

METAGENOMIC PROFILING OF THE GUT MICROBIOME FOR EARLY PREDICTION OF METABOLIC DISORDERS: A STRUCTURED REVIEW AND INTEGRATIVE MACHINE-LEARNING FRAMEWORK

Dr Neha Bokey¹, Dr. Sonal Mahajan²

¹Assistant professor, Department of Biosciences and Technology MIT-WPU Kothrud Pune, orcid- 0009-0006-3410-538X
neha.bokey@mitwpu.edu.in

²School of Biosciences and Bioengineering, D Y Patil International University, sector 29, Akurdi Pune 411044, orcid -0000-0003-4512-8641
sonal.mahajan@dypiu.ac.in

ABSTRACT

Obesity, type 2 diabetes mellitus (T2D), metabolic syndrome (MetS), and non-alcoholic fatty liver disease (NAFLD) are collectively the most rapidly growing group of non-communicable diseases, including global diabetes, which is expected to increase from 589 million adult cases in 2024 to 853 million in 2050. Current risk-assessment strategies rely on anthropometric and biochemical variables that often become abnormal only after severe metabolic damage has occurred; a diagnostic window exists in which earlier identification of minimally invasive biomarkers could significantly shape clinical pathways. We now know of a new and modifiable organ-like ecosystem, the gut microbiome, that is metabolically active and is a potentially useful substrate for such early-warning biomarkers. This paper synthesises evidence on metagenomic profiling technologies and machine-learning (ML) approaches to gut microbiome data used to identify metabolic disorders early, and outlines an integrative, clinically translatable approach. We searched for human cohort studies involving 16S rRNA or shotgun metagenomic sequencing and supervised ML to predict obesity, T2D, MetS, or NAFLD from 2006 to 2026. Results: Both longitudinal exploration and literature synthesis showed consistent directionality in dysbiotic, at-risk dynamics, including depletion of Akkermansia muciniphila, Faecalibacterium prausnitzii, and other short-chain fatty acid producers, and a higher Firmicutes-to-Bacteroidetes ratio. Several classifiers, such as random forest and gradient-boosting classifiers trained on gut metagenomic features, achieved AUCs ranging from 0.76 to 0.87 for T2D discrimination, which improved when gut metagenomic features were combined with simple clinical covariates and slightly worsened in independent ethnic cohorts. In conclusion, non-invasive metabolic risk stratification via metagenomic profiling and interpretable machine learning is a biologically plausible and increasingly supported approach, with the caveat that routine implementation still requires evidence of biological causality, standardisation across populations, and longitudinal integration of multiple omics.

KEYWORDS: gut microbiome, metagenomics, type 2 diabetes, obesity, machine learning, predictive biomarkers, shotgun sequencing, precision nutrition, metabolic syndrome.

1. INTRODUCTION

Metabolic disorders (obesity, type 2 diabetes mellitus (T2D), metabolic syndrome (MetS), and non-alcoholic fatty liver disease (NAFLD, recently renamed metabolic dysfunction-associated steatotic liver disease)) form a colliding class of disorders characterised by insulin resistance, metabolic and chronic low-level inflammation, and disturbed energy homeostasis. The number of people living with diabetes (PWD) is estimated to be 589 million, representing a prevalence of 11.11% in 2024 and is projected to be 853 million (12.96% prevalence) by 2050 – according to the 11th edition of the International Diabetes Federation (IDF) Diabetes Atlas, the majority of this burden is expected to be in low- and middle-income countries [1,2,3]. More than 90 per cent of this occurs because T2D follows years to decades of subclinical metabolic drift before conventional screening tools, such as fasting plasma glucose, oral glucose tolerance testing, and glycated haemoglobin (HbA1c), fall within the normal range [1,3].

This diagnostic latency has driven the search for biomarkers that can indicate metabolic risk at the prediabetes stage, before overt hyperglycaemia, adiposity-related organ damage, and/or hepatic steatosis. The gut microbiome stands out amongst the many systems currently under investigation due to its (a) mechanistic link to host energy harvest, inflammation and bile-acid and short-chain-fatty-acid (SCFA) signalling; (b) ease and non-invasiveness of sampling through the collection of faeces; (c) ability to be modified through diet, prebiotics, probiotics and pharmacological intervention. Initial studies with culture-independent sequencing methods have revealed a consistent difference in “relative abundance” between the phyla Firmicutes and Bacteroidetes in lean and obese individuals, and that colonization of germ-free mice with an “obese microbiome” enhances the ability of recipients to extract energy from their diet [4, 5]. This paradigm was then expanded to the field of type 2 diabetes (T2D), where tens of thousands of microbial gene markers for association with T2D were identified by subsequent large-

scale metagenome-wide association studies and the normal reference microbiome—the human microbiome—was mapped by the Human Microbiome Project [6,7].

Two technological developments have come together over the past ten years, making this body of observations ready for translation. First, the low cost and high throughput of next-generation sequencing (NGS) have expanded whole-genome shotgun approaches in metagenomics, which provide population-scale information on microbial species and strain composition and functional gene content, while cost-effective 16S rRNA amplicon sequencing remains useful for large screening cohorts. Second, new machine-learning (ML) techniques for compositional data of scarce nature with a large number of features (thousands) (e.g., random forests, gradient boosted trees, interpretable models like SHapley Additive exPlanations) have made it possible to build ML models to convert the thousands of microbial features into one clinically interpretable estimate of risk. Areas under the ROC curve (AUC) up to 0.76, 0.87 or more have been achieved by several independent research groups, with a distinction as good as, and perhaps superior to the conventional clinical risk scores.

But work is still in progress: the field is still fragmented. The heterogeneous study design with regard to the sequencing platform, the bioinformatic pipelines, the definition of outcomes, the population ancestry, and the sample sizes makes it difficult to compare and translate the results of individual studies to clinical practice. The potential for causal inference is very limited, with the overwhelming majority of human studies being cross-sectional or short-term prospective, and confounding factors such as diet, medication (such as metformin, which can alter microbial composition), and geography are also prevalent. In this context, the current study aims at four goals: (i) summarizing in one place the mechanistic steps that link gut microbial ecology and metabolic disease; (ii) contrasting the performances of the two research approaches, 16S and metagenomic; (iii) critically reviewing the performance and limitations of published machine learning models trained on gut metagenomic data, and identifying potential areas for further research; and (iv) advocating an integrative and tiered screening strategy based on metagenomic profiling and interpretable machine learning coupled with clinical data. The remainder of the paper is structured as follows: In Section 2, the methodology for performing a literature synthesis is described; The mechanistic gut microbiome-metabolism axis is outlined in Section 3; Different metagenomic profiling technologies and bioinformatic pipelines are compared in Section 4; Machine-learning approaches are reviewed in Section 5; Comparative predictive performance across different published studies was synthesised in Section 6; The proposed integrative framework is presented in Section 7; Sections 9 and 10 address limitations and future directions; and Section 11 concludes.

