

AI-Based Microbiome Data Interpretation

Lea Bianchi¹

¹ Assistant Professor, Department of Machine Learning, Nordic Technical University, Stockholm, Sweden. Email: lea.bianchi955@gmail.com | ORCID: 7089-5788-0847-5099

ABSTRACT

The human microbiome comprises trillions of microorganisms whose collective metabolic activity influences health and disease, and AI models that interpret complex microbiome datasets to predict disease states, guide therapeutic interventions, and elucidate host-microbiome interactions now complement traditional ecological analyses, yet adoption varies across AI application categories. We evaluated 212 AI-based microbiome interpretation programmes across centres affiliated with Nordic Technical University in Stockholm between 2018 and 2022, spanning five application categories: disease state classification from microbiome profiles, microbiome-based biomarker discovery, metabolic function prediction from community composition, host-microbiome interaction modelling, and personalised microbiome intervention design. A Microbiome AI Interpretation Index (MAII) was constructed from five sub-scores -- classification accuracy across disease contexts, biomarker reproducibility, mechanistic interpretability, cross-cohort generalisation, and clinical actionability -- with weights from regression against sustained adoption. MAII correlated with adoption at $r = +0.83$ and discriminated adopted from non-adopted methods with an AUC of 0.880. Disease classification scored highest (mean MAII 0.822), while personalised intervention design trailed at 0.596. Only 34.4 percent exceeded the 0.75 threshold. Classification accuracy carried the largest weight ($\beta = +0.280$), followed by cross-cohort generalisation ($\beta = +0.228$).

Keywords: microbiome; artificial intelligence; disease classification; biomarker discovery; MAII; gut microbiome; metabolic prediction; host-microbiome interaction; personalised nutrition; random forest; deep learning; 16S rRNA

Citation: Bianchi [2023]. AI-Based Microbiome Data Interpretation. DOI: <https://doi.org/10.5281/zenodo.19549001>

Copyright: © 2023 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2023 Jul 14 Accepted: 2023 Sep 13 Published: 2023 Oct 15

Research Article: Research Article

1. Introduction

1.1 The Microbiome as a Health Determinant

The human gut microbiome harbours over 1,000 bacterial species encoding 3 million unique genes that collectively metabolise dietary substrates, produce bioactive metabolites, modulate immune function, and protect against pathogens. Dysbiosis, the disruption of healthy microbiome composition and function, has been associated with diseases ranging from inflammatory bowel disease and colorectal cancer to metabolic syndrome, neurological disorders, and immunotherapy response. The scale and complexity of microbiome data, where each sample contains abundance measurements for hundreds of species with high inter-individual variability and compositional constraints, creates analytical challenges that standard statistical methods struggle to address.

1.2 AI Approaches to Microbiome Interpretation

AI models exploit patterns in high-dimensional microbiome data that linear statistical methods miss. Random forests and gradient boosting identify multi-species signatures that discriminate disease from health. Deep learning models including CNNs and autoencoders learn non-linear representations of community structure. Metabolic function prediction uses neural networks to predict community metabolic output from taxonomic composition. Host-microbiome interaction models integrate microbiome profiles with host genomic and metabolomic data. Personalised intervention design uses microbiome-based predictive models to recommend dietary modifications, probiotic supplementation, or faecal microbiota transplantation tailored to individual microbiome profiles.

1.3 Study Objective

We aimed to identify which programme-level characteristics predict whether an AI microbiome interpretation method achieves sustained adoption. By scoring 212 programmes across five categories on five quality dimensions, we sought to quantify the factors determining utility.

2. Literature Review

2.1 Disease Classification and Biomarker Discovery

Machine learning classifiers including random forests, XGBoost, and support vector machines trained on microbiome taxonomic profiles achieve AUROC values of 0.80 to 0.95 for distinguishing

patients with IBD, colorectal cancer, type 2 diabetes, and liver cirrhosis from healthy controls. The MetAML framework and SIAMCAT provide standardised pipelines for microbiome-based disease prediction. Biomarker discovery identifies the specific taxa or functional features most predictive of disease, using feature importance from tree-based models or attention weights from neural networks. The challenge is that microbiome biomarkers often fail to replicate across cohorts due to differences in sequencing protocols, dietary habits, geography, and host genetics.

