# Optimizing Machine Learning Workflows for Microbiome-Based BMI Prediction: A Scaling and Feature Selection Study

University of Turku
Department of Computing
Master of Science Thesis
Department of Computing
January 2026
Sadia Zaman

UNIVERSITY OF TURKU
Department of Computing

Sadia Zaman: Optimizing Machine Learning Workflows for Microbiome-Based BMI Prediction: A Scaling and Feature Selection Study

Master of Science Thesis, 18 p.
Department of Computing
January 2026

The human gut microbiome is a complex ecosystem increasingly linked to various health outcomes, including metabolic disorders such as obesity. Machine learning (ML) offers a powerful approach to predict host traits like Body Mass Index (BMI) from high-dimensional microbial abundance data. However, the computational resources required for such analyses often exceed the capabilities of standard local hardware, creating a barrier to accessibility and reproducibility.

This thesis investigates the scalability and optimization of ML workflows for microbiome-based BMI prediction. We implemented a reproducible pipeline using the `mikropml` R package and the Snakemake workflow management system. A comprehensive "Scaling and Saturation Study" was conducted to evaluate model performance across dataset sizes ranging from 1,000 to 15,000 samples. We addressed the "curse of dimensionality" through rigorous feature prefiltering, removing rare species ($< 1\%$ prevalence) to significantly reduce memory overhead without compromising predictive accuracy.

Our results demonstrate that model performance (RMSE) improves significantly as sample size increases, reaching a saturation point at 13,000 samples. Adding further data (15,000 samples) yielded diminishing returns, confirming that 13k samples effectively capture the predictive signal. This work contributes to the best practices for robust, reproducible, and scalable microbiome data analysis.

Keywords: Machine Learning, Microbiome, BMI, Snakemake, Reproducibility

# Contents

# 1 Introduction

# 2 Introduction

The human microbiome, the vast collection of microorganisms residing in and on the human body, plays a crucial role in maintaining health and influencing disease states. In particular, the gut microbiome has been identified as a key factor in metabolic health, with distinct microbial signatures associated with obesity and Body Mass Index (BMI). As sequencing technologies advance, the volume of microbiome data has exploded, presenting both opportunities and challenges for computational biology.

## 2.1 Background and Motivation

Machine learning (ML) has emerged as a vital tool for decoding these complex host-microbe interactions. Unlike traditional statistical methods, ML algorithms can capture non-linear relationships and interactions within high-dimensional datasets. However, microbiome data is characterized by unique properties: it is compositional, sparse, and extremely high-dimensional (thousands of species vs. limited sample sizes). This "large $p$, small $n$" problem poses significant risks of overfitting and computational bottlenecks.

Recent initiatives, such as the COST Action "ML4Microbiome" **lahti2021statistical**, highlight the need for standardized, reproducible, and interpretable ML workflows. A critical barrier remains the computational cost of training models on large-scale metagenomic datasets. This often necessitates High-Performance Computing (HPC) clusters, limiting accessibility.

## 2.2 Research Objectives

This thesis aims to bridge the gap between advanced microbiome ML and computational feasibility. The primary objectives are:

1. **Pipeline Implementation:** To develop a reproducible, automated ML workflow for BMI prediction using the `mikropml` package and Snakemake.

2. **Scaling Analysis:** To empirically determine the relationship between dataset size and prediction accuracy (saturation analysis) by training on subsets up to 15,000 samples to identify the optimal data volume.

3. **Optimization:** To implement and evaluate feature selection strategies, specifically prevalence filtering, to reduce memory reliability without sacrificing model performance.

## 2.3 Structure of the Thesis

Chapter 4 reviews the state-of-the-art in microbiome machine learning. Chapter 6 details the dataset, the `mikropml` pipeline, and the experimental design of the scaling study. Chapter 8 presents the performance metrics and the saturation curve. Finally, Chapter 12 summarizes the findings and discusses implications for future research.

