Can Self-Supervised Pretraining Reveal Patterns in Microbiome Data?

VECTOR INSTITUTE

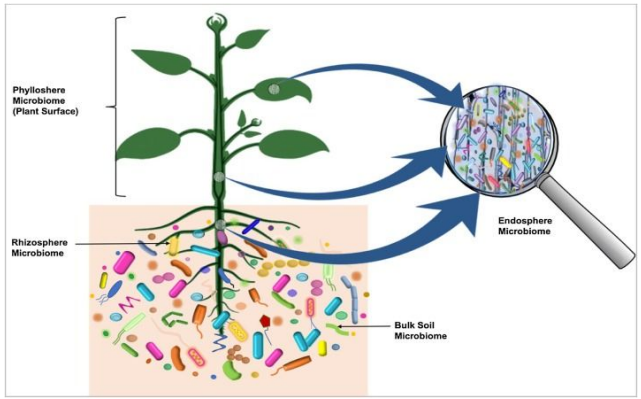
Kevin Chen

Contributors: Kevin Deng, David Pellow, Rahul G. Krishnan, Michael Brudno

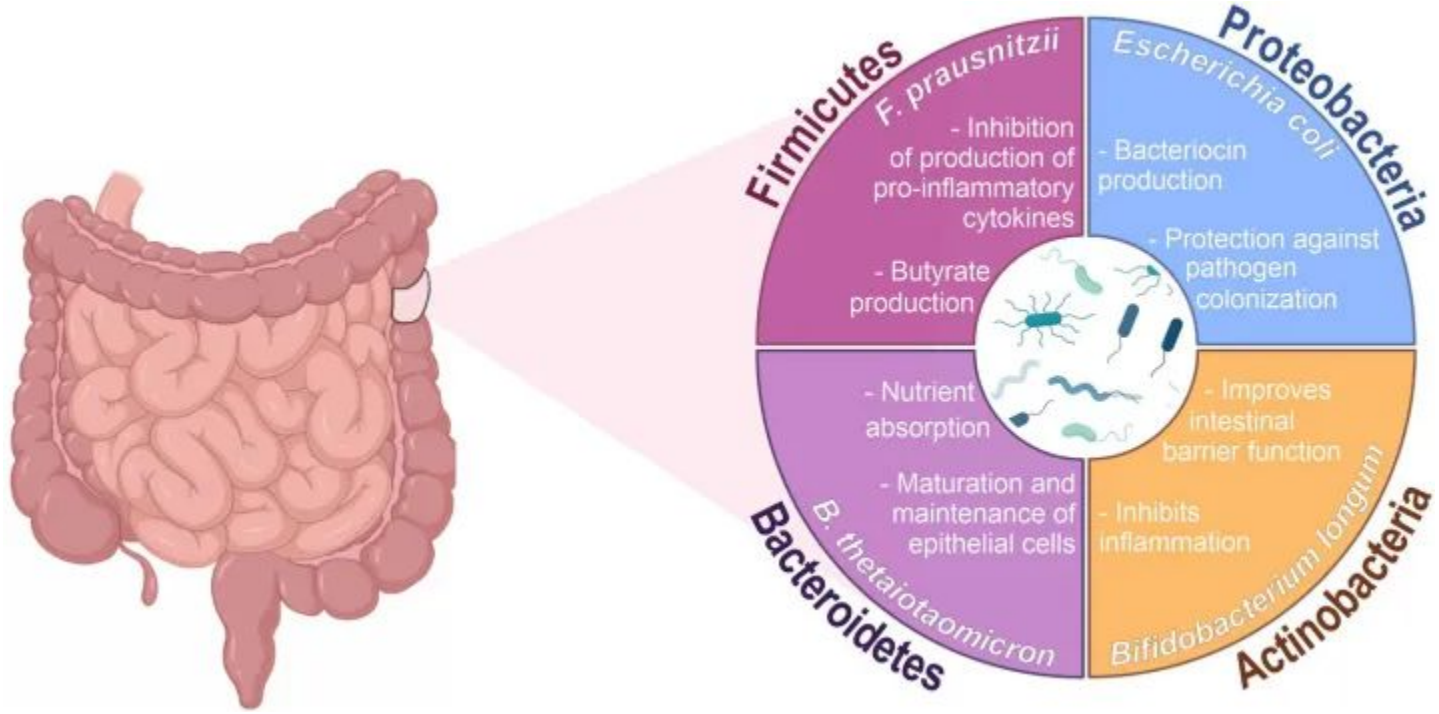
What is a **Microbiome?**

A **Collection of microorganisms** living in a community

Examples:
Kitchen Sink, Soil,
Human Gut Microbiome



What is the **Human Gut Microbiome?**



doi.org/10.3389/fmicb.2023.1098386

What is the **Human Gut Microbiome?**

Relevant for:

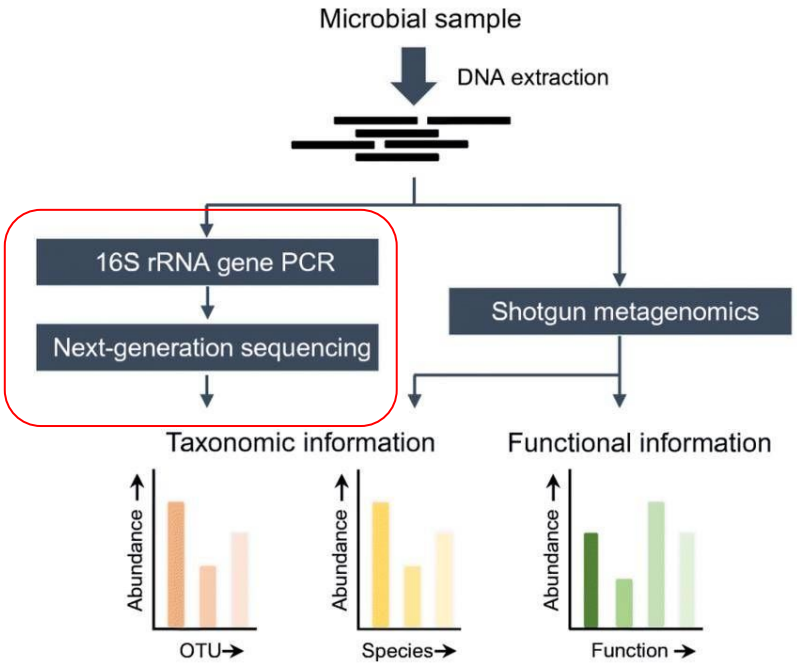
Immune System

Autoimmune Conditions (e.g. Inflammatory Bowel Disease)

Nervous System

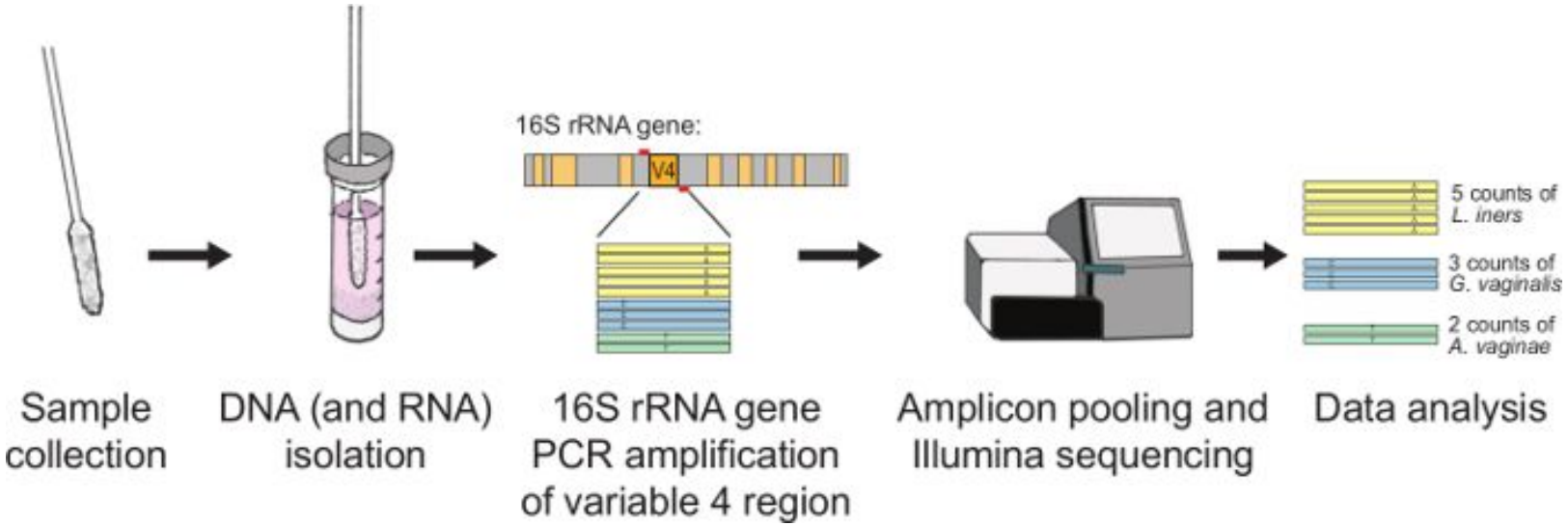
doi.org/10.3389/fmicb.2023.1098386

How is Human Gut Microbiome **Measured?**



https://encyclopedia.pub/entry/56798

How is Human Gut Microbiome **Measured?**



https://www.jove.com/t/53939/efficient-nucleic-acid-extraction-16s-rrna-gene-sequencing-for

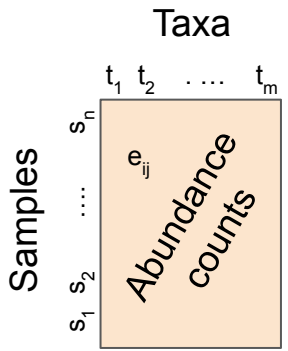
Characteristics of Microbiome Data

High Dimensional

Sparse

Tabular/
Unordered

After sequencing, data is presented in a **SAMPLE x TAXA** format



Characteristics of ~~Microbiome~~ Data

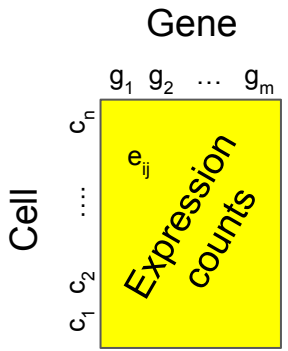
Single-Cell RNAseq

High Dimensional

Sparse

Tabular/
Unordered

After sequencing, data is presented in a **CELL x GENE** format



Previous Work in Single Cell Transcriptomics

nature machine intelligence

Article

https://doi.org/10.1038/s42256-022-00534-z

scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data

Received: 3 February 2022

Accepted: 19 August 2022

Fan Yang[1,7], Wenchuan Wang[1,2], Junzhou Huang[5], Hui Lu[2,6]

nature methods

Article

https://doi.org/10.1038/s41592-024-02201-0

scGPT: toward building a foundation model for single-cell multi-omics using generative AI

Received: 12 July 2023

Accepted: 30 January 2024

Haotian Cui[1,2,3,8], Chloe Wang[1,2,3,8], Hassaan Maan[1,3,4], Kuan Pang[2,3], Fengning Luo[2,3], Nan Duan[5] & Bo Wang[1,2,3,4,6,7]

What is a **Foundation Model?**

Self-supervised training, **Downstream Task** evaluation

Scalability is essential

Various models **exist** in different tasks: Language and omics

Previous Foundation Model work in Metagenomics

PLOS Computational Biology

Browse Publish About

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Learning a deep language model for microbiomes: The power of large scale unlabeled microbiome data

Quintin Pope, Rohan Varma, Christine Tataru, Maude M David, Xiaoli Fern

Published: May 7, 2025 • https://doi.org/10.1371/journal.pcbi.1011353

Article | Authors | Metrics | Comments | Media Coverage | Peer Review

- Abstract
- Author summary
- 1 Introduction
- 2 Materials and methods
- 4 Conclusion and future work
- Supporting information
- References
- Reader Comments
- Figures

Abstract

We use open source human gut microbiome data to learn a microbial "language" model by adapting techniques from Natural Language Processing (NLP). Our microbial "language" model is trained in a self-supervised fashion (i.e., without additional external labels) to capture the interactions among different microbial taxa and the common compositional patterns in microbial communities. The learned model produces contextualized taxon representations that allow a single microbial taxon to be represented differently according to the specific microbial environment in which it appears. The model further provides a sample representation by collectively interpreting different microbial taxa in the sample and their interactions as a whole. We demonstrate that, while our sample representation performs comparably to baseline models in in-domain prediction tasks such as predicting Irritable Bowel Disease (IBD) and diet patterns, it significantly outperforms them when generalizing to test data from independent studies, even in the presence of substantial distribution shifts. Through a variety of analyses, we further show that the pre-trained, context-sensitive embedding captures meaningful biological information, including taxonomic relationships, correlations with biological pathways, and relevance to IBD expression, despite the model never being explicitly exposed to such signals.

bioRxiv

THE PREPRINT SERVER FOR BIOLOGY

HOME | SUBMIT | FAQ | BLOG | ALERT

Search

New Results

Follow this preprint Previous

BiomeGPT:A foundation model for the human gut microbiome

Posted January 05, 2026.

