Microbiome Metabolome Integration Platform (MMIP): a web-based platform for microbiome and metabolome data integration and feature identification.

Anupam Gautam[1,2,3,$], Debaleena Bhowmik[4,5,$], Sayantani Basu[6], Wenhuan Zeng[1], Abhishake Lahiri[5,7,8], Daniel H. Huson[1,2,3], Sandip Paul[8]*

$: Equal Contribution

[1]Algorithms in Bioinformatics, Institute for Bioinformatics and Medical Informatics, University of Tübingen, Tübingen, Germany.

[2]International Max Planck Research School "From Molecules to Organisms," Max Planck Institute for Biology Tübingen, Tübingen, Germany.

[3]Cluster of Excellence: EXC 2124: Controlling Microbes to Fight Infection, University of Tübingen, Tübingen, Germany.

[4]Cell Biology and Physiology Division, CSIR-Indian Institute of Chemical Biology, Kolkata, India.

[5]Academy of Scientific and Innovative Research (AcSIR), Ghaziabad- 201002, India.

[6]Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801, United States

[7]Infectious Diseases and Immunology Division, CSIR-Indian Institute of Chemical Biology, Kolkata, India.

[8]Centre for Health Science and Technology, JIS Institute of Advanced Studies and Research Kolkata, JIS University, West Bengal, India

*To whom correspondence should be addressed: Sandip Paul, JIS Institute of Advanced Studies and Research, Arch Waterfront, Kolkata 700091, JIS University, West Bengal, India, Phone: +91-8583916488, Email: websandip@gmail.com; sandipp@jisiasr.org

Abstract

A microbial community maintains its ecological dynamics via metabolite crosstalk. Hence knowledge of the metabolome, alongside its populace, would help us understand the functionality of a community and also predict how it will change in atypical conditions. Methods that employ low-cost metagenomic sequencing data can predict the metabolic potential of a community, that is, its ability to produce or utilize specific metabolites. These, in turn, can potentially serve as markers of biochemical pathways that are associated with different communities. We developed MMIP (Microbiome Metabolome Integration Platform), a web-based analytical and predictive tool that can be used to compare the taxonomic content, diversity variation and the metabolic potential between two sets of microbial communities from targeted amplicon sequencing data. MMIP is capable of highlighting statistically significant taxonomic, enzymatic and metabolic attributes as well as learning-based features associated with one group in comparison with another. Further MMIP can predict linkages among species or groups of microbes in the community, specific enzyme profiles, compounds or metabolites associated with such a group of organisms. With MMIP, we aim to provide a user-friendly, online web-server for performing key microbiome-associated analyses of targeted amplicon sequencing data, predicting metabolite signature, and using learning-based linkage analysis, without the need for initial metabolomic analysis, and thereby helping in hypothesis generation.

Introduction

Microorganisms are omnipresent and are important players in the global bio-geochemical cycles. They are also extensive symbionts for all major clades of life (1–4). The study of collective genetic material of microorganisms in any biome is the focus of microbiome research. This varied and intensifying research field is exploring the holistic influence of the microbial community in a diverse set of environments including ocean, soil, human body etc. In case of human health, the critical role of the microbiome has been fully recognized (5, 6). The microbial community structure and function is found to be governed by environmental factors, host physiology and interactions among the microbes (7–9). For a long time, the primary focus of microbiome studies has been to explore the differences in diversities across various sample groups (10). However, the current focus of microbiome research lies in gaining mechanistic insights into the functional capabilities of the microbiome and their impact on host physiology. The delicate balance of host-bacteria crosstalk depends on various factors, including microbial-origin molecules with hormone-like features, bacteria secreted factors like specific peptides and effector proteins, as well as host secreted factors such as IgA and antimicrobial peptides (11, 12). To understand the complex interactions between the microbial community and host physiology, the integration of metabolomics and microbiome analysis has emerged as a powerful approach. By examining the production, modification, or degradation of bioactive metabolites, this integrated analysis explores the intricate role of microbial communities in shaping host physiology. Disease-specific studies have utilized microbiome-metabolome integrative analysis followed by functional characterization to identify disease-associated microbial compounds (13–17). Integrating microbiome and metabolome data is crucial for researchers to unravel the complex web of interactions, identify disease-associated biomarkers, and gain a deeper understanding of the underlying mechanisms involved.

Several computational tools have emerged that attempt to unite the microbiome and metabolome data to either predict the metabolomic yield of a given microbial community or to study the intricate microbe-microbe or host-microbe interactions (18–22). Motivated by the wealth of existing approaches, here we introduce MMIP (Microbiome-Metabolome Integration Platform), an online platform designed to integrate and analyze microbiome and metabolome data. It is aimed at delineating differential community-level information and the metabolic potential of various microbial communities from targeted amplicon sequencing data. Furthermore, we attempt to predict connections among the three layers of data (taxonomy abundance, predicted enzyme profiles and metabolite potential score) that discriminate between two microbial communities and possess probable biological significance.

Materials and Methods

The MMIP web interface utilizes recent versions of bootstrap (v4.3.1) and HTML for a user-friendly experience. JavaScript (ECMA Script 2021) enables interactive plots and downloadable images. MMIP consists of two modules: Module-I for general metagenomic analyses (taxonomic profile, diversity, predicted enzyme profiles, metabolite potential, and feature identification) between two groups and Module-II for correlating predicted metabolites with user-generated metabolomic data. Figure 1 illustrates the MMIP workflow. Users can select the appropriate module based on their experimental design and data. New users are encouraged to consult the MMIP manual, example data, and results. A job queue system notifies users via email with a job-based ID and access link for analysis results.

MMIP data inputs

For MMIP Modules I and II, the required file types are the OTU/ASV abundance table from 16S rRNA amplicon sequencing data and sample-wise metadata file in QIIME format (23). The abundance table can be generated by closed-reference otu-picking using Greengenes database (versions 13_5 and 13_8) or denoising with other database classifiers (such as Greengenes, SILVA, GTDB); the latter requires representative sequences and taxonomy files. Users can apply abundance-based filtering to remove low-count features. The MMIP feature selection model (Support Vector Machine, Random Forest, or Decision Tree) can be specified by the user. For Module-II, users need to upload the metabolic feature table in a tab-delimited "*.txt*" file with sample names in columns and compound names in rows.