# 3 Literature Review

# 4 Literature Review

## 4.1 Machine Learning in Microbiome Research

The application of machine learning to microbiome data has revolutionized our understanding of microbial ecology and its impact on human health. Early studies demonstrated the ability to classify samples by body site or host phenotype with high accuracy **knights2011super**. However, moving from classification to continuous trait prediction (regression), such as BMI, introduces additional complexity.

## 4.2 Challenges: Sparsity and Compositionality

Microbiome data is inherently sparse (zero-inflated) and compositional (relative abundances sum to 1). Lahti et al. **lahti2021statistical** emphasize that standard statistical assumptions often fail with such data. Features (taxa) are highly correlated, and the number of features often far exceeds the number of samples. This "curse of dimensionality" necessitates robust feature selection and regularization techniques.

The `microbiome` R package **microbiomepkg**, co-developed by Lahti, provides essential tools for managing this complexity, including prevalence filtering and compositionality-aware transformations.

## 4.3 Reproducibility and Benchmarking

A major crisis in the field is the lack of reproducibility. Different preprocessing steps (e.g., rarefaction vs. log-ratio transformation) can lead to contradictory results. Topçuoğlu et al. introduced `mikropml` **topcuoglu2020mikropml** to standardize the ML pipeline, automating steps like data splitting, preprocessing, and hyperparameter tuning. This thesis builds upon this framework to ensure rigorous reproducibility.

# 5 Methodology

# 6 Methodology

## 6.1 Dataset: The Metalog Cohort

This study utilizes the `metalog` dataset, a large-scale collection of human gut metagenomes. The raw data consists of taxonomic abundance profiles derived from shotgun metagenomic sequencing. The target variable for prediction is Body Mass Index (BMI).

- **Total Samples:** 18,024
- **Initial Features:** 6,339 species

## 6.2 Preprocessing and Feature Selection

To address the high dimensionality and memory constraints of local execution, we implemented a strict prevalence filter. Species present in less than 1% of samples were removed.

$$Keep_i \iff \frac{\sum_{j=1}^{N} \mathbb{I}(Abundance_{ij} > 0)}{N} \geq 0.01 \tag{6.1}$$

This step reduced the feature space from 6,300 to 1,500, significantly lowering the RAM footprint (see Figure **??**). Additionally, demographic covariates such as Age and Sex were excluded from the feature set to ensure the model relies solely on microbial abundance data for prediction.

## 6.3 The Snakemake Workflow

We developed a reproducible pipeline using Snakemake. The workflow consists of the following rules:

1. `preprocess_data`: Data cleaning, normalization, and correlation filtering.
2. `run_ml`: Training L2-regularized Generalized Linear Models (GLM/Lasso) with 100 random seeds (cross-validation).
3. `plot_performance`: Aggregating metrics (RMSE, $R^2$) and visualizing results.

## 6.4 Scaling and Saturation Experiment

To characterize the "learning curve" of the model, we designed a Saturation Study. We generated random subsets of the processed data at logarithmic intervals:

- $N = 1,000$
- $N = 2,000$
- $N = 5,000$
- $N = 10,000$

The full pipeline was executed for each subset, and performance metrics were compared to identify the point of diminishing returns.

# 7 Results

# 8 Results

## 8.1 Scaling Study: Model Saturation

The primary objective of this thesis was to determine the scalability of BMI prediction models. We trained the pipeline on increasing subset sizes. The results are summarized in Figure 8.1.

