Microbiome-Based Body Mass Index Prediction: A Scalable Machine Learning Benchmark for Large-Scale Metagenomic Data

Sadia Zaman

University of Turku
Faculty of Technology
Department of Computing
Master of Science Thesis
Master's Degree Programme in
Information and Communication Technology (ICT), Data Analytics
June 2026

Supervisors:

Leo Lahti
Geraldson Muluh

UNIVERSITY OF TURKU
Faculty of Technology
Department of Computing

SADIA ZAMAN: Microbiome-Based Body Mass Index Prediction: A Scalable Machine Learning Benchmark for Large-Scale Metagenomic Data

Master of Science Thesis
Master's Degree Programme in
Information and Communication Technology (ICT), Data Analytics
June 2026, 88 p.

The human gut microbiome is closely linked to host metabolism. Although many studies have reported associations between specific microbial taxa and obesity, the extent to which gut microbiome composition can predict body mass index (BMI) at population scale remains uncertain. This thesis investigates microbiome-based BMI prediction using 17,903 human gut metagenomes from the Metalog consortium and evaluates whether predictive performance improves through non-linear machine-learning models, large sample sizes, and substantial feature reduction.

The analysis was implemented as a reproducible Nextflow pipeline designed for portability and feasibility on consumer-grade hardware; an independent Snakemake implementation was additionally developed as a reproducibility check, although the two pipelines were not run with identical procedures and so are not compared one-to-one. Linear models (Elastic Net and linear SVM) were compared against non-linear tree-based methods. Random Forest consistently achieved the strongest performance, reaching an RMSE of 5.06 kg/m^2 ($R^2 = 0.387$), whereas the linear baselines remained substantially weaker. Saturation analyses indicated a region of diminishing returns between approximately 12,000 and 14,000 samples, with only a small further gain at the largest tested size (N=16,000). A descriptive analysis showed that the extreme BMI values reflect participant age and clinical cohort (infants at the low end; bariatric, elderly, and metabolic-disease cohorts at the high end) rather than data-entry error, and that top-ranked species show only weak, distributed associations with BMI.

To address the high dimensionality and sparsity of the data, a 1% prevalence filter and a variance-based top-500 filter were evaluated; these substantially lowered runtime and memory requirements while preserving most of the predictive signal. Feature-ranking analyses identified a mixture of named species and uncharacterised species-level genome bins among the highest-ranked predictors, including *Bifidobacterium longum*, *Veillonella* species, *Ruminococcus gnavus*, and several uncharacterised SGBs (e.g., GGB9512_SGB14909) — consistent with MetaPhlAn 4's improved profiling of previously uncharacterised taxa. Overall, the thesis establishes a scalable and reproducible baseline for microbiome-based BMI prediction while clarifying methodological trade-offs in evaluation design, preprocessing, and feature selection.

Keywords: Microbiome, Body Mass Index, Machine Learning,
Metagenomics, Nextflow, Random Forest

Acknowledgements

I would like to express my sincere gratitude to my supervisors, Leo Lahti and Geraldson Muluh, for their guidance, patience, and constructive feedback throughout the preparation of this thesis. Their expertise in microbial data science and machine learning provided an invaluable foundation for this work, and their detailed comments helped sharpen both the analysis and the writing at every stage.

I am grateful to the University of Turku and the Department of Computing for providing the academic environment and resources that made this research possible, and to the Metalog consortium for making the metagenomic data available for research purposes.

I would also like to thank my family and friends for their continued encouragement and support during my studies in Finland. Their patience and understanding made it possible to complete this work under demanding circumstances.

This thesis was written using open-source tools and publicly available workflow systems, and I gratefully acknowledge the developers of Nextflow, Snakemake, the mikropml R package, and the broader bioinformatics community whose software infrastructure made this computational research possible.

Turku, June 2026

Sadia Zaman

Artificial Intelligence Use Declaration

AI-based tools were used only for language support and limited programming assistance. All research questions, analytical decisions, code implementation, interpretation of results, and conclusions are the author's own. All cited literature was read and evaluated by the author.

Contents

List of Figures

List of Tables

List of acronyms

BMI Body Mass Index

CLR Centred Log-Ratio

CV Cross-Validation

DOI Digital Object Identifier

GGB Genus-Level Genome Bin

GLMNet Generalized Linear Model via Penalized Maximum Likelihood

HPC High-Performance Computing

MetaPhlAn Metagenomic Phylogenetic Analysis

ML Machine Learning

RAM Random Access Memory

RDS R Data Serialization

RF Random Forest

RMSE Root Mean Squared Error

SCFA Short-Chain Fatty Acid

SGB Species-Level Genome Bin

SVM Support Vector Machine

WHO World Health Organization

XGBoost Extreme Gradient Boosting

1 Introduction

1.1 The Global Metabolic Crisis and the Microbial Frontier

Obesity is one of the major public-health challenges of the twenty-first century. The World Health Organization (WHO) reports that, in 2022, 2.5 billion adults aged 18 years and older were overweight, including more than 890 million adults living with obesity [1]. Adult obesity has more than doubled since 1990, and obesity among children and adolescents has increased even more sharply over the same period [1]. These figures are significant because elevated BMI is associated with increased risk of type 2 diabetes, cardiovascular disease, several cancers, and other non-communicable diseases. The global burden of obesity is therefore not only a clinical issue, but also a social, economic, and computational research problem.

The classical explanation of weight gain is an energy-balance model: body mass increases when energy intake exceeds energy expenditure over time. This model is necessary, but it is not sufficient to explain the full biological variation observed among individuals. Individuals exposed to broadly similar dietary and lifestyle environments can exhibit different metabolic outcomes, and the same dietary intervention may produce different responses across individuals. Such variation has motivated growing interest in systems that mediate the relationship between diet, host physiology, inflammation, and energy harvest. One such system is the human gut microbiome.

The gut microbiome is a dense and dynamic community of bacteria, archaea, viruses, and microbial eukaryotes residing in the gastrointestinal tract. Microbial metabolism

contributes to the fermentation of dietary fibers, the production of short-chain fatty acids, the transformation of bile acids, vitamin synthesis, and the metabolism of dietary compounds that human enzymes cannot process directly [2]. The microbiome therefore functions as a metabolic interface between external exposures and host physiology. Experimental work has also shown that obesity-associated microbial communities can influence host energy harvest in model systems [3]. These findings do not imply that the microbiome alone determines body weight, but they do justify studying whether gut microbial composition contains measurable information about BMI.

Despite this biological plausibility, translating microbiome associations into robust prediction remains difficult. Many studies report statistically significant taxon-phenotype associations, but such associations often vary across cohorts, sequencing protocols, geographic regions, and preprocessing choices [4], [5]. This thesis is positioned within that gap. Rather than asking whether a single microbe is "the cause" of obesity, it asks a more computationally precise question: how well can BMI be predicted from species-level gut microbiome profiles alone, and what methodological choices make that prediction more or less reproducible?

This distinction between association, prediction, and explanation is important for the whole thesis. Association studies ask whether a taxon differs between groups or correlates with a phenotype. Prediction studies ask whether a model trained on one part of the data can estimate an outcome in held-out samples. Mechanistic studies ask whether changing a microbial exposure changes the host phenotype through a defined biological pathway. The present work is primarily a computational benchmark study — it evaluates prediction and workflow design rather than biological mechanisms. It uses biological literature to interpret the most important taxa, but it does not claim that the trained models demonstrate causality.

The research gap can therefore be stated directly. Previous studies have shown that gut microbiome composition is associated with BMI and related metabolic phenotypes, and several studies have applied machine-learning models to microbiome data [4], [5]. However, the specific combination examined in this thesis, namely a cohort of this scale together with an explicitly reproducible, workflow-managed design and a practical benchmark on consumer-grade hardware, appears to be less often documented. This thesis addresses that gap by treating predictive performance, workflow reproducibility, and local computational feasibility as part of the same methodological problem rather than as separate concerns.

This framing also clarifies what the thesis is not trying to do. The goal is not to claim that BMI can be reduced to microbial composition, nor to argue that species-level abundance profiles alone are sufficient for precision metabolic medicine. Rather, the goal is to establish a careful computational baseline. If a microbiome-only benchmark already captures measurable variation in BMI, that result is informative about the signal present in large taxonomic datasets. If the benchmark also reveals where performance plateaus and which workflow choices dominate runtime, then it becomes useful not only biologically but also methodologically. That combination of predictive benchmarking and reproducible computational design is what gives the thesis its contribution.

1.2 The Computational Bottleneck in Metagenomics

Shotgun metagenomic sequencing has made it possible to profile microbial communities at high taxonomic resolution. Tools such as MetaPhlAn 4 estimate species-level relative abundances by mapping reads to clade-specific marker genes, including markers derived from characterized and uncharacterized species-level genome bins [6]. This improves biological resolution, but it also produces large and sparse data

matrices. A cohort with tens of thousands of samples may contain thousands of species-level columns, many of which are absent in most individuals.

This scale creates two linked problems. The first is computational: memory usage, training time, and input/output overhead increase quickly as both sample size and feature count grow. A model that is straightforward to train on a few hundred samples can become difficult to execute on consumer-grade hardware when the matrix contains thousands of microbial features and more than 10,000 individuals. The second problem is statistical: high-dimensional, sparse, compositional data can encourage overfitting, unstable feature selection, and overly optimistic validation if preprocessing is not reported carefully [4].

The computational bottleneck is therefore also a reproducibility bottleneck. If an analysis can only be run once because it takes several days or repeatedly fails under memory pressure, then it becomes difficult to test alternative preprocessing choices, repeat random splits, or inspect sensitivity to feature filtering. For a master's thesis project, this practical constraint is not incidental; it shapes the research design. A central aim of the work is to identify a workflow that is scientifically honest about its limitations while still being feasible on the hardware available to a student researcher.

For microbiome machine learning, computational feasibility and statistical validity cannot be separated. A pipeline that is too expensive to repeat makes it difficult to estimate uncertainty. A pipeline that filters features without documenting where filtering occurs makes it difficult to assess possible data leakage. A pipeline that executes only on one machine is difficult to reproduce. These issues motivated the reproducible, workflow-managed design of this thesis. The primary analysis was implemented as a Nextflow pipeline developed to test scalability and portability under tight memory constraints. An independent Snakemake implementation was additionally developed as a reproducibility check; because the two pipelines were

not run with an identical procedure, however, the thesis treats Nextflow as the workflow of record and does not present the two as a one-to-one comparison (see the Limitations and Future Work sections of Chapter 5).

1.3 Theoretical Framework: Tipping Elements and Bistability

A second motivation for this thesis is the possibility that microbiome-phenotype relationships are not purely linear. Lahti et al. proposed that parts of the human intestinal ecosystem may exhibit "tipping elements", meaning that some taxa may occupy alternative abundance states rather than varying smoothly across a single continuum [7]. In such a setting, associations between taxa and host phenotype may involve thresholds, state dependence, or interactions among microbes.

This framework does not mean that a machine-learning model can prove ecological bistability from prediction performance alone. A higher R^2 for a Random Forest model is not direct evidence that the gut ecosystem has crossed a tipping point. However, the framework provides a useful modeling rationale. If BMI-related microbial signal is partly non-linear or interaction-rich, then linear models may underfit the data, whereas tree-based models may recover more structure through recursive splitting and implicit interaction terms [8]. The thesis therefore compares linear baselines with non-linear tree-based methods while treating any theoretical interpretation cautiously.

1.4 Research Objectives

The aim of this thesis is to evaluate the feasibility, limits, and computational requirements of BMI prediction from gut microbiome composition. To make that aim

operational, the study is organized around three linked research objectives that correspond to the main modeling, scaling, and reproducibility questions of the project.

RQ1: Predictive capacity and linearity. The first objective is to determine whether species-level gut microbiome composition alone contains enough information to support meaningful out-of-sample prediction of BMI. This objective also evaluates whether the predictive structure is approximated adequately by linear baselines or whether non-linear models provide a consistent advantage, a finding that would be more compatible with interaction-rich or threshold-like relationships in the data.

RQ2: Saturation and sample size. The second objective is to characterize how predictive performance changes as sample size increases across a large metagenomic cohort. Rather than assuming that more samples always produce proportionate gains, this objective tests whether a practical saturation region can be identified, beyond which additional metagenomes contribute only limited improvement under the available feature representation.

RQ3: Scalability and feature reduction. The third objective is to assess whether prevalence filtering and variance-based feature reduction can lower computational burden enough to make large-scale microbiome machine learning feasible on consumer-grade hardware without materially degrading predictive accuracy. This objective connects the statistical problem of high-dimensional sparse data with the practical question of reproducible execution under realistic local-computing constraints.

1.5 Contributions

This thesis makes four main contributions. First, it provides a reproducible, workflow-managed Nextflow benchmark for microbiome-based BMI prediction

on consumer-grade hardware, with an independent Snakemake implementation released as a reproducibility check. Second, it provides a learning-curve analysis that identifies an approximate saturation region for BMI prediction from taxonomic profiles. Third, it evaluates practical feature-reduction strategies for making large-scale species-level modeling feasible on local machines. Fourth, it characterises the BMI target across the cohort — showing that its extreme values reflect participant age and clinical cohort rather than data error — and identifies recurrent high-importance taxa together with the direction of their (weak) association with BMI, while keeping the biological interpretation explicitly exploratory.

The contribution is therefore primarily methodological and computational. The thesis does not claim to establish causal microbial mechanisms for obesity. Instead, it provides a reproducible benchmark for what can be achieved with species-level abundance data, local computing resources, and transparent evaluation design.

2 Literature Review

2.1 The Gut Microbiome in Metabolic Health

The human gastrointestinal tract contains a dense microbial ecosystem that participates in digestion, immune regulation, and metabolic signaling. Gut microbes metabolize dietary fibers into short-chain fatty acids, transform bile acids, synthesize or modify vitamins, and produce metabolites that can interact with host tissues [2]. These functions make the gut microbiome relevant to metabolic-health research, including studies of obesity, insulin resistance, and inflammatory phenotypes. Early germ-free mouse experiments demonstrated that the gut microbiota can function as an environmental factor directly regulating host fat storage, providing foundational experimental evidence for the microbiome-adiposity link [9]. Host-gut microbiota metabolic interactions span multiple biological systems and are mediated through signalling molecules, immune modulation, and substrate processing, illustrating why the microbiome is relevant to a broad spectrum of metabolic phenotypes [10].

Integrated host-microbiome analyses have also shown that microbial composition can be linked with circulating lipid profiles and other metabolic readouts, illustrating that taxonomic data can carry information about host metabolic state even when the underlying mechanisms remain incompletely understood [11]. This type of integrative result is relevant for the present thesis because it shows why microbiome-based prediction can be informative even when the target phenotype, such as BMI, is only one imperfect component of metabolic health.

Recent syntheses of the obesity literature emphasize that the microbiome should be viewed as one component within a broader regulatory system rather than as a single

dominant cause of body-weight variation [12]. Microbial communities can affect substrate availability, gut-brain signaling, inflammation, and host energy metabolism, but these effects are intertwined with diet quality, medication exposure, physical activity, endocrine regulation, and host behavior. This systems view is relevant for the present thesis because it helps explain why microbiome-based BMI prediction may be informative without being exhaustive.

The microbiome is not an isolated system. Its composition is shaped by diet, medication, age, geography, host genetics, sequencing protocol, and many other exposures. This complexity is important for BMI prediction. If the microbiome reflects many correlated host and environmental factors, a predictive model may learn useful signals without those signals being directly causal. For this reason, predictive modeling must be interpreted differently from mechanistic inference. A strong model can show that information is present in the data, but it cannot by itself establish the biological pathway that produced that information.

Large reference projects have repeatedly shown that healthy individuals can vary substantially in microbial composition even when no overt disease is present. The Human Microbiome Project described substantial inter-individual variation across body sites, and global gut microbiome studies have shown that age, geography, diet, and related population context can shape the structure of the gut community [13], [14]. This variation is scientifically useful because it provides signal for prediction, but it is also a source of difficulty because the same taxon can have different implications in different population contexts.

BMI is also a coarse phenotype. It is straightforward to compute and widely available in large metadata tables, but it is not a direct measure of adiposity, diet, insulin sensitivity, or metabolic disease. Two individuals with the same BMI may differ in body composition, medication exposure, physical activity, and metabolic health. This means that even a well-trained microbiome model should not be expected to

explain all BMI variation. In this thesis, BMI is treated as a practical continuous outcome for benchmarking population-scale prediction, not as a complete description of metabolic status.

This caution is consistent with recent reviews of the obesity field. Contemporary syntheses emphasize that the gut microbiota can influence weight-related biology through nutrient processing, intestinal barrier function, metabolite production, inflammation, and appetite-related signaling, but they also stress that these pathways operate within a broader host and environmental system [12], [15]. In human populations, dietary composition, medication exposure, socioeconomic context, and behavior all interact with the microbiome. As a result, even a biologically meaningful microbiome-BMI association may be partial, context dependent, and difficult to transport across cohorts without careful validation.

This heterogeneity also affects how "obesity" should be understood as a modeling target. Weight gain, maintenance, and loss can arise through different behavioral and physiological routes, and similar BMI values may reflect heterogeneous combinations of diet, body composition, treatment history, and metabolic risk [12], [15]. In practice, this means that microbiome signatures associated with one subset of high-BMI individuals may not look identical in another subset. The literature therefore argues against expecting a single universal obesity microbiome signature that is stable across every population. For predictive modeling, this does not eliminate the value of BMI. It means that BMI should be treated as a broad population-scale phenotype whose microbiome signal is likely to be diffuse, mixed with background heterogeneity, and sensitive to cohort composition.

2.1.1 Historical Context: The Firmicutes/Bacteroidetes Ratio

Early obesity-microbiome studies often searched for broad taxonomic signatures. A pivotal study reported that the gut microbiota of obese individuals differed in its proportional representation of major bacterial phyla, with an elevated Firmicutes-to-Bacteroidetes ratio compared with lean counterparts [16]. Subsequent experimental work showed that obesity-associated microbiota had increased capacity for energy harvest and could transfer this capacity to germ-free recipients [3]. A foundational twin study further confirmed that these community-level differences were reproducible and transmissible [17]. These findings collectively motivated the hypothesis that microbial community structure could contribute to host adiposity through altered substrate metabolism. A common simplification in later work was the Firmicutes/Bacteroidetes ratio, which was sometimes presented as a marker of obesity-associated dysbiosis.

The Firmicutes/Bacteroidetes ratio is now understood to be an unreliable standalone marker. Reviews and meta-analyses have shown that its direction and magnitude vary across cohorts and experimental designs [18]. A later reanalysis across multiple human microbiome datasets likewise concluded that obesity-associated microbial signals are generally modest relative to between-study variability and that many individual studies are underpowered to detect them consistently [19]. This does not diminish the importance of the early work. Rather, it shows why microbiome prediction requires more careful modeling than a single phylum-level ratio. Species-level profiles, feature selection, and multivariate learning are better suited to the heterogeneity of human microbiome data.

2.1.2 Species-Level and Functional Associations

Contemporary microbiome research increasingly focuses on species-level, strain-level, and functional signals. This shift reflects two observations. First, broad taxonomic groups can contain organisms with different ecological roles. Second, functionally similar capabilities can be distributed across multiple taxa. A BMI-associated pattern may therefore appear as a weak, distributed signal across many features rather than as one dominant species.