Nicholas A. Medearis, Siyao Zhu, Ali R. Zomorrodi

doi: https://doi.org/10.64898/2026.01.05.697599

This article is a preprint and has not been certified by peer review [what does this mean?].

Download PDF

Print/Save Options

Abstract | Full Text | Info/History | Metrics | Preview PDF

Abstract

The human gut microbiome encodes rich information about host health, yet current analysis pipelines remain narrowly optimized for individual tasks. This limits our ability to gain a thorough view of how the microbiome impacts health and disease. Here we introduce BiomeGPT, a transformer-based foundation model pretrained on over 13,300 human gut metagenomes spanning 32 phenotypes—including healthy and 31 diverse diseases—to learn context-aware, species-level gut microbiome representations. The model captures quantitative compositional structure and intricate cross-species dependencies embedded within community profiles. When fine-tuned for predicting host health status, BiomeGPT accurately distinguishes healthy from diseased microbiomes and resolves individual disease states across a broad

Subject Area

Bioinformatics

Reviews and Co

2 Comment

0 TRIP Peer Re

0 Community

1 Automated S

0 Blogs/Media

Doubts…

Research Article

Systematic benchmarking of foundation models and classical baselines for microbiome-based disease prediction

Jin Mu, Zheng-Zheng Tang, Guanhua Chen

Conclusions: In this large-scale benchmark, current foundation-model approaches offer, at best, modest gains over strong classical baselines for microbiome-based disease prediction. Our results highlight that standard numerical representations remain difficult to beat, general-purpose tabular foundation models can provide

https://www.researchsquare.com/article/rs-8912605/v1*

* preprint, currently under review.

Doubts…

Effects of data transformation and model selection on feature importance in microbiome classification data

Research | Open access | Published: 04 January 2025

Volume 13, article number 2, (2025) Cite this article

You have full access to this open access article

Download PDF Save article

Zuzanna Karwowska, Oliver Aasmets, Estonian Biobank research team, Tomasz Kosciolek & Elin Org

Microbiome data transformations can significantly influence feature selection but have a limited effect on classification accuracy. Our findings suggest that while classification is robust across different transformations, the variation in feature selection necessitates

Investigative Goals

Is predictive signal in microbiome data fundamentally limited?

Do complex models meaningfully outperform strong baselines?

Data we use: HMC

Human Microbiome Compendium (HMC):

- Collection of **~450** human gut microbiome **studies**
- After pre-processing: **~120K samples**
- **16S** sequencing, **uniform data processing pipeline**
- Resolved to genus level only (**~1500 taxa** after filtering)
- Varying **metadata**, by study (most studies include at least geographic location)

Assessment of performance - use metadata for prediction tasks

➢ e.g. predict the geographic region (continent) the sample came from

Human Microbiome Compendium

Cell
A Cell Press journal

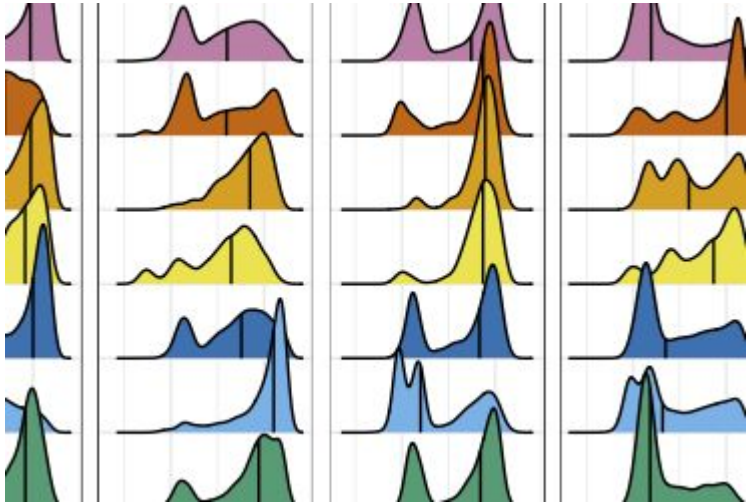
This journal Journals Publish News & events About Cell Press

RESOURCE · Volume 188, Issue 4, P1100-1118.E17, February 20, 2025 · *Open Access* Download Full Issue

Integration of 168,000 samples reveals global patterns of the human gut microbiome

Richard J. Abdill [1,7] · Samantha P. Graham [2,7] · Vincent Rubinetti [3,4] · … · Casey S. Greene [3,4] · Sean Davis [3,4] · Ran Blekhman [1,8] … Show more

Affiliations & Notes Article Info



Data we use: Human Microbiome Compendium

PRJNA436359

PRJNA436359

PRJNA436359

…

~450 Studies

Bacteria.Pseudomonadota .Alphaproteobacteria…	Bacteria.Bacillota.B acilli…	Bacteria.Bacillota .Cloistridia…	…
16	0	0	…
0	0	2	…
…			

~120k samples

Data we use: Human Microbiome Compendium

16S Sequencing with Uniform Pipeline

Resolved To Genus Level

~120k samples

Bacteria.Pseudomonadota.Alphaproteobacteria…	Bacteria.Bacillota.Bacilli…	Bacteria.Bacillota.Cloistridia…	…
16	0	0	…
0	0	2	…
…			

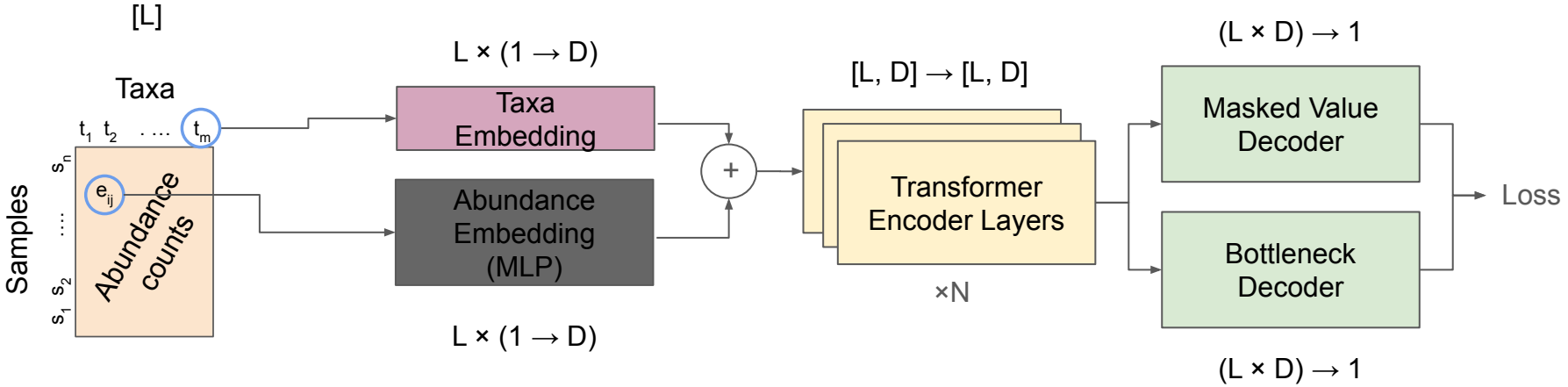
Experiment 1: Masked Autoencoder Architecture