MMIP Algorithms

Estimation of community-wise metabolic potential (CMP):
We predicted metagenome content from 16S rRNA amplicon data using PICRUSt (24) or PICRUSt2 (25). We normalized the OTU/ASV abundance table by predicted 16S copy number and performed metagenomic function prediction. This involved a gene content inference step, generating a table (matrix A_{nk}, where n-rows represent the predicted genes, KEGG Orthology or KO identifiers, while k-columns represent all the samples) of predicted gene family counts for each sample. The metagenome inference step provided values representing the contribution of mapped OTUs/ASVs to their gene families. For community-wise metabolite prediction, we used the method by Noecker *et al*. (19), which requires the predicted gene list (KO identifiers) as input. The KO identifiers are mapped to the KEGG reaction map using a surjective function and based upon the directionality of the mapped reactions. Further, these reactions are mapped to the metabolic pathways. A matrix B_{mn} of dimensions $m \times n$ was developed using the reaction stoichiometry, where m (rows) are the compound identifiers and n (columns) represents the KO

identifiers (enzymes/genes) and each of the cells stores the stoichiometric information; it implies that if a compound is present at the reactant side, its stoichiometry will be placed in the respective cell of the table with a negative sign and if it lies at the product side its stoichiometry will be placed with a positive sign. If a particular compound is used by a KEGG enzyme in other pathways then its stoichiometric values will be added or subtracted in the table, depending on the position of the compound in that particular reaction (reactant/product side). The stoichiometric matrix B_{mn} is then normalized, such that the sum total of positive vectors in a row is +1 and that of the negative vectors is -1 resulting in the matrix C_{mn}.

The multiplication of C_{mn} ($m \times n$) matrix with that of the predicted metagenome function matrix A_{nk} ($n \times k$) results in a Community Metabolic Potential (CMP) score matrix D_{mk} having community-wise metabolic potential for each of the metabolites/compounds in the corresponding samples, where m (rows) and k (columns) represent the metabolites and the sample names respectively. The CMP score, much like PRMT (18), is unit-less that only predicts the relative capacity of the microbial community to produce or to consume each metabolite. The mathematical expression of the CMP calculation is shown in supplementary Figure 1.

Feature prediction using machine learning:
The CMP matrix, as mentioned previously, is formed at the backend, and is then used along with the OTU/ASV table and the predicted metagenome table, to predict features that differentiate between the sample groups. MMIP allows the user freedom to opt for any model for the feature selection – including SVM, Random Forest, and Decision Tree, which are popular models used for microbiome data (26–28). This is the most unique attribute of MMIP and is explained in Figure 2.

A Support Vector Machine (SVM) was used to predict the hyperplane separating the data points in a multi-dimensional space. For each dataset, a gene prediction table, a compound metabolic potential score table and a metagenome abundance table (OTU/ASV table) with sample-specific data were generated and all were concatenated sample-wise in order to perform classification. The mathematical formulation of SVM can be stated according to Equation 1 with constraints as per Equation 2 where x represents the variables, y is the class outcome where y∈{-1,+1} (for an SVM model having two classes), C is a hyperparameter used for regularization, w represents the coefficient vector, b is a constant, ξ is the set of slack variables and φ represents the kernel used for data transformation.

$$min_{\mathbf{w},b,\xi_{1:N}} \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{N}\xi_i \quad (1)$$

$$y_i(\mathbf{w}^T\varphi(x_i) + b) \geq 1 - \xi_i \; for \; i = 1, \dots, N, \xi_i \geq 0 \quad (2)$$

Random forest (29) is a model where a set of decision trees is used on a dataset. The model is set up so that the error converges with increase in the number of trees in the forest. Random forest models use feature selection to decide the split at every node. A decision tree (30) is the simpler version of random forest where a single tree is used with splits at every node to determine the outcome of classification based on decisions at the various nodes. For all the mention machine learning algorithm (31, 32), we use the Python implementation in scikit-learn. The decision tree in this implementation is of the CART (Classification and Regression Trees) algorithm (33) that seeks to maximize information gain at node A ($gain(A)$; Equation 3.3), which is defined as the difference between the information I ($I(p,n)$; Equation 3.1) and the expected weighted information E(A) ($E(A)$; Equation 3.2) (30).

$$I(p,n) = -\frac{p}{p+n} log_2 \frac{p}{p+n} - \frac{n}{p+n} log_2 \frac{n}{p+n} \quad (3.1)$$

$$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I(p_i + n_i) \quad (3.2)$$

$$gain(A) = I(p,n) - E(A) \quad (3.3)$$

In terms of each machine-learning algorithm, a grid search using the five-fold cross-validation approach is carried out on the whole dataset to pick the hyperparameters that enable the model to have the best performance among a number of combinations of hyperparameters, implemented mainly by sklearn Python package. More specifically, for SVM, we check the values of gamma=(1e-3,1e-4), and C=(1,10,100,1000) with linear kernel and probability=True options. For random forest, we check n_estimators=(10,50,100) and criterion=(gini, entropy). For the decision tree, we also check criterion (gini, entropy). The model configured with the determined combination of hyperparameters is further fitted on the whole dataset, in order to track the specific features that have a significant contribution to each category of the target dataset.

A nested cross-validation approach with both inner folds and outer folds of 5 was performed on the whole dataset to provide a robust generalization performance evaluation for each machine learning model, since its architecture have the advantage of preventing the model from overfitting on the dataset with limited size. The evaluation metrics including AUC, accuracy, recall, F1-score, and precision which are calculated on the merged output from each outer fold are used to reflect a single model's ability on classifying the target dataset.

MMIP analyses

Functional profiling and metabolite potential:

A Python pipeline performs sanity checks (removal of singleton OTUs/ASVs, normalization by rRNA copy number etc) and comparative metagenome prediction (PICRUSt/PICRUSt2), calculating CMP scores. An R function conducts differential analysis at KEGG functional and

CMP levels using statistical tests such as ANOVA and t-test. Predicted metagenome results, CMP compound-wise boxplots, and statistical analysis are accessible to the user with proper visualization and downloadable diagrams using Plotly (34), D3 (35), Krona Plot (36) and flat file on the web interface

Community profiling and diversity analyses:
The user-uploaded OTU/ASV table and metadata file is processed through a R pipeline, which also uses the Phyloseq package (37) for comparative taxonomic profiling, alpha diversity analyses (Shannon, Simpson diversity indices, effective numbers) and beta diversity analyses (Bray-Curtis dissimilarity, UniFrac-weighted/unweighted). Further the R function is supposed to perform the differential analysis at different taxonomic levels along with all the statistical analysis including Wilcox test, Kruskal-Wallis, ANOVA, t-test etc. All results are available for the user visualization and download.

