# Multipurpose Denoising Autoencoder based framework for understanding taxa specific states in the gut microbiome

By
SHRUTI KHARE

Under the Supervision of

DR. TARINI SHANKAR GHOSH

Assistant Professor

Department of Computational Biology

Submitted

in partial fulfillment of the requirements for the degree of

Master of Technology

to

Indraprastha Institute of Information Technology Delhi

JUNE, 2025

# CERTIFICATE

This is to certify that the thesis titled "Multipurpose Denoising Autoencoder based framework for understanding taxa specific states in the gut microbiome" being submitted by **Shruti Khare** to the Indraprastha Institute of Information Technology Delhi, for the award of the Master of Technology, is an original research work carried out by her under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results in this thesis have not been submitted in part or full to any other university or institute for the award of any degree/diploma.

JUNE, 2025

25-06-2025

Dr. Tarini Shankar Ghosh
Assistant Professor
Department of Computational Biology
Indraprastha Institute of Information Technology
(A State University Established by Govt. of Delhi)
Okhla Phase III, New Delhi-110020

**Dr. Tarini Shankar Ghosh**
Department of Computational Biology
Indraprastha Institute of Information Technology Delhi
New Delhi 110 020

# ACKNOWLEDGEMENT

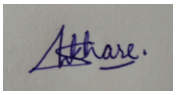
First, I would like to bow in reverence to God, whose blessing provided me with the vigorous passion, uninterrupted strength and patience needed to begin my work and complete it successfully.

I express my humble gratitude towards **Dr.Tarini Shankar Ghosh,** my professor, for offering me the opportunity and mentorship throughout the process. I am truly thankful for his guidance, expertise, insightful feedback and continuous encouragement all which played a crucial role in shaping the outcome of this project.

I would like to thank **Sourav Goswami**, a dedicated Ph.D. scholar, whose constant support has been invaluable throughout my journey. No matter how small my doubts were, he was always there to help with patience. I had the privilege of spending most of my project work under his guidance and I am deeply grateful for his ability to make most challenging work less burdensome.

I extend my sincere thanks to all members of Microbiome Informatics Lab for their knowledge and assistance in various aspects of my thesis work.Their intellectual inputs and feedback was invaluable in improving the quality of research. Special thanks to Omprakash Shete who, as a senior, always supported and guided me in the right direction.

I am grateful to my friends for making my stay very pleasant and fruitful. It would have been an uphill task without their endless love and care. Their presence made a world of difference.



Shruti Khare
MT23238

# CONTENT

# ABSTRACT

Microbiome data is often challenged by the fact that the obtained microbiome compositional profile may not be either the true or the ideal one. There is a problem certain that species may not be detected because of being either too low in number or their genomic content being inefficiently extracted during sample processing. This results in artefactual zero values. There is also a likelihood of sub-optimality where the actual state of the microbiome may be different from the ideal state depending upon the community-composition of the resident members. While the issue of artefactual zeros may be pronounced for rarer low abundant taxa, the prevalent or the core-gut-associated taxa may be more affected by the issue of sub-optimality. While addressing the first issue can help to improve biological interpretations from microbiome data, investigating the second aspect can have multiple benefits, ranging from identifying unstable microbiome states, identifying response to probiotic therapies as well as in discerning microbiome-associated disease associations.Here, we propose a deep learning-based imputation framework designed explicitly for microbiome data using a Denoising Autoencoder (DAE) to address both these issues. This approach leverages neural networks to capture non-linear patterns and complexity within species abundance distributions and effectively predict missing values for rarer species and sub-optimal or over-representation values for core-associated species. The framework was generated and trained on a large microbiome dataset (Abundance profile) with 44,943 samples and 354 features representing microbial species. For the imputation function, the framework's performance was evaluated using statistical approaches including Bray-Curtis similarity and Spearman correlation and compared with current existing imputation methods, including GemIMP and DeepImpute using both a validation as well as simulated microbiome datasets. The DAE consistently outperformed these alternative strategies and obtained the strongest correlation. For investigating unstable and responsive-ness to probiotic interventions, we extended our analysis by introducing species-specific receptive-scores and overall-microbiome state-stability-scores, derived from the DAE framework. Using data from an intervention trial involving a Bifidobacterium longum probiotic, we show that the DAE framework is able to distinguish between ‘Persisters’ (or Responders) or ‘Non-Persisters’ (or Non-Responders) using B. longum receptive scores. Furthermore, using data from a population-level longitudinal Swedish cohort, we show the ability of the microbiome-state-stability-scores to differentiate between stable and unstable microbiome states. The research summarizes findings of a practical

deep learning approach to categorize single time-point microbiome states and facilitate imputation of rarer species. It is scalable, can enhance the data quality for subsequent studies and contribute to more reliable and precise microbiome analyses.

## LIST OF FIGURES AND TABLES

# 1. INTRODUCTION

Microbiome research has gained significant attention due to its crucial role in human health, environmental studies, and disease associations. The entire genomes of microorganisms that inhabit a particular habitat, such as soil, seawater, or the human gut, are examined in microbiome investigations. Numerous studies have established the significance of microbiomes in human bodies and natural habitats. For instance, recent research has demonstrated microbes' critical roles in complex illnesses like cancer, diabetes, obesity, and pulmonary disease. These investigations have shown the potential of human microorganisms as therapeutic targets for the treatment of disease or as biomarkers for disease detection (Amato, 2017).

Analyzing microbiome data provides valuable insights into microbial community structures, interactions, and functional potential. However, one of the significant challenges in microbiome data analysis is the high sparsity in abundance matrices, where a substantial proportion of values are recorded as zeros in taxon counts. These zeros can arise due to sequencing limitations, biological absence, or detection thresholds. This issue complicates statistical analysis, impacts downstream modeling, and may lead to unreliable or biased conclusions (Tsilimigras & Fodor, 2016).

Various imputation techniques have been explored to overcome the challenge of estimating missing or zero values in microbiome datasets. Conventional techniques have been used, including giving pseudo-count to the abundance of species, probabilistic models, k-nearest neighbors (KNN), and mean imputation. Despite being simple, these methods can unwittingly introduce noise into the data. They frequently bring errors into the data and overlook intricate microbiological connections. The reliance on these traditional methods hinders in capturing the inherent complexity and non-linearity of microbiome data. Therefore, by identifying patterns in massive datasets and producing physiologically meaningful imputations, deep learning-based techniques have recently demonstrated increased performance in managing sparse microbiome data (Liu et al., 2022).