Experiment 1: Components

1. Architecture
2. Evaluation
3. Results

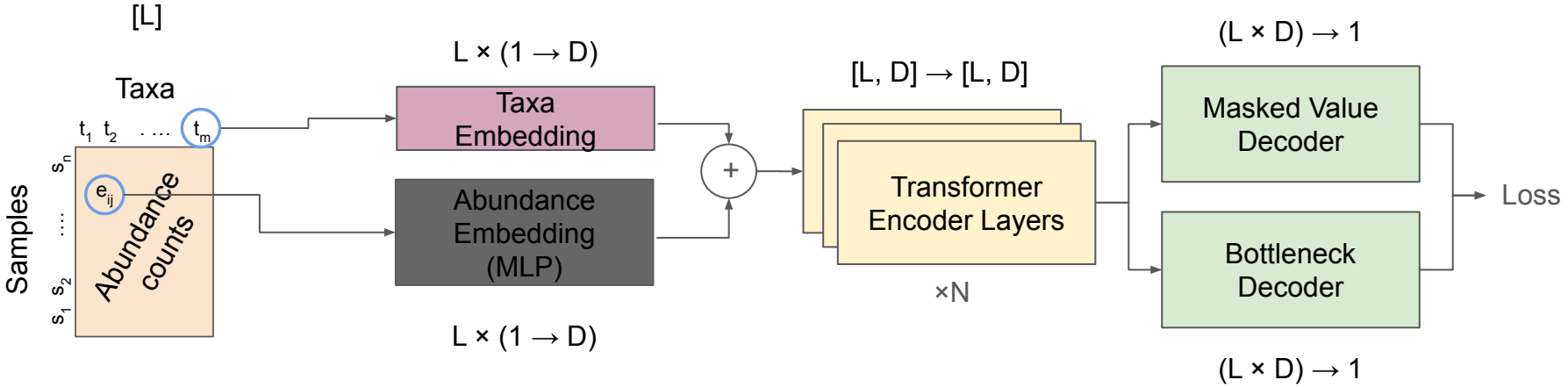
Architecture

Architecture in Detail

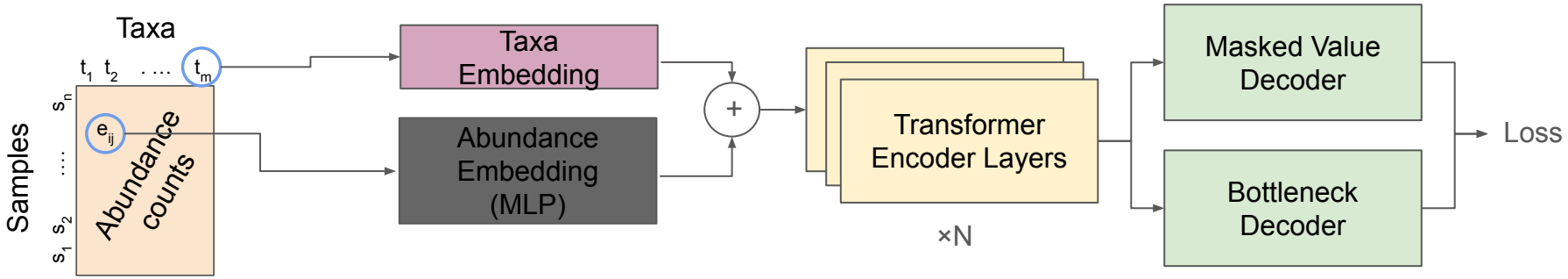


scGPT-style architecture

Architecture in Detail

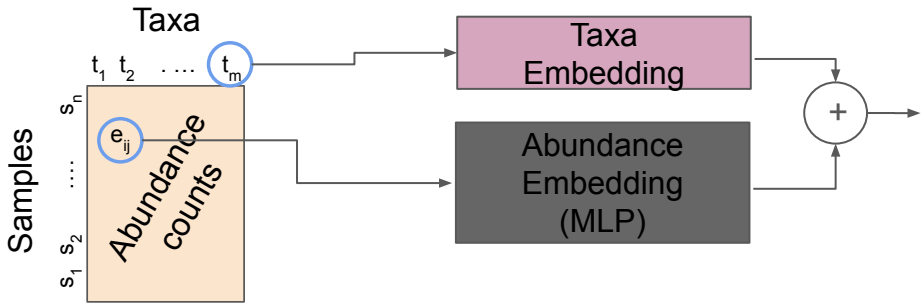


Architecture in Detail

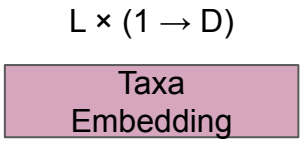


scGPT-style architecture

Architecture in Detail



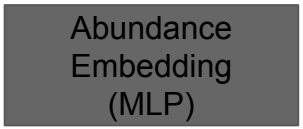
Architecture in Detail



Stores each taxa as a d-dimensional vector

Currently: lookup table

Architecture in Detail

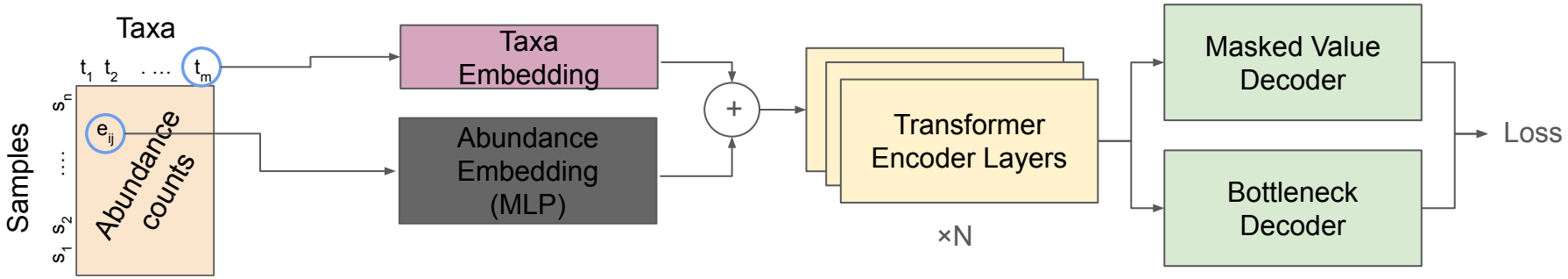


$L \times (1 \rightarrow D)$

Maps abundance values to D-dimensional vector.

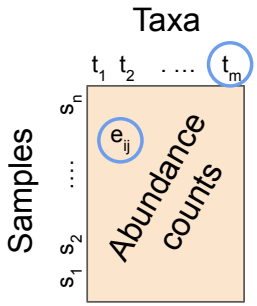
Data Normalization: **Quantile Binning**

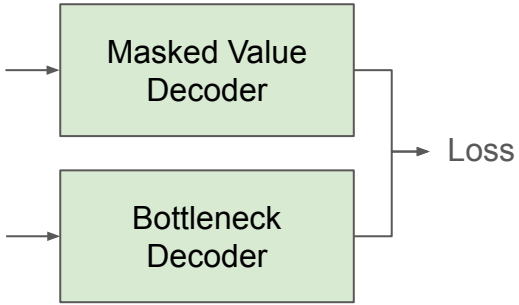
Architecture in Detail



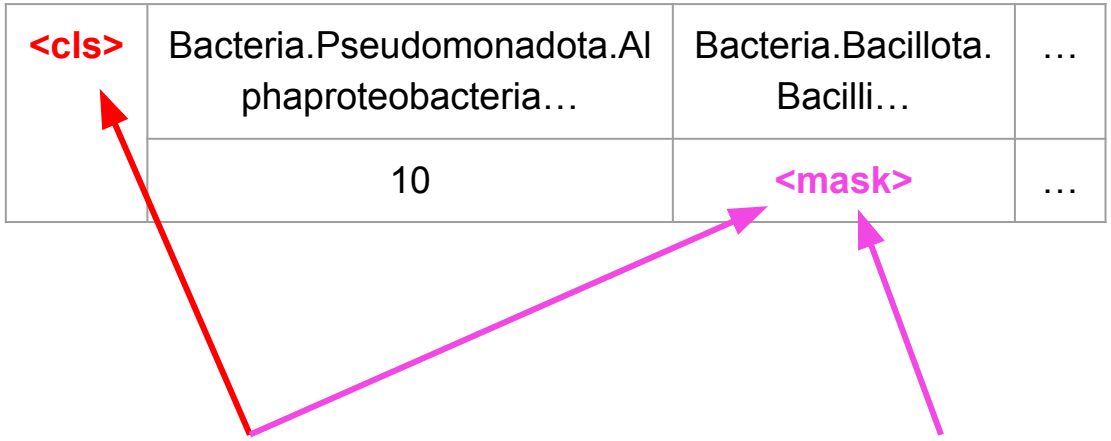
scGPT-style architecture

Architecture in Detail





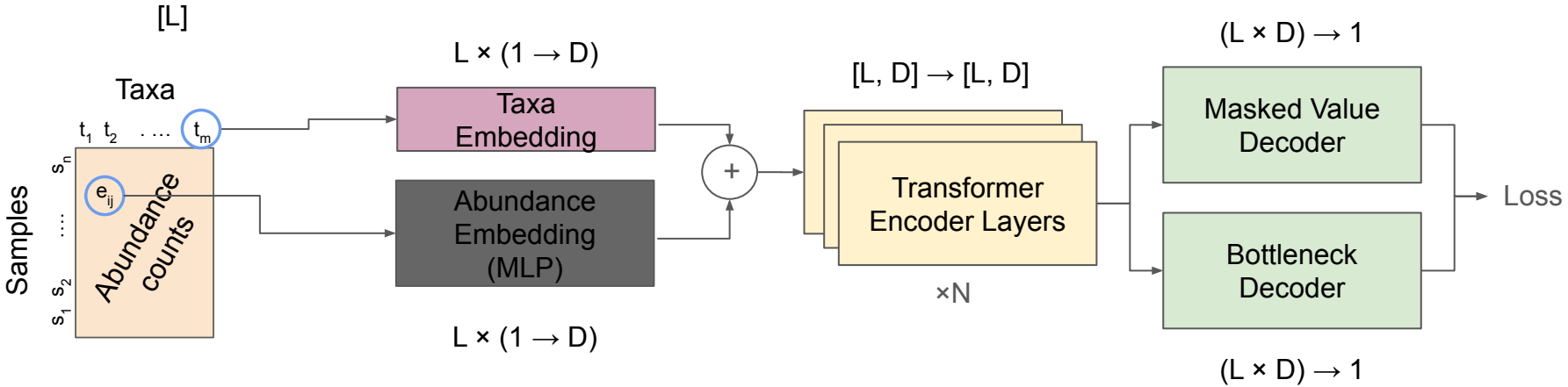
Self-Supervised Pretraining Loss Functions



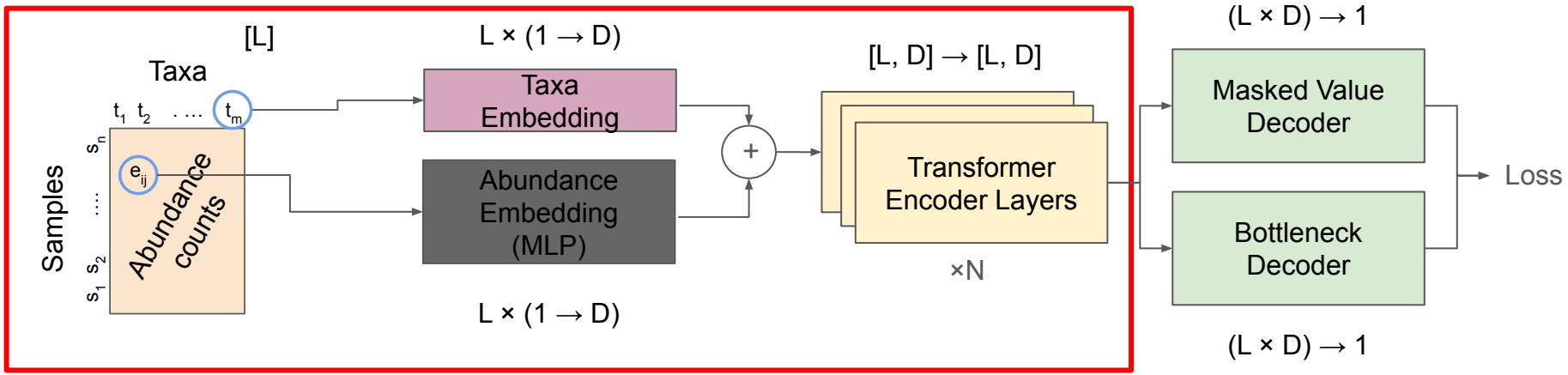
Masking with bottleneck representation: Masked Value Decoder with bottleneck at <cls> token

Masked Value Decoder: Some taxa abundance values are masked, and reconstruct missing abundance values.

Architecture in Detail

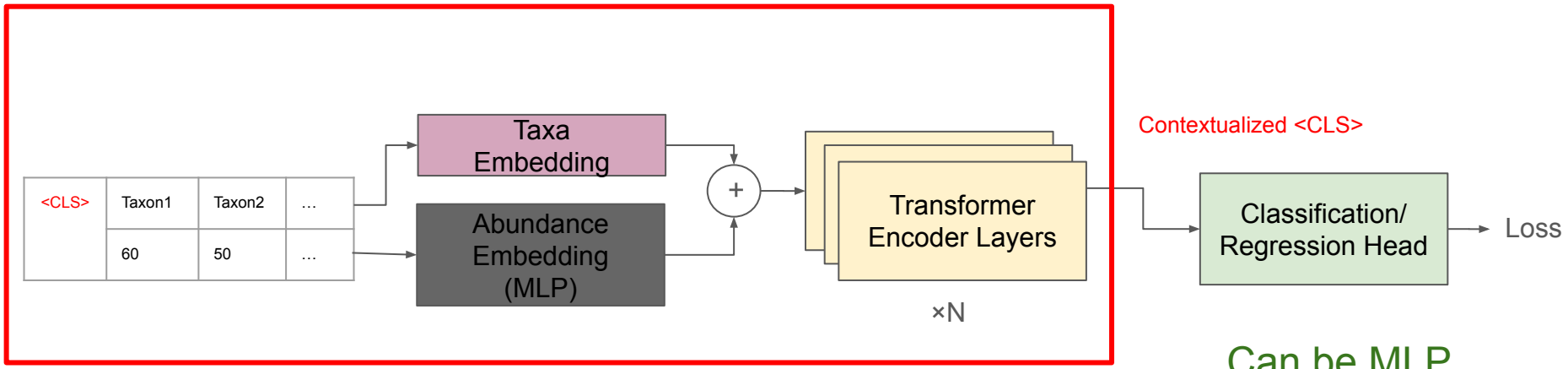


Architecture in Detail



Kept after pretraining finishes

Finetuning Procedure



From Pretraining

Can be MLP, Random Forest, etc.

