use among medicare fee-for-service vs medicare advantage beneficiaries. *JAMA* 310, 155. doi: 10.1001/jama.2013.7837

Maximum-likelihood method. (2001). *Encyclopedia of Mathematics* (EMS Press). Available online at: https://encyclopediaofmath.org/wiki/Maximum-likelihood_method

McCulloch, W. S., and Pitts, W. (1990). A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biol*. 52, 99–115; discussion: 73–97. doi: 10.1016/S0092-8240(05)80006-0

Mendes-Soares, H., Raveh-Sadka, T., Azulay, S., Edens, K., Ben-Shlomo, Y., Cohen, Y., et al. (2019). Assessment of a personalized approach to predicting postprandial glycemic responses to food among individuals without diabetes. *JAMA Netw. Open* 2, e188102. doi: 10.1001/jamanetworkopen.2018.8102

Micha, R., Wallace, S. K., and Mozaffarian, D. (2010). Red and processed meat consumption and risk of incident coronary heart disease, stroke, and diabetes mellitus: a systematic review and meta-analysis. *Circulation* 121, 2271–2283. doi: 10.1161/CIRCULATIONAHA.109.924977

Miele, L., Giorgio, V., Alberelli, M. A., De Candia, E., Gasbarrini, A., and Grieco, A. (2015). Impact of gut microbiota on obesity, diabetes, and cardiovascular disease risk. *Curr. Cardiol. Rep*. 17, 120. doi: 10.1007/s11886-015-0671-z

Miyamoto, J., Igarashi, M., Watanabe, K., Karaki, S.-I., Mukouyama, H., Kishino, S., et al. (2019). Gut microbiota confers host resistance to obesity by metabolizing dietary polyunsaturated fatty acids. *Nat. Commun*. 10, 4007. doi: 10.1038/s41467-019-11978-0

Moraes Lopes, M. H. B., Ferreira, D. D., Ferreira, A. C. B. H., da Silva, G. R., Caetano, A. S., and Braz, V. N. (2020). "Use of artificial intelligence in precision nutrition and fitness," in *Artificial Intelligence in Precision Health: From Concept to Applications*, ed D. Barh (London; Cambridge; Oxford; San Diego, CA: Elsevier Science), 465–496. doi: 10.1016/B978-0-12-817133-2.00020-3

Moreira, M. W. L., Rodrigues, J. J. P. C., Korotaev, V., Al-Muhtadi, J., and Kumar, N. (2019). A comprehensive review on smart decision support systems for health care. *IEEE Systems Journal* 13, 3536–3545. doi: 10.1109/JSYST.2018.2890121

Moreno-Indias, I., Lahti, L., Nedyalkova, M., Elbere, I., Roshchupkin, G., Adilovic, M., et al. (2021). Statistical and machine learning techniques in human microbiome studies: contemporary challenges and solutions. *Front. Microbiol*. 12, 635781. doi: 10.3389/fmicb.2021.635781

Movsisyan, N. K., Vinciguerra, M., Medina-Inojosa, J. R., and Lopez-Jimenez, F. (2020). Cardiovascular diseases in central and eastern Europe: a call for more surveillance and evidence-based health promotion. *Ann. Glob. Health* 86, 21. doi: 10.5334/aogh.2713

Munos, B., Baker, P. C., Bot, B. M., Crouthamel, M., de Vries, G., Ferguson, I., et al. (2016). Mobile health: the power of wearables, sensors, and apps to transform clinical trials. *Ann. N. Y. Acad. Sci*. 1375, 3–18. doi: 10.1111/nyas.13117

Myers, A., Johnston, N., Rathod, V., Korattikara, A., Gorban, A., Silberman, N., et al. (2015). "Im2calories: towards an automated mobile vision food diary," in *IEEE International Conference on Computer Vision (ICCV)* (Santiago). doi: 10.1109/ICCV.2015.146

Neiburga, K., Vilne, B., Bauer, S., Bongiovanni, D., Ziegler, T., Lachmann, M., et al. (2021). Vascular tissue specific miRNA profiles reveal novel correlations with risk factors in coronary artery disease. *Biomolecules* 11, 1683. doi: 10.3390/biom11111683

Nelson, C. P., Goel, A., Butterworth, A. S., Kanoni, S., Webb, T. R., Marouli, E., et al. (2017). Association analyses based on false discovery rate implicate new loci for coronary artery disease. *Nat. Genet*. 49, 1385–1391. doi: 10.1038/ng.3913

Ni, Y., Li, J., and Panagiotou, G. (2015). A molecular-level landscape of diet-gut microbiome interactions: toward dietary interventions targeting bacterial genes. *mBio* 6, e01263–15. doi: 10.1128/mBio.01263-15