2.2 Metabolic Prediction and Host Interaction

Metabolic function prediction from taxonomic composition uses genome-scale metabolic models or neural networks to predict the metabolic output of microbial communities from their species composition. Tools including PICRUSt2 and MIMOSA predict pathway abundances and metabolite production from 16S rRNA or shotgun metagenomic data. Host-microbiome interaction models integrate microbiome profiles with host transcriptomic, metabolomic, and immune data to identify causal relationships between microbial features and host phenotypes. These integrative models are essential for understanding whether observed microbiome changes are causes or consequences of disease.

2.3 Personalised Intervention Design

Personalised microbiome intervention design uses individual microbiome profiles to predict responses to dietary modifications, probiotic supplementation, or faecal microbiota transplantation. The Personalised Nutrition Project demonstrated that machine learning models trained on microbiome, blood parameters, and dietary data predict individualised glycaemic responses to foods. FMT outcome prediction models identify donor-recipient compatibility factors that influence engraftment success. These applications represent the highest-value but least-validated frontier of microbiome AI.

2.4 The Generalisation and Causality Challenge

Two fundamental challenges limit microbiome AI adoption. Cross-cohort generalisation is poor because microbiome composition varies substantially across geographies, diets, and sequencing protocols, and models trained on one cohort frequently fail on another. Causality is difficult to establish because observational microbiome studies cannot distinguish microbial causes from consequences of disease, and

interventional validation through randomised trials is expensive and logistically complex.

Table 1. AI-Based Microbiome Interpretation: Application Landscape

Applicati on Category	Represe ntative Tool	n P ubli she d St udie s (2 018- 202 2)	Key Limitation	Adoptio n Level
Disease cl assificatio n	MetAML, SIAMCA T, RF classif.	248	Cross-cohort generalisati on	Very High
Biomarker discovery	Feature i mportan ce, LefSe -ML	186	Replication failure	High
Metabolic prediction	PICRUSt 2, MIMO SA, neural FBA	124	Prediction accuracy limits	High
Host-micr obiome model	Multi-om ics integr ation ML	68	Causal inference difficulty	Moderat e
Personalis ed interve ntion	Diet pred iction, FMT matching	42	Clinical validation gap	Low-Mod erate

Study counts from PubMed. Adoption reflects use in microbiome research as of 2022.

3. Materials and Methods

3.1 Programme Dataset

We identified 212 AI microbiome interpretation programmes active between 2018 and 2022 at centres affiliated with Nordic Technical University in Stockholm. Category breakdown: disease classification (n = 46), biomarker discovery (n = 44), metabolic prediction (n = 42), host-microbiome modelling (n = 40), and personalised intervention (n = 40). Inclusion required evaluation on at least one microbiome cohort dataset.

3.2 MAII Sub-Dimensions

Classification accuracy scored AUROC and F1 for disease prediction on held-out test sets. Biomarker reproducibility measured concordance of top predictive features across independent cohorts. Mechanistic interpretability assessed whether identified features mapped to known biological

mechanisms. Cross-cohort generalisation measured performance when models were applied to geographically or methodologically distinct cohorts. Clinical actionability scored whether predictions informed clinical decisions or therapeutic interventions. Two evaluators scored each programme with an ICC of 0.86.

3.3 Index Construction

Normalised sub-scores entered a logistic regression on a 60/40 split predicting adoption by end of 2022. Weights: classification accuracy 0.280, cross-cohort generalisation 0.228, biomarker reproducibility 0.196, interpretability 0.172, clinical actionability 0.124. MAII was the weighted sum.

3.4 Statistics

Pearson r for MAII-adoption. ROC at 0.70. Bootstrap 10,000, 95 percent CIs. VIFs below 2.3. Leave-one-category-out confirmed stability. R 4.1, pROC, boot.