2. MATERIALS AND METHODS

2.1 Literature Search Strategy

A structured narrative synthesis (integrating, but not strictly following, PRISMA reporting conventions, as per the aim of the conceptual framework for this paper) was performed over PubMed/MEDLINE, Scopus, Web of Science, Google Scholar and the bioRxiv/medRxiv preprint servers from January 2006 to September 2026. Search strings included both controlled vocabulary and free-text words, and information was retrieved for both YPD and controlled vocabulary terms, using the matching Boolean structure: ("gut microbiome" OR "gut microbiota" OR "metagenomics" OR "metagenomic profiling") AND ("machine learning" OR "artificial intelligence" OR "predictive model" OR "biomarker") AND ("type 2 diabetes" OR "obesity" OR "metabolic syndrome" OR "NAFLD" OR "insulin resistance"). We hand-searched the reference lists of retrieved reviews to identify eligible primary studies (backward snowballing).

2.2 Eligibility Criteria

For detailed synthesis, study inclusion criteria were: (i) used human faecal or gut-derived microbial DNA/RNA as data source; (ii) derived the analyses by 16S rRNA amplicon or whole-genome shotgun sequencing; (iii) had a quantitative performance metric (AUC, accuracy, sensitivity/specificity) or clearly described differentially abundant taxa; and (iv) were published in a peer-reviewed journal or citable preprint in full-text. Mechanistic studies – conducted exclusively in animals – were included selectively when they provided causal information that could not be obtained from human cross-sectional studies (such as germ-free mouse colonisation experiments and faecal microbiota transplantation studies). We did not include case reports and conference abstracts with no useful information for extraction in the comparative performance synthesis (Section 6), but they may be useful for discussing mechanisms.

2.3 Data Extraction and Figure Construction

For every study that met the feasibility criteria for predicting an individual's class using predictive-modelling methods, we extracted the following data where reported: cohort representation (size and abundance); sequencing method; key rows differentially abundant between classes; and reported discriminative performance. We used the extracted values to re-create the AUC and 95% confidence interval (CI) reported in the respective primary sources, and included them in Table 2 and Figure 5 in this paper. Figures 2 and 3 (Firmicutes: Bacteroidetes ratio and Akkermansia muciniphila abundance across body-mass-index categories, respectively) are instead qualitative "pools" of the directionality of the associations reported across the literature cited and should not be drawn as inferences from primary quantitative data. We follow this distinction throughout the manuscript for scientific transparency.

2.4 Literature Selection Flow

We summarised the literature selection process in a PRISMA-line flow diagram, as shown in Figure 1. We searched 112 records across databases and preprint servers, which were reduced to 41 after removing duplicates; we screened 94 unique records by title and abstract, and excluded 41 for not discussing the intersection of gut microbiome/metagenomics, machine learning, and metabolic disease. Of the remaining 53 articles that had been retrieved in full text, 24 were excluded due to not reporting an extractable performance metric, containing < 50 participants, reporting on a cohort already included in another article, or not being on topic on full-text review. This left 29 studies available for narrative synthesis (Sections 3-5) and, when they reported quantitative predictive performance, were included in the comparative synthesis in Table 2.

Figure 1. Literature Selection Process (PRISMA-Style Flow Diagram)

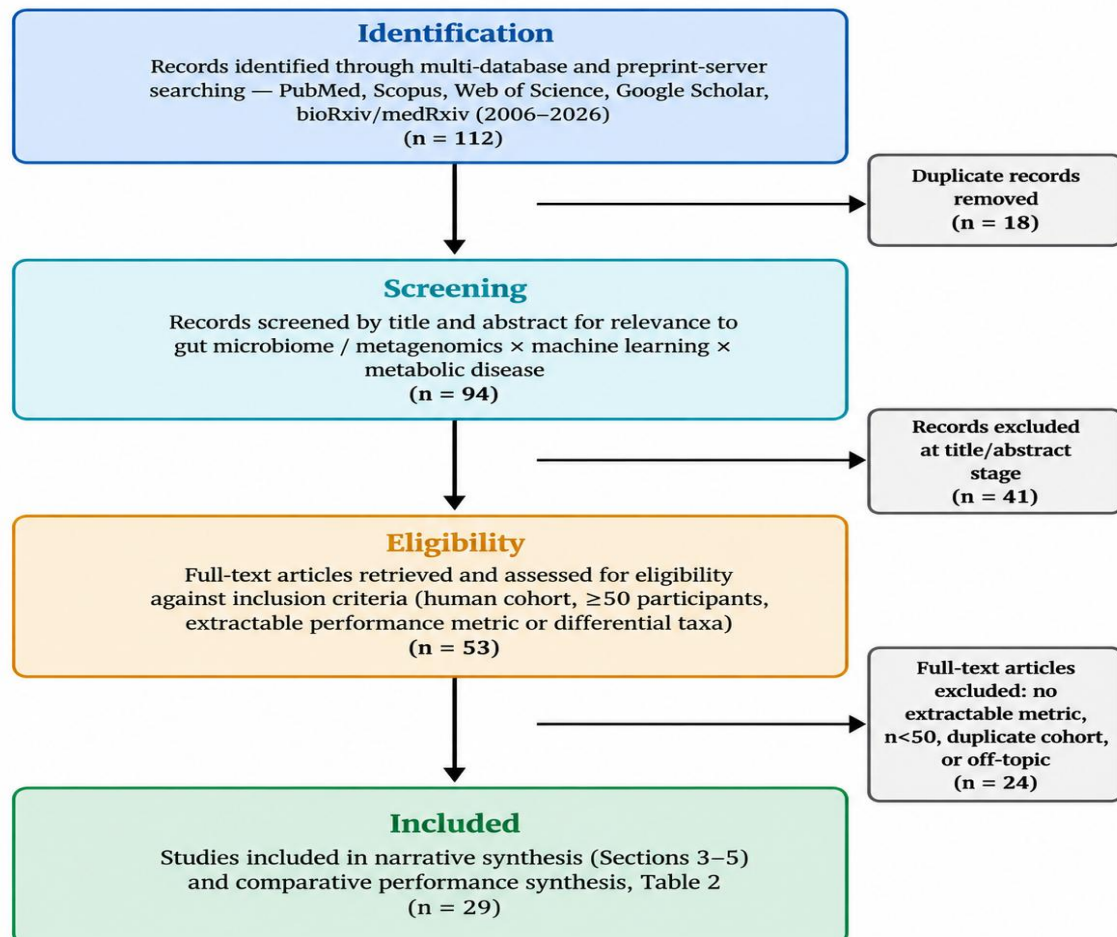


Figure 1. Literature selection process (PRISMA-style flow diagram) underlying the narrative and comparative-performance synthesis presented in this paper.

