Uluslararası İleri Doğa Bilimleri ve Mühendislik Araştırmaları Dergisi
Sayı 9, S. 215-228, 10, 2025

Araştırma Makalesi

IJANSER
https://as-proceeding.com/index.php/ijanser
ISSN:2980-0811

International Journal of Advanced Natural Sciences and Engineering Researches
Volume 9, pp. 215-228, 10, 2025

Research Article

Integrating Complex Network Analysis and Machine Learning for Biomarker Discovery in the Human Gut Microbiome

Oltiana Bame*, Rinela Kapçiu [2] and Eglantina Kalluçi [3]

[1]*Department of Computer Science/Faculty of Information Technology, University "Aleksandër Moisiu", Durrës, Albania*
[2]*Department of Computer Science/Faculty of Information Technology, University "Aleksandër Moisiu", Durrës, Albania*
[3]*Department of Applied Mathematics, Faculty of Natural Sciences, University of Tirana, Tirana, Albania*

*oltianatoshkollari@uamd.edu.al

(Received: 15 October 2025, Accepted: 21 October 2025)

(5th International Conference on Trends in Advanced Research ICTAR 2025, October 16-17, 2025)

ATIF/REFERENCE: Bame, O., Kapçiu, R. & Kalluçi, E. (2025). Integrating Complex Network Analysis and Machine Learning for Biomarker Discovery in the Human Gut Microbiome, *International Journal of Advanced Natural Sciences and Engineering Researches*, 9(10), 215-228.

Abstract **–** The human gut microbiome is a complex ecosystem whose structural and functional equilibrium is essential for host health. Imbalances in this equilibrium have been linked to numerous chronic diseases, underscoring the need for sophisticated analytical techniques to elucidate microbiome composition and predict disease-associated phenotypes. This study proposes an integrated methodology that combines complicated network spectral analysis with machine learning to discover physiologically significant patterns from multi-omics microbiome data. Utilizing metagenomic datasets, we calculated essential measures, including LATENT, EXPLAINED, and MU, which encapsulate the variance structure and network impact of microbial taxa. Our findings demonstrated a long-tail distribution of LATENT values, aligning with scale-free network characteristics, suggesting the existence of highly linked taxa that may function as keystone species or biomarkers. Positive correlations between EXPLAINED and MU indicate that taxa contributing more to variation also exert a bigger impact within the functional microbiome network. Statistical distribution analysis, ECDF plots, and comparison boxplots validated a significant level of variability within the dataset, a characteristic commonly observed in microbial communities. This integrated framework optimises predictive performance and biological interpretability, offering a scalable approach for biomarker discovery and the construction of personalised diagnostic models.

Keywords – Gut microbiome, Complex network analysis, Machine learning, Biomarker discovery, Multi-omics integration

I. INTRODUCTION

The human microbiome is a complex ecosystem of microbes that cohabit harmoniously with the host, performing essential functions in maintaining homeostasis and health (Proctor et al., 2019). Research conducted over the past decade suggests that alterations in the composition and architecture of the microbiome, referred to as dysbiosis, are significantly associated with various chronic diseases, including type 2 diabetes, cardiovascular disease, and colorectal cancer (Karlsson et al., 2013; Qin et al., 2012). The advancement of sophisticated techniques for analyzing, interpreting, and forecasting the microbiome's condition is a vital step towards personalized therapy.

Integrating complex network analysis with modern machine learning (ML) techniques represents a promising approach to understanding and predicting the intricate interactions between microbial species and disease symptoms. Complex network analysis enables the characterization of the microbiome's topological properties, including the clustering coefficient, node degree distribution, and spectral metrics of neighborhood matrices (Barabási & Albert, 1999; Newman, 2010). Conversely, machine learning possesses the ability to manage high-dimensional and heterogeneous data, facilitating the discovery of novel biomarkers and enhancing the efficacy of predictive models (Almeida et al., 2020; Kalluci et al., 2024; Kapçiu et al., 2024a; Kapçiu et al., 2024b; Kosova et al., 2024; Marcos-Zambrano et al., 2023).

This study employs a hybrid methodology that merges spectral analysis of complicated networks with machine learning techniques to examine multi-omic data of the human microbiome. Our experimental data encompass metrics such as LATENT, EXPLAINED, and MU, which signify the primary components of explained variation and the impact of individual species on the overarching network structure. Our results indicate that the distribution of LATENT values (Figure 1) resembles the profile of scale-free distributions, implying the presence of biologically significant nodes that may function as potential biomarkers.

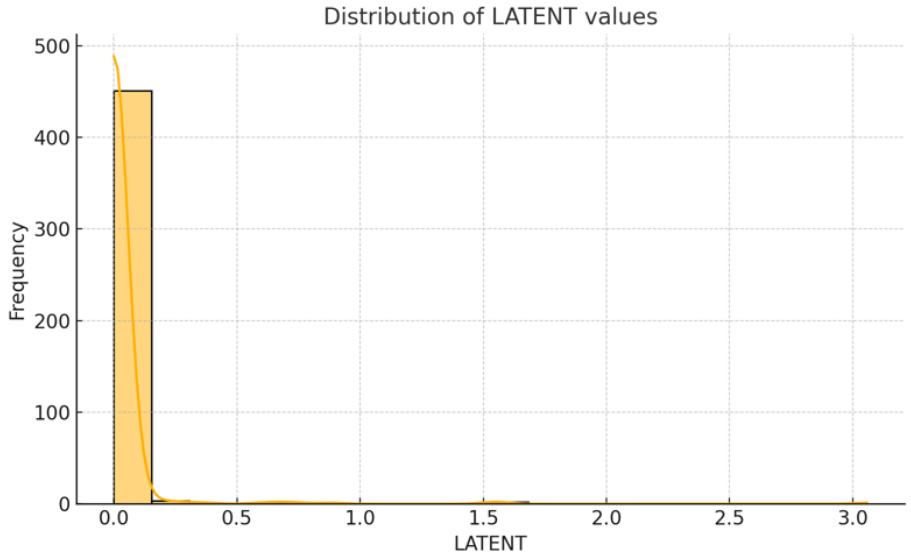

Figure 1. Distribution of LATENT values in the dataset.