Evaluation Pipeline

6 Microbiome Tasks

Regression

Evaluation: R^2

- Age
- BMI

Binary Classification

Evaluation: AUROC

- Human Diet
- Sex

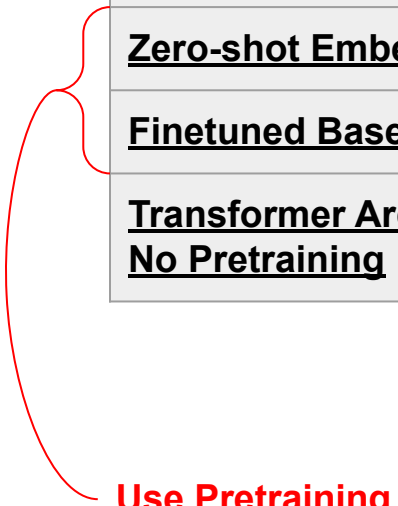
Multiclass Classification

Evaluation: Macro-F1

- Location
- Supplement Taken

7 Evaluation Strategies

	Random Forest	MLP
Using Raw Data	1	2
Zero-shot Embeddings	3	4
Finetuned Base Model	5	6
Transformer Architecture, No Pretraining		7

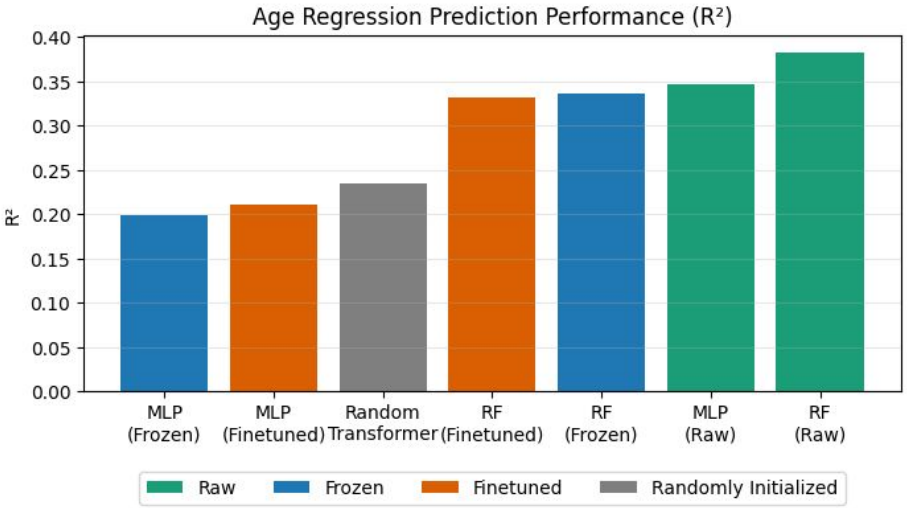


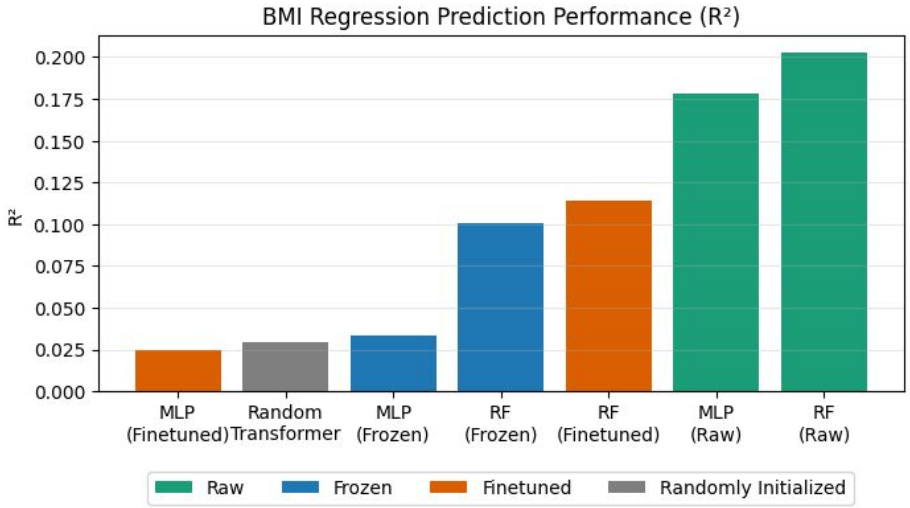
Rationale

1. Compare to classical baselines
2. Determine best way to use model
3. Determine benefit of pretraining

Experiment 1: Results

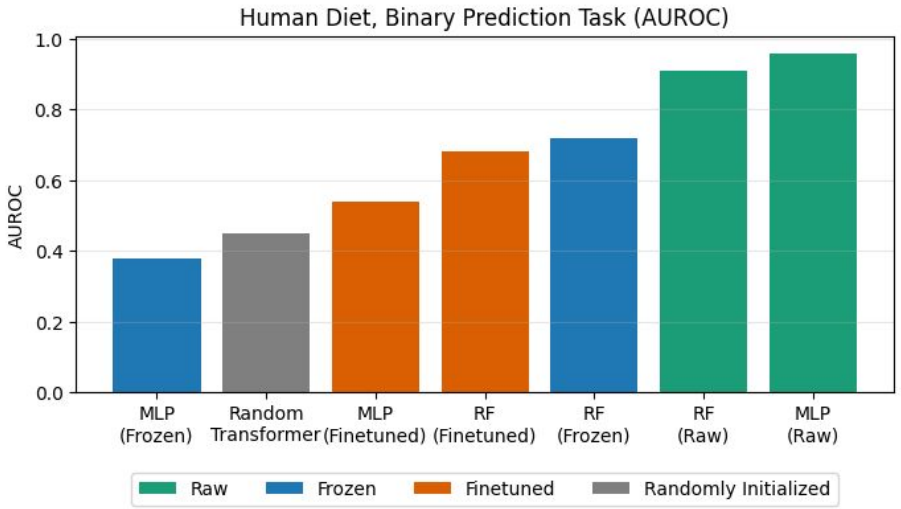
Regression Results

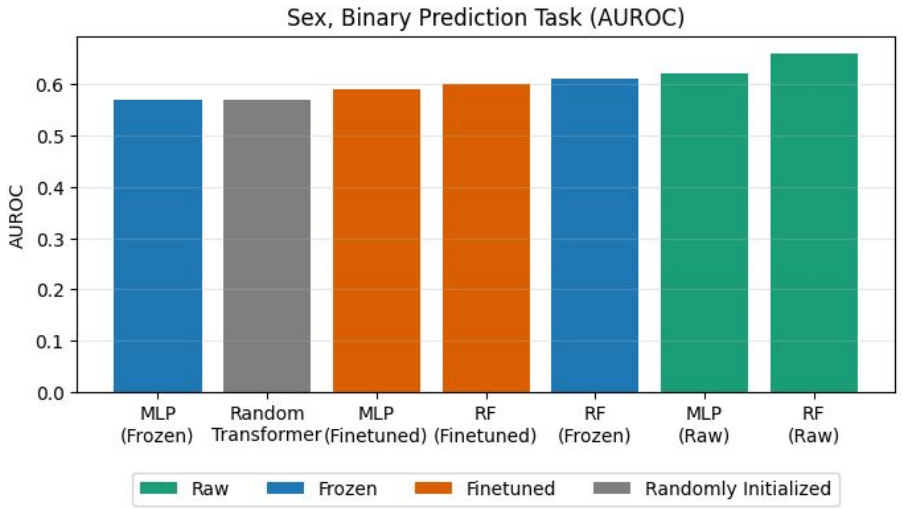




Performance with **Raw Data Dominates**
Random Forest outperforms MLP

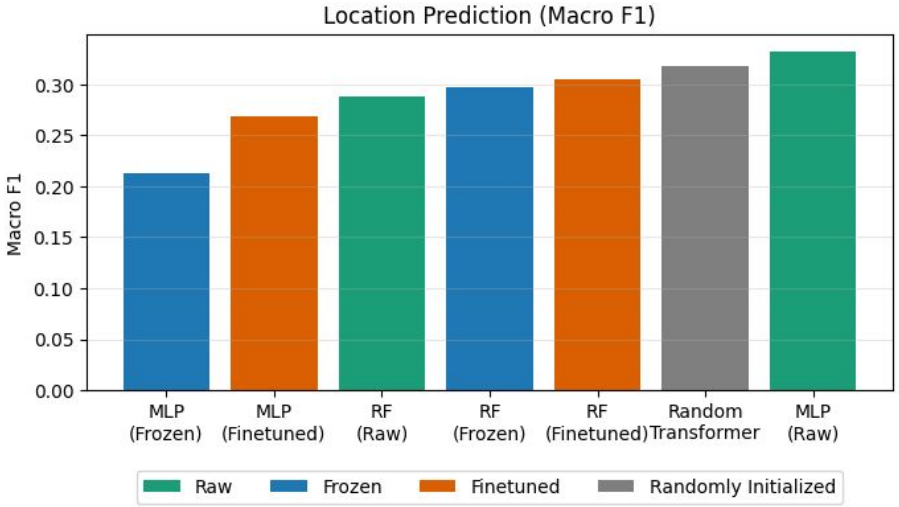
Binary Classification Results

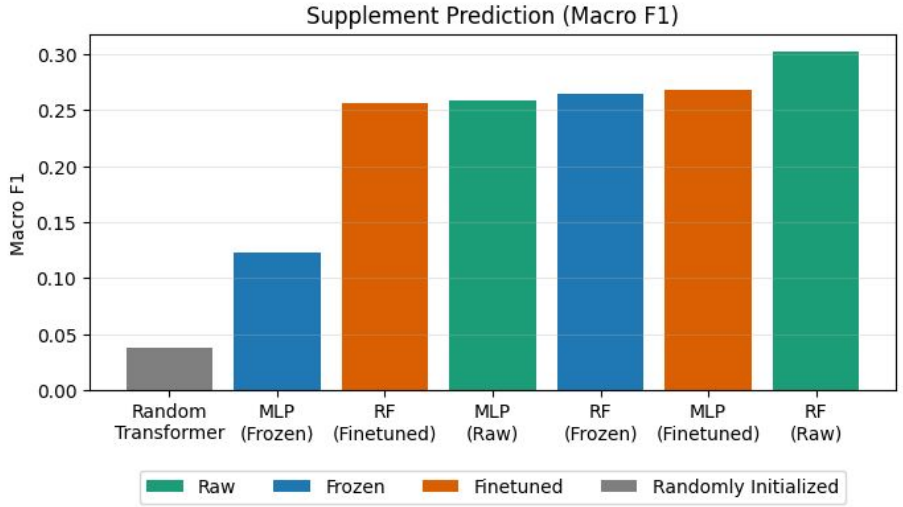




Performance with **Raw Data Dominates**
Random Forest outperforms MLP

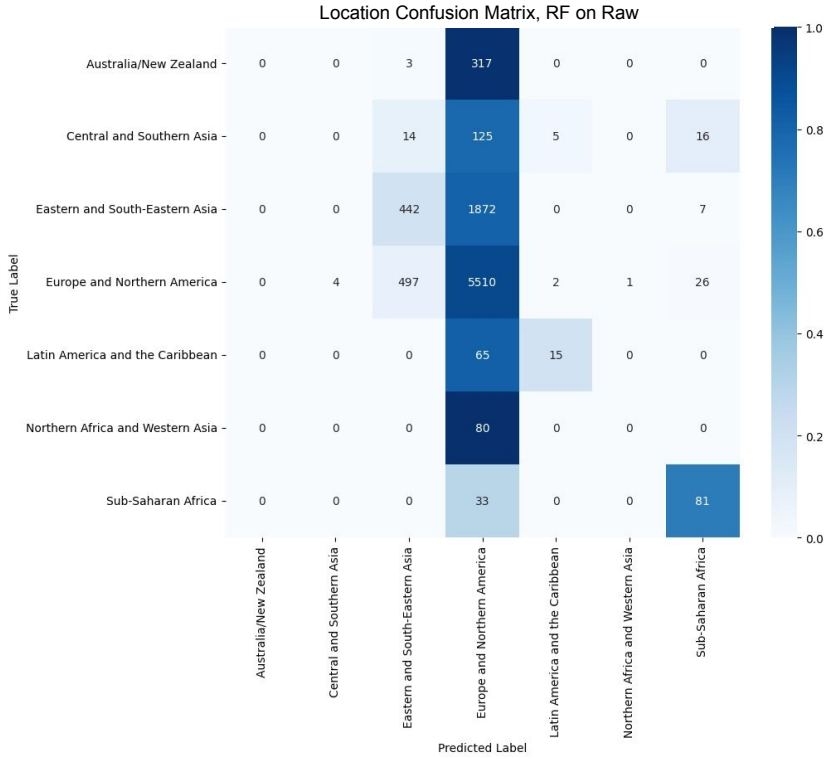
Multiclass Classification Results

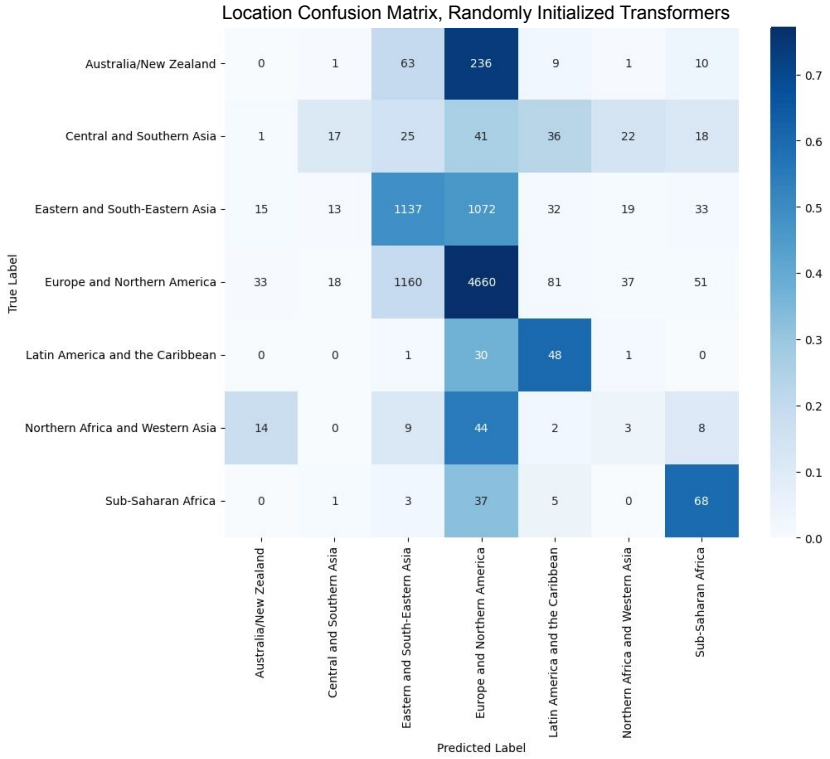




Mixed Results