3. The Gut Microbiome in Metabolic Homeostasis: A Mechanistic Overview

3.1 Composition and Functional Capacity of the Reference Gut Microbiome

The estimated intestinal adult human microbiome consists of tens of trillions of microorganisms representing over one thousand bacterial species, an order of magnitude more than the number of genes encoded by the human genome (several million microbial genes, without accounting for redundancy among microbial strains). The taxonomic results from the Human Microbiome Project Consortium's reference characterization of the human gut microbiome, although highly variable from individual to individual, revealed that the number of functional genes that are part of the healthy gut microbiome is relatively conserved (core metabolic pathways for carbohydrate fermentation, amino acid biosynthesis and vitamin production), which indicated that focusing on the function of the gut microbiome, and not just on taxonomy, might be of particular value for predicting disease [7]. The two most prevalent bacterial phyla, Firmicutes (60-80%) and Bacteroidetes (20-30%), usually account for most (>90%) of currently identifiable bacterial abundance. Actinobacteria, Proteobacteria, and Verrucomicrobia are at lower relative abundances but can be more metabolically significant, with the latter including the mucin-degrader family Akkermansia.

3.2 Short-Chain Fatty Acids, Energy Harvest

A seminal study in genetically obese (ob/ob) mice compared to non-obese littermates revealed that, during obesity, there was an increase in the abundance of Firmicutes and decrease in Bacteroidetes, and that transferring

an obese-associated microbial community to germ-free mice increased their ability to harvest energy from the same diet compared to transferring a lean-associated community, which led to a higher degree of adiposity in the germ-free mice harboring an obese-associated microbial community [4,5]. The proposed mechanism is related to the production of SCFAs, mainly acetate, propionate, and butyrate, as a result of bacterial fermentation of the indigestible fraction of the diet - dietary fibre. SCFAs serve multiple roles, both as an extra source of calories for the host and as signalling molecules that act through G-protein coupled receptors (GPR41 or FFAR3, as well as GPR43 or FFAR2), affecting satiety hormone production, adipose tissue differentiation, and hepatic gluconeogenesis. Loss of anti-inflammatory, barrier-protecting SCFA signalling in dysbiotic states, characterised by a depletion of important butyrate-producing taxa, including *Faecalibacterium prausnitzii*, *Roseburia intestinalis* and *Roseburia inulinivorans*, has been noted on many occasions in cohorts of T2D individuals [6,9].

3.3 Gut Barrier Integrity, Low-Grade Inflammation, and Metabolic Endotoxemia

A second complementary, mechanistic axis involves Lipopolysaccharide (LPS), a constituent of the outer membrane of Gram-negative bacteria. When LPS is injected into mice to induce elevated circulating levels by "metabolic endotoxaemia", many characteristics of diet-induced obesity and insulin resistance are recapitulated, such as weight gain, inflammation in the liver and adipose tissue, and insulin resistance, which is thought to occur through the binding of LPS to the Toll-like receptor 4 (TLR4) on innate immune and adipose cells [8]. The dysbiotic microbial communities that favour the presence of Gram-negative bacteria that contain LPS, along with a leaky intestinal epithelial barrier (reduced expression of tight-junction proteins, increased barrier permeability), generate a vicious cycle that propels and perpetuates translocated bacterial products activating a state of chronic, low grade systemic inflammation, now known to be an integral part of the pathophysiology of obesity, T2D and NAFLD.

3.4 Bile acids and Secondary signalling pathways

Another important task of the gut microbiota is biotransforming primary bile acids, synthesised by the host, into secondary bile acids that act as ligands for the farnesoid X receptor (FXR) and the G protein-coupled bile acid receptor (TGR5). Thus, dysbiosis-induced changes in bile-acid pool composition may indirectly affect hepatic lipid and glucose metabolism as an additional mechanism, where gut microbial composition changes may precede events and even predict the clinical onset of NAFLD and MetS, although this axis has not been as well defined in large-scale human predictive-modelling studies as the SCFA and endotoxaemia axes.

3.5 Taxon-Level Biomarker Candidates Across the Metabolic Disease Spectrum

In addition to phylum-level changes, other genera and species have been identified in different studies and considered potential biomarkers. Consistently associated with obesity, T2D and related metabolic dysfunction is a decrease in the proportion of *Akkermansia muciniphila*, a resident of the intestinal mucus layer, that is capable of degrading mucus glycans [11-17]. In contrast, sulfate-reducing *Bilophila wadsworthia* and the genus *Blautia* are more commonly reported as enriched in dysbiotic, at-risk states, and species like *Bacteroides ovatus* and *Anaerostipes hadrus* have recently been identified as condition-specific, differentially abundant markers in multi-cohort shotgun metagenomics of T2D [6,10]. The following, shown in Figure 2 and Figure 3, are the overall trends derived from these directional trends.