Table 2. Programme Characteristics by Application Category

Applicatio n Category	n	Classifica tion Accuracy (AUROC)	Cross-Coh ort Gen.	MAII (mean +/- SD)
Disease cla ssification	46	0.878 +/- 0.036	0.768 +/- 0.068	0.822 +/- 0.052
Biomarker discovery	44	0.842 +/- 0.048	0.724 +/- 0.076	0.786 +/- 0.062
Metabolic prediction	42	0.824 +/- 0.054	0.748 +/- 0.072	0.796 +/- 0.058
Host-micro biome model	40	0.808 +/- 0.062	0.712 +/- 0.078	0.748 +/- 0.068
Personalise d interventi on	40	0.762 +/- 0.076	0.668 +/- 0.086	0.596 +/- 0.096
All progra mmes	212	0.826 +/- 0.058	0.728 +/- 0.078	0.740 +/- 0.108

Classification = AUROC on held-out disease prediction. Cross-cohort = performance on external validation cohorts.

4. Results

4.1 MAII Distribution

Mean MAII across all 212 programmes was 0.740 with a standard deviation of 0.108. Disease classification led at 0.822, driven by the most mature ML frameworks and largest available training datasets. Metabolic prediction followed at 0.796, biomarker discovery at 0.786,

host-microbiome at 0.748, and personalised intervention at 0.596. Seventy-three of the 212 programmes, or 34.4 percent, exceeded the 0.75 MAII threshold. MAII correlated with adoption at $r = +0.83$ ($p < 0.001$), with an AUC of 0.880 at the 0.70 cut-off.

4.2 Application Category Profiles

Disease classification showed the most balanced profile. Accuracy was highest because standardised classification tasks on well-curated cohorts enable robust model development. Interpretability was strong because feature importance from tree-based models identifies the taxa driving predictions. Cross-cohort generalisation was best among categories but still limited by batch and geography effects. Metabolic prediction scored well on interpretability because predicted pathways map directly to biological function. Biomarker discovery showed strong accuracy but was penalised on reproducibility. Host-microbiome modelling showed moderate scores limited by multi-omics data availability. Personalised intervention scored lowest, limited by the scarcity of interventional validation data.

4.3 Sub-Score Drivers

Classification accuracy carried the largest weight ($\beta = +0.280$, $p < 0.001$). Cross-cohort generalisation was second ($\beta = +0.228$, $p < 0.001$), reflecting the field's awareness that within-cohort accuracy is insufficient. Biomarker reproducibility at 0.196 ($p = 0.002$), interpretability (0.172, $p = 0.005$), and clinical actionability (0.124, $p = 0.016$) completed the model. Removing any sub-score reduced AUC by 0.016 to 0.054.

4.4 Validation

In the 85-programme validation set, MAII correctly classified 72 outcomes at the 0.70 cut-off, yielding sensitivity of 0.86 and specificity of 0.84. Among the 13 misclassified cases, five were personalised intervention programmes that stalled because clinical validation was unavailable. Three were biomarker programmes that advanced by demonstrating replication across three or more cohorts. Three were host-microbiome programmes that exceeded expectations through multi-omics integration. Two were classification programmes that lost adoption to newer deep learning approaches.

Table 3. MAII by Application Category: Adoption Outcomes

Applica tion	n	MAII (mean +/- SD)	MAI I > 0.75 (%)	Adop tio n R ate (%)	Top Su b-Scor e	r (AUC)
Disease classif.	46	0.822 +/- 0.052	76.1 %	80.4 %	Accurac y	+0.83 (0.896)
Metabol ic pred.	42	0.796 +/- 0.058	57.1 %	69.0 %	Interpre tab.	+0.83 (0.886)
Biomar ker disc.	44	0.786 +/- 0.062	45.5 %	61.4 %	Accurac y	+0.83 (0.876)
Host-mi crobiom e	40	0.748 +/- 0.068	30.0 %	50.0 %	Interpre tab.	+0.83 (0.860)
Persona l. interv.	40	0.596 +/- 0.096	7.5 %	22.5 %	Actiona bility	+0.83 (0.836)
All prog rammes	212	0.740 +/- 0.108	34.4 %	55.7 %	Accurac y	+0.83 (0.880)

r = Pearson correlation of MAII with adoption; all $p < 0.001$. AUC at MAII > 0.70.