Performance on Location Task Driven by Low Prediction Diversity





Suggests model overfitting to majority class instead of true prediction

Why does Random Forest work so well for Microbiome Data?

- Limited Sample Size
- Sparsity
- Biological Signal Problem
  - due to measurement error, uncertainty in classification
  - The **Bayes Error is too high!**

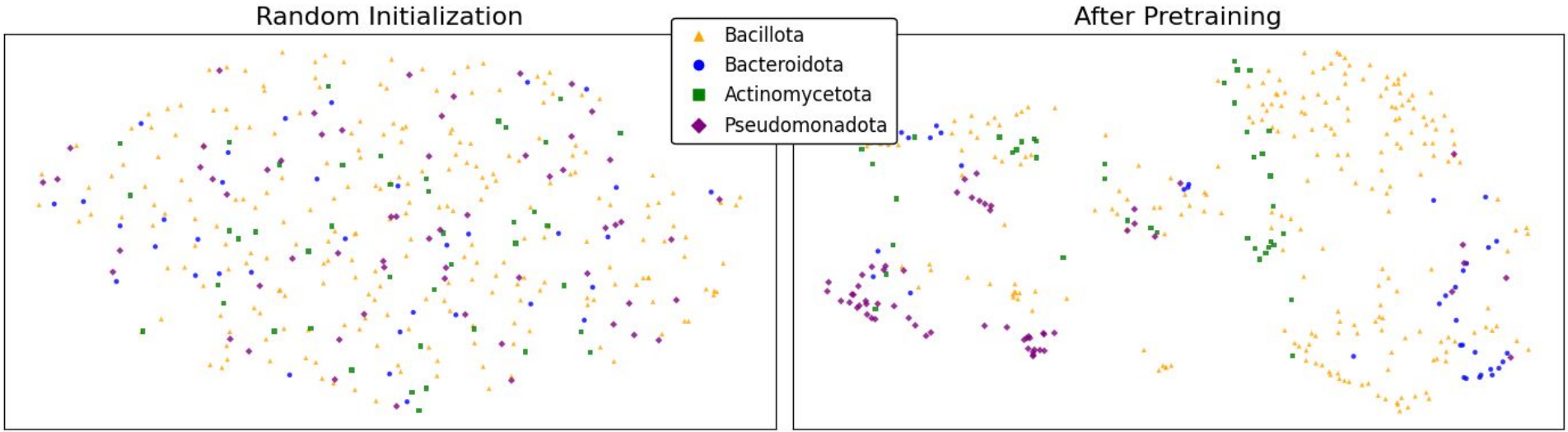
https://blekhman.substack.com/p/ai-keeps-failing-at-microbiome-prediction

Is there anything we can do?

Observation: **Clustering** of Taxa Embeddings

Taxa Embedding

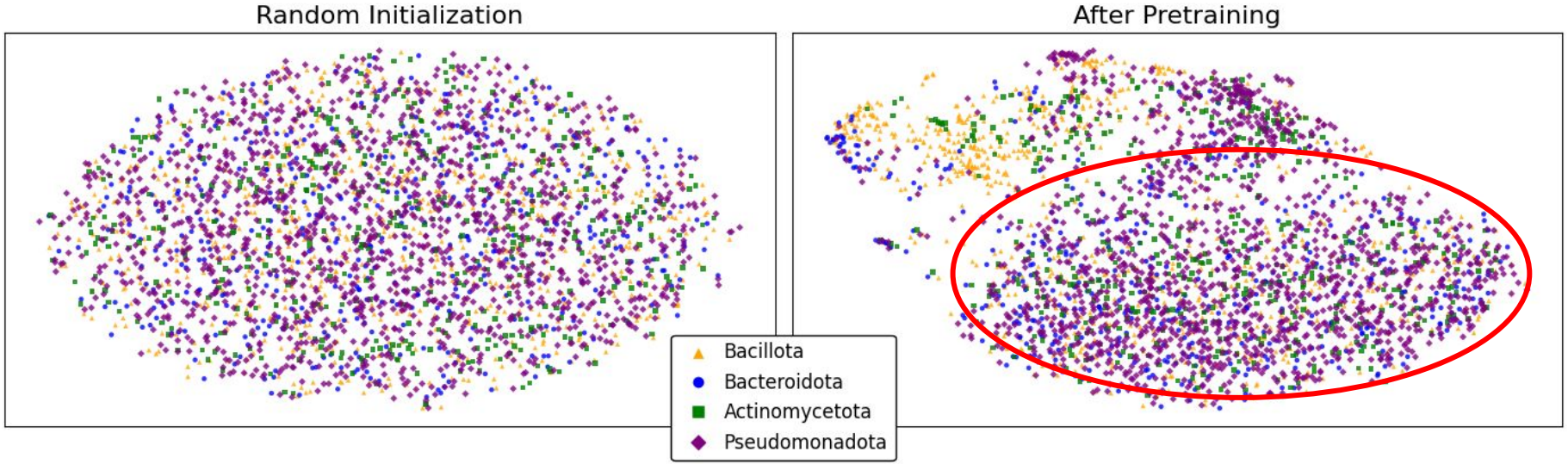
Pre-transformer embeddings store independent



Taking the **weights** of the **taxa embedding layer** for UMAP shows **related taxa clustering** despite it not being an explicit objective

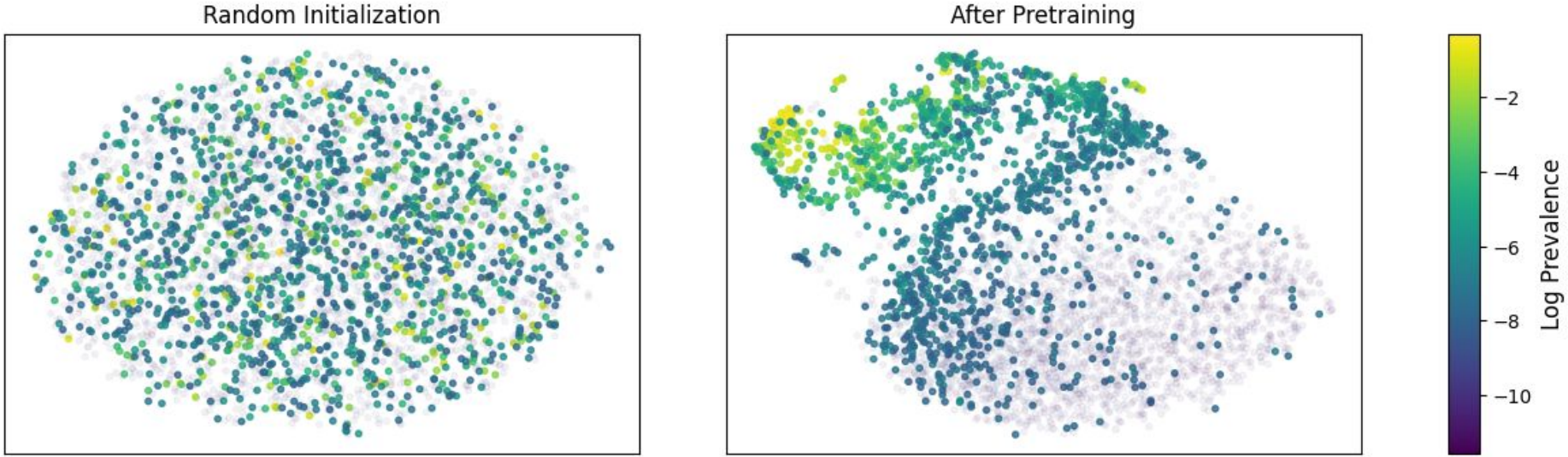
Taxa Embedding

Considering All Taxa…



Why?

Only Prevalent Taxa are updated



Hypothesis:

1. Rare Taxa do not appear in training enough to update weights.

Part 1: Conclusion

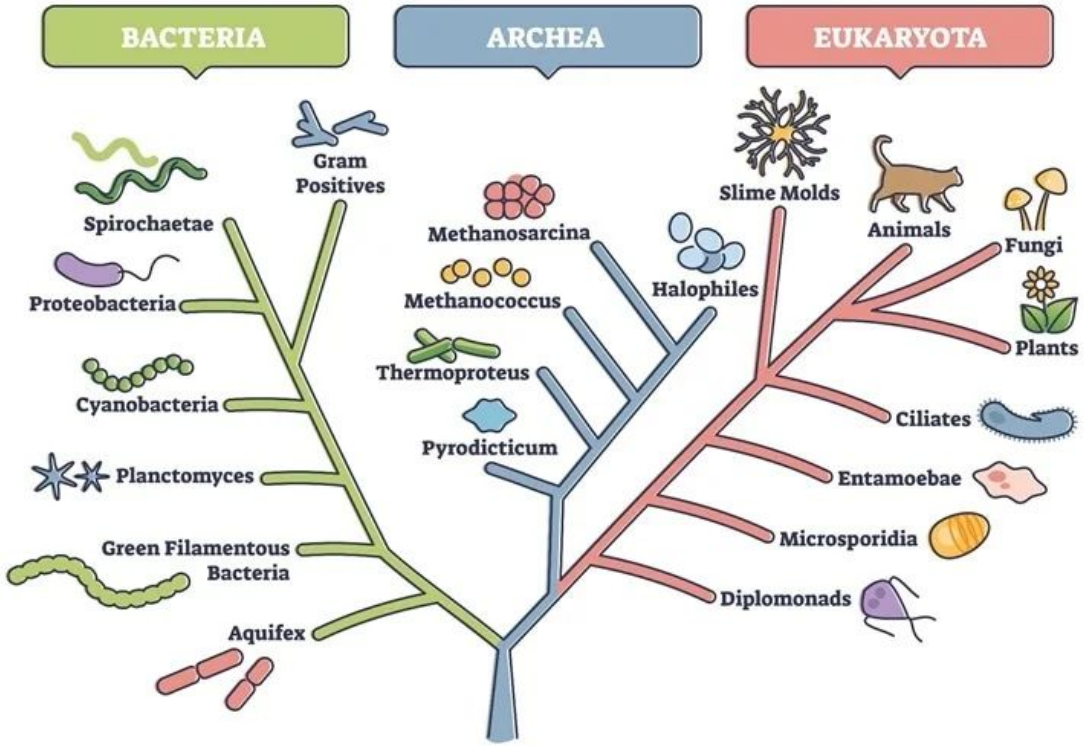
1. Foundation Models show some **predictive power,** but struggle to outperform **classical baselines.**