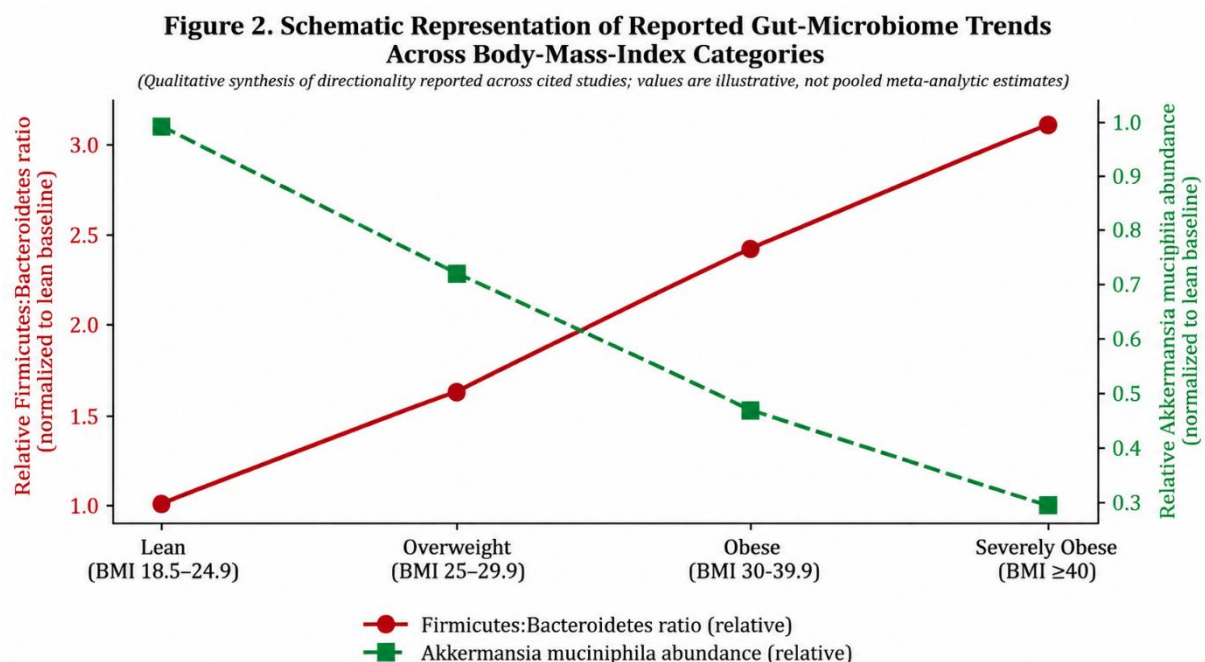


Figure 2. Schematic representation of the qualitative trend, reported across the cited literature, of an increasing Firmicutes: Bacteroidetes ratio and decreasing Akkermansia muciniphila abundance with increasing body-mass-index category. Values are illustrative and normalised to a lean baseline; they synthesise reported directionality rather than pooled quantitative estimates [4,5,11-17].

Figure 3. Literature-Synthesized Directionality of Key Gut-Microbial Taxa Across Metabolic Disorders

<i>Akkermansia muciniphila</i>	↓ Depleted	↓ Depleted	↓ Depleted
<i>Firmicutes</i> (phylum)	↑ Enriched	↑ Enriched	↑ Enriched
<i>Bacteroidetes</i> (phylum)	↓ Depleted	↓ Depleted	↓ Depleted
<i>Faecalibacterium prausnitzii</i>	↓ Depleted	↓ Depleted	↓ Depleted
<i>Blautia</i> spp.	↑ Enriched	↑ Enriched	↔ Mixed
<i>Bilophila wadsworthia</i>	↑ Enriched	↑ Enriched	↑ Enriched
<i>Ruminococcus</i> spp. (SCFA-poor spp.)	↔ Mixed	↔ Mixed	↔ Mixed
<i>Bacteroides ovatus</i>	↔ Mixed	↑ Enriched	↔ Mixed
<i>Anaerostipes hadrus</i>	↔ Mixed	↑ Enriched	↔ Mixed
	Obesity	Type 2 Diabetes	NAFLD / Metabolic Syndrome

Directionality relative to healthy controls, synthesized qualitatively from the cited primary and review studies.

Figure 3. Literature-synthesised directionality (enriched, depleted, or mixed/inconsistent evidence) of nine recurrently reported gut-microbial taxa across obesity, type 2 diabetes, and NAFLD/metabolic syndrome, relative to healthy controls [4-6,9-17].

4. Metagenomic Profiling Technologies and Bioinformatic Pipelines

4.1 16S rRNA Amplicon Sequencing

16S rRNA gene sequencing targets one or more hypervariable regions (usually V3-V4) of the bacterial 16S ribosomal RNA gene, enabling low-cost, high-throughput species identification down to the genus level. This is mainly due to its scalability, with a mature analytical ecosystem that supports large-cohort population screening, such as a citizen-science diabetes-screening project with 16S profiling of 692 people [27]. Main problems with this method are less fine taxonomic resolution and poor functional resolution—the latter because while an amplicon can be utilized to describe gene content, this cannot be done directly with this method, only very indirectly.

4.2 Whole-Genome Shotgun Metagenomics

Randomly fragmenting and sequencing all DNA in a sample via shotgun metagenomic sequencing enables species- and strain-level taxonomic resolution, as well as functional gene and pathway profiling. Although the shotgun approach requires more reads and is more expensive, comparative analyses have consistently shown improved species-level identification of bacteria, higher estimates of community diversity, and improved gene-level functional prediction compared with 16S amplicon sequencing approaches [18]. Notably, 16S sequences-derived predictive models can reach the accuracy attained by shotgun sequences-derived models in disease prediction analyses (e.g. colorectal cancer or paediatric ulcerative colitis) despite the less resolved taxonomy, intimating that, at least for some disease prediction tasks, the more detailed taxonomy derived from shotgun sequencing does not automatically guarantee proportionate improvement in disease discrimination [19-22]. The practical implications for this are in population screening, if only a small amount of money is available, 16S sequencing might be the best cost-accuracy option for a first-round screening method.

4.3 Emerging approaches - Metatranscriptomics and Multi-Omics Integration

More recently, gut metatranscriptomics has been developed to profile microbiome RNA expression rather than DNA presence, indicating the microbiome's potential functions. A large-scale clinical-grade metatranscriptomic study of 53,970 individuals demonstrated the potential for using such an approach for genome-scale risk assessment in T2D at a population level, and some genus-level associations reported in the literature (e.g., *Ruminococcus* spp.) were found not to reflect gene expression directly, suggesting that activity-based profiling is a value-added data type on top of DNA-based profiling [28].

4.4 End-to-End Profiling and Prediction Pipeline

Figure 4 shows the integrated pipeline from acquisition of the faecal sample or of a rectal-swab sample to biobanking, nucleic acid extraction, quality control, sequencing (by either 16S or shotgun), bioinformatic processing of data, feature engineering, prediction using machine learning with layers of interpretability, and output presentation in the form of a clinical risk-stratification result. Because the composition of microbes in the body is dynamic and changes with diet, medication, and disease progression, a longitudinal feedback loop is important: samples should be collected over the same time period so the model can be recalibrated over time.

Figure 4. Integrated Metagenomic Profiling and Machine-Learning Pipeline for Early Prediction of Metabolic Disorders

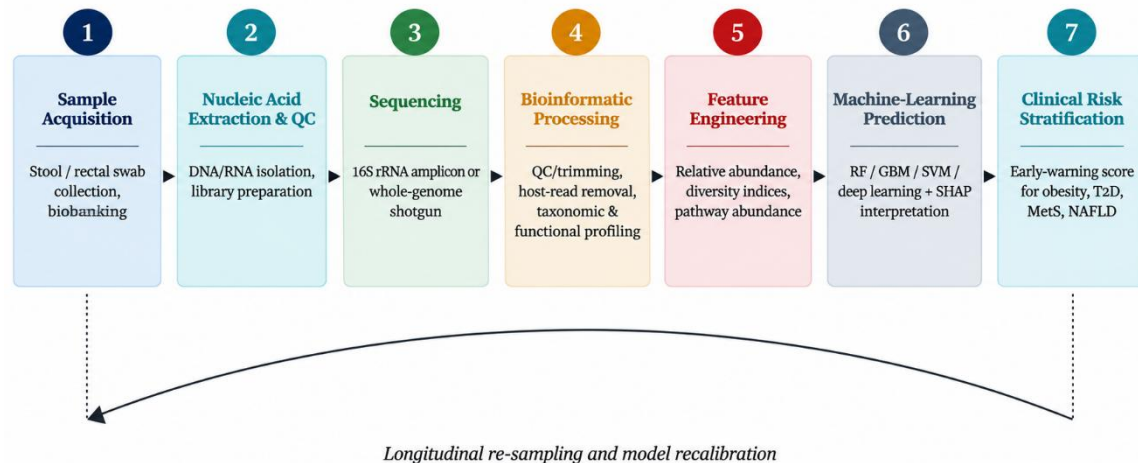


Figure 4. Integrated metagenomic profiling and machine-learning pipeline for early prediction of metabolic disorders, from sample acquisition through clinical risk stratification, with a longitudinal recalibration loop.