Nikpay, M., Goel, A., Won, H.-H., Hall, L. M., Willenborg, C., Kanoni, S., et al. (2015). A comprehensive 1,000 genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat. Genet*. 47, 1121–1130. doi: 10.1038/ng.3396

Pallazola, V. A., Davis, D. M., Whelton, S. P., Cardoso, R., Latina, J. M., Michos, E. D., et al. (2019). A clinician's guide to healthy eating for cardiovascular disease prevention. *Mayo Clin. Proc. Innov. Qual. Outcomes* 3, 251–267. doi: 10.1016/j.mayocpiqo.2019.05.001

Pasolli, E., Truong, D. T., Malik, F., Waldron, L., and Segata, N. (2016). Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. *PLoS Comput. Biol*. 12, e1004977. doi: 10.1371/journal.pcbi.1004977

Qi, L. (2012). Gene-diet interactions in complex disease: current findings and relevance for public health. *Curr. Nutr. Rep*. 1, 222–227. doi: 10.1007/s13668-012-0029-8

Reel, P. S., Reel, S., Pearson, E., Trucco, E., and Jefferson, E. (2021). Using machine learning approaches for multi-omics data analysis: a review. *Biotechnol. Adv*. 49, 107739. doi: 10.1016/j.biotechadv.2021.107739

Reynolds, K., Go, A. S., Leong, T. K., Boudreau, D. M., Cassidy-Bushrow, A. E., Fortmann, S. P., et al. (2017). Trends in incidence of hospitalized acute myocardial infarction in the cardiovascular research network (CVRN). *Am. J. Med*. 130, 317–327. doi: 10.1016/j.amjmed.2016.09.014

Rigdon, J., and Basu, S. (2019). Machine learning with sparse nutrition data to improve cardiovascular mortality risk prediction in the USA using nationally randomly sampled data. *BMJ Open* 9, e032703. doi: 10.1136/bmjopen-2019-032703

Roger, V. L., Sidney, S., Fairchild, A. L., Howard, V. J., Labarthe, D. R., Shay, C. M., et al. (2020). Recommendations for cardiovascular health and disease surveillance for 2030 and beyond: a policy statement from the american heart association. *Circulation* 141, e104–e119. doi: 10.1161/CIR.0000000000000756

Samani, N. J., Erdmann, J., Hall, A. S., Hengstenberg, C., Mangino, M., Mayer, B., et al. (2007). Genomewide association analysis of coronary artery disease. *N. Engl. J. Med*. 357, 443–453. doi: 10.1056/NEJMoa072366

Sanchez-Rodriguez, E., Egea-Zorrilla, A., Plaza-Diaz, J., Aragón-Vela, J., Munoz-Quezada, S., Tercedor-Sánchez, L., et al. (2020). The gut microbiota and its implication in the development of atherosclerosis and related cardiovascular diseases. *Nutrients* 12, 605. doi: 10.3390/nu12030605

Schunkert, H., Konig, I. R., Kathiresan, S., Reilly, M. P., Assimes, T. L., Holm, H., et al. (2011). Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nat. Genet*. 43, 333–338. doi: 10.1038/ng.784

Schunkert, H., von Scheidt, M., Kessler, T., Stiller, B., Zeng, L., and Vilne, B. (2018). Genetics of coronary artery disease in the light of genome-wide association studies. *Clin. Res. Cardiol*. 107, 2–9. doi: 10.1007/s00392-018-1324-1

Solares, J. R. A., Raimondi, F. E. D., Zhu, Y., Rahimian, F., Canoy, D., Tran, J., et al. (2020). Deep learning for electronic health records: a comparative review of multiple deep neural architectures. *J. Biomed. Inform*. 101, 103337. doi: 10.1016/j.jbi.2019.103337

Spanakis, G., Weiss, G., Boh, B., Lemmens, L., and Roefs, A. (2017). Machine learning techniques in eating behavior e-coaching. *Pers. Ubiquit. Comput*. 21, 645–659. doi: 10.1007/s00779-017-1022-4

Spector, T., Berry, S., Valdes, A., Drew, D., Chan, A., Franks, P., et al. (2019). Integrating metagenomic information into personalized nutrition tools: the PREDICT I study (p20-005-19). *Curr. Dev. Nutr*. 3(Suppl 1), nzz040.P20-005-19. doi: 10.1093/cdn/nzz040.P20-005-19

Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., et al. (2015). Big data: astronomical or genomical? *PLoS Biol*. 13, e1002195. doi: 10.1371/journal.pbio.1002195

Suykens, J. A., Vandewalle, J., and De Moor, B. (2001). Optimal control by least squares support vector machines. *Neural Netw*. 14, 23–35. doi: 10.1016/S0893-6080(00)00077-0