This creates a challenge for interpretation. A model may identify a species as useful for prediction because it is linked to diet, geography, cohort origin, medication, or another unmeasured factor. Conversely, a species may have a plausible metabolic role but still fail to generalize as a predictor across cohorts. The literature therefore suggests a cautious approach: machine-learning results can nominate taxa for discussion, but those taxa should still be interpreted in the light of external biological evidence and independent validation before being treated as mechanistic drivers [4], [5].

Studies using gene-count based microbiome profiling have also shown that low microbial gene richness is associated with adverse metabolic markers including higher adiposity, insulin resistance, and dyslipidaemia, suggesting that overall microbiome diversity carries information about metabolic health beyond individual taxa [20]. Large-scale metagenome-wide association studies have demonstrated that microbiome composition carries disease-relevant signal for phenotypes such as type 2 diabetes, further establishing the analytical feasibility of machine-learning approaches on high-dimensional metagenomic data [21]. Population-based analyses have additionally shown that gut microbiome composition is shaped by a broad range of host and environmental factors simultaneously, underscoring the complexity of any microbiome-phenotype association [22].

The distinction between taxonomic and functional interpretation is especially relevant for feature-importance analysis. A species name is useful as a narrative anchor, but the model is not directly measuring microbial metabolism. It observes relative abundance estimates derived from metagenomic profiling. A high-ranking species may be important because of its own metabolic activity, because it marks a broader community state, or because it correlates with a dietary or geographic exposure. Functional redundancy also means that a pathway may matter even when no single taxon is consistently present across all individuals. This is one reason why the discussion chapter treats top taxa as plausible signals rather than confirmed mechanisms.

The choice of species-level input is therefore best understood as a compromise between biological detail and scalable availability. For the purposes of this thesis, species-level taxonomic profiles were used as a practical common denominator because they are routinely produced by standard metagenomic pipelines and are easier to compare across large collections than richer functional or metabolomic data. Functional pathway profiles and metabolomic measurements may be closer to mechanism, but they often require additional harmonization and were not uniformly available for this benchmark. At the same time, the literature suggests that future gains in prediction may depend on moving beyond taxonomy alone and combining taxonomic, functional, and host-level data where possible [2], [13], [15].

2.2 Theoretical Foundations: Stability and Tipping Points

The concept of tipping elements in the intestinal ecosystem provides a useful ecological background for this thesis. Lahti et al. reported that some gut taxa show abundance patterns compatible with alternative states, where observations cluster near low or high abundance rather than following a simple unimodal distribution [7].

Such patterns may reflect ecological thresholds, although observational abundance distributions alone cannot prove the underlying mechanism.

From a modeling perspective, the tipping-elements framework motivates testing non-linear models. Linear regression and linear SVMs are strong baselines because they test whether the signal can be approximated by additive effects. However, if BMI-related microbial associations depend on interactions or thresholds, then linear models may miss part of the structure. Random Forest models can represent threshold-like behavior through decision-tree splits and can combine multiple features without requiring interaction terms to be specified in advance [8].

The important distinction is between consistency and proof. If Random Forest outperforms linear baselines, the result is consistent with non-linear predictive structure. It does not demonstrate ecological bistability, nor does it identify causal tipping points. This distinction is central to the interpretation used later in the Discussion chapter.

This cautious framing is also consistent with the limits of cross-sectional metagenomic data. Tipping-element theory concerns ecological dynamics and possible alternative stable states. Demonstrating such dynamics would require additional evidence, such as longitudinal observations, perturbation studies, or explicit modeling of state transitions. The present thesis uses the theory in a narrower way: it motivates the comparison between linear and non-linear models and provides language for discussing why threshold-like predictive structure would be biologically plausible.

2.3 Machine Learning in Metagenomics

Machine learning is widely used in microbiome research because sequencing studies generate more features than can be interpreted reliably through univariate testing. Pasolli et al. demonstrated that large metagenomic datasets can support cross-study

machine-learning analyses and can reveal biological patterns when preprocessing and validation are handled carefully [5]. At the same time, methodological reviews emphasize that microbiome machine learning is vulnerable to leakage, confounding, and poor generalization if models are evaluated casually [4], [23].

More recent best-practice reviews from the ML4Microbiome network extend this argument by emphasizing that microbiome machine-learning performance depends on the full analytical pipeline rather than on algorithm choice alone [24]. Choices about filtering, normalization, feature representation, resampling, performance estimation, and interpretation all shape what the final model can reasonably claim. This perspective aligns closely with the design of the present thesis, where workflow transparency and preprocessing order are treated as substantive methodological elements rather than as mere implementation details.

2.3.1 Linear Models

Penalized linear models are common in high-dimensional biological data. Elastic Net combines L1 and L2 regularization, allowing the model to shrink coefficients and reduce variance while retaining implicit feature-selection behavior [25]. In this thesis, Elastic Net functions as an interpretable baseline. If Elastic Net performs close to Random Forest, then the microbiome-BMI signal may be largely additive. If it performs much worse, then the data may contain structure that is not well captured by a linear representation.

Linear support vector machines provide a second linear baseline. The SVM framework is well established for statistical learning [26]. A linear SVM is used here as a second linear baseline for testing whether BMI variation can be approximated adequately in the original feature space. Non-linear kernels could, in principle, capture more complex structure, but they become computationally expensive at the scale of tens of thousands of samples and thousands of microbial features.

2.3.2 Tree-Based Models

Tree-based models are attractive for microbiome prediction because they can handle non-linear patterns and interactions without requiring an explicit parametric form. Random Forest aggregates many decision trees and reduces the instability of individual trees through bagging and feature subsampling [8]. Gradient boosting methods such as XGBoost can also perform strongly on structured prediction tasks [27]. However, boosting often requires more careful hyperparameter tuning and can be more sensitive to software dependencies and hardware compatibility. In this project, XGBoost was explored but not retained as a stable main result in the final Nextflow interpretation because of platform-specific problems.

Tree-based models also provide variable-importance outputs, which are useful but should be interpreted carefully. Importance scores depend on the model, the preprocessing, the feature set, and correlations among predictors. If two taxa are strongly correlated, importance may be split between them or assigned preferentially to one of them. Therefore, feature importance is best used as a ranking for follow-up interpretation, not as a direct measure of causal effect size. This thesis follows that approach by comparing whether the same taxa recur across independent workflows and by interpreting those taxa through existing biological literature.

2.3.3 Dimensionality, Sparsity, and Compositionality

Microbiome data are not ordinary tabular data. They are high-dimensional because each sample can contain thousands of species-level features. They are sparse because many taxa are absent, rare, or below the detection threshold in most samples. They are compositional because sequencing-based abundance tables are usually expressed as relative values constrained by the total number of reads or by normalization to a constant sum [28], [29]. These properties complicate both statistical inference and machine learning.

High dimensionality increases the risk that a model will learn accidental patterns rather than stable biological signal. Sparsity increases instability because rare features may be present in only a small number of samples. Compositionality means that an apparent increase in one taxon can be partly induced by decreases in other taxa. These constraints do not make microbiome machine learning impossible, but they require explicit preprocessing and cautious interpretation.

Feature filtering is one practical response to these challenges. A prevalence filter removes features that are too rare to support stable population-scale modeling, while variance filtering removes features with little inter-individual variation. These steps reduce the number of candidate predictors before model training. The statistical trade-off is that filtering can improve stability and feasibility but may remove rare subgroup-specific signals. The engineering trade-off is that without filtering, large metagenomic matrices may exceed memory limits on local hardware. This thesis evaluates that trade-off empirically rather than assuming that more features are always better.

2.3.4 Validation and Leakage

Evaluation design is one of the most important issues in microbiome machine learning. Topcuoglu et al. argue that supervised microbiome analyses should report preprocessing, feature selection, data partitioning, and model-tuning choices clearly because each step can influence performance estimates [23]. Leakage can occur not only when the outcome variable is accidentally included as a predictor, but also when feature selection or transformation parameters are estimated using information from held-out samples.

This point is especially relevant in large sparse feature matrices. Prevalence filtering may seem biologically harmless because it removes rare taxa, but if the filter is computed on all samples before train/test splitting, the held-out data have still

influenced which features are available to the model. The effect may be small when the cohort is large and the threshold is simple, but it should still be reported. This thesis therefore distinguishes between direct outcome leakage, which was prevented by excluding BMI-derived or non-microbial variables, and global prefiltering, which remained a limitation of the implemented workflows.

The strongest validation design would nest all preprocessing, feature selection, hyperparameter tuning, and model fitting inside the resampling procedure. In that design, each training fold would learn its own filters and transformations, and those learned operations would then be applied to the corresponding held-out fold. This design is more computationally expensive, but it gives the cleanest estimate of generalization. The workflows in this thesis move toward reproducible evaluation, but they do not fully implement nested preprocessing. Reporting this limitation explicitly is part of the methodological contribution because it makes the interpretation of the performance estimates more transparent.

2.3.5 Cross-Cohort Generalization and Portability

A recurring concern in microbiome machine learning is that within-cohort performance can look stronger than cross-cohort performance. Models may learn patterns that reflect not only biology, but also cohort composition, geography, diet, sequencing workflow, and taxonomic-processing choices [4], [5]. This issue is especially relevant for human gut microbiome studies because the baseline composition of the gut community can differ substantially across populations [13], [14].

This distinction between internal prediction and external portability is important for interpreting reported accuracy. A model can perform well on held-out samples from the same aggregated dataset and still generalize poorly to a different cohort.

Therefore, a large sample size does not automatically guarantee broad transportability. Heterogeneity can help a model learn richer patterns, but it can also expose how sensitive those patterns are to cohort composition and preprocessing choices.

From a data-science perspective, this problem can be described as one of external validation and distribution shift. Predictive performance estimated within a pooled dataset may not transfer reliably to independent cohorts if the underlying data distributions differ because of geography, diet, medication exposure, sequencing protocol, preprocessing choices, or cohort composition. Strong internal validation therefore does not by itself establish that a model captures broadly generalizable structure. External transportability is a stricter standard for assessing whether predictive signal reflects reproducible phenotype-related information rather than dataset-specific patterns.

For this reason, reproducibility in microbiome machine learning should be understood at more than one level. There is code reproducibility, meaning that the same scripts and workflows can be rerun. There is analytical reproducibility, meaning that similar conclusions emerge under related but not identical implementations. And there is external generalization, meaning that the predictive pattern survives beyond the dataset on which the model was developed. The present thesis is strongest on the first two levels, while also identifying the third as an important target for future work.

2.3.6 Prediction, Interpretation, and Clinical Use

The literature also raises a broader conceptual point: predictive usefulness is not equivalent to clinical readiness. A model may achieve non-trivial out-of-sample performance and still be unsuitable for decision-making if its transportability, dependence on upstream preprocessing, and behavior across different populations are

not well understood. This is especially relevant in microbiome studies, where taxonomic estimates depend on sequencing depth, reference databases, and bioinformatic choices in addition to underlying biology [4], [23].

For that reason, benchmark studies serve an important role even when they do not deliver deployable predictors. They establish what can be extracted from a given representation of the data, show where performance seems to plateau, and reveal which parts of the workflow dominate sensitivity and computational cost. In a field where biological interpretation, software reproducibility, and statistical validity are tightly coupled, this kind of benchmarking is itself a substantive contribution rather than merely a preliminary exercise.

This perspective is useful for the present thesis. A microbiome-only BMI benchmark should not be judged by whether it immediately enables clinical obesity screening. It should be judged by whether it measures the available signal honestly, reports the workflow assumptions clearly, and identifies where future studies would gain more from richer data or stricter validation than from repeating the same analysis at greater scale. That is why the thesis treats model performance, saturation behavior, and workflow transparency as co-equal outcomes rather than as separate technical appendices.

2.4 Workflow Management and Reproducibility

Reproducibility is a central requirement in bioinformatics. A computational analysis is difficult to evaluate if the software environment, file dependencies, and execution order are not explicit. Workflow managers address this problem by formalizing the steps of an analysis and by recording how intermediate outputs are produced. Nextflow is widely used in scalable bioinformatics because it supports portable dataflow pipelines and deployment across local, cluster, and cloud systems [30]. Snakemake uses a rule-based approach in which files and dependencies define

the workflow graph. It has become a widely used workflow engine in bioinformatics [31]. Software environment management through Conda and community-curated channels such as Bioconda [32] further supports reproducibility by making dependency specifications explicit and shareable.

This thesis uses Nextflow as its primary workflow manager, chosen for portability and scalable execution under tight memory constraints. An independent Snakemake implementation, built around the mikropml framework, was also developed as a reproducibility check. Because the two implementations were not configured to run an identical procedure, the thesis does not treat them as a numerical replication of one another: the Snakemake implementation is reported only as supplementary evidence that the analysis can be reconstructed in a second workflow system, and the question of fully synchronising the two is left to future work (Chapter 5).

Workflow systems are also valuable because they make negative results and constraints easier to document. In this thesis, the XGBoost instability on Apple Silicon is not treated as a biological result, but it is still relevant to reproducibility. A method that is theoretically attractive may be less useful if dependency conflicts prevent stable execution on the target platform. Recording these constraints helps other researchers understand why the final interpreted model set differs from the initial plan.

2.5 Scalability and Saturation

Large microbiome cohorts raise a practical question: how many samples are enough for stable prediction? Learning-curve analysis addresses this question by measuring predictive performance across increasing sample sizes and identifying the point at which additional data yield diminishing returns. If performance keeps improving sharply, the task may be limited by sample size. If performance plateaus, the limiting

factor may instead be the information content of the available features, unmeasured confounders, model class, or noise in the outcome.

For BMI prediction, saturation is particularly important because BMI is influenced by many factors not captured in species-level relative abundance data. Diet, medication, geography, host genetics, physical activity, and sequencing batch can all shape the relationship between microbiome composition and body size. A plateau in predictive performance should therefore not be interpreted as a universal limit of the microbiome. It is more precisely a limit of the present representation, dataset, workflows, and models.

This thesis contributes to the literature by combining saturation analysis with local-computing constraints. The goal is not only to ask whether prediction improves with more data, but also whether the required analysis can be performed reproducibly on ordinary hardware through careful filtering and workflow design.

This framing is deliberately pragmatic. A saturation threshold estimated from one dataset should not be treated as a universal sample-size rule for all microbiome-BMI studies. It depends on the phenotype definition, taxonomic resolution, preprocessing, model class, cohort heterogeneity, and validation design. However, it can still provide useful planning information. If performance gains become small after a certain point, then future studies may benefit more from richer metadata, better external validation, longitudinal sampling, or multi-omics integration than from simply adding more similar samples.

The same logic applies to feature engineering. If a learning curve flattens while using species-level relative abundance profiles, the plateau may reflect the limits of that representation rather than the limits of the microbiome itself. In that sense, the result is consistent with the possibility that richer biological resolution, such as diet, metabolite production, host heterogeneity, or microbial function, may eventually be

more informative than simply adding more rows of the same input representation [12], [15]. This makes saturation analysis valuable not only as a sample-size tool, but also as a guide for deciding when additional biological resolution may be more informative than additional rows in the same matrix.

2.6 Research Gap

Taken together, the existing literature motivates the present thesis but does not fully resolve the methodological combination addressed here. Prior work has shown that microbiome composition can be associated with BMI and other host phenotypes, and machine-learning studies have demonstrated that these data can support prediction under some conditions [4], [5], [23]. However, the combination emphasized here, namely large-sample microbiome-BMI prediction together with explicit workflow reproducibility and consumer-hardware feasibility, appears to be less often documented in the existing literature.

This gap matters because predictive accuracy alone is not enough for a strong computational thesis. If a result cannot be reproduced across workflows, if the preprocessing order is not explicit, or if the computation is impractical outside specialized environments, then the scientific usefulness of the benchmark is reduced. The present thesis addresses that gap by treating scalability, reproducibility, and evaluation design as central components of the research question, not merely as technical afterthoughts.

3 Methodology

3.1 Dataset: The Metalog Consortium

This study used a merged analysis matrix derived from the Metalog database [33], combining harmonised human metadata with species-level gut metagenomic taxonomic profiles. Species abundances were produced using MetaPhlAn 4, which profiles known and uncharacterised microbial taxa using clade-specific marker genes [6].

Source data. The working matrix for the primary (Nextflow) analysis was created by merging `human_extended_wide_2025-12-14.tsv`, containing harmonised metadata, with `human_metaphlan4_species_2025-12-14.tsv`, containing species-level taxonomic profiles (10,701 distinct species). The independent Snakemake reproducibility implementation (Section 3.2.2) was built from a near-identical later snapshot (2025-12-28); the two snapshots are nested, with the later release a superset of the earlier one, so they describe essentially the same data.

Sample count. The raw merge yielded 18,024 samples. After quality-control and overlap checks, the final analysis set contained 17,903 human gut metagenomes. Table 3.1 summarises the key dataset characteristics.

Table 3.1: Summary of the analysis dataset derived from the Metalog consortium. Feature counts refer to the primary Nextflow preprocessing pipeline.

Characteristic	Value
Raw samples after merge	18,024
Samples excluded (QC or missing BMI)	121
Final analysis samples	17,903
BMI range (kg/m^2)	10.0–76.7
BMI mean ± SD (kg/m^2)	24.59 ± 6.21
Species features before filtering	10,701
Features after 1% prevalence filter	1,799
Features after variance top-500 filter	500
Outcome variable	BMI (continuous)
Predictor scope	Species-level relative abundance only
Excluded metadata	Age, sex, weight, sample identifier

Data access and ethics. This study used publicly available, anonymised secondary metagenomic data from the Metalog resource [33]. Data were accessed and analysed for academic research purposes in accordance with the repository's licensing and reuse conditions. Because the study involved only secondary analysis of pre-existing anonymised records and no new collection of human participant data, no additional ethics review was required for this thesis. The full Metalog data tables, including the harmonised human-sample metadata and the MetaPhlAn 4 taxonomic profiles used in this work, are released under a Creative Commons (CC-BY) licence and are downloadable from `https://metalog.embl.de/downloads`.

Outcome. Body mass index (BMI) was modeled as a continuous regression target. BMI was selected because it is widely available in large human metadata collections

and is therefore suitable for population-scale regression benchmarking, even though it remains a simplified proxy for metabolic status.

Predictor scope. The predictive task was intentionally restricted to microbiome composition. Two types of variables were excluded from the predictor matrix before model fitting: those that directly duplicated or trivially reconstructed the outcome (e.g., weight), and non-microbial metadata columns (e.g., age, sex). The modelling tables therefore retained a sample identifier, BMI, and species-abundance features only; where an intermediate table contained additional non-microbial fields such as `age_years` or `sex`, these were removed before model fitting. The identifier column was retained only for bookkeeping during matrix construction and was dropped before training: in the Nextflow saturation script `sample_id` is explicitly removed before model fitting, and in the processed Snakemake object it is absent from `dat_transformed`. Thus, neither implementation treated sample identifiers or non-microbial metadata as predictive features.