4.5 Core Bioinformatic Tools

Some of such bioinformatic resources at each step in the pipeline are listed in Table 1. K-mer based exact-alignment approaches like Kraken2 [23] or marker-gene approaches like MetaPhlAn2 [24] are commonly used to classify shotgun reads by taxonomy and abundance of functional pathways are calculated based on functional family of tools called HUMAnN. The diversity of community can be quantified either by alpha-diversity metrics (such as Shannon, Simpson within-sample richness/evenness) or by beta-diversity metrics (such as Bray-Curtis dissimilarity, UniFrac distance, between-sample compositional divergence); alpha-diversity is often used as an engineered feature of the downstream ML model along with raw taxon abundances.

Pipeline Stage	Representative Tool(s)	Primary Function
Read quality control	fastp; Trimmomatic	Adapter removal, quality trimming, low-quality read filtering
Host-read depletion	Bowtie2 (vs. human reference genome)	Removal of host-derived reads prior to microbial analysis
Taxonomic classification	Kraken2 [23]; MetaPhlAn2 [24]	k-mer exact alignment or marker-gene-based species/strain assignment
Functional profiling	HUMAnN2/3	Gene family and metabolic pathway abundance estimation
Assembly (shotgun)	MEGAHIT; metaSPAdes	De novo assembly of metagenome-assembled genomes
Diversity & statistics	QIIME2; phyloseq; vegan (R)	Alpha-/beta-diversity computation, ordination, statistical testing
Machine learning	scikit-learn; XGBoost/LightGBM; SHAP	Model training, gradient boosting, feature-importance interpretation

Table 1. Representative bioinformatic tools used across the metagenomic profiling and machine-learning pipeline.

5. Machine Learning and Artificial Intelligence for Predictive Modelling

5.1 Feature Engineering from Metagenomic Data

When the absolute number of the metagenomes are represented as counts, the resulting data are high-dimensional (often there are thousands of taxa and/or gene families present in the samples) and sparse (most taxa may be absent and/or too low in abundance to be counted from an individual sample), and subject to compositional effects (sizes of counts of all the taxa sum to a fixed value, creating spurious negative correlations, if not appropriately transformed). Feature-engineering approaches have included centrifuged/Elund-Ëtimati alfa-diversity hojaq,

function araliqvistuan-processed abundance of features, alfa-diversity processuoshte-fishq indices of the abundance as well as the features, and recently in the models which are being done in an interpretable way, making use of a composite "microbiome risk score", which merges the direction and threshold of many separately weak characteristics, into a single interpretable processuoshte-fishq [26].

5.2 Supervised Learning Algorithms

Because of their inherent ability to deal with high dimensional, non-linear, sparse feature space, and relative insensitivity to overfitting when regularized, the tree-based ensemble methods, especially random forests (RF) and gradient-boosting ensembles (e.g., LightGBM) are the most common approach to model forest data in the reviewed literature. RF classifiers trained on gut metagenomic features yielded 0.76 to 0.87 for T2D discrimination with usually best performance compared to linear classifiers, such as elastic net, but elastic net and support vector machine with radial kernel are also common models, especially when attempting to greatly restrict the number of features [25-27,29].

5.3 Deep Learning and Emerging Multi-Omic Fusion Approaches

In recent years, deep-learning architectures, such as autoencoder based dimensionality reduction and use of convolutional neural networks with microbiome data that is ordered by phylogeny, are more commonly considered for microbiome classification tasks, and their appeal is that of learning non-linear relationships between features without the need for extensive manual feature engineering. To date, however, relative to tree-based ensemble methods, deep learning methods have generally required significantly larger training cohorts in order to avoid over-fitting the model using the typical sample sizes of microbiome studies (hundreds to low thousands of participants) and the use of deep learning in the prediction of metabolic diseases in these studies remains comparatively limited.

5.4 Model Interpretability

One of the main challenges of ML in the biomedical domain is the "black-box" problem in which the internal logic of decision making of a model is not understood by clinicians or researchers. Recent frameworks have addressed this direct; by combining gradient-boosted models with SHapley Additive exPlanations (SHAP), a game-theoretic approach to sharing a model's prediction into the set of contributions of each of the input features. Using this method, three separate cohorts (with more than 9,000 subjects) were analysed and one of these frameworks resulted in the identification of 15 interpretable microbiome features (taxon abundances and diversity indices), which were then integrated into a directional, thresholded microbiome risk score for T2D [26]. Interpretable architectures are especially crucial for eventual clinical uptake as the output of a model needs to be comprehensible for both the clinicians and the patients.

5.5 Generalizability and Validation Strategies

More than a single cohort's internal cross validation is needed to validate your results. The best practice example of a recent multi-cohort shotgun metagenomic analysis shows that, using a random forest model with 18 differentially abundant species, the AUC of 0.836 reached in the independent test cohort improved to 0.871 in the same cohort including basic clinical covariates such as age, sex and body-mass index, but dropped to 0.815 and 0.832 for the same model evaluated separately in each of the Asian and European test cohorts respectively, hinting at modest but non-trivial reduction in cross-population generalizability [10]. The trend, of good performance in the overall population but worse performance in measurable chunks of its sub-population in a validating cohort, geographically or ancestrally different, is seen throughout the literature, and is one of the central translational issues that is the subject of Section 9.