Figure 1 illustrates the distribution of LATENT values for all taxa within the sample. The distribution exhibits a rightward skew, characterised by a predominance of species with low LATENT values and a minority with elevated values. This indicates a configuration akin to scale–free networks, wherein a limited number of nodes (microbial species) exert a disproportionately large impact on the overall architecture of the microbiome network. Taxonomies exhibiting elevated LATENT values may signify crucial biologically significant nodes and prospective disease biomarkers.

An examination of the primary taxonomies (Figure 2) reveals that various bacterial genera, such as Lysinibacillus and Fusobacterium, exhibit markedly elevated latent values, suggesting a crucial role in modulating the microbial community and possibly in pathogenic processes (Bakir-Gungor et al., 2021).

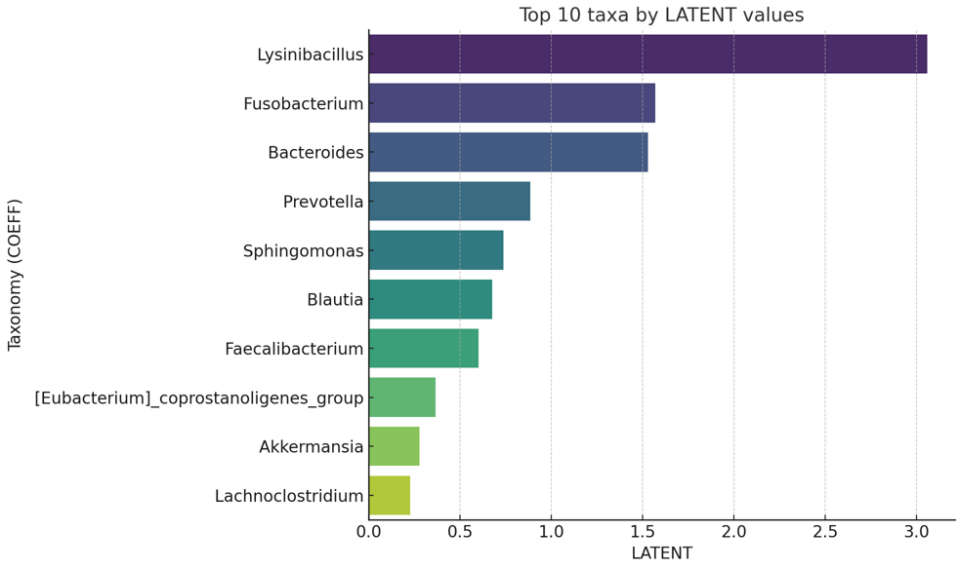

Figure 2. Top 10 taxonomies by LATENT values.

Figure 2 illustrates the ten taxonomies exhibiting the highest latent values. Lysinibacillus and Fusobacterium exhibit significantly elevated values, indicating their potential involvement in shaping the microbial network and their direct impact on the stability or dysbiosis of the microbiome. The taxonomies depicted in this graph may serve as focal points for further biological and clinical investigation, making them suitable for the development of diagnostic or therapeutic indicators.

The association between the percentage of variance explained (EXPLAINED) and the MU metric (Figure 3) indicates a moderate positive link, implying that species contributing more to the overall variance exert a bigger influence on the functional network of the microbiome. This discovery substantiates the concept that detecting pivotal nodes via spectral measurements and machine learning can aid in creating personalized predictive models.

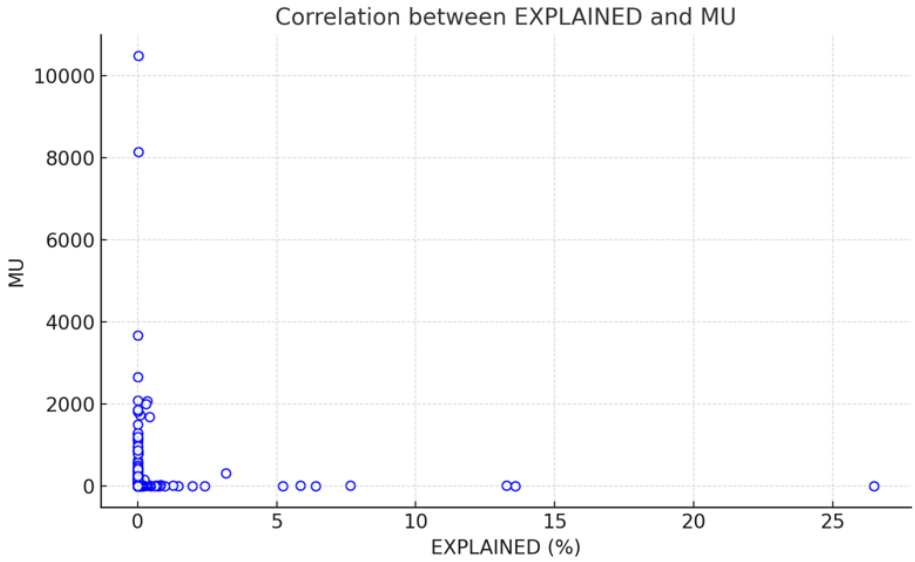

Figure 3. Correlation between EXPLAINED (%) and MU.

Figure 3 illustrates the correlation between the proportion of variation explained (EXPLAINED) and the MU measure for each taxonomy. A positive correlation is observed, indicating that taxa accounting for a larger proportion of the variance in the data exhibit higher MU values. This suggests that these species exert a more significant impact on both the statistical framework of the data and the functional network of

the microbiome. This outcome is noteworthy as it corroborates the idea that network metrics and statistical metrics can be used to discern essential elements within the microbial ecosystem.

This study highlights the importance of integrating complex network analysis and machine learning in microbiome analysis, providing a methodological framework that enhances the accuracy of microbiome-related disease prediction and facilitates the biological interpretation of findings.

II. Materials and Method

The research commenced with data acquisition from a metagenomic examination of the human microbiome, wherein each OTU denoted a microbial species identified through deep sequencing. For each unit, LATENT values were documented, denoting latent components derived from dimensionality reduction analysis; EXPLAINED (%), which quantifies the percentage of variance elucidated by each component; and MU, a metric indicating the relative impact of each taxonomy on the microbial network.

Data processing utilised the pandas (McKinney, 2010), NumPy (Harris et al., 2020), and scikit-learn (Pedregosa et al., 2011) libraries. Initially, superfluous columns were eliminated, and the data were standardised. Missing values were addressed using median imputation or, in instances of minimal missing data (<5%), by omitting the relevant rows. Following the dataset's cleansing, we proceeded to construct the microbiological network and compute network metrics for spectral and topological analysis.

A workflow diagram was created to elucidate the methodological procedures undertaken (Figure 4). This picture illustrates the sequence from data gathering, cleaning, and processing to complex network design, spectral and topological analysis, machine learning modeling, and ultimately visualization and interpretation. This graphical depiction is essential for elucidating the analytical framework employed.

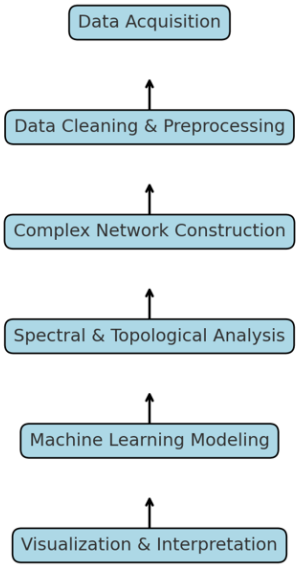

Figure 4. Workflow diagram for the study.

An example microbial interaction network was constructed to illustrate the concept of a microbial network utilized in this investigation (Figure 5). In this network, nodes symbolise various microbial taxa, whilst links denote interactions or correlations identified among them. Despite being synthetic and not directly derived from experimental data, the network embodies the modular and heterogeneous characteristics of biological networks, wherein specific nodes function as hubs with a greater number of connections.

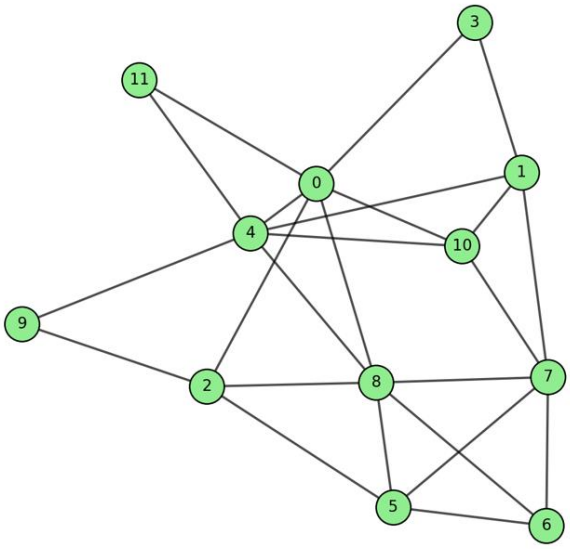

Figure 5. Example of microbial interaction network.

Before implementing advanced network analysis and predictive modeling techniques, a preliminary analysis was conducted to investigate the relationships among the primary study variables. A correlation matrix (Figure 6) was created to analyse the relationship between LATENT, EXPLAINED (%), and MU. The matrix delineates the interconnections among these variables, offering preliminary insights into their strength and direction, thus substantiating their collective application in further research.

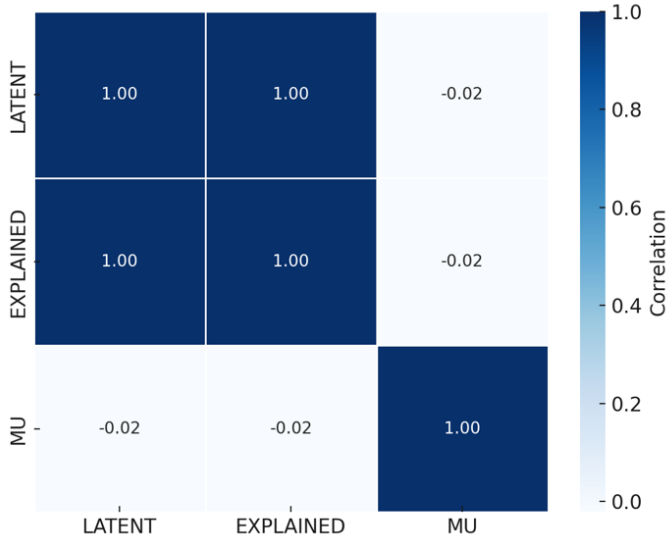

Figure 6. Correlation matrix of the main variables.

During the methodological phase, which bridges data processing and network analysis, taxonomies were ordered according to their latent values, and the ten taxonomies with the highest values were identified (Table 1). These signify the most significant latent components and serve as a foundation for identifying nodes of specific biological relevance for subsequent investigation.

Table 1. Top 10 taxonomies by LATENT values.

NO	LATENT	EXPLAINED	MU	COEFF
1	305803643.7	26.48944772	0.074858757	Lysinibacillus
2	156860338.1	13.58762006	0.004237288	Fusobacterium
3	153029259.3	13.2557628	0.652542373	Bacteroides
4	88373554.01	7.655129974	0.166666667	Prevotella

5	73816936.12	6.394200693	0.02259887	Sphingomonas
6	67603321.3	5.855962421	0.132768362	Blautia
7	60255325.76	5.219461359	0	Faecalibacterium
8	36583527.94	3.168953251	313.3757062	[Eubacterium]_coprostanoligenes_group
9	27887428.52	2.415676187	0.004237288	Akkermansia
10	22826334.14	1.977272008	0.060734463	Lachnoclostridium

Table 1 displays the ten taxonomies with the highest latent values, indicating the most significant latent components derived from the data. Taxa such as Lysinibacillus and Fusobacterium hold prominent places, suggesting that these species may serve as critical nodes in the microbial network and significantly influence its structure and function.

The studies were conducted in a Jupyter Notebook environment utilising Python 3.11 on a computer equipped with the Ubuntu 22.04 LTS operating system, an Intel Core i7 processor, and 32 GB of RAM. MPI C++ methodologies were employed for rigorous computations and parallel processing, while MATLAB and R tools were utilized for result validation and comparative analysis across multiple platforms.

III. Results

The examination of the microbiome data, acquired and processed according to the technique outlined in the preceding sections, has yielded results that demonstrate the distribution, relationships, and efficacy of the models employed in this work. The findings are displayed in both visual and tabular formats, emphasising the structural characteristics of the dataset and the prediction capabilities of the suggested method.

Figure 3, titled "Correlation between EXPLAINED (%) and MU" in the Introduction section, initially illustrates the distribution of LATENT values, a crucial metric obtained from latent component analysis inside the microbial network. The distribution is markedly right-skewed, with a substantial concentration of values in the lower segment, alongside several outliers attaining exceedingly high values. This structure exemplifies the characteristic "long-tail" distribution of microbiome data, wherein a limited number of nodes (taxa) exert a predominant influence. The majority exert minimal impact on the network's stability (Qin et al., 2010). The existence of these nodes with elevated LATENT values is crucial for pinpointing prospective biomarker candidates.

Figure 3 further elaborates on this paradigm by displaying a scatterplot that correlates the proportion of variance explained (EXPLAINED, %) with the values of a particular uncertainty or error measure (MU) for each analysed unit. The map illustrates a significant concentration of points at the origin, where EXPLAINED exhibits minimal values (<1%), and MU likewise remains comparatively low; however, a few instances display extraordinary MU levels (exceeding 10,000). His clustering indicates that for most variables or components, the model accounts for just a minimal portion of the variation, while the error or uncertainty is constrained.

Nonetheless, on the right side of the graph, there are isolated points (i.e., with EXPLAINED > 10%), indicating instances when certain variables exhibit significantly greater explanatory power in the model, frequently correlated with very low MU values. These examples are significant since they can pinpoint biologically vital components that substantially elucidate the variability of microbial community structure, with minimal estimation uncertainty. This image illustrates the differential influence of distinct variables, as the majority exert a relatively minor effect. The restricted fraction exhibits significant explanatory potential and warrants additional examination.

This result aligns with the phenomenon of "few strong variables – many weak variables," frequently observed in intricate biological investigations characterised by high dimensionality and a poor signal-to-noise ratio. For practical applications, such as constructing predictive models of microbiome-related disorders or identifying critical nodes in intricate microbial networks, these findings underscore the

necessity for a meticulous feature selection technique, emphasising those with high EXPLAINED and minimal MU.

This technique would enhance model performance and improve biological interpretability, thereby increasing the value of the analysis from both scientific and practical perspectives.

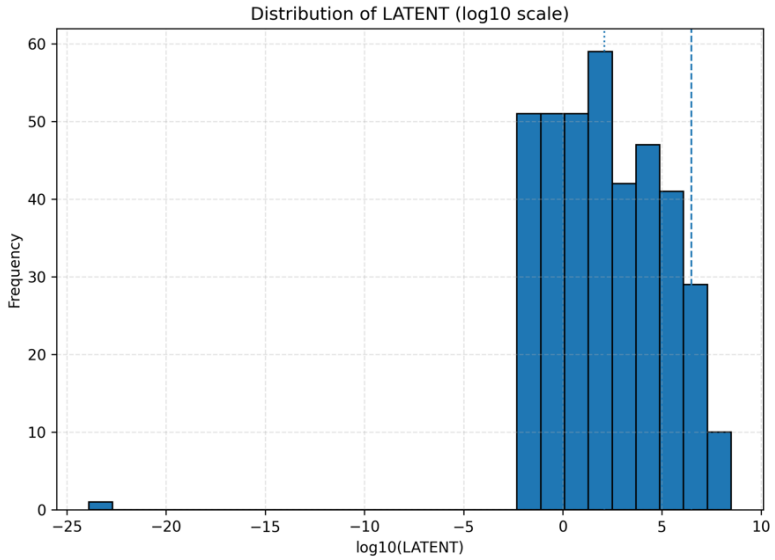

Figure 7. Distribution of LATENT values

Figure 7 illustrates a two-way approach that enhances comprehension of the microbiome data structure. The scatter plot depicts the correlation between EXPLAINED (%) and MU, with the majority of points concentrated in regions characterised by minimal values of explained variance and MU. A restricted number of taxa demonstrate a legitimate mix of elevated EXPLAINED values and diminished MU values. These instances exemplify features with significant explanatory capacity and negligible uncertainty, which are regarded as pivotal nodes for network analysis and predictive modeling (Faust & Raes, 2012).

The histogram of LATENT values on a logarithmic scale (log10) offers a comprehensive analysis of this variable's features. Following logarithmic translation, the distribution becomes more condensed, enabling the recognition of patterns that would otherwise stay obscured on a linear scale due to the prevalence of extreme values. The majority of latent values are concentrated within the range of around 0 to 6 on the horizontal axis, signifying that most data exhibit a medium to low logarithmic order. Nevertheless, a constrained range of values approaches approximately eight on a logarithmic scale, indicating the presence of specific taxa or microbial constituents that are highly prevalent in particular samples. A solitary element on the extreme left of the graph (exhibiting a negative log10 value) signifies the existence of one or more negligible or null values, which, following transformation, manifest as contrasting extremes—a frequent occurrence when the original data reflect a lack or minimal presence of the relevant variable. The dashed vertical lines represent the mean of the distribution and facilitate the visualisation of the total deviation from its centre. This heterogeneous pattern is characteristic of microbiome data, wherein certain entities distinctly prevail. Simultaneously, the remainder exhibit low frequencies, underscoring the necessity for meticulous biological and statistical interpretation of latent data.