This study presents an imputation technique for microbiome data based on deep neural networks. Our approach utilizes autoencoders to learn the underlying distribution and non-linearity of microbial abundances and predict missing values with high accuracy. Robust learning of

microbial abundance patterns was ensured by training the algorithm on a large dataset. Using this approach on a validation(unseen) dataset, we show that our imputation approach maintains the biological relevance of the data while drastically reducing sparsity. Comprehensive tests demonstrate that our method achieves the best Spearman correlation with actual data, outperforming traditional imputation strategies. To evaluate the model's performance on real world data, a simulated dataset was generated whose ground truth values were known and the Canberra distance was computed across various subsampling reads. This was done to compare our model with different existing methods namely, GemImpute (Sun & Song, 2024) and Deepimpute (Arisdakessian et al., 2019), to check the ability in restoring abundance of species in varying sequencing depth. We particularly examined the model's ability to impute probiotic species, which are biologically significant. Our solution outperformed other approaches and successfully restored their abundance characteristics.

# 2. OBJECTIVE

The key objectives of this study are as follows:

To address the widespread sparsity in microbiome abundance data by creating a deep learning-based imputation technique that is both reliable and biologically meaningful.

To assess the model's performance against current techniques and enhance the caliber of downstream analyses.

To assess the model’s effectiveness in capturing the original data’s correlation to varying sequencing depths.

Another important goal was to evaluate the model's practical usefulness by determining how well it could impute the abundance of well-established probiotic species, thereby guaranteeing the biological relevance and applicability of the imputed data.

Investigate colonization by using receptive score analysis to differentiate between persistor and non-persistor samples.

Assess microbial dynamics over time using Longitudinal data.

To facilitate the method's easy adoption by the research community, the project also aims to create an intuitive Python package.

# 3. METHODOLOGY

The approach utilizes the inherent sparsity of microbiome datasets and separates features based on abundance characteristics to allow targeted training of deep learning models for better recovery of missing values. Our workflow comprises five main steps: **Preprocessing**, **Network Architecture**, **Training Procedure**, **Imputation, and Evaluation.**

## 3.1. Preprocessing

Given the high dynamic range and sparsity of microbiome data, it was important to reduce it by feature selection. The raw abundance matrix used for this model was highly sparse which comprises 44,943 samples and 7436 species. So the goal was to select species based on detection and abundance threshold. To find the core set of species which are abundant and consistently present across different studies, a threshold framework was developed. For each species a detection matrix was constructed based on its non-zero abundance across samples within each study. Now to check the overall contribution of particular species to community composition, cumulative abundance was calculated. Species were selected based on a combination of detection and study threshold . For each threshold pair we calculated the number of species and assessed their representation by calculating the proportion of studies where these species accounted $\geq$90% or <70% of total cumulative abundance. Finally we chose the optimal number of species that were detected in a sufficient percentage of samples and represent $\geq$90% of the community in the majority of studies ensuring that the contribution is less in <70% representation.

We apply a log transformation to stabilize the variance and improve the numerical behavior of the training process. We then categorize the features (species) into two groups based on their median abundance values. Features with a median value below a defined threshold (1e-5) are classified as low-abundance, while others are treated as high-abundance features. This separation enables training specialized models to learn more effectively from distinct abundance regimes.

### 3.2. Network Architecture

For each feature group, categorized as low-abundance and high-abundance species—we design and train a separate denoising autoencoder (DAE) model to better capture the distinct characteristics of each subset. Each DAE model comprises a tailored architecture beginning with an input layer that matches the number of features within the respective subgroup. This is followed by a fully connected hidden layer with 128 neurons employing a Tanh activation function, designed to model non-linear relationships among microbial features. To ensure stable and efficient learning, batch normalization is applied immediately after the hidden layer. To mitigate the risk of overfitting, a dropout layer with a rate of 20% is introduced. Finally, the output layer mirrors the input dimension and uses Tanh activation to reconstruct the original input vector, effectively allowing the model to learn and denoise feature-specific patterns.This architecture captures explicitly nonlinear dependencies in microbiome abundance profiles while minimizing error.

Although traditional denoising autoencoders inject artificial noise during training, we do not explicitly corrupt or add noise to the input data; however, microbiome data's **intrinsic sparsity** and **zero-inflation** act as a natural noise source. Thus, our autoencoder learns to reconstruct clean and complete abundance profiles from naturally corrupted inputs. This aligns with the motivation behind denoising autoencoders to capture accurate signals from noisy observations.

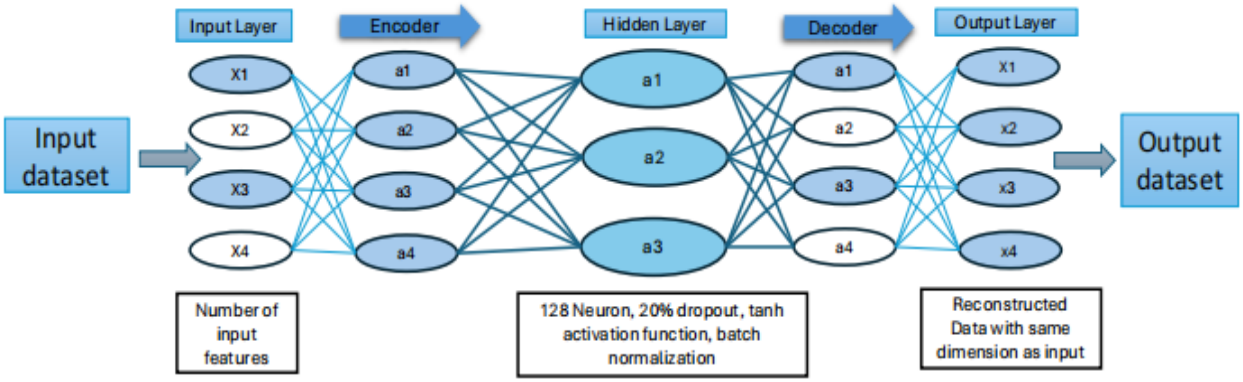



Fig. 3.2 Model Architecture

### 3.3. Training Procedure

The training data for each feature group is divided into training (80%) and testing (20%) subsets. We train each model using the **Adam optimizer** with a learning rate of 0.001 and a **batch size 32**, minimizing the **mean squared error (MSE)** between the input and reconstructed output.