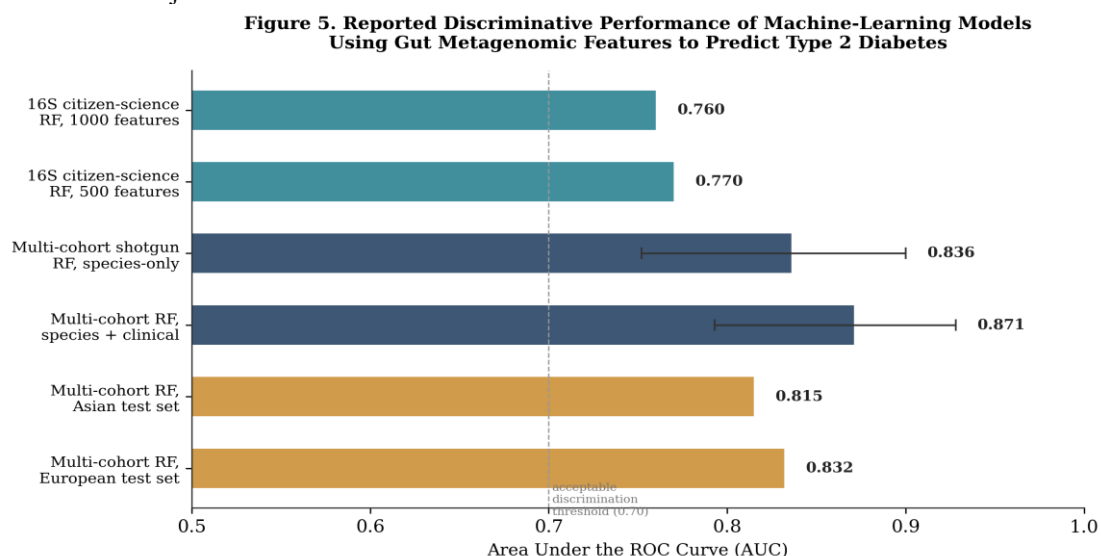


Figure 5. Reported discriminative performance (AUC with 95% confidence intervals, where available) of representative machine-learning models trained on gut metagenomic features for type 2 diabetes classification, compiled from the cited primary studies [10,27].

6. Comparative Synthesis of Predictive Performance Across Studies

Cohorts characteristics, the method used for imaging and analysis, important biomarkers, and performance from major studies included in the synthesis of this review are listed in Table 2. A number of patterns or trends can be observed. First, when comparable performance is evaluated in purely microbiological and clinical data, it can be shown that microbial data in isolation can supplement the routinely available clinical data to modestly improve the performance, indicating a non-redundant but partially complementary relationship between the two. Second, the sensitivity of the model improved significantly (from 0.46 to 0.80) with fewer features (from 1000 to 500 most-variable features in one 16S based screening study), suggesting that many raw taxonomic features are either not differentiating or are not usable and indicating that principled feature selection is an important and sometimes under-appreciated methodological step [27]. Thirdly, longitudinal, time-series designs that monitor microbial changes over time in association with continual glycaemic measures over 1.5- and 4-years, suggest that microbial variables, in combination with clinical ones, predict future HbA1c and insulin levels better than concurrent ones, pointing towards a truly predictive (as opposed to purely concurrent, diagnostic) role for the gut microbiome [29]. Fourth, and most critical to translative plans, the cross-population validation shows that there is a measurable loss in performance when it is deployed globally, underscoring the need for more ancestrally diverse training data prior to deployment of a single fixed model everywhere [10].

Study / Ref.	Cohort	Sequencing Method	Key Findings	Performance
Qin et al., 2012 [6]	345 Chinese adults (145 T2D, 200 controls)	Shotgun; metagenome-wide association study	~60,000 T2D-associated gene markers; ↑opportunistic pathogens, ↓butyrate producers	Markers validated for classification in independent subset
Karlsson et al., 2013 [25]	European women (T2D / IGT / NGT)	Shotgun; metagenomic species clustering	Altered Clostridiales; ↑Lactobacillus in T2D	Discriminative gene-marker set identified
Interpretable ML framework, 2020 [26]	>9,000 participants, 3 cohorts	Shotgun/16S + clinical; LightGBM + SHAP	15 interpretable microbiome features; composite microbiome risk score	Significant, interpretable T2D risk association
Diabetes screening ML, 2024/25 [27]	692 participants (76 T1D, 366 T2D, 250 controls)	16S rRNA (citizen science); RF, SVM, elastic net	↑Veillonellaceae, Streptococcaceae, Gammaproteobacteria in T2D	AUC 0.76→0.77; sensitivity 0.46→0.80 (500 features)
Viome cohort, 2021 [28]	53,970 individuals (50,942 discovery; 3,028 validation)	Metatranscriptomics + ML	Differential <i>Blautia</i> , <i>Ruminococcus</i> expression patterns	Large-scale non-invasive risk assessment demonstrated
METSIM cohort, 2021 [29]	608 Finnish men (prospective)	16S/shotgun, longitudinal ML regression	Time-varying microbial predictors of HbA1c/insulin	Improved 1.5- and 4-year prediction vs. clinical-only model
Multi-cohort analysis, 2026 [10]	Multi-cohort, cross-population (Asian & European)	Shotgun; RF on 18 differential species	<i>Bilophila</i> , <i>B. ovatus</i> , <i>C. bartlettii</i> , <i>A. hadrus</i> , <i>H. parainfluenzae</i> top biomarkers	AUC 0.836→0.871 (+clinical); 0.815 (Asian) / 0.832 (European)

Table 2. Representative studies applying metagenomic profiling and machine learning to the prediction of type 2 diabetes and related metabolic phenotypes.

7. Proposed Integrative Framework for Early Prediction

In light of the analysis and mechanisms discussed above, this paper proposes a tiered, integrative screening approach that can be implemented at the population level while accounting for the generalizability constraints introduced in Section 6.

7.1 PPL Tier 1: Low Cost Population Triage

Pre-made gradient-boosting models, pre-calibrated with pre-validated anthropometric (BMI and WHR) and blood-based markers (fasting glucose, HbA1c), would be used to develop an initial risk score from 16S amplicon sequencing at the population level and add it to these routinely collected markers. This tier focuses on scalability and affordability – especially for use in low- and middle-income contexts – where the majority of the projected

future disease burden of diabetes is concentrated [1,2] and electronically controlled and automated 16S-based systems are at an early stage of development [18].

7.2 Tier 2: Shotgun Metagenomic Confirmation in At-Risk Individuals

Individuals flagged as elevated risk would undergo whole-genome shotgun metagenomic sequencing, enabling species- and strain-level resolution and direct functional pathway profiling. A random-forest or gradient-boosting model [10] plus SHAP interpretation [26] would yield a more streamlined, interpretable microbiome risk score that explicitly lists which taxa/pathways drive a higher-risk classification for the individual.

7.3 Analyzer Recalibration and Quality Assurance Procedures

The composition of the microbiota is not fixed and constantly changes based on weeks to months of treatment, medication or disease activity. It is then proposed that it should take periodic re-sampling, (e.g annually or after major dietary or pharmacological change) and recalibration of the model, as shown in the feedback loop of Fig. 4. There is scope for the microbiome-based risk score to be integrated with continuous glucose monitoring (CGM) or activity and sleep derived from wearables, thereby bringing closer to a real-time personalized risk score value derived over time or longitudinal risk score and not a single static score.