3.1.1 Study Population and BMI Quality Control

Because BMI is the prediction target, its distribution across the pooled cohort was examined directly before modelling (the descriptive analysis is reported in Section 4.1). The modelled BMI values span 10.0–76.7 $\mathrm{kg/m^2}$, a range whose tails are physiologically extreme and therefore require explanation. Three checks were used to decide whether these values represent data-entry errors or genuine cohort variation. First, for the 7,248 samples that also recorded weight and height, BMI was recomputed from those fields; the recomputed and recorded values agreed almost exactly (median absolute difference 0.003 $\mathrm{kg/m^2}$, Pearson $r = 0.99$, 96% within one BMI unit), including at the extreme tails, indicating that the extreme values are internally consistent rather than transcription errors. The underlying height and weight values are themselves physically plausible (Figure 3.1): across the 5,670 adults with both measurements, height ranged 139–200 cm and weight 33–171 kg

with no impossible combinations, the highest BMIs arising from adults of ordinary height carrying very high weight (for example 160 kg at 156 cm, BMI 66) and the lowest from newborns (for example 3.2 kg at 53 cm, BMI 11). This rules out the possibility that height, weight, and BMI were jointly mis-recorded. Second, the extreme records were linked back to their study-level metadata (age, study, geography, disease status). The low tail (BMI < 15, $n = 291$) is dominated by infants and young children — 82% are younger than 18 years and 61% younger than 5 — drawn largely from an infant cystic-fibrosis cohort, a paediatric Zimbabwean cohort, and an anorexia-nervosa cohort, all settings in which low BMI is expected. The high tail (BMI > 50, $n = 149$) consists entirely of adults (median age 68 years) drawn from bariatric-surgery, elderly, and metabolic-disease (NAFLD, type 2 diabetes) cohorts, in which severe obesity is expected. The extremes are therefore attributable to participant age and clinical cohort, not to measurement error.

On this basis the full cohort of 17,903 samples was retained for the primary analysis, because the extreme values are genuine and excluding them would discard real biological variation. To confirm that the headline results are not driven by the age-related low tail, an adult-only sensitivity analysis (age ≥ 18, $n = 15{,}368$; Section 4.1) was also examined; restricting to adults removes most of the implausible-looking low tail (samples with BMI < 15 fall from 291 to 35) while leaving the overall distribution and conclusions essentially unchanged. It should be noted that BMI is not a conventional measure of adiposity below approximately two years of age, where weight-for-length and age-specific growth references are normally used instead; the 1,288 infant samples (age < 2 years) are therefore retained only as part of the pooled, population-scale benchmark rather than as clinically interpreted BMI values, and the adult-only sensitivity analysis isolates the adult microbiome–BMI signal from this group.

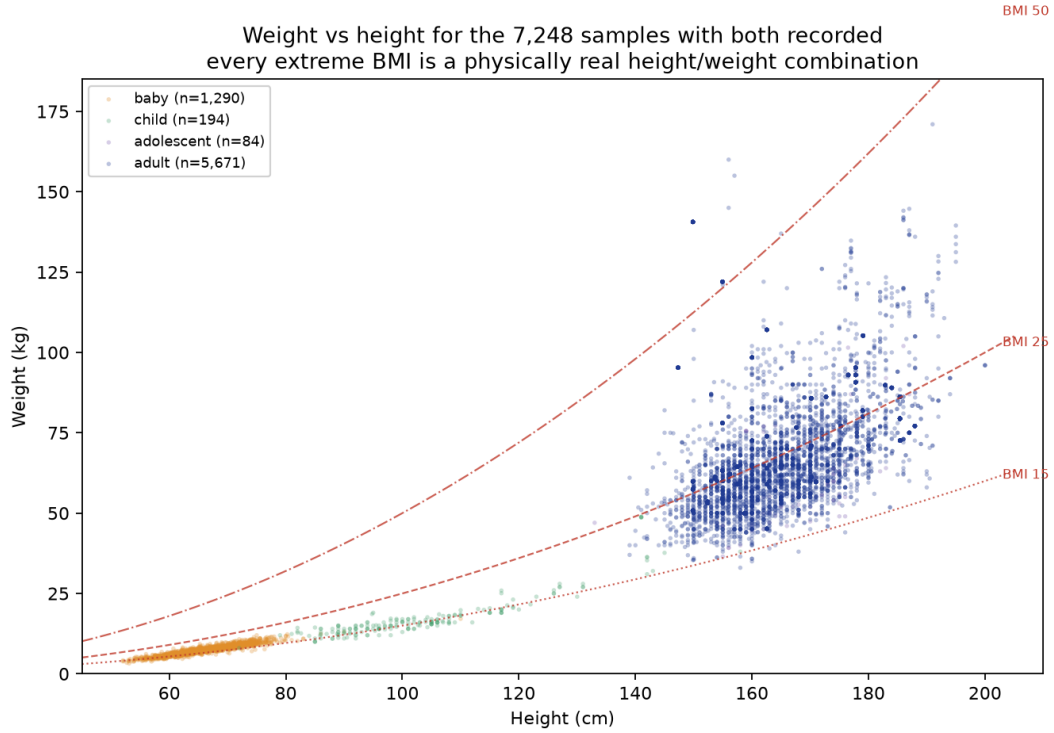

Figure 3.1: Recorded weight against height for the 7,248 samples with both measurements, coloured by age category, with lines of constant BMI (15, 25, 50 $\mathrm{kg/m^2}$). Infants and young children form the low-height, low-weight cluster (a low BMI is normal at this age), while adults span ordinary heights (139–200 cm). The most extreme BMIs correspond to physically real height/weight combinations — newborns of a few kilograms at the low end, and adults of ordinary height carrying very high weight at the high end — with no impossible combinations, confirming that the BMI extremes are not the product of jointly mis-entered height and weight.

3.1.2 Data Matrix Construction

The practical input to the machine-learning workflows was a sample-by-feature matrix. Rows corresponded to individual metagenomic samples and columns corresponded to the regression outcome, optional sample identifier, and microbial species features. The metadata and species-abundance tables were merged by sample identifier so that each retained row had both BMI metadata and taxonomic abundance information. Samples without a usable BMI value or without matching taxonomic data were excluded during construction of the working matrix.

The final analysis used 17,903 human gut metagenomes. This number is slightly smaller than the raw merge because the analysis required complete overlap between metadata and species profiles after quality checks. This distinction is important for reproducibility: the raw merged table and the final modeling table are not identical objects. The thesis therefore reports both the raw merge count and the analysis count.

3.1.3 Outcome Definition and Predictor Exclusion

BMI was treated as a continuous outcome rather than converted into categorical obesity classes. This choice preserves more information than dichotomization and avoids arbitrary threshold effects in the target variable. The regression task asks whether microbiome composition can estimate variation in BMI across individuals, not whether it can classify individuals into clinical BMI categories.

The exclusion of non-microbial metadata was deliberate. Age and sex can improve BMI prediction in many datasets, and weight directly defines BMI when height is available. Including such variables would make the prediction task less specific to microbiome composition. Therefore, they were removed from the predictor matrix even though their exclusion may lower absolute predictive performance. The aim was to benchmark the microbiome-only signal.

3.1.4 Rationale for a Microbiome-Only Benchmark

This design choice deserves explicit justification because it shapes how the results should be interpreted. The thesis does not attempt to build the strongest possible BMI predictor from all available variables. If that had been the goal, demographic covariates and outcome-adjacent anthropometric variables would have been retained, and additional metadata would likely have improved predictive performance. Instead, the narrower question was whether large-scale species-level metagenomic data

alone contain enough signal to support meaningful BMI prediction under transparent computational constraints.

There are two reasons for treating this as a benchmark problem. First, microbiome studies often differ widely in which non-microbial metadata are available or harmonized. A microbiome-only setup therefore provides a more transferable baseline across datasets than a richer model that depends on covariates not consistently recorded in every cohort. Second, a restricted benchmark makes it easier to interpret what the models are learning. If age, sex, and weight are excluded, then improved performance by a flexible model can be discussed primarily in relation to microbial composition rather than to obvious non-microbial predictors. This does not remove confounding, but it makes the benchmark question more coherent.

3.2 Computational Architecture

The primary analysis was implemented as a single Nextflow pipeline, which emphasised portability, multi-model comparison, and memory-aware feature reduction on an Apple Silicon laptop. The pipeline is summarised in Figure 3.2. An independent Snakemake implementation was additionally developed as a reproducibility check; it is described briefly in Section 3.2.2. Because the two implementations were not configured to run an identical procedure, the Nextflow pipeline is treated throughout as the workflow of record, and the two are not presented as a one-to-one comparison.

3.2.1 The Nextflow Pipeline

The Nextflow workflow was implemented using Nextflow DSL2 [30]. The codebase contains a main workflow file, an environment configuration, and R/Python scripts for data merging, preprocessing, model training, saturation analysis, feature extraction, plotting, and exploratory XGBoost/SHAP analyses. The workflow was run on an Apple MacBook with an ARM64 architecture and an 8 GB RAM constraint.

The main Nextflow pipeline first calls a preprocessing process that reads the merged Metalog CSV and produces a preprocessed RDS object. The preprocessing script protects metadata columns, identifies species features, and computes prevalence across the full input matrix. It then applies a 1% prevalence threshold and retains the top 500 most variable species if more than 500 features remain. This variance-based top-500 filter was designed to reduce the feature space aggressively while retaining the most informative inter-individual variation. The script then removes near-zero-variance features and centers/scales numeric predictor columns, excluding the outcome column.

The saturation script reads the preprocessed RDS object and evaluates sample sizes from 2,000 to 16,000 in increments of 2,000. For each sample-size step, the script draws one random subset using a fixed seed (`seed = 42` for all subset draws). It then trains selected models with `mikropml::run_ml()` [23], [34]. In this implementation, an 80/20 training/test split (`training_frac = 0.8`), 3-fold cross-validation (`kfold = 3`), and one cross-validation repeat (`cv_times = 1`) were used. The separate `train_ml.R` script, which produced the saved `rf_model.rds` and its feature-importance output (Figure 4.11), used `seed = 100` and the full preprocessed dataset; this model yielded $R^2 = 0.351$, which is lower than the saturation-peak result of 0.387 because the full cohort introduces more heterogeneity than the N=16,000 subset. Random Forest, linear SVM, and decision-tree results were retained for interpretation. XGBoost was explored in the code through `xgbTree` calls wrapped in a `tryCatch` block, but platform-specific compatibility and stability issues on Apple Silicon prevented it from yielding a stable retained benchmark output. As a result, the Nextflow workflow functioned primarily as a multi-model comparative and scaling benchmark rather than as a repeated-seed uncertainty study.

This setup was chosen because the Nextflow workflow had to balance several objectives simultaneously: testing multiple models, evaluating a sample-size ladder,

producing figures, and staying within local memory limits. The top-500 filter was therefore not only a statistical choice but also an execution strategy. Without aggressive feature reduction, model fitting on the full feature matrix created long runtimes and a higher risk of memory failure.

The Nextflow codebase also supports reproducibility beyond the final PDF. The workflow file records process order, the R scripts record filtering and modeling settings, and the repository provides a minimal execution structure for re-running the pipeline when the required input data are available. This is why the Methods chapter reports implementation details that may seem technical: they are necessary for another researcher to understand what was actually evaluated.

An additional consequence of this design is that the Nextflow workflow acts as a stress test for methodological feasibility under constrained resources. It asks not only which model performs best, but also which combination of filtering, splitting, and model fitting remains executable often enough to support comparison. That emphasis is important for the thesis because a benchmark that cannot be rerun reliably is less scientifically useful than a slightly simplified benchmark that can be audited, repeated, and extended.

3.2.2 Independent Reproducibility Implementation (Snakemake)

To check that the analysis was not tied to a single workflow system, the pipeline was also implemented independently in Snakemake [31], using a rule-based structure around the mikropml package and run on a separate Lenovo ThinkPad (x86-64). This second implementation reproduced the same overall logic — merge, 1% prevalence filtering, and model training with `mikropml::run_ml()` — and confirmed that interpretable microbial structure and a comparable saturation pattern could be recovered in a different software stack on different hardware.

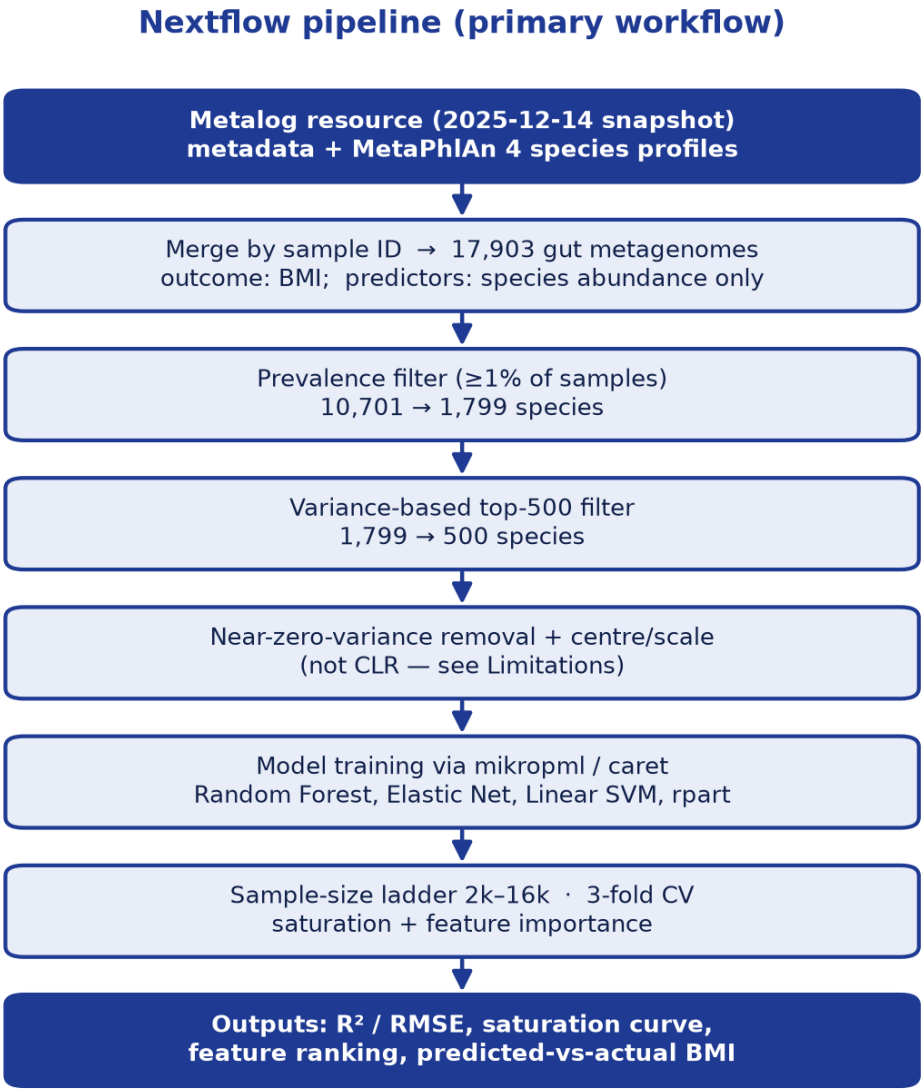

Figure 3.2: The primary Nextflow pipeline. Starting from the merged Metalog matrix (BMI as outcome, species relative abundances as the only predictors), the pipeline applies a 1% prevalence filter (10,701 → 1,799 species), a variance-based top-500 filter (1,799 → 500 species), near-zero-variance removal and column-wise centring/scaling (not a centred log-ratio transform; see Section 3.3.1 and the Limitations), and then trains the benchmarked models with three-fold cross-validation across a 2,000–16,000 sample-size ladder, producing the performance, saturation, feature-importance, and predicted-versus-actual outputs reported in Chapter 4.

This Snakemake implementation was not, however, configured to run an *identical* procedure to the Nextflow pipeline: it used a near-identical later Metalog snapshot, did not apply the variance-based top-500 filter, used a different set of fixed sample-size subsets, and used `kfold = 2` with repeated seeds rather than the Nextflow `kfold = 3` single-draw design. For that reason the two implementations are not compared one-to-one, and the Snakemake outputs are not presented as a parallel result set in Chapter 4. The Snakemake implementation is reported here only as evidence of cross-workflow reproducibility; aligning the two pipelines into a single identical procedure to isolate the effect of the workflow manager itself is identified as future work (Chapter 5). The code remains publicly available (Section 3.6).

3.3 Preprocessing and Feature Selection

The raw species-abundance matrix was high dimensional and sparse. The Nextflow pipeline therefore combined two feature-reduction strategies: a 1% prevalence filter followed by a more aggressive variance-based top-500 filter, in order to support multi-model benchmarking under a tight memory ceiling.

3.3.1 Prevalence Filtering

The prevalence filter removed taxa that were present at non-zero abundance in fewer than 1% of samples. In the Nextflow pipeline this step reduced the species feature space from 10,701 to 1,799 columns, before the top-500 variance step (the independent Snakemake implementation applied the same 1% threshold). The rationale was computational and statistical: extremely rare taxa contribute to sparsity and increase memory use, while offering limited support for regression in this cohort. This does not mean that rare taxa are biologically unimportant; it means they were not the primary target of this predictive benchmark.

The 1% threshold corresponds to roughly 180 individuals in a cohort of about 18,000 samples. Features observed in fewer individuals than this can still be biologically interesting, but they are difficult to use reliably in a broad regression benchmark because the model has little information about their relationship to BMI. The threshold was therefore used as a pragmatic population-scale filter rather than as a claim about biological relevance.

3.3.2 Variance-Based Feature Selection

The Nextflow workflow included a more aggressive variance-based feature selection step. After prevalence filtering, if more than 500 species remained, features were ranked by variance across the input matrix and only the 500 most variable species were retained. This variance-based top-500 filter was developed because the full matrix could not be modeled comfortably under the MacBook memory constraint. The goal was to retain inter-individual variation while reducing runtime and memory pressure.

Variance filtering was chosen because it is unsupervised with respect to BMI. It does not rank features by their correlation with the outcome; it ranks them by how much they vary across samples. This reduces the risk of direct outcome leakage compared with supervised feature selection. Nevertheless, because the variance was computed before the final train/test split in the implemented Nextflow workflow, the step is still reported as global prefiltering rather than as fully nested feature selection.

3.3.3 Transformation and Leakage Considerations

Direct outcome leakage was addressed by excluding variables that directly define or trivially reconstruct BMI, including weight where present, and by excluding non-microbial metadata such as age and sex from the predictor matrix. This forced

the models to learn from microbiome composition rather than from demographic or anthropometric covariates.

However, leakage prevention was not complete across all preprocessing stages. In the Nextflow pipeline, prevalence filtering, variance filtering, near-zero-variance filtering, and centering/scaling were applied to the full matrix before downstream train/test partitioning and cross-validation (the independent Snakemake implementation likewise applied its prevalence filter globally before evaluation). Therefore, the reported performance estimates should be interpreted as held-out evaluation after global prefiltering, not as estimates from a fully nested feature-selection pipeline.

This limitation is important but bounded. The global filters used here were unsupervised with respect to BMI: they were based on feature presence, variance, and numerical scaling rather than on feature-outcome association. Nevertheless, because held-out samples influenced the retained feature set and transformation parameters, a fully nested implementation would be preferable for final confirmatory benchmarking.