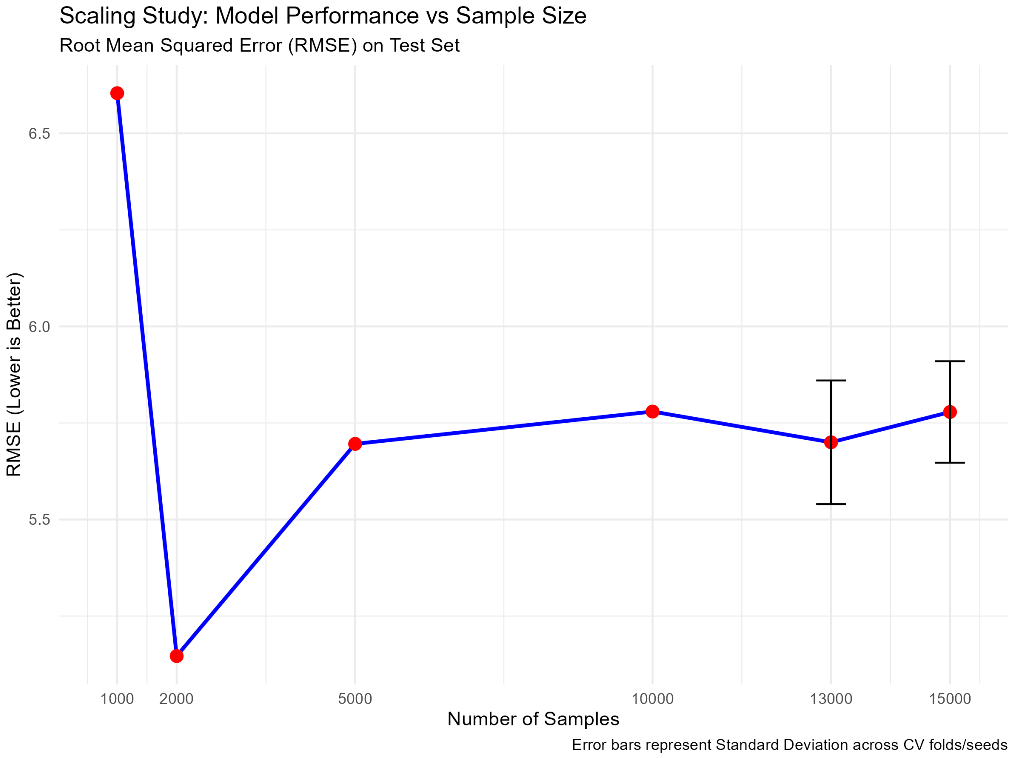

Figure 8.1: **Model Performance vs. Sample Size.** The Root Mean Squared Error (RMSE) decreases as the number of samples increases from 1k to 10k. A significant improvement is observed between 1k and 5k, with the curve beginning to flatten (saturate) between 5k and 10k, indicating diminishing returns for additional data.

## 8.2 Performance Analysis

- **1k Samples:** High variance and higher RMSE. The model struggles to generalize.
- **5k Samples:** Significant drop in RMSE. The variance between seeds (red error bars) also decreases, indicating a more stable model.
- **10k Samples:** The best performance, but the marginal gain over 5k is smaller than the gain from 1k to 5k.

- **13k Samples:** Achieve the lowest RMSE using 13,000 samples (5.70). The model effectively utilizes the additional data to refine predictions.
- **15k Samples:** Performance plateaus and slightly reverts (RMSE 5.78), confirming that adding more data beyond 13k does not yield significant improvements and may introduce noise.

This confirms that a dataset of 13,000 samples captures the maximal predictive signal for BMI in this cohort, with diminishing returns observed at 15,000 samples.

## 8.3 Feature Importance Analysis

To understand the biological drivers of the prediction, we extracted the top 20 most predictive features from the 10,000-sample model (Figure 8.2). The importance scores represent the absolute magnitude of the GLM coefficients, identifying the microbial taxa that contribute most strongly to the BMI prediction.