To prevent overfitting and reduce training time, we use **early stopping** with a patience of 10 epochs based on validation loss, and a **learning rate scheduler** that reduces the learning rate by half if the validation loss plateaus. Each model is trained for a maximum of 500 epochs, although training often converges earlier due to early stopping.

The final trained models are saved separately for the low-abundance and high-abundance feature groups to facilitate their use during the imputation phase. Alongside the models, the feature categorizations for both groups are also stored. This allows the imputation process first to determine the abundance category of each species to be imputed and then select the corresponding trained model, either for low or high abundance, to perform the imputation. This targeted approach ensures that each species is imputed using the model best suited to its abundance characteristics, enhancing imputation accuracy and preserving biological relevance.

### 3.4. Imputation

Once trained, the DAE models impute missing or zero values in the **test data**. The trained models independently predict reconstructed abundance profiles for each feature group. All features(species) are listed down from the data. Any feature that belongs to the low category is imputed using a low DAE model whereas a feature belonging to the high category is imputed using a high model.

All negative values in the imputed matrix are zero to ensure biological plausibility, as negative microbial abundances are not biologically meaningful or interpretable. The imputed outputs from the low- and high-abundance models are then concatenated and reordered to align with the original feature structure. To maintain the general integrity and structure of the data, the original values of features that are missing from the trained models are kept as it is. Lastly, the complete

dataset undergoes an exponential transformation to undo the previous log transformation and return the abundance values to their initial scale.

The fully imputed dataset,whose structure will be the same as the original dataset, is saved for analyses.

### 3.5. Evaluation

#### 3.5.1 Correlation Between Original and Imputed Values

To evaluate the accuracy and reliability of the imputation process, we assessed the closeness of imputed values to original values using Spearman correlation. Spearman correlation is a non-parametric test that examines whether two variables are correlated with one another or not. Since microbiome data does not follow normal distribution and contain outliers, this test was perfect for evaluation. We could assess how well the model preserves the underlying data structure by computing the correlation between the original (non-zero values that were intentionally made zero) and their corresponding imputed values for each species. A higher correlation coefficient indicates a stronger agreement between the imputed and actual values, reflecting the model's ability to generate consistent imputations across different abundance levels.

#### 3.5.2 Bray-Curtis Similarity Calculation

For another statistical analysis, we calculated the Bray-Curtis similarity between the original and imputed abundance profiles. Bray-Curtis similarity is a metric used mostly to quantify the similarity between two samples. When comparing community compositions, such as species abundances in various environments, it is quite helpful. For microbiome data this could be a well-suited measure ,as it accounts for both the presence and relative abundances of taxa, making it a robust choice for evaluating changes introduced during imputation. The range of this similarity score is 0 to 1 and high value indicates that the overall community composition was well-preserved during the imputation process.

### 3.5.3 Validation of DAE Method with Other Imputation Methods

To check the performance of the proposed Denoising Autoencoder (DAE)-based imputation approach, we conducted a comparative analysis with other established methods commonly used for imputing biological and microbiome data. Literature review provided two representative methods that were selected for comparison: **GemIMP** and **DeepImpute**. These techniques were created especially to deal with the complexity and sparsity of high-dimensional biological information.

Each method, including our DAE approach, was used on the same data under identical preprocessing and evaluation conditions. **Spearman correlation** and **Bray-Curtis similarity** between the original and imputed values were used to assess the quality of imputation in each method. Results suggested that the DAE method consistently achieved **higher correlation coefficients and similarity scores** than GemIMP and DeepImpute. This proves that our method is more effective in capturing the underlying structure of microbial abundance data and generating biologically meaningful imputations. These findings demonstrated the robustness and effective performance of the DAE-based method.

### 3.5.4 Evaluation Using Simulated Microbiome Data

The next goal was to assess the effectiveness of the model in maintaining the originality of data. The simulated microbiome sequencing data was generated under controlled conditions,using *InSilicoSeq*. This is widely used as a sequencing simulator that mimics the profile of real Illumina sequencing platform. This is a tool in python which requires known genomes and their abundance to be simulated.

#### a) Generation of Simulated Dataset

We selected microbial species with available genome sequences to simulate realistic data. These species were selected having a mix of low and high abundance observed in microbiome samples. Microbiome selection was done from five diseases: Parkinson's Disease (PD),Type 2 Diabetes (T2D), Colorectal Cancer (CRC), Cardiovascular Disease (CVD), and Inflammatory Bowel Disease (IBD). Ten microbiomes were obtained by selecting the top two samples for each disease condition based on the highest percentage of observed abundance. Using *InSilicoSeq*,

paired-end sequencing reads were simulated for each sample by predefined detected abundance levels. Each microbiome produced 20 million reads, each having a certain number of reads according to its relative abundance (20 million × abundance of species).The master data was created containing reads of these 10 samples. This simulation ensured access to ground truth abundance values, enabling assessment of the imputation model.

**b) Profiling with MetaPhlAn 3**

Generated simulated reads by *InSilicoSeq* were profiled using **MetaPhlAn 3**, a tool for species-level microbial community profiling based on clade-specific marker genes. MetaPhlAn 3 was chosen due to its high specificity and accuracy, especially in low-abundance contexts. The tool was executed to generate relative abundance profiles for the simulated species.

**c) Role of Simulated Data in Evaluation**

The simulated dataset served as a benchmark for model evaluation, as the actual abundance values were known. The master dataset was subsampled at three different levels: 10%, 20%, and 30% of the original read depth to replicate real-world sequencing restrictions. The DAE model's performance was assessed using this subsampled dataset and subjected to imputation using the GemIMP and DAE techniques.