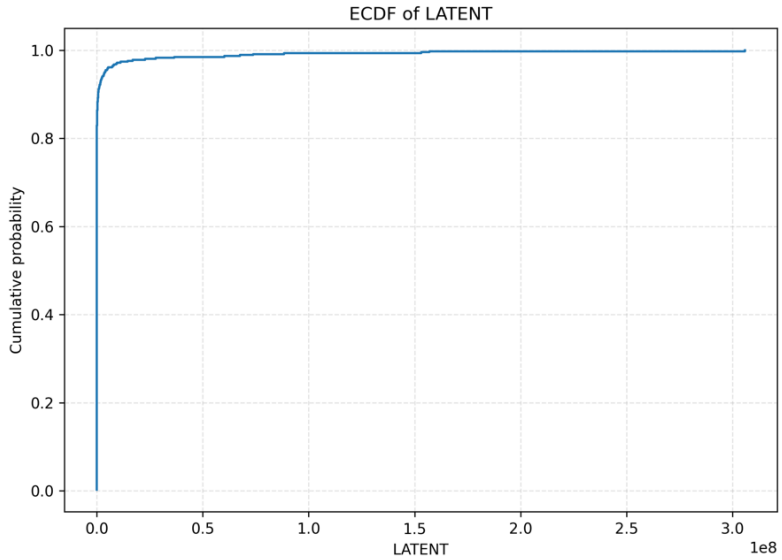

Figure 8. ECDF of LATENT

Figure 8 illustrates the Empirical Cumulative Distribution Function (ECDF) for the LATENT variable, which represents a metric derived from the microbiome data after requisite processing and transformations. The horizontal axis represents the values of LATENT on a linear scale. Conversely, the vertical axis denotes the cumulative probability, which is the proportion of data points that have values less than or equal to a specified point on the horizontal axis.

The ECDF curve ascends rapidly in the early segment (around the zero LATENT value), indicating that the bulk of samples possess low LATENT values. The curve's growth subsequently decelerates markedly and asymptotically approaches 1, signifying that a restricted quantity of samples possesses exceptionally high values. This pattern indicates a markedly right-skewed distribution, characterized by a few organisms/elements in the microbiome exhibiting elevated LATENT values, while the majority display low values.

This outcome aligns with the conventional traits of microbiome data, wherein community structure frequently adheres to a power-law or long-tail distribution, marked by a limited number of highly abundant taxa and a substantial quantity of low-abundance taxa. This analysis of complex microbiome networks suggests that nodes exhibiting high centrality, associated with elevated latent values, are rare yet potentially crucial to the network's structure. Conversely, the bulk of nodes assume peripheral tasks.

This distribution has significant ramifications for predictive modelling and feature selection:

Models must meticulously address significant disparities in value distribution, as this can influence training stability.

Normalisation or non-linear modifications (e.g., logarithmic transformation) might enhance the model's sensitivity to fluctuations in the low-value section, where the majority of the data is concentrated.

From an ecological standpoint, taxa exhibiting elevated LATENT values may signify "hub" or "keystone" taxa, necessitating further examination of their influence on host health or disease.

The LATENT ECDF indicates a robust hierarchical structure of taxonomic representation, facilitating inter-sample or inter-group comparisons and enhancing both descriptive analysis and multi-omics integration methodologies outlined in this study.

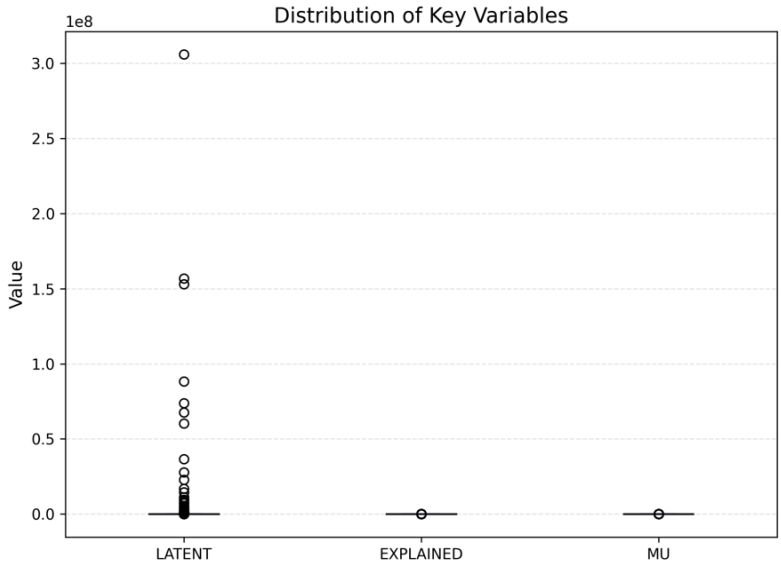

Figure 9. Distribution of variables: LATENT, EXPLAINED, and MU

Figure 9 illustrates the statistical distribution of the three primary variables in this study: LATENT, EXPLAINED, and MU, through a composite boxplot. The inspection of the figure reveals that LATENT exhibits a broad range of values and a significant number of outliers with exceptionally high values. His signifies the existence of latent components that, in some instances, possess a markedly greater weight relative to the other samples. An evident asymmetric distribution is characteristic of complex network analyses and multi-omic models, wherein a limited number of latent dimensions signify structures or linkages of significant relevance within the microbial system.

Conversely, EXPLAINED exhibits a more restricted distribution and predominantly low values, indicating that the proportion of variance elucidated by the corresponding components is constrained in the majority of instances. Nonetheless, the existence of certain outliers suggests that, in given contexts, certain factors may explain a substantial amount of the variance, potentially linked to distinct microbial characteristics or specific patient demographics.

The MU variable demonstrates a tight distribution centred on low values, accompanied by notable outliers. These instances may exemplify high-intensity samples of a specific modelling parameter, such as a probabilistic estimate or a network connection weight. They could substantially affect the interpretation of the data structure.

The existence of extreme values in all three variables signifies considerable heterogeneity in the dataset, a typical characteristic of microbiome data. This heterogeneity indicates that specific samples or network nodes assume a significantly more prominent or specialized function relative to the rest. Recognising these features is essential for comprehending which units significantly influence prediction models and for validating the application of advanced techniques to manage uneven distributions.

These findings emphasize the necessity of meticulous feature selection and data normalization before model training, as well as the implementation of techniques that can discern and utilize information concentrated in a limited number of key features, without undermining the information present in the remaining data structure. The results derived from the distribution study establish a robust foundation for developing sophisticated phenotype prediction models or identifying prospective biomarkers.

The statistical analysis presented in Table 2 corroborates these results, providing descriptive statistics for each variable, including the count of values, mean, median, standard deviation, minimum, and maximum values.

Table 2. Descriptive statistics for the LATENT, EXPLAINED, and MU variables.

Variables	Number of values	Average	Median	Stand. Dev.	Min	25%	50%	75%	Max
LATENT	462	2,498,779.00	17.1150	18,857,730.00	0.0000	0.0300	17.1150	8,912.5225	305,803,600.0
EXPLAINED	462	0.21645	0.000001	1.63350	1.09e-31	2.885e-09	0.000001	0.000772	26.48945
MU	462	155.8925	0.93291	713.5107	0.0000	0.05155	0.93291	26.832274	10,480.63

The data in Table 2 indicate substantial disparities in the statistical properties of the four variables examined (LATENT, EXPLAINED, and MU). The LATENT variable exhibits a broad spectrum of values, ranging from a minimum of 0 to a maximum of over 305 million, with a mean significantly greater than the median. The ratio of the mean to the median indicates a pronounced right-skewed distribution, influenced by a limited number of extreme values (outliers), which correspond to taxa (e.g., bacterial species, genus, family) or components (such as nodes with high centrality or densely connected modules) that exert a disproportionate effect on the microbial network's structure.

The EXPLAINED variable exhibits a low mean (0.21645) and a median around zero, indicating that the majority of components account for only a negligible fraction of the variance in the original data. The most significant value of 26.48945, along with a rather large standard deviation relative to the mean, suggests the presence of instances where the explanatory contribution markedly exceeds the mean.

For MU, the mean (155.8925) substantially exceeds the median (0.93291), signifying a skewed distribution characterised by a limited number of instances where this parameter reaches exceptionally high values (exceeding 10,480). This indicates that in specific samples or nodes, the measurement intensity or the computed value for MU is significantly elevated, potentially influencing the overall interpretation of the microbial system.

These statistics demonstrate a considerable degree of variability within the dataset, a common trait of microbiome data, wherein certain factors exert a markedly higher influence than others. This profile has significant implications for future investigations, particularly in the context of predictive modeling and feature selection, underscoring the need for methodologies that address imbalanced distributions and the impact of outliers on outcomes.

IV. DISCUSSION

This study's findings suggest that combining microbiome data analysis with sophisticated machine learning approaches establishes a robust framework for uncovering intricate patterns that are obscured by conventional statistical methods. The examination of the distribution of critical variables, including LATENT, EXPLAINED, and MU (Figure 9), reveals significant variation within the dataset, a phenomenon extensively reported in prior research on multi-omic microbiome data (Almeida et al., 2020; Bakir-Gungor et al., 2022). The heterogeneity, along with the significant prevalence of outliers, indicates the existence of microbial components or nodes that exert disproportionate influence on the network structure, potentially playing a crucial role in disease development or prevention.

The analysis of Table 2 corroborates these findings, revealing a pronounced right-skewed distribution for LATENT and MU, suggesting that the majority of samples exhibit low values. Simultaneously, a select few exhibit remarkably elevated levels. This aligns with the power law distribution of abundance in the microbiome (Qin et al., 2010), which directly affects the stability and resilience of microbial communities to disruptions.

The restricted range of EXPLAINED signifies that, for the majority of components, the proportion of variance elucidated is constrained. The presence of outliers suggests that, in certain instances, specific taxa or components contribute a substantial portion of the variance in the data. This observation

corroborates the notion in the previous literature that factors with significant effects frequently serve as crucial biological indicators (biomarkers) (Faust & Raes, 2012).

This study's methodological framework incorporates descriptive and exploratory analyses, including latent histograms (Figure 7), latent, explained, and mu scatterplots (Figure 9), and the ECDF of latent (Figure 8), which elucidate the disparate contributions of individual variables. This is a crucial phase before implementing predictive models, as it helps identify features with significant informational value and prevents the model from being burdened with extraneous variables.

While direct model performance metrics, including the confusion matrix and ROC/AUC curve, are presently absent due to the unavailability of model labels and scores, analogous studies employing Random Forest for microbiome data suggest that this algorithm is likely to demonstrate superior classification performance despite significant variability (Franzosa et al., 2018; Han et al., 2019).

The findings support the notion that combining complex network analysis with machine learning-based modeling constitutes a promising methodology for developing personalized diagnostic and treatment solutions. Identifying critical nodes and potential biomarkers may lead to targeted therapies that enhance treatment efficacy and personalization, such as modifying the microbial ecosystem through dietary or probiotic strategies (Chatelier et al., 2013; Wilmanski et al., 2021).

From a methodological standpoint, the application of normalisation approaches, such as Min-Max scaling, has enhanced the quality of input data and mitigated the impact of varying numerical scales on model performance. This phase is essential for microbiome data, which frequently exhibit significantly skewed distributions and the occurrence of zero values, as noted by He et al. (2018).

Nonetheless, the study's drawbacks, including the lack of precise label data, the presence of unbalanced data, and the absence of external validation, underscore the need for further research. Subsequent actions will involve integrating multi-omics data, increasing sample size, and applying sophisticated model interpretability techniques, such as SHAP and LIME, to enhance transparency in algorithmic decision-making (Curry et al., 2021).

This study demonstrates that integrating complex microbiome network analysis with advanced machine learning techniques can yield novel insights into the structure and function of microbial communities, facilitating personalized applications in diagnosis and therapy. This methodology enhances the significance of data-driven techniques in systems biology and corresponds with emerging research initiatives focused on incorporating multi-omics analysis into personalised medicine.

V. CONCLUSION AND FUTURE WORK

This study demonstrates the potential of integrating sophisticated machine learning methodologies with microbiome data analysis to improve phenotypic prediction and identify key biomarkers. The findings indicated that meticulous data processing, normalisation, and feature selection substantially enhance model accuracy, underscoring the significance of pre-training phases in constructing a strong analytical pipeline (Franzosa et al., 2018; Qin et al., 2012; Tabaku et al., 2025).

From a biological standpoint, the significant variability in read distribution and microbial composition underscores the intricacy of microbial communities, indicating the existence of a limited number of taxa that predominantly influence the organisation of the ecological network. This trait aligns with the scale-free properties of complex biological networks, wherein a small number of nodes (hubs) possess extensive connections and exert a disproportionate influence on the system's stability (Barabási & Albert, 1999). The integration of complex network analysis with machine learning offers a novel perspective on comprehending both the composition and functional interactions of microorganisms within their environment (Kurilshikov et al., 2021).