The thesis therefore separates three different leakage questions. First, direct outcome leakage was prevented by removing variables that encode BMI or non-microbial metadata. Second, supervised feature-selection leakage was avoided because the filtering steps did not use BMI to rank features. Third, unsupervised global-preprocessing leakage remained because filters and scaling parameters were estimated before the final split. This third issue is the one acknowledged in the Limitations section. It does not invalidate the workflow as an exploratory benchmark, but it means the performance values should not be presented as final clinical-grade generalization estimates.

For a future implementation, the stricter design would be: split the data first, compute prevalence and variance filters on the training fold only, estimate scaling

parameters on the training fold only, train the model, and then apply the learned filtering/scaling operations to the held-out fold. Repeating that procedure inside each resampling iteration would remove the main global-preprocessing limitation.

Compositionality and CLR transformation. Metagenomic relative abundance data are compositional in the statistical sense: values in each sample sum to a constant, meaning that the apparent increase of one taxon is mathematically coupled to changes in others [28], [29]. A principled response to this is centred log-ratio (CLR) transformation, which maps relative abundances into an unconstrained space suitable for standard regression. In this thesis, however, CLR transformation was not applied: the Nextflow pipeline (confirmed in the preprocessing script `preprocess.R`) instead applied standard centring and scaling to the prevalence- and variance-filtered abundance columns, and the independent Snakemake implementation used mikropml's default centring/scaling. This is acknowledged as a limitation. Its practical impact is bounded for the headline result, because the best-performing model, Random Forest, is tree-based and splits on one feature at a time, which makes it relatively insensitive to the monotonic rescaling that distinguishes centred/scaled from CLR-transformed inputs; the effect would be larger for the linear baselines (Elastic Net and linear SVM). CLR or a related isometric log-ratio transform would nonetheless provide a more rigorous treatment of the compositional structure and is recommended for future implementations, ideally estimated inside each training fold to avoid preprocessing leakage.

3.4 Machine-Learning Models

The analysis compared model families with different assumptions about the microbiome-BMI relationship.

Elastic Net (GLMNet). Elastic Net combines L1 and L2 penalties and is a strong linear baseline for high-dimensional data [25]. It is useful for testing whether the signal can be captured by additive feature effects.

Linear SVM. Linear support vector machines provide another linear baseline grounded in statistical learning theory [26]. In this thesis, linear SVM was used to test whether a simple hyperplane-like representation was adequate.

Random Forest. Random Forest is a non-linear ensemble of decision trees [8], [35]. It was the main flexible benchmark because it can capture threshold-like structure and interactions without requiring explicit interaction terms.

Decision Tree (rpart). A single decision tree (rpart) was included as a simple non-linear baseline. It helps distinguish the value of ensemble averaging from the value of recursive partitioning alone.

XGBoost. XGBoost was explored because gradient boosting can perform strongly in structured prediction tasks [27]. However, XGBoost was not retained as a stable final interpreted result: in the Nextflow pipeline, repeated Apple Silicon compatibility and stability problems prevented a consistent retained benchmark output. XGBoost is therefore discussed as an explored but ultimately non-interpreted model family.

3.4.1 Hyperparameter Handling

All models were trained through the mikropml package [34], which wraps the caret framework [36]. The model-tuning strategy was intentionally conservative because the project was constrained by local hardware and because the primary aim was benchmarking model classes and sample sizes rather than exhaustive algorithm optimization. In the Nextflow saturation analysis the models were run through mikropml with a lightweight cross-validation setup and the mikropml/caret defaults. This

made the analysis feasible across multiple sample sizes, but it means the reported performance should be interpreted as benchmark performance under practical tuning constraints, not as the maximum possible performance of each algorithm.

This choice affects interpretation. Random Forest outperforming the linear baselines is informative because the gap is large, but the exact numerical ranking among all possible models could change with more extensive hyperparameter tuning, additional model families, or HPC-scale experimentation. The thesis therefore emphasizes robust qualitative patterns rather than claiming that the selected Random Forest configuration is globally optimal.

3.5 Experimental Design: Scaling and Saturation

3.5.1 Sample-Size Ladder

The Nextflow saturation workflow evaluated 2,000 to 16,000 samples in increments of 2,000. This ladder was designed to locate a practical saturation region for BMI prediction from species-level taxonomic profiles.

3.5.2 Cross-Validation Depth and Its Trade-Offs

The Nextflow workflow used `kfold = 3` with one subset draw per sample-size step, prioritising multi-model comparison across a wide sample-size ladder under a strict 8 GB RAM constraint. Higher-fold cross-validation (5-fold, 10-fold) was evaluated but exceeded the computational constraints of the consumer-grade platform used in this benchmark. With 17,903 samples, even 3-fold cross-validation allocates approximately 11,935 training samples per fold — a training set substantially larger than most published microbiome ML studies. At this scale, the variance-reduction benefit of additional folds is marginal relative to the linear increase in computational cost. This trade-off is itself a methodological finding: it characterises the practical

limits of cross-validation depth on accessible hardware for large-scale metagenomic datasets. To partially validate the 3-fold design, 5-fold cross-validation was applied to a randomly drawn subset of 5,000 samples on an external compute resource. Results were nearly identical to those obtained using 3-fold CV on the same subset ($R^2 = 0.264$, RMSE = 5.28 for 3-fold CV; $R^2 = 0.265$, RMSE = 5.27 for 5-fold CV), supporting the stability of the adopted evaluation design.

3.5.3 Metrics

Model performance was summarized using Root Mean Squared Error (RMSE) and the coefficient of determination (R^2). RMSE reports error in BMI units, while R^2 reports the proportion of variance explained relative to a baseline model. Training time and memory feasibility were also considered because computational feasibility was one of the research questions.

RMSE was included because it is directly interpretable in outcome units (kg/m^2): lower RMSE indicates that predicted BMI values are, on average, closer to measured values. R^2 was included because it allows comparison of explained variance across models. In a heterogeneous cross-sectional population, a moderate R^2 can still be meaningful if the model uses only microbiome features and excludes demographic or anthropometric covariates. However, neither metric establishes causality. They measure predictive accuracy under the specified evaluation design.

Runtime and memory feasibility were treated as secondary methodological outcomes because the thesis explicitly asks whether large-scale microbiome modeling can be done on consumer-grade hardware. These engineering quantities are not auxiliary implementation details; they determine whether repeated benchmarking is even possible. A workflow that yields slightly stronger accuracy but cannot be rerun reliably on the target hardware is less useful for this thesis than a workflow that is somewhat simpler but reproducible in practice.

3.6 Code Availability and Reproducibility

Both implementations have been archived as formally citable software artefacts: the primary Nextflow workflow [37] and the independent Snakemake reproducibility implementation [38], each released under the Artistic 2.0 open-source licence and assigned a permanent Digital Object Identifier (DOI) through Zenodo, following established best practices for reproducible scientific computing [39]. These repositories contain the workflow definitions, preprocessing scripts, model settings, result-generation logic, and supporting documentation needed to inspect preprocessing order, sample-size settings, cross-validation parameters, model choices, and generated figures. The large Metalog-derived data tables are not redistributed in the public repositories because of dataset size and data-governance constraints.

Public code availability improves the reproducibility of the thesis in two ways. First, it allows another researcher to inspect the exact implementation of the filtering order, subset generation, model calls, plotting routines, and workflow dependencies rather than relying only on the narrative description in the thesis text. Second, it makes it easier to extend the benchmark in future work, for example by adding nested preprocessing, external-validation cohorts, or alternative model families under the same workflow structure.

The public repositories should therefore be understood as reproducibility materials rather than as complete data releases. They expose the analytical logic of the thesis and make the implementation auditable, but full re-execution still requires access to the underlying Metalog-derived inputs and to a compatible software environment for the R- and workflow-based dependencies.

To make the repository state more specific at the time of thesis finalization, the workflows can also be identified by fixed Git commit hashes. The Nextflow workflow snapshot [37] corresponds to commit `f565f74d6b5127bd92e42edf94dd9f9118f2b766`

(abbreviated form: `f565f74d`), and the Snakemake workflow snapshot [38] corresponds to commit `6185a3749310228a72ac84d4bab9a6372d58f41a` (abbreviated form: `6185a374`). These hashes identify the exact public repository states used as the code references for the thesis, even without a separate release archive.

Taken together, these repositories allow the analytical logic of the thesis to be inspected at the level of workflow definitions, preprocessing scripts, model calls, and plotting routines. They also make it possible to inspect how the analysis was reconstructed in a second, independent workflow system. In that sense, the repositories serve not only as code archives but also as methodological documentation supporting reproducibility, extension, and critical re-evaluation of the benchmark.

Nextflow workflow: GitHub `https://github.com/Sadia12345/Microbiome_BMI_Thesis_Nextflow`; archived release `https://doi.org/10.5281/zenodo.20569792`.

Snakemake workflow: GitHub `https://github.com/Sadia12345/Thesis-BMI-Prediction-Snakemake`; archived release `https://doi.org/10.5281/zenodo.20569828`.

4 Results

4.1 Study Population and the BMI Target

Before modelling, the BMI target was described across the full analysis cohort of 17,903 samples. BMI ranged from 10.0 to 76.7 $\mathrm{kg/m^2}$ (mean 24.59, SD 6.21, median 23.7). Figure 4.1 shows the distribution: it is unimodal and right-skewed, concentrated in the normal-to-overweight range (a clear mode near 23–25 $\mathrm{kg/m^2}$) with long tails reaching implausibly low and high values. The pooled Metalog cohort combines 90 contributing studies from 39 countries (Figure 4.2); the most represented geographic origins are the USA (6,960 samples), China (3,272), the Netherlands (1,088), Germany (938), and Japan (716), and the most represented studies include an elderly cohort, an infant cystic-fibrosis cohort, and the Human Microbiome Project. This breadth of contributing studies and countries is the direct reason the BMI target spans such a wide range, as the following section examines. Most samples are from adults (age category: adult 16,151; baby 1,290; child 286; adolescent 163), with recorded ages spanning 0.1 to 107 years (median 47).

4.1.1 The Extreme BMI Values: Age and Clinical Cohort, Not Error

The tails of the BMI distribution were investigated directly, because values near 10 or above 50 $\mathrm{kg/m^2}$ could in principle indicate data-entry error. Three lines of evidence show that the extremes are genuine rather than artefactual:

- **Internal consistency.** For the 7,248 records that also carried weight and height, BMI recomputed from those fields matched the recorded BMI almost

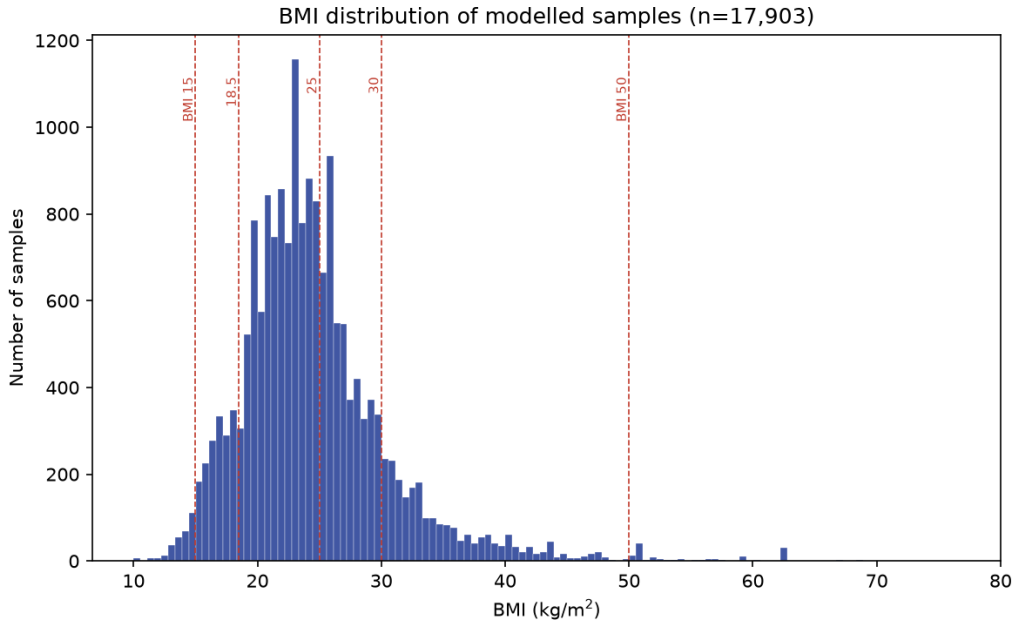

Figure 4.1: Distribution of BMI across the 17,903 modelled samples. Dashed lines mark conventional reference points (BMI 15, 18.5, 25, 30, and 50 $\mathrm{kg/m^2}$). The distribution is concentrated in the normal-to-overweight range with a long right tail; a small number of samples fall in implausible-looking tails below 15 and above 50 $\mathrm{kg/m^2}$, which are characterised in Section 4.1.1.

exactly (median absolute difference 0.003 $\mathrm{kg/m^2}$; Pearson $r = 0.99$; 96% within one BMI unit), including within the extreme tails. The extreme values are therefore internally consistent and not transcription errors.

- **The low tail is infants and clinical cohorts.** Of the 291 samples with BMI < 15, 82% are younger than 18 years and 61% younger than 5; they come predominantly from an infant cystic-fibrosis cohort (*Hayden 2020*, 140 samples), a paediatric Zimbabwean cohort (*Osakunor 2020*, 36), and an anorexia-nervosa cohort (*Fan 2023*, 25). A low BMI is physiologically expected in infants and in these conditions.

- **The high tail is severe-obesity and metabolic cohorts.** All 149 samples with BMI > 50 are adults (median age 68 years), drawn mainly from a bariatric-surgery cohort (*Penney 2022*, 44 samples), an elderly cohort (*Larson*

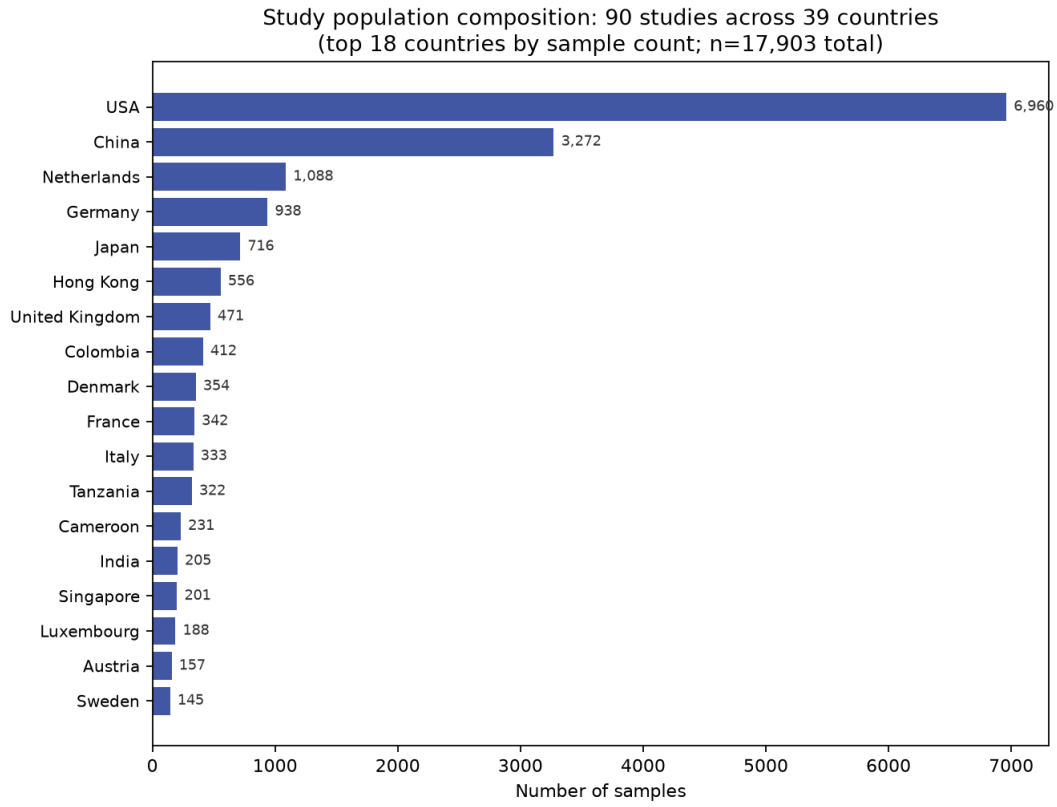

Figure 4.2: Composition of the analysis cohort: 17,903 samples pooled from 90 studies across 39 countries (top 18 countries by sample count shown). Because the dataset aggregates many independent studies of very different populations — from infant cohorts to elderly and clinical cohorts — the BMI target spans a correspondingly wide range.

2022, 59), and NAFLD / type-2-diabetes cohorts. Severe obesity is expected in these groups.

Figure 4.3 makes the age structure explicit: the low-BMI points form a distinct cluster at ages near zero, confirming that the implausible-looking low tail is driven by infants rather than by adult measurement error. The extreme BMI values are therefore attributable to participant age and clinical-cohort composition — an expected consequence of pooling many heterogeneous studies — rather than to data corruption.

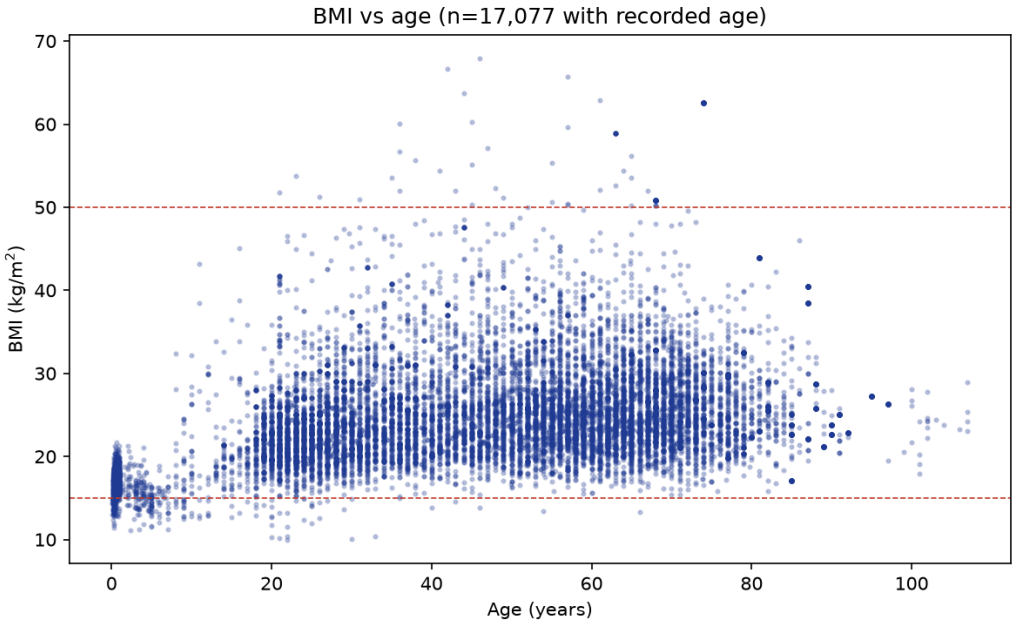

Figure 4.3: BMI versus age for the 17,077 modelled samples with a recorded age. Dashed lines mark BMI 15 and 50 kg/m^2. The cluster of very low BMI values at ages near zero shows that the low tail of the BMI distribution corresponds to infants and young children, not to adult measurement error.