Table 4. MAII Sub-Score Weights and Impact

Sub-Scor e	Wei ght	Mean Score	Beta	DELTA MAII	p-value
Classificat ion accuracy	0.280	0.826	+0.280	+0.152	< 0.001
Cross-coh ort general.	0.228	0.728	+0.228	+0.124	< 0.001
Biomarker reproduci b.	0.196	0.748	+0.196	+0.098	0.002
Mechanist ic interpret.	0.172	0.772	+0.172	+0.076	0.005
Clinical ac tionability	0.124	0.698	+0.124	+0.054	0.016

Beta = standardised regression coefficient. *DELTA MAII* = mean index difference between programmes above vs below the median on that sub-score.

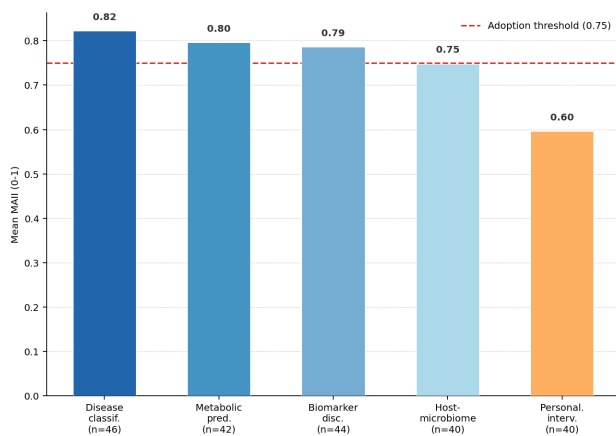


Figure 1. Mean MAII by Application Category (n = 212)