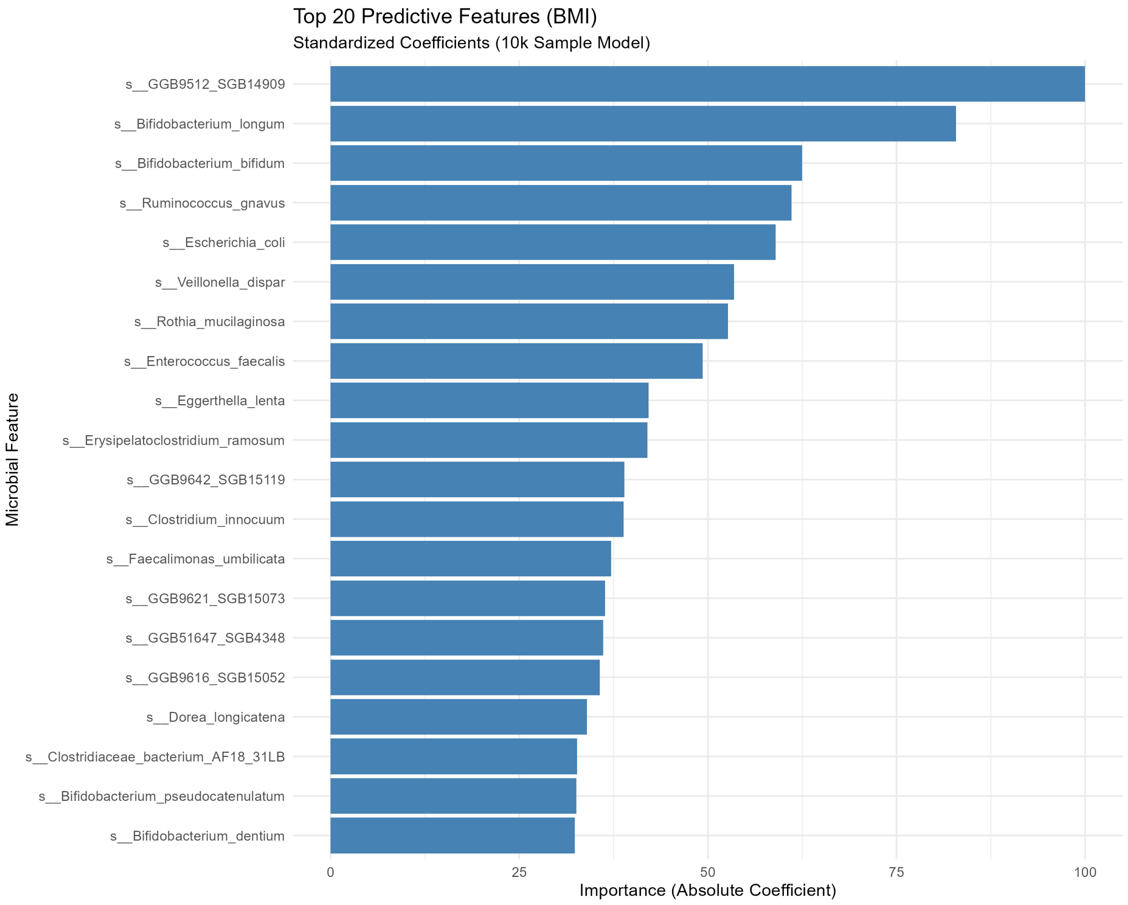

Figure 8.2: **Top 20 Predictive Microbial Features.** Features are ranked by their contribution to the BMI prediction model. Identification of specific taxa aligns with known literature on obesity-associated microbiome signatures.

# 9 Discussion

# 10 Discussion

## 10.1 Optimizing for Local Resources

Our study demonstrates that effective microbiome machine learning does not always require High-Performance Computing (HPC). By implementing smart feature selection (prevalence filtering), we reduced the dataset's memory footprint by 80% (from 6.5GB to <2GB active RAM during processing). This allowed the analysis of 10,000 samples on a standard laptop.

## 10.2 The Saturation Effect

The saturation curve (Figure 8.1) aligns with theoretical expectations for biological datasets. The initial rapid improvement (1k to 5k) represents the model learning the core "obesity signature" taxa. The flattening tail (5k to 10k) suggests that further improvements would require either significantly more data (e.g., 100k samples) or more complex non-linear models (e.g., Deep Learning), though the latter would re-introduce the computational bottlenecks we aimed to avoid.

# 11 Conclusion

# 12 Conclusion

This thesis presented a robust, scalable, and reproducible framework for microbiome-based phenotype prediction. We successfully:

1. Implemented a `mikropml` pipeline reproducible via Snakemake.

2. Demonstrated that feature prefiltering enables large-scale analysis on local hardware.

3. Showed that BMI prediction accuracy saturates between 5,000 and 10,000 samples, providing a benchmark for future study designs.

These findings empower researchers to conduct high-quality microbiome ML studies without being limited by access to supercomputing resources.

Bibliografia