Feature selection and predictive linkage identification:
The top ten positive and negative CMP features were used along with contribution file in order to link them with corresponding KEGG enzymes and OTUs/ASVs. Through multiple iterations a threshold of 0.7 was selected for the presence of OTUs/ASVs in a certain number of samples. This implies that the corresponding mean OTU/ASV contributed would reflect a certain quantity only if the number of samples for that OTU/ASV exceeded the threshold value, otherwise the value in the mean columns will be zero. The features are then linked according to their feature ranking obtained earlier. The three-tuples that have positive ranks for all three features (the linked CMP, KEGG enzyme and OTU/ASV) have a greater probability of contributing to the positive side of the hyperplane and hence represent collective features for this particular group. Likewise, it can be said for the negatively ranked three-tuples that represent collective features for the other group considered for classification.

Correlation of predicted vs real time metabolites:
With Module-II, users can easily correlate predicted metabolic information with real-time metabolite data (LC-MS, GC-MS feature table). The framework identifies common compounds between predicted and real-time data, applying the Mantel test on their distance matrices. It determines whether the predicted and matched data are positively or negatively correlated based on the correlation coefficient (-1 to +1 range). This allows users to identify the most probable microbial source for the compounds or metabolites detected in the real-time metabolite data.

Dataset selection for validation of our webtool

In order to validate our platform, we selected three studies with publicly available amplicon sequencing data coupled with metabolome data. These validation datasets are from the studies of Srinivasan *et al.* (2015) (dataset-I), Dhakan *et al*. (2019) (dataset-II), and Jacobs *et al*. (2023)

(dataset-III) (38–40). Amplicon sequencing data (accession numbers PRJNA275907, PRJNA397112, PRJNA812699) and paired mass spectrometry data were obtained. All the amplicon data was re-analysed using standard QIIME/QIIME2 pipeline (23, 41), and the OTU/ASV table was obtained either by closed-reference OTU picking against Greengenes v13.8 database (42, 43) or denoised with DADA2 q2 plugin and then assigned taxonomy using SILVA v138 classifier (44). The OTU/ASV table and associated metadata in QIIME format were uploaded to MMIP for analysis. MMIP predicted metabolites were compared with reported data to assess concordance with the prediction model. We used Module-I for dataset–I and Module-II for dataset-II and dataset-III. Metabolomics data were processed and annotated with KEGG IDs separately for dataset II and III (Additional File 1).

Result:

Comparative community-wise profile analyses for validation dataset-I:

The cohort B validation dataset-I (Srinivasan *et al.* (2015) study) contains 40 Bacterial Vaginosis (BV) samples and 20 healthy (H) samples. The alpha diversity (Shannon) analysis by MMIP accurately reflects significantly increased microbial diversity in BV samples compared to the H (Supplementary Figure 2a). Further, the beta diversity of the samples, shown in supplementary Figure 2b, depicts statistically significant separation between the Healthy (H) and the BV samples ($p<0.001$).

The taxonomic distribution showed concordance in the findings generated from MMIP, our standardised analysis and the study itself (supplementary Figure 3). It is known that *Lactobacillus* belonging to the Phylum Firmicutes, is the most abundant microbial genus found in a healthy vagina (45, 46) and it maintains the acidic environment by the production of lactic acid or lactate. MMIP results show a similar trend as reported in the study where the genus *Lactobacillus* and Phylum Firmicutes is over-abundant in H group, while a higher percentage of diverse bacteria (*Megasphaera*, *Gardnerella*, *Prevotella* and *Sneathia*, and also *Atopobium vaginae*) were present in BV samples, indicating a dysbiotic vaginal environment.

Comparative pathway and predicted metabolite analysis for validation dataset-I:

Srinivasan *et al*. (2015) reported 108 metabolites with KEGG annotations; MMIP predicted metabolites overlapped with 52 of the reported metabolites (Supplementary Figure 10a). In case of metabolites prediction, MMIP provides a CMP score corresponding to each metabolite, which signifies the capacity of a bacterial community to consume or accumulate that metabolite. It showed that the relative capacity of a particular metabolite, cadaverine, being accumulated in the BV samples was significantly higher than in the healthy samples, and which agrees with the findings of the study.

The study revealed a number of metabolites including amino acids, dipeptides, and metabolites associated with carbohydrate metabolism. They reported 19 amino acids, out of which 18 were found to be in lower amounts in BV samples as compared to healthy. MMIP also predicted decreased levels of alanine, cysteine, histidine, tryptophan, phenylalanine and the dipeptide, alanylalanine, in BV samples. For metabolites related to carbohydrate metabolism the study demonstrated lowering of simple sugars, intermediate metabolite like glucose-6-phosphate, and sugar alcohols like sorbitol and mannitol in BV samples with respect to healthy. MMIP not only was able to predict all the above-mentioned compounds but also highlighted their tendency to be lower in the BV samples. In addition to that, the study reported lowering of glycerol, ethanolamine and glycerophosphorylcholine (GPC), all of which are associated with lipid metabolism, in BV samples and they too were correctly predicted by MMIP.

Feature link prediction for validation dataset-I:

MMIP predicts links between OTUs, KEGG IDs (enzyme profiles), and metabolites. It is noteworthy that CMP scores in MMIP does not indicate net usage or synthesis of metabolites; positive score suggests increased utilization and negative score implies more accumulation when comparing sample groups.

Such positive and negative feature links obtained from MMIP on -validation dataset-I is depicted in supplementary Figure 4a and 4b respectively. The positive link is predicted for the compound C00089 (sucrose), where it is associated with the enzyme, sucrose phosphorylase (K00690) and further downstream to *Lactobacillus* species. *Lactobacillus* is abundant in the healthy (H) samples, it can thus be expected that a high *Lactobacillus* population causes an increased utilization of the metabolite via the sucrose phosphorylase enzyme. The negative link is predicted for the compound C00436 (N-Carbamoylputrescine), which is associated with the enzyme agmatine deiminase (K10536) and further with the bacterial family *Coriobacteriaceae*, harbouring genera like *Eggerthella* sp. and *Atopobium* sp., specific to the BV samples according to the study. The MMIP link thus suggests that increase in these particular bacterial species in BV group would mean increased action of the enzyme agmatine deiminase, yielding a higher amount of the metabolite Carbamoylputrescine in the samples.