2. Data Signal is Inherently Limited.

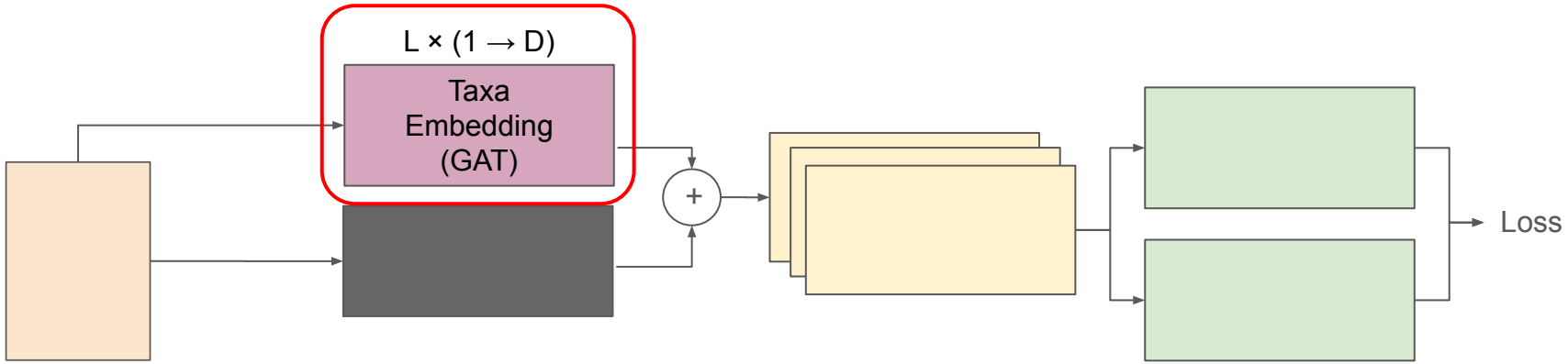
3. Data Sparsity causes many **weights to be not updated** during pretraining.

Experiment 2: Graph Neural Network for Taxa Embeddings

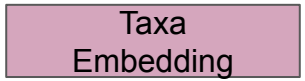
PHYLOGENETIC TREE

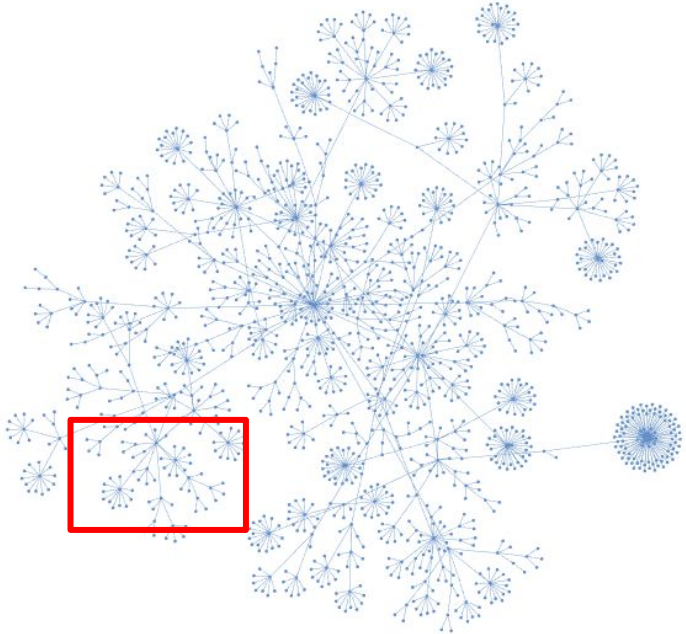


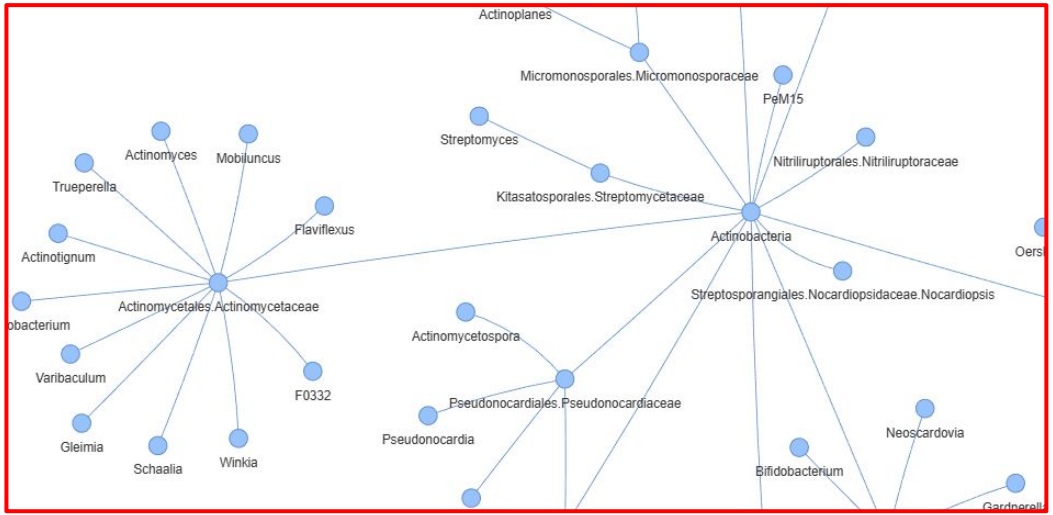
https://www.azolifesciences.com/article/What-is-Molecular-Phylogenetics.aspx


L × (1 → D)
Taxa
Embedding
(GAT)
+
Loss

Architecture in Detail

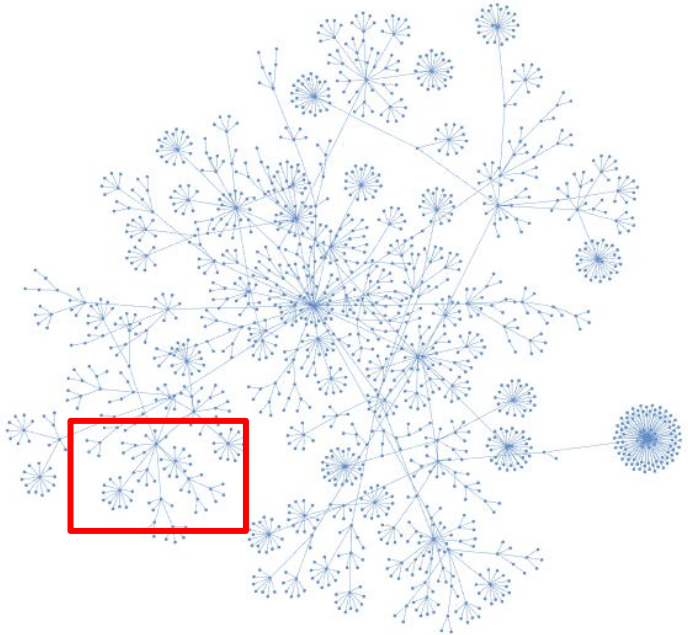




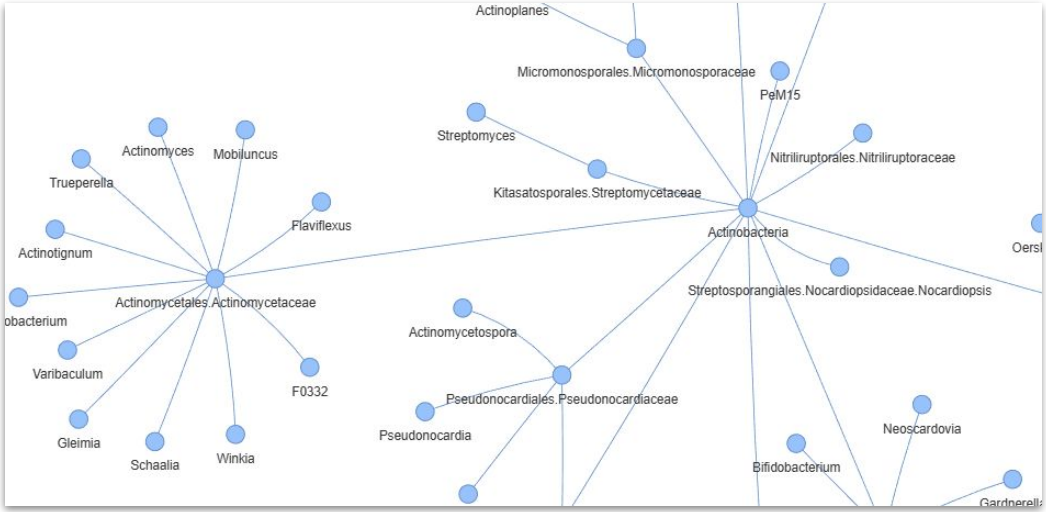

Actinoplanes
Micromonosporales.Micromonosporaceae
PeM15
Streptomyces
Nitriliruptorales.Nitriliruptoraceae
Actinomyces
Mobiluncus
Trueperella
Kitasatosporales.Streptomycetaceae
Flaviflexus
Actinobacteria
Actinotignum
Streptosporangiales.Nocardiopsidaceae.Nocardiopsis
Actinomycetales.Actinomycetaceae
Actinomycetospora
Varibaculum
F0332
Pseudonocardiales.Pseudonocardiaceae
Pseudonocardia
Neoscardovia
Gleimia
Schaalia
Winkia
Bifidobacterium

Architecture in Detail

Taxa Embedding



Graph Attention Network (GAT) for embeddings

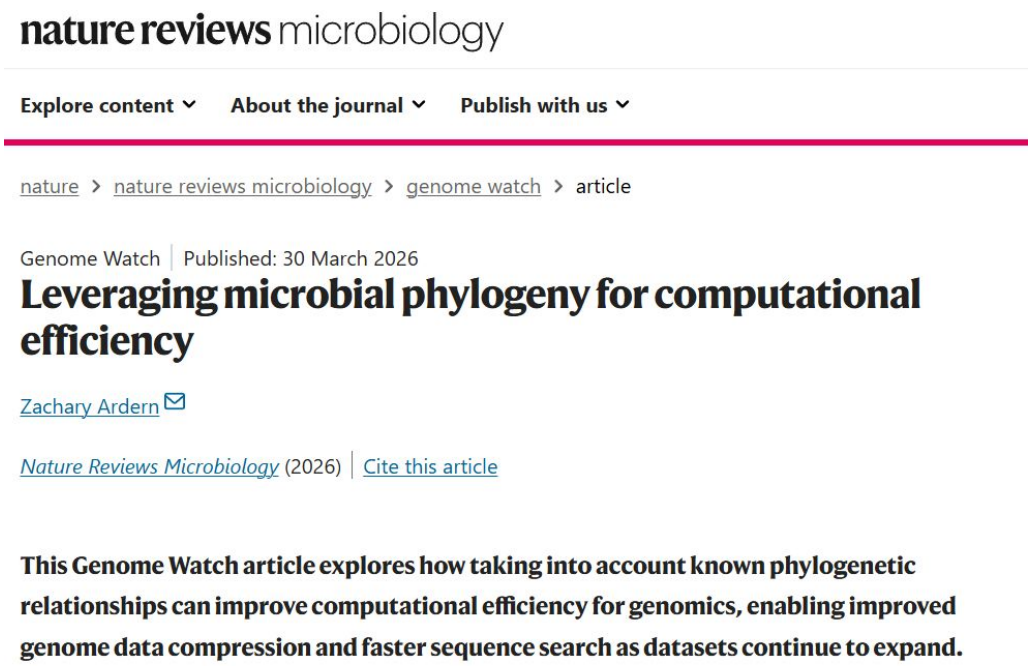


Why?