The Canberra distance between the imputed abundances and the original (ground truth) dataset for each approach was calculated. The ability of each model to retain the initial distribution of species abundance at low sequencing depth was measured by this parameter. When compared to GemIMP, the DAE displayed smaller Canberra distances, suggesting improved accuracy and robustness of the method.

### 3.6 Understanding Microbiome States

#### 3.6.1 Evaluation of Imputation on Probiotic Species

To further assess the effectiveness of the imputation strategy, we conducted a targeted evaluation on a subset of species known to be probiotics. For this purpose probiotic species were listed down and the model was again trained by ensuring these species are present in the data so that

our model can learn the abundance pattern in probiotic species. Then, we extracted the taxonomic abundance profiles for these probiotic species from the validation dataset and artificially introduced zeros to mimic sparsity.

The imputation process used separately pre-trained denoising autoencoder (DAE) models for low- and high-abundance features including all probiotics.

Spearman correlation and P value were computed between the original (pre-zeroed) and imputed values for each probiotic species. This provided insights into how well the model could reconstruct accurate abundance signals under sparsity conditions.

This focused evaluation on probiotic species demonstrated the practical utility of the imputation method in improving downstream biological interpretation, especially in health-related microbiome research where accurate quantification of beneficial microbes is critical.

**3.6.2 Receptive Score-Based Identification of *Bifidobacterium longum* Persistence Using Rank-Scaled Microbiome Profiles:** To investigate the colonization dynamics of *Bifidobacterium longum* in the human gut microbiome, we utilized data from GomezM_2016 study, which includes microbiome samples collected at baseline (timepoint zero). The original microbial abundance profiles were first processed to obtain rank-scaled values. This transformation normalized abundance values between 0 and 1 based on their ranks, enabling comparison across samples. We applied this scaling separately to both the original and imputed datasets. The imputed values were generated using our custom-trained model designed to reconstruct missing or sparse entries in microbiome datasets.Focusing on the species *Bifidobacterium longum*, we extracted its rank-scaled values from both the original and imputed datasets. A new metric, termed the **receptive score**, was then calculated as the difference between the imputed and original rank-scaled values for each sample. This score quantifies how much a species' inferred abundance increases upon imputation, indicating its potential to be underrepresented in the original profile due to sparsity or dropout effects.We hypothesized that samples in which *B. longum* genuinely persists (termed **persistors**) would exhibit **higher receptive scores**, while those where the species is transient or non-colonizing (**non-persistors**) would show lower scores. Using predefined sample groups based on metadata (e.g., samples

starting with IDs A, B, G, H, I, J), we classified persistors and non-persistors and visualized their receptive scores via a box plot.

**3.6.3 Evaluation of Microbiome State Stability Using Variance in Receptive Scores of Longitudinal data :**We used longitudinal microbiome data from a population-level Swedish cohort to assess the microbiome-state-stability scores' capacity to differentiate between stable and unstable microbial communities. The fluctuation in each person's receptive scores over several time points was calculated to determine the stability score. This variation provides a quantifiable indicator of microbiome stability at the individual level by reflecting the degree to which a species' responsiveness (as measured by the receptive score) varies over time.

**6. Software Availability of Imputation Framework for Public Use :**

We have developed a Python package on the Python Package Index (PyPI) . This package lets users easily and minimally set up the Denoising Autoencoder (DAE)-based imputation technique on their microbiome datasets. Following installation with a straightforward pip install command, users can utilize a command-line interface or a single-line function call to impute their input data and get the missing values imputed in their abundance matrices. The package has user-friendly architecture and includes built-in preprocessing processes, model loading, features loading and imputation code. Instructions are also offered to help users to incorporate this technique into their workflows for microbiome investigation. The imputation approach is readily adopted across research by making it a reusable and open-source tool, guaranteeing uniformity and reproducibility in preparing microbiological data.

# 4. RESULTS

**4.1 Scatter Plot between original and imputed values of Test dataset: S**catter plot comparing the imputed and original values on the test dataset was created to evaluate the accuracy of the imputation carried out by the model. Each point in the scatter plot represents a species abundance value, with the x-axis showing the initial (non-zero) values and the y-axis showing the matching imputed values**.** The Spearman correlation coefficient was computed to measure this visual agreement, indicating a strong monotonic relationship. The capacity of the model to maintain the biological hierarchy of microbial abundances during the imputation is further supported by this.

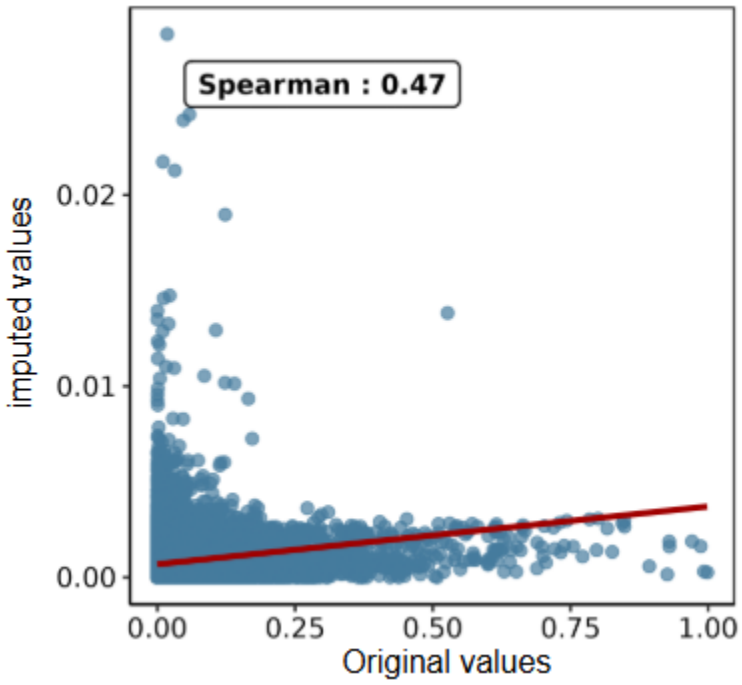



Fig.4.1 Scatter plot between original and imputed values of Test Data

**4.2 Comparison with Existing Imputation Methods:** To determine our model's performance, we compared it with GemImpute, a technique designed for microbiome data, and DeepImpute, which was initially created for scRNA-seq but is based on a similar deep learning architecture. Our DAE model performed better than both approaches, as seen by scatter plots and Spearman correlations. It achieved a greater correlation and showed tighter alignment with the original values, indicating higher accuracy in reconstructing biologically essential patterns.