Comparative community-wise profile analyses for validation dataset-II:

Validation dataset-II (Dhakan *et al*. (2019) study) comprised 110 individuals with omnivorous and vegetarian diet, belonging to the Southern and the North-Central states of India, respectively. We filtered out samples with less than 1% of maximum OTU, resulting in a final dataset of 57 omnivorous and 35 vegetarian samples and uploaded to MMIP. The genus-level taxonomic plot generated by MMIP demonstrates a clear difference between the two groups (Supplementary

Figure 5). The significant OTUs obtained from MMIP for this dataset correlated with the findings of the original study. The results showed that the omnivorous group was enriched in *Ruminococcus* and *Faecalibacterium* (including *F. prausnitzii*), and *Roseburia* and *Akkermensia muciniphila*; while the vegetarian group had bacterial genera *Prevotella* and *Megasphaera.*

Comparative pathway and predicted metabolite analysis for validation dataset-II:

Dhakan *et al*. (2019) highlighted enrichment of 5928 KEGG pathways forming two clusters associated with two different diet-groups. MMIP also generated KEGG pathway level feature prediction distinguishing between the two groups and 4918 of those overlapped with that of the study. The study also identified 13 annotated metabolite clusters, 6 of which overlapped with MMIP results (Supplementary Table 1). We processed the metabolomic data separately, identify 306 metabolites, out of which 103 could be annotated and 53 matched with our metabolite prediction (Supplementary Figure 10b, Supplementary Table 2).

While analysing the individual trends of the predicted metabolites, MMIP predicted many of the metabolites as mentioned in the study, including serine (C00065), butyrate (C00246) and valine (C00183), all of which were found to be associated with that of the omnivorous diet group. For the metabolite serine, MMIP assigned a high negative CMP score in the omnivorous group, implying that the omnivorous group associated microbes have a relatively higher potential of serine synthesis; the same holds true for valine as well. SFCAs (short chain fatty acids) like butyrate, propionate, correlate with that of omnivorous diet, and butyrate is found to have a negligible CMP score in vegetarian group, while many of the samples of the omnivorous group show a negative CMP score concluding that again this group has a higher potential of synthesizing butyrate. All the above-mentioned results from MMIP follow the same trend as reported in the study.

Feature link prediction for validation dataset-II:

MMIP predicted dietary habit specific links for validation dataset-II. The positive link is predicted for the compound manninotriose (C05404), and then linked with the enzyme alpha-galactosidase (K07406) and the members of *Ruminococcaceae* family and of genus *Ruminococcus* (Supplementary Figure 6a). These microbes are related to the omnivorous diet group and could imply the increased utilization of the associated metabolite. Similarly, another link was predicted for the compound sucrose-6-phosphate (C02591) associated with the enzyme sucrose-6F-phosphate phosphohydrolase (K07024) and in turn with microbial genera *Ruminococcus* and *Faecalibacterium*, found in omnivorous diet group, shown in supplementary Figure 6b. This linkage too shows increased utilization of the metabolite via the enzyme in the given set of microbes.

Comparative community-wise profile analyses for validation dataset-III:

Validation dataset-III, Jacobs *et al*. (2023) investigated patients with Irritable Bowel Syndrome (IBS) and healthy controls. We specifically selected the cross-sectional samples that included paired metabolomic data without repeats for our analysis. Our final dataset consisted of 181 and 134 samples from individuals with IBS and control subjects respectively. MMIP analysis with ASV table yielded non-significant differences in alpha diversity, similar to the findings in the study, although, differences in beta diversity were significant and the R^2 values were comparable, as also found in the study (Supplementary Figure 7).

The significant taxa identified in MMIP were consistent with the findings of the study. In the IBS group, there was a significant abundance of *Actinomyces*, *Bacteroides*, *Alistipes*, *Faecalibacterium*, and *Blautia hydrogenotrophica* species. Furthermore, the family *Oscillospiraceae* and *Ruminococcaceae*, along with the class Clostridia, exhibited increased population. In contrast, the genera *Bilophila*, *Roseburia*, and *Subdoligranulum* were significantly reduced in the IBS group. Refer to supplementary Figure 8 for a taxonomic plot comparing the genera between the two groups.

Comparative predicted metabolite analysis for validation dataset-III:

Jacobs *et al*. (2023) conducted an extensive analysis of metabolite profiles, and we successfully matched KEGG compound IDs to 306 of those metabolites. By comparing MMIP predicted metabolites with those matched ones, we found 134 overlapping compounds (Supplementary Figure 10c, Supplementary Table 3). Especially, tyramine, which is known for its association with promoting IBS, and fumarate, were found to be elevated in IBS samples. Conversely, hydrocinnamate and 2-oxoarginine were observed in lower quantities compared to control samples.

Feature link prediction for validation dataset-III:

MMIP predicted connections between metabolites, enzymes, and microbial community features that were congruent with IBS-related characteristics found in the study. The first link involved Hematoside (G00108), Hexosaminidase (K12373) enzyme, and *Actinomyces sp*., *Alistipes sp*., *Bacteroides fluxus*, and *Prevotella timonensis* (Supplementary Figure 9a). The study reported an abundance of these bacteria, elevated levels of the enzyme, and increased glycan metabolism in IBS samples. The reaction shown in this link also is part of the glycan metabolism catalyzed by the Hexosaminidase enzyme. The second link connected Lactosylceramide (C01290), Beta-galactosidase (K01190) enzyme, and *Actinomyces* sp., *Faecalibacterium* sp., *Bacteroides* sp., and *Blautia hydrogenotrophica* (Supplementary Figure 9b). The feature compound in the first link had higher accumulation in the IBS group, while the second link's feature compound had

lower accumulation. These linked bacterial groups, as well as the Beta-galactosidase enzyme, were specific to IBS, suggesting increased forward reaction in the IBS samples.

Run time, model performance and comparison with other online tools:

The runtime of MMIP's Module-I depends on the biom file size and number of samples. Feature density in the biom file also affects the processing time. We selected two datasets, Srinivasn *et al*. (2015) and Armstrong *et al*. (2018) (47), created subsets, processed reads with QIIME1 and QIIME2, and uploaded the data individually. Supplementary Table 4 provides the time requirements for different file types, and shows that time increases linearly with sample count.

For feature prediction in the validation datasets, SVM algorithm was used as it excels in high-dimensional spaces. Evaluation of model performance shows that SVM and RF perform similarly, with SVM outperforming for the third dataset. Further we have included metrics for all the models (Supplementary Table 5), ROC curves (Supplementary Figure 11), and confusion matrices (Supplementary Figure 12) to showcase overall performance.

We performed a qualitative comparison of MMIP with few other online microbiome analysis tools - EasyMAP (48), M2IA (49), MicrobiomeAnalyst 2.0 (50, 51) and VAMPS (52), highlighting their different features and functionalities (Supplementary Table 6). While these tools offer comprehensive microbiome and statistical analysis, they lack knowledge-based integration of microbiome and metabolome data. We also performed benchmarking of MMIP with MicrobiomeAnalyst 2.0 using all three datasets (Supplementary Table 7).