Desired Effect: Inductive Bias that **phylogenetically close** species should be **represented similarly**

Justification: Rare taxa initialized with embedding of **more common neighbour**

Evidence of this strategy working for Similar Tasks



nature reviews microbiology

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › nature reviews microbiology › genome watch › article

Genome Watch | Published: 30 March 2026

Leveraging microbial phylogeny for computational efficiency

Zachary Ardern

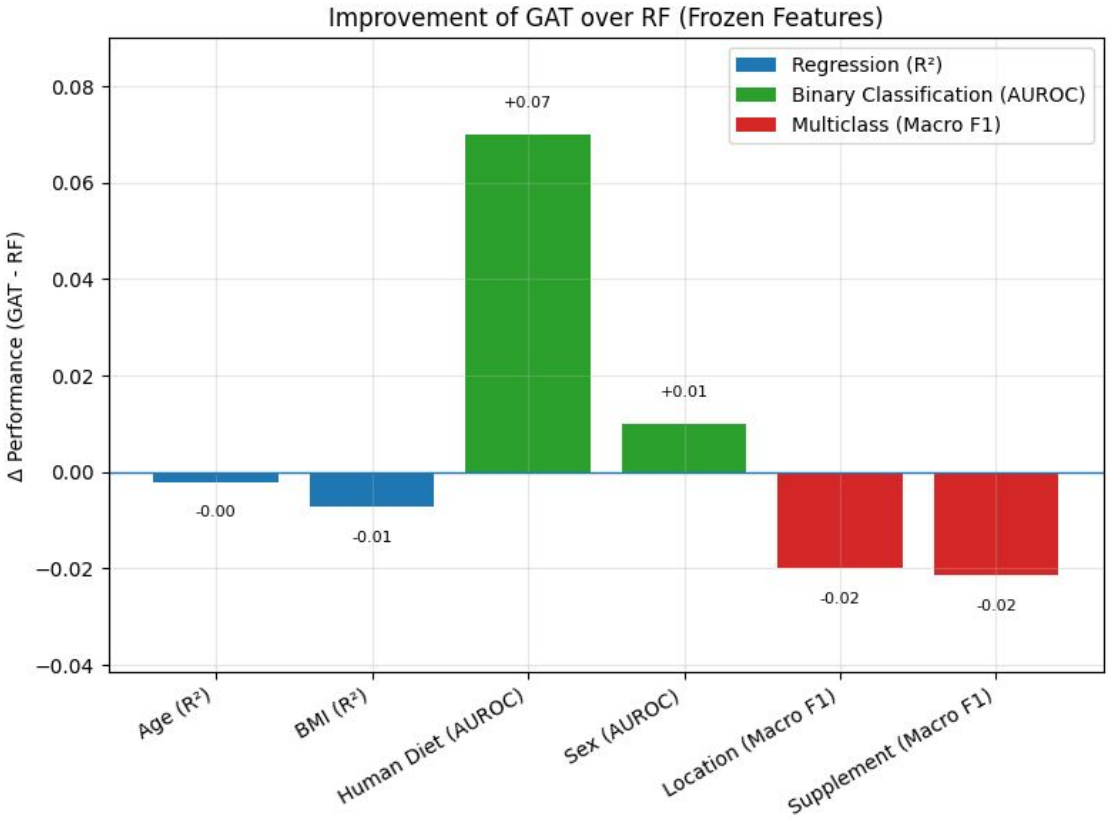
Nature Reviews Microbiology (2026) | Cite this article

This Genome Watch article explores how taking into account known phylogenetic relationships can improve computational efficiency for genomics, enabling improved genome data compression and faster sequence search as datasets continue to expand.

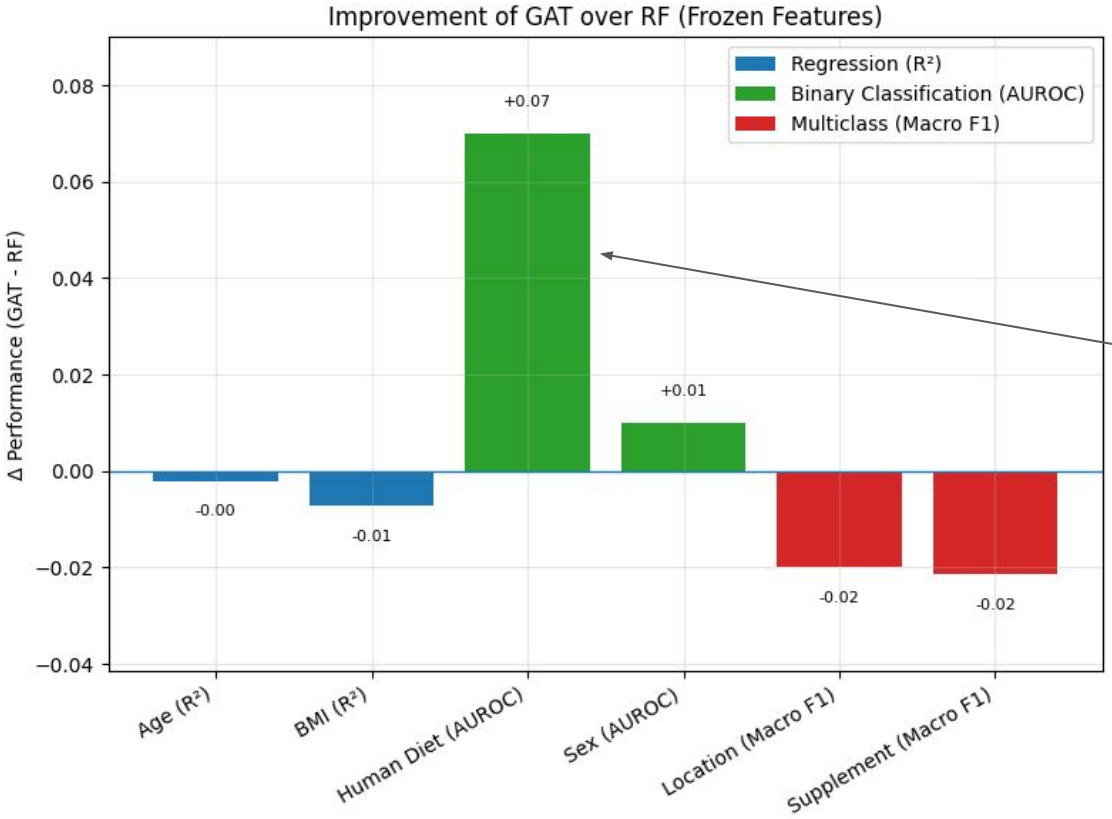
https://www.nature.com/articles/s41579-026-01303-0

Experiment 2: Results

Mixed Performance gains



Mixed Performance gains



Anomaly due to majority class prediction rather than true generalization

Part 2: Conclusion

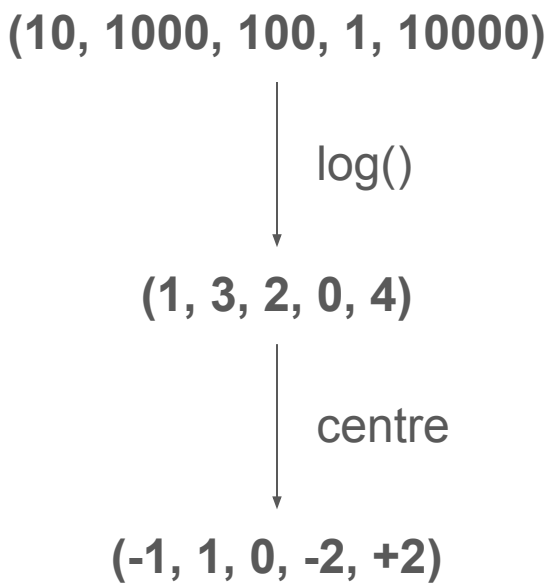
1. GNN offers **Limited Benefits**

   <u>Potential explanations:</u>
   - Taxa should not necessarily cluster in embedding space
   - Taxa clustering was driven by confounding factors, e.g. primer choice

2. Alternatively, incorporating phylogeny may require an **alternate strategy** e.g. Evo2

Experiment 3: Centre Log Ratio

What is the Centre-Log-Ratio Transform?



What is the Centre-Log-Ratio Transform?

$$\mathrm{clr}(\mathbf{x}) = \left(\log \frac{x_1}{g(\mathbf{x})}, \log \frac{x_2}{g(\mathbf{x})}, \ldots, \log \frac{x_D}{g(\mathbf{x})} \right)$$

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

Why?

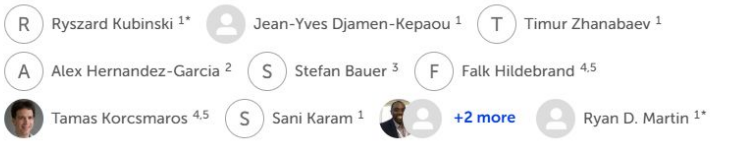
Desired Effect: Standard Preprocessing step for **compositional data** (know ratios only, not absolute amounts)

Justification: Sometimes found to increase model performance

Why?

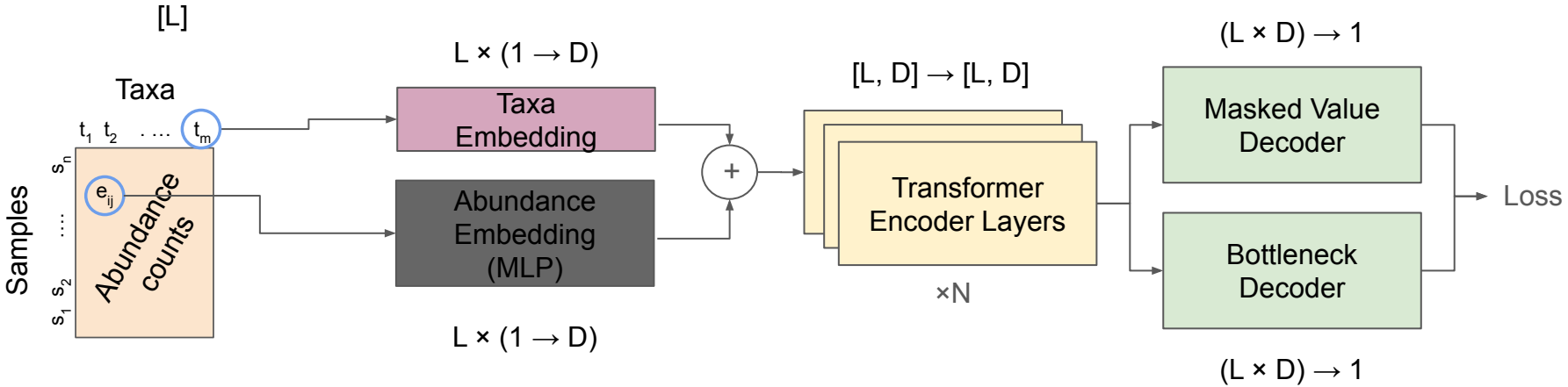
Justification:
Sometimes found to increase model performance

Benchmark of Data Processing Methods and Machine Learning Models for Gut Microbiome-Based Diagnosis of Inflammatory Bowel Disease

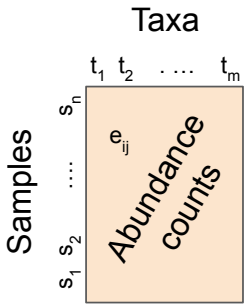


The compositional transformation methods were the most generalizable across non-linear models (**Figure 4A**), with median ILR F1 score of 74.3 (71.4–76.9) and MCC of 58.7 (53.6–63.1) and median CLR F1 score 74.2 (71.5–76.9) and MCC 58.5 (53.8–63.1). The compositional transformations were followed by the variance/distribution modifiers ARS (F1 score of 72.5 (69.8–75.8) and MCC of 56.0 (50.8–61.8)) and VST (F1 score 72.0 (65.9–75.0) and MCC 54.8 (44.1–59.9)). Lastly, TSS [F1 score 69.8 (63.8–73.9) and MCC 51.9 (40.5–58.7)] and LOG[F1 score 68.9 (64.1–73.5) and MCC 51.3 (40.4–58.6)] were consistently the lowest performing normalization. Furthermore, the compositional methods led to significantly better F1 score and MCC than the other normalization type (**Figure 4B**), whereas the variance/distribution modifiers and scaling method were significantly better than no normalization.