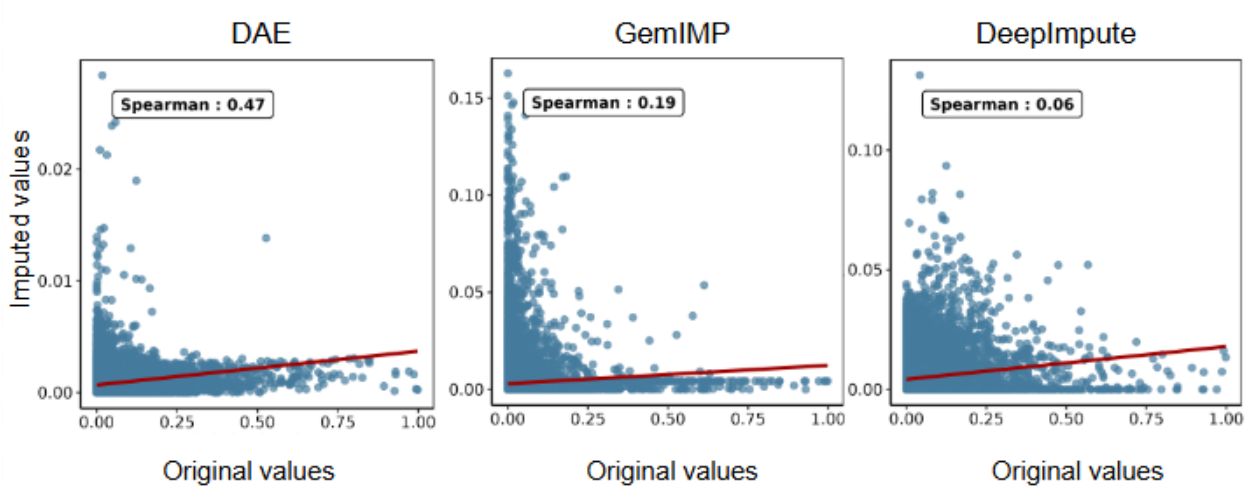



Fig. 4.2 Scatter plot showing comparison between different methods

**4.3 Bray-Curtis Similarity for All Three Methods:** The similarity between the original and imputed values was computed for each method to assess the imputation quality. This similarity metric measures the compositional similarity between two samples and is frequently employed in ecological and microbial studies. The results demonstrated the DAE model's capacity to maintain the general structure and biological composition during imputation, consistently obtaining higher Bray-Curtis similarity scores than the other two approaches.

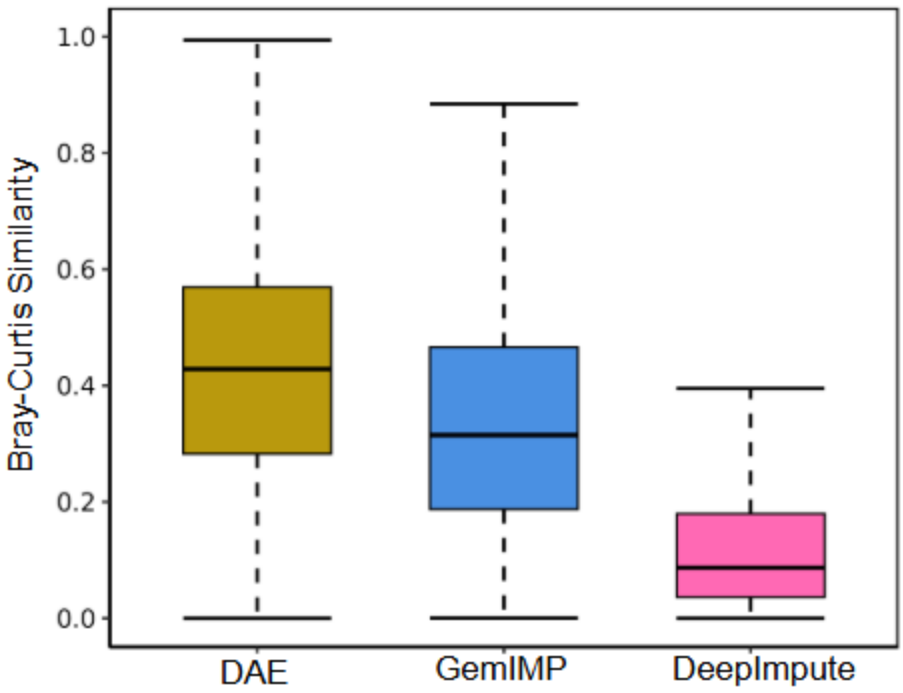



Fig.4.3 Box Plot for Bray-Curtis Similarity

**4.4 Scatter Plot and Correlation on Unseen Validation Dataset**: To evaluate the model's generalizability, imputation was performed using an unseen validation dataset entirely different from the training data. The model was applied straight to this dataset. The performance was assessed by creating a scatter plot between the imputed and original values. Spearman correlation was also computed to measure the link between the actual and imputed abundances. The strong correlation and aligned scatter distribution validated the model's ability to impute missing values. It has also proved that our model can perform better than other existing imputation methods when it comes to explicitly different dataset which is highly sparse. Our model is scalable so irrespective of the dimension of the dataset,it effectively imputes the missing values.

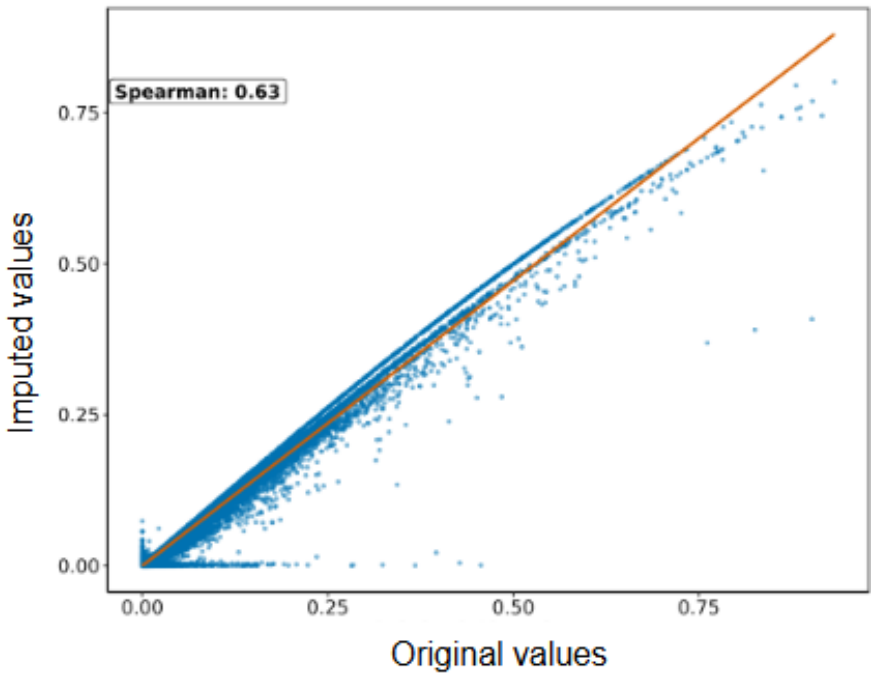



Fig.4.4 Scatter Plot for Validation dataset

**4.5 Bray-Curtis Similarity Across Different Abundance Ranges for All Methods:** Bray-Curtis similarity was computed for several ranges of original abundances for each of the three methods to obtain a better understanding of the imputation performance across varied abundance levels. We assessed how effectively each approach maintained community composition in low and high abundance values.

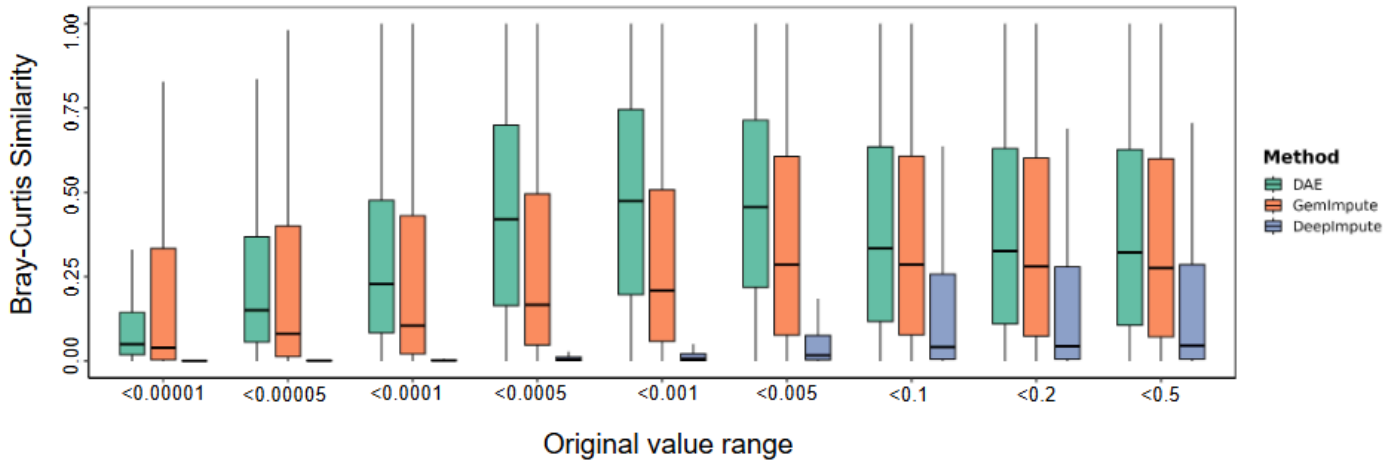



Fig.4.5 Bray-Curtis Similarity for different ranges

**4.6 Relationship Between Species Sparsity & Mean Imputed:** The correlation of species sparsity and the mean imputed values was computed to investigate how data sparsity affects the imputation outcomes. This approach to estimate values for rarely observed species is reflected in the tendency for species with higher sparsity to have lower mean imputed value. It is supported by this relationship since low abundance shouldn't be overinflated during imputation.

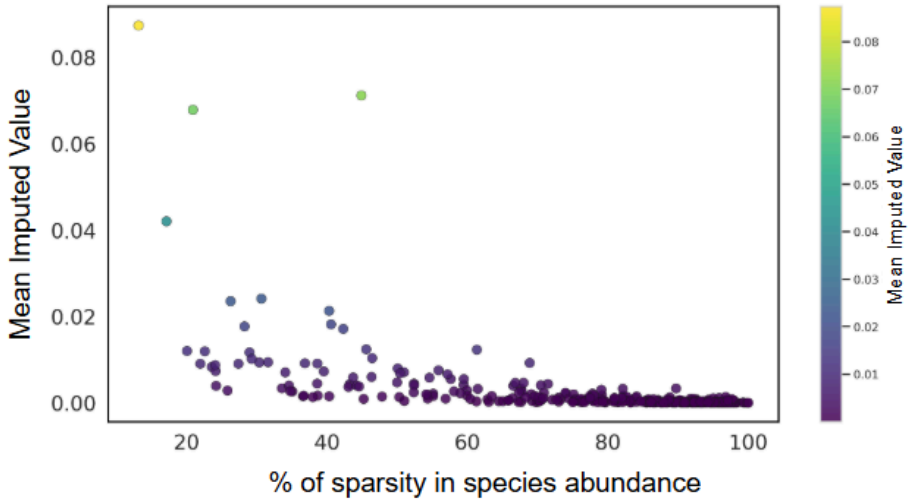



Fig. 4.6 Scatter Plot for the relationship between Sparsity & Imputed Abundance

**4.7 Evaluation of DAE Across Varying Sequencing Depths Using Simulated Data:** 10%, 20%, and 30% subsampling of reads from the simulated data were used for analyses. We used both GemIMP and DAE for imputation. Further, we computed the Canberra distance between the imputed and ground truth data. DAE consistently produced smaller distances, demonstrating better efficacy in recovering microbial abundance under sequencing situations with restricted resources.