Discussion

We validated the functionality of MMIP using datasets from three different studies. Using data from Srinivasan *et al* (2015) (38), MMIP predicted 52 metabolites that overlapped with real time metabolites from the study, along with microbial profiling and diversity indices comparison. Supplementary Figure 4a explains how the positive link could be interpreted biologically and biochemically. Sucrose (C00089) is a disaccharide made of fructose and glucose and the enzyme sucrose phosphorylase (K00690, part of carbohydrate metabolism reaction map), breaks it down to form fructose and glucose-1-phosphate (KEGG reaction ID R00803). This enzyme is further associated with *Lactobacillus iners* and other *Lactobacillus* species, according to MMIP. As per the study, these microbes are associated with the healthy group. Increase in the utilization of sucrose (C00089) is shown by the high positive CMP score in healthy group, suggesting that the forward reaction is taking place and hence there would be an increased formation of fructose, which was found to be less in BV group in the study. Similarly, the reaction in the supplementary Figure 4b suggests, the increase in the population of BV associated bacteria might be associated with an increase in the enzyme action of agmatine deiminase (K10536), along with

the entire Arginine degradation pathway, allowing the forward reaction and hence more production of N-Carbamoylputrescine (C00436), a precursor for putrescine, which is a compound elevated in BV samples in the study. The CMP score of N-Carbamoylputrescine shows that this compound is highly accumulated in BV. It can thus be concluded that this reaction is possibly quite important for BV associated bacteria.

In our second validation dataset by Dhakan *et al*. (2019) (39), the taxonomic distribution between the two groups by MMIP was at par with that of the study, showing significant presence of bacteria *Roseburia* and *Akkermensia muciniphila* in the omnivorous group and that of *Prevotella* and *Megasphaera* in the vegetarian group. The study also checked pathways associated with the two diet groups and when compared with MMIP results they showed remarkable overlap with our pathway analysis. MMIP also predicted the feature linkages, one of them describing the increase in the omnivorous diet specific microbes, which is related to higher activity of the alpha-galactosidase enzyme (K07406), thus favouring forward reaction that depletes manninotriose while increasing D-Galactose (C00124) in the system (Supplementary Figure 6a). Another feature points out that omnivorous group is enriched in Phylum Firmicutes, genus *Ruminococcus*, and species *Faecalibacterium prausnitzii*, implying an increased action of the associated enzymes. Thus, an elevated conversion from sucrose–6-phosphate (C02591) to sucrose (C00089) and eventually to D-Glucose (C00031) and D-Fructose (C00095), both of which were reported to be increased in omnivorous diet (Supplementary Figure 6b).

The final validation study was by Jacob *et al*. (2023) (40) and MMIP was able to reproduce the microbial population signatures mentioned in the study, including *Actinomyces*, *Bacteroides*, *Alistipes*, *Faecalibacterium* and *Blautia hydrogenotrophica* being significantly abundant in IBS samples while lowering of *Bilophila*, *Roseburi*a and *Subdoligranulum*. With respect to metabolites, MMIP was able to predict many of the IBS associated metabolites including tyramine which increased in IBS and hydrocinnamate and 2-oxoarginine, which were lowered. Finally, the MMIP predicted links that highlighted two compounds, one of which accumulated in IBS samples while the other depleted. The first link shows that the bacteria associated with IBS (like *Bacteroides fluxus*, *Prevotella timonensis*) were linked with the Hexosaminidase enzyme (K12373) which drives the forward reaction changing Ganglioside (G00109) to Hematoside (G00108), causing an accumulation of the latter in IBS samples, as indicated by the lower CMP score (Supplementary Figure 9a). The fact that the IBS associated bacteria were reported to have increased Glycan metabolism and the reaction mentioned in the link (KEGG reaction ID R06004), is part of the Glycan metabolism pathway, proves the validity of the link. The second link associates another set of IBS bacteria, *Actinomyces*, *Bacteroides* and *Blautia hydrogenotrophica*, with Beta-galactosidase enzyme, which drives the reaction (KEGG reaction ID R03355) forward, depleting the compound Lactosylceramide (C01290), and which is also shown by the higher negative CMP score for IBS (Supplementary Figure 9b).

When using MMIP, it is important to consider key points and limitations to ensure proper interpretation of the results. MMIP provides the potential value of metabolite utilization within the microbial community, rather than absolute content or metabolism rate. The predicted metabolites in MMIP require experimental validation. Data pre-processing using standardized pipelines is necessary for reliable results. Limitations and biases associated with the KEGG reference database should be considered. Predicted features and linkages may lack statistical significance and should be interpreted cautiously. Users should consider the sample size, as statistical tests do not work for very small sample numbers. It is important to understand that the feature links identified by MMIP can help to generate preliminary hypotheses about the types of reactions occurring in a microbial community. However, further experimental validation and rigorous investigation are necessary to confirm these hypotheses.

This web-resource offers diversity analysis, taxonomy profiling, metagenome prediction, and metabolite potential measurement and a supervised learning-based approach to connect taxonomy abundance, enzyme profiles, and metabolite potential. It provides the user an initial idea of the probable metabolite signature from targeted amplicon sequencing data without the need for metabolomic analysis. MMIP could also help the user to form a preliminary idea of the metabolic potential or the enriched pathways of the microbial communities; and facilitates the formation of biologically valid hypotheses for comparing sample groups.

Data and code availability:

The web-server is available at https://github.com/websandip/mmip and the data for validation was from already published studies which can be accessed by NCBI BioProject through Project ID mentioned in the main text. Further detailed manual on how to use MMIP is available at https://github.com/websandip/mmip/blob/main/manual.pdf

Supplementary data:

All supplementary data (including files, figures and tables) will be available in the published version of this paper.

Acknowledgment:

The research was supported by the Ramanujan Fellowship Grant of the Science and Engineering Research Board (SERB), Government of India, awarded to SP. SP also acknowledges the support of the Bioinformatics Centre, JISIASR. DB received fellowship support from the Indian Council of Medical Research, Government of India. AG acknowledge support by the High Performance and Cloud Computing Group at the Zentrum für Datenverarbeitung of the University of Tübingen, the state of Baden-Württemberg through bwHPC, the German Research

Foundation (DFG) through grant no. INST 37/935-1 FUGG. AG and WZ also acknowledge support of the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) (031A532B, 031A533A, 031A533B, 031A534A, 031A535A, 031A537A, 031A537B, 031A537C, 031A537D, 031A538A). We would like to express our gratitude to Prof. Dr. Boris Maček, Dr. Daniel Petras, and Dr. Timo Sachsenberg for their valuable discussions on GC-MS data analysis. Additionally, we extend our thanks to Prof. Dr. Nico Pfeifer and Jonas Ditz for their insights on machine learning methods.