This Modification is on **Data**, not **Model**

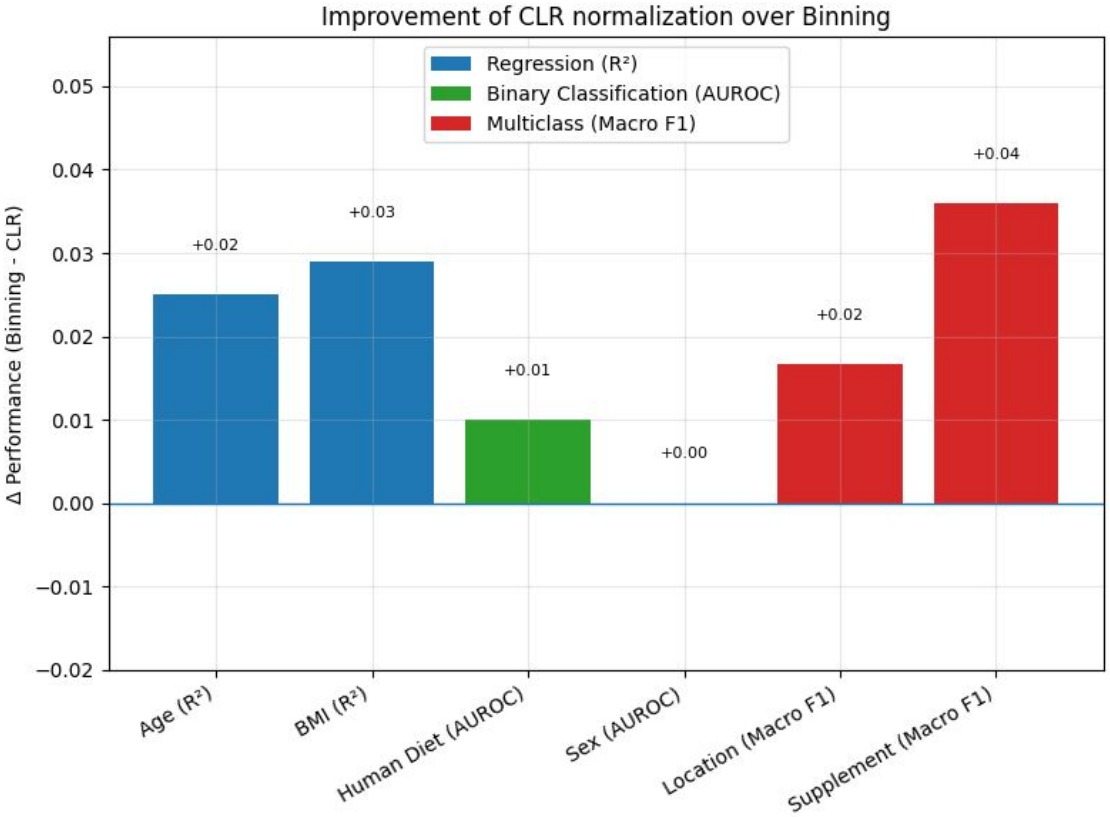


This Modification is on **Data**, not **Model**



Experiment 3: Results

Consistent Performance Gains



Part 3: Conclusion

1. Using a CLR ratio transform, instead of Binning as is done in scGPT, **improves performance**

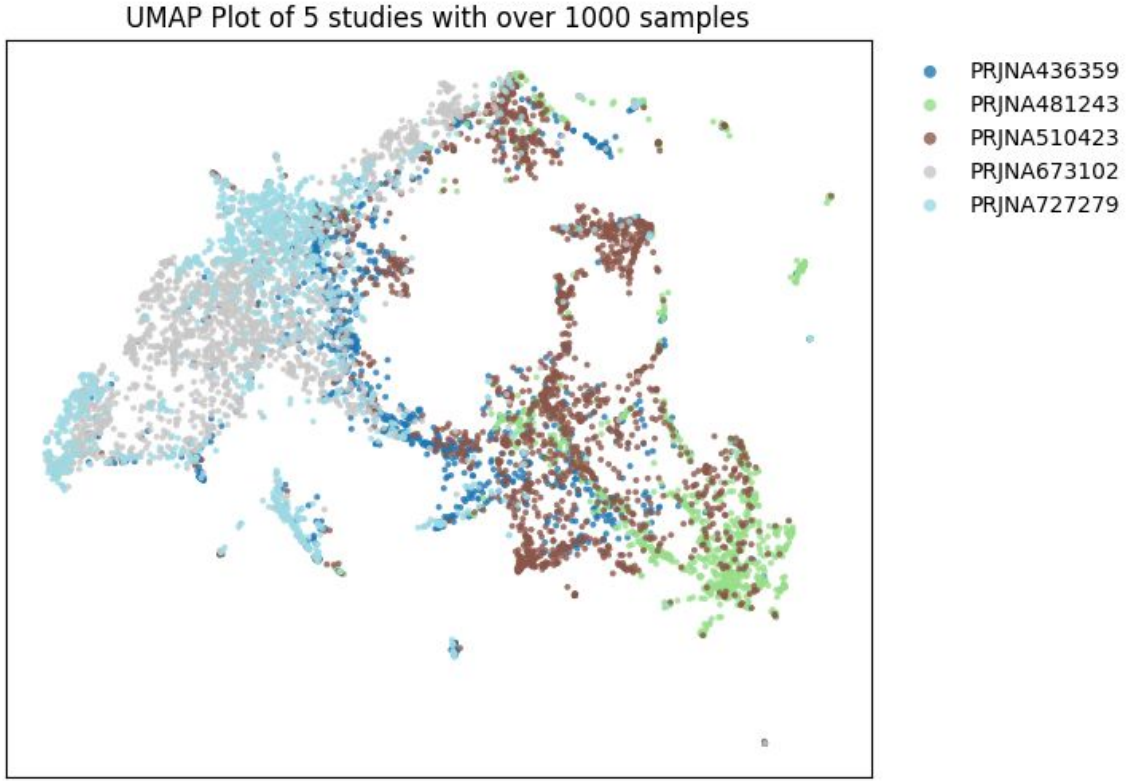
2. Performance is **similar to classical baselines**

3. This suggests a bottleneck with the **representation**, not the **model capacity**.

4. These results are **consistent with literature**

Challenges and Limitations

Heterogeneity Between Studies



Sources of Heterogeneity Between Studies

Biological Variation

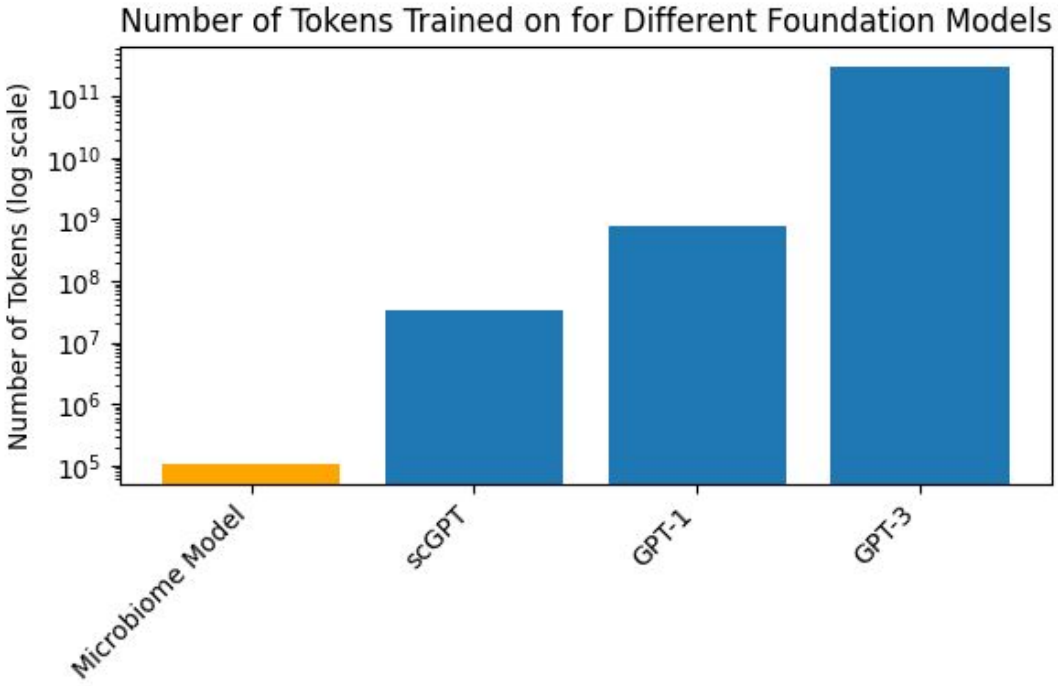
Primer Choice

DNA Extraction Method

It is difficult to **disentangle** these variables.

It is an **active area** of research.

Scaling

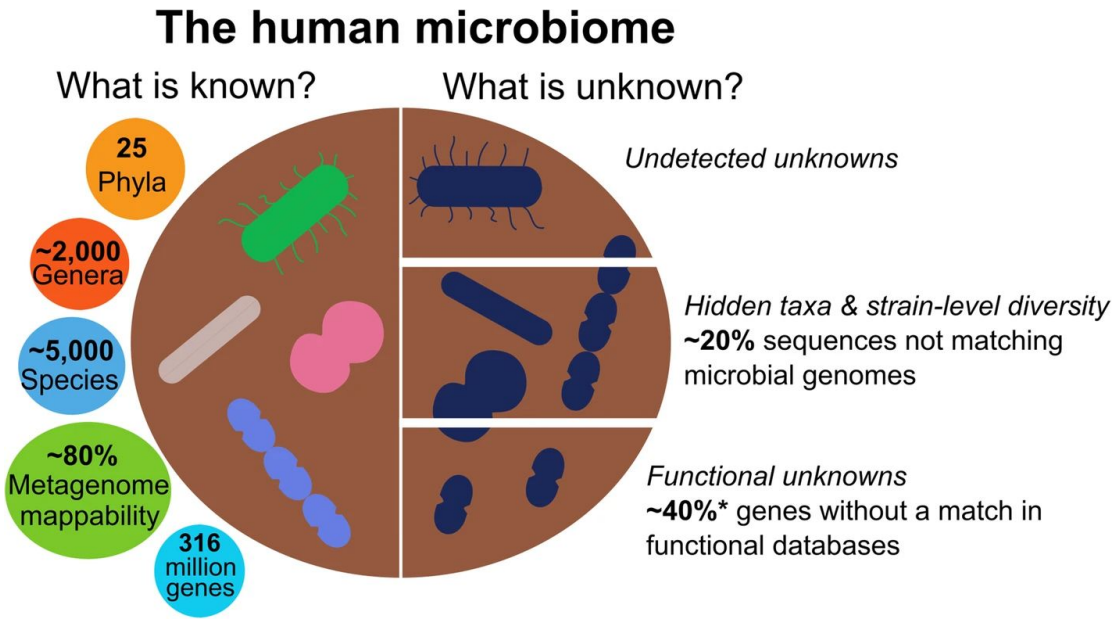


Infeasible to generate **scale** of data required

Scaling is limited by **wet-lab** work

Success will depend on **Multimodal Integration** to **expand dataset size**

“Unknowns”



https://link.springer.com/article/10.1186/s12915-019-0667-z/figures/1

More **accurate** microbiome profiling would **improve computational analysis**

Performance Saturation Suggests Data-Limited Regime

- Multiple model classes converge to **similar performance** (RF, MLP, Transformer)
- Increasing **model capacity** / **finetuning** does not yield consistent gains
- **Strong baselines** (tree-based) match or outperform deep models

Findings / Takeaway

We are likely near the **Bayes error of this dataset**

Next Steps:

Add signal: better **measurements** or **multimodality**