Tang, W. H. W., Kitai, T., and Hazen, S. L. (2017). Gut microbiota in cardiovascular health and disease. *Circ. Res*. 120, 1183–1196. doi: 10.1161/CIRCRESAHA.117.309715

Tregouet, D.-A., Konig, I. R., Erdmann, J., Munteanu, A., Braund, P. S., Hall, A. S., et al. (2009). Genome-wide haplotype association study identifies the slc22a3-lpal2-lpa gene cluster as a risk locus for coronary artery disease. *Nat. Genet*. 41, 283–285. doi: 10.1038/ng.314

Trøseid, M., Andersen, G. Ø., Broch, K., and Hov, J. R. (2020). The gut microbiome in coronary artery disease and heart failure: current knowledge and future directions. *eBiomedicine* 52, 102649. doi: 10.1016/j.ebiom.2020.102649

Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., et al. (2009). A core gut microbiome in obese and lean twins. *Nature* 457, 480–484. doi: 10.1038/nature07540

van der Harst, P., and Verweij, N. (2018). Identification of 64 novel genetic loci provides an expanded view on the genetic architecture of coronary artery disease. *Circ. Res*. 122, 433–443. doi: 10.1161/CIRCRESAHA.117.312086

Venkataraman, A., Sieber, J. R., Schmidt, A. W., Waldron, C., Theis, K. R., and Schmidt, T. M. (2016). Variable responses of human microbiomes to dietary supplementation with resistant starch. *Microbiome* 4, 33. doi: 10.1186/s40168-016-0178-x

Verma, M., Hontecillas, R., Tubau-Juni, N., Abedi, V., and Bassaganya-Riera, J. (2018). Challenges in personalized nutrition and health. *Front. Nutr*. 5, 117. doi: 10.3389/fnut.2018.00117

Vilne, B., Meistere, I., Grantina-Ievina, L., and Kibilds, J. (2019). Machine learning approaches for epidemiological investigations of food-borne disease outbreaks. *Front. Microbiol*. 10, 1722. doi: 10.3389/fmicb.2019.01722

Vilne, B., and Schunkert, H. (2018). Integrating genes affecting coronary artery disease in functional networks by multi-omics approach. *Front. Cardiovasc. Med*. 5, 89. doi: 10.3389/fcvm.2018.00089

Vilne, B., Skogsberg, J., Foroughi Asl, H., Talukdar, H. A., Kessler, T., Bjorkegren, J. L. M., et al. (2017). Network analysis reveals a causal role of mitochondrial gene activity in atherosclerotic lesion formation. *Atherosclerosis* 267, 39–48. doi: 10.1016/j.atherosclerosis.2017.10.019

Webb, T. R., Erdmann, J., Stirrups, K. E., Stitziel, N. O., Masca, N. G. D., Jansen, H., et al. (2017). Systematic evaluation of pleiotropy identifies 6 further loci associated with coronary artery disease. *J. Am. Coll. Cardiol*. 69, 823–836. doi: 10.1016/j.jacc.2016.11.056

Weber, I., and Achananuparp, P. (2016). Insights from machine-learned diet success prediction. *Pac. Symp. Biocomput*. 21, 540–551. doi: 10.1142/9789814749411_0049

Weber, I., and Achananuparp, P. (2016). Insights from machine-learned diet success prediction. *Pac. Symp. Biocomput*. 21, 540–551.

Westerlund, A. M., Hawe, J. S., Heinig, M., and Schunkert, H. (2021). Risk prediction of cardiovascular events by exploration of molecular data with explainable artificial intelligence. *Int. J. Mol. Sci*. 22:10291. doi: 10.3390/ijms221910291

Wilson, P. W., D'Agostino, R. B., Levy, D., Belanger, A. M., Silbershatz, H., and Kannel, W. B. (1998). Prediction of coronary heart disease using risk factor categories. *Circulation* 97, 1837–1847. doi: 10.1161/01.CIR.97.18.1837

Zeevi, D., Korem, T., Zmora, N., Israeli, D., Rothschild, D., Weinberger, A., et al. (2015). Personalized nutrition by prediction of glycemic responses. *Cell* 163, 1079–1094. doi: 10.1016/j.cell.2015.11.001

Zhao, Z., Yang, Y., Zeng, Y., and He, M. (2016). A microfluidic exosearch chip for multiplexed exosome detection towards blood-based ovarian cancer diagnosis. *Lab Chip* 16, 489–496. doi: 10.1039/C5LC01117E

Zhu, Q., Gao, R., Zhang, Y., Pan, D., Zhu, Y., Zhang, X., et al. (2018). Dysbiosis signatures of gut microbiota in coronary artery disease. *Physiol. Genomics* 50, 893–903. doi: 10.1152/physiolgenomics.00070.2018

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership