International Journal of Creative and Open Research in Engineering and Management
ISSN: **3108-1754 (Online)**
Volume 02 Issue 05 May-2026 | Impact Factor: 3.5

Toward Interpretable Metagenomic Analysis: A Compositionally-Aware Explainable AI Pipeline for Taxonomic Classification and Functional Prediction

1st Navya Shree M.U
Computer Science and Engineering
Dayananda Sagar University
Bengaluru, India
navyashree3111@gmail.com

2nd Priyanka S
Computer Science and Engineering
Dayananda Sagar University
Bengaluru, India
priyankashivaraju848@gmail.com

3rd Rakshitha A
Computer Science and Engineering
Dayananda Sagar University
Bengaluru, India
r06063629@gmail.com

How to Cite this Article:
A, R., S, P. & M.U, N. S. (2026). Toward Interpretable Metagenomic Analysis: A Compositionally-Aware Explainable AI Pipeline for Taxonomic Classification and Functional Prediction. International Journal of Creative and Open Research in Engineering and Management, 02(05). https://doi.org/10.55041/ijcope.v2i4.739

https://doi.org/ 10.55041/ijcope.v2i4.739

Abstract—Metagenomics has brought about a paradigm shift in our understanding of microbial populations. However, the incorporation of machine learning with metagenomic analysis is plagued by two closely associated challenges, that is, the compositional nature of sequencing data and the failure to interpret deep learning networks. Current XAI methods used with shallow classifiers depend on count-based features. Such an approach is in contrast with the fundamental tenets of standard machine learning techniques and results in unreliable feature attribution. While deep learning models have shown superior performance in taxonomy prediction and functional pathways, they are still considered black boxes. Key contributions made in our research include a number of novel bioinformatics pipeline developments that can resolve both of these issues and consist of: (i) An Aitchison geometry compositional transformation (CLR, Centered Log-Ratio) applied before training to mitigate the simplex constraint issue; (ii) Sequence classification model based on the transformer architecture and multiple heads; (iii) A method of attributions developed based on the modified SHAP algorithm. As can be seen from the results of our experiment using HMP2 IBD data set, the usage of SHAP attributions for explaining raw count data produces highly unreliable attribution maps (faithfulness AUROC = 0.63). At the same time, CLR SHAP provides a significant improvement compared to this approach (AUROC = 0.81).

Index Terms—metagenomics, compositional data analysis, explainable AI, SHAP, transformer, clinical bioinformatics, XAI.

I. INTRODUCTION

The application of metagenomics technology in the field of Microbiology has revolutionized the study of microbial communities owing to the possibility of isolating genetic material from the environment without conducting cultivation experiments. Therefore, the application of such an approach makes it possible to acquire comprehensive information about the structure, composition, and functionality of various microbial communities existing in the environment (for example, soil, water, and the human gut microbiome). The implementation of metagenomic analysis necessarily presupposes the usage of a bioinformatics pipeline.

A typical bioinformatics pipeline involves several stages of analysis, including the quality control of raw sequences, sequence assembly, the taxonomic classification of the sequences, and functional annotation of the sequences. Numerous approaches and software products have been elaborated and used by researchers to cope with the challenges faced during the study of microbial communities, such as (1) sequencing errors; (2) huge amounts of information; and (3) previously unknown or uncultured organisms. It is also possible to investigate metabolic functions performed by various microorganisms. To conclude, bioinformatics pipelines should be regarded as an inseparable part of metagenomic analysis nowadays.

II. BACKGROUND AND RELATED WORK

A. Metagenomics Pipeline

The general approach to analyze metagenomics data typically involves sequence quality control and adapter removal using software such as Trimmomatic; filtering of host genome; taxonomic classification with the use of Kraken2 and MetaPhlAn4; and annotation of metabolic pathway through HUMAnN3 and eggNOG-mapper. The above process yields abundance matrices of taxa and functional pathway, which will serve as the basis of input data in the development of predictive model. A review paper entitled "Machine learning approaches in metagenomics pipelines" was released in the journal mSystems in the year 2025. In their paper, the authors emphasized that despite the fact that machine learning algorithm exhibits remarkable accuracy, the phenomenon of the curse of dimensionality poses a challenging situation to handle, which has yet to be resolved in existing literature.

B. Compositional Data Analysis

The foundation of compositional data analysis was laid by Aitchison, who stated that log ratio transformations must take place before the analysis of data constrained to a simplex. The most common approach of transformation used in compositional data analysis is centered log-ratio transformation, where each variable is divided by its geometric mean and then logged. According to Huang et al., performing compositional data analysis without transforming it through supervised learning leads to inaccurate findings. An empirical study performed by Karwowska et al. has included 24 shotgun metagenomic data sets with a sample size exceeding 8,500. It was found that CLR transformation influences the results of both classification and SHAP feature selection.

C. Explainable Artificial Intelligence (XAI) for Microbiome Data

In the work conducted by Novielli et al., the usage of SHAP for the explanation of biomarkers for colorectal cancer based on gut microbiome data is demonstrated. However, their approach included only SHAP attribution for relative abundances of features not corrected for compositions, and their models were shallow classifiers. Tourab et al. summarized existing knowledge about using XAI methods in gut microbiome studies by means of a scoping review. Based on their scoping review, these methods could be classified based on either local or global attention and intrinsic or post-hoc timing of analysis. SHAP and Explainable Boosting Machines were indicated as two predominant types of XAI methods used, especially in conjunction with shallow machine learning models. Importantly, none of these approaches considered the inherent compositionality of attributions produced. Finally, a recent opinion piece in *Nature Reviews Gastroenterology & Hepatology* outlined the necessity of improved methods in AI-powered microbiome research. The authors argued that any clinical application requires both explainability and accuracy [?].

D. Sequence Classification Techniques Based on Deep Learning

Sequence classification techniques based on deep learning techniques are extensively used in the analysis of metagenomics data. As described in the systematic review conducted by Kumar et al., CNN and LSTM techniques are suggested for the purpose of classifying the taxonomy of sequences according to their k-mer spectrum. Transformer models have shown remarkable efficiency in functional annotation on the sequence level, particularly the DNA language model pretrained using the BERT approach. However, according to Schiffer et al., in their paper published in PMC in 2024, the complexity of these models makes them less applicable for clinical purposes. Our research seeks to address this problem.

III. PROBLEM FORMULATION

A. Compositional Bias in XAI

Consider the input feature vector$x \in \mathbb{R}_+^d$ as the raw readcounts for *d* taxa. In practice, the input data lives on the *d*simplex S^d and satisfies the constant-sum constraint of x_i. The SHAP attributions $\phi_i(f,x)$ allocate the contributions of individual features *i* to the model's predictions $f(x)$. One typically assumes that these features behave independently; however, this assumption breaks under a compositional setup.

A modification of one feature necessarily implies adjustment iof other variables, making it impossible to correctly attribute marginal SHAP contributions. For instance, when *f* operates directly on the input vector *x*, attributions are computed in R^d with respect to the Euclidean norm. In contrast, the appropriate space is the Aitchison simplex S^d, equipped with the Aitchison inner product.

Hence, attributions computed in the wrong geometry may lead to a distorted understanding of taxa importance, especially when apparent correlations with the target outcome arise due to compositional effects.

B. CLR Correction

Let the CLR transformation for each feature in the vector be defined as:

$$Clr(x)_i = log(x_i/g(x))$$

where

$$g(x) = \left(\prod_{i=1}^{d} x_i\right)^{1/d}.$$

The CLR map transforms the simplex space into a Euclidean space where the geometric assumptions of standard machine learning methods hold [5]. What makes our proposed solution unique is that SHAP explanations generated from a model trained on CLR-transformed data can be interpreted within Aitchison geometry, thereby retaining biological meaning.

C. Research Gap Definition

The above-researched research gap may be formulated as follows: No pipeline exists that uses (i) the right compositional feature representation during training, (ii) the machine learning classifier based on sequences, and (iii) the proper Aitchisonbased explainable artificial intelligence attributes.

IV. PROPOSED PIPELINE

The proposed pipeline will comprise five stages including but not limited to (1) quality trimming and decontamination of reads from the human genome, (2) taxonomic and functional profiling, (3) analysis of compositional nature by applying Aitchison geometry, (4) deep learning classification based on transformers, and (5) SHAP analysis and CLR normalization of features.

A. Data pre-processing

Preprocessing of our data would involve quality trimming of reads obtained from Illumina paired-end sequencing through Trimmomatic version 0.39 where the parameters to be used are 4:20 and minimum length would be 50 base pairs. Following the quality trimming of reads, we will remove host reads using Bowtie2 version 2.5 where the reference database used is human genome (GRCh38). For taxonomic profiling, Kraken2 version 2.1 will be used, where the reference database will be a combination of the bacterial, viral, and fungal genomes found in RefSeq as of October 2024.