Two further descriptive views reinforce this conclusion. Figure 4.4 summarises BMI within each recorded age category: median BMI rises from infants and children towards adults, the sub-15 values are confined almost entirely to babies and children, and only the adult category extends above 50 kg/m^2. Figure 4.5 breaks the distribution down by the fourteen largest contributing studies: the implausible-looking extremes are not scattered uniformly, as random data-entry error would be, but cluster within specific cohorts — the infant cystic-fibrosis cohort (*Hayden 2020*) occupies the low end, whereas the elderly and metabolic cohorts (for example *Larson 2022*) carry the high outliers. Both views are consistent with the interpretation that the BMI extremes are genuine age- and cohort-related variation rather than measurement error.

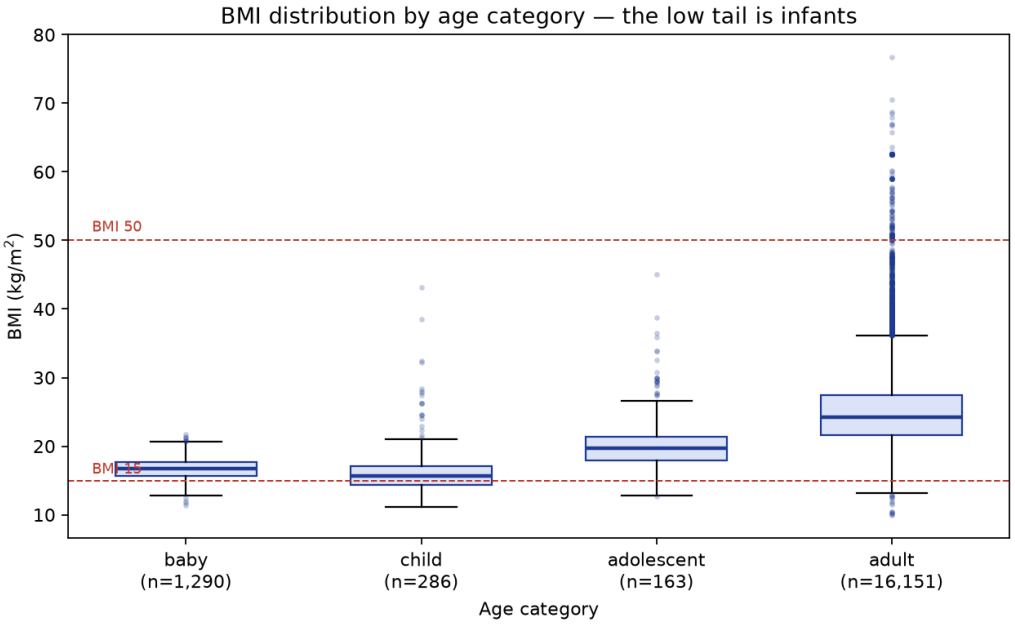

Figure 4.4: BMI distribution within each recorded age category (baby, child, adolescent, adult). Boxes show the interquartile range and median; dashed lines mark BMI 15 and 50 $\mathrm{kg/m^2}$. Median BMI increases with age category, the values below 15 $\mathrm{kg/m^2}$ are dominated by babies and children, and only the adult category reaches the severe-obesity range above 50 $\mathrm{kg/m^2}$.

4.1.2 Adult-Only Sensitivity Check

Because the full cohort mixes infants, children, and adults, an adult-only subset (age ≥ 18) was examined to confirm that this age mixing does not distort the main analysis. Restricting to adults retains 15,368 of 17,903 samples (86%) and removes most of the implausible low tail: samples with BMI < 15 fall from 291 to 35, while the clinically genuine high tail is largely retained (BMI > 50: 104 samples). The adult-only BMI distribution is otherwise close to the full-cohort distribution (median 24.0 versus 23.7 $\mathrm{kg/m^2}$). The full cohort was therefore retained for the primary models, with the adult-only subset used as a sensitivity reference; its effect on the per-species BMI associations is reported in Section 4.5.

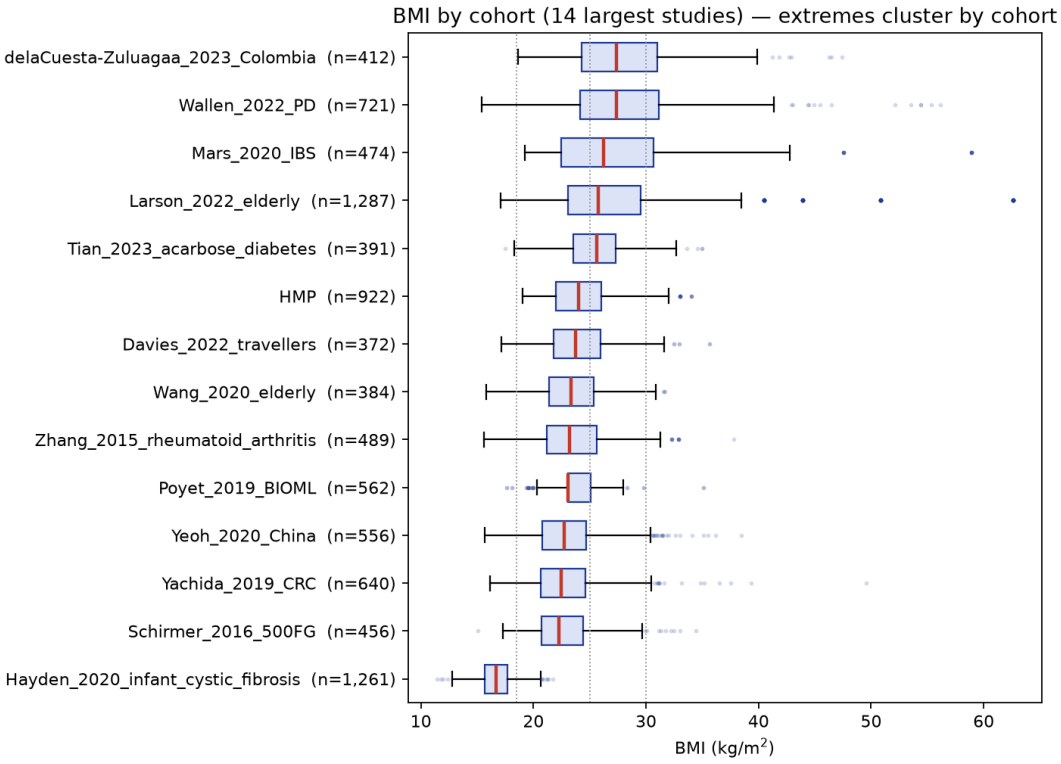

Figure 4.5: BMI distribution for the fourteen largest contributing studies, ordered by median BMI. Dotted lines mark the conventional reference points 18.5, 25, and 30 $\mathrm{kg/m^2}$. The extreme BMI values cluster within specific cohorts rather than appearing as uniform noise: the infant cystic-fibrosis cohort (*Hayden 2020*) sits at the low end, while the elderly cohort (*Larson 2022*) carries the highest outliers. This cohort structure is consistent with the extremes being genuine rather than data-entry error.

4.2 Model Benchmarking: The Non-Linear Advantage

The model benchmark compared linear and non-linear approaches for BMI regression from species-level gut microbiome profiles. Table 4.1 summarizes the main performance results.

The benchmark should be read as a comparison of practical modeling strategies under the implemented Nextflow pipeline, not as a claim that all possible versions of each algorithm were exhaustively optimized. The models were evaluated under the preprocessing and cross-validation settings described in Chapter 3. This framing

is important because the purpose of the chapter is to report observed performance patterns, while broader biological interpretation is reserved for the Discussion.

Table 4.1: Model Performance Comparison (Peak Performance at N=16,000). Random Forest, Decision Tree (Rpart), and Linear SVM values are read from the N=16,000 step of the Nextflow saturation analysis. †Elastic Net (GLMNet) was not included in the saturation ladder and was instead evaluated as a separate full-preprocessed-dataset run (seed=100); its R^2 and RMSE are therefore reported from the dedicated GLMNet output rather than from the N=16,000 saturation row.

Model	R^2	RMSE	Model class
Random Forest	0.387	5.06	Non-linear ensemble
Elastic Net (GLMNet)†	0.192	5.83	Linear regularized
Decision Tree (Rpart)	0.10	6.02	Simple non-linear
Linear SVM	0.04	7.47	Linear baseline

Random Forest produced the strongest predictive performance in the main benchmark, with $R^2 = 0.387$ and RMSE = 5.06. Elastic Net was the strongest linear baseline but reached only $R^2 = 0.192$. The single decision tree performed below Random Forest, indicating that the ensemble structure contributed meaningfully beyond a single tree. Linear SVM performed weakest among the reported models.

The RMSE values provide a complementary view of the same pattern. Random Forest had the lowest RMSE, meaning that its predicted BMI values were closest to the observed values among the compared models. Elastic Net and Rpart had intermediate RMSE values, while Linear SVM had the largest error. The gap between the single decision tree and Random Forest is also informative: both use recursive partitioning, but the ensemble substantially outperformed the single tree, confirming the value of bagging and feature subsampling for this task.

Figure 4.6 isolates the clearest two-model contrast within the broader benchmark summarized in Table 4.1. It is included as a visual comparison between the strongest retained non-linear model and the strongest retained linear baseline, rather than as a full four-model uncertainty plot.

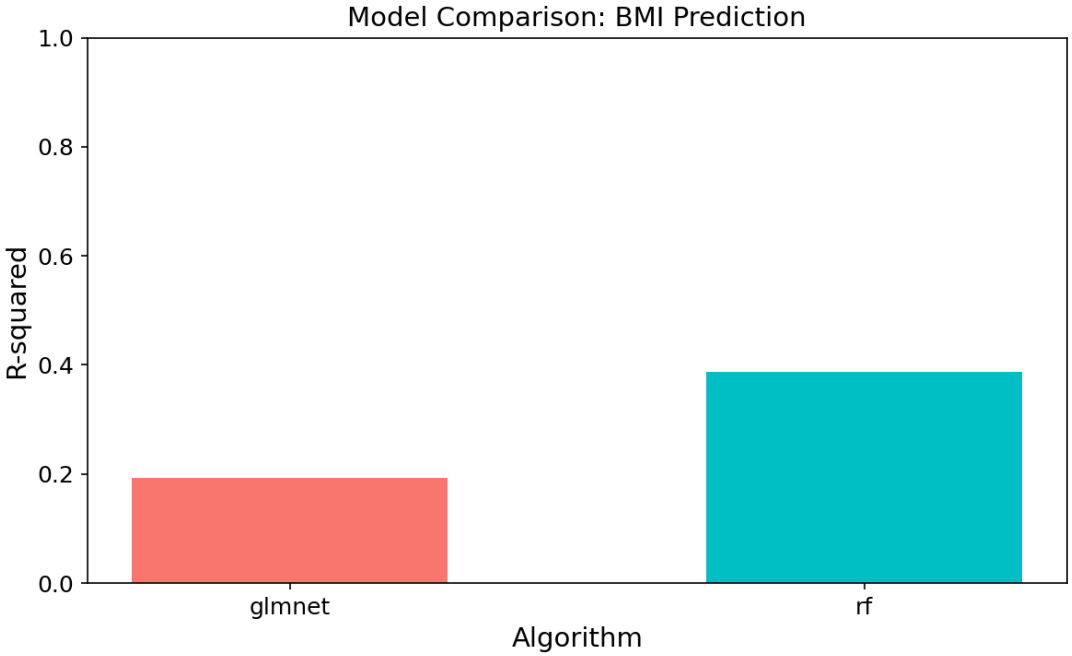

Figure 4.6: Model comparison at N = 16,000 under the retained Nextflow benchmark settings (80/20 training/test split: training set N = 12,800; test set N = 3,200). The figure shows Random Forest and Elastic Net as representative models from the broader four-model comparison reported in Table 4.1. Random Forest achieved $R^2 = 0.387$, compared with $R^2 = 0.192$ for Elastic Net. Values are point estimates only (one subset draw per sample-size step, `cv_times = 1`).

These results show a clear performance gap between Random Forest and the linear baselines. The result is consistent with the hypothesis that the microbiome-BMI relationship contains non-linear or interaction-rich predictive structure. It should not be interpreted as proof of a specific ecological mechanism. The particularly weak Linear SVM performance ($R^2 = 0.04$) may partly reflect the use of default caret hyperparameters without grid search, as linear SVMs can be sensitive to the regularisation parameter C at this data scale; a more comprehensive tuning sweep could recover stronger linear SVM performance.

Figure 4.7 shows a scatterplot of Random Forest predicted versus actual BMI values on the held-out test set, providing a direct visual representation of model accuracy at the individual level.

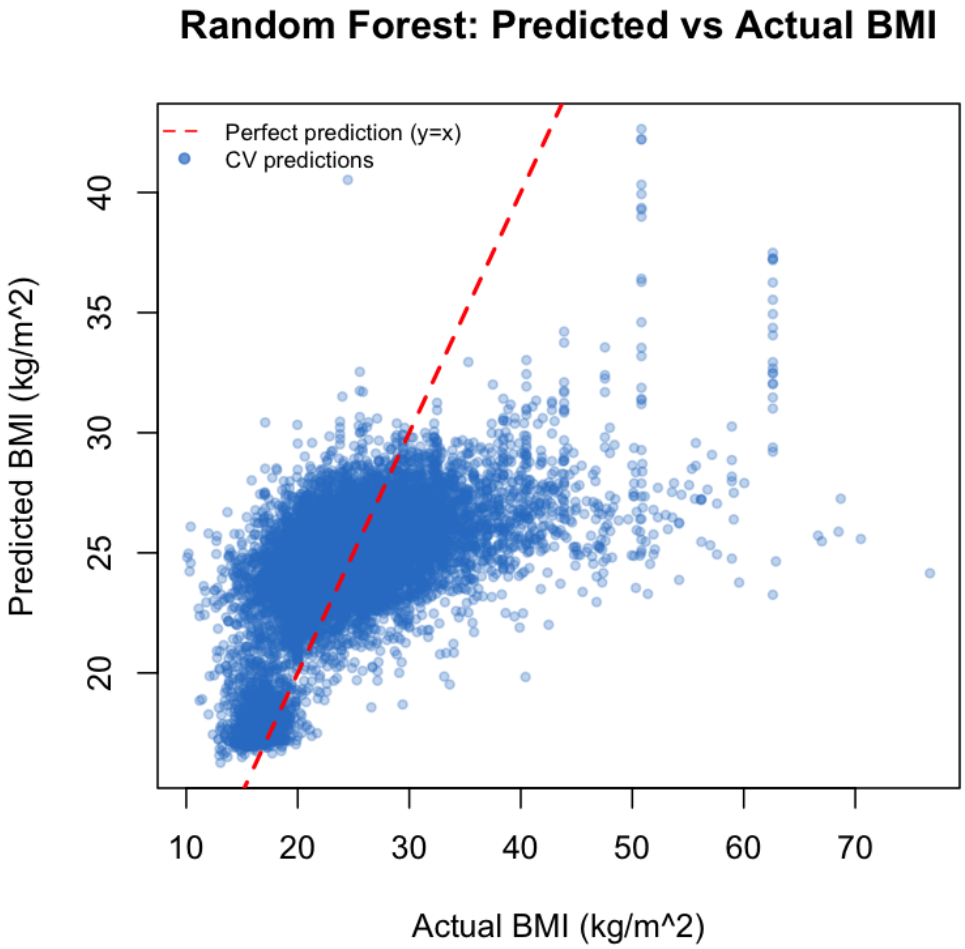

Figure 4.7: Predicted versus actual BMI values for the final saved Random Forest model trained on the full preprocessed dataset. Predictions were obtained using 3-fold cross-validation on the training partition of the full dataset (N = 17,903; training set N = 14,322; held-out test set N = 3,581; best `mtry` = 16; seed = 100). The dashed line indicates perfect prediction ($y = x$). The model captures the central BMI range more closely than the extreme BMI values.

The full-dataset Random Forest model achieved held-out $R^2 = 0.351$ and test-set RMSE = 5.37 kg/m^2, with cross-validation RMSE = 5.10 kg/m^2. This is lower than the saturation-peak result of $R^2 = 0.387$ at N = 16,000 reported in Table 4.1, likely

because the full cohort introduced additional heterogeneity and was evaluated using a different random seed.

Beyond these summary metrics, Figure 4.8 shows the full distribution of the cross-validated prediction errors (predicted minus actual BMI) underlying Figure 4.7. The errors are centred near zero with a mean absolute error of 3.45 kg/m^2: approximately 79% of samples are predicted within 5 BMI units and 42% within 2. The distribution is mildly left-skewed, reflecting the under-prediction of the highest-BMI individuals — a direct consequence of the regression-toward-the-mean behaviour expected when the predictors carry only a partial, distributed signal.

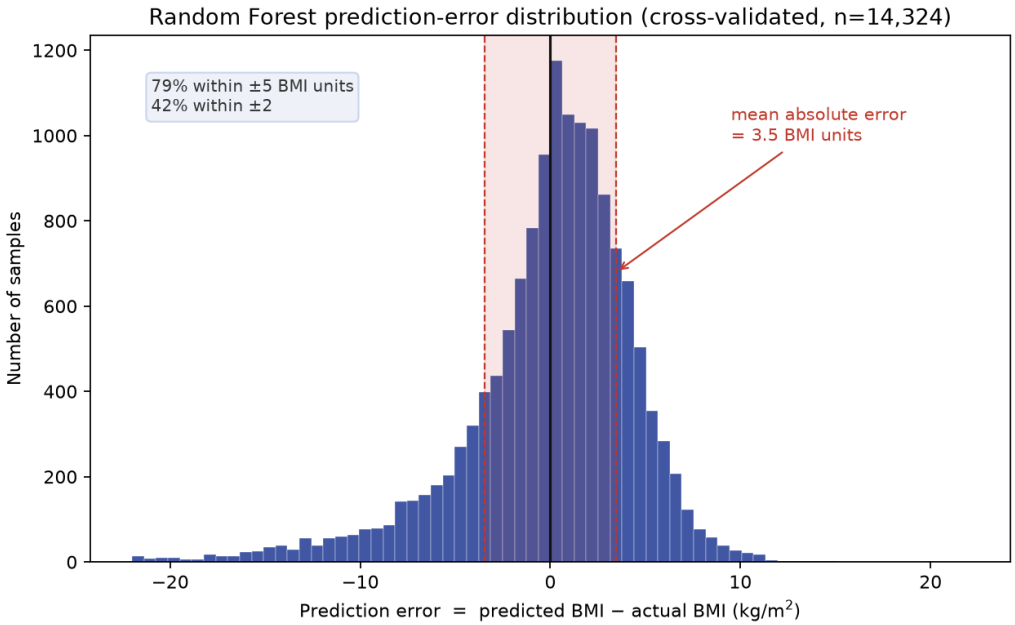

Figure 4.8: Distribution of the cross-validated Random Forest prediction errors (predicted minus actual BMI) for the 14,324 training-partition samples underlying Figure 4.7 (seed = 100, best `mtry` = 16). The shaded band marks the mean absolute error ($\pm 3.45\,kg/m^2$); about 79% of samples fall within 5 BMI units and 42% within 2. The mild left skew corresponds to the systematic under-prediction of the highest-BMI individuals.