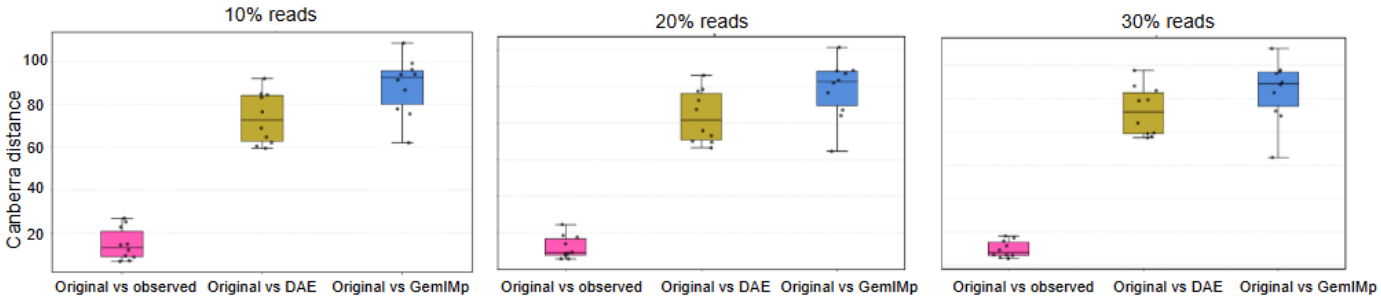



Fig.4.7 Canberra distance Across Varying Sequencing Depths

**4.8 Correlation and P value calculation for Probiotic Species:** Spearman correlation between the original and imputed values for chosen probiotic species to assess the model's biological significance. Compared to GemIMP, the DAE model produced more significant p-values and stronger correlation coefficients, suggesting that we better recover these biologically critical microorganisms' meaningful abundance patterns.

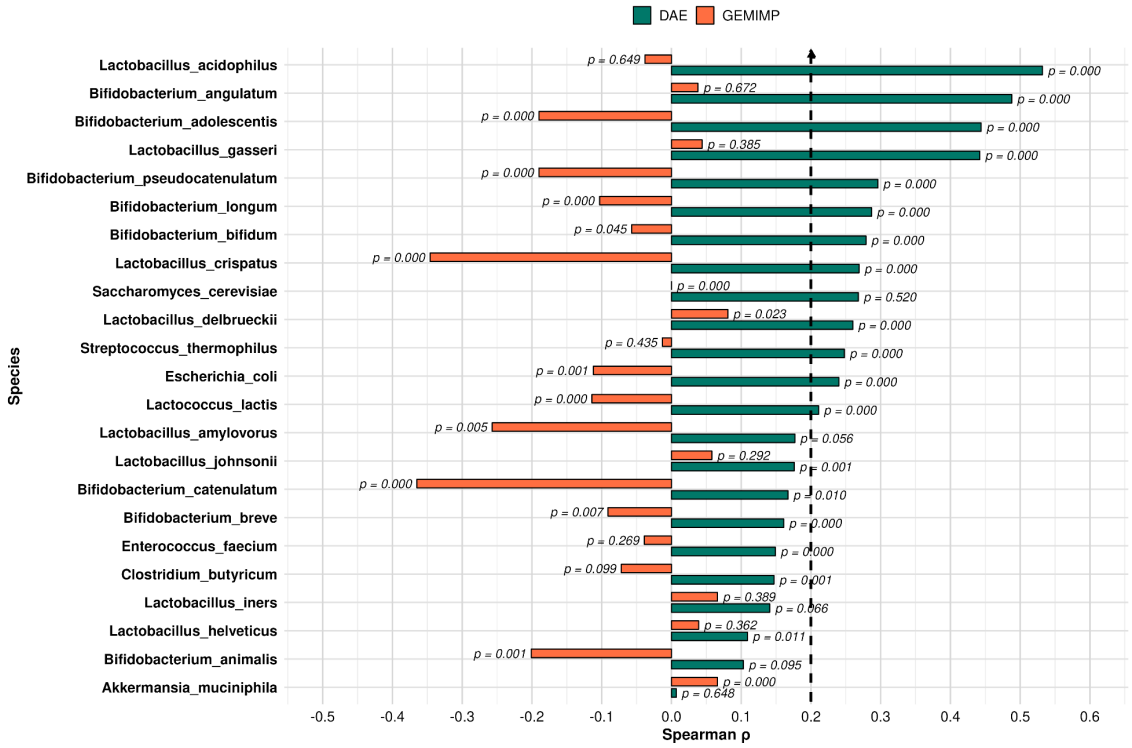



Fig.4.8 Spearman correlation and P value comparison between DAE and GemImp

**4.9 Differentiation of *Bifidobacterium longum* Persistence Using Receptive Scores Derived from Rank-Scaled Microbiome Profile:** A box plot comparing receptive scores of persistors and non-persistors revealed a clear distinction between the two. Persistor samples showed consistently higher receptive scores, indicating that *B. longum* was more substantially "recovered" or inferred by the imputation model in these individuals. In contrast, non-persistors exhibited lower receptive scores, suggesting minimal change upon imputation and thus potentially transient or absent presence of *B. longum*.

To statistically validate this observation, we performed a Wilcoxon rank-sum test (Mann–Whitney U test), which confirmed a significant difference between the two groups ($p <$

0.05). These findings demonstrate that receptive scores derived from rank-scaled imputed data can serve as an effective marker for distinguishing between persistent and non-persistent microbial colonization, reinforcing the potential of our approach in microbiome-based stratification and probiotic evaluation**.**