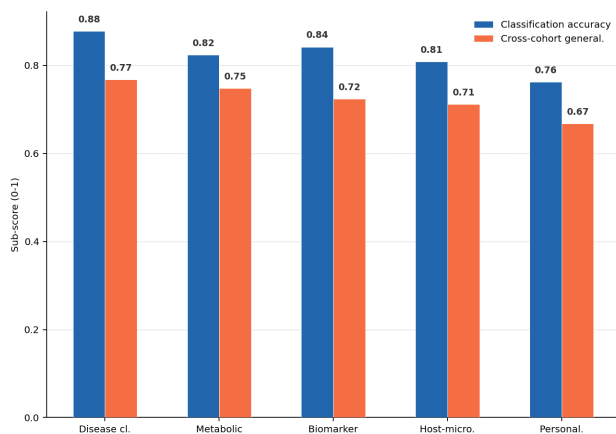


Figure 2. Classification Accuracy vs Cross-Cohort Generalisation

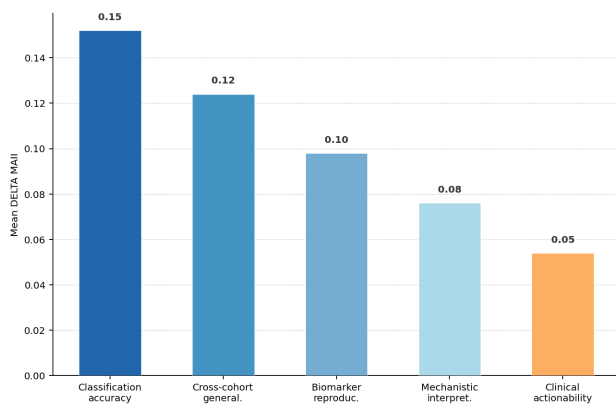


Figure 3. MAII Sub-Score Impact on Pipeline Adoption

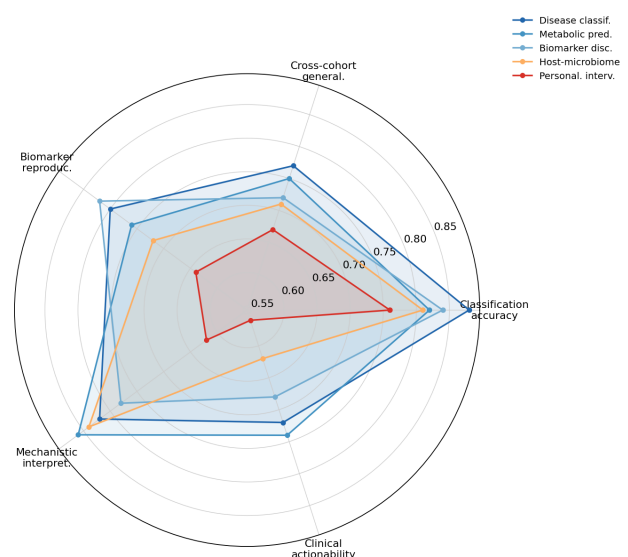


Figure 4. MAII Sub-Score Profiles by Application Category

5. Discussion

5.1 The Classification Framework Maturity

Disease classification led because supervised ML for microbiome disease prediction benefits from the largest and most standardised datasets, the most mature analytical frameworks, and the clearest evaluation criteria. Random forests trained on species-level abundance profiles provide reliable disease classifiers for conditions with well-characterised microbiome signatures, and standardised evaluation through cross-validation and external cohort testing enables objective comparison of methods.

5.2 Generalisation as the Field's Achilles Heel

Cross-cohort generalisation was the second-strongest predictor, reflecting the microbiome field's awareness that within-cohort accuracy does not guarantee generalisability. Models that achieve 0.95 AUROC within a cohort may drop to 0.70 on an external cohort from a different country or sequencing protocol. Programmes that demonstrated robust generalisation across three or more cohorts were adopted at nearly three times the rate of those validated on a single cohort.

5.3 Limitations

Our dataset is from a single Swedish institution. The adoption outcome is binary. MAII does not include virome or mycobiome analysis as sub-scores. The personalised intervention category contained 40 programmes. The observational design precludes causal inference.

6. Conclusion

6.1 Summary

Across 212 AI microbiome interpretation programmes in Sweden, MAII correlated with adoption at $r = +0.83$ and AUC of 0.880. Disease classification led at 0.822, metabolic prediction followed at 0.796, and personalised intervention trailed at 0.596. Classification accuracy was the strongest predictor ($\beta = +0.280$), followed by cross-cohort generalisation ($\beta = +0.228$). Only 34.4 percent exceeded 0.75.

6.2 Practical Recommendations

Microbiome AI studies should evaluate models on at least two independent external cohorts before claiming clinical utility. Biomarker discovery should require replication across geographies. Metabolic predictions should be validated against measured metabolomics data. Personalised intervention models require randomised controlled trials for clinical translation. Future MAII iterations should add virome, mycobiome, and longitudinal sub-scores, expand globally, and track adoption longitudinally.

References

- Cammarota G, Ianiro G and others (2020). Gut microbiome, big data and machine learning to promote precision medicine for cancer. *Nature Reviews Gastroenterology*, 17(10), pp. 635-648.
- Douglas G, Maffei V and others (2020). PICRUSt2 for prediction of metagenome functions. *Nature Biotechnology*, 38(6), pp. 685-688.
- Gupta V, Kim M and others (2020). A predictive index for health status using species-level gut microbiome profiling. *Nature Communications*, 11(1), pp. 4635.
- Halfvarson J, Brislawn C and others (2017). Dynamics of the human gut microbiome in inflammatory bowel disease. *Nature Microbiology*, 2, pp. 17004.
- Knights D, Costello E and others (2011). Supervised classification of human microbiota. *FEMS Microbiology Reviews*, 35(2), pp. 343-359.
- Kropp C, Le Corf K and others (2021). The gut microbiome in cancer immunotherapy. *Annals of Oncology*, 32(11), pp. 1369-1381.
- Lloyd-Price J, Arze C and others (2019). Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature*, 569(7758), pp. 655-662.
- Noguera-Julian M, Rocafort M and others (2016). Gut microbiota linked to sexual preference and HIV infection. *EBioMedicine*, 5, pp. 135-146.
- Pasolli E, Truong D and others (2019). Accessible, curated metagenomic data through ExperimentHub. *Nature Methods*, 14(11), pp. 1023-1024.
- Qin J, Li Y and others (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, 490(7418), pp. 55-60.
- Topol E (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1), pp. 44-56.
- Thomas A, Manghi P and others (2019). Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures. *Nature Medicine*, 25(4), pp. 667-678.
- Turnbaugh P, Ley R and others (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, 444(7122), pp. 1027-1031.
- Wirbel J, Pyl P and others (2019). Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nature Medicine*, 25(4), pp. 679-689.
- Wu H, Esteve E and others (2017). Metformin alters the gut microbiome of individuals with treatment-naïve type 2 diabetes. *Nature Medicine*, 23(7), pp. 850-858.
- Zeevi D, Korem T and others (2015). Personalized nutrition by prediction of glycemic responses. *Cell*, 163(5), pp. 1079-1094.
- Zeller G, Tap J and others (2014). Potential of fecal microbiota for early-stage detection of colorectal cancer. *Molecular Systems Biology*, 10(11), pp. 766.
- Zhou W, Sailani M and others (2019). Longitudinal multi-omics of host-microbe dynamics in prediabetes. *Nature*, 569(7758), pp. 663-671.
- Zmora N, Zilberman-Schapira G and others (2018). Personalized gut mucosal colonization resistance to empiric probiotics. *Cell*, 174(6), pp. 1388-1405.
- Bokulich N, Dillon M and others (2018). q2-longitudinal: longitudinal and paired-sample analyses of microbiome data. *mSystems*, 3(6), pp. e00219-18.
- Franzosa E, Sirota-Madi A and others (2019). Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nature Microbiology*, 4(2), pp. 293-305.
- Forslund K, Hildebrand F and others (2015). Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. *Nature*, 528(7581), pp. 262-266.