The model comparison also establishes the practical baseline used for the rest of the Results chapter. Because Random Forest was the best-performing model in the main comparison, the filtering and scalability results focus on whether this stronger

model can be made computationally feasible without losing most of its predictive performance.

Summary. Across the benchmarked models, Random Forest provided the strongest predictive performance for BMI, while linear models captured substantially less of the observed signal.

4.3 Saturation Analysis: The Law of Diminishing Returns

The saturation analysis evaluated whether predictive performance continued to improve as more samples were added. This question matters because metagenomic datasets are expensive to process and because larger cohorts do not necessarily improve prediction indefinitely.

4.3.1 The Saturation Threshold

The Nextflow learning curve showed rapid improvement at smaller sample sizes followed by a flatter region at larger sample sizes (Figure 4.9). Random Forest performance improved steadily at smaller sample sizes, showed a flattened region between approximately 12,000 and 14,000 samples, and then recorded a further small improvement at N=16,000. The overall trajectory is therefore best described as a region of diminishing — but not exhausted — returns within the tested range. Increasing the sample size to 16,000 produced only limited additional improvement. This pattern should be interpreted specifically as the output of the Nextflow scaling workflow, which combined one subset draw per sample-size step with 3-fold cross-validation, rather than as the output of a repeated-seed design. The two summary metrics also do not peak at exactly the same sample size: Random Forest RMSE was lowest near N=14,000 (RMSE = 4.84), whereas R^2 was highest at N=16,000

($R^2 = 0.387$, RMSE = 5.06). Because each step used a single subset draw (`cv_times = 1`), R^2 is sensitive to the BMI variance of the particular draw; the N=16,000 R^2 peak should therefore be read together with this single-draw caveat rather than as strictly monotonic improvement, which reinforces the diminishing-returns interpretation above.

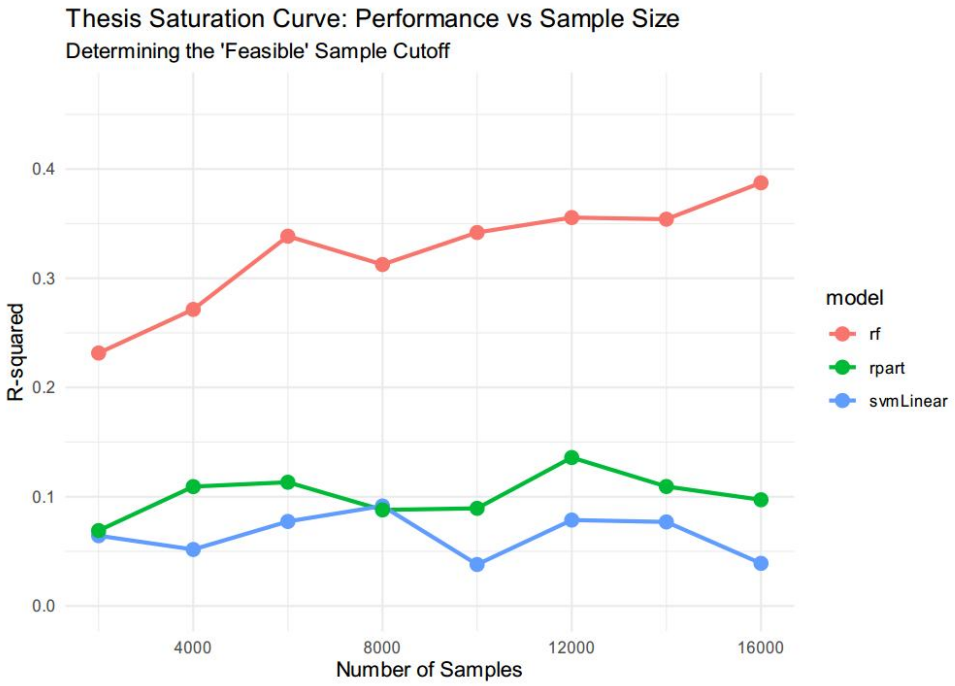

Figure 4.9: Nextflow workflow saturation figure showing R^2 as a function of sample size for Random Forest, Decision Tree (Rpart), and Linear SVM. Each point reflects the Nextflow scaling design, in which one random subset was drawn at each sample-size step and evaluated with 3-fold cross-validation (`kfold = 3`). At each step, an 80/20 training/test split was applied (e.g., training: N=12,800, test: N=3,200 at the N=16,000 step). Point estimates only; no error bars are shown because one subset draw was used per step (`cv_times = 1`). The figure shows Random Forest outperforming the other models and exhibiting diminishing gains after roughly 12,000 samples.

This pattern suggests that much of the recoverable predictive information available from species-level taxonomic abundance profiles is largely captured by approximately 12,000–14,000 samples in this dataset. This does not mean that larger studies are unnecessary. Rather, it suggests that simply adding more taxonomic profiles may

yield diminishing returns unless additional information, such as diet, metabolomics, medication, geography, or longitudinal measurements, is incorporated.

An additional evaluation at the full dataset size (N=17,903) confirmed the plateau trend, yielding $R^2 = 0.351$ — slightly lower than the N=16,000 peak value of 0.387, consistent with added heterogeneity at the full cohort extent rather than a continuing improvement. This result supports the interpretation that the saturation region has been reached within the tested range.

The saturation finding is therefore a study-design result. It indicates that, under these workflows and with the available feature representation, the largest gains occurred before the final sample-size steps. The result does not imply that 12,000–14,000 samples are sufficient for every microbiome phenotype or every validation design. It is specific to BMI prediction from species-level relative abundances in this Metalog-derived analysis.

Summary. The Nextflow saturation analysis indicates a practical region of diminishing returns around 12,000–14,000 samples for BMI prediction from the available taxonomic profiles, confirmed by the full-cohort (N=17,903) evaluation.

4.3.2 Non-Monotonicity and Cohort Heterogeneity

The Nextflow learning curve is not strictly monotonic: as noted above, RMSE is lowest near N=14,000 while R^2 peaks at N=16,000, and the full-cohort evaluation sits slightly below the N=16,000 peak. One contributing factor is increasing cohort heterogeneity as the sample broadens. A smaller subset may, by chance, contain a more homogeneous group of individuals, making prediction temporarily easier, whereas larger subsets introduce additional geographic, dietary, technical, and cohort-level diversity. Because the learning curve was not explicitly stratified by cohort, geography, or sequencing batch, this remains a hypothesis rather than a direct finding, but it is consistent with the descriptive cohort structure reported in Section 4.1.

Summary. The minor non-monotonicity is best treated as a hypothesis-generating pattern attributable to cohort heterogeneity, while the overall trajectory still supports a region of diminishing returns.

4.4 Scalability Optimization: The Efficacy of Filtering

Feature filtering was evaluated because the full species matrix was difficult to train on local hardware. Table 4.2 summarizes the observed effect of progressively stronger feature reduction in the Nextflow workflow, which is the workflow in which the variance-based top-500 filter was implemented.

Table 4.2: Impact of prevalence and variance-based feature filtering on Random Forest feasibility and performance in the Nextflow workflow. Runtime refers to the total saturation pipeline execution (all sample-size steps combined). All configurations used an 80/20 training/test split. †Run did not complete: memory pressure on the 8 GB MacBook caused failure before the saturation analysis could finish, so no Random Forest R^2 is reported for this configuration. ‡Approximate extrapolation from the confirmed top-500 runtime; runtime does not scale strictly linearly with feature count, so this figure is indicative only. No separate Random Forest evaluation was archived for this intermediate filtering configuration because the preprocessing pipeline applies prevalence and variance filtering in a single sequential step ("n/a" in the table). *Verified from Nextflow trace file: preprocessing took 38.6 s and the full saturation analysis (all sample-size steps) took 9 h 15 min. §Reported only for configurations that produced an archived Random Forest run. The top-500 value shown ($R^2 = 0.387$) is the saturation peak at N=16,000 from `saturation_results.csv`; the same top-500 configuration on the full dataset (N=17,903, seed=100) yielded $R^2 = 0.351$ (`rf_results.csv`), and the mean across the eight saturation steps is $R^2 = 0.324$.

Configuration	Feature count	RF R^2§	Total pipeline runtime
No filter	10,701	—	Did not complete†
1% prevalence	1,799	n/a	∼16 hours‡
Variance-based top-500 filter	500	0.387	9 h 15 min*

The 1% prevalence filter reduced the feature space sharply, and the variance-based top-500 filter reduced it further so that model training remained feasible on the memory-limited MacBook environment (Figure 4.10). Because the preprocessing pipeline applies these two filters in a single sequential step, only the top-500 configuration produced an archived Random Forest evaluation; the unfiltered run did

not complete on the available hardware, and no separate Random Forest result was stored for the intermediate 1,799-feature state.

Expressed as a feature-count reduction, the 1% prevalence filter retained approximately 17% of the original Nextflow feature space (1,799 of 10,701 features). The top-500 configuration retained less than 5% of the original feature count. Despite this aggressive reduction, the top-500 Random Forest reached its saturation peak of $R^2 = 0.387$ at N=16,000 and $R^2 = 0.351$ on the full dataset (Figure 4.9). The observed trade-off therefore favored filtering for this workflow: the variance-based top-500 configuration completed the full saturation analysis in 9 h 15 min (verified from the Nextflow trace), whereas the unfiltered run could not be completed on the available hardware and a rough extrapolation suggested that on the order of 48 hours or more would have been required.

The result should not be read as evidence that rare taxa are biologically irrelevant. Instead, it shows that for this BMI regression benchmark, the features most useful for prediction were concentrated among prevalent and highly variable species. This is a practical finding about the modeling task and hardware constraint.

It is also worth noting the computational cost of individual models. The Nextflow trace data indicate that Linear SVM at N=16,000 required approximately 15,132 seconds (roughly 4.2 hours) for a single model fit. This exceptionally long SVM runtime — far exceeding Random Forest at the same scale — reflects the well-known quadratic scaling of SVM with sample size. It provides additional practical justification for preferring ensemble tree-based methods over kernel machines in large-scale microbiome benchmarks run under local hardware constraints.

Summary. Aggressive feature reduction preserved most of the observed predictive performance while substantially reducing runtime and memory burden.

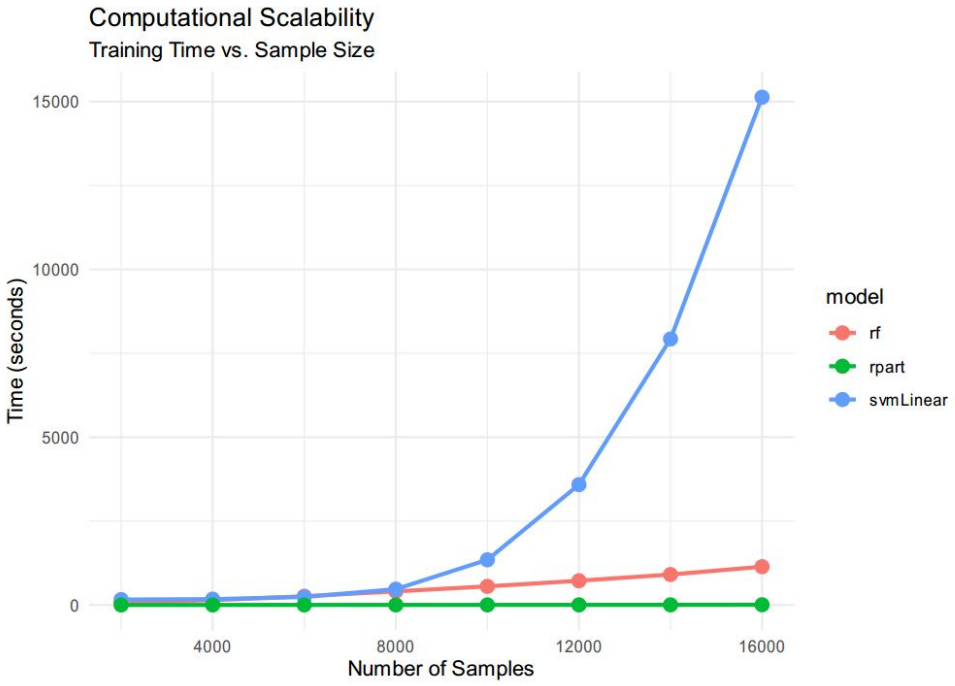

Figure 4.10: Nextflow workflow computational-scalability figure showing training time as a function of sample size for the benchmarked models under the Nextflow scaling design. Each point comes from the same one-subset-per-step, 3-fold cross-validation runs used for the Nextflow saturation analysis. Training times reflect model fitting on the training partition (80% of each sample-size step; e.g. at N=16,000 the training set was N=12,800 and the test set was N=3,200). The figure reflects the memory-limited MacBook analysis and illustrates why prevalence filtering and the variance-based top-500 filter were important for practical execution on local hardware. The unfiltered runtime estimate (∼48 hours) is based on partial execution; memory constraints on the 8 GB MacBook prevented stable completion of that configuration.

4.5 Feature Importance Analysis

Feature-ranking analysis identified biologically interpretable taxa and unclassified genome bins in the Random Forest output. The main Nextflow Random Forest model, sorted by descending IncNodePurity, placed an uncharacterised species-level genome bin (*s__GGB9512_SGB14909*) first, followed by oral and opportunistic species (*Enterococcus faecalis*, *Cutibacterium acnes*, *Rothia mucilaginosa*, several *Streptococcus* and *Veillonella* species), *Escherichia coli*, *Bifidobacterium longum*, and *Ruminococcus gnavus*, together with additional unclassified genome bins.

The feature-importance figure was used to identify candidate signals rather than to estimate causal effects. This is especially important because correlated microbial features can share information. If two taxa co-occur, the model may assign importance unevenly between them even when both are part of the same broader community pattern. The most reliable descriptive result is therefore the persistence of interpretable microbial structure, not the exact rank of a single feature.

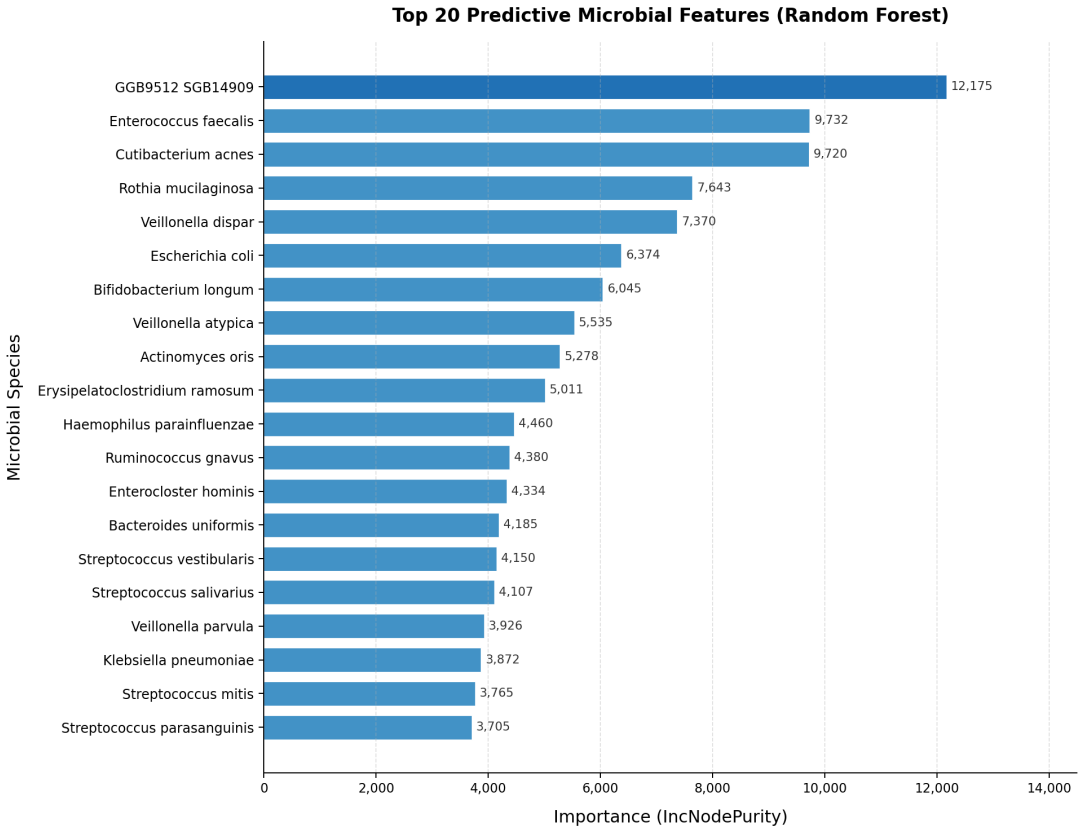

Figure 4.11: Nextflow workflow feature-importance figure derived from the saved `rf_model.rds`, trained by `train_ml.R` on the full preprocessed dataset (seed = 100; training set: N=14,322; held-out test set: N=3,581 [approximately 20% of the full cohort N=17,903]; $R^2 = 0.351$). Importance scores are IncNodePurity from the underlying `randomForest` final model. The figure shows the global top-20 features by IncNodePurity. The ranking is dominated by an uncharacterised species-level genome bin (*s__GGB9512_SGB14909*), oral and opportunistic species (*Enterococcus faecalis*, *Cutibacterium acnes*, *Rothia mucilaginosa*, several *Streptococcus* and *Veillonella* species), *Escherichia coli*, and *Bifidobacterium longum*. The figure is descriptive rather than inferential and these taxa should be interpreted as nominated predictors rather than confirmed biomarkers.

The named taxa are useful for interpretation because they can be connected to existing biological literature. However, several unclassified SGBs and GGBs also appeared among important features, indicating that under-characterized microbial diversity may contribute to prediction. This pattern is consistent with one rationale for using MetaPhlAn 4, namely its improved profiling of previously uncharacterized species-level genome bins [6].

4.5.1 Direction of Association: Bacterial Abundance versus BMI

Feature-importance scores (Figure 4.11) are undirected: they indicate that a species helped the model split the data, but not whether higher abundance is associated with higher or lower BMI. To make the direction explicit, the relative abundance of the eight highest-ranked species from the importance figure (its top seven by IncNodePurity plus *Ruminococcus gnavus*) was plotted against BMI (Figure 4.12), with a fitted trend line and Spearman rank correlation. Spearman correlation is reported because the abundance distributions are sparse and zero-inflated.

Two observations follow. First, the associations are uniformly weak: all fall at or below $|\rho| \approx 0.22$. The strongest is the top-ranked uncharacterised genome bin *GGB9512_SGB14909*, which is weakly *positive* ($\rho = +0.22$); the strongest negative associations are *Enterococcus faecalis* ($\rho = -0.19$) and *Veillonella dispar* ($\rho = -0.19$), followed by *Escherichia coli* (-0.13), *Rothia mucilaginosa* (-0.09), *Ruminococcus gnavus* (-0.07), and *Bifidobacterium longum* (-0.06); *Cutibacterium acnes* is weakly positive ($+0.09$). This pattern is consistent with — and helps explain — the model-benchmarking result: because no individual species carries more than a weak signal, additive linear models capture little, whereas the Random Forest can combine many weak, distributed signals through interactions, which is why it substantially outperforms the linear baselines.