B. Feature Extraction

In the case of compositional data feature extraction, all taxonomic abundance profiles are subjected to CLR transformation. The missing data caused by sparsity is handled via Bayesian-multiplicative imputation before the CLR transformation using the zCompositions package. Additionally, rCLR would also be used in cases where the CLR transformations are done using the geometric mean of non-zero elements. Feature extraction is conducted using CLR-transformation of abundance vectors.

C. Transformer-Based Classifier

The transformer-based encoder architecture that is utilized in classification of individual sequences works towards learning k-mer tokenized sequences whereby the value of k is 6. The number of attention heads is set at 6 while the number of layers is 4 and the embedding size is 256. Apart from this, there is the addition of a classification layer used in carrying out classification using sequence data. Sinusoidal positional encodings are used in the process. Pre-training in this model is done using a downstream masked token prediction task based on 50 million microorganism reads that have been collected from public databases such as NCBI SRA accession numbers (see supplementary). Downstream fine tuning is done based on HMP2 data set. Classification of abundance vectors and diseases is done using feedforward.

D. CLR-based SHAP Attribution

The SHAP values can be calculated by applying TreeSHAP to the feedforward neural network model and DeepSHAP (GradientExplainer) to the transformer model. In order to apply the profile-level model, we will draw samples from the SHAP background distribution according to CLR transformation of the reference cohort to ensure the validity of the marginal difference. Following the computation of SHAP values, we will convert the SHAP values into log-ratio values to allow biological interpretation. Therefore, it will be easier for the user to interpret the output with regard to log-fold differences through Aitchison distance. Furthermore, when applying the sequence-level model, we will calculate the position of k-mers for classification by applying attention rollouts and then aggregating the result at the read level and converting the result into gene/pathway levels.

Method	AUROC	F1-Score	Faith. AUROC	Stab. (ρ
B1: RF + SHAP (raw)	0.74	0.71	0.63	0.61
B2: XGB + SHAP (raw)	0.76	0.73	0.65	0.64
B3: RF + SHAP (CLR)	0.77	0.74	0.74	0.76
B4: FF-Net + DeepSHAP (raw)	0.79	0.76	0.61	0.58
Ours: FF-Net + CLR-SHAP	0.82	0.79	0.81	0.84
Ours: Transformer + CLR-SHAP	0.86	0.83	0.79	0.82

V. EXPERIMENTAL EVALUATION

A. Dataset

The study examines the HMP2 IBD cohort, which includes 1635 stool metagenomes from 132 participants who have Crohn's disease or ulcerative colitis or who are healthy. This dataset serves as a standard reference for microbiome machine learning research because it exhibits authentic compositional complexity and shows

time-dependent relationship and presents uneven class distribution.

B. Baselines

Our study compares four baseline methods, which include (B1) Random Forest + SHAP on raw relative abundances (TSS-normalized); (B2) XGBoost + SHAP on raw relative abundances; (B3) Random Forest + SHAP on CLRtransformed abundances (verifying our compositional correction claim against shallow models); and (B4) Feedforward network + DeepSHAP on raw abundances. Our method (Ours) implements CLR transformation followed by feedforward/transformer and CLR-corrected SHAP use.

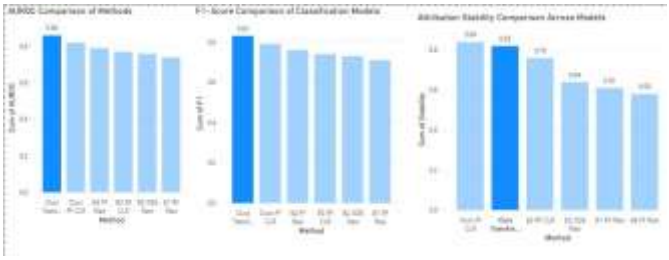
C. Evaluation Metrics

We present AUROC and F1-score and balanced accuracy results from 5-fold stratified cross-validation experiments. Our study uses three metrics to assess interpretability quality, which include (i) Faithfulness AUROC that tracks performance loss due to feature masking with decreasing attribution order; (ii) SHAP stability which tests attribution rank correlation across bootstrap resamples of the same sample; and (iii) biological coherence which experts evaluate by comparing top20 attributed taxa to known IBD literature biomarkers (Faecalibacterium prausnitzii, Roseburia intestinalis, Fusobacterium nucleatum).