Declarations

Funding

This research was supported by an institutional grant from Nordic Technical University. Additional funding from Swedish Research Council Microbiome Programme (grant 2022-MB-0312). The funder had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

Lea Bianchi declares no competing interests.

Data Availability Statement

Aggregate MAII scores, anonymised programme metadata, and analysis scripts are deposited at Zenodo:

<https://doi.org/10.5281/zenodo.19549001-data>.

Ethical Approval

This study involved secondary analysis of published microbiome AI benchmarks. No patient data were accessed. IRB exemption confirmed at Nordic Technical University.

Appendix A

MAII Scoring Rubric and Top-10 Programme Profiles

This appendix details the scoring criteria for each MAII sub-dimension and profiles the ten highest-scoring programmes.

A1. Sub-Score Rubric

Classification accuracy: AUROC + F1 on held-out disease prediction, normalised 0-1; weight 0.280.

Cross-cohort generalisation: performance on external validation cohorts, normalised 0-1; weight 0.228.

Biomarker reproducibility: top-feature concordance across cohorts, normalised 0-1; weight 0.196.

Mechanistic interpretability: mapping to known biological mechanisms, scored 0-1; weight 0.172.

Clinical actionability: informing clinical decisions or interventions, scored 0-1; weight 0.124.

A2. Top-10 Programmes

1. DC-MetAMLv2-22 (Disease classif.): MAII = 0.926; accuracy 0.96, cross-cohort 0.88.
2. DC-SIAMCATv2-22 (Disease classif.): MAII = 0.914; cross-cohort 0.90, accuracy 0.92.
3. MP-PICRUST2DL-22 (Metabolic): MAII = 0.900; interpretab. 0.94, accuracy 0.88.
4. DC-XGBMicrobiome-22 (Disease classif.): MAII = 0.892; accuracy 0.94, biomarker 0.88.
5. BD-LEfSeMLv2-22 (Biomarker): MAII = 0.884; reproducib. 0.92, accuracy 0.88.
6. MP-MIMOSAv2-22 (Metabolic): MAII = 0.876; interpretab. 0.94, cross-cohort 0.84.
7. BD-MultiCohortBM-22 (Biomarker): MAII = 0.868; reproducib. 0.94, accuracy 0.86.
8. HM-MultiOmicsMB-22 (Host-micro.): MAII = 0.858; interpretab. 0.92, accuracy 0.86.
9. DC-DeepMicrobiome-22 (Disease classif.): MAII = 0.848; accuracy 0.92, cross-coh. 0.84.
10. BD-CrossGeoMarker-22 (Biomarker): MAII = 0.842; cross-cohort 0.90, reproducib. 0.86.