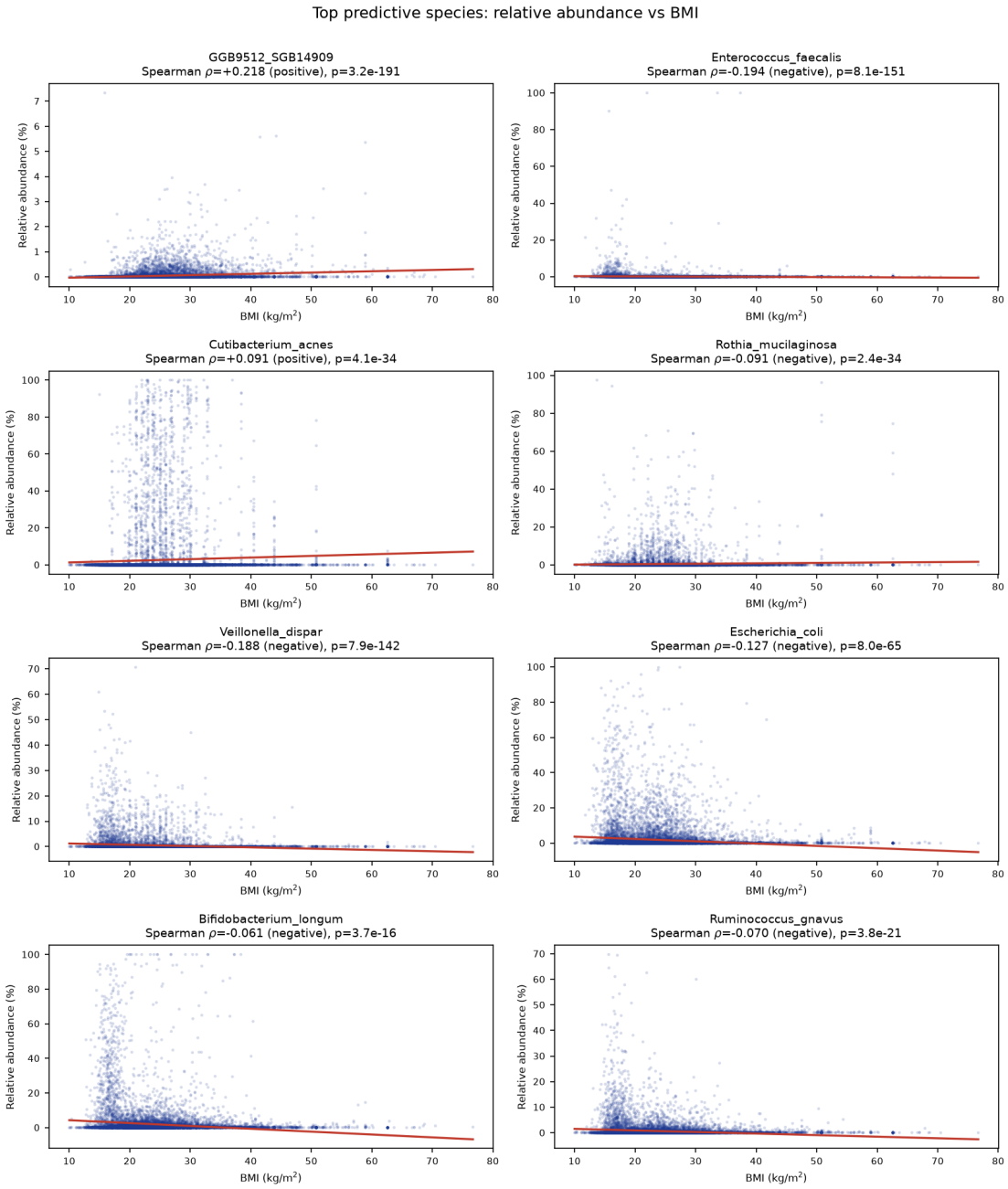

Figure 4.12: Relative abundance (%) of the eight highest-ranked predictive species (the same taxa as the IncNodePurity figure, Figure 4.11) plotted against BMI across the 17,903 modelled samples, with a fitted linear trend (red) and the Spearman rank correlation ρ annotated per panel. *Bifidobacterium longum*, named during the thesis examination, is included. The distributions are sparse and zero-inflated, and all associations are weak ($|\rho| \leq 0.22$): the top-ranked uncharacterised bin *GGB9512_SGB14909* is the strongest and is positive ($\rho = +0.22$), while most named species are weakly negative. These are descriptive, candidate associations, not biomarkers and not evidence of causation.

Second, several of the negatively associated species (*Veillonella*, *Enterococcus*, *Escherichia*) are oral- or infant-associated taxa, and the lowest-BMI samples are infants (Section 4.1.1). To test whether these associations reflect adult BMI biology or the infant subpopulation, the same correlations were recomputed on the adult-only subset (age $\geq$ 18; Section 4.1.2). The negative associations largely collapse toward zero in adults (*Enterococcus faecalis* $-0.19 \rightarrow +0.02$; *Veillonella dispar* $-0.19 \rightarrow -0.04$; *Escherichia coli* $-0.13 \rightarrow -0.02$; *Bifidobacterium longum* $-0.06 \rightarrow +0.00$; *Ruminococcus gnavus* $-0.07 \rightarrow +0.00$), confirming that much of the apparent species–BMI signal is carried by the age structure of the pooled dataset rather than by a direct adult microbiome–BMI relationship. In contrast, the top-ranked *GGB9512_SGB14909* association persists in adults ($+0.22 \rightarrow +0.18$), making it the most robust single candidate and giving a concrete (if still weak) direction to the otherwise uncharacterised feature that the model ranked first. This reinforces the interpretation of the named taxa as undirected, distributed predictors rather than individual biomarkers, while flagging the leading uncharacterised bin for follow-up. The biological context of the top taxa is discussed in Chapter 5.

Summary. The top predictive species show only weak associations with BMI ($|\rho| \leq 0.22$); most named taxa are weakly negative and largely reflect the infant subpopulation (they vanish in adults), whereas the top-ranked uncharacterised bin *GGB9512_SGB14909* is weakly positive and persists in adults — consistent with a non-linear model aggregating many weak, distributed signals rather than any single biomarker.

5 Discussion

5.1 Interpreting the Non-Linear Performance Gap

The clearest computational result of the thesis is the gap between Random Forest and the tested linear baselines. Random Forest achieved the strongest reported predictive performance, whereas Elastic Net and linear SVM were substantially weaker. This pattern is consistent with the presence of non-linear or interaction-rich predictive structure in the microbiome-BMI relationship.

The result should still be interpreted cautiously. Better performance by Random Forest does not prove ecological bistability, does not identify tipping points, and does not establish a causal mechanism. It shows that, in this dataset and under these preprocessing choices, a flexible tree-based ensemble recovered more predictive information than the tested linear baselines. This is compatible with the tipping-elements framework of Lahti et al. [7], but it remains an indirect computational observation rather than a mechanistic demonstration.

This distinction matters because overclaiming is easy in microbiome machine learning. A model can use signals produced by diet, geography, medication, sequencing batch, or cohort structure. If those signals correlate with BMI, they may improve prediction without representing direct microbial effects on body mass. Therefore, the most defensible interpretation is that the observed performance gap supports the pragmatic use of non-linear models for microbiome-based BMI prediction, while motivating further validation rather than closing the biological question.

The linear baselines are useful precisely because they define this contrast. Elastic Net and linear SVM were not included merely as weaker alternatives. They test

whether the signal can be approximated adequately by additive linear effects. Their weaker performance suggests that the predictive structure in the available species-abundance matrix is not summarized well by a simple linear representation. That is a modeling conclusion, not a mechanistic one, but it is still important for future workflow design.

There is also an interpretability trade-off embedded in this result. Random Forest is more flexible and therefore more capable of recovering complex predictive structure, but it does not provide the same direct coefficient-based interpretation as a penalized linear model. In that sense, the thesis illustrates a common tension in microbiome machine learning: the models that predict better are not always the models that are easiest to explain biologically. This is one reason why the discussion emphasizes caution when moving from predictive performance to ecological or clinical interpretation.

It is also worth being explicit about what the reported accuracy means in practical terms. An R^2 of 0.387 is not high enough to support stand-alone clinical prediction of an individual's metabolic health from microbiome composition alone. At the same time, it is not trivial, especially because the benchmark intentionally excluded age, sex, weight, and other non-microbial covariates that would ordinarily strengthen a predictive model. The result is therefore best read as evidence that species-level gut metagenomes contain measurable but incomplete information about BMI. That is a meaningful computational finding without being a claim of clinical readiness.

5.2 Biological Plausibility of the Top Predictors

The main Nextflow Random Forest feature-importance output, when sorted by descending IncNodePurity, places a diverse mixture of taxa at the top: an uncharacterised species-level genome bin (*s__GGB9512_SGB14909*) ranks first, followed by several oral and opportunistic species, gut commensals with established

metabolic roles, and additional unclassified genome bins. As shown in the direction-of-association analysis (Section 4.5.1), these top predictors individually carry only weak correlations with BMI ($|\rho| \leq 0.22$): most of the named species are weakly negative and largely disappear once infants are excluded, whereas the top-ranked uncharacterised bin is weakly positive and persists in adults. This reinforces the view that the predictive signal is distributed across many weak features rather than concentrated in any single biomarker. Throughout, the thesis remains predictive rather than mechanistic: the literature can help explain why particular taxa appear in a BMI model, but it cannot convert a feature-importance ranking into causal evidence.

5.2.1 Oral and Opportunistic Species Among the Top Predictors

Several taxa near the top of the IncNodePurity ranking are commonly associated with oral or opportunistic niches: *Enterococcus faecalis* (rank 2), *Cutibacterium acnes* (rank 3), *Rothia mucilaginosa* (rank 4), *Actinomyces oris* (rank 9), *Haemophilus parainfluenzae* (rank 11), and multiple *Streptococcus* and *Veillonella* species occupying ranks 5, 8, 15–17, and 19–20. In the gut microbiome literature, oral taxa detected in stool are frequently interpreted as markers of oral-gut translocation, which can accompany disrupted mucosal barriers, antibiotic exposure, or broader community dysbiosis. Their presence as high-ranked predictors in a BMI model is therefore plausible as an indirect marker of host physiological state rather than as a direct driver of adiposity.

It is important to emphasise, however, that the present study did not measure oral microbiome composition, antibiotic history, gut permeability, inflammatory markers, or any other variable that would confirm translocation as the mechanism. Moreover, oral taxa detected in stool are subject to analytical ambiguity: MetaPhlAn 4 may

assign reads from partially characterised species to oral-associated clades even when the organism is a distinct gut-resident lineage. The most defensible interpretation is therefore that the elevated IncNodePurity scores of these taxa indicate that their abundance profiles correlate with BMI variation in the training data, and that this correlation merits follow-up rather than mechanistic extrapolation.

5.2.2 *Bifidobacterium longum* and *Ruminococcus gnavus*

Bifidobacterium longum (rank 7, IncNodePurity $\approx$ 6,045) is among the most extensively studied gut commensals. It is associated with dietary fibre fermentation, production of short-chain fatty acids, and modulation of host immune responses. Several population-level studies have linked *Bifidobacterium* abundance to metabolic markers including glucose tolerance and body weight, making its appearance near the top of a BMI prediction ranking biologically plausible. In the present data its direct association with BMI is weak and slightly negative (Spearman $\rho = -0.06$, falling to near zero in adults; Section 4.5.1), broadly in line with reports of lower *Bifidobacterium* abundance in obesity — though the effect here is small, and the feature-importance ranking is undirected rather than a causal claim.

Ruminococcus gnavus (rank 12, IncNodePurity $\approx$ 4,380) has been discussed in the context of both inflammatory bowel disease and metabolic phenotypes. Its mucus-degrading and immunomodulatory properties have attracted interest as potential mediators of gut-host crosstalk. Elevated *R. gnavus* abundance has been reported in association with obesity-related and inflammatory phenotypes in several cohorts, although the direction and consistency of these associations depend on population context.

Neither taxon was causally tested in this thesis. The study did not measure fermentation metabolites, dietary intake, inflammatory markers, or any mechanistic

variable. The findings should be read as predictive associations that motivate more focused biological inquiry.

5.2.3 Unclassified Species-Level Genome Bins

The single highest-ranked feature by IncNodePurity is itself an unclassified genome bin, *s__GGB9512_SGB14909* (≈ 12,175), a result that strengthens the general observation made throughout this thesis: a meaningful share of predictive signal in MetaPhlAn 4-profiled data comes from organisms that are not yet assigned conventional species names. These predictors are harder to discuss biologically, but they represent real and reproducible microbial variation. Their prominence is consistent with the use of MetaPhlAn 4, which substantially expands profiling of previously uncharacterised species-level genome bins [6]. Excluding such features because they resist easy narrative framing would clean the story at the cost of discarding genuine signal. The direction-of-association analysis (Section 4.5.1) adds weight to this: of all the top-ranked taxa, *s__GGB9512_SGB14909* shows both the strongest correlation with BMI (Spearman $\rho = +0.22$, positive) and the only one that clearly persists when the analysis is restricted to adults ($\rho = +0.18$), whereas the named oral and infant-associated taxa largely reflect the infant subpopulation. The thesis therefore reports this bin descriptively, treating its high IncNodePurity and robust, adult-persistent positive association as the leading candidate for targeted follow-up characterisation rather than as an interpretable biological finding in its current form.

5.3 Scalability and Practical Feature Reduction

One of the most practical results of the thesis is the finding that aggressive dimensionality reduction can make large-scale microbiome modeling feasible on consumer-grade hardware. In the Nextflow workflow, a 1% prevalence filter followed by a variance-based top-500 filter reduced feature count and runtime dramatically while

preserving most of the observed predictive performance. This directly addresses the computational bottleneck described in the Introduction.

The practical interpretation is not that discarded taxa are meaningless. Rare taxa can be biologically important, especially in disease contexts, subgroup analyses, or mechanistic studies. Rather, for this particular BMI regression task, most of the model-usable signal appeared to be concentrated among prevalent and variable species. This suggests that aggressive filtering can be a defensible engineering choice when the goal is scalable baseline prediction under local hardware constraints.

The result is also relevant for accessibility. Large-scale metagenomic machine learning is often assumed to require high-performance computing infrastructure. This thesis shows that, with transparent filtering and workflow management, meaningful benchmark analyses can be run on ordinary laptops. That does not remove the need for careful validation, but it does make exploratory research more accessible to students and smaller research groups.

There is still a trade-off. Local feasibility required compromise in several parts of the workflow, including aggressive feature reduction and lighter cross-validation settings. These choices were reasonable for the aims of the project, but a future HPC-scale implementation could explore a broader hyperparameter space and fully nested preprocessing with lower practical cost. The value of the present result is therefore not that local analysis replaces large-scale infrastructure, but that it establishes what can be done reproducibly before that level of infrastructure is available.

5.4 What the Saturation Result Means

The saturation result is one of the most conceptually useful findings in the thesis because it links predictive performance to study design. The Nextflow saturation analysis suggests that most of the recoverable microbiome-only BMI signal in this

benchmark is captured by approximately 12,000–14,000 samples. Beyond that range, the gains become small relative to the computational effort required to add more data. This does not imply that additional samples are uninformative. It implies that, under the present phenotype definition and feature representation, sample size ceases to be the dominant bottleneck.

This point matters because it changes what "more data" should mean in future work. If the learning curve had continued to climb steeply, the obvious next step would be to prioritize larger cohorts of the same general type. A plateau suggests a different conclusion: additional progress may depend more on richer measurement than on simply increasing n. In practice, this means that metadata quality, external validation cohorts, diet information, metabolomics, longitudinal follow-up, or pathway-level representations might now offer greater marginal value than adding more similarly processed taxonomic profiles to the same benchmark.

The saturation finding should also be interpreted in light of cohort heterogeneity. A large pooled cohort is valuable because it exposes the model to more variability, but this same variability can reduce apparent gains if the phenotype is only weakly expressed in the available input representation. Therefore, the apparent plateau should not be read as evidence that the microbiome has little relationship to body weight. It is more accurately interpreted as a limit of the specific combination of BMI, species-level relative abundance, cross-sectional aggregation, and practical workflow constraints used here.

It should also be noted that the Nextflow learning curve does not show a complete performance plateau within the tested sample-size range. The rate of improvement slows after approximately 12,000 samples, but the curve does not fully flatten. The saturation interpretation is therefore tentative and should be understood as a region of diminishing returns rather than a hard ceiling. Testing beyond 16,000 samples or

with an independent cohort would be necessary to confirm whether a true plateau has been reached.

From a methodological standpoint, this result is still highly useful. It provides a planning signal for future computational studies: beyond a certain sample-size region, it may be more efficient to invest effort in cleaner external-validation design or richer biological resolution rather than in ever-larger taxonomic matrices. For a thesis concerned with scalability as well as prediction, that is an important outcome in its own right.

5.5 Reproducibility and Workflow Design

The thesis was built around a single primary workflow (Nextflow), chosen for portable, scalable dataflow execution under tight memory constraints. To check that the analysis was not tied to one software system, it was also implemented independently in Snakemake, which is built around explicit file dependencies and rule-based reproducibility. This independent implementation confirmed that interpretable microbial structure and a comparable saturation pattern could be recovered in a different software stack on different hardware.

It is important to be explicit about the scope of this check. The two implementations were not configured to run an *identical* procedure: they used a near-identical (but not the same) Metalog snapshot, different sample-size designs, a different cross-validation fold count, and feature-reduction steps that differ (the variance-based top-500 filter was applied in the Nextflow pipeline only). Because both the workflow manager and the analytical procedure differ at once, the two implementations cannot be used to isolate the effect of the workflow manager itself, and their numerical results are therefore *not* compared one-to-one. The Snakemake implementation is treated only as evidence that the pipeline can be reconstructed in a second system,

not as a parallel result set, and the headline results in this thesis are reported solely from the Nextflow pipeline.

This distinction is useful because reproducibility has several layers. Code reproducibility asks whether the same scripts can be executed again. Computational reproducibility asks whether the required environment, dependencies, and hardware allow the analysis to run. Analytical reproducibility asks whether similar high-level conclusions appear under reasonable alternative implementations. The present thesis is strongest on code and computational reproducibility for the Nextflow pipeline; a rigorous analytical-reproducibility test — two managers running an identical procedure so that only the manager differs — was not completed here and is identified as future work.

The workflow audit also improved the written thesis. Inspecting the actual pipeline code showed that global prefiltering occurred before the final held-out evaluation. Revising the Methods and Limitations chapters to state this explicitly makes the thesis more credible, because the code and the written claims are now aligned. That is an important part of reproducible academic writing.

This point is worth emphasizing because workflow design affects scientific interpretation, not only execution convenience. When preprocessing order, subset generation, and model calls are encoded explicitly in workflow scripts, methodological assumptions become inspectable. That transparency made it possible in this thesis to distinguish more carefully between direct outcome leakage, supervised feature selection, and global unsupervised prefiltering. In other words, the workflow infrastructure did not merely run the analysis; it helped clarify what the analysis actually meant.