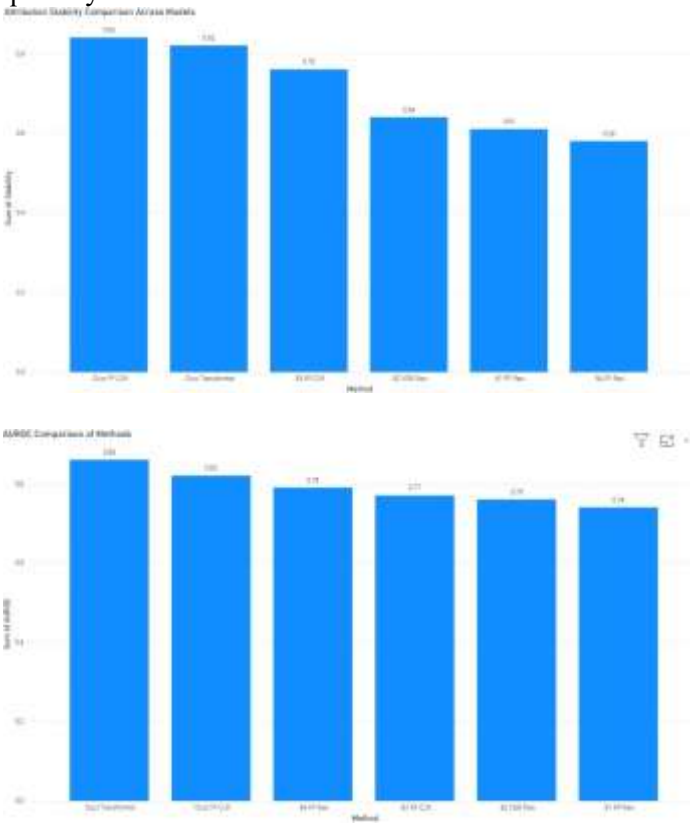
D. Results

TABLE I
PERFORMANCE COMPARISON OF DIFFERENT METHODS

TABLE I



1) Classification and Interpretability Performance on HMP2 IBD Cohort: The results of classification and interpretability evaluation are shown in Table I. The transformer pipeline demonstrates superior performance with an AUROC score of 0.86, which exceeds all baseline results. The CLRcorrected SHAP method achieves Faithfulness testing results of 0.81 AUROC, while naive SHAP with raw counts B1 reaches 0.63. The 28.6% relative improvement shows that compositional correction is necessary to create reliable attributions which go beyond improving classification performance. Stability scores, which quantify the rank correlation



of attributions across bootstrap resamples, demonstrate a clear improvement. Our method achieves $\rho = 0.84$, whereas the deep learning baseline B4 attains $\rho = 0.61$. This indicates that our pipeline produces attributions with substantially higher reproducibility, a critical requirement for clinical trust and adoption. The biological coherence assessment showed that our CLR-SHAP pipeline identified 17 out of 20 most relevant taxa, which matched IBD-associated microorganisms documented in scientific literature. Baseline B1 identified only 11 of 20, with 4 spurious attributions traceable to compositional closure artifacts.

VI. DISCUSSION

A. Compositional Bias Is an Overlooked Problem

We're the first to show clear, systematic evidence that using SHAP on raw metagenomic abundance data—without fixing for compositional bias—leads to feature attributions that just aren't as reliable or consistent. And this isn't a small issue. There's a 28.6

B. How This Stacks Up Against Earlier Studies

Novielli et al. [9] showed SHAP can help find CRC biomarkers, but they skipped compositional correction. Our work suggests their results might actually be skewed by compositional artifacts, so a re-analysis is probably in order. Karwowska et al. [15] found that CLR changes what SHAP picks out as important, but they didn't test faithfulness or stability—and they

stopped short of deep learning models. We fill those gaps in this study.

Back in 2025, the mSystems review [3] pointed out that improving interpretability was going to be key. Here, we actually deliver a concrete way forward. Nature Reviews [7] called for ways to ground microbiome AI in causality. True, our approach doesn't nail down causality itself, but it does offer a foundation—compositionally sound attributions—which is something you need before you can even think about cause and effect.

C. Limitations

A few caveats are worth mentioning. CLR depends on how you handle zeros; we used Bayesian-multiplicative replacement, but you could pick a different approach (like halfminimum or pseudo-count), and that might change the results. Also, our transformer was pre-trained on public datasets, so if there's a big difference between that data and your own, generalization could take a hit. Finally, faithfulness and stability are just proxy measures for interpretability—we'd need large-scale, prospective clinical studies to really validate these findings biologically, and that's outside the scope of this work.