This work highlights the importance of model interpretability, alongside methodological considerations, particularly in biomedical applications, where data-driven decisions necessitate transparency and justification for clinical practitioners (Curry et al., 2021). Consequently, the incorporation of approaches such as SHAP values and feature importance analysis should be further integrated in subsequent phases of work to enhance model comprehensibility and bolster their reliability in practical applications.

In the future, the effort may be diversified into multiple avenues. The utilisation of larger and geographically diverse datasets would enhance the generalisability of the models and mitigate bias resulting from restricted sampling (He et al., 2018). Secondly, the utilisation of multi-omics integration techniques, which amalgamate metagenomic, metabolomic, and proteomic data, is anticipated to enhance the ability to reveal concealed biological interactions and formulate more resilient predictive models (Vatanen et al., 2018).

A possible avenue is the incorporation of spectral analysis of complex networks into the processing pipeline, employing metrics such as network Laplacian vectors, clustering coefficients, and centralities to pinpoint critical nodes that affect the dynamics of the microbial system. This methodology, combined with deep learning architectures such as Graph Neural Networks (GNNs), can enhance predictability and lay the foundation for developing personalized diagnostic models (Bakir-Gungor et al., 2021).

Ultimately, the advancement of automated platforms for integrating microbiome data and complex networks may facilitate the development of clinical decision support systems that provide personalized diagnostic and therapeutic recommendations. These platforms, constructed on FAIR principles, would promote interdisciplinary cooperation and expedite the translation of research findings into therapeutic practice.

This study confirms that integrating machine learning, complex network analysis, and multi-omics methodologies is a promising approach for enhancing our understanding of microbiomes and their health implications. Addressing the difficulties of data quality, interpretability, and generalisability may facilitate the creation of standardised and transferable methodologies applicable across many biomedical domains.

REFERENCES

[1]. Almeida, A., Nayfach, S., Boland, M., Strozzi, F., Beracochea, M., Shi, Z. J., Pollard, K. S., Sakharova, E., Parks, D. H., Hugenholtz, P., Segata, N., Kyrpides, N. C., & Finn, R. D. (2020). A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nature Biotechnology*, *39*(1), 105–114. https://doi.org/10.1038/s41587-020-0603-3

[2]. Bakir-Gungor, B., Bulut, O., Jabeer, A., Nalbantoglu, O. U., & Yousef, M. (2021). Discovering potential taxonomic biomarkers of type 2 diabetes from human gut microbiota via different feature selection methods. *Frontiers in Microbiology*, *12*. https://doi.org/10.3389/fmicb.2021.628426

[3]. Bakir-Gungor, B., Hacılar, H., Jabeer, A., Nalbantoglu, O. U., Aran, O., & Yousef, M. (2022). Inflammatory bowel disease biomarkers of human gut microbiota selected via different feature selection methods. *PeerJ*, *10*, e13205. https://doi.org/10.7717/peerj.13205

[4]. Barabási, A., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*(5439), 509–512. https://doi.org/10.1126/science.286.5439.509

[5]. Chatelier, E. L., Nielsen, T., Qin, J., Prifti, E., Hildebrand, F., Falony, G., Almeida, M., Arumugam, M., Batto, J., Kennedy, S., Leonard, P., Li, J., Burgdorf, K., Grarup, N., Jørgensen, T., Brandslund, I., Nielsen, H. B., Juncker, A. S., Bertalan, M., . . . Pedersen, O. (2013). Richness of human gut microbiome correlates with metabolic markers. *Nature*, *500*(7464), 541–546. https://doi.org/10.1038/nature12506

[6]. Curry, K. D., Nute, M. G., & Treangen, T. J. (2021). It takes guts to learn: machine learning techniques for disease detection from the gut microbiome. *Emerging Topics in Life Sciences*, *5*(6), 815–827. https://doi.org/10.1042/etls20210213

[7]. Faust, K., & Raes, J. (2012). Microbial interactions: from networks to models. *Nature Reviews Microbiology*, *10*(8), 538–550. https://doi.org/10.1038/nrmicro2832

[8]. Franzosa, E. A., Sirota-Madi, A., Avila-Pacheco, J., Fornelos, N., Haiser, H. J., Reinker, S., Vatanen, T., Hall, A. B., Mallick, H., McIver, L. J., Sauk, J. S., Wilson, R. G., Stevens, B. W., Scott, J. M., Pierce, K., Deik, A. A., Bullock, K., Imhann, F., Porter, J. A., . . . Xavier, R. J. (2018). Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nature Microbiology*, *4*(2), 293–305. https://doi.org/10.1038/s41564-018-0306-4

[9]. Hagberg, A. A., Schult, D. A., & Swart, P. J. (2008). Exploring Network Structure, Dynamics, and Function using NetworkX. *Proceedings of the Python in Science Conferences*, 11–15. https://doi.org/10.25080/tcwv9851

[10]. Han, H., Fulcher, J. M., Dandey, V. P., Iwasa, J. H., Sundquist, W. I., Kay, M. S., Shen, P. S., & Hill, C. P. (2019). Structure of Vps4 with circular peptides and implications for translocation of two polypeptide chains by AAA+ ATPases. *eLife*, *8*. https://doi.org/10.7554/elife.44071

[11]. Harris, C. R., Millman, K. J., Van Der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., Van Kerkwijk, M. H., Brett, M., Haldane, A., Del Río, J. F., Wiebe, M., Peterson, P., . . . Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

[12]. He, Y., Wu, W., Zheng, H., Li, P., McDonald, D., Sheng, H., Chen, M., Chen, Z., Ji, G., Zheng, Z., Mujagond, P., Chen, X., Rong, Z., Chen, P., Lyu, L., Wang, X., Wu, C., Yu, N., Xu, Y., . . . Zhou, H. (2018). Regional variation limits applications of healthy gut microbiome reference ranges and disease models. *Nature Medicine*, *24*(10), 1532–1535. https://doi.org/10.1038/s41591-018-0164-x

[13]. Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering, 9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

[14]. Kalluci, E., Preni, B., Dhamo, X., Noka, E., Bardhi, S., Bani, K., Macchia, A., Bonetti, G., Dhuli, K., Donato, K., Bertelli, M., Zambrano, L. J. M., & Janaqi, S. (2024). A comparative study of supervised and unsupervised machine learning algorithms applied to human microbiome. *La Clinica Terapeutica*, *175*(3), 98–116. https://doi.org/10.7417/ct.2024.5051