Conflict of Interest:

None declared.

References

1. Falkowski PG, Fenchel T, Delong EF. 2008. The Microbial Engines That Drive Earth's Biogeochemical Cycles. Science 320:1034–1039.

2. Oh D-C, Poulsen M, Currie CR, Clardy J. 2009. Dentigerumycin: a bacterial mediator of an ant-fungus symbiosis. Nat Chem Biol 5:391–393.

3. McFall-Ngai M. 2014. Divining the Essence of Symbiosis: Insights from the Squid-Vibrio Model. PLOS Biol 12:e1001783.

4. Kawaguchi M, Minamisawa K. 2010. Plant–Microbe Communications for Symbiosis. Plant Cell Physiol 51:1377–1380.

5. Carabotti M, Scirocco A, Maselli MA, Severi C. 2015. The gut-brain axis: interactions between enteric microbiota, central and enteric nervous systems. Ann Gastroenterol Q Publ Hell Soc Gastroenterol 28:203–209.

6. Thaiss CA, Zmora N, Levy M, Elinav E. 2016. The microbiome and innate immunity. Nature 535:65–74.

7. Bonder MJ, Kurilshikov A, Tigchelaar EF, Mujagic Z, Imhann F, Vila AV, Deelen P, Vatanen T, Schirmer M, Smeekens SP, Zhernakova DV, Jankipersadsing SA, Jaeger M, Oosting M, Cenit MC, Masclee AAM, Swertz MA, Li Y, Kumar V, Joosten L, Harmsen H, Weersma RK, Franke L, Hofker MH, Xavier RJ, Jonkers D, Netea MG, Wijmenga C, Fu J, Zhernakova A. 2016. The effect of host genetics on the gut microbiome. Nat Genet 48:1407–1412.

8. Dyhrman S, Ammerman J, Van Mooy B. 2007. Microbes and the Marine Phosphorus Cycle. Oceanography 20:110–116.

9. Krause E, Wichels A, Giménez L, Lunau M, Schilhabel MB, Gerdts G. 2012. Small Changes in pH Have Direct Effects on Marine Bacterial Community Composition: A Microcosm Approach. PLOS ONE 7:e47035.

10. Birtel J, Walser J-C, Pichon S, Bürgmann H, Matthews B. 2015. Estimating Bacterial Diversity for Ecological Studies: Methods, Metrics, and Assumptions. PLOS ONE 10:e0125356.

11. Small molecules from the human microbiota | Science. https://www.science.org/doi/10.1126/science.1254766. Retrieved 11 July 2023.

12. Zhang C, Liu H, Sun L, Wang Y, Chen X, Du J, Sjöling Å, Yao J, Wu S. 2023. An overview of host-derived molecules that interact with gut microbiota. iMeta 2:e88.

13. Lee-Sarwar KA, Lasky-Su J, Kelly RS, Litonjua AA, Weiss ST. 2020. Metabolome–Microbiome Crosstalk and Human Disease. Metabolites 10:181.

14. Visconti A, Le Roy CI, Rosa F, Rossi N, Martin TC, Mohney RP, Li W, de Rinaldis E, Bell JT, Venter JC, Nelson KE, Spector TD, Falchi M. 2019. Interplay between the human gut microbiome and host metabolism. 1. Nat Commun 10:4505.

15. Yachida S, Mizutani S, Shiroma H, Shiba S, Nakajima T, Sakamoto T, Watanabe H, Masuda K, Nishimoto Y, Kubo M, Hosoda F, Rokutan H, Matsumoto M, Takamaru H, Yamada M, Matsuda T, Iwasaki M, Yamaji T, Yachida T, Soga T, Kurokawa K, Toyoda A, Ogura Y, Hayashi T, Hatakeyama M, Nakagama H, Saito Y, Fukuda S, Shibata T, Yamada T. 2019. Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. 6. Nat Med 25:968–976.

16. Franzosa EA, Sirota-Madi A, Avila-Pacheco J, Fornelos N, Haiser HJ, Reinker S, Vatanen T, Hall AB, Mallick H, McIver LJ, Sauk JS, Wilson RG, Stevens BW, Scott JM, Pierce K, Deik AA, Bullock K, Imhann F, Porter JA, Zhernakova A, Fu J, Weersma RK, Wijmenga C, Clish CB, Vlamakis H, Huttenhower C, Xavier RJ. 2019. Gut microbiome structure and metabolic activity in inflammatory bowel disease. 2. Nat Microbiol 4:293–305.

17. Kostic AD, Gevers D, Siljander H, Vatanen T, Hyötyläinen T, Hämäläinen A-M, Peet A, Tillmann V, Pöhö P, Mattila I, Lähdesmäki H, Franzosa EA, Vaarala O, de Goffau M, Harmsen H, Ilonen J, Virtanen SM, Clish CB, Orešič M, Huttenhower C, Knip M, Xavier RJ. 2015. The Dynamics of the Human Infant Gut Microbiome in Development and in Progression towards Type 1 Diabetes. Cell Host Microbe 17:260–273.

18. Larsen PE, Collart FR, Field D, Meyer F, Keegan KP, Henry CS, McGrath J, Quinn J, Gilbert JA. 2011. Predicted Relative Metabolomic Turnover (PRMT): determining metabolic turnover from a coastal marine metagenomic dataset. Microb Inform Exp 1:4.

19. Noecker C, Eng A, Srinivasan S, Theriot CM, Young VB, Jansson JK, Fredricks DN, Borenstein E. 2016. Metabolic Model-Based Integration of Microbiome Taxonomic and Metabolomic Profiles Elucidates Mechanistic Links between Ecological and Metabolic Variation. mSystems 1:e00013-15.

20. Noecker C, Eng A, Muller E, Borenstein E. 2022. MIMOSA2: a metabolic network-based tool for inferring mechanism-supported relationships in microbiome-metabolome data. Bioinforma Oxf Engl 38:1615–1623.

21. Mallick H, Franzosa EA, Mclver LJ, Banerjee S, Sirota-Madi A, Kostic AD, Clish CB, Vlamakis H, Xavier RJ, Huttenhower C. 2019. Predictive metabolomic profiling of microbial communities using amplicon or metagenomic sequences. Nat Commun 10:3136.

22. Shaffer M, Thurimella K, Quinn K, Doenges K, Zhang X, Bokatzian S, Reisdorph N, Lozupone CA. 2019. AMON: annotation of metabolite origins via networks to integrate microbiome and metabolome data. BMC Bioinformatics 20:614.

23. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R. 2010. QIIME allows analysis of high-throughput community sequencing data. Nat Methods 7:335–336.

24. Langille MGI, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, Clemente JC, Burkepile DE, Vega Thurber RL, Knight R, Beiko RG, Huttenhower C. 2013. Predictive

functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat Biotechnol 31:814–821.

25. Douglas GM, Maffei VJ, Zaneveld JR, Yurgel SN, Brown JR, Taylor CM, Huttenhower C, Langille MGI. 2020. PICRUSt2 for prediction of metagenome functions. 6. Nat Biotechnol 38:685–688.

26. Statnikov A, Henaff M, Narendra V, Konganti K, Li Z, Yang L, Pei Z, Blaser MJ, Aliferis CF, Alekseyenko AV. 2013. A comprehensive evaluation of multicategory classification methods for microbiomic data. Microbiome 1:11.

27. Zhou Y-H, Gallins P. 2019. A Review and Tutorial of Machine Learning Methods for Microbiome Host Trait Prediction. Front Genet 10.

28. Rudar J, Golding GB, Kremer SC, Hajibabaei M. 2023. Decision Tree Ensembles Utilizing Multivariate Splits Are Effective at Investigating Beta Diversity in Medically Relevant 16S Amplicon Sequencing Data. Microbiol Spectr 11:e02065-22.

29. Breiman L. 2001. Random Forests. Mach Learn 45:5–32.

30. Quinlan JR. 1986. Induction of decision trees. Mach Learn 1:81–106.

31. sklearn.ensemble.RandomForestClassifier. Scikit-Learn. https://scikit-learn/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html. Retrieved 11 July 2023.

32. sklearn.tree.DecisionTreeClassifier. Scikit-Learn. https://scikit-learn/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html. Retrieved 11 July 2023.

33. Breiman L. 2017. Classification and Regression Trees. Routledge, New York.

34. Inc PT. 2015. Collaborative data science. Plotly Technologies Inc., Montreal, QC. https://plot.ly.

35. Bostock M, Ogievetsky V, Heer J. 2011. D3: Data-Driven Documents. IEEE Trans Vis Comp Graph Proc InfoVis.

36. Ondov BD, Bergman NH, Phillippy AM. 2011. Interactive metagenomic visualization in a Web browser. BMC Bioinformatics 12:385.

37. McMurdie PJ, Holmes S. 2013. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. PLOS ONE 8:e61217.

38. Srinivasan S, Morgan MT, Fiedler TL, Djukovic D, Hoffman NG, Raftery D, Marrazzo JM, Fredricks DN. 2015. Metabolic signatures of bacterial vaginosis. mBio 6:e00204-15.

39. Dhakan DB, Maji A, Sharma AK, Saxena R, Pulikkan J, Grace T, Gomez A, Scaria J, Amato KR, Sharma VK. 2019. The unique composition of Indian gut microbiome, gene catalogue, and associated fecal metabolome deciphered using multi-omics approaches. GigaScience 8:giz004.

40. Jacobs JP, Lagishetty V, Hauer MC, Labus JS, Dong TS, Toma R, Vuyisich M, Naliboff BD, Lackner JM, Gupta A, Tillisch K, Mayer EA. 2023. Multi-omics profiles of the intestinal microbiome in irritable bowel syndrome and its bowel habit subtypes. Microbiome 11:5.

41. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Brown CT, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope EK, Da Silva R, Diener C, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibbons SM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley GA, Janssen S, Jarmusch AK, Jiang L, Kaehler BD, Kang KB, Keefe CR, Keim P, Kelley ST, Knights D, Koester I, Kosciolek T, Kreps J, Langille MGI, Lee J, Ley R, Liu Y-X, Loftfield E, Lozupone C, Maher M, Marotz C, Martin BD, McDonald D, McIver LJ, Melnik AV, Metcalf JL, Morgan SC, Morton JT, Naimey AT, Navas-Molina JA, Nothias LF, Orchanian SB, Pearson T, Peoples SL, Petras D, Preuss ML, Pruesse E, Rasmussen LB, Rivers A, Robeson MS, Rosenthal P, Segata N, Shaffer M, Shiffer A, Sinha R, Song SJ, Spear JR, Swafford AD, Thompson LR, Torres PJ, Trinh P, Tripathi A, Turnbaugh PJ, Ul-Hasan S, van der Hooft JJJ, Vargas F, Vázquez-Baeza Y, Vogtmann E, von Hippel M, Walters W, Wan Y, Wang M, Warren J, Weber KC, Williamson CHD, Willis AD, Xu ZZ, Zaneveld JR, Zhang Y, Zhu Q, Knight R, Caporaso JG. 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. 8. Nat Biotechnol 37:852–857.

42. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL. 2006. Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. Appl Environ Microbiol 72:5069–5072.

43. McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, Andersen GL, Knight R, Hugenholtz P. 2012. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. 3. ISME J 6:610–618.

44. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 41:D590–D596.

45. Antonio MAD, Hawes SE, Hillier SL. 1999. The Identification of Vaginal Lactobacillus Species and the Demographic and Microbiologic Characteristics of Women Colonized by These Species. J Infect Dis 180:1950–1956.

46. Atassi F, Brassart D, Grob P, Graf F, Servin AL. 2006. Lactobacillus strains isolated from the vaginal microbiota of healthy women inhibit Prevotella bivia and Gardnerella vaginalis in coculture and cell culture. FEMS Immunol Med Microbiol 48:424–432.

47. Armstrong AJS, Shaffer M, Nusbacher NM, Griesmer C, Fiorillo S, Schneider JM, Preston Neff C, Li SX, Fontenot AP, Campbell T, Palmer BE, Lozupone CA. 2018. An exploration of Prevotella-rich microbiomes in HIV and men who have sex with men. Microbiome 6:198.

48. Frontiers | EasyMap - An Interactive Web Tool for Evaluating and Comparing Associations of Clinical Variables and Microbiome Composition. https://www.frontiersin.org/articles/10.3389/fcimb.2022.854164/full. Retrieved 11 July 2023.

49. Ni Y, Yu G, Chen H, Deng Y, Wells PM, Steves CJ, Ju F, Fu J. 2020. M2IA: a web server for microbiome and metabolome integrative analysis. Bioinforma Oxf Engl 36:3493–3498.

50. Dhariwal A, Chong J, Habib S, King IL, Agellon LB, Xia J. 2017. MicrobiomeAnalyst: a web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data. Nucleic Acids Res 45:W180–W188.