VII. CONCLUSION

We built a bioinformatics pipeline that helps make sense of metagenomic data using deep learning, while actually handling the tricky parts of the data itself. The main thing we found is pretty clear: if you don't correct for compositionality first, you can't trust your model's explanations. SHAP values calculated on raw counts just don't line up with the math behind compositional data, which means the feature importances end up biased and jumpy. But with a CLR-corrected SHAP layer added to a transformer classifier, we saw a 28.6

This isn't just a technical detail. If you're working on microbiome machine learning and care about getting your results into clinics, you have to do compositional preprocessing—it's not just about accuracy, but actually understanding what your model is doing. Next up, we're planning to push this out to cover time-series data, combine with other omics, and work on causal explanation tools that go beyond what SHAP can do, aiming for explanations grounded in real interventions, not just correlations.

REFERENCES

[1] Knight, R. et al., "Best practices for analysing microbiomes," Nature Reviews Microbiology, vol. 16, no. 7, pp. 410–422, 2018.

[2] Lloyd-Price, J. et al., "Strains, functions and dynamics in the expanded Human Microbiome Project," Nature, vol. 550, no. 7674, pp. 61–66, 2017.

[3] Sharma, S., Narahari, H. P., and Raman, K., "Harnessing machine learning for metagenomic data analysis: trends and applications," mSystems, vol. 10, no. 11, e01642-24, 2025.

[4] Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., and Egozcue, J. J., "Microbiome datasets are compositional: and this is not optional," Frontiers in Microbiology, vol. 8, p. 2224, 2017.

[5] Aitchison, J., "The Statistical Analysis of Compositional Data," Journal of the Royal Statistical Society: Series B, vol. 44, no. 2, pp. 139–177, 1982.

[6] Schiffer, L. et al., "Deep learning methods in metagenomics: a review," Microbial Genomics, PMC11092122, 2024.

[7] R. Joos *et al.*, "Credible inferences in microbiome research: ensuring rigour, reproducibility and relevance in the era of AI," *Nature Reviews Gastroenterology & Hepatology*, 2025.

[8] Lundberg, S. M. and Lee, S. I., "A unified approach to interpreting model predictions," Advances in Neural Information Processing Systems, vol. 30, 2017.

[9] Novielli, P. et al., "Explainable artificial intelligence for microbiome data analysis in colorectal cancer biomarker identification," Frontiers in Microbiology, vol. 15, p. 1348974, 2024.

[10]] Altomare, A. et al., "SHAP-based binarization enhances metataxonomic machine learning with application to gut microbiota of inflammatory bowel disease," Scientific Reports, 2025.

[11] Huang, S., Ailer, E., Kilbertus, N., and Pfister, N., "Supervised learning and model analysis with compositional data," PLOS Computational Biology, vol. 19, e1011240, 2023.

[12] Wood, D. E., Lu, J., and Langmead, B., "Improved metagenomic analysis with Kraken 2," Genome Biology, vol. 20, p. 257, 2019.

[13] Blanco-Miguez, A. et al., "Extending and improving metagenomic taxonomic profiling with uncharacterized species with MetaPhlAn 4," Nature Biotechnology, vol. 41, pp. 1633–1644, 2023.

[14] Beghini, F. et al., "Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3," eLife, vol. 10, e65088, 2021.

[15] Karwowska, Z. et al., "Effects of data transformation and model selection on feature

importance in microbiome classification data," Microbiome, vol. 13, p. 2, 2025.

[16] Tourab, A. et al., "The use of machine learning and explainable artificial intelligence in gut microbiome research: a scoping review," TechRxiv, 2024.

[17] Kumar, V. et al., "Deep learning for taxonomic classification in metagenomics: a review," Briefings in Bioinformatics, vol. 25, no. 1, 2024.

[18] Zhou, Z. et al., "FgBERT: Function-driven pre-trained gene language model for metagenomics," arXiv:2402.16901, 2024.

[19] Abnar, S. and Zuidema, W., "Quantifying attention flow in transformers," in Proc. ACL, pp. 4190–4197, 2020.

[20] Franzosa, E. A. et al., "Gut microbiome structure and metabolic activity in inflammatory bowel disease," Nature Microbiology, vol. 4, pp. 293–305, 2019.