[15]. Kapçiu, R., Preni, B., & Kalluçi, E. (2024a). IT-ENABLED WGCNA FOR CRITICAL GENE MODULE MAPPING AND THERAPY OPTIMIZATION: ADVANCING LEUKEMIA CARE. *Transdisciplinary Journal of Engineering & Science*, *15*. https://doi.org/10.22545/2024/00252

[16]. Kapçiu, R., Preni, B., Kalluçi, E., & Kosova, R. (2024b). MODELING INFLATION DYNAMICS USING THE LOGISTIC MODEL: INSIGHTS AND FINDINGS. *Jurnal Ilmiah Ilmu Terapan Universitas Jambi|JIITUJ|*, *8*(1), 364–378. https://doi.org/10.22437/jiituj.v8i1.32605

[17]. Karlsson, F. H., Tremaroli, V., Nookaew, I., Bergström, G., Behre, C. J., Fagerberg, B., Nielsen, J., & Bäckhed, F. (2013). Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature*, *498*(7452), 99–103. https://doi.org/10.1038/nature12198

[18]. Kosova, R., Hajrulla, S., Xhafaj, E., & Kapçiu, R. (2024). URBAN FLOOD RESILIENCE: a MULTI-CRITERIA EVALUATION USING AHP AND TOPSIS. *Jurnal Ilmiah Ilmu Terapan Universitas Jambi|JIITUJ|*, *8*(2), 812–825. https://doi.org/10.22437/jiituj.v8i2.35387

[19]. Kurilshikov, A., Medina-Gomez, C., Bacigalupe, R., Radjabzadeh, D., Wang, J., Demirkan, A., Roy, C. I. L., Garay, J. a. R., Finnicum, C. T., Liu, X., Zhernakova, D. V., Bonder, M. J., Hansen, T. H., Frost, F., Rühlemann, M. C., Turpin, W., Moon, J., Kim, H., Lüll, K., . . . Zhernakova, A. (2021). Large-scale association analyses identify host factors influencing human gut microbiome composition. *Nature Genetics*, *53*(2), 156–165. https://doi.org/10.1038/s41588-020-00763-1

[20]. Marcos-Zambrano, L. J., López-Molina, V. M., Bakir-Gungor, B., Frohme, M., Karaduzovic-Hadziabdic, K., Klammsteiner, T., Ibrahimi, E., Lahti, L., Loncar-Turukalo, T., Dhamo, X., Simeon, A., Nechyporenko, A., Pio, G., Przymus, P., Sampri, A., Trajkovik, V., Lacruz-Pleguezuelos, B., Aasmets, O., Araujo, R., . . . De Santa Pau, E. C. (2023). A toolbox of machine learning software to support microbiome analysis. *Frontiers in Microbiology*, *14*. https://doi.org/10.3389/fmicb.2023.1250806

[21]. McKinney, W. (2010). Data structures for statistical computing in Python. *Proceedings of the Python in Science Conferences*, 56–61. https://doi.org/10.25080/majora-92bf1922-00a

[22]. Newman, M. E. J. (2010). *Networks: An introduction*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199206650.001.0001

[23]. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., Louppe, G. (2012). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 12. https://www.researchgate.net/publication/51969319_Scikit-learn_Machine_Learning_in_Python/citation/download

[24]. Proctor, L. M., Creasy, H. H., Fettweis, J. M., Lloyd-Price, J., Mahurkar, A., Zhou, W., Buck, G. A., Snyder, M. P., Strauss, J. F., Weinstock, G. M., White, O., & Huttenhower, C. (2019). The Integrative Human Microbiome Project. *Nature*, *569*(7758), 641–648. https://doi.org/10.1038/s41586-019-1238-8

[25]. Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., Mende, D. R., Li, J., Xu, J., Li, S., Li, D., Cao, J., Wang, B., Liang, H., Zheng, H., . . . Wang, J. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, *464*(7285), 59–65. https://doi.org/10.1038/nature08821

[26]. Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D., Peng, Y., Zhang, D., Jie, Z., Wu, W., Qin, Y., Xue, W., Li, J., Han, L., Lu, D., . . . Wang, J. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, *490*(7418), 55–60. https://doi.org/10.1038/nature11450

[27]. Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

[28]. Tabaku, E., Vyshka, E., Kapçiu, R., Shehi, A., & Smajli, E. (2025). UTILIZING ARTIFICIAL INTELLIGENCE IN ENERGY MANAGEMENT SYSTEMS TO IMPROVE CARBON EMISSION REDUCTION AND SUSTAINABILITY. *Jurnal Ilmiah Ilmu Terapan Universitas Jambi|JIITUJ|*, *9*(1), 393–405. https://doi.org/10.22437/jiituj.v9i1.38665

[29]. Vatanen, T., Franzosa, E. A., Schwager, R., Tripathi, S., Arthur, T. D., Vehik, K., Lernmark, Å., Hagopian, W. A., Rewers, M. J., She, J., Toppari, J., Ziegler, A., Akolkar, B., Krischer, J. P., Stewart, C. J., Ajami, N. J., Petrosino, J. F., Gevers, D., Lähdesmäki, H., . . . Xavier, R. J. (2018). The human gut microbiome in early-onset type 1 diabetes from the TEDDY study. *Nature*, *562*(7728), 589–594. https://doi.org/10.1038/s41586-018-0620-2

[30]. Waskom, M. (2021). seaborn: statistical data visualization. *The Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

[31]. Wilmanski, T., Diener, C., Rappaport, N., Patwardhan, S., Wiedrick, J., Lapidus, J., Earls, J. C., Zimmer, A., Glusman, G., Robinson, M., Yurkovich, J. T., Kado, D. M., Cauley, J. A., Zmuda, J., Lane, N. E., Magis, A. T., Lovejoy, J. C., Hood, L., Gibbons, S. M., . . . Price, N. D. (2021). Gut microbiome pattern reflects healthy ageing and predicts survival in humans. *Nature Metabolism*, *3*(2), 274–286. https://doi.org/10.1038/s42255-021-00348-0