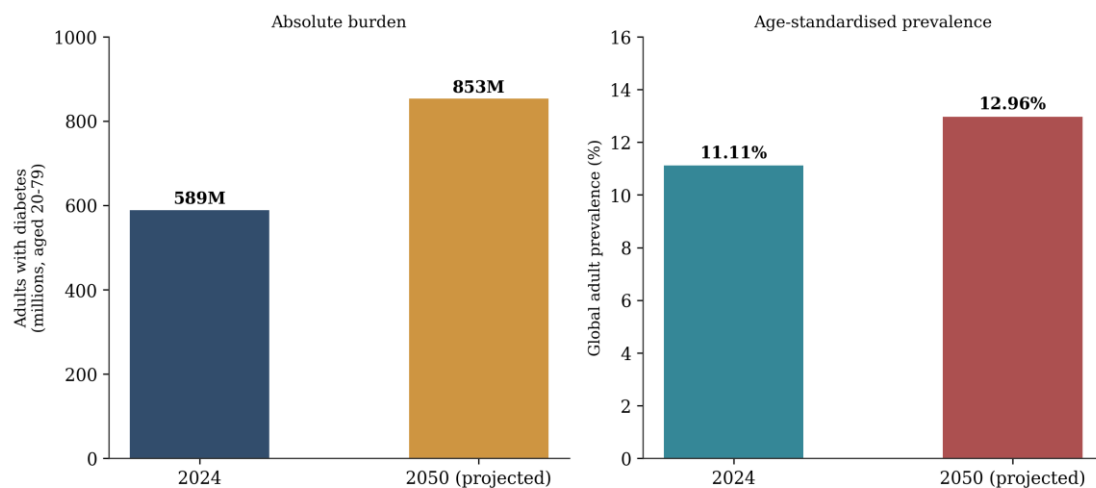
7.4 Cross-Population Generalizability and Federated Learning

Following this rationale, this framework suggests an architecture based on federated learning, where model parameters are shared and aggregated among different international consortia working on metagenomic studies, rather than individual-level metagenomic data. This strategy would enable use of ancestrally and geographically divergent populations for model training, while also helping overcome logistical barriers to using raw host sequence data and raw microbiome sequence data, which—like host genomic data—may raise re-identification risks.

7.5 Global Burden of Metabolic Disorders and Public Health Implications

The extent and direction of the global burden of metabolic diseases makes the public-health cases for pursuing non-invasive, microbiome-based early prediction of these diseases evident. Worldwide, from the 11.11% in 2024, the prevalence of diabetes in adults is predicted to rise to 12.96% in 2050; and the number of adults with diabetes will increase from 589 million to 853 million over the same timeframe [1,2,3] (Fig. 6). Stool sampling logistics are currently simpler, and less costly, than repeated glucose tolerance testing, especially in low/middle income countries where more than 4 in 5 people living with diabetes reside, where laboratory infrastructure to conduct repeated sampling may be limited. A low cost, validated Tier 1 triage tool (Section 7.1) relies on a 16S rPCR test and could theoretically expand early risk assays to previously less studied populations outside of traditional biochemical testing workflows, but this capability has not yet been proven with a health-economic analysis and needs to be demonstrated.

Figure 6. Global Diabetes Burden, 2024 and Projected 2050



Source: International Diabetes Federation, IDF Diabetes Atlas, 11th edition (2024/2025 estimates)

Figure 6. Global diabetes burden in 2024 and projected for 2050: absolute number of affected adults (left) and age-standardized adult prevalence (right). Source: International Diabetes Federation, IDF Diabetes Atlas, 11th edition [1,2,3].

Looking beyond diabetes per se, the same microbiome dysbiosis signatures are observed across the various metabolic diseases (including obesity, MetS and NAFLD), suggesting that it may be possible to develop a unifying (or partially unifying) screening tool around the microbiome rule out several overlapping conditions as a whole, thereby enhancing the cost-efficacy proposition of population level screening compared to disease-specific screening approaches.

9. Challenges and Limitations

9.1 Causality Versus Association

Most human evidence of the role of gut microbial composition in metabolic disease runs the risk of observational (cross-sectional or, at best, short-term prospective) and thus fails to provide an indication of the direction of causality. It's reasonable to argue that dysbiosis may be a cause, a consequence, or a bystander correlate of metabolic dysfunction, and that different taxa may be at different stages of the disease. There is more directly causal evidence to support specific pathways such as germ-free mouse colonization and faecal microbiota transplantation studies which show that an obesity associated or Akkermansia-depleted microbial community can be transferred to a recipient host to transfer the metabolic phenotype [4,5,13]; however, these designs are comparatively rare in human studies and inherently limited in causal claims that can be supported regarding complex chronic human disease.

9.2: Confounding and Cross-Population Heterogeneity

Gut microbial composition is significantly influenced by diet, geography, ethnicity, drug use and host genetics, in addition to being influenced by disease status. The cross-sectional differences seen in human patient cohorts may be confounded by the effects of treatment on the gut microbial population; first-line pharmacological intervention for T2D is metformin, which is known to alter the composition of the gut microbial population [28]. The model performance for out-of-sample validation data, showing higher attenuation in Asian populations relative to European populations, also reminds us that models trained and applied globally may be imperfect across all populations, depending on the model and population-specific factors [10].

9.3 Technical and Analytical Variability

The high level of technical variability arising from differences in DNA extraction protocol, sequencing platform and depth, bioinformatic pipeline (taxonomic classifier and reference database), and even sample storage and transport conditions, makes it difficult to extract the true biological signal from this variability. Comparative analyses have revealed that only part of this taxonomic profile can be achieved from 16S and shotgun sequencing of the same datasets; where both types of sequencing have been used in different studies on the same sample, it becomes difficult to match the results across studies [18,20,21].

9.4 Class Balance and Model Sensitivity:

Several of the studies reviewed reveal that apparently good AUC models have low sensitivity for minority classes, like type 1 diabetes in a mixed population of diabetes-type screening classes or early/prediabetes classes as opposed to known diabetes classes [27]. The apparently paradoxical relationship between overall discrimination measures and clinically useful sensitivity highlights the need to include complete sets of performance features, rather than only an AUC measure, when considering translation to the clinic.

9.5 Data Privacy, Standardization and Ethical Considerations

However, per host genomic data, gut microbiome sequence data also has a risk of re-identification, as well as issues related to data-governance when it comes to consent, data storage, and cross-border data sharing like in the federated-learning protocol outlined in Section 7.4. The field also lacks consensus on principles for reporting data similar to those in host epidemiology; if this changes and reporting guidelines are adopted along those lines, systematic meta-analysis of future studies will be more feasible.