5.6 Limitations

External validation. The thesis used held-out evaluation within the Metalog-derived data. This is appropriate for internal benchmarking, but external validation on geographically and methodologically independent datasets would provide a stronger test of generalization. Without such validation, the results should be interpreted as an internally evaluated baseline rather than as a deployed predictive model.

This limitation is especially important in microbiome research because external cohorts often differ in diet, geography, sequencing workflow, and case-mix even when the nominal phenotype is the same. A model can therefore perform well on internally held-out samples while still relying partly on patterns that do not transport well outside the development dataset. The absence of external validation does not remove the value of the current benchmark, but it does place a clear boundary around the strength of the claims that can be made from it.

Global prefiltering before final evaluation. In the Nextflow pipeline, prevalence filtering, top-500 variance filtering, near-zero-variance filtering, and centering/scaling were applied to the full matrix before downstream splitting and cross-validation. These steps were unsupervised with respect to BMI, but the design is still less conservative than fully nested fold-wise preprocessing. Reported performance may therefore be slightly optimistic relative to a fully nested pipeline.

Cohort heterogeneity. As a pooled multi-cohort dataset, the Metalog consortium introduces heterogeneity from multiple sources, including participant age, geographic origin, dietary patterns, medication exposure, disease status, sequencing protocols, and batch effects. The descriptive analysis of the BMI target (Section 4.1) made this concrete: the extreme BMI values are not data-entry errors but reflect the cohort composition — infants and clinical paediatric cohorts at the low end,

and bariatric, elderly, and metabolic-disease cohorts at the high end. This heterogeneity exposes the model to broader human variation, but it also means that observed associations may partly reflect cohort-level structure rather than universal microbiome-BMI biology; the adult-only sensitivity analysis (Section 4.1.2), in which several top species–BMI associations weakened markedly, is a direct illustration. A specific instance is age: BMI is not conventionally defined for children under two years, where weight-for-length and age-specific growth references are normally used, so the 1,288 infant samples in the pooled cohort should be read as part of the population-scale benchmark rather than as standard BMI measurements. The adult-only sensitivity analysis, which excludes this group, mitigates their influence and confirms that the main conclusions do not depend on them. Future work should explicitly stratify analyses by age and cohort source and test whether the saturation pattern is consistent within individual cohorts.

Cross-validation fold count. The Nextflow workflow used `kfold = 3`, lower than the more conventional 5-fold or 10-fold settings. Higher-fold cross-validation was evaluated but exceeded the computational constraints of the consumer-grade hardware platform used in this benchmark. With 17,903 samples, 3-fold cross-validation allocates approximately 11,935 training samples per fold, which is substantially larger than most published microbiome ML studies. To partially validate the 3-fold design, 5-fold cross-validation was applied to a randomly drawn 5,000-sample subset; results were consistent with those from 3-fold CV on the same subset (see Section 3), supporting the stability of the adopted evaluation design.

Cross-sectional design. The Metalog dataset is observational and cross-sectional. The study cannot determine whether microbial features contribute to BMI, result from BMI-related physiology, or reflect correlated lifestyle and environmental factors.

Workflow comparison not isolated. An independent Snakemake implementation was developed as a reproducibility check, but it was not configured to run an identical procedure to the primary Nextflow pipeline. It used a near-identical later Metalog snapshot (2025-12-28 versus 2025-12-14; nested and essentially the same data), a different set of fixed sample-size subsets, two-fold rather than three-fold cross-validation, and it did not apply the variance-based top-500 filter. Because the workflow manager and the analytical procedure differ at the same time, the two implementations cannot isolate the effect of the workflow manager itself, and their numerical outputs are therefore not compared one-to-one; only the Nextflow pipeline is used for the reported results. Fully synchronising the two pipelines into a single identical procedure — so that the only difference is the workflow manager — was not feasible within the project's time and consumer-hardware constraints, particularly because the unfiltered configuration did not complete on the available 8 GB machine. This is left as future work, and both implementations remain publicly available (Section 3.6) as a starting point.

Limited covariate and mechanistic metadata. The analysis intentionally excluded age, sex, and weight from the predictor matrix in order to focus on microbiome-only prediction. This design isolates microbial signal but does not remove all confounding. Diet, medication, geography, sequencing batch, and cohort origin may still influence the microbiome-BMI relationship. The study also lacked dietary, metabolomic, and longitudinal measurements needed for mechanistic interpretation.

Hardware-specific constraints. XGBoost was explored but not retained as a stable final model in the Nextflow benchmark because of software-dependency compatibility issues on the ARM64 (Apple Silicon) platform. A more comprehensive model comparison on HPC infrastructure could include boosted trees, non-linear kernels, and nested hyperparameter tuning more systematically.

BMI as a simplified phenotype. BMI is useful for large-scale benchmarking because it is widely available, but it compresses complex physiology into one number. It does not distinguish lean mass from fat mass, does not measure fat distribution, and does not directly capture insulin resistance or metabolic health. This limits the biological interpretation of prediction performance.

Relative abundance input. The analysis used species-level relative abundance profiles. Relative abundance is standard in metagenomic profiling, but it is affected by compositional constraints [28], [29]. Absolute microbial load, strain-level variation, and functional pathway abundance were not modeled directly. These omissions may limit both predictive performance and mechanistic interpretation.

5.7 Future Work

Future work should first implement fully nested preprocessing, where prevalence filtering, variance filtering, and scaling are estimated inside each training fold and then applied to held-out data. This would provide a stricter generalization estimate and directly address the main leakage limitation.

Second, external validation on independent cohorts is needed. The present thesis evaluates held-out performance within the Metalog-derived dataset, but generalization across cohorts is the stronger test for microbiome prediction. Stratified validation by geography, cohort source, sequencing protocol, or diet would also help determine whether the intermediate-sample dip reflects heterogeneity.

Third, future models should integrate additional data modalities. BMI is unlikely to be predicted optimally from species-level relative abundances alone. Dietary variables, medication history, metabolomics, host genetics, and longitudinal sampling would make it possible to distinguish predictive association from plausible mechanism.

Fourth, future benchmarking could compare additional model families under a common nested evaluation design. In particular, boosted trees, carefully tuned Elastic Net variants, and models built on transformed compositional features could help determine whether some of the current performance gap reflects model class, input representation, or workflow constraint.

Fifth, the two workflow implementations could be aligned into a single identical procedure — the same Metalog snapshot, feature-selection rule, model set, sample-size design, and cross-validation configuration — so that the only variable is the workflow manager itself. This would convert the present cross-system reproducibility check into a clean analytical-reproducibility test and isolate any effect of the manager. The independent Snakemake implementation released with this thesis provides a natural starting point for that work.

Finally, the top-ranked taxa identified here should be treated as candidates for follow-up, not as confirmed therapeutic targets. Their value lies in motivating more focused biological analyses, including metabolite measurements, diet-aware validation, and experimental follow-up. The direction-of-association analysis (Section 4.5.1) further suggests that adult-only or age-stratified modelling would help separate genuine adult microbiome–BMI signal from associations driven by the infant subpopulation.

5.8 Overall Interpretation

Taken together, the results support a measured conclusion. Species-level gut microbiome profiles contain measurable predictive information about BMI, but the signal is partial and sensitive to methodological choices. Non-linear modeling improves performance relative to the tested linear baselines, but the result does not prove ecological bistability. Large sample sizes help, but gains diminish after the main

saturation region. Feature reduction makes local computation feasible, but global prefiltering creates a limitation that should be corrected in future work.

This balanced interpretation is the main intellectual outcome of the revised thesis. The final text no longer treats model superiority as computational validation of the tipping-elements theory, and it no longer implies that feature-importance taxa are causal biomarkers. Instead, the thesis presents a reproducible computational baseline for microbiome-based BMI prediction, useful for understanding what can be predicted from current taxonomic profiles, useful for planning future cohorts, and useful for identifying where methodological improvement is still needed.

6 Conclusion

This thesis examined whether body mass index can be predicted from gut microbiome composition at population scale and under realistic local-computing constraints. Using 17,903 human gut metagenomes from the Metalog consortium, the study implemented a reproducible Nextflow benchmark — with an independent Snakemake implementation as a reproducibility check — and evaluated model class, sample size, and feature reduction as determinants of predictive performance. BMI was used as the target because it is widely available in large human metadata collections and therefore provides a practical phenotype for large-scale regression benchmarking, even though it remains an imperfect proxy for metabolic status.

Three main conclusions emerge from the work. First, the Random Forest model substantially outperformed the tested linear baselines, suggesting that the predictive structure in the available species-abundance data is not well captured by a simple additive linear representation. This supports the use of non-linear models for this task, but it does not prove ecological bistability or any specific tipping-point mechanism.

Second, the saturation analyses suggest that predictive gains diminish around 12,000–14,000 samples for the available taxonomic profiles. This provides a practical benchmark for future cohort planning. In the context of this thesis, the result indicates that simply adding more samples of the same type may yield smaller returns than improving validation design, adding richer covariates, or integrating additional omics layers.

Third, strong feature reduction made large-scale modeling feasible on consumer-grade hardware. A 1% prevalence filter and the variance-based top-500 filter sharply

reduced runtime and memory requirements while preserving most of the observed predictive performance. This is primarily a computational contribution, but it also lowers the barrier to reproducible exploratory analysis for resource-constrained research groups.

The thesis also clarifies several methodological limits that shape how the results should be interpreted. The workflows prevented direct outcome leakage by excluding BMI-defining or non-microbial metadata from the predictor matrix, but they did not fully nest all filtering and scaling inside the final evaluation loop. The reported performance should therefore be read as internally evaluated benchmark performance after global unsupervised prefiltering, not as a final clinical-grade estimate of generalization. Making this limitation explicit strengthens the credibility of the work rather than weakening it.

The biological interpretation remains necessarily cautious. In the main Nextflow Random Forest analysis, the highest-ranked predictors included an uncharacterised species-level genome bin (GGB9512_SGB14909), oral and opportunistic taxa, *Bifidobacterium longum*, and *Ruminococcus gnavus*. These should be viewed as candidates for further investigation rather than as validated causal biomarkers or therapeutic targets.

The broader significance of the thesis is therefore methodological as much as biological. By aligning the written interpretation with the actual workflow code, the project makes clear that predictive benchmarking in metagenomics depends not only on model choice but also on feature filtering order, validation scope, computational limits, and transparency about what was and was not nested inside evaluation. This is relevant for future student-led and research-group workflows. A benchmark that is modest but reproducible is often more useful than an apparently stronger result whose assumptions are underreported or whose implementation cannot be rerun easily outside a specialized environment. In that sense, the thesis contributes a

workable template for how large-scale microbiome machine-learning studies can be framed honestly when resources are limited but scientific standards still need to be high.

Overall, the thesis establishes a reproducible computational baseline for microbiome-based BMI prediction. Its main contribution is to show what can be achieved with species-level taxonomic profiles, transparent workflow design, and local computing resources, while also identifying the methodological improvements needed for stronger confirmatory studies. In that sense, the work is best understood as a careful and scalable benchmark that can support future microbiome prediction research rather than as a final answer to the biological mechanisms linking the gut microbiome and BMI.

References

[1] World Health Organization, *Obesity and overweight*, Fact sheet, updated 8 December 2025, 2025. Accessed: Apr. 25, 2026. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight.

[2] I. Rowland et al., "Gut microbiota functions: Metabolism of nutrients and other food components", *European Journal of Nutrition*, vol. 57, no. 1, pp. 1–24, 2018. DOI: 10.1007/s00394-017-1445-8.

[3] P. J. Turnbaugh et al., "An obesity-associated gut microbiome with increased capacity for energy harvest", *Nature*, vol. 444, no. 7122, pp. 1027–1031, 2006. DOI: 10.1038/nature05414.

[4] I. Moreno-Indias et al., "Statistical and machine learning techniques in human microbiome studies: Contemporary challenges and solutions", *Frontiers in Microbiology*, vol. 12, p. 635 781, 2021. DOI: 10.3389/fmicb.2021.635781.

[5] E. Pasolli et al., "Machine learning meta-analysis of large metagenomic datasets: Tools and biological insights", *PLOS Computational Biology*, vol. 12, no. 7, e1004977, 2016. DOI: 10.1371/journal.pcbi.1004977.

[6] A. Blanco-Miguez et al., "Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4", *Nature Biotechnology*, vol. 41, no. 11, pp. 1633–1644, 2023. DOI: 10.1038/s41587-023-01688-w.

[7] L. Lahti et al., "Tipping elements in the human intestinal ecosystem", *Nature Communications*, vol. 5, p. 4344, 2014. DOI: 10.1038/ncomms5344.

[8] L. Breiman, "Random forests", *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. DOI: 10.1023/A:1010933404324.

[9] F. Bäckhed et al., "The gut microbiota as an environmental factor that regulates fat storage", *Proceedings of the National Academy of Sciences*, vol. 101, no. 44, pp. 15 718–15 723, 2004. DOI: 10.1073/pnas.0407076101.

[10] J. K. Nicholson et al., "Host-gut microbiota metabolic interactions", *Science*, vol. 336, no. 6086, pp. 1262–1267, 2012. DOI: 10.1126/science.1223813.

[11] L. Lahti et al., "Associations between the human intestinal microbiota, Lactobacillus rhamnosus GG and serum lipids indicated by integrated analysis of high-throughput profiling data", *PeerJ*, vol. 1, e32, 2013. DOI: 10.7717/peerj.32.

[12] M. Van Hul and P. D. Cani, "The gut microbiota in obesity and weight management: Microbes as friends or foe?", *Nature Reviews Endocrinology*, vol. 19, pp. 258–271, 2023. DOI: 10.1038/s41574-022-00794-0.

[13] Human Microbiome Project Consortium, "Structure, function and diversity of the healthy human microbiome", *Nature*, vol. 486, no. 7402, pp. 207–214, 2012. DOI: 10.1038/nature11234.

[14] T. Yatsunenko et al., "Human gut microbiome viewed across age and geography", *Nature*, vol. 486, no. 7402, pp. 222–227, 2012. DOI: 10.1038/nature11053.

[15] R. N. Carmody and J. E. Bisanz, "Roles of the gut microbiome in weight management", *Nature Reviews Microbiology*, vol. 21, pp. 535–550, 2023. DOI: 10.1038/s41579-023-00888-0.

[16] R. E. Ley et al., "Microbial ecology: Human gut microbes associated with obesity", *Nature*, vol. 444, no. 7122, pp. 1022–1023, 2006. DOI: 10.1038/4441022a.

[17] P. J. Turnbaugh et al., "A core gut microbiome in obese and lean twins", *Nature*, vol. 457, no. 7228, pp. 480–484, 2009. DOI: 10.1038/nature07540.

[18] F. Magne et al., "The Firmicutes/Bacteroidetes ratio: A relevant marker of gut dysbiosis in obese patients?", *Nutrients*, vol. 12, no. 5, p. 1474, 2020. DOI: 10.3390/nu12051474.

[19] M. A. Sze and P. D. Schloss, "Looking for a signal in the noise: Revisiting obesity and the microbiome", *mBio*, vol. 7, no. 4, e01018–16, 2016. DOI: 10.1128/mBio.01018-16.

[20] E. Le Chatelier et al., "Richness of human gut microbiome correlates with metabolic markers", *Nature*, vol. 500, no. 7464, pp. 541–546, 2013. DOI: 10.1038/nature12506.

[21] J. Qin et al., "A metagenome-wide association study of gut microbiota in type 2 diabetes", *Nature*, vol. 490, no. 7418, pp. 55–60, 2012. DOI: 10.1038/nature11450.

[22] A. Zhernakova et al., "Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity", *Science*, vol. 352, no. 6285, pp. 565–569, 2016. DOI: 10.1126/science.aad3369.

[23] B. D. Topçuoğlu et al., "A framework for effective application of machine learning to microbiome-based classification problems", *mBio*, vol. 11, no. 3, e00434–20, 2020. DOI: 10.1128/mBio.00434-20.

[24] G. Papoutsoglou et al., "Machine learning approaches in microbiome research: Challenges and best practices", *Frontiers in Microbiology*, vol. 14, p. 1 261 889, 2023. DOI: 10.3389/fmicb.2023.1261889.

[25] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net", *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005. DOI: 10.1111/j.1467-9868.2005.00503.x.

[26] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer, 1995.

[27] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system", in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794. DOI: 10.1145/2939672.2939785.

[28] J. Aitchison, "The statistical analysis of compositional data (with discussion)", *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 44, no. 2, pp. 139–177, 1982. DOI: 10.1111/j.2517-6161.1982.tb01195.x.

[29] G. B. Gloor et al., "Microbiome datasets are compositional: And this is not optional", *Frontiers in Microbiology*, vol. 8, p. 2224, 2017. DOI: 10.3389/fmicb.2017.02224.

[30] P. Di Tommaso et al., "Nextflow enables reproducible computational workflows", *Nature Biotechnology*, vol. 35, no. 4, pp. 316–319, 2017. DOI: 10.1038/nbt.3820.

[31] J. Köster and S. Rahmann, "Snakemake–a scalable bioinformatics workflow engine", *Bioinformatics*, vol. 28, no. 19, pp. 2520–2522, 2012. DOI: 10.1093/bioinformatics/bts480.

[32] B. Grüning et al., "Bioconda: Sustainable and comprehensive software distribution for the life sciences", *Nature Methods*, vol. 15, no. 7, pp. 475–476, 2018. DOI: 10.1038/s41592-018-0046-7.

[33] M. Kuhn et al., "Metalog: Curated and harmonised contextual data for global metagenomics samples", *bioRxiv*, 2025, Preprint; published in Nucleic Acids Research, doi:10.1093/nar/gkaf1118. Database available at https://metalog.embl.de/. DOI: 10.1101/2025.08.14.670324.

[34] B. D. Topçuoğlu et al., "Mikropml: User-friendly R package for supervised machine learning pipelines", *Journal of Open Source Software*, vol. 6, no. 61, p. 3073, 2021. DOI: 10.21105/joss.03073.

[35] A. Liaw and M. Wiener, "Classification and regression by randomForest", *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[36] M. Kuhn, "Building predictive models in R using the caret package", *Journal of Statistical Software*, vol. 28, no. 5, pp. 1–26, 2008. DOI: 10.18637/jss.v028.i05.

[37] S. Zaman, *Microbiome bmi thesis: Nextflow workflow for random forest, Elastic Net, SVM, and decision-tree benchmarking of BMI prediction from gut metagenomic profiles*, version f565f74d, Artistic 2.0 licence; archived snapshot of the public GitHub repository at the cited commit hash., 2026. DOI: 10.5281/zenodo.20569792. [Online]. Available: https://github.com/Sadia12345/Microbiome_BMI_Thesis_Nextflow.

[38] S. Zaman, *Thesis bmi prediction: Snakemake workflow for repeated-seed Elastic Net saturation analysis of BMI prediction from gut metagenomic profiles*, version 6185a374, Artistic 2.0 licence; archived snapshot of the public GitHub repository at the cited commit hash., 2026. DOI: 10.5281/zenodo.20569828. [Online]. Available: https://github.com/Sadia12345/Thesis-BMI-Prediction-Snakemake.

[39] G. Wilson et al., "Good enough practices in scientific computing", *PLOS Computational Biology*, vol. 13, no. 6, e1005510, 2017. DOI: 10.1371/journal.pcbi.1005510.