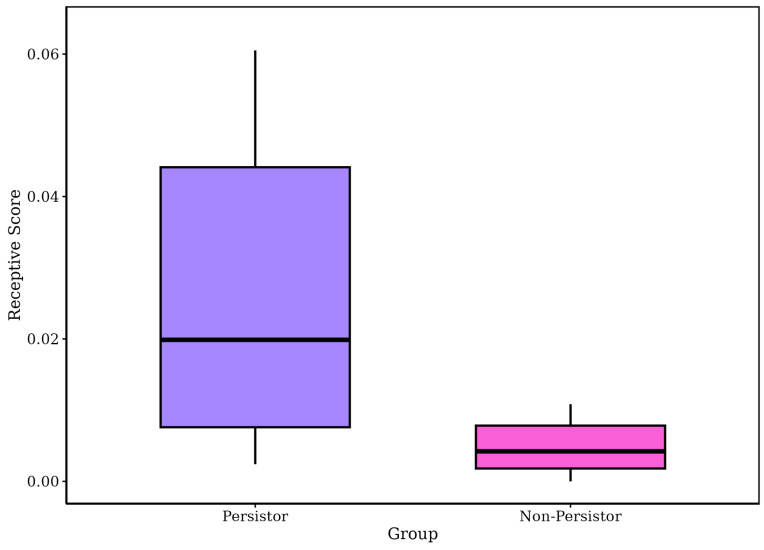



Fig.4.9 Receptive score distribution distinguishing persistent and non - persistent

**4.10 Relationship between Receptive Score Variance and bray curtis distance in longitudinal data :** We evaluated the variance in receptive scores between time periods with Bray-Curtis dissimilarities to evaluate the connection between changes in species-level responsiveness and general differences in microbiome composition. To investigate this relationship, a box plot visualization was created.

Individuals with higher shifts in microbial responsiveness also showed greater compositional instability over time which showed a strong positive association between the variance in

receptive scores and Bray-Curtis distances. Both the 16S and WGS datasets showed this tendency, demonstrating the state-stability score's resilience and cross-platform applicability.

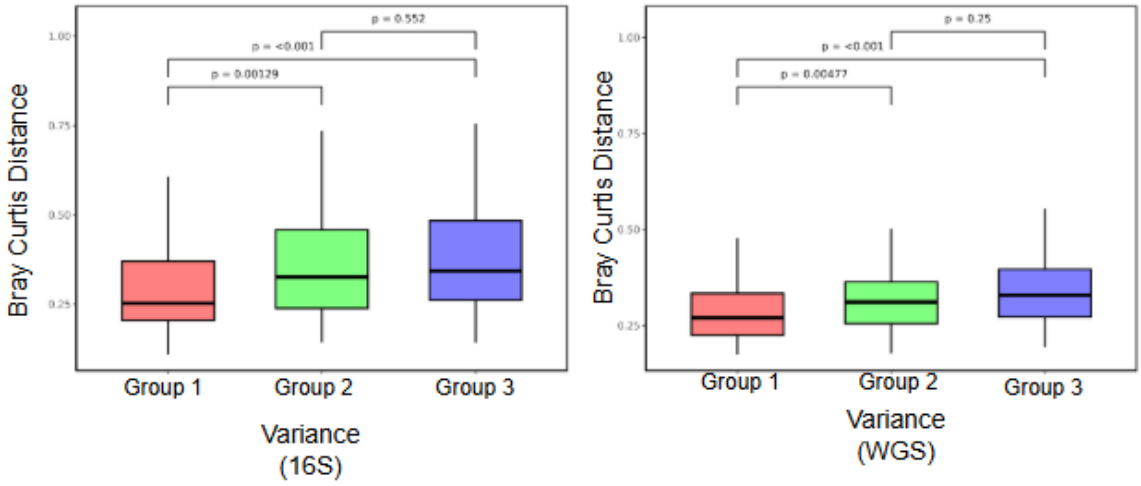



Fig. 4.10 Box plot showing relationship between variance and bray curtis distance

## 5. Discussion and Conclusion

In this study we have addressed a fundamental challenge in microbiome data analysis. The high sparsity of abundance species profile, uneven sampling depths and challenge in recovering the abundance of taxonomic units limits and hinders meaningful biological interpretation. This restricts functional biological interpretation in different fields of microbiome, and is a fundamental problem in data analysis. To address these issues we have developed a deep learning-based imputation method and maintained biological consistency while significantly enhancing the quality and completeness of species profiles. Our model, based on Denoising Autoencoder (DAE) architecture, showed excellent generalizability across various datasets including test and validation dataset and successfully described the intricate structure of microbial communities. Multiple assessments were used to confirm the model's performance, including comparisons with current imputation methods that are used in imputation of biological data and correlation analyses with recognized biological signals like probiotic species. Our strategy consistently outperformed competing techniques to restore lost biological signals and maintain species abundance connections throughout all evaluations. Simulations replicating real-world sequencing settings further demonstrated the model's efficacy, maintaining high accuracy even at lower read depths. This work offers a reliable and expandable imputation method that is explicitly designed for microbiome research. Our model's improved data quality can help with more trustworthy downstream analysis, including microbiome-based diagnostic and differential abundance studies. To gain a better understanding of microbial dynamics and host interactions, future research could investigate extending this imputation technique to longitudinal microbiome datasets or incorporating it into multi-omics pipelines.

# 6 . References


1. Amato, K. R. (2017). An introduction to microbiome analysis for human biology applications. *American Journal of Human Biology*, *29*(1), e22931.
2. Tsilimigras, M. C., & Fodor, A. A. (2016). Compositional data analysis of the microbiome: fundamentals, tools, and challenges. *Annals of epidemiology*, *26*(5), 330-335.
3. Liu, J., Pan, Y., Ruan, Z., & Guo, J. (2022). SCDD: a novel single-cell RNA-seq imputation method with diffusion and denoising. *Briefings in Bioinformatics*, *23*(5), bbac398.
4. Jiang, R., Li, W. V., & Li, J. J. (2021). mbImpute: an accurate and robust imputation method for microbiome data. *Genome biology*, *22*(1), 192.
5. Arisdakessian, C., Poirion, O., Yunits, B., Zhu, X., & Garmire, L. X. (2019). DeepImpute: an accurate, fast, and scalable deep neural network method to impute single-cell RNA-seq data. *Genome biology*, *20*, 1-14.
6. Sun, Z., & Song, K. (2024). GEMimp: An Accurate and Robust Imputation Method for Microbiome Data Using Graph Embedding Neural Network. *Journal of Molecular Biology*, *436*(23), 168841.
7. He, M., Zhao, N., & Satten, G. A. (2024). MIDASim: a fast and simple simulator for realistic microbiome data. *Microbiome*, *12*(1), 135.