51. Lu Y, Zhou G, Ewald J, Pang Z, Shiri T, Xia J. 2023. MicrobiomeAnalyst 2.0: comprehensive statistical, functional and integrative analysis of microbiome data. Nucleic Acids Res 51:W310–W318.

52. Huse SM, Mark Welch DB, Voorhis A, Shipunova A, Morrison HG, Eren AM, Sogin ML. 2014. VAMPS: a website for visualization and analysis of microbial population structures. BMC Bioinformatics 15:41.

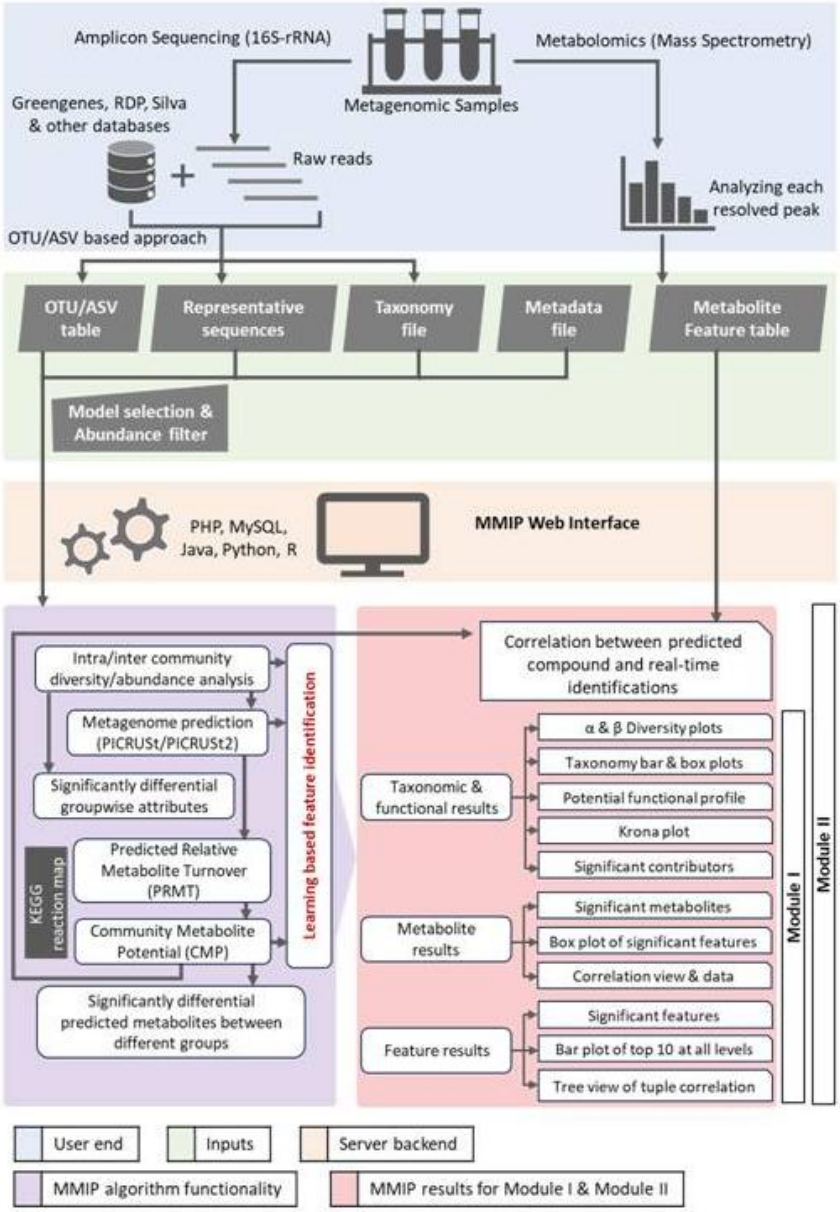

Figure 1: The entire workflow that runs at the backend of MMIP. Keys define different categories for each section of workflow.

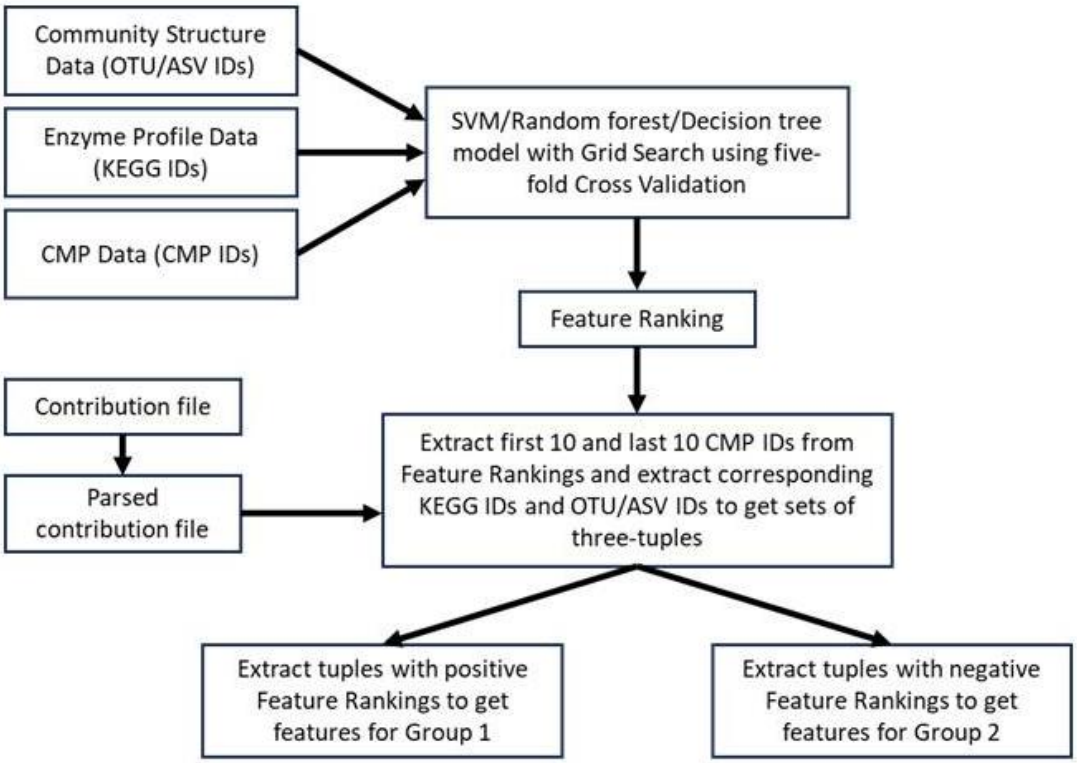

Figure 2: Flowchart for the classification and feature extraction of the two groups for MMIP.