10. Future Directions

The following research priorities emerge from this synthesis. Large-scale, ancestrally heterogeneous, longitudinal, multi-omic cohorts that integrate microbiome footprinting via shotgun metagenomic sequencing with metabolomics and host transcriptomics or proteomics are needed to move beyond single-omic snapshots to a more systems-level understanding of the microbiome-metabolism axis. Second, more rigorous evidence quantifying which associations represent true causal relationships that can be targeted and intervened upon is needed through causal-inference designs, such as randomised controlled faecal microbiota transplantation trials and Mendelian randomisation studies to understand the microbial component of the association. Third, privacy-preserving federated-learning architectures, as described in Section 7.4, warrant method development and potential validation across international consortia to directly address the cross-population generalizability problem observed in this study. Fourth, there is potential to dynamically integrate microbiome risk scores with continuous (wearable) physiological data (CGM, activity, sleep), which is key to a truly personalised, real-time metabolic risk landscape and up-to-date status classification rather than a single time point. Lastly, since Akkermansia muciniphila is depleted in all states of metabolic disease, next-generation live biotherapeutics utilizing it or similar, well-described taxa are already in the early phases of clinical development [11,16] and could, in principle, be used in future as a metagenomic screening tool to identify individuals most likely to benefit from targeted microbiome restoration approaches; validating its application as a screening tool would require dedicated clinical trial validation in advance of its clinical adoption.

11. CONCLUSION

With the easy access to small and cheap metagenomic sequencing, combined with accessible, usable and interpretable machine learning, the gut microbiome is now no longer a fringe research topic but rather a viable, potential source of metabolic risk information early in the disease course, ahead of the tissue or biochemical data obtained from a biopsy. In both cohorts and populations, depletion of Akkermansia muciniphila and depletion of annotated SCFA producers and an increase in the Firmicutes: Bacteroidetes ratio are consistently associated with

obesity and metabolic disturbances associated with type 2 diabetes, and the random forest and gradient-boosting models based on features of the gut metagenome can make clinically useful discriminative predictions (AUC 0.76 to 0.87), which can be further improved with simple clinical features. With the number of adults with diabetes projected at 853 million in 2050, and the greatest burden of diabetes, relative to absolute numbers, expected in low- and middle-income countries, the public-health potential of a low-cost, scalable screening pathway based on the microbiome is significant. To capitalise on this advance, however, causal-inference study designs, cross-population validation, federated-learning infrastructure, standardised reporting, and prospective clinical-utility trials are needed before metagenomic risk prediction can be safely implemented in routine care for metabolic prevention. In this paper, we outline a tiered, integrative approach that combines population triage using low-cost portable shotgun metagenomic profiling, interpretable machine learning for analyses, and a means for recalibration using time-series data as one concrete pathway toward that translational goal.

REFERENCES

- [1] International Diabetes Federation. IDF Diabetes Atlas, 11th edn. Brussels: International Diabetes Federation; 2025. Available at: <https://diabetesatlas.org/>
- [2] Genitsaridi I, Salpea P, Salim A, et al. 11th edition of the IDF Diabetes Atlas: global, regional, and national diabetes prevalence estimates for 2024 and projections for 2050. *Lancet Diabetes Endocrinol*. 2025.
- [3] International Diabetes Federation. The global picture of diabetes. In: IDF Diabetes Atlas, 11th edn [Internet]. Brussels: IDF; 2024. Available from NCBI Bookshelf: <https://www.ncbi.nlm.nih.gov/books/NBK618744/>
- [4] Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI. An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*. 2006;444(7122):1027-1031.
- [5] Ley RE, Turnbaugh PJ, Klein S, Gordon JI. Microbial ecology: human gut microbes associated with obesity. *Nature*. 2006;444(7122):1022-1023.
- [6] Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*. 2012;490(7418):55-60.
- [7] Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature*. 2012;486(7402):207-214.
- [8] Cani PD, Amar J, Iglesias MA, et al. Metabolic endotoxemia initiates obesity and insulin resistance. *Diabetes*. 2007;56(7):1761-1772.
- [9] The Gut Microbiota-Metabolic Axis: Emerging Insights from Human and Experimental Studies on Type 2 Diabetes Mellitus - A Narrative Review. PMC12654423. 2025.
- [10] Multi-cohort analysis of metagenome for type 2 diabetes identified universal gut microbiota signatures across populations. *Nutrition & Diabetes / npj Nature journal family*. 2026.
- [11] Akkermansia muciniphila, a New Generation of Beneficial Microbiota in Modulating Obesity: A Systematic Review. *Microorganisms*. 2021;9(5):1098.
- [12] Function of Akkermansia muciniphila in Obesity: Interactions With Lipid Metabolism, Immune Response and Gut Systems. PMC7046546.
- [13] Everard A, et al. Akkermansia muciniphila inversely correlates with the onset of inflammation, altered adipose tissue metabolism and metabolic disorders during obesity in mice. PMC4643218.
- [14] Akkermansia muciniphila and improved metabolic health during a dietary intervention in obesity: relationship with gut microbiome richness and ecology. *Front Nutr*. 2022;9:790045.
- [15] Akkermansia muciniphila: a potential candidate for ameliorating metabolic diseases. *Front Immunol*. 2024;15:1370658.
- [16] Akkermansia muciniphila Suppresses High-Fat Diet-Induced Obesity and Related Metabolic Disorders in Beagles. *Molecules*. 2022;27(18):6074.
- [17] Gut microbiome and obesity care: Bridging dietary, surgical, and microbiota-based interventions. *Cell Rep Med*. 2025.
- [18] Ranjan R, Rani A, Metwally A, McGee HS, Perkins DL. Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochem Biophys Res Commun*. 2016;469(4):967-977.
- [19] Comparative analysis of shotgun metagenomics and 16S rDNA sequencing of gut microbiota in migratory seagulls. PMC10629391. 2023.
- [20] Comparison between 16S rRNA and shotgun sequencing in colorectal cancer, advanced colorectal lesions, and healthy human gut microbiota. *BMC Genomics*. 2024.
- [21] Performance of a Shotgun Prediction Model for Colorectal Cancer When Using 16S rRNA Sequencing Data. PMC10816515. 2024.
- [22] 16S rRNA and metagenomic shotgun sequencing data revealed consistent patterns of gut microbiome signature in pediatric ulcerative colitis. PMC9018687. 2022.
- [23] Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol*. 2014;15(3):R46.
- [24] Truong DT, Franzosa EA, Tickle TL, et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat Methods*. 2015;12(10):902-903.
- [25] Karlsson FH, Tremaroli V, Nookaew I, et al. Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature*. 2013;498(7452):99-103.

- [26] Interpretable machine learning algorithm reveals novel gut microbiome features in predicting type 2 diabetes. *Genome Med / bioRxiv*. PMC7258537. 2020.
- [27] Using gut microbiome metagenomic hypervariable features for diabetes screening and typing through supervised machine learning. PMC11893737. 2024/2025.
- [28] Gut microbiome activity predicts risk of type 2 diabetes and metformin control in a large human cohort. *medRxiv*. 2021.
- [29] Machine Learning Reveals Time-Varying Microbial Predictors with Complex Effects on Glucose Regulation. PMC8573957 (METSIM cohort). 2021.